Hochschule für Technik Stuttgart

Studiengang: Digitale Prozesse und Technologien

Modul: Data Analytics

Projektbericht Gruppe 6:
Obesity-Dataset

Eingereicht von: Sera Can (1002692)
Dorian Schimmel (1002525)
Michael Held (1001993)

Betreuende Lehrperson: Prof. Dr. Sebastian Speiser

Abgabetermin am: 09. Januar 2025

Inhalt

1. Motivation

Adipositas ist ein wachsendes globales Gesundheitsproblem, das sowohl die Lebensqualität der Betroffenen als auch die Ressourcen des Gesundheitssystems erheblich belastet (Phelps et al. 2024). Weltweit sind immer mehr Menschen übergewichtig oder fettleibig, was schwerwiegende gesundheitliche Folgen wie Herz-Kreislauf-Erkrankungen, Typ-2-Diabetes und bestimmte Krebsarten nach sich ziehen kann. Die Verbesserung des allgemeinen Gesundheitszustands der Bevölkerung ist daher von entscheidender Bedeutung (Carlberg 2023).

Ein besseres Verständnis der Faktoren, die das Adipositasrisiko beeinflussen, ist für die Entwicklung wirksamer Präventionsstrategien von entscheidender Bedeutung. Änderungen des Lebensstils wie Ernährungsumstellung, mehr körperliche Aktivität und bewusste Ernährung bieten ein hohes Potenzial zur Verringerung des Adipositasrisikos. Dabei sollte eine Kombination mehrerer dieser Ansätze die besten Ergebnisse erzielen. Ziel ist es, durch gezielte Maßnahmen nicht nur das individuelle Wohlbefinden zu steigern, sondern langfristig auch die gesellschaftliche und wirtschaftliche Belastung durch Übergewicht zu reduzieren (World Health Organization 2022).

2. Problemstellung

Die zunehmende Prävalenz von Adipositas stellt eine große Herausforderung für die öffentliche Gesundheit dar. Trotz zahlreicher Bemühungen ist die Prävalenz übergewichtiger und adipöser Menschen in den letzten Jahrzehnten weiter angestiegen. Dieser Trend deutet darauf hin, dass die bestehenden Maßnahmen nicht wirksam genug oder nicht auf die individuellen Bedürfnisse der Betroffenen zugeschnitten sind. Die Komplexität der Adipositas wird durch eine Vielzahl von Einflussfaktoren verstärkt, darunter genetische Veranlagung, demografische Merkmale, Ernährungsgewohnheiten und Verhaltensweisen (World Health Organization 2022).

Ein zentraler Aspekt ist die Frage, wie wirksam bestimmte Änderungen des Lebensstils bei der Prävention von Adipositas sind. Veränderungen wie eine ausgewogene Ernährung und regelmäßige körperliche Aktivität können dabei signifikante Vorteile bieten (World Health Organization 2022). Dabei ist allerdings nicht bekannt, welche Kombination von Maßnahmen zu den nachhaltigsten Ergebnissen führt.

Darüber hinaus erfordert die Entwicklung zielgerichteter Präventionsprogramme eine solide Datengrundlage. Der vorliegende Datensatz zur Adipositas, der demographische Merkmale, Ernährungs- und Lebensgewohnheiten sowie verhaltensbezogene Einflüsse umfasst, bietet die Möglichkeit, vertiefte Erkenntnisse über die Zusammenhänge zwischen Lebensstil und Adipositas zu gewinnen. Dieses Wissen könnte die Grundlage für Ansätze zur Prävention und langfristigen Gewichtskontrolle bilden. Die Umsetzung solcher Programme würde nicht nur die Zahl der Übergewichtigen reduzieren, sondern auch das Gesundheitssystem entlasten (World Health Organization 2022).

3. Lösungsansatz

Ein datenbasierter Ansatz zur Prävention von Adipositas setzt auf eine umfassende Analyse der vorhandenen Daten (Koklu und Sulak 2024), um zentrale Risikofaktoren und Zusammenhänge zu identifizieren. Der Datensatz zur Fettleibigkeit ist durch eine Online-Umfrage mit 1610 Teilnehmern erstellt worden. Er enthält Informationen zu verschiedenen Faktoren, die das Risiko von Fettleibigkeit beeinflussen könnten. Dazu gehören demografische Merkmale, Ernährungs- und Lebensgewohnheiten. Die einzelnen Studienteilnehmer werden in die Klassen Untergewichtig, Normalgewichtig, Übergewichtig und Adipös eingeteilt.

Zu Beginn ist es notwendig, die demografischen Merkmale wie Alter oder Geschlecht zu untersuchen, da diese Faktoren oft eine Rolle beim Adipositasrisiko spielen. Ergänzend dazu werden Verhaltensmuster wie der Konsum von Fast Food, die Häufigkeit des Gemüseverzehrs, das Bewegungsverhalten sowie der Umgang mit Technologie analysiert. Zusätzlich werden weitere Merkmale berücksichtigt wie Zusammenfassungen des Lebensstils, eine Gesamtbewertung des Ernährungsverhaltens sowie aggregierte Werte wie der BMI oder das Gewicht, um eine variablere Skalierung und präzisere Analysen durchzuführen. Mithilfe deskriptiver Statistiken werden Muster wie Häufigkeiten und Verteilungen erfasst, um festzustellen, welche Bevölkerungsgruppen besonders anfällig für ungesunde Verhaltensweisen sind. Zum Beispiel können Zusammenhänge zwischen dem Fast-Food-Konsum und dem Gewicht erkannt werden.

Basierend auf dieser Analyse wird im nächsten Schritt ein prädiktives Modell entwickelt, das das individuelle Risiko für Adipositas auf Grundlage der Eingabeparameter vorhersagen kann. Hierfür kommen fortgeschrittene Algorithmen des maschinellen Lernens in Frage wie beispielsweise Entscheidungsbäume, SVM, XGBoost, logistische Regression oder Random-Forest-Modelle. Das Modell wird trainiert und getestet, um eine hohe Vorhersagegenauigkeit sicherzustellen. Mithilfe von Regressionsanalysen wird der Einfluss einzelner Faktoren wie körperliche Aktivität oder Essgewohnheiten auf die Wahrscheinlichkeit einer höheren BMI-Kategorie quantifiziert. Clusteranalysen dienen dazu, Gruppen mit ähnlichen Verhaltensmustern zu identifizieren, etwa solche mit hoher körperlicher Aktivität und gesundem Essverhalten. Diese methodischen Ansätze bieten eine detaillierte Grundlage zur Identifikation und Bewertung der wichtigsten Einflussfaktoren.

Die datenbasierte Analyse bildet somit die Grundlage, um tiefgreifende Erkenntnisse über die zugrundeliegenden Einflussfaktoren von Adipositas zu gewinnen. Sie erlaubt

eine strukturierte Betrachtung der vielfältigen Daten, wodurch Korrelationen und Muster sichtbar werden, die zur weiteren Untersuchung dienen. Die gewonnenen Informationen liefern ein Bild über die Effekte einzelner Einflussfaktoren auf das Körpergewicht der Bevölkerung. Im weiteren Verlauf werden die zentralen Ergebnisse und Schlussfolgerungen dieser Analyse detailliert dargestellt.

4. Haupterkenntnisse aus den Daten

Die Haupterkenntnisse der Befragung zeigen, dass ältere Menschen im Vergleich zu jüngeren aktiver sind. Körperliche Aktivität hat jedoch, auf die Gesamtheit betrachtet, keinen signifikanten Einfluss auf das Gewicht. Die Mehrheit der Teilnehmer sind jüngere Personen mit einem Altersbereich von 18 bis 54 Jahren, einem Durchschnittsalter von etwa 33 Jahren und einem Median von 32 Jahren. Ältere Menschen konsumieren im Durchschnitt mehr Hauptmahlzeiten als jüngere, während nur etwa 25% der Befragten angeben, regelmäßig Fast Food zu konsumieren. Unter- und normalgewichtige Teilnehmer sind im Durchschnitt etwa 26 Jahre alt, während übergewichtige und adipöse Teilnehmer ein durchschnittliches Alter von 38 bis 39 Jahren aufweisen. Es gibt kaum unter- oder normalgewichtige Personen in der älteren Altersgruppe. Die Darstellung der Daten als Zahlen kann teilweise irritieren, z.B. bei der Frage, ob jemand raucht. Aus den Antworten 1 und 2 ist keine Aussage direkt ablesbar.

Um eine präzise Datenanalyse für die weiteren Aufgaben zu ermöglichen, wurden zusätzliche Variablen in die Untersuchung integriert. Diese umfassen den Body-Mass-Index (BMI), die körperliche Aktivität (Activity), das Ernährungsverhalten (FoodConsumption) sowie das Körpergewicht (Weight). Die neuen Datensätze wurden dabei aus den bestehenden Daten generiert.

Die Daten lassen sich anhand der entwickelten Merkmale mithilfe des k-Means-Algorithmus in vier Cluster gliedern: Cluster A zeichnet sich durch hohe körperliche Aktivität, ungesunde Ernährungsgewohnheiten und eine große Gewichtsspanne aus, während Cluster B niedrige Aktivität, durchschnittlichen Lebensmittelkonsum und ein geringes Gewicht aufweist. Cluster C kombiniert hohe Aktivität mit gesundem Ernährungsverhalten und einem niedrigen bis mittleren Gewicht, während Cluster D durch unterdurchschnittliche Aktivität, ungesunde Ernährungsgewohnheiten und ein hohes Gewicht charakterisiert ist. Diese Einteilung verdeutlicht, dass der vorliegende Datensatz eine sinnvolle Gruppierung ermöglicht.

Die Zielvariable für die Klassifikation ist Class, was die Gewichtsklasse darstellt. Um den besten Algorithmus zur Vorhersage des Zielwertes zu finden, wurden SVM, Random Forest, Logistic Regression und XGBoost gegeneinander getestet. Die durchgeführte Analyse zeigt, dass das Random-Forest-Modell und XGBoost die besten Ergebnisse bei der Vorhersage der Gewichtsklasse (Class) liefern. Während Random Forest das Alter als das wichtigste Merkmal identifiziert, erkennt XGBoost die Anzahl der täglichen Hauptmahlzeiten als entscheidendes Feature. Beide Modelle erreichen eine

hohe Gesamtgenauigkeit, wobei das Random-Forest-Modell die Gewichtsklasse mit einer Wahrscheinlichkeit von 86% korrekt vorhersagt.
Normalgewicht zeigt mit einem F1-Score von 0,92 die beste Leistung, was auf die hohe Repräsentation dieser Klasse im Datensatz zurückzuführen ist (Support: 132). Untergewicht erzielt die höchste Genauigkeit (0,92), weist jedoch eine niedrigere Trefferquote (Recall: 0,73) auf, was auf eine unzureichende Erkennung vieler Fälle hindeutet. Adipositas zeigt mit einer Genauigkeit von 0,73 die schwächste Leistung, da es zu einer erhöhten Anzahl von falsch positiven Vorhersagen kommt.
Da es sich bei der Gewichtsklassifikation um ein sensibles Thema handelt, ist der Recall-Wert entscheidender als die Präzision, da ein niedriger Recall insbesondere bei kritischen Kategorien wie Adipositas und Untergewicht dazu führen kann, dass potenzielle Gesundheitsrisiken übersehen werden. Trotz der vergleichsweise guten Gesamtleistung ist das Modell aufgrund der niedrigen Recall-Werte für diese Klassen nicht praxistauglich. Verbesserungen in der Erkennung dieser Kategorien sind notwendig, um eine zuverlässige Anwendung zu gewährleisten.
Um die Datei übersichtlicher zu gestalten, wurde mit der Principal Component Analysis (PCA) eine Dimensionsreduktion durchgeführt. Dabei erklären bereits 10 Hauptkomponenten etwa 85 % der Varianz der ursprünglichen 9 Dimensionen. Dies zeigt, dass ein Großteil der Information auch mit weniger Dimensionen erhalten bleibt. Der reduzierte Datensatz kann nun für effizientere Modelle oder eine vereinfachte Datenanalyse genutzt werden.
Die Modellqualität einer Regression wird anhand des R^2-Werts, der p-Werte und der Koeffizienten beurteilt. Der R^2-Wert zeigt, welcher Anteil der Varianz in den Zieldaten durch das Modell erklärt werden kann. Die p-Werte geben an, ob die Einflussgrößen statistisch signifikant sind, und die Koeffizienten beschreiben die Richtung und Stärke ihres Einflusses. Die Analyse der Daten zeigt, dass der durchschnittliche BMI mit 25,99 den Bereich des Übergewichts überschreitet Außerdem sind mehr als die Hälfte der untersuchten Personen übergewichtig oder adipös. Im linearen Regressionsmodell wurde die Annahme formuliert, dass eine geringere Aktivität zu einem höheren BMI führt. Der R^2-Wert von 0,001 zeigt jedoch, dass die erklärte Varianz minimal ist und die Korrelation somit äußerst schwach ausfällt. Der p-Wert für „Activity“ liegt bei 0,23 und übersteigt die Signifikanzschwelle von 0,05, weshalb die Nullhypothese, dass „Activity“ keinen Einfluss auf den BMI hat, nicht abgelehnt werden kann. Der Koeffizient weist

darauf hin, dass ein Anstieg von „Activity“ um eine Einheit den BMI um durchschnittlich 0,08 senkt, jedoch ist dieser Effekt statistisch nicht signifikant.
Wird das Modell um Variablen wie „Age“, „Height“, „Weight“ und „FoodConsumption“ erweitert, steigt der R^2-Wert auf 0,99, was auf eine sehr starke Korrelation hinweist. Gleichzeitig liegen die p-Werte aller Variablen unter 0,05, wodurch die statistische Signifikanz des Modells bestätigt wird. Insbesondere tragen „Height“ und „Weight“ entscheidend zur Modellgüte bei, was nachvollziehbar ist, da der BMI direkt aus die-sen Variablen berechnet wird. Sobald „Height“ und „Weight“ aus dem Modell ausgeschlossen werden, verschlechtert sich die Modellgüte. Mit einem R^2-Wert von 0,43 ist die Korrelation jedoch weiterhin stark. Außerdem bleiben „Age“, „Activity“ und „FoodConsumption“ signifikante Prädiktoren (p-Wert jeweils unter 0,05).
Der Modellvergleich mittels ANOVA zeigt zudem, dass das Modell dem R^2-Wert von 0,99 statistisch signifikant besser ist als das Modell mit dem R^2-Wert von 0,43, welches wiederum statistisch signifikant besser ist als das Modell mit dem R^2-Wert von 0,001. Die Haupter-kenntnisse unterstreichen, dass die Modellgüte stark von der Wahl der Variablen ab-hängt.
Der Vergleich mit der Frequenzbaseline dient als wichtige Referenz zur Beurteilung der Modelllleistung. Eine Frequenzbaseline stellt ein einfaches Modell dar, das lediglich die häufigste Klasse („Normalgewicht“, F = 0,41) vorhersagt. Mit F-Scores von 0,67 (SVM), 0,86 (Random Forest) und 0,68 (Logistic Regression) können die Modelle relevante Muster in den Daten identifizieren und sind somit in der Lage, differenzierte Vorhersagen zu treffen. Weiterhin wurde das Modell Logistic Regression mit einer Lernkurve und LIME genauer analysiert. Bei der Lernkurvenanalyse verliefen die Trainings- und Validierungskurve nahezu parallel und stabilisierten sich auf einem ähnlichen Niveau von etwa 0,7. Dies zeigt, dass das Modell weder zu Over- noch Underfitting neigt. Die Modellinterpretation mit LIME ergab Einblicke in die einflussreichsten Merkmale. Bei der Vorhersage der Klasse „Normalgewicht“ (Vorhersagewahrscheinlichkeit 95,6%) hatte das Alter den stärksten positiven Einfluss (+58,1%), gefolgt von dem Geschlecht, der Körpergröße und der Häufigkeit des Gemüseverzehrs. Für die Klasse „Übergewicht“ hatte das Alter hingegen einen negativen Einfluss (-42,8%), während die Menge an täglichen Hauptmahlzeiten und die Körpergröße geringe positive Beiträge leisteten.

Ein wiederkehrendes Muster in der Analyse war, dass das Alter in fast allen Vorhersagen den stärksten Einfluss hatte. Jüngere Personen wiesen tendenziell einen geringeren BMI auf, was im Modell klar reflektiert wurde. Dies macht das Alter zu einem entscheidenden Faktor in der Vorhersage und unterstreicht dessen zentrale Bedeutung im Datensatz.

5. Haupterkenntnisse aus der Umsetzung

Der Datensatz umfasst insgesamt 16 Dimensionen und stellt auf den ersten Blick eine solide Basis für die Analyse der Probanden dar. Die hohe Anzahl an Datensätzen (1610 Werte und die Varianz der Datensätze ermöglicht aussagekräftige Ana-lysen.
Im Laufe der Zeit traten jedoch Einschränkungen auf, die komplexere Analysen erschwerten und Anpassungen erforderlich machten.
Ein zentrales Problem war die unzureichende Kategorisierung einiger Dimensionen, zum Beispiel Gewichtsangaben. Der Datensatz war lediglich in vier Gewichtsklassen unterteilt, was eine präzise und differenzierte Analyse erheblich erschwerte. Insbesondere bei der Clusteranalyse, die auf detaillierteren Daten basiert, erwies sich diese grobe Einteilung als hinderlich. Um diesen Mangel zu beheben, wurden die Gewichtsdaten überarbeitet und um genauere Werte ergänzt.
Im Rahmen des Feature-Engineerings wurden vier zusätzliche Dimensionen eingeführt: Körpergewicht (Weight), BMI, körperliche Aktivität (Activity) und Ernährungs-gewohnheiten (Food Consumption). Diese Erweiterungen ermöglichten eine detailliertere Betrachtung der Probanden und trugen dazu bei, Zusammenhänge zwischen verschiedenen Faktoren besser zu erkennen. Die Werte wurden auf Basis der vorhandenen Daten berechnet und die detaillierte Vorgehensweise ist im zugehörigen Feature-Engineering-Dokument beschrieben. Diese neuen Dimensionen verbesserten die Qualität der Analyse erheblich, insbesondere bei komplexeren Verfahren wie der Clusteranalyse, da sie die Identifizierung aussagekräftiger Muster erleichterten.
Die Vielzahl der Dimensionen brachte jedoch auch Herausforderungen mit sich. Für spezifische Analysen mussten die relevantesten Dimensionen ausgewählt werden, was eine sorgfältige Abwägung in Bezug auf Qualität und Quantität erforderte. Bei der K-Means-Clusterung wurden zum Beispiel verschiedene Dimensionen getestet und die Ergebnisse verglichen. Wichtig war es hier nicht nur die Ergebnisse sicherzustellen, sondern auch die Anschaulichkeit zu gewährleisten. Da das 3D Modell et-was unübersichtlich sein kann, wurde sich hier entschieden zusätzlich zwei 2D Mo-delle einzubauen.
Für die Regression wurden aussagekräftige Variablen ausgewählt: Age, Height, Weight, BMI, FoodConsumption und Activity. Andere Spalten des Datensatzes wurden ausgeschlossen, da sie für die Modellierung und Analyse weniger relevant waren. Nach der Erstellung des linearen Regressionsmodells (LR) konnten die deskriptiven Statistiken interpretiert werden, was mit den vorliegenden Daten sehr gut funktionierte.

Das LR-Modell wurde anschließend interpretiert, wobei eine zentrale Hypothese aufgestellt wurde: Geringe körperliche Aktivität führt zu einem höheren BMI. Diese Hypothese wurde anhand von drei Schlüsselmaßen überprüft: P-Wert, R^2-Wert und Koeffizienten. Diese Metriken wurden auch verwendet, um die Qualität des Modells zu bewerten. Mit den generierten Daten war die Überprüfung einfach und effektiv. Im nächsten Schritt wurden weitere Abhängigkeiten zu den Hypothesen hinzugefügt, darunter FoodConsumption, Age, Weight und Height. Es zeigte sich eine starke Kor-relation, die jedoch höchstwahrscheinlich auf die Abhängigkeit von Weight und Height zurückzuführen ist, da diese Variablen den BMI direkt beeinflussen. Für ein aussagekräftiges Modell mussten sie außer Betracht gezogen werden.

6. Fazit

Abschließend kann festgestellt werden, dass der verwendete Datensatz sowie die im Laufe des Projektes vorgenommenen Erweiterung eine Vielzahl aussagekräftiger Analysen ermöglichte. Bei der Untersuchung der verschiedenen Themenbereiche konnten relevante Zusammenhänge zwischen den Daten identifiziert werden. Es zeigte sich, dass mit zunehmendem Alter und geringerer körperlicher Aktivität das Risiko für Übergewicht steigt. Die Clusteranalyse ergab zudem, dass bestimmte Variablen, wie etwa der Lebensmittelkonsum, mit dem Körpergewicht korrelieren. In vielen Fällen führt ein höherer Lebensmittelkonsum zu einem erhöhten Gewicht.

Die größte Einschränkung des Projekts war jedoch der Ausgangsdatensatz, der keine genauen Gewichtsangaben enthielt. Stattdessen standen nur grobe Gewichtsklassen zur Verfügung, was die Genauigkeit der Analyse einschränkte. Obwohl die Gewichtsdaten nachträglich geschätzt wurden, sind diese Schätzungen nicht so genau wie die tatsächlichen Daten. Maschinelles Lernen und andere komplexe Algorithmen wurden erfolgreich eingesetzt, aber es konnten keine verlässlichen Ergebnisse für eine vollumfängliche und genaue Risikoabschätzung für Adipositas erzielt werden.

Die Ergebnisse dieses Projekts zeigen, dass ein detaillierterer Datensatz und der umfassende Einsatz fortgeschrittener Algorithmen in zukünftigen Analysen zu genaueren und aussagekräftigeren Ergebnissen führen können. Trotz dieser Einschränkungen bieten die gewonnenen Daten bereits wertvolle Einblicke in die vielfältigen Einflussfaktoren auf das Körpergewicht. Darauf aufbauend könnte ein Präventionsprogramm entwickelt werden, das gezielte Handlungsstrategien zur Bekämpfung der Adipositas aufzeigt.

Literaturverzeichnis

Carlberg, Carsten (2023): Fettleibigkeit und Diabetes. In: Carsten Carlberg (Hg.): Die molekulare Basis von Gesundheit: Wie Epigenetik und Ernährung unser Leben beeinflussen. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 157–196.

Koklu, Nigmet; Sulak, Süleyman Alpaslan (2024): Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals' Social and Physical Activities. In: *Sinop Üniversitesi Fen Bilimleri Dergisi* 9 (1), S. 217–239. DOI: 10.33484/sinopfbd.1445215.

Phelps, Nowell H.; Singleton, Rosie K.; Zhou, Bin; Heap, Rachel A.; Mishra, Anu; Bennett, James E. et al. (2024): Worldwide trends in underweight and obesity from 1990 to 2022: a pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. In: *The Lancet* 403 (10431), S. 1027–1050. DOI: 10.1016/S0140-6736(23)02750-2.

World Health Organization (2022): Obesity in the WHO European region factsheet. Unter Mitarbeit von Nutrition, Physical Activity & Obesity (NAO), Office for Prevention & Control of NCDs (MOS). Hg. v. World Health Organization.

Anhang

Gemeinschaftlicher Beitrag

Im Rahmen der Projektarbeit haben wir gemeinsam die gesamten Aufgaben bis zur Datentransformation bearbeitet. Der Fokus lag dabei auf der Datenaufbereitung und -exploration, einschließlich des Ladens, Bereinigens und der Erstellung erster Visualisierungen wie Line-Charts und Scatter-Plots zur besseren Datenanalyse. Diese Schritte dienten uns als Grundlage für die weiteren Analysen.
Das Schreiben des Berichts am Ende des Projekts geschah ebenfalls in Zusammenarbeit.

Individueller Beitrag von Sera Can

Die Aufgaben zur Regression sowie deren Evaluation und Interpretation habe ich eigenständig durchgeführt. Dabei habe ich eine lineare Regressionsanalyse angewendet, um den Zusammenhang zwischen Aktivitätsniveau und BMI zu untersuchen. Zur Bewertung der Modellgüte habe ich statistische Kennzahlen wie R^2-Werte, p-Werte und Koeffizienten berechnet und interpretiert. Anschließend habe ich das Modell um Variablen wie Alter, Größe, Gewicht und Ernährungsverhalten erweitert und Modellvergleiche zur Analyse der verbesserten Güte durchgeführt. Zusätzlich habe ich mittels ANOVA die Verbesserungen der Modellgüte überprüft.

Des Weiteren habe ich mich mit der Bewertung eines Klassifikationsmodells zur Vorhersage von Gewichtskategorien beschäftigt. Zunächst habe ich die Frequenzbaseline definiert, die als einfaches Modell die häufigste Klasse im Datensatz vorhersagt, und habe diese als Vergleich für die Leistungsfähigkeit meines Modells aufgestellt.

Ich habe die Leistung verschiedener Klassifikationsmodelle, darunter Support Vector Machines (SVM), Random Forest und logistische Regression, anhand des F-Scores und der Lernkurven verglichen. Die Lernkurvenanalyse habe gezeigt, dass das Modell weder zu Overfitting noch zu Underfitting neige, was auf eine ausgewogene Modellkomplexität hinwies.

Abschließend habe ich LIME zur Interpretation des Modells genutzt, um die wichtigsten Einflussgrößen auf die Vorhersagen zu identifizieren und zu verstehen, wie Merkmale wie Alter, Größe und Ernährungsgewohnheiten die Vorhersage des Modells beeinflussten.

Individueller Beitrag von Dorian Schimmel

Die Clusterung und die Dimensionsreduktion habe ich bearbeitet. Bei der Clusterung habe ich mich auf die effiziente Implementierung des K-Means-Algorithmus konzentriert. Die ersten Ergebnisse mit den Originaldaten waren jedoch nicht zufriedenstellend, da die Werte nur eine geringe Varianz aufwiesen. Nachdem das Kapitel Feature Engineering abgeschlossen war, standen nun Dimensionen mit größerer Skalierbarkeit zur Verfügung. Aus diesem Grund führte ich das Clustering mit drei der neuen Dimensionen durch, was zu deutlich besseren Ergebnissen führte. Nach der Implementierung habe ich verschiedene Parameter und Werte angepasst, um sowohl die optimale Anzahl der Cluster als auch die notwendigen Iterationen zu bestimmen. Für die erneute Visualisierung habe ich nach Rücksprache mit Michael Held neben dem 3D-Modell auch zwei 2D-Modelle integriert, um eine bessere Übersichtlichkeit der Cluster zu gewährleisten. Für die Dimensionalitätsreduktion entschied ich mich für den Einsatz der Principal Component Analysis (PCA). Nach der erfolgreichen Implementierung experimentierte ich mit den Hauptkomponenten, um ein aussagekräftiges Ergebnis zu erhalten. Ich wählte einen Schwellenwert von 85% erklärter Varianz und integrierte diese Entscheidung in den Prozess mit Hilfe einer Pipeline. Zur besseren Visualisierung habe ich dann ein Diagramm erstellt, das die Projektion der Daten auf die ersten beiden Hauptkomponenten zeigt, wobei die Klassen farblich dargestellt werden, um einen möglichen Zusammenhang zwischen den Daten zu erkennen. Da dieses Diagramm jedoch nur ca. 36% der Gesamtvarianz darstellt, dient es hauptsächlich der Veranschaulichung und ist nicht repräsentativ für die gesamte Datenstruktur. Außerdem habe ich bei der Umsetzung der Anforderung Klassifikation geholfen. Dies betraf vor allem den allgemeinen Ansatz und die Auswahl der Trainingsdaten.

Individueller Beitrag von Michael Held

Bei der Bearbeitung der ersten Aufgaben fiel auf, dass die aus den Daten resultierenden Diagramme nur eine unbefriedigende Menge an Informationen wiedergeben konnten. Dies lag vor allem an der eingeschränkten Skalierung der einzelnen Features, welche mit Ausnahme der demographischen Daten nur ganzzahlige Werte zwischen 1 bis maximal 5 enthielten. Aufgrund der Natur der Daten (Anonyme Onlinebefragung) bestand keine Möglichkeit die Daten mit externen Daten anzureichern. Demensprechend habe ich aufgrund der Gewichtsklassifikationen die BMI-Werte sowie ein darauf basierendes Gewicht extrahiert. Die Skalierung der Werte für Aktivitäten und Essensgewohnheiten konnte erreicht werden, indem die einzelnen dazu zugehörigen Features aufgrund von eigenen Erfahrungswerten aufeinander addiert oder voneinander subtrahiert worden sind.

Zusätzlich habe ich im Rahmen der Clusterung bei der Visualisierung und der Analyse der Cluster geholfen.

Um den besten Klassifikationsalgorithmus zu bestimmen, habe ich die Modelle SVM, Random Forest, Logistic Regression und XGBoost getestet. Ich habe eine 80/20 Aufteilung der Daten vorgenommen, um Overfitting zu vermeiden und eine fünffache Kreuzvalidierung durchgeführt. Die Merkmalsauswahl wurde mit SelectFromModel und einer linearen SVM durchgeführt. Für jedes Modell identifizierte ich die wichtigsten Merkmale und bewertete ihre Leistung. Schließlich analysierte ich die Ergebnisse in Bezug auf Genauigkeit, F1-Score und Recall, wobei ich auf die Schwächen des Recalls einging.

Mein Beitrag umfasste zusätzlich die Zusammenstellung des Jupyter-Notebooks. Außerdem bin ich für die Qualitätssicherung bis zu den Klassifikationsaufgaben verantwortlich, indem ich sichergestellt habe, dass die verwendeten und dargestellten Daten logisch konsistent und interpretierbar sind, um aussagekräftige Ergebnisse zu gewährleisten.