[Identifying Default of Credit Card Payments]

Vikas Virani – s3715555

Master of Data Science, RMIT University,
Melbourne, VIC 3000
(s3715555@student.rmit.edu.au)

Salina Bharthu – s3736867

Master of Data Science, RMIT University,
Melbourne, VIC 3000
(s3736867@student.rmit.edu.au)

COSC 2670/2738 | Practical Data Science
Assignment 2: Data Modelling and Presentation

Date: 30-05-2019

Table of Content

1. **Abstract……3**
2. **Introduction……3**
3. **Method……..3**
   - **3.1. Data Retrieval……3**
   - **3.2. Data Preparation……4**
   - **3.3. Data Exploration……4**
   - **3.4. Data Modelling……..10**
4. **Results……11**
5. **Discussion……..12**
6. **Conclusion……..12**
7. **References……..12**

Abstract:

Nowadays, granting money through the credit card is one of the leading domains in the financial business. At whatever point banks issue credit cards to clients, it includes hazard with regards to the likelihood of the client's inability to pay the bill. The fundamental objective of this study is to determine the ability of customer to repay the utilized credit of the bank. By considering individual data and record of payments, the likelihood of default payment can be inferred. By foreseeing precisely which clients are most plausible to default, the bank can improve its credit card services for the mutual benefit of clients and the business itself.

Key words: Default payment, Classification, KNN classification, Decision tree classification, Feature Selection, Resampling, Parameter Tuning.

Introduction:

While demand of credit card is increasing exponentially, it has become essential to confirm the eligibility of customer by issuer, as the crisis resulted by inability of client to repay credits can effectively challenge both banks and clients. Moreover, the credit limit should also be determined precisely in order to avoid future hazards. In pursuit to find out whether customer is eligible to pay the bill or not, the personal details along with history (payment of last 6 months) needs to be taken into consideration.

This report outlines the investigation performed on financial details such as balance limit and repayment records and customer transactions as well as personal data such as sex, education, marital status and age. There are several hypothesis determined from available data which endorses the significance of attributes on default payment. This analysis is followed by modelling the available data to emphasis the conclusion by classifying it into default payment categories. This would likewise enable the issuer to have a prior knowledge of the potential of present and potential clients, which would illuminate their future methodology.

Method:

To model the dataset to predict the chances of default payment, initially, data needs to be retrieved from data source followed by data preparation and data exploration.

1. Data Retrieval:

Firstly, data is extracted in the form of csv file from the below mentioned data source.

Data Source -> UCI machine learning repository (https://archive.ics.uci.edu/ml/machine-learning-databases/00350/)

Data description:

The dataset used for this investigation beholds 30,000 observations spread across 24 variables. It contains detail of individual customer such as gender, education, marital status, age and financial details such as credit limit, payment status of last 6 months, bill amount and amount paid in last 6 months and default payment status.

As per data source, below is the list of 25 variables of dataset:

X1: Amount of the given credit (NT dollar)
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age in years.
X6 - X11: History of past payment from April to September (X6 = the repayment status in September, 2005 … X11 = the repayment status in April, 2005 where -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above)
X12-X17: Amount of bill statement from April to September (NT dollar) (X12 = amount of bill statement in

September, 2005 … X17 = amount of bill statement in April, 2005.)
X18-X23: Amount of previous payment from April to September (NT dollar) (X18 = amount paid in September, 2005 … X23 = amount paid in April, 2005.)

Y - Default payment (Yes = 1, No = 0)

While importing the data, the suitable variable names are provided to dataset columns. Also, after the retrieval, dataset is validated against source data.

2. Data Preparation:

After data retrieval, data is checked against missing values and datatypes of columns. There are no missing values present in any variable. To handle outliers present in data, outliers of each columns are removed based on absolute Z-score value approach. If the value of Z-score is greater than 3, it is filtered out from the dataset. Moreover, the further analysis is done on categorical variables for impossible data. There are few impossible values present in categorical variables, which are handled by replacing it with nearest possible value.

For education variable, there are 4 categories (1 to 4) whereas, there are impossible values such as 5,6 and 0 present in dataset. This values are labelled to others category (category 4). Similarly, marital status having 0 value is labelled to 3(others). The payment status variables for all 6 months are also having such impossible values (-2 and 0) which are replaced by labelling it to -1(duly paid).

Additionally, after converting nominal variables from numeric to categorical, the summary statistics (such as minimum, maximum, mean etc.) of numerical variables and summary (unique values, highest occurring value and its frequency) of categorical variables are provided.

3. Data Exploration:

Exploration of single variables:

As data exploration provides important insights regarding hidden trends of data, the exploration of each variable of dataset is performed. Initially, for all categorical variables (sex, education, marital status, Payment status of 6 months (that is Pay Sep, Pay Aug, Pay July, Pay June, Pay May, Pay Apr), default payment), bar chart is plotted to represent the values divided across different categories.

Following trends of the dataset can be inferred from bar plots:

- Dataset contains higher number of female customers.
- Customers having university education are relatively higher followed by graduates.
- There are more single customers as compared to married.
- Relatively higher number of customers have duly paid bill amount each month.
- Approximately 30% customers are defaulters.

While Histogram best represent the density and distribution of continuous numerical variable, boxplot precisely depicts variance and outliers. To represent numerical variables of dataset, histogram along with boxplot is used.

Following insights can be derived from visualization of numerical variables:

- Though the age range of customers is 20 to 65, it is evident that most of the customers are between age of 25 to 45.
- The balance limit varies from 10000 to 550000. However, there are maximum customers falls within the range of 50000 to 250000. There are few customers having balance limit falling in range of 250000 to 550000.
- The bill amount of all 6 months follows relatively similar curve ranging from -95000 to 260000. The frequency of customers is higher within the range of 1000 to 55000.

- Similarly, the amount paid in previous 6 months follows relatively similar curve ranging from 0 to 70000. The highest frequency is under 0 to 10000. There are few customers falling beyond the range of 3rd quartile.

Exploration of pairs of variables to find relationship:

After exploring single variables, pair of variables are explored to determine the relationships and investigate hypothesis.

Below mentioned hypothesis are explored:

- **Does education have any impact on chances of default?**
  To compare the categories of education by default payment, the count graph is plotted.

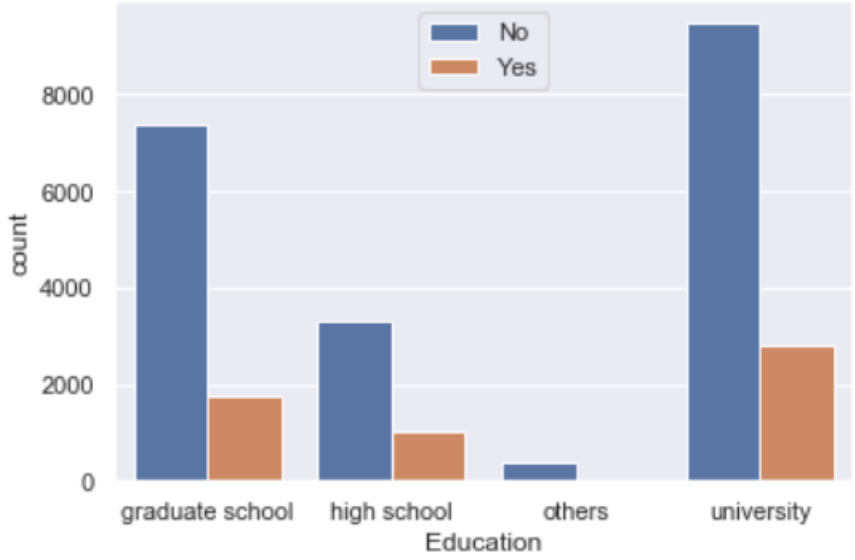


Figure 1- Observation of default payment by education

From the above visualization of education and default payment, it is inferred that, there are higher chances of default for university graduate as compared to school and high school graduate respectively. Therefore, the education of card holder does not impact the chances of default payment.

- **Does Bill amount for default cards is higher?**
  To infer the relation between bill amount and default payment, the bar chart is plotted.

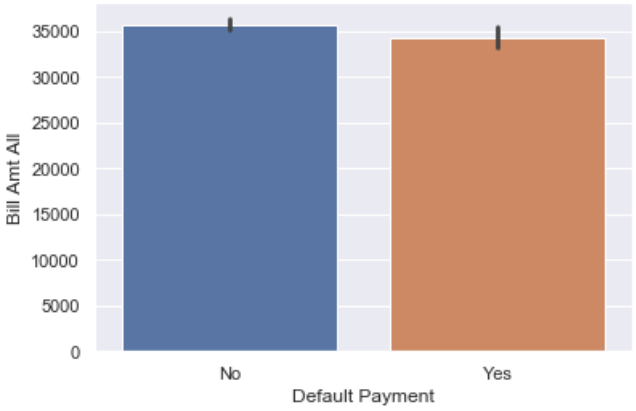


Figure 2- Observation of Bill amount by default payment

As comprehended from the above bar plot of Bill amount and Default payment, the bill amount is slightly higher for non-default cards as compared to default cards.

- **Do the chances of default vary in different age group?**

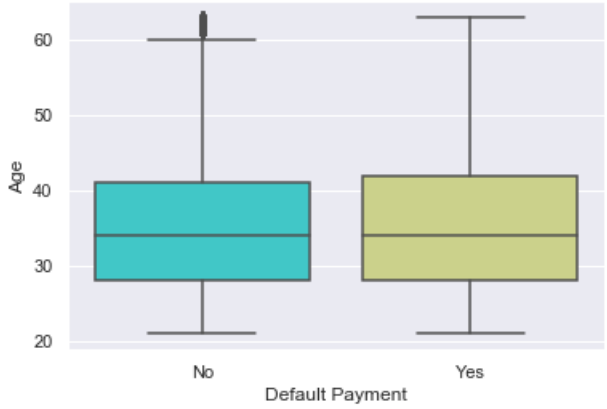


Figure 3- Box plot of default payment by age

It can be inferred from the box plot of Default payment by age that, the chances of default are more in age group of 25-45. This observation is very similar for non-default card holders as well, which concludes that there is very loose relationship between age group and chances of default.

- **Does marital status have any impact on chances of default?**

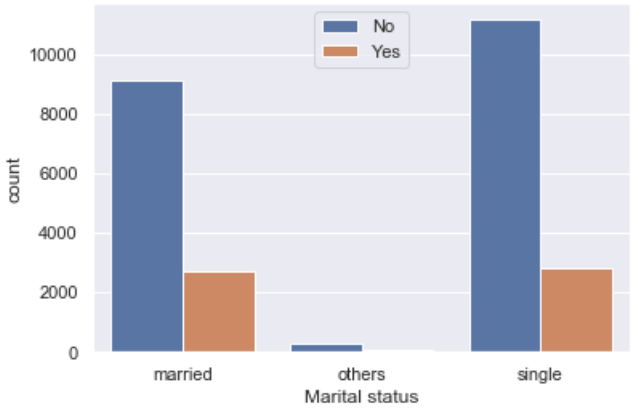


Figure 4- Observation of marital status by default payment

As depicted from the count plot, the frequency of married defaulters and single defaulters is almost similar. Therefore, there is no major impact of marital status on chances of default.

- **Does having a delay in previous payment impact chances of default?**

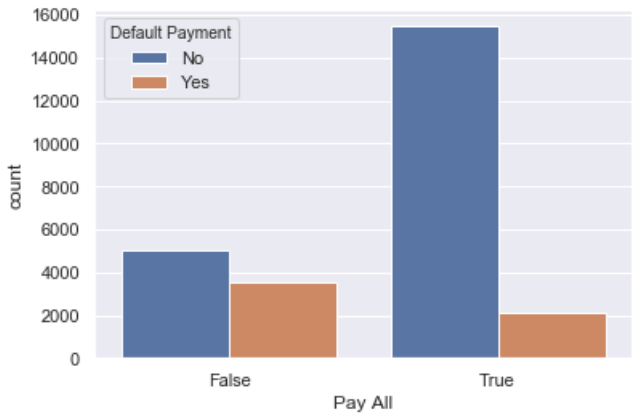


Figure 5- Observation of default payment by previous payment status

Above plot shows 'False' value of Pay All for delay in previous payment and 'True' value for duly paid previous amount status. While comparing the status of Payment with Default payment, it can be observed that, the proportion of Defaulters is much higher when payment status is 'False'. Therefore, the chances of default are increased when there is a delay in previous payments for even 1 month.

- **Does the chance of default vary in male and female?**

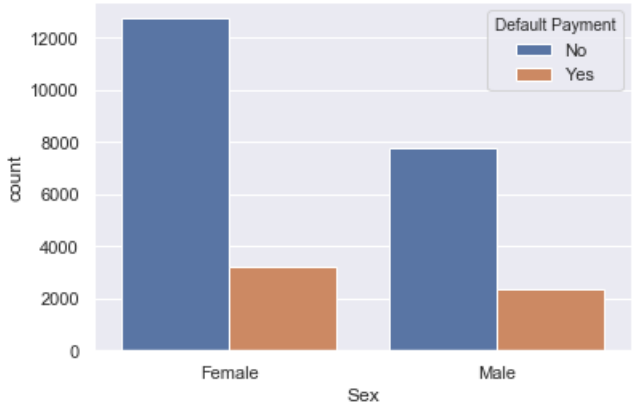


Figure 6- Observation of default payment by gender

As inferred from the bar plot above, number of female customers are more probable to default as compared to male.

- **Do married people spend more than others?**

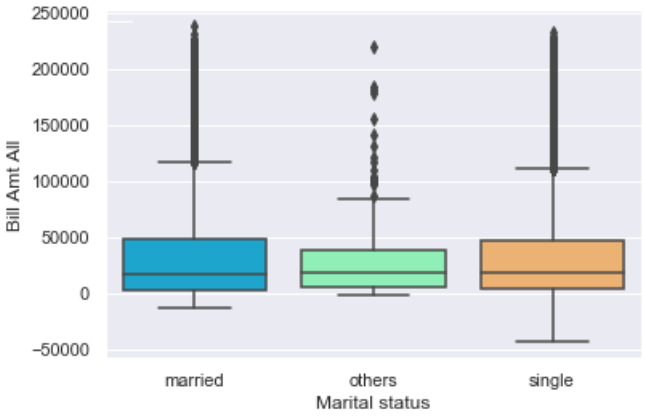


Figure 7- Boxplot of Marital status by bill amount

As depicted from the boxplot of marital status by bill amount of 6 months, bill amount range of married card holders is slightly higher than single and other card holders respectively. This concludes that married people tend to spend more than others.

- **Do women spend more money than men?**

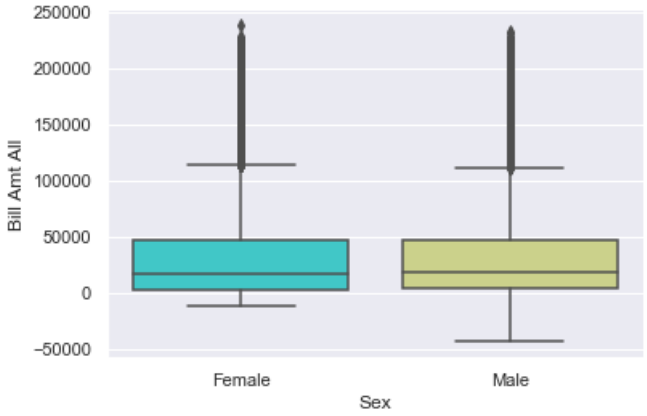
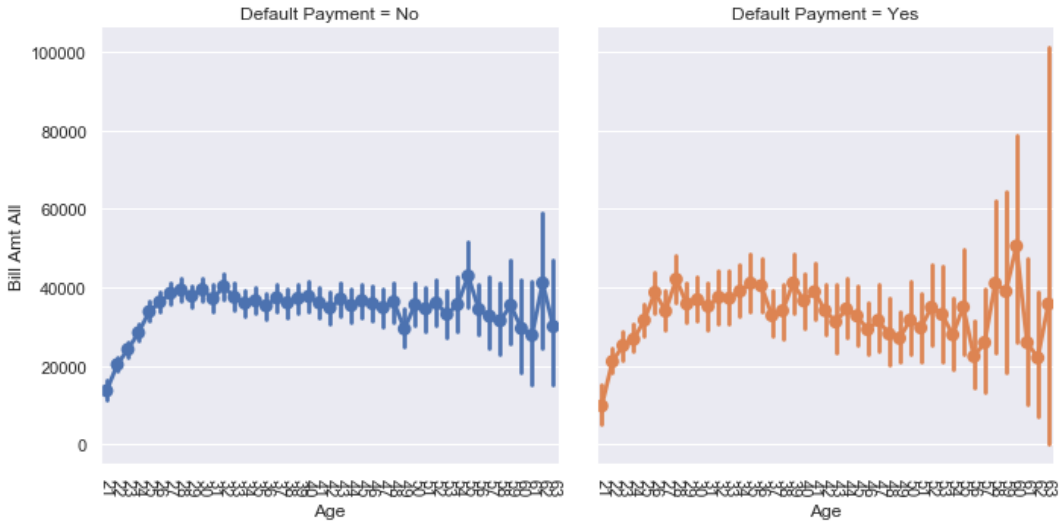


Figure 8- Boxplot of gender by bill amount

Above boxplot compares gender and bill amount of 6 months. The range of bill amount of men is approximately same as women, which shows that spending habits does not vary in gender.

- **Do spending habits varies in different age group?**



As can be seen from the graphs, people below the age of 24 tends to spend less money overall. For rest of the people, spending habits are equally distributed except for some high spending by higher age group.

- A New Attribute, "**Pay All"** is created, the value of which will be **TRUE** if a customer has made payments on time for all the 6 months otherwise **FALSE**.

- **Do Payment delays are dependent on Age, Gender or Marital Status of the Person?**

  To investigate the relationship between payment delays and individual details (age, gender and marital status) bar plot and box plot are used.

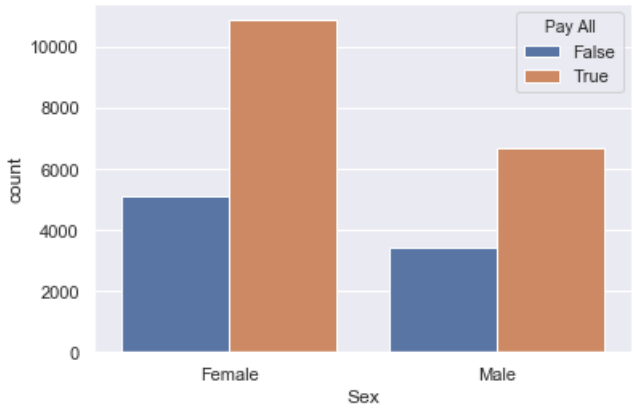


Figure 9- Observation of payment status by gender

It is comprehended from bar plot of gender and payment delays that, more women have paid the bill in time for all last 6 months.

Figure 10- Observation of payment status by marital status

Moreover, while comparing marital status with payment delays, there are higher number of single card holders having made payment on time for all 6 months as compared to married and others.

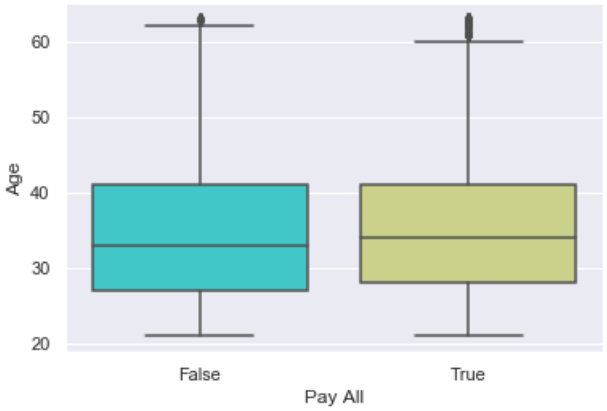


Figure 11- boxplot of payment status by age

However, payment status follows similar trend in all age groups, there are few card holders beyond the age of 60 having made all payments on time.

4. Data Modelling:

As a part of data modelling, we are building a model which can predict probability of default payment for unseen data. As classification aims at categorizing the target data to which category it belongs, there are two different classification algorithms used for data modelling.

We have used a **F1-score feature selection technique** to identify the best features in the dataset & plotted the graph for the same. Based on the graph, we have selected best 13 features from the dataset and applied following techniques to build a model.

To model the data, **KNN classification and Decision Tree classification** algorithms are used. Initially the dataset is split into predictors and dependent variable followed by dividing data into two groups, one for training the model and one for validating the model. The data is divided in different proportion of training and test data (that is, 80% data for training and 20% data for testing, 60% data for training and 40% data for testing and 50% data for training and 50% data for testing), in order to achieve higher accuracy. The splitted training data is further fed into below mentioned classification algorithms.

- **KNN Classification:**

KNN classification algorithm classify each observation using similarity of the surrounding observations.

- **Decision Tree Classification:**

Decision tree algorithm works by splitting training data into subsets in such a way that each subset holds identical value or characteristic for some attribute. This process is recursively executed until the leaf node holds classified data.

As our dataset was imbalanced in terms of target values, we applied **resampling technique; oversampling**, on our train split of data to make our data balanced in each iteration of train test split. We have up sampled the minority class in the ratio of **1:2 of minority class to majority class**. It is important to apply resampling techniques only to train data **after splitting the data into train & test**, as applying this to whole dataset may lead to having **same observations in train and test** dataset, which might cause **overfitting**.

As the performance of classification model is determined by the choice of parameters, the precise combination of classifiers' parameter needs to be chosen. The algorithm is tuned by using **GridSearchCV** that aimed to maximize the prediction accuracy by selecting the **best suitable values for classifiers' parameters** to do **Parameter Tuning**. Therefore, by applying classifier along with GridSearchCV, the model is trained to handle unseen data. Now, test data is fed into model and target value is predicted. To validate the performance of this classification model, the confusion matrix is created. The confusion matrix is derived by analysing values of predicted data with the values of test data of dependent variable and it shows count of correctly and incorrectly identified defaulters. To further validate the model, classification report is prepared. The report beholds **precision** (ratio of true positives to the sum of true and false positives), **recall** (ability of a classifier to find all positive instances), **F1 score** (mean of precision and recall) and support (actual occurrences of each category). Also, **accuracy score as well as error rate** for data model is also generated in order to find efficiency of model.

Results:

To concur the requirement of preparing data model to successfully identify probability of default, the KNN classification and Decision tree algorithm is used. As mentioned in the method section, to validate the model, **confusion matrix, classification report, accuracy score and error rate** is generated.

Below table shows the accuracy score and error rate of classification models for different proportion of train- test data:

Classification Algorithm	Training- test data proportion	Accuracy score	Error rate
K nearest neighbour	80% - 20%	0.742	0.258
K nearest neighbour	60% - 40%	0.726	0.274
K nearest neighbour	50% - 50%	0.732	0.268
Decision Tree	80% - 20%	0.800	0.200
Decision Tree	60% - 40%	0.803	0.197
Decision Tree	50% - 50%	0.789	0.211

It can be inferred from the model we explored that, **Decision Tree gives higher accuracy** than K nearest neighbour in our dataset.

Discussion:

We scaled the data using **MinMaxScaler()** to set all features in the same scale. **F1 Score technique** among various feature selection techniques was used to select the best features of dataset. We have used **GridSearchCV()** to build a model using Decision Tree & K nearest neighbour. As GridSearchCV internally uses a cross validation, we didn't need to do it separately and also we identified the best combination of values for "criterion" and "max_depth" for Decision tree & "n_neighbors", "weights" and "p-value" for KNN. By observing the performance of both models based on various metrics like "Confusion Matrix", "Error Rate", "Accuracy", "Precision", "Recall" and "F1-Score", we identified that, overall Decision Tree is more suitable than KNN for our Dataset.

Conclusion:

This investigation deals with effectively classifying unseen data to find probability of default payment. This document has discussed KNN and Decision tree classification algorithms in order to achieve the high accuracy in classifying data. From the result of models we explored, it can be **concluded that Decision tree algorithm gives higher accuracy** as compared to KNN. As decision tree works best with larger datasets and is more deterministic as compared to KNN, it is highly suitable for our dataset with large volume.

In general, **Tree based algorithms** or **Ensemble** techniques like **boosting**, works **best for imbalanced dataset** as their hierarchical structure gives patterns for all types of targets regardless of their proportions.

References:

[1] Matplotlib library: https://matplotlib.org-index.html

[2] https://python-graph-gallery.com-24-histogram-with-a-boxplot-on-top-seaborn-

[3] Seaborn library: https://seaborn.pydata.org-index.html

[4] Scikit-learn library: https://scikit-learn.org/stable/