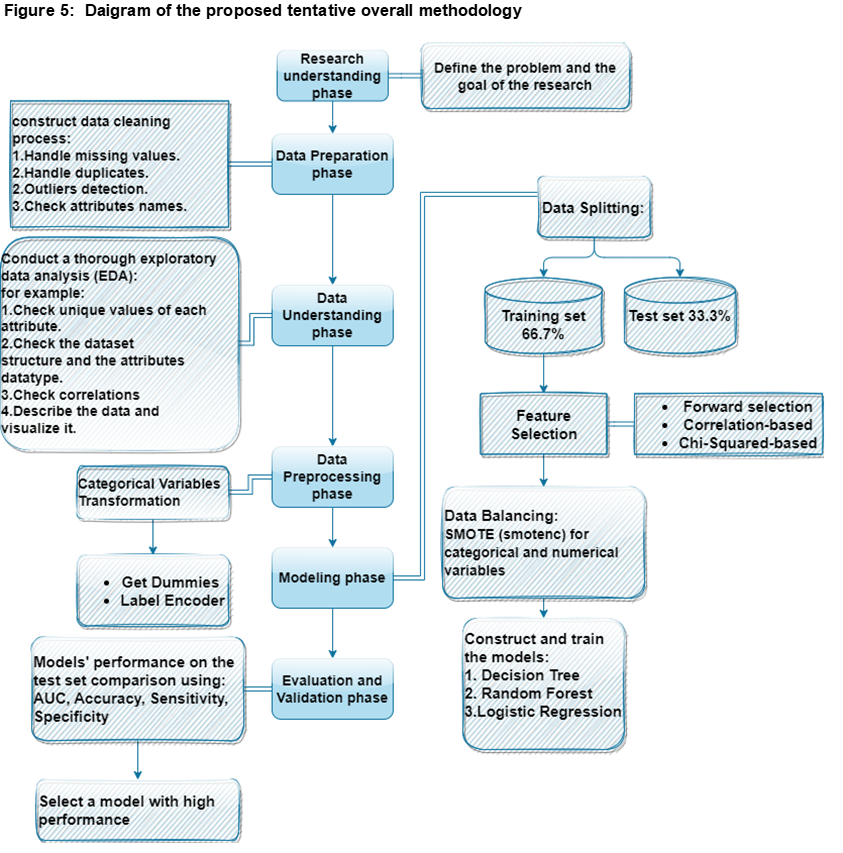
# Methodology



## Research understanding Phase

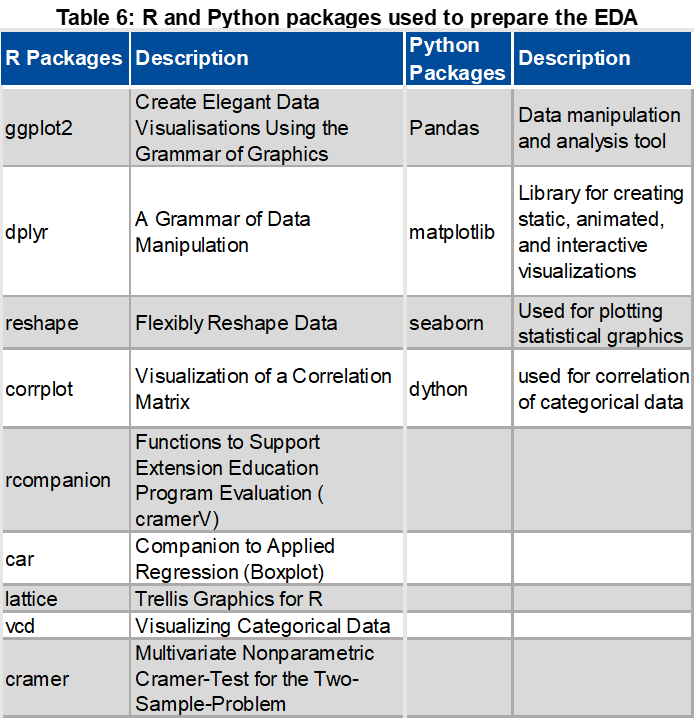
In this study we analyzed a Portuguese banking dataset for a direct marketing campaign (phone calls campaign) for their term deposits product. The data is available on: UCI Machine Learning Repository: Bank Marketing Data Set. The main goal of this study is to reduce the direct marketing campaign cost and time by finding a model to predict the subscriber for the term deposits, but an approximately number of successes (clients subscribing the deposit) should be kept.

## Data Preparation Phase

The preparation of the dataset is done as data was accessed and imported as an CSV file using R and Python software. The data doesn’t contain any missing values nor duplicates, outliers and imbalance classes were detected so that two balancing techniques applied in the preprocessing phase (SMOTE and Random-Under sampling).

## Data Understanding Phase

Exploratory Data Analysis (EDA) is applied as an approach for data analysis that provides maximum insight into a data set. R and Python packages and methods were applied to employ various types of exploratory analysis. Table 6 shows a brief description of each used package.



## Data Preprocessing Phase

In this Phase, we know that we have an imbalanced dataset. Therefore, there is a need to prepare the data for the next step where we balance the data using the mentioned two techniques (SMOTE and Random Undersampling). In the stage a categorical variables transformation will be applied using two functions in python, get dummies () which is one-Hot-Encoding method to convert categorial column to numerical one and LabelEncoder () in sklearn package in Python which can also be used to transform non-numerical labels (if they are hashable and comparable) to numerical labels. LabelEncoder () function will be used to transform the response variable (y) and the Education variable.

## Modeling Phase

### Cross validation method

Cross-validation is a data resampling method to assess the generalization ability of predictive models and to prevent overfitting (Berrar, 2019). The aim in cross-validation is to guarantee that every item in the original dataset has the same chance of appearing in the training and test set. The training set is used to train the model and the test set is used to estimate the error rate of the trained model.

In this work, the splitting technique is holdout method cross validation which will be applied on the data as train set (66.7%of the dataset) and test set (33.4% of the dataset). Prusty (2013) and Moro, Laureano and Cortez (2011) used the same technique on this dataset, In addition to that it is one of the simplest ones and fit well when the data is large. I decided to take the same ratio as Moro (2011)

### Feature Selection

In this step the feature selection techniques that will be performed are, forward selection, correlation based and the Chi-Squared-based.

In forward selection, we choose our independent variables to be used in the predictive model, starting with no variables at all, and then add the variables and add one variable at a time, by choosing the one which has the highest R-squared. By the number R-squared, we mean the number that indicates the proportion of the variance in the dependant variable which is predictable for that specific independent variable. on other word, let’s say if the independent variable has R-squared equals to 1, that means that it explains all the variations in the dependant variable. If R-squared is equal to zero, this means that it explains none of the variation in the dependent variable. at the second step, we select the independent variable which increases the R-squared the most and we keep adding independent variables until none of the remaining variables are significant and don’t add any increasing in the R-squared.

The reason of choosing this method that it fits perfectly with the regression problems. As the response variable “y” in the Portuguese dataset can be transformed to a binary discrete variable and deal with the problem as a regression problem. Furthermore, it is one of Wrapper-based feature selection techniques that I want to compare it with the Filter-based feature selection technique (correlation based and the Chi-Squared-based). From the name “Correlation based” we can tell that the feature selection will be based on the correlation with the response variable. Chi-Squared-based is a statistical test of independence to determine the dependency of two variables. It shares similarities with coefficient of determination R-squared, However, chi-square test is only applicable to categorical or nominal data while R-squared is only applicable to numeric data (Jiarpakdee, Tantithamthavorn, and Treude 2018).

### Data Balancing

As we mentioned earlier the Smote and the random under sampling techniques will be adopted to balance the data. Smote is technique that generates artificial examples of the minority class by interpolating existing instances that lie close together. A certain number of instances from the k nearest neighbors are randomly chosen based on the amount of oversampling.

Smote proposed by Chawla, Bowyer, Hall, and Kegelmeyer, (2002) to solve the overfitting problem that might happen when applying the random over-sampling (ROS). ROS is the simplest strategy to expand the minority class, which corresponds to a non-heuristic method that balances the class distribution through a random replication of positive examples.

Random under sampling techniques selects randomly from the majority class until the data is balanced. It reduces the run time, but it could lead to loss of useful information Mishra, (2017).

### Classification Algorithms

Classification algorithms work by predicting the best group to which a data point belongs to by “learning” from labeled observations. It uses a set of input features for the “learning” process. Classification algorithms are good for grouping data that are never seen before into their various groupings and are therefore extensively used in machine learning tasks. Some of the well-known classification algorithms used in this paper are briefly discussed below:

Decision Trees: it is basically partitioning the data samples into two or more subsets so that the samples within each subset are more homogeneous than in the previous subset. There are two kinds of decision trees: classification trees and regression trees. A decision tree can be described as a flow-chart like structure in which internal node represents test on an attribute, each branch represents outcome of the test and each leaf node represent decision taken after computing all attributes or a response after computing all given attributes. Two reasons for the popularity of decision tree techniques; the procedures are relatively straightforward to understand and explain, and the procedures address a number of data complexities, such as nonlinearly and interactions, that commonly occur in real data (Derrig and Francis, 2006).

Logistic Regression: it is a linear model that studies the association between a categorical dependent and a set of independent variables. Logistic regression is used when the dependent variable is nominal binary Elsalamony (2014). LR predicts the probability that an observation falls into one of the two categories of the dependent variables. In a classification problem it predicts the class of the dependent variable.

Random Forest: it is a supervised classification algorithm as it is an ensemble which made from decision trees. The advantage of it that it increases the accuracy of the model as it combines the simplicity of the decision trees with flexibility. Random Forest can be created by first create a bootstrapped dataset from the original dataset by randomly adding rows (samples) from the original dataset to the bootstrapped dataset. It is important to mention that the bootstrapped dataset size should be equal to the original dataset size. Also, it allows adding rows (samples) more than once (duplicates) into the bootstrapped dataset. After that, A creation of number of decisions trees will done by using random subset of variables(columns) in each step to decide what variable will be picked for each node in each tree. After all trees are built up, the classification of any new input will be run on all the trees as each tree will give a classification (output). The result of all trees will be aggregated then the forest will choose the classification with the majority of the “votes”. When using Random Forest for regression, the forest picks the average of the outputs of all trees. Random forest technique is called bagging because it uses a combination of bootstrapping and aggregation methods. We can improve the performance of the random forest by tunning the number of variables used to create the subset that will be used to pick a variable for each node.

### Evaluation and Validation Phase

The performance of the models on the test set will be measured using different approaches such as an Accuracy, Sensitivity, Precision, and Area Under the Curve (ROC).

Accuracy is calculated as the number of all correct predictions divided by the total number of the dataset. It is important to mention that the accuracy measure doesn’t suite the imbalanced dataset. Accuracy ranges from 0 to 1, where 1 is the best and 0 is the worst.

\mathrm{ACC = \displaystyle \frac{TP +TN}{TP + TN + FN + FP} = \frac{TP + TN}{P + N}}

Sensitivity is also known as Recall or true positive rate (TPR) which calculates the true positive rate by dividing the correct positive predictions over the actual total number of positives. Sensitivity ranges from 0 to 1, where 1 is the best and 0 is the worst. In our model it will measure the performance of the model to predict the subscriber to term deposits.

\mathrm{SN = \displaystyle \frac{TP}{TP + FN} = \frac{TP}{P}}

Specificity is also called true negative rate (TNR) that defines the true negative rate by dividing the correct negative predictions over the actual total number of negatives. Specificity ranges from 0 to 1, where 1 is the best and 0 is the worst. In our model it will measure the performance of the model to predict the non-subscriber to term deposits.

\mathrm{SP = \displaystyle \frac{TN}{TN + FP} = \frac{TN}{N}}

Last but not least, Area Under the Curve (ROC). The Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity which is True Positive Rate) is plotted in function of the false positive rate for different cut-off points. False positive rate is the proportion of the samples which wrongly classified as positives by the classifier. It is also calculated by 1-Specificity. Each point on the ROC curve represents a sensitivity/(1-specificity) pair corresponding to a particular decision threshold, so it will let you know what threshold would perform better. Where the Area Under Carve ROC can be used to decide which model is better. This area is always represented as a value between 0 to 1. Better AUC closer to 1.

For example, if we have two classifiers A and B, by comparing the AUC we can tell that classifier A performs better than B.

