DESTATIS
Statistisches Bundesamt

M. Beck, F. Dumpert (Universität Bayreuth), J. Feuerhake

Proof of Concept Machine Learning

Abschlussbericht

Wiesbaden, 31. Juli 2018
Statistisches Bundesamt

Herausgeber: Statistisches Bundesamt (Destatis)

Internet: www.destatis.de

Ihr Kontakt zu uns:

www.destatis.de/kontakt

Zentraler Auskunftsdienst

Tel.: +49 (0) 611 / 75 24 05

Verfasser: Martin Beck (E1)
Florian Dumpert (Universität Bayreuth)
Joerg Feuerhake (E105)

Version 1.0

Erschienen im Juli 2018

Fotorechte:

Deckblatt © Statistisches Bundesamt (Destatis)

Inhalt

Managementfassung

1. In der Leitungsklausur vom 13.–15. November 2017 haben die Amtsleitung und die Abteilungsleitungen 59 Maßnahmen der Digitalen Agenda nach Nutzen und Aufwand bewertet und priorisiert. Die Durchführung eines „Proof of Concept Machine Learning“ wurde hoch priorisiert und als eines von vier Leuchtturmprojekten der Digitalen Agenda eingestuft. Der Termin für den Projektabschluss wurde auf den Juni 2018 festgelegt, die Verantwortlichkeit E1 übertragen.
2. Die Themen „Machine Learning“ und noch mehr „Künstliche Intelligenz“ sind sehr umfassend und kaum noch zu überschauen. Das Projektteam verständigte sich darauf, sich ausschließlich auf das maschinelle Lernen zu konzentrieren und dessen Anwendbarkeit in allen Fachstatistiken in den Fokus zunehmen. Künstliche Intelligenz und Big Data wurden aus der Analyse (weitgehend) ausgeklammert.
3. Nach Abschluss des Proof of Concept sollte eine Übersicht der potenziellen Anwendungen in den Fachstatistiken vorliegen, die über eine Hausumfrage gewonnen werden sollte. Um hierfür eine Informationsbasis zu schaffen wurden folgende vorbereitende Schritte durchgeführt:
   - Einrichtung einer Informations- und Austauschplattform für alle an Machine Learning interessierten Mitarbeiterinnen und Mitarbeiter im Statistischen Bundesamt.
   - Bestandsaufnahme der Anwendung von Machine-Learning-Verfahren im Statistischen Bundesamt sowie in nationalen und internationalen Statistikinstitutionen.
   - Durchführung einer Kurzveranstaltung zum Thema „Machine Learning“.
4. Das sogenannte statistische maschinelle Lernen zeichnet sich dadurch aus, dass auf Grundlage endlich vieler Beobachtungen ein Zusammenhang zwischen Eingabevariablen und Ausgabevariable erlernt wird, der anschließend auf neue, ggf. noch nicht bekannte Eingabewerte angewendet werden kann, um den unbekannten Ausgabewert zu schätzen.
5. Hinsichtlich des grundsätzlichen Vorgehens unterscheidet man zwischen überwachtem (supervised learning) und unüberwachtem (unsupervised learning) maschinellen Lernen. Das überwachte Lernen wird für Klassifikation, d. h. die Zuordnung eines Objektes zu einer von endlich vielen (häufig genau zwei) Klassen, und für die Regression eingesetzt und ist im Falle des maschinellen Lernens i.d.R. nichtparametrisch. Benötigt werden geeignete Trainings- und Testdaten, für die das wahre Ergebnis der Klassifikation (bzw. der Regression) bekannt ist. Unüberwachtes Lernen wird z. B. für Ausreißeridentifikation oder Clustering verwendet.
6. Gängige, auch in der Praxis statistischer Institutionen verbreitete Methoden des maschinellen Lernens sind Support Vector Machines, Random Forests und Neuronale Netze.
7. Machine Learning hat in den letzten Jahren eine explosionsartige Entwicklung genommen. Dies gilt sowohl für die Forschung als auch die mediale Wahrnehmung. Die Zahl der

wissenschaftlichen Veröffentlichungen ist in den 2010er Jahren exponentiell gewachsen. Die Forschungsergebnisse fließen in konkrete Anwendungen ein, die zunehmend im Alltag Bedeutung erlangen. Dementsprechend verstärkt sich auch die Berichterstattung in den Medien. Spätestens 2018 ist das Thema maschinelles Lernen/Künstliche Intelligenz auch in der deutschen Politik angekommen. Die entsprechenden Vereinbarungen im Koalitionsvertrag werden von der Bundesregierung nun in politisches Handeln umgesetzt. Beispielhaft dafür stehen die Einsetzung einer Enquetekommission am 28. Juni 2018 und das am 18. Juli 2018 vom Bundeskabinett verabschiedete Eckpunktepapier zur Künstlichen Intelligenz. Bis Dezember 2018 soll die „Strategie Künstliche Intelligenz“ vorgelegt werden.

8. Im Rahmen des Proof of Concept Machine Learning wurden die 14 Statistischen Landesämter, weitere 18 nationale Statistikorganisationen (überwiegend ONAs) sowie 39 internationale Statistikinstitutionen hinsichtlich des Einsatzes von maschinellen Lernverfahren befragt. Außerdem wurden vorliegende Dokumentationen ausgewertet und Recherchen auf Webseiten durchgeführt. Mit Ausnahme von vier internationalen Statistikinstitutionen (Bulgarien, Frankreich, Griechenland, Malta) liegen für alle Institutionen Informationen vor.
9. In den Statistischen Landesämtern spielt Machine Learning noch keine Rolle. Lediglich das Hessische Statistische Landesamt meldete eine Anwendung. Das Statistische Bundesamt kann also nicht von den Erfahrungen in den Statistischen Landesämtern profitieren.
10. Auf die Abfrage bei den nationalen Institutionen gab es positive Rückmeldungen von dem Bundesamt für Migration und Flüchtlinge (BAMF), der Deutschen Bundesbank, dem GESIS – Leibniz-Institut für Sozialwissenschaften, dem Institut für Arbeitsmarkt- und Berufsforschung (IAB), dem Robert-Koch-Institut (RKI) und dem Zentrum für europäische Wirtschaftsforschung (ZEW). Die Antwortenden meldeten insgesamt 36 Projekte, in denen Maschine-Learning-Verfahren eingesetzt werden. Fünf dieser Anwendungen sind im Produktivbetrieb. Bei weiteren 21 Projekten handelt es sich um laufende Forschungsprojekte im engeren Sinne. Zehn weitere Projekte sind Machbarkeitsstudien und Entwicklungen von Prototypen, die Potenzial für den Produktivbetrieb haben. In den Projekten wurden Machine-Learning-Verfahren am häufigsten zur Identifikation und Klassifikation von Einheiten eingesetzt. Als Methoden werden neben Random Forests zumeist andere entscheidungsbaumbasierte Verfahren eingesetzt. Auch Neuronale Netze und Support Vector Machines werden häufig verwendet.
11. Die internationalen Statistikinstitutionen setzten in insgesamt 119 Projekten Machine Learning ein. Statistics Canada meldete bei weitem die meisten Projekte dieser Art, gefolgt vom Australian Bureau of Statistics, Stats NZ, dem Bundesamt für Statistik der Schweiz sowie Institutionen in den Vereinigten Staaten. Die Maßnahmen sind überwiegend

statistikübergreifend, aber auch die Haushalts- und Unternehmensstatistiken werden häufig als Einsatzbereiche genannt. Machine-Learning-Verfahren werden dabei häufig zur Klassifikation, Identifikation und Imputation verwendet. Unter den am häufigsten genannten Machine-Learning-Methoden finden sich Random Forests, Verfahren, die Neuronale Netze nutzen, Support Vector Machines sowie weitere entscheidungsbaumbasierte Verfahren. 18 der genannten Projekte sind bereits im Produktiveinsatz, weitere 25 sind in der Entwicklung hin zum Produktiveinsatz. 51 Projekte sind im Experimentierstadium und weitere 25 sind zurzeit als Idee formuliert.

12. Die genannten Projekte lassen sich den Phasen des GSBPM zuordnen. Machine-Learning-Verfahren werden vor allem bei der Datengewinnung, Datenaufbereitung und der Ergebnisanalyse eingesetzt. Weiterhin wurden Projekte in der Statistikkonzeption, der Organisation des Nutzerservice und in der Evaluation genannt.
13. Fazit: Aus den Abfragen bei den Statistischen Ämtern der Länder, den nationalen und internationalen statistikproduzierenden Institutionen lassen sich folgende Schlüsse ziehen: Projekte, die maschinell Merkmale klassifizieren, Werte imputieren und Einheiten identifizieren und dazu Random Forests oder ähnliche baumbasierte Verfahren, Support Vector Machines oder Neuronale Netze einsetzen, sind zurzeit weit verbreitet. Anwendungsfälle können in fast allen Fachstatistiken gefunden werden. In der Regel werden Prozesse in der Statistikkonzeption, der Datengewinnung und -analyse sowie der Statistikverbreitung und -evaluation mit Machine-Learning-Verfahren unterstützt.
14. Um mögliche Einsatzgebiete für Machine-Learning-Verfahren in den Fachstatistiken zu identifizieren, wurde bei den Gruppen des Statistischen Bundesamtes eine Hausabfrage durchgeführt. Für die Abfrage wurden am 9. Mai 2018 alle 29 Gruppen angeschrieben, die auch geantwortet haben. Amtsleitung, Leitungsstab, Abteilungs- und Referatsleitungen erhielten die entsprechende Mail in Kopie, so dass sie sich ggf. in die Beantwortung einbringen konnten.
15. Aus den abgefragten Gruppen wurden 16 Fehlanzeigen gemeldet, dies in der Regel mit dem Hinweis, dass derzeit keine geeigneten Anwendungsmöglichkeiten ersichtlich sind. Fünfmal wurde auch fehlende Expertise bzw. fehlende Zeit zum Aufbau von Expertise bzw. zur Umsetzung von Projekten als Grund genannt. Informationsdefizite, scheinen nicht zu bestehen. Insgesamt wird auch bei Gruppen, die keine Projekte oder Projektideen meldeten, grundsätzlich Potenzial für entsprechende Anwendungen gesehen.
16. Die verbleibenden 13 Gruppen meldeten 31 Projektideen. Bei 25 der Meldungen handelt es sich um Ideen, wie Machine-Learning-Verfahren eingesetzt werden könnten. Sechs Projekte sind bereits in der Experimentier- bzw. Testphase. Da viele Vorschläge erste Ideen sind, kann noch keine abschließende Aussage über verwendete Verfahren gemacht werden. Fast alle

Gruppen, die Projektideen meldeten, gaben auch an, dass sie (gruppen-)externe Expertise bei der eventuellen Umsetzung benötigen werden. Die Projektideen zielen oft auf die maschinelle Klassifikation von Merkmalen oder die Identifikation von Einheiten (Dubletten, Ausreißer). Dies sind auch die Anwendungen, die bei den Rückmeldungen aus den nationalen und internationalen statistikproduzierenden Institutionen sehr häufig genannt wurden.

17. Zusammenfassend kann festgehalten werden, dass es viele vielversprechende Ansätze und Ideen für den Einsatz von Machine Learning im Statistischen Bundesamt gibt. Im Moment scheint es aber einen Engpass beim Aufbau bzw. der Bereitstellung von Expertise in diesem Feld zu geben.
18. Neben den mit der Hausumfrage eruierten geplanten Maßnahmen gibt es mehrere Projekte, die bereits umgesetzt sind oder derzeit durchgeführt werden. Es handelt sich um fünf Projekte in E1 und eines in D306. Die Maßnahmen von E1 sind in der folgenden Übersicht zusammengefasst:

Statistik	Problem	Methode	Stand	Ergebnis
Unternehmens-register	Zuordnung von Unternehmen zum 3. Sektor	Support Vector Machine	abge-schlossen	+
Handwerkszählung	Erkennung nicht relevanter Unternehmen in den Handwerksstatistiken	Random Forest Support Vector Machine	abge-schlossen	++
Verdienst-strukturerhebung	Schätzung einer Erwerbsunterbrechung von Frauen in der Verdienststrukturerhebung	Support Vector Machine	abge-schlossen	+/–
Verdienst-strukturerhebung	Schätzung der Staatsbürgerschaft von Beschäftigten in der Verdienststrukturerhebung	Support Vector Machine)	abge-schlossen	(–)
Verdienst-strukturerhebung	Anreicherung der Integrierten Erwerbsbiografien (IEB) von BA/IAB um Informationen zur Mindestlohnbetroffenheit aus der Verdienststrukturerhebung	Random Forest	laufend	+/–

Im Referat D306 wird aktuell anhand von Marktforschungsdaten die Nutzung von Scannerdaten zur Berechnung von Preisindizes getestet.

19. Für die Realisierung von Projekten, die Machine-Learning-Verfahren verwenden, muss eine geeignete Infrastruktur vorliegen. Bezüglich der notwendigen Software führt kein Weg an R und Python vorbei. Hinsichtlich der erforderlichen Hardware lässt sich aufgrund der

bisherigen Erfahrungen mit Machine-Learning-Projekten in E1 eindeutig sagen, dass die Standard-PCs nicht ausreichen. Ziel sollte es sein, für verschiedene Machine-Learning-Verfahren zeitlich begrenzt ausreichend dimensionierte virtuelle Rechenkraft bereitzustellen. Hierzu müsste ein geeignetes Konzept entwickelt und zeitnah umgesetzt werden.

20. Das amtsinterne Fortbildungsprogramm sollte bedarfsgerecht fortgeführt und ggf. ausgebaut werden (z. B. ergänzend in Richtung Python).
21. Um die in den Fachgruppen geplanten Vorhaben zum Einsatz von Machine-Learning-Verfahren umsetzen zu können, bedarf es nicht nur der geschilderten Infrastruktur, sondern auch einer kompetenten, kontinuierlichen und strukturierten Beratung in Methodenfragen. Ab dem 1. August 2018 ist eine entsprechende Umorganisation in der Gruppe C1 „Mathematisch-statistische Methoden, Forschungsdatenzentrum“ vorgesehen. Die Zuständigkeit für „Machine Learning“ wird dem noch aufzubauenden neuen Referat „C103 Maschinelles Lernen und Imputationsverfahren“ übertragen.

22. Aus den Erkenntnissen des Proof of Concept Machine Learning werden folgende zehn Handlungsempfehlungen abgeleitet (Kurzbeschreibung):

Handlungsempfehlung	Zuständigkeit	Termin
E1: Einrichtung eines Kompetenzzentrums in der Gruppe C1 „Mathematisch-statistische Methoden, Forschungsdatenzentrum“ *(neues Referat C103)*	Abt. A ; C1	01.08.2018
E2: Sicherung bzw. Schaffung der notwendigen Infrastruktur für Machine-Learning-Projekte *(Hardware; Software, wie R und Python; Informations- und Austauschplattform)*	C2; C3; C103; ITZ-Bund	31.12.2018
E3: Schulungen zu „Machine Learning“ *(Durchführung; bedarfsgerechte Weiterentwicklung)*	A203	fortlaufend
E4 bis E7 beinhalten Vorschläge, die die Information verschiedener Interessengruppen über die Durchführung und die Ergebnisse des Proof of Concept Machine Learning betreffen		
E4: Informationen für Führungskräfte *(in ALB und Gruppen-/Referatsleitungsforum)*	L E1; C103	17.09.2019 21.09.2018
E5: Allgemeine Information aller Mitarbeiterinnen und Mitarbeiter des Statistischen Bundesamtes *(Intranetmeldung und Veröffentlichung Abschlussbericht)*	LS; B306; L E1	Ende 09/2018 sowie November 2018
E6: Information der Politik und weiterer Stakeholder *(Exzellenz (Digital) Show für Ressorts; Statistische Woche; Kolloquium; Aufsätze)*	i-Punkt; L E1; C103; B305	bis Ende 2018
E7: Thematisierung in Statistikgremien und Einbindung der Statistischen Landesämter *(in den Sitzungen des ALG FS und der ALK)*	Amtsleitung; LS; L E1; C103	offen
E8: Zusammenarbeit mit Hochschulen, Forschungs- und Statistikinstitutionen *(Forschungskooperation; Personalgewinnung)*	C103; A201	fortlaufend
E9: Generierung neuer Projektideen *(gruppeninterne Workshops)*	Fachgruppen; ggf. unterstützend E1	bis März 2019
E10: Behandlung im Jahresarbeitsplanungsgespräch 2019 *(Thematisierung der Machine-Learning-Projekte)*	Amtsleitung; A102- Controlling; Alle Abteilungen	Anfang 2019

23. Fazit: Der Proof of Concept Machine Learning bestätigte, dass es in den Fachstatistiken Potenzial für die Anwendung von maschinellem Lernen gibt. Neben den in E1 bereits abgeschlossenen Projekten, die als Piloten angesehen werden können, wurden seitens der Fachgruppen 31 Ideen für Machine-Learning-Anwendungen genannt. Bei sechs der genannten Ideen werden bereits Tests durchgeführt. 25 weitere müssen nun auf

Umsetzbarkeit geprüft werden. Machine-Learning ist jedoch kein Allheilmittel und nicht jede Fachaufgabe kann damit erfolgreich gelöst werden. Dies lässt sich jedoch im Vorhinein nur schwer beurteilen. Ob eine Maßnahme erfolgversprechend ist, lässt sich im Allgemeinen nur feststellen, wenn man sie durchführt. Zum Einsatz von Machine Learning in den Fachstatistiken sollte ermutigt werden, ohne das erwartet wird, dass alle Projekte erfolgreich sein werden. Führungskräfte und Beschäftigte müssen offen sein für Veränderungen, die mit dem Einsatz von Machine Learning einhergehen. Sie müssen im Sinne einer Fehlerkultur einkalkulieren und akzeptieren, dass mit Innovationen und Veränderungen auch Misserfolge einhergehen können.

24. Das Projektteam sieht mit der Vorlage dieses Abschlussberichts den Proof of Concept Machine Learning und somit auch den Auftrag aus der Leitungsklausur als erfolgreich abgeschlossen an.
25. Der Abschlussbericht über den „Proof of Concept Machine Learning“ wird am 17. September 2018 in der ALB vorgestellt.

1 Auftrag

1.1 Digitalisierungsworkshop und Leitungsklausur

Am 10. Oktober 2017 erfolgte in einem Digitalisierungsworkshop der Startschuss für die Erarbeitung einer Digitalen Agenda des Statistischen Bundesamtes. Eines von vielen Themen, die in dem Workshop intensiv diskutiert und danach vertieft wurden, war das sogenannte „Machine Learning“. Es fand als Maßnahme Eingang in die erste Version der Digitalen Agenda, die am 27. Oktober 2018 an Staatssekretär Klaus Vitt im BMI gesandt wurde. Im Handlungsfeld „D. Datenanalyse automatisieren und neue Analysemethoden“ wurde die Maßnahme wie folgt umschrieben: „PoC Machine Learning – Proof-of-Concept für Machine Learning (Künstliche Intelligenz) aufsetzen, z. B. in den Unternehmensstatistiken, um automatische Kategorisierung durchzuführen und Analysepotenzial zu verbessern“. In der Leitungsklausur vom 13.–15. November 2017 haben die Amtsleitung und die Abteilungsleitungen 59 Maßnahmen der Digitalen Agenda nach Nutzen und Aufwand bewertet und priorisiert. Der Proof of Concept Machine Learning wurde hoch priorisiert und als eines von vier Leuchtturmprojekten der Digitalen Agenda eingestuft. Die inhaltliche Vorgabe wurde nur geringfügig modifiziert in „D1: Proof of Concept Machine Learning – Proof of Concept für Machine Learning (Künstliche Intelligenz) einrichten, z. B. in den Unternehmensstatistiken, um eine automatische Kategorisierung durchzuführen und das Analysepotenzial zu verbessern“. Der Termin für den Projektabschluss wurde auf den Juni 2018 festgelegt, die Verantwortlichkeit E1 übertragen.

1.2 Eingrenzung des Auftrags

Die Themen „Machine Learning“ und noch mehr „Künstliche Intelligenz“ sind sehr umfassend und kaum noch zu überschauen (siehe z. B. die Darstellung in Kapitel 3). Damit das Projekt „handhabbar“ blieb und mit den vorhandenen Kapazitäten in der zur Verfügung stehenden Zeit bearbeitet werden konnte, war es zunächst notwendig, die Aufgabenstellung einzugrenzen und zu operationalisieren. Das Projektteam verständigte sich darauf, sich ausschließlich auf das maschinelle Lernen zu konzentrieren und dessen Anwendbarkeit in den Fachstatistiken in den Fokus zunehmen, letzteres jedoch nicht beschränkt auf die Unternehmensstatistiken. Künstliche Intelligenz (zur Abgrenzung siehe Abschnitt 3.2) und deren Anwendung werden nicht (umfassend) behandelt. Dies gilt auch für Anwendungen von maschinellem Lernen und Künstlicher Intelligenz zur Aufgabenerledigung im Statistischen Bundesamt außerhalb der Fachstatistiken, z. B. in der Verwaltung. Auch die mit dem Einsatz von maschinellem Lernen/künstlicher Intelligenz verbundenen ethischen Fragen werden nicht näher untersucht.[1] Letztlich wird die Aufgabe Proof of Concept Machine Learning wie folgt interpretiert: Überprüfung der Einsetzbarkeit von maschinellem Lernen in den Prozessen der Fachstatistiken. Nach

[1] Sie wurden jedoch in den Kurzveranstaltungen angesprochen (siehe auch O’Neil 2017)

Abschluss des Proof of Concept sollte eine Übersicht der potenziellen Anwendungen in den Fachstatistiken vorliegen.

2 Vorgehensweise

Zu Projektbeginn lagen im Statistischen Bundesamt bereits konkrete Erfahren zum Einsatz von Machine-Learning-Verfahren in der Gruppe E1 vor.[2] Um deren Einsatzmöglichkeiten jedoch flächendeckend untersuchen zu können, sollten den Fachstatistikern zunächst Informationen über die Methoden selbst und die potenziellen Einsatzgebiete zur Verfügung gestellt und ein fachlicher Austausch ermöglicht werden. Auf dieser Wissensbasis sollte dann eine Hausumfrage zur potenziellen Verwendung von Methoden des maschinellen Lernens durchgeführt werden. Die konkrete Vorgehensweise war wie folgt geplant und wurde im Wesentlichen auch so umgesetzt:

- Einrichtung einer Informations- und Austauschplattform PCML TRAC bis Ende Februar 2018. Deren Zweck ist es, einerseits Informationen über Methoden, Anwendungen, Software etc. zur Verfügung zu stellen und andererseits den fachlichen Austausch aller an Machine Learning interessierten Mitarbeiterinnen und Mitarbeiter im Statistischen Bundesamt zu ermöglichen.
- Bestandsaufnahme der Anwendung von Machine-Learning-Verfahren im Statistischen Bundesamt sowie in nationalen und internationalen Statistikinstitutionen.
- Durchführung einer Kurzveranstaltung zum Thema „Machine Learning“, in der die Erkenntnisse der o.g. Bestandsaufnahme einflossen und über die Methoden selbst sowie über „prominente“ Anwendungsfälle informiert wurde.
- Durchführung und Auswertung der Hausabfrage.

[2] Streng genommen war damit der Proof of Concept bereits erbracht.

3 Was ist Machine Learning?

3.1 Der Begriff

Es gibt eine Vielzahl von Ansätzen, den Begriff des maschinellen Lernens zu definieren oder dessen Wesen zu erfassen. Arthur Samuel definierte maschinelles Lernen 1959 als Chance, Problemlösungsmethoden nicht mehr exakt implementieren zu müssen:

> „The studies reported here have been concerned with the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning. […] Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort."
>
> (Samuel 1959)

Herbert Simon schreibt 1983 hingegen:

> „Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time."
>
> (Simon 1983)

Und der Technologiejournalist Thomas Ramge fasst 2018 zusammen:

> „Bei Maschinellem Lernen erkennen Computersysteme Muster in Beispielen und können ihre ‚Erkenntnisse' auf andere Beispiele übertragen. So lernen sie, aus Daten immer genauere Schlüsse zu ziehen und Entscheidungen abzuleiten."
>
> (Ramge 2018)

Eine scharfe Abgrenzung zwischen Methoden der klassischen Statistik und Machine-Learning-Verfahren ist nicht möglich. Beispielsweise wird die logistische Regression, die seit dem frühen 20. Jahrhundert in der heutigen Form bekannt ist und daher (wie auch aus inhaltlichen Gründen) zu den Methoden der klassischen Statistik gezählt werden müsste, in (Lehr-)Büchern und entsprechendem Videomaterial häufig auch dem maschinellen Lernen zugeschlagen. Als Kriterien zur Unterscheidung von klassischer Statistik und Machine Learning können herangezogen werden:

- Flexibilität: Während die klassische Statistik häufig einschränkende Annahmen an das zugrundeliegende Modell oder die Daten macht, entfallen diese Annahmen zunehmend beim maschinellen Lernen.
- Interpretierbarkeit: Einschränkende Annahmen erlauben in der klassischen Statistik jedoch die sehr gute Interpretierbarkeit der Modelle. Mit zunehmender Flexibilität bei

Verfahren des Machine Learnings geht die Interpretierbarkeit zunehmend verloren (Black-Box-Problematik).

- Zielsetzung: Mit der Interpretierbarkeit einher geht die Frage nach der Zielsetzung. Während die klassische Statistik an der Erklärung von Zusammenhängen, idealerweise von Ursache-Wirkungs-Ketten, interessiert ist, spielen diese Aspekte beim maschinellen Lernen keine wichtige Rolle. Hier geht es stattdessen um möglichst gute Schätzungen für den Wert der Outputvariable gegeben die Ausprägungen der Inputvariablen (und es ist zunächst unerheblich, ob oder wie man deren Zustandekommen interpretieren kann).

Lernen kann auf verschiedene Arten beschrieben werden. Hier soll das sogenannte *statistische* maschinelle Lernen betrachtet werden, das es ermöglicht, den Vorgang und den Erfolg des Lernens mathematisch zu fassen. Das Ziel besteht stets darin, zu gegebenen Eingabewerten (auch als Werte oder Realisierungen von Eingabemerkmalen, Eingabevariablen, erklärenden Variablen oder Inputvariablen bezeichnet) einen brauchbaren Ausgabewert (auch als Wert des zu erklärenden Merkmals, der zu erklärenden Variable, der Ausgabevariable oder der Outputvariable bezeichnet) zu schätzen. Es muss also ein Zusammenhang zwischen den Eingabevariablen und der Ausgabevariable erlernt werden, der anschließend auf neue, ggf. noch nicht bekannte Eingabewerte angewendet werden kann. Dieses Lernen muss statistisch geschehen, also auf Basis von nur endlich vielen Beobachtungen von Eingabe- und ggf. Ausgabewerten. Ausgerichtet muss es aber auf das Ziel sein, Aussagen über die Grundgesamtheit zu treffen, d. h. für alle denkbaren Kombinationen von Eingabe- und Ausgabewerten, also alle denkbaren Realisierungen der Eingabe- und Ausgabevariablen.

Jedes Berechnen einer Regressionsfunktion o.ä. kann im weiteren Sinne als statistisches Lernen bezeichnet werden, erfasst es doch die Informationen im Datensatz, um später zu neuen Beobachtungen entsprechende Outputwerte zu schätzen. Damit wird auch deutlich, dass die Methoden des Machine Learnings bessere Ergebnisse als die der klassischen Statistik liefern können, aber nicht müssen. Dass man von statistischem *maschinellen* Lernen spricht, liegt darin begründet, dass einige der heute unter diesem Begriff firmierenden Methoden vor der Entwicklung entsprechend leistungsstarker Computer zwar theoretisch denkbar, praktisch jedoch ohne maschinelle Unterstützung nicht oder nicht für große Datenmengen durchführbar waren.

Alle im Rahmen dieses Dokuments beschriebenen Projekte sind im Feld des sogenannten überwachten Lernens (supervised learning) anzusiedeln, bei dem der Trainingsdatensatz (d. h. die Menge der zum Lernen zur Verfügung stehenden Daten) sowohl Eingabe- als auch zugehörige Ausgabewerte enthält. Spezieller betrachtet wird die binäre Klassifikation, also die Schätzung der Zugehörigkeit zu einer von zwei Klassen. Letzteres stellt keine praxisrelevante Einschränkung dar. Die vorhandenen Softwarelösungen sind zum überwiegenden Teil in der Lage, auch

Mehrklassenprobleme zu lösen. Dies kann durch explizite Mehrklassen-Algorithmen geschehen oder durch einen Rückgriff auf binäre Klassifikationsprobleme wie beispielsweise bei sogenannten One-versus-One- oder One-versus-All-Ansätzen. Wurde der Zusammenhang einmal gelernt, so kann die zu erfüllende Aufgabe im Sinne von Simon (im Folgenden die Klassifikation) effizienter (d. h. im vorliegenden Fall ressourcenschonender) und/oder effektiver gelöst werden. Maschinelles Lernen in diesem Sinne kann somit nur funktionieren, wenn hinreichend gutes Material vorhanden ist, anhand dessen gelernt werden kann. Maschinelles Lernen kann darüber hinaus scheitern, wenn sich die Aufgabe nach dem Lernen nicht mehr auf die gleiche Grundgesamtheit bezieht, ein früher gelernter Zusammenhang also möglicherweise gar nicht mehr zutrifft.

Die mathematische Beschreibung des statistischen maschinellen Lernens ist ein Verdienst von Vladimir Vapnik, der das allgemeine „Problem zu lernen“ in drei Komponenten zerlegt (Vapnik 1995):

(1) Zunächst gibt es einen Mechanismus, welcher die Eingabewerte gemäß einer dem Anwender unbekannten (aber sich nicht verändernden) Verteilung erzeugt.
(2) Anschließend tritt ein Mechanismus in Kraft, der zu jedem Eingabewert einen Ausgabewert erzeugt. Dies geschieht aufgrund von dem Anwender unbekannten, aber wiederum sich nicht verändernden bedingten Verteilungen für die Ausgabe gegeben die Eingabe.
(3) Aufgabe des maschinellen Lernens ist es nun, aus einer geeignet zu fassenden Fülle an möglichen funktionalen Zusammenhängen zwischen Eingabe- und Ausgabevariablen diejenige Funktion auszuwählen, die das in Schritt (2) betrachtete Verhalten am besten approximiert. Gesucht wird also derjenige Prädiktor, der die bedingten Verteilungen der Ausgabewerte gegeben die Eingabewerte am besten reproduziert.

Maschinelles Lernen bedient sich meist nichtparametrischer Methoden. Im Unterschied zu parametrischen Verfahren unterstellen solche Methoden nicht bereits von vorneherein ein bestimmtes Modell von Verteilungen, dem die Schritte (1) und (2) in Vapniks Zerlegung folgen. Die Zusammenhänge zwischen Eingabe- und Ausgabevariablen werden durch die nichtparametrischen Methoden erst noch entdeckt. Somit lösen nichtparametrische Verfahren dieses Grundproblem der Statistik, Informationen über die zugrunde liegende Verteilung anhand der gegebenen Daten zu schätzen, wesentlich allgemeiner als parametrische Verfahren.

Andere Formen des maschinellen Lernens haben keine Informationen über die wahren Ausgabewerte oder das Problem selbst enthält in seiner Formulierung keine Ausgabevariablen. Dies ist beispielsweise bei der Ausreißeridentifikation, der Schätzung der Hauptmasse einer Verteilung oder beim Clustering der Fall. Diese Kategorie wird unüberwachtes Lernen

(unsupervised learning) genannt. Eine andere Form, die jedoch mutmaßlich eher wenig Anwendung in der amtlichen Statistik finden wird, ist das bestärkende Lernen (reinforcement learning).

3.2 Gängige Methoden des Machine Learning

Maschinelles Lernen schlägt sich in verschiedenen Methoden nieder. Wegen deren breiter Verwendung sollen hier die Grundzüge von Neuronalen Netzen, Random Forests und Support Vector Machines dargestellt werden.

Die theoretischen Grundlagen zu Neuronalen Netzen, die im Prinzip die „Aktivitäten" des menschlichen Gehirns nachbilden, wurden bereits in den 1950er Jahren gelegt (Rosenblatts Perceptron, Rosenblatt 1958), wieder aufgegriffen wurden diese Ideen jedoch erst in den 1980er Jahren mit der Verbreitung erster Hochleistungscomputer. Während das Perceptron nur lineare Zusammenhänge zwischen Eingabevariablen und Ausgabevariable darstellen kann, sind Neuronale Netze in der Lage, nichtlineare Zusammenhänge zu modellieren. Dies wird durch das Einziehen sogenannter verdeckter Schichten (hidden layers) mit ggf. unterschiedlich vielen Neuronen möglich. Ein Perceptron ist äquivalent zu einem zweischichtigen (eine Input- und eine Outputschicht), gerichteten Neuronalen Netzwerk. Dabei kommt es weniger auf die Neuronen selbst als vielmehr auf die Verbindungen zwischen diesen an. Die Gewichte dieser Verbindungen sind die in optimaler Weise zu ermittelnden Parameter von Neuronalen Netzen. Der sogenannte Backpropagation-Algorithmus stellt das meistgenutzte Verfahren zum Lernen von Neuronalen Netzen dar. Er geht nach folgendem Schema vor: Nach dem Berechnen des Ausgabewertes (sogenanntes Vorwärtspropagieren) für ein Element des Trainingsdatensatzes wird der Fehler festgestellt. Das ist im supervised learning stets möglich. Dieser wird dann beim sogenannten Rückwärtspropagieren genutzt, um von Schicht zu Schicht rückwärts die Gewichte anzupassen. Diese Schritte werden nun auf alle Beobachtungen im Trainingsdatensatz angewendet und so lange wiederholt, bis sich beispielsweise die Gewichte nicht mehr oder nicht mehr stark ändern.

Support Vector Machines und Random Forests sind im Vergleich zu Neuronalen Netzen relativ jung und wurden 1992 (Boser et al. 1992) respektive 2001 (Breiman 2001) erstmals vorgestellt.

Random Forests basieren auf sogenannten Klassifikations- oder Regressionsbäumen (auch als Decision Trees bezeichnet), die wiederum bereits in Breiman et al. (1984) veröffentlicht wurden, und kombinieren diese einerseits mit der Idee wiederholter Ziehungen aus dem vorhanden Datenmaterial, sogenannte Bootstrap-Ziehungen, andererseits mit einer zufälligen Auswahl relevanter Variablen während der Berechnung des Modells. Auf diese Weise können verschiedene (meist 500 oder 1.000), nur sehr schwach korrelierte Bäume gelernt werden. Ein Random Forest ordnet einen neuen Datenpunkt anhand der Klassifikationsentscheidung der Mehrheit der ihm zugrundeliegenden Bäume einer Klasse zu. Beim Einsatz von Random Forests

für Regression würde ein Mittelwert der Regressionswerte der Bäume gebildet werden. Bäume selbst nehmen Klassifikationen auf Basis einer immer feiner werdenden achsenparallelen Aufteilung des Eingaberaums vor (recursive partitioning), wobei beispielsweise eine vom Nutzer gesetzte Maximalzahl der Teilräume verhindert, dass sich ein Baum zu gut an einen Trainingsdatensatz anpasst und somit nicht mehr für neue Datenpunkte geeignet ist. Der WISTA-Aufsatz Feuerhake und Dumpert (2016) enthält eine ausführliche Darstellung dieser Methoden.

Support Vector Machines stellen im ursprünglichen Sinne eine Klassifikationsmethode basierend auf der Idee dar, die Trennung zwischen den Klassen abstandsmaximierend und linear, d. h. je nach Wahl der Inputvariablen durch eine Gerade, eine Ebene oder eine Hyperebene vorzunehmen. Analytische Formulierungen des Optimierungsproblems erlauben darüber hinaus auch die Behandlung von Regressionsmodellen oder die Ausreißeridentifikation. Wesentlicher Bestandteil von Support Vector Machines ist der sogenannte Kern-Trick, mit dessen Hilfe in der Beispielsituation, dass es zwei erklärende Variablen und eine vorherzusagende Klasse gibt, die Eingabevariablen, die sich im 2D-Bild nicht durch eine gerade Linie trennen lassen, in einen höheren Raum abgebildet werden. Diese Hinzunahme von Dimensionen ermöglicht eine (abstandsmaximale) lineare Trennung der Datenpunkte. Anschließend erfolgt die Rückprojektion, woraus nichtlineare Trennungen im 2D-Bild entstehen. Details dieser Methode wurden bereits in WISTA-Aufsätzen beschrieben: Die analytische Herleitung findet sich in Dumpert et al. (2016), die geometrische in Feuerhake und Dumpert (2016).

3.3 Abgrenzung zu Künstlicher Intelligenz und zu Big Data

Wenngleich die öffentliche Debatte bisweilen die Begriffe „Künstliche Intelligenz“ und „Maschinelles Lernen“ synonym verwendet, gibt es doch wesentliche Unterschiede. Spricht man über Künstliche Intelligenz, so ist zunächst die Entstehung dieses Begriffs in Erinnerung zu rufen. 1955 beantragte der damalige Assistenzprofessor für Mathematik in New Hampshire, John McCarthy, die Förderung eines zweimonatigen Sommercamps für zehn ausgewählte Forscher im Jahr 1956. Ziel des Camps war es, herauszufinden, „wie Maschinen dazu gebracht werden können, Sprache zu benutzen, Abstraktionen und Begriffe zu bilden, Probleme zu lösen, die zu lösen bislang dem Menschen vorbehalten sind, und sich selbst zu verbessern.“ (übersetzt aus McCarthy et al. 1955) Der Titel dieses Forschungsprojektes war „Artificial Intelligence“. Überlegungen in diese Richtung gab es selbstverständlich schon zuvor; die Bezeichnung des Forschungsgebietes als „Künstliche Intelligenz“ wurde aber erstmalig hier aktenkundig. Wann eine Maschine als intelligent gilt, ist eine bis heute ungeklärte Frage, bleibt doch stets offen, was Intelligenz als solche ist. 1950 schlug Alan Turing vor, die Frage nicht direkt zu beantworten, sondern stattdessen zu prüfen, ob eine Maschine bei einem Menschen den Eindruck erwecken kann, intelligent zu sein (Turing-Test). Oder kurz: „Ziel der KI ist es, Maschinen zu entwickeln, die sich verhalten, als verfügten sie über Intelligenz“ (übersetzt aus McCarthy et al. 1955).

Manchmal wird eher von Kognition als von Intelligenz gesprochen, was das Problem aber nur verschiebt, da Kognition im engeren Sinne als bewusstes, reflektiertes und sprachlich gefasstes Denken verstanden wird. Hiervon sind Maschinen jedoch weit entfernt. Interessanterweise ist auch die Wahrnehmung von Menschen in diesen Belangen verzerrt. Wer beispielsweise eine komplizierte Division oder die Berechnung einer allgemeinen Wurzel schnell und fehlerfrei vollziehen kann, gilt häufig als intelligenter Mensch; dennoch würde man keinem Taschenrechner Intelligenz zusprechen. Darüber hinaus wird menschliche Intelligenz meist durch unterbewusste Prozesse gesteuert; eine direkte Übertragung von Vorgängen oder Routinen auf Maschinen ist somit schwer möglich. Menschen verstehen beispielsweise selbst nicht genau, weshalb sie ein bekanntes Gesicht in einer Menschenmenge wiedererkennen und einem Namen oder einem Erlebnis zuordnen können. Und so bietet sich eine weitere mögliche, zeitlose Definition von Künstlicher Intelligenz an (Rich 1983): „Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better."

Russell und Norvig fassen notwendige Komponenten Künstlicher Intelligenz in ihrem Buch „Künstliche Intelligenz – Ein moderner Ansatz" (Russell und Norvig 2012) unter folgenden Schlagworten zusammen: Verarbeitung natürlicher Sprache (zur Kommunikation), Wissensrepräsentation (Abspeichern und Organisieren vorhandener Informationen), automatisches logisches Schließen (Schlussfolgerungen und Beantwortung von Fragen), maschinelles Lernen (Anpassung an neue Umstände, Mustererkennung, Extrapolation), Computervision (Wahrnehmung von Objekten), Robotik (Manipulation und Bewegung von Objekten).

Auf eine Kleine Anfrage der Fraktion Bündnis 90/Die Grünen an die Bundesregierung, siehe Deutscher Bundestag (2018b), wie sie „Künstliche Intelligenz" (auch in Abgrenzung zu maschinellem Lernen, Deep Learning etc.) definiert, antwortete diese, siehe Deutscher Bundestag (2018c):

> „Künstliche Intelligenz (KI) ist ein Teilgebiet der Informatik, welches sich mit der Erforschung von Mechanismen des intelligenten menschlichen Verhaltens befasst. Dabei geht es darum, technische Systeme so zu konzipieren, dass sie Probleme eigenständig bearbeiten und sich dabei selbst auf veränderte Bedingungen einstellen können. Diese Systeme haben die Eigenschaft, aus neuen Daten zu „lernen" und mit Unsicherheiten umzugehen, statt klassisch programmiert zu werden. Die etablierten Forschungsgebiete der KI sind:
>
> 1. Deduktionssysteme, maschinelles Beweisen: Ableitung formaler Aussagen aus logischen Ausdrücken, Systeme zum Beweis der Korrektheit von Hardware und Software;

2. Wissensbasierte Systeme: Methoden zur Modellierung und Erhebung von Wissen; Software zur Simulation menschlichen Expertenwissens und Unterstützung von Experten (früher: „Expertensysteme“); eng verbunden mit Psychologie und Kognitionswissenschaften;
3. Musteranalyse und Mustererkennung: Induktive Analyseverfahren allgemein, damit auch maschinelles Lernen;
4. Robotik: Autonome Steuerung von Robotik-Systemen;
5. Intelligente multimodale Mensch-Maschine-Interaktion: Analyse und „Verstehen“ von Sprache (in Verbindung mit Linguistik), Bildern, Gestik und allen anderen Formen menschlicher Interaktion und Reaktion darauf.

Maschinelles Lernen ist eine Methodik im Gebiet der Künstlichen Intelligenz mit einem besonderen Schwerpunkt im Bereich Mustererkennung und bei Anwendungen in der Robotik. Deep Learning ist ein spezielles Verfahren des Maschinellen Lernens.“

Sowohl die gängigen Aufzählungen in Lehrbüchern als auch die Auffassung der Bundesregierung enthalten das maschinelle Lernen als notwendige Voraussetzung für ein System, um als Künstliche Intelligenz eingestuft werden zu können. Betrachtet man Fragestellungen der amtlichen Fachstatistik, so betreffen diese aus obigen Aufzählungen hauptsächlich die Wissensrepräsentation und sowie das maschinelle Lernen; die übrigen Komponenten erscheinen wenig zweckmäßig. Insofern sollte im Hinblick auf neue statistische Methoden von Machine Learning und nicht von Künstlicher Intelligenz gesprochen werden.

Big Data bezeichnet eine Situation, in der bestimmte Daten verfügbar sind, und beschreibt somit etwas anderes als Künstliche Intelligenz und Machine Learning. Im Allgemeinen wird Big Data durch die drei Vs charakterisiert: Volume, Velocity und Variety (siehe beispielsweise die entsprechenden Aufsätze in König et al. (2017)). Das aktuelle Buch von Suthaharan (2016) vergleicht verschiedene Datensituationen in folgender Graphik:

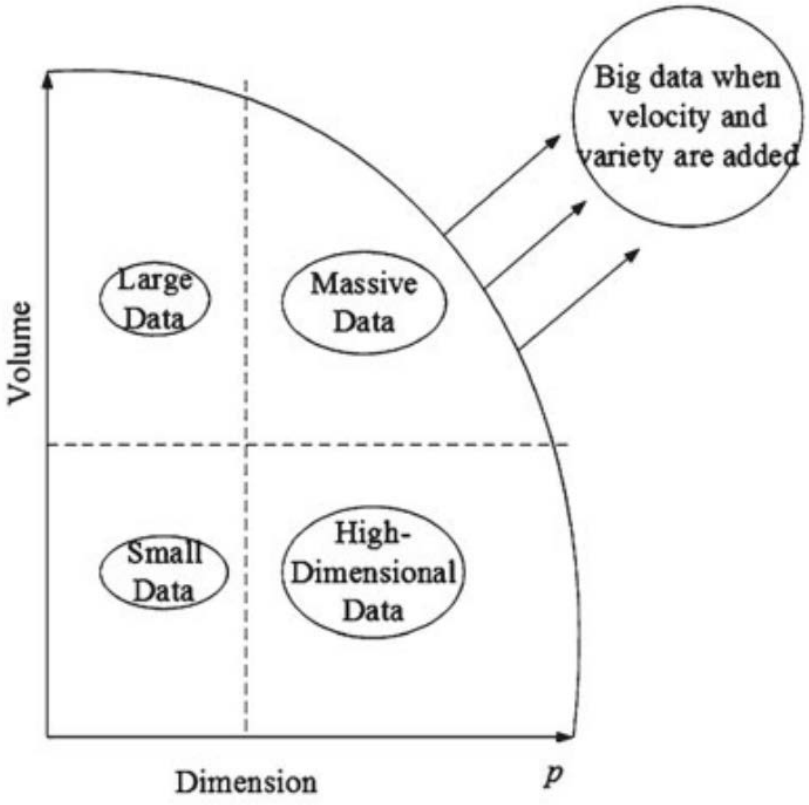


Abbildung 1: Abgrenzung von Big Data

„Dimension“ (p) ist dabei die Anzahl der Variablen. Während Volume in der Umschreibung von Big Data die schiere Menge an Beobachtungen bei gleichzeitig vielen Variablen beschreibt, steht Velocity für das ständige und schnelle Nachströmen neuer Beobachtungen, häufig schneller als statistische Verfahren damit umgehen können. Variety bringt darüber hinaus eine weitere Schwierigkeit ins Spiel: Die Daten könne von höchst unterschiedlichen Skaleniveaus und bar jeglicher Struktur sein. In einem früheren Artikel gibt Suthaharan (2014) eine Definition: „It means that some point in time, when the volume, variety and velocity of the data are increased, the current techniques and technologies may not be able to handle storage and processing of the data. At that point the data is defined as Big Data.“ Hin und wieder werden noch die Vs „Value“ (im Sinne von (meist monetären) Mehrwert generierend) und „Veracity“ (im Sinne von vagen und ungenauen Daten) genannt. Ohne Frage eignen sich einige Methoden aus dem Bereich des maschinellen Lernens zum Umgang mit Big Data, ggf. sind aber Anpassungen von Nöten. Beispielsweise tun sich Support Vector Machines sehr schwer in Big Data Situationen. Dennoch gibt es hier Ansätze (z. B. distributed learning, regionalization, online learning). Auch bei Random Forests braucht es Modifikationen, da die bei dieser Methode eingesetzten Bootstrapsamples nicht mehr zu verarbeiten sind. Auswege bieten hier gestufte Modelle (zunächst eine die betrachteten Beobachtungen reduzierende Zufallsstichprobe, dann erst die Bootstrapstichproben daraus; und dieser Vorgang wird einige Male wiederholt). Die Vorzüge von

Machine-Learning-Methoden, bislang unbekannte Muster in den Daten zu erkennen, sind jedoch gerade in der Situation von Big Data hervorzuheben.

3.4 Einbettung in aktuelle Entwicklungen

Maschinelle Lernverfahren finden breite Anwendung in verschiedensten Bereichen von Wirtschaft, Verwaltung und Forschung. Als Beispiele können hier die Bestimmung von Wahrscheinlichkeiten, wann und in welcher Straße einer Stadt potenziell eine Straftat begangen wird (Predictive Policing), die Auswertung von Bild- und Videomaterial (Erkennung von Gegenständen, Personenidentifikation, Rückschluss auf personenbezogene Merkmale wie beispielsweise die sexuelle Orientierung, …) sowie die Klassifikation von Dokumenttypen im Posteingang von Unternehmen genannt werden. Weitere Anwendungen finden sich beim Erkennen von Versicherungsbetrug, bei der Schätzung von potenziell erzielbaren Mietpreisen bei Immobiliensowie bei (Sprach-) Dialogsystemen. Um die Vielfalt der Einsatzmöglichkeiten aufzuzeigen, seien weiterhin genannt: Die Muster- oder Zeichenerkennung (Erkennen von z. B. handgeschriebenen oder gescannten Buchstaben und Ziffern), der Einsatz in der Bioinformatik sowie das Erkennen physikalischer Phänomene, z. B. das Erkennen von Phasen und Phasenübergängen bei Materie, oder die Lösung der Schrödingergleichung in Vielteilchensystemen; daneben die Kreditrisikoanalyse, Analyse von sozialen Netzwerken, Erkennung von Steroiden im Antidopingkampf, Vorhersage der Wasserqualität, Schätzung der Bodenversiegelung aufgrund von Fernerkundungsdaten, Demokratiemessung, Go- und Schachspiele und viele mehr. Die Vielfalt der Anwendungen spricht dafür, dass maschinelles Lernen auch in der Fachstatistik erfolgreich eingesetzt werden kann. Auf den Einsatz maschineller Lernverfahren durch Statistik-Produzenten wird in Kapitel 5 eingegangen. Zunächst soll jedoch untersucht werden, ob maschinelles Lernen eine Modeerscheinung ist oder ob es so nachhaltig ist, dass eine intensive Auseinandersetzung damit für die amtliche Statistik lohnend erscheint.

Exkurs: Machine Learning/Künstliche Intelligenz für nicht-statistische Anwendungen im Statistischen Bundesamt

Auch wenn die nicht-statistischen Anwendungsgebiete nicht im Vordergrund standen, wurden bei der Durchführung des Proof of Concept folgende (kommerzielle) Produkte „entdeckt“, die grundsätzlich auch im Statistischen Bundesamt eingesetzt werden könnten:

DeepL: Hierbei handelt es sich um ein nicht auf Vokalbelisten und Grammatikregeln basierendes Übersetzungsprogramm, das den Sprachendienst in B105 bei Übersetzungstätigkeiten unterstützen könnte. In der Leistungsfähigkeit ist DeepL z. B. Google Translator deutlich überlegen. Eine kostenfreie Version von DeepL steht im Internet unter https://www.deepl.com/translator zur Verfügung.

Precire: Dieses Produkt wird von renommierten Unternehmen bei der Auswahl von Führungskräften eingesetzt. Die Software erkennt durch Sprachanalyse die Bewerberpersönlichkeit. Grundlage ist ein Telefonat des Bewerbers mit einem Computer. Die (Vor-)Auswahl von Bewerbungen kann schneller und kostengünstiger als durch Assessmentcenter erfolgen. Eine Einsatzmöglichkeit bestünde eventuell im Bewerbungsbüro in A201.

Google Duplex: Anfang Mai 2018 hat Google erstmals den Sprachassistenten Google Duplex vorgestellt. Dabei geht es Google darum, eine Konversation zwischen Mensch und Computer in natürlicher Sprache zu ermöglichen, bei der der Computer nicht mehr als solcher erkannt wird. Die Technik befindet sich noch im Experimentierstadium. Google Duplex kann Restauranttische bestellen und Friseurtermine vereinbaren. Langfristig könnten solche Systeme jedoch Telefonate für unterschiedliche Aufgaben übernehmen, im Statisischen Bundesamt z. B. eventuell im Auskunftsdienst.

3.5 Entwicklung der Forschungsaktivitäten sowie der öffentlichen und politischen Wahrnehmung

Das maschinelle Lernen gewinnt zunehmend auch an Universitäten und Hochschulen an Bedeutung. Exemplarisch sei hier eine Ausschreibung der Universität Göttingen vom Mai 2018 genannt: Vier Professuren für Data Science (W3) und Künstliche Intelligenz/Maschinelles Lernen (W3, W2 t.t. W3). Hinzu kommen regelmäßig Ausschreibungen für Doktoranden- und Postdoc-Stellen in diesem Bereich.

Auch in der Forschung lässt sich eine dynamische Entwicklung aufzeigen. Fachzeitschriften aus dem Bereich Machine Learning verzeichnen eine zunehmende Zahl an Einreichungen. Beispielsweise verdreifachten sich die Zahlen beim Journal of Machine Learning Research seit 2005. Die sehr häufig (seit 1991 insgesamt 1,4 Mio. Publikationen) für Preprints genutzte Internetseite https://arxiv.org der Cornell University enthält auch Metadaten der zum Download zur Verfügung gestellten Aufsätze, die im weiteren Sinne überwiegend den Naturwissenschaften zuzuordnen sind. Die Angaben zum Erscheinungsjahr und zur wissenschaftlichen Disziplin sowie zum verwendeten Machine-Learning-Verfahren wurden für den Proof of Concept per Webscraping ausgelesen und anschließend ausgewertet. Die Ergebnisse werden im Folgenden graphisch aufbereitet dargestellt. Dabei ist jeweils zu berücksichtigen, dass das Jahr 2018 nur bis Mitte Mai ausgewertet wurde. Mit Blick auf Machine Learning zeigt sich ein extremer Anstieg der Zahl der wissenschaftlichen Beiträge in den letzten sechs Jahren, insbesondere im Bereich Computer Science (Informatik) sowie in etwas geringerem Maße auch im Bereich Statistik (siehe Abbildung 2). Die eher theoretisch orientierten, mathematischen Beiträge stiegen ebenfalls an, jedoch nur auf weit geringere Zahlen. In der Tat ist bei den Forschungsaktivitäten (und ebenso bei den Veröffentlichungen) im Bereich des maschinellen Lernens zu trennen zwischen

- theoretischen Arbeiten, die deduktiv Eigenschaften (z. B. Existenz von Lösungen, Eindeutigkeitsaussagen, Konsistenz, Unverzerrtheit, Lernraten, Fehlerschranken, Konfidenzbereiche usw.) maschineller Lernverfahren herleiten;
- Arbeiten, die Algorithmen theoretisch verbessern und die Ergebnisse empirisch überprüfen;
- Arbeiten, die rein empirisch Eigenschaften von Verfahren aufzudecken oder zu überprüfen versuchen.

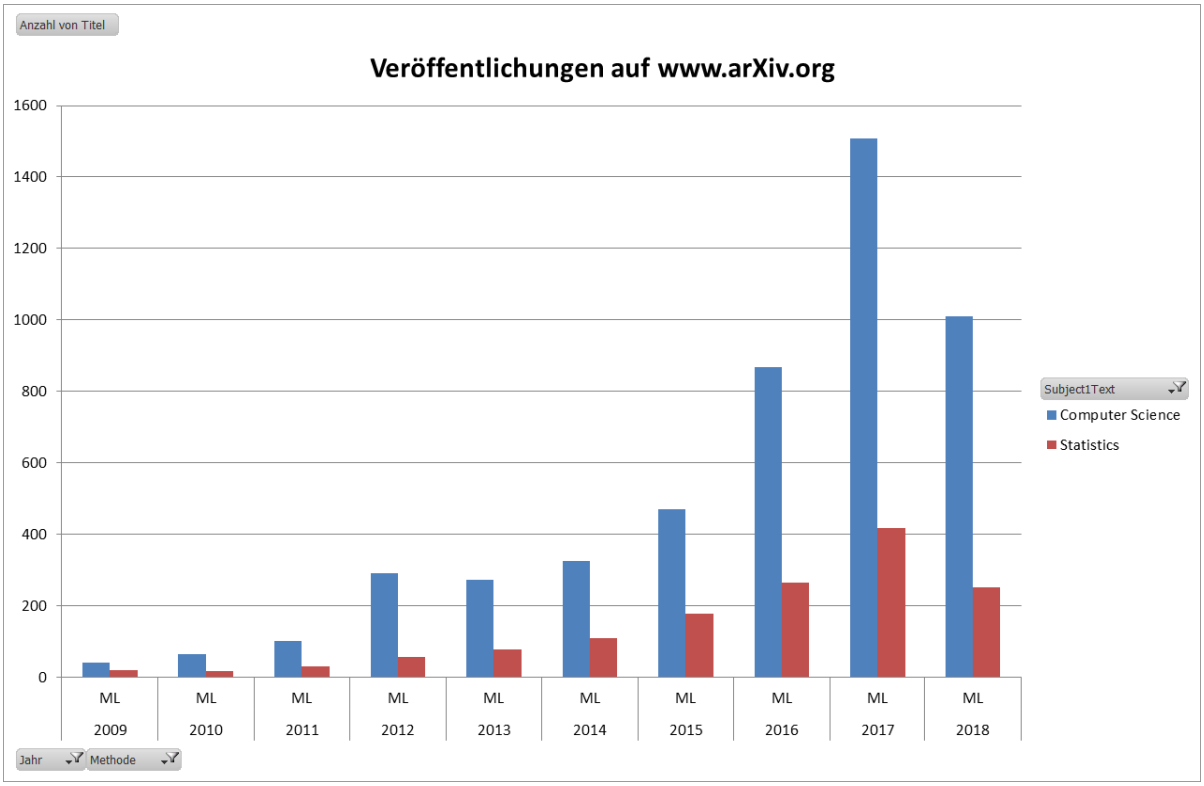


Abbildung 2: Veröffentlichungen auf arXiv.org – Statistik und Informatik

Betrachtet man die weiteren Einsatzgebiete von Machine-Learning-Verfahren, so ergibt sich das Bild in Abbildung 3. Artikel zu Machine-Learning-Verfahren finden zunehmende Verbreitung in der Mathematik, der Astrophysik, der Quantenphysik, weiteren Bereichen der Physik sowie in der quantitativen Biologie.

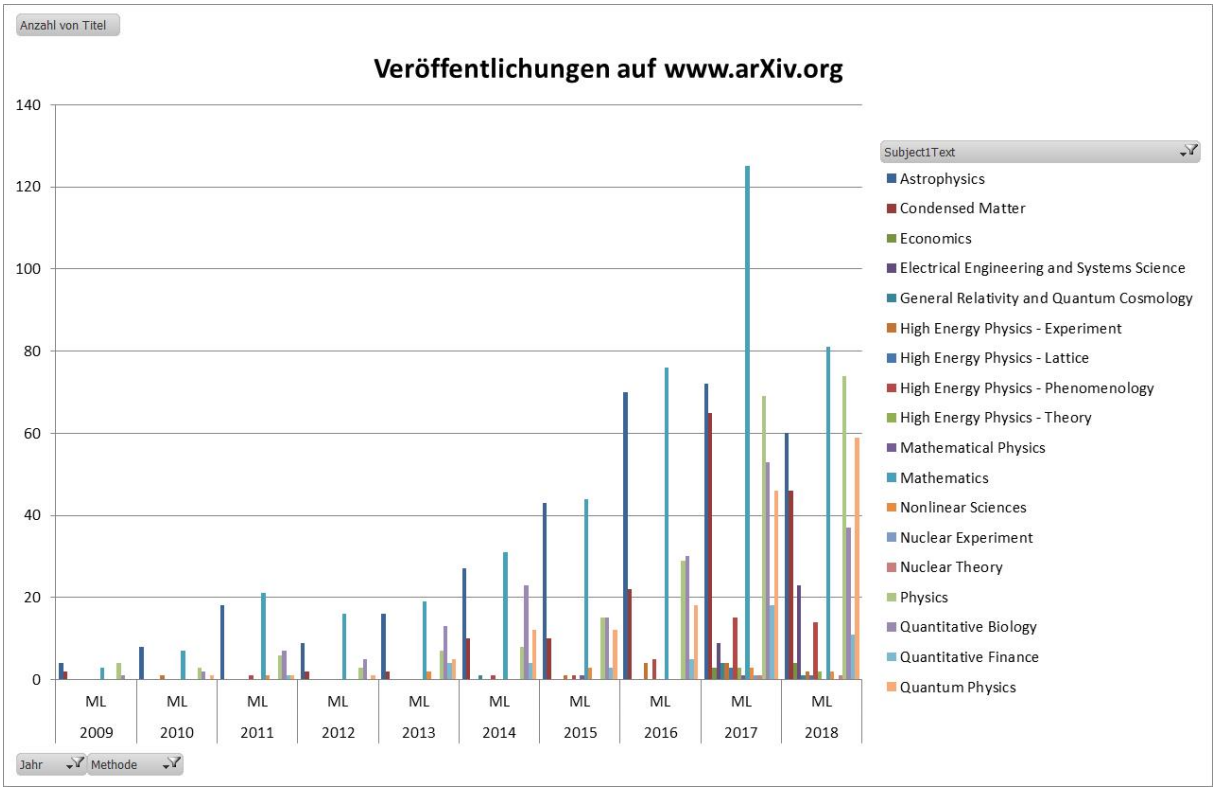


Abbildung 3: Veröffentlichungen auf arXiv.org – andere Forschungsfelder

Sortiert man die Beiträge nach gängigen Verfahren des maschinellen Lernens, erkennt man deutlich die aktuelle Dominanz von Veröffentlichungen zu Neuronalen Netzen (NN), gefolgt von Support Vector Machines (SVM) und Random Forests (RF) (siehe Abbildung 4). Die Zahl der Veröffentlichungen ist in den 2010er Jahren explosionsartig angestiegen. Dies hängt vermutlich mit einen verbesserten Datenzugang, der Verfügbarkeit von geeigneten Softwarelösungen sowie der Entwicklung ausreichend leistungsfähiger Hardware zusammen.

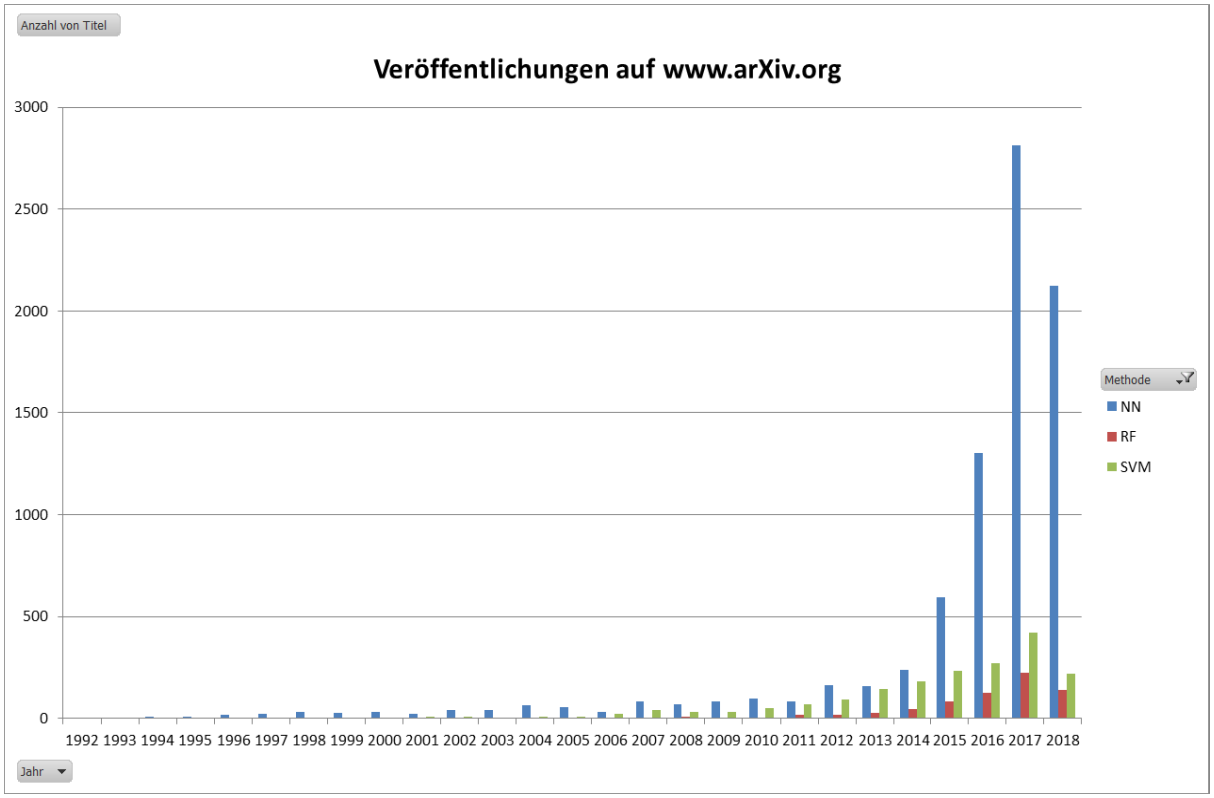


Abbildung 4: Veröffentlichungen auf arXiv.org - Machine Learning Verfahren

Das ifo-Institut hat im April 2018 eine Studie zur Verwendung verschiedener empirischer Methoden in wirtschaftswissenschaftlichen Veröffentlichungen publiziert, die auf rund 777.000 Einträgen der EconLit-Datenbank basiert (Huber et al. 2018). Die Autoren stellen fest, dass seit 2013 das stärkste Wachstum beim Einsatz vom maschinellem Lernen zu erkennen ist. 2013 steigt dessen prozentualer Anteil an allen Publikationen sprunghaft an, die Entwicklung hält seitdem an (siehe Abbildung 5).

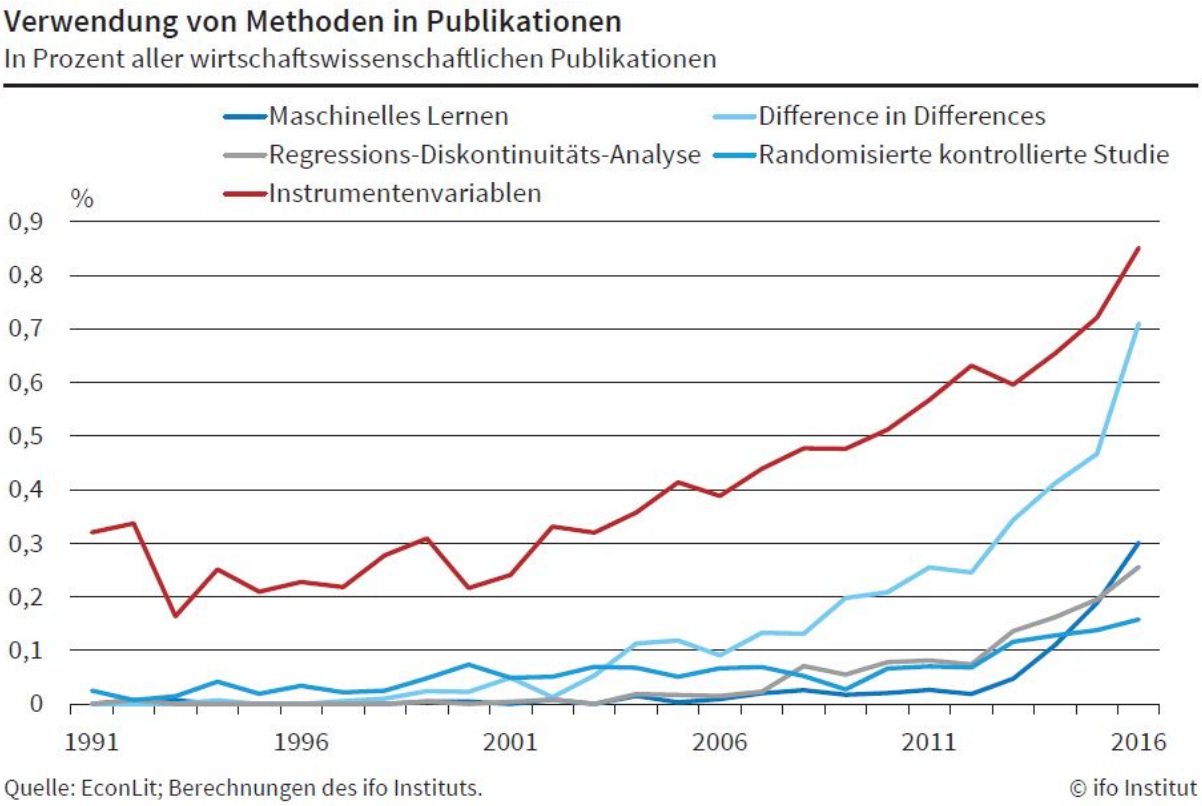


Abbildung 5: Veröffentlichungen auf EconLit – Maschinelles Lernen

Die deutliche Steigerung der Forschungsaktivitäten und die erzielten Erfolge schlagen sich auch in der öffentlichen Wahrnehmung nieder. So zeigt eine Suche im F.A.Z.-Archiv der Frankfurter Allgemeinen Zeitung zum Stichwort „Künstliche Intelligenz“ eine Zunahme der entsprechenden Artikel von 31 in 2013 auf 673 im Jahr 2017 und fast genau soviel im ersten Halbjahr 2018. (siehe Abbildung 6) Auch eine Recherche von SPIEGEL, Süddeutscher Zeitung und Handelsblatt zeigt ein ähnliches Bild.

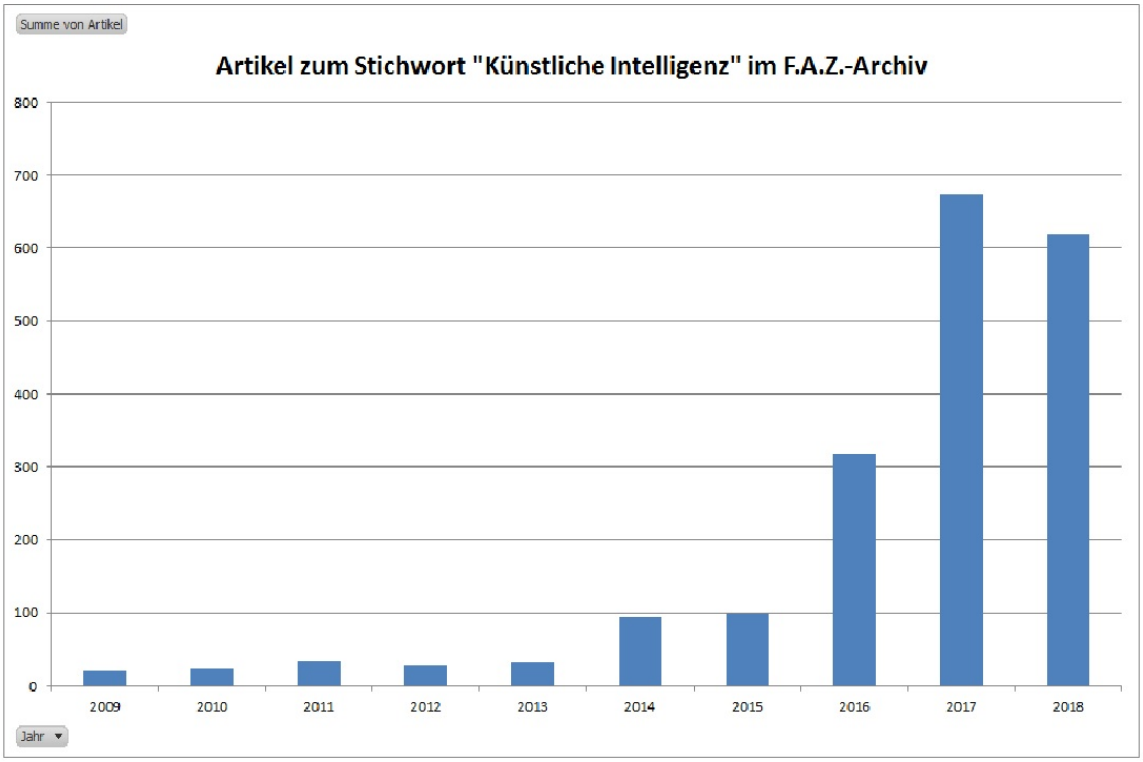


Abbildung 6: Artikel zum Stichwort „Künstliche Intelligenz“ im F.A.Z.-Archiv

Ein weiteres Indiz für die breite und zunehmende öffentliche Wahrnehmung liefert eine Analyse von Google Trends zum Stichwort „Machine Learning“. Auch hier ist eine deutliche und sich schnell beschleunigende Entwicklung der Suchanfragen seit 2013 erkennbar (siehe Abbildung 7).

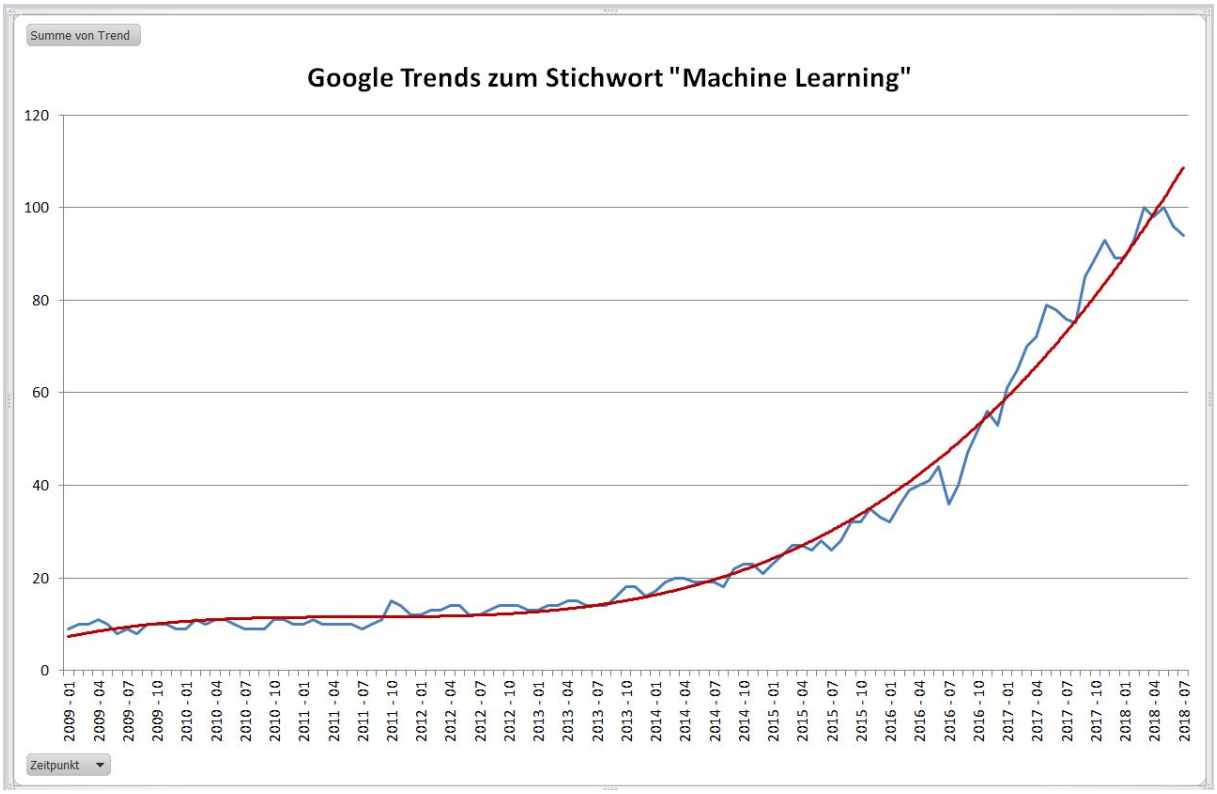


Abbildung 7: Google Trends zum Stichwort „Machine Learning“

Spätestens mit der letzten Bundestagswahl hat das Thema maschinelles Lernen/Künstliche Intelligenz auch Eingang in die politische Agenda gefunden. Im Koalitionsvertrag von CDU, CSU und SPD vom 7. Februar 2018 werden künstliche Intelligenz und maschinelles Lernen mehrfach angesprochen. Auf S. 35 heißt es: „Technologische Basis und Triebfeder der Digitalisierung sind Mikroelektronik, moderne Kommunikationstechnik, künstliche Intelligenz, Robotik, Datenwissenschaften, IT-Sicherheit und Quantentechnologien. Wir wollen die Forschung zu diesen Schlüsseltechnologien intensiv fördern, inklusive sozialer und geisteswissenschaftlicher Begleitforschung. Insbesondere wollen wir Deutschland zu einem weltweit führenden Standort bei der Erforschung von künstlicher Intelligenz machen. Hierzu wollen wir aus der Plattform Lernende Systeme heraus ein Nationales Forschungskonsortium für künstliche Intelligenz und maschinelles Lernen aufbauen und konsequent auf Anwendungen in allen Feldern der Forschungs- und Innovationsstrategie ausrichten. Wir werden gemeinsam mit unseren französischen Partnern ein öffentlich verantwortetes Zentrum für künstliche Intelligenz errichten. Gemeinsam mit Polen wollen wir ein Zentrum für digitale Innovationen in der Systemforschung einrichten.“ Außerdem plant die Bundesregierung einen „Masterplan ‚Künstliche Intelligenz‘ auf

nationaler Ebene“. Das gemeinsam mit Frankreich geplante Zentrum kann auf dem von dem französischen Mathematiker und Parlamentsabgeordneten Cédric Villani im Auftrag des Parlaments erarbeiteten und im März 2018 veröffentlichten Strategiepapier „For a Meaningful Artificial Intellience. Towards a French and European Strategy“ (Villani 2018) aufbauen.

Die Bundesregierung hat inzwischen konkrete Maßnahmen eingeleitet:

- Am 9. April 2018 kündigte das Bundesministerium für Bildung und Forschung (BMBF) zum Thema „Maschinelles Lernen“ an (Bundesministerium für Bildung und Forschung 2018): „Um neue Erkenntnisse aus großen und komplexen Datenmengen ableiten zu können, müssen die Daten durch leistungsfähige wissenschaftliche Analysemethoden aufbereitet werden. Ein wichtiges Werkzeug ist das maschinelle Lernen: Es dient dazu, Muster in Daten zu erkennen oder Daten erst auf eine Weise zu segmentieren, die eine weitere Bearbeitung ermöglicht. Das Bundesforschungsministerium unterstützt mit einer Fördermaßnahme die Verbesserung der Qualifizierung im Bereich des Maschinellen Lernens durch die Verbindung von Algorithmenentwicklung für Forschungsfragen mit akademischer Ausbildung. Weiterhin sollen vier Kompetenzzentren in Berlin, Dortmund/St. Augustin, München und Tübingen für die praxisrelevante Anwendung von Maschinellem Lernen eingerichtet werden. Damit eine breite Anwendbarkeit gesteigert und neue disruptive Anwendungen und Technologien ermöglicht werden, werden zudem ab 2018 Forschungsvorhaben für die praxisrelevante Anwendung von Verfahren des Maschinellen Lernens gefördert.“
- Am 18. Juli 2018 hat das Bundeskabinett die Eckpunkte für eine Strategie Künstliche Intelligenz der Bundesregierung beschlossen. Die Kabinettsvorlage wurde gemeinsam vom Bundesministerium für Wirtschaft und Energie, vom Bundesministerium für Bildung und Forschung sowie vom Bundesministerium für Arbeit und Soziales eingebracht. Die Eckpunkte stellen die Grundlage für die Strategie Künstliche Intelligenz dar, die in den nächsten Monaten erarbeitet wird. Dazu wird die Bundesregierung in den nächsten Wochen einen Konsultationsprozess mit bundesweit arbeitenden Organisationen, Verbänden und Institutionen sowie Expertenworkshops und Fachforen durchführen. Auf dem Digitalgipfel am 3. und 4. Dezember 2018 soll die Strategie vorgestellt werden. Das Bundesministerium für Wirtschaft und Energie hat zeitgleich eine Studie „Potenziale der Künstlichen Intelligenz im produzierenden Gewerbe in Deutschland“ veröffentlicht (iit 2018), in der u. a. die Bedeutung des maschinellen Lernens hervorgehoben wird. Des Weiteren betonte in einem Interview mit dem SPIEGEL am 21. Juli 2018 („Chefsache“) Kanzleramtsminister Helge Braun die strategische, politische und ökonomische Bedeutung Künstlicher Intelligenz.

Die Strategie der Bundesregierung steht in einer Reihe mit den KI-Initiativen weiterer Länder, die in den letzten beiden Jahren vorgelegt wurden (siehe Abbildung 8).

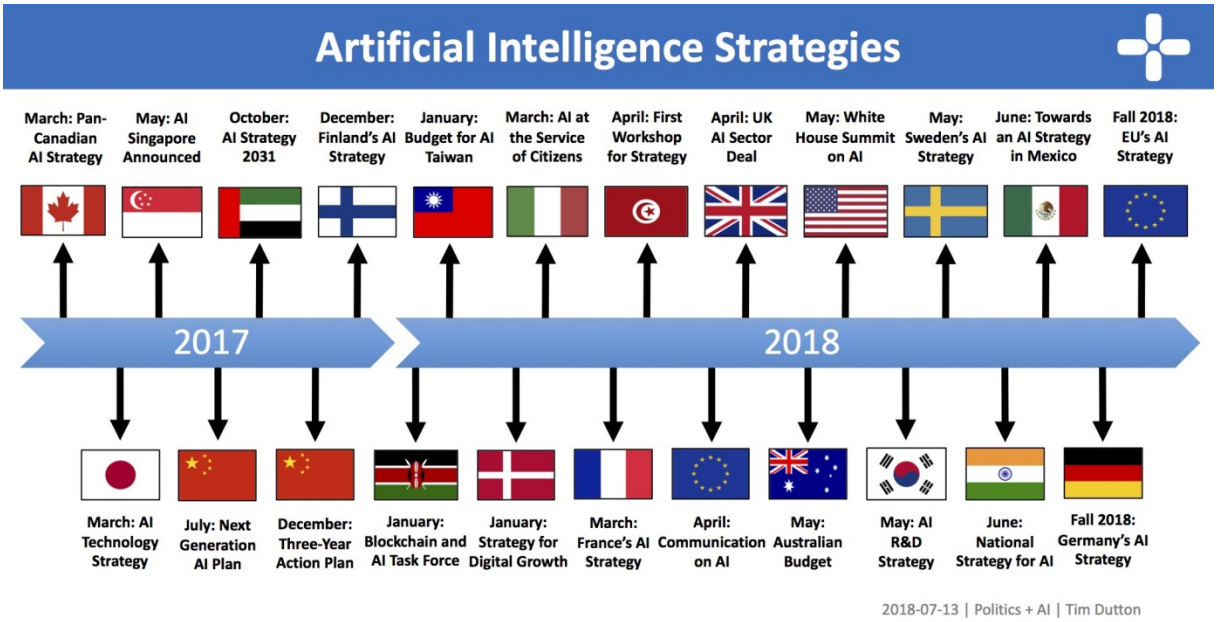


Abbildung 8: Aktivitäten wichtiger Staaten im Bereich KI[3]

Auch das Parlament hat sich inzwischen intensiv dem Thema Künstliche Intelligenz gewidmet und am 28. Juni 2018 eine Enquete-Kommission „Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche Potenziale“ eingesetzt. „Die Kommission hat den Auftrag, Handlungsempfehlung im Umgang mit Künstlicher Intelligenz (KI) zu formulieren. Sie soll unverzüglich eingesetzt werden und nach der parlamentarischen Sommerpause 2020 ihren Abschlussbericht mitsamt Handlungsempfehlungen vorlegen. Ihr gehören 19 Mitglieder des Bundestages sowie 19 Sachverständige an.“ (Deutscher Bundestag 2018d)

Darüber hinaus stellte die Abgeordnete Saskia Esken im Januar 2018 schriftliche Fragen an die Bundesregierung, inwieweit „algorithmen-basierte Entscheidungsprozesse, automatisierte Mustererkennung und künstliche Intelligenz“ eingesetzt werden (Deutscher Bundestag 2018a). Die Fraktion Bündnis 90/Die Grünen richtete am 4. April 2018 mit Bezug auf die Vereinbarungen im Koalitionsvertrag eine kleine Anfrage zum Thema „Konkrete Ziele und Vorhaben der Bundesregierung im Bereich Künstliche Intelligenz“ an die Bundesregierung (Deutscher Bundestag 2018b).

Insgesamt kann festgestellt werden, dass Machine Learning kein Hype ist, sondern wissenschaftlich auf einer soliden und sich dynamisch fortentwickelnden Basis aufsetzt und medial, gesellschaftlich, ökonomisch und politisch zunehmend an Bedeutung gewinnt. Die

[3] Dutton, T. (2018). An Overview of National AI Strategies, https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd, zuletzt abgerufen am 26.07.2018.

intensive Befassung mit dem Thema seitens des Statistischen Bundesamtes ist somit nur folgerichtig.

4 Informationsverbreitung zu Machine Learning

Teil des Auftrags des Proof of Concept Machine Learning ist es, die Mitarbeiterinnen und Mitarbeiter des Statistischen Bundesamtes über Methoden, Entwicklungen und mögliche Anwendungen zu informieren und so eine Basis für den sinnvollen und wirtschaftlichen Einsatz von Machine Learning in den Fachgruppen des Statistischen Bundesamtes zu schaffen. Hierbei wurden vom Projektteam mehrere Strategien verfolgt.

4.1 Informations- und Austauschplattform

Um interessierten Mitarbeiterinnen und Mitarbeitern des Statistischen Bundesamtes die Möglichkeit zu geben, relevante Informationen zu finden und auch bereitzustellen, wurde eine Informations- und Austauschplattform eingerichtet. Sie besteht aus einem TRAC-Forum, in dem Diskussionen zu bestimmten Fragestellungen möglich sind, und einem angegliederten Wiki, das für die Verbreitung von Informationen rund um das Thema Machine Learning allgemein und speziell zum Proof of Concept Machine Learning im Statistischen Bundesamt genutzt wird. Die Plattform findet sich unter der Adresse https://appweb.stba.testa-de.net/trac/pcml/. Zurzeit sind dort über 90 Nutzer zugelassen, die teils auf sehr hohem fachlichen Niveau Wissen, Anwendungen und Erfahrungen austauschen.

Das Werkzeug „TRAC“ wird in der amtlichen Statistik hauptsächlich zur Abwicklung von IT-Projekten verwendet und ist auf die Kommunikation zwischen Projektbeteiligten ausgerichtet (z. B. Auftraggeber, Programmierer, Testpersonen). Es kann jedoch auch leicht an andere Zwecke angepasst werden und ist wenig aufwendig in der Administration. Da aus anderen Statistiken (Insolvenzstatistik; Verdiensterhebung; Unternehmensregister) bereits positive Erfahrungen vorlagen und TRAC kurzfristig zur Verfügung stand, wurde im Rahmen des Proof of Concept dieses Kommunikationstool gewählt. Der technischen Basis der Austauschplattform fehlt es noch an einigen wünschenswerten Funktionen. So können zum Beispiel momentan wegen Hardwarebeschränkungen nur Dateien mit weniger als 1 MB Größe verwaltet werden.

4.2 Kurzveranstaltungen

Um Mitarbeiterinnen und Mitarbeitern aller Laufbahngruppen das Thema Machine Learning näher zu bringen und Methoden sowie mögliche Anwendungsfälle vorzustellen, die auch für die Arbeit im Statistischen Bundesamt relevant sein könnten, wurde im Rahmen des Proof of Concept Machine Learning zu Kurzveranstaltungen eingeladen. Die Resonanz war erheblich, so dass die ursprünglich vorgesehene eine Kurzveranstaltung nicht ausreichte. Um allen interessierten Mitarbeiterinnen und Mitarbeitern die Möglichkeit zu geben, die Veranstaltung zu besuchen, wurden insgesamt drei Termine angeboten, zwei in Wiesbaden (20. April und 29. Mai 2018) und

einer in Bonn (14. Juni 2018). Die Veranstaltungen wurden insgesamt von rund 250 Beschäftigten besucht. Außerdem wurde am 30. Mai 2018 eine gemeinsame Veranstaltung mit der Deutschen Bundesbank durchgeführt, an der rund 50 Personen teilnahmen. So wurde im Haus eine Wissensbasis geschaffen, auf der aufbauend eine Hausabfrage zu möglichen Einsatzgebieten von maschinellen Lernverfahren durchgeführt werden konnte (siehe Kapitel 6 Hausumfrage).

Die große Beteiligung an den Veranstaltungen und auch die Rückmeldungen von Teilnehmerinnen und Teilnehmern im Nachgang zeigen ein hohes Interesse am Thema und punktuell auch erhebliches Vorwissen. Dieses Potential im Haus sollte zeitnah genutzt werden, um Projekte mit maschinellen Lernverfahren anzustoßen.

4.3 Digitalisierung hautnah

Im März/April 2018 wurde die Veranstaltungsreihe „Digitalisierung hautnah“ durchgeführt, in der B2 und C2 gemeinsam über die bevorstehende Digitalisierung informierten. Eines der Themen war „Maschinelles Lernen“. In 35 Veranstaltungen wurden über 1.300 Kolleginnen und Kollegen erreicht.

4.4 Andere Informationskanäle

Bereits im Vorfeld und darüber hinaus zeitlich parallel zur Durchführung des Proof of Concept Machine Learning wurden einer Reihe weiterer Informationskanäle genutzt, um Stakeholder und Mitarbeiterinnen und Mitarbeiter über das Thema „Maschinelles Lernen“ zu informieren:

Datum	Anlass
20.09.2017	Vortrag auf der Statistischen Woche „Einsatz von Machine-Learning-Verfahren in der amtlichen Unternehmensstatistik“
27./30.11.2017	Marktplatz Leitungsklausur
1.12.2017	Vortrag „Einsatz von Machine-Learning-Verfahren in der amtlichen Statistik“ in der Bund-Länder-Arbeitsgruppe „Neue digitale Daten“
11.12.2017	Vortrag „Einsatz von Machine-Learning-Verfahren in der amtlichen Statistik“ in der Sitzung des Hausnetzwerkes „Neue digitale Daten“
7.02.2018	Vortrag „Einsatz von Machine-Learning-Verfahren in der amtlichen Unternehmensstatistik“ in der Kurzveranstaltung zur Statistischen Woche
26.04.2018	Vortrag „Machine Learning“ anlässlich des gemeinsamen Treffens von Statistik Austria und dem Statistischen Bundesamt zum Thema „Neue digitale Daten für amtliche Statistiken“
14.05.2018	TOP 5 „Künstliche Intelligenz zur automatisierten Plausibilisierung in den Verdienststatistiken“ des Statistischen Beirates sowie Vergabe des Innovationspreises
11.07.2018	Information des Gesamtpersonalrates

Übersicht 1: Informationsverbreitung zu Machine Learning

Weitere Veranstaltungen sind vorgesehen (Vortrag auf der Statistischen Woche vom 11.–14. September 2018; Gruppen- und Referatsleitungsforum am 21. September 2018; Vortrag auf der CESS vom 18.–19. Oktober 2018; 27. Wissenschaftliches Kolloquium „Mehr Zahlen, bessere Entscheidungen? Neue digitale Daten und Methoden in der empirischen Analyse und Beratung“ am 22./23. November 2018; Exzellenz (Digital) Show für Ressorts am 5.12. in Berlin).

5 Abfrage bei Statistikinstitutionen

5.1 Vorgehensweise

Um ein Bild über aktuell verfolgte Anwendungsfälle für maschinelles Lernen in nationalen und internationalen statistikproduzierenden Institutionen zu erhalten, wurde eine Abfrage durchgeführt. Ziel war es, Informationen über angewendete Methoden und relevante fachstatistische Fragestellungen zu gewinnen, aus denen für das weitere Vorgehen im Statistischen Bundesamt gelernt werden kann. In Deutschland wurden am 13. März 2018 die 14 Statistischen Landesämter sowie am 23. März 2018 18 weitere Statistikproduzenten (überwiegend sogenannte ONAs – Other National Authorities, siehe Übersicht 2) mittels eines strukturierten Fragenkatalogs (Excel-Datei) angeschrieben. Die Adressaten wurden darauf hingewiesen, dass das Statistische Bundesamt einen Proof of Concept Machine Learning durchführt und in diesem Kontext Umfragen bei nationalen und internationalen Statistikinstitutionen durchführt, um die Einsatzmöglichkeiten von maschinellem Lernen in der amtlichen Statistik einschätzen zu können. Für jedes Projekt, in dem Machine-Learning-Verfahren eingesetzt werden, wurden Informationen zu folgenden Stichworten erbeten:

Institution:	Name oder Abkürzung der statistikproduzierenden Organisation
Projektbezeichnung:	Bezeichnung des Machine-Learning-Projekts
Beschreibung:	Kurze Beschreibung des Machine-Learning-Projekts (z.B. Ziel, Herangehensweise, Methoden)
Anwendung:	Was soll mit dem Machine-Learning-Verfahren erreicht werden, z.B. Klassifikation, Regression, Clustering, etc.?
Status:	Status des Projekts - Produktiv - Experiment - Test - Idee
Methode:	Welche ML-Methode wurde eingesetzt: - Support Vector Machine (SVM) - Entscheidungsbäume - Random Forest - Neuronale Netze
Software:	Welche Software wird eingesetzt (z.B. R, Python, etc.)?
Quelle (Link):	Gibt es bereits schriftliche Quellen?

Alle Institutionen haben auf die Anfrage geantwortet.

Institution	Rückmeldung	Fehlanzeige
Bundesagentur für Arbeit		x
Bundesamt für Migration und Flüchtlinge (BAMF)	x	
Bundesamt für Verbraucherschutz und Lebensmittelsicherheit (BVL)		x
Bundesamt für Wirtschaft und Ausfuhrkontrolle (BAFA)		x
Bundesanstalt für Landwirtschaft und Ernährung (BLE)		x
Deutsche Bundesbank	x	
FDZ Deutsche Rentenversicherung		x
FDZ Institut zur Zukunft der Arbeit		x
FDZ SOEP		x
GESIS	x	
Institut für Arbeitsmarkt- und Berufsforschung (IAB)	x	
Julius Kühn-Institut		x
Kraftfahrtbundesamt		x
Robert-Koch-Institut	x	
Stifterverband|Wissenschaftsstatistik		x
Thünen-Institut		x
Umweltbundesamt		x
Zentrum für Europäische Wirtschaftsforschung (ZEW)	x	

Übersicht 2: Umfrage bei nationalen Statistikinstitutionen

In gleicher Weise wurden die Statistikämter der 27 EU-Mitgliedstaaten, der vier EFTA-Staaten, von sechs ausgewählten außereuropäischen Staaten sowie Eurostat und die OECD angeschrieben. Der Rücklauf gestaltete sich sehr schwierig. Trotz dreier Erinnerungen beteiligten sich (Stand: 23. Juli 2018) sieben Ämter leider nicht an der Umfrage. 19 meldeten Machine-Learning-Projekte (zwei angekündigte Rückmeldungen stehen noch aus) und elf erstatteten Fehlanzeige (siehe Übersicht 3).

Land	Rückmeldung	Fehlanzeige	Rückmeldung angekündigt	Keine Rückmeldung
Belgien	x			
Bulgarien				x
Dänemark	x			
Estland		x		
Finnland	x			
Frankreich				x
Griechenland				x
Irland		x		
Italien				x
Kroatien		x		
Lettland	x			
Litauen		x		
Luxemburg	x			
Malta				x
Niederlande	x			
Österreich	x			
Polen		x		
Portugal	x			
Rumänien	x			
Schweden	x			
Slowakei		x		
Slowenien		x		
Spanien	x			
Tschechien		x		
UK				x
Ungarn		x		
Zypern		x		
Eurostat				x
Island	x			
Liechtenstein		x		
Norwegen	x			
Schweiz	x			
Australien	x			
Israel			x	
Japan	x			
Kanada	x			
Neuseeland	x			
OECD			x	
USA	x			

Übersicht 3: Umfrage bei internationalen Statistikinstitutionen

Als weitere Datenquelle wurde die Machine Learning Documentation Initiative herangezogen, die aktuell von Valentin Todorov (UNIDO) fortgeführt wird (Chu und Poirier 2015). Außerdem wurde

versucht, für die Länder, die auf die Umfrage nicht geantwortet haben und auch in der Machine Learning Documentation Initiative nicht aufgeführt werden, Informationen über eine Recherche auf deren Website zu gewinnen. Trotz aller Bemühungen gelang es nicht, Angaben für Bulgarien, Frankreich, Griechenland und Malta zu gewinnen. Gleichwohlliefern die zusammengetragenen Informationen ein breites Bild über den internationalen Einsatz von Machine-Learning-Methoden in der amtlichen Statistik.

5.2 Ergebnisse der Statistischen Ämter der Länder

Im Rahmen der Abfrage wurden alle 14 Statistischen Ämter der Länder angeschrieben. 13 meldeten Fehlanzeige. Das Hessische Statistische Landesamt testet als einziges ein Verfahren, bei dem mittels Webscraping gewonnene Daten zu den Unternehmen des Unternehmensregisters (URS) u. a. durch Methoden des maschinellen Lernens ausgewertet werden, um statistikübergreifende Kohärenzprüfungen durchzuführen sowie neue Merkmale zu generieren.

Als Fazit lässt sich festhalten, dass das Statistische Bundesamt nicht von den Erfahrungen in den Statistischen Landesämtern profitieren kann.

5.3 Ergebnisse anderer nationaler Institutionen

Auf die Abfrage bei den nationalen Institutionen gab es positive Rückmeldungen von dem Bundesamt für Migration und Flüchtlinge (BAMF), der Deutschen Bundesbank, dem GESIS – Leibniz-Institut für Sozialwissenschaften, dem Institut für Arbeitsmarkt- und Berufsforschung (IAB), dem Robert-Koch-Institut (RKI) und dem Zentrum für europäische Wirtschaftsforschung (ZEW). Die Antwortenden meldeten insgesamt 36 Projekte, in denen Maschine-Learning-Verfahren eingesetzt werden. Fünf dieser Anwendungen sind im Produktivbetrieb. Dabei handelt es sich um Verfahren zur Imputation fehlender Information zur Arbeitszeit, zur Vorhersage der Arbeitslosigkeitsdauer und zur Typisierung von Arbeitsmarktregionen, die beim Institut für Arbeitsmarkt- und Berufsforschung (IAB) durchgeführt werden. Das Robert-Koch-Institut nutzt maschinelles Lernen produktiv für die automatische Erkennung von Ausbruchsereignissen bei Infektionskrankheiten und für die Analyse molekularer Daten.

Bei weiteren 21 Projekten handelt es sich um laufende Forschungsprojekte im engeren Sinne. Zehn weitere Projekte sind Machbarkeitsstudien und Entwicklungen von Prototypen, die Potenzial für den Produktivbetrieb haben.

Institut	Anzahl der Anwendungen
GESIS	16
IAB	8
Deutsche Bundesbank	5
RKI	4
ZEW	2
BAMF	1
Insgesamt	36

Tabelle 1: Machine-Learning-Anwendungen nach Institution

Verwendete Machine-Learning-Methoden (Mehrfachnennungen möglich)	Anzahl
Entscheidungsbaummethoden[4]	13
Random Forest	13
Neuronale Netze	11
SVM	10
Sonstige	12
Insgesamt	59

Tabelle 2: Verwendete Machine-Learning-Methoden

Art der Anwendung (Mehrfachnennungen möglich)	Anzahl
Klassifikation	19
Identifikation	11
Clustering	6
Text-Analyse	9
Regression	4
Sonstige	7
Insgesamt	56

Tabelle 3: Art der Anwendung

In den Projekten wurden Machine-Learning-Verfahren am häufigsten zur Identifikation und Klassifikation von Einheiten eingesetzt. Ein weiteres wichtiges Einsatzgebiet ist die Regression. Als Methoden werden neben Random Forests andere entscheidungsbaumbasierte Verfahren wie etwa Gradient Boosting am häufigsten eingesetzt. Auch Neuronale Netze und Support Vector Machines werden häufig verwendet.

Obwohl die antwortenden Institutionen sehr unterschiedliche Fragestellungen verfolgen und mithin sehr unterschiedliche Projektziele haben, kann doch festgehalten werden, dass häufig ähnliche Probleme mit Machine-Learning-Verfahren angegangen werden. Auch die eingesetzten Methoden sind häufig ähnlich.

Die ausführlichen Ergebnisse sind im Anhang 10.2.2 dargestellt.

5.4 Ergebnisse internationaler Institutionen

Unter den befragten bzw. recherchierten internationalen, statistikproduzierenden Institutionen waren überwiegend nationale Statistikbehörden. Es zeigt sich, dass diese in unterschiedlichem Ausmaß Projekte mit Machine-Learning-Verfahren betreiben. Statistics Canada meldete bei

[4] Die Anzahl 13 bei Entscheidungsbaummethoden und Random Forest ist zufällig identisch. Die Entscheidungsbaummethoden sind immer explizit nicht Random Forest.

weitem die meisten Projekte dieser Art. Auch das Australian Bureau of Statistics, Stats NZ, das Bundesamt für Statistik der Schweiz sowie Institutionen in den Vereinigten Staaten betreiben relativ viele Projekte.

Institution	Anzahl der Projekte
Statistics Canada	36
U.S. Bureau of Labor Statistics	11
Stats NZ	9
U.S. Department of Agriculture NASS	7
Australian Bureau of Statistics	6
Federal Statistical Office of Switzerland	6
National Institute of Statistics Romania	4
Statistics Austria	4
Statistics Netherlands	4
Statistics Portugal	4
Statistics Spain (INE)	3
Statistics Sweden	3
Eurostat	2
STATEC (Luxemburg)	2
Statistics Finland	2
Statistics Iceland	2
Statistics Poland	2
Bureau of Economic Analysis (USA)	1
Central Statistical Bureau of Latvia	1
Central Statistics Office of Ireland	1
Hungarian Central Statistical Office	1
Italian National Institute of Statistics	1
National Statistics Center, Japan	1
OECD	1
ONS (UK)	1
Statistics Belgium	1
Statistics Denmark	1
Statistics Norway	1
U.S. Census Bureau	1
Insgesamt	119

Tabelle 4: Umfrage und Recherche bei internationalen Statistikorganisationen, Anzahl der Projekte nach Institution

Von besonderem Interesse ist die Frage nach den Problemen, die mit Machine-Learning-Verfahren gelöst werden sollen, und den Statistikbereichen, in denen sie eingesetzt werden. Es zeigt sich, dass der Großteil der Projekte statistikübergreifend ist, also für mehrere Statistiken einsetzbare Verfahren liefert oder liefern soll. Häufig genannte Statistikbereiche sind die Haushalts- und Unternehmensstatistiken.

Machine-Learning-Verfahren werden dabei häufig zur Klassifikation, Identifikation und Imputation verwendet. Die Identifikation von Einheiten wird hierbei oft im Zusammenhang mit Mikrodatenverknüpfung genannt. Bei der Bewertung der einzelnen Projekte ist auch der Status relevant. 18 der genannten Projekte sind bereits im Produktiveinsatz, weitere 25 sind in der

Entwicklung hin zum Produktiveinsatz. 51 Projekte sind im Experimentierstadium und weitere 25 sind zurzeit als Idee formuliert.

Statistik	Anzahl der Anwendungen
Statistikübergreifend	24
Arbeitsmarkt	14
Haushaltsstatistik	13
Landwirtschaftsstatistik	10
Unternehmensstatistik	8
Zensus	8
WZ-Klassifikation	6
Preisstatistik	5
Verkehrsstatistik	3
Weitere	28
Insgesamt	119

Tabelle 5: Machine-Learning-Anwendungen nach Fachstatistik

Art der Anwendung (Mehrfachnennungen möglich)	Anzahl
Klassifikation	66
Imputation	20
Mikroverknüpfung	15
Clustering	9
Text-Analyse	6
Regression	5
Identifikation	4
Dimensionsreduktion	2
Sonstige	15
Insgesamt	142

Tabelle 6: Art der Anwendungen

Projektstatus	Anzahl der Anwendungen
Idee	25
Experiment	51
In Entwicklung	25
Produktiv	18
Insgesamt	119

Tabelle 7: Anzahl der Anwendungen nach Projektstatus

Verwendete Machine-Learning-Methoden (Mehrfachnennungen möglich)	Anzahl
Random Forest	28
Neuronale Netze	20
SVM	18
Entscheidungsbaummethoden	14
Nearest-Neighbour-Ansätze	11
Bayes-Ansätze	6
Natural Language Processing	4
Clusterverfahren	2
Sonstige	43
Insgesamt	146

Tabelle 8: Verwendete Machine-Learning-Methoden

Unter den am häufigsten genannten Machine-Learning-Methoden finden sich Random Forests, Verfahren, die Neuronale Netze nutzen, Support Vector Machines sowie weitere entscheidungsbaumbasierte Verfahren.

Zusammenfassend lässt sich sagen, dass in den internationalen Statistischen Ämtern in vielen Statistikbereichen Machine Learning eingesetzt oder getestet wird. In der Regel werden Klassifikations-, Identifikations- oder Imputationsaufgaben angegangen. Es werden häufig entscheidungsbaumbasierte Methoden, Neuronale Netze oder Support Vector Machines eingesetzt.

Die ausführlichen Ergebnisse sind im Anhang 10.2.3 dargestellt.

5.5 Gesamtschau

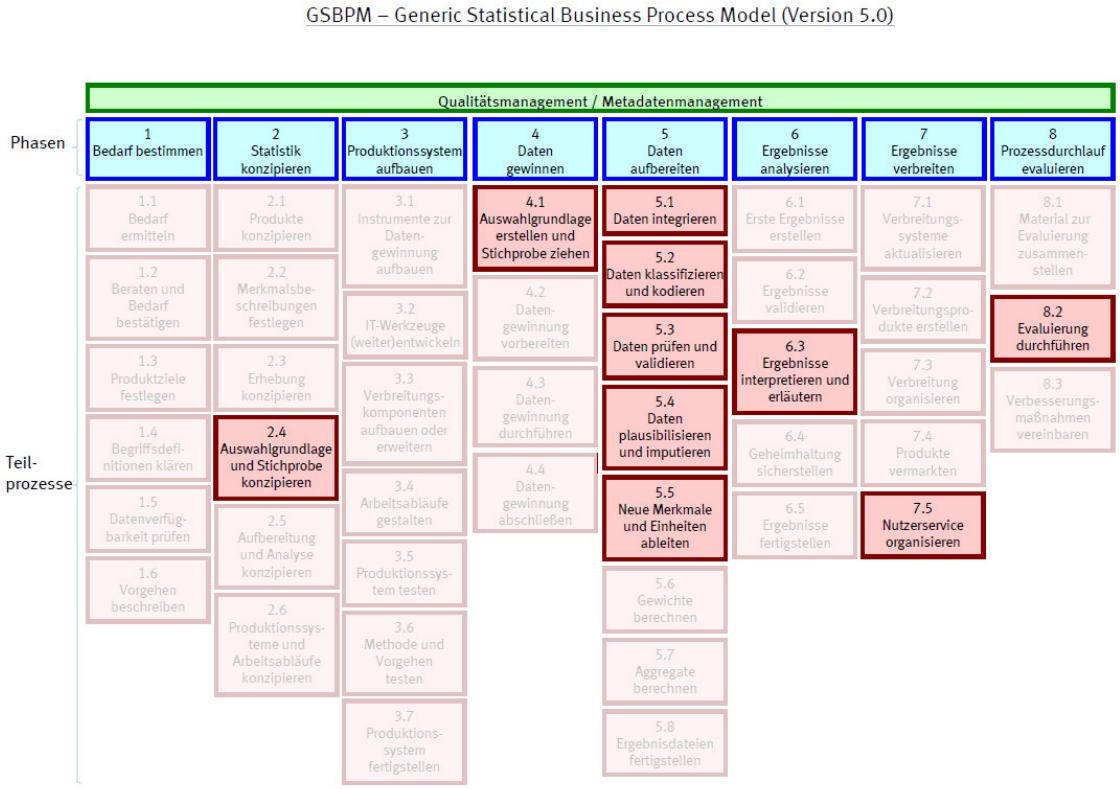


Abbildung 9: Teilprozesse des GSBPM mit Machine-Learning-Projekten aus den Abfragen bei nationalen und internationalen Institutionen

Da die einbezogenen nationalen und internationalen Institutionen Statistik produzieren, lassen sich die genannten Projekte den Phasen des GSBPM zuordnen. Abbildung 9 zeigt die Teilprozesse des GSBPM, die durch die genannten Projekte unterstützt werden. Es ist zu erkennen, dass Machine-Learning-Verfahren vor allem bei der Datengewinnung, Datenaufbereitung und der Ergebnisanalyse eingesetzt werden. Weiterhin wurden Projekte in der Statistikkonzeption, der Organisation des Nutzerservice und in der Evaluation genannt. Im Zusammenhang mit dem GSBPM listet das Papier „Machine-Learning in Surveys Steps“ von Statistics Canada zusätzliche Anwendungsfälle auf, bei denen Machine-Learning-Verfahren eingesetzt werden könn(t)en (Statistics Canada 2018).

Eine ausführliche, an dem GSPBM orientierte Darstellung der Ergebnisse, die für die hausinternen Kurzveranstaltungen erstellt wurde, findet sich in Anhang 10.2.1.

Abbildung 10: Wortwolke auf Basis der Projektbeschreibungen der antwortenden Institutionen

Aus den Abfragen bei den Statistischen Ämtern der Länder, den nationalen und internationalen statistikproduzierenden Institutionen lassen sich folgende Schlüsse ziehen: Projekte, die maschinell Merkmale klassifizieren, Werte imputieren und Einheiten identifizieren und dazu Random Forests oder ähnliche baumbasierte Verfahren, Support Vector Machines oder Neuronale Netze einsetzen, sind zurzeit weit verbreitet. Anwendungsfälle können in fast allen Fachstatistiken gefunden werden. In der Regel werden Prozesse in der Statistikkonzeption, der Datengewinnung und -analyse sowie der Statistikverbreitung und -evaluation mit Machine-Learning-Verfahren unterstützt.

6 Hausumfrage

6.1 Vorgehensweise

Um mögliche Einsatzgebiete für Machine-Learning-Verfahren in den Fachstatistiken zu identifizieren, wurde bei den Gruppen des Statistischen Bundesamtes eine Hausabfrage durchgeführt. Für die Abfrage wurden am 9. Mai 2018 alle 29 Gruppen angeschrieben[5], die auch geantwortet haben. Amtsleitung, Leitungsstab, Abteilungs- und Referatsleitungen erhielten die entsprechende Mail in Kopie, so dass sie sich ggf. in die Beantwortung einbringen konnten. Als Hintergrundinformation wurde den Gruppenleitungen eine Reihe von Dokumenten zur Verfügung gestellt:

a) Eine konsolidierte Zusammenstellung der Ergebnisse der Umfrage bei nationalen und internationalen Statistikinstitutionen zum Einsatz von Machine-Learning-Verfahren. Diese können und sollen auch als Anregung zur Identifikation von Einsatzmöglichkeiten im Statistischen Bundesamt dienen. Die Darstellung orientiert sich an den Phasen des Generic Statistical Business Process Model (GSBPM).

b) Das Papier „Machine-Learning in Surveys Steps“ von Statistics Canada, das anhand des GSBPM Anwendungsfälle auflistet, bei denen Machine-Learning-Verfahren eingesetzt werden können.

c) Ein einfaches Prüfschema, ob Machine-Learning-Verfahren in Frage kommen (siehe Anhang 10.3).

d) Die für die Kurzveranstaltung zum Machine Learning am 20. April 2018 erstellte Präsentation.

e) Einen Aufsatz, in dem die Projekte von E1 zusammenfassend dargestellt sind.

Außerdem wurde auf weitere Informationen zum Thema Machine Learning auf der hausinternen Informations- und Austauschplattform PCML TRAC hingewiesen.

[5] Die Organisationseinheiten F-REB, B1-Recht, B1-Int, H1-Soziales, H1-Gesundheit und Wahlen werden zu den Gruppen gezählt.

Der Abfrage lag ein strukturierter Fragenkatalog zugrunde (Excel-Datei). Für den Fall, dass in der jeweiligen Gruppe Machine-Learning-Verfahren eingesetzt werden oder dies geplant ist, sollten nach Möglichkeit Angaben zu folgenden Sachverhalten gemacht werden:

1. EVAS-Nr. (sofern vorhanden)
2. Statistik- oder Projektbezeichnung
3. Organisationseinheit/Projektteam
4. Ansprechperson
5. Inhaltliche Kurzbeschreibung des Projektes
6. Erwarteter Nutzen (z. B. Zeitgewinn bei der Aufgabenerledigung; Aktualitätsgewinn; Effizienzsteigerung; Qualitätsverbesserung; Personaleinsparung; bessere Analysemöglichkeiten; Chance, neue Aufgaben zu erledigen; ...)
7. Status [erste Idee; Konzeptionsphase; Experiment; Test; Echtbetrieb; Projekt verworfen]
8. Gibt es bereits eine schriftliche Dokumentation des Projektes o.ä.? Falls ja: Wo ist sie einsehbar? (z. B. Intranet, auf Anfrage, beigefügt, veröffentlicht in ...)
9. Wird Unterstützung benötigt (z. B. durch C1 oder seitens der Wissenschaft)?
10. Soll das Projekt in Kooperation mit Dritten (gleichrangige Partner) durchgeführt werden?
11. Steht ausreichend Personal zu Verfügung (quantitativ und qualitativ)?
12. (Geplanter) Projektbeginn
13. (Geplantes) Projektende
14. Geplante Methodik (z. B. SVM, Random Forest, ...)
15. Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
16. Hardware (Auf welchen Rechnern wird die Anwendung laufen?)

Andernfalls waren folgende Fragen zu beantworten:

1. Warum ist kein Einsatz von Machine-Learning-Verfahren in Ihrer Gruppe geplant? [Mögliche Gründe: Keine geeignete Anwendungsmöglichkeit; fehlende Expertise; keine Zeit für konzeptionelle Arbeiten; ...]
2. Wurde das beigefügte Prüfschema angewandt?
3. Liegen Ihnen hinreichende Informationen zu den Anwendungsmöglichkeiten für Machine-Learning-Verfahren vor? (Falls nein: Welche weiteren Informationen benötigen Sie?)

6.2 Ergebnisse

Aus den abgefragten Gruppen wurden 16 Fehlanzeigen gemeldet, dies in der Regel mit dem Hinweis, dass derzeit keine geeigneten Anwendungsmöglichkeiten ersichtlich sind. Fünfmal wurde auch fehlende Expertise bzw. fehlende Zeit zum Aufbau von Expertise bzw. zur Umsetzung von Projekten als Grund genannt. Ein Informationsdefizit, was Machine Learning allgemein angeht, wurde jedoch kaum als Begründung genannt. Vielmehr wurden explizit die Informationsveranstaltungen im Haus als sehr hilfreich beschrieben. Es wurde auch angeregt, weiterhin über neue Entwicklungen auf dem Gebiet Machine Learning zu informieren. Insgesamt

wird auch bei Gruppen, die keine Projekte oder Projektideen meldeten, grundsätzlich Potenzial für entsprechende Anwendungen gesehen.

Die verbleibenden 13 Gruppen meldeten 31 laufende Projekte oder Projektideen. Bei 25 der Meldungen handelt es sich um Ideen, wie Machine-Learning-Verfahren eingesetzt werden könnten. Sechs Projekte sind bereits in der Experimentier- bzw. Testphase. Da viele Vorschläge erste Ideen sind, kann noch keine abschließende Aussage über verwendete Verfahren gemacht werden. Fast alle Gruppen, die Projektideen meldeten, gaben auch an, dass sie (gruppen-) externe Expertise bei der eventuellen Umsetzung benötigen werden.

Die Projektideen zielen oft auf die maschinelle Klassifikation von Merkmalen oder die Identifikation von Einheiten (Dubletten, Ausreißer). Dies sind auch die Anwendungen, die bei den Rückmeldungen aus den nationalen und internationalen statistikproduzierenden Institutionen sehr häufig genannt wurden.

Zusammenfassend kann festgehalten werden, dass es viele vielversprechende Ansätze und Ideen für den Einsatz von Machine Learning im Statistischen Bundesamt gibt. Im Moment scheint es aber einen Engpass beim Aufbau bzw. der Bereitstellung von Expertise in diesem Feld zu geben.

Die ausführlichen Ergebnisse der Hausumfrage sind im Anhang 10.2.4 dargestellt.

6.3 Bereits realisierte Vorhaben

Neben den mit der Hausumfrage eruierten geplanten Maßnahmen gibt es mehrere Projekte, die bereits umgesetzt sind oder derzeit durchgeführt werden. Es handelt sich um fünf Projekte in E1 und eines in D306, die nachfolgend näher beschrieben werden. Die Maßnahmen von E1 sind in der folgenden Übersicht vorab zusammengefasst:

Statistik	Problem	Methode	Stand	Ergebnis
Unternehmensregister	Zuordnung von Unternehmen zum 3. Sektor	Support Vector Machine	abge-schlossen	+
Handwerkszählung	Erkennung nicht relevanter Unternehmen in den Handwerksstatistiken	Random Forest Support Vector Machine	abge-schlossen	++
Verdienst-strukturerhebung	Schätzung einer Erwerbsunterbrechung von Frauen in der Verdienststrukturerhebung	Support Vector Machine	abge-schlossen	+/–
Verdienst-strukturerhebung	Schätzung der Staatsbürgerschaft von Beschäftigten in der Verdienststrukturerhebung	Support Vector Machine (u. a.)	abge-schlossen	(–)
Verdienst-strukturerhebung	Anreicherung der Integrierten Erwerbsbiografien (IEB) von BA/IAB um Informationen zur Mindestlohnbetroffenheit aus der Verdienststrukturerhebung	Random Forest	laufend	+/–

Übersicht 4: Anwendung von Machine Learning in den Unternehmensstatistiken

6.3.1 Sektorkennzeichnung im Unternehmensregister

Bereits im Jahr 2013 stellte die Gruppe E1 erste Überlegungen und Tests an, maschinelle Lernverfahren im Bereich der (Dritt-)Sektorkennzeichnung zu erproben. Die amtliche Statistik hatte zu diesem Zeitpunkt noch keine Erfahrungen mit maschinellem Lernen gesammelt und musste sich hier Neuland erschließen. Mit Herrn Dumpert, wissenschaftlicher Mitarbeiter am Lehrstuhl für Stochastik der Universität Bayreuth, konnte ein Experte gewonnen werden, der im Bereich maschineller statistischer Lernverfahren, insbesondere Support Vector Machines, forscht. Er beriet und unterstützte den Fachbereich Unternehmensregister im Rahmen eines Werkvertrages maßgeblich bei der Durchführung des Projektes zur (Dritt-)Sektorkennzeichnung (und auch bei weiteren Projekten sowie bei der Durchführung des Proof of Concept).

Ziel dieses Projektes war es, a priori nicht eindeutig zuzuordnende Unternehmen im statistischen Unternehmensregister (URS) hinsichtlich ihrer Zugehörigkeit zum sogenannten Dritten Sektor (auch Non-Profit-Sektor genannt) zu klassifizieren – idealerweise derart, dass das Verfahren auch

in späteren Jahren einsetzbar ist und somit auf aufwändige Einzelfallrecherchen verzichtet werden kann. Als Klassifikationsmethode sollten Support Vector Machines erprobt werden. Die Zuordnung von Unternehmen zum Dritten Sektor ermöglicht die Bestimmung der Anzahl der Unternehmen und der darin sozialversicherungspflichtig Beschäftigten in diesem Bereich sowie die zugehörige Bruttowertschöpfung. Ein großer Teil der im URS erfassten Unternehmen kann aufgrund fachstatistischer Überlegungen und mithilfe eines daraus resultierenden regelbasierten Algorithmus automatisiert als dem Dritten Sektor zugehörig oder nicht zugehörig klassifiziert werden. Für ca. 45.000 Unternehmen, für die der Algorithmus keine eindeutige Zuordnung zuließ, waren für das Berichtsjahr 2007 hingegen zeit- und personalintensive Einzelfallrecherchen notwendig. A priori nicht eindeutig zuzuordnende Unternehmen sind demnach solche, für welche der Standard-Algorithmus keine eindeutige Klassifikation ermöglicht.

Aufgrund der oben genannten Einzelfallrecherchen standen für ca. 45.000 a priori nicht eindeutig zuzuordnende Unternehmen sowohl die üblichen Merkmale des URS als auch als verlässlich eingestufte Klassifizierungen bezüglich der Zugehörigkeit zum Dritten Sektor zur Verfügung. Diese Situation stellte hinsichtlich des Umfangs der Daten als auch hinsichtlich der Vollständigkeit der Informationen eine sehr gute Ausgangslage für statistisches maschinelles Lernen und die Anwendung von Support Vector Machines dar. Zu klären war, ob eine SVM in der Lage wäre, weitere, über die im Standard-Algorithmus bereits abgebildeten Regeln hinausgehende Strukturen in den Trainings- bzw. Testdaten aufzufinden (vgl. Dumpert et al. 2016; Dumpert und Beck 2017). Die Ergebnisse, die hierbei erzielt werden konnten, wurden in Bezug auf die vorliegende Datenbasis als erfolgreich gewertet.

Beim Einsatz der SVM zur Abgrenzung des Dritten Sektors im Echtbetrieb mit Daten des Berichtsjahres 2014 kam es zu einer unerwartet hohen Missklassifikationsrate bei der Zuordnung von Unternehmen zum Dritten Sektor, während die Zuordnung zum „Nicht-Drittsektor“ zufriedenstellende Ergebnisse lieferte. Die durch nachgelagerte Einzelfallrecherchen aufgedeckte Missklassifikation betraf insbesondere Unternehmen im Bereich der Gastronomie. Der vorgelagerte Standard-Algorithmus konnte aufgrund der Erkenntnisse, die mit der Anwendung der SVM gewonnen wurden, verbessert werden.

Weitere Untersuchungen im Nachgang zum SVM-Einsatz für Berichtsjahr 2015 zeigten, dass eine Dritt-Sektor-Zuordnung für die Gruppe der natürlichen Personen nicht sinnvoll ist. Auch diese wurden aus der SVM-Klassifikation herausgenommen und über den regelbasierten Algorithmus zugeordnet. Hierdurch konnte die Missklassifikationsrate deutlich gesenkt werden.

Nachdem die SVM zur Dritt-Sektor-Klassifikation über drei Berichtsjahre hinweg im Echtbetrieb eingesetzt, analysiert und die von ihr zu klassifizierende Datenbasis optimiert wurde, lässt sich festhalten, dass sie als Instrument zur Qualitätsverbesserung im Arbeitsprozess fest installiert

ist. Sie ermöglicht eine relativ sichere Aussteuerung von nicht dem Dritten Sektor zuzuordnenden Einheiten. Einheiten, die die SVM dem Dritten Sektor zuordnet, werden zur Qualitätssicherung durch Bearbeiter manuell überprüft. Ein vollständiger Verzicht auf Einzelfallrecherchen ist nicht möglich, die Fallzahl der zu überprüfenden Einheiten konnte jedoch deutlich reduziert werden und ist mit den vorhandenen Ressourcen bewältigbar.

6.3.2 Kennzeichnung nicht relevanter Handwerksunternehmen

Die Handwerksstatistiken werden aktuell vollständig aus Verwaltungsdaten ermittelt. Für die Aufbereitung der Statistiken über das Handwerk ist das statistische Unternehmensregister von zentraler Bedeutung. Dort werden einmal jährlich Lieferungen der Handwerkskammern verarbeitet. Die Handwerkskammern liefern für diesen Zweck den handwerklichen Gewerbezweig und Hilfsmerkmale, wie Gewerbesteuernummer und Adressinformationen, anhand derer die Unternehmen identifiziert werden können. Nach der Verarbeitung stehen die Handwerksinformationen im URS für statistische Auswertungen zur Verfügung.

Es sind jedoch nicht alle Unternehmen, die in den Lieferungen der Handwerkskammern enthalten sind, relevant für die Handwerksstatistiken. Laut § 2 Handwerkstatistikgesetz (HwStatG) sind in den Handwerksstatistiken nur selbstständige Handwerksunternehmen zu erfassen. Daneben gibt es aber eine Gruppe von Unternehmen, die selbst nicht Handwerker sind, aber handwerkliche innerbetriebliche Abteilungen oder handwerkliche Nebenbetriebe unterhalten. Dies sind zum Beispiel Speditionen, die eigene Kfz-Werkstätten haben, Supermärkte mit Fleischer- oder Bäckertheken oder Energieversorgungsunternehmen, die Lehrwerkstätten für bestimmte Handwerksberufe betreiben. Diese Unternehmen müssen identifiziert werden, weil sie bei der Aufbereitung der Handwerksstatistiken nicht einbezogen werden sollen.

Die Handwerkskammern sind anhand der ihnen vorliegenden Informationen nicht in der Lage, diese Unternehmen zu identifizieren. Diese Aufgabe wird deswegen jährlich von den Fachbereichen Handwerk der Statistischen Ämter der Länder übernommen. Da die Arbeiten zur manuellen Klassifizierung in erheblichem Maße personelle Ressourcen binden, stellte sich die Frage, ob Unternehmen mit ausreichender Genauigkeit maschinell klassifiziert werden können (vgl. Feuerhake und Dumpert 2016; Dumpert und Beck 2017). Für die Klassifikation werden zurzeit maschinelle Lernalgorithmen eingesetzt, um mit relativ geringem personellen Aufwand Unternehmen zu klassifizieren. Dabei werden Support Vector Machines in Verbindung mit Random Forests eingesetzt.

Das Training der Verfahren erfolgt auf Basis von Informationen aus dem Unternehmensregister. Es werden rund 600.000 Einheiten verwendet. Nach einigen Schritten zur Dimensionsreduktion wird eine Support Vector Machine auf einen Datensatz mit ca. 80.000 Einheiten mit 30 Merkmalen angepasst.

Mit Hilfe des trainierten Modells werden nun schon in zwei aufeinanderfolgenden Jahren die Arbeiten zur Kennzeichnung der Handwerksrelevanz in den Statistischen Ämtern der Länder unterstützt. Für die aktuell laufende Prüfung des Bezugsjahrs 2017 wurden über 11.000 Einheiten maschinell klassifiziert.

6.3.3 Verbesserung der Schätzung des bereinigten Gender Pay Gaps

Um sich an das für die Schätzung des bereinigten Verdienstunterschiedes zwischen Männern und Frauen (Gender Pay Gap) fehlende Merkmal „wahre Berufserfahrung einer Arbeitnehmerin“ annähern zu können, sollten die Daten der Verdienststrukturerhebung (VSE), auf deren Basis der bereinigte Gender Pay Gap geschätzt wird, um eine frauenspezifische Variable „Erwerbsunterbrechung aufgrund der Geburt eines Kindes (ja/nein)“ angereichert werden. Damit würde es dem Statistischen Bundesamt ermöglicht, diese Form der Erwerbsunterbrechung bei der Erklärung von Verdienstunterschieden zu berücksichtigen (vgl. Finke et al. 2017; Dumpert und Beck 2017). Der Mikrozensus 2012 enthält das (wenngleich freiwillig zu beantwortende) Merkmal „Haben Sie Kinder geboren?“, das als Indiz für das Vorliegen einer Erwerbsunterbrechung gewertet wurde. Weitere Gründe für eine Erwerbsunterbrechung wie beispielsweise Arbeitslosigkeit oder Erwerbsunterbrechungen bei Männern wurden nicht untersucht. Anhand eines Kranzes von erklärenden Merkmalen, die in identischer oder ähnlicher Form sowohl im Mikrozensus 2012 als auch in der VSE vorhanden sind, sollte nun ein Zusammenhang zwischen der Ausprägung dieses Merkmals des Mikrozensus und den erklärenden Merkmalen gefunden werden. Anschließend konnten die in der VSE vorliegenden Daten zu Arbeitnehmerinnen um das so geschätzte Merkmal „Erwerbsunterbrechung ja/nein“ ergänzt werden. Der Mikrozensus diente dazu, die benötigten Trainings- und Testdaten zu generieren.

Eine reine Optimierung der Missklassifikationsrate führte dazu, dass im Testdatensatz ca. 20 % der Frauen hinsichtlich der Mutterschaft falsch klassifiziert wurden. Darüber hinaus war eine starke Disparität zwischen den Fehlerarten zu beobachten: Der Fehler, dass eine Mutterschaft geschätzt wurde, obwohl eine solche nicht vorlag, trat ungefähr dreimal so häufig auf wie der entgegengesetzte Fehler (irrtümliches Schätzen, dass keine Mutterschaft vorliegt).

Im bislang in der amtlichen Statistik verwendeten Ansatz zur Erklärung der Verdienstunterschiede zwischen Männern und Frauen werden alle Frauen so behandelt, als läge keine Erwerbsunterbrechung vor. Das fälschliche Unterstellen einer Mutterschaft ist daher der schwerer wiegende Fehler, wenn man einen konservativen Ansatz bei der Verfeinerung des bisherigen Verfahrens verfolgt. Es wurden daher weitere Anpassungen vorgenommen, mit dem Ziel, einen Ausgleich zwischen den beiden beschriebenen Fehlerarten zu schaffen. Diese

Anpassungen waren erfolgreich, führten jedoch zu einem Anstieg der Missklassifikationsrate insgesamt.

Es stellte sich heraus, dass die SVM-Methodik ihre Stärke, nämlich die Erkennung von Mustern (gegebenenfalls unter Inkaufnahme langer Rechenzeiten), nicht ausspielen konnte. Vergleichsrechnungen mit Random Forests lieferten bei deutlich kürzeren Berechnungsdauern vergleichbar gute Ergebnisse hinsichtlich der Missklassifikationsraten. Allerdings wiesen die Random Forests ein deutlich höheres Ungleichgewicht zwischen den auftretenden Fehlern als die Support Vector Machines auf. Sowohl Random Forests als auch Support Vector Machines lagen hinsichtlich der Missklassifikationsraten im Bereich parametrischer Verfahren wie der logistischen Regression.

Die experimentelle Schätzung der Mutterschaft bei weiblichen Beschäftigten, die in der VSE-Stichprobe für 2014 erfasst sind, erbrachte Resultate, die in das Modell zur Berechnung des bereinigten Gender Pay Gap einflossen, diesen jedoch nur im Bereich der sowieso vorhandenen statistischen Unsicherheit verringerten.

6.3.4 Prognose der Staatsangehörigkeit (dichotom) in der Verdienststrukturerhebung

Ein weiteres Merkmal, das derzeit nicht durch die Verdienststrukturerhebung erfasst wird, ist die Staatsangehörigkeit von Beschäftigten. Fragestellungen wie beispielsweise die Verteilung von Beschäftigten unterschiedlicher Staatsbürgerschaften auf verschiedene Wirtschaftszweige oder Verdienstklassen können derzeit nicht anhand der Daten der VSE beantwortet werden. Analog zur Schätzung der Mutterschaft von Arbeitnehmerinnen auf Basis des Mikrozensus wurde versucht, auch die verschiedenen Ausprägungen der Staatsangehörigkeit auf die Verdienststrukturerhebung in Form einer binären Klassifikation zu übertragen (vgl. Dumpert 2018).

Dabei trat folgende Schwierigkeit auf: Die Gruppe der Befragten im Mikrozensus mit rein nicht-deutscher Staatsbürgerschaft ist so klein, dass jedes Klassifikationsverfahren diese Gruppe „opfert“, um eine geringe Missklassifikationsrate zu erreichen; mit anderen Worten: Alle Ausländer werden als Deutsche klassifiziert. Es wurden daher verschiedene Herangehensweisen zum Umgang mit sogenannten unausgeglichenen Daten (unbalanced data) geprüft. Das Ziel bestand dabei darin, zum einen den Anteil der als Ausländer erkannten Beschäftigten unter den ausländischen Beschäftigten (Spezifität), zum anderen aber auch den Anteil der richtigerweise als Ausländer erkannten Beschäftigten unter allen als Ausländern klassifizierten Beschäftigten (Segreganz) auf ein für diese Fragestellung annehmbares Niveau zu heben. Als Zielkriterium wurde das harmonische Mittel aus diesen beiden Anteilen genutzt (F-Maß). Letztlich stellte sich heraus, dass klassische Verfahren zur Klassifikation (insbesondere die lineare Diskriminanzanalyse) ausreichen, um das Klassifikationsproblem zu lösen. Die Ergebnisse sind

jedoch nicht überzeugend: Offensichtlich sind sich Ausländer und Deutsche im Mikrozensus zu „ähnlich“, als dass sie mit den bislang verwendeten klassischen Methoden und Machine-Learning-Verfahren auf Basis des gemeinsamen Merkmalskranzes von VSE und Mikrozensus hinreichend gut klassifiziert werden können. Untersuchungen anhand anderer Datenquellen wurden bislang nicht angestellt.

6.3.5 Übertragung der Eigenschaft „Mindestlohnbetroffenheit“ auf die Daten der IEB

Die Mindestlohnkommission hat zur Verbesserung ihrer Datengrundlage in ihrem ersten Bericht die Verknüpfung der Verdienststrukturerhebung mit den Integrierten Erwerbsbiografien (IBE) der Bundesagentur für Arbeit angeregt. Die Paneldaten der IEB sollen auf diese Weise mit ansonsten fehlenden Angaben zum Bruttostundenverdienst bzw. zur Mindestlohnbetroffenheit aus der VSE angereichert werden, um so die Analysemöglichkeiten zu verbessern. Aufgrund gesetzlicher Vorgaben kommen hier nur statistische Verfahren in Frage, z. B. maschinelle Lernverfahren wie Random Forests. Getestet wurden Random Forests auf Basis der VSE 2014. Ziel ist es, ein Random-Forest-Modell aufzubauen, das die Klassifikation der Beschäftigten in „vom Mindestlohn betroffen“ und „vom Mindestlohn nicht betroffen“ mit hinreichender Verlässlichkeit ermöglicht. Diese Information soll dann auf die Daten der IEB „übertragen“ werden. Als Eingabevariablen werden Merkmale verwendet, die sowohl in der VSE als auch in den IEB vorliegen (vgl. Himmelreicher et al. 2017; Dumpert und Beck 2017).

Die Testrechnungen auf der Basis eines relativ einfachen Random-Forest-Modells ergaben für Vollzeitbeschäftigte gute Klassifikationsergebnisse. Die Missklassifikationsrate liegt bei rund 1 %. Bei den Teilzeitbeschäftigten ist der Fehler deutlich höher. Die Missklassifikationsrate beträgt rund 10 %, wobei der Großteil auf diejenigen Beschäftigten entfällt, die vom Mindestlohn betroffen sind, für die das Random-Forest-Modell aber das Gegenteil voraussagt. Da die vom Mindestlohn betroffenen nur rund 11 % der Teilzeitbeschäftigten ausmachen, liegt hier (ähnlich wie bei der Staatsbürgerschaft, vgl. Abschnitt 6.3.4) ein Unbalanced-Data-Problem vor. Noch schwieriger ist die Klassifikation von geringfügig Beschäftigten in vom Mindestlohn Betroffene und davon nicht Betroffene. Bei den Vollzeit- und Teilzeitbeschäftigten ist der Bruttomonatsverdienst die mit Abstand einflussreichste erklärende Variable. Bei den geringfügig Beschäftigten ist dies nicht der Fall, da der Bruttomonatsverdienst bei 450 Euro „gedeckelt“ ist und somit weniger Varianz aufweist. Hinsichtlich der verbleibenden erklärenden Variablen sind sich die vom Mindestlohn betroffenen bzw. nicht betroffenen geringfügig Beschäftigten zu „ähnlich“, als dass einfache Random-Forest-Modelle sie erfolgreich klassifizieren können. Zur Verbesserung des Modells kommen Maßnahmen, die dem Ungleichgewicht der Klassen bei den Teilzeitbeschäftigten entgegenwirken (z. B. Undersampling oder Oversampling) bzw. die Hinzunahme weiterer erklärender Variablen und der Einsatz eines „hybriden“ Ansatzes aus

Random Forest und Expertenschätzung bei den geringfügig Beschäftigten in Frage. Diese Ansätze sollen weiterverfolgt werden.

6.3.6 Scannerdaten in der Preisstatistik

Im Rahmen eines Eurostat-Projektes wird im Referat D306 aktuell anhand von Marktforschungsdaten die Nutzung von Scannerdaten zur Berechnung von Preisindizes getestet. Um zukünftig Rohdaten von den Einzelhandelsunternehmen verwerten zu können, müssen die unterschiedlichen Produkte der Klassifikation des Verbraucherpreisindexes zugeordnet werden. Eurostat hat sich auf übergeordneter Ebene diesem Problem gewidmet und ein Unternehmen beauftragt, im Zuge eines Proof of Concepts einen Prototyp zu entwickeln, mit dessen Hilfe die einzelnen Produkte den jeweiligen Positionen der ECOICOP (European Classification of Individual Consumption by Purpose) zugeordnet werden können. In Zusammenarbeit mit C303 passt D306 aktuell diesen Prototyp für einen Einsatz im Haus an.

7 Notwendige Infrastruktur

7.1 Software

Auswahl und Einsatz verlässlicher, idealerweise auch effizienter Software ist von entscheidender Bedeutung bei der Verwendung von Machine-Learning-Verfahren.

Da sowohl im Statistischen Bundesamt als auch im statistischen Verbund SAS Standard für Ad-hoc-Auswertungen ist, liegt es nahe, zuerst zu prüfen, ob SAS in der aktuell im statistischen Verbund verwendeten Form geeignet ist, Machine-Learning-Verfahren einzusetzen. SAS in der aktuell vorliegenden Version bietet keine vorgefertigten Methoden um SVM, baumbasierte Verfahren oder Neuronale Netze anzuwenden. Hierzu müssten zusätzliche oder gänzlich neue Werkzeuge angeschafft werden. Das Unternehmen SAS bietet die Produkte Enterprise Miner und Visual Analytics an. Beide enthalten Methoden, die die gängigen Machine-Learning-Algorithmen implementieren. Da diese Produkte im Hause noch nicht lizensiert sind, kann im Rahmen dieses Berichts keine Aussage zu Leistung und Einsetzbarkeit gemacht werden.

Natürlich besteht stets die Möglichkeit, die theoretisch verfassten Algorithmen händisch zu programmieren; die Wahl der Programmiersprache ist diesbezüglich dem Programmierer überlassen. Legt man alleine Wert auf Effizienz, bieten sich hier die noch sehr maschinennahen Sprachen C und Fortran an. Auch mit SAS ließen sich entsprechende Methoden theoretisch „from scratch“ entwickeln.

Aus naheliegenden Gründen sollte das Statistische Bundesamt jedoch auf bereits implementierte, erprobte Software zurückgreifen, die idealerweise zusätzlich Spielraum für eigene Adaptionen lässt, wesentliche Fragen der Programmierung (z. B. Speicher- und Objektverwaltung sowie Input- und Outputroutinen) jedoch bereits gelöst hat.

Im Bereich des maschinellen Lernens werden vornehmlich drei „Sprachen“ genannt, die diese Kriterien erfüllen: R, Python und MATLAB. In jüngerer Zeit tritt mit Julia eine weitere Sprache auf.

R ist eine kostenlos verfügbare Open-Source-Software und zusätzlich eine interpretierte Programmiersprache für Datenanalyse und deren graphische Umsetzung, bestehend aus grundlegenden Funktionen (base) und ergänzenden, ebenfalls kostenfreien Packages; letztere werden vom Nutzer nach Bedarf installiert. Die Software besitzt einen zentralen Internetauftritt: https://www.r-project.org. R-Packages zu Machine Learning stehen in großer Zahl und Vielfalt zur Verfügung. Zur Interoperabilität mit SAS ist festzuhalten, dass SAS grundsätzlich in der Lage ist, R-Prozeduren auszuführen und die Resultate der R-Prozeduren weiterzuverarbeiten[6].

Wie bei R handelt es sich bei Python um eine kostenlose, interpretierte Open-Source-Programmiersprache mit Standardbibliothek. Diese Standardbibliothek bietet grundlegende Funktionalitäten wie Input-Output-Routinen, einfache mathematische Operationen etc. Da Python selbst, als abstrakte Programmiersprache, nur eingeschränkte Möglichkeiten zur maschinennahen Programmierung bietet, müssen eigene, auf Effizienz ausgelegte Erweiterungen beispielsweise in C geschrieben werden. Aufgrund dieser Erweiterbarkeit können Python-Programmierer jedoch aus einem großen Bestand an Bibliotheken schöpfen, für Machine Learning-Anwendungen beispielsweise TensorFlow oder scikit-learn. Python besitzt einen zentralen Internetauftritt: https://www.python.org. Auch Python-Code lässt sich aus SAS ausführen. Eine grundsätzliche Interoperabilität mit SAS ist gegeben.[7]

MATLAB ist im Unterschied zu Python und R ein kostenpflichtiges Programmpaket, das ursprünglich für den effizienten Umgang mit Matrizen geschrieben wurde. Nicht alle Quellcodes sind einsehbar. Der Code wird in der Regel interpretiert; Compiler nach C existieren jedoch. Erweiterungen für spezielle Themengebiete, unter anderem Machine Learning, werden in sogenannten (in der Regel ebenfalls kostenpflichtigen) Toolboxen (z. B. die Statistics and Machine Learning Toolbox) angeboten; es steht Anwendern jedoch frei, eigene Toolboxen oder Programmroutinen zu verfassen und öffentlich zur Verfügung zu stellen. Für MATLAB stehen keine speziellen Schnittstellen für den Einsatz mit SAS zur Verfügung. Da MATLAB Batch-Verarbeitung unterstützt, kann es mit Abstrichen beim Input/Output aus SAS heraus verwendet werden.

Julia ist eine compilierbare, kostenfreie Open-Source-Programmiersprache, die erst 2012 entwickelt wurde, um die Effizienz von C und Fortran mit der Eleganz und einfachen Anwendbarkeit von beispielsweise Python und R zu kombinieren. Dabei wurde Julia speziell für sogenanntes High-Performance-Computing entwickelt, findet also sowohl innerhalb als auch außerhalb der Statistik Anwendung. Speziell für statistische Verfahren (inkl. Machine Learning)

[6] https://communities.sas.com/t5/General-SAS-Programming/Run-R-code-inside-SAS-easily/td-p/210116
[7] https://communities.sas.com/t5/SAS-Data-Management/How-to-RUN-R-or-Python-in-SAS/td-p/330170

findet sich auf juliastats.github.io eine Übersicht über vorhandene Packages, unter anderem „MLBase“ für Methoden „rund um das Machine Learning“ (data preprocessing, performance evaluation, cross validation, model tuning). Konkrete Implementierungen von Machine-Learning-Verfahren sind in diesem Package nicht enthalten. Für Support Vector Machines gibt es beispielsweise das eigene Package „SVM“. Weitere Packages im (erweiterten) Kontext Machine Learning sind „Clustering“, „MultivariateStats“, „GLMNet“ und „RegERMs“. Ebenso wie bei R wird die Anzahl der Packages im Laufe der Zeit zunehmen. Julia wird derzeit über github zur Verfügung gestellt: https://julialang.org oder https://github.com/JuliaLang/julia. Zur Interoperabilität mit SAS liegen keine Erfahrungen vor.

Während R speziell für Fragestellungen aus Statistik und Datenanalyse entwickelt wurde und weite Verbreitung in der statischen Lehre und Forschung findet, sind Python, MATLAB und Julia für weit allgemeinere Zwecke konzipiert. Im Rahmen der Projekte im Statistischen Bundesamt, die bereits Machine-Learning-Methoden verwendet haben, wurde R bzw. eine von Eurostat bereitgestellte Anwendung in Java/Javascript eingesetzt.

7.2 Hardware

Machine-Learning-Verfahren stellen häufig relativ hohe Anforderungen an die Hardware, auf der sie gerechnet werden sollen. In der Regel sind die Verfahren nur sinnvoll einsetzbar, wenn sie parallel auf mehreren Prozessorkernen gleichzeitig gerechnet werden können. Zum einen fordert dies ausreichend Kerne. Weiterhin steigt mit jedem parallel laufenden Prozess auch der benötigte Arbeitsspeicher.

Kleinere Projekte zum Testen einer Methode lassen sich zwar auf Arbeitsplatz-PCs (mit i.d.R. lediglich vier Kernen[8]) rechnen. Wenn man Verfahren jedoch produktiv einsetzen will, müssen leistungsstärkere Rechner mit ausreichend Arbeitsspeicher und Prozessoren genutzt werden. Zum Beispiel ließ sich die SVM zur Klassifikation der Relevanz von Handwerksunternehmen, bei der ca. 80.000 Einheiten zum Training verwendet wurden, nur auf Rechnern mit mehr als 40 parallel zur Verfügung stehenden Kernen in annähernd angemessenem zeitlichem Rahmen trainieren. Auf einem Arbeitsplatzrechner liefen Teilprozesse, die im produktiven Einsatz ca. 100 mal wiederholt werden müssen, bereits rund zwei Wochen.

Hinzu kommt die Entwicklung, dass einige Deep-Learning-Verfahren, i. d. R. Neuronale Netze, oft nur sinnvoll auf sehr spezieller Hardware einsetzbar sind. Hier sind beispielsweise Rechner zu nennen, die mit speziellen GPUs ausgerüstet sind, um besonders aufwändige Rechenoperationen effizient abzuarbeiten.

[8] Intel® Core™ i5-4570 Prozessor: https://ark.intel.com/de/products/75043/Intel-Core-i5-4570-Processor-6M-Cache-up-to-3_60-GHz

7.3 Beschäftigte – Know How und Skillsets

Neben passender Soft- und Hardware, müssen auch Mitarbeiterinnen und Mitarbeiter in die Lage versetzt werden, Machine-Learning-Verfahren einzusetzen. Hierzu sind bei den relevanten Personen Kenntnisse im Umgang mit den jeweils zu verwendenden Softwarepaketen Voraussetzung. In vielen Fällen ist sie schon vorhanden und in weiteren Fällen besteht die Bereitschaft, sich entsprechende Kenntnisse anzueignen.

Für die Gruppe der Beschäftigten, die bereits erste Erfahrungen gesammelt haben, müssen ausreichend Möglichkeiten bestehen, Erfahrungen und Informationen auszutauschen. Entsprechende Plattformen und Möglichkeiten sind bereitzustellen.

Für Beschäftigte, die noch keine Kenntnisse oder Erfahrung mit den relevanten Softwarepaketen und Verfahren haben, sollten entsprechende Schulungen angeboten werden.

8 Handlungsempfehlungen

Aus den Erkenntnissen des Proof of Concept Machine Learning werden folgende zehn Handlungsempfehlungen abgeleitet:

E1: Einrichtung eines Kompetenzzentrums in der Gruppe C1 „Mathematisch-statistische Methoden, Forschungsdatenzentrum“

Bisher ist das strategisch bedeutsame Thema „Machine Learning“ im Statistischen Bundesamt in der Aufbauorganisation noch nicht verankert. Dies wäre jedoch erforderlich, um die in den Fachgruppen geplanten Vorhaben zum Einsatz von Machine-Learning-Verfahren kompetent, kontinuierlich und strukturiert methodisch begleiten und unterstützen zu können. Ab dem 1. August 2018 ist eine entsprechende Umorganisation in der Gruppe C1 „Mathematisch-statistische Methoden, Forschungsdatenzentrum“ vorgesehen. Die Zuständigkeit für „Machine Learning“ wird dem noch aufzubauenden neuen Referat „C103 Maschinelles Lernen und Imputationsverfahren“ übertragen. Dabei ist es essentiell, dass das neue Referat personell so aufgestellt wird, dass es die Einführung und Anwendung von maschinellem Lernen sowie die Weiterentwicklung der Methodik quantitativ und qualitativ leisten und somit einen wesentlich Beitrag zur Umsetzung der Digitalen Agenda erbringen kann. Der Kompetenzaufbau in C103 (neu) muss möglichst rasch erfolgen und die Personalausstattung sollte hinsichtlich Anzahl und Kompetenz der künftigen Mitarbeiterinnen und Mitarbeiter der strategischen Bedeutung des Referates angemessen sein. Aus Sicht der Fachgruppen muss sichergestellt sein, dass C103 (neu) eine proaktive Beratungstätigkeit hinsichtlich problemadäquater Methoden wahrnehmen kann und dass Projekte in den Fachstatistiken zeitnah unterstützt werden. Flaschenhälse und Bearbeitungsstaus sind zu vermeiden, damit die Motivation, maschinelles Lernen/Künstliche

Intelligenz in fachstatistische Prozesse zu integrieren und die damit erhofften Effizienzgewinne zu realisieren, gestärkt wird.

E2: Sicherung bzw. Schaffung der notwendigen Infrastruktur für Machine-Learning-Projekte

Um Machine-Learning-Verfahren praktisch einsetzen zu können, bedarf es einer angemessenen Infrastruktur (Hardware; Software; Kommunikationstechnologie), die im Statistischen Bundesamt erst in Ansätzen vorhanden ist.

Hinsichtlich der erforderlichen Hardware lässt sich aufgrund der bisherigen Erfahrungen mit Machine-Learning-Projekten in E1 eindeutig sagen, dass die Standard-PCs nicht ausreichen. Es wird sicherlich wirtschaftlich nicht sinnvoll sein, für jedes rechenintensive statistische Verfahren – Machine Learning oder nicht – eigene Hardware zu beschaffen. Da Trainings- und Klassifikationsläufe in den meisten Fällen periodisch (einmal im Jahr, im Quartal, im Monat) laufen, muss die jeweils nötige Hardwareausstattung nicht dauerhaft bereitstehen. Vielmehr sollte sie zeitlich flexibel bereitzustellen sein. Da Server in Rechenzentren meistens virtuell gefahren werden, sie also aus verschiedenen physisch vorhandenen Maschinen zusammengestellt werden, sollte dies auch möglich sein. Ziel sollte es sein, für verschiedene Machine-Learning-Verfahren zeitlich begrenzt ausreichend dimensionierte virtuelle Rechenkraft (ggf. auch mit GPUs ausgestattete Rechner) bereitzustellen. Hierzu müsste von C2 und C3 unter Beteiligung von C103 (neu) sowie des ITZ-Bund ein geeignetes Konzept entwickelt und zeitnah umgesetzt werden.

Bezüglich der notwendigen Software führt kein Weg an R und Python vorbei. In der Digitalen Agenda ist diesbezüglich bereits das Projekt „D2: Einführung neuer Auswertungstools – Nutzung von R und Python zur Datenauswertung ermöglichen“ vorgesehen. Die konkrete Umsetzung erfolgt im Zuge der Artemis-Maßnahme 2400 bis 12/2018 durch die Gruppe C2. In der Zwischenzeit hat sich im Statistischen Bundesamt auch bereits eine informelle Gruppe von R-Anwender(inne)n gebildet (https://appweb.stba.testa-de.net/trac/ipunkt).

Prinzipiell bewährt hat sich trotz gewisser technischer Restriktionen die Informations- und Austauschplattform PCML TRAC, die auch nach Abschluss des Proof of Concept weitergeführt werden sollte. Dies gilt zumindest solange, bis im Statistischen Bundesamt moderne und leistungsstarke Alternativen zur Verfügung stehen. Es sollte geprüft werden, ob z. B. die derzeit pilotierten bzw. evaluierten Anwendungen „Confluence“ und „Jira“ als Kollaborations- und Wissensmanagementsoftware in Frage kommen.

E3: Schulungen und Fortbildung zu „Machine Learning“

Im September 2018 wird zum zweiten Mal ein Schulungskurs zum Thema „Maschinelles Lernen“ (ST 48) durchgeführt. Außerdem bietet die amtsinterne Fortbildung neu den Kurs „Einführung in das Statistikprogramm R“ (IS 80) an. Das amtsinterne Fortbildungsprogramm sollte bedarfsgerecht fortgeführt und ggf. ausgebaut werden (z. B. ergänzend in Richtung Python). Insbesondere für Mitarbeiterinnen und Mitarbeiter, die mit der Konzeption und Durchführung von Machine-Learning-Verfahren im Statistischen Bundesamt betraut werden, sollte die Möglichkeit geschaffen bzw. ausgebaut werden, an Fachtagungen und einschlägigen externen Schulungen teilzunehmen sowie sich mit Experten bei anderen Statistikproduzenten (national und international) auszutauschen. Zugriff auf einschlägige Fachzeitschriften sollte ermöglicht werden.

E4: Informationen für Führungskräfte

Die im Rahmen des Proof of Concept Machine Learning durchgeführten Kurzveranstaltungen und die gemeinsame Veranstaltung mit der Deutschen Bundesbank waren mit zusammen annähernd 300 Teilnehmerinnen und Teilnehmern sehr erfolgreich. Die Führungskräfte des Statistischen Bundesamtes, die letztlich über den Einsatz von Machine-Learning-Verfahren entscheiden bzw. diesen fördern sollten, konnten durch diese Veranstaltungsform jedoch nur zum Teil erreicht werden. Daher ist vorgesehen, das Thema „Machine Learning“ in der Abteilungsleitungsbesprechung (17. September 2018) und im Gruppen-/Referatsleitungenforum (21. September 2018) vorzustellen. Es wird empfohlen, den Themenkomplex „Machine Learning/Künstliche Intelligenz“ auch für die Führungskräftetagung 2019, ggf. mit externen Keynote-Speakern, einzuplanen.

E5: Allgemeine Information aller Mitarbeiterinnen und Mitarbeiter des Statistischen Bundesamtes

Alle Mitarbeiterinnen und Mitarbeiter des Statistischen Bundesamtes sollten über die Ergebnisse des Proof of Concept Machine Learning und die darauf aufsetzenden Umsetzungsmaßnahmen in geeigneter Weise informiert werden. Angeregt wird daher eine entsprechende Information im Intranet im Anschluss an die Behandlung in der Abteilungsleitungsbesprechung sowie eine hausinterne Veröffentlichung dieses Abschlussberichts, z. B. auf der Sonderseite zur „Digitalisierung“ im Intranet. Eine weitere Möglichkeit wäre die Gestaltung eines Informationsstandes, z. B. als Showcase oder im Rahmen des Marktplatzes zu den Ergebnissen der Leitungsklausur 2018.

E6: Information der Politik und weiterer Stakeholder

Über die Durchführung und die Ergebnisse des Proof of Concept Machine Learning soll im Rahmen der vom i-Punkt vorbereiteten Exzellenz (Digital) Show für Ressorts am 5. Dezember 2018 in Berlin berichtet werden. Zudem könnten auf diesem Wege die Bemühungen unterstützt werden, die Kompetenz des Statistischen Bundesamtes in die KI-Strategie des Bundes einzubringen.

Des Weiteren sind Vorträge auf der Statistischen Woche vom 11.–14. September 2018, auf der CESS vom 18.–19. Oktober 2018 sowie auf dem 27. Wissenschaftlichen Kolloquium „Mehr Zahlen, bessere Entscheidungen? Neue digitale Daten und Methoden in der empirischen Analyse und Beratung“ am 22./23. November 2018 in Wiesbaden vorgesehen.

Darüber hinaus wird empfohlen, den vorliegenden Abschlussbericht in gekürzter Form, z. B in AStA Wirtschafts- und Sozialstatistisches Archiv, zu veröffentlichen, um einen größeren Leserkreis zu erreichen. Für Berichte über abgeschlossene Machine-Learning-Projekte sollte ein WISTA-Sonderheft für ca. Mitte 2020 vorgesehen werden.

E7: Thematisierung in Statistikgremien und Einbindung der Statistischen Landesämter

Aufgrund seiner Methodenkompetenz hat das Statistische Bundesamt auch die Aufgabe, Anwendungsfälle im statistischen Verbund zu identifizieren und die eventuelle Umsetzung zu begleiten. Die künftige Anwendung von Machine-Learning-Verfahren in dezentralen Statistiken erfordert die frühzeitige Information und Einbindung der Statistischen Landesämter. Insbesondere sollte zunächst deren Leitungsebene vom Einsatz maschineller Lernverfahren überzeugt werden, um so bereits im Vorfeld der notwendigen Beratungen auf der Arbeitsebene die Unterstützung durch die Führungsebene sicherzustellen. Hierzu bietet es sich an, das Thema „Machine Learning“ jeweils auf die Tagesordnung des Abteilungsleitungsgremiums Fachstatistik (ALG FS) sowie der Amtsleitungskonferenz (ALK) zu setzen. Es empfiehlt sich allerdings, zuvor die Ergebnisse der geplanten Zuwendung für Forschung und Beratung zum „Einsatz von Verfahren der künstlichen Intelligenz in der amtlichen Statistik“ abzuwarten.

Die 2018 bereits begonnene Berichterstattung und Diskussion im Statistischen Beirat sollte 2019 fortgesetzt werden. Darüber hinaus bietet sich die Behandlung im „Arbeitskreis für mathematisch-statistische Methoden“ an.

E8: Zusammenarbeit mit Hochschulen, Forschungs- und Statistikinstitutionen

Die intensive Zusammenarbeit mit Hochschulen und Forschungsinstitutionen ist unerlässlich, da sich die Forschung zum Thema maschinelles Lernen rasant entwickelt. Nur in Kooperation mit der Wissenschaft kann sichergestellt werden, dass praxisrelevante methodische

Weiterentwicklungen schnell erkannt und für Zwecke der amtlichen Statistik nutzbar gemacht werden. Diese Aufgabe wäre künftig bei C103 (neu) anzusiedeln. Sie wird derzeit schon bei C1 wahrgenommen.

Darüber hinaus können intensive Kontakte zur Wissenschaft auch die Gewinnung talentierter Nachwuchskräfte in einem zunehmend härter werdenden Wettbewerb begünstigen. Das Statistische Bundesamt sollte im Bereich „Machine Learning“ gezielt Praktikumsplätze, Bachelor- und Masterarbeiten sowie Promotionsarbeitsplätze anbieten. Außerdem sollte das Statistische Bundesamt seine einschlägigen Projekte regelmäßig auf den entsprechenden Fachtagungen präsentieren. Für interessierte Mitarbeiterinnen und Mitarbeiter sind hierfür die entsprechenden Möglichkeiten zu schaffen.

Das gesamte Feld Machine Learning scheint in den in- und ausländischen Statistikinstitutionen aktuell hoch innovativ zu sein, so dass kurz- und mittelfristig weitere Entwicklungen beobachtet, getestet und bewertet werden müssen. Sinnvoll scheint es in diesem Zusammenhang zu sein, Kontakt zu den Ämtern mit vielen Projekten aufzunehmen oder zu halten, um von dort gemachten Erfahrungen zu profitieren. Im außereuropäischen Ausland wären dies insbesondere Kanada, USA, Neuseeland und Australien.

E9: Generierung neuer Projektideen

Die Hausumfrage hat gezeigt, dass es zwar viele Projektideen gibt, eine Reihe von Fachgruppen jedoch aus unterschiedlichen Gründen Fehlanzeige meldeten. Die Gruppe E1 befasst sich bereits seit mehreren Jahren mit der Anwendung von Machine-Learning-Verfahren. Um noch neue Projektideen zu generieren, wurde ein „hierarchiefreier“ gruppeninterner Workshop durchgeführt, in dem alle interessierten Mitarbeiter und Mitarbeiterinnen unabhängig von deren Funktion und konkreter Aufgabe Anregungen zur Anwendung von Machine-Learning-Verfahren einbringen und mitentwickeln konnten. Dabei ging es zunächst ausschließlich um die Generierung von Ideen. Ob diese überhaupt umsetzbar sind und ob dies effizient wäre, spielte in dem Workshop noch keine Rolle, sondern soll bewusst erst zu einem späteren Zeitpunkt geprüft werden. Auf diese Weise konnten in E1 trotz vorheriger intensiver Befassung mit den Einsatzmöglichkeiten von Machine Learning noch acht neue Vorschläge erarbeitet werden, die nun näher zu untersuchen sind.

Es wird angeregt, ähnliche Herangehensweisen in anderen interessierten Gruppen anzuwenden. E1 wäre bereit, hierbei bei Bedarf zu unterstützen.

E10: Behandlung im Jahresarbeitsplanungsgespräch 2019

Es wird angeregt, dass die geplanten, in der Hausumfrage genannten Machine-Learning-Projekte bzw. die Gründe, die dazu führen, dass keine solchen Maßnahmen verfolgt werden, in den

Jahresarbeitsgesprächen 2019 mit den (Fach-)Abteilungen thematisiert werden. Ziel ist es, den sinnvollen und effizienten Einsatz von Machine-Learning-Verfahren breit angelegt zu fördern.

Die Handlungsempfehlungen sind in Übersicht 5 noch einmal kompakt zusammengefasst und ergänzt um Zuständigkeiten und Termine.

Handlungsempfehlung	Zuständigkeit	Termin
E1: Einrichtung eines Kompetenzzentrums in der Gruppe C1 „Mathematisch-statistische Methoden, Forschungsdatenzentrum“ *(neues Referat C103)*	Abt. A ; C1	01.08.2018
E2: Sicherung bzw. Schaffung der notwendigen Infrastruktur für Machine-Learning-Projekte *(Hardware; Software, wie R und Python; Informations- und Austauschplattform)*	C2; C3; C103; ITZ-Bund	31.12.2018
E3: Schulungen zu „Machine Learning“ *(Durchführung; bedarfsgerechte Weiterentwicklung)*	A203	fortlaufend
E4 bis E7 beinhalten Vorschläge, die die Information verschiedener Interessengruppen über die Durchführung und die Ergebnisse des Proof of Concept Machine Learning betreffen		
E4: Informationen für Führungskräfte *(in ALB und Gruppen-/Referatsleitungsforum)*	L E1; C103	17.09.2019 21.09.2018
E5: Allgemeine Information aller Mitarbeiterinnen und Mitarbeiter des Statistischen Bundesamtes *(Intranetmeldung und Veröffentlichung Abschlussbericht)*	LS; B306; L E1	Ende 09/2018 sowie November 2018
E6: Information der Politik und weiterer Stakeholder *(Exzellenz (Digital) Show für Ressorts; Statistische Woche; Kolloquium; Aufsätze)*	i-Punkt; L E1; C103; B305	bis Ende 2018
E7: Thematisierung in Statistikgremien und Einbindung der Statistischen Landesämter *(in den Sitzungen des ALG FS und der ALK)*	Amtsleitung; LS; L E1; C103	offen
E8: Zusammenarbeit mit Hochschulen, Forschungs- und Statistikinstitutionen *(Forschungskooperation; Personalgewinnung)*	C103; A201	fortlaufend
E9: Generierung neuer Projektideen *(gruppeninterne Workshops)*	Fachgruppen; ggf. unterstützend E1	bis März 2019
E10: Behandlung im Jahresarbeitsplanungsgespräch 2019 *(Thematisierung der Machine-Learning-Projekte)*	Amtsleitung; A102- Controlling; Alle Abteilungen	Anfang 2019

Übersicht 5: Zusammenfassung der zehn Handlungsempfehlungen

9 Fazit

Der Proof of Concept Machine Learning bestätigte, dass es in den Fachstatistiken Potenzial für die Anwendung von maschinellem Lernen gibt. Neben den in E1 bereits abgeschlossenen Projekten, die als Piloten angesehen werden können, wurden seitens der Fachgruppen 31 Ideen für Machine-Learning-Anwendungen genannt. Bei sechs der genannten Ideen werden bereits Tests durchgeführt. 25 weitere müssen nun auf Umsetzbarkeit geprüft werden. Machine-Learning ist jedoch kein Allheilmittel und nicht jede Fachaufgabe kann damit erfolgreich gelöst werden. Dies lässt sich jedoch im Vorhinein nur schwer beurteilen. Ob eine Maßnahme erfolgversprechend ist, lässt sich im Allgemeinen nur feststellen, wenn man sie durchführt. Zum Einsatz von Machine Learning in den Fachstatistiken sollte ermutigt werden, ohne das erwartet wird, dass alle Projekte erfolgreich sein werden. Führungskräfte und Beschäftigte müssen offen sein für Veränderungen, die mit dem Einsatz von Machine Learning einhergehen. Sie müssen im Sinne einer Fehlerkultur einkalkulieren und akzeptieren, dass mit Innovationen und Veränderungen auch Misserfolge einhergehen können.

Das Projektteam sieht mit der Vorlage dieses Abschlussberichts den Proof of Concept Machine Learning und somit auch den Auftrag aus der Leitungsklausur als erfolgreich abgeschlossen an.

Der Abschlussbericht über den „Proof of Concept Machine Learning“ wird am 17. September 2017 in der ALB vorgestellt.

10 Anhang

10.1 Literatur

Zitierte Literatur

Boser, B. E., Guyon, I. M., Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory, 144–152.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A. (1984). Classification and Regression Trees. CRC Press.

Braun, H. (2018). Chefsache. Interview in DER SPIEGEL, Nr. 30, 21.7.2018.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32.

Bundesministerium für Bildung und Forschung (2018). Künstliche Intelligenz. Beitrag auf der Website vom 9. April 2018. https://www.bmbf.de/de/kuenstliche-intelligenz-5965.html, zuletzt abgerufen am 26.07.2018.

Bundesregierung (2018). Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz Stand: 18. Juli 2018. https://www.bmbf.de/files/180718%20Eckpunkte_KI-Strategie%20final%20Layout.pdf, zuletzt abgerufen am 26.07.2018,

Chu, K., Poirier, C. (2015). Machine Learning Documentation Initiative. UNECE Conference of European Statisticians, Workshop on the Modernisation of Statistical Production Meeting, 15-17 April 2015. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2015/Topic3_Canada_paper.pdf, zuletzt abgerufen am 14.07.2018. Mittlerweile fortgeführt durch V. Todorov (UNIDO).

Deutscher Bundestag (2018a). Bundestagsdrucksache 19/605. http://dip21.bundestag.de/dip21/btd/19/006/1900605.pdf, zuletzt abgerufen am 26.07.2018.

Deutscher Bundestag (2018b). Bundestagsdrucksache 19/1525. https://dip21.bundestag.de/dip21/btd/19/015/1901525.pdf, zuletzt abgerufen am 10.07.2018.

Deutscher Bundestag (2018c). Bundestagsdrucksache 19/1982. https://dip21.bundestag.de/dip21/btd/19/019/1901982.pdf, zuletzt abgerufen am 10.07.2018.

Deutscher Bundestag (2018d). Enquete-Kommission zur künstlichen Intelligenz eingesetzt. https://www.bundestag.de/dokumente/textarchiv/2018/kw26-de-enquete-kommission-kuenstliche-intelligenz/560330, zuletzt abgerufen am 26.07.2018.

Dumpert, F. (2018). Abschlussbericht zum Projekt „Prüfung und Bewertung von Optionen zur Schätzung der Staatsbürgerschaft in der Verdienststrukturerhebung (VSE)“. Bericht liegt dem Statistischen Bundesamt vor.

Dumpert, F., Beck, M. (2017). Einsatz von Machine-Learning-Verfahren in amtlichen Unternehmensstatistiken. AStA Wirtschafts- und Sozialstatistisches Archiv, 11, 83–106.

Dumpert F., von Eschwege K., Beck M. (2016). Einsatz von Support Vector Machines bei der Sektorzuordnung von Unternehmen. WISTA Nr. 1/2016, 87–97.

Dutton, T. (2018). An Overview of National AI Strategies, https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd, zuletzt abgerufen am 26.07.2018.

Feuerhake, J., Dumpert, F. (2016). Erkennung nichtrelevanter Unternehmen in den Handwerksstatistiken. WISTA Nr. 2/2016, 79–94.

Finke, C., Dumpert, F., Beck, M. (2017). Verdienstunterschiede zwischen Männern und Frauen. WISTA Nr. 2/2017, 43–61.

Himmelreicher, R. K., vom Berge, P., Fitzenberger, B., Günther, R., Müller, D. (2017). Überlegungen zur Verknüpfung von Daten der Integrierten Erwerbsbiographien (IEB) und der Verdienststrukturerhebung (VSE). RatSWD Working Papers 262/2017.

Huber, M., Schüller, S., Stöckli, M., Wohlrabe, K. (2018). Maschinelles Lernen in der ökonomischen Forschung. Ifo-Schnelldienst Nr. 7/2018, 50-53.

iit – Institut für Innovation und Technik (2018). Potenziale der Künstlichen Intelligenz im produzierenden Gewerbe in Deutschland. Studie im Auftrag des Bundesministeriums für Wirtschaft und Energie, https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/potenziale-kuenstlichen-intelligenz-im-produzierenden-gewerbe-in-deutschland.pdf?__blob=publicationFile&v=16, zuletzt abgerufen am 26.07.2018.

Koalitionsvertrag 2018. Ein neuer Aufbruch für Europa - Eine neue Dynamik für Deutschland - Ein neuer Zusammenhalt für unser Land. Koalitionsvertrag zwischen CDU, CSU und SPD. https://www.bundesregierung.de/Content/DE/_Anlagen/2018/03/2018-03-14-koalitionsvertrag.pdf;jsessionid=CAC83E2BE85B3BB247EF228FDEC04C6C.s3t1?__blob=publicationFile&v=6, zuletzt abgerufen am 26.07.2018.

König, C., Wiegand, E., Schröder, J. (Eds.) (2017). Big Data: Chancen, Risiken, Entwicklungstendenzen. Springer.

Ramge, T. (2018) Mensch fragt, Maschine antwortet. Aus Politik und Zeitgeschichte Nr. 6–8/2018, 15–21.

McCarthy , J., Minsky, M. L., Rochester, N., Shannon, C. E. (1955). Proposal for the Dartmouth summer research project on artificial intelligence. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf, zuletzt abgerufen am 28.07.2018.

O'Neil, C. (2017). Angriff der Algorithmen, Carl Hanser Verlag.

Rich, E. (1983) Artificial Intelligence. McGraw-Hill.Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Reviews. 65, 386–408.

Russell, S., Norvig, P. (2012) Künstliche Intelligenz - Ein moderner Ansatz, 3. Auflage. Pearson.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Simon, H. A. (1983). Why should machines learn? In: Michalski R. S., Carbonell J. G., Mitchell T. M. (Eds.) Machine Learning: An Artificial Intelligence Approach, 25–38. Tioga Press.

Statistics Canada (2018). Machine Learning in Surveys Steps.

Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Performance Evaluation Review, 41(4), 70-73.

Suthaharan, S. (2016). Machine Learning Models and Algorithms for Big Data Classification. Springer.

Vapnik, V. N. (1995). The nature of statistical learning theory. Springer.

Villani, Cédric (2018). For a Meaningful Artificial Intelligence. Towards a French and European Strategy. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf, zuletzt abgerufen am 26.07.2018.

Überblicksliteratur zum Thema Machine Learning allgemein

Alpaydin, E. (2014). Introduction to Machine Learning, Third Edition. MIT Press.

Caruana, R., Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

Fernandez-Delgado, M., Cernadas, E., Barro, S., Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research, 15(Oct), 3133–3181.

Ghatak, A. (2017). Machine Learning with R. Springer.

Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep learning. MIT press.

Hastie, T., Tibshirani, J., Friedman, R. (2008). The Elements of Statistical Learning, Second Edition. Springer.

James, G., Witten, D., Hastie, T., Tibshirani, J. (2013). An Introduction to Statistical Learning (with applications in R). Springer.

Lantz, B. (2015). Machine learning with R. 2nd edition. Packt Publishing.

Mueller, J.P., Massaron, L. (2016): Machine Learning for Dummies.

Mullainathan, S., Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87–106.

Raschka, S., Mirjalili, V. (2017a). Python machine learning, 2nd edition. Packt Publishing.

Raschka, S., Mirjalili, V. (2017b). Machine Learning mit Python und Scikit-Learn und TensorFlow. MITP.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Literatur zu Support Vector Machines

Bennett, K. P., Campbell, C. (2000). Support Vector Machines: Hype or Hallelujah? SIGKDD Explorations, 2(2), 1–13.

Cristianini, N., Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge university press.

Hamel, L. (2009). Knowledge Discovery with Support Vector Machines. Wiley.

Hsu, C. W., Chang, C. C., Lin, C. J. (2003, Last updated: 2016). A practical guide to support vector classification, https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, zuletzt abgerufen am 14.07.2018.

Steinwart, I., Christmann, A. (2008). Support Vector Machines. Springer.

Literatur zu Trees und Random Forests

Criminisi, A., Shotton, J., Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations and Trends in Computer Graphics and Vision, 7(2–3), 81–227.

Fawagreh, K., Gaber, M. M., Elyan, E. (2014). Random forests: from early developments to recent advancements. Systems Science & Control Engineering, 2, 602–609.

Wyner, A. J., Olson, M., Bleich, J., Mease, D. (2017). Explaining the success of adaboost and random forests as interpolating classifiers. Journal of Machine Learning Research, 18(1), 1558–1590.

Sonstige Literatur

Rekatsinas, T., Chu, X., Ilyas, I., Ré, C. (2017). HoloClean: Holistic Data Repairs with Probabilistic Inference, in: Proceedings of the VLDB Endowment, Vol. 10, No. 11, 1190-1201.

Rekatsinas, T., Ilyas, I., Ré, C. (2017). HoloClean: Weakly Supervised Data Repairing. https://hazyresearch.github.io/snorkel/blog/holoclean.html, zuletzt abgerufen am 26.07.2018.

Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier, https://arxiv.org/abs/1602.04938v3 [cs.LG] 9 Aug 2016.

Silver, D. et al. (2017): Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, https://arxiv.org/abs/1712.01815v1 [cs.AI] 5 Dec 2017.

Silver, D. et al. (2017): Mastering the game of Go without human knowledge, in: Nature, 550, 354–359.

Varian, H.R. (2014): Big Data: New Tricks for Econometrics, in: Journal of Economic Perspectives, 28 (2), 3-28.

10.2 Rückmeldung aus nationaler/internationaler Abfrage

10.2.1 Kurzauswertung der Abfragen

DESTATIS
Statistisches Bundesamt

UMFRAGE ZUM EINSATZ VON MACHINE LEARNING VERFAHREN – AUSWERTUNG

Proof of Concept Machine Learning

wissen.nutzen.

Wie setzen Statistikproduzenten Machine Learning ein?

Im April 2018 wurde eine Umfrage zum Einsatz von Machine Learning bei nationalen und internationalen Statistikproduzenten durchgeführt.

Beteiligung	StLÄ	National	International	Insgesamt
Angeschrieben	14	18	39	71
Rückmeldung mit Angaben	1	6	19	26
Anzahl der Anwendungen	**1**	**36**	**101**	**138**
Fehlanzeige	13	12	11	36
Keine Rückmeldung	0	0	9	9

Stand: 30.07.2018; bei der Anzahl der Anwendungen wurden nur Rückmeldungen aus der Umfrage berücksichtigt.

Die Antwortenden haben sehr unterschiedliche Machine Learning Anwendungen genannt ...

2

wissen.nutzen.

Grafik erzeugt mit Word Cloud Generator Copyright © Jason Davies

wissen.nutzen.

Erstes Fazit

Es gibt bereits eine Vielzahl von Einsatzmöglichkeiten für Machine Learning Verfahren ...

... in vielen Phasen des Statistikprozesses ...

... und in nahezu allen Bereichen der amtlichen Statistik.

Welche Anwendungen sind zurzeit am wichtigsten?

Grafik erzeugt mit Word Cloud Generator Copyright © Jason Davies

wissen.nutzen.

Grafik erzeugt mit Word Cloud Generator Copyright © Jason Davies

wissen.nutzen.

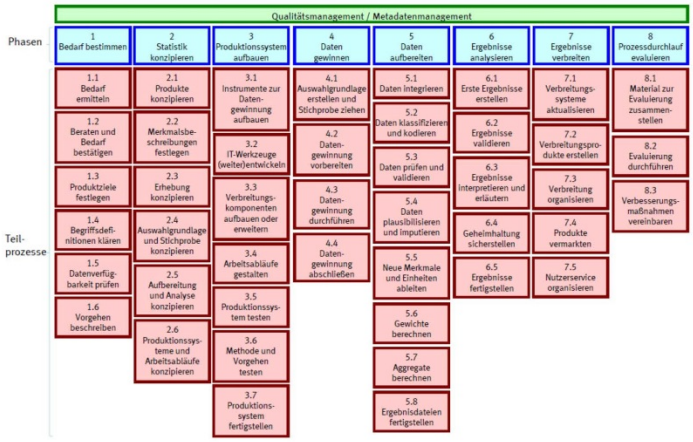
In welchen Phasen der Statistikproduktion kommt Machine Learning in Statistikinstitutionen bereits zum Einsatz?

Grafik erzeugt mit Word Cloud Generator Copyright © Jason Davies

wissen.nutzen.

Geschäftsprozess-modell Amtliche Statistik

GSBPM – Generic Statistical Business Process Model (Version 5.0)



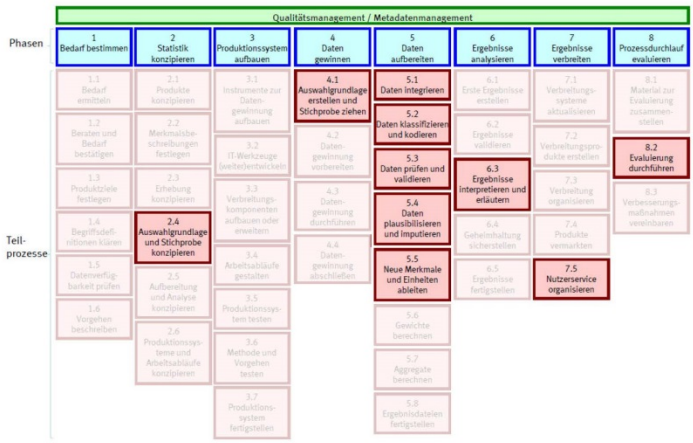
Quelle: GMAS Geschäftsprozessmodell Amtliche Statistik - StaNet

wissen.nutzen.

Machine Learning Einsatzgebiete

Laut Umfrage wird in den hervorgehobenen Phasen des GMAS bereits Machine Learning eingesetzt

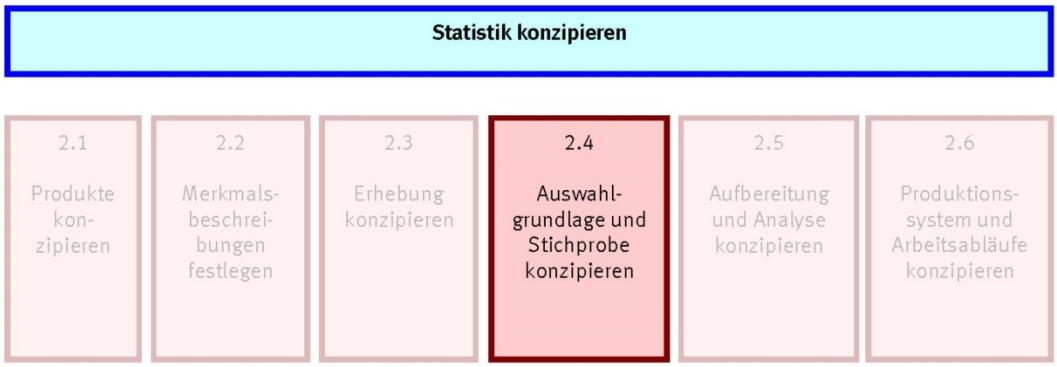
GSBPM – Generic Statistical Business Process Model (Version 5.0)



Quelle: GMAS Geschäftsprozessmodell Amtliche Statistik - StaNet

wissen.nutzen.

P02 Statistik konzipieren



Quelle: GMAS Geschäftsprozessmodell Amtliche Statistik - StaNet

wissen.nutzen.

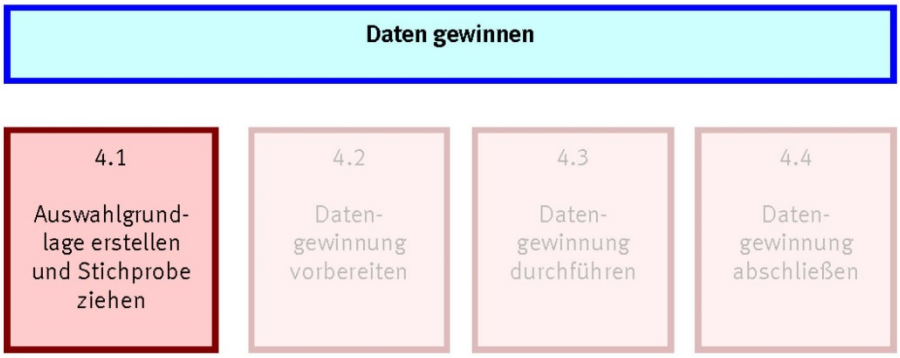
2.4 Auswahlgrundlage und Stichprobe konzipieren

Machine Learning wird zur Modellierung der zu erwartenden Antwortausfälle anhand vorliegender Informationen aus Vorperioden eingesetzt.

Ziel: Optimierung der Stichprobenallokation

wissen.nutzen.

P04 Daten gewinnen



Quelle: GMAS Geschäftsprozessmodell Amtliche Statistik - StaNet

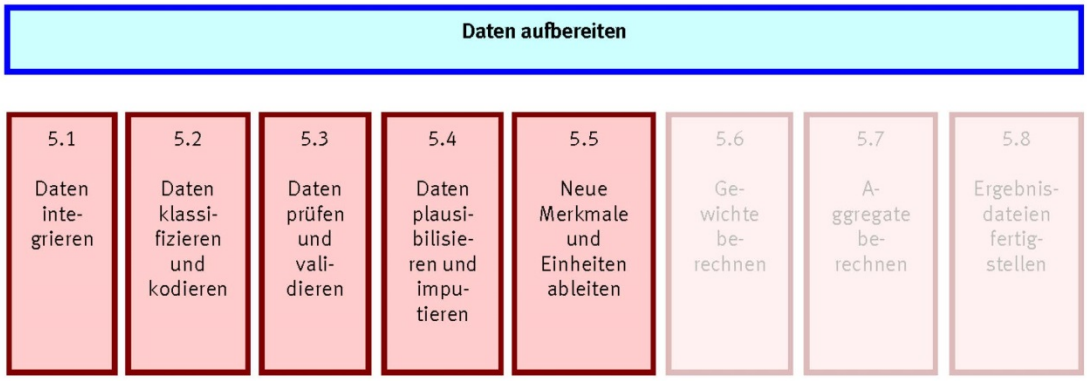
wissen.nutzen.

4.1 Auswahlgrundlage erstellen und Stichprobe ziehen

Machine Learning wird zur Identifikation von Einheiten verwendet, die für die Auswahlgrundlage relevant bzw. nicht relevant sind.

wissen.nutzen.

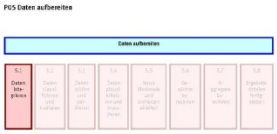
P05 Daten aufbereiten



Quelle: GMAS Geschäftsprozessmodell Amtliche Statistik - StaNet

wissen.nutzen.

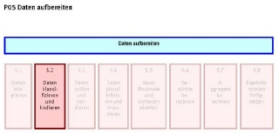
5.1 Daten integrieren



Bei der Mikrodatenverknüpfung werden Machine Learning Algorithmen eingesetzt, um relevante Merkmale für die Suche nach möglichen Verknüpfungen zu finden sowie um Zuordnungskonflikte effizient zu behandeln.

wissen.nutzen.

5.2 Daten klassifizieren und kodieren



Beim Vorliegen unstrukturierter Texte z.B. aus Webscraping oder freien Antwortfeldern, aber auch bei der Auswertung von Satellitenbildern können Klassifizierungs- und Kodierungsaufgaben (Berufe, Produkte, Wirtschaftszweige, etc.) automatisiert werden.

» **... dies ist der am häufigsten genannte Anwendungsfall für Machine Learning**

 15

wissen.nutzen.

5.3 Daten prüfen und validieren

Machine Learning Algorithmen werden bereits in verschiedenen Projekten erfolgreich zur Identifikation von Ausreißern eingesetzt.

 16

wissen.nutzen.

5.4 Daten plausibilisieren und imputieren

Sowohl die Identifikation von unplausiblen Meldungen als auch die Imputation plausibler Werte bei fehlenden oder unplausiblen Angaben wird häufig als Einsatzgebiet für Machine Learning Algorithmen genannt.

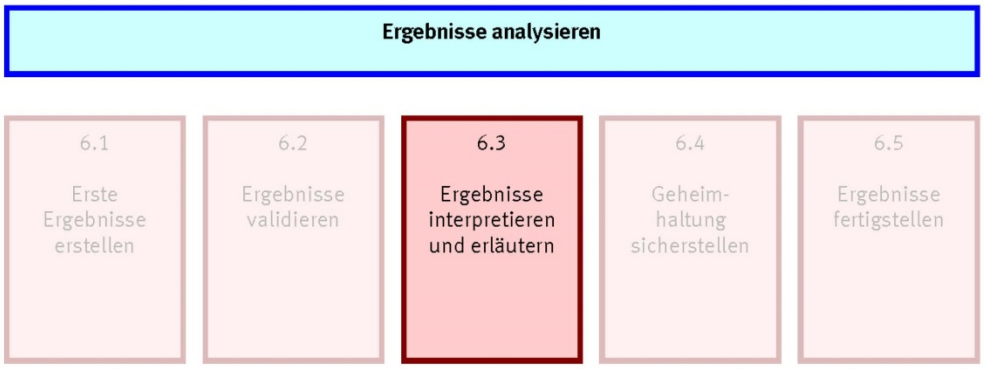
wissen.nutzen.

5.5 Neue Merkmale und Einheiten ableiten

Mit Machine Learning Verfahren lassen sich zusätzliche Merkmale schätzen, die z.B. in Datenbeständen aus Vorperioden vorliegen oder aus anderen Quellen ergänzt werden können.

wissen.nutzen.

P06 Ergebnisse analysieren



Quelle: GMAS Geschäftsprozessmodell Amtliche Statistik – StaNet

wissen.nutzen.

6.3 Ergebnisse interpretieren und erläutern

Machine Learning Verfahren werden häufig angewandt, um in Datenbeständen Prognosen bzw. Schätzungen zu ergänzen, die zum Zeitpunkt der Erhebung nicht vorliegen.

Beispiel: Prognose der Dauer von Arbeitslosigkeit aus Informationen der Meldung.

wissen.nutzen.

P07 Ergebnisse verbreiten

Quelle: GMAS Geschäftsprozessmodell Amtliche Statistik – StaNet

wissen.nutzen.

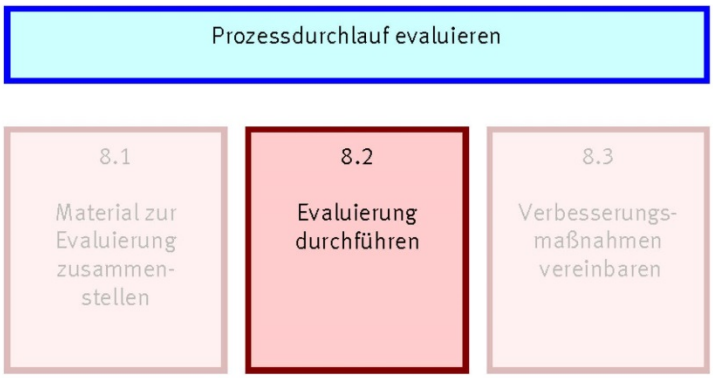
7.5 Nutzerservice organisieren

Machine Learning Verfahren können eingesetzt werden, um Webseiten-Nutzer anhand ihres Klickverhaltens als „erfolglos suchende Nutzer“ zu identifizieren und gezielt per Webseiten-Chat anzusprechen.

Ziel: Effizient die Nutzerzufriedenheit steigern, indem man vermutlich unzufriedenen Nutzern gezielt Hilfestellung anbieten kann.

wissen.nutzen.

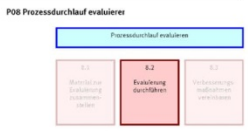
P08 Prozessdurchlauf evaluieren



Quelle: GMAS Geschäftsprozessmodell Amtliche Statistik - StaNet

wissen.nutzen.

8.2 Evaluierung durchführen

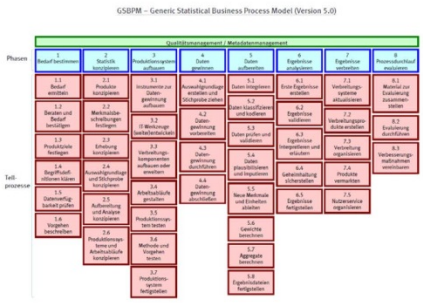


Automatisierte Auswertung der Kommentare der Berichtenden in den Fragebögen und Online-Erhebungswerkzeugen. Identifikation besonders relevanter Handlungsfelder bei der Verbesserung der Erhebungswerkzeuge.

Automatisierte Auswertung des Surf-Verhaltens auf den Webseiten zur Identifikation von Usability-Potential.

wissen.nutzen.

Weitere Anwendungsfälle



Das Papier „Machine-Learning in Surveys Steps“ von Statistics Canada listet anhand des GSBPM Anwendungsfälle auf, bei denen Machine Learning Verfahren eingesetzt werden können.

Quelle: GMAS Geschäftsprozessmodell Amtliche Statistik - StaNet

wissen.nutzen.

Welche Statistiken sind geeignet?

wissen.nutzen.

Genannte Statistiken und Schlagworte

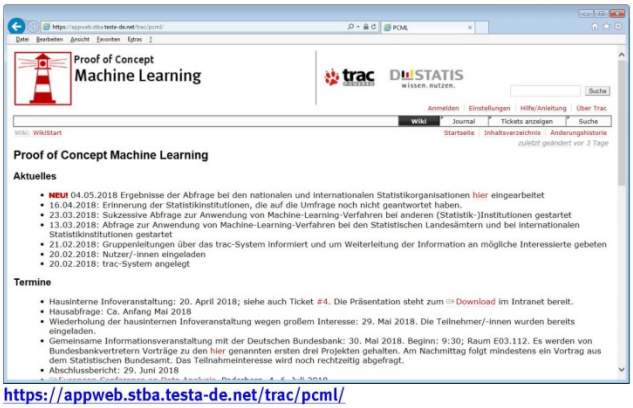
In der Umfrage wurden Projekte aus den verschiedensten Bereichen der Statistik genannt:

… Verbraucherpreisindex, Scannerdaten, Einkommens- und Verbrauchserhebung, EU-SILC, Verwaltungsdaten, Baugenehmigungen, Einzelhandel, Konsum, Tourismus, Landwirtschaft, Zensus, Wirtschaftsstatistiken, Todesursachenstatistik, neue digitale Daten, Nutzer- und Berichtendenbeziehungen …

wissen.nutzen.

Austausch- und Informationsplattform

Auf der Austausch- und Informationsplattform zum Proof of Concept Machine Learning finden sie weitere Informationen



https://appweb.stba.testa-de.net/trac/pcml/

wissen.nutzen.

Kontakt

» Martin Beck
» martin.beck@destatis.de
» Tel.: 0049 611 75 4460

» Joerg Feuerhake
» joerg.feuerhake@destatis.de
» Tel.: 0049 611 75 4116

10.2.2 Rückmeldungen nationaler Institutionen

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
RKI / ZBS 6	Identifizierung, Differenzierung und Klassifizierung von biolog. Proben wie Zellen, Gewebe und Mikroorganismen auf Basis von spektroskopischen und spektrometrischen Daten	Im Fachgebiet ZBS 6 des RKI werden Verfahren des maschinellen Lernens und der künstlichen Intelligenz (KI) seit über 20 Jahren eingesetzt. Zur Anwendung kommen hierbei u.a. Varianten Künstlicher Neuronaler Netze (ANNs), welche Mustererkennung anhand von Spektrendaten, z.B. aus der Vibrationsspektroskopie oder Massenspektrometrie ermöglichen. Spektroskopische und spektrometrische Daten werden von komplexen biologischen Proben wie Zellen, Gewebe u.ä. aufgezeichnet; Ziel der Anwendung von ANNs ist die schnelle, objektive und kostengünstige Identifizierung, Differenzierung und Klassifizierung in der Zytologie, Histologie und Mikrobiologie im Rahmen von Forschungsaktivitäten bzw. zur Methodenentwicklung in der Diagnostik	Differenzierung Identifizierung Klassifizierung	Forschungsprojekte	ANN SVM Strategien zur Optimierung	NeuroDeveloper Matlab Matlab NN Toolbox Biotools ga_ors tooldiag SNNS

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
RKI / MF1	Bioinformatik / Analyse molekularer Daten	Verschiedene Verfahren des Maschinellen Lernens werden zur Analyse von großen Datensätzen aus Omics-Experimenten wie der Genomsequenzierung automatisierte Assistenzsysteme entwickelt und eingesetzt. Diese Datensätze sind so groß (teilweise bis zu einer Milliarde Genomfragmenten aus einem einzelnen Experiment), dass eine manuelle Analyse gerade bei zeitkritischen Vorgängen nicht vollumfänglich sinnvoll oder möglich ist. Einsatz findet dies bspw. zur Charakterisierung von bakteriellen oder viralen Erregern oder zur Erregersuche.Mit Hilfe der Assistenzsysteme werden insbesondere irrelevante Messungen für den menschlichen Entscheider identifiziert, also bspw. Genomfragemente, die nur mit niederer Qualität gemessen wurden, die einen biologisch geringen Informationsgehalt haben oder Kontaminationen aus Messung, Umwelt oder einem Wirtsorganismus beinhalten.Ferner werden besondere Gefahrenpotentiale einzelner Messungen (bspw. Nähe zu bekannten Erregern oder relevante potenzielle Phänotypen wie Virulenz oder Resistenz) für den menschlichen Entscheider hervorgehoben.	Klassifikation (binär und Mehrklassenprobleme), Regression, Clustering	Forschungsprojekte und Produktivbetrieb	Random Forests, Deep Learning	R, Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
RKI / FG-31	Signale - automatische Ausbruchserkennung für Infektionskrankheiten	Wir entwickeln und implementieren Methoden des maschinellen Lernens zur Vorhersage von Fallzahlen infektiöser Krankheiten und zur Erkennung von Auffälligkeiten in Surveillance Daten. Weiterhin verwenden wir sogenannte "supervised learning" Ansätze um diese Algorithmen zu vergleichen und optimieren. Im Rahmen des "Signale 2.0" Projekts ist weitere Anwendungsforschung hinsichtlich Natural Language Processing zur Verarbeitung unstrukturierter Daten (z.B. Protokolle, Veröffentlichungen) geplant, die extrahierten Informationen sollen zur Verfügung gestellt und ebenfalls in die Ausbruchserkennung einfliessen.	Regression, Klassifikation und Informationsgewinnung	Forschungsprojekte und Produktivbetrieb	HMM, GLM, NLP, Klassifizierungsalgorithmen (Log. Reg., Random Forest, SVM, Neuronale Netzwerke)	R, Python
RKI / P4	Anwendung maschineller Lernverfahren im RKI Gesundheitsmonitoring (KiGGS)	Im Rahmen der KiGGS-Studie hat das RKI rund 4000 Variablen zur Gesundheit von über 17.000 Kindern und jungen Erwachsenen erhoben. Durch die Größe und Heterogenität des Datensatzes stoßen klassische statistische Methoden teilweise an konzeptionelle Grenzen. Als Pilotprojekt für den Einsatz Maschineller Lernverfahren im Public-Health-Monitoring soll dieses Projekt aufzeigen, wie neuartige Methoden des Maschinellen Lernens in der Arbeit des RKI genutzt werden können. Verschiedene Methoden sollen getestet, um Strukturen und Zusammenhänge in den Daten zu erkennen und Vorherhagen zum Auftreten chronischer Krankheiten zu treffen.	Clusterbildung, binäre und mehrklassige Klassifikation	Experiment	Clusterverfahren, Entscheidungsbäume, Neuronale Netze	Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Zentrum für Europäische Wirtschaftsforschung (ZEW)	TOBI - Textdaten-basierte Output-Indikatoren als Basis einer neuen Innovationsmetrik	Im Verbundprojekt werden neue Output-Indikatoren zu Innovationsaktivitäten entwickelt. Dabei kommen computerlinguistische Verfahren zum Einsatz, die auf große Mengen von Textdaten angewandt werden. Die Entwicklung der Methoden und Validierung der generierten Indikatoren erfolgt dabei arbeitsteilig am ZEW in Mannheim und der Justus-Liebig-Universität Gießen. Am ZEW erfolgt die Analyse auf Basis von Textinhalten aus Unternehmenswebseiten, die automatisiert und regelmäßig über einen Webscraper gesammelt werden. Mittels Text Data Mining (z. B. Topic-Modelle) werden dann aus diesen Texten Informationen zu Innovationen identifiziert und daraus Innovationsindikatoren abgeleitet. Der Zugriff auf die Webseiten erfolgt auf Basis der am ZEW vorhandenen Datenbanken. Diese erlauben das fortlaufende Monitoring der Webseiten des aktuellen deutschen Unternehmensbestandes und die Berücksichtigung umfangreicher Metadaten (z. B. Branche und Standort des Unternehmens). Zusätzlich können die neu generierten Innovationsindikatoren über die ZEW Datenbanken mit konventionellen Innovationsindikatoren verglichen werden.	Text-Analyse und Klassifikation	Entwicklung	Diverse	Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Zentrum für Europäische Wirtschaftsforschung (ZEW)	Science4KMU	Einschätzung der Kooperationsbereitschaft von Unternehmen mittels eines Neuronalen Netzwerkes basierend auf Umfragedaten des Community Innovation Survey (CIS). Das Modell soll die Kontaktaufnahme von Technology Transfer Offices (TTO) zu Unternehmen effizienter gestallten.	Probabilistisches Scoring-Modell	Prototyp	Neural Network	Stata
Hessisches statistisches Landesamt	Webscraping von Unternehmenswebseiten	Mittels Webscraping gewonne Daten zu den Unternehmen des URS sollen u. a. durch Methoden des maschinellen Lernens ausgewertet werden, um statistikübergreifende Kohärenzprüfungen durchzuführen sowie neue Merkmale zu generieren.	Datenextraktion, Kohärenzprüfungen	Testbetrieb	Webscraping, graphenbasierte neuronale Netze	Java, R, Mysql, Apache Spark
Deutsche Bundesbank	Record Linkage in RIAD	Objektidentifizierung für den Aufbau und für die Aktualisierung von RIAD-BBk.	Klassifikation	Prototype	Random Forrests	Python
Deutsche Bundesbank	SIMBA - System zur integrierten Meldebearbeitung im Außenwirtschaftsverkehr	SIMBA wird die neue Plattform für die integrierte Meldebearbeitung der Außenwirtschaftsstatistiken und in den nächsten 24 Monaten (bis etwa Mitte 2020) in einer Grundstufe ausgeprägt. Im Zuge des geplanten weiteren Ausbaus des Systems kann man sich perspektivisch den Einsatz von ML-Verfahren vorstellen zur Steigerung der Datenqualität, beispielweise in einer systemgestützten Analyse welche Meldungen bearbeitet werden sollten, zur Ermittlung von Vorschlagswerten etc.	binäre Klassifikation	Projekt (Durchführung der Grundstufe des Systems)	RF (Random Forest)	Java/ DB2

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Deutsche Bundesbank	Klassifikation von Holdinggesellschaften im Pool der nichtfinanziellen Einzelabschlussstatistik	Der Jahresabschlussdatenpool der Deutschen Bundesbank der nicht nanziellen deutschen Einzelabschlüsse beinhaltet Daten von durchschnittlich 130.000 Unternehmen jahrlich von 1997 bis 2015. Jeder Merkmalsträger verf ugt über eine Vielzahl von Positionen zur Bilanz, Gewinn- und Verlustrechnung, sowie zur Wirtschaftszweigzuordnung nach der Klassifkation der Wirtschaftszweige 2008. Manuelle Überpr ufungen haben ergeben, dass bei den Wirtschaftszweigen teilweise Fehlzuordnungen vorliegen. In diesem Kontext ist die Klassi kation von Holdinggesellschaften zun achst von besonderem Interesse. Erfahrungsgem a unterscheiden sich Holdinggesellschaften von anderen Unternehmen in einer Reihe von Eigenschaften. Um Fehlzuordnungen bei den Wirtschaftszweigen zu identi zieren, wird ein maschinelles Lernprogramm angewandt.	binäre Klassifikation	Testbetrieb	Logit/Probit/Neuronale Netze	NA
Deutsche Bundesbank	Verknüpfung von Firmendaten der Bundesbank mit Hilfe von Maschinellem Lernen	Dr. Christopher-Johannes Schild, Abstract Statistische Woche: We present a method of automatically linking data sets on companies based on supervised machine learning. When different data sources do not have common unique identifying keys, alternative identifying variables such as firm names, addresses or balance sheet figgures can be used to identify different representations of the same real-world unit. Since alternative identifying variables are often not standardized and erroneous, they can be compared using approximate comparison measures such as different string distance metrics. In the presence of training data, and given a number of different similarity measures, an ensemble of several supervised machine learning classification algorithms	Record-Linkage	Testbetrieb	Random Forest, Extra Trees, Gradient Boosting	NA

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
		is used to predict the match probability for a list of match candidate pairs. The evaluation of our machine learning based method shows that the matching process yields suciently precise results as well as a suciently high coverage / recall rate to make full automation of company data linkage feasible for typical use cases in research and analytics.				
Deutsche Bundesbank	Verbesserung des DQM mit Machine Learning	Im Rahmen des Datemqualitätsmanagement sollen Prüfwürdige Meldungen deutscher Banken zur Wertpapierhaltung identifiziert werden.	Identifikation unplausibler Fälle	Testbetrieb	random Forest	NA
IAB FB B2	Typisierung von Arbeitsmarktregionen (Vergleichstypen SGB III / SGB II)	In der zweiten Stufe des Typisierungsverfahrens werden durch den Einsatz von Unsupervised Learning Arbeitsagentur- bzw. Jobcenter-Bezirke geclustert, die ähnliche regionale Arbeitsmarktcharakteristika aufweisen. Hierbei wird das Ward-Verfahren mit einem k-Means-Algorithmus kombiniert.	Clustering	Produktivbetrieb	Clustering nach Ward / k-Means Algorithmus	Stata

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
IAB & Uni Mannheim	Neue Methoden zur Berufsverkodung	Unter Berufsverkodung versteht man die Zuordnung von Freitextantworten aus Umfragen in offizielle Berufsklassifikationen. Die Erzeugung von Vorschlägen zur manuellen Kodierung geschieht bisher teils automatisch, was auf Basis eines Verzeichnisses von Berufsbenennungen und mithilfe von String-Matching-Algorithmen erfolgt. Im Projekt wird getestet, ob alternativ Trainingsdaten von vorherigen Umfragen verwendet werden können um bessere Vorschläge zu generieren. Weiterhin sollen die Vorschläge direkt während des Interviews eingeblendet werden, so dass befragte Personen selber die passendste Kategorie auswählen können.	Klassifikation mit mehreren Klassen	in Entwicklung	Vergleich verschiedener Algorithmen, Stacking, Boosting	R
IAB	Korrektur der Ausbildungsinformation in administrativen Daten	Im Rahmen des Projekts soll ein datengetriebenes Verfahren zur Korrektur der Bildungsinformation in administrativen Individualdaten entwickelt und mit gängigen deterministischen Methoden verglichen werden.	Klassifikation mit mehreren Klassen	in Entwicklung	Decision Trees, Random Forests, SVM, Boosting, Ensembles, Stacking	R
IAB	Vorhersage der Dauer in Arbeitslosigkeit	Im Zuge des Projekts soll die Dauer in Arbeitslosigkeit von Kunden der Bundesagentur für Arbeit vorhergesagt werden. Das Projekt ist explorativ angelegt. Vorläufiges Ziel ist es daher, das Potential von Machine Learning im genannten Kontext zu evaluieren.	Regression	Idee / Machbarkeitsstudie	Decision Trees, Random Forests, SVM, Neuronale Netze, Boosting, Ensembles, Stacking	R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
IAB	Vorhersage der Dauer in Arbeitslosigkeit und Evaluation von Maßnahmen der Vermittlung auf Basis eines Feldexperimentes mit Machine Learning Methoden	Im Rahmen des DFG-Projektes werden kontextbezogene Möglichkeiten für die Prognose der Arbeitslosigkeit mit maschinellen Lernen getestet sowie diese zur Evaluation der Effekte eines Feldexperimentes genutzt.	Regression und Klassifikation	laufendes DFG-Projekt	Decision Trees, Random Forests, SVM, Boosting, Ensembles, Stacking	R
IAB	Umschätzung von Wirtschaftszweigen unter Verwendung von CART Verfahren	Regelmäßige Änderungen in der Klassifikation der Wirtschaftszweige verursachen Probleme bei der Auswertung über längere Zeiträume. Im Rahmen des Projekts werden für das Betriebshistorikpanel des IAB einzelne Klassifikationen über den gesamten Zeitraum des Panels fortgeschrieben, so dass Auswertungen auf einer konsistentn Klassifikation ermöglicht werden	Klassifikation mit mehreren Klassen	in Entwicklung	Decision Trees	R
IAB FB C1 und Universität Bristol	Vorhersage der Arbeitslosigkeitsdauer	Ziel des Projektes ist es mittels Machine Learning Methoden die Arbeitslosigkeit vorherzusagen sowie bereits durchgeführte Experimente über verschiedene Aspekte der Zuweisungsprozesse auf heterogene Treatment Effekte, auch unter Bezug von Machine Learning Methoden zu untersuchen.	Klassifikation und Vorhersage	Produktivbetrieb	Decision Trees, Logit, Lasso, SVM	R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
IAB	Imputation fehlender Informationen zur Arbeitszeit	Infolge der Umstellung des Tätigkeitsmerkmals der Meldungen zur Sozialversicherung (DEÜV) fehlte die Vollzeit/Teilzeit-Angabe in den Beschäftigtenmeldungen in den Jahren 2011 und 2012. Sie wurden mit Hilfe von Classification trees imputiert	Klassifikatin	Abgeschlossen	Classification Trees, cross validation	R
Bundesamt für Migration und Flüchtlinge	Profilanalyse	Mit der Durchführung eines Proof of Concepts und anschließender Entwicklung eines Pilotsystems werden mit dem Projekt Profilanalyse erstmalig Methoden des Machine Learning im Bundesamt für Migration und Flüchtlingen erprobt. Dabei wird eine semantische Textanalyse der Anhörungsniederschriften im Rahmen des Asylverfahrens vorgenomen. Ziel ist es, relevante Textpassagen, die auf sicherheitsrelevante Informationen hindeuten, hervorzuheben und nach vorgegebenen Sicherheitskrieterien zu klassifizieren. Die Analyseergebnisse werden anwenderfreundlich aufbereitet und unterstützen somit die zuständigen Sachbearbeiter der Meldepflicht des BAMF nachzukommen.	Aggregation und Klassifizierung von Textpassagen	Entwicklung eines Pilotsystems	Semantische Textanalyse	Watson Explorer (WEX)

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	Election Campaigning on Social Media: Politicians, Audiences and the Mediation of Political Communication on Facebook and Twitter	Abstract: "Although considerable research has concentrated on online campaigning, it is still unclear how politicians use different social media platforms in political communication. Focusing on the German federal election campaign 2013, this article investigates whether election candidates address the topics most important to the mass audience and to which extent their communication is shaped by the characteristics of Facebook and Twitter. Based on open-ended responses from a representative survey conducted during the election campaign, we train a human-interpretable Bayesian language model to identify political topics. Applying the model to social media messages of candidates and their direct audiences, we find that both prioritize different topics than the mass audience. The analysis also shows that politicians use Facebook and Twitter for different purposes. We relate the various findings to the mediation of political communication on social media induced by the particular characteristics of audiences and sociotechnical environments."	Topic Modeling	Forschungspublikation	semi-supervised classification	Julia

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	When populists become popular: comparing Facebook use by the right-wing movement Pegida and German political parties	Abstract: "Previous research has acknowledged the use of social media in political communication by right-wing populist parties and politicians. Less is known, however, about its pivotal role for right-wing social movements which rely on personalized messages to mobilize supporters and challenge the mainstream party system. This paper analyzes online political communication by the right-wing populist movement Pegida and German political parties. We investigate to which extent parties attract supporters of Pegida, to which extent they address topics similar to Pegida and whether their topic use has become more similar over a period of almost two years. The empirical analysis is based on Facebook posts by main accounts and individual representatives of these political groups. We first show that there are considerable overlaps in the audiences of Pegida and the new challenger in the party system, AfD. Then we use topic models to characterize topic use by party and surveyed crowdworkers to which extent they perceive the identified topics as populist communication. The results show that while Pegida and AfD talk about rather unique topics and smaller parties engage to varying degrees with the topics populists emphasize, the two governing parties CDU and SPD clearly deemphasize those. Overall, the findings indicate that the considerable attention devoted to populist actors and shifts in public opinion due to the refugee crisis have left only moderate marks in political communication within the mainstream party system."	Topic Modeling	Forschungspublikation	LDA	Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	Towards Quantifying Sampling Bias in Network Inference	Abstract: "Relational inference leverages relationships between entities and links in a network to infer information about the network from a small sample. This method is often used when global information about the network is not available or difficult to obtain. However, how reliable is inference from a small labelled sample? How should the network be sampled, and what effect does it have on inference error? How does the structure of the network impact the sampling strategy? We address these questions by systematically examining how network sampling strategy and sample size affect accuracy of relational inference in networks. To this end, we generate a family of synthetic networks where nodes have a binary attribute and a tunable level of homophily. As expected, we find that in heterophilic networks, we can obtain good accuracy when only small samples of the network are initially labelled, regardless of the sampling strategy. Surprisingly, this is not the case for homophilic networks, and sampling strategies that work well in heterophilic networks lead to large inference errors. These findings suggest that the impact of network structure on relational classification is more complex than previously thought."	Relational Classification	Forschungspublikation	Bayes + Relaxation + Collective Inference	Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	How Users Explore Ontologies on the Web: A Study of NCBO's BioPortal Usage Logs	Abstract: "Ontologies in the biomedical domain are numerous, highly specialized and very expensive to develop. Thus, a crucial prerequisite for ontology adoption and reuse is effective support for exploring and finding existing ontologies. Towards that goal, the National Center for Biomedical Ontology (NCBO) has developed BioPortal---an online repository containing more than 500 biomedical ontologies. In 2016, BioPortal represents one of the largest portals for exploration of semantic biomedical vocabularies and terminologies, which is used by many researchers and practitioners. While usage of this portal is high, we know very little about how exactly users search and explore ontologies and what kind of usage patterns or user groups exist in the first place. Deeper insights into user behavior on such portals can provide valuable information to devise strategies for a better support of users in exploring and finding existing ontologies, and thereby enable better ontology reuse. To that end, we study and group users according to their browsing behavior on BioPortal and use data mining techniques to characterize and compare exploration strategies across ontologies. In particular, we were able to identify seven distinct browsing types, all relying on different functionality provided by BioPortal. For example, Search Explorers extensively use the search functionality while Ontology Tree Explorers mainly rely on the class hierarchy for exploring ontologies. Further, we show that specific characteristics of ontologies influence the way users explore and interact with the website. Our results may guide the development of more user-oriented systems for ontology exploration on the Web."	Clustering, dimensionality reduction	Forschungspublikation	K-means + PCA	Python, R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	Practical collapsed stochastic variational inference for the HDP	In this work we explore a collapsed stochastic variational Bayes inference for the Hierarchical Dirichlet process (HDP). The proposed online algorithm is easy to implement and accounts for the inference of hyper-parameters.	Bayesian non-parametric mixed-membership clustering	Forschungspublikation	PCSVB0	Julia
GESIS - Leibniz-Institut für Sozialwissenschaften	Polylingual Labeled Topic Model	Development and evaluation of the Polylingual Labeled Topic Model	Topic Model	Forschungspublikation	PLL-TM	Julia, CML
GESIS - Leibniz-Institut für Sozialwissenschaften	Predicting structured metadata from unstructured metadata	Framework to predict structured metadata terms from unstructured metadata for improving quality and quantity of metadata, using the Gene Expression Omnibus (GEO) microarray database	Metadata prediction	Forschungspublikation	LDA, SVM	Scala, Python
GESIS - Leibniz-Institut für Sozialwissenschaften	Enriching ontologies with encyclopedic background knowledge for document indexing	Using encyclopedic background knowledge for enriching domain-specific ontologies for document classification	Document classification	Forschungspublikation	LLDA, SVM	Scala, Python
GESIS - Leibniz-Institut für Sozialwissenschaften	Measuring Motivations of Crowdworkers: The Multidimensional Crowdworker Motivation Scale	Presents the Multidimensional Crowdworker Motivation Scale (MCMS), a scale for measuring the motivation of crowdworkers on micro-task platforms.	Scale development	Forschungsarbeit	Strukturgleichungsmodelle	R, MPlus, Python, CML

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	A System for Probabilistic Linking of Thesauri and Classification Systems	Presents a system which creates and visualizes probabilistic semantic links between concepts in a thesaurus and classes in a classification system.	Concept linking	Forschungsarbeit	PLL-TM	Julia, CML, D3

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	Why we read Wikipedia	Abstract: "Wikipedia is one of the most popular sites on the Web, with millions of users relying on it to satisfy a broad range of information needs every day. Although it is crucial to understand what exactly these needs are in order to be able to meet them, little is currently known about why users visit Wikipedia. The goal of this paper is to fill this gap by combining a survey of Wikipedia readers with a log-based analysis of user activity. Based on an initial series of user surveys, we build a taxonomy of Wikipedia use cases along several dimensions, capturing users' motivations to visit Wikipedia, the depth of knowledge they are seeking, and their knowledge of the topic of interest prior to visiting Wikipedia. Then, we quantify the prevalence of these use cases via a large-scale user survey conducted on live Wikipedia with almost 30,000 responses. Our analyses highlight the variety of factors driving users to Wikipedia, [...]. Finally, we match survey responses to the respondents' digital traces in Wikipedia's server logs, enabling the discovery of behavioral patterns associated with specific use cases. For instance, we observe long and fast-paced page sequences across topics for users who are bored or exploring randomly, whereas those using Wikipedia for work or school spend more time on individual articles focused on topics such as science. Our findings advance our understanding of reader motivations and behavior on Wikipedia and can have implications for developers aiming to improve Wikipedia's user experience, editors striving to cater to their readers' needs, third-party services (such as search engines) providing access to Wikipedia content, and researchers aiming to build tools such as recommendation engines."	Classification	Forschungspublikation	Gradient Boosting	Python, Spark

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	Predicting Genre Preferences from Cultural and Socio-economic Factors for Music Retrieval	Abstract: "In absence of individual user information, knowledge aboutlarger user groups (e.g., country characteristics) can be exploited forderiving user preferences in order to provide recommendations to users.In this short paper, we study how to mitigate the cold-start problem on acountry level for music retrieval. Specifically, we investigate a large-scaledataset on user listening behavior and show that we can reduce the errorfor predicting the popularity of genres in a country by about 16.4% overa baseline model using cultural and socio-economics indicators."	Regression	Forschungspublikation	Gradient Boosting, Random Forests	Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	Discovering and Characterizing Mobility Patterns in Urban Spaces: A Study of Manhattan Taxi Data	Abstract: "Nowadays, human movement in urban spaces can be traced digitallyin many cases. It can be observed that movement patternsare not constant, but vary across time and space. In this work,we characterize such spatio-temporal patterns with an innovativecombination of two separate approaches that have been utilized forstudying human mobility in the past. First, by using non-negativetensor factorization (NTF), we are able to cluster human behaviorbased on spatio-temporal dimensions. Second, for characterizingthese clusters, we propose to use HypTrails, a Bayesian approachfor expressing and comparing hypotheses about human trails. Toformalize hypotheses, we utilize publicly available Web data (i.e.,Foursquare and census data). By studying taxi data in Manhattan,we can discover and characterize human mobility patterns that cannotbe identified in a collective analysis. As one example, we finda group of taxi rides that end at locations with a high number ofparty venues on weekend nights. Our findings argue for a morefine-grained analysis of human mobility in order to make informeddecisions for e.g., enhancing urban structures, tailored traffic controland location-based recommender systems."	Clustering	Forschungspublikation	Tensor Factorization	Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	Extracting Semantics from Random Walks on Wikipedia: Comparing Learning and Counting Methods.	Abstract: "Semantic relatedness between words has been extracted from a variety of sources. In this ongoing work, we explore and compare several options for determining if semantic relatedness can be extracted from navigation structures in Wikipedia. In that direction, we first investigate the potential of representation learning techniques such as DeepWalk in comparison to previously applied methods based on counting co-occurrences. Since both methods are based on (random) paths in the network, we also study different approaches to generate paths from Wikipedia link structure. For this task, we do not only consider the link structure of Wikipedia, but also actual navigation behavior of users. Finally, we analyze if semantics can also be extracted from smaller subsets of the Wikipedia link network. As a result we find that representa- tion learning techniques mostly outperform the investigated co-occurrence counting methods on the Wikipedia network. However, we find that this is not the case for paths sampled from human navigation behavior."	Semantic Relatedness	Forschungspublikation	Deep Learning	Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	Text Categorization for Deriving the Application Quality in Enterprises Using Ticketing Systems	Abstract: "Today's enterprise services and business applications are often centralized in a small number of data centers. Employees located at branches and side offices access the computing infrastructure via the internet using thin client architectures. The task to provide a good application quality to the employers using a multitude of different applications and access networks has thus become complex. Enterprises have to be able to identify resource bottlenecks and applications with a poor performance quickly to take appropriate countermeasures and enable a good application quality for their employees. Ticketing systems within an enterprise use large databases for collecting complaints and problems of the users over a long period of time and thus are an interesting starting point to identify performance problems. However, manual categorization of tickets comes with a high workload.In this paper, we analyze in a case study the applicability of supervised learning algorithms for the automatic identification of relevant tickets, i.e., tickets indicating problematic applications. In that regard, we evaluate different classification algorithms using 12,000 manually annotated tickets accumulated in July 2013 at the ticketing system of a nation-wide operating enterprise. In addition to traditional machine learning metrics, we also analyze the performance of the different classifiers on business-relevant metrics."	Textklassifikation	Forschungspublikation	SVM, Decision Tree, etc...	Java

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
GESIS - Leibniz-Institut für Sozialwissenschaften	RDF Vocabulary Term Recommendation	Abstract: "Deciding which RDF vocabulary terms to use when modeling data as Linked Open Data (LOD) is far from trivial. In this paper, we propose TermPicker as a novel approach enabling vocabulary reuse by recommending vocabulary terms based on various features of a term. These features include the term's popularity, whether it is from an already used vocabulary, and the so-called schema-level pattern (SLP) feature that exploits which terms other data providers on the LOD cloud use to describe their data. We apply Learning To Rank to establish a ranking model for vocabulary terms based on the utilized features. The results show that using the SLP-feature improves the recommendation quality by 29–36 % considering the Mean Average Precision and the Mean Reciprocal Rank at the first five positions compared to recommendations based on solely the term's popularity and whether it is from an already used vocabulary."	Learning To Rank	Forschungspublikation	Various Learning To Rank algorithms from the RankLib library	RankLib Library (Java)

10.2.3 Rückmeldungen internationaler Institutionen

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Australian Bureau of Statistics	Automated coding for complex classifications		multilevel multiclass classification	productive	SVM, bootstrap aggregation, text models	c, java
Australian Bureau of Statistics	Linking and grouping of statistical entities by their pattern of connections		Graph structure classification, motif detection	experiment	SVM, graph kernel models	R
Australian Bureau of Statistics	Land utilisation and crop prediction		multiclass classification	experiment	CNN	R
Australian Bureau of Statistics	Editing administrative datasets		multiclass prediction, anomaly detection	experiment	probabilistic graphical models	R
Australian Bureau of Statistics	Predicting Census occupancy		binary classification	experiment	tree models	R
Australian Bureau of Statistics	Linking mutliple datasets		Entitly resolution creation of a linking spine	experiment	empirical bayes	R, Scala
Central Statistical Bureau of Latvia	Number of population and key demographic indicators		binary classification	productive	Logistic regression (GLM) is used in production. Stochastic Gradient Boosting (GBM), Support Vector Machines (SVM), Regularized Discriminant Analysis (RDA), Multi-Layer Perceptron (MLP), Radial Basis Function Network (RBF) were tested during the implementation	R (data.table)

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
					stage.	
Central Statistics Office of Ireland	Automatic coding via open-source indexing utility	An automatic coding system for Classification of Individual Consumption by Purpose (COICOP) assignment for their Household Budget Survey, using previously coded records as training data. Their method is based on the open-source indexing and searching tool Apache Lucene (http://lucene.apache.org).	multi-class classification	Development		Apache Lucene / Python
Eurostat	Categorical data imputation via neural networks and Bayesian networks	Eurostat compared imputation results for missing categorical data (voting intensions) based on two machine learning methods (neural networks and Bayesian networks) against one of the current prevailing statistical imputation methods (multiple imputation using logistic regression).	imputation	experiment	Logistische Regression, Neural Networks, Bayesian Networks	(SAS)
Federal Statistical Office of Switzerland	Modelling of the non-response mechanisme	Classification trees are used to model the behaviour of the non-respondents in order to diminish non-response bias in the results.	binary classification of response homogeneity groups	productive	CHAID	SAS

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Federal Statistical Office of Switzerland	Detection of suspicious responses	Several machine learning algorithms are tested to detect anomalies in the data. Detected units are contacted to check the data again. These methods are applied in a field where no or very few edit rules are available.	binary classification	test	Generalized boosted models, Random forests, Neural networks, Naive Bayes, Tree algorithmes, ...	R
Federal Statistical Office of Switzerland	Turnover breakdown from grouped answers to the enterprises	This project is at its beginning. It is planned to use random forests to learn from the Turnover statistics and to apply it to the units not in the Turnover statistics. There is no other data available to breakdown grouped VAT data.	classification	test	Random forests	R
Federal Statistical Office of Switzerland	Automatisation of the NOGA (Swiss NACE) coding	This project is at its very beginning. It is planned to test several machine learning algorithms which are not yet fixed.	classification	test	several	R
Federal Statistical Office of Switzerland	Automatisation of land cover and land use codes based on aerial images	This project is at its beginning. It is planned to use convolutional neural networks for coding or for change detection.	classification	test	CNN	R/Python
Federal Statistical Office of Switzerland	Clustering of the "careers" in the social security system	This project is at its beginning. It is planned to test whether it is possible to detect similar "careers" in the social security system automatically.	clustering	test	several	R
Hungarian Central Statistical Office	Tax evader detection	Detecting self-employed proprietors who are tax evaders	classification	experiment	k-NN	
Italian National Institute of Statistics	Substitutes for surveys via internet scraping	Research regarding the possibility of substituting (fully or partially) surveys by collecting data via internet scraping and extracting information therein using machine learning methods.		experiment	Naive Bayes und andere	R
National Institute of Statistics Romania	Use of administrative data in business statistics	Efficient integration of administrative data into the statistical process implies finding and resolving data quality issues. Tree-based methods are used for imputation of the turnover variable for businesses from the value added tax (VAT) administrative data.	data imputation	experimental	Tree-Based Methods: Regression Trees, Random Forests, Boosting	R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
National Institute of Statistics Romania	Price index estimation using web scraped data	Levenshtein distance between two strings is the number of deletions, insertions or substitutions required to transform source string into target string. This string matching technique is necessary for automatic classification of products names into categories and across periods.	string matching	experimental	Levenshtein distance	R
National Institute of Statistics Romania	Action plan for EU-SILC improvements	The data matching methods have been used in order to increase the quality of EUSILC sample bringing together information from different data sources: sample survey and administrative registers.	poverty indicators	productive	Random forests	R
National Institute of Statistics Romania	Modeling the potential human capital on the labor market	Creating the profile of two categories of potential human capital by modelling the relationship between economically inactive persons who are seeking for a job, but are not immediately available to start working, respectively economically inactive persons who are not seeking for a job, but are immediately available to start working, and some socio-economic predictors. The aim is to identify the impediments which determine inactive people not to become active on the labour market.	classification	eigentlich kein ML; experiment	logistic regression	R
National Statistics Center, Japan	Supervised multiclass classifier for an autocoding for the Family Income and Expenditure survey	Multiclass classifier that can classify Japanese short text descriptions according to their corresponding classification codes has been developed for the Family Income and Expenditure survey in Japan. The concept of the naïve Bayes classifier is borrowed for the algorithm of the classifier. The classifier is also applicable to English text descriptions and other classifications tasks.	multiple classification	experiment	naïve Bayes	Perl / R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
STATEC (Luxemburg)	Scanner Data	Machine Learning Ranking (MLR) is used in the Scanner Data project to classify the individual items into COICOP (Classification of individual consumption according to purpose), which is the standard classification used for compiling a CPI. It allows to use a larger sets of incoming data without increasing costs of manuel processing.	classification	productive	MLR	Solr / Java / Talend
STATEC (Luxemburg)	Business Enterprise Research and Development (BERD)	An in-house developed ensemble classifier assesses for each survey respondent the probability of performing intra-mural R&D activities during the reference year. The probability is conditional on the available data, which includes current and past survey data (including unstructured text) as well as administrative sources. The model results are used in the survey data validation process to spot item non-response and inaccurate responses on the intra-mural R&D variables. Such respondents are then individually contacted by the statistical analysts, if necessary. Albeit the use of an ensemble classifier, the model results remain fully interpretable.	classification	productive	model stacking	KNIME, R
Statistics Austria	Imputation	In various surveys and also in projects with administrative data missing values should be imputed.	imputation	productive	kNN	R
Statistics Austria	Statistical matching	On the basis of common variables two data sets are matched (very similar to imputation.)	Statistical matching	productive	random forest, kNN	R
Statistics Austria	Estimating AROPE for the Austrian rich frame		Estimation/Classification	productive	Boosting Trees Algorithmus	R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Austria	Estimating a SILC-like income based on adminstrative data	For a couple of years now, the household income in the ICT survey is not asked for bute estimated based on adminstrative data and the SILC data.	Estimation/Regression	productive	random forest	R
Statistics Belgium		Machine learning is used to predict the NACE code of a job vacancy based on the job description. We use administrative databases (from national jobs portals) to model the link between words in the description and the NACE code. We expect to apply this model to scrap job vacancies from internet job portals in Belgium.	binary classification	Test	SVM,	R (RtxtTools)
Statistics Canada	Consumer Prices scanner data use	Consumer Prices: retail scanner data classification to the Consumer Price Index (CPI) commodities classification, used to suggest CPI product substitutions. Currently, in production parallel run.	Classification	In production parallel run	SVM	Python
Statistics Canada	Retail scanner data use for Monthly retail trade survey and Quarterly retail commodity survey	Machine learning text classification is used to obtain the NAPCS code of each product sold within the retail scanner data, and obtain aggregate sales for each NAPCS, aggregate sales by area/ postal code. Proof of Concept (PoC) completed.	Classification	Transitioning to production	XGBoost linear, with bag of words character n-grams model	R
Statistics Canada	International Trade data Outliers Detection	Outliers are a major problem in the international trade data, in particular for the quantity variable. Errors include unit errors, 0 or 1 entered instead of a proper quantity,etc. Current system is based on unit value (UV) clipping, manual checking (including "unclipping"), and an approval process. Machine learning (ML) is used to automate this process. ML is also used to reconstruct/ impute the original value of the flagged points.	Outlier detection	The outlier detection experiment is completed. The imputation work is in progress.	XGBoost tree model	R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Canada	Business Activity, Expenditure and Output (BAEO) survey comments text mining	The BAEO survey receives close to 9000 comments. ML was used to classify the comments into 8 action categories.	Classification	Experiment completed	Linear SVM, with bag of words model	R
Statistics Canada	Payments data feasibility project	ML is used to classify the payment transactions into standard statistical classification concepts (NAICS, COICOP, etc). The project's goal is to determine if the data can be used to produce information related to retail sales, household consumption, digital transactions, and tourism statistics. ML might be also used for imputation of misisng values.	Classification, Imputation	At experimentation stage	SVM	R
Statistics Canada	Transport statistics: Trucking Commodity Origin and Destination Survey (TCOD)	ML will be used to classify the electronically reported data. In the context of the TCOD survey redesign, the use of auxiliary sources such as GPS data, satellite imagery is evaluated as a new source for trucking data analysis, including linkage to a specific business on the BR.	Text classification, image classification, predictive modelling for misisng information based on auxiliary data, record linkage	At experimentation stage	recently started	R, Python
Statistics Canada	Entreprise statistics	Web scraping for large enterprises to complement the survey data sources with on-line data (newsfeed, financial reports, company Web sites) to enhance data coherence, improve profiling, prepare the Entreprise Portfolio Managers' visits to companies.	Data extraction, Natural Language Processing, Record Linkage	At experimentation stage	recently started	R, Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Canada	Web scraping to enhance the research and development survey frames	Web scraping is used to supplement the R&D survey data with on-line data to identify adequately the survey population, and develop complementary indicators for innovation.	Data extraction, Natural Language Processing, Record Linkage	At experimentation stage	recently started	R, Python
Statistics Canada	Retail statistics	Web scraping and google map for retail store information	Data extraction, Natural Language Processing, Record Linkage	At experimentation stage	web scraping libraries	Python
Statistics Canada	Manufacturing industries and producer prices commodities	Web scraping to obtain the list of commodities being produced by manufacturing companies from catalogues on their Web sites or other existing on-line data bases. Code the products to the NAPCS classification using ML.	Data extraction, Classification to NAPCS	At test stage, experimentation to start in May'18	TBD	R, Python
Statistics Canada	Agriculture crop yield estimates	Phase 1 : The model-based crop estimates provide provincial and national yield and production estimates for principal field crops in Canada. The model utilizes data from low resolution satellite imagery, historical field crop survey estimates, and agroclimatic information. Phase 2: Develop in season crop area and yield estimates, combining crop insurance, remote sensing and other business intelligence including supply and disposition data and historical patterns to identify likely crop cover at the field level and then employing historical data, trends and weather data to produce yield estimates.	Predictive modelling	Phase 1 in production since 2015 Phase 2 (November crop yield and area) at Idea stage, to start in May'18	Phase 1 : LASSO and others Phase 2 : TBD	mostly SAS for phase 1; TBD for phase 2

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Canada	Canadian Housing Statistics Program, project on citizenship and country of birth	Partial information about citizenship and country of birth can be found in different databases such as the Census of Population, the Social Insurance Number Registry, immigration data, tax data, and so on. Standardize and integrate the existing information, use ML to impute missing information.	Classification, Imputation	Idea, to start in May'18	TBD	TBD
Statistics Canada	Exploration of machine learning to create indicators on tourism spending	Model a set of early indicators on international tourism spending in Canada based on survey data and payment processor data from debit and credit cards.	regression	idea	TBD	TBD
Statistics Canada	Economic Analysis: Identifying high growth firms and firm failures	Administrative data over a three-year time span is commonly used to identify high growth firms. This project uses ML techniques to develop more timely estimates of high growth firms. Similarly, firm peformance is hypothesized to deteriorate years before exit (shadow of death effect). ML techniques are also applied to develop more timely estimates of firm exits.	Classification	Idea, to start in May 2018	Random Forest	R
Statistics Canada	Economic Analysis: Identifying firm networks	Using ML to identify firm networks from Business Register and origin and destination of shipments data so that supply chains and intra- and intra-firm trade can be analysed	Classification/Network Analysis	Initial assessment completed. Additional data needed to improve algorithm.	TBD	TBD
Statistics Canada	Communications and Dissemination: improving user experience on the web	Live chat featured for visitors based on their navigation patterns	chatbots	Idea	TBD	TBD

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Canada	Communications and Dissemination: improving user experience on the web	Use AI to analyze content web site visitors based on their past content consumption / navigation, or to offer live chat.	TBD	Idea, to start in April 2018	TBD	TBD
Statistics Canada	Predicting Mortality Rates	Investigate how sociodemographic and health-related factors contribute to life expectancy. Constructing well-specified predictive models can be extremely time consuming and labor intensive. Data driven approaches, like machine learning techniques, are gaining popularity. These algorithms may be able to gain insightful information about what predicts an outcome by iteratively learning from data, instead of being explicitly directed by theory ML solution: Trial various AI/machine learning techniques, including support vector machines and neural networks, to predict mortality in the Census linked CANCHEC cohorts. These are large cohorts that contain health outcomes, but not necessarily health exposure data. As Statistics Canadas holdings of such data increase we need to explore alternative methods to construct robust predictive models/analytic tools, when we are lacking key health exposure variables.	Predictive modelling	Idea to start in FY 2018	TBD	TBD

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Canada	Harvesting key data on causes of death from narrative descriptions	Harvesting key data on causes of death and abuse from narrative descriptions Exploring narrative descriptions to harvest statistical information (cause of death in coroners' reports; abuse case in social workers' narrative description) A machine learning application is being used to mine more rapidly and efficiently the unstructured narratives that coroners include in their reports and that detail the circumstances of the deaths. These narratives are included in our Canadian Coroners and Medical Examiners Database (CCMED). Its first goal is to identify opioid-related deaths. In the longer-term, it is hoped that machine learning/Artificial Intelligence would permit to rapidly recognize any patterns that would point to another crisis or specific circumstances that affect many investigated death cases.	Classification	Exploration part 1 completed (promising approach with limited success due to weak training data); Literature review completed for future exploration	various (SVM, NN, adabost Naïve Bayes)	Python
Statistics Canada	Victimisation studies (Harvesting key data from narrative descriptions)	Exploring narrative descriptions to harvest statistical information (abuse case in social workers' narrative description)	Classification of narrtive description	Idea, to start in Sprin/Summer 2018 (data dependent)	Text recognition (possibly), Natural language processing, TBD	TBD
Statistics Canada	R&D for generalised systems	Use of genetic algorithm (AI based on natural selection) for sample allocation	sample allocation/selection	Proof of concept completed	Genetic Algorithm	R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Canada	Census	Immigration admission category variables were added to the 2016 Census through record linkage rather than collecting them from respondents. Data were not available for some individuals and had to be inferred from other characteristics provided by them. Machine Learning was used to identify the best combination of characteristics to make these inferences.	Imputation	Completed and used for Census 2016 production	ReliefF algorithm for feature selection and weigthing for neirest neighbor imputation	R for feature selection and weigthing; model used in "regular" production system (CANCEIS)
Statistics Canada	Study of comments submitted during the 2016 Census content consultation	Used text mining software to analyze over 1.1 million comments compiled from the 2016 Census content consultation to inform possible content and questionnaire changes for 2021.	key term identification and binary classification	productive	Natural Language Processing	SAS JMP
Statistics Canada	Census and others	Exploring information provided in comments box in the Census with focus on non-binary gender self-identification in open text/comment box	Classification	Exploration	Unsupervised ML to cluster comments	SAS Enterprise Miner
Statistics Canada	R&D for imputation method	Hackathon for advanced imputation methods using AI. Use use is imputation of a large admin files where adata can be entered for generic or detailed financial item. When generic is used, data must be imputed for detailed item.	Imputation	Exploration via hackathon planned in May 2018	various	various
Statistics Canada	Exploration of machine learning for coding of industry and occupation text descriptions	Opportunities to use Machine Learning and Artifical Intelligence to improve the effectiveness of automated coding of survey responses is being investigated. This would have a number of applications in survey and administrative data programs, including the Labour Force Survey. For example, new methods could permit the collection of additional open-ended information related to task descriptions of jobs and skills profiles of individuals.	statistical coding	exploratory / idea generation	various	various

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Canada	Census Program Transformation Project	Exploring how AI could be applied in data linkage processes in the building of statistical registers.	to be defined	Idea	TBD	TBD
Statistics Canada	Creating synthetic data for microsimulation	Microsimulation models require complete data. Currently in the Population Health Microsimulation Model (POHEM) we do not model measured data such as that collected Canadian Health Measures Survey (CHMS). Having measured data in POHEM is a desirable attribute if the model is to be used for evidence-based policy analysis around cardiovascular disease and other chronic diseases. AI solution: Use AI/machine learning techniques to match data from the CHMS to the Canadian Community Health Survey (CCHS) in order to initialize a starting POHEM population with measured health data.	Classification, matching	Idea to start in FY 2018	Likely K nearest neighbours technique	R, kknn package
Statistics Canada	Synthetic Data File	Examination of machine learning algorithms for creation of synthetic data for disclosure control	synthetic data (disclosure control)	Exploration	CART models and random forests	R
Statistics Canada	Priorisation score for collection	Creating a Priorization Score in CATI Surveys: Analysis of the Bayesian Hierarchical Rule Modeling		Exploration / proof of concept in progress	Bayesian Hierarchical	SAS

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Canada	Simulation of administrative data	Modeling of administrative data which can incorporate both the relationships between variables (correlations) and the longitudinal dependences are requirements to provide realistic simulated data for statistical purposes. Current methodology and available tools apply Gibbs Sampling or Bayesian networks to model all the conditional distributions. Machine learning algorithms may also provide a solution. For example, neural networks are capable of modeling complex correlations amongst variables. Such networks could be extended to deep networks in the case of large, complicated data sets.	Modelling	Idea	TBD	TBD
Statistics Canada	Automated extraction of features for record-linkage	The goal is to automate the selection of features for record-linkage. Currently these features are selected manually based on expert knowledge.	classification, feature selection	Initial exploration	supervised, unsupervised, dimension reduction techniques (e.g. PCA)	SAS
Statistics Canada	Development of the record linkage software G-Link	Supervised and Semi-supervised Machine Learning methods are used for the develop automatic thresholds in the Felleigi-Sunter methodology. The k-means method and the two-step method (k-means and probit) will be implemented in the version 3.4 of G-Link. In this way, Statistics Canada is creating an opportunity to replace costly searches carried out by humans through manual review by machine learning techniques.	Probabilistic record linkage with Fellegi-Sunter methodology	Development	Unsupervised: K-Means Supervised: two-step method K-means + Probit	G-Link (coded in SAS)

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Canada	Machine Learning for record linkage	Similarities metrics such as Jaro, Jaro-Winkler, Fuzzy Winkler, Fuzzy Jaro are developed in G-Link and ready to be used by Machine Learning methods outside of G-Link. We are exploring the use of the SAS enterprise Miner software. In this way, Statistics Canada is creating an opportunity to examine and compare the Fellegi-Sunter classifier with alternative methods (SVM, neural networks etc.)	Binary classification	Exploration	SVM, Neural network, Supervised Logistic	SAS EM
Statistics Canada	NAICS/NOC autocoder	Automatic data classification (various classifications including occupation and industry)	Classification to NAICS/NOC	Refining the model for increased accuracy, to be implemented in Python	Bag of words and statistical classifier	Perl, Python
Statistics Canada	Exploration of machine learning to identify driving under the influence by type of substance	Social media scrapping for estimating the prevalence of driving under the influence by type of substance.	classification	idea	TBD	TBD
Statistics Denmark	Imputation of educational status for immigrants		imputation		Random Forest	R (missForest)
Statistics Finland	Machine reading accident reports	Free-text road traffic accident reports from the Police are classified to those which caused personal injuries and those which did not	binary classification	productive	tf-idf + logistic regression	Python
Statistics Finland	Automatic coding of industry and occupation	Random forest classifiers are used to automatically classify Finnish Labour Force Survey respondents to correct industry and occupation (NACE, ISCO) based on combined register and survey data	multi-class classification	development	random forest	Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Iceland	Assignment of fine-grained product ids based on product descriptions		classification	in review	Random forest	Python
Statistics Iceland	Increased automation in data processing		Varied	planning	N/A	N/A
Statistics Netherlands		Clustering and classification of SMEs (small and medium enterprises) based on website descriptions	Clust	research	k-means	
Statistics Netherlands		Classification of clothing for CPI	class	test	Random Forests	
Statistics Netherlands		Imputation of economic data	imp	research	Random Forests	
Statistics Norway	Exploration on Random Forest for editing purposes in register based salary statistics		classification		Random Forest	R
Statistics Poland	Detection of agricultural crops	The goal is to identify crop types based on Sentinel-1 and Sentinel-2 satellite images. Different algorithms have been tested for the detection of crop types, including Support Vector Machine (SVM), Decision Trees (DT), K-Nearest Neighbours (KNN) with the following classification parameters: Sigma, Entropy, Alfa, multi-temporal indicators, Wishard distribution but the highest accuraccy is based on KNN with wishard distribution.	segmentation	pilot	KNN, SVM	MTSar, ArcGIS

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Poland	Life satisfaction	The goal of the use case is to deliver data on life satisfaction - 1.happy, 2.neutral, 3.calm, 4.upset, 5.depressed and 6.discouraged. The goal is to support the data from EU-SILC survey with more recent data. The major drawback from this case study is that the dataset may not be representative. The methodology includes Machine learning – supervised learning and Web scraping – we use Twitter API to gather and process the data.	classification	pilot	NB	Python, MongoDB, Apache Spark
Statistics Portugal	Big Data ESSNet		multiclass classification	test	SVM (linear kernel)	Python (SciKit-Learn library in a Jupyter Notebook environment)
Statistics Portugal	Big Data ESSNet		multiclass classification	test	Perceptron (linear)	Python (SciKit-Learn library in a Jupyter Notebook environment)
Statistics Portugal	Big Data ESSNet		multiclass classification	test	Neural Network Model for language identification	R Package “cld3” (Google's Compact Language Detector 3)
Statistics Portugal	Identification of error-containing records via classification trees	A method based on classification trees for error detection in foreign trade transaction data collected by the Portuguese Institute of Statistics.	error detection	experiment	Decision Tree	
Statistics Spain (INE)	UFAES (new methodology)	Random forests are used to model questionnary data from admin data. This model assists the design-based estimator in a probability sampling survey in order to reduce the sample size.	algorithm-assisted survey sampling	research with real data	Random Forests	R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Statistics Spain (INE)	Selective Editing of Quantitative Variables	Different predictive techniques are to be explored to predict anticipated values in the optimization approach to selective editing of quantitative variables developed at Statistics Spain (INE)	prediction	Planning stage	Random Forests, SVMs, Spline Regression, kNN Regression	R
Statistics Spain (INE)	Selective Editing of Qualitative Variables	Different classification techniques are to be explored to classify influential/non-influential units in the optimization approach to selective editing of qualitative variables under development at Statistics Spain (INE)	binary classification	Planning stage	Random Forests, SVMs, Logistic Regression, kNN Regression	R
Statistics Sweden	Essnet big data WP2 - Web scraping enterprise characteristics - Use case of Job advertisement		binary classification	Under development	SVC, DecisionTree NavieBayes Keras sequential NN	Python
Statistics Sweden	Essnet big data WP2 - Web scraping enterprise characteristics - Use case of NACE		multilabel classification	Under development	Keras sequential NN	Python
Statistics Sweden	Automatic coding of occupation title using machine learning methods		binary classification	implementation in production	k Nearest Neigbour	C# and R
Stats NZ	Assignment of geographic region to individuals	Decision Trees and random forest methods used to assign the correct Territorial Authority to individuals given a range of administrative data sources	Classification	Test	Decision Tree	R
Stats NZ	Classification of Building Consents with Natural Language Processing	A generalised linear model has been trained on historical consents data to perform classification of building consents into multiple classes. The model uses a bag of words approach, and is currently being brought into production.	multi-level classification	Tested, not yet in production	Supervised Learning: Generalised Linear Model	R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Stats NZ	Classification of building consents	We used manually coded data to train a model that categorises building consents based on a free text field (job description, as supplied by the applicant). We predict several variables, including building type (which has 29 possible values).	classification	experiment	Generalised linear model using vectorised n-grams	R (glmnet and text2vec)
Stats NZ	Coding of industry classification	Early days in thinking	Classification	Idea		
Stats NZ	Automatic coding of census variables via Support Vector Machines	Investigation of the potential of using Support Vector Machines (SVM) to improve coding of item responses in their Census. They applied SVM to code the variables Occupation and Post-school Qualification, using two disjoint sets of observations, each of size 10,000, from Census 2013 data for training and testing.	multi-class classification	experiment	SVM	
Stats NZ	Imputation via Classification and Regression Trees	Investigation in the use of CART to predict two binary variables based on Census 2013 data. The binary variables were 1. the missingness of the income variable, and 2. the response to the question of whether the respondent has moved since the previous census. Results of this investigation are being evaluated.	classification / imputation	experiment	Decision Tree	

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
Stats NZ	Determination of imputation matching variables	Statistics New Zealand is redesigning the editing and imputation methodology of their Household Economic Survey (HES). Their current proposed methodology will use the Canadian Census Edit and Imputation System (CANCEIS). The imputation module of CANCEIS is based on the Nearest Neighbour Imputation Methodology, which requires user specification of a distance measure of pairs of units based on a number of "matching variables" as well as weights which defines the relative importance of these matching variables. The weight of a matching variable should reflect its strength as a predictor for the variables to be imputed. Statistics New Zealand has reported promising results in using Random Forests to select the set of matching variables for CANCEIS, as well as their weights.	imputation / variable selection	experiment	Random Forest	
Stats NZ	Creation of homogeneous imputation classes	Comparison of two methods (CART, predictive mean stratification) for creating homogeneous imputation classes.	imputation	experiment	Decision Tree	R
Stats NZ	Derivation of edit rules	Investigation of the potential of using association analysis to derive additional edit rules to enhance the processing of census data.	editing	experiment	association analysis	
U.S. Bureau of Economic Analysis	Nowcasting of Source Data for Advance GDP Estimates	This pilot effort aims to reduce revisions in key GDP components by improving trending methods. By training an ensemble of ML models (e.g. LASSO, Ridge, Random Forest, etc) using a variety of alternative data, nowcasted predictions are produced for certain source data series in time for the advance estimate of national GDP. Note that this project is currently being evaluated in parallel with the current estimate process.	regression	pilot	Ensemble	R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
U.S. Bureau of Labor Statistics	Automatic coding of worker injury narratives for the Survey of Occupational Injuries and Illnesses	Each year the Survey of Occupational Injuries and Illnesses (SOII) collects hundreds of thousands of written narratives describing work related injuries and illnesses. In order to produce statistics from this information each of these narratives receives 6 of several thousand possible codes to indicate the occupation of the worker and various characteristics of the incident. To improve the consistency and efficiency of this manual effort BLS developed and evaluated a variety of automated approaches, settling initially on regularized multinomial logistic regression. BLS found automated coding with regularized logistic regression produced more accurate coding than trained human coders, even after the human codes had the benefit of several layers of review. As a result BLS began using this technique to automatically assign codes starting with 2014 data. BLS is now automatically assigning nearly two-thirds of all SOII codes and has recently developed new deep neural network models that provide even better performance. These neural network models are currently used to identify suspicious codes for review, and are likely to be deployed for production autocoding later this year.	multinomial text classification	production	regularized logistic regression, deep neural networks	Python, scikit-learn, tensorflow, keras

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
U.S. Bureau of Labor Statistics	Automatic coding and review of occupation narratives for the Occupational Requirements Survey	One of the key data elements collected by the Occupational Requirements Survey (ORS) is the occupation classification, which is recorded in part, in text format. To more efficiently and effectively validate this data BLS is investigating the application of the same machine learning techniques now successfully being used for the SOII to assist in the review and validation of ORS occupation data. Initial results show promise. BLS will also be investigating whether occupation data from other surveys (like the SOII, and the National Compensation Survey), can be used to improve the performance of the ORS automatic reviewing system.	multinomial text classification	research	regularized logistic regression	Python, scikit-learn
U.S. Bureau of Labor Statistics	Automatic extraction of benefits information from Summary of Benefits and Coverage documents.	Health insurance benefits are an important component of worker's compensation and are often described in a semi-structured document called the "Summary of Benefits and Coverage". This project aims to use machine learning to automatically extract benefits information from these documents with the goal of eventually using this information to augment the National Compensation Survey. Initial results demonstrate that we can automatically extract some of this information at very high accuracy.	text classification, information extraction	research	random forests	Python, scikit-learn

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
U.S. Bureau of Labor Statistics	Automatic linkage of fatal injury case information to OSHA records	To produce statistics about fatal occupational injuries in the U.S., the Census of Fatal Occupational Injuries collects and combines information from a wide variety of sources including local, state, and federal government agencies and U.S. media. One of the biggest sources of official information is data from the Occupational Safety and Health Administration, which often investigates fatal work related injuries. One of the key challenges in incorporating this information is figuring out whether an OSHA record corresponds to a record already in the master file, and if so, which one. Often, this must be accomplished even without imperfect identifiers like decedent and establishment name. To address this issue this project uses machine learning, trained on previously linked documents, to automatically determine whether an OSHA investigation document should be linked to an already partially collected case, or represents a new case that should be added to the master file. By combining a variety of noisy and sometimes missing signals including information about the age of the decedent, the date of injury, the location of the incident, and the description of the incident, we can successfully automatically link OSHA records to the master file even without typical identifiers like decedent and establishment name. When this information is available however, our model can link these documents even more effectively. The system is likely to get production use later this summer.	record linkage	research	random forests	Python, scikit-learn

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
U.S. Bureau of Labor Statistics	Automatic linkage of fatal injury case information to webpage articles.	To produce statistics about fatal occupational injuries in the U.S., the Census of Fatal Occupational Injuries collects and combines information from a wide variety of sources including online media such as news articles. One of the key challenges in incorporating this information is simply finding and matching it to existing case information, often in the absence of even flawed identifiers like decedent name. To address this issue this project uses machine learning, trained on previously linked documents, to automatically determine whether an automatically collected webpage article should be linked to a case already in the master file, or represents a new case that should be added. By combining a variety of noisy signals and by automatically extracting name, date, and establishment information from the article, we hope to be able to conduct this linkage automatically. Systems have already been built to automatically collect these webpages, separate the article text from the rest of the webpage, and automatically extract information like the names of people and companies mentioned in the articles, but more work remains to be done to automatically separate relevant and irrelevant articles.	record linkage	research	random forests	Python, scikit-learn, spacy, Apache Tika
U.S. Bureau of Labor Statistics	Occupational Employment Statistic (OES) Occupation Autocoding	Division of Occupational of Employment Statistics (OES) uses Multinomial Logistic Regression with Stochastic Gradient Descent to develop a model to assign occupation codes to job titles received with employer survey responses.	Job title and other available information into one of many occupation codes	development and testing	LR using SGD	Python

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
U.S. Bureau of Labor Statistics	Analysis of nonresponse to the Occupational Employment Statistics (OES) Survey	To gain insight into how characteristics of an establishment are associated with nonresponse, a recursive partitioning algorithm is applied to the Occupational Employment Statistics survey data to build a regression tree. The tree models an establishment's propensity to respond to the survey given certain establishment characteristics. It provides mutually exclusive cells based on the characteristics with homogeneous response propensities. This makes it easy to identify interpretable associations between the characteristic variables and an establishment's propensity to respond, something not easily done using a logistic regression propensity model. This representation is then used along with frame-level administrative wage data linked to sample data to investigate the possibility of nonresponse bias. We show that without proper adjustments the nonresponse does pose a risk of bias and is possibly nonignorable.	Nonresponse propensity modeling	research	regression trees	R

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
U.S. Bureau of Labor Statistics	Analysis of nonresponse to the Occupational Employment Statistics (OES) Survey	Auxiliary information can increase the efficiency of survey estimators through an assisting model when the model captures some of the relationship between the auxiliary data and the study variables. Despite their superior properties, model-assisted estimators are rarely used in anything but their simplest form by statistical agencies to produce official statistics. This is due to the fact that the more complicated models that have been used in model-assisted estimation are often ill suited to the available auxiliary data. Under a model-assisted framework, we propose a regression tree estimator for a finite population total. Regression tree models are adept at handling the type of auxiliary data usually available in the sampling frame and provide a model that is easy to explain and justify. The estimator can be viewed as a post-stratification estimator where the post-strata are automatically selected by the recursive partitioning algorithm of the regression tree. We establish consistency of the regression tree estimator and a variance estimator, along with asymptotic normality of the regression tree estimator. We then compare the performance of our estimator and the coverage of the confidence intervals using our variance estimator to other survey estimators using US Bureau of Labor Statistics Occupational Employment Statistics Survey data.	model assisted estimation	research	regression trees	R (mase package)

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
U.S. Bureau of Labor Statistics	Analysis of nonresponse to the Longitudinal Occupational Employment Statistics (OES) Survey	This article introduces and discusses a method for conducting an analysis of nonresponse for a longitudinal establishment survey using regression trees. The methodology consists of three parts: analysis during the frame refinement and enrollment phases, common in longitudinal surveys; analysis of the effect of time on response rates during data collection; and analysis of the potential for nonresponse bias. For all three analyses, regression tree models are used to identify establishment characteristics and subgroups of establishments that represent vulnerabilities during the data collection process. This information could be used to direct additional resources to collecting data from identified establishments in order to improve the response rate.	Nonresponse propensity modeling	research	regression trees with linear models	R (rpms package)
U.S. Bureau of Labor Statistics	Outlier Detection Using Unsupervised Learning Under Informative Sampling	A Bayesian hierarchical modeling approach was developed and applied to Current Employment Statistics survey. This approach is an enhanced k-means method, and it was used to find potential outliers.	Outlier Detection	research	K-means	R
U.S. Bureau of Labor Statistics	Text Analysis of Interviewer Notes	Using text analysis and clustering (unsupervised learning) to extract information and themes from survey interviewer notes, based on data from the Consumer Expenditure Survey. We also connected the themes from interviewer notes with sample unit behavior as captured in the Contact History Instrument.	Clustering of interviewer notes	research	Model-based clustering, k-means clustering, Bayesian hierarchical clustering	MATLAB and R
U.S. Census Bureau	Using an Autocoder to Code Industry and Occupation in the American Community Survey	Every year the American Community Survey (ACS) collects industry and occupation data on nearly 2.5 million individuals. The text write-in information must then be coded, or converted to an industry or occupation numeric category code.	multi-class classification	test	Logistische Regression	SAS

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
U.S. Department of Agriculture NASS	Informing Sample Design for the Census of Agriculture	Random forests with boosting were applied to predict the probability of the non-respondents after mailing to respond to the Census of Agriculture. The predicted probabilities were used as one of the stratifying variables in the sample design of non-respondents.	Developing response propensity scores to inform sampling design	The response propensity scores have been used in developing a sampling plan for non-respondents. Data collection is in progress.	Random forests with boosting	SAS JMP
U.S. Department of Agriculture NASS	Informing imputation for the Census of Agriculture	Random forests with boosting are used to inform imputation of the demographics section of the Census of Agriculture	Developing response propensity scores and informing imputation	A first imputation model has been implemented, and efforts are underway to improve it.	Random forests with boosting	SAS and SAS JMP
U.S. Department of Agriculture NASS	Crop Prediction Based on the Cropland Data Layer, Administrative Data, and Survey Data	An effort to bring all available data together to enhance crop forecasts of crop yield	Blending diverse data to produce the best possible prediction of crop yield	Foundational work is being undertaken	TBD, probably Bayesian	TBD, likely R or SAS with others Source (link)

Institution	Projektbezeichnung	Beschreibung	Anwendung	Status	Methode	Software
U.S. Department of Agriculture NASS	Machine learning for the Census of Agriculture	Machine learning methods are applied to computation needs in the production of Census of Agriculture, specifically sampling and imputation	Developing response propensity scores and informing imputation	The response propensity scores have been used in developing a sampling plan for non-respondnets. Currently, machine learning techniques are being used to inform imputation.	Random forests with boosting	SAS JMP
U.S. Department of Agriculture NASS	Questionnaire consolidation	Each state had its own questionnaire version (different states were surveyed on different items at different frequencies), as it was believed that this approach reduced respondent burden.	Questionnaire consolidation	experiment	hierarchical clustering	(SAS)
U.S. Department of Agriculture NASS	Non-respondent prediction	Non-response adjustment for the Census of Agriculture. Farms within each state in the US were partitioned into groups of “homogeneous response propensity” using a classification tree model. Non-response adjustments were performed within each such group based on the response rate within that group.	classification	experiment	Decision Tree	SAS
U.S. Department of Agriculture NASS	Analysis of reporting errors	Prediction of respondents likely to make reporting errors based on sampling frame data. Results of this analysis could suggest reasons for the reporting errors, types of respondents to be included in questionnaire testing, and editing strategies after data collection.	classification	experiment	Decision Tree	SAS

10.2.4 Rückmeldungen Hausabfrage

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
Digitale Bewerbungs-prozesse	A2	Heinz-Christoph Herbertz	A2 plant im Rahmen der JAP-Maßnahme "Digitale Bewerbungsprozesse" im Hinblick auf die angestrebte Automatisierung des Bewerbungsverfahrens auch die Einführung von Bots zur Kommunikation mit (potentiellen) Bewerber/innen.	Idee	nicht relevant	nicht relevant
Auskunftserteilung/ Anfragenbearbeitung	B 303, B i-Punkt	Ilka Willand, Daniel O'Donnell	Chatbots bzw. Live Chat Funktion im Rahmen des Zensus 2021 und vielleicht darüber hinaus	Idee, in Vorbereitungs- bzw. Konzeptionsphase	Noch unklar	Noch unklar
Machine Learning Methodik	C104	Lydia Spies	Bewertung von Machine Learning Verfahren als Möglichkeit zur Umsetzung einer automatisierten Datenaufbereitung und Datenanalyse	Tests	kNN, Naive Bayes, Random Forest, SVM, ANN, ...	R, Python, HoloClean

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
BIP-Abstimmung	D103	Tanja Mucha	Bei der BIP-Abstimmung könnte das Ergebnis der "Maschine" als zusätzliche Säule eingeführt werden. Trainings- und Testdaten stünden hierfür quartalsweise ab 1991 zur Verfügung. Berücksichtigt man die Ergebnisse der einzelnen Aggregate der Verwendungsrechnung und die einzelnen Wirtschaftsbereiche der Entstehungsrechnung, sowie alle angefallenen Revisionen, ergäbe sich ein umfangreicher Datenbestand.	Idee	nicht relevant	nicht relevant
BIP-Schnellschätzung	D103	Tanja Mucha	Als zweiter Einsatzbereich könnte Machine Learning bei BIP t+10 eingesetzt werden. Trainings- und Testdaten müssten hierfür aber zunächst erzeugt werden, sodass der Aufwand deutlich größer wäre. Wir können nicht einschätzen, ob der Nutzen des Einsatzes die Kosten rechtfertigen würde.	Idee	nicht relevant	nicht relevant

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
Klassifizierung von Fallgruppen bei der Leistungsberechnung der Krankenhäuser	D109	Susanne Goldhammer	In einer referatsinternen Diskussion haben wir festgestellt, dass zur Auswertung bestimmter Informationen Machine-Learning-Verfahren hilfreich sein könnten (Hier beispielsweise für die zeitintensive Bearbeitung von Datenmaterial der AOK bezüglich diagnosebezogener Fallgruppen für die Leistungsabrechnung in Krankenhäusern. Wir nutzen diese Auswertung um am Ende einen Deflator für Krankenhäuser zu errechnen.). Wenn wir uns nun aber am Prüfschema der Gruppe E1 orientieren, sind wir nicht sicher, ob dieses Problem hinreichend komplex ist. Es geht in unserem Fallbeispiel darum, dass wir etwa 1300 Datensätze prüfen und zuordnen müssen. Dies wird derzeit mit Excel-Tools und manuell erledigt. Mit Machine-Learning-Verfahren würde man hier sicher Ressourcen sparen, wir können aber nicht einschätzen, ob der Nutzen des Einsatzes die Kosten rechtfertigen würde. Wir haben zudem derzeit nicht das nötige Know-How.	Idee (externe Expertise erforderlich)	nicht relevant	nicht relevant

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
Nutzung von Webscraping in der Verbraucherpreisstatistik	D304	Daniel Seeger, Christian Blaudow	Produktnamen und Produkt-beschreibungen werden durch Webscraping automatisiert erhoben. Mit der Hilfe von Machine Learning-Techniken (vor allem "Text Mining" und "Natural Language Processing") können Produkte anhand ihrer Namen und Beschreibungen einer COICOP Klassifikation zugeordnet werden.	Idee	nicht relevant	nicht relevant
Nutzung von Scannerdaten in der Verbraucherpreisstatistik	D306	Timm Behrmann	Zuordnung von Artikeln über deren Produktbeschreibungen zur ECOICOP-Klassifikation. Eurostat hat ein Tool programmieren lassen, das an die deutschen Gegebenheiten angepasst werden soll.	Test	Im Tool stehen drei Verfahren zur Verfügung, die in der Weiterführung getestet werden sollen: SVM, Random Forest, Logistische Regression.	Das Tool ist in Java und Ember.js als Webservice programmiert.
Im Unternehmensprofiling Internetdokumente automatisiert auswerten.	E101	Simon Rommelspacher	Das manuelle (Unternehmens-) Profiling ist eine sehr komplexe Aufgabe, die aus vielen Recherche- und Analyseschritten besteht. Durch Webscraping, Text Mining und anschließenden Machine Learning-Verfahren könnten einige bisher manuell durchgeführten Schritte unterstützt oder automatisiert werden.	Idee	ist noch zu prüfen	ist noch zu prüfen

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
Automatic Profiling	E101	Matthias Redecker	Automatic Profiling bedeutet die Zerlegung von kleinen und mittelkomplexen Unternehmensgruppen in sogenannte statistische Untenehmen (Clusterung) und die Zuordnung der unternehmensgruppenzugehörigen rechtlichen Einheiten zu diesen Unternehmen (Klassifikation). Beide Prozesse werden gegenwärtig auf recht rudimentären Annahmen und auf Basis der vorliegenden Daten im URS durchgeführt. Der Blick auf die realen Tatbestände fehlt größenteils. Offen ist, ob die Erkenntnisse aus dem manuellen Profiling im Sinne eines maschinellen Lernens verwendet werden können, um die Abgrenzung und Zuordnung zu verbessern.	Idee	ist noch zu prüfen	ist noch zu prüfen

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
Klassifikation der Tätigkeiten in den WZ 64.20/70.10 bzw. Klassifizierung der verschiedenen Holding-gesellschaften nach ihren Funktionen	E101/E102	Katja von Eschwege, Matthias Redecker	In den genannten WZ-Klassen gibt es häufig Fehlklassifikationen, u.a. aufgrund eingeschränkter Rück-meldungen aus den Erhebungen (WZ 64.20), unterschiedlichem Verständnis der "Holding"-Begriff-lichkeiten in den verschiedenen Verwaltungsdaten, durch die das URS gepflegt wird. Ziel ist eine korrekte und trennscharfe Abgrenzung der Tätigkeiten bzw. der Holding-Einheiten mit ihren jeweiligen Funktionen in der deutschen Unternehmensland-schaft. Dies ist für URS (Profiling), Erhebungen (Abgrenzung der Stichproben) und VGR (Sektorklassifizierung) wichtig.	Idee	Unsupervised Learning, d.h. Clusterung der WZ bzw. der Holding-Einheiten. Derzeitig sind keine guten Trainingsdaten vorhanden, um eine direkte Klassifizierung zu unterstützen.	ist noch zu prüfen
Zuordnungsvorschläge der AVS im URS verbessern; Rechtsformermittlung im Rahmen der AVS verbessern.	E101/E102	Philipp Hadeball, Katja von Eschwege	Die AVS (Adressverarbeitungs-software) im URS-Neu normiert/-standardisiert Adressen hinsichtlich verschiedene Merkmale (Straße, Hausnummer, Postleitzahl, Rechtsform) und wird u.a. auch zur Generierung von Abgleichen/Zuordnungen zwischen Einheiten verschiedener Quellen (Kern- und Adminregister) verwendet.Da die AVS nicht immer optimale Ergebnisse liefert, könnte geprüft werden ob mittels geeigneter Verfahren die Ergebnisse der AVS nachträglich "verbessert" werden könnten.	Idee	ist noch zu prüfen	Für die Bestimmung/ Validierung des Verfahrens können exportierte Daten verwendet werden. Im Anwendungsfall muss die Methode jedoch in das URS-Neu migriert werden.

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
Überprüfung der WZ-Signierung im URS	E102	Katja von Eschwege	Im Unternehmensregister (URS) werden WZ-Angaben aus Erhebungen als verlässlich eingestuft. Die WZ-Angaben von BA und Finanzverwaltung, die ins URS einfließen, können fehlerhaft sein und müssten überprüft werden. Angesichts der Fallzahl fehlen hierfür Personalkapazitäten. Idee: Trainieren eines Modells anhand der Daten aus Erhebungen und Anwendung auf die restlichen Daten. Missklassifikationen ergeben Hinweise auf mögliche Fehlklassifikationen in den Ausgangsdaten der BA und der Finanzverwaltung und somit manuell zu prüfende Fälle.	Idee	SVM, Random Forest	R

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
International Sourcing Survey: relevante Einheiten identifizieren.	E103	Wolfhard Kaus	Bei der aktuellen Probeerhebung für die neue FRIBS-Erhebung zu "International Sourcing" zeigt sich, dass der Großteil der Fragen nur von einem kleinen Teil der Unternehmen in der Grundgesamtheit (~2%) sinnvoll beantwort werden können, weil nur wenige Unternehmen Geschäftsbereiche/wirtschaftliche Aktivitäten in das Ausland verlagern. Das Vorliegen von Verlagerungsaktivitäten kann weder aus Verwaltungsdaten noch aus dem statistischen Unternehmensregister abgeleitet werden. Ziel ist es betroffene Einheiten vor der Befragung zu identifizieren, um diese bei der Stichprobenziehung gezielt überproportional zu berücksichtigen.	Idee	ist noch zu prüfen	ist noch zu prüfen

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
Ausreißerdentifikation: Isolation Forest	E103	Philipp Leppert	Innerhalb des Projekts "Umsetzung des EU-Unternehmensbegriffs in den Unternehmensstruktur-statistiken" wird zur Datengewinnung ein spender-basiertes Imputationsverfahren verwendet. Um unzureichend plausibilisierte oder "extreme" Datenpunkte aus dem Spenderpool zu entfernen, wurden neben parametrischen Ausreißeridentifikationsmethoden auch ein ML-basiertes Verfahren getestet. Isolation Forest bietet geringen Aufwand bei der Implementierung und hohe Effizienz bzgl. der Rechenleistung auch im Umgang mit großen (strukturierten) Datenbeständen.	Test	ist noch zu prüfen	R / Paket: isofor; Python / Paket: scikit-learn
Gewerbeanzeigenstatistik: Freitext der Gewerbeanzeige nach WZ-2-Steller klassifizieren.	E105	Jenny Neuhäuser	Bei einer Gewerbemeldung ist von den Gewerbetreibenden mit einem Freitext eine Beschreibung der wirtschaftlichen Tätigkeit anzugeben. Diese Angaben werden zur Gewerbeanzeigenstatistik geliefert und bei den Statistischen Ämtern der Länder wird hieraus manuell der WZ-2-Steller für das Unternehmen bzw. den Betrieb ermittelt. Dieser Arbeitsschritt könnte mit maschinellen Lernverfahren unterstützt und so weit wie möglich automatisiert werden.	Idee	ist noch zu prüfen	ist noch zu prüfen

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
Proof of Concept automatisierte Plausibilisierung (in den Verdienststatistiken)	E107	Paul Mätzig	Der Einsatz der Software HoloClean zur automatisierten Plausibilisierung der Daten der Verdienststatistiken soll geprüft werden. HoloClean verfügt über ein Fehler- und Ausreißererkennungsmodul und lernt auf Grundlage mehrerer Datenquellen ein Wahrscheinlichkeitsmodell, das die Imputation (Reparatur) der fehlerhaften Daten übernimmt.	Experiment	HoloClean (zugrundeliegendes Wahrscheinlichkeitsmodell als factor graph, Inferenz via Gibbs-Sampling)	Python (PyTorch), Apache Spark, PostgreSQL
Merkmal "Urlaubstage“ (approximiert als Urlaubsanspruch) schätzen.	E109	Simone Scharfe	Im Rahmen des "Verzahnungsprojektes" ist eine Überlegung, das Merkmal Urlaubstage in der VSE nicht mehr zu erheben, sondern für Zwecke der Datenlieferung an Eurostat ab dem Berichtsjahr 2022 (Lieferung an Eurostat Mitte 2014) zu schätzen. Als Quellen stehen einerseits historische VSE-Ergebnisse zur Verfügung und andererseits Angaben zum aktuellen Urlaubsanspruch aus der Tarifdatenbank (Herausforderung: nicht alle Beschäftigten sind tarifgebunden, das Merkmal der individuellen Tarifgebundenheit liegt u.U. nach der Verzahnung nicht mehr direkt im Datensatz komplett vor).	Idee	ist noch zu prüfen	ist noch zu prüfen

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
Überprüfung der Signierung der Teilzeitbeschäftigung im SV-Schlüssel der BA in der Verdienststruktur-erhebung (VSE)	E109	Paul Mätzig	In der VSE wird auch der SV-Schlüssel der BA erhoben, in dem auch die Teilzeitbeschäftigung codiert ist. Fehlerhafte Angaben werden im Rahmen der Plausibilitätskontrollen anhand von Informationen zur Arbeitszeit korrigiert. Ziel: Mit den korrigierten Werten wird ein Modell zur Prognose der Teilzeittätigkeit gelernt und dann auf die fehlerbehafteten BA-Daten angewandt. Die Missklassifika-tionen ergeben dann Hinweise auf Fälle, in denen die Teilzeit-beschäftigung in den BA-Daten falsch codiert ist.	Test	Random Forest	R (Ranger)
Musterprozess bei zentralen Unternehmensstatstiken	E2	Daniel Vorgrimler, René Söllner, Jens Dechent	Ziel der Leitungsklausur-Maßnahme ist die Digitalisierung und Standardisierung der statistischen Produktionsprozesse in den zentralen Unternehmensstatistiken. Hierzu gehört auch der Einsatz von Machine Learning Verfahren zur Einzeldatenplausibilisierung und zur Ergebnisvalidierung.	erste Idee	zu klären	zu klären
Finden von Mehrfachfällen	F1	Stefan Dittrich	In den Melderegistern existieren Fehrfachfälle. Viele dieser Fälle können mit einfachen Gleichheitsoperationen ermittelt werden, aber es gibt auch Fälle, die schwerer zu erkennen sind.	Idee	nicht relevant	nicht relevant

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
Kodieren von Berufen etc.	F1	Stefan Dittrich	Beim Zensus wird auch Freitext erhoben, der i.d.R. kodiert werden muss.	Idee	nicht relevant	nicht relevant
Codierung des Geburtsstaats	F202	Claire Grobecker, Rabea Mundil-Schwarz, Bettina Sommer	Bei fehlender Angabe zum Geburtsstaat soll der Geburtsstaat anhand weiterer Angaben im Datensatz codiert werden.	Idee	zu klären	zu klären
Klassifizieren von Online-Stellenanzeigen	F205	Chris-Gabriel Islam	Gescrapte Online-Stellenanzeigen liegen meist in unstrukturiertem Volltext vor. Oft fehlen gesonderte Angaben, wie der Wirtschaftszweig, das gewünschte Bildungsniveau oder ob die Anzeige von einem Personalvermittler geschaltet wurde oder nicht. Ein ML-Ansatz soll derartige Klassifizierungen automatisiert vornehmen.	Test	KNN, MN-Naive-Bayes	Python (NLTK, SKlearn), R
Berichtskreiserstellung in der Erhebung der Waren, Bau- und Dienstleistungen für den Umweltschutz (GMAS Phase 2.4)	G203	Gesine Petzold	Die Zielgruppe dieser Erhebung sind Betriebe, die Waren, Bau- oder Dienstleistungen für den Umweltschutz erstellen bzw. anbieten. Im Betrieb kann eine Spezialisierung vorliegen oder nur ein Teil des Produktportfolios für den Umweltschutz dienlich sein. Ausgehend von der technischen Art und dem Hauptverwendungszweck der Ware oder Dienstleistung wird bestimmt, ob es sich um eine Umweltschutzmaßnahme handelt oder nicht. Wenn ja, so ist der Betrieb Teil der Zielgruppe der Erhebung. Sowohl die amtliche statistische Klassifi-	Idee	nicht relevant	nicht relevant

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
			kation der Wirtschaftszweige (WZ, Ausgabe 2008) als auch die Informationen aus dem Unternehmensregister (URS) geben kaum bis keine Hinweise auf potentielle Umweltbetriebe, die Güter oder Dienstleistungen für den Umweltschutz erstellen. Auch gibt es keine externen Datenbanken, mit deren Hilfe die Berichtseinheiten eindeutig identifiziert werden können. a) Die Frage ist, ob sich mithilfe der Informationen der zurückliegenden Berichtsjahre und der dazugehörigen Angaben aus dem URS Regeln ableiten lassen, mit deren Hilfe potentielle Umweltbetriebe identifiziert werden. b) Eine weitere Fragestellung ist, ob man für externe Datenbanken wie UMFIS Regeln definieren kann, die potentielle Umweltbetriebe in dieser Datenbank identifizieren. Hintergrund: In dieser Datenbank gibt es Selbstselektion. Nicht immer stimmt die Einschätzung der Betriebe mit der Definition Umweltschutz der amtlichen Statistik überein.			
Datenvalidierung (GMAS TP 5.3/5.4 Item non-response)	G203	Gesine Petzold	In allen drei Erhebungen des Referats G203 werden monetäre Merkmale erfragt unter dem Aspekt Umweltschutz. Die monetären Werte wie Investitionen, Umsätze,	Idee	nicht relevant	nicht relevant

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
			laufende Aufwendungen unterliegen Buchführungsregeln und lassen sich von Unternehmen und Betrieben aus den Büchern ermitteln. Im Rahmen der umweltökonomischen Erhebungen interessiert jedoch nur der Teil dieser Werte, der die Definition "für den Umweltschutz" erfüllt. Dazu bedarf es oft einer Abstimmung zwischen der Buchhaltung, dem Umweltmanagement, Ingenieuren oder dergleichen. Die Zielgruppe der Erhebung der laufenden Aufwendungen für den Umweltschutz sind Unternehmen, die durch den Betrieb von Umweltschutzanlagen oder die Inanspruchnahme von Umweltschutzleistungen wiederkehrend Aufwendungen haben. Generell haben Unternehmen immer Kosten durch die Entsorgung des Abwassers und der Abfälle. Entscheidender sind jedoch die Aufwendungen durch den Betrieb von Umweltschutzanlagen. Aufgrund der Stichprobe und der Periodizität (dreijährig) der Statistik ist ein Abgleich der Meldedaten mit der Vorerhebung nur bedingt möglich. Item non-response wird auf die Art geprüft, dass die Angaben der Unternehmen auf ihrer Website			

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
			recherchiert werden. Da das Image der Unternehmen zunehmend auch durch ihre Aktivitäten im Bereich Umweltschutz geprägt wird, werben Unternehmen auf ihrer Website verstärkt damit. a) Wünschenswert wäre der Abgleich der Meldedaten mit den Informationen zum Umweltschutz auf der dazugehörigen Unternehmensseite. Neben der technischen Schwierigkeit gibt es eine inhaltliche: Die Definition Umweltschutz im Sinne der amtlichen Statistik – und entsprechend die zu definierenden Kriterien für einen solchen technischen Abgleich – muss nicht mit der Definition Umweltschutz des Unternehmens übereinstimmen.			
Außenhandel nach Unternehmensmerkmalen (Trade by enterprise characteristics (TEC))	G 301	Ilda Duarte Fernandes	Die TEC Daten werden erstellt, indem die Außenhandelsdaten mit dem Unternehmensregister verknüpft werden. Dabei können nicht alle Außenhandelsdaten auf Unternehmensebene verknüpft werden: Bei der Anzahl der Unternehmen liegt die Matching-Rate zwischen 55%-93%, bei Wert zwischen 85% und 91%. Durch ML sollen fehlenden Strukturinformationen aus dem URS (WZ, Anzahl der Beschäftigten) bei den nicht verknüpfte Daten imputiert	in Planung	zu klären	zu klären

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
			werden.			
Verbesserung der PL	H201	Stefan Brings	Grundsätzlich scheinen Machine Learning Verfahren geeignet, langfristig die von den StLÄ durchgeführte PL zu den aus Verwaltungsunterlagen erstellten Hochschulstatistiken zu automatisieren und effizienter zu gestalten. Der Einsatz von ML-Verfahren ist derzeit nicht prioritär. Es sollte zunächst der Proof of Concept „automatisierte Plausibilitätskon-trolle (in den Verdienststatistiken)“ abgewartet werden.	Idee	nicht relevant	nicht relevant

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
EVS 2023	H301	Birgit Lenuweit	Signierung von Klassifikationsitems auf Basis von Klartextangaben: Für die EVS 2023 wird ein elektronisches Erhebungsinstrument entwickelt, das als Desktopanwendung sowie auf mobilen Endgeräten laufen soll. Zudem wird das Verwaltungs- und Erfassungsprogramm neu programmiert. In beiden Anwendungen sollen die Befragten/Erfasser Ausgaben nach der SEA-Klassifikation codieren. Hierfür wird eine Suchfunktion entwickelt bzw. die Suchfunktion des Klassifikationsservers in die zu programmierenden Anwendungen eingebunden (Offline-Komponente). Die Anwendungen sollen anhand der durch die Befragten/-Erfasser angegebenen Klartexte "lernen" und die Suchfunktionalität hierdurch verbessert werden. Die Zuordnung nicht codierter Klartexte soll zudem ggf. durch eine maschinelle Codierung unterstützt werden. Es soll geprüft werden, ob Methoden des Machine Learning eingesetzt werden können.	Idee	zu klären	zu klären

Statistik- oder Projekt-bezeichnung	Organisations-einheit/ Projektteam	Ansprechperson	Inhaltliche Kurzbeschreibung des Projektes	Status	Geplante Methodik (z. B. SVM, Random Forest, ...)	Einzusetzende Softwarelösung (z. B. R, Python, SAS, ggf. auch Nennung einzelner Packages)
ZVE 2021/22	H303	Holger Breiholz	Signierung von Klassifikationsitems auf Basis von Klartextangaben Weitere Erläuterung und Begründung wie bei „EVS 2023“.	Idee	zu klären	zu klären
Mikrozensus	H305	Birte Tiedemann	Signierung von Klassifikations-items auf Basis von Klartext-eingaben: Mit den Erhebungs-instrumenten (Laptopanwendung, IDEV, Erfassung der Selbstaus-füllerbögen) des Mikrozensus ab 2020 erfolgt die Zuordnung zu einem Klassifikationsitem (z.B. Beruf, Wirtschaftszweig) nach Klartexteingabe, Ausgabe von Suchtreffern des Klassifikations-servers und Auswahl eines Items. Problem: Nicht immer kann eine zutreffende Sigierung erfolgen. Mögliche Gründe: Zu ungenaue Angaben des Befragten, Ausgabe von keinen oder unpassenden Suchtreffer, fehlerhafte Vercodung aus Mangel an Alternativen. Es soll geprüft werden, ob hier mit Methoden des Machine Learning eine Verbesserung der Zuordnung erzielt werden kann.	Idee	zu klären	zu klären

10.3 Prüfschema

wissen.nutzen.

"Prüfschema" für mögliche Anwendungen in den Fachstatistiken

- Welche Voraussetzungen müssen für den Einsatz von Machine-Learning-Verfahren erfüllt sein?
  - Für Machine-Learning-Verfahren geeignetes (und hinreichend komplexes) Problem
  - Ausreichendes Know-how (Methodik, Software, Fachstatistik, ...)
  - Auswahl eines oder mehrerer angemessener Machine-Learning-Verfahren
  - Geeignete, verfügbare und einsetzbare Software
  - Ausreichend dimensionierte Hardware
  - Trainings- und Testdaten

wissen.nutzen.

Trainings- und Testdaten: Fall 1

Fall 1: Trainings-, Test- und zu klassifizierende Daten liegen zum Zeitpunkt t aus der Statistik X vor
Anwendung: bisher keine (Idee: WZ-Klassifikation im URS)

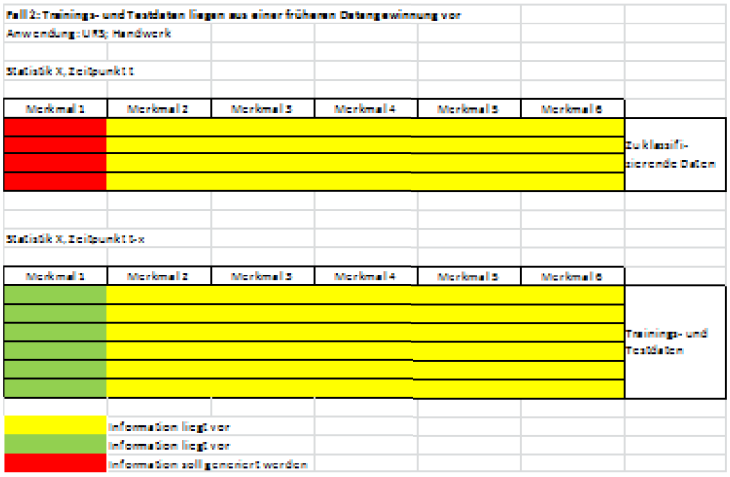
Statistik X, Zeitpunkt t

Merkmal 1	Merkmal 2	Merkmal 3	Merkmal 4	Merkmal 5	Merkmal 6	
						Zu klassifi-zierende Daten
						Trainings- und Testdaten

Information liegt vor
Information liegt vor
Information soll generiert werden

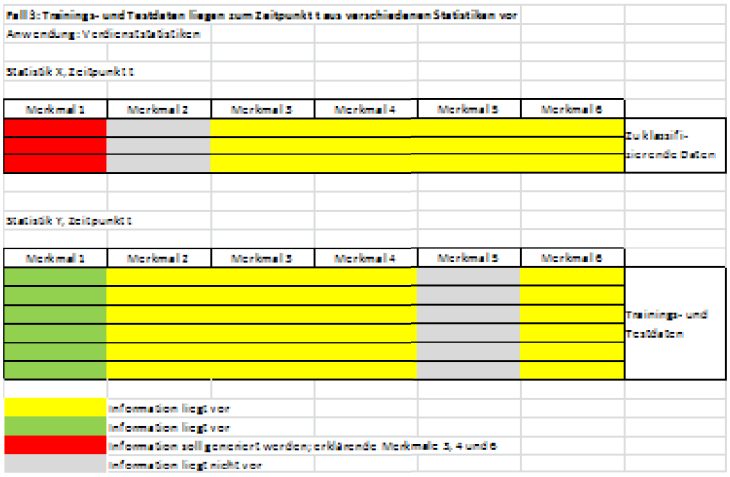
wissen.nutzen.

Trainings- und Testdaten: Fall 2



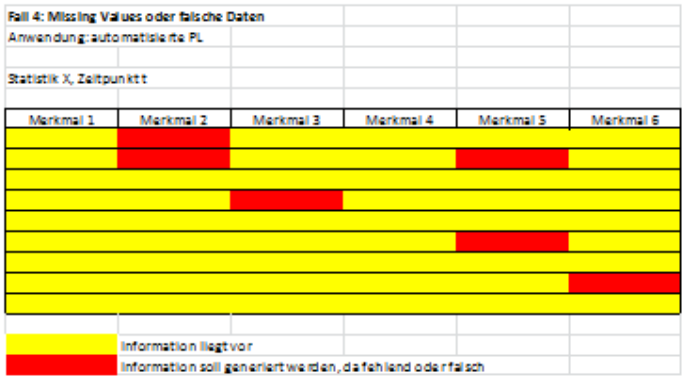
wissen.nutzen.

Trainings- und Testdaten: Fall 3



wissen.nutzen.

Trainings- und Testdaten: Fall 4



10.4 Glossar

Baum (auch Klassifikations- oder Regressionsbaum; im ersten Fall auch Decision Tree, im zweiten auch Regression Tree): Auf dem Prinzip rekursiver Partition des Eingaberaums beruhendes Verfahren aus dem Machine Learning. Der Eingaberaum wird dabei anhand geeigneter Splitvariablen (aus der Menge der Eingabevariablen) immer weiter (jeweils achsenparallel) aufgeteilt, bis die verbleibenden Regionen in sich hinreichend homogen sind.

Klassifikation: Teilgebiet des supervised learnings, bei dem die Ausgabevariable höchstens endlich viele Werte (z. B. die Klasse, zu der ein Datenpunkt gehört) annehmen kann. Ziel einer Klassifikationsmethode ist die korrekte Schätzung der Klassenzugehörigkeit eines Datenpunktes.

Neuronales Netz: Methode, die in verschiedenen Ausprägungen in allen Bereichen des maschinellen Lernens eingesetzt wird. Ein (künstliches) Neuronales Netz besteht aus Neuronen und gewichteten Verknüpfungen. Die zu lernenden Gewichte der Verknüpfungen entscheiden über den Output eines Neuronalen Netzes, z. B. über die Schätzung der Klassenzugehörigkeit eines Datenpunktes.

Random Forest: Klassifikations- und Regressionsmethode aus dem Bereich des maschinellen Lernens basierend auf der Idee, einen Mehrheitsentscheid (bei Klassifikation) bzw. eine Durchschnittsbildung (bei Regression) heranzuziehen, um für einen neuen Datenpunkt die Ausprägung der Outputvariable zu schätzen. Mehrheit bzw. Durchschnitt werden bzgl. vieler Bäume (trees) gebildet.

Reinforcement Learning (bestärkendes Lernen): Form des maschinellen Lernens, das die Schätzfunktion immer wieder an neu hinzukommende Daten anpasst, indem der Verlauf beobachtet und schlechte Vorhersagen einen negativen Beitrag, gute Vorhersagen einen positiven Beitrag zu einem über die Gesamtzeit zu maximierenden Gewinn liefern. Letztlich wird also die Vorhersagestrategie optimiert.

Supervised Learning (überwachtes Lernen): Das Machine-Learning-Problem besteht aus Eingabe- und Ausgabevariablen, deren Zusammenhang die Methode (zumindest approximativ) nachbilden soll. Zum Lernen stehen dabei Trainingsdaten zur Verfügung, für die sowohl die Ausprägungen der Eingabevariablen als auch der Ausgabevariable bekannt sind. Beispiele: Klassifikation und Regression.

Support Vector Machine: Methode des maschinellen Lernens für Fragestellungen im supervised und im unsupervised learning. Ursprünglich zum Zwecke der binären Klassifikation entworfen; dort auf dem Grundprinzip aufbauend, die trennende Hyperebene so zu wählen, dass die zu den Klassen gehörigen Punktwolken abstandsmaximal getrennt werden.

Unsupervised Learning (unüberwachtes Lernen): Das Machine Learning Problem besteht darin, Charakteristika der Daten zu erkennen. Ausgabevariablen gibt es in dieser Form des maschinellen Lernens nicht. Beispiele: Clusteranalyse, Feature Selection, Ausreißererkennung.

10.5 Aufsatz Dumpert/Beck (2017)

Dumpert, F., Beck, M. (2017). Einsatz von Machine-Learning-Verfahren in amtlichen Unternehmensstatistiken. AStA Wirtschafts- und Sozialstatistisches Archiv, 11, 83–106.

AStA Wirtsch Sozialstat Arch
https://doi.org/10.1007/s11943-017-0208-6

ORIGINALVERÖFFENTLICHUNG

Einsatz von Machine-Learning-Verfahren in amtlichen Unternehmensstatistiken

Florian Dumpert · Martin Beck

Eingegangen: 28. Juli 2017 / Angenommen: 28. September 2017

Zusammenfassung Aufgabe der amtlichen Unternehmensstatistiken ist die Bereitstellung von Informationen über Struktur und Entwicklung der Wirtschaft, die sie durch Erhebungen, die Nutzung von Verwaltungsdaten, den Zukauf kommerzieller Daten und die Verknüpfung von Mikrodaten gewinnt. In jüngster Zeit wurde darüber hinaus auch der Einsatz von Machine-Learning-Verfahren in amtlichen Unternehmensstatistiken experimentell erprobt, und zwar bei Zuordnungsentscheidungen und der Generierung neuer Informationen. In diesem Beitrag wird das Vorgehen im Überblick dargestellt. Dazu werden zunächst die Methodik des maschinellen Lernens in den Grundzügen dargestellt, bisherige Anwendungsgebiete außerhalb und in der amtlichen Statistik beschrieben sowie die in der Unternehmensstatistik experimentell eingesetzten Verfahren erläutert. Anschließend wird die praktische Anwendung von Support Vector Machines und Random Forests auf fünf konkrete Aufgabenstellungen in ausgewählten Unternehmensstatistiken dargestellt. Abschließend werden die bisherigen Erfahrungen zusammenfassend bewertet und potenzielle weitere Aufgabenstellungen sowie absehbare Weiterentwicklungen der maschinellen Lernverfahren aufgezeigt.

Schlüsselwörter Maschinelles Lernen · Random Forest · Support Vector Machine · Unternehmensstatistik

F. Dumpert
Fakultät für Mathematik, Physik und Informatik, Lehrstuhl für Stochastik, Universität Bayreuth, 95440 Bayreuth, Deutschland
E-Mail: florian.dumpert@uni-bayreuth.de

M. Beck (✉)
Gustav-Stresemann-Ring 11, Statistisches Bundesamt, 65189 Wiesbaden, Deutschland
E-Mail: martin.beck@destatis.de

Published online: 12 October 2017

Use of machine learning in official business statistics

Abstract The task of the official business statistics is to provide information on the structure and development of the economy, which is gained through surveys, the use of administrative data, the purchase of commercial data and the linking of micro data. Recently, the use of machine learning methods in official business statistics has also been experimentally tested in the case of classification decisions and the generation of new data. This article provides an overview of the proceeding. To this end, the methodology of machine learning is first presented in the basic principles, previous fields of application are described outside and in official statistics, and the methods used experimentally in the business statistics are explained. Subsequently, the practical application of Support Vector Machines and Random Forests is presented in five concrete tasks in selected business statistics. Finally, the experience gained so far is summarized and potential further tasks as well as foreseeable further developments of the machine learning methods are presented.

Keywords Machine learning · Random Forest · Support Vector Machine · Business statistics

1 Einführung

Während die empirische Wissenschaft frei ist in der Wahl ihrer Methoden, sind die Aufgaben und die Verfahren der amtlichen Statistik in weiten Teilen gesetzlich vorgegeben. Die wesentlichen grundsätzlichen Bestimmungen enthält das Bundesstatistikgesetz, das in § 1 die Zwecke der Bundesstatistik regelt, in § 9 den Regelungsumfang vorgibt und in § 5 Abs. 1 die Erfordernis der gesetzlichen Anordnung von Bundesstatistiken festschreibt. Einzelne Datenerhebungen werden somit i. d. R. in Spezialgesetzen angeordnet, die die konkrete Durchführung im Detail regeln. Das betrifft unter anderem die Festlegung der Erhebungsmerkmale und die Periodizität, die somit in der Praxis auch bei neu auftretenden Informationsbedarfen nicht kurzfristig angepasst werden können.

Aufgabe der amtlichen Unternehmensstatistiken ist die Bereitstellung von Informationen über Struktur und Entwicklung der Wirtschaft. Dabei stellt sich die grundsätzliche Frage, wie und bei wem die Informationen gewonnen werden sollen. Die Antwort unterliegt im Zeitablauf Veränderungen, die mit der Verfügbarkeit und den Zugangsmöglichkeiten zu Daten sowie methodischen Weiterentwicklungen zusammenhängen. Hauptsächlich fußt die Informationsgewinnung auf der gesetzlich angeordneten Befragung von Unternehmen bzw. genauer „rechtlichen Einheiten“ durch die Statistischen Ämter. Ergänzend bzw. ersetzend sind in den letzten Jahren die Verwendung von Verwaltungsdaten (Lorenz und Opfermann 2017), der Ankauf kommerzieller Daten (Kleber et al. 2010) sowie die Verknüpfung vorhandener Datenquellen (Jung und Käuser 2016; Kaus und Leppert 2017) als alternative Möglichkeiten der Informationsgewinnung hinzugekommen. Ausschlaggebend hierfür waren die Notwendigkeit und das Ziel, Unternehmen von Bürokratiekosten zu entlasten. Bevor Befragungen von Unternehmen vorgesehen oder ausgeweitet werden,

ist zunächst zu prüfen, ob und inwieweit benötigte Informationen auch auf anderem Wege ressourcenschonend gewonnen werden können. Hierzu wurde in jüngster Zeit auch der Einsatz von Machine-Learning-Verfahren in amtlichen Unternehmensstatistiken experimentell erprobt. Dabei ging es vornehmlich um folgende grundsätzliche Aufgabenstellungen:

- Zuordnungsentscheidungen.
  Häufig ist zu entscheiden, ob ein Unternehmen zu einem relevanten Ausschnitt der Volkswirtschaft gehört, über den eine Statistik Informationen bereitstellen soll. Diese Frage stellt sich u. a. bei der Zuordnung von Unternehmen zu den Sektoren der Volkswirtschaftlichen Gesamtrechnung (VGR) bzw. zum sogenannten „dritten Sektor“ aber auch bei der Abgrenzung des Handwerks in der Handwerkszählung. Die bisherige Lösung besteht in der händischen Überprüfung der Zuordnung von Zweifelsfällen und ist damit zeit- und personalaufwändig.
- Generierung von neuen Informationen.
  Erhebungen beruhen i. d. R. auf Gesetzen, die u. a. die Erhebungsmerkmale regeln. Dies schafft einerseits Rechtssicherheit für die Statistischen Ämter und die zur Auskunft verpflichteten Unternehmen. Andererseits können Erhebungen nicht schnell an neue Informationsbedürfnisse angepasst werden, da entsprechende Gesetzesänderungen notwendig sind. Aktuelle Beispiele hierfür sind neue Informationsanforderungen zum Verdienstunterschied von Frauen und Männern (Gender Pay Gap), zu den Wirkungen des zum 01.01.2015 eingeführten allgemeinen gesetzlichen Mindestlohns und zur Arbeitsmarktintegration von Migranten. Die traditionelle Lösung würde in solchen Fällen die Erhebung neuer bzw. zusätzlicher Daten vorsehen. Dies wäre für die auskunftspflichtigen Unternehmen mit erhöhten Bürokratiekosten und für die Statistischen Ämter mit zusätzlichem Personalaufwand verbunden.

Die amtliche Statistik ist gehalten, nach möglichen Alternativen zu suchen. Da sich die beschriebenen Aufgabenstellungen formal als Klassifikationsprobleme formulieren lassen, stellen Machine-Learning-Verfahren eine Option dar, die in jüngster Zeit in der Unternehmensstatistik aufgegriffen wurde. Das konkrete Vorgehen soll in diesem Beitrag im Überblick dargestellt werden. In Abschn. 2 werden zunächst die Methodik des maschinellen Lernens in den Grundzügen dargestellt, bisherige Anwendungsgebiete beschrieben sowie die in der Unternehmensstatistik experimentell eingesetzten Verfahren erläutert. In Abschn. 3 wird dann die praktische Anwendung von Support Vector Machines und Random Forests auf fünf konkrete Aufgabenstellungen in ausgewählten Unternehmensstatistiken dargestellt. In Abschn. 4 werden die bisherigen Erfahrungen resümiert und potentielle weitere Aufgabenstellungen sowie absehbare Weiterentwicklungen der maschinellen Lernverfahren aufgezeigt.

2 Was Machine-Learning-Verfahren sind, welche es gibt und welche wir nutzen

Thema dieses Abschnitts sind Verfahren, die in der Fachwelt dem sogenannten Machine Learning (oder deutsch: maschinellen Lernen) zugeordnet werden. Vornean

steht die Klärung der Begriffe, anschließend werden kurz Einsätze derartiger Verfahren innerhalb und außerhalb der amtlichen Statistik beleuchtet. Zum Abschluss dieses Abschnitts werden die eingesetzten Methoden genauer vorgestellt.

2.1 Grundlegende Einführung

Es gibt eine Vielzahl von Ansätzen, den Begriff des maschinellen Lernens zu definieren oder dessen Wesen zu erfassen. Samuel (1959) definierte maschinelles Lernen als Chance, Problemlösungsmethoden nicht mehr exakt implementieren zu müssen:

> The studies reported here have been concerned with the programming of a digital computer to behave in a way which, if done by human beeings or animals, would be described as involving the process of learning. [...] Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort.

Steinwart und Christmann (2008) beziehen sich im einführenden Kapitel ihrer Monographie auf Simon (1983), der schreibt:

> Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time.

Lernen kann auf verschiedene Arten beschrieben werden. In diesem Aufsatz soll das sogenannte *statistische maschinelle Lernen* betrachtet werden, das es ermöglicht, den Vorgang und den Erfolg des Lernens mathematisch zu fassen. Das Ziel besteht stets darin, zu gegebenen Eingabewerten (auch als Werte oder Realisierungen von Eingabemerkmalen, Eingabevariablen, erklärenden Variablen oder Inputvariablen bezeichnet) einen brauchbaren Ausgabewert (auch als Wert des zu erklärenden Merkmals, der zu erklärenden Variable, der Ausgabevariable oder der Outputvariable bezeichnet) zu schätzen. Es muss also ein Zusammenhang zwischen den Eingabevariablen und der Ausgabevariable erlernt werden, der anschließend auf neue, ggf. noch nicht bekannte Eingabewerte angewendet werden kann. Dieses Lernen muss statistisch geschehen, also auf Basis von nur endlich vielen Beobachtungen von Eingabe- und Ausgabewerten. Ausgerichtet muss es aber auf das Ziel sein, Aussagen über die Grundgesamtheit zu treffen, d. h. für alle denkbaren Kombinationen von Eingabe- und Ausgabewerten, also alle denkbaren Realisierungen der Eingabe- und Ausgabevariablen. Die Gesamtheit der zur Verfügung stehenden Beobachtungen heißt Trainingsdatensatz. Statistiker und Informatiker oder sogenannte Data Scientists verwenden allerdings zum Teil unterschiedliche Begriffe für die gleichen Objekte, die Wasserman (2004, S. xi) einleitend gegenüberstellt.

Dieser Aufsatz beschränkt sich auf das sogenannte überwachte Lernen (supervised learning), bei dem der Trainingsdatensatz sowohl Eingabe- als auch zugehörige

Ausgabewerte enthält.[1] Spezieller betrachtet wird die binäre Klassifikation, also die Schätzung der Zugehörigkeit zu einer von zwei Klassen.[2]

Wurde der Zusammenhang einmal gelernt, so kann die zu erfüllende Aufgabe im Sinne von Simon (im Folgenden die Klassifikation) effizienter (d. h. im vorliegenden Fall ressourcenschonender) und/oder effektiver gelöst werden. Statistisches maschinelles Lernen in diesem Sinne kann somit nur funktionieren, wenn hinreichend gutes Material vorhanden ist, anhand dessen gelernt werden kann. Statistisches maschinelles Lernen kann darüber hinaus scheitern, wenn sich die Aufgabe nach dem Lernen nicht mehr auf die gleiche Grundgesamtheit bezieht, ein früher gelernter Zusammenhang also möglicherweise gar nicht mehr zutrifft.

Die mathematische Beschreibung des statistischen maschinellen Lernens ist ein Verdienst von Vapnik (1995, S. 15 f.), der das allgemeine „Problem zu lernen“ in drei Komponenten zerlegt:

1. Zunächst gibt es einen Mechanismus, welcher die Eingabewerte gemäß einer dem Anwender unbekannten (aber sich nicht verändernden) Verteilung erzeugt.
2. Anschließend tritt ein Mechanismus in Kraft, der zu jedem Eingabewert einen Ausgabewert erzeugt. Dies geschieht aufgrund von dem Anwender unbekannten, aber wiederum sich nicht verändernden bedingten Verteilungen für die Ausgabe gegeben die Eingabe.
3. Aufgabe des statistischen maschinellen Lernens ist es nun, aus einer geeignet zu fassenden Fülle an möglichen funktionalen Zusammenhängen zwischen Eingabe- und Ausgabevariablen diejenige Funktion auszuwählen, die das in Schritt (2) betrachtete Verhalten am besten approximiert. Gesucht wird also derjenige Prädiktor, der die bedingten Verteilungen der Ausgabewerte gegeben die Eingabewerte am besten reproduziert.

Unweigerlich stellt sich nun die Frage, wie man misst, ob eine Approximation die bestmögliche ist. Hierzu ist es notwendig, den Unterschied zwischen vorhergesagter Ausgabe und beobachteter (d. h. verrauschter wahrer) Ausgabe zu quantifizieren. Dies geschieht im statistischen maschinellen Lernen mithilfe sogenannter globaler oder lokaler Verlustfunktionen, die entsprechend dem zu lösenden Problem (beispielsweise Regression oder Klassifikation) konstruiert werden. Dabei ist festzuhalten, dass eine Approximation gesucht wird, die für alle denkbaren Eingabe- und Ausgabewerte brauchbar ist und nicht nur für die dem Anwender bekannten Werte aus dem Trainingsdatensatz. Es bedarf daher weiterer Überlegungen, um eine Überanpassung des Prädiktors an den Trainingsdatensatz zu vermeiden. Im Allgemeinen wird ein Teil der zur Verfügung stehenden Daten vom Trainingsdatensatz abgetrennt und dient nach dem Lernen als Testdatensatz, um die Güte des gelernten Prädiktors (beispielsweise anhand der Missklassifikationsrate) schätzen zu können.

[1] Andere Bereiche des maschinellen Lernens sind das nichtüberwachte Lernen (unsupervised learning), das halbüberwachte Lernen (semisupervised learning) und das ver- oder bestärkende Lernen (reinforcement learning), siehe beispielsweise Russel und Norvig (2012, S. 811).

[2] Die nachfolgenden Beschreibungen sind im Wesentlichen jedoch auch für Regression oder die Klassifikation mit mehr als zwei möglichen Optionen gültig.

Dies ist möglich, da ja im Fall des überwachten Lernens auch für den Testdatensatz die (verrauschten) wahren Ausgabewerte bekannt sind.

Statistisches maschinelles Lernen bedient sich meist nichtparametrischer Methoden. Im Unterschied zu parametrischen Verfahren unterstellen solche Methoden nicht bereits von vorneherein ein bestimmtes Modell von Verteilungen, dem die Schritte (1) und (2) in Vapniks Zerlegung folgen. Die Zusammenhänge zwischen Eingabe- und Ausgabevariablen werden durch die nichtparametrischen Methoden erst noch entdeckt. Somit lösen nichtparametrische Verfahren dieses Grundproblem der Statistik, die zugrunde liegende Verteilung anhand der gegebenen Daten zu schätzen, wesentlich allgemeiner als parametrische Verfahren.[3]

2.2 Einsatzgebiete außerhalb der amtlichen Statistik

Methoden des statistischen maschinellen Lernens werden heutzutage in nahezu allen Wissenschaftsgebieten eingesetzt, deren Objekte in geeigneter Weise quantifizierbar sind. Beispiele sind dabei klassisch die Muster- oder Zeichenerkennung (Erkennen von z. B. handgeschriebenen oder gescannten Buchstaben und Ziffern; siehe u. a. LeCun et al. 1998), der Einsatz in der Bioinformatik (beispielsweise Baldi und Brunak 2001) sowie das Erkennen physikalischer Phänomene, beispielsweise Carrasquilla und Melko (2017) und Wang (2016) für das Erkennen von Phasen und Phasenübergängen bei Materie oder Carleo und Troyer (2017) für die Lösung der Schrödingergleichung in Vielteilchensystemen. Weitere Einsatzfelder sind die Kreditrisikoanalyse (beispielsweise Yu et al. 2008), Analyse von sozialen Netzwerken (beispielsweise Murty und Raghava 2016), Steganalyse (beispielsweise Schaathun 2012), Erkennung von Steroiden im Antidopingkampf (van Renterghem et al. 2013), Vorhersage der Wasserqualität (Singh et al. 2011), Schätzung der Bodenversiegelung aufgrund von Fernerkundungsdaten (Bachofer et al. 2009), Demokratiemessung (Gründler und Krieger 2015) und viele mehr.

2.3 Einsatzgebiete in der amtlichen Statistik

Die amtliche Statistik außerhalb Deutschlands nutzt ebenfalls maschinelle Lernverfahren. Eine von Statistics Canada ins Leben gerufene Machine Learning Documentation Initiative (Chu und Poirier 2015) hat das Ziel, einen Überblick über durchgeführte oder geplante Einsätze maschinellen Lernens in Statistischen Ämtern weltweit

[3] Bei genauerer Betrachtung ist zu erkennen, dass bislang hauptsächlich auf den Begriff des statistischen Lernens eingegangen wurde. Jede Berechnung einer Regressionsfunktion o. ä. kann im weiteren Sinne als statistisches Lernen bezeichnet werden, erfasst es doch die Informationen im Datensatz um später zu neuen Beobachtungen entsprechende Outputwerte vorherzusagen. Damit wird auch deutlich, dass die vorgestellten Methoden bessere Ergebnisse als die altbewährten liefern können, aber nicht müssen. Der Anteil des „Maschinellen“ wurde hingegen noch nicht verdeutlicht. Dass man von statistischem *maschinellen* Lernen spricht liegt darin begründet, dass einige der heute unter diesem Begriff firmierenden Methoden vor der Entwicklung entsprechend leistungsstarker Rechner zwar theoretisch denkbar, praktisch jedoch ohne maschinelle Unterstützung nicht oder nicht für große Datenmengen durchführbar waren. Brücken zu den Forschungsgebieten „Big Data“, „Data Mining“, „Künstliche Intelligenz“ und im Hinblick auf die Algorithmen auch zur Informatik könnten an dieser Stelle ohne weiteres geschlagen werden, sollen aber nicht Bestandteil des Aufsatzes sein.

zu schaffen. Als Beispiele werden unter anderem genannt: Klassifikation von Berufen mit dem sogenannten Naive-Bayes-Algorithmus, Klassifikation von kategorialen Zensus-Variablen mittels SVMs, Imputationen mittels Neuronalen Netzwerken, Bäumen, Clusteranalyse und Random Forests, Record Linkage mittels Bäumen oder Support Vector Machines und Erkennung von Steuerhinterziehung mittels k-Nearest-Neighbour-Verfahren. Da die die Daten generierenden Verteilungen in der amtlichen Statistik in der Regel unbekannt sind, erscheint die Erprobung und ggf. der Einsatz von Methoden des statistischen maschinellen Lernens angebracht.

Die Statistischen Ämter des Bundes und der Länder nutzten unseres Wissens mit Ausnahme der in diesem Aufsatz vorgestellten Ansätze bislang keine statistischen maschinellen Lernverfahren.

2.4 Argumente für den Einsatz von Support Vector Machines und Random Forests

Unter einer Vielzahl von Machine-Learning-Methoden stechen zwei besonders heraus, wenn man als Kriterium ihren Erfolg bei Klassifikationsaufgaben, also bei der Schätzung, zu welcher Klasse eine neue Beobachtung gehört, betrachtet: Support Vector Machines und Random Forests. Sie erreichen in der Regel wenigstens die gleiche Güte wie andere Klassifikationsverfahren, häufig übertreffen sie die anderen Methoden (Bennett und Campbell 2000; Caruana und Niculescu-Mizil 2006; Kotsiantis 2007; Caruana et al. 2008; Fernández-Delgado et al. 2014; Wainberg et al. 2016). Beide Verfahren sind nichtparametrisch-statistische Methoden, setzen also nicht voraus, dass der Zusammenhang zwischen Eingabe- und Ausgabevariablen einer bestimmten Verteilungsklasse folgt. Statt nur die Parameter einer Verteilung zu schätzen, kann der grundlegende Zusammenhang selbst durch die nichtparametrisch-statistische Methode ermittelt werden.

Allerdings ist zu beachten, dass die von uns verwendeten Implementierungen von SVM und Random Forest im Unterschied zur logistischen Regression aktuell keine Survey-Gewichte berücksichtigen.

2.5 Beschreibung von Random Forests

Random Forests, als solche benannt und eingeführt durch Breiman (2001), basieren auf den ebenfalls von Breiman et al. (1984) eingeführten Klassifikations- oder Entscheidungsbäumen. Einen komprimierten Überblick über die Random-Forest-Methodik liefert Boulesteix et al. (2012).

Die Herangehensweise eines Klassifikationsbaumes besteht darin, die Grundgesamtheit der Eingabevariablen in sich nicht überlappende Regionen aufzuteilen. Zunächst wird hierzu eine der erklärenden Variablen ausgewählt sowie ein Schwellenwert bestimmt. Beobachtungen, die hinsichtlich dieser Variable unterhalb des Schwellenwertes liegen, werden einer Teilregion zugewiesen, die übrigen einer anderen. Dieses Verfahren wiederholt sich im Folgenden sukzessive für die so entstehenden Teilregionen. Kriterium für die Auswahl der Variable und des Schwellenwertes ist, dass die Beobachtungen innerhalb derselben Region große Übereinstimmung hinsichtlich der Ausprägung der Ausgabevariable und somit eine geringe Varianz

aufweisen (also homogen sind). In jedem Schritt, d. h. an jedem Knoten des Baumes, ist daher ein Optimierungsproblem zu lösen, das davon abhängt, wie der Anwender den Begriff der Homogenität definiert. Im Falle der Klassifikation ist dies in der Regel die Reinheit bzgl. der Klassenzugehörigkeit der Beobachtungen in einer Region. Verschiedene Maße für die Homogenität im Klassifikationsfall werden beispielsweise in James et al. (2013, S. 312) diskutiert. Der Algorithmus agiert anschaulich für die am jeweiligen Knoten betrachtete Region wie nachfolgend beschrieben: Die Region wird probeweise in alle möglichen und im Hinblick auf die erklärenden Variablen sinnvollen Paare von Teilregionen aufgeteilt. Anschließend werden die dadurch entstandenen neuen Reinheitsgrade berechnet. Die Aufteilung, welche die stärkste Zunahme an Reinheit erzielt, wird ausgewählt (Breiman et al. 1984, S. 26 ff.). Eine später neu hinzukommende, zu klassifizierende Beobachtung wird anhand ihrer Eingabewerte einer Region zugewiesen. Als Schätzung für die Klassenzugehörigkeit der neuen Beobachtung wird dann die in der Region vorherrschende Klasse herangezogen. Das Verfahren, die Regionen in immer kleinere Teilregionen aufzuteilen, könnte nun so lange fortgesetzt werden, bis jedes Element des Trainingsdatensatzes eine eigene Region definiert. Dies wäre ein Beispiel für eine Überanpassung (overfitting) des Baumes an die Trainingsdaten und birgt die Gefahr schlechter Resultate bei der Klassifizierung neuer Objekte. Der Baum darf daher zunächst sehr weit verzweigt, also sehr komplex sein, muss anschließend aber wieder „zurückgeschnitten“ werden (pruning).[4] Als erster Ansatz bietet sich hier folgendes Vorgehen an: Für jeden möglichen Teilbaum, der durch Wiedervereinigen von Regionen entsteht, wird die Verschlechterung in der Klassifikationsgüte geschätzt; der Teilbaum mit der kleinsten Verschlechterung wird schließlich ausgewählt. Dieses Vorgehen erscheint jedoch als sehr aufwändig, wenn man die Anzahl der möglichen Teilbäume berücksichtigt. Meist wird daher das sogenannte Cost Complexity Pruning eingesetzt, das im Wesentlichen darauf beruht, nicht alleine die Varianz innerhalb einer Region zu minimieren, sondern darüber hinaus auch noch die Anzahl der Endknoten, also die Zahl der Blätter des Baumes. Der Trade-Off zwischen Homogenität der Regionen und Komplexität des Baumes wird somit in das Optimierungsproblem integriert. Als Nebeneffekt erhält man für jeden gelernten Baum eine Übersicht darüber, welche Variablen den größten Beitrag zur Verbesserung der Homogenität in den verschiedenen Knoten liefern. Dies ermöglicht eine Sortierung der Eingabevariablen nach deren Wichtigkeit für die Klassifikation. Ergänzend zu oder anstelle einer zunächst sehr tiefen Verzweigung und anschließendem Zurückschneiden des Baumes sind auch Abbruchkriterien für die Aufteilungsschritte denkbar. Beispielsweise könnte ein Knoten nicht weiter verzweigt werden, d. h. eine Region nicht weiter aufgeteilt werden, wenn bereits eine Höchstzahl an Knoten vorhanden ist oder die Anzahl an Beobachtungen in einer Region andernfalls eine Mindestanzahl unterschreiten würde.

Bei Random Forests soll nun nicht nur ein einzelner Baum die Klassifikation vornehmen, d. h. die Zugehörigkeit zu einer Klasse schätzen, sondern ein ganzes Ensemble von Bäumen per (relativem) Mehrheitsentscheid. Empirisch zeigt sich

[4] Auf das pruning kann und soll verzichtet werden, wenn der Baum nicht alleine die Klassifikation vornimmt, sondern Eingang in einen Random Forest findet (Wyner et al. 2017).

dieser Ansatz gegenüber der Nutzung eines einzelnen Baumes meist überlegen. Kleinere Veränderungen in den Daten (Hinzunahme eines neuen oder Streichung eines alten Datenpunktes, Messungenauigkeiten etc.) können nämlich zu gänzlich veränderten Baumstrukturen führen, was die Klassifikationsergebnisse in solchen Bereichen des Eingaberaums, in dem sich die beiden Klassen überlappen, beeinflussen, ggf. sogar willkürlich erscheinen lassen kann. Die Klassifikation auf Grundlage eines einzigen Baumes erscheint unter diesem Gesichtspunkt ungünstig. Es gibt verschiedene Verfahren, wie (anhand eines festen, gegebenen Trainingsdatensatzes) viele sich voneinander unterscheidende Bäume gelernt werden können. Random Forests generieren die B verschiedenen Bäume zum einen dadurch, dass nicht anhand des Trainingsdatensatzes (mit Umfang n_{train}) selbst, sondern anhand von B sogenannten Bootstrap-Stichproben gelernt wird, d. h. anhand von Datensätzen mit Umfang n_{train}, die durch n_{train}-faches zufälliges Ziehen mit Zurücklegen aus dem Trainingsdatensatz entstehen. Zum anderen werden die $mtry$ für die Auswahl der optimalen Splitvariable in Frage kommenden Variablen (bei kategorialen Variablen nach Aufteilung in Dummy-Variablen also die Ausprägungen der kategorialen Variable) für jeden Split zufällig aus allen $m(> mtry)$ erklärenden Variablen gezogen. Es stehen also für verschiedene Bäume an verschiedenen Knoten unterschiedliche Eingabevariablen zur Verfügung, um die Ausgabe zu erklären. Diese zweite Variation in den Bäumen wirkt sich je Baum positiv auf die Rechenzeit aus, da pro Knoten nicht mehr alle Variablen als Splitvariable in Frage kommen. Außerdem sind die Bäume dadurch weniger stark korreliert, was wünschenswert ist, wenn man – wie im Falle von Random Forests – Informationen aus mehreren Bäumen gemeinsam für die Klassifikation heranziehen möchte. Die Information über die für die Klassifikation wichtigen Variablen kann aus den Bäumen in einen daraus entstandenen Random Forest übertragen werden, siehe James et al. (2013, S. 319).

Hinsichtlich der Skalen der erklärenden Variablen sind Random Forests sehr flexibel: nominal skalierte Merkmale können ebenso verarbeitet werden wie kontinuierliche Variablen.

Klassifikationsbäume optimieren in jedem Schritt, ohne ein globales Optimum (Auffinden des „besten“ Baumes) anzustreben. Das Auffinden eines „global besten“ Baumes ist NP-schwer (Hyafil und Rivest 1976), in praktischen Anwendungen aufgrund extrem langer Laufzeiten also unmöglich umzusetzen. Wenngleich Random Forests viele der damit verbundenen Nachteile auszugleichen versuchen, bieten die nachfolgend erläuterten Support Vector Machines eben dies: Eine „global beste“ Lösung.

2.6 Beschreibung von Support Vector Machines

Eingeführt durch die Arbeiten von Boser et al. (1992) und Cortes und Vapnik (1995) sind Support Vector Machines (SVMs) Lösungen des folgenden Optimierungsproblems: Minimiere den mittleren, geeignet gemessenen Abstand zwischen geschätztem und wahrem Ausgabewert für alle möglichen Eingabewerte. Da der geschätzte Ausgabewert eine Funktion des jeweiligen Eingabewertes ist, lässt sich das Optimierungsproblem auch anders formulieren: Finde diejenige Funktion, welche den mittleren, geeignet gemessenen Abstand zwischen geschätztem und wahrem Aus-

gabewert für alle möglichen Eingabewerte minimiert. Nicht überraschend kommt dies der mathematischen Beschreibung des statistischen maschinellen Lernens durch Vapnik sehr nahe. Zu SVMs gibt es inzwischen eine große Fülle an Literatur. Eine allgemeinverständliche Einführung in SVMs bietet Hamel (2009); eine genauere Betrachtung aus statistischer Sicht liefert Steinwart und Christmann (2008).

Im Falle der Klassifikation für zwei Klassen liegt es nahe, den Abstand zwischen geschätztem und wahrem Ausgabewert als 1 (oder eine beliebige andere positive Zahl) festzulegen, falls Vorhersage und wahrer Wert nicht übereinstimmen, und als 0 für den Fall einer richtigen Vorhersage. Aus mathematischer Sicht bietet diese Definition jedoch den Nachteil, dass die so festgelegte Missklassifikationsverlustfunktion L (ein Abstand zwischen Vorhersage und wahrem Wert bedeutet einen Verlust in der Vorhersagekraft des Verfahrens) nicht konvex ist, was Existenz und Eindeutigkeit sowie die algorithmische Auffindbarkeit eines Minimums des durchschnittlichen Abstandes bzw. des durchschnittlichen Verlusts in Frage stellt. Stattdessen definiert man den Verlust beispielsweise als das Maximum aus 0 und $1 - \mathrm{yf(x)}$, also $\mathrm{L(y, f(x))} = \max\{0, 1 - \mathrm{yf(x)}\}$, wobei y die wahre Klassenzugehörigkeit (codiert als –1 oder +1), f ein Zusammenhang zwischen Eingabe- und Ausgabewerten und x ein Eingabewert ist. Das f(x) stellt dann die vorhergesagte Klasse dar, wobei aus technischen Gründen f(x) nicht auf –1 und 1 beschränkt ist, sondern beliebige reelle Werte annehmen kann. Negative Werte von f(x) würden dann als Vorhersage für die mit –1 codierte Klasse interpretiert, alle anderen Werte von f(x) als Vorhersage für die mit +1 codierte Klasse. Eine so definierte Abstandsfunktion wird als Hinge-Verlustfunktion bezeichnet und ist konvex. Man kann zeigen, dass die Lösung des Optimierungsproblems bzgl. der Hinge-Verlustfunktion die Lösung bzgl. der Missklassifikationsverlustfunktion bereits impliziert. Es gibt weitere Möglichkeiten, einen Abstand bei Klassifikation sinnvoll zu definieren, beispielsweise, wenn nicht nur Konvexität, sondern auch Differenzierbarkeit der Verlustfunktion gefordert ist. Ein Beispiel für diesen Fall bietet die logistische Verlustfunktion für Klassifikation: $\mathrm{L(y, f(x))} = \ln(1 + \exp(-\mathrm{yf(x)}))$.

Aus Sicht eines Anwenders der SVM-Methodik wäre es wünschenswert, wenn alle Vorhersagen zuträfen, der mittlere Verlust also 0 wäre. Ob dies bei Betrachtung aller möglichen Eingabe- und deren zugehöriger Ausgabewerte überhaupt erreichbar ist, ist unbekannt, da zum Zeitpunkt des Lernens nur eine Stichprobe und nicht die gesamte Population zur Verfügung steht. Support Vector Machines versuchen daher, alle Informationen aus den Daten zu nutzen und minimieren den mittleren Verlust auf Basis des Trainingsdatensatzes zuzüglich eines Strafterms, das heißt die SVM $f_{L,\lambda}$ minimiert

$$\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{L}(y_i, \mathrm{f}(x_i)) + \lambda \|\mathrm{f}\|^2$$

über alle zulässigen Funktionen f. Welche Funktionen zulässig sind, entscheidet der Anwender meist durch Wahl eines Kernparameters. Häufig werden sogenannte Gauß-Kerne genutzt, um den Raum der zulässigen Funktionen festzulegen. Im Hintergrund steht hierbei die Theorie der reproduzierenden Kern-Hilberträume; für Details hierzu siehe beispielsweise Schölkopf und Smola (2002) oder Steinwart und

Christmann (2008). $\|f\|$ ist die Norm einer Funktion f in einem solchen reproduzierenden Kern-Hilbertraum, also eine reelle Zahl, die der Funktion f aufgrund deren Eigenschaften zugewiesen wird. Je größer die Norm ist, desto komplexer ist f. Beispielsweise hat eine die Daten stark interpolierende Funktion eine größere Norm, eine die Daten allenfalls ausgleichende Funktion eine kleinere Norm. Der Parameter λ bestimmt nun den Einfluss der Komplexität der Funktion auf das Minimierungsproblem im Vergleich zu dem durch diese Funktion entstehenden mittleren Verlust. Die Wahl von λ obliegt dem Anwender innerhalb gewisser Grenzen, welche die statistischen Eigenschaften der SVMs garantieren. Wäre $\lambda = 0$, würde man also auf den Strafterm verzichten, so würde hohe Komplexität nicht bestraft und es bestünde die Gefahr der Überanpassung an den Trainingsdatensatz. Für die endlich vielen bekannten Datenpunkte wäre die Anpassung in diesem Fall hervorragend, bei neu zu klassifizierende Beobachtungen würde eine ohne Strafterm gelernte SVM potentiell schlecht abschneiden. Numerisch resultiert die Optimierung bei der Hinge-Verlustfunktion in einem quadratischen Programm mit Box-Constraints, also einem Optimierungsproblem mit sehr guten Konvergenzeigenschaften; die Optimierung bei der logistischen Verlustfunktion liefert hingegen lediglich ein konvexes Programm mit schlechteren Konvergenzeigenschaften.

Wie Random Forests können auch SVMs sowohl mit nominal skalierten als auch mit kontinuierlichen Eingabevariablen umgehen.

2.7 Kurze Diskussion von Random Forests und Support Vector Machines

Wägt man SVMs und Random Forests gegeneinander ab, so ist festzuhalten, dass hinsichtlich der Geschwindigkeit des Lernens und der Auswertung Random Forests SVMs überlegen sind, ebenso hinsichtlich der Interpretierbarkeit des resultierenden Prädiktors (zumindest sind die den Random Forests zugrunde liegenden Bäume der SVM-Methode diesbezüglich überlegen; ein Random Forest selbst nur noch bedingt). Bezüglich garantierter statistischer Eigenschaften ist es umgekehrt: Hier sind SVMs die sichere Wahl. Support Vector Machines können zum Teil auch hinsichtlich der Güte bessere Ergebnisse liefern als Random Forests, was gemeinsam mit den garantierten statistischen Eigenschaften wie insbesondere Robustheit im geeigneten Sinne ihren Einsatz rechtfertigt.

3 Konkrete Anwendungen

3.1 Zuordnung von Unternehmen zum Dritten Sektor[5]

Ziel der hier beschriebenen Untersuchung war es, a priori nicht eindeutig zuzuordnende Unternehmen im statistischen Unternehmensregister (URS), siehe Sturm und Tümmler (2006), hinsichtlich ihrer Zugehörigkeit zum sogenannten Dritten oder Non-Profit-Sektor zu klassifizieren – idealerweise derart, dass das Verfahren auch in späteren Jahren einsetzbar ist und somit auf aufwändige Einzelfallrecherchen ver-

[5] Zu Hintergründen und weiteren Details zu nachfolgendem Abschnitt siehe Dumpert et al. (2016).

zichtet werden kann. Als Klassifikationsmethode sollten Support Vector Machines erprobt werden. Der Begriff des Dritten Sektors bezeichnet dabei nicht-gewinnorientierte, gemeinnützige Organisationen außerhalb des privaten und öffentlichen Sektors, die keine privaten Haushalte sind. Die Zuordnung von Unternehmen zum Dritten Sektor ermöglicht die Bestimmung der Anzahl der Unternehmen und der darin sozialversicherungspflichtig Beschäftigten in diesem Bereich sowie die zugehörige Bruttowertschöpfung. Hintergrund der Unterscheidung nach Zugehörigkeit zum Dritten Sektor ist zum einen ein Interesse an der Bemessung von dessen volkswirtschaftlicher Bedeutung, zum anderen die Vorgabe der EU, den institutionellen Sektor nach dem Europäischen System Volkswirtschaftlicher Gesamtrechnungen als Merkmal von Unternehmen im Unternehmensregister zu speichern. Das Merkmal „institutioneller Sektor" erlaubt u. a. die Abgrenzung des „Staates" sowie die Unterscheidung von Unternehmen, die marktaktiv bzw. nicht marktaktiv sind. Ersteres ist beispielsweise für die Berechnung von Defizitkriterien bedeutsam, letzteres für die Abgrenzung von Grundgesamtheiten, aus denen Stichproben derjenigen Unternehmensstatistiken gezogen werden, die sich auf marktaktive Unternehmen beschränken.

Ein großer Teil der im URS erfassten Unternehmen kann aufgrund inhaltlicher Überlegungen und mithilfe eines daraus resultierenden Algorithmus als dem Dritten Sektor zugehörig oder nicht zugehörig klassifiziert werden. Für ca. 45.000 Unternehmen, für die der Algorithmus keine eindeutige Zuordnung zuließ, waren für das Berichtsjahr 2007 Einzelfallrecherchen notwendig. A priori nicht eindeutig zuzuordnende Unternehmen sind demnach solche, für welche der Standard-Algorithmus keine eindeutige Klassifikation ermöglicht. Details zu den allgemeinen Voraussetzungen werden in Rosenski (2012) aufgeführt.

Aufgrund der oben genannten Einzelfallrecherchen standen für ca. 45.000 a priori nicht eindeutig zuzuordnende Unternehmen sowohl die üblichen Merkmale des URS als auch als verlässlich eingestufte Klassifizierungen bezüglich der Zugehörigkeit zum Dritten Sektor zur Verfügung. Diese Situation stellte sowohl hinsichtlich des Umfangs der Daten als auch hinsichtlich der Vollständigkeit der Informationen eine sehr gute Ausgangslage für statistisches maschinelles Lernen und die Anwendung von Support Vector Machines dar.

Es wurde untersucht, ob Support Vector Machines geeignet sind, die bislang erforderlichen, personal- und zeitaufwändigen Einzelfallrecherchen für a priori nicht eindeutig zuzuordnende Unternehmen zu ersetzen. Zu klären war also, ob eine SVM in der Lage wäre, weitere, über die im Standard-Algorithmus bereits abgebildeten Regeln hinausgehende Strukturen in den Trainings- bzw. Testdaten aufzufinden. Hierzu musste zunächst ermittelt werden, welche Eingabevariablen beim Lernen der SVM genutzt werden sollten. Plausible Kriterien für eine solche Auswahl sind:

- Optimalität: Es ist diejenige Variablenkombination zu wählen, welche im Mittel (über verschiedene zufällige Aufteilungen des Datensatzes in Trainings- und Testdatensatz) das geringste Missklassifikationsrisiko liefert.
- Stabilität: Es ist diejenige Variablenkombination zu wählen, welche im Mittel die geringste Schwankung des Missklassifikationsrisikos aufweist.

- Recheneffizienz bzw. Variableneffizienz: Es ist diejenige Variablenkombination zu wählen, welche den geringsten Rechenaufwand verursacht bzw. welche die geringste Anzahl an erklärenden Variablen benötigt.

Testrechnungen ergaben, dass die Nutzung der erklärenden Variablen „Wirtschaftszweig des betrachteten Unternehmens“, „Status als öffentliche Einheit“, „Rechtsform des Unternehmens“ und „Bundesland des Sitzes des Unternehmens“ sowie der Präklassifikatoren auf Basis des Wirtschaftszweigs und der Rechtsform obige drei Aspekte am besten erfüllen. Die Präklassifikatoren enthalten fachliche Einschätzungen, ob der jeweils betrachtete Wirtschaftszweig bzw. die jeweils betrachtete Rechtsform dem Dritten Sektor mutmaßlich zuzuordnen ist, möglicherweise zuzuordnen ist oder eher nicht zuzuordnen ist.

Insgesamt lagen 45.662 Einheiten vor, davon wurden 27.398 zum Trainieren und 9133 zur Wahl der beiden Parameter (Kernparameter und λ) der SVM genutzt. Die verbleibenden 9131 Einheiten bildeten den Testdatensatz.

Es ergaben sich die Werte in Tab. 1 und 2.

Tab. 1 Geschätzte Missklassifikationsraten bei Dritter-Sektor-Zuordnung in % (aus Dumpert et al. (2016))

	Fälschlich nicht dem Dritten Sektor zugeordnet	Fälschlich dem Dritten Sektor zugeordnet	Summe	Saldo bezüglich der Zuordnung zum Dritten Sektor
	(1)	(2)	(1) + (2)	(2) – (1)
Anzahl Unternehmen	5,0	8,9	13,9	+3,9
Steuerbarer Umsatz	11,6	7,9	19,5	–3,7
Sozialversicherungspflichtig Beschäftigte	6,1	8,9	15,0	+2,8

Tab. 2 Geschätzte Sensitivitäten und Spezifitäten bei Dritter-Sektor-Zuordnung in % (aus Dumpert et al. (2016))

	Anteil der durch die SVM dem Dritten Sektor zugeordneten Unternehmen an den tatsächlichen Dritt-Sektor-Unternehmen (Sensitivität)	Anteil der durch die SVM dem Dritten Sektor nicht zugeordneten Unternehmen an den tatsächlichen Nicht-Dritt-Sektor-Unternehmen (Spezifität)
Anzahl Unternehmen	90,5	81,5
Steuerbarer Umsatz	78,6	82,8
Sozialversicherungspflichtig Beschäftigte	87,7	82,5

3.2 Identifikation für die Handwerksstatistik relevanter Unternehmen[6]

Seit dem Berichtsjahr 2008 werden die Handwerksstatistiken vollständig aus Verwaltungsdaten erstellt, d. h. aus bereits durch andere Stellen (beispielsweise den Handwerkskammern, der Bundesagentur für Arbeit oder der Finanzverwaltung) erhobenen Daten. Dieses die Handwerksbetriebe entlastende Verfahren enthält jedoch die Schwierigkeit, dass der Handwerksbegriff des Handwerksstatistikgesetzes nicht mit dem Handwerksbegriff beispielsweise der Handwerkskammern übereinstimmt. Diese Abweichung führt dazu, dass nicht alle dem Statistischen Bundesamt von den Handwerkskammern gemeldeten Handwerksbetriebe auch für die Handwerksstatistik relevant sind und aus diesem Grund identifiziert und aussortiert werden müssen. Die Handwerkskammern als Datenlieferanten verfügen nicht über die hierzu notwendigen Informationen, sodass diese Aufgabe durch die Fachbereiche Handwerk der Statistischen Ämter der Länder erfüllt wird. Dazu werden Informationen aus unterschiedlichen Quellen wie beispielsweise dem Internet, gegebenenfalls vorliegenden Geschäftsberichten oder dem Handelsregister, recherchiert. Diese manuellen Recherchen binden in erheblichem Umfang personelle Ressourcen, weshalb im Rahmen eines Projektes der Einsatz von maschinellen Lernverfahren untersucht werden sollte.

Zur besseren Einschätzbarkeit des Problems: Von den rund 600.000 Unternehmen, die für das Jahr 2012 von den Handwerkskammern als Handwerksunternehmen gemeldet wurden, waren 1,3 % nicht relevant für die Handwerksstatistik. Sie repräsentierten jedoch 19 % der sozialversicherungspflichtig Beschäftigten und erzielten 35 % der Umsätze. Diese Zahlen sowie die Verteilung der Umsätze in Handwerks- und Nichthandwerksunternehmen legen nahe, dass vorwiegend große Unternehmen nicht relevant für die Handwerksstatistiken sind.

Zu den Handwerksunternehmen liegen Informationen über den Wirtschafts- und den Gewerbezweig vor. Eine Mehrheit der auftretenden Kombinationen aus Wirtschafts- und Gewerbezweig lässt sich relativ leicht hinsichtlich der statistischen Zugehörigkeit zum Handwerk klassifizieren, weil dort entweder nur für die Handwerksstatistiken relevante oder nur nicht relevante Unternehmen vorkommen. Bei einer Minderheit sind jedoch manuelle Zuordnungen erforderlich. Im Berichtsjahr 2012 wurden 6100 von den Handwerkskammern erstmals gemeldete Unternehmen dem Handwerk oder Nichthandwerk (im statistischen Sinne) durch Einzelfallrecherchen der Statistischen Ämter der Länder zugeordnet.

Als Datenmaterial zum Trainieren und Testen des statistischen maschinellen Lernverfahrens stand ein Auszug aus dem statistischen Unternehmensregister zur Verfügung, der verschiedene Variablen bzgl. der Größe, der Handwerkseigenschaft und der Struktur der zu klassifizierenden Unternehmen enthielt. Der Umfang des Datensatzes betrug ca. eine Million Beobachtungen handwerksrelevanter Unternehmen aus dem statistischen Unternehmensregister der Bezugsjahre 2011 und 2012. Dabei konnten sehr viele trivial zu klassifizierende Beobachtungen entfernt werden, was die Handhabbarkeit des Problems erheblich verbesserte. Nach Ausschluss dieser Unter-

[6] Zu Hintergründen und weiteren Details zu nachfolgendem Abschnitt siehe Feuerhake und Dumpert (2016).

Tab. 3 Geschätzte Missklassifikationsraten (bei SVM) bei Handwerkszugehörigkeit in % (aus Feuerhake und Dumpert (2016))

	Fälschlich nicht dem Handwerk zugeordnet	Fälschlich dem Handwerk zugeordnet	Summe	Saldo bezüglich der Zuordnung zum Handwerk
	(1)	(2)	(1) + (2)	(2) – (1)
Anzahl Unternehmen	0,3	3,4	3,7	+3,1
Steuerbarer Umsatz	0,3	1,5	1,8	–1,2

Tab. 4 Geschätzte Sensitivitäten und Spezifitäten (bei SVM) bei Handwerkszugehörigkeit in % (aus Feuerhake und Dumpert (2016))

	Anteil der durch die SVM dem Handwerk zugeordneten Unternehmen an den tatsächlichen Handwerksunternehmen (Sensitivität)	Anteil der durch die SVM dem Handwerk nicht zugeordneten Unternehmen an den tatsächlichen Nicht-Handwerksunternehmen (Spezifität)
Anzahl Unternehmen	99,7	48,5
Steuerbarer Umsatz	99,5	95,7

nehmen verblieben ca. 68.000 Einheiten, die zu klassifizieren waren. Der reduzierte Datensatz enthielt 6,5 % für die Handwerksstatistik nicht relevante Unternehmen, die 34,6 % der Umsätze repräsentierten.

Eine große Zahl der Eingabevariablen ist untereinander korreliert. Dies stellt für Methoden des statistischen maschinellen Lernens im Unterschied zu beispielsweise der logistischen Regression kein Hindernis dar. Aus Effizienzgründen (Rechenzeit und Speicherkapazität) ist man jedoch daran interessiert, die Anzahl der Variablen nach Möglichkeit zu reduzieren. Für diesen ersten Schritt wurden Random Forests gelernt und auf den Trainingsdatensatz angewandt. Wie oben beschrieben, bieten Random Forests die Möglichkeit, Informationen über für die Klassifikation entscheidende Variablen zu erhalten. Diese Eigenschaft von Random Forests wurde bei dieser Untersuchung genutzt.[7] (Zwar liefert die Random-Forest-Methode auch selbst Klassifikationsergebnisse, die im Folgenden eigenständig genutzt werden könnten, doch waren im vorliegenden Fall die Ergebnisse bei einer Klassifikation durch SVMs besser, siehe Tab. 1 in Feuerhake und Dumpert (2016).)

Nach Reduzierung der Zahl der Eingabevariablen auf die gemäß dem Random Forest relevantesten 30 wurde in einem zweiten Schritt eine SVM anhand des Trainingsdatensatzes und des so verringerten Merkmalskranzes gelernt. Es ergaben sich die in Tab. 3 und 4 dargestellten Ergebnisse.

[7] Der Ansatz, Random Forests zur Identifizierung der für den späteren SVM-Algorithmus heranzuziehenden Variablen zu nutzen, wurde unter anderem auch von Löw et al. (2013) gewählt.

3.3 Schätzung der Muttereigenschaft (zur Verbesserung der Schätzung des bereinigten Gender Pay Gaps)[8]

Das Ausmaß der durchschnittlichen Lohnunterschiede von Männern und Frauen (unbereinigter Gender Pay Gap) sowie die Frage, ob diese ggf. auf Lohndiskriminierung von Frauen zurückzuführen sind, ist seit Jahrzehnten ein wichtiges Thema in Politik und Forschung. Das Statistische Bundesamt hat auf Basis der vierjährlich durchgeführten Verdienststrukturerhebung (VSE) 2006, 2010 und 2014 den sogenannten bereinigten Gender Pay Gap geschätzt, der Unterschiede im durchschnittlichen Bruttostundenverdienst von Männern und Frauen mit vergleichbaren Qualifikationen, Tätigkeiten und Erwerbsbiographien misst und den jährlich berechneten unbereinigten Gender Pay ergänzt (Finke 2011; Finke et al. 2017).

Anders als bei der unter anderem mit Daten des sozioökonomischen Panels arbeitenden Studie von Boll und Leppin (2015), stehen dem Statistischen Bundesamt bislang keine Informationen über Erwerbsunterbrechungen von Frauen im Berufsleben zur Verfügung. Zur Erklärung von Verdienstunterschieden zwischen Männern und Frauen im Rahmen der Schätzung des bereinigten Gender Pay Gaps können solche Informationen daher nicht herangezogen werden. Sowohl für Männer als auch für Frauen wird stattdessen die sogenannte potentielle Berufserfahrung aus Lebensalter und Qualifikationsniveau geschätzt. Erwerbsunterbrechungen aufgrund von Geburt und Erziehung von Kindern werden hierbei mangels Kenntnis darüber nicht berücksichtigt. Somit ist die potentielle Berufserfahrung gegebenenfalls größer als die tatsächliche. Ein Arbeitgeber weiß in der Regel um diese fehlende Berufserfahrung und bezahlt eine Arbeitnehmerin entsprechend. Aus Sicht der amtlichen Statistik wirkt dies, als ob ein zusätzliches Jahr potentieller Berufserfahrung bei Frauen geringer honoriert würde als bei Männern. Bei einer Berücksichtigung der tatsächlichen anstelle der potentiellen Berufserfahrung in den Schätzmodellen würde der bereinigte Gender Pay Gap niedriger ausfallen. Die Ausprägungen der beiden Variablen unterscheiden sich durch die Dauer der Erwerbsunterbrechungen.

Um dieses Defizit zu verkleinern, sollten die Daten der Verdienststrukturerhebung, siehe Statistisches Bundesamt (2016), auf deren Basis der bereinigte Gender Pay Gap geschätzt wird, um eine frauenspezifische Variable „Erwerbsunterbrechung aufgrund der Geburt eines Kindes (ja/nein)“ angereichert werden. Damit wird es dem Statistischen Bundesamt ermöglicht, diese Form der Erwerbsunterbrechung bei der Erklärung von Verdienstunterschieden zu berücksichtigen. Der Mikrozensus 2012, siehe Statistisches Bundesamt (2012), enthält das (freiwillig zu beantwortende) Merkmal „Haben Sie Kinder geboren?“, das als Indiz für das Vorliegen einer Erwerbsunterbrechung genutzt wurde. (Weitere Gründe für eine Erwerbsunterbrechung wie beispielsweise Arbeitslosigkeit oder Erwerbsunterbrechungen bei Männern wurden nicht untersucht.) Anhand eines Kranzes von erklärenden Merkmalen, die in identischer oder ähnlicher Form sowohl im Mikrozensus 2012 als auch in der VSE vorhanden sind, sollte nun ein Zusammenhang zwischen der Ausprägung dieses Merkmals des Mikrozensus und den erklärenden Merkmalen gefunden werden.

[8] Zu Hintergründen und weiteren Details zu nachfolgendem Abschnitt siehe Finke et al. (2017).

Tab. 5 Geschätzte Missklassifikationsraten bei der Muttereigenschaft in % (aus Finke et al. (2017))

	Fälschlich nicht als Mutter klassifiziert	Fälschlich als Mutter klassifiziert	Summe	Saldo bezüglich der Zuordnung der Muttereigenschaft
	(1)	(2)	(1) + (2)	(2) – (1)
Frauen im Mikrozensus	14	12	26	–2

Tab. 6 Geschätzte Sensitivitäten und Spezifitäten bei der Muttereigenschaft in % (aus Finke et al. (2017))

Anteil der durch die SVM der Gruppe der Mütter zugeordneten Frauen an den Frauen, die tatsächlich Mutter sind (Sensitivität)	Anteil der durch die SVM nicht der Gruppe der Mütter zugeordneten Frauen an den Frauen, die tatsächlich auch kein Kind geboren haben (Sensitivität)
78	68

Anschließend konnten die in der VSE vorliegenden Daten zu Arbeitnehmerinnen um das so geschätzte Merkmal „Erwerbsunterbrechung ja/nein“ ergänzt werden.

Als Datenmaterial, anhand dessen die Klassifikation gelernt werden sollte, standen prinzipiell alle Datensätze von Frauen im Alter von 15 bis 65 Jahren aus dem Mikrozensus 2012 zur Verfügung, mithin 228.151 Datensätze. Eine große Zahl davon hatte allerdings keine Ausprägungen von für die Untersuchung potentiell wichtigen Merkmalen, sodass der nutzbare Datenbestand auf 152.842 Datensätze reduziert werden musste. Klare Zuordnungen im Hinblick auf das Merkmal „Haben Sie Kinder geboren?“ enthielten 147.599 Datensätze. Mit diesen wurde die Untersuchung schließlich durchgeführt. Es gilt zu beachten, dass eine Überprüfung der Güte der Schätzung nur anhand eines Testdatensatzes aus dem Datenmaterial des Mikrozensus möglich war, nicht jedoch auf dem Datenmaterial der VSE.

Als wichtigste Merkmale zur Schätzung, ob eine Arbeitnehmerin wenigstens ein Kind geboren hat oder nicht, stellten sich das Alter der Frau sowie die Größe des Betriebes, in dem sie zum Zeitpunkt der Befragung für den Mikrozensus 2012 beschäftigt war, heraus. Eine reine Optimierung der Missklassifikationsrate führte dazu, dass im Testdatensatz ca. 20 % der Frauen hinsichtlich der Mutterschaft falsch klassifiziert wurden. Darüber hinaus war eine starke Disparität zwischen den Fehlerarten zu beobachten: Der Fehler, dass eine Mutterschaft geschätzt wurde, obwohl eine solche nicht vorlag, trat ungefähr dreimal so häufig auf wie der entgegengesetzte Fehler (irrtümliches Schätzen, dass keine Mutterschaft vorliegt).

Im bislang in der amtlichen Statistik verwendeten Ansatz zur Erklärung der Verdienstunterschiede zwischen Männern und Frauen werden alle Frauen so behandelt, als läge keine Erwerbsunterbrechung vor. Das fälschliche Unterstellen einer Mutterschaft ist daher der schwerer wiegende Fehler, wenn man einen konservativen Ansatz bei der Verfeinerung des bisherigen Verfahrens verfolgt. Es wurden daher weitere Anpassungen vorgenommen, mit dem Ziel, einen Ausgleich zwischen den beiden beschriebenen Fehlerarten zu schaffen. Diese Anpassungen waren erfolgreich, führten jedoch zu einem Anstieg der Missklassifikationsrate. Es stellten sich die in Tab. 5 und 6 dargestellten Ergebnisse ein.

Tab. 7 Ergebnisse für die Klassifikation der Staatsangehörigkeit in %

	Support Vector Machine	Logistische Regression
Missklassifikationsrate	14	13
Vorhersage Deutscher obwohl Ausländer	4	4
Vorhersage Ausländer obwohl Deutscher	10	9
Anteil der Deutschen, die als Deutsche klassifiziert wurden	89	91
Anteil der Ausländer, die als Ausländer klassifiziert wurden	38	34
Anteil der tatsächlich Deutschen unter den als Deutsche klassifizierten	95	95
Anteil der tatsächlichen Ausländer unter den als Ausländer klassifizierten	21	22
F-Maß der Gruppe der Deutschen	92	93
F-Maß der Gruppe der Ausländer	26	26

Es stellte sich heraus, dass die SVM-Methodik ihre Stärke, nämlich die Erkennung von Mustern (gegebenenfalls unter Inkaufnahme langer Rechenzeiten), nicht ausspielen konnte. Vergleichsrechnungen mit Random Forests lieferten bei deutlich kürzeren Berechnungsdauern vergleichbar gute Ergebnisse hinsichtlich der Missklassifikationsraten. Allerdings wiesen die Random Forests ein deutlich höheres Ungleichgewicht zwischen den auftretenden Fehlern als die Support Vector Machines auf. Sowohl Random Forests als auch Support Vector Machines lagen hinsichtlich der Missklassifikationsraten im Bereich parametrischer Verfahren wie der logistischen Regression. Die Hinzunahme weiterer erklärender Variablen lieferte keine Verbesserung der Ergebnisse.

3.4 Schätzung der Staatsangehörigkeit (deutsch vs. ausländisch)

Ein weiteres Merkmal, das derzeit nicht durch die Verdienststrukturerhebung erfasst wird, ist die Staatsangehörigkeit von Arbeitnehmern. Fragestellungen wie beispielsweise die Verteilung von Arbeitnehmern unterschiedlicher Staatsbürgerschaften auf verschiedene Wirtschaftszweige oder Verdienstklassen können derzeit nicht anhand der Daten der VSE beantwortet werden. Analog zur Schätzung der Mutterschaft von Arbeitnehmerinnen auf Basis des Mikrozensus wurde versucht, auch die verschiedenen Ausprägungen der Staatsangehörigkeit auf die Verdienststrukturerhebung zu übertragen. Dabei ergibt sich eine zusätzliche Schwierigkeit im Vergleich zu Abschn. 3.3: Die Gruppe der Befragten im Mikrozensus mit rein nicht-deutscher Staatsbürgerschaft ist so klein, dass jedes Klassifikationsverfahren diese Gruppe „opfert“, um eine geringe Missklassifikationsrate zu erreichen; mit anderen Worten: Alle Ausländer werden als Deutsche klassifiziert. Es wurden daher verschiedene Herangehensweisen zum Umgang mit sogenannten unausgeglichenen Daten (unbalanced data; auch: imbalanced data) geprüft, nämlich Fehlergewichtung und einfaches Oversampling der nicht-deutschen Arbeitnehmer. Das Ziel bestand dabei darin, zum einen den Anteil der als Ausländer erkannten Arbeitnehmer unter den ausländischen Arbeitnehmern (Spezifität), zum anderen aber auch den Anteil der richtigerweise als Ausländer erkannten Arbeitnehmer unter allen als Ausländern klassifizierten Ar-

beitnehmern (Segreganz) auf ein für diese Fragestellung annehmbares Niveau zu heben. Als Zielkriterium wurde das harmonische Mittel aus diesen beiden Anteilen genutzt (F-Maß, siehe Lewis und Gale 1994; van Rijsbergen 1979)[9] Dabei zeigten sich SVMs gegenüber Random Forests überlegen[10], bewegten sich aber im Rahmen der Ergebnisse der logistischen Regression (Tab. 7).

Offensichtlich sind sich Ausländer und Deutsche im Mikrozensus zu „ähnlich", als dass sie mit den bislang verwendeten Machine-Learning-Verfahren hinreichend gut klassifiziert werden können. Daher sind weitere Testrechnungen mit neuen, jüngst für die Analyse unausgeglichener Daten entwickelten Verfahren (z. B. Gong und Kim 2017) vorgesehen, siehe auch Abschn. 4.

3.5 Anreicherung der Integrierten Erwerbsbiografien (IEB) um Informationen aus der Verdienststrukturerhebung 2014

Aufgabe der Mindestlohnkommission ist es u. a., die Auswirkungen des zum 01.01.2015 eingeführten allgemeinen gesetzlichen Mindestlohns von 8,50 € pro Arbeitsstunde zu evaluieren und der Bundesregierung hierüber zu berichten. Hierzu benötigt die Mindestlohnkommission geeignete Daten. Sie hat daher in ihrem ersten Bericht die Verknüpfung der Verdienststrukturerhebung mit den Integrierten Erwerbsbiografien (IEB) der Bundesagentur für Arbeit gefordert (Mindestlohnkommission 2016, S. 32). Die Paneldaten der IEB sollen auf diese Weise mit ansonsten fehlenden Angaben zum Bruttostundenverdienst bzw. zur Mindestlohnbetroffenheit aus der VSE angereichert werden, um so die Analysemöglichkeiten zu verbessern. Neben einem Record Linkage und einem Statistical Matching kommen hierfür auch statistische maschinelle Lernverfahren, wie Support Vector Machines und Random Forests in Frage (Himmelreicher et al. 2017, S. 19 f.). Das Statistische Bundesamt testet derzeit Random Forests auf Basis der VSE 2014. Ziel ist es, ein Random-Forest-Modell aufzubauen, das die Klassifikation der Beschäftigten in „vom Mindestlohn betroffen" und „vom Mindestlohn nicht betroffen" mit hinreichender Verlässlichkeit ermöglicht. Diese Information könnte dann auf die Daten der IEB „übertragen" werden. Als Eingabevariablen werden Merkmale verwendet, die sowohl in der VSE als auch in den IEB vorliegen.

Die bisherigen Testrechnungen auf der Basis eines relativ einfachen Random-Forest-Modells ergaben für Vollzeitbeschäftigte gute Klassifikationsergebnisse. Die Missklassifikationsrate liegt bei rund 1 %. Bei den Teilzeitbeschäftigten ist der Fehler deutlich höher. Die Missklassifikationsrate beträgt rund 10 %, wobei der Großteil auf diejenigen Arbeitnehmer entfällt, die vom Mindestlohn betroffen sind, für die das Random-Forest-Modell aber das Gegenteil voraussagt. Da die vom Mindestlohn betroffenen nur rund 11 % der Teilzeitbeschäftigten ausmachen, liegt hier (ähnlich wie bei der Staatsbürgerschaft, vgl. Abschn. 3.4) ein Imbalanced-Data-Problem vor. Noch schwieriger ist die Klassifikation von geringfügig Beschäftigten in vom Min-

[9] Ein alternatives Zielkriterium stellt das sogenannte G-Maß dar, das geometrische Mittel aus Spezifität und Sensitivität (Kubat et al. 1997).

[10] In Tab. 7 werden die Mittelwerte für die Berechnungen über zehn verschiedene Aufteilungen des Gesamtmaterials in Trainings- und Testdatensatz angegeben.

destlohn Betroffene und nicht Betroffene. Bei den Vollzeit- und Teilzeitbeschäftigten ist der Bruttomonatsverdienst die mit Anstand einflussreichste erklärende Variable. Bei den geringfügig Beschäftigten ist dies nicht der Fall, da der Bruttomonatsverdienst bei 450 € „gedeckelt" ist und somit weniger Varianz aufweist. Hinsichtlich der verbleibenden erklärenden Variablen sind sich die vom Mindestlohn betroffenen bzw. nicht betroffenen geringfügig Beschäftigten zu „ähnlich", als dass einfache Random-Forest-Modelle sie erfolgreich klassifizieren können. Es sind weitere Untersuchungen notwendig, ob und wie die Klassifikation der Teilzeit- und geringfügig Beschäftigten hinreichend verbessert werden kann. In Frage kommen Maßnahmen, die dem Ungleichgewicht der Klassen bei den Teilzeitbeschäftigten entgegenwirken (z. B. Undersampling oder Oversampling) bzw. die Hinzunahme weiterer erklärender Variablen und der Einsatz eines „hybriden" Ansatzes aus Random Forest und Expertenschätzung bei den geringfügig Beschäftigten.

4 Fazit und Ausblick

Die bisherigen Erfahrungen mit dem Einsatz von Machine-Learning-Verfahren in den Unternehmensstatistiken sind ambivalent. Bei der Abgrenzung des Dritten Sektors kam es im Echtbetrieb mit Daten des Berichtsjahres 2014 zu einer unerwartet hohen Missklassifikationsrate bei der Zuordnung von Unternehmen zum Dritten Sektor, während die Zuordnung zum Nichtdrittsektor zufriedenstellende Ergebnisse lieferte. Die durch nachgelagerte Einzelfallrecherchen aufgedeckte Missklassifikation betraf insbesondere Unternehmen im Bereich der Gastronomie. Derzeit werden mögliche Ursachen geprüft und Weiterentwicklungsmöglichkeiten des SVM-Modells, auch auf der Basis aktuellerer und umfassenderer Trainingsdaten, untersucht. Als positiv festzuhalten ist, dass der vorgelagerte Standard-Algorithmus aufgrund der Erkenntnisse, die mit der Anwendung der SVM gewonnen wurden, verbessert werden konnte. Die Zuordnung von Unternehmen zum Handwerk mittels SVM verlief im Test ermutigend, so dass geplant ist, das Verfahren ab dem Berichtsjahr 2016 in den Aufbereitungsprozess der Handwerkszählung zu integrieren. Die experimentelle Schätzung der Mutterschaft bei weiblichen Beschäftigten, die in der VSE-Stichprobe für 2014 erfasst sind, erbrachte Resultate, die in das Modell zur Berechnung des bereinigten Gender Pay Gap einflossen. Ob diese Vorgehensweise auch für die nächste VSE 2018 umgesetzt werden soll, ist noch offen. Inwieweit mittels SVM belastbare Informationen zur Staatsbürgerschaft von Beschäftigten in der VSE-Stichprobe generiert werden können, bedarf weiterer Untersuchungen. Auch das Vorhaben, die Integrierten Erwerbsbiografien um aus der VSE 2014 mittels Random Forest gewonnene Information zur Betroffenheit vom Mindestlohn anzureichern und so die Analysemöglichkeiten hinsichtlich der Wirkungen der Einführung des Mindestlohnes zu verbessern, erfordert noch weitergehende Analysen.

Gleichwohl sind potentielle weitere Aufgabenstellungen denkbar, bei deren Lösung Machine-Learning-Verfahren eingesetzt werden können. In Frage kommen beispielsweise die Imputation fehlender oder fehlerbehafteter Werte im Rahmen einer verstärkt automatisierten Validierung erhobener Daten. In diesem Kontext ebenfalls relevant ist die Identifikation von Ausreißern (vgl. zu neueren methodischen Ent-

wicklungen Rousseeuw und van den Bossche 2016). Die „outlier detection“ mit Hilfe von Ansätzen des „unsupervised learning“ ist auch für analytische Fragestellungen in den Verdienststatistiken relevant. Darüber hinaus können maschinelle Lernverfahren eine bedeutende Rolle bei der Verwendung neuer digitaler Datenquellen (Big Data) in der amtlichen Statistik spielen.

Auf dem Gebiet des statistischen maschinellen Lernens wird weiterhin intensiv geforscht, sodass den Anwendern laufend verbesserte Methoden zur Verfügung stehen, mit denen die in diesem Aufsatz geschilderten Schwierigkeiten beim Einsatz in den Unternehmensstatistiken womöglich überwunden werden können.

Random Forests, die aus den Klassifikationsbäumen entwickelt wurden, um deren Nachteile zu überwinden, unterliegen selbst dem wissenschaftlichen Fortschritt. Einen Überblick über Weiterentwicklungsschritte in diesem Feld bietet Fawagreh et al. (2014). Interessante Ansätze für den Bereich der amtlichen Statistik bieten die Arbeiten von Bader-El-Den und Gaber (2012), die eine Kombination von genetischen Algorithmen mit Random Forests vorschlagen, sowie von Xu et al. (2012b) zur Klassifikation hochdimensionaler Daten (im Sinne von großen, beispielsweise mehr als 100 Eingabevariablen) mittels Random Forests, wobei hier ein besonderer Wert auf der effizienten Auswahl von für das Modell wichtigen erklärenden Variablen liegt: Weighted sampling im Unterschied zu Breimans Ansatz, die Menge der für einen Split zu betrachtenden Eingabevariablen als einfache Zufallsstichprobe aus allen möglichen Eingabevariablen zu ziehen. Des Weiteren wurden von Xu et al. (2012a) sogenannte Hybrid-Random-Forests erprobt, also solche, bei welchen die zugrundeliegenden Bäume mit zunächst verschiedenen, somit in Konkurrenz um das beste Klassifikationsergebnis stehenden Verfahren (CART, CHAID oder C4.5) gebildet werden. Insofern liegt also eine weitere Variation zu Breimans Random Forests vor.

Die Methodik der Support Vector Machines wird ebenfalls weiterentwickelt. Neben neuen Implementierungen, beispielsweise liquidSVM (Steinwart und Thomann 2017), und der Untersuchung statistischer Eigenschaften in neuen Anwendungsfeldern, beispielsweise im sogenannten Pairwise-Learning (Christmann und Zhou 2016b) oder der Bestimmung von Lernraten von Support Vector Machines (beispielsweise Christmann und Zhou 2016a oder Meister und Steinwart 2016) unter geeigneten Voraussetzungen, wird derzeit auch an in unterschiedlicher Weise lokalisierten Verfahren geforscht. Hable (2013) betrachtet eine Kombination der k-Nearest-Neighbour-Methodik mit Support Vector Machines, Chang et al. (2017) bietet Resultate im Bereich des distributed learning, Meister und Steinwart (2016) sowie Dumpert (2017) untersuchen das Lernen auf in einem vorgelagerten Schritt zu ermittelnden Regionen. Aufgrund der vorhandenen, gegebenenfalls sehr großen Datensätze könnten Ansätze wie das distributed learning oder das Lernen auf Regionen zukünftig auch in der amtlichen Statistik Anwendung finden.

Ein großes Problem, für das alle Klassifikationsmethoden empfindlich sind, sind Grundgesamtheiten und Stichproben, bei welchen die beiden Klassen sehr ungleich häufig auftreten. Auch Random Forests und Support Vector Machines sind hiervon betroffen. Wie in Abschn. 3.4 bereits erläutert, neigen Klassifikationsverfahren in solchen Fällen dazu, die sich in der Minderheit befindende Klasse zu übergehen und alle Beobachtungen der Mehrheitsklasse zuzuordnen. Die Missklassifikations-

rate wird in diesem Fall auf diese Weise zwar minimiert, jedoch kann ein solches Ergebnis nicht Sinn des Einsatzes eines Klassifikationsverfahrens ein. Zum Teil ist es möglich, dass die Verfahren die verschiedenen Fehlzuordnungen mit verschiedenen Gewichten versehen, sodass die Fehlzuordnung einer Beobachtung der Minderheitsklasse im Rahmen der Optimierung stärker bestraft wird als die entgegengesetzte Missklassifikation. Somit ist die direkte Gewichtung der Fehler in der zu optimierenden Funktion zwar ein möglicher Ausweg aus dem Problem unbalanced/imbalanced data, häufig ist dieses Verfahren jedoch zu grob und die Klassifikationsmethode weist ab einem bestimmten Gewichtsverhältnis alle Beobachtungen der Minderheitsklasse (statt wie zuvor alle der Mehrheitsklasse) zu. Andere Ansätze greifen daher bereits bei der Auswahl bzw. der Produktion des Trainingsdatensatzes. Wichtige Beispiele hierfür liefern das Undersampling der Mehrheitsklasse, das Oversampling der Minderheitsklasse oder Kombinationen hieraus. Während beim Undersampling die Gefahr besteht, wichtige Informationen aus dem Trainingsdatensatz zu eliminieren und diese somit im Folgenden nicht mehr zum Lernen heranziehen zu können, erhöht das Oversampling den Rechenaufwand und birgt die Gefahr einer Überanpassung. Eine Übersicht über die Entwicklung der Ansätze, mit unbalanced data umzugehen, bieten Lin und Chen (2012) sowie Dubey et al. (2014). Eine Möglichkeit, bisherige Ansätze zu kombinieren stellt das RHSBoost-Verfahren (Gong und Kim 2017) dar, welches die Mehrheitsklasse einem zufälligen Undersampling unterwirft, die Minderheitsklasse einem fortgeschrittenen Oversampling. Die Notwendigkeit, nur in deutlicher Minderheit vertretene Klassenzugehörigkeiten schätzen zu müssen, stellt die amtliche Statistik wie bei der Schätzung der Staatsangehörigkeit (siehe Abschn. 3.4) vor Herausforderungen, zu deren Bewältigung die genannten Ansätze beitragen könnten.

Auch vor dem Hintergrund der dargestellten Weiterentwicklung der Methoden kann zusammenfassend festgestellt werden, dass Machine-Learning-Verfahren in der amtlichen Unternehmensstatistik mit Erfolg eingesetzt werden können und in Zukunft eine zunehmend bedeutsame Rolle spielen werden.

Literatur

Bachofer F, Esch T, Klein D (2009) Ableitung von Versiegelungsgraden basierend auf hochaufgelösten Fernerkundungsdaten mittels Support Vector Machines. In: Strobl J, Blaschke T, Griesebner G (Hrsg) Angewandte Geoinformatik. Wichmann, Heidelberg, S 432–441

Bader-El-Den M, Gaber M (2012) GARF: Towards self-optimised random forests. In: Huang T, Zeng Z, Li C, Leung C-S (Hrsg) ICONIP 2012, Part II. Springer, Berlin, S 506–515

Baldi P, Brunak S (2001) Bioinformatics. MIT Press, Cambridge

Bennett KP, Campbell C (2000) Support vector machines: Hype or hallelujah? SIGKDD Explor Newsl 2:1–13

Boll C, Leppin JS (2015) Die geschlechtsspezifische Lohnlücke in Deutschland: Umfang, Ursachen und Interpretation. Wirtschaftsdienst 95:249–254

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. Fifth Annual ACM Workshop on Computational Learning Theory, S 144–152 (Proceedings)

Boulesteix A-L, Janitza S, Kruppa J, König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 2:493–507

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall/CRC, Boca Raton

Carleo G, Troyer M (2017) Solving the quantum many-body problem with artificial neural networks. Science 355:602–606
Carrasquilla J, Melko RG (2017) Machine learning phases of matter. Nat Phys 13:431–434
Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. 23rd International Conference on Machine Learning, S 161–168 (Proceedings)
Caruana R, Karampatziakis N, Yessenalina A (2008) An empirical evaluation of supervised learning in high dimensions. 25th International Conference on Machine Learning, S 96–103 (Proceedings)
Chang X, Lin S-B, Zhou D-X (2017) Distributed semi-supervised learning with Kernel Ridge regression. J Mach Learn Res 18:1–22
Christmann A, Zhou D-X (2016a) Learning rates for the risk of kernel based quantile regression estimators in additive models. Analysis Appl 14:449–477
Christmann A, Zhou D-X (2016b) On the robustness of regularized pairwise learning methods based on kernels. J Complex 37:1–33
Chu K, Poirier C (2015) Machine learning documentation initiative. Statistics Canada. https://statswiki.unece.org/download/attachments/63931489/Machine-Learning-documentation-initiative-v10.docx. Zugegriffen: 3. Juli 2017
Cortes C, Vapnik VN (1995) Support-vector networks. Mach Learn 20:273–297
Dubey R, Zhou J, Wang Y, Thompson PM, Ye J (2014) Analysis of sampling techniques for imbalanced data. Neuroimage 87:220–241
Dumpert F (2017) Universal consistency and robustness of localized support vector machines. https://arxiv.org/abs/1703.06528. Zugegriffen: 11. Juli 2017
Dumpert F, von Eschwege K, Beck M (2016) Einsatz von Support Vector Machines bei der Sektorzuordnung von Unternehmen. WISTA Wirtschaft Stat 2016(1):87–97
Fawagreh K, Gaber MM, Elyan E (2014) Random forests: From early developments to recent advancements. Syst Sci Control Eng 2:602–609
Fernández-Delgado M, Cernadas E, Barro S (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15:3133–3181
Feuerhake J, Dumpert F (2016) Erkennung nicht relevanter Unternehmen in den Handwerksstatistiken. WISTA Wirtschaft Stat 2016(2):79–94
Finke C (2011) Verdienstunterschiede zwischen Männern und Frauen. Wirtsch Stat 2011(1):36–48
Finke C, Dumpert F, Beck M (2017) Verdienstunterschiede zwischen Männern und Frauen. WISTA Wirtschaft Stat 2017(2):43–62
Gong J, Kim H (2017) RHSBoost: Improving classification performance in imbalance data. Comput Stat Data Analysis 111:1–13
Gründler K, Krieger T (2015) Using support vector machines for measuring democracy. https://www.wiwi.uni-wuerzburg.de/fileadmin/12010400/DP_130.pdf. Zugegriffen: 3. Juli 2017 (Discussion Paper)
Hable R (2013) Universal consistency of localized versions of regularized kernel methods. J Mach Learn Res 14:111–144
Hamel L (2009) Knowledge discovery with support vector machines. John Wiley & Sons, Hoboken
Himmelreicher R, vom Berge P, Fitzenberger B, Günther R, Müller D (2017) Überlegungen zur Verknüpfung von Daten der Integrierten Erwerbsbiographien (IEB) und der Verdienststrukturerhebung (VSE). RatSWD Working Papers, Bd. 262.
Hyafil L, Rivest RL (1976) Constructing optimal binary decision trees is NP-complete. Inf Process Lett 5:15–17
James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning with applications in R. Springer, New York
Jung S, Käuser S (2016) Herausforderungen und Potenziale der Einzeldatenverknüpfung in der Unternehmensstatistik. WISTA Wirtschaft Stat 2016(2):95–106
Kaus W, Leppert P (2017) Außenhandelsaktive Unternehmen in Deutschland: neue Perspektiven durch Micro data Linking. WISTA Wirtschaft Stat 2017(3):22–38
Kleber B, Sturm R, Tümmler T (2010) Ergebnisse zu Unternehmensgruppen aus dem Unternehmensregister. Wirtsch Stat 2010(6):527–536
Kotsiantis SB (2007) Supervised machine learning: A review of classification techniques. Informatica 31:249–268
Kubat M, Holte R, Matwin S (1997) Learning when negative examples abound. In: van Someren M, Widmer G (Hrsg) Machine Learning: ECML-97 1224, S 146–153
LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86:2278–2324

Lewis DD, Gale WA (1994) A sequential algorithm for training text classifiers. In: Croft WB, von Rijsbergen CJ (Hrsg) Proceedings Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Springer, London, S 3–12
Lin W-J, Chen JJ (2012) Class-imbalanced classifiers for high-dimensional data. Brief Bioinformatics 14:13–26
Lorenz R, Opfermann R (2017) Verwaltungsdaten in der Unternehmensstatistik. WISTA Wirtschaft Stat 2017(1):49–66
Löw F, Michel U, Dech S, Conrad C (2013) Impact of feature selection on the accuracy and spatial uncertainty op per-field crop classification using support vector machines. ISPRS J Photogramm Remote Sens 85:102–119
Meister M, Steinwart I (2016) Optimal learning rates for localized SVMs. J Mach Learn Res 17:1–44
Mindestlohnkommission (2016) Erster Bericht zu den Auswirkungen des gesetzlichen Mindestlohns. Bericht der Mindestlohnkommission an die Bundesregierung nach § 9 Abs. 4 Mindestlohngesetz
Murty MN, Raghava R (2016) Support vector machines and perceptrons. Springerbriefs Comput Sci. https://doi.org/10.1007/978-3-319-41063-0
van Renterghem P, Sottas P-E, Saugy M, van Eenoo P (2013) Statistical discrimination of steroid profiles in doping control with support vector machines. Anal Chim Acta 768:41–48
van Rijsbergen CJ (1979) Foundation of evaluation. J Documentation 30:365–373
Rosenski N (2012) Die wirtschaftliche Bedeutung des Dritten Sektors. Wirtsch Stat 2012(3):209–217
Rousseeuw PJ, van den Bossche W (2016) Detecting deviating data cells. https://arxiv.org/abs/1601.07251. Zugegriffen: 24. Juli 2017
Russel S, Norvig P (2012) Künstliche Intelligenz, 3. Aufl. Pearson, München
Samuel AL (1959) Some studies in machine learning using the game of checkers. Ibm J 3:210–229
Schaathun HG (2012) Machine learning in image steganalysis. John Wiley & Sons, Chichester
Schölkopf B, Smola AJ (2002) Learning with Kernels. MIT Press, Cambridge
Simon HA (1983) Why should machines learn? In: Michalski RS, Carbonell JG, Mitchell TM (Hrsg) Machine learning: An artificial intelligence approach. Tioga Press, Palo Alto, S 25–38
Singh KP, Basant N, Gupta S (2011) Support vector machines in water quality management. Anal Chim Acta 703:152–162
Statistisches Bundesamt (2012) Mikrozensus 2012 Qualitätsbericht. https://www.destatis.de/DE/Publikationen/Qualitaetsberichte/Bevoelkerung/Mikrozensus2012.pdf?__blob=publicationFile. Zugegriffen: 3. Juli 2017
Statistisches Bundesamt (2016) Verdienststrukturerhebung Qualitätsbericht. https://www.destatis.de/DE/Publikationen/Qualitaetsberichte/VerdienstеArbeitskosten/VerdienststrukturerhebungVSE_2014.pdf. Zugegriffen: 3. Juli 2017
Steinwart I, Christmann A (2008) Support vector machines. Springer, New York
Steinwart I, Thomann P (2017) liquidSVM: A fast and versatile SVM package. https://arxiv.org/abs/1702.06899. Zugegriffen: 11. Juli 2017
Sturm R, Tümmler T (2006) Das statistische Unternehmensregister – Entwicklungsstand und Perspektiven. Wirtsch Stat 2006(10):1021–1036
Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
Wainberg M, Alipanahi B, Frey BJ (2016) Are random forests truly the best classifiers? J Mach Learn Res 17:1–5
Wang L (2016) Discovering phase transitions with unsupervised learning. Phys Rev B 94:195105-1–195105-5
Wasserman L (2004) All of Statistics. Springer, New York
Wyner AJ, Olson M, Bleich J (2017) Explaining the success of AdaBoost and random forests as interpolating classifiers. J Mach Learn Res 18:1–33
Xu B, Huang JZ, Williams G, Li MJ, Ye Y (2012a) Hybrid random forests: Advantages of mixed trees in classifying text data. In: Tan P-N, Chawla S, Ho CK, Bailey J (Hrsg) PAKDD 2012, Part I. Springer, Berlin, S 147–158
Xu B, Huang JZ, Williams G, Wang Q, Ye Y (2012b) Classifying very high-dimensional data with random forests built from small subspaces. Int J Data Warehous Min 8:44–63
Yu L, Wang S, Lai KK, Zhou L (2008) Bio-inspired credit risk analysis. Springer, Berlin