OSTBAYERISCHE
TECHNISCHE HOCHSCHULE
REGENSBURG

Projektarbeit im Rahmen der DDMI Vorlesung von Prof. Dr. Schicker

Data Mining Projekt zur Analyse von Essgewohnheiten, Bewegungsmustern und dem den Body Mass Index von Menschen in Südamerika

Name: Magnus Jachnik
Matrikelnummer: 3177039
Datum: 14.01.2024

Inhaltsverzeichnis

1. Abstract

Der vorliegende Bericht stellt eine Dokumentation eines Data Mining Projekts im Rahmen der Vorlesung DDMI von Prof. Dr. Schicker im WiSe 23/24 an der OTH Regensburg dar. Der bearbeitete Datensatz lautet „Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico“.

Es wurden die Supervised Learning Methoden Entscheidungsbaum, Random Forest und darüber hinaus ein neuronales Netz verwendet, um in Fettleibigkeitsstufen zu klassifizieren. Außerdem wurde analysiert welche Features die höchste positive Korrelation bezogen auf die jeweiligen Klassen besitzen, um so Gewohnheiten zu finden, die eine große Auswirkung auf die Fettleibigkeitsstufe der Menschen hat. Durch die Implementierung dieser Methoden konnten Genauigkeiten von bis zu 0,70 und Phi-Koeffizienten bis 0,84 erzielt werden.

2. Beschreibung des Datensatzes

Der verwendete Datensatz für das Data-Mining-Projekt stammt aus einer Studie, die in den Ländern Mexiko, Peru und Kolumbien durchgeführt wurde. Die Datensammlung erfolgte über eine Webplattform, auf der anonyme Nutzer eine Umfrage zu ihren Gewohnheiten und beantworteten und Angaben zu ihrem Körper bereitstellten. Insgesamt umfasst der Datensatz 17 Features und 2111 Instances.

Ein großer Teil der Umfrage machen die Essgewohnheiten aus. Dazu gehört der häufige Verzehr von kalorienreichen Lebensmitteln, die Häufigkeit des Verzehrs von Gemüse, die Anzahl der Hauptmahlzeiten, der Verzehr von Nahrung zwischen den Mahlzeiten, die Menge von täglich getrunkenem Wasser und der Konsum von Alkohol und Zigaretten.

Des Weiteren gaben die Teilnehmer an, ob sie ihre Kalorienzufuhr Überwachen, die Häufigkeit ihrer körperlichen Aktivität pro Woche, die Nutzung von Technologiegeräten in Stunden am Tag, die Art des Fortbewegungsmittels im Alltag und ob sie Verwandte mit Übergewicht haben. Die körperlichen Daten sind Geschlecht, Alter, Größe und Gewicht.

Damit der Datensatz zur Klassifikation verwendet werden kann, wurden die Kategorie „NObesity“ (obesity_levels) erstellt. Dabei wurden die Menschen anhand Richtlinien der Weltgesundheitsorganisation in sieben Kategorien eingeteilt: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II und Obesity Type III.

77% der Daten wurden synthetisch generiert, damit die sieben Klassen annähernd gleichverteilt im Datensatz vorkommen. Diese Maßnahme stellt sicher, dass Data Mining Verfahren die Klassen gleichermaßen effektiv erlernen können. (vgl. de la Hoz Manotas & Mendoza Palechor, 2019)

3. Data Mining Vorgehensweise

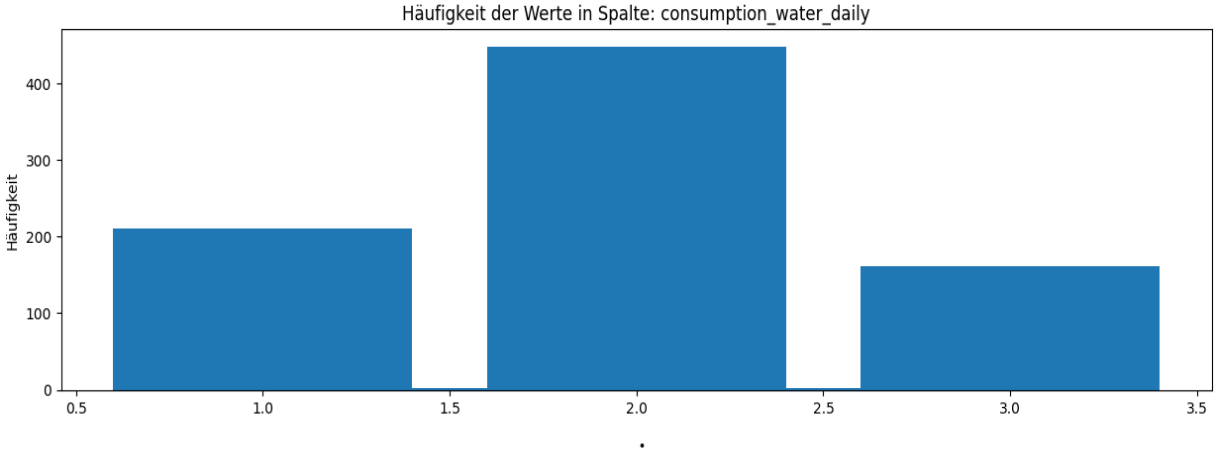
3.1 Aufbereitung und Bewertung des Datensatzes

Zu Beginn des Projekts wurde der Datensatz aus eine CSV-Datei geladen und mithilfe der Bibliothek „pandas“ als DataFrame in Python eingelesen. Um einen ersten Einblick in die Struktur und die Vollständigkeit der Daten zu erlangen, wurden die ersten fünf Einträge ausgegeben.

Da die im späteren Verlauf das Label „obesity_levels“ anhand von Verhaltensweisen vorhergesagt werden soll, wurden alle Features entfernt, die körperliche Daten enthalten oder keine Informationen über Gewohnheiten liefern. Der Datensatz beinhaltet dadurch elf Features. Anschließend erfolgte eine Umbenennung der verbleibenden Spalten, um die Lesbarkeit dieser zu verbessern.

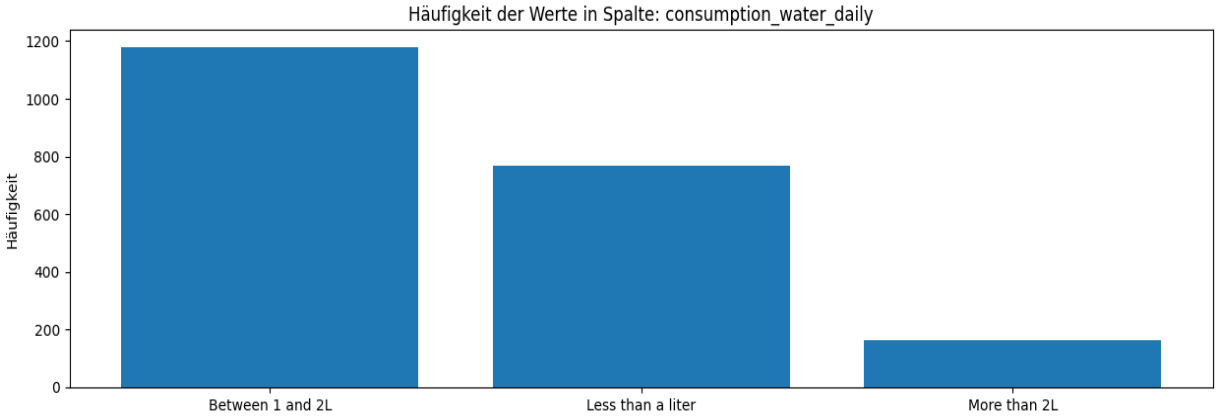
Als Nächstes wurde für jede Spalte ein Balkendiagramm erstellt, das die Häufigkeiten der vorkommenden Inhalte visualisiert. Das ermöglicht einen tieferen Einblick über die Verteilung, Korrektheit und Vollständigkeit der Daten.

Beim Analysieren der Diagramme fällt auf, dass einzelne Spalten eine Transformation der Daten erfordern.



So befinden sich beispielsweise die Inhalte der Spalte „consumption_water_daily“ im Float Format, anstatt der erwarteten kategorischen Werte. Außerdem wirken die Daten in dieser Spalte unvollständig, da die Summe der auftretenden Häufigkeiten niedriger als die Anzahl der Instanzen des Datensatzes zu sein scheint.

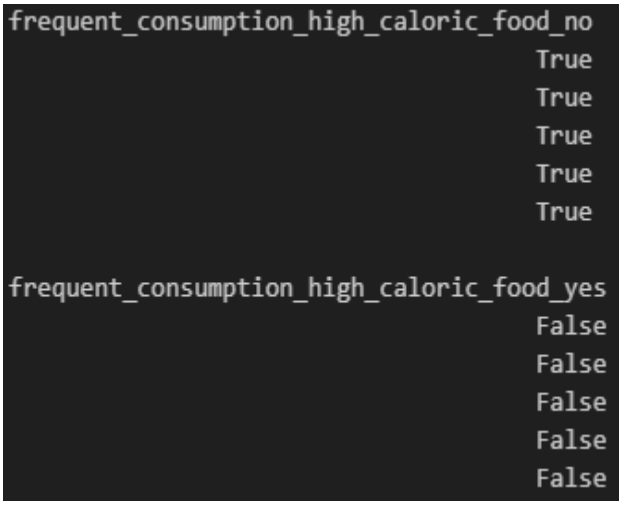
Es ergibt sich ein weiteres Diagramm nach der Umwandlung der Inhalte in Integer Variablen und der anschließenden Transformation dieser in die zugehörigen kategorischen Werte. Hier befanden sich die Inhalte der Spalte fälschlicherweise in der Einheit Liter am Tag.



Die Werte entsprechen jetzt gültigen Antwortmöglichkeiten der Umfrage, auf der der Datensatz beruht. Durch die neue Darstellung scheint es nicht mehr, als würden Instanzen mit fehlenden Einträgen in der Spalte „consumption_water_daily“ existieren. Dies wurde anschließend bestätigt, indem die CSV-Datei ausgegeben wurde und die Werte händisch auf Vollständigkeit geprüft wurden.

3.2 Untersuchung mit Learning Methoden

Zuerst wurde eine Assoziationsanalyse durchgeführt, indem der Phi-Koeffizient bezogen auf die sieben Ausprägungen der „obesity_levels“ für jedes Feature berechnet wurde. Dafür wurde der Datensatz erneut transformiert. Durch das One-Hot-Encoding wurde jeder kategorischer Inhalt jedes Features zu einer eigenen Spalte des DataFrames. Dadurch wurde die Anzahl der Spalten auf 42 erhöht.



Zum Beispiel ergaben sich aus der Spalte „frequent_consumption_high_caloric_food“ und ihren möglichen Inhalten „yes“ und „no“ zwei neue Spalten „frequent_consumption_high_caloric_food_no“ und „frequent_consumption_high_caloric_food_yes“ mit boolschen Werten.

Um den Datensatz mit Supervised Learning Methoden zu untersuchen, wurden die Verfahren „Entscheidungsbaum“ und „Random Forest“ angewendet. Die Verfahren sollten mit einer möglichst hohen Genauigkeit die „obesity_levels“ anhand der Verhaltensweisen der Menschen vorhersagen. Die Verfahren nutzen hierbei ebenfalls den Datensatz, der mittels One-Hot-Encoding erstellt wurde. Nachdem der Datensatz in Trainings- und Testdaten aufgeteilt wurde, konnte das Entscheidungsbaum- und Random Forest Modell erstellt werden. Dies wurde zunächst mit Standardparametern durchgeführt. Danach wurde die Genauigkeit der Modelle ausgegeben.

Im Anschluss sollte die Genauigkeit verbessert werden und Overfitting vermieden werden, indem die Hyperparameter der Modelle verändert wurden. So wurde die maximale Tiefe der Bäume bei beiden Verfahren auf 25 begrenzt.

Zusätzlich wurden die Parameter „n_estimators = 50“, „min_samples_split = 6“ und „min_samples_leaf = 2“ im Random Forest Modell angewandt, da diese Werte die höchsten Genauigkeiten erreichten.

Danach wurde eine Feature Selection durchgeführt, die ebenfalls sowohl die Genauigkeit verbessern als auch Overfitting vermeiden sollte. Hierfür wurde ein weiteres Entscheidungsbaummodell mit Standardparametern erstellt, das die Features mit der höchsten Wichtigkeit für den Entscheidungsprozess des Modells ausgab. Anschließend wurde die beiden Modelle auf ein DataFrame angewandt, das lediglich diese neun wichtigsten Features beinhaltete. Durch die Begrenzung war es erstmals sinnvoll das Entscheidungsbaummodell zu plotten, um eine visuelle Übersicht über die Berechnungen des Baumes zu erlangen. Allerdings ist der Baum weiterhin komplex, sodass die einzelnen Zwischenschritte auf Grund von schlechter Lesbarkeit nicht nachvollziehbar sind.

Das Supervised Learning wurde abgeschlossen, indem die Stabilität der beiden Modelle getestet wurde. Dafür wurde die Einteilung in Trainings- und Testdaten, das Trainieren der Modelle, sowie die Ausgabe der Genauigkeiten als Funktion programmiert. Beim Aufrufen der Funktion muss ein Wert für den Parameter „random_state“ übergeben werden, der dafür zuständig ist, die Daten auf bestimmte Art und Weise in Trainings- und Testdaten aufzuteilen. Durch das mehrmalige Aufrufen der Funktion mit unterschiedlichen, übergebenen Werten, können die Modelle auf Stabilität getestet werden.

3.3 Neuronales Netzwerk

Das DataFrame wurde erneut transformiert, damit ein neuronales Netzwerk mit den Daten arbeiten kann. Die Funktion „LabelEncoder“ ändert die kategorischen Werte der Spalte „obesity_levels“ in Zahlenwerte von null bis sieben.

Danach wurden die Spalten des DataFrame als Feature oder Label definiert und in Trainings- und Testdaten aufgeteilt. Das Normalisieren der Daten ist nicht notwendig, da der Datensatz aus boolschen Werten besteht und damit gut vergleichbar ist.

Das Neuronale Netzwerk wurde mit der TensorFlow-Keras API erstellt. Es besteht aus einer Eingangsschicht mit 64 Neuronen und ReLU-Aktivierung, einer versteckten Schicht mit 32 Neuronen und ReLU-Aktivierung sowie einer Ausgangsschicht mit sieben Neuronen und der Softmax-Aktivierungsfunktion. Diese sieben Neuronen sind notwendig, um die sieben möglichen unterschiedlichen Werte für „obesity_levels“ darzustellen.

Das Modell wurde mit der Categorical Crossentropy als Verlustfunktion und dem Adam-Optimizer kompiliert. Die Daten wurden über 50 Epochen mit einer Batch-Größe von 32 und einer Validierungssplitrate von 20% trainiert.

Um die Ergebnisse zu visualisieren, wurde der Loss und die Genauigkeit auf den Testdaten ausgegeben. Außerdem wurden die Lernkurven des neuronalen Netzwerks in zwei Plots dargestellt: einer für den Loss und einer für die Genauigkeit während des Trainings und der Validierung.

4. Ergebnisse

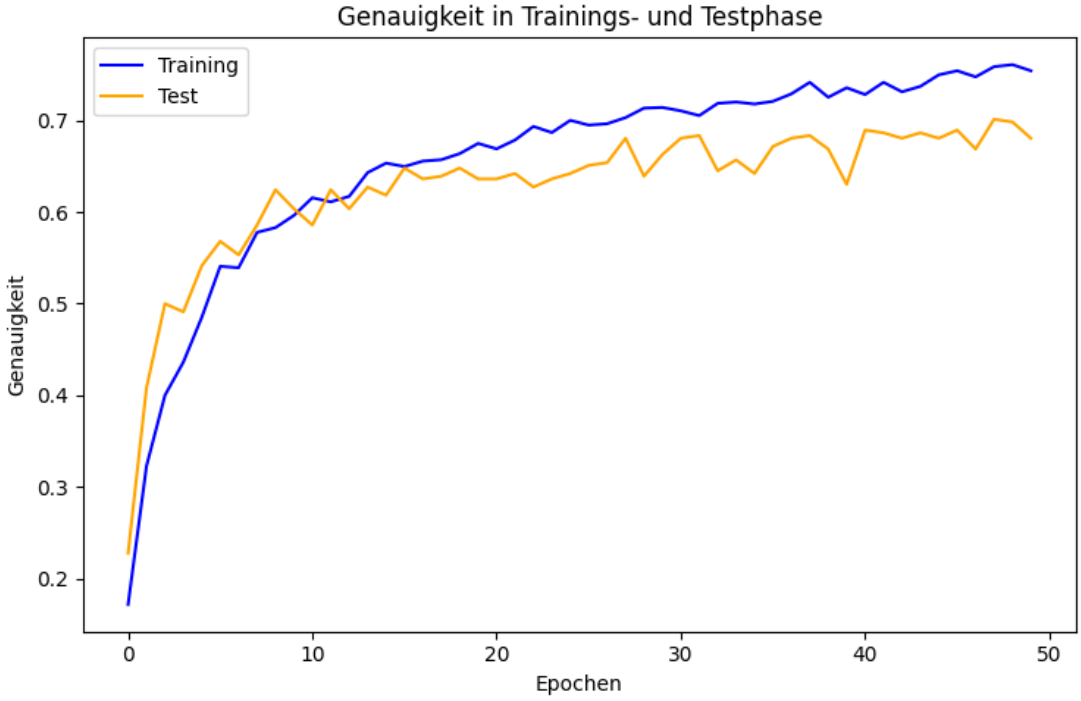
Die untersuchten Phi-Koeffizienten zeigen allgemein wenige eindeutige Zusammenhänge mit den Fettleibigkeitsklassen. Eine Ausnahme stellen die Spalten ‚frequency_consumption_vegetables_Always“ und „frequency_consumption_vegetables_Sometimes“ dar, die einen Wert von 0,84 und 0,72 bezogen auf die Klasse „Obesity_Type_III“ vorweisen. Eine weitere Auffälligkeit gibt es bei der Klasse „Insufficient_Weight“. Der Phi-Koeffizient der Spalten „consumption_food_between_meals_Frequently“ beträgt 0,58 und die Spalte „consumption_food_between_meals_Sometimes“ 0.47. Diese Werte entsprechen einer leichten poitiven Korrelation mit der Klasse Untergewicht.

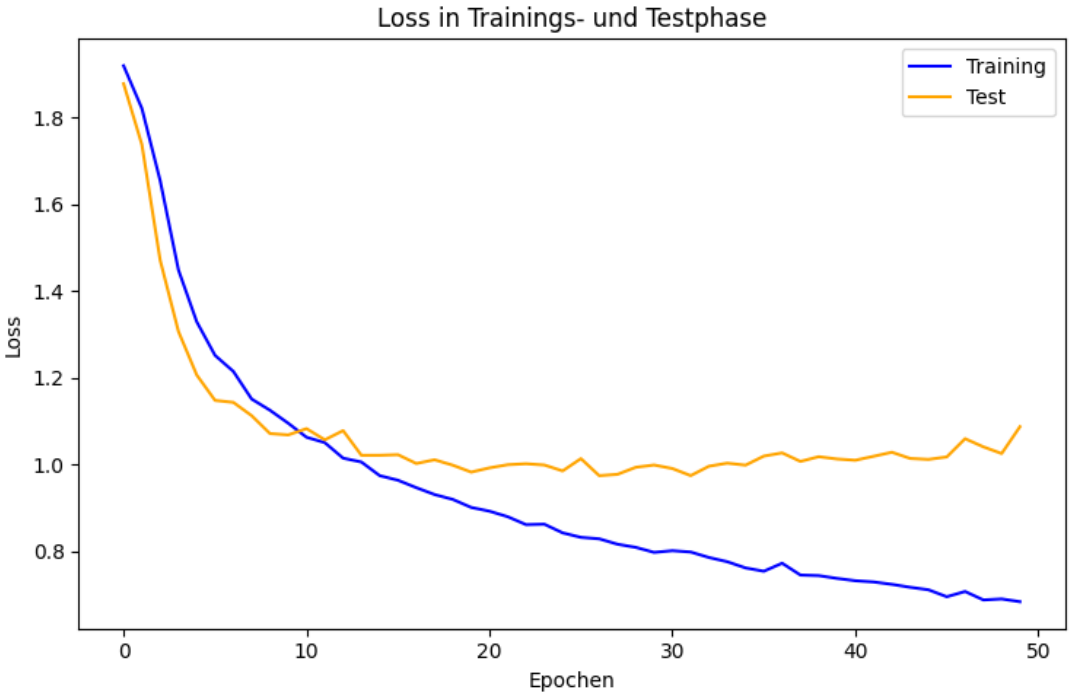
Der Entscheidungsbaum mit Standardparametern erreicht eine Genauigkeit von 0,65 und das Random Forest Modell eine Genauigkeit von 0.68. Durch das Verändern der Parameter erhöhen sich beide Genauigkeiten lediglich marginal auf ,066 und 0,70.

Durch die Feature Selection sinken die Genauigkeiten auf 0,56 (Entscheidungsbaum) und 0,55 (Random Forest).

Beide Modelle sind als stabil einzuordnen, da sich die Genauigkeiten bei Veränderungen der Trainings- und Testdaten um einen Betrag von ungefähr 0,02 von den ersten Berechnungen unterscheiden.

Das trainierte neuronale Netzwerk erreicht eine Genauigkeit von 0,63 und einen Loss von 1,13 bei 50 Epochen. Die Lernkurven der beiden Bewertungskriterien sind in den folgenden Diagrammen abgebildet.





5. Ausblick

Die Ergebnisse der Data Mining Verfahren geben Anlass zu weiteren Untersuchungen. Insbesondere die positive Korrelation von regelmäßigem Verzehr von Gemüse mit der Klasse „Obesity_III“ und das regelmäßige Essen zwischen Hauptmahlzeiten mit der Klasse „Insufficient_Weight“ sind hierbei interessante Ansätze. Auf Grund der Tatsache, dass ein Großteil der Daten synthetisch generiert sind, wären weitere Umfragen interessant, die sich diesen Zusammenhängen widmen.

Ausgewählte Features, die sich auf körperliche Daten beziehen ohne, dass sie Informationen über den Body Mass Index liefern, könnten in einem nächsten Schritt in die Trainings- und Testdaten aufgenommen werden. Danach würde untersucht werden, ob die Genauigkeiten der Modelle beispielsweise durch das Geschlecht und das Alter erhöht werden können.

Diese Arbeit und fortführende Forschungen könnten zukünftig eine Grundlage bieten, um Schüler über Verhaltensweisen aufzuklären, die bei Menschen mit Über- oder Untergewicht häufig auftreten. Dies wäre eine Möglichkeit, über die Gefahren dieser Körperbilder und die damit einhergehenden Folgeerkrankungen zu informieren und Präventionsarbeit in diesem Zusammenhang zu leisten.

6. Literaturverzeichnis

De la Hoz Manotas, A., Mendoza Palechor, F. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Science Direct.*
https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub