Fraudulent Credit Cards Transactions Detection Using Naive Bayes Classifier

Authors:

Abdulrahman Emad

Amir Hesham

Habiba Salama

Mariam Ahmed

Muhammad Mosilhy

SBE2240: Biostatistics
Prof. Dr. Ibrahim Youssef

Abstract —**In the midst of an increasingly fast-paced and interconnected world, financial transactions play a crucial role in every aspect of our daily lives, however, Credit Card fraud stands out as a threat to the economy and banking systems, As it was fielded by the Federal Trade Commission (FTC) that there were nearly 390,000 reports of credit card fraud in 2021 alone.**

On that account, things like financial fraud are considered to be an ever-growing menace in the financial industry, thus it is of high significance to detect fraudulent transactions from legitimate ones, in order to provide a safer and less risky business for companies.

This study decided to benefit from the Naive Bayes classifier as assistance in decision-making. As it predicts, depending on its input feature, whether the transaction is fraudulent or legitimate, other methods were used in the process of making the model like the Wilk-shapiro test, in order to detect if the dataset is normally distributed or not, and Tukey's IQR Method in order to identify outliers, We reached an accuracy of credit card fraud detection of 91%, then compared it to other Sckikit-Learn Built-in models.

Keywords—Naive Bayes Classifier, Binary Classifier, Machine Learning, Statistics, Python

I. Introduction

Fraud is a typical criminal activity that's defined as "*the intentional perversion of truth in order to induce another to part with something of value or to surrender a legal right*", according to merriam webster, and it poses a threat across the generations, even though the means of transferring goods, selling, and buying changed drastically, fraud in its various forms existed in all of them.

Fraudulent is a problem that undermines the foundations of trust and causes distortion in the economic system, as well as forms a significant threat that ranges from individuals to organizations and governments. so, along with the change in trading from more traditional ways to digitalization and plastic money, credit card fraud came into the picture.

The Nilson Report, which is responsible for monitoring the payments industry, forecasts the losses from card fraud in the U.S. to be a total of $165.1 billion over the next 10 years, Also according to Insider Intelligence, only one type of fraud which is the online, or over the phone and mail transactions would be accounted for an estimated loss of $5.72 billion in the U.S In 2022. on that account, a solution has to be met in order to put an end to the drastic losses caused by credit card fraud.

Machine learning technology is one of the fastest-growing trends in this decade and a powerful tool to be used in prediction and decision-making, it comes in handy in case of problems as complicated as credit card fraud.

Machine Learning has been helpful in various fields, and if we focus on the achievements of machine learning in the world of economy, to mention a few examples we would say things like detecting email spam, targeting product recommendations, and correct diagnosing, The power of machine learning resides in it's increasing power process, availability of huge information, and improvement in statistical modelings, and yet stands the most difficult issue for banks and commerce industry which is the fraud management.

As prediction models and classifiers focus on developing the codes or algorithms to analyze the data and come up with conclusions, we will be a Naive Bayes binary classifier to categorize the transactions into two classes: either fraud or legitimate, Naive Bayes would make a good choice for starters in such problem as it simplifies the problem as much as possible and assumes the features to be all independent, in addition to its high accuracy.

II. Literature Review

As previously mentioned, credit card fraud causes drastic losses to the economy, which encourages a lot of researchers to find a solution that would detect fraud transactions in order to prevent it, and several methods were proposed and tested already, upon analyzing the different detection models, there were a couple of hindrances that were found when dealing with fraud detection; Sharma and Chalapathi (2022) have discovered the issue of real-life data, which turned out to be a huge problem with the dataset used, and the reason to the limitedness of real-life data is because of privacy and sensitivity concerns. Another thing that was discovered by Sharma and Chalapathi (2022) is that the techniques used in data mining take time to execute when you're dealing with big data. Sharma and Chalapathi (2022) and Zojaji, Atani, & Monadjemi (2016). have noticed an imbalance in data or more of a skewed distribution of data, and it's because of the ratio of fraud to non-fraud when compared to each other in the transaction data.

Another major drawback that was found by Zojaji, Atani, & Monadjemi, (2016) and Zareapoor, Seeja, & Alam (2012) in the preparation of credit card and transaction data, is that the data often overlap, as legitimate transactions closely resemble fraudulent ones, and vice versa. Plus, it's tricky when you're dealing with machine learning algorithms and categorical values, because most machine learning algorithms don't support categorical values, thus, another problem that would arise is the feature selection and detection algorithms, in addition to the fact that training the algorithms takes longer than predicted, and the complexity of the problem could make things even harder. However, the key is in feature selection which is a vital element in filtering the attributes in order to describe the fraud transaction and its characteristics.

There were multiple approaches by Randhawa, Loo, Seera, Lim, & Nandi, (2018) to deal with classification problems by employing Logistic Regression (LR), which starts by turning the instance of a fraudulent transaction into a discrete domain using Gaussian Mixture Models (GMMs), and to overbear the issue of class imbalance, a technique known as the "*synthetic minority oversampling was*" used, furthermore, a sensitivity analysis on the economic value was conducted with the purpose of emphasizing the importance of estimates. Their findings substantiated that a practical approach requires minimal training and a classifier that is restrained typically round by round, had similar performances.

Risk-Based Ensemble (RBE) is another model that can handle the data that has issues and manage to give tremendous results. and a highly efficient model was used to handle the case of imbalance data, while the implicit noise in the transaction dataset has been dealt with using the Naive Bayed algorithm Akila & Reddy (2017), another major obstacle they overcame is class imbalance and scalability problems by using oversampling, which is Duplicating samples from the minority class. they also used sensitivity analysis In order to pinpoint the hyper-parameter with the highest influence

twelve standard models and hybrid methods were used by (Chee et al.), in order to reach better accuracy rates in the case of credit card fraud detection. The evaluation was carried out using both benchmark and real-world data, they also reached an evaluation of the strengths and limitations of the methods used, and the performance measure was taken by The Matthews Correlation Coefficient metric (MCC)

III. Experimental Methodology

1. *Data description*

The very first step to solving is the collection of the data we are working with and describing it, the dataset used is the credit card transactions in September 2013, by European cardholders, the transactions in the dataset occurred within the span of two days, which were 284,807 transactions and 492 of them were fraud, which would make the positive fraud cases 0.1727% of the whole transactions, and by just looking at that anyone can tell how unbalanced the data is.

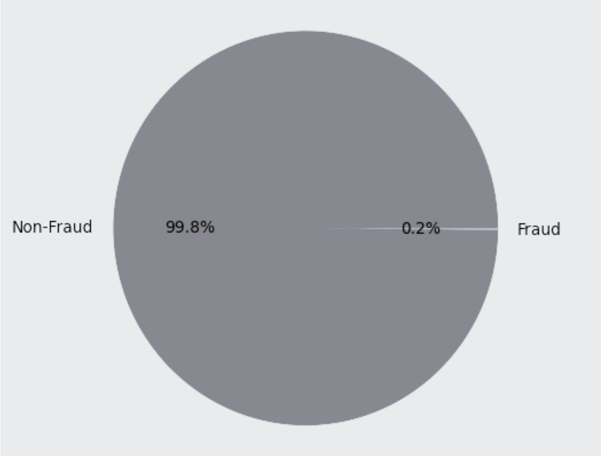


figure (1): the imbalance in Dataset

The data have undergone PCA transformation, which is a technique used with multiple purposes like; reducing the dimensions of the dataset while still maintaining the important information, simplifying complex datasets, improving model performance, and providing valuable insights.

The importance of PCA dwells in its ability to enable dimensionality reduction, feature extraction, noise reduction, visualization, and addressing multicollinearity. PCA is typically applied by analysts to increase the efficiency of analyzing high-dimensional datasets, thus, achieving better results and a deeper understanding of the data.

Performing PCA on the dataset would lead to numerical input variables, the main reason for not being able to disclose the original features and further background information, is confidentiality and privacy issues. The principal components that were obtained with PCA are our Features; (V1, V2, … V28) with the only exceptions of 'Time' and 'Amount' as they have not been transformed. The time feature is the elapsed time between each transaction and the first transaction in the dataset, and the Amount is the transaction amount that was done, plus the Feature 'Class' which is a representation of the classes output by the model (fraud, legit), where it takes the value 1 in case of fraud and 0 otherwise.

2. *Data preparation*

A. *Feature selection and Data cleaning*

One of the principal techniques used in machine learning is feature selection, which is choosing the variables that the most relevant to our given dataset and ignoring the least significant ones in order to reduce overfitting, improve the accuracy and reduce training time.

A technique that is of use is data visualization, bearing in mind the fact that the time and amount are highly varying in comparison to other data, thus this can be solved by scaling so that the features can be on the same level of magnitude, eventually, the process of feature selection finished with 28 features as Both Time and Amount don't show a significant differentiating pattern that will yield a significant predictive power and hence we will drop these features and only use the PCA transformed features V1, V2, ..., V28

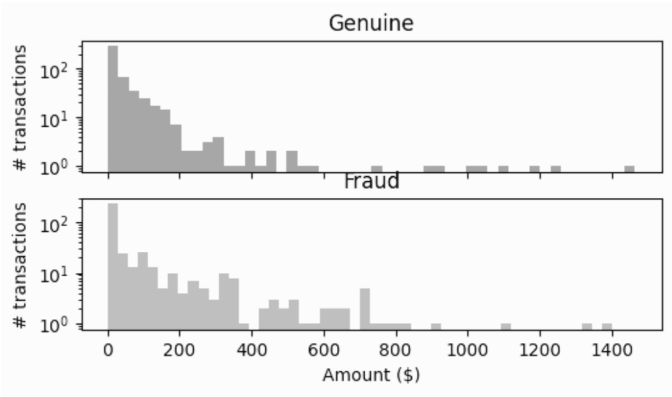


figure (2): the Amount Feature for Fraud and Genuine

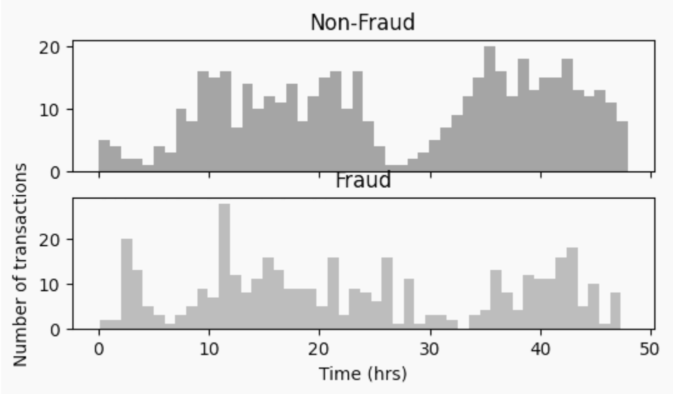


figure (3): the TimeFeature for Fraud and Genuine

B. Outlier Normalization Using IQR Method

Outliers are, generally, data that are extremely out of range unlike most of the instances that exist in the feature column, upon comparing the range of outliers against any feature in any given dataset. Outliers are undesired because they can mislead the classifier during training and may have an impact on the range of incoming testing values after deployment, that’s why, removing outliers is a necessary step that has to be taken before moving on with the model.

In order to remove outliers, you’ll have to check the range of values against each feature first, a boxplot is used to visualize the quartile ranges of all features, as shown in Figure(1)

Outlier removal should be done using an appropriate method, in our model, we used the Interquartile Range (IQR) method

The IQR method’s mathematical range is represented in Equation:

$$IQR = Q3 - Q1 \quad (1)$$

the resulting value, which is The IQR, is used to cover the feature outlier values; and it is the 75th percentile of all data values. It is calculated by arranging the data values in ascending order and then dividing them into 4 equal parts or quartiles. Then, the 1st and 3rd quartiles are found and, with their subtraction, we find an inner range of values called the IQR value. These inner values are adopted for the whole dataset and we cap all values.

figure(4) shows the feature Distribution before removing outliers.

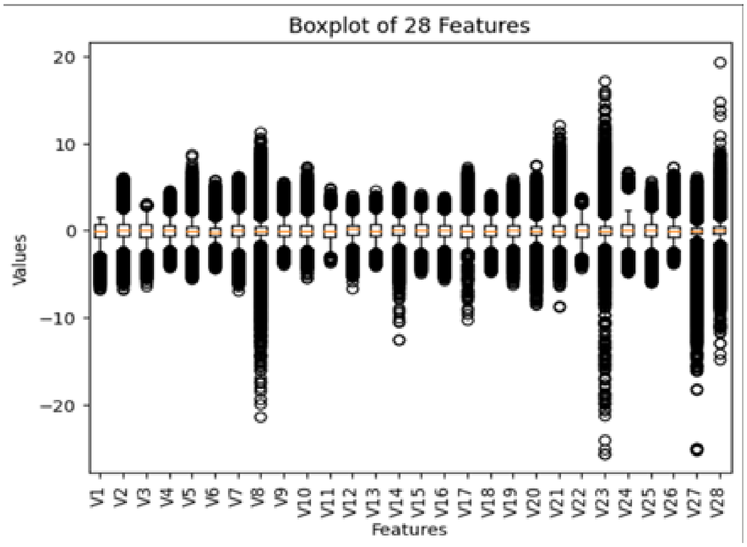


figure (4): the Feature Outliers

C. Data Undersampling:

An unbalanced dataset poses a problem in prediction, as the number of samples in one class is much higher than the other, thus having much more impact than the other in a way that would lead to inaccurate classification and poor model performance when it comes to predicting the minority class due to lack of representation, which is in our case detecting fraudulent transactions.

A proper solution to this problem would be shuffling the dataset via a sample function. then the data is sliced and only a subset of the non-fraudulent dataset is taken in order to match the number of that of the minority, which is the fraud. then the fraudulent dataset and the subset of the non-fraudulent dataset are concatenated, thus achieving a balanced dataset

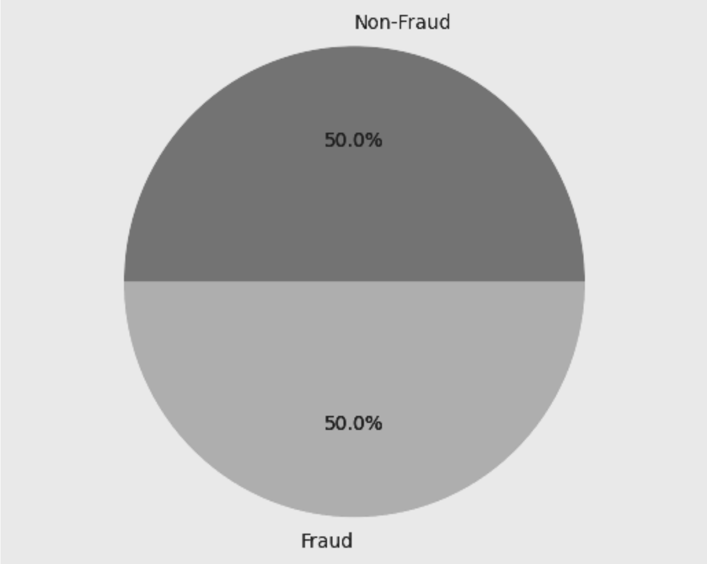


figure (5): Balancing the Dataset

D. Feature Ranking via Shapiro Method

The Shapiro Method is used in order to check whether the data is normally distributed or not, and the Shapiro-Wilk test outputs the p-values, and the ranking indicates the relative departure from normality for each feature. and it works by calculating two things: a test statistic and a p-value based on the observed data, then a comparison is made between the p-value and the significance level which is represented by alpha, our model uses a significance level of 5%.

The hypothesis of the Shapiro tests are; (1)Null hypothesis: which suggests that The data being tested is normally distributed. (2) Other hypotheses which suggest that The data does not follow a normal distribution. A lower rank suggests a lower p-value, indicating stronger evidence against the null hypothesis of normality. In contrast, a higher rank and higher p-value, indicate weaker evidence against the null hypothesis and a closer approximation to normality.

By examining the ranks, we can identify the features that deviate the most from normality (lower ranks) and those that are closer to being normally distributed (higher ranks).

figure (6) shows some features and their Distributions.

E. Measuring the Central Tendency and Dispersion

The purpose of measuring the mean, median, and standard deviation, is to gain a deeper understanding of the dataset's central tendency, variability, and distributional characteristics.

Hence, we have 28 features. It's not practical to write the values of central tendency and dispersion here in this report but the reader can find it in our Collab Notebook or Python Scripts.

F. Plotting the Conditional Distribution

Plotting the conditional distribution would allow us to explore, visualize, and analyze the relationship between two variables in a comprehensive way, as it can be shown in *figure(7)*

G. Building Classifier Model

The final step in the model is classification. There are multiple methods to classify, such as Naive Bayes Decision theorem, Gaussian distribution, Multinomial distribution, and Bernoulli distribution. each of them is used according to the distribution Naive Bayes, as the name suggests is based on Bayes theorem, is the one used in this model and it’s a supervised learning algorithm with the assumption of no dependencies between attributes, as well as Gaussian distribution, so we can use probability density function (pdf) value as an estimate for the conditional probability.

a conditional probability is the likelihood of some class/conclusion (C) given some evidence/observation(X), where a dependence relationship exists between C and X. This probability is denoted as P (C| X) where:

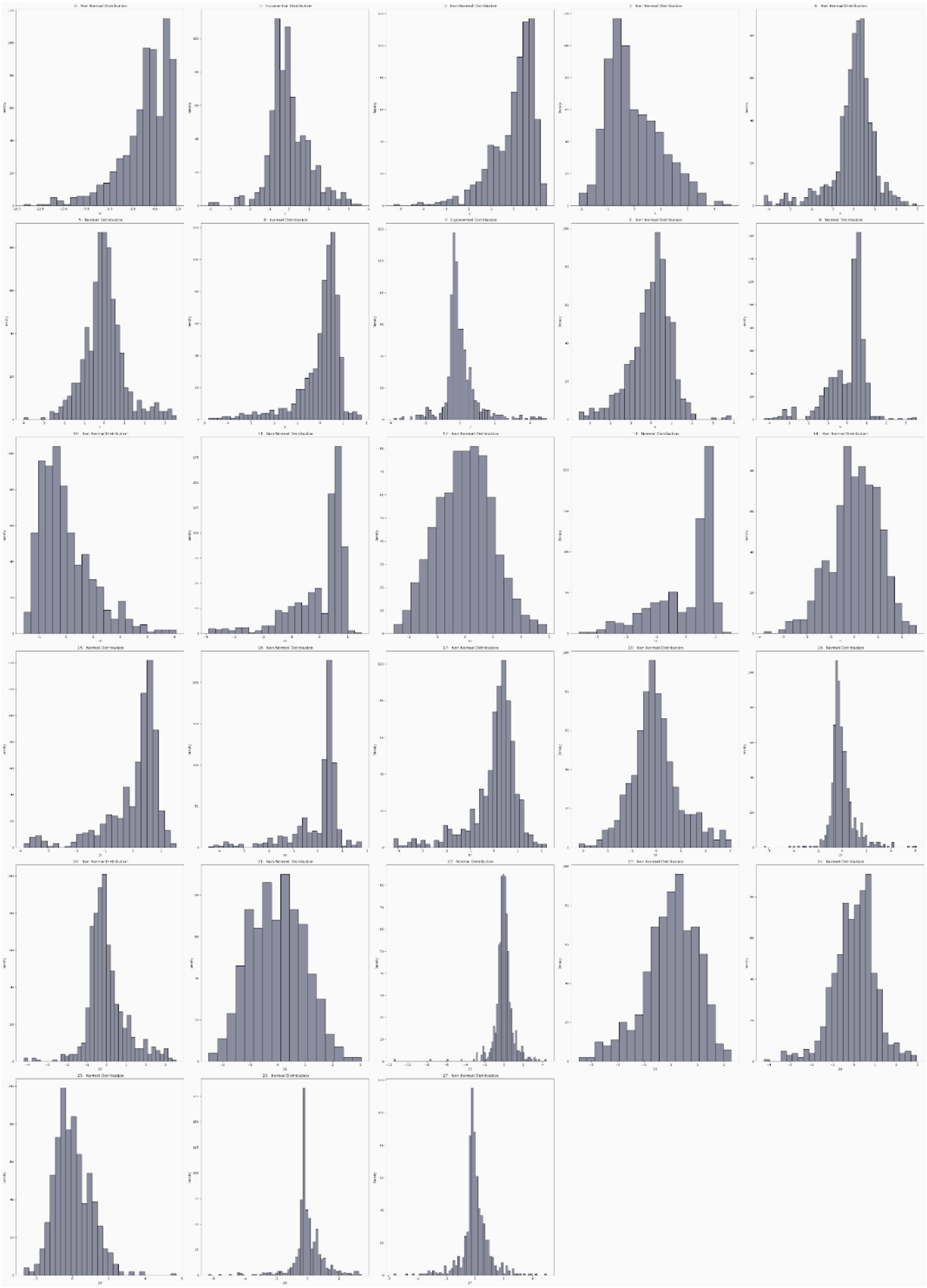
$$P(C_i|X) = \frac{(P(X|C_i)P(C_i)}{(P(X))} \quad (2)$$

upon making the naive assumption of class conditional independence is made, and given the class label of the sample Mathematically this means that:

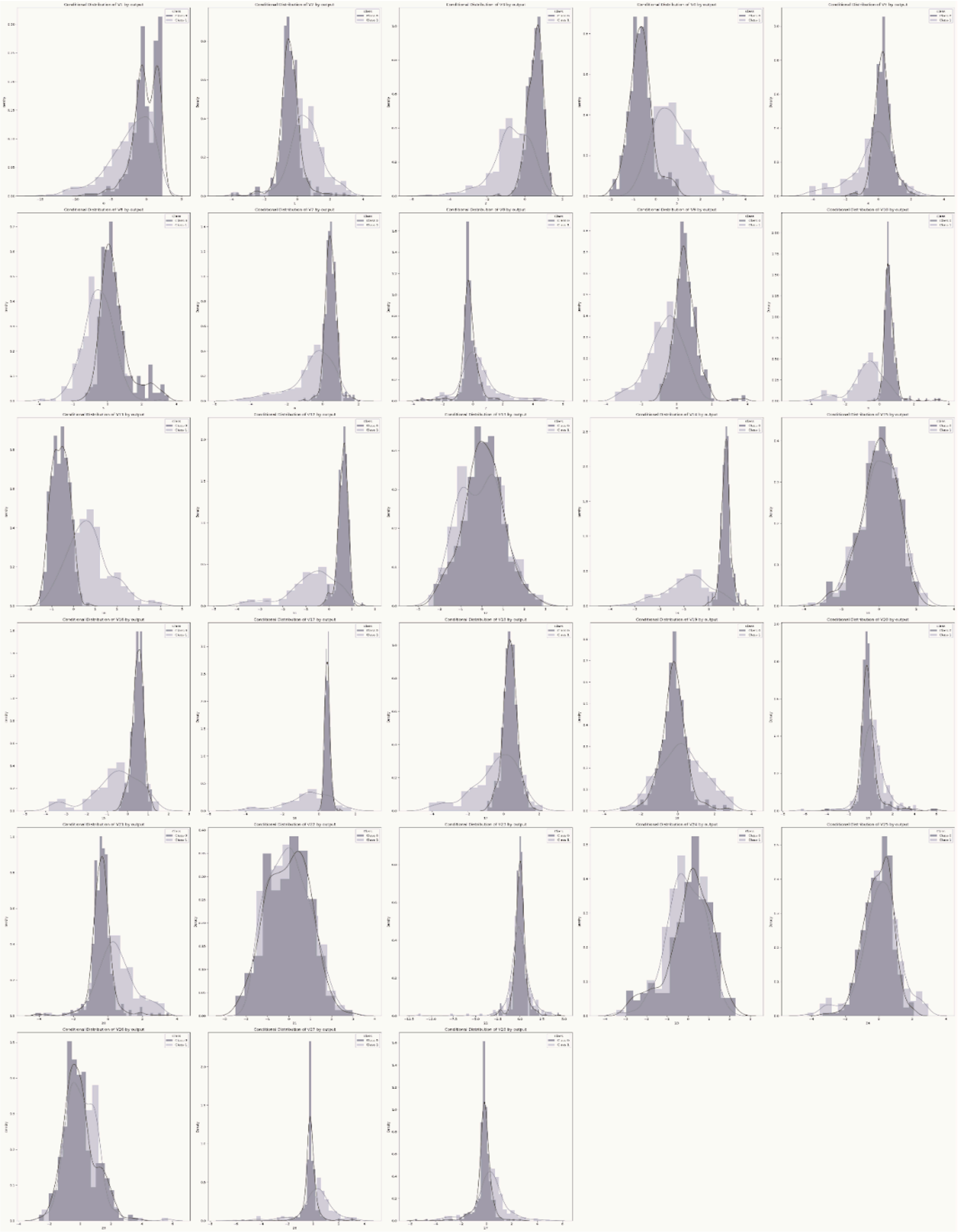
$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C) \quad (3)$$

H. Calculating the Correlation

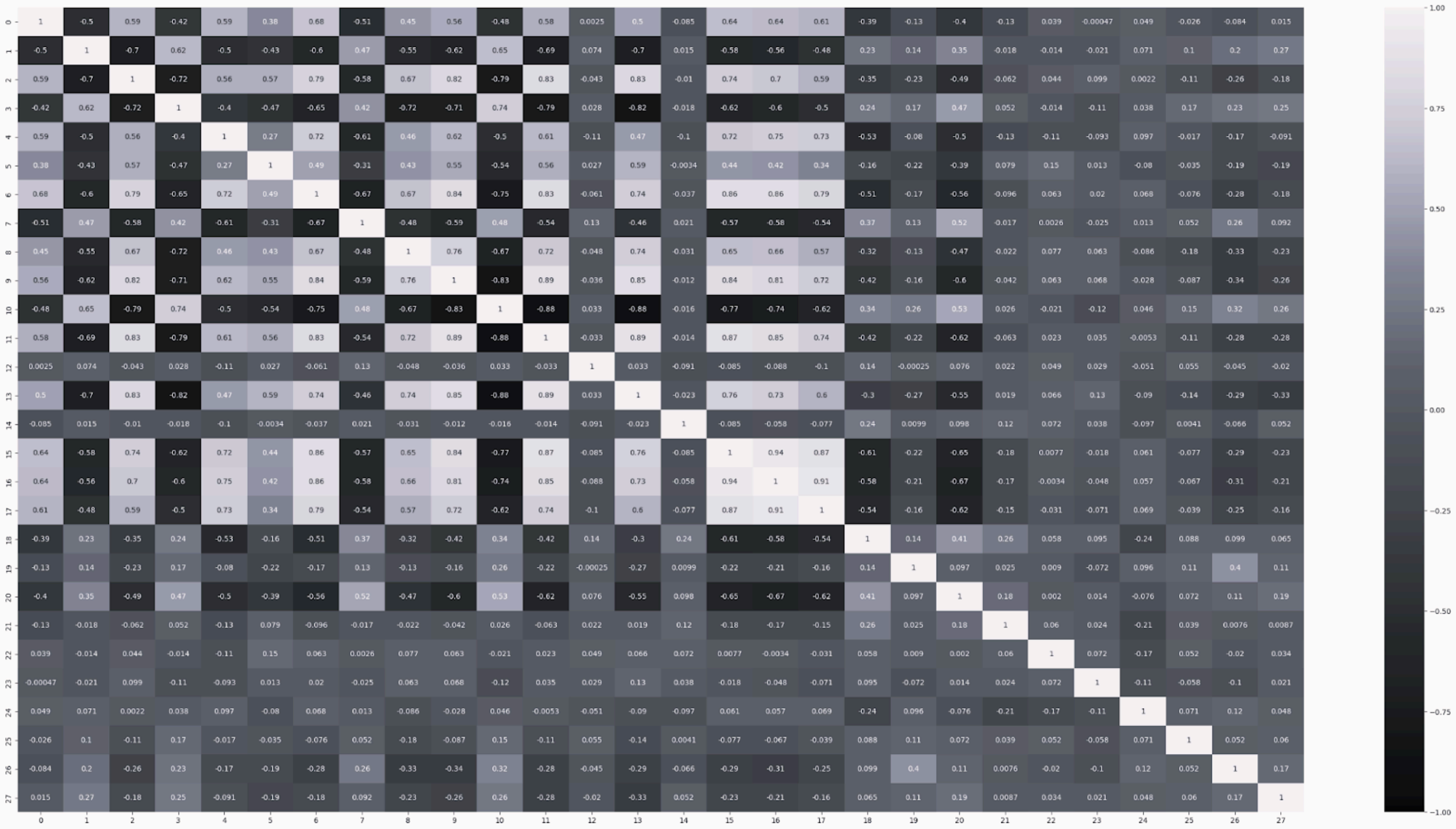
The calculation of the correlation between features in a dataset is essential for understanding the relationships and dependencies between variables. The correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It can help determine which features are highly correlated and identify potential redundancies or dependencies in the data, shown in Figure (8).



figure(6): Histogram for each Feature, showing the type of distribution



figure(7): the conditional distribution



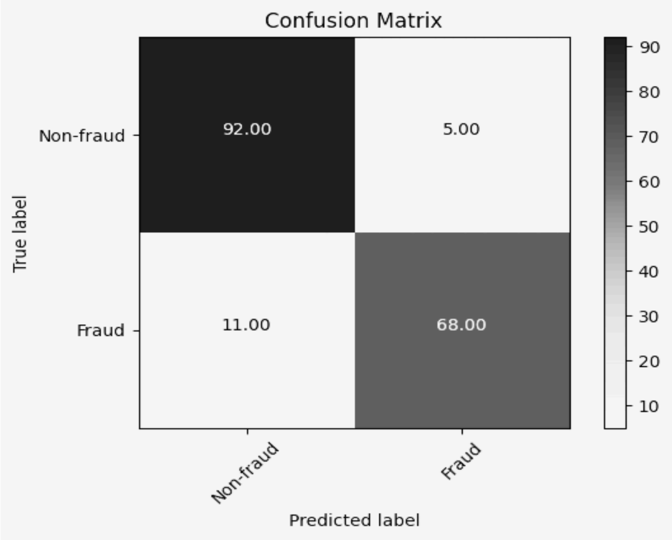
figure(8): the correlation between the features

IV. Results

In order to determine the most suitable algorithm for the problem of detecting fraudulent transactions, a comparison between algorithms was made. The measured metrics are accuracy, recall, and precision. and they are all calculated using a confusion matrix.

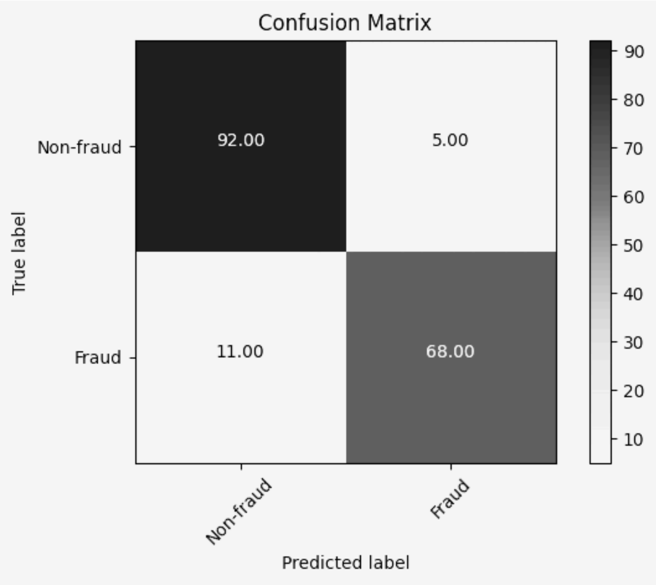
Our Custom Naive Bayes (NB) model achieved:

1. Recall score: 0.85
2. Precision score: 0.94
3. F1 score: 0.89
4. Accuracy score: 0.91



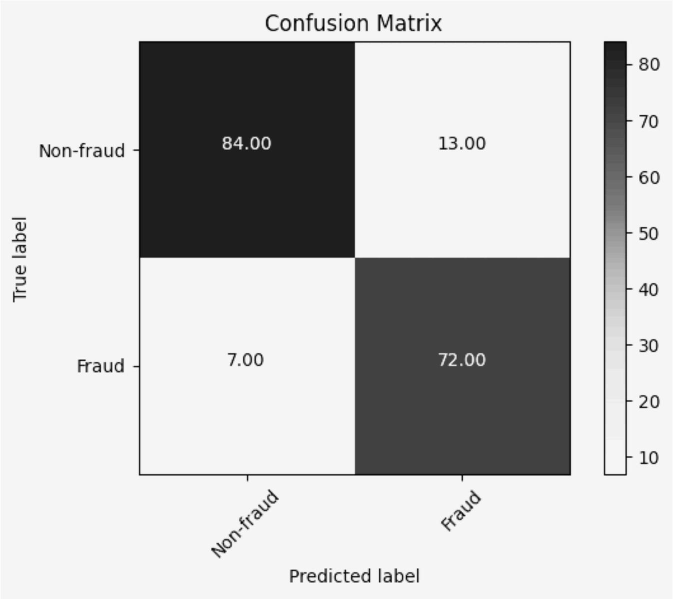
Scikit-learn Naive Bayes (NB) model achieved:

1. Recall score: 0.85
2. Precision score: 0.94
3. F1 score: 0.89
4. Accuracy score: 0.91



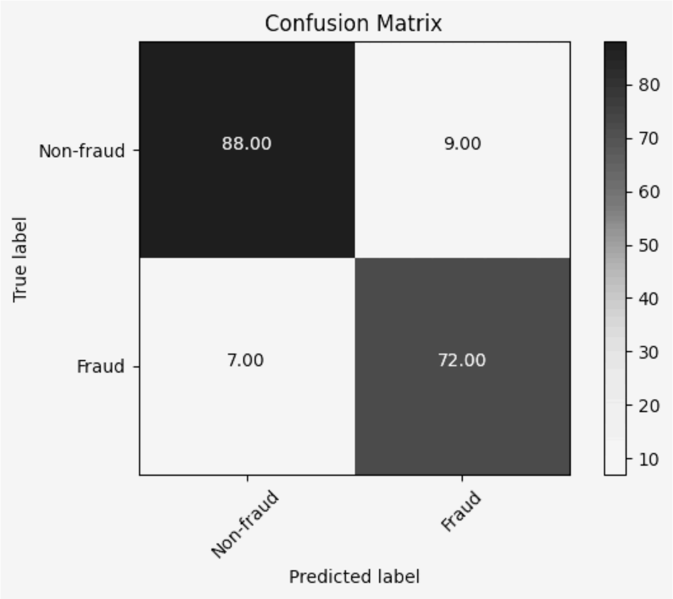
Scikit-learn Decision Tree (DT) Classifier achieved:

1. Recall score: 0.9
2. Precision score: 0.85
3. F1 score: 0.87
4. Accuracy score: 0.88



Scikit-learn Logistic Regression (LR) achieved:

1. Recall score: 0.91
2. Precision score: 1.0
3. F1 score: 0.95
4. Accuracy score: 0.96



	Custom NB	Scikit NB	Scikit DT	Scikit LR
Recall	0.85	0.85	0.9	0.91
Precision	0.94	0.94	0.85	1.0
F1	0.89	0.89	0.87	0.95
Accuracy	0.91	0.91	0.88	0.96

Summary of the findings of our model and other Sckikit-Learn built-in models, as it can be concluded, the Custom Naive Bayes made some promising accuracies and is higher than that of the Decision Tree

V. Conclusion and Future Work:

Credit card frauds impose a great threat to the world of business, and they can lead to extreme losses, on a personal scale, and on larger ones like business or even government. because of that companies highly encourage the development of new technology and ideas to detect and prevent fraud.

this paper's main aim was to try solving this problem and compare it to other machine learning algorithms that are used in the detection of fraud transactions, Hence upon comparing them it was found that Naive Bayes gave the best results, and this conclusion was reached by comparing different metrics such as recall, accuracy, and precision, and recall is important to have in high values, It’s also worth mentioning that the Feature selection and balancing of the dataset have had a significant impact on achieving such results

As for Further research, the focus should be on various machine learning algorithms such as genetic algorithms, and different types of stacked classifiers, alongside with extensive feature selection to get even better results, and hopefully, put an end to the everlasting menace of credit card fraudulent

VI. References

1. Sharma, N. D. K., & Chalapathi, N. M. M. V. (2022). A Novel Machine Learning Technique for Fraud Detection on Credit Card Financial Data. International Journal of Engineering Technology and Management Sciences, 371–378. https://doi.org/10.46647/ijetms.2022.v06i04.0060
2. Zojaji, Zahra, Reza Ebrahimi Atani, and Amir Hassan Monadjemi, “A survey of Credit Card Fraud Detection Techniques : Date and Technique Oriented Perspective, ” pp. 1-26, 2016. https://arxiv.org/pdf/1611.06439
3. Zareapoor, Masoumeh, K. R. Seeja, and M. Afshar Alam. "Analysis on credit card fraud detection techniques: based on certain design criteria." International journal of computer applications 52, no. 3, pp. 35-42, 2012. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=8c4805a11949ae979c126749ebb88d56b3b41336
4. Akila, S., and U. Srinivasulu Reddy. "Risk-based bagged ensemble (RBE) for credit card fraud detection." In 2017 International Conference on Inventive Computing and Informatics (ICICI), pp. 670-674. IEEE, 2017. https://ieeexplore.ieee.org/abstract/document/8365220/
5. Randhawa, Kuldeep, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, and Asoke K. Nandi. "Credit card fraud detection using AdaBoost and majority voting." IEEE Access 6 (2018): 14277-14284. https://ieeexplore.ieee.org/abstract/document/8292883/
6. Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare. "Credit card fraud detection using machine learning techniques: A comparative analysis." In 2017 international conference on computing networking and Informatics (ICCNI), pp. 1-9. IEEE, 2017. https://www.academia.edu/download/57586303/V4I3-1165.pdf
7. Husejinovic, Admel. "Credit card fraud detection using naive Bayesian and c4. 5 decision tree classifiers." Husejinovic, A.(2020). Credit card fraud detection using naive Bayesian and C 4 (2020): 1-5. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3521283
8. Varmedja, Dejan, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, and Andras Anderla. "Credit card fraud detection-machine learning methods." In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1-5. IEEE, 2019. https://ieeexplore.ieee.org/abstract/document/8717766/
9. Husejinovic, Admel, Credit Card Fraud Detection Using Naive Bayesian and C4.5 Decision Tree Classifiers (January 17, 2020). Husejinovic, A. (2020). Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers. Periodicals of Engineering and Natural Sciences, ISSN 2303-4521, 8 (1), 1-5, Available at SSRN: https://ssrn.com/abstract=3521283

Team Members Contributions:

Name	Efforts
Abdulrehman Emad	Loading, Cleaning and splitting the data
	Implemented Shapiro test function
	Implemented IQR method function to get the indices of the outliers
	Implemented a function to get the precision , recall and f1 score
	Collaborated on the final presentation on the following sections (Data Cleaning, Naive Baye, Results)
Amir Hesham	Organized team meetings, tasks, and reviews
	Implemented the Naive Bayes Classifier from scratch
	Selected the proper features that will go into the classifier
	Balanced the data by taking a sample with equal number of observations
	Collaborated on the final presentation on the following sections (Introduction, Data Collection, Data Cleaning, Naive Bayes)
Habiba Salama	Plotting the histogram for each feature and determine its Distribution
	Plotting the Conditional Distribution for each feature by the output Class
	Wrote the literature review in Report
	Wrote the Experimental Methodology in Report
	Wrote the Results and Conclusion
Mariam Ahmed	Co-Wrote the documentation of the code throughout the model
	Collaborated in Comparing our Model other built in models in python to see it's accuracy
	Wrote the abstract, Introduction in report
	Co-wrote the literature review, the experimental Methodology in Report
	Co-wrote the Results and Conclusion in Report
Mohamed Mosilhe	Loading, Cleaning, and splitting the data
	Plotting the histogram for each feature and determining its Distribution
	Plotting the Conditional Distribution for each feature by the output Class
	Collaborated on the final presentation on the following sections (Introduction, Data Collection, Data Description, PCA Transformation)
	Collaborated on the final presentation on the following sections (Types of Naive Bayes classifier, Advantages and limitations of Naive Bayes, Model training, and Evaluation)