Data engineering

Toelichting Practicum en opdrachten

LU2: Machine learning

Inhoud

1. Te bestuderen theorie

Literatuur 1

Titel: Machine learning Foundations – Supervised, Unsupervised, and Advanced Learning

Auteur: Taeho Jo

Uitgeverij: Springer

ISBN: 978-3-030-65899-1

Te bestuderen hoofdstukken ->

- Preface
- Part 1 Foundation
    - 1. Introduction
        - 1.1 Definition of Machine Learning
        - 1.2 Application Areas
            - 1.2.1 Classification
            - 1.2.2 regression
        - 1.3 Machine learning types
            - 1.3.1 Supervised Learning
            - 1.3.2 Unsupervised Learning
    - 4. Simple Machine Learning Algorithms
        - 4.1 Introduction
        - 4.2 Classification
            - 4.2.1 Binary Classification
            - 4.2.2 Multiple Classification
            - 4.2.3 Regression

Literatuur 2

Titel: AI-Ethics

Auteur: Paula Boddington

Uitgeverij: Springer

ISBN: 978-981-19-9381-7

Te bestuderen hoofdstukken ->

- Chapter 1: Introduction: Why AI Ethics
- Chapter 2: The Rise of AI Ethics

2. Datacamp

Module -> Data engineering – 2025-2026*

Assignment:

Associate Data scientist in Python

Stap 1: Meld je aan via de uitnodiging die je hebt ontvangen

Stap 2: Basecamp account aanmaken

Stap 3: Start with the track that Maurice van Haperen has assigned to you – Power BI -> Start (Na stap 3 zijn beide assignments acties voor de deelnemer)

Stap 4: Kies My library -> hier zie je beide courses staan -> kies continu om verder te gaan.

<*als het goed is heb je dit al gedaan aan het begin van de module, voor het uitvoeren van LU1>

2.1. Inhoud -> Associate Data Scientist in Python

1. *Introduction to Python*
2. *Intermediate Python*
3. *Project: Ivestigating Netflix Movies*
4. *Data Manipulation with Pandas*
5. *Exploring NYC Public School Test Result Scores*
6. *Joining Data with Pandas*
7. *Introduction to Statistics in Python*
8. *Introduction to Data Visualization with Matplotlib*
9. *Introduction to Data Visualization with Seaborn*
10. *Visualizing the History of Nobel Prize Winners*
11. *Introduction to Functions in Python*
12. *Python Toolbox*
13. *Data Manipulation with Python*
14. *Exploratory data Analysis in Python*
15. *Analyzing Crime in Los Angeles*
16. *Working with Categorical Data in Python*
17. *Customer Analytics: Preparing Data for Modeling*
18. *Data Communication Concepts*
19. *Introduction to Importing Data in Python*
20. *Cleaning Data in Python*
21. *Exploring Airbnb Market Trends*
22. *Working with Dates and Times in Python*
23. *Importing & Cleaning Data With Python*
24. *Writing Functions in Python*
25. *Python Programming*
26. *Introduction to Regression with statsmodels in Python*
27. *Modeling Car Insurance Claim Outcomes*
28. *Sampling in Python*
29. *Hypothesis testing in Python*
30. *Experimental Design in Python*
31. *Hypothesis Testing with Men's and Women's Soccer Matches*
32. *Supervised Learning with scikit-learn*
33. Predictive modeling for Agriculture
34. Unsupervised learning in Python
35. Clustering Antartic Penguin Species
36. Machine Learning with Tree-Based Models in Python
37. Predicting Movie Rental Durations

3. Google Colab

Colab.research.google.com

Login met een google account, heb je deze nog niet maak dan eerst een google account aan.

Google account aanmaken: accounts.google.com -> Create account

(Alternatief voor Colab is Kaggle, biedt ook de mogelijkheid voor het aanmaken van een notebook.)

4. Opdracht + Beoordeling

De uitwerking van de opdracht van de tweede leeruitkomst voor de Data engineering module wordt individueel uitgevoerd. De opdracht heeft verschillende deliverables en worden ingeleverd in Brightspace in de module Data engineering bij leeruitkomst 2.

Naast het uitvoeren van de opdrachten is er een individueel assessment (CGI, 25 minuten) aan het einde van de module. In dit assessment wordt de aangeboden theorie getoetst, daarnaast worden er vragen gesteld over de gerealiseerde deliverables.

4.1. Details Machine learning opdracht

Opdracht: Ontwerp en realiseer een Machine Learning-toepassing op basis van eigen data

4.1.1. Context

In deze opdracht ontwikkel je zelfstandig een complete Machine Learning (ML) casus, bij voorkeur gebaseerd op data afkomstig van je **werkplek**. Wanneer dat niet mogelijk is, kun je gebruikmaken van **openbare datasets** (bijv. via Kaggle, UCI Machine Learning Repository of overheid.nl/data).

Het doel is om een **praktisch toepasbare ML-oplossing** te realiseren waarin je het volledige proces doorloopt van dataverkenning tot evaluatie en reflectie.

4.1.2. Opdrachtomschrijving

Ontwerp, ontwikkel en documenteer een Machine Learning-toepassing die een realistisch probleem adresseert binnen jouw werk- of interessegebied. De oplossing dient te bestaan uit:

1. **Een dataset (voorkeur: eigen data)**
    - Zelf verzameld via werkplek of externe bron.
    - Uitgebreid met **synthetisch gegenereerde data** om de dataset te verrijken of te balanceren.
2. **Een Machine Learning-model (classificatie of regressie)**
    - Getraind op je verwerkte data.
    - Vergelijk minimaal twee verschillende modellen (bijv. Decision Tree vs. Random Forest, of Linear Regression vs. SVR).
    - Onderbouw welke het best presteert en waarom.
3. **Een softwaretoepassing of prototype**
    - Bouw een eenvoudige **applicatie of user interface** waarin gebruikers invoer kunnen geven en de **ML-voorspelling** direct kunnen zien.
    - De interface voedt het onderliggende algoritme en toont de resultaten op een begrijpelijke manier.
4. **Een begeleidend rapport (document)**
    - Beschrijft het volledige proces, van probleemkeuze tot reflectie.
    - Wordt beoordeeld op inhoudelijke kwaliteit, structuur, en naleving van rapportageregels.

4.1.3. Deliverables

1. **Python-code/notebook** waarin het volledige proces inzichtelijk is (data-invoer, training, evaluatie).
2. **Applicatie of prototype** met een eenvoudige gebruikersinterface die het ML-model aanspreekt.
3. **Rapport (PDF of Word)** conform de rapportagestandaarden.

4.1.4. Op te leveren document – Inhoudsvereisten

Je rapport bevat ten minste de volgende onderdelen:

1. **Toelichting gekozen onderwerp**
    - Beschrijf de context en motivatie voor je casus.
    - Wat is het probleem of de onderzoeksvraag die je met ML wilt oplossen?
2. **Analyse van de data**
    - Herkomst van de data (werkplek of online bron).
    - Beschrijving van de kenmerken, datatypen, ontbrekende waarden, verdelingen, etc.
    - Visualisaties ter ondersteuning.
3. **Data-voorbewerking (ML-ready maken)**
    - Beschrijf welke stappen je hebt genomen om de data geschikt te maken voor ML (opschoning, normalisatie, feature-engineering, synthetische data, train/test-split).
4. **Ontwikkeling van de ML-oplossing**
    - Kies en motiveer een geschikte ML-taak (classificatie of regressie).
    - Presenteer de gebruikte modellen, trainingsparameters en gebruikte Python-bibliotheken.
5. **Vergelijking van modellen**
    - Vergelijk minimaal twee algoritmen en bespreek de prestaties.
    - Onderbouw waarom één model beter presteert dan het andere.
6. **Resultaten en evaluatie**
    - Toon en bespreek de **confusion matrix**.
    - Bereken de volgende metrische waarden:
        - Accuracy
        - Precision
        - Recall
        - Specificiteit
        - F1-Score
    - Licht toe wat deze resultaten betekenen voor jouw casus.
7. **Reflectie**
    - Wat ging goed, wat kon beter?
    - Welke stappen of keuzes zou je anders maken bij een herhaling van deze opdracht?

4.1.5. Rapportage-eisen

Het document voldoet aan de gebruikelijke hbo-rapportage-normen:

- Voorzien van titelpagina, inhoudsopgave en hoofdstuknummering.
- Logische structuur en duidelijke schrijfstijl.

- Correct gebruik van bronvermelding (APA-stijl).
- Heldere visualisaties met figuuraanduiding en tabellen met bronvermelding en tabelaanduiding
- Enventueel bijlagen

4.1.6. Beoordelingscriteria (samenvattend)

Criterium	Toelichting
Relevantie casus	Praktische toepasbaarheid en helder probleemkader.
Data-analyse & voorbereiding	Kwaliteit van dataverkenning en preprocessing.
Modelontwikkeling & onderbouwing	Correcte toepassing van ML-modellen en motivatie van keuzes.
Evaluatie & interpretatie	Juist gebruik van metrische waarden en betekenisvolle analyse.
Software-deliverable	Werkende toepassing die ML-resultaten benut.
Reflectie & rapportagekwaliteit	Kritische zelfreflectie en heldere, professionele verslaglegging.