

School of Computing, Creative Technology and Engineering

|  |  |
| --- | --- |
| Student ID |  |
| Student Name |  |
| Module Name & CRN |  |
| Level |  |
| Assessment Name & Part No. |  |
| Project Title |  |
| Data of Submission |  |
| Course |  |
| Academic Year |  |

Table of Contents

[PART I 3](#_Toc158150264)

[PART II 5](#_Toc158150265)

This document includes information regarding report writing for each part. Remember you are not aiming to become an expert in a particular problem. The main aim of this assignment is to test if you can identify the type of the problem and use right techniques to address it, present tool (how to), then present results and communicate your outcomes.

# PART I

The project’s title – Relate to the case study

The author’s first and last names, course and year

Table of Content

**Introduction** to the problem/claim

* introduction to the problem what it is, why it is important these days, examples of some real application.
* Literature review of similar problems and machine learning/ statistical methods used to solve it. Summarising each literature, you should mention what the problem authors presented, what data they use, what technique they used and what was the performance of the methods. Also, if there were any interesting relationships or causality discussed please present it. You can comment on similarity of your case study with the one from literature based on the dataset similarity or methodology you are going to use.
* Use at least 10 scientific papers/books, conference papers
* You may find it useful to write literature review: https://www.phrasebank.manchester.ac.uk/
* **Exploratory Data Analysis**
* this is an open-ended question – report interesting facts about data, use tabular/graphical methods, explore the data as much as possible, avoid analysis of single variable for predictors – except for output variable which can be analysed as a single variable
* You have two data sets you should present both separately or you can combine them into one but remember to keep track of which records belong to which dataset to split them accordingly later on.
* You start with presenting dimension, columns and their types, some summaries with explanation or interpretation, you can mention any problems you can spot at this point out wrong datatypes for variables, missing values, 0-variance variables, unique identifier, complex variables that need to be split, transformed, or removed
* Visually explore some variable (please do not include correlation here)
* Example:

*Dataset consist of two datasets train and test.*

*> str(train)*

*….*

*>str(test)*

*…...*

*Train dataset have X samples/rows and test dataset has Z samples. Both has N number of columns that represent ….............................*

*We can notice presence of missing values what will be addressed in Section 3, unique identifiers (columns N1, N5, ….) that need to be removed.*

*The training and testing dataset were combined for a purpose of EDA analysis in the following way:*

*> R code*

*Or*

*The following analysis were presented separately for each dataset,*

*The prediction of the concrete strength is the main interest of the company delivering concrete for building sites. In the collected dataset variable describing this characteristic is presented as Strength. It is numerical variable and indicated that we will have to solve regression problem to create predictive model for it. Let's first explore Strength variable:*

*> ggplot(combine, aes(X=Strength) + geom\_histogram() + …..*

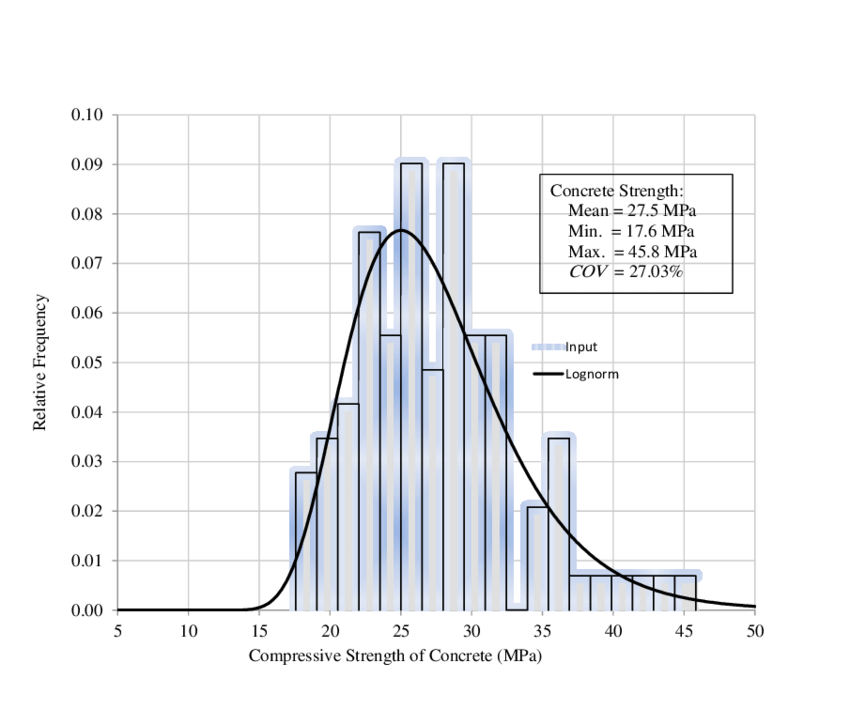


Fig2.2 Distribution of variable Strength

*Figure 2.2 presents distribution for concrete strength, the distribution is normal and symmetric with mean equals to …. we can notice some high values on right side that will be further explored in Section 3.*

* **Data Pre-processing**
* Missing values - examine column missingness and overall missingness. If the average of missing values in columns is greater than 1% impute values using mice() otherwise remove missing values. Report the steps of analysis, the code used to assess missingness, removal or imputation also report changes to the dataset and make clear statement with argumentation of your decision (impute/delete)
* Outliers - state a technique you used to detect outliers (univariate, multivariate), report the analysis steps of finding outliers (code and results) and provide your justification for the selected techniques. Make a decision if you delete outliers or not and justify it.
* Multicollinearity - once missingness sorted report correlation of predictor variables, justify your decision to delete/not to delete correlated variables. Report R code to calculate/visualise corelation and results then interpret your results.
* Scaling - justify if there is a need/or not for scaling variables values. Use a technique of your choice.
* **Bibliography**
* 10 scientific positions- at least
* if you use somebody’s analysis available online (e.g you may find some web with a nice visualisation of missingness or probability distributions and you think that may work with your data – you can apply and adjust the code to your case study but **you must** to refer to the original source. This can be done in following way in your report text e.g – this refers to req Tasks 2-5:

*“In [6] authors used a function “ name” to visualise various probability distributions. Following their approach/solution the credited card training data can be graphically presented: ….. You can notice that…….. “)*

# PART II

The project’s title – Relate to the case study

The author’s first and last names, course and year

Start with the summary of your data after the pre-process step. How many column rows meaning of the variables, summarise in one sentence what have you done as pre-process step. If you combined dataset, it is a moment where you need to split it. Do not use random split. You need to have records as they were originally.

* **Modelling:**
* Develop five predictive models: choose any model you have studies in this module. While training models, optimise their parameters and sample data to minimise Type 1 and 2 Errors. Justify sampling methods and K - fold size choice.
  + - Classification: Apply cross validation for the training method. For each model show confusion matrix: Comment on the best model for training and its parameters. Justify sampling methods and K - fold size choice. The iteration during training can be presented using ACC and kappa by bwplots().
    - Then aggregate results of the bests model's performance in a form of table, ROC curves. Combine all results in one table/graph.

The results for classification models should be aggregated in one table as:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **FP** | **FN** | **Kappa** | **AUC** |
| LR |  |  |  |  |  |  |  |
| NN |  |  |  |  |  |  |  |
| …... |  |  |  |  |  |  |  |

The results for regression models should be aggregated in one table as:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **R2** | **Adjust R2** | **MSE** | **RMSE** | **MAE** |
| LR |  |  |  |  |  |

* + Test performance of each model on test dataset. Report performance indicators. Comment on the results – similar table as above to compare models on testing dataset. All ROCs on one plot for Classification. Generate separate plots for residuals and fitted vs original and plot them as one grid for Regression.
* **Model Interpretation**

Select the best predicting model based on test. Apply interpretation method for the selected model

* **Conclusions**

Critical analysis of how your models perform in comparison to the selected ML methods discuss in Task 3 (Introduction)? Discuss the most important variables for the selected model by referring to the (Tasks 4 and 5). Do any of these variables provide a good separation/explanation for classes or are highly correlated with a target variable?

* **Bibliography**

You may include new references to support your justification so include all refernces also these from part I

Requirements:

* + short intro to the analytical step
  + include a code snippet with the realisation of the analytical task
  + include the code output in form of text/table/plots
  + interpret the output/results – plain English - just describe what you can see from your results output and comment on it.
  + Write the report in the way that whoever will take it it will be reproduce your results.