Machine Learning
Final Group Project

Thanks to: Ujjwal Sharma, Hongyi Zhu, Shuai Wang, and Ilker Birbil

1 Logistics and Instructions

This project brings together all the topics covered in the course. The project **will count towards your grade** and should be submitted through Canvas by **15.12.2023 at 23:59 (CET)**. You can get at most 20 points for these assignments, which is 20% of your final grade.

While it is perfectly acceptable to brainstorm and discuss solutions with other colleagues, please do not copy code. We will check all submissions for code similarity with each other and with openly-available solutions on the web.

Below we describe two problems, from which you should choose **only one**. In both cases, we indicate possible tasks that you could perform on the datasets. However, you are free to perform additional analysis or formulate interesting research questions on your own. While grading the projects, we will reward innovative and unconventional research questions. You can use matplotlib or other plotting libraries to visualize your findings.

2 Deliverables

There are two deliverables: a report and your code. You will write a report on your project, which explains to the reader what problem you are trying to solve, the approach to solving this problem, the results, and the implications of the results. You should also provide the code used for the experiments with your report.

Your report should not exceed **8 pages** including the figures and references. The report should be submitted in PDF format. Your final submission should consist of a single zip file with these deliverables and should be submitted through Canvas.

All datasets are available here.

3 Neighbourhood Statistics as predictors of bigger problems.

The Central Bureau voor de Statistiek (CBS) or Statistics Netherlands is a Dutch governmental organization that collects statistical data for the Netherlands. Once a year, they release the Wijk en Buurtstatistieken (neighborhood statistics) containing data on i.a. demographic, social, and geographical trends for all neighborhoods in the Netherlands. This data is publicly available and can be used to accurately predict and understand a range of societal effects that depend on these indicators.

In this project, your goal is to use the CBS data and combine it with another dataset publicly available for the Netherlands. For this you could look at a large collection of open datasets provided by the CBS, the Netherlands National Institute for Public Health and the Environment RIVM or other open data sources. The choice of this secondary dataset depends on the nature of the problem you are attempting to investigate.

Your mission, should you choose to accept it, could include:

- Building a regressor to predict:
  - *Cancer* mortality rates given demographic indicators for a region. For this, you could merge Cancer data from the RIVM with neighborhood statistics.
  - *Depression* cases and risk percentages given demographic indicators for a region. For this, you could merge Depression risk data from the RIVM with neighborhood statistics.

  Please note that these are just examples of the many possible options you can choose from.
- Performing feature importance analysis to understand what features strongly affects the depression prevalence

rates or cancer mortality rates and can be used as good predictors for similar public- health issues.

- Utilizing unsupervised learning techniques, such as clustering or outlier detection to identify different groups and anomalies in the dataset. The plot here shows a cluster of high mortality rates in the north-west of the country and could be related to a particular demographic attribute. Similar associations may hold for depression data.

Please note that these examples are just one of the many problems you can investigate. If you have a more interesting problem you'd like to link with demographics, feel free to do so.

4 Online News Popularity

The popularity of online news can depend on multiple factors like the Website where it was published, the content, etc. Websites often track viewing statistics for individual articles to better understand the type of content their users are looking for. In this project, you are provided with information for 39,797 online news articles [Fernandes et al., 2015]. For each article, you are provided with statistics on the word, topic, and sentiment level.

Your mission, should you choose to accept it, could include:

1. Classifying which articles will be popular by defining an appropriate popularity threshold.
2. Clustering articles on the types of words, topics, and general sentiment (provided as features) and examining if these clusters correspond with article popularity.

In addition, your in-depth analysis could also include:

1. Investigation into the effects of diminishing training set size and regularization strength on generalization.
2. An examination of the effect of the independent variables on all of the chosen dependent variables. You could also perform feature importance analysis to examine independent variables that strongly affect chosen dependent variables and can be used as good predictors.
3. Try experimenting with unsupervised machine learning techniques, such as clustering and outlier detection, to identify trends in the data.

5 Instructions for Tasks

These instructions are meant to serve only as pointers on how to think about the tasks and datasets described above. While previous assignments in this course centered around imparting you the required technical skills, this project will additionally test your ability to use scientific methods and observations to reach valid conclusions about the data.

6 Grading

Component	Points	Explanation
Problem Statement	5	Work Objective, problem tackled by the paper, etc.
Technical Quality	5	Experimental rigor, reproducibility, etc.
Quality, Diversity, and Novelty of Experiments	5	Breadth and quality of experiments to validate novel findings (if any), etc.
Presentation and Discussion-Quality	5	Quality of exposition (presentation, plots, usage of appropriate references wherever necessary, etc.)

References

[Fernandes et al., 2015] Fernandes, K., Vinagre, P., and Cortez, P. (2015). A proactive intelligent decision support system for predicting the popularity of online news. In *Portuguese Conference on Artificial Intelligence*, pages 535–546. Springer.