Exposé

in dem Modul

Web Mining

zu dem Thema:

Web-Scraping von Daten der ersten Bundesliga zur Vorhersage von Spielergebnissen

Vorgelegt im berufsbegleitenden Studiengang M.Sc. Data Science

von

Alfred Anselm
Matrikelnummer 30258459

Kevin Diec
Matrikelnummer 30245778

Luca Janas
Matrikelnummer 30277119

Prüfer: Prof. Dr. Christian Gawron
Im Sommersemester 2023

Eigenständigkeitserklärung

Ich erkläre hiermit, dass die vorgelegte Arbeit mein eigenes Werk ist. Alle direkt oder indirekt verwendeten Quellen sind als Referenzen angegeben. Die Arbeit wurde bisher nicht vor einem anderen Prüfungsausschuss vorgelegt und nicht veröffentlicht.

Mir ist bekannt, dass die Arbeit in digitaler Form auf die Verwendung unerlaubter Hilfsmittel überprüft werden kann, um festzustellen, ob die Arbeit als Ganzes oder darin enthaltene Teile als Plagiat zu werten sind. Für den Vergleich meiner Arbeit mit vorhandenen Quellen erkläre ich mich damit einverstanden, dass sie in eine Datenbank aufgenommen wird und dort auch nach der Prüfung verbleibt, um einen Vergleich mit künftigen eingereichten Arbeiten zu ermöglichen.

Münster, 11. Juni 2023.

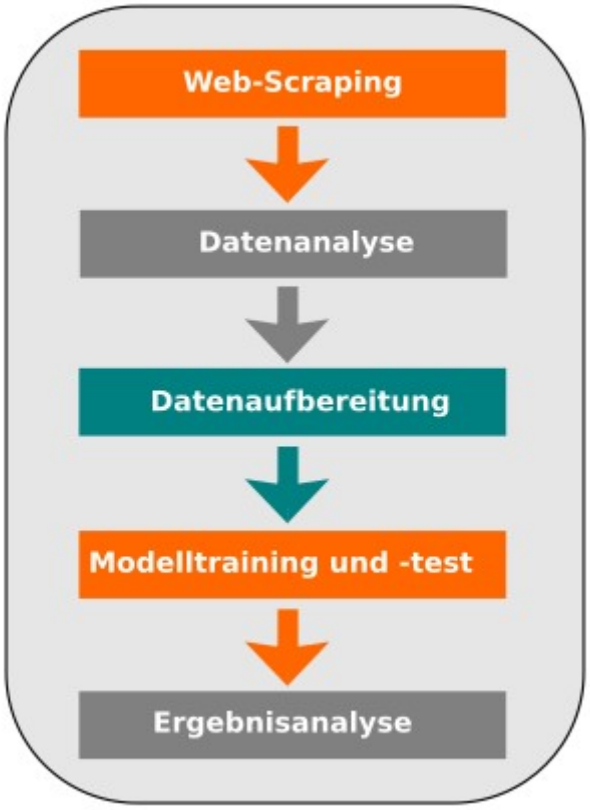
Alfred Anselm **Kevin Diec** **Luca Janas**

I Inhaltsverzeichnis

1 Projektbeschreibung

In dem hier vorgelegten Exposé werden Inhalte und der Aufbau des Projektes im Modul Web Mining im berufsbegleitenden M.Sc. Data Science an der Fachhochschule Südwestfalen beschrieben. Ziel des Projektes ist es, unter Verwendung von Daten, die auf der Webseite https://transfermarkt.de/ zur Verfügung gestellt werden, Spielergebnisse in der ersten deutschen Bundesliga vorherzusagen. Dazu werden historische Daten zu Spielen und Vereinen von transfermarkt.de durch Web-Scraping gesammelt und anschließend aufbereitet und analysiert, um einen Datensatz zu erstellen, der zur Vorhersage der Spielergebnisse verwendet werden kann. Dafür bietet Transfermarkt Informationen zu Fußballspielern, Vereinen, Marktwerten und Statistiken für verschiedene Ligen weltweit.

Das Projekt kann in folgende 5 Einzelteile aufgeteilt werden.



2 Web-Scraping

Für die Selektion der Daten zu den Mannschaften aus der ersten Bundesliga und den jeweiligen Spielergebnissen pro Spieltag und Saison wird das Verfahren des Web-Scrapings auf die Internetseite https://transfermarkt.de/ als Basis-URL angewendet. Für die jeweilige Datenselektion werden weitere URL-Pfade untersucht, die unter anderem Parameter beinhalten, welche die ausgewählte Saison und Spieltage enthalten. Über eine Loop-Funktion können alle Kombinationsmöglichkeiten daraus extrahiert werden.

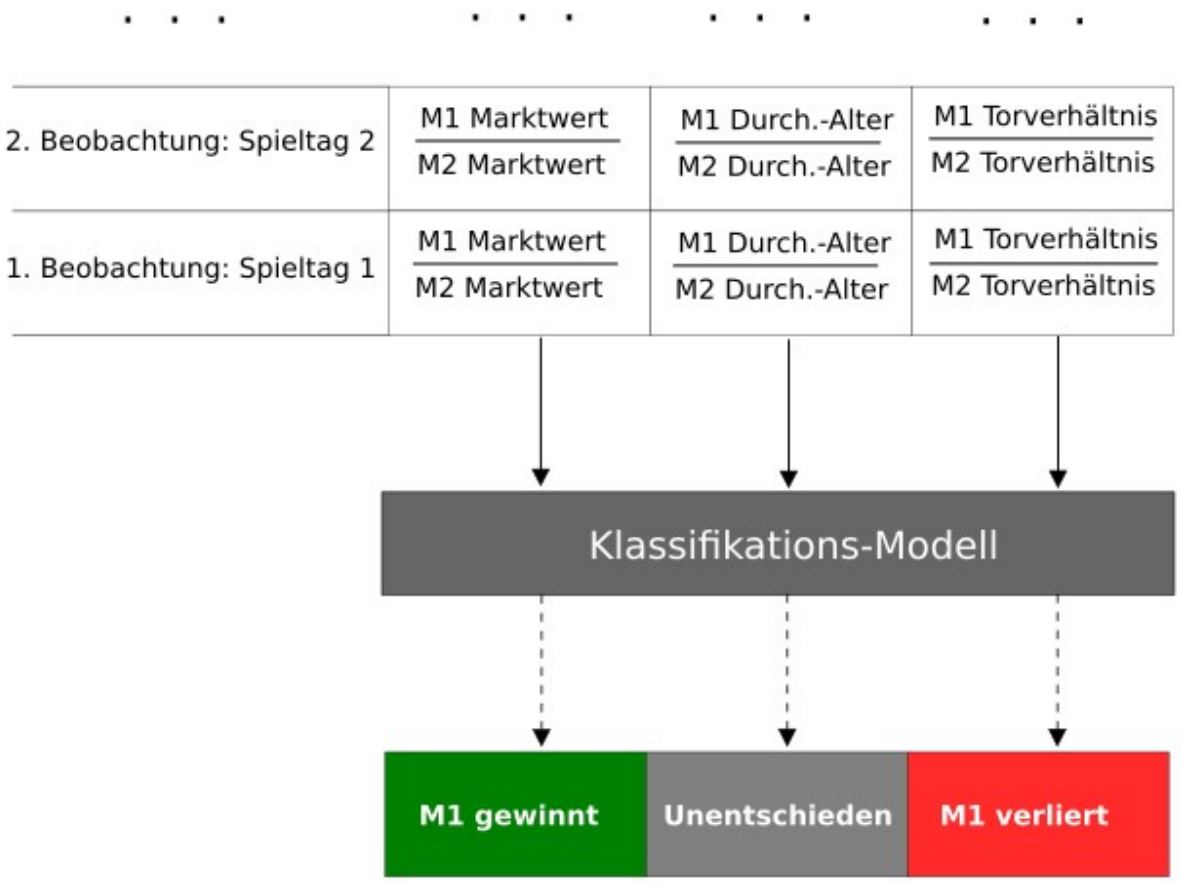
Die erste Datenselektion erfolgte mit den Python-Bibliotheken *requests*, *BeautifulSoup* und *lxml.* Mit dem Modul *requests* ist es möglich, eine Anfrage an die vorgegebene URL zu senden und die HTML-Daten der Webseite auszulesen und zu extrahieren.
Um die extrahierten Daten aus dem HTML-Quellcode zu analysieren und zu filtern, wird das Modul *BeautifulSoup* verwendet. Es handelt sich dabei um eine Python-Bibliothek, die beim Web-Scraping eine wichtige Rolle spielt. Sie dient dazu, HTML-Dokumente zu parsen und sie in einer aufbereiteten Struktur darzustellen. In dieser Ausarbeitung wird *BeautifulSoup* in Verbindung mit der „lxml"-Bibliothek als Parser verwendet. Durch die Verwendung von *lxml* kann *BeautifulSoup* von den Funktionen und Vorteilen dieser Bibliothek profitieren, um die Analyse und Manipulation von Webseiteninhalten zu optimieren. Dabei bietet sie eine Reihe von Funktionen und Methoden an, um den Inhalt von Webseiten zu analysieren, die Datenstruktur zu verstehen und spezifische Elemente wie Tabellen, Überschriften, Links oder Absätze zu identifizieren.

Um die Daten weiterzuverarbeiten und zu analysieren, wird in gesonderten Fällen die Methode *etree* aus dem Modul *lxml* verwendet. Diese Methode ermöglicht es, die HTML-Struktur der Webseite genauer zu untersuchen und die Informationen gezielt auszuwählen. Insbesondere wird hierbei auf die Struktur der XPATH-Logik mit *etree* zurückgegriffen, um Schwierigkeiten bei der eindeutigen Identifizierung der Struktur zu überwinden. Das XPATH-Format erlaubt es, bestimmte Elemente in einem HTML-Dokument basierend auf ihrer Position und Hierarchie in einer Tabelle zu identifizieren und über eine For-Loop über die Reihen und Spalten gezielt auszuwählen.

Erste Web-Scraping Selektionen sind in dem GitHub Repository einzusehen:
web-mining/Crawler.ipynb at main · lucajanas/web-mining (github.com)

3 Datenanalyse

Zur Datenanalyse gehört die Auswahl der Features, mit denen das spätere Modell trainiert werden soll. Features wie *Marktwert der Mannschaft*, *Torverhältnis*, *Durchschnittsalter* usw. könnten gute Indikatoren zur Prognose eines Spielausgangs sein.
Bei der Auswahl der Features werden unterschiedliche Möglichkeiten zum Aufbau des Datensatzes untersucht. Eine Möglichkeit wäre jedes Feature in zweifacher Ausprägung zu haben, jeweils einmal für Heim- und Auswärtsmannschaft. Eine Alternative dazu wäre ein Modell, welches auf den Verhältnissen der Features von Heim- und Auswärtsmannschaft trainiert wird.



4 Datenaufbereitung

Wenn die Datenstruktur festgelegt wurde, gilt es die Trainingsdaten aufzubereiten. Einige Features weisen eine saisonale Dynamik auf. Bei einigen Features ergibt sich die Dynamik sogar pro Spieltag. Ziel ist ein sequenziell-spielabhängiger Aufbau der Datenstruktur, dessen finales Resultat die Features (X-Werte) und dazugehörigen Labels (Y-Wert/Spielausgang) darstellen und direkt von einem Klassifikationsmodell verarbeitet werden können.

5 Modelltraining und -test

Für die Modellprognose, ob eine Mannschaft gewinnt, verliert oder unentschieden spielt, sollen bekannte Klassifikatoren wie XGBoost-Klassifikator oder SVM-Klassifikator zum Einsatz kommen. Dabei wird der Datensatz als Erstes in ein Trainings- und Testdatensatz aufgeteilt. Mit dem Trainingsdatensatz wird das Modell trainiert und mit Hilfe des Testdatensatzes wird die Performance des trainierten Modells/Modelle überprüft.

6 Ergebnisanalyse

Nach der Prognose wird mit Hilfe der Testdaten die Performance der Modelle mit Klassifikationsmetriken wie z.B. Accuracy, Precision, Recall überprüft und gemessen. Dabei wird auch untersucht, ob die Prognose für bestimmte Mannschaften besser zutrifft als für andere. Die Ergebnisse werden graphisch aufbereitet und gegenübergestellt.