Instructions for Project Submission

The participants are expected to work on a project on the topic of ‘Classification’ and submit their analyses. They are required to perform classification analysis on data and interpret the results obtained by providing business insights.

# Important Points

1. Please consider a case study and clearly mention the **motivation** behind it, its **objectives** and **outcomes**. Emphasise the business or management side of the problem by providing an appropriate business context. (Please refer to the two-wheeler case study):
   1. This should ideally be a one-page write-up clearly mentioning:
      1. The description of the problem or the business scenario,
      2. The intent/purpose/objective of the analysis with business contexts and
      3. How **classification** algorithms would help you to achieve your objective.
2. The analysis should aim to categorise the observations into one of the two categories. This means you need to develop a case study that aims at ‘**binary classification**’. That further means that the response variable that you choose should have just two classes.
3. Gather relevant data for the case study. Consider a data set with several observations and variables. Make sure some predictors are quantitative and some are categorical:
   1. For public data or data obtained from an open-source database, please mention the source and URL explicitly.
   2. For private data, it is recommended that you use a portion of it, as opposed to the full data, to avoid any copyright issues, and that too with written permission/consent from the concerned authority or organisation.

## Ideally, you should first find suitable data and then build a case study around it. Listed below are some repositories that may be useful for data selection:

1. <https://catalog.data.gov/dataset>
2. <https://www.kaggle.com/datasets>
3. <https://github.com/>
4. <https://www.nseindia.com/>

Please note that the websites mentioned above are not the only repositories for choosing data. You may feel free to collect/access relevant data from any other website, but ensure to mention the data source and URL appropriately.

## Note: You may choose the same data that you have used for the `Regression Analysis’. Make sure to change the context to a more suitable one with respect to classification and pre-process the data appropriately.

1. Clearly describe the variables in the data set that you have used in the analyses. Specify the response variable (**binary**) and the predictors/features:
   1. For example, if the data set contains a variable ‘Age’, then you must provide a description such as this.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Details** |
| AGE | Continuous | Measures the age of a person in years |

* 1. If there is a derived variable ‘Age’ that is required for analyses, then you must specify how that derived variable is calculated. The table below shows an example.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Details** |
| DATE | Date/time | Today’s date in DD-MM-YYYY |
| DOB | Date/time | Date of birth in DD-MM-YYYY |
| AGE | Continuous/derived | Measures the age of a person in years; AGE=ROUND((DATE-DOB)/365.25) |

1. Perform **descriptive** and **inferential analyses** on the data:
   1. Identify 2–3 important numeric predictors from a business perspective and compare them across the two groups of the binary response variable using measures of centers/dispersions or histograms.
   2. Perform **two-sample t-tests** between the two groups of the response variable based on the continuous predictors mentioned above and, thus, check whether any difference exists between the two groups at 5% level of significance. Before performing the two-sample t-test, please check for the normality of the observations using **Shapiro–Wilk’s** test.
2. **Logistic regression (LR):** Perform an 80%–20% split of the data into training and test sets. Fit an LR model on the training set. Next, you need to do the following:
   1. Perform a **variable selection** and choose an optimum subset of predictors. If the variable selection method rejects a specific variable that could be important from a business perspective, then decide by judgement and proper reasoning what is to be done (include/exclude) with that predictor
   2. Present the LR model with the optimal subset of predictors
   3. Define each component of the model. Also, define what does ‘odds’ signify in this case
   4. Interpret the **partial regression coefficients** based on at least one continuous and one categorical predictor and discuss their impacts on the response variable **in terms of odds**
   5. Interpret **residual deviance** and **Akaike information criterion (AIC)** values
   6. For every observation in the test set, predict the probability of (Y = 1) and plot a histogram using those probabilities
   7. Choose an appropriate threshold to decide whether an observation would belong to Y = 1 or Y = 0
   8. Using the threshold above, find and **explain** the following measures for the test data using the **confusion matrix**:
   9. Sensitivity
3. Specificity
4. Positive predictive value
5. Accuracy

**i.** Plot a receiver operating characteristic (ROC) curve and provide the area under the curve (AUC) value, and, thus, comment on the performance of the model

1. **Random forest:** Fit a random forest ensemble model on the data:
   1. Choose hyperparameters appropriately, e.g., number of trees in the random forest and number of features to be used for each tree
   2. Generate the confusion matrix and calculate accuracy
   3. Report the **out-of-bag (OOB)** error and explain what this number signifies
   4. **Tune** the hyperparameters to come up with a random forest with good accuracy (if possible)
2. **Conclude** your analyses by commenting on which model is better and how these models will be useful in fulfilling the objective of the study and help in the business. Write one paragraph on this

# Presentation

* 1. The project should be written in a Word .doc file (just like a report) and submitted as a .pdf file.
  2. The number of pages should not exceed 15 (including all outputs, tables and figures). You should, thus, try to be precise in your explanation and presentation.
  3. ‘Cambria’ font of size 10 should be used.
  4. All output (including tables and figures/plots) should have proper headings and titles. All output should be numbered, e.g., Table 1, Table 2, or Figure 1, Figure 2.
  5. Axes in plots/graphs should be labelled and titled properly.
  6. Please avoid the use of a variety of colours for graphs unless absolutely necessary. Simple black, blue, red and green are fine.

1. R-codes can be submitted separately and should be commented at every step.
2. Please follow the ‘two-wheeler’ case study in case you need help with analysis and interpretation.
3. You should strictly adhere to all the requirements/guidelines mentioned in points 1–11 in order to score full marks.