Aufgabe: Modellierung von Hauspreisen mit und ohne Ausreißer

Ziel: Entwickle Modelle zur Vorhersage der Hauspreise (Spalte `MEDV`) im Boston Housing Dataset. Vergleiche die Leistung der Modelle, wenn sie mit und ohne Ausreißer trainiert wurden.

Schritte:

1. **Datensatz vorbereiten:**
    - Lade den Boston Housing Datensatz.

```
from sklearn.datasets import fetch_openml
# Boston Housing Dataset laden
boston = fetch_openml(name='boston', as_frame=True)
```

    - Überprüfe die grundlegenden Eigenschaften des Datensatzes (z. B. `describe()` usw.).
2. **Erkennung von Ausreißern:**
    - Identifiziere potenzielle Ausreißer in der Spalte `MEDV` (Hauspreise).
    - Nutze statistische Methoden wie die **IQR-Methode (Interquartilsabstand)** oder/und visuelle Methoden wie **Boxplots**, um die Ausreißer zu identifizieren.
3. **Entfernung von Ausreißern:**
    - Erstelle zwei separate Datensätze: einen mit Ausreißern und einen ohne.
4. **Datenaufteilung:**
    - Teile beide Datensätze in Trainings- und Testdaten auf (z. B. 80% Training, 20% Test).
5. **Modellierung:**
    - Trainiere beide Datensätze (mit und ohne Ausreißer) mit **linearer Regression**, **SVR** und einem **Entscheidungsbaum für Regression** zur Vorhersage von `MEDV`.
    - Evaluiere die Modelle anhand des **Root Mean Squared Error (RMSE)**.

6. **Vergleich:**
    - Vergleiche die Modelle mit und ohne Ausreißer hinsichtlich ihrer Vorhersagequalität.
    - Analysiere, wie sich Ausreißer auf die Modellleistung auswirken.

Merkmale des Boston Housing Dataset

Der Datensatz enthält **506 Beobachtungen** (Häuser) und **13 Merkmale** sowie eine Zielvariable (`MEDV`).

Merkmale (Features):

1. **CRIM**: Verbrechensrate pro Kopf nach Stadt.
2. **ZN**: Anteil der Wohngebiete, die für Grundstücke größer als 25.000 Quadratfuß zoniert sind.
3. **INDUS**: Anteil nicht-gewerblicher Geschäftsflächen pro Stadt.
4. **CHAS**: Dummy-Variable, ob das Grundstück am Charles River liegt (1 = Ja, 0 = Nein).
5. **NOX**: Konzentration von Stickstoffoxiden (in ppm).
6. **RM**: Durchschnittliche Anzahl der Räume pro Wohneinheit.
7. **AGE**: Anteil der Gebäude, die vor 1940 gebaut wurden (in Prozent).
8. **DIS**: Gewichtete Distanz zu fünf Arbeitszentren in Boston.
9. **RAD**: Index für Erreichbarkeit von Autobahnen.
10. **TAX**: Grundsteuersatz pro $10.000.
11. **PTRATIO**: Verhältnis von Schülern zu Lehrern nach Stadt.
12. **B**: 1000 $(Bk - 0.63)^2$, wobei Bk der Anteil der Bevölkerung ist, die Schwarz ist. (**bitte nicht verwenden**)
13. **LSTAT**: Prozentsatz der Bevölkerung mit niedrigem sozioökonomischen Status.

Zielvariable:

- **MEDV**: Medianwert der Häuserpreise (in $1.000).