# Analytical Computing

Kick-off project

hogeschool Windesheim

Radboud University

# Planning

20-4-2021 **Git**

22-4-2021 **Array Handling**

28-4-2021 **Pandas data manipulation**

11-5-2021 **Visualization with MatplotLib**

12-5-2021 **StatsModels for Python statistics**

21-5-2021 **Kick-off project**

26-5-2021 **Sklearn for regression learning**

18-6-2021 **Eindpresentaties (onder voorbehoud)**

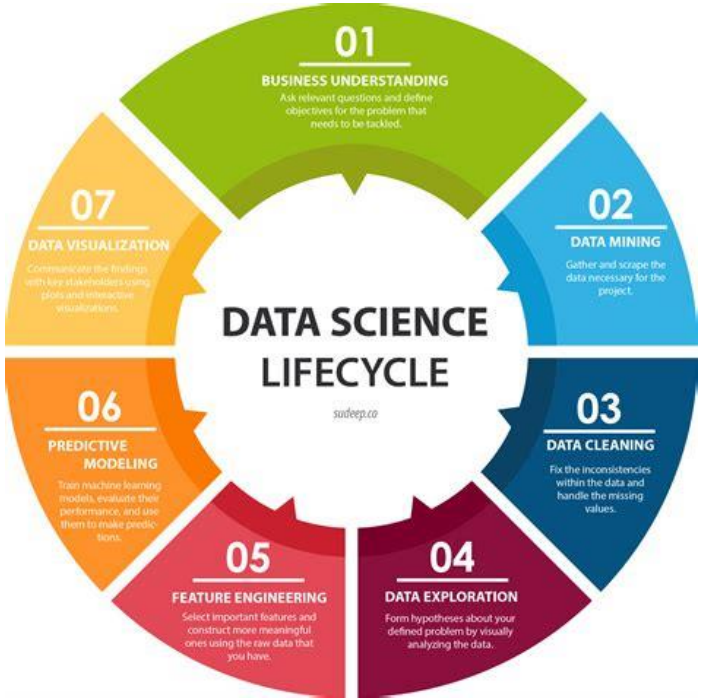
# Kick-off project

# Project

- *Trainen van een of meerdere machine learning modellen gericht op het oplossen van een regressie of classificatie probleem*
  - Regressie: continue variabele/iets dat je kan ‘tellen’ (huizenprijzen, covid-19 infecties)
  - Classificatie: categorisch/(ja/nee) (1-10 beoordeling van restaurant, wel of geen ziekte)
- Doel project:
  - Experimenteren met Pandas, NumPy, MatplotLib, StatsModels en Sklearn (behandelen in volgende les)
  - Leren werken met Git versiebeheer
  - Beheren van een efficiënte project lifecycle
  - Leren van elkaar

# Data Science lifecycle

1) **Business Understanding:** Wat wil ik bereiken met dit project? Wat willen mijn stakeholders weten?
2) **Data Mining:** Verzamel de data van de juiste bron(nen)
3) **Data Cleaning:** Verwijder missende waardes, onnodige kolommen, etc.
4) **Data Exploration:** Hoe ziet mijn data er *echt* uit, kan ik hier verbanden uit halen? (Matplotlib/Seaborn)
5) **Feature Engineering:** Zijn er kolommen (features) waar ik nieuwe features mee kan maken (bijv. categorische waardes omzetten naar binair)
6) **Predictive Modeling:** Welk model ga ik trainen en welke conclusies kan ik hieruit trekken?
7) **Data Visualization:** Wat zegt mijn model over de data? Hoe kan ik mijn stakeholders overtuigen?



# Project

- Werken in teams van 2 personen (3 bij oneven aantal)
- Python
- Twee mogelijkheden voor eindopdracht
    1) Vooraf goedgekeurde dataset, eigen invulling van probleem
    2) Eigen dataset (via bijv. Kaggle) met eigen invulling van probleem, moet eerst goedgekeurd worden

# Project

- Voor beide mogelijkheden: schrijf een kort voorstel (halve pagina) over het probleem dat je aan wil pakken (wat ga je voorspellen?)
  - Wanneer je kiest voor eigen dataset, voeg naast het voorstel ook de link naar de dataset toe
  - Mail je voorstel **voor 27 mei 23:59** naar sebastiaan.ram@student.ru.nl
    - Per team stuurt 1 persoon het voorstel door
  - Voorstel niet goedgekeurd? Pas aan met feedback en zorg voor goedkeuring **voor woensdag 2 juni**
  - Geen goedkeuring na woensdag 2 juni? Op eigen verantwoordelijkheid verder met het project

# Beoordeling

- Libraries zoals zijn gebruikt in de colleges moeten zichtbaar zijn geïmplementeerd
  - Pandas voor dataset manipulatie en visualisatie
  - NumPy voor aanmaken en manipuleren van arrays
  - Matplotlib/Seaborn voor het visualiseren van data
  - StatsModels voor het vinden van correlaties tussen features in jouw data
  - Sklearn voor het opzetten en trainen van voorspellende modellen

# Beoordeling

- Git is gebruikt voor versiebeheer
  - Maak een nieuw project aan
  - Zet in de README.MD een korte beschrijving van jouw project
  - Zorg dat jouw project public is!
  - Probeer zo klein mogelijke veranderingen te pushen om veranderingen overzichtelijk te houden en geef commit messages mee om te laten zien waar je mee bezig bent geweest
  - Optioneel: aparte productie en development branches

# Beoordeling

- Eindbeoordeling
  - 100% code + conclusies (gebruik van Git en implementatie van libraries)
  - Eindpresentatie als feedback-moment, zit geen cijfer aan gebonden
- Conclusies?
  - Overtuig ons van jouw waarnemingen
  - Waarom heb je voor bepaalde features gekozen? Wat zegt de data over jouw probleem? Klopt de nauwkeurigheid van mijn model?
  - Als markdown cell onderaan jouw Python notebook en kort samengevat in jouw eindpresentatie

# Beoordeling

- Eindpresentatie
  - 18 juni (onder voorbehoud)
  - Geen cijfer, wel feedback
  - +- 10 minuten per team
  - Vragen achteraf
  - Online (misschien fysiek?)

- Eindproduct
  - 1 Python notebook (.ipynb) met code, visualisaties en conclusies
  - Inleveren op ELO (omgeving wordt aangemaakt)
  - Deadline voor inleveren: 1 dag voor eindpresentaties (**donderdag 17 juni 23:59**)

# Projectplanning

21-5-2021 **Kick-off project**

27-5-2021 23:59 **Deadline inleveren voorstel**

2-6-2021 **Deadline goedkeuring voorstel**

17-6-2021 23:59 **Deadline eindopdracht op ELO**

18-6-2021 **Eindpresentaties (onder voorbehoud)**

# Ondersteuning

- Ondersteuning door Sebastiaan
- 1 á 2 ‘werkcolleges’ per week worden ingepland voor ondersteuning, deze zijn **optioneel**
- Voor team-to-one ondersteuning kan je een 15-minuten afspraak inplannen tijdens de werkcolleges: https://calendly.com/srram/analcomp-project
- Voor korte vragen mailen naar sebastiaan.ram@student.ru.nl

# Projecten

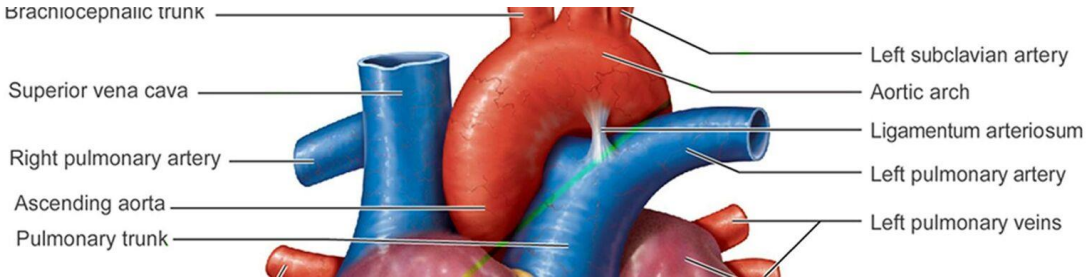
# Kaggle

- Grootste machine-learning community platform
- Ondersteunt door Google
- Duizenden datasets
- Challenges (met geldprijzen)

# Kickstarter Projects


- https://www.kaggle.com/kemical/kickstarter-projects
- Type probleem: classificatie
- 31 kolommen
- Mogelijke use-cases:
  - Voorspellen of project is geslaagd of niet

# Heart Attack Analysis & Prediction Dataset




- https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
- Type probleem: classificatie
- 15 kolommen
- Mogelijke use-cases:
  - Voorspellen of iemand een hogere kans heeft op een hartaanval (kolom ‘output’)

# Star Type Classification / NASA

- https://www.kaggle.com/brsdincer/star-type-classification
- Type probleem: classificatie
- 7 kolommen
- Mogelijke use-cases:
  - Voorspellen type ster op basis van verschillende kenmerken

# Water Quality

- https://www.kaggle.com/adityakadiwal/water-potability
- Type probleem: classificatie
- 10 kolommen
- Mogelijke use-cases:
  - Voorspellen of water veilig is voor menselijk consumeren

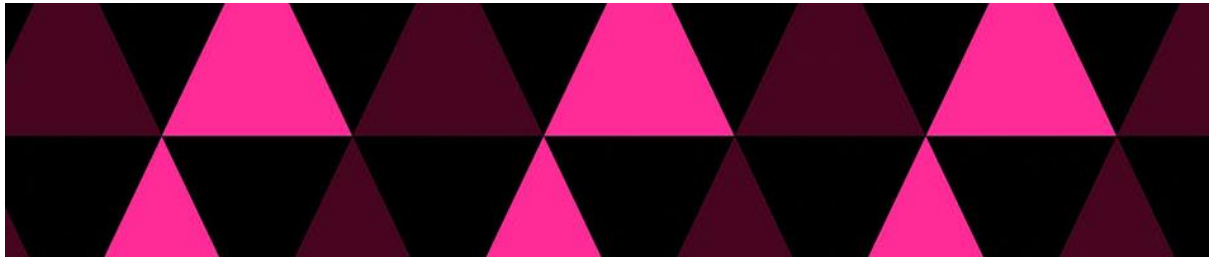
# Electric Motor Temperature


- https://www.kaggle.com/wkirgsn/electric-motor-temperature
- Type probleem: regressie
- 13 kolommen
- Mogelijke use-cases:
  - Voorspellen real-time indicatie PMSM stator en rotor temperatuur

# Beer Consumption - Sao Paulo

- https://www.kaggle.com/dongeorge/beer-consumption-sao-paulo
- Type probleem: regressie
- 33 kolommen
- Mogelijke use-cases:
  - Voorspellen cijfer(s) in verschillende periodes op basis van informatie over gezin en academische prestatie

# Student Grade Prediction



- https://www.kaggle.com/dipam7/student-grade-prediction
- Type probleem: regressie
- 7 kolommen
- Mogelijke use-cases:
  - Voorspellen bier consumptie per regio in São Paulo

# Hoe nu starten?

1) Vorm een team van 2 personen (3 bij oneven aantal studenten)
2) Kies een dataset uit de lijst (of vindt je eigen dataset via bijv. Kaggle)
3) Formuleer je probleem
4) In een paar zinnen, stuur je voorstel (1 per team) **in pdf formaat voor donderdag 27 mei 23:59** door naar sebastiaan.ram@student.ru.nl
    a) Voorstel goedgekeurd? Je kan aan de slag!
    b) Voorstel niet goedgekeurd? Zorg voor goedkeuring **voor woensdag 2 juni**