PREDICTING CREDIT CARD DEFAULTS

GROUP: H
Department of Biochemical Engineering
University College London
London, WC1E 6BT

I. INTRODUCTION

Credit card default has been a problem in Banking and Financial Services, which has caused tremendous loss to a high volume of financial institutions. There have been numerous reports on the significant adverse effects of credit card defaults on residence life quality and a country's economy. The high number of credit card defaults result in losses in the value of families savings and pensions, indirect impacts on employment rate, and a general increase in credit interest rates. The reduction and prevention in credit card defaults are essential to prevent the discussed adverse effects and the general life quality of the debtors. A wide range of solutions has been researched and implemented to minimise the violations of credit card contracts. The most novelty and efficient method is deduced from the application of Machine Learning algorithms to classify and predict credit card default risks of clients.

This report describes a novel methodology derived from a wide range of research papers with modification and additional classification algorithms to conclude and improve the general accuracy of using Machine Learning in predicting credit card defaults. The methodology used in this paper produces one of the highest accuracy results compared to recent research papers, which results from an in-depth understanding of clients' default data characteristics, accurate and necessary data transformation in pre-classification. The highest classification test result obtained is 0.82967, while the test result for using an optimal classification algorithm with high computational speed is 0.82583. An insignificant difference in test accuracy is compensated for noticeable processing time.

This report is organised as the following: Section I introduces into the toping and work done described by this paper; Section II demonstrates Exploratory Data Analysis and visualisation; Sector III illustrates and describes in detail the implemented Methodology; Section IV describes Training and Evaluating the classification model; Section V concludes results from cross-validation of classification model; Section VI demonstrates displays the final results classification methodology achieves in predicting credit card defaults; Section VII concludes findings and analysis of this paper.

II. DATA TRANSFORMATION AND EXPLORATION

When combining the training set and test set as data for EDA to help understand the distribution of data by determining a descriptive statistic of the data. It allows us find the correlation between variables and a normality test is also performed, plus a relationship between target variables and independent variables.

Observations:

Defaults have a higher proportion of Lower LIMIT_BAL values
NonDefaults have a higher proportion of Females (Sex=2)
NonDefaults have a higher proportion of MoreEducated (EDUCATION=1 or 2)
NonDefaults have a higher proportion of Singles (MARRIAGE=2)
NonDefaults have a higher proportion of people 30-40years
NonDefaults have a MUCH higher proportion of zero or negative PAY_X variables (this means that being current or ahead of payments is associated with not defaulting in the following month). This is a strong relationship as the distribution are more separated - so we expect the PAY_X to be important

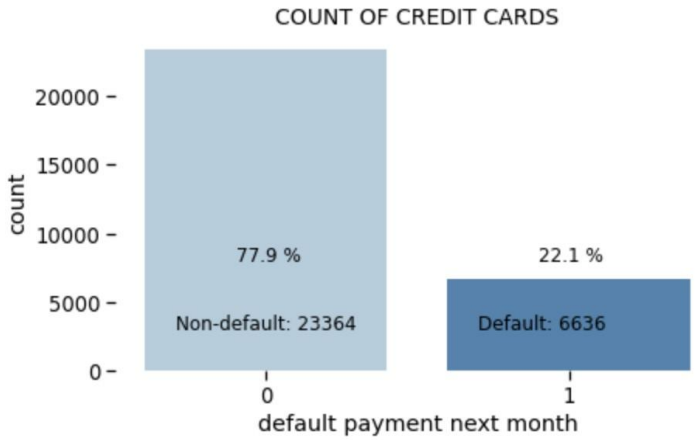

Figure 1: Credit card customer - target value.

Figure 1 shows 23364 customers (approx. 77.9% of the data set) will not default next month's payment. While 6636 customers (approx. 22.1% of the data set).

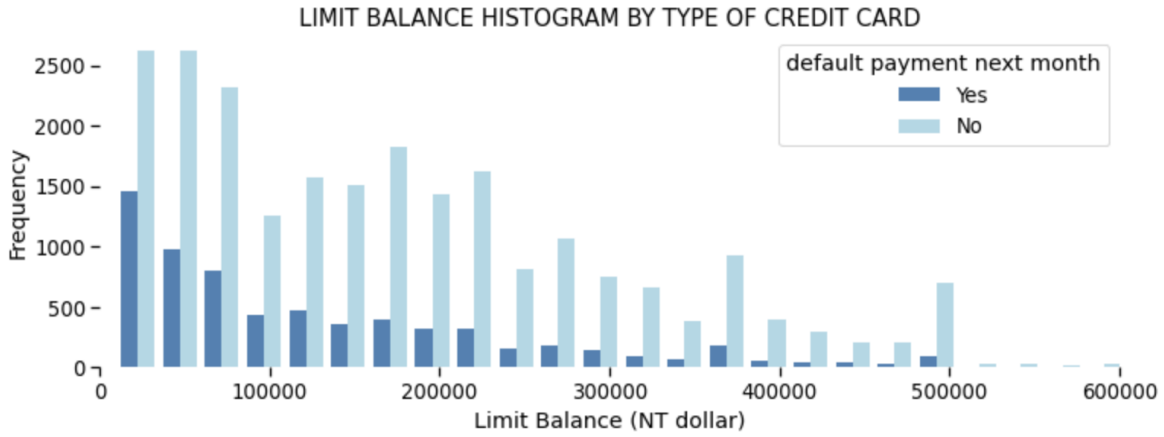

Figure 2: Non default and Default payment next month for each specific Limit group.

The LIMIT_BAL (Limit balance) is a reasonable choice for default behaviour, because generally people who have a good credit score are more likely to be given a higher credit than those with a lower credit score. Plus, people who have default payment in the past will find it harder to raise their credit lines. Thus, the number of default payments with credit lines in the top 25% should be smaller than those with low credit lines. Figure 2 represents the number of defaulters for limit balance, which supports the statement between the number of defaults and credit limit.

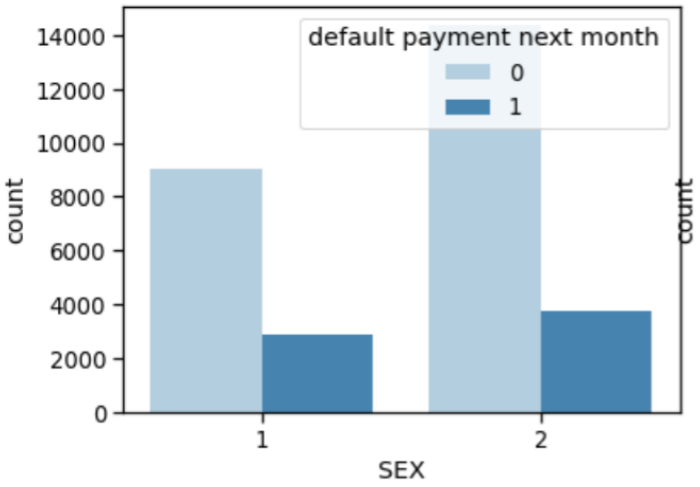

Figure 3: Default payment against Gender.

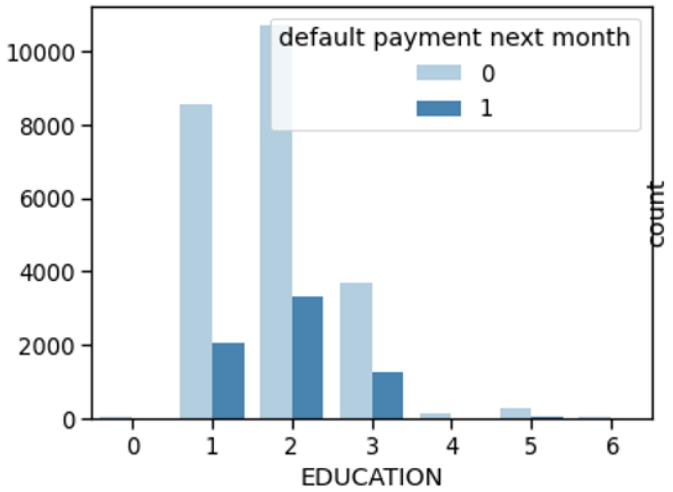

Figure 4: Default payment against education.

Next, we checked the impact of sex on the number of default payment, and from Figure 3 is tells us there are more women that default payment next month than men. However, what it doesn't tell us is the reason behind this observation, nonetheless this result is included in our training data model. From Figure 4 there's no significant evidence to identify education has a major impact on default payment next month, however there's a belief that people with higher education tend to pay their bills on time based on the assumptions they have a higher paying job.

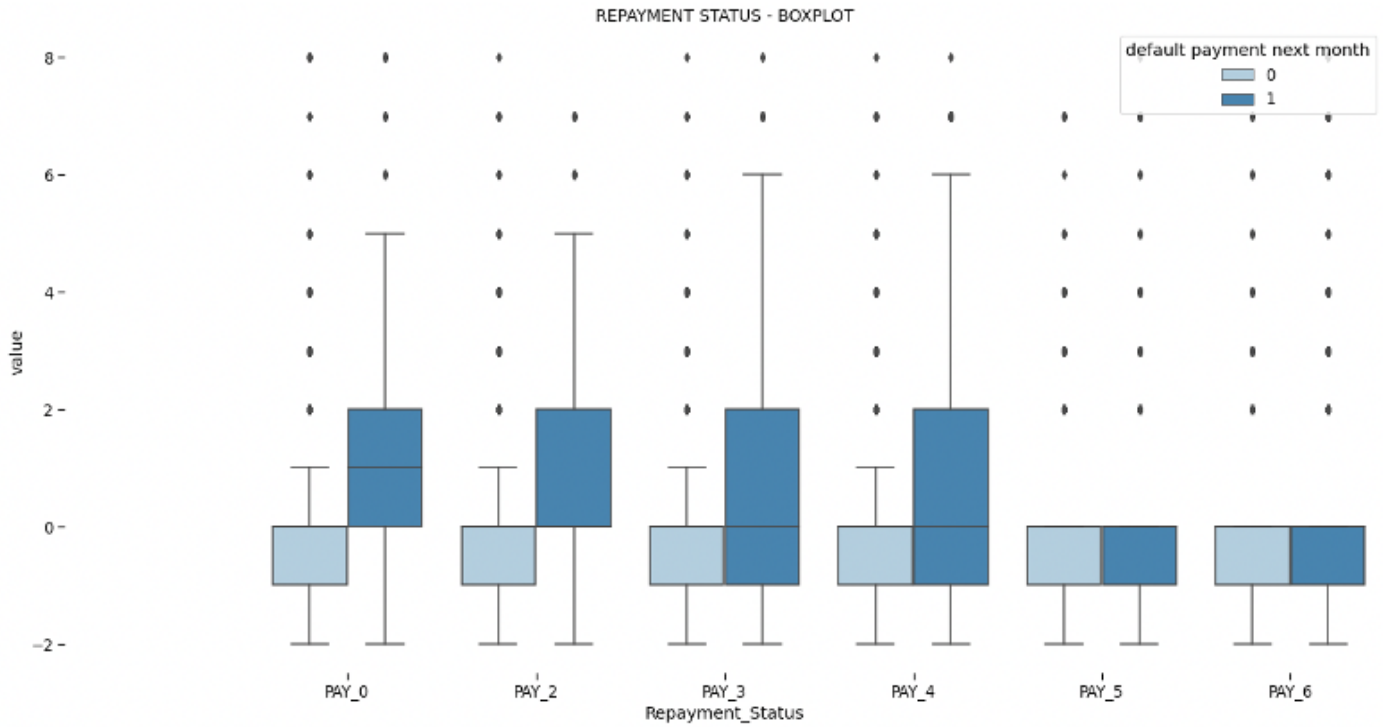

Figure 5: Payment status for default repayment next month.

Looking at PAY to give us some insight for data features for default payment, from the get-go we can see PAY_0 (repayment status in September) and PAY_2 (repayment in August) has more

discriminatory power than the rest of the months. As this is a good indicator if the customer can continue with payment for the rest of the months.

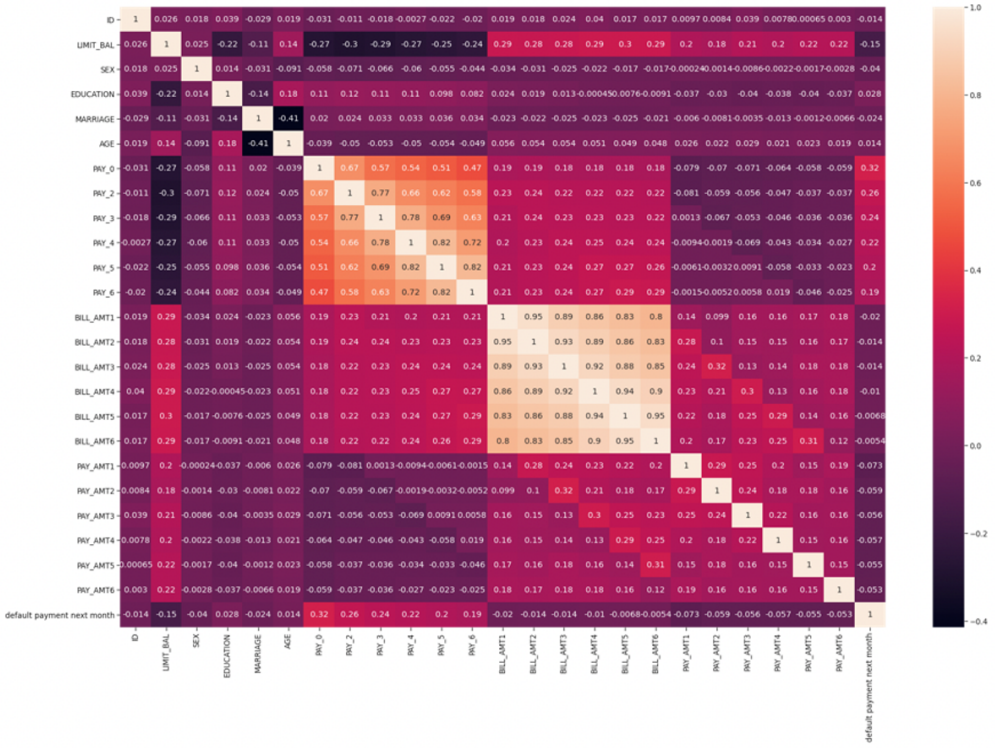

Figure 6 - Correlation between independent variable and target variable Heat map.

Now checking the correlation between the independent variables and dependent variables, and plotting a heat map, Figure 6, the target variable default payment next month depends more on PAY variables compared to the rest. However, the other features are still important and no information should be neglected.

III. METHODOLOGY

A. Overview

High accuracy in predicting credit card defaults is necessary and crucial, which is this project's main goal and objective. Novel classification approaches and methods are researched from multiple state-of-art research papers and conducted to enhance prediction accuracy and produce an excellent outlook on credit card default sectors. Furthermore, enhancement and optimisation are later attempted and applied to improve both classification test results and the data analysis process. Similar to the nature of Machine Learning, our testing accuracy is computed from the countless attempts in terms of human manual configuration and computed machine trials.

A wide range of Machine Learning classifications is applied and cross-validated to conclude the most efficient and suitable in predicting credit card defaults. The choice of classification algorithms depends on conclusions made on data visualisation/exploratory data analysis and model selection process. Due to the nature of credit card default datasets being labelled, Supervised learning classification algorithms are considered to have higher accuracy and computational speed. These are

the classifications applied to predict credit card default: Logistic Regression, Support Vector Machine, K-Nearest Neighbours, Gaussian Naive Bayes, Decision Tree and Random Forest.

Data analysis and systematic transformation is conducted to improve the classification process and results as the dataset consists of various distinct data types and characteristics. Unarguably, an unimportant and non-deciding variable *"ID"* is removed to eliminate false interference into the classification process. Nonetheless, converted categorical values in numerical representation, as *"EDUCATION"* and *"MARRIAGE"* in {1, 2, 3, 4} and {1, 2, 3}, affects classification due to ordinal relations of numerical values while the represented categorical values have no connections. Specifically, one can not argue that *"Single"*, *"2"*, is better or worse than *"Married"*, *"1"*. Hence, these numerical representations are processed using the One Hot Encoding technique to disregard the ordinal relation of representing numerical values and represent categorical values via binary variables (dummies variables). Data scaling is implemented to reduce modelling difficulties and improve classification results, as large weight values resulting from massive inputs causes the model to be unstable and sensitive to input values, which results in high generalisation error. As input variables have characteristics lightly similar to Gaussian distribution, the standardisation method of data scaling is preferred. Specifically, data standardising to mean of 0 and standard deviation of 1 is applied to solve complications induced by vast differences in ranges and characteristics of input variables. This data scaling application improves classification algorithms based on distance measurement between input values such as Support Vector Machine, K Nearest Neighbors, etc. However, algorithms classifying input data on rules, particularly tree-based algorithms, is not improved via standardisation.

B. Classification algorithms

1. Logistic Regression

Logistic Regression is a classification algorithm that applies the distinct logistic function to represent binary dependent variables [1]. The classification algorithm studies the correlation between the dependent variable, default binary output, and a dataset of independent variables and embeds it into the logistic function's weight. Regression is considered due to its high computational speed and non-correlation characteristics between independent variables. The logistic function is described as below [1]:

$$p(y = 1|x) = \frac{1}{1+e^{-w \cdot x}}$$

2. Support Vector Machine

The operation of the Support Vector Machine lies in mapping an independent dataset into a higher dimensional space that can be separated and classified [1]. Dataset separator is concluded and further developed and drawn through further data transformation using Radial bias function (RBF) kernel. SVM hyperparameter tuning (including C and Gamma parameter) is implemented through Grid search to improve the classification result, which practically brute-forces combinations of parameters on the training data. Due to the training dataset having high-dimensional characteristics, the Support Vector Machine effectively trains and predicts outputs.

3. K-Nearest Neighbours

The classification of K-Nearest Neighbours lies in the voting mechanism of its nearest k-records, in which one's vote is its class label [2]. Hence, the classified result is the class label of its trained

k-nearest records. This classification method is considered due to the high volume of trained data, covering a high area level.

4. Gaussian Naive Bayes

Naive Bayes classification operates on a probabilistic function with solid assumptions on the independence of the input variables, which is considered naive [1]. Though the datasets' variables have a high probability of not being independent, this classification algorithm is regarded as it has fast computational time and high simplicity. Naive Bayes classification is conducted based on Bayes Theorem, which is characterised as below [1]:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

5. Decision Tree

Decision Tree classification investigates its best predictor selections and best combination strategy to achieve the best model [3]. Predictors assessments are done iteratively to develop a model similar to a tree's growth process. Classifications are made with the input at the root and advancing a specific branch's leaf node, as that tree corresponds to the classification model. Decision tree classification is considered due to its simplicity in visualising and concept, which is a safe choice guaranteeing optimal accuracy.

6. Random Forest

Random Forest is an advancement made in Decision Tree classification as Random Forest operates on classifications made on multiple Decision Tree models [3]. Higher classification accuracy results from lesser sensitivity to problems, i.e., overfitting and bias, especially where individual trees are uncorrelated. Similarly to Support Vector Machine, hyperparameter tuning is required to bring the best accuracy, which is done through a cross-validation process.

IV. MODEL TRAINING AND VALIDATION

The discussed classification algorithms are cross-validated with five folds using the training dataset to assess its accuracy in predicting credit card defaults. The classification algorithm with efficient computational time and the highest precision from cross-validation is applied to the testing dataset, Support Vector Machine and Random Forest. Support Vector Machine is unarguable that it would provide the highest accuracy due to the high-dimensionality of datasets and the classification algorithm novelty. However, Support Vector Machine requires significant time to train datasets and classify, which is not the problem using the second-most accuracy algorithm, Random Forest.

V. RESULTS

The Support Vector Machine (SVM) algorithm was found to be the most accurate classification model out of all the algorithms that were implemented. Overall, SVM had a test accuracy of approximately 0.829667 and the test accuracies in several relatively new research papers were found to coincide with this value as well (~ 0.82). In order to reduce the probability of the model producing false negatives,

hyperparameter tuning was carried out. The optimal parameters for the SVM algorithm are shown below:

	C	gamma	kernel	Accuracy
0	0.1	1.0000	rbf	0.000559
1	0.1	0.1000	rbf	0.270391
2	0.1	0.0100	rbf	0.258473
3	0.1	0.0010	rbf	0.001490
4	0.1	0.0001	rbf	0.000000
5	1.0	1.0000	rbf	0.139665
6	1.0	0.1000	rbf	0.322905
7	1.0	0.0100	rbf	0.315642
8	1.0	0.0010	rbf	0.081750
9	1.0	0.0001	rbf	0.000000
10	10.0	1.0000	rbf	0.211359
11	10.0	0.1000	rbf	0.344693
12	10.0	0.0100	rbf	0.320298
13	10.0	0.0010	rbf	0.270577
14	10.0	0.0001	rbf	0.101862
15	100.0	1.0000	rbf	0.226816
16	100.0	0.1000	rbf	0.362384
17	100.0	0.0100	rbf	0.335196
18	100.0	0.0010	rbf	0.321229
19	100.0	0.0001	rbf	0.117691
20	1000.0	1.0000	rbf	0.231844
21	1000.0	0.1000	rbf	0.376536
22	1000.0	0.0100	rbf	0.339479
23	1000.0	0.0010	rbf	0.326071
24	1000.0	0.0001	rbf	0.262197

Figure 7: Table displaying the optimal values for SVM algorithm

Although the SVM algorithm's precision is rather high, the main drawback is its lengthy running time. The worst classification technique was found to be Gaussian Naive Bayes (NB) with a very underwhelming test accuracy score of 0.278875.

VI. FINAL PREDICTION ON TEST SET

Hyperparameter Tuning:

After completing the hyperparameter tuning procedure, the test score accuracy for the Support Vector Classifier had decreased, while there was an increase in the overall recall score. The Random Forest Classifier, which was shown to be the second-best model had an opposite response to the hyperparameter tuning. The test accuracy score had increased and there was a drop in the recall score. Ultimately, the client (credit card provider) will have to compromise the accuracy, F1, or precision score when selecting the best prediction model.

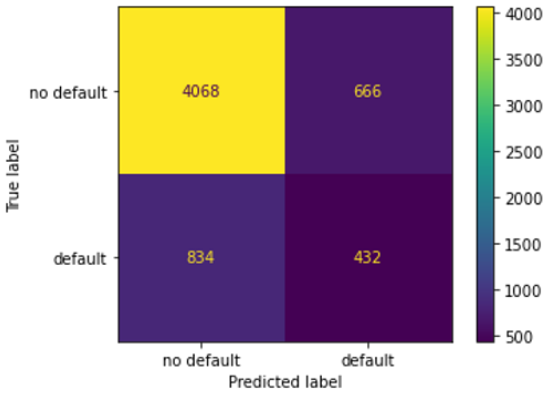

Figure 8: Confusion matrix Support Vector Classifier

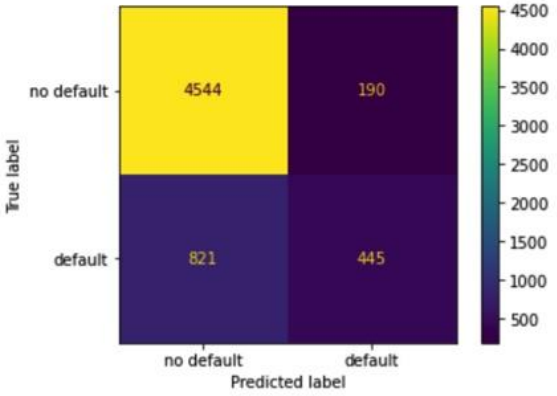

Figure 9: Confusion matrix Random Forest Classifier

From observing the confusion matrix that corresponds to the Support Vector Classifier, it can be noted that the model produced 432 true positive and 4068 true negative values and has an overall data prediction of 0.75. This means that 75% of the test data can be accurately predicted using a Support Vector Classifier that has been trained. The confusion matrix for a Random Forest Classifier was found to have 445 true positive and 4544 true negative values, and has a data prediction value of 0.8315. So, 83.15% of its test data can be predicted properly using a trained Random Forest Classifier.

VII. CONCLUSION

Compared to other research papers, the highest accuracy of 0.82967 is achieved by applying a novel classification algorithm, Support Vector Machine. Another algorithm with significantly faster training and classifying running time, Random Forest, has a slightly lower accuracy of 0.82583. This high accuracy is achieved by correct hyperparameter tuning, datasets transformation pre-classification and profound understanding of datasets through many visualisation graphs. This high accuracy in predicting credit card defaults would significantly prevent Banks and Financial Institutes from receiving losses. In the big picture, it is unarguable that a lower rate of credit card defaults would benefit the economy more or less.

VIII. REFERENCES

[1]C. Bishop, *Pattern Recognition and Machine Learning*. .
E. Alpaydin, *Introduction to machine learning*. London, England: The Mit Press, 2020.
[2]A. Joby, "What Is K-Nearest Neighbor? An ML Algorithm to Classify Data", *Learn.g2.com*, 2022. [Online]. Available: https://learn.g2.com/k-nearest-neighbor. [Accessed: 11- Jan- 2022].
[3]K. Murphy, *Machine Learning A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.