Machine Learning
Final Group Project

Ujjwal Sharma, Athanasios Efthymiou and dr. Stevan Rudinac

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 **15.12.2022 at 08:59 PM (CET)**. You must submit this assignment in groups (as registered on Canvas). 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. Submissions with high similarity will be summarily rejected and no points will be awarded.

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 in solving this problem, 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 Neighborhood 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 about the country. 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 predict and understand a wide array of societal effects connected with these indicators.

✉ u.sharma@uva.nl, a.efthymiou@uva.nl, s.rudinac@uva.nl

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 a few of the many possible options you can choose from. As a part of your analysis, you could:

- Perform 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.
- Utilize unsupervised learning techniques, such as clustering or outlier detection to identify different groups and anomalies in the dataset. For example, 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 (with its own dataset) that you'd like to link with demographics, feel free to do so.

4 Decoding Hotel Success in Europe

Europe is a popular tourist destination that hosts millions of tourists each year. Most hotels list themselves on popular reservation sites like Expedia and Booking.com to reach a larger audience. In this project, you are given a set of 515,000 reviews sourced from a similar aggregator for hotels all over Europe. This data contains the date, time, positive and negative reviews and the tags associated with those reviews.

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

1. Classifying which users/nationalities or groups are more likely to vote higher or lower than average. Can this information be used to systematically extract rating biases from these reviews?
2. Clustering hotels on attributes like client type, review types and evaluating if these clusters are semantically meaningful?
3. Clustering restaurants in a city based on their location. Use this to analyze if good and bad restaurants agglomerate in space? You could leverage libraries like Folium to plot the restaurants in interactive and rich maps (inside Jupyter Notebooks if you so prefer.

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
Problem Statement	5
Technical Quality	5
Quality, Diversity and Novelty of Experiments	5
Report Presentation and Discussion-Quality	5