iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** XI **Month of publication:** November 2026

DOI: https://doi.org/10.22214/ijraset.2026.85578

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue XI Nov 2026- Available at www.ijraset.com

Predicting Banking Customer Churn Using Machine Learning

Mrs. V. Meena[1], Dinesh M[2], Gowtham C[3], Harikaran M[4]
[1]*Assistant Professor (Project Guide),* [2, 3, 4]*B.E. Students, Department of Computer Science and Engineering, Adhiyamaan College of Engineering (An Autonomous Institution), Dr. M.G.R. Nagar, Hosur – 635130, Tamil Nadu, India*

Abstract: ***Customer churn is a major challenge in the banking industry, as customers may stop using a bank's services and move to other banks for various reasons. Identifying customers who are likely to leave is important for banks to improve customer retention and maintain long-term customer relationships. The project titled Banking Customer Churn Prediction Using Machine Learning aims to analyse customer data and predict whether a customer is likely to stay with the bank or leave. The system begins by preprocessing the dataset to prepare the data for analysis, followed by Exploratory Data Analysis (EDA) to understand customer behaviour, identify patterns, and examine the factors associated with customer churn. Customer information, such as age, credit score, account balance, tenure, number of products, geographical location, and account activity, is analysed based on the available dataset.***

Machine learning classification algorithms are trained and evaluated to identify a suitable model for predicting customer churn. The selected XGBoost model is used to estimate the probability of a customer leaving the bank. A user-friendly dashboard is developed using Streamlit to present the analysis and prediction results through graphs, a correlation heatmap, customer risk distributions, geographical comparisons, and individual customer predictions. The dashboard also allows users to enter customer details and view the predicted churn probability, helping them understand whether a customer may be at risk of leaving.

The proposed system helps identify customers who may require additional attention and provides useful insights into customer behaviour. These insights can support banks in developing appropriate customer retention strategies and making informed business decisions. Thus, the project demonstrates the practical application of Machine Learning and Data Analytics in predicting banking customer churn and supporting effective customer relationship management.

Index Terms: Customer Churn, Banking, Machine Learning, Churn Prediction, Exploratory Data Analysis, XGBoost, Customer Retention.

I. INTRODUCTION

Customer churn is an important challenge for banks because the loss of existing customers affects revenue, customer relationships and business growth. Banks often do not know in advance which customers are likely to leave, and manual analysis of customer records is time-consuming and may not identify high-risk customers early. A data-driven system that analyses customer and account-related data can therefore support timely retention decisions.

In this work, a systematic data science and machine learning workflow is followed, including data collection, data preprocessing, EDA, feature engineering, model development, model evaluation and churn prediction. The prepared dataset is used to train multiple classification models, whose performance is compared using standard metrics, and the best model is selected for predicting the churn probability of new customers. The main objectives of the work are listed below.

1) Data collection and analysis. To collect and analyse customer-related data in order to understand customer behaviour and identify the major factors that contribute to churn.
2) Data preprocessing. To handle missing values, duplicate records, categorical variables and other data-quality issues so that reliable data is prepared for model development.
3) Exploratory data analysis. To use statistical analysis and visualizations to identify patterns, relationships and important churn-related factors.
4) Model development and comparison. To develop multiple classification models and to compare them quantitatively using accuracy, precision, recall, F1-score, confusion matrix and ROC-AUC.

©IJRASET: All Rights are Reserved

5) Churn prediction interface. To use the selected model to generate churn predictions for new customer inputs through an application interface

6) Customer retention support. To provide reliable and measurable predictions that help banks identify high-risk customers at an early stage.

II. LITERATURE REVIEW

Iranmanesh et al. (2019). Iranmanesh and co-authors studied customer churn prediction using Artificial Neural Networks. Their research focused on analyzing customer behavior and identifying customers who might leave a company. This study highlights the importance of predicting customer churn to improve customer retention.

Peng et al. (2023) [6]. Peng and co-authors studied customer churn prediction using Machine Learning algorithms such as Decision Tree, KNN, LightGBM, and XGBoost. They also used SHAP to understand the factors influencing predictions. This study supports the use of multiple models and feature analysis in our project.

Singh et al. (2024). Singh and co-authors investigated banking customer churn using Machine Learning algorithms and data visualization. Their study included a visualization application to help users understand churn information. This research is related to our project because we aim to present churn predictions and analysis clearly.

Owolabi et al. (2024). Owolabi and co-authors compared Machine Learning models for customer churn prediction in banking and financial services. Their research focused on identifying customers who might leave and supporting customer-retention strategies. This study supports comparing different models in our project.

Boozary et al. (2024). Boozary and co-authors studied different Machine Learning algorithms for predicting bank customer churn. Their work included comparing classification techniques such as Random Forest and Support Vector Classifier. This research supports our approach of evaluating different models to predict customer churn.

Jin (2025). Jin studied bank customer churn prediction using Machine Learning models. The research discussed Random Forest, feature selection, and techniques for handling imbalanced data. This study highlights the importance of data preprocessing and selecting relevant features before model training.

Nisha T. N. and Dhanya Pramod (2025). The authors studied feature reduction using Genetic Algorithms for bank customer churn prediction. Their research focused on selecting useful features to support model development. This study is related to our project because feature analysis can help improve the modelling process.

Summary. The reviewed studies show that Machine Learning can help predict customer churn in the banking sector. Previous research has explored different algorithms, feature selection, data preprocessing, model evaluation, and visualization. Based on these studies, our project develops a systematic approach to analyze banking customer data, compare Machine Learning models, and predict customers who may leave the bank. The system aims to support banks in identifying at-risk customers and improving customer-retention decisions.

III. PROPOSED SYSTEM AND METHODOLOGY

A. Limitations of Existing Approaches

In existing approaches, customer churn is analysed using customer information, account details, transaction history and service-related data. Some systems rely on a single machine learning algorithm, perform limited EDA and feature engineering, report only accuracy without precision, recall, F1-score and ROC-AUC, and provide no simple interface for obtaining immediate predictions. As a result, important churn-related patterns may not be identified and the reasons behind predictions are difficult to understand.

B. Proposed System

The proposed system follows a complete data science workflow, as shown in Fig. 1. The customer dataset is first prepared by handling missing values, duplicate records and categorical data. EDA is then performed using statistical methods and visualizations to understand customer behaviour. After the relevant features are selected, multiple classification models are trained and compared, and the best-performing model is saved. Finally, a Streamlit-based interface accepts customer details and returns the churn probability, prediction status and risk category.

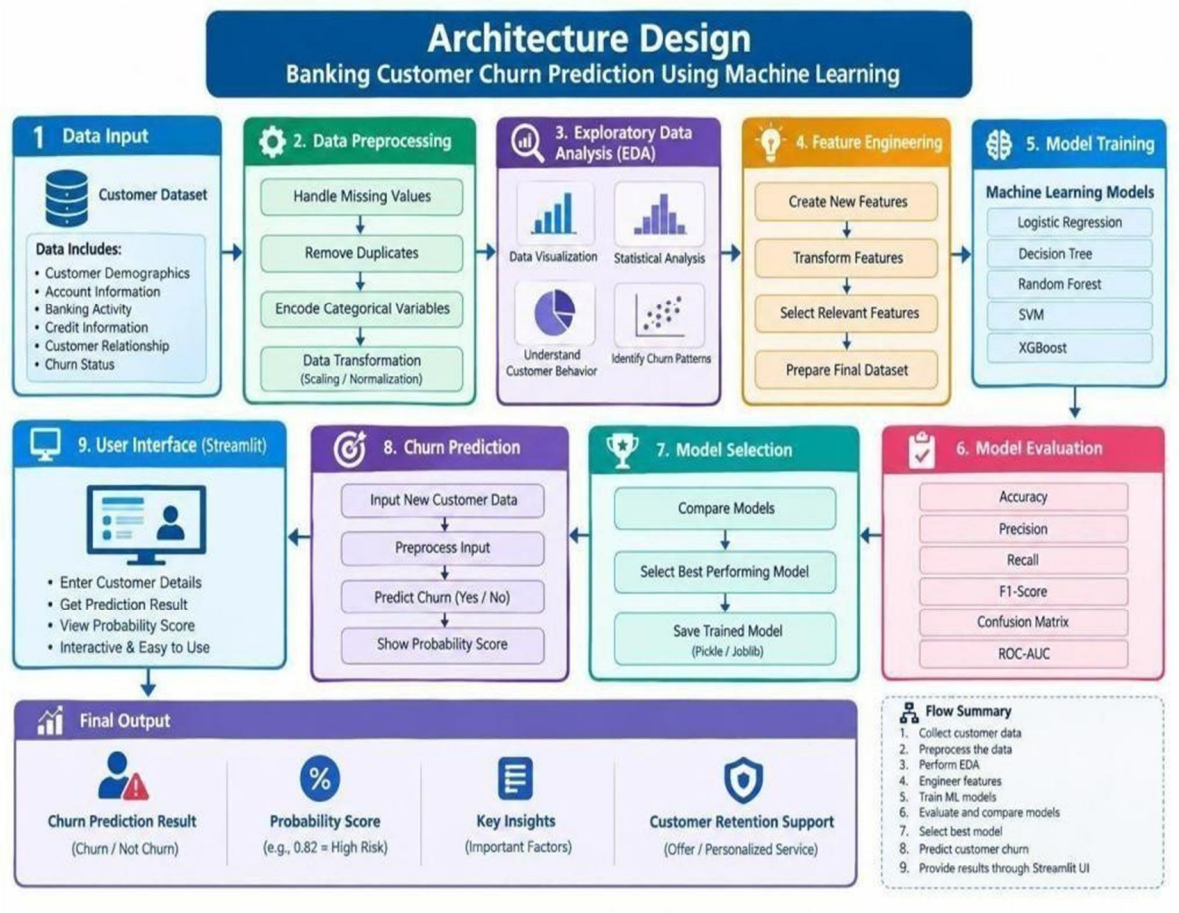


Fig. 1: Architecture of the proposed system

C. Description of Modules

1) Data input. The customer dataset (Churn_Modelling.csv) containing customer, account and banking-related attributes is loaded and inspected.

2) Data preprocessing. Missing values, duplicate records, categorical variables and inconsistent values are handled, and feature scaling is applied where required.

3) Exploratory data analysis. Distributions, correlations and differences between churn and non-churn customers are studied using statistical analysis and visualizations.

4) Feature engineering. Relevant features are selected, transformed or created, and the input features are separated from the target variable.

5) Model training and evaluation. The data is split in an 80:20 ratio using stratified sampling, and the models are trained and evaluated on unseen test data.

6) Model selection and prediction. The best model is saved and used to predict churn for new customer inputs through the Streamlit interface.

D. Data Flow

The data flow of the system is shown in Fig. 2. The dataset passes through preprocessing, EDA and feature engineering, and is divided into training and testing sets. The trained models are evaluated using accuracy, precision, recall, F1-score and ROC-AUC, and the best model is stored in a model store. During prediction, the saved model is loaded, the same preprocessing is applied to the entered customer details, and the churn probability and risk level are returned.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

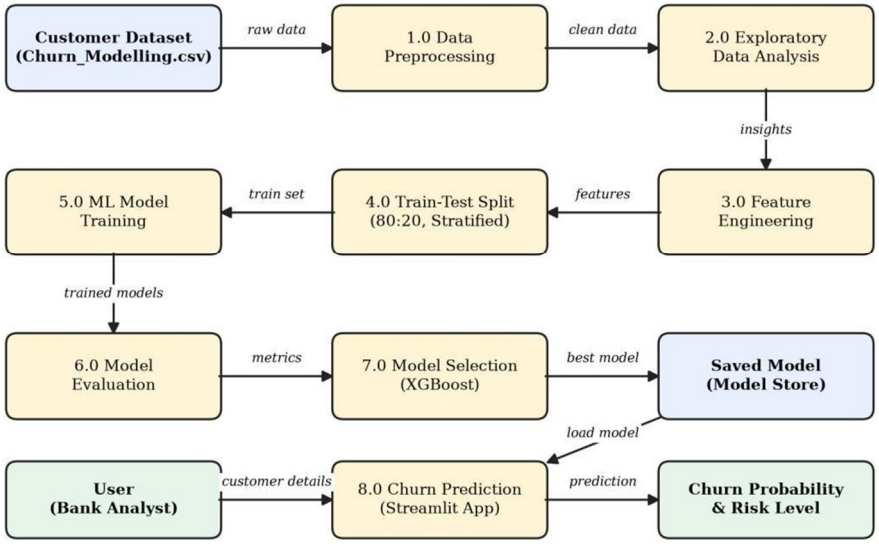


Fig. 2: Data flow diagram

IV. IMPLEMENTATION

A. *Tools and Environment*

The system is implemented in Python using Jupyter Notebook. A multi-core processor with at least 4 GB of RAM (8 GB recommended) is sufficient because the data is structured and of modest size. The libraries used are summarised in Table I.

Table I. Software tools used

Tool	Purpose
Pandas and NumPy	Data loading, cleaning and numerical processing
Matplotlib and Seaborn	Visualizations for EDA and correlation heatmap
Scikit-learn	Preprocessing, train-test splitting, models and evaluation
XGBoost	Gradient boosting classifier
Streamlit	Interactive web-based prediction interface

B. *Preprocessing and Feature Engineering*

The dataset is checked for missing values, duplicate records, incorrect data types and inconsistent values. Numerical and categorical features are identified separately; categorical features are converted into numerical form using encoding, and numerical features are scaled where required by the selected models. Identifier columns that do not contribute to prediction are removed, and the target variable (Exited) is separated from the input features.

C. *Dataset Splitting and Model Development*

The dataset is split in the ratio of 80:20, with 80% of the records used for training and 20% used for testing. Stratified splitting is applied so that the proportion of churn and non-churn customers is the same in both sets. The classification models trained are Logistic Regression, Support Vector Machine, K-Nearest Neighbours, Decision Tree, Random Forest, Extra Trees, AdaBoost, Gradient Boosting and XGBoost.

D. *Model Evaluation*

The models are evaluated on the unseen testing data. Accuracy measures the overall percentage of correct predictions, precision measures how many customers predicted as churn actually churn, and recall measures the ability to identify customers who actually churn. The F1-score balances precision and recall, the confusion matrix shows correct and incorrect predictions for each class, and ROC-AUC measures the ability of the model to distinguish between churn and non-churn customers across classification thresholds.

E. Prediction Interface

The Streamlit application loads the saved model and applies the same preprocessing used during training to the entered customer details. The result is displayed as “Likely to CHURN” or “Likely to STAY” together with the churn probability, risk category and decision threshold.

V. RESULTS AND DISCUSSION

Based on the evaluation results, XGBoost is selected as the final model. Its performance on the testing data is given in Table II and shown in Fig. 3.

Table II. Performance of the selected XGBoost model

Metric	Value (%)
Accuracy	85.75
Precision	65.72
Recall	62.65
F1-Score	64.15
ROC-AUC	87.07



Fig. 3: Model performance

The EDA results are shown in Fig. 4. The overall filtered churn rate in the analysed data is 20.4%, which indicates that a significant portion of customers belongs to the churn class. Churn is analysed by geography, and a correlation heatmap (Fig. 5) is used to study the relationships among numerical variables. It should be noted that correlation indicates directional association and is not proof of causality.



Fig. 4: EDA and customer analytics

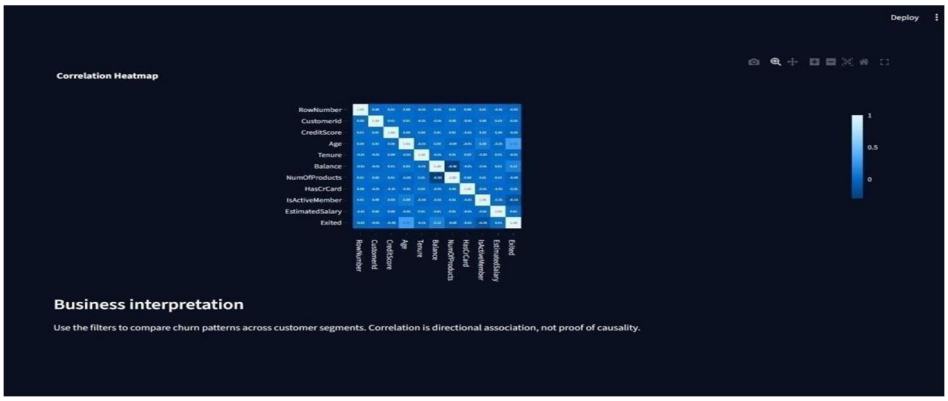


Fig. 5: Correlation heatmap

The individual customer prediction page is shown in Fig. 6. For the demonstrated input (credit score 650, age 39, balance 75000.00, tenure 5, one product, France, male), the estimated churn probability is 12.6% and the customer is classified as “Likely to STAY” with low risk at a decision threshold of 0.57.

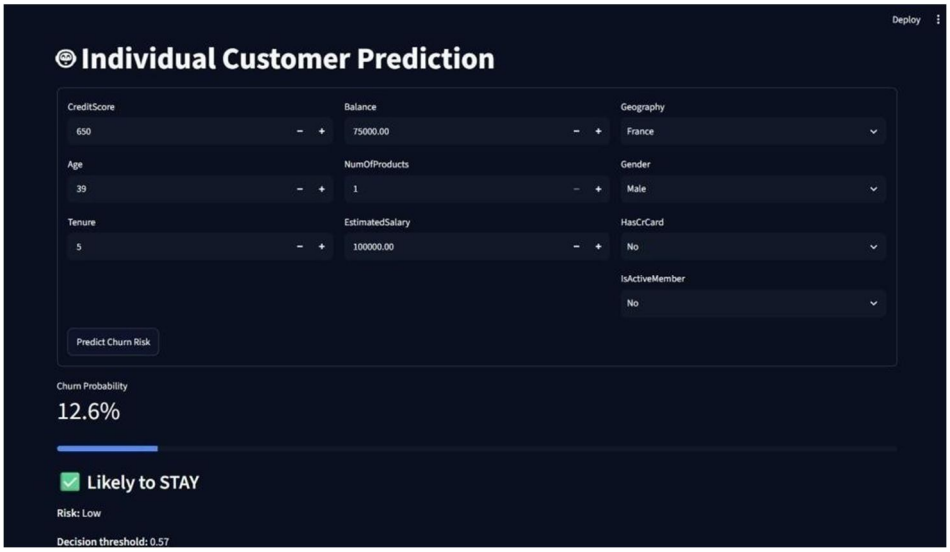


Fig. 6: Individual customer prediction

The distribution of churn probabilities across customers and the segmentation into low, medium and high risk groups are shown in Fig. 7 and Fig. 8. These views help banks to prioritise high-risk customers for retention actions.

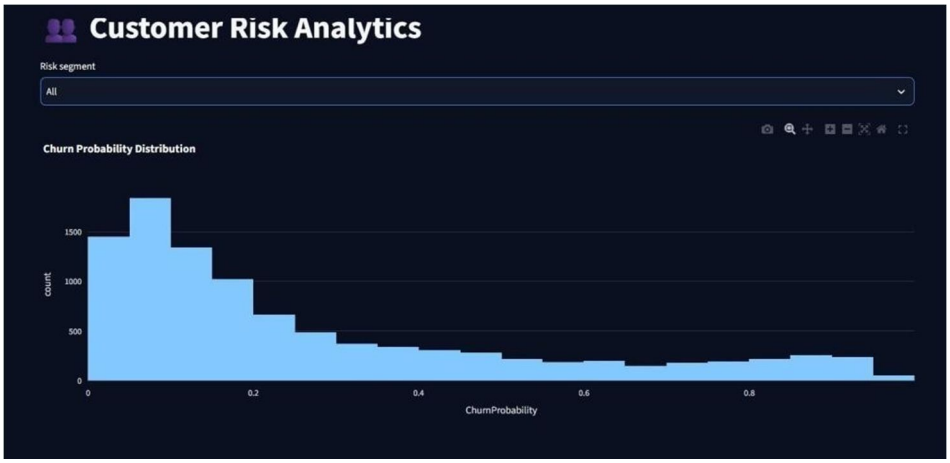


Fig. 7: Customer risk analytics

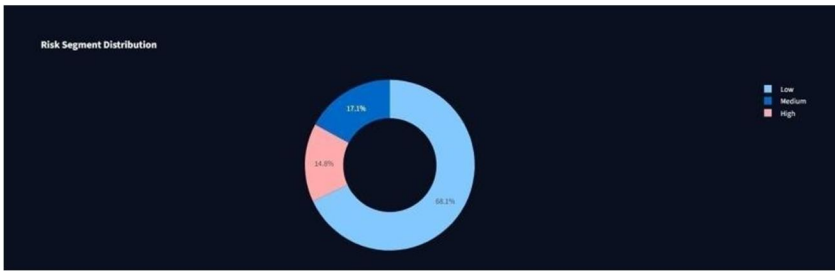


Fig. 8: Risk segment distribution

The business insights page (Fig. 9) shows that the highest churn is observed in Germany (32.4%) and in the middle-aged group (39.7%), and that inactive members churn more (26.9%) than active members (14.3%). It is recommended to prioritise high-probability customers for proactive outreach, to use inactivity and financial value as prioritisation signals rather than standalone causal explanations, and to test retention interventions with controlled experiments.



Fig. 9: Business insights and retention

VI. CONCLUSION AND FUTURE SCOPE

A data-driven system for identifying banking customers who are likely to leave has been developed, covering preprocessing, EDA, model development, model evaluation and customer-level prediction. Among the models compared, XGBoost provided the most suitable performance, with an accuracy of 85.75%, precision of 65.72%, recall of 62.65%, F1-score of 64.15% and ROC-AUC of 87.07%. The Streamlit interface and the risk analytics make the predictions easy to interpret and apply in retention planning.

The system can be enhanced in future by including additional banking features such as transaction frequency, customer service interactions, loan information and digital banking usage. Hyperparameter optimization, ensemble learning, explainable AI and model calibration can further improve prediction and show why a customer is at high risk. A retention recommendation module, real-time data integration, cloud deployment, prediction history storage, dashboards and periodic model retraining can transform the system into a comprehensive customer-risk monitoring platform.

VII. ACKNOWLEDGMENT

Expressing profound gratitude, we wish to acknowledge the invaluable support and guidance we have received during the execution of our mini project at Adhiyamaan College of Engineering.

We show our sincere thanks to Dr. R. RADHAKRISHNAN, M.E., Ph.D., Principal, Adhiyamaan College of Engineering, Hosur, for this valuable opportunity and the environment given to us.

We extend our hearty gratitude to Dr. G. FATHIMA, M.E., Ph.D., Professor and Head of the Department, Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, for her guidance, valuable suggestions and encouragement throughout this project.

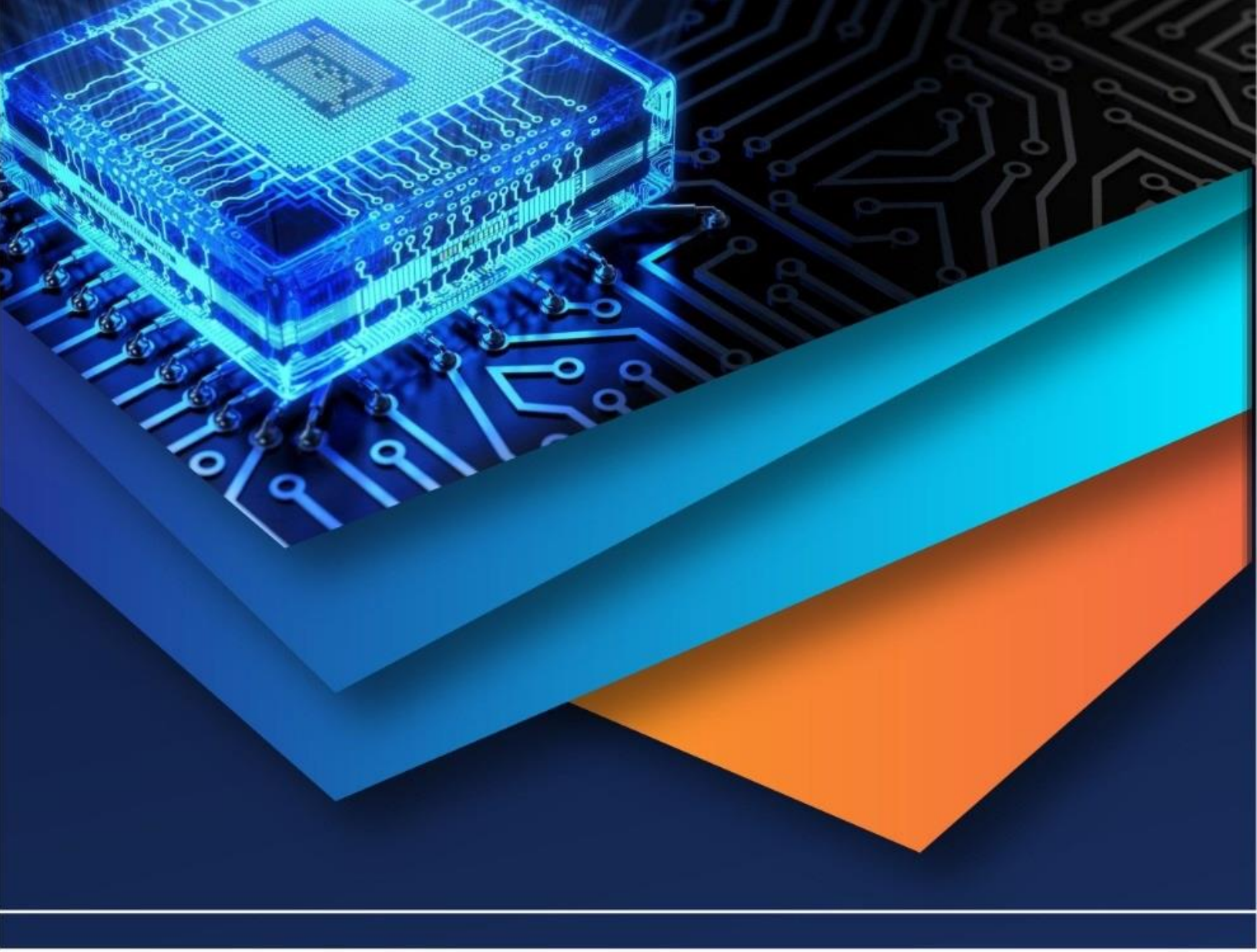
We are highly indebted to our Supervisor Mrs. V. MEENA, M.E., Assistant Professor, Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, whose immense support, encouragement and valuable guidance were responsible for completing the project successfully. We extend our hearty thankfulness to our project coordinator and all our department faculty for their support in helping us complete this project successfully.

This project would not have been possible without the collective support and guidance of these individuals and the Adhiyamaan College of Engineering community. Thank you for making this endeavour a success.

REFERENCES

[1] Waningsih et al., "Classification-Based Supervised Learning Algorithms for Accurate Prediction of Customer Churn in Banking," International Journal of Applied Technology and Information Systems (IJATIS), vol. 3, no. 1, pp. 41–49, 2026.
[2] Z. Tan, "Comparative Analysis of Several Models for Churning Customer Prediction," SHS Web of Conferences, vol. 218, Art. no. 02013, 2025.
[3] V. Gkonis and I. Tsakalos, "Deep Dive Into Churn Prediction in the Banking Sector: The Challenge of Hyperparameter Selection and Imbalanced Learning," Journal of Forecasting, vol. 44, no. 2, pp. 281–296, 2025.
[4] R. Bhuria, S. Gupta, U. Kaur, S. Bharany, et al., "Ensemble-Based Customer Churn Prediction in Banking: A Voting Classifier Approach for Improved Client Retention Using Demographic and Behavioral Data," Discover Sustainability, vol. 6, no. 1, Art. no. 28, 2025.
[5] H. Zhu, "Bank Customer Churn Prediction with Machine Learning Methods," Advances in Economics, Management and Political Sciences, vol. 69, pp. 23–29, 2024.
[6] K. Peng, Y. Peng, and W. Li, "Research on Customer Churn Prediction and Model Interpretability Analysis," PLOS ONE, vol. 18, no. 12, pp. 1–26, 2023.
[7] H. S. Hon, K. W. Khaw, X. Chew, and W. P. Wong, "Prediction of Customer Churn for ABC Multistate Bank Using Machine Learning Algorithms," Malaysian Journal of Computing, vol. 8, no. 2, pp. 1602–1619, 2023.
[8] H. D. Tran, N. Le, and V.-H. Nguyen, "Customer Churn Prediction in the Banking Sector Using Machine Learning-Based Classification Models," Interdisciplinary Journal of Information, Knowledge, and Management, vol. 18, pp. 87–105, 2023.
[9] A. Muneer, R. F. Ali, A. Alghamdi, S. M. Taib, A. Almaghthawi, and E. A. A. Ghaleb, "Predicting Customers Churning in Banking Industry: A Machine Learning Approach," Indonesian Journal of Electrical Engineering and Computer Science, vol. 26, no. 1, pp. 539–549, 2022.
[10] K. A. Amuda and A. B. Adeyemo, "Customers Churn Prediction in Financial Institution Using Artificial Neural Network," arXiv preprint, arXiv:1912.11346, 2019.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)