Data Science 1. Team Napoleon
Tim Frommknecht (6368432) tim.frommknecht@gmx.de
Duy Anh Tran (4959759) duy_anh@live.de
Anh Duong Vo (6047944) anhduongvo.m@gmail.com
Sourcecode: `https://github.com/finnelias/bundesligaWetter_teamNapoleon`

1 Einleitung

Wir vereinigten Fußballdaten (1.Bundesliga) mit zum Spieltag zugehörigen Wetterdaten. Mithilfe von Clusteranalysen sollten folgende Hypothesen untersucht werden:
T1 Schüsse/Spielsieg: Eine hohe Schussanzahl in Richtung des gegn. Tors geht mit Spielsieg einher.
T2 Teamstrategie: Bundesligateams unterscheiden sich in ihrer Spielweise anhand Ergebnis, Fouls und Karten.
T3 Niederschlag/Schüsse: Niederschlag am Spieltag geht mit einer je niedrigen Schussanzahl beider Teams einher.

2 Methodik

2.1 Datenaufbereitung

Grundlage des Projekts sind 1. Bundesligaspieldaten[1] von Janunar 2009 bis Juni 2020. Von diesem Datensatz (12 CSV-Dateien) wurden 21 Attribute ausgewählt (s. Code Kapitel 1) und zusammengeführt.
Die Wetterdaten wurden mittels der Wetter-API *Word Weather Online*[2] bezogen. Zunächst wurde das Attribut *HomeCity* aus den Heimteams generiert. Mithilfe des Datums und *HomeCity* konnten die Wetterdaten für den jeweiligen Tag um 15 Uhr mit der API abgefragt werden. Die Wetterdaten (24 Wetterattribute) wurden in CSV-Dateien ausgegeben und mit Exel-Macros zu den Bundesligadaten ergänzt. Ferner wurde die Datenqualität überprüft:

Tabelle 1: Datenqualitätsanalyse mit Bewertung von 1-5

Dimension	Bewertung	Bundesliga	Wetter
Vollständigkeit	3	Nullwert-Überprüfung der Attributswerte (s. Code Kapitel 2)	
		Tägl. Aktualisierung/Wartung d. Website	
		Alle Spiele innerhalb der Saison vollständig; Uhrzeit nicht vorhanden	Deckt den untersuchten Zeitraum & Orte ab
Redundanz	5	Spiel eindeutig durch Datum und Mannschaften	Wetter eindeutig durch Datum und Ort
		keine Duplikate	
Zuverlässigkeit	4	Primäre Quelle: Deutsche Fußball Liga GmbH	Eigene Wetterstationen & Modell, verwenden NASA-Daten; Abgleich mit *World Metrological Organisation*
Konsistenz	5	gleiches Format und Struktur d. Datenpunkte	
Korrektheit	4	Verifizierung mit Histogrammen (s. Code Kapitel 2)	

2.2 Clustering mit K-Means und Hierarchisches Clustern

Wir arbeiteten mit K-Means und Agglomeratives Hierarchisches Clustering (AHC). Folgende Parameter wählten wir aus:

Tabelle 2: Parameterwerte. HS = #Heimschüsse in Richtung gegn. Tor, HST = #Auswärtsschüsse auf gegn. Tor, AS = #Auswärtsschüsse in Richtung gegn. Tor, AST = #Auswärtsschüsse auf gegn. Tor, FTR = Endergebnis, rain = Niederschlag ja/nein

These	Parameter	K-Means und AHC
1	#Cluster	4
	Attribute	HS, HST, AS, AST
	Zusätzl. Attr. im Graph	FTR
2	#Cluster	6
	Attribute	FTR, HF, HY, HR
	Zusätzl. Attr. im Graph	HomeTeam (#Heimspiele > 100)
3	#Cluster	4
	Attribute	HS, HST, AS, AST
	Zusätzl. Attr. im Graph	rain

Bei T1 und T3 entschieden wir uns für 4 Cluster, da wir die Gruppen in Tab. 3 erwarteten. Diese könnten wir anhand von FTR (T1) bzw. von rain (T3) graphisch darstellen und mögliche Abhängigkeiten erkennen. Nach einer Analyse der Abhängigkeiten der Attribute durch Graphen und Korrelationsmatrix entschieden wir uns für die restlichen Attribute (s. Code Kapitel 3).
Bei T2 wählten wir für ein vielfältigeres Bild der Strategien 6 Cluster. Wir untersuchten dies für Heimteams, um keine Unterschiede durch Heimvorteil zu haben und nur Teams mit über 100 Heimspielen für ein aussagekräftiges Ergebnis.

Tabelle 3: Erwartete Cluster für T1 und T3

C	Eigenschaften
1	Niedrige Heim-, hohe Auswärtschüsse
2	Hohe Heim-, hohe Auswärtsschüsse
3	Niedrige Heim-, niedrige Auswärtsschüsse
4	Hohe Heim-, niedrige Auswärtsschüsse

3 Ergebnisse und Diskussion

T1 Schüsse/Spielsieg: Die Cluster C bei K-Means(Abb. 1A) entsprechen etwa der Erwartung aus Tab. 3. Mehr Schüsse eines Teams (C1/C4) treten häufiger auf, wenn das Auswärts-/Heimteam gewinnt (Abb. 1C), was unsere These bestätigt. Eine ähnliche Anzahl Schüsse des Heim- und Auswärtsteam geht mit einem Unentschieden einher (C2,C3).
Bei AHC (Abb. 1D) sind die Cluster bei allen Spielen ähnlich verteilt. Man kann hier keine Aussage zu T1 treffen, da wir keinen Cluster mit auschließlich der Eigenschaft *Entweder hohe Heim- oder Auswärtsschüsse* vorfinden.

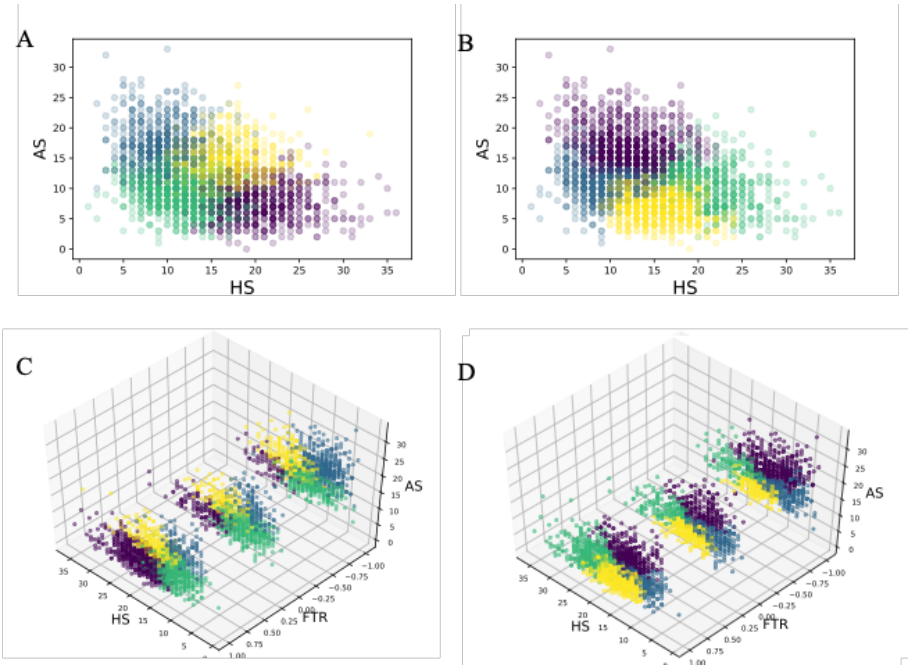

Abbildung 1: Resultate T1. **A:** HS in Abhängigkeit von AS mit K-Means mit Cluster 1 blau, 2 gelb, 3 grün, 4 violett. **B:** HS in Abhängigkeit von AS mit AHC. **C:** FTR in Abhängigkeit von HS und AS mit K-Means. **D:** FTR in Abhängigkeit von HS und AS mit AHC

T2 Teamstrategie: In Tab. 4 sind die Strategien durch das Clustering einsehbar, was T2 bestätigt. Bei K-Means (Abb. 2A) ist z.B. erkennbar, dass Bayern München oft in Cluster 4 und seltener in Cluster 3 und 6 vertreten ist.
Beim AHC variieren zwischen den Clustern Foul stärker und Spielausgang und Karten weniger. Durch die hohe Variabilität in Clustern entsteht Informationsverlust. Alle Teams sind zum Großteil in den Clustern 1 und 3 vertreten (Abb. 2B). Vermeintlich stärkere Teams sind häufiger in 3 als vermeintlich schwächere.
K-Means ist hier besser geeignet: Die Cluster sind gleichmäßiger verteilt und Mittelwerte variieren im Spielausgang mehr, was die Analyse erleichtert. (für teambezogenere Analyse und Clusterwerte s. Code Kapitel 4)

Tabelle 4: Clustereigenschaften T2

C K-Means	Eigenschaft
1	Mittel viele Fouls, viele gelbe Karten
2	Viele Fouls, wenig Karten
3	Alle Spiele mit roter Karte
4	Viele Siege, wenig Fouls/Karten
5	Wenig Siege, wenig Fouls/Karten
6	Wenig Siege, viele Fouls/Karten
AHC	
1	Durchschnitt Spielwerte aller Attribute
3	Etwas fairer und mehr Siege als Durchschnitt
2,4,5	Weniger Siege, unfair
6	mehr Siege, besonders fair

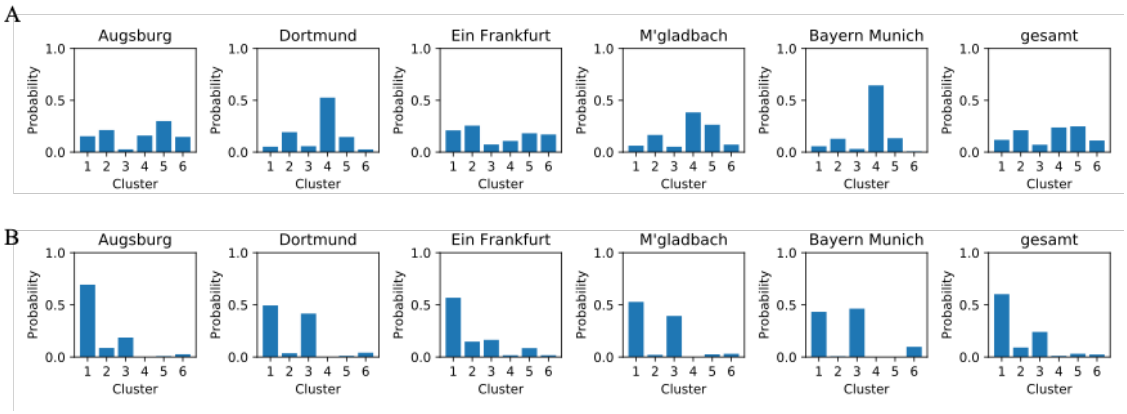

Abbildung 2: Resultate T2. Clusterzugehörigkeitsanteile ausgewählter Teams mit **A:** K-Means und **B:** AHC

T3 Niederschlag/Schüsse: Wir verwenden die Cluster aus T1 in Abhängigkeit von Regen (Abb. 3). Bei K-Means (Abb. 3A) müssten nach T3 C1, C3, C4 bei Regen verhältnismäßig größer sein, da die Teams weniger schießen. Die Clustergrößen ändern sich jedoch nicht bei Regen. Also ist keine Abhängigkeit erkennbar. AHC ist aufgrund der Clusterbildung wie in T1 erläutert nicht für unsere Analyse geeignet.
Fehler könnten aufgetreten sein, weil wir Wetterdaten von jeweils 15 Uhr nutzten, die Spiele aber zu unterschiedlichen Zeiten stattfanden. Zudem könnte der Niederschlag keinen starken Einfluss auf die Schussgenauigkeit haben. Bei Temperatur wurden auch keine nennenswerten Abhängigkeiten gefunden.

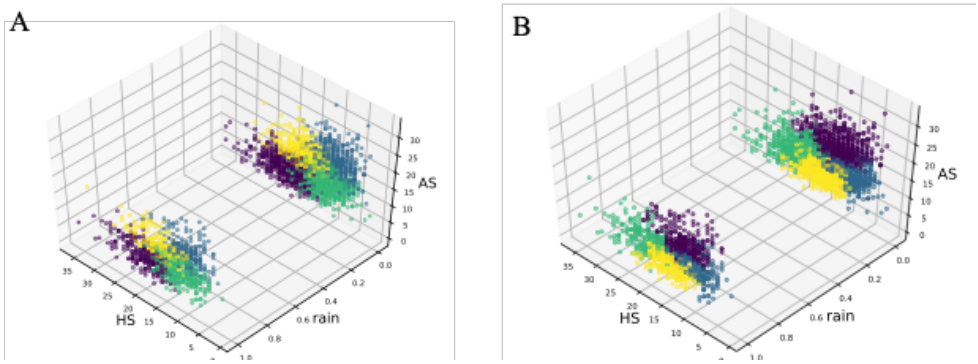

Abbildung 3: Resultate T3. **A:** rain in Abhängigkeit von HS und AS mit 4-Means-Clustering mit Cluster 1 blau, 2 gelb, 3 grün, 4 violett. **B:** rain in Abhängigkeit von HS und AS mit AHC

Laut Silhouette-Score (Tab. 5) finden beide Clustering-Verfahren keine Struktur bzw. klar definierte Gruppen ($s_C < 0.25$). Das ist für unsere Ergebnisse keine Einschränkung, da für die Verteilungsanalyse keine Clusterabgrenzung benötigt wird.

Häufigkeiten, Verteilungen, Zentroiden der Cluster usw. sind im Code Kapitel 4 einsehbar.

Tabelle 5: Silhouette-Score s_C von K-Means, AHC

These	K-Means	AHC
1	0.23	0.17
2	0.05	0.17
3	0.23	0.17

4 Fazit

Wir wandten K-Means und AHC auf vereinigte Bundesliga- und Wetterdaten an. Wir verifizierten T1 und T2, wobei K-Means mehr geeignet war. T3 wurde widerlegt. Der Einsatz eines Klassifikators (wie Regression, SVM) des Spielausgangs wäre auch sinnvoll, um ungeahnte Abhängigkeiten zu finden. Zudem könnten Uhrzeiten der Spiele förderlich sein, um Wetterdaten genauer zuzuordnen.

Literatur

[1] *football-data.co.uk*. [Online; accessed 2020-06-25]. URL: `https://www.football-data.co.uk/germanym.php`.

[2] *World Weather Online*. [Online; accessed 2020-06-25]. URL: `https://www.worldweatheronline.com/developer/`.