Case Study Projects

Guidelines for Group Project

- Each project describes a case study on machine learning.
- Four steps:
  - Task definition
  - Feature engineering, data pre-processing
  - Model building
  - Evaluation and Visualization
- You can choose any model!
- Submit:
  - Python notebook/R script
  - Report (4-5 pages)

Organization of the Report

- Title and group composition (max 5 members)
- Problem statement and objectives (task description)
- Dataset description (features, classes, instances, quality)
- Pre-processing and data representation
- Models used (including improvisations)
- Experimental results and comparison
- Visualization and discussion of results
- Conclusion

Project Descriptions

- Consumer and retail
- Energy
- Finance
- Public service
- Health care
- Telecom

P1: Retail Online Sales

- Goal: Predict the ratio of online sales as percentage of total sales of consumer good in UK for Q4, 2021 and Q1, 2022, based on historical time-series data
- Challenges: Modelling pre-covid and post-covid trends

- Data set:
- https://www.ons.gov.uk/generator?format=xls&uri=/businessindustryandtrade/retailindustry/timeseries/j4mc/drsi

P2: Bike Sharing

- Goal: Predict count of bike share requests based on weather and holiday parameters
- Challenges: Handling heterogeneous data types
- Data set: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

P3: Household Electricity Consumption

- Goal: Cluster customers of an utility company based on their household electricity consumption patterns
- Challenges: Time series clustering

- Data Set: https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

P4: Building Energy Efficiency

- Goal: Predicting heating and cooling load of a building based on its design
- Challenges: Complex prediction task

- Data Set:
- https://archive.ics.uci.edu/ml/datasets/energy+efficiency

P5: Solar Radiation for Energy Generation

- Goal: Predict amount of solar radiation based on weather parameters
- Challenges: Noise in data

- Data Set:
- https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

P6: Sentiment in Hotel Reviews

- Goal: Classify Sentiment (+/-) of textual customer reviews of an eco resort
- Challenge: Text handling

- Data Set: https://archive.ics.uci.edu/ml/datasets/Eco-hotel

P7: Credit Card Approval

- Goal: Classify credit card applications as approved or not-approved
- Challenges: Mixed attribute types

- Data Set: https://archive.ics.uci.edu/ml/datasets/Credit+Approval

P8: Stock Index Prediction

- Goal: Predict stock market prices and return on next dividend based on the Dow Jones stock index
- Challenges: Complex time series pattern

- Data Set: https://archive.ics.uci.edu/ml/datasets/Dow+Jones+Index

P9: Credit Rating

- Goal: Categorize credits into good/bad classes
- Challenges: Mixed attribute types

- Data Set: https://archive.ics.uci.edu/ml/datasets/South+German+Credit

P10: Bankruptcy Prediction

- Goal: Predict bankruptcy of Polish companies based on various attributes
- Challenges: Mixed data types, sequence data

- Data Set: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

P11: Online Shopper Buying Intention

- Goal: Predict intention of an online shopper (TRUE/FALSE)
- Challenges: Mixed data type, unbalanced classes

- Data Set: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

P12: Income Census

- Goal: Predicting if a citizen has income above/below 50K from demographic census data
- Challenges: Categorical data type, missing data
- Data Set: https://archive.ics.uci.edu/ml/datasets/Adult

P13: Gender Names

- Goal: The probability of first names (character strings) to be of a certain gender, is obtained from birth records in a number of western countries. We have to predict genders for first name character strings, whose gender field is missing in the birth records.
- Challenges: Modelling as a KNN classifier
- Data Set: https://archive.ics.uci.edu/ml/datasets/Gender+by+Name

P14: Predictive Maintenance

- Goal: The data set contains operating parameters of several machines sensed as a time series. Goal is to predict if the machine will fail within a future time frame.
- Challenges: Multivariate time series

- Data Set: https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset

P15: Influenza Outbreak Prediction

- Goal: We want to detect if there is an outbreak of influenza in a geographical state based on twitter activity of the previous week
- Challenges: Handling short text

- Data Set: https://archive.ics.uci.edu/ml/datasets/Influenza+outbreak+event+prediction+via+Twitter+data

P16: Heart Failure Risk Assessment

- Goal: To predict if a person is likely to have heart failure in future based on other health parameters.
- Challenges: Large variability in data set

- Data Set: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

P17: Drug Review Sentiment

- Goal: To predict patient satisfaction with a drug based on textual feedback comments
- Challenge: Presence of ambiguous terms in reviews

- Data Set: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

P18: Drug Discovery Using Molecular Features

- Goal: To predict binding sites of receptor molecules from studies on related molecules
- Challenges: Complex relation modelling

- Data Set: https://archive.ics.uci.edu/ml/datasets/Dorothea

P19: Telemonitoring Parkinson's Disease

- Goal: Predict Parkinson disease scores based on telemonitoring of patient voice conversations
- Challenges: Noise in the data set

- Data Set: https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

P20: Hospital Readmission Prediction

- Goal: Predict necessity hospital readmission of diabetic patients in USA based on hospital records
- Challenge: Incomplete data

- Data Set: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

P21: Customer Churn Prediction

- Goal: Predict churn for telecom customers based on calling behaviour
- Challenge: Incomplete data

- Data Set: https://archive.ics.uci.edu/ml/datasets/Iranian+Churn+Dataset

P22: Direct Marketing Campaign

- Goal: To predict whether a customer to whom a targeted marketing call is made will subscribe to a term deposit scheme
- Challenge: Heterogeneous features

- Data Set: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

P23: Industrial Fan Failure Time Prediction

- Goal: Predict time to failure of an industrial fan motor from sensor data
- Challenges: Sequence data

- Data Set: https://archive.ics.uci.edu/ml/datasets/Accelerometer

P24: Industrial Emission Prediction

- Goal: To predict emission of CO and NOx gases from sensor data in a power plant
- Challenges: Noise in data

- Data Set: https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set

P25: Music Recommendation Engine

- Goal: Build a simple music recommendation system based on past user choices
- Challenge: Cold start problem

- Data Set: Last.FM music recommendation data set
- http://ocelma.net/MusicRecommendationDataset/lastfm-1K.html

Thank You!