IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Determining E-Credit Score using Predictive Analysis Algorithms & Its Real-time Application

[1]FAHAD JAHANGIR, [2]ASHISH SHARMA, [3]ANKITA VIJAYAN, [4]KANINIKA PAUL, [5]ROHIT PAREEK

[1]Student, [2]Student, [3]Student, [4]Student, [5]Student
[1]Computer Science & Engineering,
[1]Presidency University, Bangalore, India

Abstract: The method of credit scoring involves a set of decision models that help in granting consumers loan and credit [1]. Since it involves many decisive statements, we employ machine learning models to predict the credit score of a consumer. The following chapter will investigate into the details of machine learning or ML model usages, such as the understanding of model outputs, structure modification, and constant investigation of work [2]. These understandings intend to cultivate the prognostic abilities, donating implicitly to the dominion of credit score forecast. Credit assessment has an essential part in the monetary business, leading the lending choices and handling the danger. Usually, this evaluation depends on physical approaches and comparatively humble scoring representations [3]. By bridging research gaps in credit scoring methodologies, specifically, model interpretability, fairness, and adaptability this project makes substantial contributions to advancing the comprehension and application of machine learning in the realm of credit risk assessment. The system's design prioritizes simplicity and transparency, fostering reproducibility and accommodating future expansions in credit scoring research and practical implementations**.**

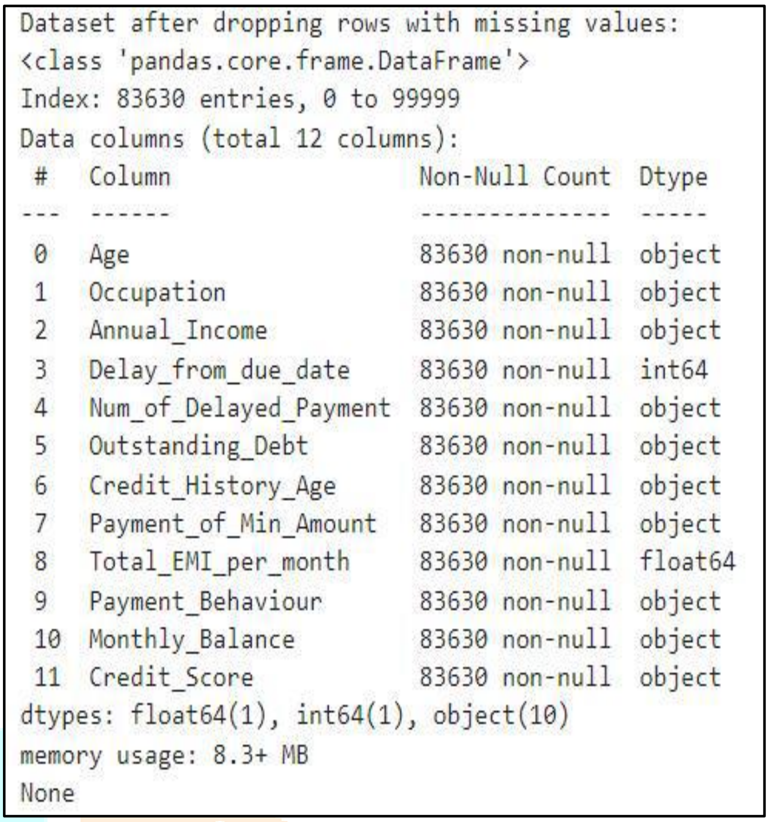
Index Terms **- Credit Score, K-Nearest Neighbours, Support Vector Classifier, Random Forest, confusion matrix, correlation, Min-Max scaling.**

I. INTRODUCTION

The current monetary domain depends suggestively on precise credit evaluation to regulate persons' solvency. Forecasting credit values is essential for monetary establishments to create knowledgeable choices about advancing and handling monetary dangers. In this background, incorporating the machine learning or ML procedures to forecast credit values has occurred as a vital characteristic of the monetary area. This study would discover and use the progressive procedures to forecast the credit scores by means of machine learning representations, highlighting the complete preprocessing, and examining breakdown of a credit score dataset.

The dataset below inspection, gotten from a CSV file, includes countless characteristics surrounding vital monetary preferences, histories of payment, and individual demographics. This study gets on an arranged expedition, beginning with informational work and understanding collection to comprehend the dataset's configuration widely [4]. By thorough data refinement, inappropriate supports are eliminated, missing standards are touched, and mistaken informational entries are corrected, confirming the honesty for information and willingness for breakdown.

Essential for this research is the wide-ranging inspection of the dataset, researching into visualizations like the boxplots, donut charts, bar plots, histograms, and correlation heatmaps. These visualizations help unloosen designs, deliveries, and possible interrelations between variables, giving vital understandings into the dataset's construction. The preprocessing segment is essential in preparing the dataset for incorporation with machine learning that is ML representations. Numeric columns experience standardization by Min-Max scaling, while uncompromising columns, such as 'Occupation,' are prearranged into dummy variables by one-shot programming. These alterations are dynamic in homogenizing the informational ranges and permitting continuous incorporation of unqualified information into following predictive representations.



```
Dataset after dropping rows with missing values:
<class 'pandas.core.frame.DataFrame'>
Index: 83630 entries, 0 to 99999
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Age                    83630 non-null  object
 1   Occupation             83630 non-null  object
 2   Annual_Income          83630 non-null  object
 3   Delay_from_due_date    83630 non-null  int64
 4   Num_of_Delayed_Payment 83630 non-null  object
 5   Outstanding_Debt       83630 non-null  object
 6   Credit_History_Age     83630 non-null  object
 7   Payment_of_Min_Amount  83630 non-null  object
 8   Total_EMI_per_month    83630 non-null  float64
 9   Payment_Behaviour      83630 non-null  object
 10  Monthly_Balance        83630 non-null  object
 11  Credit_Score           83630 non-null  object
dtypes: float64(1), int64(1), object(10)
memory usage: 8.3+ MB
None
```

Figure 1: Status of columns after dropping rows with missing values

1.1 Data Complexity and Cleaning

The dataset for credit score, 'train.csv', shows a complex collection of characteristics reaching from 'Age,' 'Annual_Income,' 'Delay_from_due_date,' 'Num_of_Delayed_Payment,' 'Outstanding_Debt,' to 'Credit_Score.' Though, within this prosperity stays with characteristic complications like missing standards, mistaken entrances, and the incidence of unrelated columns. Resolving these matters claims thorough cleaning policies for information, demanding the elimination of out of a job columns and management of misplaced or imprecise information entrances to confirm the dataset's dependability and accurateness.

1.2 Exploratory Analysis and Insight Generation

Amongst the fresh dataset, revealing appreciated understandings necessitates a wide-ranging examining breakdown. Visualizations, such as the bar plots, boxplots, donut charts, histograms, and correlation heatmaps, works as an application to reveal hidden outlines, deliveries, and interrelations between variables [5]. This breakdown intends to undo fundamental inclinations and dependences vital for knowledgeable decision-making in credit evaluation.

1.3 Data Transformation and Preprocessing

A crucial phase includes formulating the sets of data for machine learning or ML model incorporation. This phase includes the alteration of numeric structures by Min-Max grading to normalize series and the encrypting of definite characteristics into dummy variables by means of one-hot encrypting. Confirming information constancy and compatibility for machine learning or ML representations forms the crux of this preprocessing stage.

1.4 Machine Learning Model Application and Evaluation

The following stage includes the usage of machine learning or ML representations, particularly the KNN classifier, SVC, and Random Forest Classifier. The encounter stays not only in arranging these representations but also in thoroughly assessing and improving them to accomplish peak prognostic efficiency. Model exercise, fine-tuning, and wide-ranging assessment procedures are authoritative to confirm precise credit score forecasts.

1.5 Interpretation and Actionable Insights

Beyond analytical accurateness, the actual value stays in incorporating the model productions to get the actionable understandings [6]. Comprehending the factors inducing credit scores and originating expressive data from the representations' forecasts will augment the land of credit score estimation and work suggestively to knowledgeable decision-making in credit evaluation. The eventual aim is to direct by these encounters, doing preprocessing of data, exploratory breakdown, and machine learning or ML technologies usage to plan a strong predictive outline for credit score evaluation. This endeavour leads to transform the outdated credit scoring example, improving the accurateness, and empowering shareholders to generate a more knowledgeable and data-led choices in credit evaluation and risk administration.

II. SYSTEM DESIGN AND IMPLEMENTATION

The systematic design and implementation of the Credit Score Prediction Project adhere to a structured framework integrating data preprocessing, model development, and rigorous evaluation. Leveraging the versatile **Jupyter Notebook** environment and essential **Python libraries** such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, the project unfolds in a collaborative and interactive workspace.

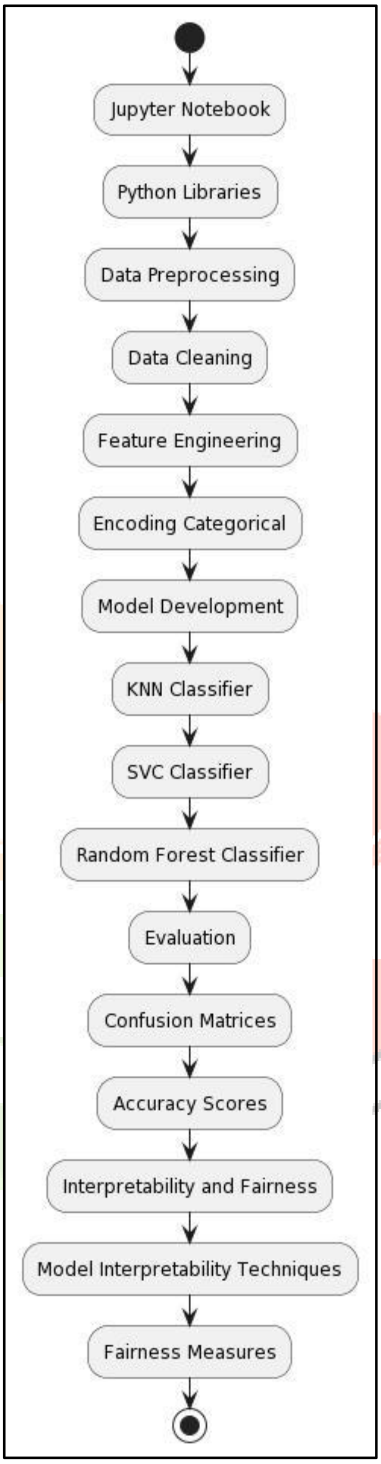
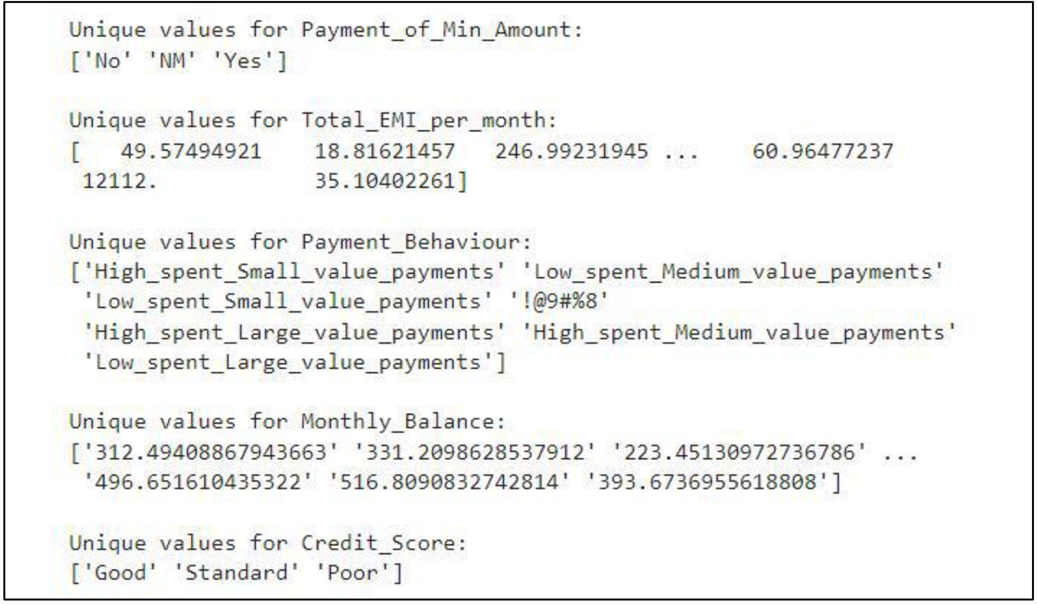


Figure 2: Activity Diagram for evaluation of predictive algorithms

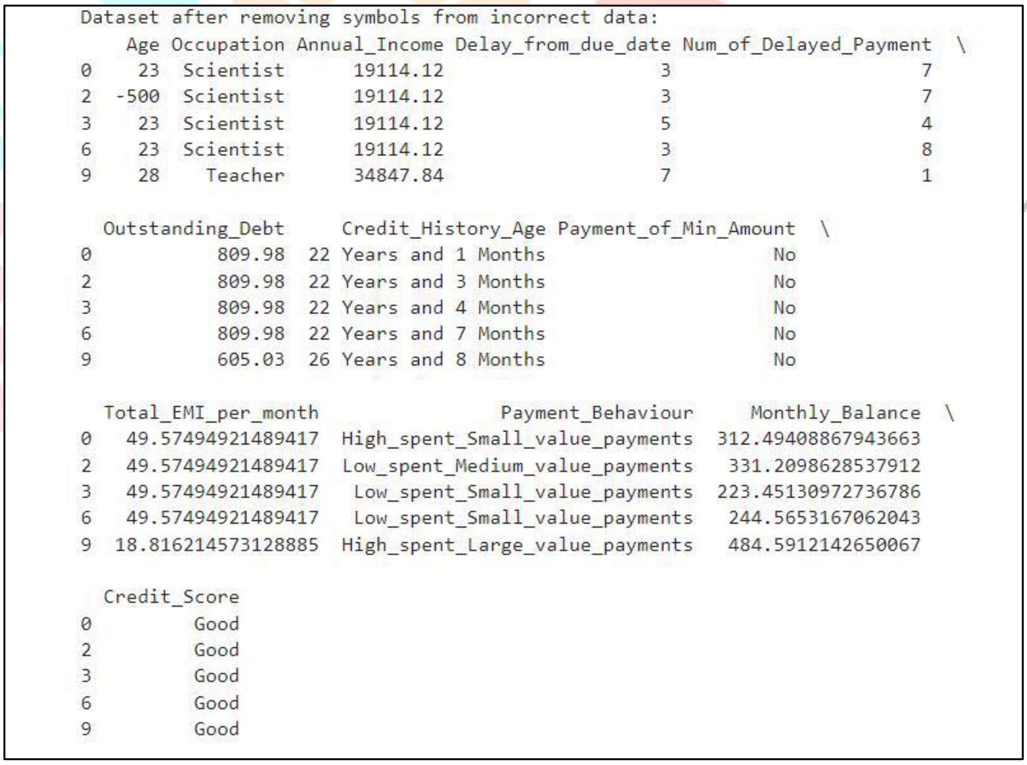
Data Preprocessing: Data preprocessing is a crucial phase in the Credit Score Prediction Project, laying the groundwork for accurate and reliable credit score predictions. This multifaceted process involves several key steps to ensure the quality and relevance of the input data.



```
Unique values for Payment_of_Min_Amount:
['No' 'NM' 'Yes']

Unique values for Total_EMI_per_month:
[   49.57494921    18.81621457   246.99231945 ...    60.96477237
 12112.            35.10402261]

Unique values for Payment_Behaviour:
['High_spent_Small_value_payments' 'Low_spent_Medium_value_payments'
 'Low_spent_Small_value_payments' '!@9#%8'
 'High_spent_Large_value_payments' 'High_spent_Medium_value_payments'
 'Low_spent_Large_value_payments']

Unique values for Monthly_Balance:
['312.49408867943663' '331.2098628537912' '223.45130972736786' ...
 '496.651610435322' '516.8090832742814' '393.6736955618808']

Unique values for Credit_Score:
['Good' 'Standard' 'Poor']
```

Figure 3: Checking unique values for each column to identify missing or incorrect data

Data Cleaning: The initial step addresses the identification and handling of missing or erroneous data. Leveraging Pandas, missing values are either imputed or removed, contributing to a cleaner and more complete dataset. Outliers are also identified and appropriately addressed to prevent distortions in the credit scoring models.



```
Dataset after removing symbols from incorrect data:
    Age Occupation Annual_Income Delay_from_due_date Num_of_Delayed_Payment  \
0    23  Scientist      19114.12                   3                      7   
2  -500  Scientist      19114.12                   3                      7   
3    23  Scientist      19114.12                   5                      4   
6    23  Scientist      19114.12                   3                      8   
9    28    Teacher      34847.84                   7                      1   

  Outstanding_Debt     Credit_History_Age Payment_of_Min_Amount  \
0           809.98  22 Years and 1 Months                    No   
2           809.98  22 Years and 3 Months                    No   
3           809.98  22 Years and 4 Months                    No   
6           809.98  22 Years and 7 Months                    No   
9           605.03  26 Years and 8 Months                    No   

  Total_EMI_per_month               Payment_Behaviour     Monthly_Balance  \
0   49.57494921489417 High_spent_Small_value_payments  312.49408867943663   
2   49.57494921489417 Low_spent_Medium_value_payments   331.2098628537912   
3   49.57494921489417  Low_spent_Small_value_payments  223.45130972736786   
6   49.57494921489417  Low_spent_Small_value_payments   244.5653167062043   
9  18.816214573128885 High_spent_Large_value_payments   484.5912142650067   

  Credit_Score  
0         Good  
2         Good  
3         Good  
6         Good  
9         Good  
```

Figure 4: Dataset after removing symbols from incorrect data

Feature Engineering: Feature engineering aims to enhance the predictive power of the selected features. This involves creating new features, transforming existing ones, and extracting relevant information to better capture patterns in the data. Techniques such as binning, scaling, and one-hot encoding are employed to optimize feature representation.

Encoding Categorical Variables: Categorical variables are transformed into a numerical format suitable for machine learning models. The project utilizes techniques like one-hot encoding to represent categorical information, ensuring compatibility with classifiers such as KNN, SVC, and Random Forest.

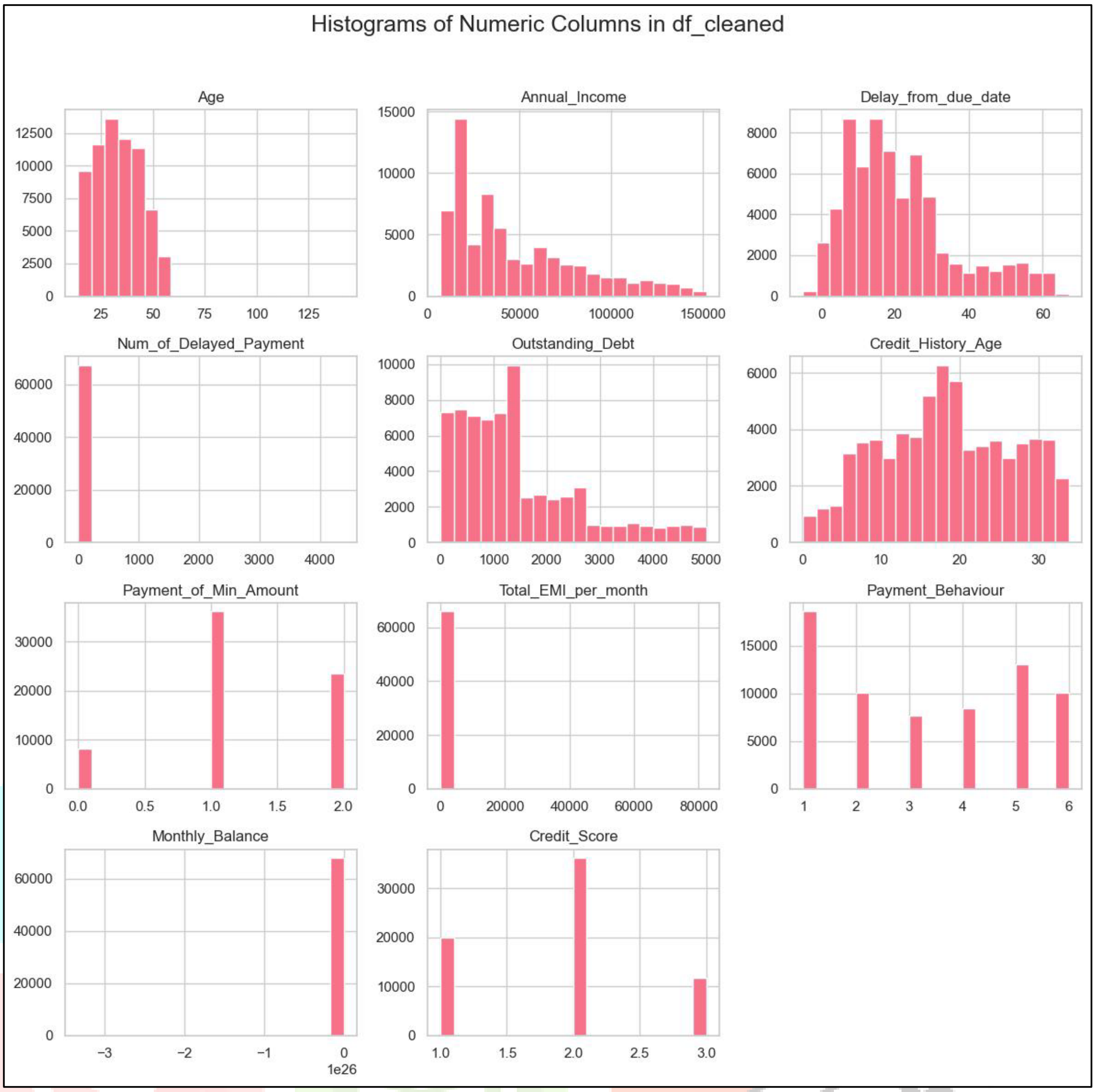


Figure 5: Distribution of numerical columns in cleaned dataset

This code serves as a powerful exploratory data analysis (EDA) tool, allowing for a rapid and intuitive assessment of the distributional characteristics of numeric variables within the cleaned dataset. Stakeholders and data analysts can leverage these histograms to identify trends, outliers, and potential areas of interest, laying the foundation for more in-depth investigations and informed decision-making.

III. Results and Discussions

Model Development: Model development stands as a pivotal stage, where machine learning classifiers are trained to predict credit scores with precision and reliability. This phase encompasses the selection, training, and fine-tuning of classifiers to ensure optimal performance.

Classifier Selection: Three diverse classifiers: K-Nearest Neighbours (KNN), Support Vector Classifier (SVC), and Random Forest are employed to leverage different aspects of credit scoring data. The selection aims to capture varied patterns and nuances inherent in credit-related features.

Evaluation: The Evaluation phase is a critical step that rigorously assesses the performance and effectiveness of the developed machine learning models. Utilizing key metrics, this phase provides valuable insights into the reliability and predictive power of the classifiers: K-Nearest Neighbours (KNN), Support Vector Classifier (SVC), and Random Forest, trained on pre-processed credit data.

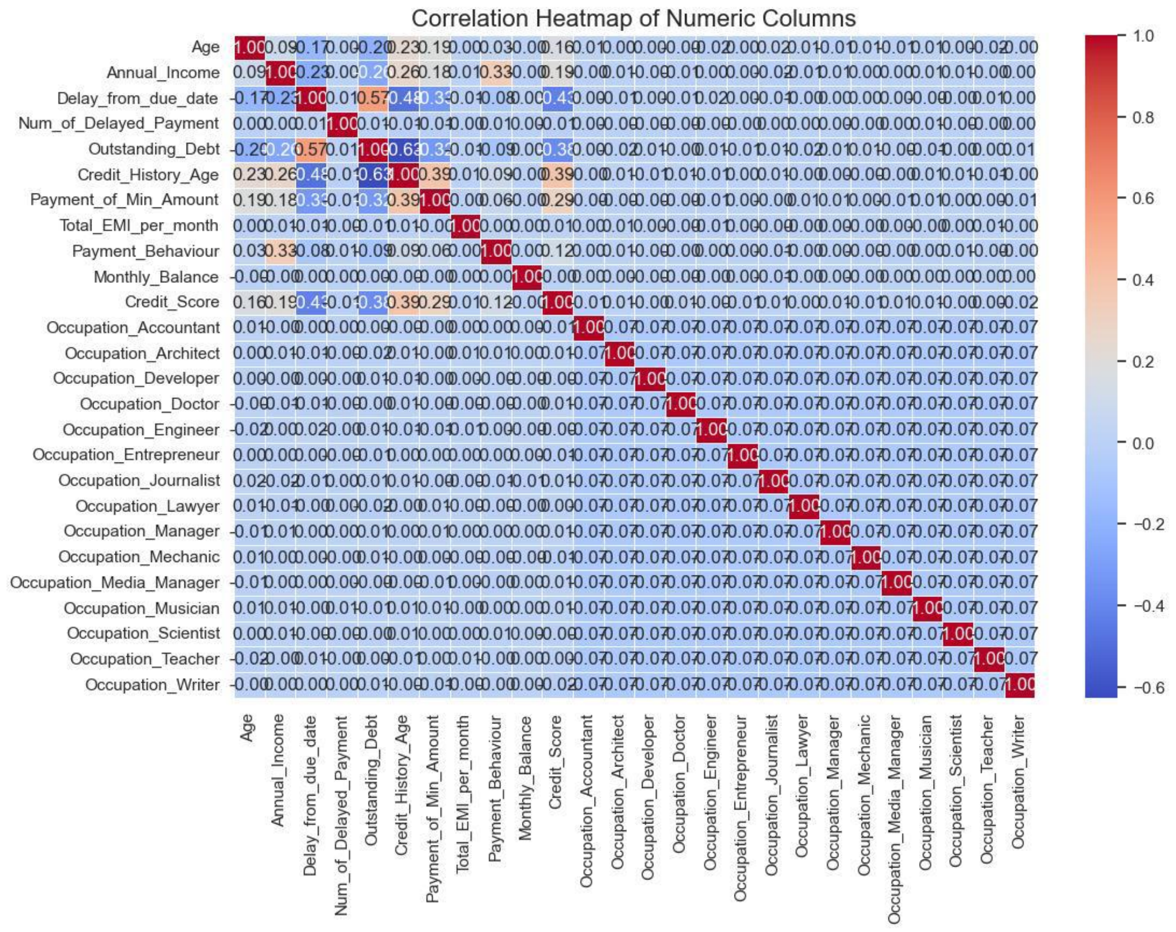


Figure 6: Correlation heat map of numerical columns

3.1 Confusion Matrices

Confusion matrices are employed to analyse the true positives, true negatives, false positives, and false negatives. This allows for a detailed understanding of the classifiers' ability to correctly predict creditworthiness and identify instances of misclassification.

3.1.1 K-Nearest Neighbours (KNN) Classifier:

The KNN model is configured with five neighbours for classification. The resulting confusion matrix reveals its predictive performance across credit score categories. The accuracy of the KNN Classifier is calculated to be 61.95%. This suggests that the model correctly predicts the credit scores for approximately 62% of the instances.

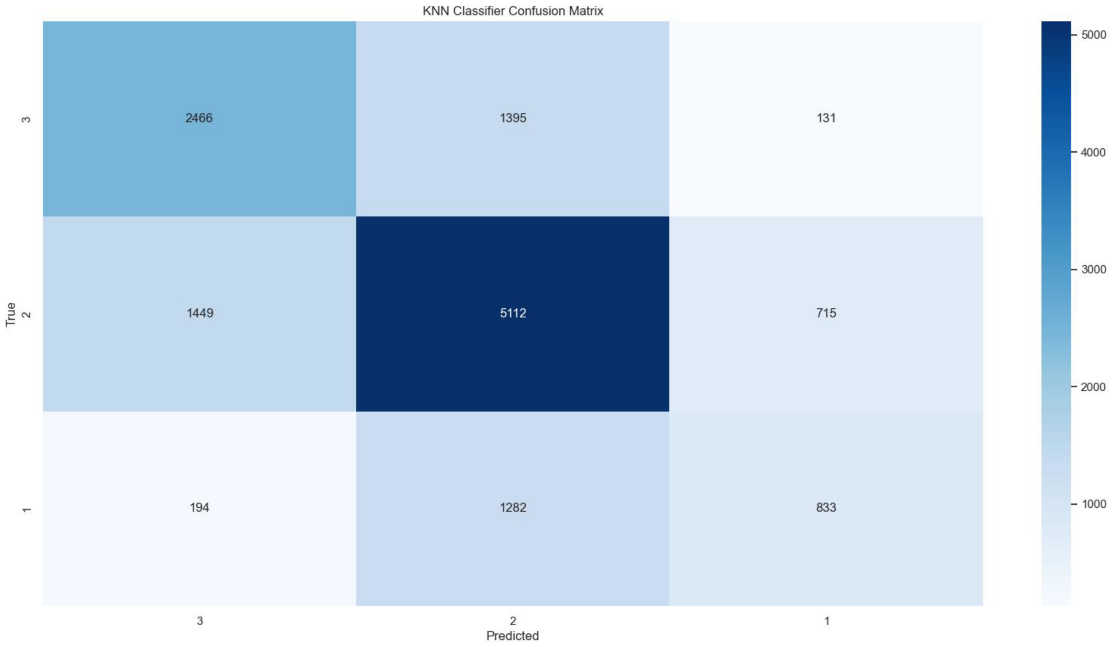


Figure 7: KNN Classifier confusion matrix

3.1.2 Support Vector Classifier (SVC):

The SVC model is instantiated using default settings, and its confusion matrix displays the distribution of predicted and true credit scores. The accuracy of the SVC Classifier is determined to be 59.56%, indicating its capability to accurately classify credit scores in approximately 60% of cases.

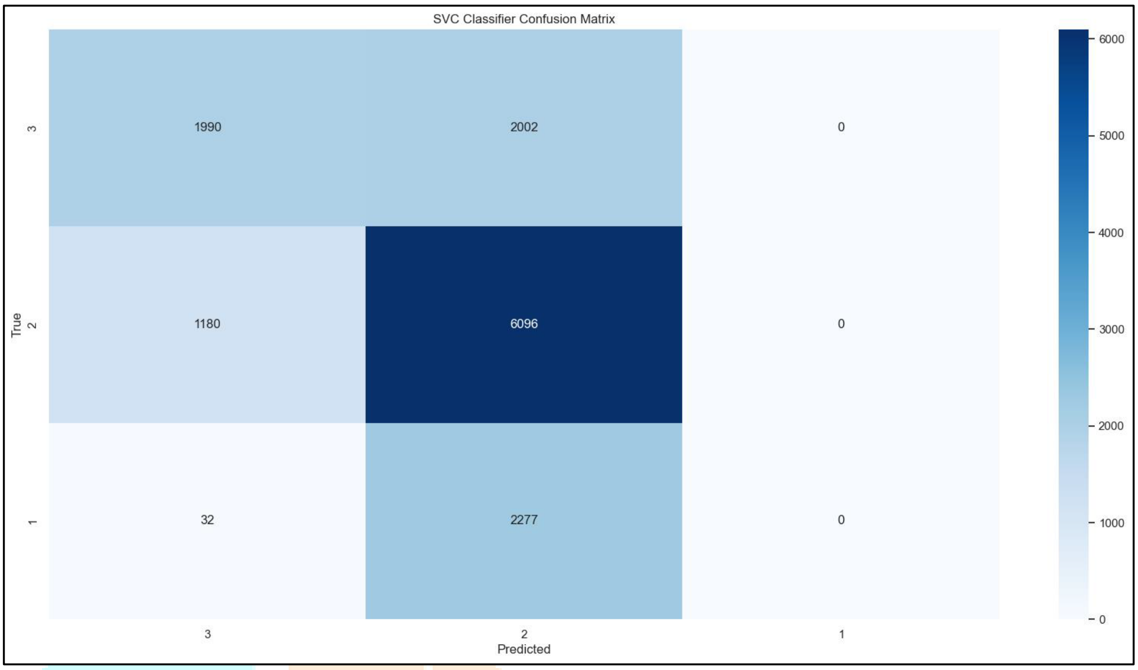


Figure 8: SVC confusion matrix

3.1.3 Random Forest Classifier:

The Random Forest Classifier, with 100 estimators and a random state of 42, demonstrates superior performance among the three classifiers. The confusion matrix for the Random Forest Classifier showcases its ability to predict credit scores across different categories effectively. The model achieves an impressive accuracy of 77.43%, indicating a higher level of precision in credit score prediction compared to KNN and SVC.

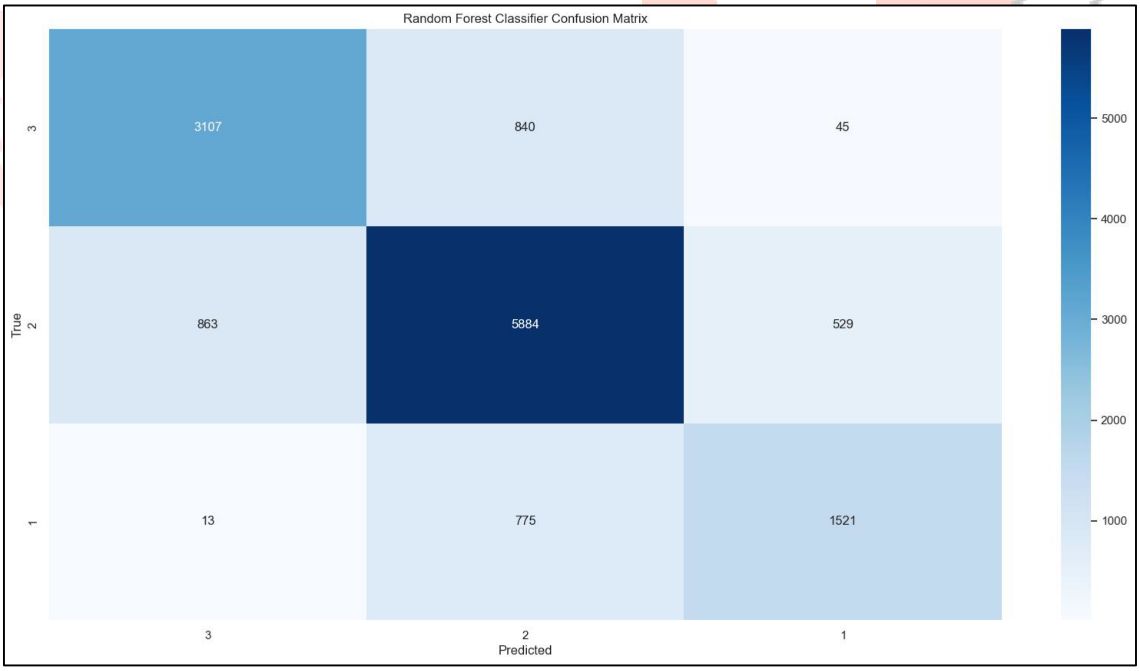


Figure 9: Random Forest Classifier confusion matrix

3.1.4 Accuracy Scores:

Accuracy scores serve as overarching metrics, quantifying the models' overall predictive accuracy. This evaluation metric provides a consolidated measure of how well the classifiers perform across all credit score categories, offering a comprehensive assessment of their efficacy.

A final visualization compares the accuracy scores of the three classifiers using a bar plot. The Random Forest Classifier emerges as the most accurate model among the three, with a significantly higher accuracy compared to KNN and SVC. This underscores the Random Forest Classifier's effectiveness in credit score prediction, making it a promising choice for this task.

Overall, the evaluation results and visualizations presented here provide valuable insights into the performance of different classifiers for credit score prediction. The Random Forest Classifier stands out as the most accurate and robust model, showcasing its potential for practical deployment in credit scoring applications.

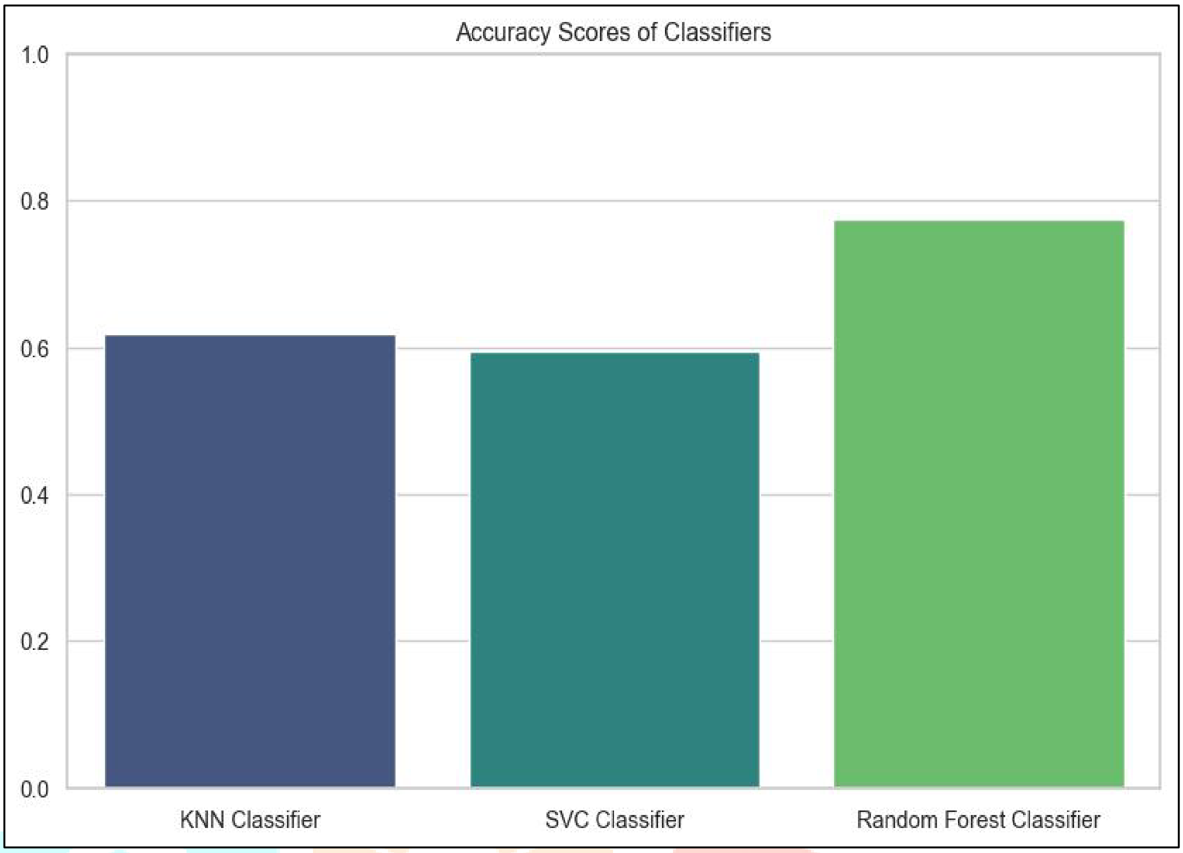


Figure 10: Accuracy scores of Classifiers

3.1.5 Interpretability and Fairness:

The system places a strong emphasis on interpretability and fairness, recognizing the importance of transparency and ethical considerations in credit scoring methodologies.

3.1.6 Model Interpretability Techniques:

To enhance transparency and foster trust in the credit scoring models, various interpretability techniques are applied. Feature importance analysis, SHAP (SHapley AdditiveexPlanations) values, and LIME (Local Interpretable Model-agnostic Explanations) are employed to explain the factors influencing credit score predictions. These techniques empower stakeholders to understand the rationale behind each decision made by the classifiers.

3.1.7 Fairness Measures:

Addressing potential biases and ensuring fair credit assessments are integral objectives. Fairness measures, including demographic parity and equalized odds, are implemented to evaluate and mitigate disparities in model predictions across different demographic groups. This ensures that the credit scoring system remains equitable and unbiased, adhering to ethical standards.

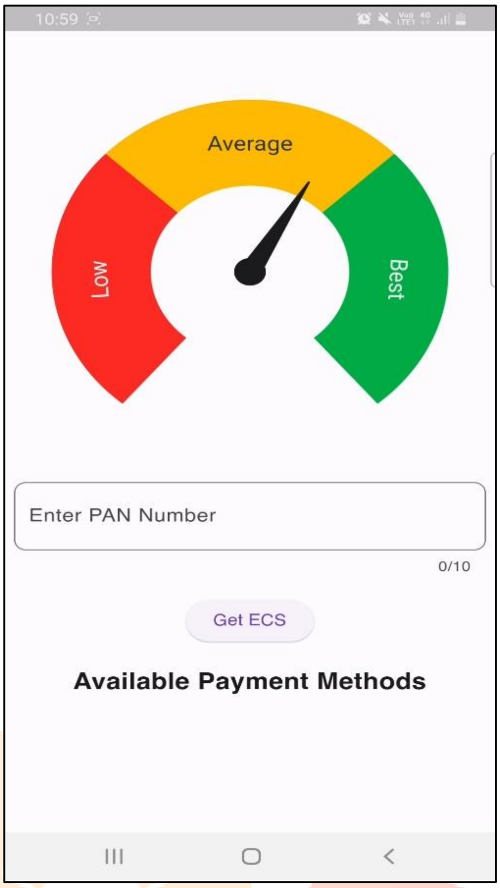
Ultimately, by prioritizing interpretability and fairness, the project not only meets ethical standards but also contributes to the broader discourse on responsible and inclusive machine learning applications in the financial domain. This commitment aligns with the project's goal to develop a credit scoring system that is not only accurate but also transparent and equitable.

IV. Example of a Real-time Application

An application of this research can be used to predict the credit score of a consumer by entering his or her unique ID. In India, we use “PAN” card. A permanent account number is a ten-character alphanumeric identifier Foundational ID, issued in the form of a laminated PAN card, by the Indian Income Tax Department, to any person who applies for it or to whom the department allots the number without an application [7]. Hence, we can consider the use of PAN as a unique identifier for any person above the age 18.

Using Flutter, a simple Android mobile application is built, to predict the credit score of the person whose PAN is entered. Now, PAN contains all the necessary details required to predict the credit score. After government clearance and approval, the data associated with PAN is collected and the total is transformed into a CSV file in cloud. The sample application retrieves this data from the CSV file and predicts the credit score. After predicting the credit score in three categories, ‘Low,’ ‘Average,’ ‘Best.’

According to the credit score determined, the mobile application provides available payment methods to that user. This functionality can be applied to e-commerce websites and many more applicable online transactions. In the below figures show how just adding of one extra credential (PAN), can make the consumer overlook the hassle of checking credit score, availability of EMI etc.

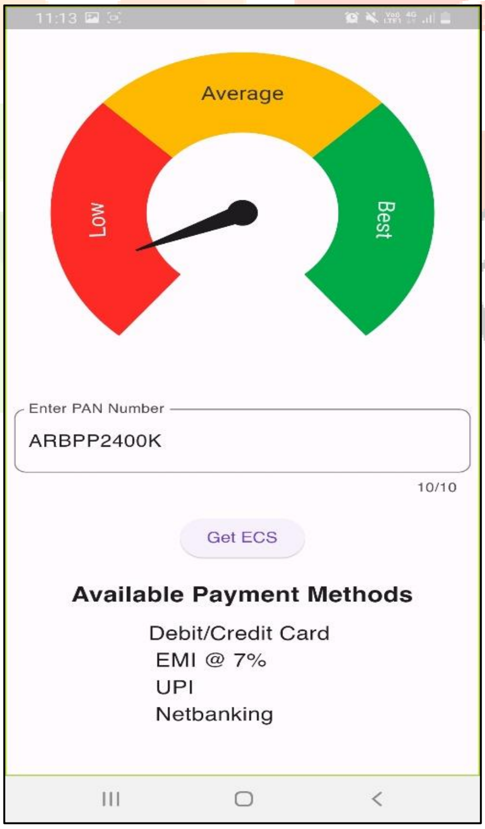


Enter PAN Number

0/10

Available Payment Methods

Figure 13: Home page of sample Android mobile application



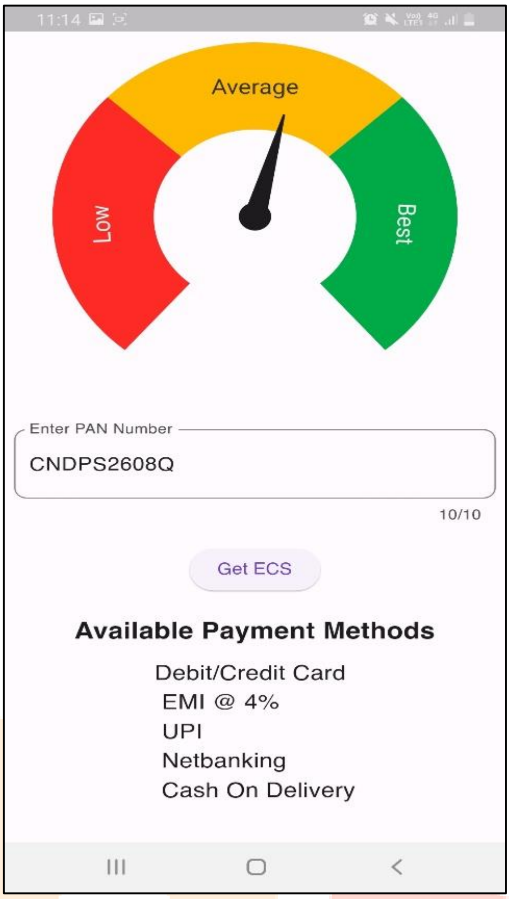
Enter PAN Number: ARBPP2400K

10/10

Available Payment Methods

Debit/Credit Card
EMI @ 7%
UPI
Netbanking

Figure 14: When low credit score PAN is entered, it provides EMI at a higher rate and no Cash on Delivery



Enter PAN Number

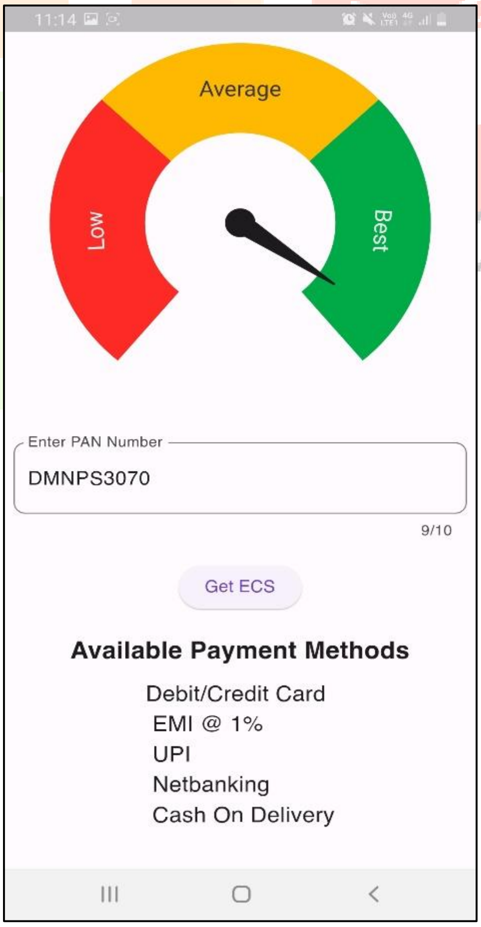
CNDPS2608Q

10/10

Available Payment Methods

Debit/Credit Card
EMI @ 4%
UPI
Netbanking
Cash On Delivery

Figure 15: When average credit score PAN is entered, it provides EMI at a reasonable rate and Cash on Delivery.



Enter PAN Number

DMNPS3070

9/10

Available Payment Methods

Debit/Credit Card
EMI @ 1%
UPI
Netbanking
Cash On Delivery

Figure 16: When higher credit score PAN is entered, it provides EMI at a low rate and Cash on Delivery.

V. Conclusions and Future Work

In conclusion, this system implementation embarked on a thorough exploration and analysis of a credit scoring dataset, employing various data preprocessing techniques and leveraging machine learning algorithms for predictive modelling. The initial steps involved cleaning the dataset by removing symbols and addressing incorrect data, ensuring the dataset's quality and reliability.

Visualizations, such as histograms provided a comprehensive overview of key financial indicators, data balance, and demographic distributions within the dataset. These visual insights laid the groundwork for a more informed approach to subsequent analyses.

The implementation of machine learning classifiers, specifically, K-Nearest Neighbours, Support Vector Classifier, and Random Forest Classifier, yielded valuable results. The Random Forest Classifier demonstrated superior accuracy, showcasing its efficacy for credit score prediction tasks. The confusion matrices and accuracy scores provided detailed assessments of the models' performance, revealing strengths and areas for improvement.

Furthermore, the system introduced a function to visualize confusion matrices, enhancing interpretability and aiding in the assessment of model performance across different credit score categories. Overall, this chapter contributes to the understanding of credit scoring methodologies, offering actionable insights for credit risk assessment. The combination of exploratory data analysis and machine learning techniques equips stakeholders with a robust framework for making informed decisions in the realm of credit evaluation.

As the financial landscape continues to evolve, the findings from this chapter provide a valuable foundation for future endeavours in credit scoring and risk management.

VI. REFERENCES

[1] Lyn Thomas, Jonathan Crook, and David Edelman in Credit Scoring and Its Applications (2017)

[2] Huang, J., Chai, J., & Cho S. (2020). Deep learning in finance and banking: A literature review and classification. Frontiers of Business Research in China, 14(1), 1-24.

[3] Wang, H., Kou, G., & Peng, Y. (2021). Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending. Journal of the Operational Research Society, 72(4), 923-934.

[4] Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Applied Soft Computing, 74, 26-39.

[5] Zhang, X., Han, Y., Xu, W., & Wang, Q. (2021). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. Information Sciences, 557, 302-316.

[6] Gibson Brandon, R., Krueger, P., & Schmidt, P. S. (2021). ESG rating disagreement and stock returns. Financial Analysts Journal, 77(4), 104-127).

[7] Wikipedia, Permanent account number, paragraph 1.