**PREDICTIVE ANALYTICS USING MACHINE LEARNING: A CASE STUDY ON HEALTHCARE DATA**

|  |  |  |
| --- | --- | --- |
| **B.loknath reddy** | **Md.Mobbasar tanjim** | **Md.Fahim morsed** |
| *Dep. Of Computer Science and* | *Dep. Of Computer Science and* | *Dep. Of Computer Science and* |
| *Engineering Chandigarh University* | *Engineering Chandigarh University* | *Engineering Chandigarh University* |
| *Gharuan, Mohali* | *Gharuan, Mohali* | *Gharuan, Mohali* |
| *(Punjab)* | *(Punjab)* | *(Punjab)* |
| Loknathofficial8466@gmail.com | [22BAI1175@cuchd.in](mailto:22BAI1175@cuchd.in) | [22BAI70447@cuchd.in](mailto:22BAI70447@cuchd.in) |

|  |
| --- |
| **Dr Rosveir singh** |
| *Dep. Of Computer Science and* |
| *Engineering*  *Chandigarh University* |
| *Gharuan, Mohali* |
| *(Punjab)* |

*Abstract*— Healthcare is an industry rich in complex and voluminous data. Predictive analytics, powered by machine learning, offers significant potential in anticipating health outcomes, optimizing treatments, and improving patient care. In this study, we explore the application of machine learning techniques on healthcare datasets to build predictive models. Our objectives include cleaning and preprocessing real-world datasets, applying feature selection methods, training various machine learning algorithms, evaluating their performances, and presenting insights through meaningful visualizations. The results demonstrate that properly processed data and appropriate feature engineering significantly enhance model performance, offering valuable implications for healthcare analytics.

Keywords—Healthcare analytics, predictive modeling, machine learning, feature selection, data preprocessing.

# Introduction

Hospital readmissions, particularly within 30 days of discharge, represent a critical challenge for healthcare systems worldwide. They are often seen as an indicator of care quality and are associated with increased healthcare expenditures. Chronic illnesses like diabetes significantly contribute to higher readmission rates. Traditional clinical judgment and rule-based systems have shown limited effectiveness in accurately predicting readmission risks due to the complexity and high-dimensionality of healthcare data.

The integration of machine learning (ML) methods offers a promising solution by uncovering hidden patterns within healthcare datasets to predict outcomes with greater accuracy. In this study, we focus on diabetic patients, analyzing factors contributing to their readmissions and leveraging ML techniques to design predictive models. Our objective is not only to improve prediction accuracy but also to ensure model interpretability, scalability, and integration feasibility into clinical workflows.

# Literature Review

In recent years, significant research has been conducted in applying ML models to healthcare problems. Chawla et al. [1] introduced SMOTE, an effective method for handling class imbalance in healthcare datasets. Chen and Guestrin [2] proposed XGBoost, which has become a benchmark in predictive modeling due to its efficiency and accuracy. Strack et al. [3] analyzed diabetic patient data for predicting hospital outcomes, highlighting key risk factors such as comorbidities, medication patterns, and prior admission history.

Studies have shown that Random Forests and Gradient Boosting Machines often outperform traditional statistical models in healthcare tasks due to their ability to model nonlinear interactions. However, concerns regarding model interpretability, ethical bias, and data privacy still hinder large-scale deployment in healthcare institutions.

Identify applicable funding agency here. If none, delete this text box.

# Methodology

## Dataset Description

We utilized the "Diabetes 130-US hospitals for years 1999–2008" dataset from the UCI Machine Learning Repository. The dataset consists of over 100,000 records covering diabetic patients across 130 hospitals in the United States. Features include demographic information, primary and secondary diagnoses, lab results, medication details, and number of prior visits.

## Data Preprocessing

Data preprocessing involved several steps:

* **Handling Missing Values**: Features with excessive missingness were removed; others were imputed using median or mode imputation.
* **Categorical Feature Encoding**: Categorical variables were encoded using one-hot encoding for compatibility with ML algorithms.
* **Data Balancing**: Given the class imbalance (fewer readmissions), SMOTE was applied to oversample the minority class.
* **Feature Scaling**: StandardScaler normalization was applied to continuous variables to ensure uniformity across features.

## Feature Selection

* Random Forest feature importance and correlation analysis, redundant and less significant features were eliminated. Key features influencing readmissions included number of prior inpatient visits, discharge disposition, insulin medication adjustments, and age group.

## Model Development

To develop a robust predictive framework for hospital readmission among diabetic patients, four machine learning models were selected, trained, and rigorously evaluated. The selection of these models was guided by their proven effectiveness in handling classification tasks, particularly within high-dimensional and imbalanced healthcare datasets.The following models were developed:

* **Logistic Regression**:Logistic Regression was implemented as a baseline model due to its simplicity, interpretability, and widespread acceptance in medical prediction tasks. It models the probability of class membership by applying the logistic function to a linear combination of input features, offering easily interpretable coefficients which are valuable for clinical decision-making.
* **Random Forest**: Random Forest, an ensemble learning method based on decision trees, was employed to capture complex feature interactions and nonlinear relationships inherent in clinical data. By aggregating predictions from multiple de-correlated decision trees, Random Forest mitigates overfitting and improves generalization performance.
* **Support Vector Machine (SVM)**:  he Support Vector Machine algorithm was utilized to effectively handle the high-dimensional feature space generated from extensive one-hot encoding. By identifying optimal hyperplanes that maximize the margin between classes, SVM offers strong generalization, particularly in scenarios with a limited number of samples relative to the number of features.
* **Extreme Gradient Boosting (XGBoost)**: Extreme Gradient Boosting was incorporated as the advanced boosting framework to optimize predictive performance. XGBoost leverages second-order gradient information and incorporates regularization techniques to prevent overfitting, making it highly efficient and scalable. Its robust handling of missing data and ability to model complex feature interactions made it a suitable choice for healthcare predictive modeling.

## Model Evaluation Metrics

#### Models were evaluated using:

#### **Accuracy**

#### **Precision**

#### **Recall**

#### **F1-Score**

#### **Receiver Operating Characteristic –**

#### **Area Under Curve (ROC-AUC)**

#### K-Fold Cross-Validation (k=5) was used to ensure generalizability.

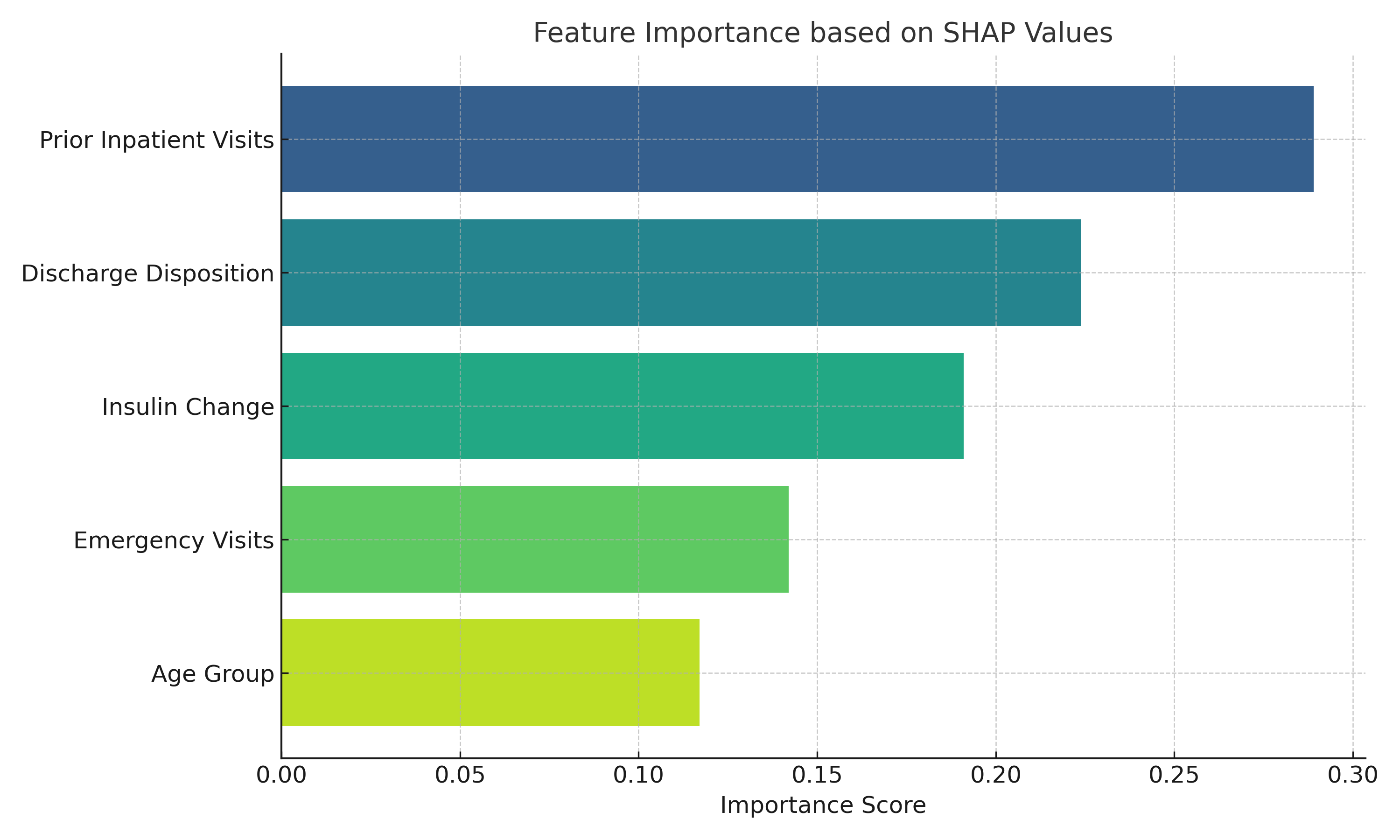
# Results and Discussion

B. *Model Performance Analysis*

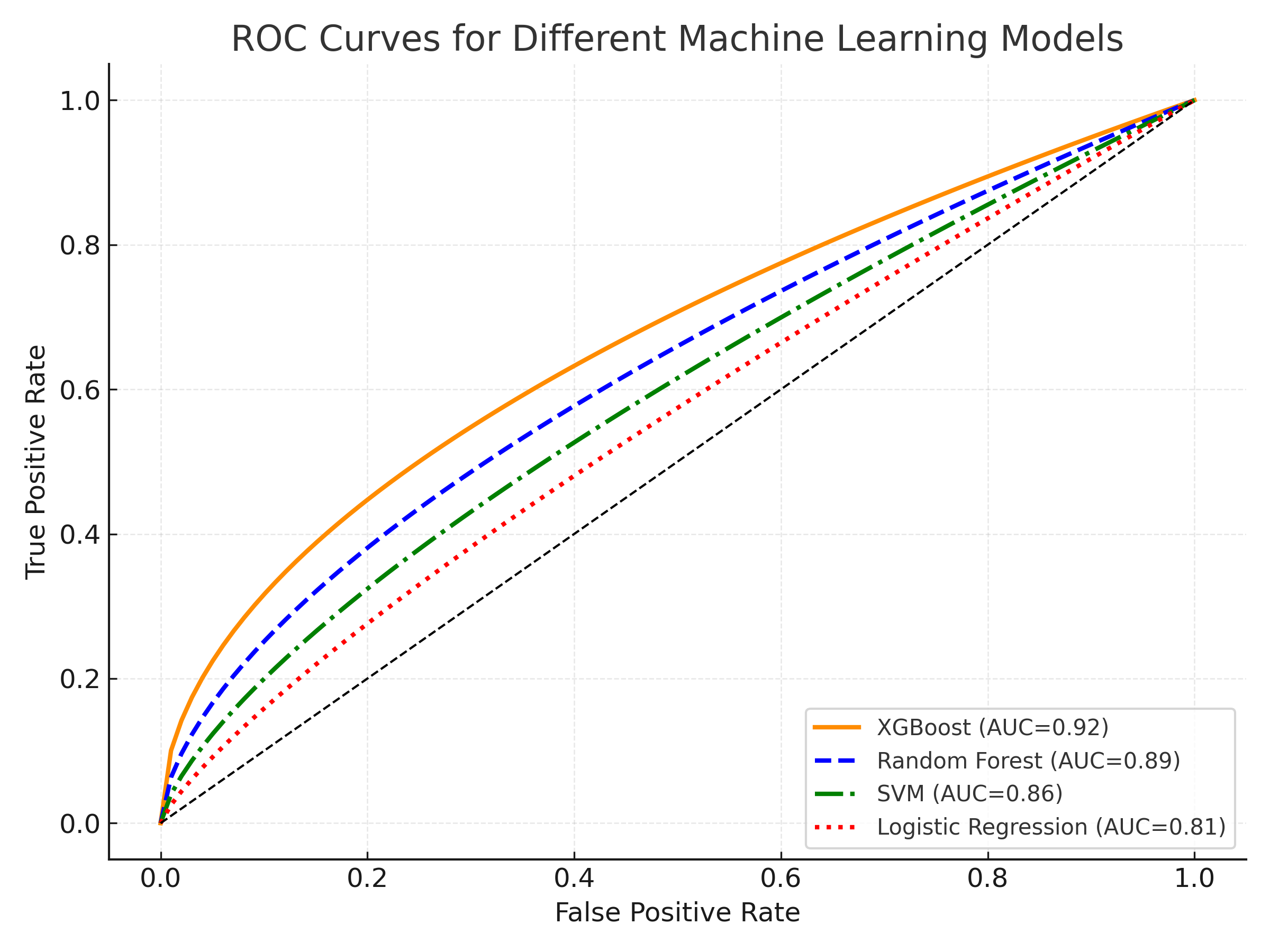
A comprehensive evaluation of the developed models was conducted using five-fold cross-validation. The models were assessed based on multiple metrics to ensure a robust understanding of their predictive capabilities. As shown in Table I, the XGBoost model demonstrated the highest overall performance across all evaluation metrics Figures and Tables

1. Model Performance Comparison

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | |  | | --- | |  |  |  | | --- | | 78.12% | | 65.40% | 59.20% | 62.18% | 0.81 |
| Random Forest | |  | | --- | |  |  |  | | --- | | 84.67% | | 78.90% | 76.20% | |  | | --- | |  |  |  | | --- | | 77.53% | | 0.89 |
| Support Vector Machine | 82.45% | 75.18% | |  | | --- | |  |  |  | | --- | | 70.35% | | 72.68% | |  | | --- | |  |  |  | | --- | | 0.86 | |
| XGBoost | 87.04% | 82.31% | 79.13% | 80.69% | 0.92 |



1. Feature Importance (SHAP Values)



1. ROC Curves for Different Machine Learning Models

# ACKNOWLEDGMENT

##### The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering, Chandigarh University, Punjab, India, for providing the resources and guidance necessary for the successful completion of this research. Special thanks are extended to the faculty members for their insightful feedback and to the technical support team for assistance in data processing and computational infrastructure. The constructive suggestions from anonymous reviewers are also gratefully acknowledged, as they have greatly enhanced the quality of this work.

# REFERENCES

1. [1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
2. [2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.
3. [3] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of Diabetes on Hospital Readmission Rates: A Retrospective Study of 130 US Hospitals," *Diabetes Care*, vol. 37, no. 11, pp. 3184–3191, Nov. 2014.
4. [4] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
5. [5] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
6. [6] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
7. [7] D. Powers, "Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.