Datensatz „Fish species“

Regression:

Welche numerische Werte in deinem Datensatz, die du mit einem Regression Modell vorhersagen kannst?

Im Datensatz sind die numerischen Spalten „weight“ (Gewicht), „length“ (Länge) und „weight/length“ enthalten. Jedes dieser Attribute könnte potenziell als Zielvariable für eine Regression verwendet werden.

Welche Vorverarbeitungsschritte sind notwendig?

- **Daten sortieren**: Sortieren der Daten hilft in der Regel bei zeitlichen oder geordneten Abhängigkeiten. Bei einer Regression auf Basis einzelner Variablen könnte es jedoch vorrangig sein, Daten mit fehlenden Werten zu entfernen oder zu ergänzen, die Daten zu skalieren oder zu normalisieren, falls verschiedene Skalen vorkommen.
- **Feature Engineering**: Das Erstellen neuer Variablen, wie z.B. quadratische oder logarithmische Transformationen der Länge oder des Gewichts, könnte die Modellleistung verbessern.
- **Train-Test-Split**: Eine Unterteilung des Datensatzes in Trainings- und Testdaten ist wichtig, um das Modell auf unabhängigen Daten zu evaluieren.
- **Standardisierung/Normalisierung**: Numerische Werte sollten standardisiert oder normalisiert werden, damit das Modell nicht durch die unterschiedlichen Größenordnungen beeinflusst wird.

Wie gut funktioniert das Modell? Erläutere was die einzelnen Werte aus der Confusion Matrix bedeuten.

Bei einer Regression werden keine Confusion-Matrix-Werte erzeugt, da dies eine Klassifikationsmetrik ist. Stattdessen sind Metriken wie **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)** oder **R^2 (Bestimmtheitsmaß)** geeigneter, um die Modellleistung zu bewerten. Diese Metriken geben Auskunft über die durchschnittliche Abweichung der Vorhersagen vom tatsächlichen Wert.

Decision Tree:

Welche Klassen für die Vorhersagen gibt es in deinem Datensatz? Ist dein Datensatz ausgeglichen?

- Wenn der Datensatz zur Klassifikation verwendet wird, müssen diskrete Klassen vorhanden sein. Im aktuellen Datensatz sind jedoch nur kontinuierliche Werte wie Gewicht und Länge vorhanden.

- Eine mögliche Klassenbildung könnte durch das Diskretisieren des Gewichts oder des Verhältnisses von Gewicht zu Länge erfolgen, z.B. Einteilung in Kategorien wie „leicht“, „mittel“, „schwer“ basierend auf Gewichtsklassen.

Um zu prüfen, ob der Datensatz ausgeglichen ist, müssten wir die Häufigkeit jeder Klasse (z.B. Gewichts- oder Größenkategorien) betrachten. Ein unausgeglichener Datensatz könnte zu Verzerrungen im Entscheidungsbaum führen.

Welche Vorverarbeitungsschritte sind notwendig?

- **Diskretisierung**: Da die Werte wie Gewicht und Länge kontinuierlich sind, ist eine Diskretisierung notwendig, um sinnvolle Klassen zu erstellen.
- **Feature Scaling**: Obwohl Entscheidungsbäume unempfindlich gegenüber Skalierungen sind, ist es trotzdem ratsam, Daten konsistent zu transformieren, um bessere Interpretationen und Visualisierungen zu ermöglichen.
- **Kategorisierung**: Die Klasseneinteilung muss sinnvoll und auf den Anwendungsfall abgestimmt sein.

Wie gut funktioniert das Modell? Erläutere was die einzelnen Werte aus der Confusion Matrix bedeuten.

Für die Evaluierung eines Entscheidungsbaum-Modells wird eine **Confusion Matrix** verwendet, die die tatsächlichen und vorhergesagten Klassen vergleicht. Sie zeigt:

- **True Positives (TP)**: Anzahl der korrekt vorhergesagten positiven Klassen.
- **True Negatives (TN)**: Anzahl der korrekt vorhergesagten negativen Klassen.
- **False Positives (FP)**: Anzahl der fälschlicherweise als positiv vorhergesagten negativen Klassen.
- **False Negatives (FN)**: Anzahl der fälschlicherweise als negativ vorhergesagten positiven Klassen.

Aus der Confusion Matrix lassen sich Metriken wie **Genauigkeit (Accuracy)**, **Präzision**, **Recall** und der **F1-Score** berechnen, um das Modell umfassend zu bewerten.