Labo 2 Edge AI:

Deel 2 op 2: data exploratie en data preprocessing

Voor het tweede deel van het labo over data exploratie en data preprocessing in Python, zullen we de populaire Titanic-dataset gebruiken. De Titanic-dataset bevat gegevens over passagiers aan boord van de Titanic, inclusief informatie over wie de ramp heeft overleefd en wie niet. Deze dataset is goed gedocumenteerd en ideaal voor data exploratie.

Doel van de opdracht
In deze opdracht zal je leren data te verkennen, analyseren en voorbereiden voor machine learning-toepassingen met behulp van Python. Je zal ook leren omgaan met ontbrekende gegevens, categorische variabelen en feature scaling."

Stappen

Stap 1: Data verzamelen en importeren
Download de Titanic-dataset in CSV-formaat van deze locatie: Link naar de dataset
Je moet een Kaggle-account hebben om toegang te krijgen tot de dataset. Importeer de dataset in Python.

Stap 2: Data verkennen
Voer de volgende data exploratie stappen uit:

a. Bekijk de eerste rijen van de dataset om een idee te krijgen van de beschikbare variabelen.
Tip: je kan hiervoor de functie *head()* toepassen op de dataset.

b. Bereken statistieken zoals het gemiddelde, de mediaan, standaardafwijking en percentielen voor numerieke variabelen.
Tip: maak een summary van de datatypes met de functie *describe()*

c. Visualiseer de verdeling van numerieke variabelen met behulp van histogrammen en boxplots.
Tip: gebruik de passende histogrammen en plots hiervoor. Voorbeelden zijn: *histplot()* en *boxplot*.

d. Onderzoek de correlatie tussen variabelen met behulp van een correlatiematrix en een heatmap.
Tip: gebruik hiervoor de functie *heatmap*.

Tip: vergeet niet om alle visualisaties te voorzien van een titel

Stap 4: Data Preprocessing
Voer data preprocessing uit:

a. Behandel ontbrekende gegevens door middel van imputatie (bijvoorbeeld door het invullen met het gemiddelde of de modus).
Tip: hiervoor kan je bijvoorbeeld de functie *fillna()* toepassen om NA waarden te behandelen.

b. Codeer categorische variabelen in numerieke vorm (bijvoorbeeld met behulp van one-hot encoding).
Tip: gebruik de functie *dummies()* om bijvoorbeeld de variabelen *Sex* en *Embarked* te encoden.

c. Schaal numerieke variabelen indien nodig (bijvoorbeeld met behulp van z-score normalisatie of min-max schaling).

Tip: importeer *StandardScaler* van *sklearn.preprocessing* om een scaler te maken aan de hand van de functie *StandardScaler()*. Kies vervolgens 2 variabelen die je kan scalen.

Tip: laat het resultaat zien van uw dataverwerking door opnieuw de *head()* functie te gebruiken en het resultaat te printen.