Take-home Assessment

1 Description

This final practical is based on the material covered in lectures and previous practicals. You should write a practical report that should include a description and evaluation of the work done of not more than 2500 words excluding tables, graphs and images. The final practical will contribute 80% of the final mark. The deadline for handing in completed reports to student administration is **Friday 6th December 2019, 5pm**.

Additionally, you will need to submit your code – Jupyter notebook(s) or python script(s) – to the Moodle webpage. Assessors may run your code, but you will **not** be assessed on the quality of code writing, **nor** will you be assessed on the basis of where your system's results rank amongst others. The assessment will be based on the report itself and on clarity of description of the work done, evaluation performed and insights gained.

2 Dataset

The data set is a subset of DIABETES 130-US HOSPITALS FOR YEARS 1999-2008 DATA SET .[1] This data set contains around 50 various attributes representing patient and hospital outcomes for 10 years ($1999 - 2008$) of clinical care at 130 US hospitals.

You are encouraged to look further into:

- Original dataset: `http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008`
- Publications accompanying it: e.g., `https://www.hindawi.com/journals/bmri/2014/781670/`
- Description of attributes: `https://www.hindawi.com/journals/bmri/2014/781670/tab1/`

You will be working with the `diabetes/diabetic_data_balanced.csv` file. Following modifications have been made to the original data:

- to speed processing up, only about 10% of the original data is used in the practical;
- the data was extracted in such a way as to provide enough training instances for all classes.

Note that the original dataset is included in the same folder for your reference – see `diabetes/diabetic_data_original.csv`.

[1] `http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008`

3 Your task

Your task is to build a machine learning pipeline to predict the target value based on the other variables in the dataset. The target value represents whether a patient is readmitted to the hospital, and if they are, how many days it takes. Original dataset includes three types of outcomes: "<30" if the patient was readmitted in less than 30 days, ">30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.

Your implementation and report should include the following steps:

- *Data exploration*: note that the dataset contains a combination of categorical and numerical-valued features. It also contains a number of missing values. Explore different features in the dataset (e.g., you might want to remove features with mostly missing values), gain insights from the data and report your findings.
- *Machine learning algorithms implementation*: apply machine learning algorithms that you learned about in the previous practicals. Find out which algorithm works best and report your results.
- *Evaluation*: look into different ways and measures for evaluating the algorithms. You may consider looking into: RMSE, accuracy, precision, recall, F_1, trade-offs, ROC curves and AUC. Report your results and present your findings for the best-performing ML algorithm.
- *Visualisation and dimensionality reduction*: look into dimensionality reduction. For instance, you may consider using PCA on a selected set of features, plotting a scatter plot of the components and colour-coding the points by a selected categorical feature.