**DEDS Week 10 - Machine Learning**

# **Decision Tree & Random Forest**:

Dit is het eerst gemaakte model, hierom maakt het nog geen gebruik van de later ingevoerde SQLite methodes.

A screenshot of a computer program

Description automatically generated with medium confidence

Vervolgens wordt de data opgeschoond en gesplitst in de nodige tabellen. Ook wordt er binnen de dataset een split gemaakt in de data die gebruikt wordt voor het trainen en voor het testen.

A picture containing text, font, screenshot

Description automatically generated

Het algorithme wordt getrained en de tree wordt gegenereerd

A screenshot of a computer

Description automatically generated

Om de effectieviteit van het algorithme te testen genereren we een confusion matrix

A screenshot of a computer program

Description automatically generated with medium confidence

Hieruit kunnen we opmaken dat het model niet al te best werkt.

Om deze score toch omhoog te krijgen gebruiken we een random forest, ook hiervoor genereren we een matrix.

A screenshot of a computer

Description automatically generated with medium confidence

We zien hier dat het random forest een stuk beter presteerd dan de enkele decision tree en zelfts rond de 98% accuraat is! Dit komt doordat het random forest de resultaten van meerdere descision trees combineerd waardoor overfitting binnen het model beter afgehandeld wordt.

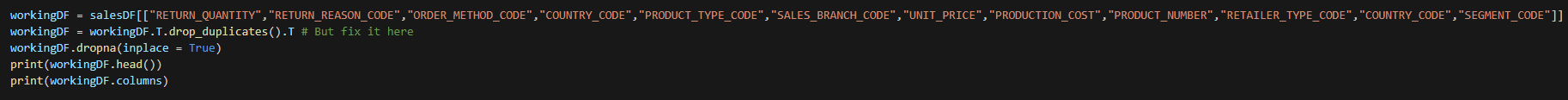
# **Linear Regression**:

Beginnend met de data import, model importeert alle data die ergens gerelateerd zou kunnen zijn aan retouren. Dit keer gebruikmakend van SQLite.

A screen shot of a computer program

Description automatically generated with low confidence

Vervolgens worden alle numerieke, relevante colommen geselecteerd en word de data opgeschoond.



Hier stopt het goede nieuws echter want na een zorgvuldige analyse van de correlatiematrix blijkt dat bijna geen enkele colom invloed uit te oefenen over de hoeveelheid retouren.

A screenshot of a computer

Description automatically generated with medium confidence

Om het model alsnog te testen, splitsen we het weer in een test en training dataset en gebruiken we deze om voorspellingen te doen op de testdata.

A screenshot of a computer program

Description automatically generated with medium confidence

Zowel de Mean Squared error als de Mean Absolute error zijn redelijk hoog. Vooral de MSE is hoog omdat deze grotere fouten in het model erger aanrekend terwijl de MAE fouten allemaal gelijk beschouwd. Een groot verschil tussen deze getallen is hierom wel te verwachten.

# Clustering