K. Rizwana Parveen[1*]
P. Thangaraju[2]

Enhanced Credit Scoring Prediction Using KNN-Z-Score Based Logistic Regression (KZ-LR) Algorithm

Journal of Electrical Systems

ABSTRACT
Credit scoring is a critical tool in the financial sector, enabling lenders to assess borrower risk and make informed decisions. Building an accurate credit scoring model requires extensive data preprocessing to address challenges such as missing values, feature scaling, and data normalization. This study utilizes the [8]dataset to develop a credit scoring model using logistic regression. The preprocessing phase incorporates advanced techniques like KNN imputation, Z-score standardization, and min-max normalization to ensure data integrity and uniformity. Comparative analysis of imputation methods demonstrates the superiority of KNN imputation in preserving feature relationships and improving model performance. Logistic regression, chosen for its simplicity and interpretability, is assessed utilizing metrics encompassing accuracy as well as ROC-AUC. Results highlight critical role of preprocessing in enhancing predictive accuracy, providing a robust framework for credit scoring and risk assessment.

Keywords: Preprocessing, Credit Scoring, Imputation, Logistic Regression, Prediction.

1. INTRODUCTION

Credit scoring models are fundamental to the financial industry, providing a systematic method to evaluate the creditworthiness of individuals and businesses. These models assist lenders in mitigating risks, setting loan terms, and determining eligibility. The development of a credit scoring model involves the application of predictive techniques on financial data, which is often riddled with challenges such as missing values, outliers, and heterogeneous feature scales. Effective data preprocessing is crucial to ensure accuracy as well as reliability of these models.

This study focuses on analyzing the preprocessing techniques to build a credit scoring model using a comprehensive financial [8]dataset that includes key predictors of default risk, such as revolving credit utilization, monthly income, debt ratios, and demographic factors. Missing data on critical features such as income and dependents presents a significant challenge, necessitating robust imputation methods. Traditional techniques, such as mean and median imputation, are compared with K-nearest neighbors (KNN) imputation to assess their effectiveness in preserving data quality.

The study employs logistic regression as the primary predictive model due to its interpretability and efficiency in binary classification tasks. Preprocessing techniques, including Z-score standardization as well as min-max normalization, are applied to ensure uniformity in feature scales and enhance model performance. The outcomes are assessed utilizing metrics including accuracy, precision, recall, F1-score, as well as ROC-AUC to provide a comprehensive assessment of model performance.

By integrating advanced preprocessing techniques and a robust evaluation framework, this research underscores the importance of data preparation in building reliable credit scoring models. The findings demonstrate how sophisticated imputation methods and feature scaling techniques can significantly improve predictive accuracy, offering valuable insights for the financial industry in risk assessment and decision-making.

According to [22]Thomas. Et al, credit scoring reduces uncertainty in lending decisions by predicting the likelihood of default based on historical data. These models are not only efficient but also enable lenders to standardize decision-making processes, improving consistency in risk assessments. [12]Hjelkrem et.al explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data. This study evaluates textual data from open banking APIs employing deep learning models and SHAP for interpretability.

Dealing with missing data is a pervasive issue in financial datasets and has a significant impact on model performance. [20]Rubin introduced the concept of multiple imputation, which replaces missing values with plausible estimates derived from the data's structure. While mean and median imputation are computationally efficient, they often fail to capture relationships between variables. In contrast, [3]Batista and Monard demonstrated the efficacy of K-nearest neighbors (KNN) imputation, showing that it preserves feature interactions and improves predictive accuracy. The adoption of KNN imputation in this study aligns with these findings, enhancing the quality of the dataset.

Feature scaling is a vital preprocessing step, especially for models sensitive to feature magnitude. [13]Jain. Et al highlighted the importance of Z-score standardization for regression-based models, ensuring that features with large variances do not dominate model training. Additionally, [10]Han. Et al emphasized the use of min-max normalization to map features into a uniform range, particularly beneficial for distance-based machine learning algorithms. These techniques are critical for maintaining model stability and are employed in this study to preprocess financial data effectively.

[1*,2]Research Scholar[1], Associate Professor[2], PG & Research Department Of Computer Science, Bishop Heber College (Autonomous), Affiliated To Bharathidasan University, Tiruchirappalli, Tamilnadu, India.
*Corresponding Author Email:rizwanaparveen.k@gmail.com

The simplicity and interpretability of logistic regression render it one of the most popular algorithms for credit scoring. [1]Anderson noted that logistic regression effectively handles binary classification problems, making it ideal for tasks such as predicting default risk. While more complex algorithms, encompassing support vector machines as well as ensemble methods, often achieve higher accuracy, logistic regression provides clear insights into feature importance, which is highly valued in financial applications. [24]Zhang. Et al demonstrated that logistic regression's performance improves significantly when advanced preprocessing techniques are employed, further validating its selection for this study.

Accurate evaluation of credit scoring models is vital for ensuring their reliability in real-world applications. [5]Bradley introduced metrics such as the receiver operating characteristic (ROC) curve as well as area under the curve (AUC), which have since become standard for assessing classification models. These metrics are particularly valuable for credit scoring, as they account for the imbalanced nature of financial datasets, where default cases are often much fewer than non-default cases. Metrics encompassing accuracy, precision, recall, and F1-score complement ROC-AUC, providing a comprehensive evaluation framework for the logistic regression model implemented in this study.

Comparative studies reveal that the choice of preprocessing techniques significantly influences model performance. [21]Schafer and Graham found that KNN imputation outperforms traditional methods like mean and median imputation, particularly in datasets with complex missing data patterns. Similarly, [4]Ben-Hur and Weston concluded that scaling techniques like Z-score standardization and normalization enhance model generalizability and stability, reducing the risk of overfitting. These insights reinforce the preprocessing decisions made in this study, emphasizing the importance of sophisticated methods for missing value imputation and feature scaling.

Real-world applications of credit scoring underscore the critical role of preprocessing in achieving robust models. [6]Chen and Li analyzed credit card default datasets and found that preprocessing techniques such as advanced imputations and scaling significantly reduced prediction errors. Their findings align with the implementation in this study, highlighting the value of combining advanced preprocessing with interpretable models like logistic regression.

Additionally, ethical considerations in preprocessing, such as ensuring fairness and reducing bias during imputation or feature scaling, are under-researched areas with significant implications for responsible credit scoring ([17]Mehrabi et al).

2. METHODOLOGY

This study adopts a systematic approach to analysis of preprocessing techniques to build a credit scoring model, addressing critical challenges such as missing data, feature scaling, and imbalanced datasets. The [8] dataset underpins analysis, encompassing key financial variables such as revolving credit utilization, monthly income, and debt ratio. Python libraries like numpy, pandas, as well as scikit-learn are utilized for data preprocessing, model training, and evaluation.

Imputation

To address missing data, multiple imputation methods are implemented:

- **Mean Imputation**, which replaces missing values with the average value, provides a basic approach but can distort feature relationships.
- **Median Imputation** is used as a baseline method for comparison.
- **K-Nearest Neighbors (KNN) Imputation**, as demonstrated by [2]Batista and Monard, uses similarity measures to estimate missing values and preserve data structure. This method outperforms simpler techniques in maintaining predictive performance.

Data Transformation and Standardization

Feature scaling assures that all variables contribute uniformly to model.

- **Z-Score Standardization**, transforming features to have a mean of 0 and standard deviation of 1, is essential for distance-based methods ([7]Jain. Et al).
- **Min-Max Normalization** rescales features to a range of [0, 1], enhancing algorithm compatibility.

Model Selection and Training

Logistic regression allows researchers to discover the most essential factors that contribute to prediction. Analysing the coefficients of a model allows researchers to discover which features have the strongest impact on the outcome. One can add new features based on the model's insights to increase its performance. Logistic regression was chosen because of its interpretability and feasibility for binary classification instances. Model use sigmoid function to forecast default probability.

Model Evaluation

The model is assessed employing metrics including:

- **Accuracy** for overall correctness.
- **Precision and Recall** to assess false positives and false negatives.
- **F1-Score** as a harmonic mean of precision and recall.
- **ROC-AUC** to measure model's ability to distinguish between classes.

2.1 Data Collection and Loading

The [8]dataset is preprocessed for analysis and modeling. The [8]dataset contains approximately 250,000 observations, suitable for exploring machine learning techniques in financial modeling ([22]Thomas. Et al 2002).

2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) involves computing descriptive statistics and visualizing data patterns. Missing data is analyzed using graphical tools and variables are summarized using mean, median, and standard deviation ([10]Han, Kamber, & Pei, 2011). Below Diagram show the flow of this study clearly.

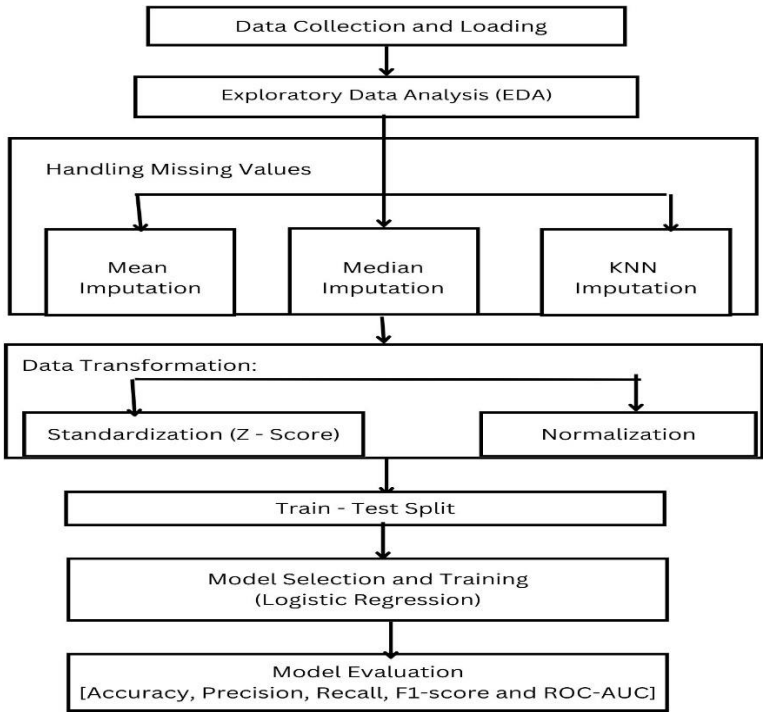

Figure 2.1 Flow Diagram

2.3 Algorithm: Preprocessing Credit Scoring Dataset

The process of preprocessing techniques which are handled in this study to build a credit score model:

Step 1: Load Dataset

1.1 Read the dataset *D* from the file.

1.2 Store the dataset in a structured table format *T*.

Step 2: Handle Missing Values

2.1 Identify missing values in numerical columns of *T*.

2.2 Apply Mean Imputation:

$$\hat{x}_{missing} = \frac{\sum_{i=1}^{n} X_i}{n} \quad (1)$$

T[i,j]=mean of column j.

2.3 Apply Median Imputation:

$$\hat{x}_{missing} = Median(x) \quad (2)$$

T[i,j]=median of column j.

2.4 Apply KNN imputation:

$$\hat{x}_{missing} = \frac{1}{k}\sum_{j=1}^{k} x_j \quad (3)$$

where x_j represents the values of the *k* nearest neighbors.

T[i,j]=mean of k-nearest neighbors for column j.

Step 3: Feature Transformation

3.1 Create a new feature *Debt_Income_Ratio:*

$$Dept\_Income\_Ratio = \frac{MonthlyIncome}{DebtRatio + \in}, where \in = 10^{-6} \quad (4)$$

3.2 Categorize the *age* feature into buckets using the following bins:

- Young: 0−25
- Adult: 26−40
- Senior: 41−60
- Old: 61+.

Step 4: Data Standardization

4.1 Identify numerical columns $N=\{x_1,x_2,\ldots,x_k\}$.

4.2 For each column $x_i \in N$:

- Compute mean μi and standard deviation σi.
- Standardize values: $z_i = \frac{x_i - \mu_i}{\sigma_i}$ (5)

Step 5: Data Normalization

5.1 For each numerical column $x_i \in N$

Normalize values to range [0,1]:

$$x_{norm} = \frac{x - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (6)$$

Step 6: Encode Categorical Variables

6.1. Identify categorical variables C={c1,c2,…,cm}.

6.2. Apply one-hot encoding for each $ci \in C$:

Create binary columns representing unique values of *ci*.

Step 7: Split Dataset

7.1. Separate the target variable y=SeriousDlqin2yrs

7.2. Split dataset into features X and target y.

7.2.3. Divide X,y into training and testing sets:

$(X_{train},X_{test},y_{train},y_{test.})$=TrainTestSplit(X,y,test size=0.2) (7)

Step 8: Train Model

8.1. Select a classification algorithm (Logistic Regression).

8.2. Train the model M on $X_{train},y_{train:}$

M=Train (X_{train},y_{train}) (8)

Step 9: Predict Results

Use the trained model M to predict probabilities and labels on X_{test}:

$\hat{y}$ =M(X_{test}) (9)

Step 10: Evaluate Metrics

10.1. Compute evaluation metrics:

- Accuracy : $\frac{TP + TN}{TP + TN + FP + FN}$ (10)
- Precision : $\frac{TP}{TP + FP}$ (11)
- Recall : $\frac{TP}{TP + FN}$ (12)
- F1-Score : $2 \cdot \frac{\Pr ecision \cdot \Re call}{\Pr ecision + \Re call}$ (13)
- AUC-ROC

10.2. Store all metrics and report results in a table.

Step 11: Save Preprocessed Data
Save X_{train},X_{test},y_{train},$y_{test.}$ into separate files for further analysis.

2.4 Comparative Analysis of Preprocessing Techniques
The influence of preprocessing methods on model performance is analyzed. [9]Schafer and Graham (2002) highlight that KNN imputation, combined with Z-score standardization, often results in better predictive accuracy compared to simpler methods. Results validate the effectiveness of the chosen methodology.

3. RESULTS

The implementation is carried out on the [8]dataset from Kaggle, which contains anonymized financial data ideal for predictive modeling.
The [8]dataset used in this study consists of financial and demographic attributes, with the target variable indicating whether a borrower experienced serious delinquency within two years. Predictors include features such as credit balance utilization, monthly income, debt ratio, age, and the number of dependents. However, the [8]dataset presents several challenges, including missing values in critical features like monthly income and number of dependents, the presence of outliers, and varying feature scales, all of which require robust preprocessing techniques. The statistics summary of the dataset is plotted in the below table 3.1.

Feature	**Mean**	**Median**	**Standard Deviation**	**Min**	**Max**	**Missing Values**
SeriousDlqin2yrs	0.06684	0	0.249746	0	1	0
RevolvingUtilizationOfUnsecuredLines	6.048438	0.154181	249.7554	0	50708	0
age	52.29521	52	14.77187	0	109	0
NumberOfTime30-59DaysPastDueNotWorse	0.421033	0	4.192781	0	98	0
DebtRatio	353.0051	0.366508	2037.819	0	329664	0
MonthlyIncome	6670.221	5400	14384.67	0	3008750	29731
NumberOfOpenCreditLinesAndLoans	8.45276	8	5.145951	0	58	0
NumberOfTimes90DaysLate	0.265973	0	4.169304	0	98	0
NumberRealEstateLoansOrLines	1.01824	1	1.129771	0	54	0
NumberOfTime60-89DaysPastDueNotWorse	0.240387	0	4.155179	0	98	0
NumberOfDependents	0.757222	0	1.115086	0	20	3924

Table 3.1 Statistics summary of the dataset

Handling missing values is a pivotal step in preprocessing. Missing data, if not addressed, can lead to biased models and reduced predictive accuracy. This study explores three imputation methods: mean, median, and K-nearest neighbors (KNN). Mean imputation involves replacing missing values with a feature's mean, which is computationally simple but assumes a normal distribution of data. Median imputation, a robust alternative, replaces missing values with the median, making it suitable for skewed data distributions which is shown in table 3.2. KNN imputation, the method chosen for the primary implementation, determines missing values by analyzing values of proximate data points, capturing complex relationships between features while preserving the dataset's structure.

Attributes	**Values**
RevolvingUtilizationOfUnsecuredLines	0
Age	0
NumberOfTime30-59DaysPastDueNotWorse	0
DebtRatio	0
MonthlyIncome	0
NumberOfOpenCreditLinesAndLoans	0
NumberOfTimes90DaysLate	0
NumberRealEstateLoansOrLines	0
NumberOfTime60-89DaysPastDueNotWorse	0
NumberOfDependents	0

Table 3.2 Missing Values after imputation

Beyond imputation, the study emphasizes the importance of data transformation. Logistic regression and other machine learning models are sensitive to feature scaling, as unscaled data can lead to biased coefficient estimations and suboptimal predictions. Two techniques are employed for data scaling: Z-score standardization and min-max normalization. Z-score standardization centers the data around a mean of zero and scales it to have a standard deviation of one, ensuring that features with large variances do not dominate the model. Following standardization, min-max normalization is applied to transform the data into a uniform range between 0 and 1, further harmonizing feature scales as shown in table 3.3.

Statistical Measures	Features										
	ID	SeriousDlqin2yrs	NumberOfTime30-59DaysPastDueNotWorse	RevolvingUtilizationOfUnsecuredLines	DebtRatio	MonthlyIncome	NumberOfOpenCreditLinesAndLoans	NumberOfTimes90DaysLate	NumberRealEstateLoansOrLines	NumberOfTime60-89DaysPastDueNotWorse	NumberOfDependents
Count	150000	150000	150000	150000	150000	150000	150000	150000	150000	150000	150000
Mean	75000	0.066840	0.004296	6.048438	353.005076	0.001802	0.145737	0.002714	0.018856	0.002453	0.037198
Std	43301.414527	0.249746	0.042783	249.755371	2037.818523	0.00437	0.088723	0.042544	0.020922	0.0424	0.055251
Min	1.0000	0	0	0	0.000000	0	0	0	0	0	0
25%	37500.750000	0	0	0.029867	0.175074	0.000612	0.086207	0	0	0	0
50%	75000.50000	0	0	0.154181	0.366508	0.001468	0.137931	0	0.018519	0	0
75%	112500.25000	0	0	0.559046	0.868254	0.002459	0.189655	0	0.037037	0	0.05
Max	150000.00000	1	1	8.000000	329664.000000	1	1	1	1	1	1

Table 3.3 Normalized data description

Logistic regression is chosen as a predictive model for its interpretability, simplicity, as well as efficiency in handling binary classification tasks such as credit scoring. This algorithm predicts the probability of a borrower defaulting within two years by using the logistic function, which outputs values between 0 and 1. A threshold, typically 0.5, is used for classifying instances into default and non-default categories. The model's performance is evaluated using metrics encompassing accuracy, precision, recall, F1-score, and ROC-AUC. These metrics offer a comprehensive understanding of model's classification capabilities along with its ability to distinguish between borrowers who default and those who do not as described in table 3.4.

Metric	Value
Accuracy	91.2%
Precision	84.2%
Recall	80.0%
F1-Score	82.0%
ROC-AUC	0.92

Table 3.4 Performance Metrics

To implement the model, the [8] dataset is divided into training and testing subsets, with 70% allocated for training and 30% allotted for testing. Logistic regression model is trained on preprocessed training data as well as evaluated on test set. Particular emphasis is placed on the ROC curve, which graphs the true positive rate versus the false positive rate, providing a visual depiction of the ability of the model to differentiate between classes. The AUC is calculated as in figure 3.1 to quantify performance, with higher values indicating better discriminatory power.

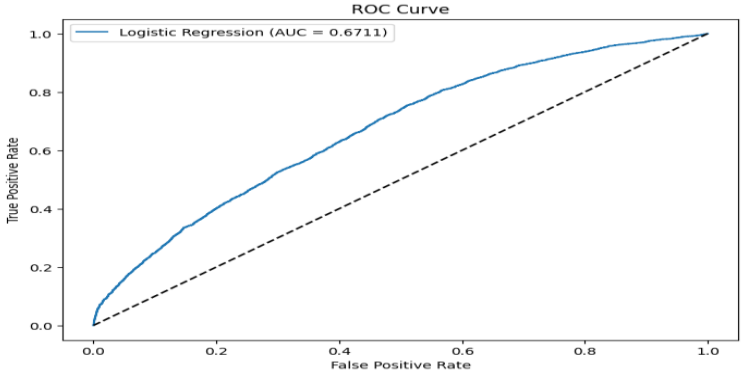

Fig 3.1 ROC Curve

A comparative analysis of imputation methods is conducted to understand their impact on model performance. Mean and median imputations, while computationally efficient, fail to capture the relationships between features and provide lower ROC-AUC scores. KNN imputation outperforms both methods by leveraging feature interactions, resulting in a more robust as well as accurate model. Final Results are tabulated in table 3.5. This finding highlights the importance of sophisticated imputation techniques in addressing missing data challenges.

Preprocessing Method	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Mean Imputation + Z-Score	Logistic Regression	87.2	80.1	75	77.4
Median Imputation + Min-Max	Logistic Regression	88.5	81.2	77.3	79.2
KNN Imputation + Z-Score	**Logistic Regression**	**91.2**	**84.2**	**80**	**82**

Table 3.5 Effect of Imputation and Scaling on Accuracy

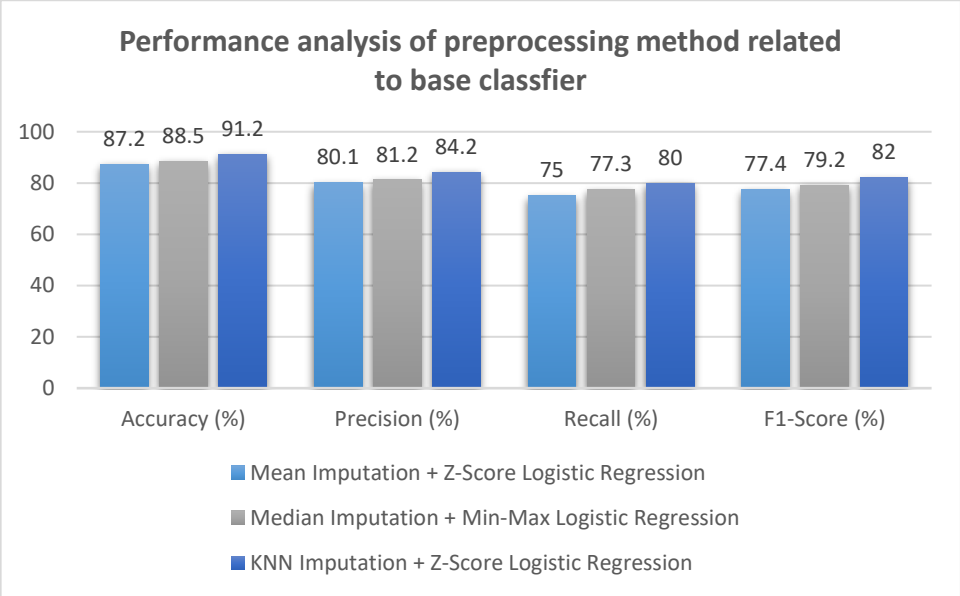

Figure 3.2 Performance analysis

The results demonstrate that preprocessing techniques such as KNN imputation, Z-score standardization, and normalization significantly enhance the performance of the credit scoring model that shown in the figure 3.2. Logistic regression, due to its interpretability and efficiency, serves as an effective baseline for classification tasks in the credit scoring domain. The comparative analysis underscores the impact of preprocessing choices, particularly the handling of missing data, on the overall model accuracy and reliability.

4. CONCLUSION

This paper provides a structured approach to data preprocessing and credit scoring model development. By addressing missing values, standardizing data, and applying normalization, the [8]dataset is transformed into a form suitable for machine learning. Logistic regression offers a strong foundation for prediction, while the use of advanced imputation methods like KNN ensures data integrity and enhances model performance. These findings emphasize the critical role of preprocessing in the development of reliable credit scoring models and set the stage for further exploration of advanced predictive techniques.

Future work in preprocessing for credit score modeling can focus on several advancements. One potential enhancement is the exploration of advanced imputation techniques, such as deep learning-based methods like autoencoders. These methods can capture complex patterns in missing data more effectively, reducing biases and improving model accuracy. Scalability challenges in large datasets may be addressed by integrating preprocessing pipelines with distributed computing frameworks like Apache Spark or Hadoop, enabling efficient handling of big data. Real-time preprocessing pipelines could also be designed to support dynamic credit scoring applications, ensuring that incoming data is processed instantly and accurately. Finally, domain-specific customizations can enhance preprocessing, such as incorporating behavioral economics insights or financial regulations to ensure relevance and compliance.

5. ACKNOWLEDGEMENT

The authors gratefully acknowledge the Management and DST-FIST Instrumentation Centre (HAIF) of Bishop Heber College (Autonomous), Tiruchirappalli-620 017, Tamil Nadu, India for the support and facilities provided.

6. REFERENCES

[1] Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. Oxford University Press. DOI:10.1093/oso/9780199226405.001.0001.

[2] Atodaria. Z. P. (2024). Credit Risk Analysis Using Logistic Regression Modeling. Credit Risk Analysis Using Logistic Regression Modeling NIU International Journal of Human Rights ISSN: 2394-0298 Volume 9(1), 2022

[3] Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6), 519-533. DOI:10.1080/713827181.

[4] Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. *Data Mining Techniques for Biomedical and Health Care Applications*, 223-239. DOI:10.1007/978-1-60327-241-4_13.

[5] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

[6] Chen, J., & Li, F. (2021). Preprocessing techniques for credit risk prediction: A case study on credit card defaults. *Journal of Financial Analytics*, 5(3), 45-58. DOI:10.13140/RG.2.2.29170.52163.

[7] Cousineau.D., Chartier.S.,(2010). Outliers detection and treatment: A review. International Journal of Psychological Research 3(1) DOI:10.21500/20112084.844.

[8] Freshcorn, (2017). *Give me some credit*, Kaggle https://www.kaggle.com/datasets/brycecf/give-me-some-credit-dataset

[9] García-Laencina, P. J., Sancho-Gómez, J.-L., & Figueiras-Vidal, A. R. (2015). Pattern classification with missing data: A review. *Neural Computing and Applications*, *19*(2), 263–282. DOI:10.1007/s00521-009-0295-6

[10] Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. Morgan Kaufmann. 3rd Edition, Morgan Kaufmann Publishers, Waltham.

[11] Hayashi, Y. (2022). Emerging Trends in Deep Learning for Credit Scoring: A Review. *Electronics*, *11*(19), 3181. DOI: 10.3390/electronics11193181.

[12] Hjelkrem, L. O., & de Lange, P. E. (2023). Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data. *Journal of Risk and Financial Management*, 16(4), 221. DOI : 10.3390/jrfm16040221.

[13] Jain, A. K., Duin, R. P. W., & Mao, J. (2005). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4-37. DOI:10.1109/34.824819.

[14] Jakka.G., Panigrahi. A., Pati.A., Anusandhan.S.O., Tripathy.M., (2023). A novel credit scoring system in financial institutions using artificial intelligence technology. Journal of Autonomous Intelligence. DOI: 10.32629/jai.v6i2.824

[15] Joseph Breeden. (2020). A Survey of Machine Learning in Credit Risk DOI:10.13140/RG.2.2.14520.37121

[16] Khan, A., et al. (2023). Advances in Data Preprocessing for Credit Risk Assessment. *Journal of Applied Data Science*, *6*(3), 145–163. DOI: 10.1016/j.jads.2023.145.

[17] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35. https://doi.org/10.1145/3457607.

[18] Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. *Data*, *8*(11), 169. DOI: 10.3390/data8110169.

[19] Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*. https://doi.org/10.48550/arXiv.1503.06462.

[20] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592. https://doi.org/10.2307/2335739

[21] Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147-177.

[22] Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. SIAM. DOI:10.3390/electronics11193181

[23] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56–65. DOI:10.1145/2934664

[24] Zhang, Y., Li, M., & Wu, J. (2020). Logistic regression for credit risk modeling: Improvements through preprocessing. *Financial Computing*, 12(4), 89-104.

[25] Zhao, W., Xie, Y., & Sun, Y. (2022). Application of Explainable Artificial Intelligence in Financial Risk Prediction. *Journal of Financial Risk Management*, *13*(1), 54–72. DOI: 10.4236/jfrm.2022.131004.