Data Analytics  
Master Digitale Prozesse und Technologien  
Hochschule für Technik Stuttgart

Projektbericht

**[Titel SDSDDDDDDDDDDDDDD]**

[ggf. Untertitel]

Vorgelegt von:

Antonino Forte   
Matrikelnummer: XXXX, Fachsemester: 2

Bruno Vidal dos Santos  
Matrikelnummer: XXXX, Fachsemester: 1

Vinko Jelić

Matrikelnummer: XXXX, Fachsemester: 2

Abgabedatum: 09.01.2024

Modul: Data Analytics

Semester: Wintersemester 2023/24

Dozenten: Prof. Dr. Sebastian Speiser, Prof. Dr. Ulrike Pado

Inhaltsverzeichnis

[Abbildungsverzeichnis I](#_Toc155450665)

[Tabellenverzeichnis II](#_Toc155450666)

[1. Motivation 1](#_Toc155450667)

[2. Problemstellung 2](#_Toc155450668)

[3. Lösungsansatz 3](#_Toc155450669)

[4. Haupterkenntnisse aus den Daten 4](#_Toc155450670)

[5. Haupterkenntnisse aus der Umsetzung inklusive Schwierigkeiten 5](#_Toc155450671)

[6. Fazit und Ausblick 6](#_Toc155450672)

[Literaturverzeichnis IV](#_Toc155450673)

[Anhang V](#_Toc155450674)

[Eidesstattliche Erklärung VI](#_Toc155450675)

# Abbildungsverzeichnis

# Tabellenverzeichnis

# 1. Motivation

Für unser Projekt liegt die Motivation in der immer weiter fortschreitenden Entwicklung der Datenanalyse und des maschinellen Lernens im Kontext von „Big Data“. Dabei gerät der Trend zur Entwicklung einer künstlichen Intelligenz durch diese Ansätze in den Hintergrund.

Statt das menschliche Denken zu simulieren, sollen spezifische Lösungen für klar definierte Probleme geschaffen werden. Setzt man Algorithmen des maschinellen Lernens ein, können entsprechende Muster aus historischen Daten erkannt und verarbeitet werden.

In der Regel greifen die Methoden des maschinellen Lernens auf objektive und umfassende Daten („Big Data“). Dadurch werden genauere und effizientere Vorhersagen ermöglicht, da die Algorithmen nach Mustern suchen, die am ehesten zur jeweiligen Situation passen.

Dementsprechend lassen sich drei Schlüsselvoraussetzungen aufstellen:

* Datenverfügbarkeit (z.B. durch IoT-Sensoren)
* Rechen- und Speicherkapazitäten
* Verfügbare Algorithmen und Tools (Python inkl. Bibliotheken)

Unser Projekt verfolgt die Zielsetzung, das Potenzial der Datenanalyse und des maschinellen Lernens auszunutzen und tiefe Einblicke in das komplexe Problemfeld zu wahren. Zum Einsatz kommen das Tool Jupyter Notebook sowie die Programmiersprache Python, unterstützt durch leistungsstarke Bibliotheken, wie „pandas“, „numpy“, „scikit-learn“ und weitere.

# 2. Problemstellung

Unser Fokus im Projekt liegt auf der prädiktiven Analyse der Schwere („Severity“) von Verkehrsunfällen in den USA. Als Grundlage nutzen wir den Datensatz „US Accidents“ (vgl. US Accidents (2016 - 2023) 2024) von Kaggle. Somit konzentrieren wir uns auf die Entwicklung von Modellen zur Vorhersage der Unfallschwere und dem zeitgleichen Einblick in die Faktoren, die zu den schwerwiegenderen Unfällen führen.

Die Hauptaspekte unserer Problemstellung werden durch folgende Punkte untermauert:

1. Zielvariable – Severity: Die Vorhersage der Unfallschwere dient als zentrales Ziel der Analyse. Die verschiedenen Kategorien der Schwere (eins bis vier) werden betrachtet, um einordnen zu können, welche Faktoren oder Zusammenhänge mit schweren Unfällen in Verbindung stehen.
2. Analyse: Die Analyse der Faktoren auf die Unfallschwere anhand der verfügbaren Daten, wie Witterungsbedingungen, Verkehrsbedingungen, Straßenzustand, geografische Koordinaten und weiterer relevanter Daten.
3. Modell: Mit Hilfe der ausgewählten Tools sollen prädiktive Modelle erstellt werden. Dabei kommen z.B. Methoden zur Klassifikation zum Einsatz, um die Unfallschwere mit Hilfe der ermittelten und relevanten Faktoren vorherzusagen.
4. Interpretation: Aus den gewonnenen Erkenntnissen aus den Modellen sollen Interpretationen abgeleitet werden. Dabei spielen z.B. mögliche Maßnahmen zur Verbesserung der Verkehrssicherheit oder Infrastruktur eine Rolle. Dazu gehören mögliche Erweiterungen oder Umverteilungen von Ressourcen in Prozessen rund um das Verkehrs- und Rettungsmanagement.

# 3. Lösungsansatz

Der Lösungsansatz besteht aus folgenden Schritten:

**Datenvorbereitung/Datenbereinigung**Die Datenvorbereitung oder Datenbereinigung beinhaltet den Prozess der systematischen Aufbereitung von Datensätzen, um fehlerhafte oder unvollständige Daten zu identifizieren und zu eliminieren. Dies kann die Entfernung von Zeilen mit fehlenden Werten oder unerwünschten Datenpunkten umfassen, um die Qualität und Zuverlässigkeit der Daten für eine anschließende Analyse oder Modellierung zu verbessern.

**Datenexploration**Die Datenexploration umfasst eine eingehende Analyse der "US Accidents" Daten, einschließlich statistischer Kennzahlen, Visualisierungen und Identifikation von Mustern, um eine fundierte Grundlage für den weiteren Analyseprozess zu schaffen.

**Datentransformation**  
In diesem Schritt werden notwendige Datentransformationen durchgeführt, um die Daten für die Modelleignung vorzubereiten. Dazu gehören die Behandlung von fehlenden Werten, Normalisierung von numerischen Variablen und Kodierung von kategorialen Variablen.

**Feature Engineering und Zeitreihen**Feature Engineering beinhaltet die Auswahl und Schaffung relevanter Merkmale, um die Vorhersagegenauigkeit zu verbessern. Zeitreihenanalyse ermöglicht die Berücksichtigung zeitlicher Aspekte in den Daten, was insbesondere für die Unfallvorhersage von Bedeutung ist.

**Clustering**Die Methoden zum Clustering werden angewendet, um homogene Gruppen von Unfällen (Schwere der Unfälle) zu identifizieren und mögliche Muster oder Trends innerhalb dieser Gruppen zu erkennen. Dies erleichtert die differenzierte Betrachtung verschiedener Unfalltypen.

**Klassifikation**Um die Unfallschwere vorherzusagen, werden Klassifikationsalgorithmen benutzt. Dadurch lassen sich Unfälle, basierend auf den identifizierten Einflussfaktoren, zu unterschiedlichen Schweregraden zuordnen.

**Dimensionalitätsreduktion**Um die Komplexität des Modells zu reduzieren und die Effizienz zu steigern, werden Methoden zur Dimensionalitätsreduktion eingesetzt. Dadurch werden nur die relevantesten Variablen in die Modelle einbezogen.

**Regression**  
Um quantitative Beziehungen zwischen verschiedenen Variablen und der Unfallschwere zu modellieren, wird die Regression genutzt. Damit erhält man eine präzise Vorhersage der Unfallschwere, welche auf kontinuierlichen Variablen basiert.

**Evaluation/Erklärbarkeit**  
In diesem Abschnitt liegt der Fokus darauf, die erzielten Ergebnisse umfassend zu interpretieren und zu erklären, wobei besonderes Augenmerk auf die Zusammenhänge und Eigenschaften der Daten gelegt wird. Im Rahmen dieser Analyse wird versucht, über die reinen Machine-Learning-Ergebnisse hinauszugehen und tiefere Einsichten und Verständnis in Bezug auf die Charakteristiken der Daten zu gewinnen.

# 4. Haupterkenntnisse aus den Daten

**Datenvorbereitung (Datenbereinigung)**Das gewählte Datenset weist zu Beginn mehr als acht Millionen Unfälle (Zeilen) auf. Damit verbunden sind einige fehlerhafte Daten in verschiedenen Spalten. Wir haben uns vorgenommen, die Daten im Vorhinein aufzubereiten. Primär umfasst die Bereinigung die Entfernung aller Unfälle bzw. Zeilen, die einen „NaN“-Wert haben. Hierzu sollen alle betroffenen Zeilen ausfindig gemacht und eliminiert werden. Nach der Bereinigung bleiben etwas mehr als sieben Millionen Zeilen übrig. Darüber hinaus werden auch, die von uns als unwichtig bewerteten Spalten, gelöscht.

**Datenexploration**In diesem Abschnitt werden verschiedene Visualisierungen und Analysen der Verkehrsunfalldaten durchgeführt. Zunächst werden Diagramme erstellt, um Trends in der Gesamtanzahl der Unfälle über die Jahre sowie die Top 10 Bundesstaaten mit den meisten Unfällen zu identifizieren. Ein Scatter-Plot auf einer Weltkarte zeigt die geografische Verteilung der Unfälle, wobei ein Zoom auf Nordamerika aufzeigt, dass die meisten Unfälle an der West- und Ostküste stattfinden.

Weiterhin werden Histogramme für die Gesamtanzahl der Unfälle pro Jahr, Monat, Wochentag und Stunde erstellt. Es folgt eine statistische Zusammenfassung der Wetterbedingungen, einschließlich Temperatur, Luftfeuchtigkeit, Druck, Sichtweite und Windgeschwindigkeit. Die Daten werden auch nach Bundesstaaten gruppiert, um Durchschnittswerte für Schweregrad und Wetterbedingungen zu erhalten. Die Daten werden schließlich in einer CSV-Datei gespeichert.

Die Daten kommen in unterschiedlichen Datentypen vor (int, float, bool, obj). Dabei handelt sich oft um Messwerte oder Beschreibungen. Die Daten beschreiben Features, die Verkehrunfälle dokumentieren. Der dabei am wichtigsten erscheinende Faktor hier ist die Schwere des Unfalls (Severity). Weiterhin können die Daten in Kategorien unterteilt werden, wie Wetter, Straßeninformationen, zeitliche Faktoren (Sonnenuntergang etc.) und räumliche Faktoren. Die Daten könnten für die Vorhersage der schwere eines Unfalles dienen. Dadurch könnte erreicht werden, dass beispielsweise Notdienste besser und schneller auf Unfälle reagieren können. Weiterhin könnten man die Analysen durchführen, um Unfälle und deren Schwere auf räumliche Faktoren und Infrastruktur zu untersuchen. Dadurch können Hotspots identifiziert werden, die zu Maßnahmen wie umbauten an einer Straße führen könnten.

**Datentransformation**Hier werden zusätzliche Features zu den Unfalldaten hinzugefügt. Wir haben dafür eine Konvertierungstabelle von Bundesstaaten zu Regionen eingebunden und in unsere Daten integriert. Damit sind drei neue Spalten hinzugekommen. Im Anschluss wird eine Pivot-Tabelle erstellt, die die Anzahl der Unfälle nach der Schwere und Wetterbedingungen aufschlüsselt.  
Eine grafische Darstellung in Form eines Kreisdiagramms verdeutlicht die Anteile der Unfälle nach verschiedenen Wetterbedingungen. Die Analyse zeigt, dass die meisten Unfälle bei heiterem Wetter („Fair“) auftreten.

**Clustering**Es wird eine Clusteringanalyse auf den Unfalldaten durchgeführt, um mögliche Muster oder Trends in den Daten herauszufinden. Zunächst werden alle verfügbaren Features für die Analyse verwendet.

Wir haben festgestellt, dass die Unfallschwere (Severity) nicht dazu beiträgt, klare Cluster in den Daten zu erkennen. Die erste Visualisierung der Cluster war daher intuitiv nicht aussagekräftig. Die Erhöhung der Anzahl der Cluster brachte keine Verbesserung, wie im zweiten Scatterplot ersichtlich. Durch die Änderung der Featureauswahl konnten jedoch gut differenzierbare Cluster identifiziert werden. Insbesondere eignet sich die Verwendung von drei Clustern im Vergleich zu einer höheren Anzahl (15).

Die Analyse der Zentroide und Cluster in der Visualisierung zeigt, dass die Cluster unterschiedliche klimatische und geografische Bedingungen repräsentieren. Diese Bedingungen korrelieren mit der Häufigkeit von Unfällen. Zum Beispiel zeigt Cluster 1 Unfälle in kälteren Temperaturen und höherer Luftfeuchtigkeit, während Cluster 2 Unfälle in wärmeren Gebieten mit hoher Luftfeuchtigkeit repräsentiert.

Des Weiteren könnte das Clustering verwendet werden, um Ausreißer, insbesondere bei den numerischen Werten (Temperature, Humidity), zu identifizieren. Eine mögliche Vorgehensweise wäre, den Abstand zwischen den Datenpunkten und dem Zentroid zu berechnen (als absoluten Wert). Die maximalen und minimalen Temperaturen könnten aus externen Quellen wie dem Internet abgelesen werden, und Datenpunkte außerhalb dieser Grenzen könnten eliminiert werden.

# 5. Erkenntnisse aus der Umsetzung inklusive Schwierigkeiten

# 6. Fazit und Ausblick

# Literaturverzeichnis

# Anhang

# Eidesstattliche Erklärung

Der Wortlaut der eidesstattlichen Erklärung wird dir meist von deinem Fachbereich vorgegeben. Wenn das nicht der Fall ist, findest du hier ein Beispiel: [Eidesstattliche Erklärung | Muster zum Download](https://www.scribbr.de/hausarbeit/eidesstattliche-erklaerung-hausarbeit/)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe, dass ich sie zuvor an keiner anderen Hochschule und in keinem anderen Studiengang als Prüfungsleistung eingereicht habe und dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die wörtlich oder sinngemäß aus Veröffentlichungen oder aus anderweitigen fremden Äußerungen entnommen wurden, sind als solche kenntlich gemacht.

\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_ \_

Ort, Datum Unterschrift