Heart Health

UNDERSTANDING HEART HEALTH AND CORONARY HEART DISEASE RISK

PROJECT 4 - GROUP 5

AAYUSHI, EMILY, MAGGIE, TYLI

Agenda

1. Importance of our topic
2. Data source & objective of the analysis
3. Data exploration/cleaning
4. Machine Learning Modules
5. Flask & HTML - Output

Importance of Heart Health

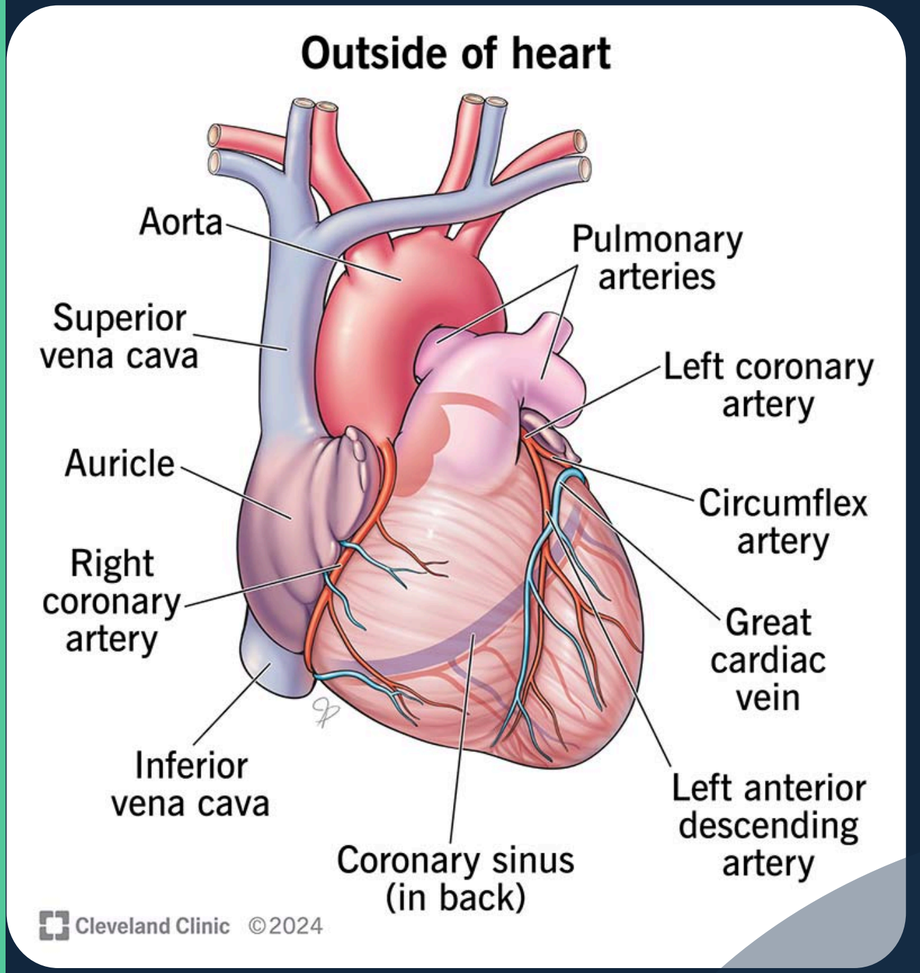
Heart Health Overview:

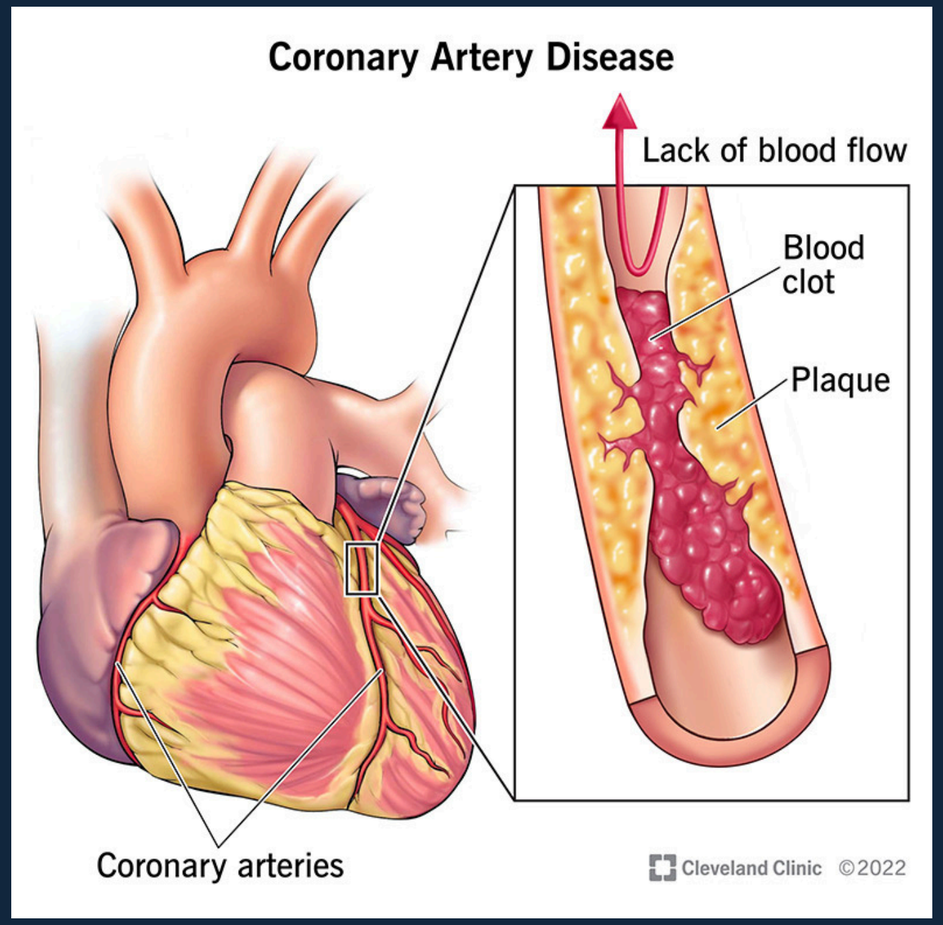
The heart is a vital organ that pumps blood throughout the body, delivering oxygen and nutrients to tissues and removing waste products.

Significance of Heart Health:

Prevalence of Heart Disease: Cardiovascular diseases are the leading cause of death globally, accounting for approximately 31% of all deaths.

Preventability: Many heart diseases are preventable through lifestyle changes, early detection, and management of risk factors such as smoking, poor diet, and lack of exercise.





Coronary Heart Disease (CHD)

- A condition where the coronary arteries become narrowed or blocked due to the buildup of atherosclerotic plaques, leading to reduced blood flow to the heart muscle.
- Symptoms: Common symptoms include chest pain (angina), shortness of breath, and heart attacks.
- CHD is the leading cause of death worldwide, responsible for millions of deaths annually.

Data Source

Heart Disease Dataset

Predictive Factors and Risk Assessment for Coronary Heart Disease(CHD)

https://www.kaggle.com/datasets/mahdifaour/heart-disease-dataset/code

Data Exploration

1

Explore & Clean Data

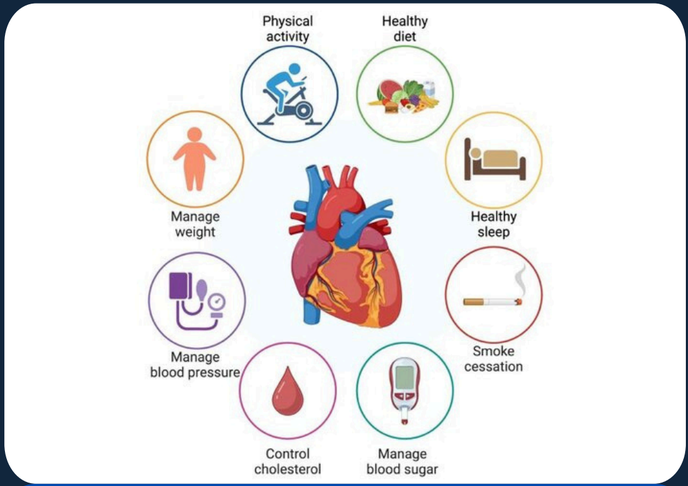
2

Create ML algothrims including Logistic Regression, Random Forest & Neural Networks

3

Adjust ML algorithms, try different Optimizations & Analyze

Objective of the Analysis



To predict the risk of coronary heart disease (CHD) based on various health and demographic factors.

To identify key factors contributing to heart disease risk.

To develop a predictive model for healthcare professionals.

Machine Learning Modules

Logistic Regression Model

Why Logistic Regression?

- Logistic regression was selected due to its simplicity, interpretability, and effectiveness in binary classification tasks.

Model Performance

- Accuracy: 66.26%
- Precision: 91% for no risk, 26% for risk
- Recall: 66% for no risk, 65% for risk
- Precision and recall scores show promising discrimination between CHD risk and no risk.
- Confusion matrix provides insight into model performance.

Interpretation

The model effectively distinguishes between individuals with and without CHD risk, albeit with moderate performance. Confusion matrix reveals 48 false positives and 259 false negatives.

Machine Learning Modules

Random Forest Model

Why Random Forest?

- Random Forest was chosen due to its ability to handle complex relationships and feature importance assessment.

Model Performance

- Accuracy: 79.89%
- Precision: 87% for no risk, 30% for risk
- Recall: 90% for no risk, 25% for risk

Feature Importance

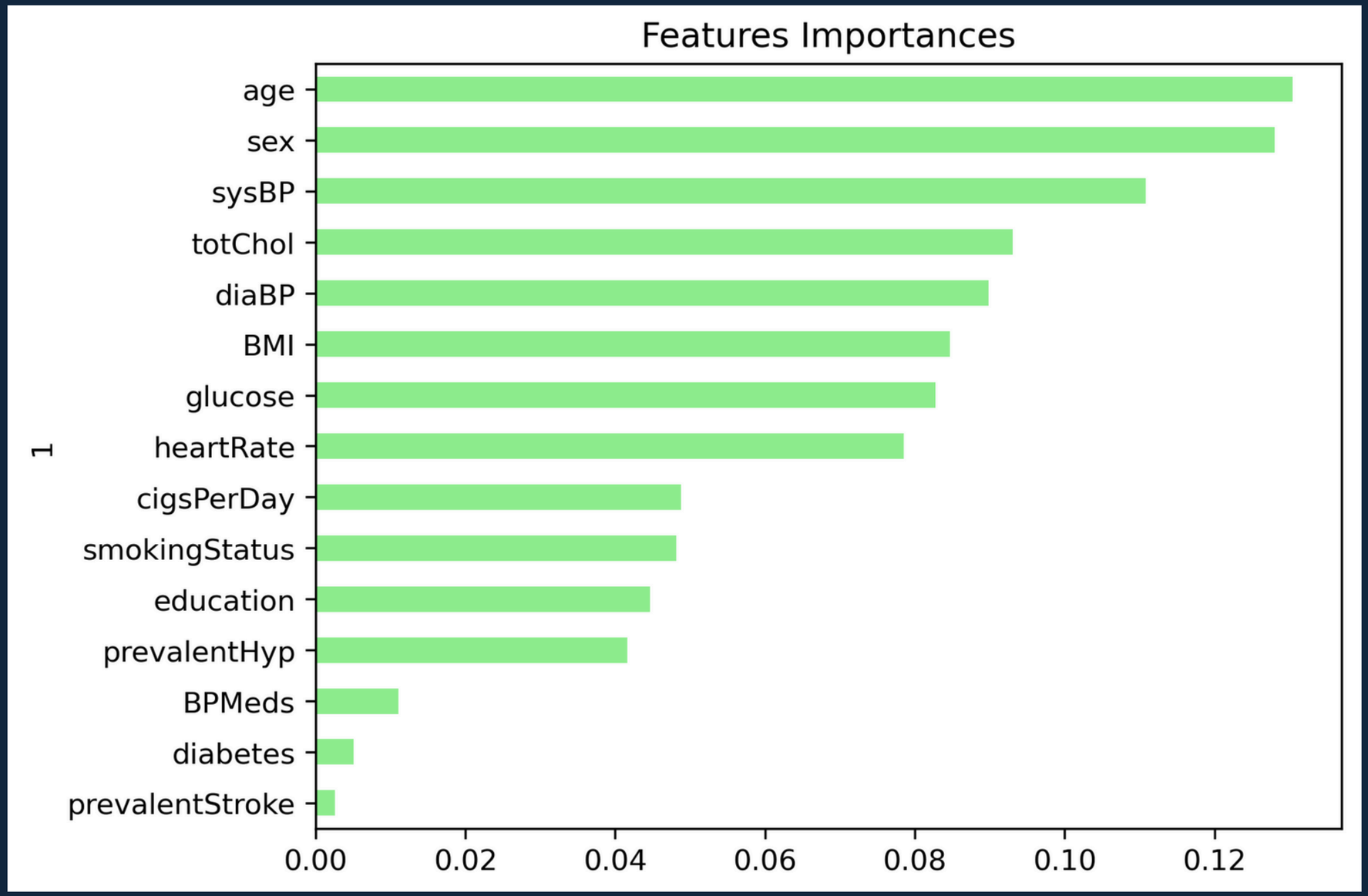
- Top features: age, sex, sysBP, totChol, diaBP

Considerations

Despite its accuracy, the model exhibits limited performance in identifying individuals at risk of CHD. Further exploration of feature engineering and parameter tuning may enhance predictive power.

Machine Learning Modules

Random Forest Model



Machine Learning Modules

Neural Network Model

Why use a Neural Network?

- A neural network was selected for its capability to capture intricate patterns in the data, especially with its ability to handle nonlinear relationships.

Model Performance

- Accuracy: 70.55%
- Recall: 35.51%

Model Architecture

- layers, each with a ReLU activation function, followed by an output layer with a sigmoid activation function.

Input Layer: Number of neurons: 64

Hidden Layers

First hidden layer: -- 32 neurons

Second hidden layer: -- 16 neurons

Third hidden layer -- 1 neuron

Output Layer: Activation function: Sigmoid

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	1,024
dense_1 (Dense)	(None, 32)	2,080
dense_2 (Dense)	(None, 16)	528
dense_3 (Dense)	(None, 1)	17

Total params: 3,649 (14.25 KB)
Trainable params: 3,649 (14.25 KB)
Non-trainable params: 0 (0.00 B)

Machine Learning Modules

Neural Network Model

Model Performance

- The neural network achieves an accuracy of 70.55% and a recall of 35.51%, indicating its effectiveness in predicting coronary heart disease (CHD) risk.

Training Duration

- The model is trained over 100 epochs, showing consistent improvement in accuracy and recall over time.

Imbalanced Data Handling:

- Synthetic Minority Over-sampling Technique (SMOTE) was employed to balance the target variable conditions in the training data, ensuring robust performance in predicting both positive and negative CHD risk cases.

Further Exploration

Hyperparameter Tuning: Fine-tuning hyperparameters such as the number of neurons, learning rate, and batch size could potentially enhance model performance.

Model Optimizations

Analysis & Conclusion

Model Optimizations

PCA ANALYSIS

PCA Exploration

We conducted Principal Component Analysis (PCA) to determine which features had low PCA loadings to inform our model optimization. PCA reduces feature dimensionality while retaining critical information.

Key Findings:

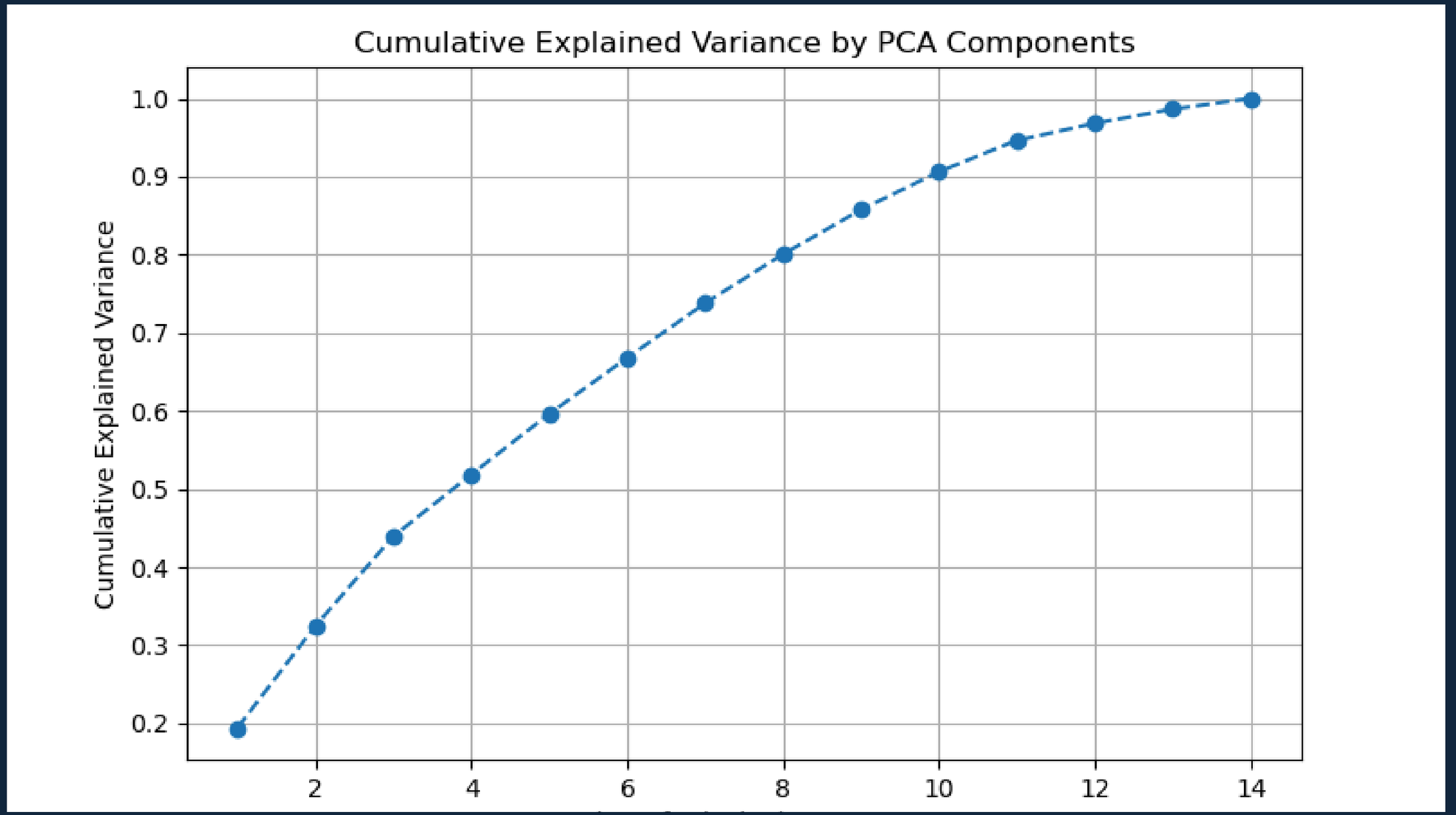
- **Explained Variance:** PC1 explains 19.39% of variance, with the top 5 components capturing 59.61%.
- **Feature Loadings**: Age exhibits strong loadings across multiple components, while Diabetes and glucose contribute notably to specific components.
- **Cumulative Variance:** About 95% of variance is explained by approximately 10 principal components.

Implications:

- **Dimensionality Reduction:** PCA mitigates dimensionality issues while preserving information.
- **Feature Importance:** Loadings help identify critical features.
- **Model Interpretability:** Reduced dimensionality aids model interpretability.

Model Optimizations

PCA Analysis



Model Optimizations

Random Forest Model – Optimization # 1

Objective

- To optimize the random forest model, we dropped features with low PCA loadings (sex, BPMeds, prevalentStroke, diabetes).

STEPS:

- Dropped features with low PCA loadings.
- Split the dataset into training and testing sets, Applied Synthetic Minority Over-sampling Technique (SMOTE) to balance the target variable in the training data.
- Instantiated and trained the Random Forest model using resampled and scaled data.
- Predictions were made and evaluated using the testing data.
- Confusion matrix, accuracy, and classification report were generated to evaluate model performance.

RESULTS:

Accuracy: 78.9%

Precision: 87% for class 0 (no risk), 29% for class 1 (at risk)

Recall: 88% for class 0 (no risk), 27% for class 1 (at risk)

F1-score: 0.88 for class 0 (no risk), 0.28 for class 1 (at risk)

Model Optimizations

Random Forest Model - Optimization # 2

Objective

- Binary features with high correlation to other variables (smokingStatus, prevalentHyp, and diabetes) were dropped.

STEPS:

- Dropped features noted above
- Split the dataset into training and testing sets, Applied Synthetic Minority Over-sampling Technique (SMOTE) to balance the target variable in the training data.
- Instantiated and trained the Random Forest model using resampled and scaled data.
- Predictions were made and evaluated using the testing data.
- Confusion matrix, accuracy, and classification report were generated to evaluate model performance.

RESULTS:

Accuracy: 80.22%

Precision: 87% for class 31% for class 1

Recall: 90% for class 0, 24% for class 1

F1-score: 0.89 for class 0, 0.27 for class 1

Model Optimizations

Random Forest Model - Optimization # 4

Objective

- To optimize the Random Forest model by combining systolic and diastolic blood pressure mathematically to produce Mean Arterial Pressure (MAP). This was an attempt to reduce the potential multicollinearity of these variables.

STEPS:

- MAP was calculated from systolic and diastolic blood pressure, and systolic and diastolic BP were dropped from the dataset.
- Split the dataset into training and testing sets, applied Synthetic Minority Over-sampling Technique (SMOTE) to balance the target variable in the training data.
- The Random Forest classifier with 500 estimators was instantiated and trained using the resampled and scaled data.
- Confusion matrix, accuracy score, and classification report were generated to assess the model's performance.
- Feature importances were calculated to understand the relative importance of different features in predicting the target variable.

Model Optimizations

Random Forest Model - Optimization # 4

Feature Importances:

- MAP, age, and sex were the top three features by importance.

RESULTS:

Accuracy: 80.99%

Precision: 87% for class 0, 33% for class 1

Recall: 91% for class 0, 24% for class 1

F1-score: 0.89 for class 0, 0.28 for class 1

Conclusion:

Replacing systolic and diastolic blood pressure with Mean Arterial Pressure (MAP) did not notably improve the accuracy and overall performance of the Random Forest model.

Model Optimizations

Neural Network Optimization

Objective

- Optimize a neural network model for binary classification on cardiovascular disease prediction while ensuring high accuracy and recall.

STEPS:

- Split the dataset into training and testing sets
- Constructed a neural network architecture with multiple hidden layers and activation functions such as ReLU and Sigmoid.
- Evaluated the trained model using the testing data to assess its performance.
- Tuned hyperparameters like the number of hidden layers, units per layer, and learning rate to optimize model performance.

Model Optimizations

Neural Network Model

RESULTS

- Accuracy: 72.75%
- Loss: 1.57
- Recall: 24.64%

CONCLUSION:

Overall, the neural network achieved relatively high accuracy but with a lower recall rate for class 1, indicating a potential imbalance in the model's ability to correctly predict positive cases of cardiovascular disease. Further fine-tuning and exploration of different architectures or techniques like class weighting may improve the model's performance, particularly in capturing positive cases.

Best Performing Model:

Optimization Attempt 3: To enhance the Random Forest (RF) model's performance, features with low importances, namely diabetes, BPMeds, and prevalentStroke, were dropped.

Steps:

1. Features with low importances were identified and removed from the dataset.
2. The data was split into training and testing sets, with the training set balanced using Synthetic Minority Over-sampling Technique (SMOTE).
3. Feature scaling was applied to ensure uniformity across variables.
4. The RF model was instantiated and trained using the resampled and scaled training data.
5. Predictions were and evaluated using the testing data, and model performance was evaluated using metrics like accuracy, precision, recall, and F1-score.

Best Performing Model:

Optimization Attempt 3: To enhance the Random Forest (RF) model's performance, features with low importances, namely diabetes, BPMeds, and prevalentStroke, were dropped.

Results:

- **Accuracy: 81.98%**
- **Precision: 87% for no risk, 37% for at risk**
- **Recall: 92% for class no risk, 26% for at risk**

Conclusion:

- The RF model achieved a high accuracy rate with fewer features than the baseline model.
- Although the recall for the minority class (class 1) improved compared to the baseline, it remains relatively low, indicating a potential imbalance issue.
- Further adjustments, such as threshold modification, were explored to improve recall without significantly sacrificing overall accuracy. However, there is still room for improvement in capturing positive cases of cardiovascular disease.

Best Performing Model:

Why is this our best model?

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1 Score (Class 0)	F1 Score (Class 1)
Optimization 1	78.9%	87%	29%	88%	27%	88%	28%
Optimization 2	80.22%	87%	31%	90%	27%	89%	27%
Optimization 3 (Best Model)	81.98%	87%	37%	92%	26%	89%	27%
Optimization 4	80.99%	87%	33%	91%	24%	89%	28%
Baseline Model	79.8%	87%	30%	90%	25%	88%	27%

RESULTS FOR OPTIMIZATION 3:

Results:

- Accuracy: 81.98%.
- Precision: 87% for class 0 (no risk), 37% for class 1 (risk)
- Recall: 92% for class 0(no risk) , 26% for class 1 (risk).
- F1 Score: 89% for class 0(no risk), 27% for class 1 (risk).

VS

RESULTS FOR ORIGINAL RANDOM FOREST:

-> Accuracy: 79.89%

-> Precision: 87% for class 0, 25% for class 1

-> Recall: 90% for class 0, 25% for class 1

-> F1 Score: 88% for class 0, 27% for class 1

Best Performing Model:

Why is this our best model?

Accuracy: Optimization 3 exhibits the highest accuracy among all attempts, indicating superior overall predictive performance.

Precision: While precision for class 1 slightly varies across optimizations, optimization 3 showcases the highest precision for class 1 (at risk) which came in at 37%, implying fewer false positives.

Recall: Although optimization 3 demonstrates a lower recall for class 1 compared to some other attempts, it maintains the highest recall for class 0 (no risk group) at 92%, crucial for capturing true positives in cardiovascular disease prediction.

Balance: Optimization 3 achieves a commendable balance between precision and recall, ensuring both the identification of positive cases (recall) and the accuracy of those identifications (precision).

Overall Performance: Considering accuracy, precision, and recall, optimization 3 emerges as the most well-rounded and effective model for cardiovascular disease prediction.

Considerations:

- While we were able to produce a model with high accuracy, our recall scores for the minority class (those at risk of developing CHD) were consistently low
- Our best model was only able to correctly predict those at risk of developing CHD 26% of the time, leading to a high false negative rate
    - This has significant implications in health data, as misidentifying those at risk of developing a disease is a serious concern
- Although attempts were made to improve recall scores for the minority class (such as lowering the classification threshold), these attempts negatively affected overall model accuracy
- In order to address this issue, more data must be collected, specifically for those in the at-risk population

Recommendations:

1. Adopt Random Forest Optimization 3 for its highest accuracy (81.98%) and improved minority class precision (37%).

2. Continue excluding features with low importance (e.g., diabetes, BPMeds, prevalentStroke) to streamline models. Continue trying new optimizations until the outcome is met.

3. Adjust decision thresholds to enhance recall for the minority class, especially in critical scenarios.

4. Use SMOTE for handling class imbalance and standardize feature scaling for improved model performance.

5. Regularly monitor performance with fresh data, using a combination of metrics for a comprehensive view.

6. Future Enhancements:
Explore additional hyperparameter tuning and advanced optimization techniques for further improvement. Explore other binning techniques for columns like age, BMI, Total Cholesterol etc.

Conclusion

In our CHD risk prediction analysis, Random Forest Optimization 3 stands out with the highest accuracy of 81.98% and improved precision for the minority class. By removing less significant features, balancing the dataset, and adjusting decision thresholds, we achieved a robust model. Continued evaluation and fine-tuning will ensure its effectiveness in real-world applications, enhancing our predictive capabilities and contributing to better healthcare outcomes.

Consult us for Questions

Thank you..!!