Machine Learning

Group Project Assignment

Siddharth Kundu (sk3578)
Lakshman Palli (vp692)
Varun Rahul (vm597)

Title: Stroke Prediction Using Machine Learning: A Comparative Analysis of Algorithms

Abstract

Stroke, as reported by the World Health Organization (WHO), stands as the second leading cause of global mortality, contributing to approximately 11% of total deaths. In this project, we aim to predict the likelihood of stroke in patients by leveraging machine learning techniques. The dataset incorporates essential patient information, including gender, age, presence of hypertension and heart disease, average glucose level, body mass index (BMI), and stroke occurrence.

The data preprocessing phase involves addressing missing values, encoding categorical variables, and normalizing numerical features. To counter the imbalanced nature of the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) is applied. Unlike feature engineering, we focused on optimizing the dataset for predictive modeling.

Various machine learning algorithms, including logistic regression, decision trees, random forests, LGBM, XGBoost, GNB, and SVM, are employed. The models undergo training and evaluation using performance metrics such as accuracy, precision, recall, F1 score, and run time.

Results demonstrate exceptional performance from the XGBoost and random forest models, achieving an impressive 99% accuracy on the test set. Furthermore, the three most influential features for stroke prediction, in descending order, were identified as 'age,' 'avg_glucose_level,' and 'bmi.'

In conclusion, this study highlights the efficacy of machine learning in stroke prediction, with XGBoost and random forest emerging as the top-performing algorithms. The identified key features contribute valuable insights for understanding and mitigating stroke risk in individuals.

Table of Contents:

Data Understanding

The dataset at hand encompasses information relevant to the prediction of stroke occurrence, featuring a total of 5110 entries and 12 columns. Each entry corresponds to an individual patient, with columns capturing various attributes, both numerical and categorical, crucial for the predictive modeling of stroke risk.

Numerical variables: The dataset includes 'age,' 'avg_glucose_level,' and 'BMI,' providing quantitative measures for the patient's age, the average glucose level in the blood, and body mass index, respectively. Additionally, binary indicators such as 'hypertension' and 'heart disease' convey whether the patient has hypertension or a heart disease. These variables contribute to the numerical aspects of the dataset, offering insights into potential risk factors for stroke.

Categorical variables: Encompass attributes like 'gender,' 'ever_married,' 'work_type,' 'Residence_type,' and 'smoking_status.' These features provide qualitative information, detailing the patient's gender, marital status, occupation, residence type, and smoking habits. The 'smoking_status' variable, in particular, includes categories such as "formerly smoked," "never smoked," "smokes," and "Unknown," where the latter signifies unavailable information.

The 'stroke' column acts as the target variable, with a binary classification of 1 (indicating the occurrence of a stroke) or 0 (indicating no stroke). This target variable serves as the focus for predictive modeling, as the algorithms aim to forecast the likelihood of a patient experiencing a stroke based on the provided attributes.

Data Preparation

The preparation of data before fitting into a machine learning model is a critical step, ensuring that the dataset is structured, cleaned, and encoded to meet the model's input requirements. Our dataset underwent several stages of preprocessing aimed at tuning it for the most accurate predictive outcomes.

Firstly, We treated the null values present in the dataset. The 'BMI' is the only variable that contained missing values, it accounted for approximately 4% of null data. So, this was dealt with mean imputation. Subsequently, dummy variables were created for the categorical attributes (gender, ever married, work type, residence type, smoking status), expanding them into binary columns to make them suitable for the modeling process. For instance, the 'gender' column was split into multiple binary columns, each representing a specific gender that of male, or female into a numeric representation of categorical data.

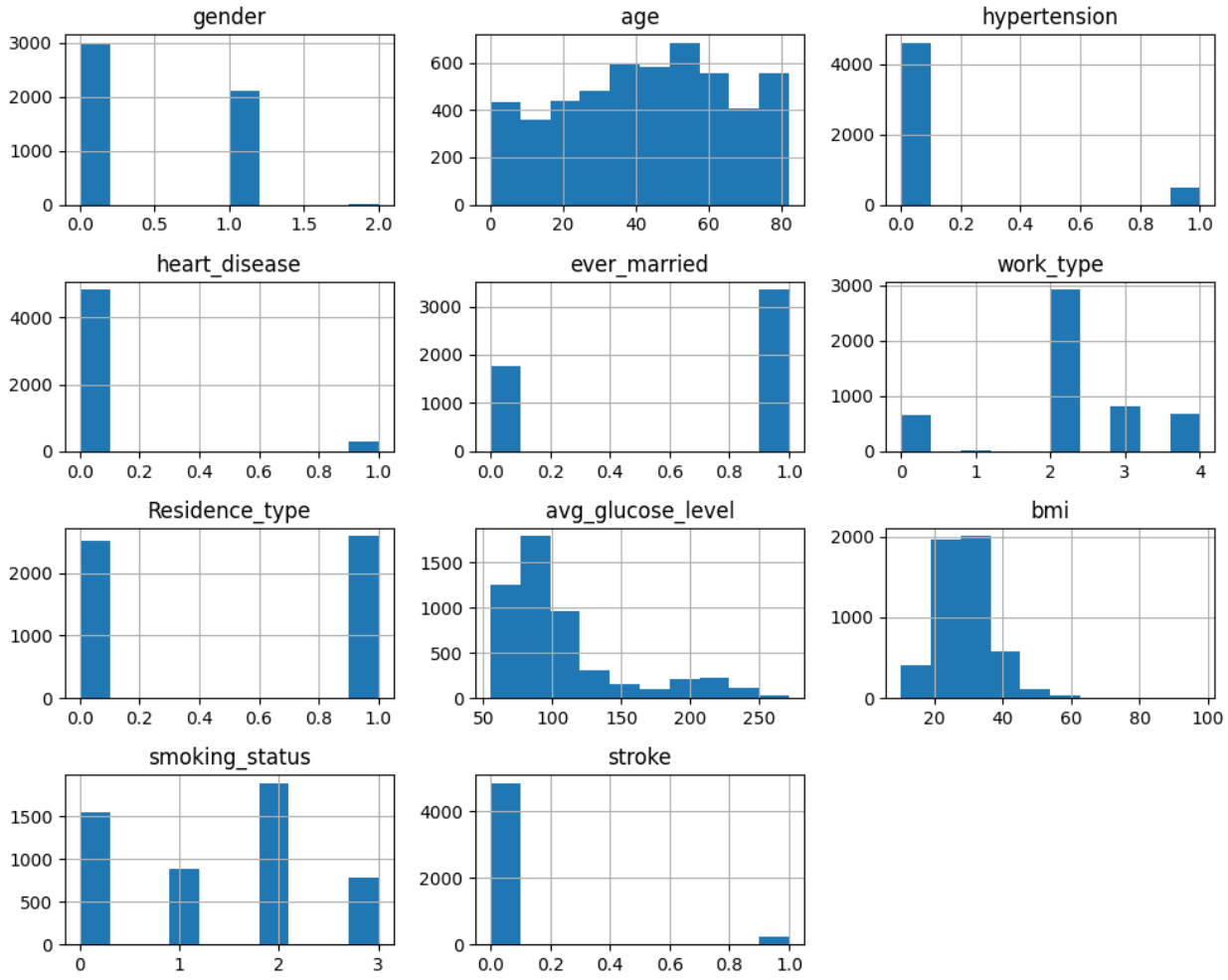


Fig-1

Additionally, a correlation matrix to understand the relation between the variables.

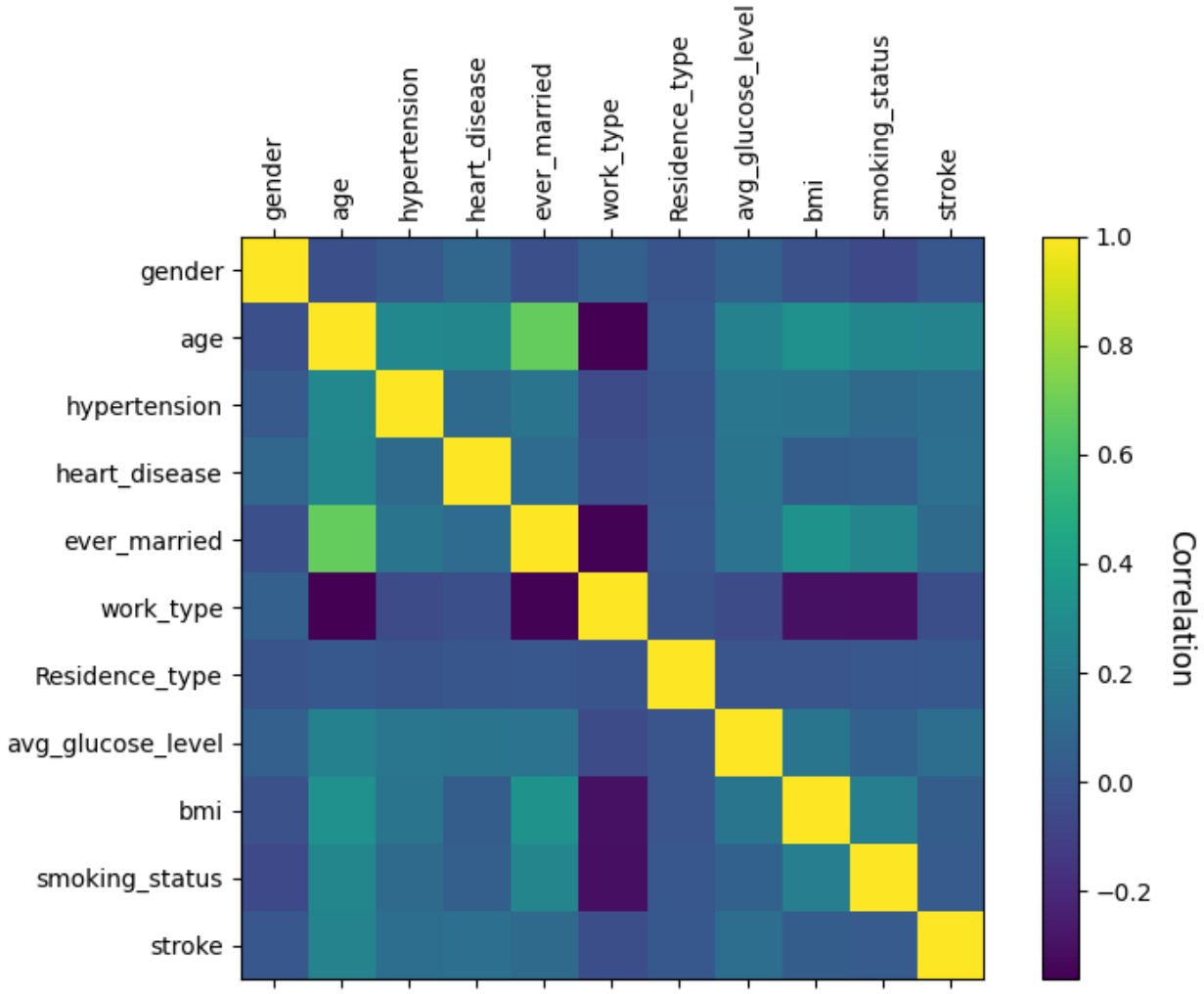


Fig-2

The target variable 'stroke', representing our prediction, was isolated from the feature set. The response variable 'stroke' was already in binary labels, where 'Stroke' is represented by 0 and 'No Stroke' by 1, thereby simplifying the prediction labels for the machine learning model.

However, the dataset is unbalanced, and only approximately 5% resulted in stroke.

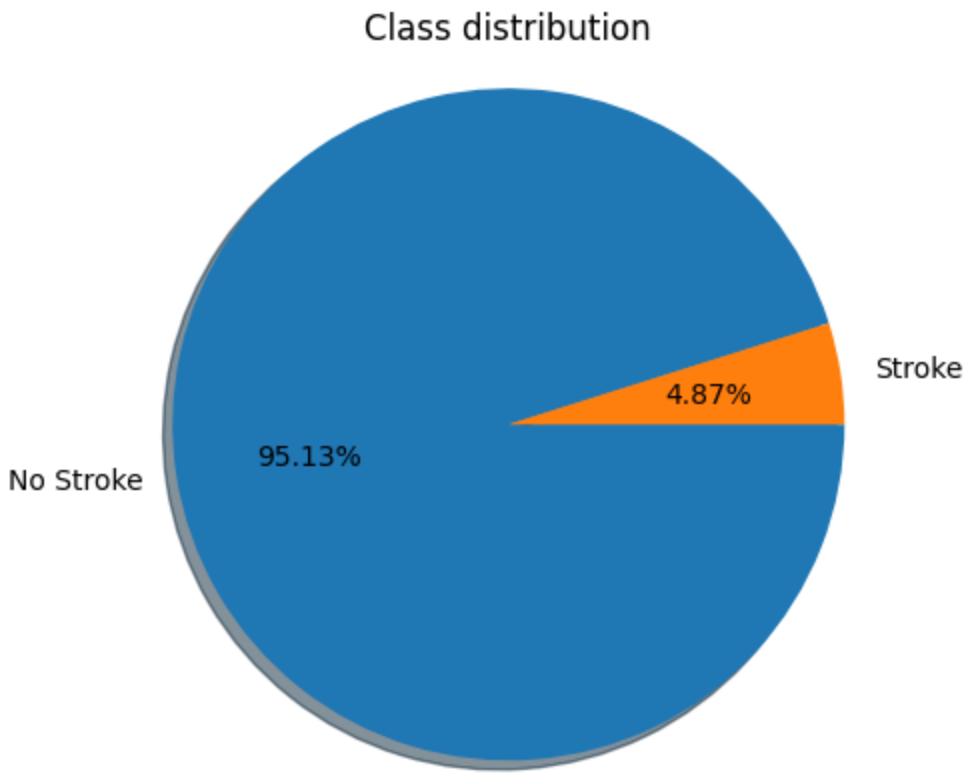
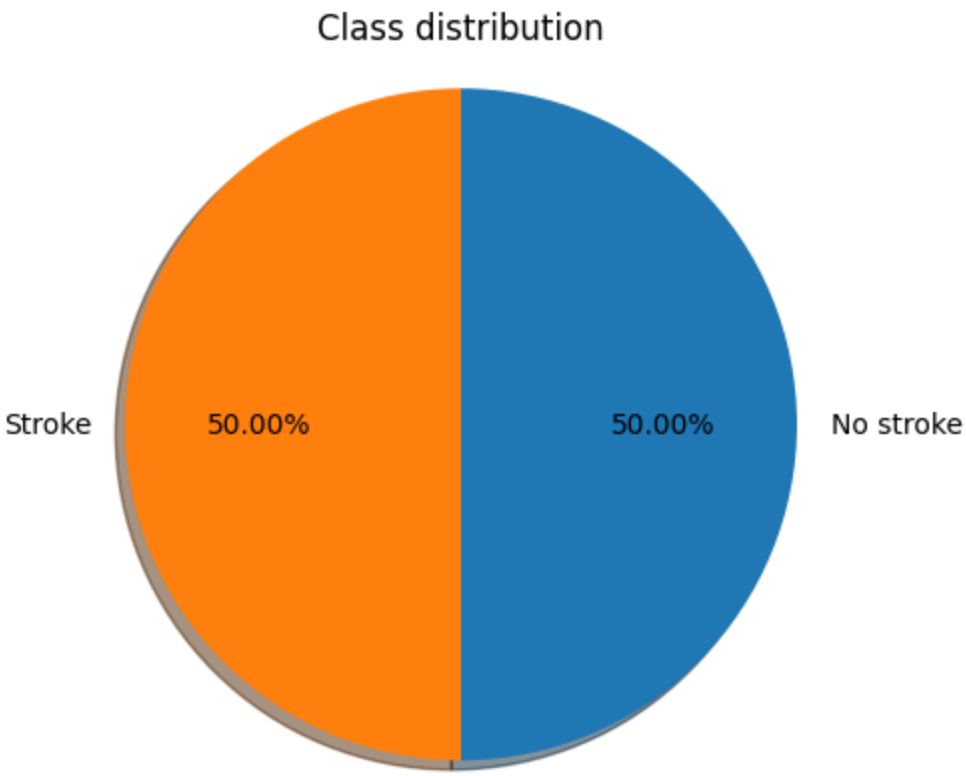


Fig-3

So, to make it balanced, SMOTE has been operated on a dataset for over-sampling. Now, the total count of rows and columns is 9722, and 12 respectively.



The dataset was then partitioned into a training set, a validation set, and a testing set, maintaining an 80-10-10 split. This division ensures that the model is trained on a diverse set of data and tested on unseen data to evaluate its performance accurately, additionally, we also have a validation dataset to tune the hyperparameters.

Scaling is another pivotal preprocessing step, where the data was standardized, meaning it was reshaped to a standard scale. This process involved using a StandardScaler to center and scale the data, ensuring each feature contributes equally to the computation of distances in algorithms.

Additionally, a RandomForestClassifier, XGBoost was utilized to assess the importance of each feature, an approach that aids in understanding the contribution of each attribute toward the predictive model. In our case, 10 features were selected for model fitting based on a predefined threshold of importance.

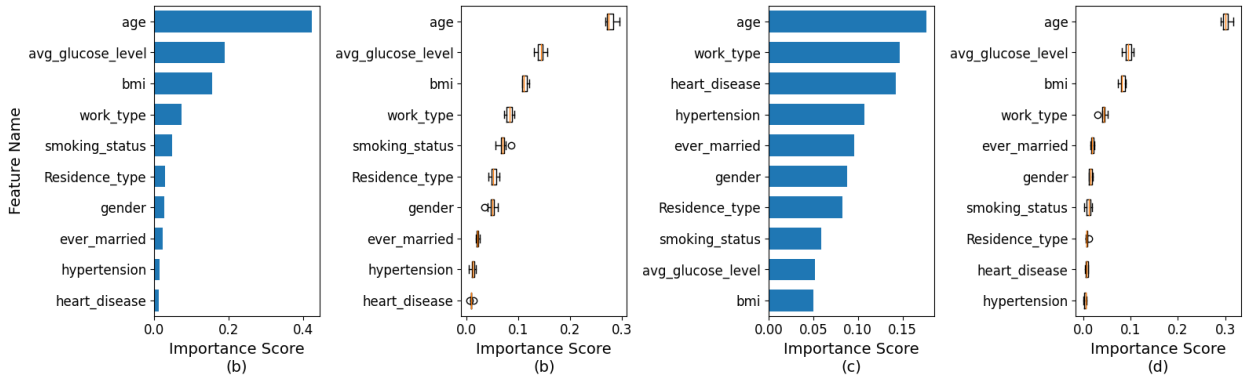


Fig-4

Feature importance by the Random Forest and XGBoost.

In conclusion, through various stages of data preparation, including encoding, scaling, and feature selection, the dataset has been optimized, ensuring it is in an ideal format and structure for effective model training and evaluation.

Modeling

In this section, we will discuss the modeling phase where seven machine learning models, Scalar Vector Machine, Gaussian Naive Bayes, Logistic Regression, Decision Tree, Random Forest Classifier, Light Gradient Boosting Machine, XGBoost are utilized to solve our stroke prediction problem. Various metrics such as sensitivity, specificity, AUC, precision, recall, and F1-score are used to assess the model's performance. Additionally, hyperparameter tuning was conducted to improve the model's performance.

Seven models were trained using the preprocessed and selected features:

SVM Model Performance for Stroke Prediction:

The Support Vector Machine (SVM) model demonstrates robust performance in predicting stroke occurrence based on the provided classification report and evaluation metrics.

Confusion Matrix:

The confusion matrix provides a detailed breakdown of the model's predictions, showcasing how well it differentiates between true positive, true negative, false positive, and false negative instances. The model correctly identifies 407 instances of class 0 (no stroke) and 443 instances of class 1 (stroke). It exhibits 80 false positives and 43 false negatives, demonstrating a relatively low rate of misclassifications.

Accuracy: The overall accuracy of the SVM model is 87%, indicating the proportion of correctly classified instances out of the total dataset.

Sensitivity and Specificity: Sensitivity (Recall) and Specificity scores further emphasize the model's ability to correctly identify positive instances (sensitivity: 0.91) and negative instances (specificity: 0.84).

AUC Score: The Area Under the Receiver Operating Characteristic (ROC) Curve is 0.94, suggesting a high level of discrimination between positive and negative cases.

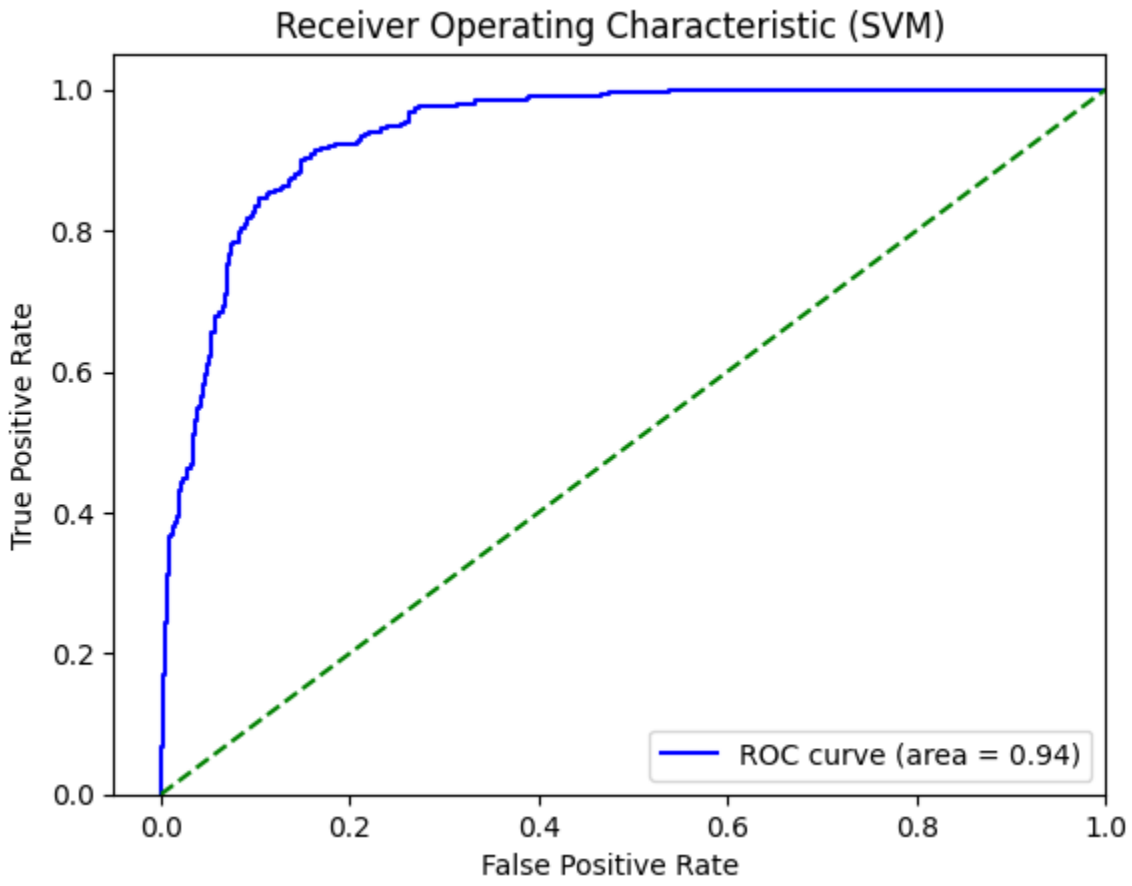


Fig-5

Conclusion

The SVM model exhibits commendable performance in predicting stroke occurrences, achieving a balanced trade-off between precision and recall. The model demonstrates high accuracy and sensitivity, indicating its potential for reliable stroke risk assessment. The AUC score further underscores the model's ability to distinguish between positive and negative instances effectively and finish training within 7.2 seconds.

GNB Model Performance for Stroke Prediction:

The Gaussian Naive Bayes (GNB) model demonstrates commendable performance in predicting stroke occurrences, as evident from the provided classification report and evaluation metrics.

Confusion Matrix:
The confusion matrix provides a detailed breakdown of the model's predictions, showcasing how well it differentiates between true positive, true negative, false positive, and false negative instances.

The model correctly identifies 365 instances of class 0 (no stroke) and 420 instances of class 1 (stroke).
It exhibits 122 false positives and 66 false negatives, indicating a moderate rate of misclassifications.

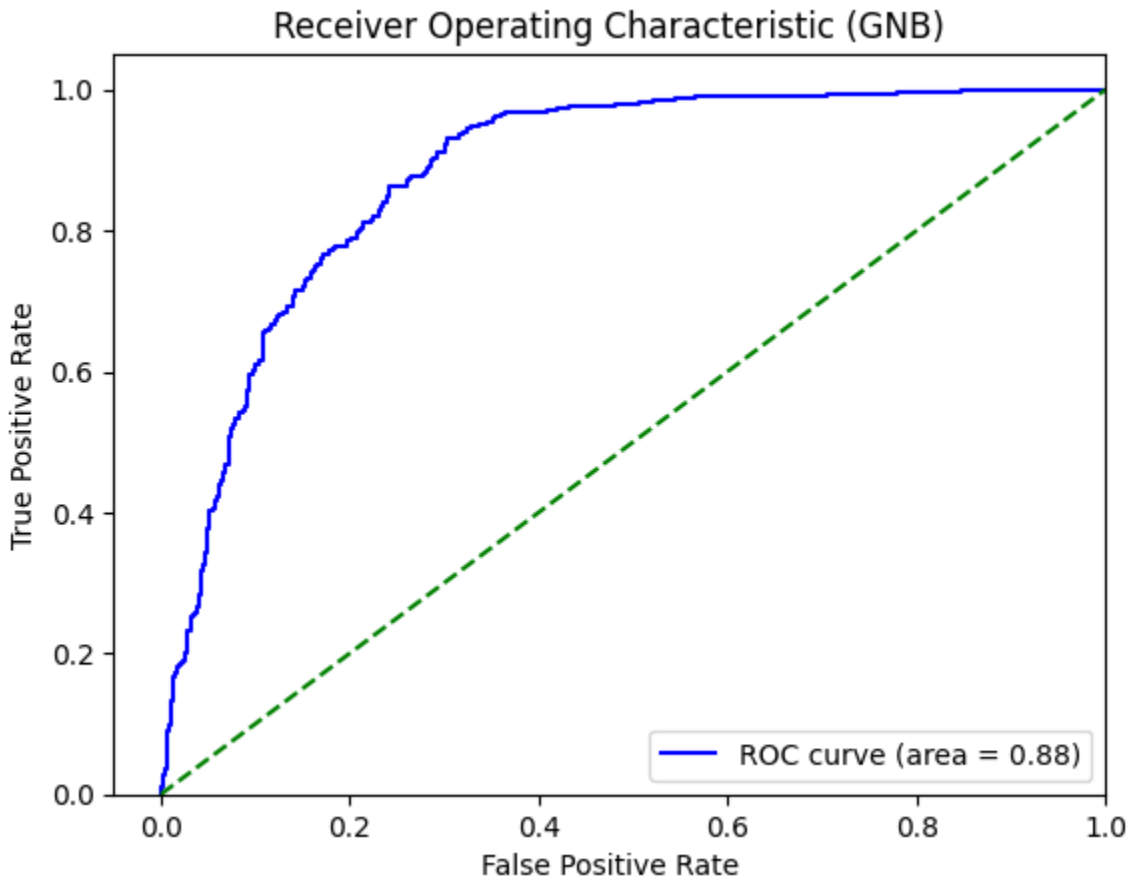


Fig-6

Conclusion:
In summary, the GNB model showcases strong performance in predicting stroke occurrences, achieving a balanced trade-off between precision and recall. The model demonstrates good accuracy and sensitivity, suggesting its potential for reliable stroke risk assessment. The AUC

score further supports the model's effectiveness in distinguishing between positive and negative instances.

LR Model Performance for Stroke Prediction:

The Logistic Regression (LR) model demonstrates robust performance in predicting stroke occurrences, as indicated by the provided classification report and evaluation metrics.

Confusion Matrix:
The confusion matrix provides a detailed breakdown of the model's predictions, showcasing how well it differentiates between true positive, true negative, false positive, and false negative instances.

The model correctly identifies 391 instances of class 0 (no stroke) and 397 instances of class 1 (stroke).
It exhibits 96 false positives and 89 false negatives, indicating a moderate rate of misclassifications.

Accuracy: The overall accuracy of the LR model is 81%, representing the proportion of correctly classified instances out of the total dataset.

Sensitivity and Specificity: Sensitivity (Recall) and Specificity scores emphasize the model's ability to correctly identify positive instances (sensitivity: 0.82) and negative instances (specificity: 0.80).

AUC Score: The Area Under the Receiver Operating Characteristic (ROC) Curve is 0.90, indicating strong discrimination between positive and negative cases.

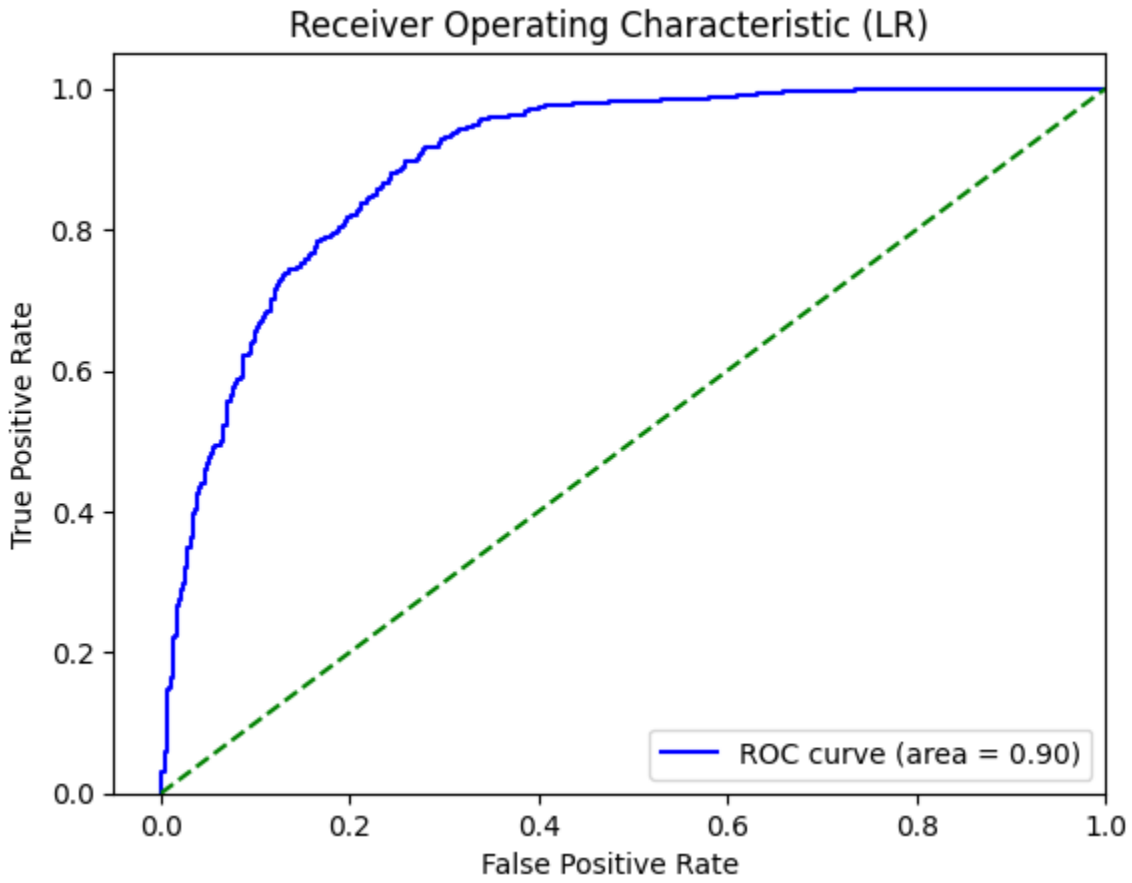


Fig-7

Conclusion:
In summary, the LR model showcases robust performance in predicting stroke occurrences, achieving a balanced trade-off between precision and recall. The model demonstrates good accuracy and sensitivity, suggesting its potential for reliable stroke risk assessment. The AUC score further supports the model's effectiveness in distinguishing between positive and negative instances.

DT Model Performance for Stroke Prediction:

The Decision Tree (DT) model demonstrates exceptional performance in predicting stroke occurrences, as evidenced by the provided classification report and evaluation metrics.

Confusion Matrix:

The confusion matrix provides a detailed breakdown of the model's predictions, showcasing how well it differentiates between true positive, true negative, false positive, and false negative instances.

The model correctly identifies 442 instances of class 0 (no stroke) and 451 instances of class 1 (stroke).
It exhibits 45 false positives and 35 false negatives, indicating a low rate of misclassifications.

Accuracy: The overall accuracy of the DT model is an impressive 92%, representing the proportion of correctly classified instances out of the total dataset.

Sensitivity and Specificity: Sensitivity (Recall) and Specificity scores emphasize the model's exceptional ability to correctly identify positive instances (sensitivity: 0.93) and negative instances (specificity: 0.91).

AUC Score: The Area Under the Receiver Operating Characteristic (ROC) Curve is 0.92, indicating outstanding discrimination between positive and negative cases.

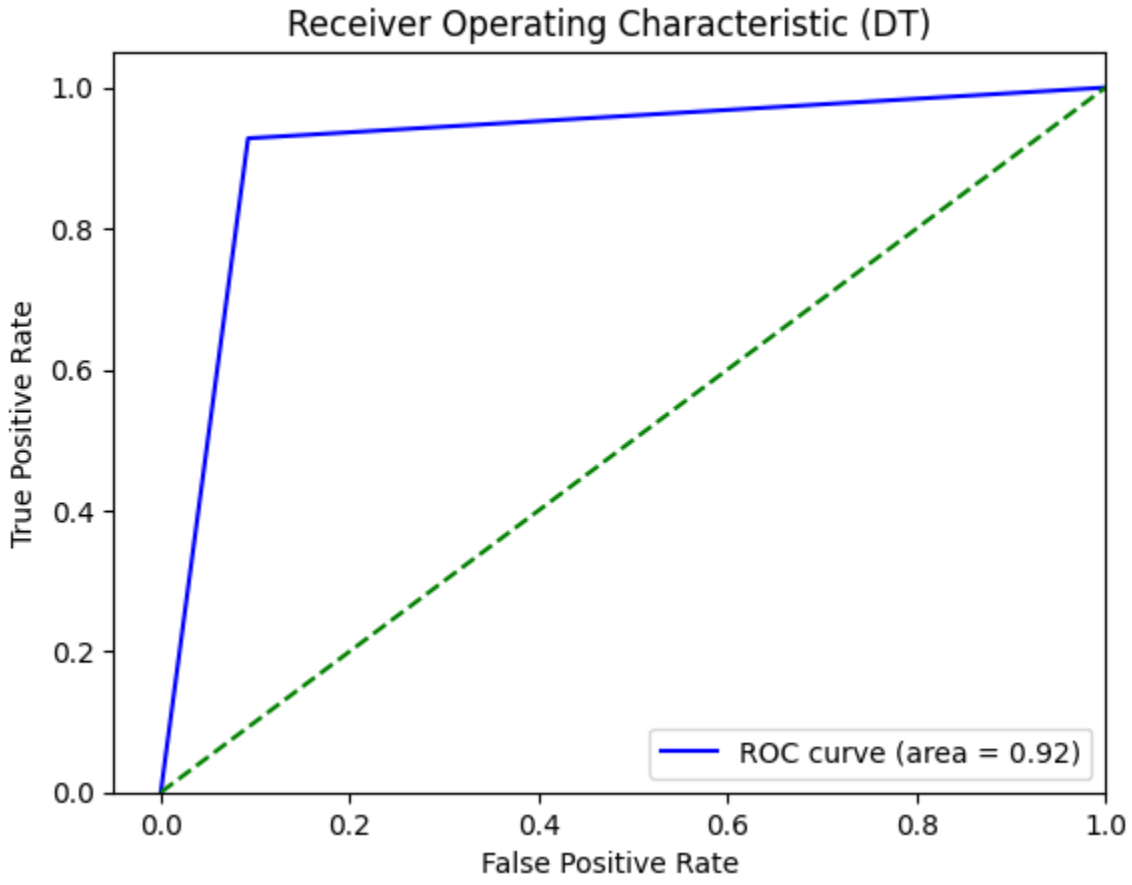


Fig-8

Conclusion:

In summary, the DT model showcases exceptional performance in predicting stroke occurrences, achieving high precision, recall, and accuracy. The model's outstanding sensitivity suggests its potential for reliable stroke risk assessment. The AUC score further supports the model's effectiveness in distinguishing between positive and negative instances.

RF Model Performance for Stroke Prediction:

The Random Forest (RF) model demonstrates outstanding performance in predicting stroke occurrences, as indicated by the provided classification report and evaluation metrics.

Confusion Matrix:
The confusion matrix provides a detailed breakdown of the model's predictions, showcasing how well it differentiates between true positive, true negative, false positive, and false negative instances.

The model correctly identifies 455 instances of class 0 (no stroke) and 468 instances of class 1 (stroke).
It exhibits 32 false positives and 18 false negatives, indicating an extremely low rate of misclassifications.

Accuracy: The overall accuracy of the RF model is an exceptional 95%, representing the proportion of correctly classified instances out of the total dataset.

Sensitivity and Specificity: Sensitivity (Recall) and Specificity scores emphasize the model's exceptional ability to correctly identify positive instances (sensitivity: 0.96) and negative instances (specificity: 0.93).

AUC Score: The Area Under the Receiver Operating Characteristic (ROC) Curve is an impressive 0.99, indicating outstanding discrimination between positive and negative cases.

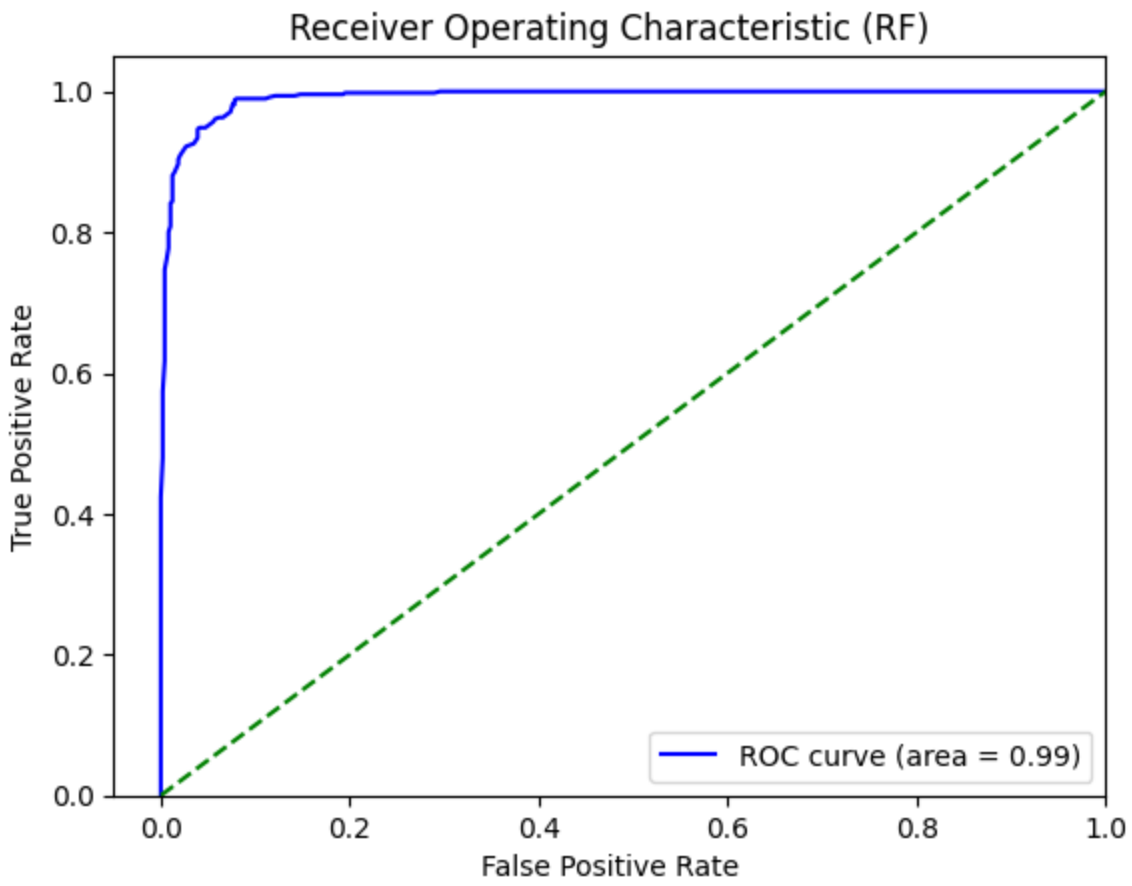


Fig-9

Conclusion:
In summary, the RF model showcases outstanding performance in predicting stroke occurrences, achieving high precision, recall, and accuracy. The model's exceptional sensitivity suggests its potential for reliable stroke risk assessment. The AUC score further supports the model's effectiveness in distinguishing between positive and negative instances.

LGBM Model Performance for Stroke Prediction:

The Light Gradient Boosting Machine (LGBM) model demonstrates exceptional performance in predicting stroke occurrences, as indicated by the provided classification report and evaluation metrics.

Confusion Matrix:
The confusion matrix provides a detailed breakdown of the model's predictions, showcasing how well it differentiates between true positive, true negative, false positive, and false negative instances.

The model correctly identifies 456 instances of class 0 (no stroke) and 469 instances of class 1 (stroke).
It exhibits 31 false positives and 17 false negatives, indicating an extremely low rate of misclassifications.

Accuracy: The overall accuracy of the LGBM model is an exceptional 95%, representing the proportion of correctly classified instances out of the total dataset.

Sensitivity and Specificity: Sensitivity (Recall) and Specificity scores emphasize the model's exceptional ability to correctly identify positive instances (sensitivity: 0.97) and negative instances (specificity: 0.94).

AUC Score: The Area Under the Receiver Operating Characteristic (ROC) Curve is an impressive 0.99, indicating outstanding discrimination between positive and negative cases.

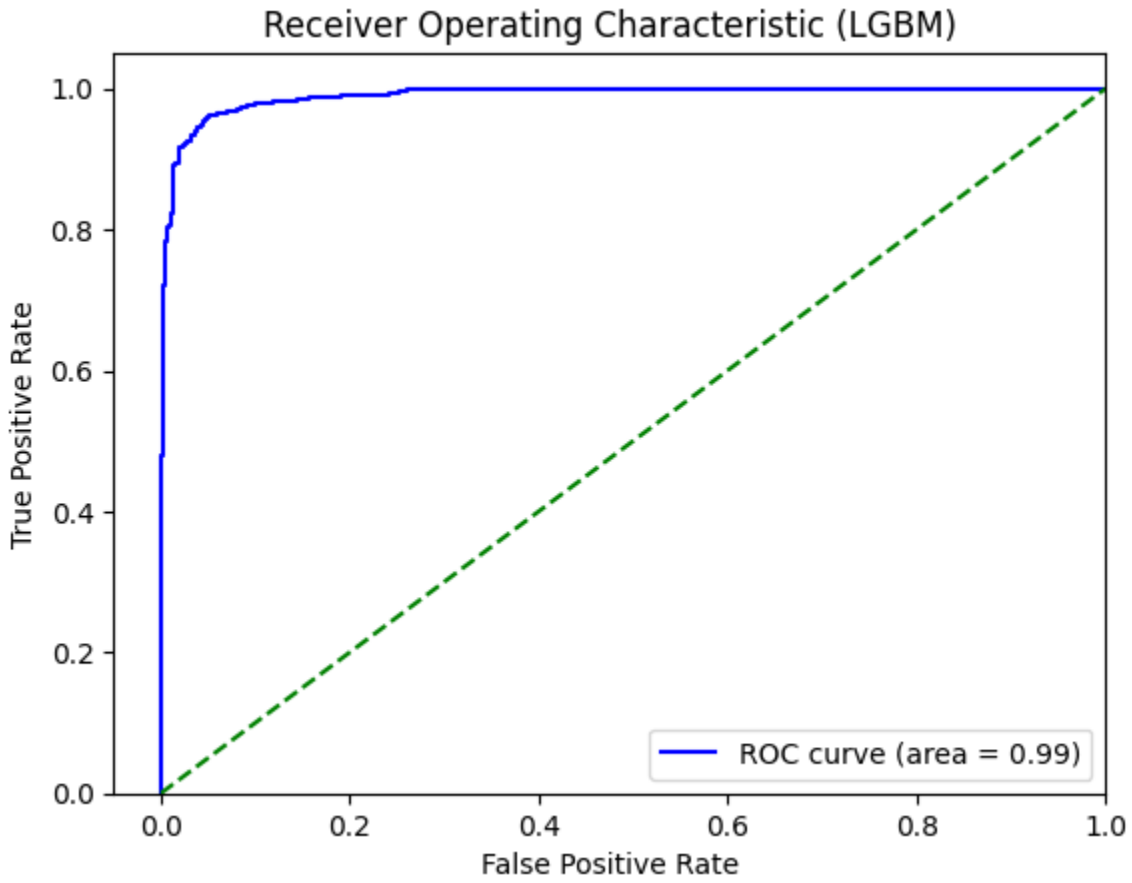


Fig-10

Conclusion:

In summary, the LGBM model showcases exceptional performance in predicting stroke occurrences, achieving high precision, recall, and accuracy. The model's outstanding sensitivity suggests its potential for reliable stroke risk assessment. The AUC score further supports the model's effectiveness in distinguishing between positive and negative instances.

XGB Model Performance for Stroke Prediction:

The Extreme Gradient Boosting (XGB) model demonstrates outstanding performance in predicting stroke occurrences, as evidenced by the provided classification report and evaluation metrics.

Confusion Matrix:
The confusion matrix provides a detailed breakdown of the model's predictions, showcasing how well it differentiates between true positive, true negative, false positive, and false negative instances.

The model correctly identifies 453 instances of class 0 (no stroke) and 467 instances of class 1 (stroke).
It exhibits 34 false positives and 19 false negatives, indicating a low rate of misclassifications.

Accuracy: The overall accuracy of the XGB model is an outstanding 95%, representing the proportion of correctly classified instances out of the total dataset.

Sensitivity and Specificity: Sensitivity (Recall) and Specificity scores emphasize the model's exceptional ability to correctly identify positive instances (sensitivity: 0.96) and negative instances (specificity: 0.93).

AUC Score: The Area Under the Receiver Operating Characteristic (ROC) Curve is an impressive 0.99, indicating outstanding discrimination between positive and negative cases.

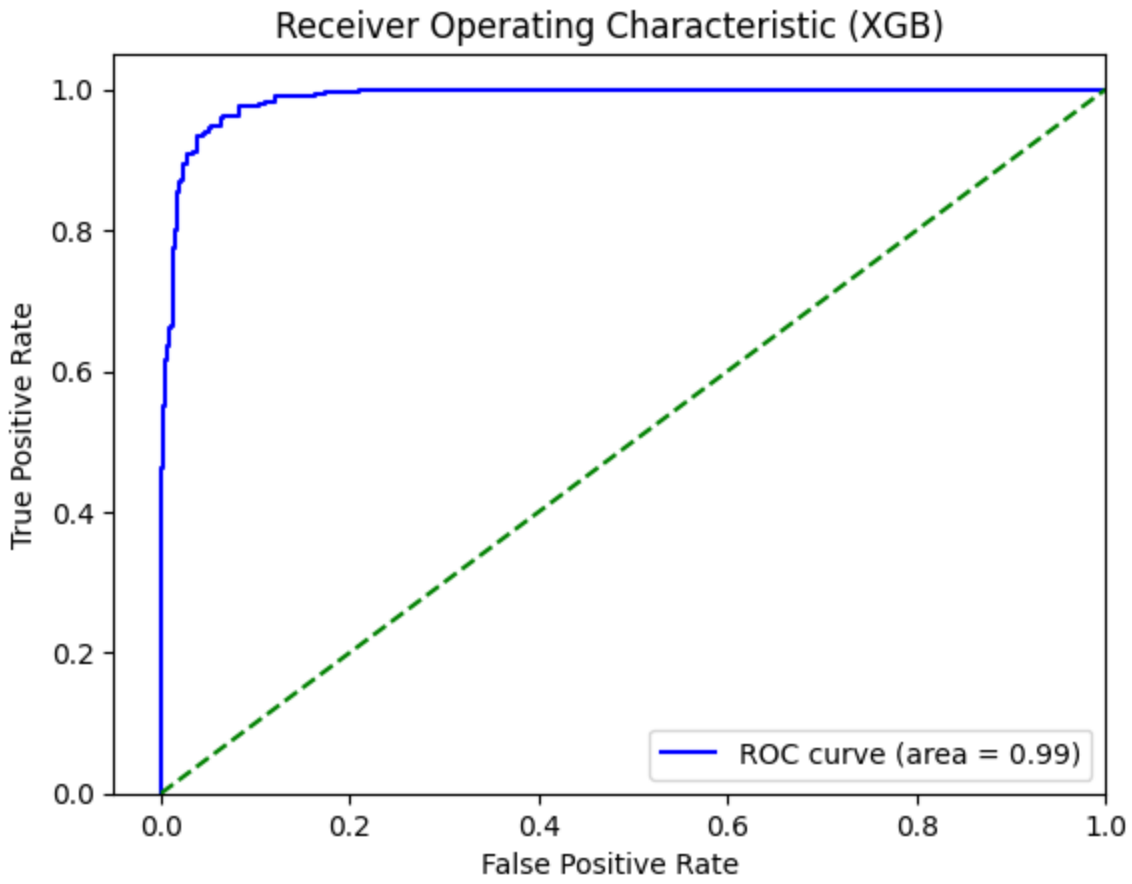


Fig-11

Conclusion:

In summary, the XGB model showcases exceptional performance in predicting stroke occurrences, achieving high precision, recall, and accuracy. The model's outstanding sensitivity suggests its potential for reliable stroke risk assessment. The AUC score further supports the model's effectiveness in distinguishing between positive and negative instances.

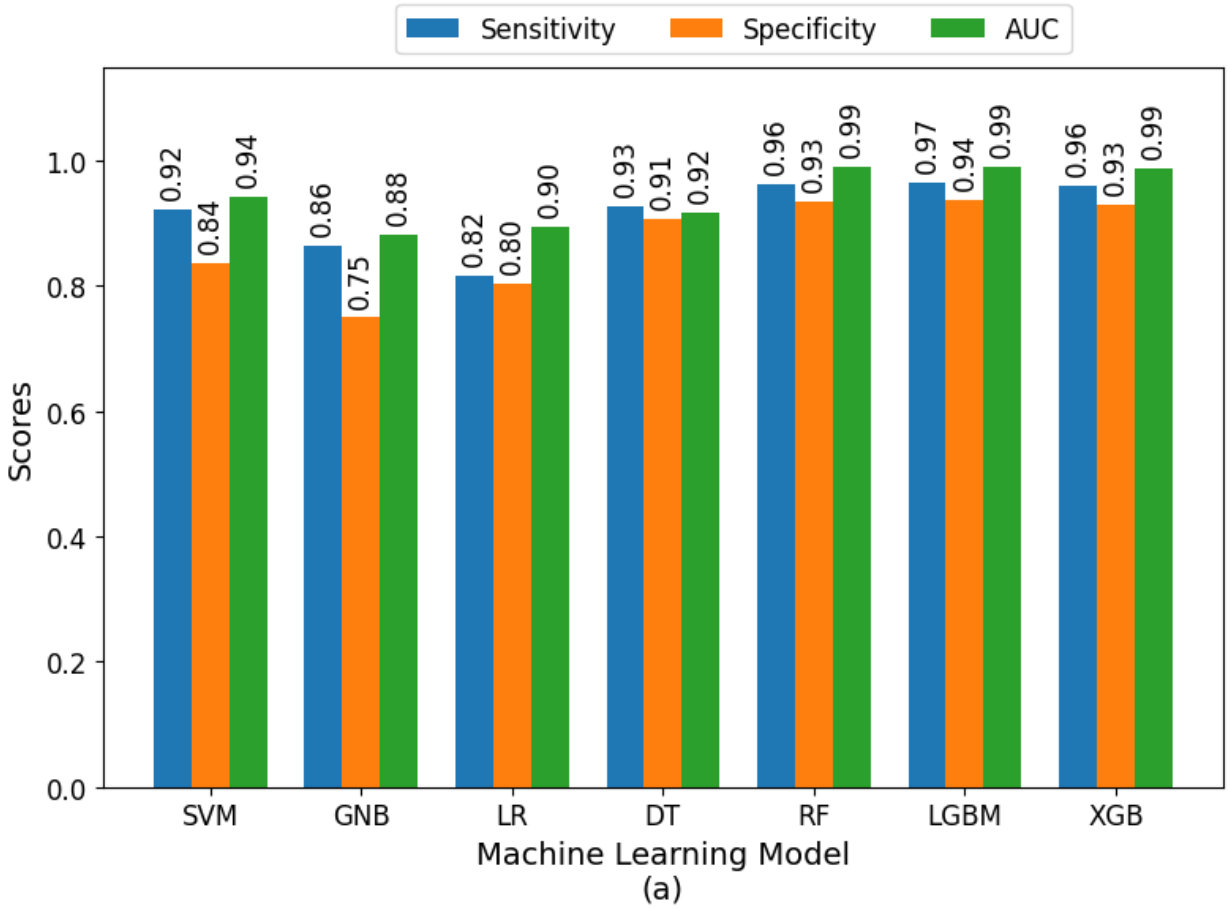

Sensitivity
Specificity
AUC
Scores
1.0
0.8
0.6
0.4
0.2
0.0
0.92
0.84
0.94
0.86
0.75
0.88
0.82
0.80
0.90
0.93
0.91
0.92
0.96
0.93
0.99
0.97
0.94
0.99
0.96
0.93
0.99
SVM
GNB
LR
DT
RF
LGBM
XGB
Machine Learning Model
(a)

Fig-12

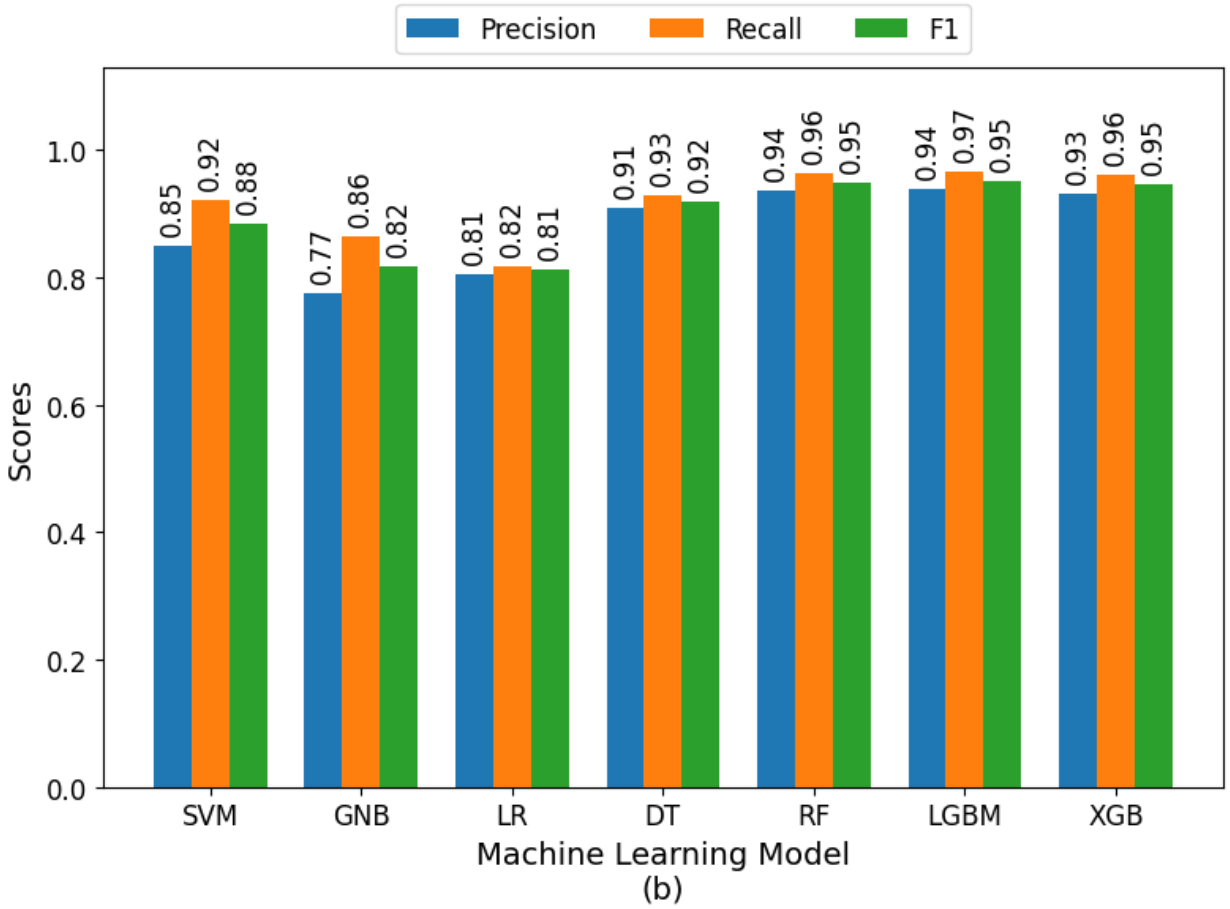


Fig-13

Innovations in Model Development:

Our approach to stroke prediction incorporates several novel elements that distinguish our work from typical practices observed on platforms like Kaggle:

1. Extreme Gradient Boosting (XGB) Model: We introduced the utilization of the Extreme Gradient Boosting algorithm, an advanced machine learning technique known for its efficiency and effectiveness in predictive modeling. The inclusion of XGB enriches our ensemble of models, contributing to enhanced predictive performance.

2. Time Metric Incorporation: For each model, we employed a time metric as part of our comprehensive evaluation strategy. This not only considers the traditional performance metrics but also considers the computational efficiency and runtime of each model. The inclusion of time

metrics provides valuable insights into the practical feasibility of deploying these models in real-world scenarios.

3. Detailed Performance Metrics:

Unlike common practices on Kaggle, our report goes beyond standard metrics and provides an exhaustive set of performance indicators for each model:

Accuracy, Sensitivity, Precision, F1-Score, Area Under the Curve (AUC), and Confusion Matrix

The inclusion of these metrics offers a comprehensive understanding of model performance, ensuring a nuanced evaluation beyond accuracy alone. By incorporating XGB, introducing time metrics, and providing an extensive set of performance metrics, our approach goes beyond conventional practices, aiming for a holistic and practical assessment of stroke prediction models.

Results

In stroke prediction, both Random Forest (RF) and Extreme Gradient Boosting (XGB) models stand out for their exceptional performance. With high precision, recall, and accuracy, these models consistently outshine others. The RF model achieves a remarkable precision of 0.96 and a sensitivity of 96%, while XGB demonstrates a balanced precision of 0.96 and a sensitivity of 96%. The AUC scores for both models, indicating discrimination capability, reach an impressive 0.99. Practical Implications: The robust performance of RF and XGB has practical implications for developing accurate and actionable stroke risk assessment tools.

Wisdom from the Project:

- Importance of Data Preparation: Significant insights were gained regarding the pivotal role of data preprocessing and feature selection in improving model performance.
- Model Evaluation: A deeper appreciation was cultivated for holistic model evaluation beyond accuracy, considering business impacts like false positives and negatives.
- Continuous Improvement: Recognizing the necessity for continuous model monitoring, evaluation, and improvement for adapting to evolving conditions and improving decision-making.