Portfolio assignment 101 & 102

Tristan Goossens (2188507)

Avans Hogeschool

Contents

[1.0 Wat is data science 2](#_Toc129640419)

[2.0 Doel 3](#_Toc129640420)

[3.0 Data vergaren 3](#_Toc129640421)

[3.1 Voorbeeld portfolio 3](#_Toc129640422)

[4.0 Databronnen integreren 4](#_Toc129640423)

[5.0 Opschonen van data 4](#_Toc129640424)

[6.0 Verkennende data analyse 4](#_Toc129640425)

[6.2 Voorbeelden portfolio 4](#_Toc129640426)

[7.0 Model trainen, valideren en verbeteren 7](#_Toc129640427)

[7.2 Voorbeeld portfolio 7](#_Toc129640428)

# 1.0 Wat is data science

Data science is een interdisciplinair vakgebied dat wetenschappelijke methoden, processen, algoritmen en systemen combineert om inzichten en kennis uit gegevens te halen. Het combineert elementen van statistiek, informatica en domeinspecifieke kennis om complexe datasets te analyseren, interpreteren en communiceren.

In het algemeen omvat data science de volgende stappen:

* Gegevensverzameling: Verzamelen van relevante gegevens uit verschillende bronnen, zoals databases, API's en sensoren.
* Gegevensvoorbereiding: Het reinigen, transformeren en voorbereiden van de gegevens voor analyse.
* Verkennende data-analyse: Het onderzoeken en visualiseren van de gegevens om patronen, correlaties en andere inzichten te identificeren.
* Modelbouw: Het creëren van statistische of machine learning-modellen om de gegevens te analyseren en voorspellingen te doen.
* Model evaluatie: Het beoordelen van de prestaties van de modellen en het verfijnen ervan indien nodig.
* Implementatie: Het implementeren van de modellen in real-world omgevingen en het gebruiken ervan om besluitvorming te informeren.

Data science wordt veel gebruikt in industrieën zoals financiën, gezondheidszorg, marketing en productie, evenals in onderzoeksgebieden zoals sociale wetenschappen en natuurwetenschappen.

In de komende hoofdstukken zal het proces die een data scientist doorloopt om zijn doel te bereiken en wat data science inhoud beschreven worden aan de hand van een aantal voorbeelden uit het portfolio.

Diagram

Description automatically generated

Figuur 1 Proces data scientist

# 2.0 Doel

Het doel van de analyse is om meer inzicht te krijgen over een dataset naar mijn keuze. Het doel dat ik gesteld heb is dat ik de verschillen wil zien in de data van een spel op laag niveau en een pot op erg hoog niveau.

# 3.0 Data vergaren

De dataset die ik voor mijn portfolio heb gebruikt heb ik binnengehaald via Kaggle. Kaggle is een website waar duizenden datasets te vinden zijn die gemaakt zijn door gebruikers van de website. ([Chess dataset](https://www.kaggle.com/datasets/datasnaek/chess)). De gekozen dataset is via Kaggle als .csv bestand gedownload

## 3.1 Voorbeeld portfolio

In figuur 2 is te zien hoe de data in een Jupyter notebook is geladen en is enkele informatie over de kolommen getoond

Graphical user interface, text, application

Description automatically generated

Figuur 2 Vergaren data

# 4.0 Databronnen integreren

De databron is zoals in kop 2.0 beschreven gedownload via Kaggle. Ook is in hoofdstuk 2 te zien hoe de databron al in een Jupyter notebook ingeladen kon worden

# 5.0 Opschonen van data

Het opschonen van data was in deze opdracht niet van toepassing. Ook had ik het geluk dat de data consistent was en er waren nergens lege kolommen te vinden. In figuur 2 is te zien hoe elke kolom evenveel ‘Non-null’ waarden had

# 6.0 Verkennende data analyse

Als onderzoeker weet je nog niet helemaal welke resultaten je gaat vinden en welke kant je onderzoek opgaat. Je gebruikt deze analyse om beter bekend te raken met je dataset.

De verschillende tools binnen een verkennende analyse zijn bijvoorbeeld:

* Scatter plot
* Bar plot
* Box plot
* Pie charts

## 6.2 Voorbeelden portfolio

Graphical user interface, application

Description automatically generated

Figuur Univariate analyse categorische data

Chart, box and whisker chart

Description automatically generated

Figuur Univariate analyse m.b.v een boxplot

A picture containing graphical user interface

Description automatically generated

Figuur Bivariate analyse m.b.v een scatterplot

Chart, scatter chart

Description automatically generated

Figuur Correlatie in een scatterplot

# 7.0 Model trainen, valideren en verbeteren

Een data scientist kan een model trainen op basis van bestaande gegevens om zo nieuwe meetwaarden te categoriseren. Het model maakt dan een ‘gok’ op basis van nauwkeurigheid

## 7.2 Voorbeeld portfolio

Graphical user interface, text, application, email

Description automatically generated

Figuur getrainde data model met evaluatie