Opdrachtomschrijving

Deze opdracht is een toets om de module Machine Learning af te sluiten (beoordelingsopdracht). Het resultaat is een individueel cijfer.

De deelopdrachten volgen het *machine-learning workflow* zoals in college is benoemd/behandeld.

Lees dit document nauwkeurig door, svp.

Uitvoering

1. Je maakt de opdracht samen met anderen (team van 4 personen (maximum)).
2. Opleveringsformat: één Jupyter notebook (Python kernel).
3. Samenstelling notebook: combinatie van model, Python code en uitleg (beschrijvende tekst, grafieken).
4. Je levert de Jupyter notebook **en de bijbehorende dataset** als een ZIP-file op ELO (inleveropdracht), uiterlijk op de aangegeven deadline. Zie ELO en Studiehandleiding (nieuwe regel sinds 2019-2020: **ieder teamlid levert individueel de ZIP-file op**.
5. Je presenteert de resultaten samen met je teamgenoten aan een panel, bestaande uit docent(en) en expert(s).

> Tijd: max.30 minuten, inclusief vragen vanuit panel. Minimaal 15 minuten presentatie model, code en uitleg.

Criteria

1. Je maakt gebruik van minimaal de volgende libraries: `numpy`, `pandas`, `matplotlib` (`seaborn`) en `scikit-learn`. **Optioneel**: maak ook gebruik van de libraries, die in de colleges zijn behandeld of benoemd.
2. Alle code werkt zonder run-time errors (warnings zijn acceptabel).
3. Je hebt minimaal 2 machine learning algorithm/technieken toegepast. Je toont hiermee aan inzicht te hebben in de situatie en dat je niet slechts één machine learning techniek gebruikt, die je dan 'toevallig' kent.
4. Je hebt een verantwoording beschreven voor de `feature selection`. M.a.w. je kan uitleggen de keuzen voor de `features` en wat je ermee gedaan hebt voor aanvang van de analysis (zoals weglaten, mee laten wegen met bepaalde factor, etc.).
5. Elke (deel)opdracht is in het resultaat aanwezig.
6. Elk teamlid is in staat een relevant antwoord te geven op vragen van panel.

Beoordeling

- Wanneer aan de criteria is voldaan is het resultaat VOLDOENDE, anders ONVOLDOENDE.
- Het cijfer wordt bepaald door het panel op basis van hun ervaring en expertise, waarbij de volgende onderwerpen een rol spelen:

  > - netheid/verzorging Jupyter notebook.
  > - diepgang en complexiteit van de opdracht.
  > - diepgang van de door jou gegeven antwoorden.

- Het gegeven team-cijfer is een individueel cijfer, tenzij er redenen zijn om af te wijken.

Voorbeeld uitwerking

Voorbeeld van een mogelijke uitwerking staan bij [Kaggle tutorials (https://www.kaggle.com/tags/tutorial)](https://www.kaggle.com/tags/tutorial) w.o. de [Titanic: Machine Learning from Disaster (getting started) (https://www.kaggle.com/startupsci/titanic-data-science-solutions#)](https://www.kaggle.com/startupsci/titanic-data-science-solutions#).

1. Probleem/behoefte definitie

Opdracht

- Formuleer je vraagstuk. Doe een kort vooronderzoek om te bepalen welke vraagstuk je gaat uitwerken.
- Bespreek vraagstuk met docent/expert voor toestemming.

Alternatief: tijdens de exploratory fase formuleer je het vraagstuk dat je wilt uitzoeken.

Resultaat

Formulering vraagstuk, mogelijk in de vorm van een serie vragen, waarvoor je toestemming hebt van docent/expert.

Bronnen

- [Kaggle competities (https://www.kaggle.com/competitions)](https://www.kaggle.com/competitions)
- [UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/)](http://archive.ics.uci.edu/ml/)
- ....

2. Dataset

Zoek een dataset uit die je gaat uitwerken in deze casus.

Dataset bronnen

- [Kaggle (https://www.kaggle.com/datasets)](https://www.kaggle.com/datasets), mogelijk op basis van een competitie.
- [UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/)](http://archive.ics.uci.edu/ml/).
- **zelfgekozen repository**, overleg van tevoren met de docent/expert.

3. Exploratory phase

Opdrachten

- Onderzoek de dataset zowel **textueel** (beschrijvend) als **grafisch** (exploratief). Gebruik de libraries `pandas`, `numpy` en `matplotlib` (`seaborn`).
- Beschrijf elke `feature`: betekenis en datatype.
- Geef minimaal twee 'voorlopige vragen' die je mogelijk gaat uitgewerkt.
- Bespreek uit-te-werken vraagstuk/vragen met docent/expert.

Milestone resultaat

Formulering vraagstuk, mogelijk in de vorm van een serie vragen, waarvoor je toestemming hebt van docent/expert.

4. Cleaning phase

Opdrachten

- Benoem ontbrekende data voor `features`, zoals `NaN`-data, data dat in ander data-type beter past bij de analyses, of met een ander gewicht meetelt in de analysis.

  > Gebruik o.a. de libraries `pandas` en `numpy`.

- Benoem wat je met die `features` gaat doen en beargumenteer je besissing.

5. Model construction phase

Opdrachten

- Formuleer het algorithm (model) dat je gaat toepassen op de dataset en beargumenteer je keuze van het model.
- Werk het model/algorithm uit, zodat je antwoord krijgt op de eerder gestelde vragen of vraagstuk.

> Gebruik de libraries `pandas`, `numpy`, `matplotlib` en `scikit-learn`, en eventueel andere data-science libraries.

- Formuleer en werk uit een 2de machine learning algorithm.
- De uitwerking bevat

> - uitleg over model en strategie van de `feature selection`
> - grafieken
> - Python code
> - conclusie(s)

6. Model evaluation phase

Opdrachten

- Discusieer de waarschijnlijkheid van de antwoorden. Eventueel vergelijk je de resultaten met een ander, tevens uitgewerkt model/algorithm.
- **Optioneel**: Benoem de operationele vereisten voor het monitoren en updaten van het model.

7. Rapportage phase

Opdrachten

- Maak je Jupyter notebook opleverings-waardig.

> - check/verbeter spelfouten.
> - check/verbeter dat alle code werkt.
> - check/toon in de 1ste code-cell welke versies van de libraries je gebruikt hebt.
> - check/benoem de referenties, die je gebruikt hebt in de opdrachten. Referentie conform APA.
> - check/benoem wat-je-verder-van-belang-vindt met onderbouwing.

- Lever de Jupyter notebook op ELO, uiterlijk op de deadline.

- Presenteer samen met je team de resultaten voor het panel.

  - Zorg dat de presentatie van de Jupyter notebook correct werkt op de beoordelings-sessie.
  - Zorg dat de Jupyter notebook met de resultaten binnen drie minuten getoond kan worden.
  - Presenteer de resultaten (model, code, uitleg) in minimaal 15 minuten.
  - Tijd om aan te tonen dat jouw resultaat voldoet aan de cesuur is 30 minuten.

Success!

Stefan en Peter