Comprehensive Analysis of Diabetic Readmissions

Milestone: Model Performance Evaluation and Interpretation

Group 18

Mallika Gaikwad
Divya Babulal Shah

gaikwad.mal@northeastern.edu
shah.divyab@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%
Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Mallika Gaikwad
Signature of Student 2: Divya Babulal Shah

Submission Date: 15th March 2024

Model Performance Evaluation and Interpretation

1. Data Splitting:

- We divided the dataset into a training set and a test set to ensure unbiased model evaluation and prevent overfitting.

2. Imbalanced Data Observation:

- Upon analysis, we noticed that the dataset was imbalanced, with a significantly lower number of instances belonging to the positive class (readmissions within 30 days) compared to the negative class.

3. Under sampling Technique:

- To address the imbalance, we implemented under sampling specifically on the training data. This involved randomly selecting a subset of instances from the majority class (no readmissions within 30 days) to balance the class distribution in the training set.

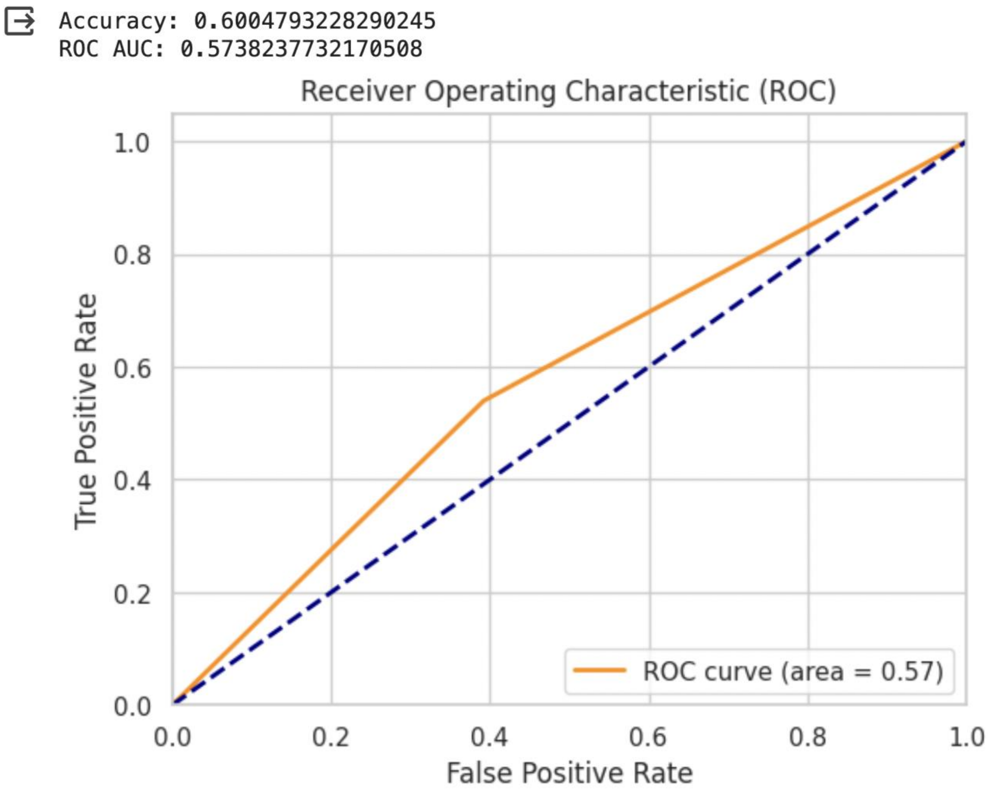
4. Model Training:

- After under sampling the training data, we proceeded to train various models using the balanced dataset. The models included logistic regression, decision trees, random forests, SVM and neural networks suitable for binary classification tasks.

Baseline Model: Logistic Regression serves as a reference point against which more complex models are compared.

Interpretation of Metrics

- Precision: Precision measures the proportion of true positive predictions among all positive predictions made by the model. In this case, the precision for predicting readmissions within 30 days (class 1) is 0.15, indicating that out of all instances predicted as readmissions, only 15% are true positives.
- Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive instances that were correctly predicted by the model. The recall for predicting readmissions within 30 days (class 1) is 0.53, indicating that the model captures 53% of all actual positive instances.
- F1-score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both precision and recall. It is particularly useful when dealing with imbalanced classes. The F1-score for predicting readmissions within 30 days (class 1) is 0.23.
- Accuracy: Accuracy measures the overall correctness of the model's predictions across all classes. In this case, the model has an accuracy of 0.61, meaning that it correctly predicts the class label for 61% of the instances in the dataset.
- ROC AUC: The ROC AUC (Receiver Operating Characteristic Area Under the Curve) is a metric that evaluates the performance of a binary classification model across different threshold values. A higher ROC AUC value indicates better discrimination between positive and negative instances. In this case, the ROC AUC is 0.5764, suggesting moderate discrimination ability of the model.



```
Accuracy: 0.6004793228290245
ROC AUC: 0.5738237732170508
```

ROC Curve for Logistic regression model

Confusion Matrix:

C1 : Diabetes Readmission within 30 days

C2 : Diabetes Readmission > 30 days and No Readmission

	Predicted C1	Predicted C2
Actual C1	8	3
Actual C2	48	41

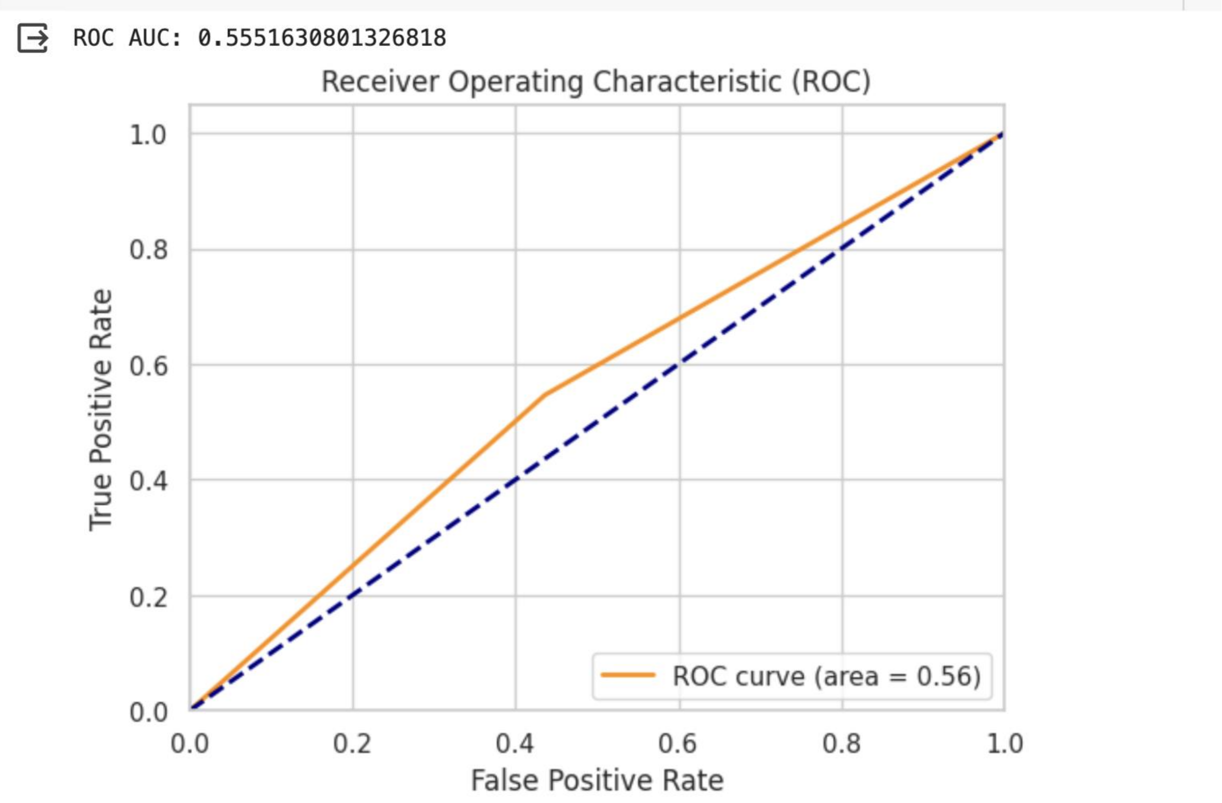
Decision Tree Model

Interpretation of Metrics

- **Accuracy**: The model correctly predicts diabetic readmission within 30 days for 56% of the cases, indicating moderate overall accuracy.
- **Precision**: With a precision of 0.13, the model correctly identifies around 13% of the actual diabetic readmissions within 30 days among all instances it predicts as positive.
- **Recall (Sensitivity)**: The model captures about 55% of all actual diabetic readmissions within 30 days, indicating its ability to detect positive instances.
- **ROC AUC**: An ROC AUC of 0.556 suggests that the model's ability to distinguish between positive and negative instances is slightly better than random chance.

The feature importance values provide insights into which features or variables contribute the most to the decision-making process of the decision tree model in predicting diabetic readmission within 30 days. Here is an interpretation of the feature importance values:

- **Time in Hospital (0.145369)**: This feature has the highest importance, indicating that the duration of a patient's stay in the hospital significantly influences the model's prediction of diabetic readmission within 30 days (about 4 and a half weeks). Longer hospital stays may suggest more severe health conditions or complications, leading to higher chances of readmission.
- **Age (0.040010)** and **Insulin (0.027367):** While age and insulin usage contribute to the model's decision-making, their importance ranks lower compared to other features. Age can still be a factor as older patients may have different healthcare needs and risks. Insulin usage may be indicative of diabetes severity, which is directly related to readmission risks.



```
ROC AUC: 0.5551630801326818
```

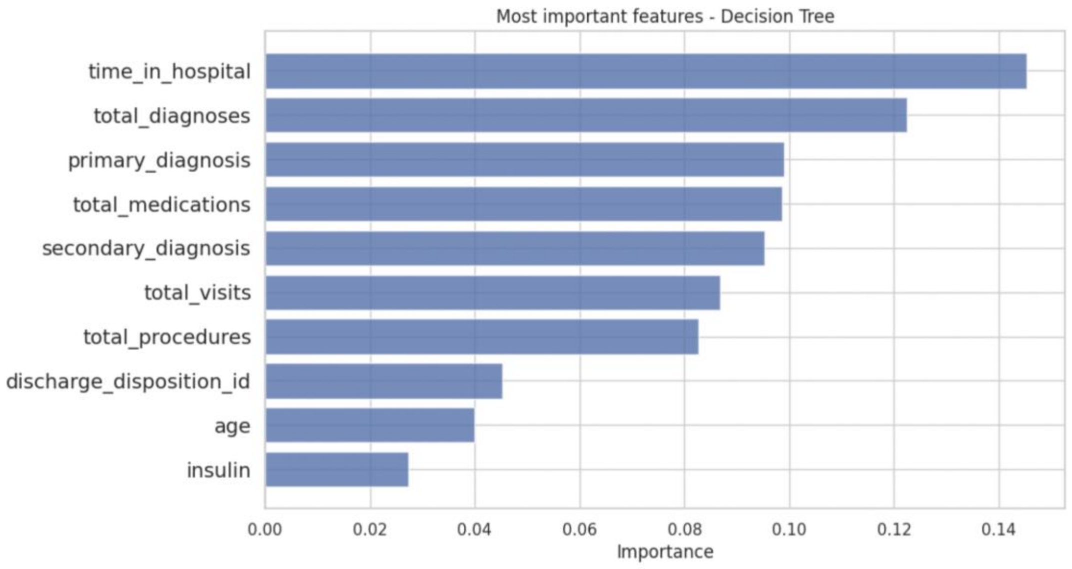
ROC curve for Decision Tree Algorithm

Confusion Matrix:

C1 : Diabetes Readmission within 30 days

C2 : Diabetes Readmission > 30 days and No Readmission

	Predicted C1	Predicted C2
Actual C1	202	227
Actual C2	1447	1860



10 Most Important Features

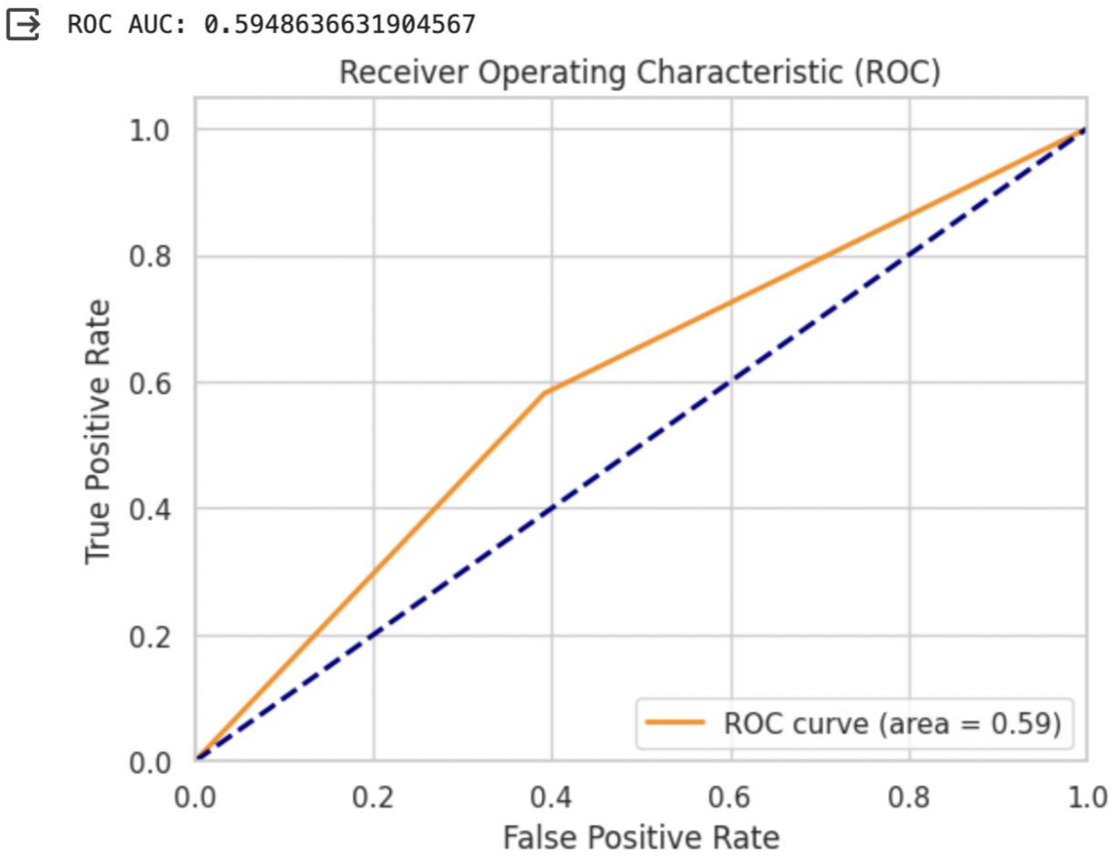
Random Forest Algorithm Model

Interpretation of Metrics

- **Precision:** Precision measures the accuracy of positive predictions. The precision for predicting diabetic readmission within 30 days (class 1) is 0.15, indicating a high false positive rate.
- **Recall:** Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives correctly identified by the model. The recall for predicting diabetic readmission within 30 days is 0.58, indicating moderate performance in identifying actual positive cases.
- **F1-score:** The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall. The F1-score for predicting diabetic readmission within 30 days is 0.24, indicating that the model's performance needs improvement in both precision and recall.
- **Accuracy:** The overall correctness of the predictions made by the model is measured by accuracy. The accuracy of this random forest model is 0.61, indicating a moderate level of overall correctness in predictions.
- **ROC AUC:** The ROC AUC (Receiver Operating Characteristic Area Under the Curve) is a metric that evaluates the model's ability to distinguish between classes. A value of 0.594 indicates some ability to discriminate between positive and negative cases.

Interpretation of feature importance values:

- **Total Diagnoses** has the highest importance, suggesting that the number of diagnoses a patient has is a crucial factor in predicting diabetic readmission within 30 days. Patients with a higher number of diagnoses may have more complex medical conditions or require additional care, influencing their readmission risk.
- **Admission Type ID** has the lowest importance among the features considered, suggesting that the type of admission may have a minimal direct impact on predicting diabetic readmission within 30 days in this model.



```
ROC AUC: 0.5948636631904567
```

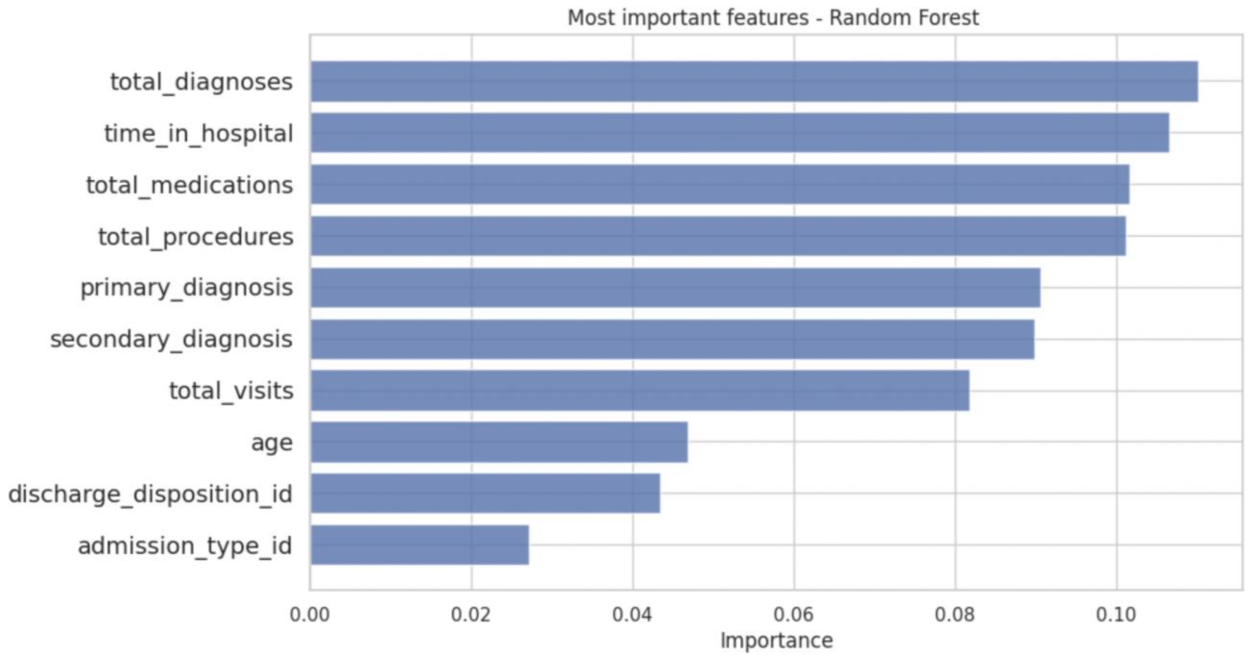
ROC Curve for Random Forest Algorithm

Confusion Matrix:

C1 : Diabetes Readmission within 30 days

C2 : Diabetes Readmission > 30 days and No Readmission

	Predicted C1	Predicted C2
Actual C1	176	253
Actual C2	1370	1937

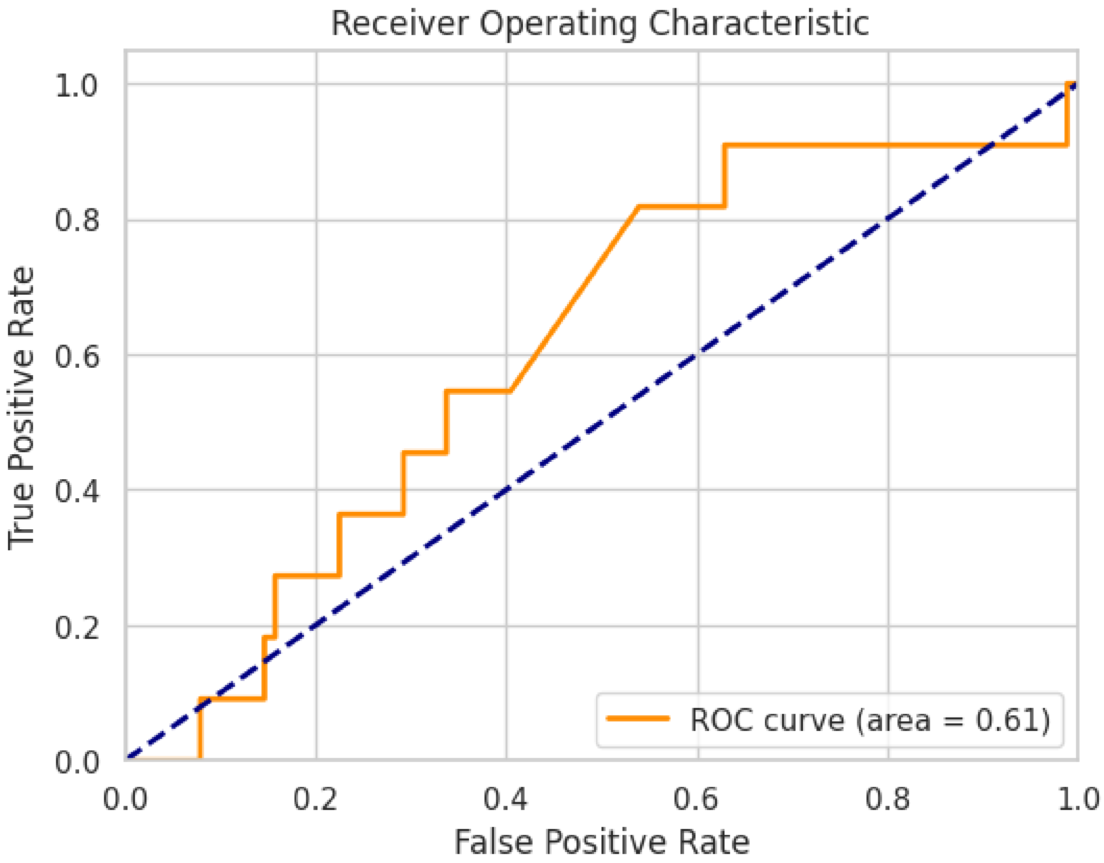


10 Most Important Features for Random Forest Algorithm

Support Vector Machines (SVM)

Interpretation of Metrics

- **Precision**: A precision of approximately 14.29% indicates that the model's ability to correctly predict diabetic readmissions within 30 days is relatively low.
- **Recall**: The recall of around 72.73% suggests that the model is good at identifying actual diabetic readmissions within 30 days.
- **Accuracy**: The model achieved an accuracy of 49%, indicating that its overall predictive performance is limited.



ROC curve for SVM Model

Confusion Matrix:

C1 : Diabetes Readmission within 30 days

C2 : Diabetes Readmission > 30 days and No Readmission

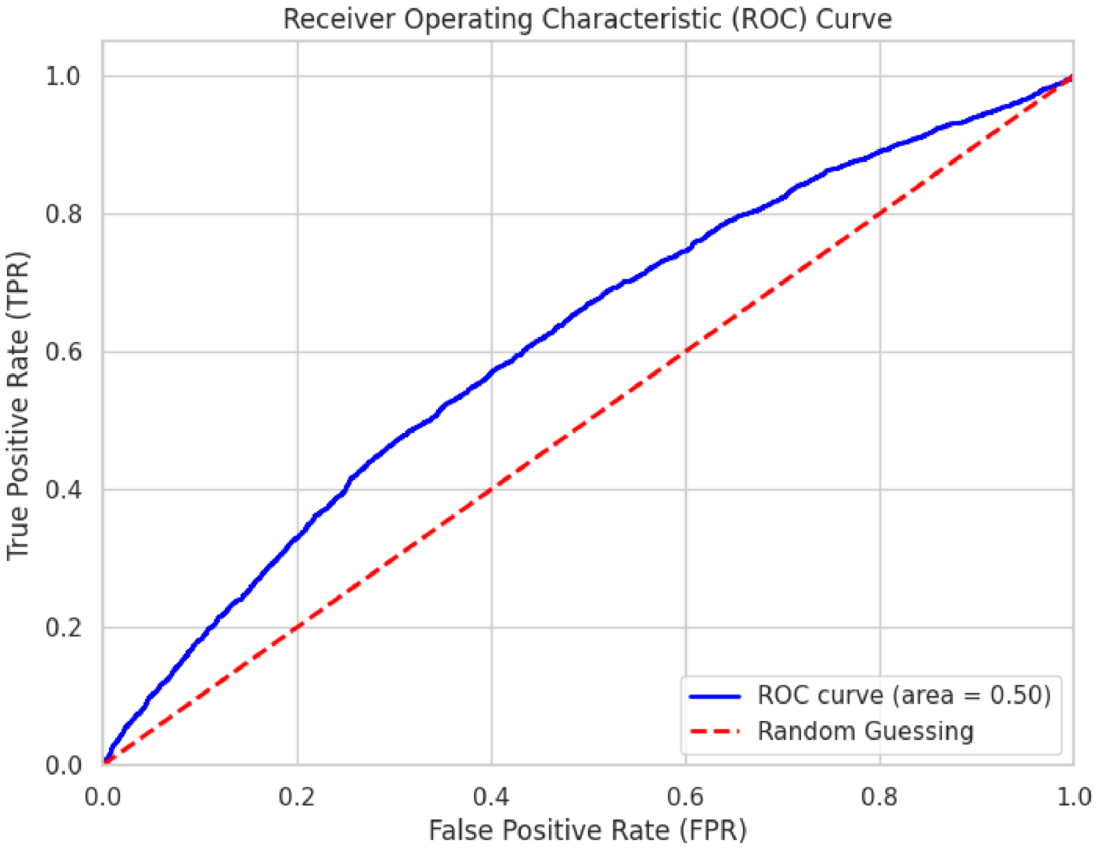
	Predicted C1	Predicted C2
Actual C1	8	3
Actual C2	48	41

Neural Networks

Interpretation of Metrics

- ROC AUC (Receiver Operating Characteristic Area Under the Curve): This metric measures the model's ability to distinguish between positive and negative classes. An ROC AUC score of 0.613 indicates a moderate level of discrimination.
- Accuracy: The accuracy of the model is 0.8496, which means that it correctly predicts the outcome (either readmission or non-readmission) approximately 85% of the time.
- Precision: Precision is the ratio of true positive predictions to the total number of positive predictions (true positives + false positives). A precision of 0.1895 means that when the model predicts readmission, it is correct about 19% of the time.

- Recall: Recall (also known as sensitivity or true positive rate) measures the model's ability to correctly identify positive instances out of all actual positive instances. A recall of 0.1187 indicates that the model captures about 12% of all actual readmissions.
- F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall. The F1 score of 0.146 suggests that the model's overall performance is relatively low in terms of capturing both true positives and minimizing false positives.



ROC Curve for Neural Networks Model.

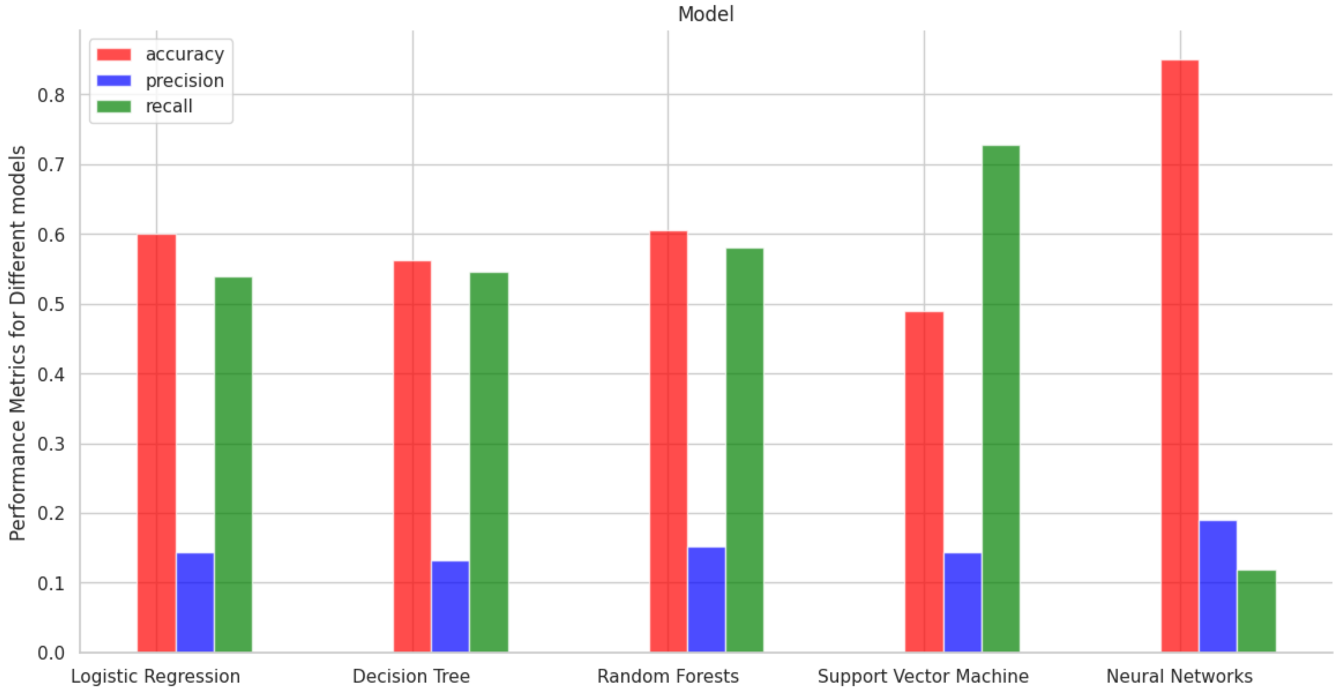
Confusion Matrix:

C1 : Diabetes Readmission within 30 days

C2 : Diabetes Readmission > 30 days and No Readmission

	Predicted C1	Predicted C2
Actual C1	252	1871
Actual C2	1078	16410

Comparison of Models based on Accuracy, Precision, Recall



Comparison of (Accuracy, Precision, Recall) of all the models

Recommended Model: Prioritizing Precision with a Balanced Approach

Given the potential cost of unnecessary interventions but also the seriousness of missing high-risk patients, a model that prioritizes precision with a balanced approach to recall is recommended. Here's why:

- **Focusing on Reducing False Positives:** Unnecessary interventions due to false positives can strain resources (staff time, medication costs) and potentially cause patient anxiety. A model with high precision minimizes these issues.
- **Maintaining Adequate Recall:** Completely missing high-risk patients could lead to complications and readmissions. The model shouldn't miss a significant number of true positives.

Conclusion:

While **Random Forest Algorithm** model performs better on precision than others indicating its ability to make accurate positive predictions:

- There is still a notable bias towards predicting negative cases (class 0), indicating the need for model optimization.
- Further analysis and refinement of the model are recommended to enhance its predictive capabilities for diabetic readmission within 30 days.
- Utilize cross-validation techniques such as k-fold cross-validation to evaluate model performance robustly and ensure generalizability.
- Revisit feature importance analysis to identify and prioritize the most influential features for predicting readmission within 30 days.

- Consider dropping less important features or combining correlated features to simplify the model without sacrificing predictive power.
- Explore and engineer new features that could capture more relevant information related to readmission risk, such as patient history, comorbidities, socioeconomic factors, and previous healthcare utilization patterns.