|  |  |  |  |
| --- | --- | --- | --- |
| **COMP1816** | **Machine Learning** | **Faculty Header ID:** | **Contribution: 60% of course** |
| **Course Leader: Dr Jing Wang** | **COMP1816**  **Coursework** |  | **Deadline:**  **23:30, Monday 11/03/2024** |
| This coursework should take an average student who is up to date with tutorial work approximately 50 hours   Feedback and grades are normally made available within 21 calendar days of the coursework deadline | | | |
| **Learning Outcomes:**   1. Demonstrate understanding of the major approaches of machine learning. 2. Understand different types of machine learning algorithms. 3. Choose appropriate learning algorithms to solve specific real-world applications and evaluate the choice of algorithms. 4. Critically evaluate and analysis the merits and limitations of machine learning algorithms. | | | |

Plagiarism is presenting somebody else's work as your own. It includes copying information directly from the Web or books without referencing the material; submitting joint coursework as an individual effort; copying another student's coursework; stealing coursework from another student and submitting it as your own work. Suspected plagiarism will be investigated and if found to have occurred will be dealt with according to the procedures set down by the University. Please see your student handbook for further details of what is / isn't plagiarism.

**All material copied or amended from any source (e.g. internet, books) must be referenced correctly according to the reference style you are using.**

**Your work will be submitted for plagiarism checking. Any attempt to bypass our plagiarism detection systems will be treated as a severe Assessment Offence.**

# Coursework Submission Requirements

* Uploaded on the Deadline of **23:30 Monday 11/03/2024** using the link on the coursework Moodle page for COMP1816.
* For this coursework you must submit a **single PDF document** and a **zip file your code both as a. ipynb workbook and as a PDF exported** from collab.  In general, any text in the document must not be an image (i.e. must not be scanned) and the file should be generated from **LaTeX**.
* Maximum 2000 words (exclude the title of the report, student name, student number). No page limitation.
* Make sure that any files you upload are virus-free and not protected by a password or corrupted otherwise they will be treated as null submissions.
* You must NOT submit a paper copy of this coursework.
* All coursework must be submitted as above. Under no circumstances can they be accepted by academic staff

The University website has details of the current Coursework Regulations, including details of penalties for late submission, procedures for Extenuating Circumstances, and penalties for Assessment Offences. See [http://www2.gre.ac.uk/current-](http://www2.gre.ac.uk/current-students/regs) [students/regs](http://www2.gre.ac.uk/current-students/regs)

**Detailed Specification**

***This Coursework is to be completed individually.***

# Datasets

Two datasets are provided for the two tasks, respectively. The **California Housing dataset** is for the regression task, and the **Titanic dataset** is for the classification task.

**California Housing dataset**: The data provided is a modified version of benchmark California housing dataset. In this regression task, please use the first 800 datapoints of the given dataset for training and the rest 200 datapoints for test. You are required to build predictive models **to predict median\_house\_value** in the test datasets, and then evaluating the performance of different models based on evaluation metrics.

The columns of the dataset are: No., longitude, latitude, housing\_median\_age, total\_rooms,total\_bedrooms,population,households,median\_income,median\_house\_value, ocean\_proximity.

**Titanic dataset**: The data provided is a modified version of benchmark Titanic dataset. The sinking of the Titanic is one of the most infamous shipwrecks in history. In this classification task, please use the first 700 datapoints of the given dataset for training and the rest 190 datapoints for test. You are required to build a predictive model using passenger data (ie name, age, gender, socio-economic class, etc) **to predict whether passenger in test datasets will survive or not (i.e., Survival)**, and then evaluating the performance of different models based on evaluation metrics.

Data Dictionary

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Key |
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st class, 2 = 2nd class, 3 = 3rd class |
| Name | Passanger’s name |  |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | Number of siblings / spouses aboard the Titanic |  |
| parch | Number of parents / children aboard the Titanic |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

### Variable Notes

pclass: A proxy for socio-economic status (SES)  
 1st = Upper  
 2nd = Middle  
 3rd = Lower  
  
 age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5  
  
 sibsp: The dataset defines family relations in this way...  
 sibling = brother, sister, stepbrother, stepsister  
 spouse = husband, wife (mistresses and fiancés were ignored)  
  
 parch: The dataset defines family relations in this way...  
 parent = mother, father  
 child = daughter, son, stepdaughter, stepson  
 Some children travelled only with a nanny, therefore parch=0 for them.

To complete this assignment, you must use the data provided on Moodle at the following URL:

<https://moodlecurrent.gre.ac.uk/course/view.php?id=81057>

Submissions based on other data will not be marked and will receive 0 marks.

# Report

# This coursework is to implement a selection of appropriate machine Learning methods for the two tasks (Regression and Classification). For each task, at least THREE machine learning models (one main model and at least two baseline model) are implemented. The main model is the model that you think would be most suitable for the task. You need to clarify the reason for choosing the model and describe it in detail. The baseline models are the models to be compared with the main model. The implemented models MUST contain one or more models that you have learned in the module COMP1816. You need to evaluate these models and analyze the reason why the main model is better/worse than the baseline models. This must be presented in the form of a report to your line manager. This report must be split into the FOUR parts below (shown with the corresponding mark weighting), with marks also being rewarded for the presentation and language in your report. Your report should contain images, tables, and equations to help you demonstrate your work in Parts 2 and 3. It is down to your discretion offences. how many images/tables you wish to include, what they should show and how to split them across the different parts.

# Part 1 – Introduction (5 Marks)

# This should be a brief summary of what the report contains – the problem you are solving, the algorithms you have implemented, results you have obtained and your conclusions. This should only be between 100-200 words. Marks will be deducted if words more than 300.

# Part 2 – Regression (40 Marks)

# You are required to import the data and implement at least two regression models to predict housing prices.

# You must describe all relevant details of the implementation and clarify the reason for choosing the model and evaluation metrics.

# Part 2-1: Pre-processing (10 Marks)

# You are required to give a brief introduction to the dataset and describe the detailed procedure that how you pre-processed the dataset (e.g., splitting features and target, drop features, the reason should be clearly stated).

# Part 2-2: Methodology (15 Marks)

# You are required to clarify the reason for choosing the regression model as the main model. Also, you need to describe the model in detail (e.g., by using mathematical equations).

# Part 2-3: Experiment (15 Marks)

# Part 2-3-1: Experimental Settings You are required to describe the experimental settings in detail, e.g., baseline models, and hyperparameters tuning of all the models.

# Part 2-3-2: Results You are required to select at least one regression evaluation metric and show comprehensive results. The reason for choosing the metric should be clarified. Note that 0 marks will be given if only results are given but no explanation of experimental settings in the report.

# Part 2-3-3: Discussion You are also required to compare the results of different models, and analyse the results (e.g., the reason that the main outperforms baselines).

# Part 3 – Classification (40 Marks)

# You are required to import the data and implement at least three classification models to predict passengers will survive or not.

# You must describe all relevant details of the implementation and clarify the reason for choosing the model and evaluation metrics.

# Part 3-1: Pre-processing (10 Marks)

# You are required to give a brief introduction to the dataset and describe the detailed procedure that how you pre-processed the dataset (e.g, dealing with missing values).

# Part 3-2: Methodology (15 Marks)

# You are required to clarify the reason for choosing the classification model as the main model. Also, you need to describe the model in detail (e.g., by using mathematical equations).

# Part 3-3: Experiment (15 Marks)

# Part 3-3-1: Experimental Settings You are required to describe the experimental settings in detail, e.g., baseline models, and hyperparameters tuning of all the models.

# Part 3-3-2: Results You are required to select at least one classification evaluation metric and show comprehensive results. The reason for choosing the metric should be clarified. Note that 0 marks will be given if only results are given but no explanation of experimental settings in the report.

# Part 3-3-2: Discussion You are also required to compare the results of different models, and analyse the results (e.g., the reason that the main outperforms baselines).

# Part 4 – Conclusion (5 Marks)

# Provide a brief summary of the work done and discuss the limitations and potential improvement.

# Report presentation and language (10 Marks)

# The report should be presented in a professional manner with a neat and clear layout and all writing in proper English using good grammar.

# The report must be written using the following Latex template (make your own copy to work on):

* + URL: <https://www.overleaf.com/read/hbqfxbwkrrys>

# To obtain full marks your report should adhere to the following:

# Follow the template page layout (Latex):

# No changes to font type, size or colour

# No changing the title of the report and page layout (e.g. margin, orientation, size)

# Text should be split into sections with titles and proper paragraphs

# Images should be properly cropped and at an appropriate size and resolution.

# Correct spelling, punctation and grammar should be used in any text.

# Marks may also be removed at markers discretion for other issues with presentation not listed above.

# You should not include any screenshots of your code or the loaded data.

# Python Code

You should implement a solution to the above task all within a single Python Notebook using Google Colab. Any data shown in the text/tables or images used in your report should be generated using this Python Notebook.

This source code must also be provided as both an downloaded notebook (.ipynb) and exported as a PDF from Google Colab.

Markers should be able to run your provided code without any errors requiring debugging. If they are unable to reproduce any results in your report, you will be docked marks for that section.

However, the markers may not actually view your source code, and will only do so if they feel they need to test or otherwise check something. Your marks are based on the report and not the code, so ensure that everything required to complete the tasks are in the report. Even if your code provides a good solution, anything not shown/explained in the report will not be marked.

# Deliverables

The following two files should be uploaded using the appropriate Moodle link:

* + A PDF report (exported from Latex)
  + A zip file containing

1. Your supporting source code in Python Notebook format (.ipynb) implemented using Google Colab
2. A PDF copy of this source code downloaded from Google Colab

# Grading Criteria

Marks will be allocated according to the following rubric:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **0-29% Fail** | **30-49% Fail** | **50-59% Good** | **60-69% Very Good** | **70-79% Excellent** | **80-100% Exceptional** |
| **Introduction**  **(5 Marks)** | A completely unsatisfactory summary of the work done which fails to cover almost any of the basic points. | A poor summary of the work done which fails to cover all the basic points. | A good summary of the work done covering all the basic points. | A very good summary of the work done covering all points in detail. | An excellent summary of the work done lacking only minor details. | An outstanding summary with very little that could be improved. |
| **Regression** | A completely | A poor attempt to | A good implementation of | A very good | An excellent | An outstanding |
| **(40 Marks)** | unsatisfactory attempt to | implement a regression | regression with a model | implementation of binary | implementation of | implementation of |
|  | implement a regression | model which has limited | chosen with some | classification choosing a | regression choosing a | regression with very little |
|  | model which has no | experimental justification | evidence of | model with clear evidence | model with evidence of | that could be improved. |
|  | experimental justification | and confused explanation | experimentation and some | of experimentation and | systematic |  |
|  | and incorrect explanation | of regression accuracy | explanation of regression | good explanation of | experimentation and |  |
|  | of regression accuracy | metrics. The data used | accuracy metrics. Data has | classification accuracy | excellent explanation of |  |
|  | metrics. The data used | may not have been | been successfully loaded | metrics. Data has been | regression accuracy |  |
|  | may not have been | correctly loaded, split or | and utilised by the model. | successfully loaded and | metrics. Data has been |  |
|  | correctly loaded, split or | otherwise applied to the |  | utilised by the model. | successfully loaded and |  |
|  | otherwise applied to the | model. |  |  | utilised by the model. |  |
|  | model. |  |  |  |  |  |
| **Binary** | A completely | A poor attempt to | A good implementation of | A very good | An excellent | An outstanding |
| **Classification** | unsatisfactory attempt to | implement a binary | binary classification with a | implementation of binary | implementation of binary | implementation of binary |
| **(40 Marks)** | implement a binary | classification model which | model chosen with some | classification choosing a | classification choosing a | classification with very little |
|  | classification model which | has limited experimental | evidence of | model with clear evidence | model with evidence of | that could be improved. |
|  | has no experimental | justification and confused | experimentation and some | of experimentation and | systematic |  |
|  | justification and incorrect | explanation of | explanation of | good explanation of | experimentation and |  |
|  | explanation of | classification accuracy | classification accuracy | classification accuracy | excellent explanation of |  |
|  | classification accuracy | metrics. The data used | metrics. Data has been | metrics. Data has been | classification accuracy |  |
|  | metrics. The data used | may not have been | successfully altered to a | successfully altered to a | metrics. Data has been |  |
|  | may not have been | adequately altered for a | form fit for classification | form fit for classification | successfully altered to a |  |
|  | adequately altered for a | classification problem |  |  | form fit for classification |  |
|  | classification problem |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Conclusion (5 Marks)** | A completely unsatisfactory conclusion that fails to adequately summarise the prior work done and no discussion on the limitation of the work | A poor conclusion that only loosely summarises the work done and very poor discussion on the limitation of the work | A good conclusion summarising work done and a basic discussion on the limitation of the work | A very good conclusion providing a clear summary of the work done, a clear discussion on the limitation of the work | An excellent conclusion which provides an in-depth summary of the work done and a clear discussion on the limitation of the work with a rough idea is provided to improve the work. | Outstanding conclusion with little to improve.  It concludes an in-depth summary of the work done and a very clear discussion on the limitation of the work with a very clear idea provided to improve the work. |
| **Report Presentation and language (10 marks)** | A completely unsatisfactory report layout and language, often being disjointed and confusing to read. | Poor language and layout overall, with many mistakes in the report layout and language that is often hard to follow, | Good layout and language overall, though many minor mistakes in report layout and quality of language | Very good layout and language. Could be improved with better grammar and minor fixes to presentation mistakes. | Excellent language and layout, could be improved with better grammar. | Outstanding layout and language used throughout. Report presentation follows guidelines perfectly  and language clear and professional throughout. |

# If you are unsure about any of these instructions, then please email your module leader or make an appointment to see your module leader as soon as possible.