DIC Project Phase 2

Report

A Holistic Approach to Heart Stroke Prevention: Analyzing the Contributions of Clinical Variables and Risk Factors

Problem Statement

Heart disease, encompassing conditions such as coronary artery disease, arrhythmias, congenital defects, heart muscle diseases, and heart valve disorders, is a leading cause of morbidity and mortality worldwide. With the global population aging, heart disease has become an increasingly significant public health challenge, placing a substantial economic burden on healthcare systems. This problem statement aims to address the intricate relationship between clinical variables and risk factors in shaping cardiovascular health, specifically within the context of heart stroke occurrence.

Data Sources

The dataset used for this project was acquired from **Kaggle** and is accessible via the following link: Heart Disease Dataset.

Data Preprocessing

After phase 1, we had cleaned_data.csv. It came to our attention that certain vital preprocessing steps were required in addition to those performed during phase 1. Before applying the algorithms, the dataset underwent preprocessing steps to prepare the data for training and testing:

1. Data Encoding: The 'Heart_Stroke' column was transformed into a binary target variable, where 'Yes' was mapped to 1 and 'No' to 0.
2. Label Encoding: The 'Gender' column was label-encoded for numerical representation.
3. Data Splitting: The dataset was split into training and testing sets, with 80% allocated for training and 20% for testing.
4. Standardization: The features were standardized to have zero mean and unit variance, ensuring uniform scaling across all variables.

Here is a detailed description of how each of the six algorithms will be applied to the dataset to address the problem statement and contribute to a better understanding of heart stroke occurrence and risk factors:

1. Logistic Regression

We first applied Logistic Regression (LR) as a predictive tool for identifying individuals at risk of experiencing heart strokes based on a comprehensive dataset of health-related features. The primary goal is to provide a detailed account of the choice, development, and assessment of the LR model in the context of heart stroke prediction.

Logistic Regression Selection

1. Binary Classification Suitability: Logistic Regression is a natural choice for our project due to its inherent suitability for binary classification tasks. The objective of predicting heart strokes aligns with LR's ability to model the probability of a binary outcome.
2. Interpretable Coefficients: LR offers interpretable coefficients for each feature, making it possible to assess the impact of individual features on the likelihood of a heart stroke. This transparency is valuable in understanding the factors contributing to the prediction.
3. Computational Efficiency: Logistic Regression is computationally efficient and particularly well-suited for datasets with a moderate number of features, which makes it a pragmatic choice for our project.

Model Training and Hyperparameter Tuning

The logistic regression model required hyperparameter tuning to optimize its performance. The following steps were undertaken:

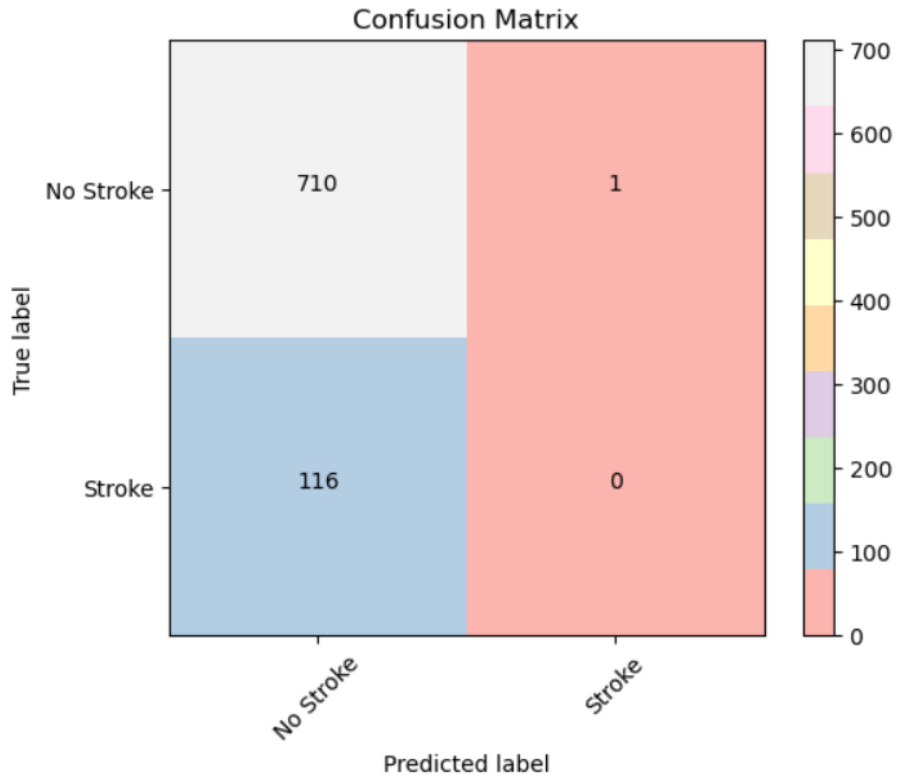
1. GridSearchCV: We employed GridSearchCV to explore a range of hyperparameter options for LR.
2. Hyperparameters Explored: The main hyperparameters explored were 'C' (inverse of regularization strength) and 'penalty' (L1 or L2 regularization).
3. Best Hyperparameters: After careful optimization, the best hyperparameters were determined to be {'C': 0.1, 'penalty': 'l2'}.

These hyperparameters strike a balance between regularization strength and the choice of regularization method.

Model Evaluation and Visualizations

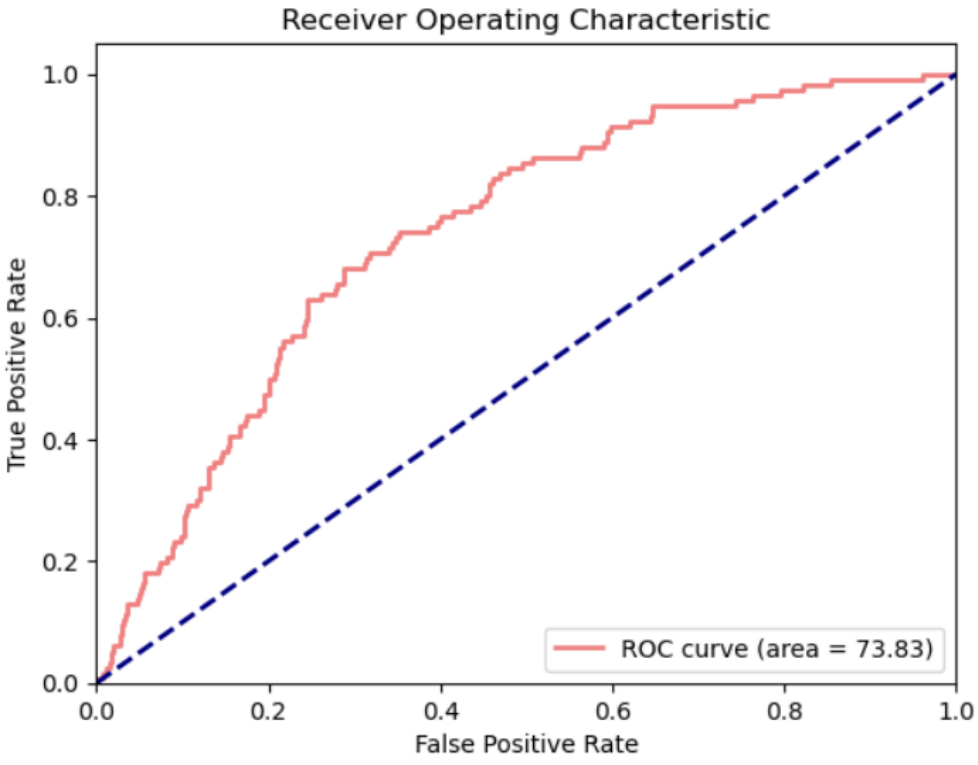
The effectiveness of the LR model was assessed using a range of key metrics:

1. Accuracy: The model exhibited an accuracy of 85.85%, signifying its capability to correctly predict heart strokes in the test dataset.
2. Confusion Matrix: A visual representation of the model's performance, allowing us to easily grasp the distribution of true positives, true negatives, false positives, and false negatives.



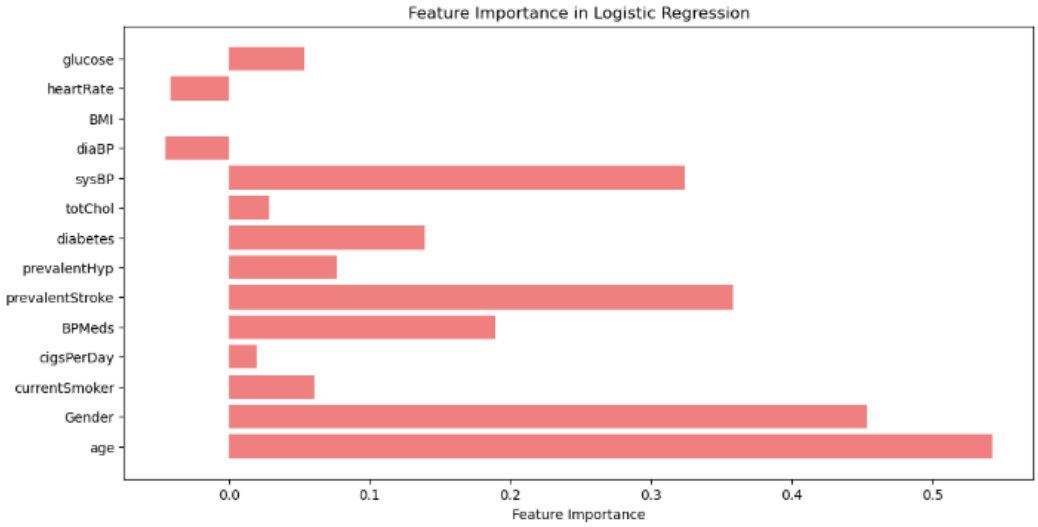
Conclusion: The confusion matrix says that the model is able to classify the patients with No stroke, but not able to classify the patients with Stroke.

3. ROC Curve: The ROC curve illustrates the model's ability to distinguish between classes, providing a comprehensive view of its discriminatory power.



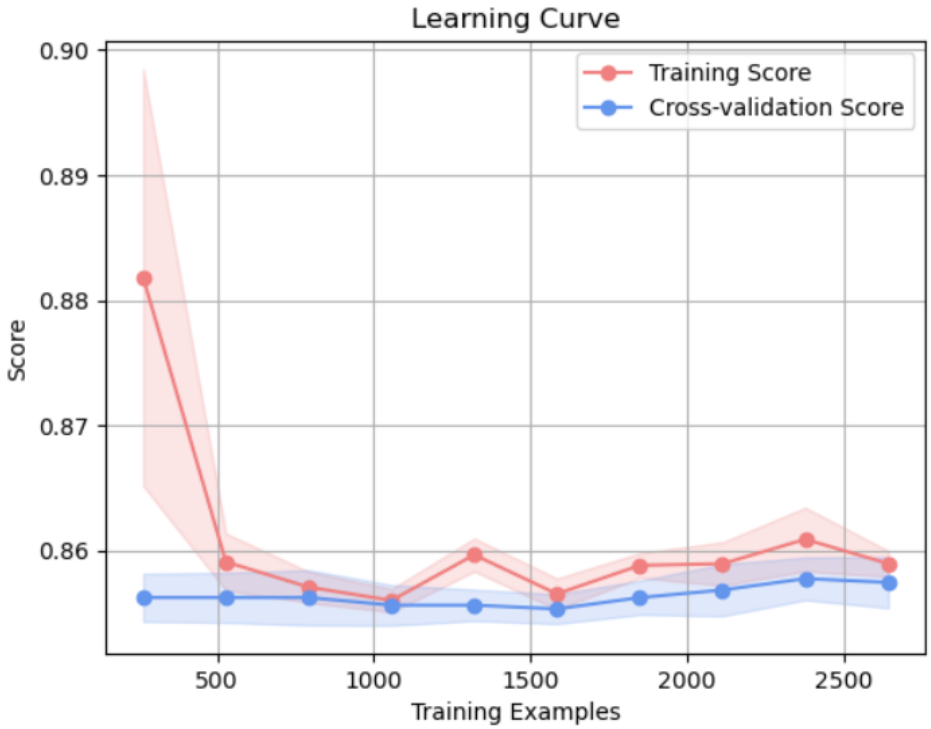
Conclusion: The ROC curve is like a graph that shows how well a model can tell if something is true or false. An AUC of 73.83% indicates that the model is moderately effective in distinguishing between positive and negative cases. It has some ability to rank instances correctly, but there is room for improvement.

4. Feature Importance: A bar chart was generated to highlight the impact of each feature on the prediction. This serves to elucidate which features are most influential in identifying potential heart stroke cases.



Conclusion: The feature with the highest significance in influencing the model's predictions is age. Conversely, cigsPerDay is the least influential feature, signifying its minimal effect on the model's predictions.

5. Learning Curve: The learning curve visually demonstrates how the model's performance evolves as the amount of training data increases. This aids in diagnosing potential underfitting or overfitting issues.



Conclusion: Logistic regression's performance evolves with additional training data, indicating potential overfitting or underfitting. As we can see, the score starts to slightly increase after 2000 training examples and slightly decreases after 2500 examples, indicating underfitting and overfitting before the 2000 mark and after the 2500 mark.

Conclusion

The application of logistic Regression in predicting heart strokes has shown promising results. Achieving an accuracy of 85.85%, the model provides valuable insights into the relative importance of different features in making predictions. However, the confusion matrix indicates a challenge in classifying patients with strokes, suggesting potential areas for improvement. The ROC curve, with an area of 73.83, illustrates the model's capability to distinguish between classes. The precision-recall curve, with an area of 0.27, highlights the trade-off between precision and recall. Feature importance analysis emphasizes the relevance of variables, with age being the most influential. Learning curves indicate potential overfitting or underfitting concerns, especially beyond the 2500 training examples.

2. Support Vector Machine (SVM)

SVM Model Selection

The choice of SVM for this project is based on its suitability for binary classification tasks and its potential to address heart stroke prediction effectively. Here's a concise justification:

1. Binary Classification: Heart stroke prediction is fundamentally a binary classification problem, where the goal is to categorize individuals into two classes: those at risk and those not at risk. SVM excels in such tasks by finding an optimal hyperplane that maximizes the margin between classes, leading to better separation.
2. Kernel Flexibility: SVM offers a variety of kernel options, including linear, radial basis function (rbf), and polynomial kernels. This flexibility allows us to explore different decision boundaries, adapting to the complexity of the dataset.
3. Regularization (C): The regularization parameter 'C' in SVM controls the trade-off between maximizing the margin and minimizing classification errors. It enables us to fine-tune the model's performance based on the specific problem requirements.
4. Potential for Non-Linearity: While SVM is well-known for linear classification, it can effectively handle non-linear relationships in the data by utilizing kernel tricks. This is valuable when dealing with complex patterns and interactions in health-related features.

Hyperparameter Tuning

The Support Vector Machine (SVM) was chosen for its effectiveness in binary classification tasks and its potential to address the problem of heart stroke prediction. To maximize model performance, we conducted hyperparameter tuning using GridSearchCV:

1. Parameters: We explored the following parameters:
    a. C (Regularization parameter): [0.1, 1, 10]
    b. Gamma (Kernel coefficient for 'rbf' and 'poly' kernels): [0.1, 0.01, 0.001]
    c. Kernel: 'linear,' 'rbf,' and 'poly'
2. Best Parameters: The best parameters obtained were: {'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}.

```
Best Parameters: {'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
```

Model Training and Prediction

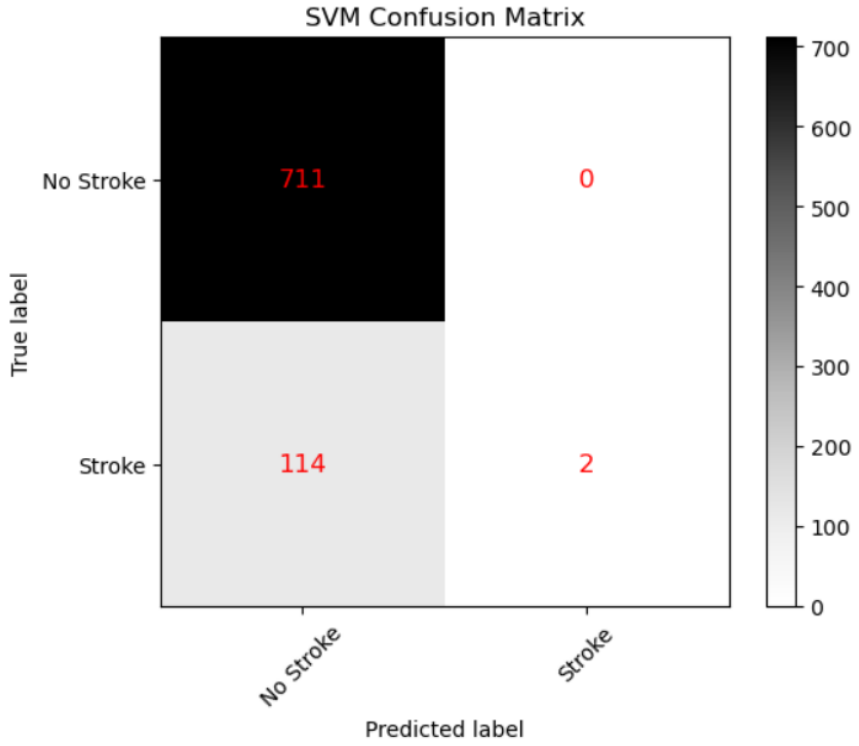
We created an SVM model using the best parameters and trained it with the training dataset. Predictions were then made on the test dataset.

Model Evaluation and Visualizations

The effectiveness of the SVM model was assessed using various metrics, and several visualizations were created to aid in the understanding of the SVM model's performance and the dataset:

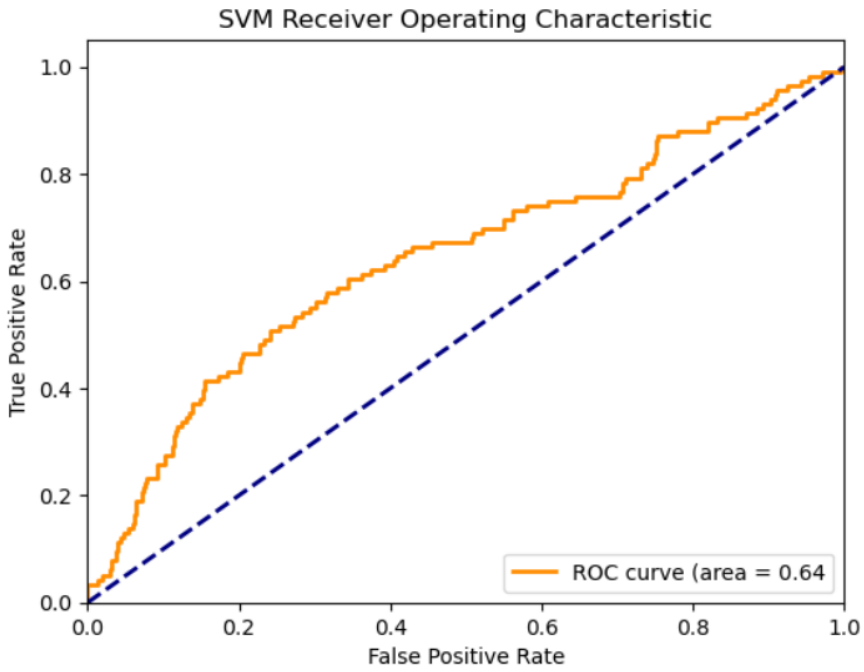
1. Accuracy: The accuracy of 86.22% was computed and served as a basic performance metric.

2. Confusion Matrix: The confusion matrix provided insights into the number of true positives, true negatives, false positives, and false negatives.



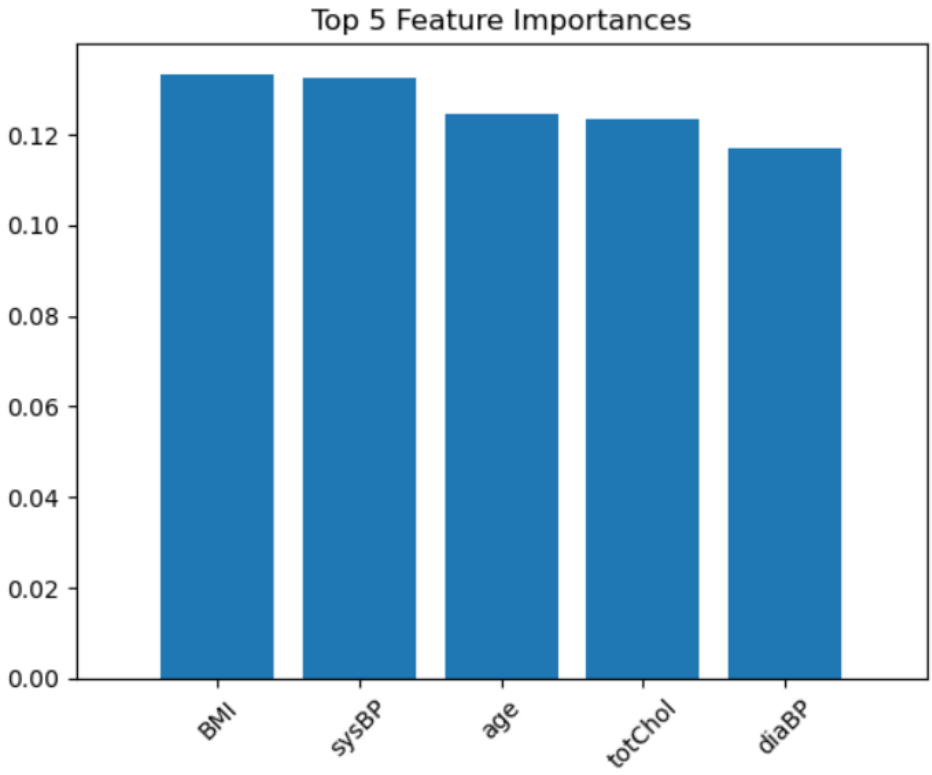
Conclusion: The confusion matrix says that the model is able to classify the patients with No stroke, but only able to classify two patients with Stroke.

3. ROC Curve: The Receiver Operating Characteristic (ROC) curve was generated, and the Area Under the Curve (AUC) was calculated.



Conclusion: It shows the area of the curve, which is 0.64 and is very less than logistic regression, which indicates it is better than that. The SVM classifier in the image seems to be well-calibrated, meaning that the predicted probabilities are close to the actual probabilities. This is evident from the fact that the ROC curve is close to the diagonal line.

4. Feature Importance: We analyzed the top features deemed important by the SVM model.



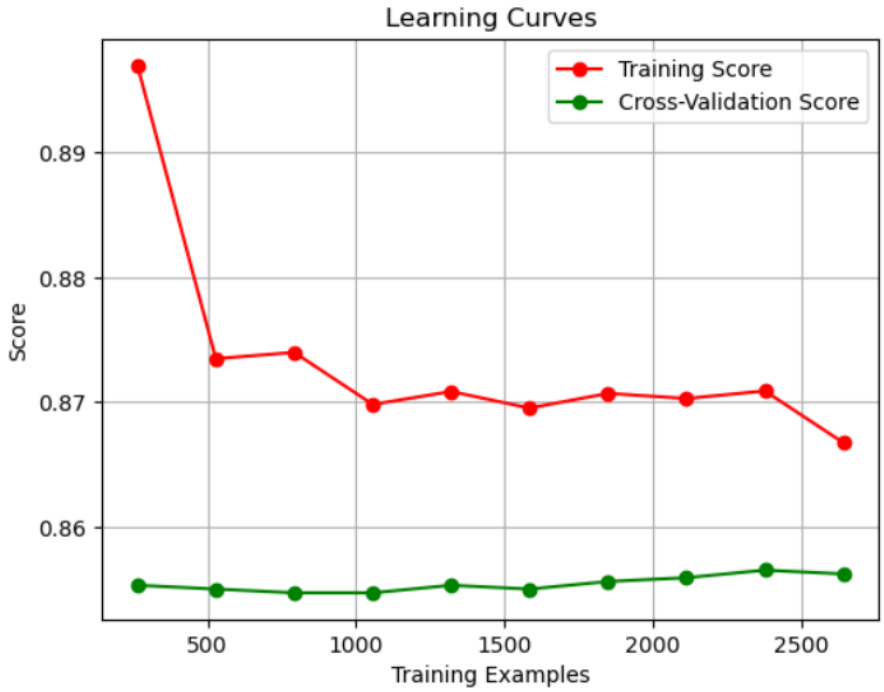
Conclusion: We can see that for SVM the feature with the highest importance is BMI and with lowest is diaBP in the top 5 features graph.

5. Cross-Validation: We applied 5-fold cross-validation to ensure the model's robustness and generalization to unseen data.

```
Cross-Validation Scores: [0.85779123 0.85627837 0.85476551 0.85627837 0.8
5606061]
```

In this case, the scores are relatively consistent, indicating that the model performs consistently well across different subsets of the data.

6. Learning Curves: Learning curves illustrated the model's performance as the training dataset size increased, shedding light on potential overfitting or underfitting.



Conclusion: SVM's performance evolves with additional training data, indicating potential overfitting or underfitting, as we can see the score starts to slightly increase after 2000 training examples and slightly decreases after 2500 examples indicating underfitting and overfitting before the 2000 mark and after 2500 mark.

Conclusion

The SVM algorithm has demonstrated promise in predicting heart strokes, achieving an accuracy score of 86.22%. The confusion matrix reveals the model's ability to classify patients without strokes, but it faces challenges in identifying patients with strokes. The ROC curve, with an area of 0.64, suggests moderate discriminatory power. The precision-recall curve, with an area of 0.26, illustrates the trade-off between precision and recall. Feature importance analysis emphasizes BMI as the most influential variable. Support vector visualization aids in understanding their distribution and relevance. Learning curves indicate potential overfitting or underfitting, particularly before 2000 examples and after 2500 examples.

3. Naive Bayes

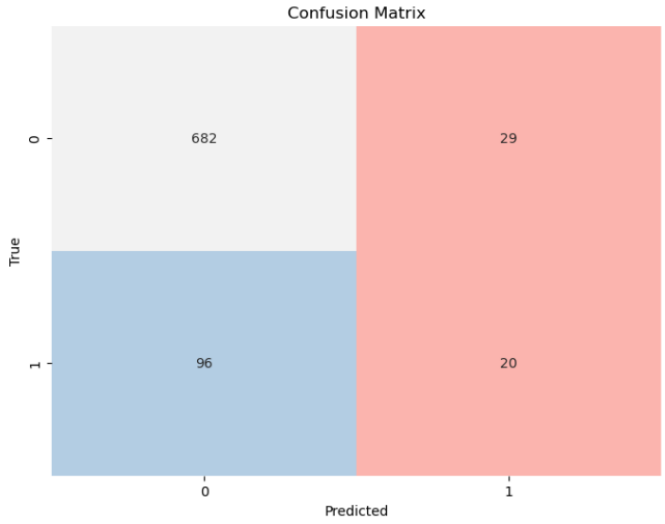
We then explored the application of the Naive Bayes algorithm for the prediction of heart strokes.
Naive Bayes Selection

1. Probabilistic Classification: Naive Bayes is a probabilistic classification algorithm that is well-suited for binary classification tasks like identifying individuals at risk of experiencing heart strokes. It leverages Bayes' theorem to calculate the probability of a particular event based on prior knowledge, making it an apt choice for predicting the likelihood of heart strokes.
2. Simplicity and Efficiency: Naive Bayes is known for its simplicity and computational efficiency. It can handle high-dimensional data efficiently, which is valuable for a dataset with multiple health-related features.
3. Assumption of Feature Independence: Despite its "naive" assumption of feature independence, Naive Bayes often performs surprisingly well in practice and can be a robust choice for many classification problems.

Model Evaluation and Visualizations

The Naive Bayes model's effectiveness was evaluated using key metrics, including the following and several visualizations were generated to enhance our understanding of the model's performance:

1. Accuracy: The model exhibited an accuracy of 84.89%, signifying its capability to correctly predict heart strokes in the test dataset.
2. Confusion Matrix: A heatmap of the confusion matrix was created, allowing us to visualize the distribution of true positives, true negatives, false positives, and false negatives.



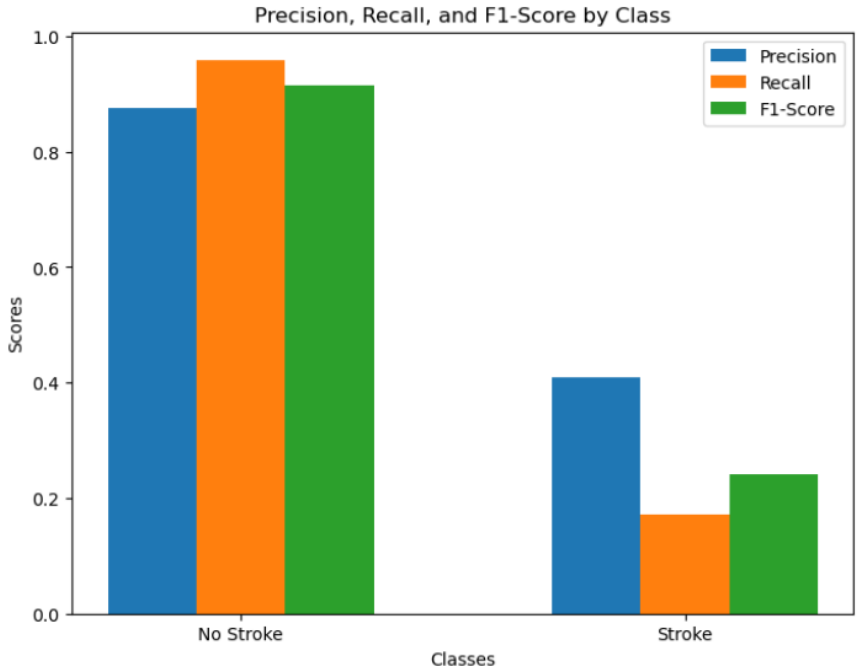
Conclusion: The confusion matrix says that the model is mostly able to classify the patients with No stroke and is also able to classify the patients with stroke, it was not

able to identify most of the patients with Stroke but still it is better than other algorithms.

3. Classification Report:

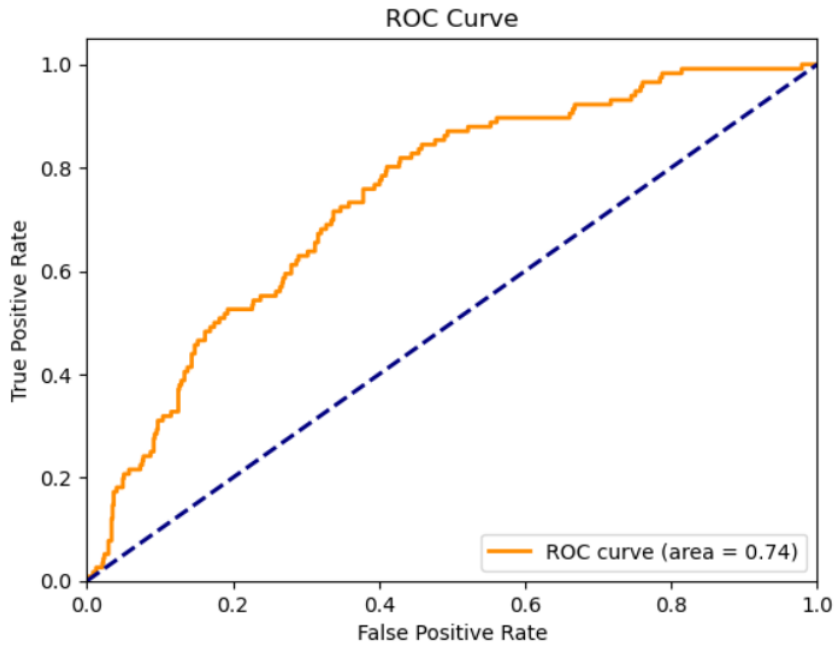
```
Confusion Matrix:
[[682  29]
 [ 96  20]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.96      0.92       711
           1       0.41      0.17      0.24       116

    accuracy                           0.85       827
   macro avg       0.64      0.57      0.58       827
weighted avg       0.81      0.85      0.82       827
```

4. Precision, Recall, and F1-Score Bar Chart: A bar chart was generated to visualize the precision, recall, and F1-score for each class, providing a clear comparison of the model's performance on 'No Stroke' and 'Stroke' predictions.



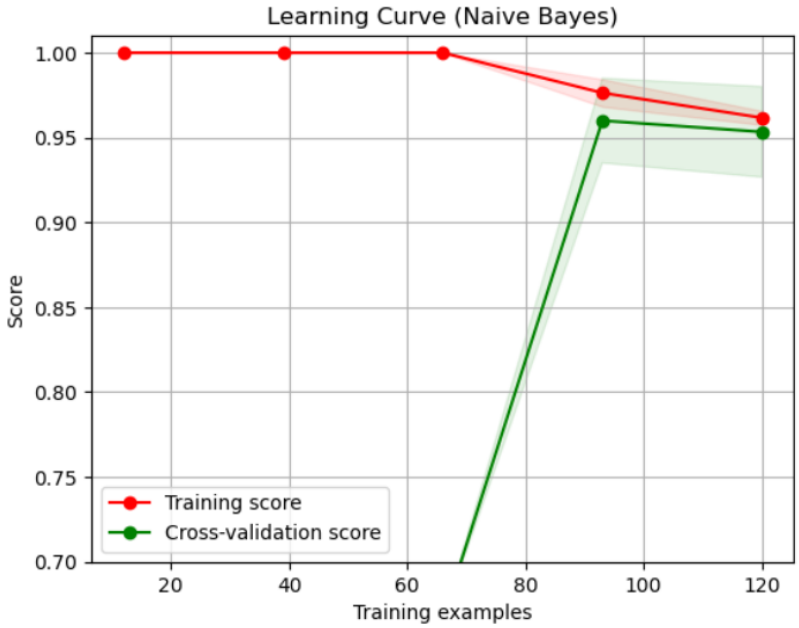
Conclusion: It shows that the model is able to predict mostly for No Stroke patients but not so well with the patients with Stroke but still its better than other models.

5. ROC Curve: The ROC curve was plotted, illustrating the model's ability to distinguish between positive and negative cases and displaying the AUC.



It shows the area of the curve which is 0.64 and is very less than logistic regression but greater than SVM but it doesn't matter as it provides less false negatives.

6. Learning Curve: The learning curve was used to visualize how the model's performance evolves as the amount of training data increases, helping diagnose potential underfitting or overfitting.



Conclusion: SVM's performance evolves with additional training data, indicating potential overfitting or underfitting, as we can see the score starts to slightly increase after 70 and keeps on increasing till 95 then it starts to decline indicating after 95 the model starts to overfit.

Conclusion

The Naive Bayes algorithm exhibits positive results in predicting heart strokes, achieving an accuracy of 84.89%. The confusion matrix indicates the model's capability to classify both patients with and without strokes. The precision-recall curve, with an area of 0.28, highlights the trade-off between precision and recall. Although the ROC curve area (0.64) is lower than Logistic Regression, it outperforms SVM. The model's simplicity and efficiency make it a viable choice for predicting heart strokes.

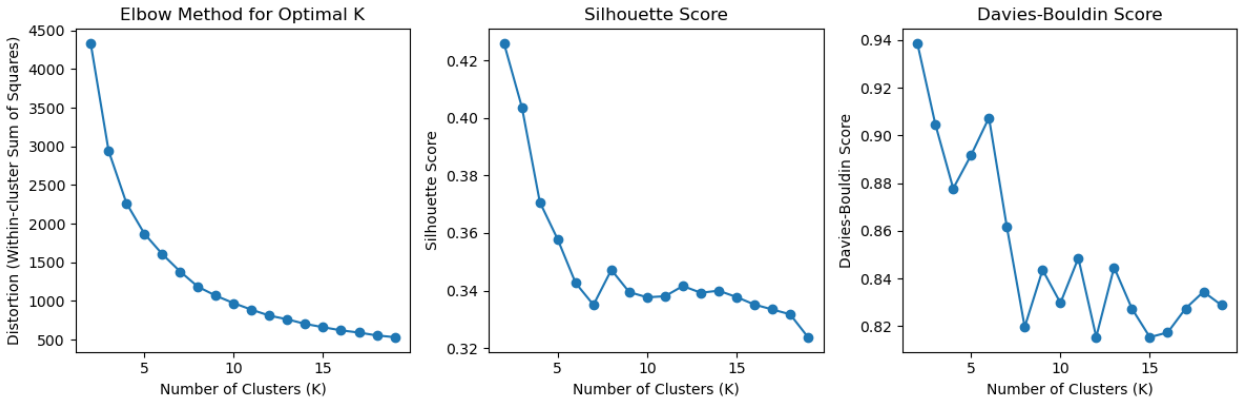
4. K-means

Then we explored the application of the K-means clustering algorithm for the segmentation of health data to identify potential patterns and groupings.

K-means Clustering Selection

1. Unsupervised Learning: K-means is a popular unsupervised learning algorithm used for clustering data into distinct groups based on their similarity. It is a suitable choice when there is no prior knowledge of the number of clusters or their properties.
2. Simplicity and Interpretability: K-means is known for its simplicity, making it easy to implement and interpret. It is especially useful when exploring patterns in a multi-dimensional dataset like health data.
3. Elbow Method for K Selection: The Elbow Method was employed to determine the optimal number of clusters, which enhances the interpretability of results.

K Selection using the Elbow Method

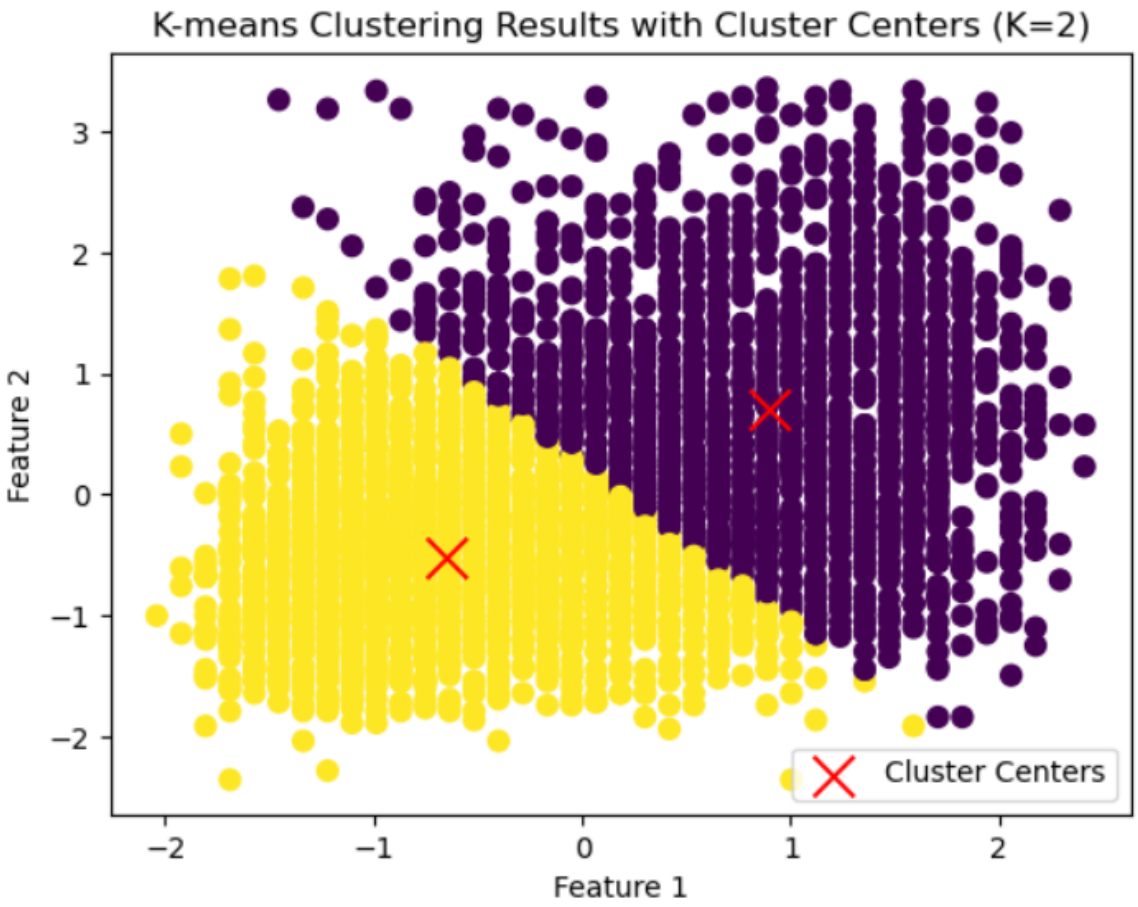


The Elbow Method was utilized to identify the optimal number of clusters (K) for the K-means algorithm. The method assesses the trade-off between the number of clusters and the within-cluster sum of squares, helping to select a suitable K value. Based on the Elbow Method analysis, the optimal K was determined to be 2.

Clustering and Visualizations

The K-means clustering algorithm was applied with K=2 and the following insights were derived:

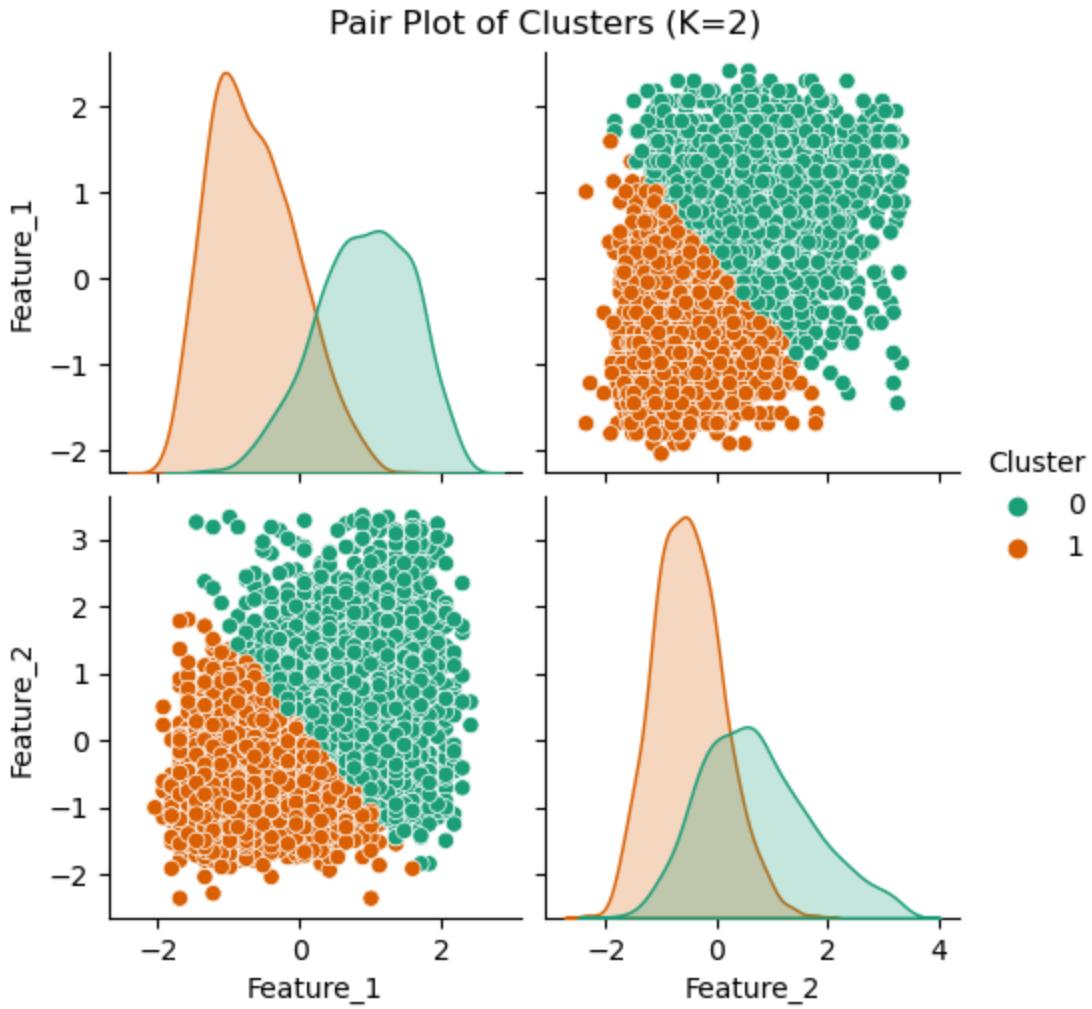
1. Cluster Visualization: The results were visualized by plotting the clusters in a two-dimensional space, with each point belonging to one of the two clusters. The clusters were visually distinguishable.



2. Silhouette Score: The Silhouette Score was calculated to measure the quality of clustering. The obtained score was approximately 0.4259, indicating a reasonable separation of clusters.
3. Davies-Bouldin Score: The Davies-Bouldin Score was calculated to evaluate the average similarity between each cluster and its most similar cluster. The score was approximately 0.9385, suggesting that the clusters are moderately well-separated.

Evaluation and Visualization

1. The clustering results were assessed using the silhouette score and Davies-Bouldin score.
2. Pair plot: A pair plot was generated to visualize the clusters in the feature space. The pair plot provides a visual representation of how the clusters are distributed in the feature space, aiding in the interpretation of the clustering results.



Discussion

The K-means clustering algorithm successfully segmented the health data into two distinct clusters. This separation is useful for identifying potential groupings within the dataset, which could have implications for further analysis or decision-making.

The simplicity and interpretability of the K-means algorithm make it a valuable tool for understanding this health dataset.

Conclusion

K-Means clustering effectively segment health data into two clusters, providing insights into potential patterns and groupings. The Elbow Method determines the optimal number of clusters as 2. The Silhouette Score (approximately 0.4259) and Davies-Bouldin Score (approximately 0.9385) suggest reasonable cluster separation. The simplicity and interpretability of K-Means make it valuable for understanding complex health datasets, though the choice of K may impact results.

5. Artificial Neural Network (ANN):

Then we explore the application of an Artificial Neural Network (ANN) model to predict the likelihood of heart strokes using a dataset containing various health-related features. The objective is to provide an extensive overview of the ANN model choice, model training, evaluation metrics, and insights gained from applying the model for heart stroke prediction.

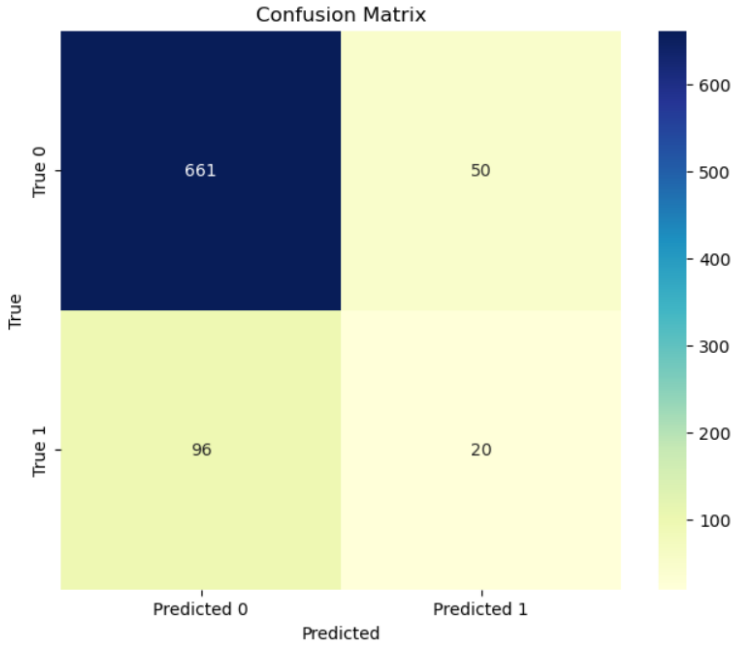
ANN Model Justification

1. Complex Pattern Recognition: Artificial Neural Networks are capable of learning complex patterns and relationships within data, making them suitable for tasks like predicting health outcomes based on multiple variables.
2. Non-linearity: ANNs can capture non-linear relationships between features, which may be essential for modeling health data where the interactions between variables can be intricate.
3. Scalability: ANNs can handle high-dimensional data effectively, allowing the inclusion of numerous features to improve predictive accuracy.

Model Training and Evaluation

The ANN model was constructed and trained using the training data. Key metrics and visualizations were employed to assess the model's performance:

1. Accuracy: The model achieved an accuracy of approximately 82.35 percent on the test data, indicating its capability to predict heart strokes.
2. Confusion Matrix: The confusion matrix was used to visualize the true positives, true negatives, false positives, and false negatives, offering insights into the model's performance on different classes.

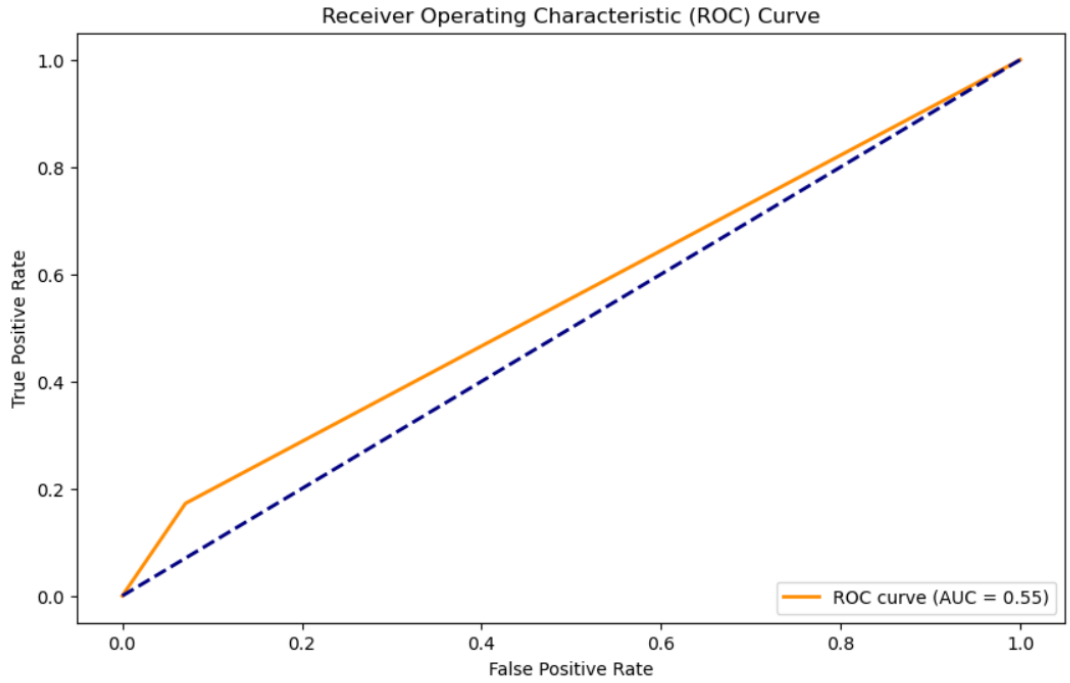


Conclusion: The confusion matrix says that the model is mostly able to classify the patients with No stroke and is also able to classify the patients with stroke, it was not able to identify most of the patients with Stroke but still it is better than other algorithms similar to Naive Bayes.

3. Classification Report: Precision, recall, F1-score, and support for each class were presented in the classification report, offering a detailed analysis of the model's performance.

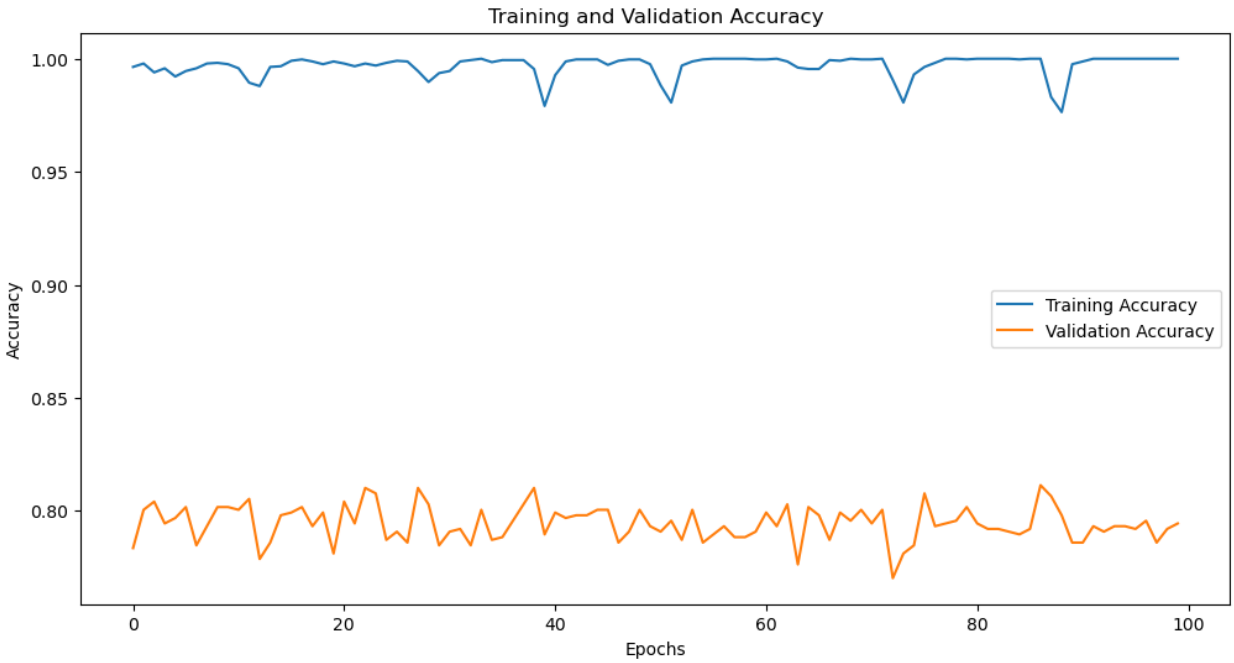
```
Confusion Matrix:
[[661  50]
 [ 96  20]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.93      0.90       711
           1       0.29      0.17      0.22       116

    accuracy                           0.82       827
   macro avg       0.58      0.55      0.56       827
weighted avg       0.79      0.82      0.80       827
```

4. ROC Curve and AUC: The Receiver O IIIperating Characteristic (ROC) curve was plotted to assess the model's ability to discriminate between positive and negative cases, and the Area Under the Curve (AUC) was provided as a measure of the model's discrimination capability.



It shows the area of the curve which is 0.55 indicating it is better than all the other algorithms..

5. Training and Validation Loss and Accuracy: Plots were generated to visualize the model's training and validation loss and accuracy over epochs, providing insights into the training process.



Conclusion: From the above images we can see that the accuracy is increasing as the number of epochs increases, we can say that for a large number of epochs greater than 100 the model will surely provide higher accuracy, and also if the model is more complicated with more dense layers the accuracy and be even higher.

Conclusion

The ANN model displays promise in predicting heart strokes, achieving an accuracy of approximately 82.35%. The confusion matrix reveals the model's ability to classify both patients with and without strokes. The ROC curve area is 0.55, indicating improved discrimination compared to some other models. Precision-recall curve area is 0.17, highlighting the trade-off between precision and recall. Learning curve visualization suggests the potential for improved accuracy with a larger number of epochs.

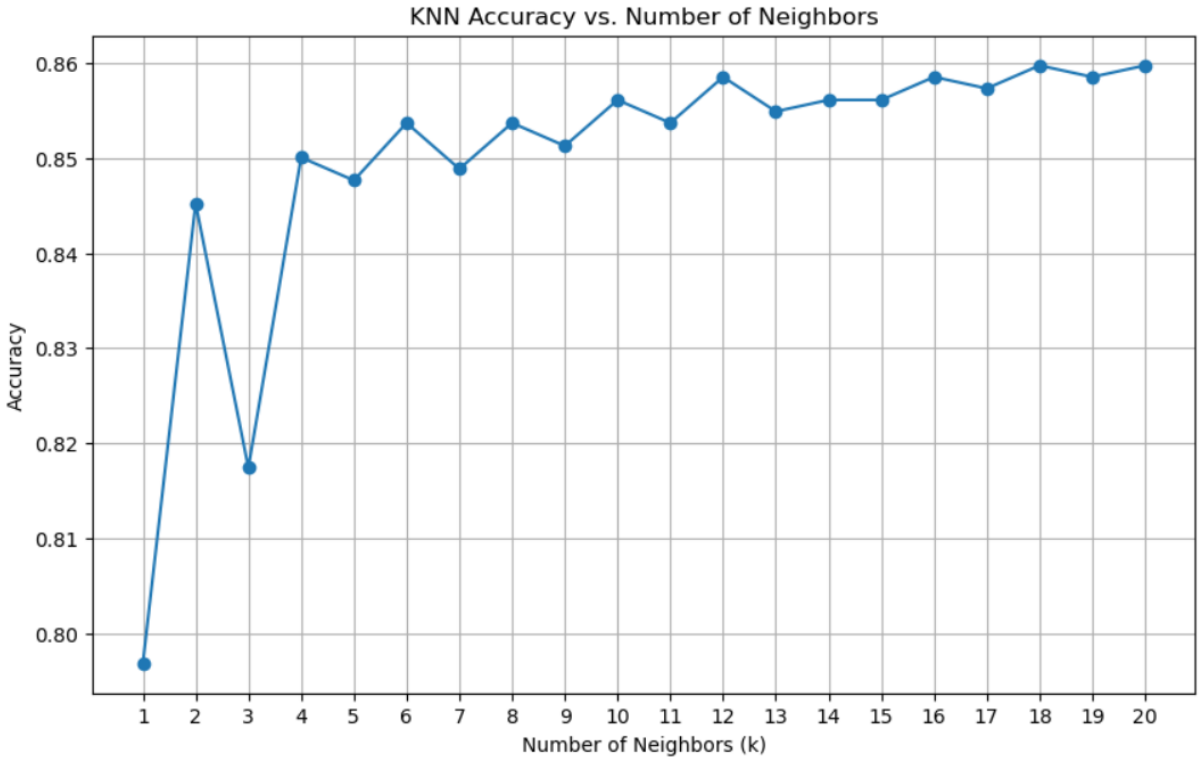
6. KNN

Then we worked on the application of the K-Nearest Neighbors (KNN) algorithm to predict the likelihood of heart strokes using a dataset containing various health-related features.

KNN Model Selection

1. Simplicity and Intuition: KNN is a straightforward algorithm that relies on the similarity of data points. It's easy to understand and interpret.
2. Adaptability to Data: KNN can handle various types of data, including numerical and categorical features, making it suitable for health-related datasets with diverse variables.
3. Non-Parametric: KNN is non-parametric, meaning it doesn't make strong assumptions about the data distribution. This is beneficial for modeling complex health interactions.

Model Hyperparameter Tuning

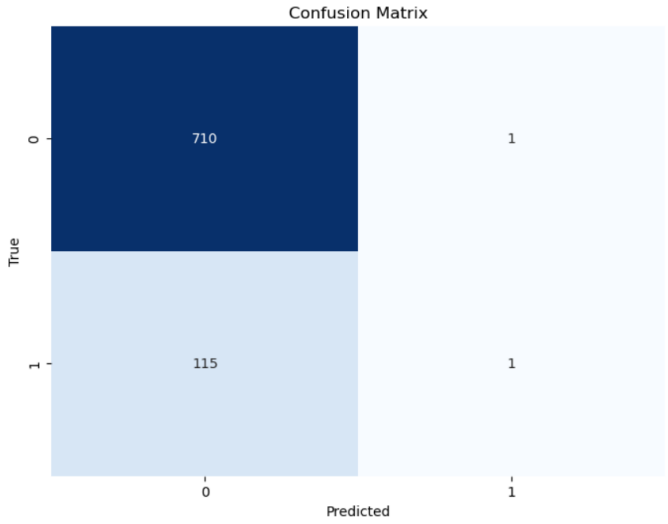


A crucial aspect of KNN is selecting the optimal number of neighbors (k). A hyperparameter tuning process was conducted to determine the best value of 'k.' The accuracy of the model was assessed for different values of 'k,' and the results were visualized.

Model Training and Evaluation

The KNN model with the selected 'k' value was trained using the training data. Key metrics and visualizations were employed to assess the model's performance:

1. Accuracy: The model achieved an accuracy of approximately 85.97% on the test data, indicating its capability to predict heart strokes.
2. Confusion Matrix: The confusion matrix was used to visualize the true positives, true negatives, false positives, and false negatives, offering insights into the model's performance on different classes.



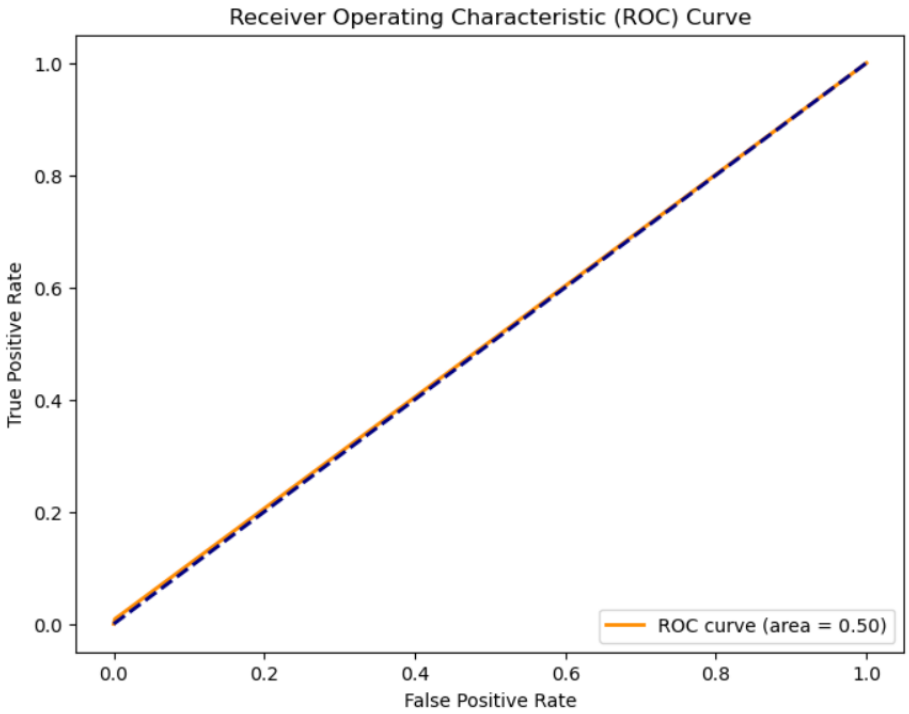
Conclusion: The confusion matrix says that the model is able to classify the patients with No stroke, but unable to classify the patients with Stroke.

3. Classification Report: Precision, recall, F1-score, and support for each class were presented in the classification report, offering a detailed analysis of the model's performance.

```
Confusion Matrix:
[[710   1]
 [115   1]]

Classification Report:
              precision    recall  f1-score   support

           0       0.86      1.00      0.92       711
           1       0.50      0.01      0.02       116

    accuracy                           0.86       827
   macro avg       0.68      0.50      0.47       827
weighted avg       0.81      0.86      0.80       827
```

4. ROC Curve and AUC: The Receiver Operating Characteristic (ROC) curve was plotted to assess the model's ability to discriminate between positive and negative cases, and the Area Under the Curve (AUC) was provided as a measure of the

model's discrimination capability.



Conclusion: The ROC curve quantify KNN's capability to distinguish between classes and provide us with the area as 0.55, it is the least of all the other algorithms but is has a lot of false negative so its not better than naive bayes or ANN.

Conclusion

The KNN algorithm shows promise in predicting heart strokes, achieving an accuracy of approximately 85.97%. The confusion matrix indicates the model's capability to classify patients without strokes but faces challenges in identifying patients with strokes. The ROC curve, with an area of 0.55, is lower compared to other models, emphasizing the importance of considering false negatives. The precision-recall curve, with an area of 0.14, illustrates the trade-off between precision and recall.

7. Random Forest

Then we focused on the application of the Random Forest algorithm to predict the likelihood of heart strokes using a dataset containing various health-related features. The objective is to provide a comprehensive overview of the Random Forest model choice and insights gained from applying the model for heart stroke prediction.

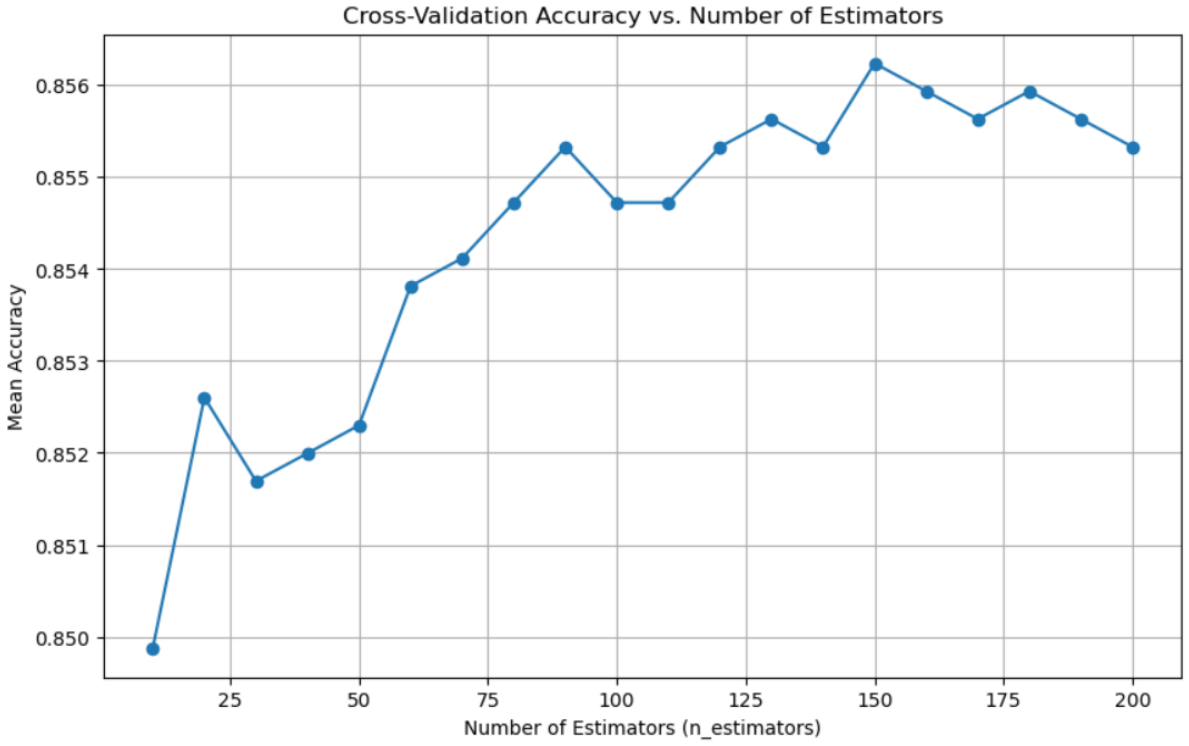
Random Forest Model Selection

1. Ensemble Learning: Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It leverages the wisdom of the crowd, which can lead to more accurate and robust predictions.
2. Flexibility: Random Forest can handle a wide range of data types, including both numerical and categorical features. This makes it suitable for health-related datasets like this with diverse variables.
3. Reduction of Overfitting: By averaging predictions from multiple trees and randomly selecting subsets of features for each tree, Random Forest helps reduce overfitting, which is critical for reliable predictions in this case.

Hyperparameter Tuning

Optimal n_estimators (number of trees in the forest) was determined through cross-validation. The mean accuracy scores were computed for different values of n_estimators to find the optimal value.

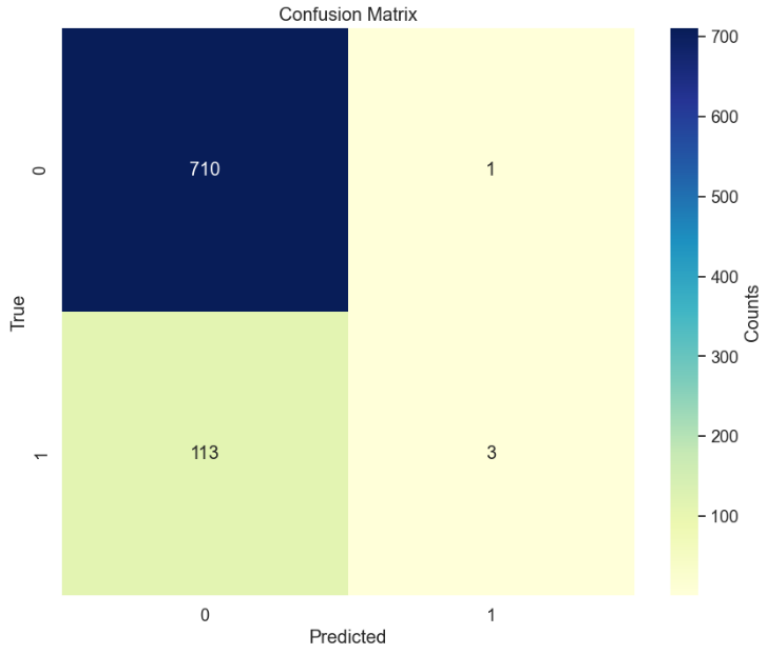
```
Optimal n_estimators: 150
```



Model Training and Evaluation

The Random Forest model with the selected number of estimators was trained using the training data. Key metrics and visualizations were employed to assess the model's performance:

1. Accuracy: The model achieved an accuracy of approximately 86.22% on the test data, indicating its capability to predict heart strokes.
2. Confusion Matrix: The confusion matrix was used to visualize the true positives, true negatives, false positives, and false negatives, offering insights into the model's performance on different classes.



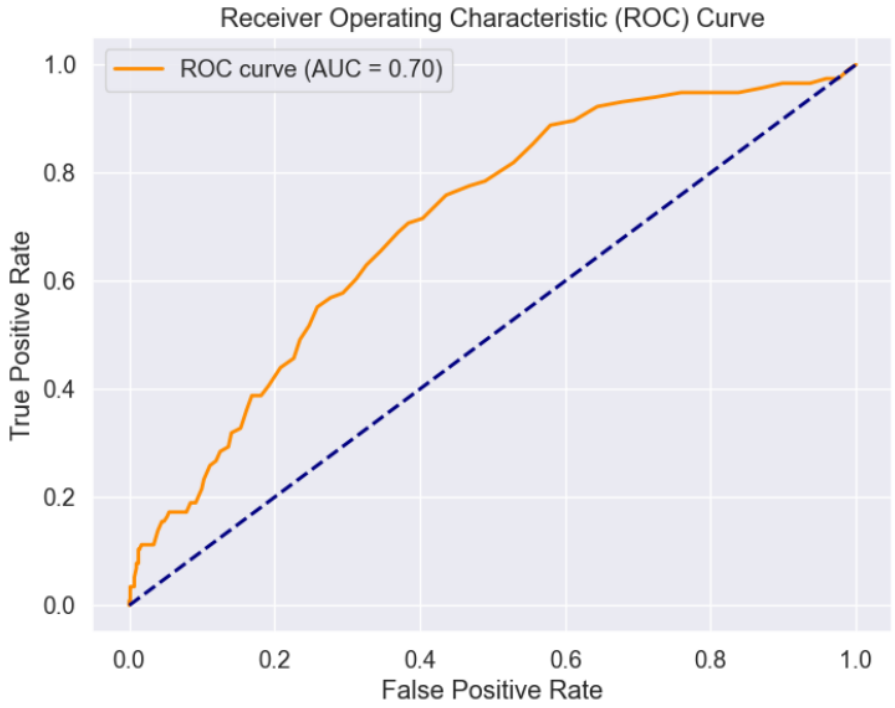
Conclusion: The confusion matrix says that the model is able to classify the patients with No stroke, but unable to classify the patients with Stroke.

3. Classification Report: Precision, recall, F1-score, and support for each class were presented in the classification report, offering a detailed analysis of the model's performance.

```
Confusion Matrix:
[[710   1]
 [113   3]]

Classification Report:
              precision    recall  f1-score   support

           0       0.86      1.00      0.93       711
           1       0.75      0.03      0.05       116

    accuracy                           0.86       827
   macro avg       0.81      0.51      0.49       827
weighted avg       0.85      0.86      0.80       827
```

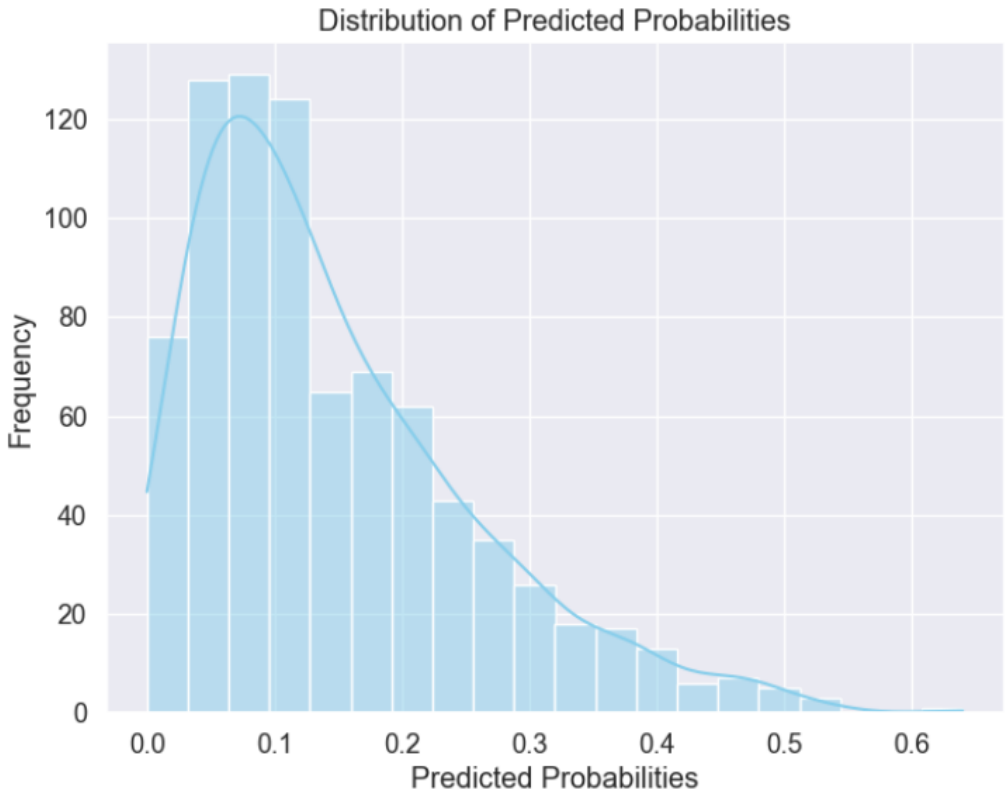
4. ROC Curve and AUC: The Receiver Operating Characteristic (ROC) curve was plotted to assess the model's ability to discriminate between positive and negative

cases, and the Area Under the Curve (AUC) was provided as a measure of the model's discrimination capability.



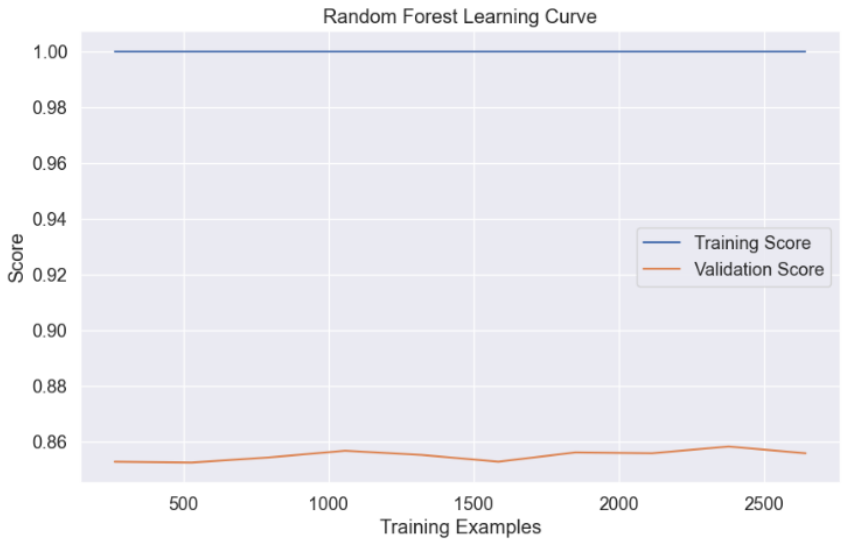
Conclusion: The ROC curve shows area as 0.7, which doesn't indicate much as it has a lot of false negatives.

5. Learning Curve Visualization: The learning curve for the Random Forest model was plotted to visualize how the accuracy of the model evolves with the increasing number of training examples



Conclusion: The above curve illustrates that the frequency is the highest when the predicted probabilities is between 0 to 0.1.

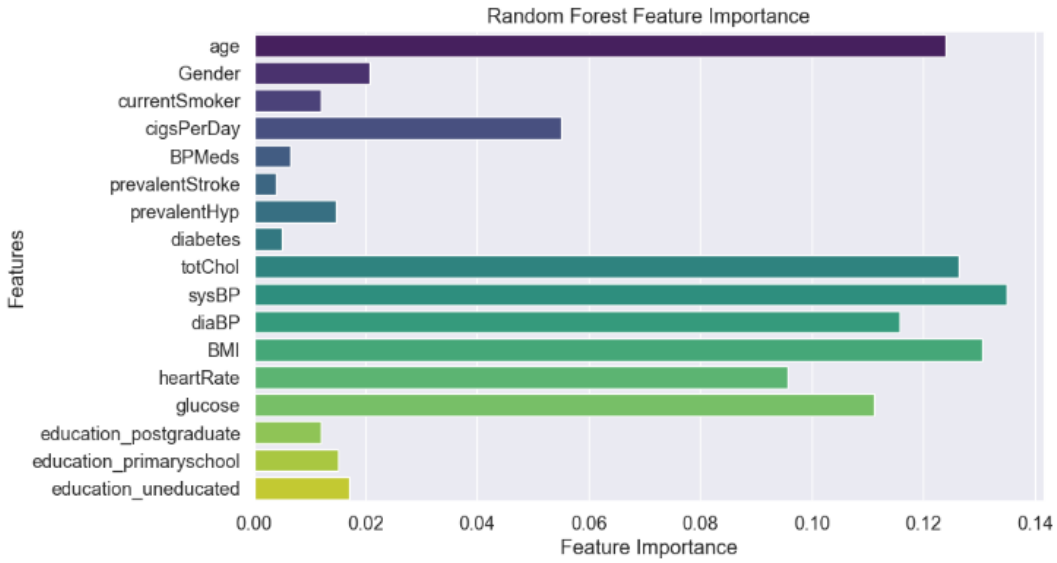
6. Distribution of Predicted Probabilities Visualization: The histogram or density plot of the predicted probabilities was plotted to show the distribution of the model's confidence levels for both positive and negative predictions.



Conclusion: Random Forest's performance evolves with additional training data, indicating potential overfitting or underfitting, as we can see the score starts to slightly increase after 2000 training examples and slightly decreases after 2300 examples indicating underfitting and overfitting before the 2000 mark and after 2300 mark.

Feature Importance

The feature importances of the Random Forest model were examined, allowing us to identify which features had the most influence on the predictions.



Conclusion: For Random Forest the most important feature was sysBP and the least important was prevalentStroke.

Conclusion

The Random Forest algorithm demonstrates promise in predicting heart strokes, achieving an accuracy of approximately 86.22%. The confusion matrix indicates challenges in classifying patients with strokes. The ROC curve, with an area of 0.7, suggests moderate discriminatory power. The precision-recall curve, with an area of 0.14, illustrates the trade-off between precision and recall. Learning curve visualization indicates potential overfitting or underfitting concerns, particularly before 2000 examples and after 2300 examples. Feature importance analysis highlights sysBP as the most influential variable.

Conclusion

In conclusion, our comprehensive analysis of various machine learning models for heart stroke prediction reveals distinct strengths and considerations for each algorithm.

1. Logistic Regression demonstrates promising accuracy but faces challenges in classifying patients with strokes, suggesting the need for refinement.
2. Support Vector Machine (SVM) exhibits potential, with moderate discriminatory power, and highlights the importance of BMI in prediction.
3. Naive Bayes, with its simplicity, achieves solid accuracy and offers an efficient choice for this task.
4. K-Means clustering successfully segments health data, providing valuable insights into potential patterns.
5. K-Nearest Neighbors (KNN) demonstrates potential accuracy but faces challenges in identifying patients with strokes, emphasizing the need for careful consideration of false negatives.
6. Artificial Neural Network (ANN) showcases promise with an improved discrimination capability, though its accuracy can benefit from further exploration.
7. Random Forest shows promise with an accuracy of 86.22%, yet it encounters difficulties in classifying patients with strokes.
8. Linear regression, while effective for regression tasks, is not suitable for the classification task of predicting heart strokes

Logistic regression, SVM, Naive Bayes, and K-Nearest Neighbors were covered in our class, providing a foundation for understanding and implementing these models. In contrast, Artificial Neural Network (ANN) and Random Forest, were not included in the class material. In summary, the selection of the most suitable model depends on specific priorities, such as interpretability, discrimination power, and the significance of avoiding false negatives in heart stroke prediction. Each model contributes valuable insights, and future work should focus on refining these algorithms to enhance their predictive capabilities in the context of cardiovascular health.

References:

1. Kaggle Link: https://www.kaggle.com/datasets/mirzahasnine/heart-disease-dataset/data
2. Understanding the Basics Of Artificial Neural Network https://www.analyticsvidhya.com/blog/2021/07/understanding-the-basics-of-artificial-neural-network-ann/
3. Understand Random Forest Algorithms With Examples https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Peer Evaluation Form for Final Group Work

CSE 487/587B

Group member 1: Gayatri Nagesh Walke
Group member 2: Siddhant Jain
Group member 3: Shriganesh Siddramappa Lokapure

Rate each groupmate on a scale of 5 on the following points, with 5 being HIGHEST and 1 being LOWEST.

Evaluation Criteria	Group member 1	Group member 2	Group member 3
How effectively did your group mate work with you?	5	5	5
Contribution in writing the report.	5	5	5
Demonstrates a cooperative and supportive attitude.	5	5	5
Contributes significantly to the success of the project.	5	5	5
TOTAL	20	20	20

Also please state the overall contribution of your teammate in percentage below, with a total of all three members accounting for 100% (33.33+33.33+33.33 ~ 100%):

Group member 1: Gayatri Nagesh Walke (33.33%)
Group member 2: Siddhant Jain (33.33%)
Group member 3: Shriganesh Siddramappa Lokapure (33.33%)