

Taak "Machine Learning"

Project "Kijkcijfers
in Vlaanderen"



Project “kijkcijfers”: situering

Als manager van een nationale zender zou je graag zicht hebben op de factoren die de kijkcijfers beïnvloeden. Meer nog: je zou graag de kijkcijfers kunnen voorspellen, ten behoeve van bijv. adverteerders of programmamakers.

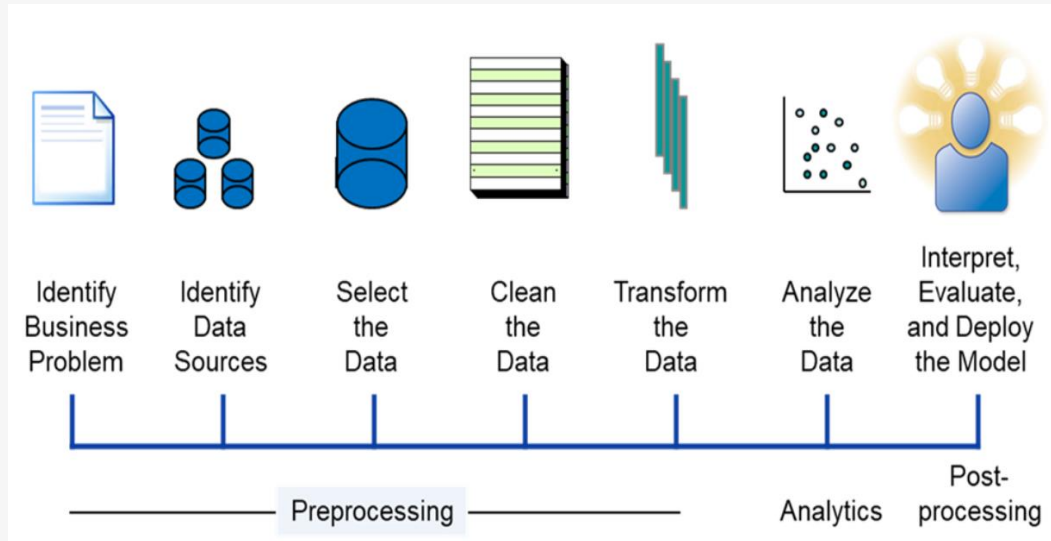
Project “kijkcijfers”: de data

- Het Centrum voor Publicatie over de Media publiceert dagelijks de top 20 van de meest bekeken Vlaamse tv-programma's, zie bijv. <https://www.cim.be/nl/televisie?type=daily&date=2025-2-2®ion=north> voor de cijfers van 2 februari 2025.
- De dagelijkse cijfers zijn ook beschikbaar via een API, bijv.: https://api.cim.be/api/cim_tv_public_results_daily_views?dateDiff=2025-02-02&reportType=north
- Je kan deze data dus via (Python-) code ophalen en opslaan.
- Beslis zelf welke data hierin relevant kunnen zijn en welke niet.
- Zoek zelf naar andere relevante, beschikbare data die de kijkcijfers zouden kunnen beïnvloeden
- Probeer met gezond verstand een onderscheid te maken tussen relevante en irrelevante data.
- Probeer zelf een aantal voorspellende “features” te construeren (*feature engineering*) uit de gekende features.
- Denk ook na of je deze features “mag” gebruiken om een goed model op te stellen, dat bijv. de kijkcijfers voor het laatavond journaal op VRT 1 van overmorgen kan voorspellen.

Taak

- Verken, zowel visueel als cijfermatig, de data met de bedoeling inzicht te verwerven in de natuur van de data.
- Probeer de data op te schonen (data cleaning!). Zorg ervoor dat foute waarden niet worden meegenomen in je analyse.
- Stel een regressiemodel op om de kijkcijfers per programma op een bepaalde zender, dag en tijdstip te voorspellen.
 - Kies een optimaal regressiemodel door alle in de les geziene modellen uit te proberen en te vergelijken volgens de "best practices".
 - Evalueer je modellen via enkele maatstaven. Kies zeker ook voor MAPE (Mean Absolute Percentage Error).
 - Visualiseer en interpreteer je resultaten.
 - Zit je voornamelijk met overschatting of onderschatting?
 - Maak een rangschikking van de factoren die het meeste bijdragen aan de kijkcijfers.
 - Verklaar eventueel waarom bepaalde modellen beter presteren dan andere indien mogelijk.
- Wekelijks mogelijkheid tot vraagstelling tijdens het vierde lesuur (vanaf week 2).
- Tijdens het project kunnen extra info of tips gegeven worden. Houd dus Chamilo in de gaten!
- Op het examen zal je je resultaten individueel presenteren aan de lector. Bij die gelegenheid krijg je een bestand met nieuwe kijkcijfergegevens. Aan de hand van een zelfgeschreven Python-applicatie maak je dan, op basis van je model voorspellingen voor de kijkcijfers. Het data formaat van de gegevens zal identiek zijn aan wat je verkrijgt via de cim API maar dan zonder de "echte" kijkcijfers. De betreffende periode zal vooraf worden meegedeeld zodat je aanvullende data kan klaarzetten. Je ontvangt dan ook een bestand met de "echte" waarden, zodat je meteen de nauwkeurigheid kan berekenen. Een voorbeeld hiervan zal ter beschikking worden gesteld.

Houd rekening met: het data mining proces



± 80-90 % van de totale projecttijd gaat naar preprocessing wegens:

- Foutieve en ontbrekende waarden.
- Data en business probleem begrijpen is meestal niet eenvoudig.
- Algoritmes zelf zijn off-the-shelf beschikbaar (meestal geen PhD nodig)
- 80 % transpiratie, 20 % inspiratie

Evaluatie

- Studiefiche:
<https://bamaflexweb.hogent.be/BMFUIDetailxOLOD.aspx?a=181008&b=5&c=1>
- Je werkt alleen.
- Je mag als voorbereiding van het project AI tools (zoals LLM's) gebruiken indien je documenteert waarvoor en hoe je deze hebt gebruikt (zie volgende slide voor meer toelichting). Dit is echter geen excuus om geen inzicht te hebben in de code.
- Presentatie van het project gedurende een mondelinge toelichting tijdens het examen.
- Tijdens het mondeling examen krijg je ook bijkomende vragen over de leerstof los van het project.
- Tweede examenkans is mogelijk: je werkt verder aan hetzelfde project

Gebruik van AI: "Full use with human oversight"

Het gebruik van AI tools is toegestaan onder volgende voorwaarden.

- Gebruik AI als assistent als je denkt dat het nodig, correct en fair is om te helpen met het uitwerken of verbeteren van jouw creatieve ideeën.
- Evalueer de output van AI-systemen kritisch.
- Werk samen met AI als “co-pilot”.
- Documenteer nauwkeurig en uitgebreid waarvoor je AI hebt gebruikt: voeg alle gebruikte prompts en chat logs toe. Door een co-pilot gegenereerde code hoeft evenwel niet tot in detail opgelijst te worden, maar zorg er wel voor dat je de gegenereerde code goed begrijpt.
- Reflecteer over het nut en de valkuilen van de gebruikte AI tools bij dit project.

De documentatie over het gebruik van AI tools staat op 15% van de punten voor het project.