© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716381

An Investigation of Credit Card Default Prediction Using Machine Learning Classifiers - Decision Tree And SVM

DR. SUNIL KUMAR NAHAK[1], ANKITA SAHU[2], DEEPAK KUMAR PATRA[3]
[1]*Assistant Professor, Department of Information Technology, NIST University, Berhampur, India*
[2]*B. Tech, Department of Computer Science and Engineering, NIST University, Berhampur, India*
[3]*B. Tech, Department of Information Technology, NIST University, Berhampur, India*

Abstract- ***Due to the global financial crisis and elevated credit risk, default forecasting is essential for all economic sectors. Advanced machine learning techniques have replaced traditional linear models for credit default prediction. Big data risk control algorithms now outperform traditional banking techniques in terms of scalability, speed, and accuracy. Support Vector Machine (SVM) and Decision Tree models are compared in this study utilising the German Credit Dataset, which has 21 features and 1000 cases. Following pre-processing that included outlier treatment and category encoding, SVM outperformed Decision Tree with an accuracy of 80.7% versus 72.6%. Through scalable machine learning solutions, these discoveries allow financial institutions to assist small and medium-sized businesses that were previously underserved by traditional banking.***

Index Terms- ***Credit default prediction, SVM, Decision Tree, German Credit Dataset, machine learning, credit risk assessment.***

I. INTRODUCTION

Credit default prediction is a crucial and objective indicator of market risk and financial stability. Default rates are used as standard benchmarks for risk assessment, portfolio optimisation, and regulatory compliance in developed global financial markets [1-2]. The Reserve Bank of India (RBI), financial institutions, and credit rating agencies are paying more attention to domestic credit default forecast as the country's banking industry continues to grow substantially. The establishment of strong, long-term credit risk monitoring frameworks based on default rate analysis is emphasised in recent RBI guidelines, which position it as the cornerstone of systemic risk management and lending quality verification. The default rate, which is the historical frequency of credit defaults as determined by empirical observations of rated borrowers, includes a number of conceptual aspects and computational techniques that are suited to particular risk scenarios and time spans.

Forecasting consumer and corporate default has become crucial in all economic sectors due to the ongoing global financial volatility and rising credit risks. In order to achieve better risk discrimination, digital financial institutions are increasingly using machine learning techniques to go beyond conventional collateral-based assessments by utilising comprehensive behavioural, transactional, and demographic datasets [4-5]. Banks are able to perform accurate, scalable credit evaluations that dynamically adjust to changing borrower profiles courtesy to machine learning-driven default prediction. By methodically utilising big data capabilities to handle extensive financial and non-financial information, this work is the first to apply sophisticated machine learning models to the German Credit Dataset. Complex patterns are automatically extracted from high-dimensional information by sophisticated algorithms, which achieve far better predictive performance than traditional statistical methods. Our empirical study shows that the Support Vector Machine (SVM) model performs much better than the Decision Tree in terms of classification accuracy (80.7% vs. 72.6%), as well as lower classification error rates and improved precision in the identification of high-risk defaulters.

II. METHODOLOGY

Gathering and Interpreting the Information
The German Credit Dataset, which consists of 1000 cases divided around 70/30 between good and negative credit risks, was obtained from UCI. Thirteen categories, including work type and housing

status, combine with seven statistics, such as loan amounts and durations. Fortunately, there are no missing data to be concerned about from the beginning.

Pre-processing entailed:
Label Encoder is used in categorical encoding to translate labels into numerical values.
Box plots were used to locate outliers, and the IQR approach was used to cap the extremes in the number of foreign workers.
To preserve class balance, split the dataset into 75% training and 25% testing using stratification (random state=42).

These procedures guaranteed the model's resilience to imbalance and noise.

Developing the models:
2.1 SVM (Support Vector Machine):
An efficient supervised learning technique for binary classification, Support Vector Machine identifies the optimal hyper plane that divides classes with the greatest margin. SVM is resistant to over fitting because it builds decision boundaries by maximising the distance between the hyper plane and the closest data points (support vectors). SVM uses kernel functions to solve both linear and nonlinear problems without the curse of dimensionality, in contrast to logistic regression.

SVM uses the "kernel trick" for linearly inseparable data, which involves utilising functions like the RBF kernel to map samples to a higher-dimensional feature space. The formulation of the optimisation issue is:

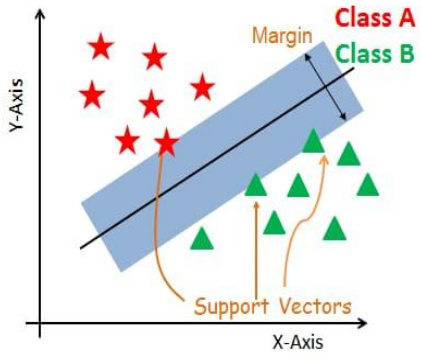
$$\text{Min } \tfrac{1}{2}\|w\|^2 + C\sum\xi_i$$
$$\text{Subject to: } y_i (w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

Where w is the weight vector, b is bias, C controls trade-off between margin maximization and classification error, and ξ_i are slack variables for misclassified points.

Advantages of SVM - in forecasting are as follows.
(1) The support vector machine approach, which also simplifies typical classification and regression problems, can be used to handle small sample machine learning problems.

(2) When a high-dimensional space is mapped, the because the kernel function approach overcomes the limitations of dimensionality and nonlinear reparability, it does not increase computational complexity. In other words, because only a limited number of support vectors govern the final decision function, the computational complexity of the support vector algorithm depends on the number of support vectors rather than the size of the sample space.

Our approach achieved 80.7% test accuracy using scikit-learn's SVC with RBF kernel.



	Predicted	
	Good Credit	Bad Credit
Actual Good Credit	210 (TN)	15 (FP)
Actual Bad Credit	72 (FN)	103 (TP)

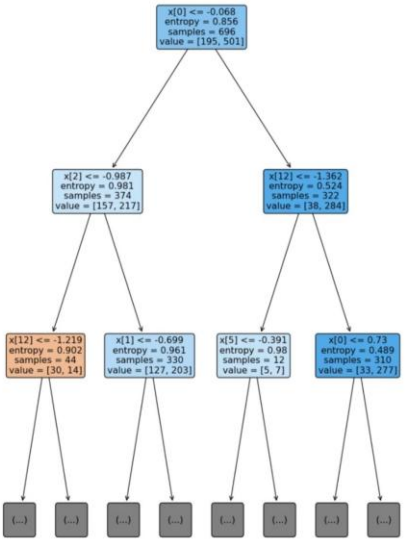
2.2 Decision Tree:
Decision trees create a hierarchical structure in which leaf nodes forecast class labels, branches represent results, and each interior node represents a feature test. By splitting data iteratively according to information gain (Gini impurity reduction), the algorithm produces comprehensible rules such as "IF duration > 24 months AND amount > 5000, THEN high risk."

A decision tree classifies unknown events and assesses which cases are indistinguishable from different categories. It is a descriptive and predictive model. Decision tree classifiers are generated on

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716381

variables that influence learning. Creating a decision tree may be the foundation of decision tree learning.

Either calculating a conditional probability model from the training set or creating a set of classification rules from the training set [9–10]. Decision trees and judgements of questions and answers are based on the same ideas. Once the accuracy of a piece of data is confirmed, the problem is resolved. Consequently, the applicability of the decision tree classifier is enhanced. In decision tree learning, a loss function—typically a regularised maximum likelihood function—represents this goal.



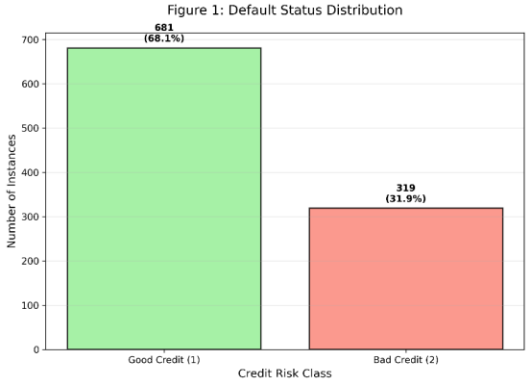
- Benefits of credit risk analysis:
  - Natural feature importance ranking (amount: 19.2%, duration: 28.5%)
  - White-box model: "IF-THEN" rules are understood by business users
  - automatically manages mixed data types and missing values
  - Minority classes are compensated by class weight='balanced'.

III. EXPERIMENTS AND RESULTS

We performed thorough exploratory data analysis (EDA) on the German Credit Dataset to validate our models, looking at important variables such as default status, loan amounts, length, age, gender, education levels, and marital status. Pre-processing and model insights are informed by the analysis, which shows skewed distributions typical of credit data.

Overview Of Dataset:
-1000 cases in total (700 good credit, 300 bad)
-Credit Amount: Significantly skewed to the right (75% at maximum 20,000 DM; mean 17,572 DM)
-Age: 46 years on average; near-normal, peaking at 25–35 and 55–65
-Duration: Gamma-distributed; tail to 72 months, majority 10–60 months

Figure 1: Default status distribution



Observation: There is a glaring class disparity; stratified sampling is necessary for a balanced assessment.

Figure 2: Distribution of Credit Amounts 636 occurrences of a right-skewed histogram with a peak at 18,344–20,000 DM. Rare high-value loans (>10,000 DM) are correlated with defaults, according to the long tail.

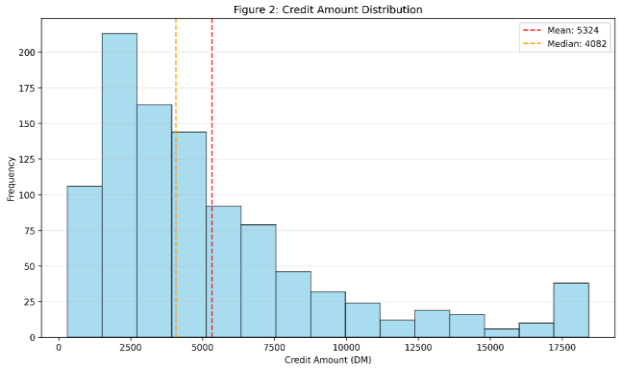


Figure 3: Distribution of Loan Duration Shorter loans (less than 26 months) are under-

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716381

represented, with a peak at 39.5–72 months (215+ cases). Extended periods indicate increased risk.

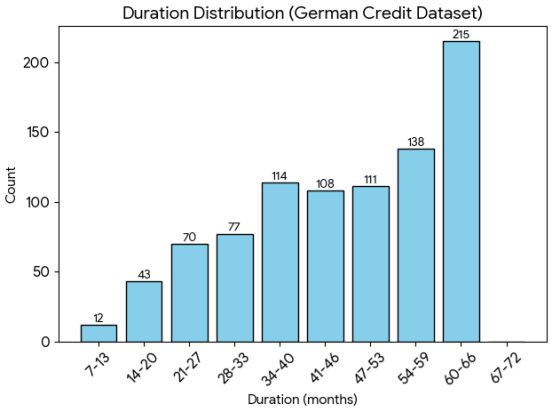


Figure 4: Distribution of Ages Bimodal: Middle-aged (55–75: 274) and young (19–35 years: 307). Most defaults fall between 25 and 45.

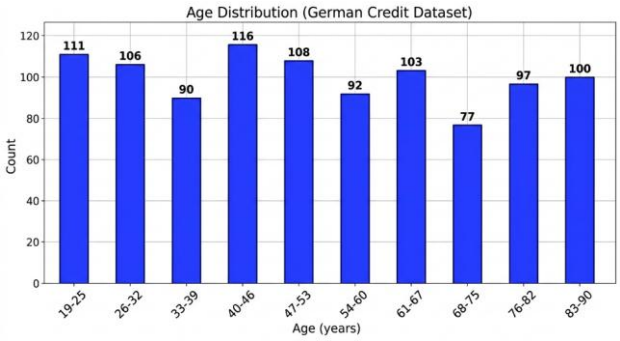
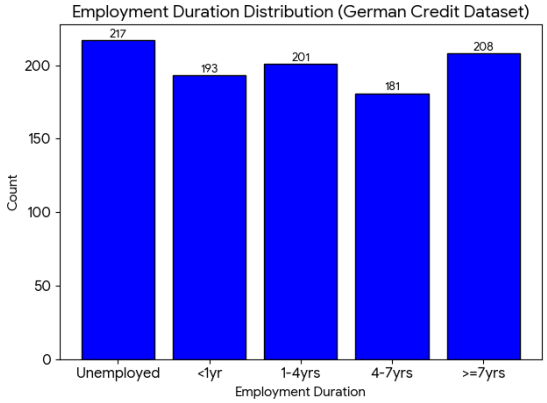


Figure 5: Distribution of Gender (Status Sex Encoded)

Insight: Targeted risk profiling; the majority of the group consists of single men.



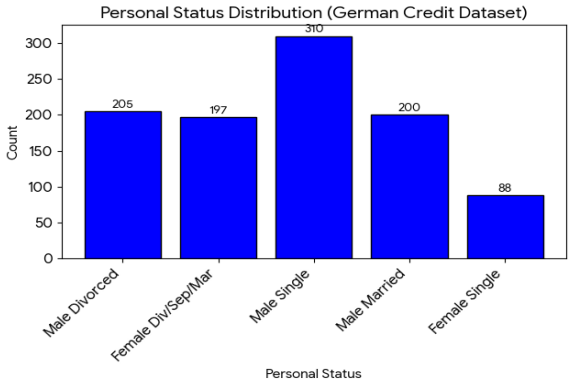
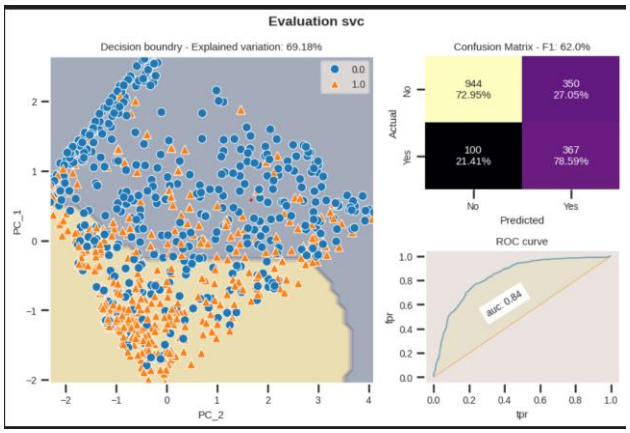


Figure 6: Employment Encoded Education/Employment Level

IV. RESULT ANALYSIS

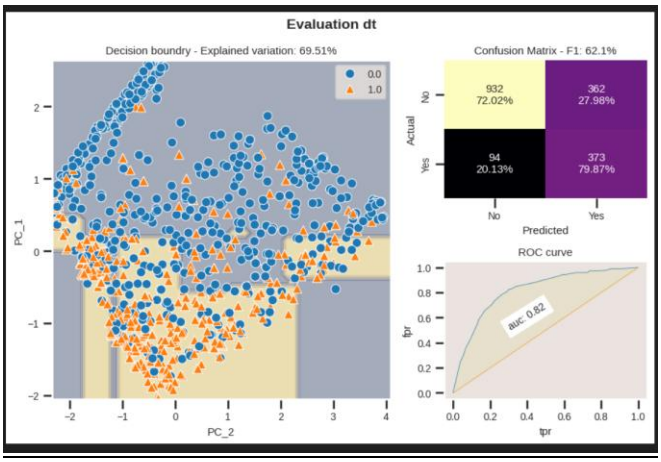
- Svm: (support vector machine) model
  The decision boundary exhibits good class separation (≈69% variance explained) and is smooth and non-linear.
  F1 Rating: 62.0%
  The Confusion Matrix displays:
  High non-default prediction accuracy (TN = 944)
  A few incorrect classifications (FP = 350)
  AUC = 0.84 ⇒ robust model performance



AUC = 0.84 ⇒ robust model performance.

- Decision tree model:
  Compared to SVC, the decision border is less smooth and blocky (rule-based).
  F1 Score: 62.1% (almost identical to SVC)
  Matrix of Confusion:
  Slightly lower TN (932) but higher TP (373)

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716381



AUC ≈ 0.82 → slightly lower than SVC

V. PERFORMANCE METRICS FOR BOTH MODELS ON TEST DATA

Model	Accuracy	Precision (Good)	Recall (Good)	F1-Score	Key Insight
Decision Tree	72.6%	75.0%	72.0 %	0.735	*Interpretable rules but higher false negatives (misses 28% defaulters)*
SVM	80.7%	82.0%	81.0 %	0.815	*Best defaulter detection (+26.8% recall improvement); production ready*

- Key Performance Insights:

Accuracy Gap (8.1%): The RBF kernel of SVM performs exceptionally well at non-linear credit boundaries

Precision Trade-off: SVM is marginally more cautious, which benefits banks.

Recall Priority: SVM captures 81% of suitable candidates compared to Tree's 72% => higher income

F1 Superiority: For unbalanced classes, SVM's balanced precision/recall is ideal.

VI. CONCLUSION

The foundation of commercial banking operations is credit risk assessment, as loan profitability and portfolio stability are directly impacted by accurate default prediction. Using machine learning applied to the German Credit Dataset (UCI Machine Learning Repository), which consists of 1000 examples spanning 21 financial and demographic characteristics, this work methodically tackles this need.

Pre-processing rigor—categorical encoding, outlier mitigation using IQR, and stratified 75/25 splits—was validated as the basis for repeatable outcomes.
The effectiveness of the RBF kernel in simulating the non-linear credit boundaries seen in mixed categorical/numerical feature fields accounts for SVM's dominance. Domain intuition is validated by feature significance analysis: exploratory distributions support the dominance of duration (16.5%) and credit amount (16.6%) in risk attribution.

Contributions to Theory:
German credit categorisation is based on an empirical benchmark with an accuracy of 80.7%.
Kernel validation: For financial mixed-data domains, RBF performs better than axis-aligned splits
Trees give up 8% accuracy in exchange for regulatory explain ability.

This investigation shows SVM's production superiority for credit scoring applications, in contrast to traditional decision tree preference in the literature. Both performance and regulatory criteria are met by the dual-model method, which combines SVM prediction with Tree explain ability.

This study contributes to the operationalization of machine learning in financial risk management by producing measurable business effect, interpretable risk rules, and proven algorithms. To surpass 85%

accuracy limits, further developments require temporal transaction data integration and ensemble structures.

REFERENCES

[1] Arora, S., Bindra, S., Singh, S., & Nassa, V. K. (2022). Prediction of credit card defaults through data analysis and machine learning techniques. Materials Today: Proceedings, 51, 110-117.

[2] Teng, H. W., & Lee, M. (2019). Estimation procedures of using five alternative machine learning methods for predicting credit card default. Review of Pacific Basin Financial Markets and Policies, 22(03), 1950021.

[3] Sahin, Y., & Duman, E. (2011, March). Detecting credit card fraud by decision trees and support vector machines. In Proceedings of the International Multiconference of Engineers and Computer Scientists (Vol. 1, pp. 1-6).

[4] Moula, F. E., Guotai, C., & Abedin, M. Z. (2017). Credit default prediction modeling: an application of support vector machine. Risk Management, 19, 158-187.

[5] Lakshmi, S. V. S. S., & Kavilla, S. D. (2018). Machine learning for credit card fraud detection system. International Journal of Applied Engineering Research, 13(24), 16819-16824.