HAW HAMBURG

BACHELORTHESIS
Thomas Jablonka

Einsatz von Machine Learning zur Erhebung von Umweltdaten

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Computer Science and Engineering
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Thomas Jablonka

Einsatz von Machine Learning zur Erhebung von Umweltdaten

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung
im Studiengang *Bachelor of Science Wirtschaftsinformatik*
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Prof. Dr. Bettina Buth

Eingereicht am: 7. August 2020

Thomas Jablonka

Thema der Arbeit

Einsatz von Machine Learning zur Erhebung von Umweltdaten

Stichworte

Machine Learning, Umweltmonitoring, KDD, Feinstaubdatensatz, Wetterdatensatz, Neuronale Netze, Klassifikation

Kurzzusammenfassung

Machine Learning ist eine Technik, um Wissen aus riesigen Daten zu ziehen. Umweltmonitoring und das damit verbundene Vorhersagen von Verschmutzung, wie etwa die Vorhersage von Feinstaubentwicklungen, ist eine riesige Herausforderung für die Wissenschaft. Studien werden vorgestellt, die sich bereits mit dieser Thematik auseinandergesetzt haben. Es wird beschrieben welche Daten Grundlage der einzelnen Studien waren, wie diese verarbeitet wurden und wie Machine Learning integriert werden kann.

Thomas Jablonka

Title of Thesis

Use of Machine Learning for collection of environmental datasets

Keywords

Machine Learning, Environmental monitoring, KDD, particulate matter dataset, weather dataset, neuronal network, classification

Abstract

Machine learning is a technique for extracting knowledge from huge amount of data. Environmental monitoring and the associated prediction of pollution, such as the prediction of particulate matter is a huge challenge for science. This work shows the possibilities which techniques can be used to accomplish this task. In addition, studies are presented that have already dealt with this topic. It describes what kind of data were the basis of the individual studies and how machine learning can be use to accomplish that goal.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

Der Bereich Machine Learning(ML) findet in unserer Gesellschaft, welche eine tägliche voranschreitende Technisierung erlebt, eine immer größer werdende Bedeutung. Schon heute ist in vielen Bereichen, sei es die Medizin, die Finanzbranche oder in der Forschung ein Arbeiten ohne jegliche Hilfestellung von Machine Learning kaum vorstellbar. Gerade letzterer Teilbereich, die Forschung, insbesondere die Klima- oder Umweltforschung, versucht stets Wege zu finden, neue Systeme mit Hilfe von ML zu entwickeln oder alte bereits bestehende Strukturen durch den Einsatz innovativer Technologien zu optimieren. In [1] beispielsweise werden ML-Methoden beschrieben, welche unter anderem dafür genutzt werden, um Naturkatastrophen wie Erdbeben, Vulkanausbrüche oder Flutkatastrophen soweit möglich vorherzusagen, um somit die daraus entstehen Schäden zu minimieren. Das ist nur eines von vielen Beispielen indem Machine Learning zu Einsatz kommen.

Über die Hälfte der Bevölkerung lebt heutzutage in Städten[2]. Außerdem wird in der selben Studie geschätzt, dass diese Zahl bis zum Jahr 2050 auf 66% ansteigen wird. Diese Urbanisierung birgt viel Potenzial stellt viele Städte gleichsam vor riesigen Herausforderungen. Eine immer weiter ansteigende Bevölkerungsanzahl und stetig wachsende Stadtstrukturen können zum Nachteil für die Umwelt werden. Auf Grundlage vieler Studien findet das Thema Umweltschutz in unserer heutigen Gesellschaft eine immer bedeutsamere Beachtung. Gerade die Luftverschmutzung wird von vielen Seiten kritisch beleuchtet. Schon längst ist die Forschung über den Punkt hinaus, Feinstaub nur als mögliche Gefahrenquelle für schwerwiegende Erkrankungen zu klassifizieren. Zahlreiche Studien belegen, dass Feinstaub akuten Einfluss auf den gesundheitlichen Zustand eines Menschen hat. Unter anderem wird in der Studie [3] über die Gefahr, die von Feinstaub ausgeht, hingewiesen. Auch ist das Thema Feinstaub und die Senkung der Grenzwerte nicht umsonst Gegenstand von zahlreichen Diskussionen der Europäischen Union. Unter anderem sind

Regelungen zur Senkungen von der Luftbelastung ein elementarer Bestandteil der europäischen Richtlinie 2008/50/EG. Auch das Umweltbundesamt in Deutschland hat sich jahrelang dem Thema angenommen und Maßnahmen zur Senkung der Belastung konzipiert [4]. Wie in der folgenden Abbildung ersichtlich, ist der Feinstaubgehalt in Raum Deutschland über die letzten 20 Jahre gesehen gesunken.

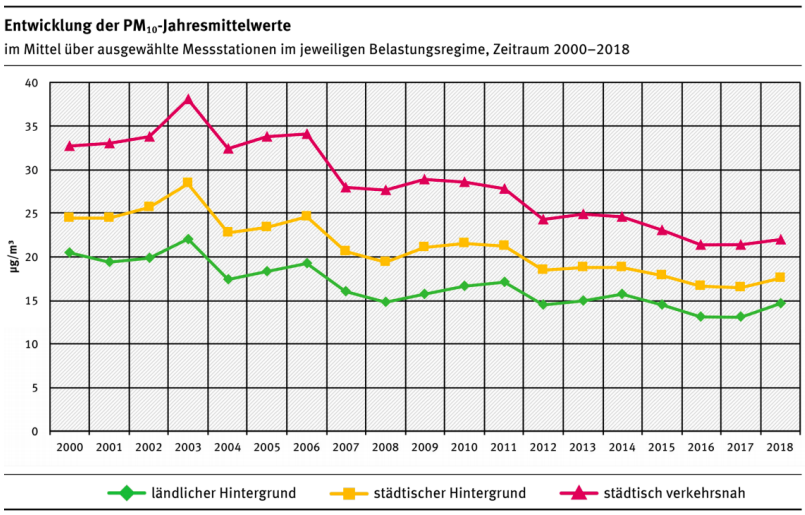

Abbildung 1.1: Die Entwicklung des Feinstaubgehalts(PM10) in Deutschland
Quelle: [5]

Dennoch gibt es weiterhin Gebiete in denen Grenzwerte über einen längeren Zeitraum überschritten werden und somit auch Bedarf an Verbesserungen. Eine komplette Vermeidung von Feinstaub ist in unserer heutigen industrialisierten Gesellschaft kaum möglich. Dennoch können Maßnahmen entwickelt werden, um die Entstehung zu reduzieren. Das Bewusstsein für eine reinere Luft hat auch in der breiten Gesellschaft Anklang gefunden. Gerade in den Ballungsgebieten Deutschlands werden verstärkt Maßnahmen zur Reduzierung ergriffen. So haben sich auch einige Initiativen gebildet um Zuarbeit für den Umweltschutz zu leisten. Eines dieser Initiativen ist das OK Lab Stuttgart. Ziel dieses Projektes ist es, die unsichtbare Gefahr Feinstaub sichtbar zu machen. Mit Hilfe von kostengünstigen und einfach zu implementierenden Messstationen haben sie es geschafft, auch die breite Masse der Gesellschaft für den Umweltschutz zu gewinnen. Dank einer regen Beteiligung an der Kampagne hat sich so ein Netz von weltweit über 9000 Messstionen gebildet. Dieses Netz an Messstationen ist kartografisch auf der Webseite des Projektes einsehbar[6]. Auch die Stadt Hamburg gilt mit seinen über 1,8 Millionen Einwohnern und besonders mit dem angebundenem Hafen als problematische Zone. Auch

hier ist die Stadt aktiv geworden und hat mit eigens implementierten Messstation neue Feinstaubdaten gewinnen können. Mithilfe von Machine Learning Methoden könnten gewisse Analysen mit den gewonnen Daten erstellt werden. Somit können Fragestellungen beantwortet und weitere Maßnahmen konstruiert werden. Hierbei könnte auf folgende Fragestellungen eingegangen werden. Wie können Methoden aus dem Bereich Machine Learning zur Bekämpfung von Feinstaub sinnvoll eingesetzt werden? In Kombination mit Wetterdaten könnten folgende unterschiedliche Fragestellungen erörtert werden. Welchen Einfluss hat die Temperatur auf den Feinstaubgehalt? Können unterschiedliche Witterungsverhältnisse den Feinstaubgehalt beeinflussen? Welchen Einfluss hat beispielsweise der aufkommende Wind auf den Feinstaubgehalt in der Luft? Im Zuge dessen könnten Maßnahmen ergriffen werden um den Feinstaubgehalt in Gebieten, wo hohe Werte auftreten, zu minimieren.

1.2 Ziele

Das Ziel dieser Arbeit ist es, die Eignung von Machine Learning zur Erhebung und Verarbeitung von Umweltdaten zu prüfen. Dies soll mit Hilfe einer theoretischen Anwendung des KDD-Prozesses auf zwei international bereits durchgeführte Studien realisiert werden. Außerdem werden Datensätze vorgestellt, die als Grundlage für ein Vorhersagemodell zur Entwicklung der Feinstaub-Belastung dienen könnten. Diese Datensätze sind zum einen die aus der in der Motivation erwähnten Feinstaubdaten der Initiative OK Lab Stuttgart und Wetterdaten des Deutschen Wetterdienstes. Genauer wird in Kapitel 3 auf die Daten eingegangen.

1.3 Aufbau der Arbeit

Diese Arbeit ist in vier logischen Kapiteln unterteilt. Zunächst erfolgt im ersten Kapitel eine vorbereitende Analyse, in der die Aufgabenstellung definiert wird. Außerdem werden die einzelnen Phasen des KDD-Prozessen sowie Methoden aus dem Bereich Machine Learning vorgestellt. Hier werden auch Studien vorgestellt, die Schnittmengen mit der vorliegenden Arbeit aufweisen. Es folgt eine Vorstellung der Datensätze und eine kurze Datenanalyse. In Kapitel 4 werden dann zwei Studien zum Thema ML und Entwicklung von Feinstaub-Belastung analysiert und evaluiert. Diese Arbeit wird anschließend mit einer Zusammenfassung der Ergebnisse, einem Fazit sowie einem Ausblick beendet.

2 Problemanalyse

Mit allen gespeicherten Daten geht die Hoffnung einher, aus ihnen Wissen und relevante Informationen zu gewinnen, um eine gegebene Anwendung zu verbessern. Bei riesigen Datenmengen hat sich vor allem das Machine Learning als nützlich erwiesen. Beim ML kommen semi-automatisierte Prozesse zum Einsatz, welche den menschlichen Aufwand stark reduzieren. Folgend wird zunächst der Bereich Umweltmonitoring etwas näher erläutert. Nachfolgend werden Studien, die sich ebenfalls mit der Thematik Umweltdaten auseinandergesetzt haben, vorgestellt. Beim Umweltmonitoring fallen meist riesige Datenmengen an. Daher wird der Bereich Big Data kurz umrissen. Anschließend werden die theoretischen Grundlagen des KDD-Prozesses und vom Machine Learning präsentiert.

2.1 Umweltmonitoring

Umweltmonitoring wird als das Beobachten und das Analysieren der Umwelt definiert[7]. Es basiert auf wissenschaftlichen Beobachtungen von Veränderungen in unserer Umwelt. Die Aufgabe der Wissenschaft besteht darin, die aufkommenden Veränderungen in unserer Umwelt zu beobachten, um die Dynamik des Naturkreislaufes analysieren und bewerten zu können. Gerade im Kampf gegen die immer steigende Umweltverschmutzung ist das Umweltmonitoring unabdingbar. Zum einen können die Auswirkungen von Umweltverschmutzung einem langsamen Prozess unterlegen sein und es bedarf daher mehrere Beobachtungen über einen gewissen Zeitraum. Zum anderen kann dieser Prozess langsam geschehen und es sind simultane Beobachtungen notwendig[8]. Beim Umweltmonitoring geht stets die Hoffnung einher, mit den gesammelten Daten wichtige Erkenntnisse gewinnen zu können. In der folgenden Abbildung ist die Idee beim Umweltmonitoring grob visualisiert.

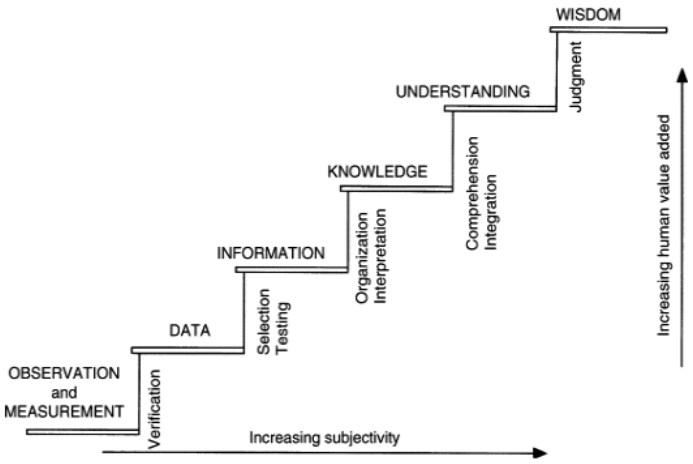

Abbildung 2.1: Wissengenerierung beim Umweltmonitoring
Quelle: [9]

Objektive Beobachtungen und Messungen produzieren Daten. Durch verschieden Auswahlprozesse werden Informationen aus den Beobachtungen und Messungen generiert. Gewonnene Informationen führen im besten Fall zu Wissen und dieses führt zu einem besseren Verständnis von der Problematik/Situation. Diese Kette steigert die Chance, konkrete Entscheidungen für bestehende Probleme/Situationen treffen zu können[9].

Beim Umweltmonitoring wird zwischen drei Haupttypen des Monitoring unterschieden. Folgend wird zwischen dem Wassermonitoring, dem Bodenmonitoring und dem Luftmonitoring unterschieden.

Wassermonitoring ist die stetige Analyse und Untersuchung von unterschiedlichen Wasserquellen(Seen, Flüsse, Mündungen oder Ozeane) und dem damit verbundenem Zustand des Wasserkörpers. Wasserqualitätsmonitoring kann die physikalischen, chemischen und biologischen Charakteristiken vom Wasserkörpern in Relation zu der Gesundheit des Menschen, dem ökolischen Zustand und dem vorgesehenem Wasserverbrauch evaluieren.

Bodenmonitoring werden Acker, Grünland und Forst langfristig beobachtet. Ziel ist es den Zustand der Böden langfristig festzustellen, die entstehenden Veränderungen zu überwachen und eventuelle Entwicklungstendenzen abzubilden[10].

Luftmonitoring ist ebenfalls ein Teilbereich des Umweltmonitoring. Hierbei wird mithilfe von unterschiedlichen Methodiken der Zustand der Luftqualität langfristig beobachtet

und auf mögliche Schadstoffe kontrolliert. Das Ziel hierbei ist saubere Luft für Mensch und Umwelt zu garantieren oder wenn nötig die Qualität durch eventuelle Maßnahmen-zu verbessern. Die vorliegende Arbeit wird sich hauptsächlich mit dem letzten Bereich auseinandersetzen.

2.2 Big Data

Wilder-James definiert in [11] Big Data als Daten, die die Prozesskapazität von konventionellen Datenbanksystemen aufgrund der Menge, Schnelllebigkeit oder inkompatiblen Struktur überschreiten. In [12] schreibt King Big Data folgende Charakteristiken zu:

1. Umfang("Volume"): eine riesige Datenmenge, die zunächst aufgenommen, anschließend analysiert und gemanagt werden muss. Hierbei steigt der Datenumfang mit Anzahl der Quellen und der höheren Auflösung.

2. Varietät ("Variety"): Die akquirierten Daten stammen meist aus neuen Quellen innerhalb und außerhalb einer Organisation. Dabei können unbekannte und unterschiedliche Datenstrukturen auftreten.

3. Schnelllebigkeit ("Velocity"): die Schnelllebigkeit referenziert auf die Geschwindigkeit, mit der die Daten produziert und verändert werden müssen. Daraus ergibt sich, dass eine schnelle Analyse sowie eine rasche Entscheidungsfindung erforderlich sind.

4. Richtigkeit ("Veracity"): die Qualität sowie die Quelle der akquirierten Daten. Hierbei wird die Qualität unter anderem von Inkonsistenz, Unvollständigkeit und Mehrdeutigkeit beeinflusst. Um eine datenbasierte Entscheidung Fällen zu können, verlangt es Nachvollziehbarkeit und Begründbarkeit.

2.3 Angrenzende Studien

Weltweit gibt es unterschiedliche Ansätze um Luftmonitoring zu realisieren. In der Vergangenheit wurden bereits unterschiedliche Studien und Projekte zu diesem Thema veröffentlicht. In [13], [14] wurden Wireless Sensor Network (WSN) basierte Lösungen realisiert um an Umweltdaten zu gelangen. Hierbei fungierten mit Sensoren ausgestattete Microcontroller als Messstationen, welche dann an ausgewählten Orten platziert wurden

und somit ein statisches Sensorenetzwerk entstehen lassen hat. Fix installierte Messstationen bringen aber den Nachteil mit sich, dass bei einer Änderung der Struktur einer Stadt einzelne Messstation unbrauchbar werden könnten. In [15], [16] wurde ein anderer Ansätz verfolgt. Beim sogenannten Vehicular Sensor Network (VSN) wird die Messstation an sich bewegenden Objekte, wie beispielsweise Autos oder Fahrräder befestigt. Dieser Ansatz gewährleistet somit ein flexibles Monitoring. Eine weitere Möglichkeit um an Daten zu kommen, ist das sogenannte Crowd Sensing. Hierbei wird versucht möglichst viele Individuen zum Sammeln von Daten zu gewinnen. In [17] wird ein solcher Ansatz beschrieben. Die in der Motivation erwähnte Initiative OK Lab Stuttgart zielt ebenfalls auf ein Crowd Sensing ab. Eine ähnliche Methode ist das Participating Sensing, welches in dieser Studie zum Einsatz gekommen ist[18].

Während die eben beschriebenden Methoden eher auf die Struktur von Messstationen eingehen und das Sammeln von Umweltdaten das Ziel ist, wird folgend auf die Anwendung von unterschiedlichen Verfahren aus dem Bereich Machine Learning eingegangen. Durch den Einsatz von Methoden aus dem Bereich ML können Analysen auf Daten angewandt werden und im besten Fall hilfreiche Erkenntnisse aus diesen gewonnen werden. Beispielsweise wurde in [19] versucht mit Hilfe von Erkenntnissen die am Vortag gezogen wurden, die maximale Konzentration an PM10, die in der Atmosphere am nächsten Tag anfallen wird, vorherzusagen. Als Parameter dienten hier meteorologische Informationen und Daten zu Schadstoffausstössen. In dieser Studie wurde geprüft welchen Einfluss die Tageszeit, der Tag der Woche, der Monat im Jahr, die Temperatur, die Windrichtung sowie die Windstärke auf die PM10 Konzentration am nächsten Tag haben könnte. Um gewisse Muster aus den benutzen Daten erkennen zu können, wurde in dieser Studie ein artificial neural network(ANN) aufgesetzt und eine Analyse der Daten gefahren. Auch in [20] wurden künstliche neuronale Netzwerke genutzt, um Vorhersagen über die Ozon-, Schwefeldioxid- und PM10-Konzentration in Yinchuan treffen zu können. In [21] habe die Autoren ein neuronales Netz genutzt, um raum-zeitliche Muster aus den Daten erkennnen zu können. Ziel der Studie war es den bodennahen Ozongehalt vorhersagen zu können. Die einzelnen Resultate der einzelnen Studien haben gezeigt, dass neuronale Netze vor allem für kurzzeitige Vorhersagen gut geeignet sind. Ein anderer Ansatz findet sich in [22]. Es wurde zwar ähnliche Datensätze wie in den vorher erwähnten Studien verwendet, allerdings wurde zur Clusteranalyse auf den k-means Algorithmus zurückgegriffen. [23] ist eine weitere Studie, welche ebenfalls Machine Learning genutzt hat, um Umweltmonitoring zu betreiben. Die VerfasserInnen dieser Studie haben hierbei auf das Modell Support Vector Regression (SVR) zurückgegriffen. Die Daten in dieser Studie

stammen aus fest installierten Messstationen der australischen Umweltbehörde. Das Ziel dieser Studie war es zu Versuchen, den Kohlenmonoxid in den überwachten Bereichen vorhersagen zu können. Auch die Studie [24] widmete sich der Entwicklung eines auf Machine Learning basiertem Modells zur Überwachung der Luftqualität. Hierbei lag der Fokus allerdings eher auf die Evaluierung von ML-Algorithmen die zum Umweltmonitoring genutzt werden können. Fünf unterschiedliche Algorithmen aus dem Bereich ML wurden auf die Daten angewandt und anschließend evaluiert. Die in dieser Studie getesteten Algorithmen waren k-nearest neighbors (KNN), support vector machine (SVM), random forest(RF), neural network(NN) und Naive-Bayesian(NB). Unabhängig davon, welcher Methodiken aus dem Bereich des Machine Learnings am Ende gewählt wurde, die Mehrheit der Studien zeigt, dass sich der Einsatz von ML im Bereich Umweltmonitoring einer großer Beliebtheit erfreut.

2.4 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) ist ein iteratives und teils automatisiertes Verfahren zur Gewinnung von Wissen aus grossen Datenbeständen. Durch die immer wachsende Datenmenge, gewinnt dieser Prozess gerade in der heutigen Zeit immer mehr an Wichtigkeit, um zuverlässige Informationen zu generieren. Unter anderem wird in [25] Knowledge Discovery in Databases (KDD) als nichttrivialer Prozess definiert um gültige, neue, potentiell nützliche und verständliche Muster in riesigen Datenbeständen zu identifizieren. Mit Hilfe der nachfolgende Abbildung soll ein Überblick sowie ein Verständnis über den KDD-Prozess gewonnen werden.

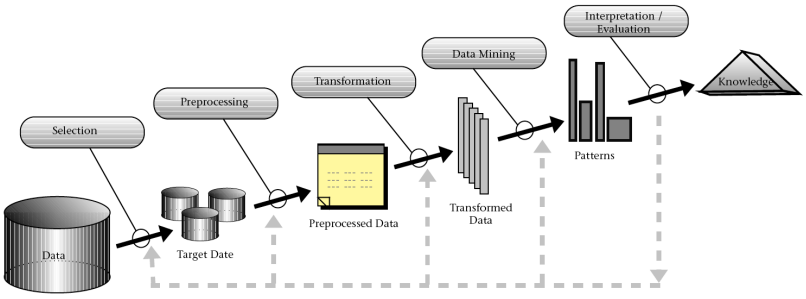

Abbildung 2.2: KDD-Prozess nach Fayyad
Quelle: [25]

Zunächst ist es von Relevanz sich ein sogenanntes Domänenverständnis anzueignen sowie die Ziele der Analyse zu definieren. Anschließend folgt die Entwicklung einer Analysebasis durch die Auswahl von Datensätzen sowie deren Bereinigung und Vorbereitung. Der nächste Schritt ist die Transformation der Daten. Hierbei werden die ausgewählten Datensätze transformiert, um eine geeignete Repräsentation zur Analyse zu gewährleisten. Der darauf folgende Schritt, ist die Anwendung verschiedener Data Mining-Methoden auf die vorverarbeiteten Daten. Im letzten Schritt erfolgt eine Interpretation und Evaluation der Ergebnisse. Folgend werden die einzelnen Schritte näher erläutert.

2.4.1 Datenselektion

Nach der Definition der Ziele und dem vorhanden Domänenwissen, werden in diesem Schritt die für die Analyse erforderlichen Daten selektiert. Die Experten bewerten hierbei die Datenverwendbarkeit. Diese Bewertung ist vor allem für die späteren Prozesse und Ergebnisse von Relevanz. Hierbei müssen unter gewissen Umständen verschiedene Informationsquellen und die darin abgelegten Daten durch geeignete Abfragen selektiert werden, da sich das Spektrum der Daten in der Erforderlichkeit unterscheidet. Anschließend müssen die Daten in für die weiteren Untersuchungen geeigneten Formate abgelegt und gespeichert werden. Diese Auslese an Daten kann auch als Dataset bezeichnet werden.[26] Sollen beispielsweise ausschließlich bestimmte Umweltdaten aus einer bestimmten Region in Deutschland analysiert, ist es sinnvoll, Informationen aus einer anderen Region nicht im weiteren Analyseprozess zu betrachten, also gar nicht erst für die weitere Analyse auszuwählen. In diesem Schritt sollte eine sorgsame Auswahl getroffen werden, da falsche Entscheidungen eine negative Beeinträchtigung des weiteren Prozesses zur Folge haben könnte.

2.4.2 Datenvorverarbeitung

Nach der Datenselektion werden häufig Daten in die Datenbank gespeichert, welche sowohl Fehler als auch Inkonsistenzen aufweisen. Es ist notwendig auftretende Fehler und Inkonsistenzen vor der Durchführung der weiteren Schritte der Analyse zu beseitigen, da dies sonst zu verfälschten Ergebnissen des Wissensentdeckungsprozesses führen könnte. Daher wird gerade in der Phase der Datenvorbereitung und der Datenbereinigung darauf geachtet, durch Einsatz von verschiedenen Methoden sowie durch eine Fehleranalyse solche Risiken zu minimieren. [26]

In [27] beschreibt Runkler die wichtigsten Aufgaben bei der Datenvorverarbeitung. Demnach ist die Erkennung und Behandlung von Fehlern, Ausreißern und Rauscherkennung sowie die Aufbereitung der Daten durch einen standardisierten Prozess als höchste Priorität anzusehen. Weiterhin erwähnt er die Wichtigkeit einer Zusammenfassung aller benötigten Daten. Dies kann beispielsweise mit Hilfe einer einzigen Datenmatrix erfolgen. Runkler unterscheidet hierbei zwischen zufälligen und systematischen Fehlern. Demnach gehören Mess- und Übertragungsfehler zu den zufälligen Fehlern, und können somit als additives rauschen modelliert werden. Eine drastische Verfälschungen eines Datensatzes werden meist durch zufällige Fehler provoziert. Hierbei zählt Runkler insbesondere Messabweichungen (oder auch als Messfehler bezeichnet) und Verarbeitungsfehler, da die Möglichkeit einer Modellierung nicht besteht.

Fayyad/Piatetsky-Shapiro/Smyth beschreiben in [25] ebenfalls, dass das Entfernen von Rauschen sowie der Umgang mit fehlenden Merkmalsauspraägung und Ausreißern als grundlegende Schritte angesehen werden sollten. In der folgenden Abbildung findet sich eine Klassifizierung von Datenqualitätsproblemen.

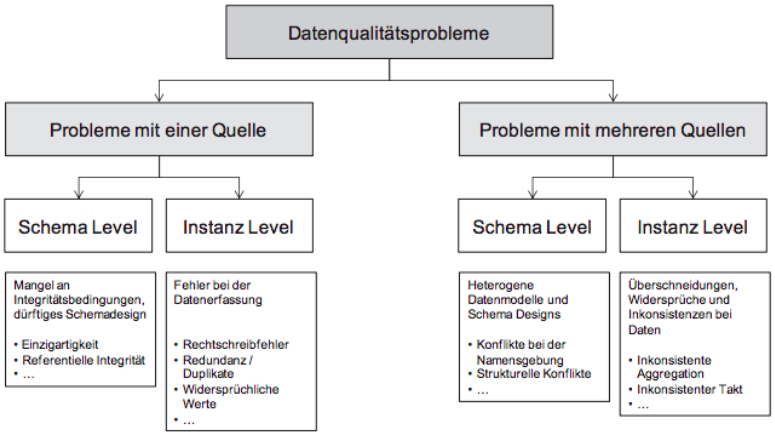

Abbildung 2.3: Klassifizierung von Datenqualitätsproblemen
Quelle: [26]

In [28] sehen Frawley/Piatetsky-Shapiro/Matheus ebenfalls die Ursachen für Datenqualitätsprobleme darin, dass die Datenbanken aus der Realität riesige Datensätze beeinhalten, welche zumeist dynamisch, unvollständig und fehlerhaft sind. Beispielsweise sind dynamische Daten zeitabhängig und die Merkmalsausprägung verändert sich dadurch. Das hat zur Folge, dass die die Ergebnisse der Analyse beeinflusst werden.

Auch die Ausreißererkennung spielt in der Phase der Datenvorverarbeitung eine tragende Rolle. Nach [29] handelt es sich bei Ausreißern um eine Beobachtung, welche eine starke Abweichung von den anderen Beobachtungen vorweist, sodass der Verdacht ensteht, diese sei durch einen anderen Mechanismus generiert worden. Weiter schreiben Koufakou/Georgiopoulos in [30], dass Ausreißer als sogenanntes Rauschen definiert werden und nicht in die weitere Betrachtung mit einfliessen sollten, da es dadurch zu einer Verzerrung der Resultate kommen könnte.

2.4.3 Datentransformation

In [25] beschreiben die Autoren die Phase der Datentransformation als eine Suche nach nützlichen Funktionen zur Visualisierung der Daten in Abhängigkeit des Ziels der Wissensentdeckung. In dem Prozess werden vor allem die Datendimensionen reduziert. Des Weiteren werden sogenannte Transformationsmethoden angewandt, damit die Anzahl der zu betrachteten Variablen ebenfalls reduziert werden kann. Auch geht es bei der Datentransformation um den Umgang mit Null-Werten, die Aggregation von Attributen zu neuen Attributen oder das Sammeln von zeitgestempelten Daten[31]. Entsprechend der Ziele werden bei der Datentransformation Attribute und Features zunächst festgelegt und anschliessend eingegrenzt. Eine Angleichung der Daten ist ebenfalls notwendig. Die Datentransformation ändert sich mit jedem zu realisierendem Projekt, da sich auch die Daten unterscheiden[26].

2.4.4 Einsatz von Data-Mining / Machine Learning

Nachdem bereits die vorherigen Prozesse auf die zu analysierenden Daten angewandt wurden, also eine Selektion, eine Bereinigung sowie eine Transformation stattgefunden haben, beginnt die eigentliche Analyse der Daten mit Hilfe von Methoden aus dem Bereich Data Mining und Machine Learning. Hierbei ist laut [32] das Hauptziel dieser Methoden, Wissen aus den vorliegenden Daten zu generieren. In [27] definiert Runkler Wissen als "[...] interessante Muster, die allgemein gültig sind, nicht trivial, neu, nützlich und verständlich."
Der Erfolg dieses Prozesses hängt stark von den vorherigen Schritten ab. In diesem Prozess kommen Data Mining-Methoden iterativ zur Anwendung, um die Daten zu analysieren. Sharafi unterscheidet in [26] hierbei zwei Arten von Zielen. Zum einen wird das Ziel der Verifikation verfolgt, zum anderen die Entdeckung. Demnach dient die Verifikation

zur Überprüfung der zuvor aufgestellten Hypothesen. Bei der Entdeckung geht es wiederum um das Vorhersagen des Verhaltens zukünfitiger Systeme sowie die Beschreibung und verständliche Darstellung von Muster [25].

2.4.5 Evaluation

Der finale Schritt des KDD-Prozesses ist die Interpretation und Evaluation der Daten. Hierbei werden die Muster, die mit Hilfe der vorherigen Prozesse entdeckt wurden, visualisiert und interpretiert. Das Wissen, welches sich dadurch generiert, sollte demnach direkt genutzt werden. Eine einfache Dokumentation oder das Verteilen an potentiell interessierte Stakeholder ist ebenfalls möglich. Außerdem erfolgt laut Sharafi in diesem Schritt ein Vergleich "[...]der Erkenntnisse mit vorher verfügbarem Wissen oder angenommenen Tatbeständen"[26]. Im Allgemeinen haben sich folgende Metriken zur Evaluation eines Modells durchgesetzt. Die Genauigkeit (precision), die Sensitivität (recall) und das F-Maß gelten als die Kennzahlen zur Bewertung von Modellen.

Die Formel für die Genauigkeit wird wie folgt abgebildet.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Hierbei steht True Positive für positive Klasse die korrekt vorhergesagt wurde und False Positive für die inkorrekte Vorhersage der positiven Klasse.

Die Formel für Sensitivität lautet wie folgt.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

False Negative gilt hier für die inkorrekt vorhergesagte negative Klasse.

Bei der Genauigkeit misst die Proportion der tatsächlich negativ Werte zu denen die auch als solches identifiziert wurden. Die Sensitivität beschreibt die Proportion der tatsächlich positiven Werten zu den die auch als solches identifiziert wurden.

Das f1-Maß entspricht einer Kombination aus der Genauigkeit und der Sensitivität.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

2.5 Machine Learning

In [33] wird Machine Learning als ein Teilbereich der Künstlichen Intelligenz definiert, der systematisch Algorithmen anwendet um eine Beziehung, Gesetzsmäßigkeiten oder auch Muster zwischen zugrunde liegende Daten und Informationen zu erkennen. Ziel einer Anwendung die mit Hilfe von Machine Leraning implementiert wurde, ist das eigenständige lösen von Problemen. ML ist in vielen Bereichen bereits eine feste Größe. So werden unterschiedliche Anwendungen in Bereichen wie der Websuche, Werbeplatzierung, Bewertung von Kreditwürdigkeit, Verhaltensanalyse, Umweltforschung und vielen weiteren Bereichen benutzt. Diese Bachelorarbeit widmet sich vor allem dem Zusammenspiel von Machine Learning und der Erhebung von Umweltdaten. Im Bereich von ML wird zwischen dem Überwachtem Lernen, dem Unüberwachtem Lernen und dem verstärktem Lernen unterschieden. In Abbildung 2.2 werden die Unterschiede der einzelnen Lernansätze angedeutet. Anschließend folgt eine genauere Beschreibung.

Abbildung 2.4: Unterschiede der einzelnen ML-Lernansätze
Quelle: [34]

2.5.1 Überwachtes Lernen

Laut [34] werden beim überwachtem Lernen gekennzeichnete Trainingsdaten gebraucht. Das heißt, dass die Trainingsdaten bereits mit einem Label oder einer Annotation versehen sind. Der Begriff überwacht soll somit vermitteln, dass die Trainingsdaten bereits mit den bekannten Ausgabewerten gekennzeichnet sind. Weiter definieren Cleve und Lämmel in [35], dass beim überwachtem Lernen Beispiele vorgegeben werden, in denen das Resultat gegeben ist. So beschreiben die Autoren ein Szenario bei dem Datensätze für Nadel- und Laubbäume gegeben sind. Diese beeinhalten zum einen Merkmale der Bäume, zum anderen aber auch die Klassifizierung in Nadel- oder Laubbaum. Somit ist bekannt, welcher Baum in welche Kategorie gehört. Folglich kann geprüft werden, ob ein Lernalgorithmus wirklich eine korrekte Klassenzuordnung liefert. Unterkategorien vom überwachtem Lernen sind das Klassifizieren und die Regression. Folgend werden beide Ansätze beschrieben. Die nachfolgende Abbildung stellt den Lernprozess beim überwachtem Lernen dar.

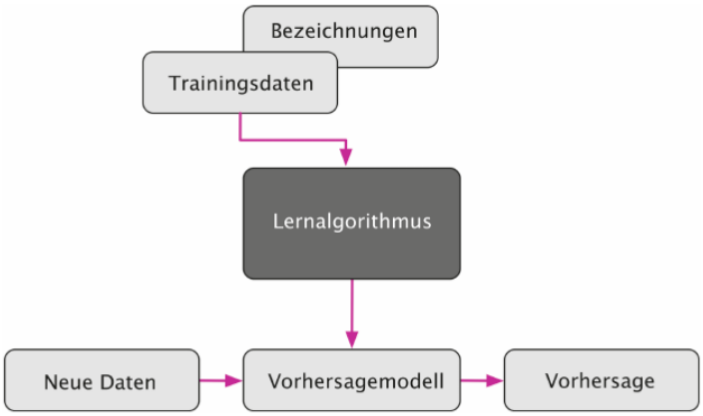

Abbildung 2.5: Lernprozess beim überwachtem Lernen
Quelle: [34]

Künstliche neuronale Netzwerke

In [36] werden künstliche neuronale Netze als ein Teilbereich des überwachten Lernens und derzeit wohl das relevanteste Teilgebiet des maschinellen Lernens beschrieben. Neuronale Netze sind in der Lage, große Mengen an unstrukturierten Daten akkurat auszuwerten und Muster in ihnen zu finden. Der Begriff stammt aus der Biologie. Das menschliche

Hirn besteht aus Neuronen, welches mit anderen Neuronen verbunden ist und durch elektrische Impulse interagiert. Das menschliche Hirn besteht aus Milliarden von Neuronen. Ein künstliches neuronales Netz fungiert ähnlich wie das menschliche Hirn. Das Neuron stellt hierbei eine mathematische Formel dar, die eine Eingabe verarbeitet und eine Ausgabe generiert. Ähnlich wie beim menschlichem Hirn, arbeiten die einzelnen Neuronen zusammen und bilden somit ein künstliches neuronales Netz.

Klassifikation

Bei der Klassifikation wird anders als bei der Regression der Ansatz verfolgt, den jeweiligen Daten Klassen zuzuordnen. Dabei ist die Anzahl der Klassen endlich und als Klassifikation auf unendliche vielen Klassen gesehen werden.

Regression

Bei der Regression wird eine Schätzung des funktionalen Zusammenhangs zwischen zwei Merkmalen bestimmt[27]. Um diesen funktionalen Zusammenhang zu bestimmen, wird hierzu eine Regressionsfunktion verwendet. Hierbei steht die Minimierung von Fehlern zwischen Vorhersagen und dem tatsächlichem Wert. So soll eine möglichst genaue Vorhersage ermöglicht werden.

2.5.2 Unüberwachtes Lernen

In [35] wird das unüberwachte Lernen als Methodik beschrieben, bei der die zu entdeckten Muster gänzlich unbekannt sind. Auch sind dem zu trainierendem Algorithmus weder Gruppierung oder Klassifikationen vorgegeben. Beim unüberwachtem Lernen versucht der Lernalgorithmus gewissen Muster oder Schemata aus den Daten zu erkennen und diese dann zu gruppieren. Dem Lernalgorithmus wird somit kein Ziel vorgegeben, auf welches hintrainiert werden soll. Beim unüberwachtem Lernen kommt meist das Clustering zum Einsatz. Folgend wird dieser Ansatz beschrieben.

Korrelationsanalyse

In [27] beschreibt Runkler die Korrelationsanalyse als Analyse, bei der Zusammenhänge zwischen einzelnen Merkmalen analysiert werden. Demnach wird die Stärke dieses Zusammenhangs als Korrelation bezeichnet. Diese Art der Analyse hilft Zusammenhänge zu erklären und bestimmte Effekte zu erzielen.

Clustering

Bei der Clusteranalyse lassen sich Klassenzugehörigkeiten aus der Struktur der Merkmalsdaten schätzen. Somit ist das Hauptziel beim Clustering Objekte eines nicht klassifizierten Merkmalsdatensatzes einer bestimmten Anzahl von Clustern zuzuordnen.[27].

2.5.3 Verstärktes Lernen

Die dritte Variante beim Machine Learning ist das verstärkte Lernen. In [34] wird das verstärkte Lernen als Methode definiert, bei der die Zielsetzung darin besteht, ein System zu entwickeln (den Agenten), welches die Leistung durch das Zusammenspiel mit seiner Umgebung weiterentwickelt. Herbei wird das System stetig durch ein sogenanntes Belohnungssignal verbessert. Bei einem Belohnungssignal handelt es sich ein Feedback oder um ein Maß, wie gut eine einzelne Aktion durchgeführt wurde. Wie gut oder schlecht eine einzelne Aktion war, wird in einer Belohnungsfunktion definiert. Als Beispiel nennen die Autoren einen Schachcomputer. Hierbei bewertet der Agent nach einer Folge von unterschiedlichen Zügen die Position auf dem Schachbrett (in dem Fall die Umgebung), und die Belohnung kann hierbei am Ende der Partie als Sieg oder Niederlage definiert werden. Die nachfolgende Abbildung stellt den Lernprozess beim verstärtem Lernen dar.

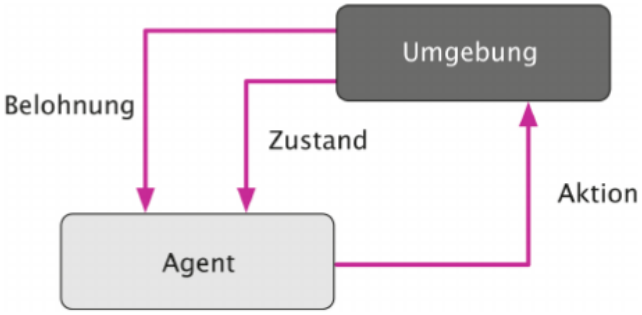

Abbildung 2.6: Lernprozess beim verstärktem Lernen
Quelle: [34]

3 Datenanalyse

Im Folgenden werden die Datensätze beschrieben, die Gegenstand der vorliegenden Arbeit sind. Es wird überprüft ob die bestehenden Datensätze für weitere Analysen mit Methodiken aus dem Bereich Machine Learning nutzbar sind und welche Methoden dafür in Frage kommen. Weiterhin wird in dem Kapitel beschrieben, welche Vorhersagen mit den Datenbeständen möglich sind.

3.1 Vorstellung Datensätze

3.1.1 Feinstaub Datensatz

Das OK Lab Stuttgart beschäftigt sich seit Jahren damit, das Bewusstsein für die akute Luftverschmutzung in Deutschland zu schärfen. Hierbei ist das Hauptziel dieser Initiative, Feinstaub, welcher in deutschen Ballungsgebieten allgegenwärtig ist, sichtbar zu machen und die Entwicklungen im Bereich Transparenz, Open Data und Citizen Science zu fördern. Erreichen will die Initiative dies, indem sie der breiten Masse der Gesellschaft den Zugang zu wissenschaftlichen Projekten ermöglicht. Konkret hat das Team von ehrenamtlichen Entwicklern die Möglichkeit erschaffen, der breiten Masse der Gesellschaft den Zugang zu kostengünstigen Feinstaub-Messstationen zu ermöglichen und somit ein Teil eines wissenschaftlichen Projektes zu werden. Spezifisch geht es in dem Projekt darum, ein Netzwerk aus möglichst vielen Feinstaub-Messstationen zu generieren, um somit an viele Daten zur Luftqualität zu kommen. Mit Hilfe dieser Daten wird dann der Feinstaubgehalt in den Orten, in denen eine Messstation installiert ist, visualisiert. Auf der Internetseite luftdaten.info des OK Labs sind die Messwerte kartographisch ersichtlich. Das Projekt hat sowohl national als auch international Anklang gefunden und es hat sich inzwischen ein Netzwerk von über 9000 Sensorstationen gebildet. Eine Messstation besteht aus einem Microcontroller, einem Sensor zur Messung des Feinstaubgehalts und

optional Sensoren zur Messung der Temperatur sowie der Luftfeuchtigkeit. Die Messstation ist außerdem mit einer Firmware des OK Labs ausgestattet und wird nach der Installation mit dem WLAN verbunden. Die Verbindung wird über die luftdaten.info-API hergestellt. Zugänglich sind die Daten jederzeit auf der Webseite des OK Labs. Sie werden unter anderem als CSV-Datein zum Download angeboten. Hierbei wird eine Datei pro Tag, Station und Sensor erstellt. Die erste Abbildung zeigt einen Feinstaubdatensatz für einen Tag, die zweite Abbildung für einen Monat.

sensor_id	sensor_type	location	lat	lon	timestamp	P1	durP1	ratioP1	P2	durP2	ratioP2
7841	SDS011	3966	51.052	13.819	2018-01-01T	29.67			16.20		
7841	SDS011	3966	51.052	13.819	2018-01-01T	25.83			16.48		
7841	SDS011	3966	51.052	13.819	2018-01-01T	21.42			14.70		
7841	SDS011	3966	51.052	13.819	2018-01-01T	20.52			12.75		
7841	SDS011	3966	51.052	13.819	2018-01-01T	16.52			9.60		
7841	SDS011	3966	51.052	13.819	2018-01-01T	16.35			9.98		
7841	SDS011	3966	51.052	13.819	2018-01-01T	41.28			20.30		
7841	SDS011	3966	51.052	13.819	2018-01-01T	22.58			13.43		
7841	SDS011	3966	51.052	13.819	2018-01-01T	16.90			9.07		
7841	SDS011	3966	51.052	13.819	2018-01-01T	10.20			6.30		
7841	SDS011	3966	51.052	13.819	2018-01-01T	17.65			10.70		
7841	SDS011	3966	51.052	13.819	2018-01-01T	11.73			7.97		
7841	SDS011	3966	51.052	13.819	2018-01-01T	11.10			6.50		
7841	SDS011	3966	51.052	13.819	2018-01-01T	11.70			6.12		
7841	SDS011	3966	51.052	13.819	2018-01-01T	6.93			4.97		
7841	SDS011	3966	51.052	13.819	2018-01-01T	13.27			7.60		
7841	SDS011	3966	51.052	13.819	2018-01-01T	17.52			11.27		
7841	SDS011	3966	51.052	13.819	2018-01-01T	19.15			12.00		
7841	SDS011	3966	51.052	13.819	2018-01-01T	8.85			5.65		
7841	SDS011	3966	51.052	13.819	2018-01-01T	8.35			5.50		

Abbildung 3.1: Feinstaubdatensatz von luftdaten.info eines Tages
Quelle: Webseite luftdaten.info

sensor_id	sensor_type	location	lat	lon	timestamp	P1	durP1	ratioP1	P2	durP2	ratioP2
7841	SDS011	3966	51.052	13.819	2018-01-01T	29.67			16.20		
7841	SDS011	3966	51.052	13.819	2018-01-01T	25.83			16.48		
7841	SDS011	3966	51.052	13.819	2018-01-01T	21.42			14.70		
7841	SDS011	3966	51.052	13.819	2018-01-01T	20.52			12.75		
7841	SDS011	3966	51.052	13.819	2018-01-01T	16.52			9.60		
7841	SDS011	3966	51.052	13.819	2018-01-01T	16.35			9.98		
7841	SDS011	3966	51.052	13.819	2018-01-01T	41.28			20.30		
7841	SDS011	3966	51.052	13.819	2018-01-01T	22.58			13.43		
7841	SDS011	3966	51.052	13.819	2018-01-01T	16.90			9.07		
7841	SDS011	3966	51.052	13.819	2018-01-01T	10.20			6.30		
7841	SDS011	3966	51.052	13.819	2018-01-01T	17.65			10.70		
7841	SDS011	3966	51.052	13.819	2018-01-01T	11.73			7.97		
7841	SDS011	3966	51.052	13.819	2018-01-01T	11.10			6.50		
7841	SDS011	3966	51.052	13.819	2018-01-01T	11.70			6.12		
7841	SDS011	3966	51.052	13.819	2018-01-01T	6.93			4.97		
7841	SDS011	3966	51.052	13.819	2018-01-01T	13.27			7.60		
7841	SDS011	3966	51.052	13.819	2018-01-01T	17.52			11.27		
7841	SDS011	3966	51.052	13.819	2018-01-01T	19.15			12.00		
7841	SDS011	3966	51.052	13.819	2018-01-01T	8.85			5.65		
7841	SDS011	3966	51.052	13.819	2018-01-01T	8.35			5.50		

Abbildung 3.2: Feinstaubdatensatz von luftdaten.info eines Monats
Quelle: Webseite luftdaten.info

Wie in der oben ersichtlichen Abbildung bietet jede Messung unterschiedliche Informationen an. Der Feinstaubgehalt kann in der Spalte P2 abgelesen werden. Die Spalte lat gibt den Breitengrad wieder, die Spalte long den Längengrad. Auch kann der Sensortyp abgelesen werden.

3.1.2 Wetterdaten

Wie unterschiedliche Studien gezeigt haben, kann das Wetter starken Einfluss auf den Feinstaubgehalt in der Luft haben. So haben die Autoren in [37] geprüft, ob es eine Korrelation zwischen dem Feinstaubgehalt in der Luft und bestimmten Wetterlagen gibt. Grundlage waren Feinstaubdaten und Wetterdaten aus zwei unterschiedlichen Städten in Ecuador. Es wurde unter anderem untersucht, welchen Einfluss Niederschlag auf die Konzentration von Feinstaub hat. Wie erwartet ist der Feinstaubgehalt in der Luft umso niedriger, je höher die Niederschlagsmenge ist. Diese Korrelation kann damit erklärt werden, dass die Atmosphäre durch den Niederschlag gereinigt wird. In der Studie [38] sind die Autoren zu ähnlichen Ergebnissen gekommen. Die Autoren sprechen hierbei von einem sogenanntem Wascheffekt. Die folgende Abbildung veranschaulicht dies. Precipitation steht hierbei für die Niederschlagsmenge.

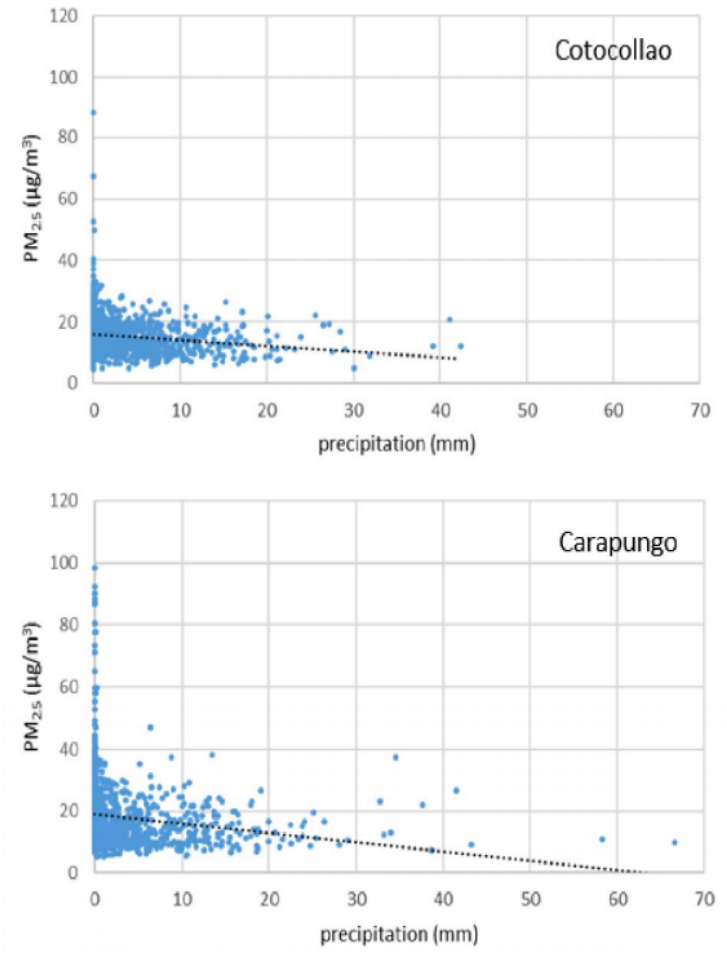

Abbildung 3.3: Korrelation zwischen Niederschlag und Feinstaubkonzentration
Quelle: [37]

Ebenfalls wurde die Korrelation zwischen Wingeschwindigkeit und Feinstaubkonzentration analysiert. In der Allgemeinheit wird vermutet, dass es sich hierbei ähnlich wie beim Niederschlag verhält. Also je höher die Windgeschwindigkeit desto niedriger der Feinstaubgehalt. Allerdings wurden in der Studie unterschiedliche Ergebnisse erzielt. Wie in den folgenden Abbildungen ersichtlich, verhält sich Feinstaub in den untersuchten Städten völlig unterschiedlich. Erklärt wurde dies durch die unterschiedliche Struktur der untersuchten Städte.

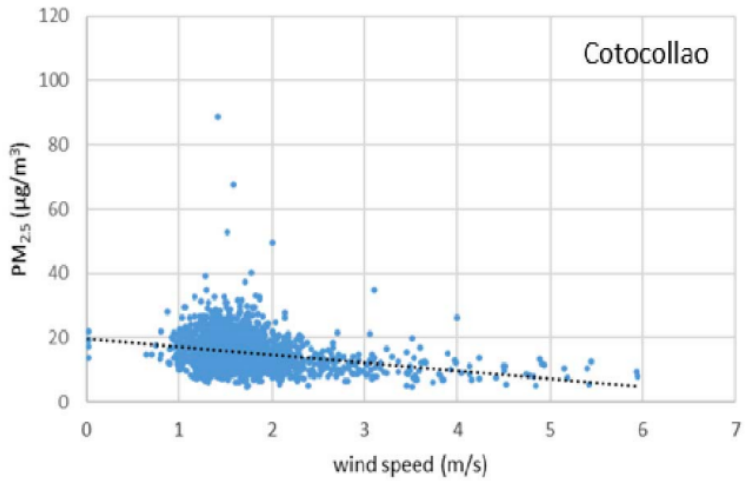

Abbildung 3.4: Korrelation zwischen der Windgeschwindigkeit und Feinstaubkonzentration

Quelle: [37]

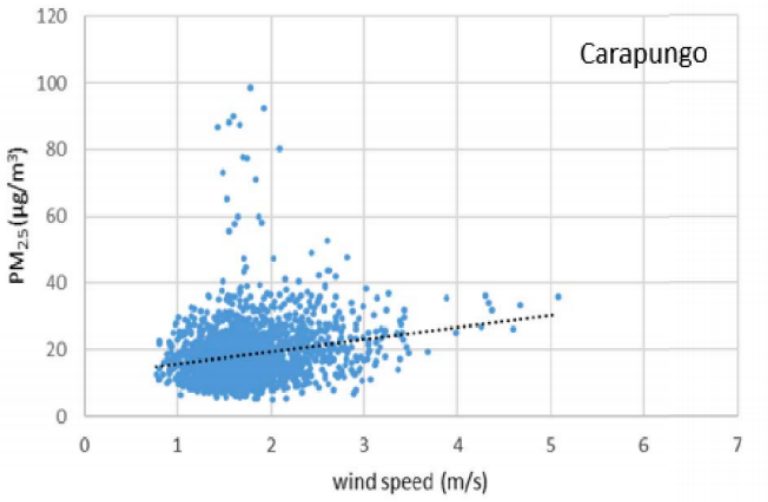

Abbildung 3.5: Korrelation zwischen der Windgeschwindigkeit und Feinstaubkonzentration

Quelle: [37]

Das Bundesumweltamt hat ebenfalls zur Witterungsabhängigkeit bei Schadstoffen geforscht. Demnach führen extreme Wetterlagen, wie etwa winterliche Hochdruckwetterlagen mit geringen Windgeschwindigkeiten dazu, dass Schadstoffe nicht abtransportiert werden können. Ähnlich ist es bei extremer sommerlicher Wetterlage[39]. Aufgrund der oben aufgeführten Ergebnisse unterschiedlicher Studien, sind Wetterdaten auch Gegenstand der vorliegenden Arbeit. Der Deutsche Wetterdienst (DWD) agiert seit dem Jahr 1952 in Deutschland und ist unter anderem für die Erbringung meteorologischer und klimatologischer Dienstleistungen zuständig. Es ist das größte Archiv für Wetterdaten in Deutschland. Das Portal Climate Data Center (CDC) bietet freien Zugriff auf Wetterdaten des DWD der letzten Jahrzehnte[40]. Wie auch bei den Daten zum Feinstaub, wurden hier ausschließlich Daten aus dem Hamburger Bereich berücksichtigt. Die Abbildung zeigt wie der Deutsche Wetterdienst Wetterdaten nach der Erhebung speichert. Sie werden in als txt-Datei auf dem CDC-Portal zu Verfügung gestellt. Zur besseren Darstellung wurde dieser Datensatz in eine CSV-Datei konvertiert.

STATIONS_ID	MESS_DATU	QN_3	FX	FM	QN_4	RSK	RSKF	SDK	SHK_TAG	NM	VPM	PM	TMK	UPM	TXK	TNK	TGK	eor
1975	20181207	3	16.1	6.5	3	8.9	6	0.000	0	8.0	11.8	999.26	10.5	92.25	11.3	6.6	5.5	eor
1975	20181208	3	18.0	8.3	3	19.2	6	0.000	0	7.1	8.4	992.32	7.1	82.67	8.7	5.3	4.0	eor
1975	20181209	3	16.3	6.0	3	10.6	6	0.267	0	7.0	8.3	990.39	6.8	83.83	8.3	5.5	3.8	eor
1975	20181210	3	16.8	6.3	3	2.0	8	1.017	0	5.0	6.7	1008.36	4.8	77.83	5.8	3.7	1.2	eor
1975	20181211	3	12.9	4.6	3	0.0	6	2.133	0	4.4	6.4	1016.50	4.1	78.33	5.6	1.8	-0.6	eor
1975	20181212	3	6.2	2.5	3	0.0	6	0.000	0	6.0	6.6	1024.70	3.2	86.46	4.9	1.3	-0.6	eor
1975	20181213	3	6.2	2.4	3	0.0	7	0.733	0	6.4	6.1	1024.01	1.6	88.88	2.8	-2.2	-4.4	eor
1975	20181214	3	6.2	2.8	3	0.0	8	0.000	0	7.4	5.5	1025.06	1.2	82.50	2.2	0.2	-0.4	eor
1975	20181215	3	11.7	4.2	3	0.0	0	0.000	0	7.2	4.6	1025.65	0.1	74.04	0.8	-1.2	-2.3	eor
1975	20181216	3	12.7	5.4	3	0.0	7	0.000	0	7.3	4.6	1012.99	-0.4	78.46	0.4	-1.7	-2.6	eor
1975	20181217	3	7.2	2.9	3	0.5	6	0.000	0	8.0	6.4	1016.67	1.5	93.50	4.9	-0.6	-0.7	eor
1975	20181218	3	8.7	3.9	3	0.0	0	0.000	0	6.5	7.9	1020.01	5.3	87.79	7.0	2.3	0.8	eor
1975	20181219	3	11.5	4.9	3	0.7	6	0.000	0	7.1	6.8	1011.49	3.4	86.83	5.3	1.8	0.3	eor
1975	20181220	3	9.1	3.4	3	1.8	6	0.000	0	7.3	8.3	1009.78	5.6	90.67	7.0	4.1	2.9	eor
1975	20181221	3	18.9	6.0	3	10.9	6	0.000	0	7.8	9.3	999.08	7.3	90.29	10.6	6.2	4.9	eor
1975	20181222	3	15.4	3.2	3	7.6	6	0.000	0	7.4	8.8	1006.01	6.4	92.00	7.4	5.5	4.2	eor
1975	20181223	3	9.4	3.9	3	5.3	6	0.000	0	7.8	8.9	1012.50	6.7	90.79	7.5	4.9	4.6	eor

Abbildung 3.6: Wetterdaten des Deutschen Wetterdienst für den Bereich Hamburg
Quelle: [40]

Wie aus der Abbildung ersichtlich besteht der Datensatz aus einer Reihe von unterschiedlichen Attributen. Hierbei ist die Temperatur auf unterschiedliche Art abzulesen. Es wird eine Durchschnittstagestemperatur(TMK), ein Tagesmaximum der Lufttemperatur(TXK) und ein Tagesminimum der Lufttemperatur(TNK) in der Tabelle geführt. Des Weiteren kann aus dem Datensatz das Tagesmaximum der Windspitze(FX) und das Tagesmittel der Windgeschwindigkeit(FM) abgelesen werden. Auch die täglichen Sonnenscheindauer(SDK), das Tagesmittel der Relativen Luftfeuchtigkeit(UPM) und das Tagesmittel des Luftdrucks(PM) bietet der Datensatz. Der Anhang zur Arbeit befindet sich auf CD und kann beim Erstgutachter eingesehen werden.

3.1.3 Weitere Datensätze

Zusätzlich zu den Feinstaub- und Wetterdaten wäre das Hinzuziehen von weiteren Datensätze eine Möglichkeit, um ein aussagekräftigeres Modell konzipieren zu können. Beispielsweise könnte das Hinzuziehen von Daten zur Ferienzeit in Erwägung gezogen werden. In der Ferienzeit steigt das Volumen an Reisen deutlich an und somit könnte dies auch unmittelbar den Feinstaubgehalt beeinflussen. In Verbindung dazu wäre das Hinzuziehen von Vekehrsdaten hilfreich. Ähnliche Annahmen könnten aufkommen, wenn es zu Großveranstaltungen kommt. Auch solch ein Datensatz könnte somit hinzugezogen werden. Informationen über die Liegezeiten von Container- und Kreuzfahrtschiffen wäre ebenfalls eine Option, da der Hamburger Hafen zu den am stärksten frequentierten Häfen Europas zählt und riesige Container- und Kreuzfahrtschiffe als Produzenten von enormen Summen von Schadstoff gelten[41]. Auch könnte die Annahme getroffen werden, dass die Heizperiode einen gewissen Einfluss auf den Schadstoffgehalt nehmen könnte. Ob an gewissen Wochentagen womöglich der Feinstaubgehalt höher ist, könnte ebenfalls überprüft werden.

3.2 Mögliche Vorhersagen mit den Datensätzen

Folgend werden einzelne Möglichkeiten genannt, um mit den zugrunde liegenden Daten Vorhersagen tätigen zu können.

3.2.1 Generelle Vorhersage des Feinstaubgehalts

Mit Hilfe der zugrunde liegende Daten könnten Vorhersagen zu dem Feinstaubgehalt in der Luft getroffen werden und ob es eventuell auch zu Überschreitungen von Grenzwerten kommen könnte. Gerade die Attribute wie Sonnenscheinstunden am Tag, Windrichtung, die Windstärke, Nierderschlagswert und die Tagestemperaturen spielen eine maßgebliche Rolle um Vorhersagen zu konzipieren. Hierbei können Modelle konzipiert werden, um langfristige oder kurzfristige Prognosen treffen zu können. So haben die Autoren in [37] unter Berücksichtigung von unterschiedlichen Wettereinflüssen, hier Windgeschwindigkeit, Windrichtung und Niederschlag ein Modell zur Vorhersage des Feinstaubgehalts konzipiert. Das Modell wurde mit Hilfe von Machine Learning Methoden implementiert.

3.2.2 Raum-zeitliche Vorhersage des Feinstaubgehalts

Weiterhin könnte mit den zugrunden liegenden Daten eine raum-zeitliche Vorhersage des Feinstaubgehalts getroffen werden. In [42] haben die Autoren ein Modell zur Vorhersage von Schadstoff in der Luft zu einer bestimmten Zeit und eines bestimmten Ortes erstellt. Auch hier ware Daten von Schadstoff-Messstationen und meteorologische Daten Grundlage zur Realisierung eines Modells. Laut der Studie sollten sowohl räumliche als auch zeitliche Informationen berücksichtigt werden, wenn eine Vorhersage der raum-zeitlichen Verteilung von Schadtstoffen in der Luft realisiert werden soll. Es besteht eine Korrelation zwischen Luftqualitätdaten für ein bestimmtes Gebiet und der Zeit. Gewisse Zustände der Luftqualität in der Vergangenheit können einen Einfluss auf den aktuellen oder zukünftlichen Zustand haben. Beispielsweise kann demnach die Luftqualität der letzten Stunde die Luftqualität der nächsten Stunde beeinflussen. Außerdem werden Schadstoffe womöglich vom Wind transportiert. Hierbei haben die Windrichtung und die Windstärke starken Einfluss. Somit kann die Luftqualität in Nebengebieten Einfluss auf die Luftqualität in lokalen Gebieten nehmen. In der Studie wurde die Quelle von Schadstoffen in der Luft in zwei wesentliche Typen unterteilt. Zum einen gibt es die lokale Quelle von Emissionen und äußere Quellen von Schadstoffen, die in den lokalen Bereich transportiert wird.

3.3 Feature Engineering

In [43] wird ein Feature als eine numerische Darstellung eines Aspekts von Rohdaten definiert. Vielmehr wird ein Feature sowohl beim maschinellem Lernen als auch bei der Mustererkennung als eine individuelle messbare Eigenschaft oder eine Eigenschaft eines beobachteten Phänomens definiert[44]. Als Feature Engineering wird die Aktion definiert, bei der Features von den Rohdaten extrahiert und in Formate umgewandelt werden, die für ML-Modelle verwertbar sind. Vielmehr wird Feature Engineering als Prozess verstanden, welcher neue Features aus den bereits bestehenden Features kreiert. Der Schritt des Feature Engineerings hat in Zusammenhang mit Machine Learning einen großen Einfluss, da die geeignete Auswahl der Features das ML-Modell vereinfachen kann und somit ebenfalls die Qualiät der Resultate gesteigert werden können[43]. Viele Rohdaten eignen sich direkt zum Trainieren eines Modells, allerdings ist es häufig notwendig, zusätzliche (entwickelte) Features für ein verbessertes Trainingsdataset zu erstellen. Nachdem

in Kapitel 3 auf die einzelnen Datensätze eingegangen wurde, kann auf dem gewonnenem Verständnis Features kreiert werden, die dabei helfen, ein zuverlässiges ML-Modell zu implementieren. Das Ergebnis vom Feature Engineering könnte wie in der folgenden Tabelle aussehen.

Tabelle 3.1: Tabelle nach dem Feature Engineering

Zeitliche Features	MeasurementYear
	MeasurementMonth
	MeasurementDay
	MeasurementHour
	MeasurementWeekday
	MeasurementTageszeit
Wetter	maximale Tagestemperatur
	minimale Tagestemperatur
	Tagesmaximum der Windspitze
	Sonnenscheindauer
	Tagesmittel der Relativen Luftfeuchtigkeit
	Tagesmittel der Luftdrucks
	Windrichtung
Raum	MeasurementArea

Quelle: Eigene Darstellung

Zusätzlich zu den zeitlichen Merkmalen, könnte das Feature `Measurement_Wochentag`, also ob es sich um einen Wochentag oder einen Wochenendtag handelt, hinzugefügt werden. Auch das Merkmal `Measurement_Tageszeit` also ob die Messung morgens, mittags oder abends stattgefunden hat, wäre ein sinnvolles Feature. Die räumlichen Merkmale Längen- und Breitengrad könnte zu einem Merkmal zusammengefasst werden. Das Feature könnte als `Measurement_Area` deklariert werden. Beim Feature-Typ Wetter wäre das Hinzuziehen oder das Reduzieren von Features nicht notwendig, da der Datensatz bereits ausreichend Informationen enthält.

4 Theoretische Anwendung des KDD-Prozesses

In der Datenanalyse wurde bereits beschrieben welchen Einfluss gewisse Wetterlagen auf den Feinstaubgehalt nehmen kann. Auf Grundlage dieser Erkenntnisse, die in verschiedenen Studien gezogen wurden, können unterschiedliche ML-Methoden genutzt werden, um eine Vorhersage des Feinstaubgehalts tätigen zu können. Klassifikation und Regression haben sich hierbei besonders hervorgetan und werden als sehr nützliche Werkzeuge in Kombination mit solchen Datensätzen beschrieben[45], [46], [47].

4.1 Einsatz von möglichen Technologien

Neben der Programmiersprache R hat sich Python in den letzten Jahren im Bereich Machine Learning zu einer festen Größe entwickelt[48]. Python bietet zahlreiche Bibliotheken zur Analyse oder Evaluierung von Datensätzen. Gerade für den KDD-Prozess wäre das Hinzuziehen von Python ein sinnvoller Schritt, da diese Programmiersprache eine große Variation von Bibliotheken bereithält, die den Prozess unterstützen könnte. Zur Datenanalyse und Datenmanipulation könnte beispielsweise die Python-Bibliothek pandas berücksichtigt werden. Diese Technologie bietet ebenfalls die Möglichkeit, die Daten in sogenannte Datenframes anzuzeigen. Um die Daten, die aus unterschiedlichen Quellen stammen, in einem Speicherort zu speichern, könnte beispielsweise eine MySQL-Datenbank hinzugezogen werden. Zur Visualiserung der Datensätze bietet Python mit Matplotlib und Seaborn hilfreiche Werkzeuge an. Für die Phasen der Transformation, des Machine-Learnings und der Evaluierung werden in diesem Kontext oft die Technologien Spark MLlib oder Scikit-Learn erwähnt. Beide erwähnten Bibliotheken bieten Hilfestellungen, um beispielsweise Cluster- oder Regressionsanalysen durchzuführen. Die Eingabeparameter die für ein Vorhersagemodell interessant wären hierbei die Wetter-Features. Wie bereits im Kapitel der Datananalyse beschrieben gibt es bereits Studien die eine

Korrelation zwischen spezifischen Wetterlagen und den Gehalt an Feinstaub in der Luft beschreiben. Besonderes Augenmerk wird hierbei auf die die Attribute Durchschnittstagestemperatur, die Windgeschwindigkeit, die tägliche Sonnenscheindauer, die Luftfeuchtigkeit und die Niederschlagswerte. Folgend wird beschrieben, wie Machine Learning auf die Datensätze angewandt werden kann, die Gegenstand dieser Arbeit sind. Speziell wird der Ansatz der Klassifikation durchleuchtet, da sich diese ML-Modell als besonders hilfreich hervorgetan hat.

4.2 Fallstudie A

4.2.1 Datenselektion

Datengrundlage dieser Studie waren Daten von einem Sensoren-Netzwerks eines kommunalen Umweltschutzvereins in Ecuador. Hierbei wurden Daten aus zwei unterschiedlichen Orten (Belisario und Cotocallao) verglichen, die 8 Kilometer auseinander lagen. Die Messstation in Belisario war angrenzend an vielbefahrenen Straßen installiert. Die Messstation in Coticallao befand sich in einer verkehrsberuhigten Lage. In der folgenden Tabelle ist ersichtlich, welche Daten die Messstation aufzeichnete.

Tabelle 4.1: Datengrundlage der Fallstudie A

Value 1	**Value 2**
Feature	Datentyp
Feinstaubkonzentration (PM 2,5)	numerisch
Windstärke	numerisch
Windrichtung	numerisch
Niederschlagsrate	numerisch
Laengengrade	numerisch
Breitengrade	numerisch
Uhrzeit	numerisch
Tageszeit	numerisch
Datum	numerisch

Quelle: [49]

4.2.2 Datenvorverarbeitung

Eine erste Analyse der Daten in der Studie hat ergeben, dass 2,8% der Daten in dem Datensatz des Orts Belisario und 2,4% des Datensatzes für den Ort Cotocollao fehlende Werte aufwiesen. In beiden Datensätze wurden die Daten, die fehlende Werte aufwiesen, gelöscht. Dadurch, dass die Daten mit fehlenden Werten ein nicht sonderlich hohen Prozentsatz des gesamten Datensatzes ausmachten, war es hier vertretbar diese Daten zu löschen. Eine andere mögliche Lösung wäre hier eine Interpolation der Daten. Außerdem wurden die Wochenendtage aus den Daten nicht für die weitere Analyse berücksichtig und ebenfalls gelöscht. Der Feinstaub-Gehalt unter der Woche(Montag-Freitag) und am Wochenende(Samstag und Sonntag) zeigen deutliche Unterschiede. An den Wochentagen waren klare Tendenzen sogenannter Rush-Hours am Morgen und am Abend zu erkennen. Am Wochenende jedoch, speziell am Samstag stieg der Gehalt an Feinstaub zwischen dem spätem Morgen und dem spätem Nachmittag. Zusätzlich konnten niedrigere Werte für Sonntage festgestellt werden. Diese Muster entstanden durch menschliche Aktivitäten und zeigte eine gewisse Konnektivität zwischen dem Feinstaub-Gehalt und dem Straßenverkehr. Die Hinzunahme der Wochenenddaten würde das Modell um weitere Dimension erweitern und dies würde die Komplexität erhöhen.

4.2.3 Datentransformation

Nachdem die Daten selektiert und vorverarbeitet wurden, wurde zudem eine Datentransformation des Features Windrichtung durchgeführt. Die Windrichtung wurde in einer linearen Skalierung angegeben (0°- 360°). Hierbei ist zu erwähnen, dass sowohl 0° als auch 360° die selbe Richtung (Nord) angeben. Es wurde eine mathematische Transformation durchgeführt. Ursprünglich wurde die Windrichtung als Punkt im polaren Koordinatensystem angegeben. Hier fand eine Transformation statt und das Feature Wetter wurde so transferiert, dass dieses Merkmal nun als Punkt im kartesischen Koordinatensystem angegeben werden kann.

$$x = sin\left(\frac{Windrichtung}{360°} \times 2\pi\right) \times Windgeschwindigkeit,$$

$$y = cos\left(\frac{Windrichtung}{360°} \times 2\pi\right) \times Windgeschwindigkeit.$$

Quelle: [49]

Diese Transformation erlaubte eine genauere Repräsentation des Wetter-Features um die Nord-Achse. Ansonsten wären Windrichtungen ein wenig über 0° und ein wenig unter 360° als zwei unterschiedliche Windrichtungen aufgefasst worden.

4.2.4 Machine Learning

Nachdem die Daten für das Machine Learning prepariert wurden, folgte nun das Erstellen von mehreren Modellen mit Hilfe unterschiedlichen ML-Ansätzen. Zunächst wurde der Ansatz Binary Classification (dt. Binäre Klassifikation) verfolgt. Anschließend wurde in dieser Studie eine Three-Class Classification durchgeführt. Zum Schluss wurde eine Regressionsanalyse auf die Daten gefahren.

Binäre Klassifikation

Bei der binären Klassifikation wurden zwei Vorhersageklassen definiert. Zum einen Feinstaubkonzentration $<15\,\mu g/m^3$ und Feinstaubkonzentration $>15\,\mu g/m^3$. Die Klassifikation wurde mit den Algorithmen Boosted Trees(BTs) und Linear Support Vector Machines(L-SVM) durchgeführt. Der ersten Klassifizierung ist zu entnehmen, dass der Datensatz für den Ort Belisario deutlich bessere Ergebnisse aufweist als die Daten für den Ort Cotocollao. Dieses ist aus der folgenden Tabelle zu entnehmen.

Außerdem ist aus beiden Datensätzen zu entnehmen, dass Daten welche eine Feinstaubkonzentration $>15\,\mu g/m^3$ aufwiesen, deutlich besser klassifiziert wurden, als diejenigen die eine Feinstaubkonzentration $<15\,\mu g/m^3$ aufwiesen. Das ist aus den nachfolgenden Tabellen zu entnehmen. Die einzelnen Werte in der Tabelle geben die True-Positive-Rate und die False-Negative-Rate wieder.

Tabelle 4.2: Erster Durchlauf der Binären Klassifikation

Model	Belisario	Cotocollao
BT	83.2%	67.6%
L-SVM	79.8%	66.3%

Quelle: [49]

Tabelle 4.3: Konfusionsmatrix der Binären Klassifikation für Cotocollao

Klasse	$<15\,\mu g/m^3$	$>15\,\mu g/m^3$
$<15\,\mu g/m^3$	51.1%	48.9%
$>15\,\mu g/m^3$	20.3%	79.7%

Quelle: [49]

Tabelle 4.4: Konfusionsmatrix der Binären Klassifikation für Belisario

Klasse	$<15\,\mu g/m^3$	$>15\,\mu g/m^3$
$<15\,\mu g/m^3$	49.0%	51.0%
$>15\,\mu g/m^3$	5.1%	94.9%

Quelle: [49]

Erklärt wurde dies in dieser Studie damit, dass bei der Datenanalyse für den Ort Belisario bereits ein Ungleichgewicht in dieser Klasse festgestellt wurde. Für den Ort Cotocollao hingegen wird dies damit erklärt, dass diese Klasse weniger ausgeprägt war und deshalb das Modell eine Optimierung dieser Klasse vollführt hat.

Die folgende Abbildung zeigt eine ROC-Kurve oder auch Grenzwertoptimierungskurve genannt. Eine ROC-Kurve ist eine Methode um Analysestrategien zu bewerten und zu optimieren. Es wird die Abhängigkeit der Effizienz mit der Fehlquote visuell dargestellt. Nachdem ein Modell zur Klassifikation für jeden Datensatz implementiert wurde, wurde ein Validierungset zur Vorhersage der Klassenlabel hinzugefügt.

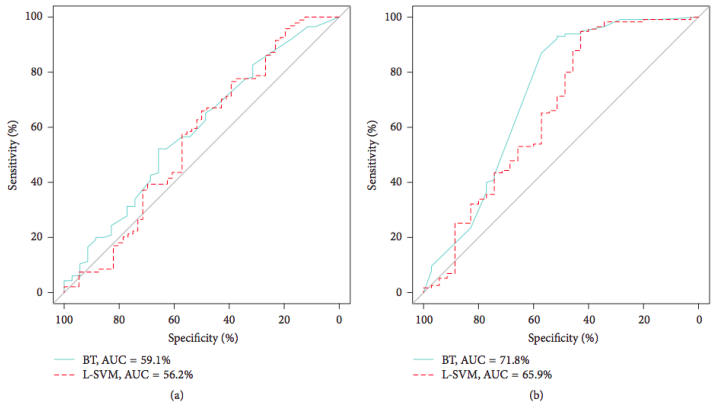

Abbildung 4.1: ROC-Kurve für den Ort Cotocollao
Quelle: [49]

ROC Kurven sind hilfreich um binäre Klassifikation zu evaluieren und um deren Performance zu vergleichen. In der Abbildung beschreibt die Achse specificity die True-Negative-Rate. Die True-Negative-Rate ist die Proportion von negativen Werten die richtig klassifiziert wurden. Die Achse sensitivity beschreibt True-Positive-Rate, also die Proportion von positive Werten die korrekt klassifiziert wurden. Wie in der Abbildung 4.1 zu erkennen, ist die Performance fü des Algorithmus BT bei der Klassifizierung der Daten für den Belisario besser, als mit der Klassifikation mittels L-SVM. Auch für den Ort Cotocollao ist die Performance des BTs bei der Klassifikation besser zu bewerten als für L-SVM. Dennoch muss erwähnt werden, dass bei der Klassifikation des Datensatzes für den Ort Belisario bessere Resultate erzielt wurden als für den Ort Cotocollao.

Three-class Klassifikation

Bei der Three-Class Klassifikation wurden drei Klassen definiert, in denen die Werte klassifiziert werden sollten. Die definierten Klassen sind aus der nachfolgenden Tabelle zu entnehmen.

Für beide Datensätze wurde entdeckt, dass die Klassen $<10\,\mu g/m^3$ und $>25\,\mu g/m^3$ unterrepräsentiert sind. Sie machen jeweils lediglich 10 % des Datensatzes aus. Infolgedessen wurde diese Unausgeglichenheit bei der Anwendung des RusBoosted Tree Algorithmus berücksichtigt. Der Ansatz beim RusBoosted Tree (RBT) war, dass eine gleiche Verteilung der Performance für alle Klassen gefunden wird, anstatt ein globales Optimum. Das führte zu einer besseren Trennbarkeit der einzelnen Klassen.

Tabelle 4.5: Definierte Klassen für die Three-class Klassifikation

Feinstaubkonzentration	**Wertebereich**
low	PM2.5 $<10\,\mu g/m^3$
moderat	$10\,\mu g/m^3$>PM2.5<$10\,\mu g/m^3$
high	PM2.5>$25\,\mu g/m^3$

Quelle: [49]

Die True-Positive-Rate(TPR) oder auch die Sensitivität, gibt den Anteil der tatsächlich Positiven Werte, die korrekt als solche identifiziert wurden, wieder. False-Negative-Rate(FNR), sind Werte die in keine der Klasse fallen und trotzdem in eine der Klasse eingestuft wurden Die folgenden Konfusionsmatrizen zeigen die Ergebnisse für beide Orte der Three-class Klassifikation mit Hilfe des RBT-Algorithmus. Die Zeilen repräsentieren die tatsächlichen Klassen und die Spalten die vorhergesagte Klassen.

Tabelle 4.6: Konfusionsmatrix nach der Three-class Klassifikation für den Ort Cotocollao

Klasse	low	moderate	high	TPR	FNR
low	76.3%	16.3%	7.4%	76.3%	23.7%
moderate	28.3%	28.8%	42.9%	28.8%	71.2%
high	6.3%	20.3%	73.4%	73.4%	26.6%

Quelle: [49]

Tabelle 4.7: Konfusionsmatrix nach der Three-class Klassifikation für den Ort Belisario

Klasse	low	moderate	high	TPR	FNR
low	84.8%	9.5%	5.7%	84.8%	15.2%
moderate	12.3%	53.5%	34.2%	53.5%	46.5%
high	6.5%	45.1%	48.4%	48.4%	51.6%

Quelle: [49]

Wie in beiden Tabellen ersichtlich, zeigt die Klasse $<10\,\mu g/m^3$ akkurate Ergebnisse. Dieses gilt für beide Orte. Dieses gilt ebenfalls für die Klassifikation der Feinstaubkonzentration $>25\,\mu g/m^3$. Allerdings nur für den Ort Cotocollao.

Regressionsanalyse

Im letzten Schritt des Machine Learnings wurde eine Regressionsanalyse basierend auf BT, L-SVM und Neuronalen Netzwerken (NN) durchgeführt. Was bei einer Regressionsanalyse geschieht, wurde in Abschnitte 2.5.1 genauer erläutert. Die allgemeine Annahme in der Studie ist, dass eine Regressionsanalyse bessere Vorhersageergebnisse erzielt, wenn die Analyse über extremen Wetterverhältnissen(starker Wind oder hohe Niederschlagsraten) geschieht. Die Eingabeparameter bei der Regressionsanalyse waren hierbei weiterhin die in Tabelle 4.1 definierten Merkmale. Ausgabeparameter war weiterhin der Feinstaubgehalt. Die Analyse über die Zeit erfolgte in der Regensaison. Die Regression wurde mit drei unterschiedlichen classifiers, basierend auf BT, SVM und NN. durchgeführt. Für jedes Modell wurden die Daten in Test- und Trainingsdaten geteilt. Das Verhältnis hierbei war 80% Trainings- und 20% Testdaten. Die Modelle wurden mit sogenannten Kreuzvalidierungsverfahren trainiert. Mit der Kennzahl mean squared error (dt. Mittlere quadratische Abweichung) wurde die jeweilige Klassenperformance ermittelt. Diese Kennzahl wurde genutzt, um die Diskrepanz zwischen den einzelnen reellen Feinstaubwerten und den einzelnen vorhergesagten Feinstaubwerten zu ermitteln. Die Formel vom mean squared error (MSE) lautet wie folgt:

$$\mathrm{MSE} = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Quelle: [49]

Eine weitere Methodik ist die Bestimmung der Fehler mit der sogenannten mean absolute percentage error (MAPE). Anders als beim MSE gibt MAPE den prozentualen relativen Abstand der wahren Messwert zu der Medellkurve an. Ermittelt wurde diese Kenngröße mit der folgenden Formel.

$$\mathrm{MAPE} = \frac{\sum_{i=1}^{n} \left| (y_i - \hat{y}_i) / y_i \right|}{n} .$$

Quelle: [49]

Folgend sind die Ergebnisse der Analyse mit Hilfe von MSE and MAPE tabellarisch dargestellt.

Tabelle 4.8: Ergebnisse MSE und MAPE

Modell	Belisario	Cotocollao
NN	22.1(26%)	40.7(40%)
L-SVM	26.8(28%)	41.8(41%)
BT	28.5(30%)	44.4(42%)

Quelle: [49]

Wie aus der Tabelle ersichtlich, liefert die Analyse mit Hilfe von NN im Vergleich die niedrigste Fehlerrate und somit die bessere Performance. Dies galt für beide Orte. Die nachfolgende Abbildung zeigt den direkten Vergleich zwischen den vorhergesagten und den tatsächlichen Werten mit Hilfe von NN für den Ort Cotocollao während einer sechs monatigen Periode.

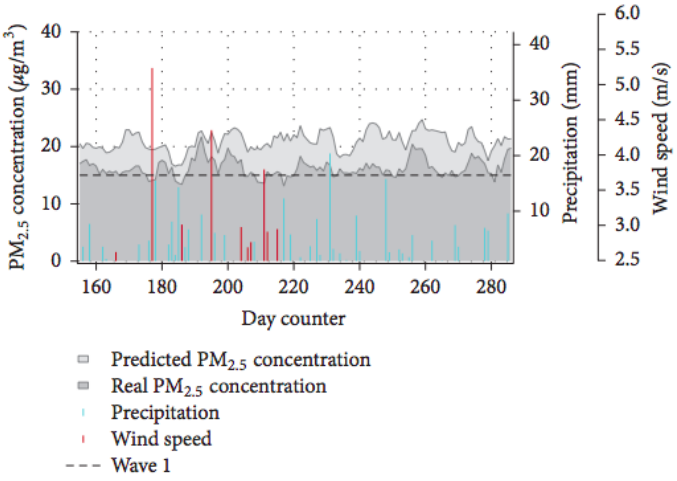

Abbildung 4.2: Vergleich zwischen den vorhergesagten und tatsächlichen Werten
Quelle: [49]

Die vorhergesagten Werte wurden hierbei in hellgrauen Farbe dargestellt, die reellen Werte im dunklen Grau. Wie aus der Abbildung ersichtlich, lieferte das Vorhersagemodell mittels NN gute Ergebnisse. Die Abbildung zeigt ebenfalls, dass bei steigendem PM2.5-Level die Fehlerrate auch zunimmt.

4.2.5 Evaluation

Zusammenfassend für diese untersuchte Studie kann beobachtet werden, dass das zugrunde liegende Modell eine zuverlässige Klassifikation für die Klassen low (<10 µg/m^3) versus high (>25 µg/m^3) und low (<10 µg/m^3) versus moderat(10 µg/m^3>PM2.5<10 µg/m^3) lieferte. In der Studie wurde außerdem erwähnt, dass eine Hinzunahme von weiteren Datenquellen eine sinnvoller Schritt wäre, um das Modell ein Stück weit aussagekräftiger zu machen. Hierbei wurde Daten wie etwa Vehrkehrsdaten als eine weitere mögliche Datenquelle erwähnt.

4.3 Fallstudie B

4.3.1 Datenselektion

Datengrundlage der Fallstudie B [50] waren Umweltdaten von dem Umweltministerium in Hong Kong. Auch hier waren meteorologische Daten und Feinstaub-Daten Gegenstand der Studie. Die Feinstaubdaten stammten hierbei teils aus vielbefahrenen Straßen und teils aus verkehrsberuhigen Orten. Bei einer ersten Datensichtung wurden die Daten in zwei Klassen aufgeteilt. Die nachfolgende Abbildung zeigt an wie vielen Tagen die Feinstaubkonzentration (<40 µg/m^3)(Low) war und an wie vielen Tagen der Feinstaubgehalt (>40 µg/m^3) (High). Der Zeitraum für diese erste Untersuchung waren die Jahre 2000-2011.

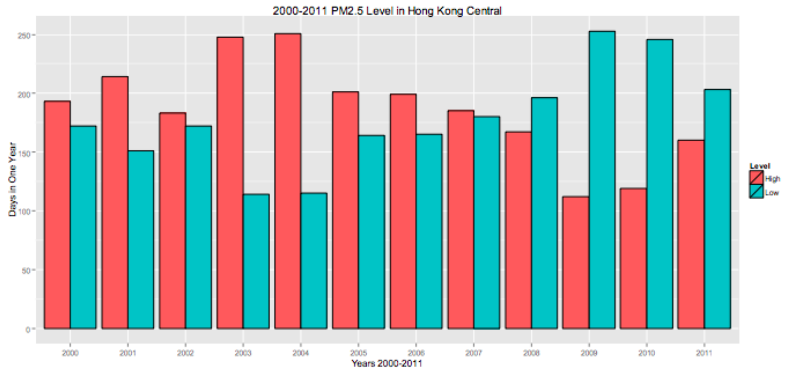

Abbildung 4.3: Feinstaubkonzentration im Zeitraum 2000-2011 in Hong Kong
Quelle: [50]

Wie aus der Abbildung ersichtlich, ist die Klasse "Low", als Werte ($<40\,\mu g/m^3$), stärker vertreten als die Klasse "High". Das Ziel dieser Studie war es, ein Modell basierend auf dem Trend der Daten zu erstellen. Daher wurde ausschließlich mit Daten aus der Zeitspanne 2008-2011 weitergearbeitet. Diese Datensätze wiesen die meisten Daten aus der Klasse 'Low' auf.

4.3.2 Datenvorverearbeitung

Das Merkmal durchschnittliche Temperature im Wetterdatensatz wies fehlende Werte auf. Dieses Merkmal wies keine Einträge auf und wurde deshalb aus dem Datensatz gelöscht. Anschließend wurde die durchschnittliche Temperatur mit der Minimaltemperatur und Maximaltemperatur neu berechnet. Des Weiteren wiesen die Umweltdaten Werte auf die als Ausreißer gekennzeichnet wurden. Hierbei wurden die Ausreißer mit approximierten Werten ersetzt.

4.3.3 Datentransformation

In dieser Studie wurden ebenfalls untersucht, ob saisonale Veränderung und meschliche Aktivitäten einen Einfluss auf den Feinstaubgehalt haben. Daher haben sich die VerfasserInnen der Studie dazu entschlossen, weitere Variablen zu berücksichtigen. Dem Modell wurden zwei Zeitvariablen(Monat, Tag der Woche) hinzugefügt. Die Abbildung 4.4 zeigt den Verlauf des Feinstaubgehalts über die Monate gesehen. Die Abbildung 4.5 zeigt den Feinstaubgehalt in den verschieden Wochentagen.

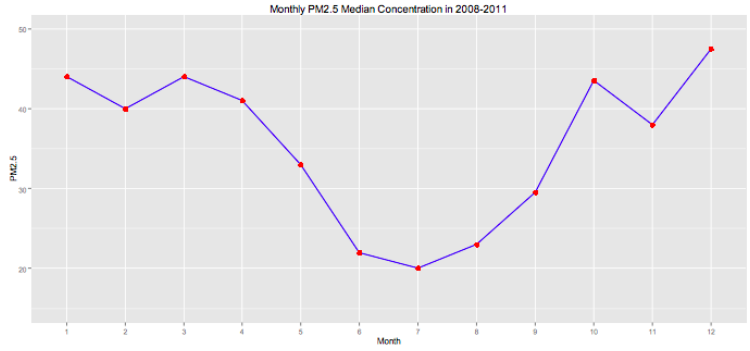

Fig. 4 Monthly $PM_{2.5}$ concentration in 2008-2011.

Abbildung 4.4: Feinstaubkonzentration im Zeitraum 2000-2011 in Hong Kong monatlicher Durchschnitt

Quelle: [50]

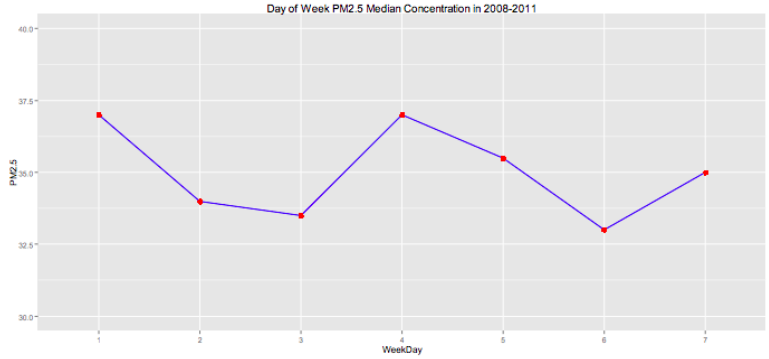

Fig. 5 Daily $PM_{2.5}$ concentration in 2008-2011.

Abbildung 4.5: Feinstaubkonzentration im Zeitraum 2000-2011 in Hong Kong fokussiert auf den Wochentag

Quelle: [50]

Beide Abbildungen verdeutlichen, dass die jeweiligen Merkmale Einfluss auf den Feinstaubgehalt haben. Daher war eine Hinzunahme dieser Variablen sinnvoll. Die finale Liste der Variablen, die für das Modell genutzt wurde, sah wie folgt aus.

Tabelle 4.9: Liste der Variablen der Fallstudie B

Value 1	**Value 2**
Feature	Datentyp
Feinstaubkonzentration (PM 2,5)	numerisch
Durchschnittlicher Luftdruck	numerisch
Maximale Temperatur	numerisch
Minimale Temperatur	numerisch
Durchschnittliche Temperatur	numerisch
Maximale Luftfeuchtigkeit	numerisch
Minimale Luftfeuchtigkeit	numerisch
Durchschnittliche Luftfeuchtigkeit	numerisch
Windrichtung	numerisch
Wingeschwindigkeit	numerisch
Monat	nominal
Wochentag	nominal

Quelle: [50]

4.3.4 Machine Learning

Im Anschluss wurde ein künstliches neuronales Netzwerk aufgesetzt und es wurde mit den Variablen aus der Tabelle 4.9 gearbeitet. Der Ausgabeparameter war hierbei der Feinstaubgehalt, die Eingabeparameter die meteorologischen Daten sowie die zeitlichen Informationen. Eines der wichtigsten Schritte beim Arbeiten mit künstlichen neuronalen Netzwerken ist das Definieren der sogenannten Hidden Layer. In dieser Studie wurde eine Kreuzvalidierung durchgeführt um die Genauigkeit des Trainingsmodells und des Testmodells zu definieren und um die optimalste Anzahl der Knoten zu finden. Wie in der nachfolgenden Tabelle ersichtlich, ist die Performance bei einer Knotengröße von 6 am optimalsten. Das Testmodell zeigte hierbei eine Genauigkeit von 78,3%.

Nodes	Training	Testing
1	0.616	0.616
2	0.616	0.616
3	0.616	0.616
4	0.801	0.759
5	0.829	0.781
6	**0.839**	**0.783**
7	0.844	0.782
8	0.849	0.779
9	0.849	0.774
10	0.852	0.777
11	0.857	0.779
12	0.851	0.772
13	0.861	0.777
14	0.837	0.778
15	0.853	0.777
16	0.829	0.777
17	0.861	0.776
18	0.835	0.779
19	0.845	0.781
20	0.841	0.776
21	0.838	0.779
22	0.835	0.775
23	0.838	0.779
24	0.831	0.780
25	0.831	0.774
26	0.853	0.775
27	0.837	0.781
28	0.839	0.773
29	0.832	0.781
30	0.852	0.776

Abbildung 4.6: Genauigkeit bei unterschiedlicher Anzahl der Hidden Nodes
Quelle: [50]

Des Weiteren kamen Support Vector Machines auch in der Fallstudie B zum Einsatz. Die besten Ergebnisse wurden hierbei erzielt, als die Kosten C bei 10 lagen und γ bei 0.01. Aus der folgenden Tabelle ist das Ergebnis abzulesen.

C	*γ*	Testing
10	**0.01**	**0.816**
10	0.001	0.806
10	0.0001	0.787
100	0.01	0.808
100	0.001	0.811
100	0.0001	0.800
1000	0.01	0.773
1000	0.001	0.814
1000	0.0001	0.804
.....	.....	.....

Abbildung 4.7: Genauigkeit der unterschiedlichen Parametern mittels SVM
Quelle: [50]

4.3.5 Evaluation

In der Fallstudie B wurde zwei unterschiedliche Ansätze erprobt. Es kamen künstliche neuronale Netze und Support Vector Machines zum Einsatz. Es wurde die Stabilität beider Ansätze erprobt. Die nachfolgende Tabelle zeigt, dass die Support Vector Machines bessere Ergebnisse lieferte. Die niedrigste Genauigkeitsrate ist hierbei höher, als die höchste der künstlichen neuronalen Netzwerke. Aus der nachfolgenden Abbildung ist die

Tabelle 4.10: Genauigkeit der evaluierten Algorithmen

Methode	Maximum	Minimum	Median
KNN	79.3%	74.6%	77.6%
SVM	82.0%	80.3%	81.1%

Quelle: [50]

Stabilität der beiden Modelle zu entnehmen.Die Violinen-Grafik zeigt, dass die Ergebnisse der Evaluation für die künstlichen neuronale eine Insatbilität aufweist. Die Ergebnisse der SVM wirken dagegen konstanter.

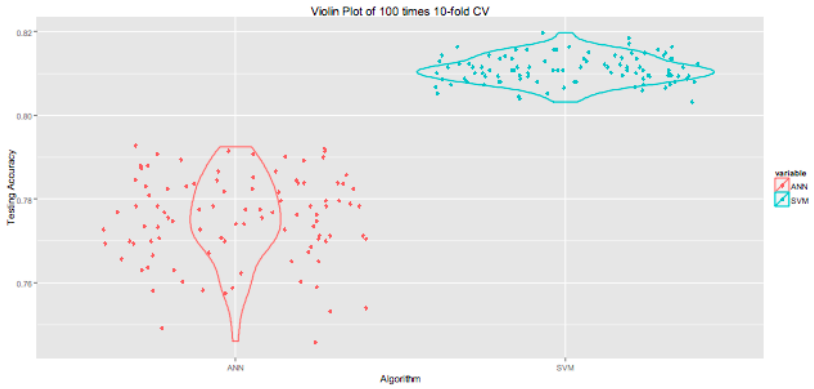

Abbildung 4.8: Violine-Grafik KNN und SVM
Quelle: [50]

4.4 Zusammenfassung

In diesem Kapitel wurde die theoretische Durchführung entlang des KDD-Prozesses anhand von zwei Fallstudien erläutert. Datengrundlage der Fallstudie A waren zum einen Feinstaubdaten und zum anderen meteorologische Daten aus dem Bereich Chile. Fallstudie B arbeitet mit Daten aus dem Bereich Hong Kong. In der Phase Datenvorverarbeitung zeigten sich Mängel in der Datenqualität bei beiden untersuchten Studien. Es haben sich sowohl syntaktische als auch semantische Ausprägungen herausgestellt. Beide Fallstudie zeigten syntaktische Charakteristiken wie etwa wechselnde Datentypen oder verändernde Spaltennamen im Laufe der untersuchten Jahre. Diese Problematik wurde durch eine Vereinheitlichung gelöst. Außerdem wurde in beiden untersuchten Fallstudien festgestellt, dass semantische Ausprägungen vorhanden waren. Die Datensätze wiesen zum einen fehlenden Werte auf. Da die fehlenden Werte keinen besonders hohen Anteil am Gesamtdatensatz hatten, wurden sie in beiden Studien nicht weitere beachtet und gelöscht. Zum anderen wurden falsche Werte identifiziert. Falsche Daten waren hierbei Ausreißer. Teilweise wurden diese Daten mit synthetischen Daten ersetzt, da sie anders als bei den fehlenden Daten, einen größeren Anteil am Gesamtdatensatz darstellten. In der Phase Datentransformation wurde in der Fallstudie A eine Transformation auf das Merkmal Windrichtung angewandt. Diese Transformation erachteten die VerfasserInnen für notwendig, da sonst Windrichtungen die wenig über 0° und ein wenig unter 360° als zwei unterschiedliche Windrichtungen aufgefasst worden wären. Diese Transformation erlaubte eine genauere Repräsentation des Wetter-Features um die Nord-Achse. In der

Fallstudie B wurde keine Datentransformation an einzelnen Merkmaler vorgenommen. Um die Genauigkeit des Modellls zu verbessern, haben die VerfasserInnen es für notwendig gehalten, weitere Merkmale dem Modell hinzuzufügen. So wurden die Zeitvariablen Monat und Tag der Woche hinzugefügt. Nachdem die Daten vorverarbeitet und transformiert wurden, kamen ML-Algorithmus zum Einsatz. Die VerfasserInnen der Fallstudie A haben zunächst eine binäre Klassifikation vorgenommen. Es wurden hierfür zwei Vorhersageklassen ($<15\,\mu g/m^3$ und Feinstaubkonzentration $>15\,\mu g/m^3$) definiert. Die Klasssifikation wurde mittels den Algorithmen Algorithmen Boosted Trees(BTs) und Linear Support Vector Machines(L- SVM) durchgeführt. Eine Klassifikations mittels BTs zeigte eine bessere Performance verglichen mit L-SVM. Bei einer anschließenden Regressionsanalyse basierenden auf den Algorithmen BTs, L-SVM und Neuronalen Netzwerken(NN), erzeugte vor allem die Analyse mittels NN Ergebnisse mit einer niedrigen Fehlerrate. In der Fallstudie B wurden die Ansätze künstliche neuronale Netzwerke(KNN) und SVM verglichen. Im Vergleich schnitt SVM mit einer stabileren Performance ab. KNN zeigte allerdings ebenfalls akkurate Ergebnisse.

5 Fazit und Ausblick

Ziel dieser wissenschaftlichen Arbeit war zunächst Machine Learning auf Tauglichkeit zur Erhebung und Verarbeitung von Umweltdaten zu überprüfen. Hierfür wurde anhand von bereits durchgeführten Studien analysiert, wie ML-Methodiken auf Umweltdatensätze angewandt wurden, um beispielsweise Vorhersagemodelle zur Entwicklung von Feinstaub-Belastung generieren zu können. Es wurden zwei Studien, welche sich mit der Entwicklung von Vorhersagemodellen zur Entwicklung der Feinstaub-Belastung auseinandergesetzt haben, näher beleuchtet. Als Datengrundlage dienten den jeweiligen Studien sowohl Feinstaubdaten als auch meteorologische Daten. Die Studien wurden mit Hilfe einer theoretischen Durchführung entlang des KDD-Prozesses erläutert. Die Ergebnisse der jeweiligen Studien haben gezeigt, dass Machine Learning ein probates Mittel zur Verarbeitung von Umweltdaten darstellt. Um wüschenswerte Ergebnisse erzielen zu können, bedarf es allerdings einer sorgfältigen Aufbereitung der im Vorfeld ausgewählten Datensätze. So haben beide Studien gezeigt, dass in den Rohdaten meist fehlerhafte Daten vorhanden sind, die Auswirkung auf die Ergebnisse des Zielmodells haben können. Zur Behandlung dieser fehlerhaften Daten können unterschiedliche Ansätze verfolgt werden. In den untersuchten Studien wurden fehlerhafte Daten gelöscht, da diese keinen wesentlichen Anteil am Gesamtdatensatz darstellten. In beiden Studien wurden unterschiedliche ML-Algorithmen auf die Datensätze angewandt. Die Evaluation der Fallstudie A hat gezeigt, dass der Einsatz von neuronalen Netzen akkurate Vorhersageergebnisse liefern kann. Verglichen wurden in dieser Studie die Ansätze neuronale Netze, Support Vector Machine und Boosted Trees. Demgegenüber stehen die Ergebnisse der Fallstudie B. In dieser Studie wurde ebenfalls der Einsatz von neuronalen Netzen und Support Vector Machine getestet. Bessere Resultate wurden hierbei, anders als in der Fallstudie A, mit Hilfe von Support Vector Machine erzielt. Dies zeigt, dass nicht einzig die Wahl der ML-Verfahren von Relevanz ist, vielmehr ist die sorgfältige Auswahl der Parameter ebenfalls ein wichtiges Kriterium, um wünschenswerte Ergebnisse zu erzielen.
Die Ergebnisse zeigen außerdem, dass sich relativ gute Prognosen mit Feinstaub- und Wetterdaten bilden lassen. Auf Basis der untersuchten Studien konnte auch gezeigt wer-

den, dass gewisse Wetterlagen den Feinstaubgehalt beeinflussen können. In den jeweiligen Studien wurden hierbei Wettermerkmale wie Windrichtung, Niederschlag oder Temperatur genauer untersucht. Außerdem wurde in dieser wissenschaftlichen Arbeit der Feinstaubdatensatz von luftdaten.info auf die Nutzbarkeit für ML-Verfahren beurteilt. Eine Beurteilung des Datensätzes wurde ausschließlich über eine Sichtung der Daten realisiert. Eine technische Analyse des Datensatzes, beispielsweise mit Hilfe von Python-Bibliotheken, war nicht Gegenstand der vorliegenden Arbeit. In Absatz 4.1 werden mögliche Technologien zur Datenanalyse erwähnt. Der Feinstaubdatensatz zeichnet sich durch eine hohe Variabilität aus. Das steigende Interesse an dem Projekt führt zu einer wachsenden Anzahl an Messstationen und somit zu vielen Datensätzen. Dadurch, dass sich die Nutzer die Frequenz, in der die Messstationen Messungen vornehmen soll, frei wählen können, birgt dieser Datensatz ein paar Probleme der Vergleichbarkeit in sich. Auch beinhaltet der Feinstaubdatensatz Aussetzer der Messungen, da eine Messstation jederzeit deaktiviert werden kann. Generell ist das Netz an Messstationen ist im Hamburger Bereich noch längst nicht so ausgeweitet, wie etwa im Raum Stuttgart oder im Ruhrgebiet. Dennoch bietet die Daten der Messstationen eine solide Grundlage, um etwa ein Vorhersagemodell für Entwicklung der Feinstaub-Belastung wie in den untersuchten Studien zu realisieren.

Durch die Analyse der Datensätze und der Studien wurden wertvolle Erkenntnisse gewonnen und können als Grundlage für weitere Studien dienen. In dieser Arbeit wurde ausschließlich analysiert, wie Machine Learning theoretisch auf Umweltdaten angewendet werden kann. Es dienten zwei Fallstudien als exemplarische Beispiele. Diese Arbeit und die hier vorgestellten Daten könnte somit als Grundlage für eine praktische Realisierung eines Vorhersagemodells zur Entwicklung der Feinstaub-Belastung dienen. In dieser Arbeit wurden ausschließlich Feinstaub- und Wetterdaten vorgestellt und als Grundlage für mögliche Vorhersagemodelle in Betracht gezogen. Zusätzlich zu den eben genannten Daten könnten Informationen über das Verkehrsvolumen, Großveranstaltungen oder auch Daten über den Schiffsverkehr im Hamburger Hafen hinzugezogen werden, um die Dimensionen des Modells zu vergrößern. Gerade das Hinzuziehen von Daten über den Schiffsverkehr wäre für die Analyse des Feinstaubgehalt, da der Hafen in Hamburg zu den am stärksten frequentiertesten Häfen Europas zählt und somit eine Anlaufstelle für eine Vielzahl von riesigen Frachtern gilt. Die vorliegende Arbeit wurde in der Zeit der Corona-Pandemie verfasst. Durch beschlossene Lockdown-Regelungen kam es in vielen Bereichen zu Einschnitten. So war auch der Straßen- und Flugverkehr arg betroffen. Ein ebenfalls interessanter Ansatz wäre zu prüfen, ob und wie sehr diese Einschnitte zu einer Veränderung der Luftqualität geführt haben. Eine weiterer interessanter Ansatz wäre das

Hinzuziehen von Daten über die Baustellensituationen im Hamburger Straßenverkehr. Gerade längeren Stehzeiten der Autos oder durch Baustellen bedingte Umleitungen führen vermutlich zu mehr Emissionen. Das Hinzuziehen von weiteren Datenquellen lässt die Dimension des Vorhersagemodells wachsen. Hierbei entsteht somit das Problem, dass durch die hohe Dimensionalität des Modells Ergebnissen erzielt werden, die vom Computer nicht erklärt werden und somit für den Menschen nicht verständlich sind. In [51] werden Ansätze beschrieben, wie Ergebnisse die mittels ML erzielt wurden, interpretierbarer für den Menschen werden. Die Feinstaubdaten, die Teil der vorliegenden Arbeit sind, entsprangen einer Crownd Sensing Initiative. Der Ansatz Crowd Sensing bietet zwar die Möglichkeit, eine Vielzahl von Daten zu generieren. Jedoch birgt diese Methodik der Datenakquise allerdings auch Probleme. Bei diesem Ansatz fehlt meist eine Kontrollinstanz und daher weisen die Daten meist ein hohes Maß an Ausreißern und Missing Data auf. Die Datenintegrität kann somit schwierig gewährleistet werden.

Literaturverzeichnis

[1] A. L. Pyayt, I. I. Mokhov, B. Lang, V. V. Krzhizhanovskaya, R. J. Meijer, *et al.*, "Machine learning methods for environmental monitoring and flood protection," *World Academy of Science, Engineering and Technology*, vol. 78, pp. 118–123, 2011.

[2] D. o. E. United Nations and P. D. Social Affairs, "World population prospects: The 2015 revision," *key findings and advance tables. New York, USA*, 2015.

[3] J. Lelieveld, K. Klingmüller, A. Pozzer, U. Pöschl, M. Fnais, A. Daiber, and T. Münzel, "Cardiovascular disease burden from ambient air pollution in europe reassessed using novel hazard ratio functions," *European heart journal*, vol. 40, no. 20, pp. 1590–1596, 2019.

[4] P. Builtjes, W. Jörß, R. Stern, and J. Theloke, "Strategien zur verminderung der feinstaubbelastung," *Zusammenfassender Abschlussbericht. Umweltbundesamt. Dessau-Roßlau (Texte, 09/2012). Online verfügbar unter http://www. uba. de/uba-info-medien/4268. html*, 2012.

[5] A. Minkos, U. Dauert, S. Feigenspan, and S. Kessinger, "Luftqualität 2018," 2019.

[6] "OK Lab Stuttgart." https://luftdaten.info/. Accessed: 2020-06-01.

[7] J. F. Artiola, I. L. Pepper, and M. L. Brusseau, "Monitoring and characterization of the environment," in *Environmental monitoring and characterization*, pp. 1–9, Elsevier Inc., 2004.

[8] J. Artiola and M. Brusseau, "The role of environmental monitoring in pollution science," in *Environmental and Pollution Science*, pp. 149–162, Elsevier, 2019.

[9] J. F. Artiola, I. L. Pepper, and M. L. Brusseau, "Monitoring and characterization of the environment," in *Environmental monitoring and characterization*, pp. 1–9, Elsevier Inc., 2004.

[10] "Webseite Bundesumweltamts." https://www.umweltbundesamt.de/themen/boden-landwirtschaft/boden-schuetzen/boden-beobachten-bewerten#boden-dauerbeobachtung-. Accessed: 2010-09-30.

[11] E. Wilder-James, *Planning for Big Data*. Ö'Reilly Media, Inc.", 2012.

[12] S. King, *Big Data - Potential und Barrieren der Nutzung im Unternehmenskontext*. Berlin Heidelberg New York: Springer-Verlag, 2014.

[13] L. G. Rios *et al.*, "Big data infrastructure for analyzing data generated by wireless sensor networks," in *2014 IEEE International Congress on Big Data*, pp. 816–823, IEEE, 2014.

[14] S. M. Saad, A. R. M. Saad, A. M. Y. Kamarudin, A. Zakaria, and A. Y. M. Shakaff, "Indoor air quality monitoring system using wireless sensor network (wsn) with web interface," in *2013 International Conference on Electrical, Electronics and System Engineering (ICEESE)*, pp. 60–64, IEEE, 2013.

[15] Y.-C. Wang and G.-W. Chen, "Efficient data gathering and estimation for metropolitan air quality monitoring by using vehicular sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 8, pp. 7234–7248, 2017.

[16] R. Rushikesh and C. M. R. Sivappagari, "Development of iot based vehicular pollution monitoring system," in *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pp. 779–783, IEEE, 2015.

[17] J. Dutta, F. Gazi, S. Roy, and C. Chowdhury, "Airsense: Opportunistic crowd-sensing based air quality monitoring system for smart city," in *2016 IEEE SENSORS*, pp. 1–3, IEEE, 2016.

[18] A. Zenonos, S. Stein, and N. R. Jennings, "Coordinating measurements for air pollution monitoring in participatory sensing settings," in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pp. 493–501, International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[19] F. Nejadkoorki and S. Baroutian, "Forecasting extreme pm10 concentrations using artificial neural networks," 2012.

[20] F. Li, "Air quality prediction in yinchuan by using neural networks," in *International Conference in Swarm Intelligence*, pp. 548–557, Springer, 2010.

[21] T. M. Chiwewe and J. Ditsela, "Machine learning based estimation of ozone using spatio-temporal data from air quality monitoring stations," in *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)*, pp. 58–63, IEEE, 2016.

[22] K. Nandini and G. Fathima, "Urban air quality analysis and prediction using machine learning," in *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, pp. 98–102, IEEE, 2019.

[23] K. Hu, V. Sivaraman, H. Bhrugubanda, S. Kang, and A. Rahman, "Svr based dense air pollution estimation model using static and wireless sensor network," in *2016 IEEE SENSORS*, pp. 1–3, IEEE, 2016.

[24] T. M. Amado and J. C. D. Cruz, "Development of machine learning-based predictive models for air quality monitoring and characterization," in *TENCON 2018-2018 IEEE Region 10 Conference*, pp. 0668–0672, IEEE, 2018.

[25] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, pp. 37–37, 1996.

[26] A. Sharafi, *Knowledge Discovery in Databases - Eine Analyse des Änderungsmanagements in der Produktentwicklung.* Berlin Heidelberg New York: Springer-Verlag, 2013.

[27] T. A. Runkler, *Data Mining - Methoden und Algorithmen intelligenter Datenanalyse.* Berlin Heidelberg New York: Springer-Verlag, 2010.

[28] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: An overview," *AI magazine*, vol. 13, no. 3, pp. 57–57, 1992.

[29] D. M. Hawkins, *Identification of outliers*, vol. 11. Springer, 1980.

[30] A. Koufakou and M. Georgiopoulos, "A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes," *Data Mining and Knowledge Discovery*, vol. 20, no. 2, pp. 259–289, 2010.

[31] K. Morik and M. Scholz, "The miningmart approach to knowledge discovery in databases," in *Intelligent technologies for information analysis*, pp. 47–65, Springer, 2004.

[32] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in knowledge discovery and data mining," American Association for Artificial Intelligence, 1996.

[33] M. Awad and R. Khanna, *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Apress, 2015.

[34] S. Raschka and V. Mirjalili, *Machine Learning mit Python und Scikit-Learn und TensorFlow - Das umfassende Praxis-Handbuch für Data Science, Predictive Analytics und Deep Learning*. Heidelberg: MITP-Verlags GmbH & Co. KG, 2017.

[35] J. Cleve and U. Lämmel, *Data Mining* -. Berlin: Walter de Gruyter, 2014.

[36] C. C. Aggarwal, *Neural Networks and Deep Learning - A Textbook*. Berlin, Heidelberg: Springer, 2018.

[37] Y. Rybarczyk and R. Zalakeviciute, "Machine learning approach to forecasting urban pollution," in *2016 IEEE Ecuador Technical Chapters Meeting (ETCM)*, pp. 1–6, 2016.

[38] W. Ouyang, B. Guo, G. Cai, Q. Li, S. Han, B. Liu, and X. Liu, "The washing effect of precipitation on particulate matter and the pollution dynamics of rainwater in downtown beijing," *Science of the Total Environment*, vol. 505, pp. 306–314, 2015.

[39] "Bundesumweltamt - Webseite des Bundesumweltamts." https://www.umweltbundesamt.de/daten/luft/feinstaub-belastung#uberschreitungssituation. Accessed: 2020-06-1.

[40] "CDC - Climate Data Center." https://opendata.dwd.de/climate_environment/CDC/. Accessed: 2020-06-1.

[41] "NABU - Artikel über die Schadstoffbelastung im Hamburger Hafen." https://www.nabu.de/umwelt-und-ressourcen/verkehr/schifffahrt/messungen/16819.html/. Accessed: 2020-06-10.

[42] J. Fan, Q. Li, J. Hou, X. Feng, H. Karimian, and S. Lin, "A spatiotemporal prediction framework for air pollution based on deep rnn," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, p. 15, 2017.

[43] A. Zheng and A. Casari, *Feature Engineering for Machine Learning - Principles and Techniques for Data Scientists*. Sebastopol: Ö'Reilly Media, Inc.", 2018.

[44] C. M. Bishop, *Pattern Recognition and Machine Learning* -. Berlin-Heidelberg: Springer New York, 2016.

[45] X. Ni, H. Huang, and W. Du, "Relevance analysis and short-term prediction of pm2. 5 concentrations in beijing based on multi-source data," *Atmospheric environment*, vol. 150, pp. 146–161, 2017.

[46] J. Chen, H. Chen, Z. Wu, D. Hu, and J. Z. Pan, "Forecasting smog-related health hazard based on social media and physical sensor," *Information Systems*, vol. 64, pp. 281–291, 2017.

[47] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," *Atmospheric Environment*, vol. 80, pp. 426–437, 2013.

[48] A. C. Müller, S. Guido, *et al.*, *Introduction to machine learning with Python: a guide for data scientists*. Ö'Reilly Media, Inc.", 2016.

[49] J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, "Modeling pm2. 5 urban pollution using machine learning and selected meteorological parameters," *Journal of Electrical and Computer Engineering*, vol. 2017, 2017.

[50] Y. Zhao, "Machine learning algorithms for predicting roadside fine particulate matter concentration level in hong kong central," *Computational Ecology and Software*, vol. 3, no. 3, p. 61, 2013.

[51] C. Molnar, *Interpretable Machine Learning* -. Raleigh, North Carolina: Lulu.com, 2020.

Erklärung zur selbstständigen Bearbeitung einer Abschlussarbeit

Gemäß der Allgemeinen Prüfungs- und Studienordnung ist zusammen mit der Abschlussarbeit eine schriftliche Erklärung abzugeben, in der der Studierende bestätigt, dass die Abschlussarbeit „— bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit [(§ 18 Abs. 1 APSO-TI-BM bzw. § 21 Abs. 1 APSO-INGI)] — ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt wurden. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich zu machen."

Quelle: § 16 Abs. 5 APSO-TI-BM bzw. § 15 Abs. 6 APSO-INGI

Erklärung zur selbstständigen Bearbeitung der Arbeit

Hiermit versichere ich,

Name: _____

Vorname: _____

dass ich die vorliegende Bachelorarbeit – bzw. bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit – mit dem Thema:

Einsatz von Machine Learning zur Erhebung von Umweltdaten

ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

_____	_____	_____
Ort	Datum	Unterschrift im Original