Apfeltalk

Ein Datenexplorationsprojekt zu der Qualität von Äpfeln

Timo Nössler
https://github.com/timonoessler/DataExplorationProject

16. April 2024

Matrikelnummer: 3561222

Inhaltsverzeichnis

1 Daten Exploration

1.1 Technische Datenanalyse

In der vorliegenden Arbeit wurden Daten über Äpfel charakterisiert. Die Charakterisierung der Daten erfolgt auf der Grundlage verschiedener statistischer und erläuternden Funktionen die auf den Datensatz angewendet werden. Dies dient dazu einen umfassenden Einblick in den Datensatz zu erhalten. Ein Överall Overviewërmöglicht einen ersten Überblick über die Daten. Hier wird ein Ausschnitt aus dem Datensatz ausgegeben (Die ersten fünf Zeilen auf ganzer Spaltenlänge, mit Werten). Ergänzt wird der Einblick durch eine "Data Summary". Hier werden Informationen über die Dimension des Datensatzes gewonnen (n x m). Die "Data Information"gibt detaillierte Informationen über die Art der Daten in jeder Spalte, einschließlich der Anzahl der Nicht-Null-Einträge (Tabelle 1, Non-Null, Dtype). Dies hilft, die Art der Daten zu verstehen und mögliche Probleme der Datenqualität zu identifizieren. Die "Data Description"liefert statistische Kennzahlen zu den numerischen Spalten des Datensatzes, wie zum Beispiel den Mittelwert, die Standardabweichung und Quartile (Tabelle 1). Dies hilft, ein besseres Verständnis für die Verteilung der Daten zu entwickeln. Zuletzt zeigt die Auswertung der "Missing Values"die Anzahl der fehlenden Werte in jeder Spalte des Datensatzes. Dies ist entscheidend, um potenzielle Datenqualitätsprobleme zu identifizieren und zu entscheiden, wie mit den fehlenden Werten umgegangen werden soll. Insgesamt bietet die Datenexploration einen fundierten Überblick über den vorliegenden Datensatz und dient als Grundlage zur weiteren Verarbeitung.

A_id	**Größe**	**Gewicht**	**Süße**	**...**	**Qualität**
Ø	-0.503015	-0.989547	-0.470479	...	—
s	1.928059	1.602507	1.943441	...	—
min	-7.151703	-7.149848	-6.894485	...	—
25%	-1.816765	-2.011770	-1.738425	...	—
50%	-0.513703	-0.984736	-0.504758	...	—
75%	0.0805525	0.030976	0.801922	...	—
max	6.406367	5.790714	6.374916	...	—
Non-Null	non-null	non-null	non-null	...	non-null
Dtype	float64	float64	float64	...	object

Tabelle 1: Zusammenfassung der Datenanalyse

1.2 Visuelle Datenanalyse

Visualisierungen werden verwendet, um die Beziehung zwischen verschiedenen Attributen von Äpfeln und deren Gesamtqualität zu untersuchen.
Der Boxplot zeigt die Verteilung der spezifischen Attribute in Bezug auf die Qualität (SSize", "Weight", etc.) der Äpfel (Abbildung 1). So wird ermöglicht einen Eindruck über die Verteilung und die Ausreißer in den Attributen

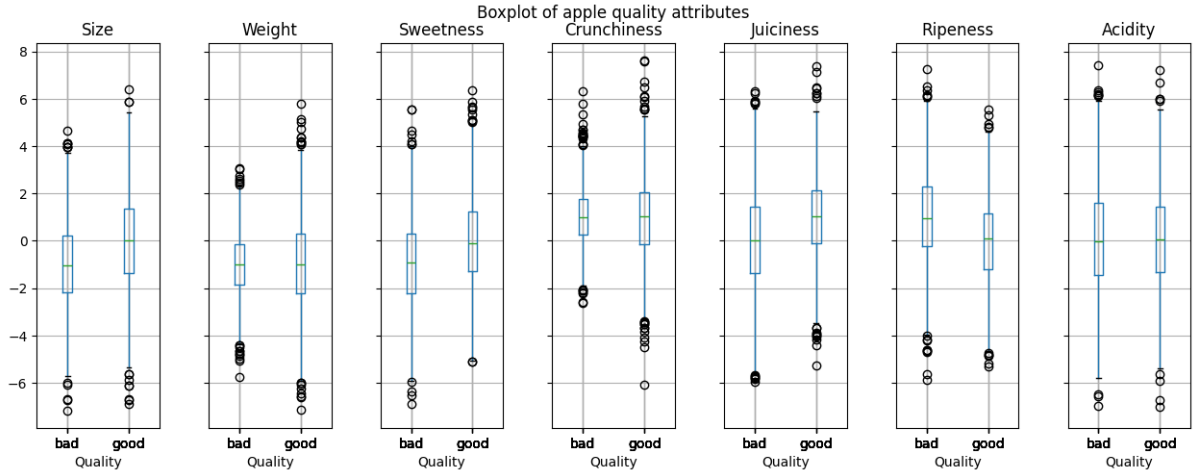

Abbildung 1: Verteilung und Ausreißer Boxplot

in Abhängigkeit von der Qualität zu visualisieren. Die Berechnung der Summe der "gutenünd ßchlechten"Werte sowie die anschließende Darstellung als Kreisdiagramm ermöglicht es, die prozentuale Verteilung von guten

und schlechten Äpfeln, in dem Datensatz darzustellen. Die Erstellung einer Heatmap in Verbindung mit einer Korrelationsmatrix gibt einen Überblick über mögliche Korrelationen zwischen den verschiedenen Attributen. Es wird gezeigt, wie stark, oder schwach, die Attribute untereinander korrelieren. Des weiteren ermöglicht es, Muster und Zusammenhänge zu identifizieren. Abschließend visualisiert der Pairplot die Beziehung zwischen den verschiedenen Merkmalen der Äpfel durch Streudiagramme und Histogramme. Durch die Farbcodierung nach Qualität können Unterschiede und Muster zwischen guten und schlechten Äpfeln identifiziert werden. Insgesamt ermöglichen diese Schritte eine umfassende Analyse der Attribute in Bezug auf die Qualität und tragen dazu bei, Muster, Zusammenhänge und Unterschiede zwischen guten und schlechten Äpfeln zu identifizieren (Abbildungen im Anhang).

1.3 Technische Merkmale

Um die technischen Merkmale des Datensatzes zu extrahieren, wurde ein *'Decision Tree Classifier'* eingesetzt. Dieser ist eine Methode im Bereich des maschinellen Lernens, das zur Klassifizierung von Daten verwendet wird. Es basiert auf dem Prinzip der rekursiven Partitionierung, bei der der Datensatz in immer kleinere Teilmengen aufgeteilt wird, um Entscheidungsregeln abzuleiten.
Die Überprüfung der Wichtigkeit einzelner Werte mittels eines *'Decision Tree Classifiers'* (Tabelle 2) ermöglicht es, die Beiträge der verschiedenen Merkmale zur Klassifikation zu verstehen und diejenigen Merkmale zu identifizieren, die entscheidend für die Klassifikation sind. Dies wiederum kann dazu beitragen, die Modellinterpretation zu verbessern.

Attribut	**Beschreibung**
Größe	22%
Gewicht	11%
Süße	13%
Knackigkeit	1%
Saftigkeit	19%
Reifegrad	20%
Säuregrad	13%

Tabelle 2: Einfluss der Attribute auf das Modell

1.4 Datensatz Split

Dieser Prozess wird als Datenpartitionierung bezeichnet und ist wichtig, um sicherzustellen, dass das Modell angemessen trainiert, validiert und getestet wird, ohne dass die Ergebnisse durch Overfitting oder Underfitting beeinflusst werden. Durch diese Aufteilung des Datensatzes in Trainings-, Validierungs- und Testsets können Modelle zunächst auf den Trainingsdaten trainiert und dann auf den Validierungsdaten optimiert werden. Anschließend kann die Leistung des Modells mit den Testdaten bewertet werden, um sicherzustellen, dass es auf neuen, unabhängigen Daten gut generalisiert.

1.5 Auswahl der Metriken

Um den Ruf des Apfelhändlers zu schützen und Geld zu sparen, ist es wichtig, dass das Klassifizierungsmodell einen ausgewogenen Score erhält. Denn wenn überreife Äpfel verkauft werden, schadet dies dem Ruf. Während das Wegwerfen zu vieler "guter"Äpfel zu finanziellen Verlusten führt. Auf dieser Grundlage wird in diesem Modell der F1-Score als Leitmetrik verwendet. Der F1-Score gibt den Mittelwert aus dem Recall-Wert und der Precision an. In dem hier erdachten Szenario ist es primär wichtig, den wirtschaftlichen Schaden so gering wie möglich zu halten.
Die Optimierung auf den F1-Score bedeutet, dass das Modell so eingestellt wird, dass es möglichst ausgeglichen klassifiziert. Dies geschieht durch die Anpassung der Entscheidungsgrenze des Modells, wodurch die Klassifizierungsschwelle verschoben wird. Die Validierung der Modelle wird sowohl auf dem F1-Score, wie auch auf Grundlage des ROC AUC-Wert durchgeführt.

2 Maschinelles Lernen

2.1 Auswahl und Beschreibung der Methode

Der k-Nearest Neighbor (k-NN) Algorithmus ist ein einfacher, überwachter Lernalgorithmus, der für Klassifizierungs- und Regressionsprobleme verwendet wird. Er verwendet eine Norm (z.B. $L_2Norm : ||x||_1 = \sum_{i=1}^{n} |x_i|$) zur Be-

rechnung der Distanz zwischen dem zu klassifizierenden Punkt und allen Trainingsdatenpunkten. Die k nächsten Nachbarn des zu klassifizierenden Punktes werden basierend auf der berechneten Distanz ausgewählt. Für Regressionsprobleme wird der Durchschnitt der k nächsten Nachbarn als Vorhersage verwendet $(\widehat{y}(x) = \frac{1}{k} \cdot \sum_{x_i \in N_k(x)} y_i)$.

2.2 Hyperparametertuning

Die Anpassung des Algorithmus erfolgt durch Hyperparametertuning. Zunächst wurde eine Annäherung an die Zielwerte per Ausschlussverfahren durchgeführt. Anschließend erfolgte die Feinjustierung mithilfe einer Gittersuche (GridSearchCV; Anhang 3). Dabei werden verschiedene Kombinationen von Parametern ausprobiert und anhand der ROC AUC-Metrik[1] bewertet, um die beste Kombination zu finden. Diese Technik ermöglicht eine schnelle Eingrenzung der optimalen Hyperparameter mittels eines manuell erstellten Parameterrasters. Die Suche nach den optimalen Parametern fand auf dem Trainingsdatensatz statt. Für den verworfenen XGBoost CLassifier (F1-Score 0.9) wurde ein Hyperparametertunig mit ML-Flow überwacht.

2.3 Evaluation und Ergebnisdarstellung

Das Modell wurde auf Grundlage der Trainingsdaten trainiert und optimiert. Der Validierungsdatensatz kommt bei dem Evaluierungsschritt auf einen Genauigkeitswert von 90%. Dieser Wert ist auf Grundlage der geringen Datenbasis ein annehmbarer Wert. Das Modell performt auf Grundlage der definierten Metriken gut.

2.4 Vorhersage - Demo

Die Vorhersage-Demonstration wurde auf Grundlage der zurückgehaltenen Testdaten durchgeführt. Bei der Vorhersage-Demonstration zeigt sich, dass das Modell auf neuen Daten nicht die Leistung, wie auf den Validierungsdaten, halten kann. Es entsteht ein Genauigkeitsgefälle von 4% auf 86%.

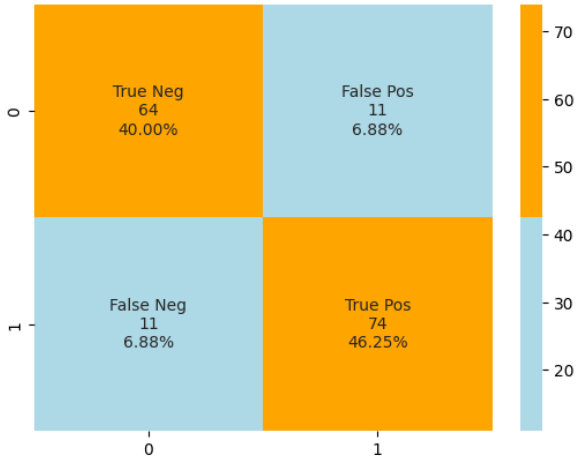

Abbildung 2: Klassifikationsmatrix der Testdaten

[1]Der ROC AUC-Wert (Receiver Operating Characteristic Area Under the Curve) ist eine Metrik zur Bewertung der Leistung eines Klassifizierungsmodells.

3 Anhang

3.1 Anhang 1

Abbildung 3: Pair Plot des Datensatz

3.2 Anhang 2

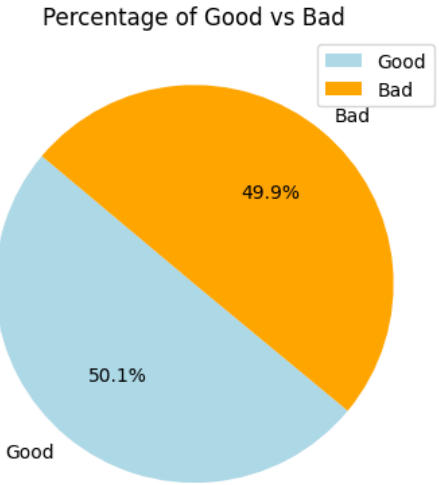

Abbildung 4: Verteilung der Qualitätsstufe

3.3 Anhang 3

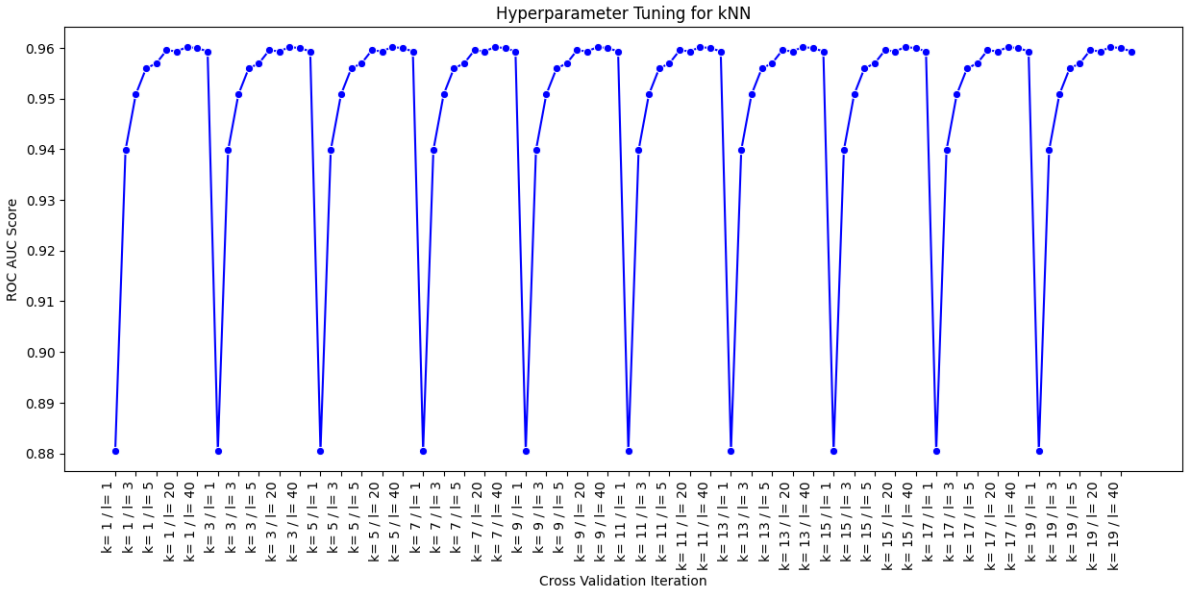

Abbildung 5: Verlauf des ROC AUC Wert kNN-Classifier