Machine Learning – Assignment

Machine learning is an active area of research with a high level of impact on real-world problems.

The objective of this assignment is to allow you to explore an interesting and relevant machine learning dataset using Scikit-Learn. More specifically you will be required to perform pre-processing, build and evaluate machine learning models and write a report on the results.

You will also be required to pick a specific area to research. This research should be integrated into your methodology and evaluation (more detail on this below).

Please note you should upload all deliverable files (the dataset, the python file and your report) into a single .zip file for submission. The submission deadline is **Tuesday Dec 18th at 20:00**.

Dataset

Your initial task will be to select an appropriate dataset. You should select either a regression or classification dataset. Ideally you should pick a dataset where machine learning algorithms have already been applied (although this is not essential). Clearly when selecting a dataset you should identify the column that will act as the classification or regression target for the model.

- Avoid using time series, text classification, image or audio data.
- Avoid datasets where you have to spend time merging a number of disparate datasets.
- I recommend that you limit the size of your dataset to 25MB in size. Just to give you an example, a 14MB file took 13 seconds to run 10 fold cross validation, a 20MB file took about 20 second to run 10 fold cross validation, a 25MB file took 25 seconds. These tests were carried out on an I5 and using a DecisionTreeModel. Clearly these times will vary significantly depending on the model and the characteristics of the data. You should keep in mind that you will need to do hyper-parameter optimization, which will take much longer. Please note this is just a recommendation and if you are really interested in adopting a bigger dataset please let me know.

Note: Before fully deciding on your proposal it is very important that you discuss the idea with me in order to validate its objectives and scope.

Distribution of Marks

This project will account for 5**0%** of your overall module grade. The marks will be broken down as follows:

- **Report - Abstract and Introduction [10%]**
- **Report - Research [15%]**
- **Report – Methodology [10%]**
- **Report – Evaluation and Conclusions [35%]**
- **Project Code [30%]**

Each of the above components in described in more detail below.

Project Overview

The project requires you to build machine learning models for your chosen dataset. You will need to perform pre-processing on your data. Follow the pre-processing steps outlined in the Scikit Learn lecture notes. You will need to build and comprehensively evaluate a range of machine learning models. The most promising models should then undergo hyper-parameter optimization.

You are also required to pick a specific topic to research and then incorporate the result of this research into your models and evaluate the impact. For example, if your dataset is imbalanced your research could focus on the techniques that are commonly used to address imbalance. You would then proceed to incorporate some of these into your evaluation and assess the impact on your results.

You should also compose a research report detailing the work you have undertaken and the overall findings. You will find a template for the research paper in the assignment folder. This template adheres to the Springer paper specification. The paper you submit should contain the following sections:

(i) Abstract
(ii) Introduction
(iii) Research
(iv) Methodology
(v) Evaluation
(vi) Conclusions and Future Work

I recommend that you do not exceed 8 pages for the research paper. I understand that some of you may have difficulty adhering to this limit. Please note that this is a recommended guideline, it is not a requirement and you will not be penalized if you do exceed that page limit. More detail on each of these sections are provide below.

1. Report - Abstract and Introduction [10%]

Your abstract should provide a short summary of the work that you undertook as part of the project. It should primarily provide an account of the main objectives and a summary of the results.

In your introduction you should provide a description of your chosen dataset. You should clearly identify the target regression or classification value that you want to predict. Describe the motivation for building models for this dataset and the objectives of the study. Your introduction should also highlight any previous academic papers that used this dataset.

2. Report - Research [15%]

The section should outline the specific topic of research that you will incorporate into your study. The objective of this section is that it allows you to select a particular stage of the pre-processing or model building process and research it in more depth and incorporate aspects of this into your methodology and results. It is also import that you describe the techniques you are using in the research section. You should demonstrate that you understand the operation of the techniques you are going to employ.

There are a broad range of topics that you could consider for your research component. For example you could look at:

- Outlier Detection (Researching a range of techniques for performing outlier detection and investigating their impact).
- Feature Selection Techniques (Research a range of feature selection techniques and investigate their impact if any on your evaluation)
- Dataset Imbalance
- Feature Encoding, etc

For many of the above areas I have demonstrated in the lecture notes a limited number of techniques. For example, with dataset imbalance we covered techniques such as random under and oversampling as well as SMOTE. If you were to select this topic of imbalance then it would be acceptable to compare the impact the techniques used in the lecture slides.

However, **to excel in the research section** you should demonstrate independent research and an ability to apply relevant techniques from literature and integrate into your process and evaluate the impact on your results. Please make sure to reference any sources you use.

3. Report - Methodology [10%]

The section should outline the sequence of pre-processing steps that you undertook in order to prepare your data and the rationale for adopting these techniques.

It should describe the range of models you used in your initial model building phase. It should describe the hyper-parameter optimization technique that you employed and the range of parameters that you examined for each of the best performing models.

If you experimented with different pre-processing techniques for a particular stage (for example, you evaluated for both one-hot-encoding and ordinal encoding) then you should also communicate that in this section.

4. Evaluation and Conclusions [35%]

This section should contain a comprehensive evaluation of your results. You should report your results for the initial model exploration as well as the optimized results after hyper-parameter optimization. Also in this section you should clearly demonstrate the impact of your chosen research on the overall results.

The results should be clearly interpreted and depicted (graphically where appropriate). You should use a range of evaluation metrics. It is important you demonstrate a clear understanding of the evaluation metrics that you use.

Also please make sure you provide an intuitive method of cross-referencing between your code and the results in the evaluation. You could for example include the section number as a comment in your code. This will allow me to easily identify the code that generated each set of results.

This section should also include a conclusion, which outlines possible areas of future work.

5. Project Code [30%]

All code should be completed using Python as the programming language. You should use Scikit Learn, NumPy and Pandas. You are free to use imported graphical libraries such as MatPlotLib or Seaborn (This is not a requirement. You can also use tools such as Excel if generating graphs). You are also free to import Scikit-Learn contribution packages such as Imbalanced Learn. If you wish to use other external libraries please check with me in advance.

Your code should have a logical structure and a high level of readability and clarity. Please comment your code and put all code into functions. You code should be efficient and should avoid duplication.