OPDRACHT 1 Regressie/Classificatie model:

In deze opdracht gaan jullie twee machine learning modellen trainen. Jullie kiezen twee algoritmes uit de lijst onderaan dit document, één om huizenprijzen mee te voorspellen en één voor het diagnosticeren van borstkanker. Jullie leveren dus twee aparte jupyter notebooks in: één met de regressie (huizenprijzen zijn een getal) en één met de classificatie methode (wel/geen borstkanker is een categorie).

Begin met het beantwoorden van een aantal vragen over het algoritme (gebruik ook plaatjes bij deze uitleg).

1. Noem een voorbeeld uit de praktijk waarin jullie algoritme wordt gebruikt.
2. Hoe werkt het algoritme conceptueel? Wat zijn de belangrijkste stappen?
3. Wat zijn de voor- en nadelen van jullie algoritme? In welke situaties werkt het heel goed en wanneer juist niet?

Nu gaan jullie het model implementeren in python. De volgorde van het python script is belangrijk. Zo zorg je bijvoorbeeld altijd dat eerst de benodigde packages worden ingeladen. Schrijf het python script daarom met een duidelijke indeling van met verschillende stappen. Bij elke stap dient toelichting worden gegeven:

1. Wat is het doel van deze stap?
2. Wat is de input van deze stap?
3. Wat is de output van deze stap en waar heb je deze output hierna nog nodig?

Regressie:
Voor de het regressie model kunnen jullie deze dataset gebruiken:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Gebruik in eerste instantie alleen het bestand train.csv. Het doel is om de prijs van huizen (variabele ***SalePrice***) te gaan voorspellen. De input die het model kan gebruiken bestaat uit de overige 80 kolommen.

- Wat voor data heb je tot je beschikking? Maak een aantal verkennende analyses om dit te verduidelijken. Welke variabelen bevatten belangrijke informatie en welke niet? Hoe zie je dat?
- Maak een enkelvoudig lineaire regressie model (dus met maar één X-variabele) net als de demo in het college. Wat is de vergelijking die hierbij hoort?
- Maak een residuen plot (de input variabele op de x-as en de fout in de voorspelling op de y-as). Wat valt je op?
- Maak nu een voorspelling met jullie gekozen machine learning model en gebruik eerst dezelfde enkele variabele.
- Voeg steeds meer variabelen toe. Wanneer ben je klaar?

Classificatie:
Voor classificatieopdracht maken jullie gebruik van de volgende dataset:
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
Jullie gaan voorspellen wie wel en wie geen diagnose borstkanker krijgt op basis van een aantal kolommen met beschikbare informatie.

- Wat voor data heb je tot je beschikking? Maak een aantal verkennende analyses om dit te verduidelijken. Welke variabelen bevatten belangrijke informatie en welke niet? Hoe zie je dat?
- Pas het gekozen machine learning algoritme toe met maar één variabele. Hoe vaak is de voorspelling correct?
- Voeg steeds meer variabelen toe. Wanneer ben je klaar?

Keuzes regressiemodellen:

1. Gradient Boosting regressie
2. Support vector machines
3. Nearest Neighbour regression
4. MLP (neural network)
5. Random Forest

Keuzes classificatiemodellen:

1. Naive Bayes
2. Random Forest
3. Gradient boosting classificatie
4. Support Vector Machine
5. Nearest Neighbours classificatie
6. MLP (neural network)