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

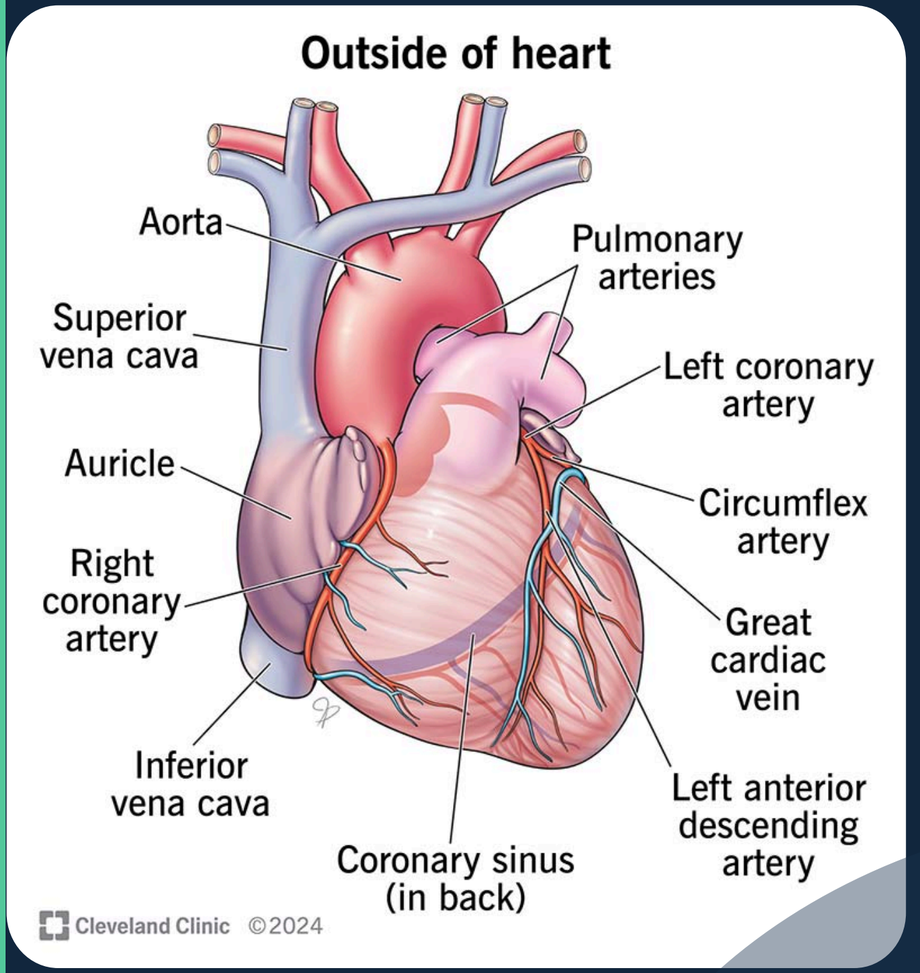
Heart Health Overview:

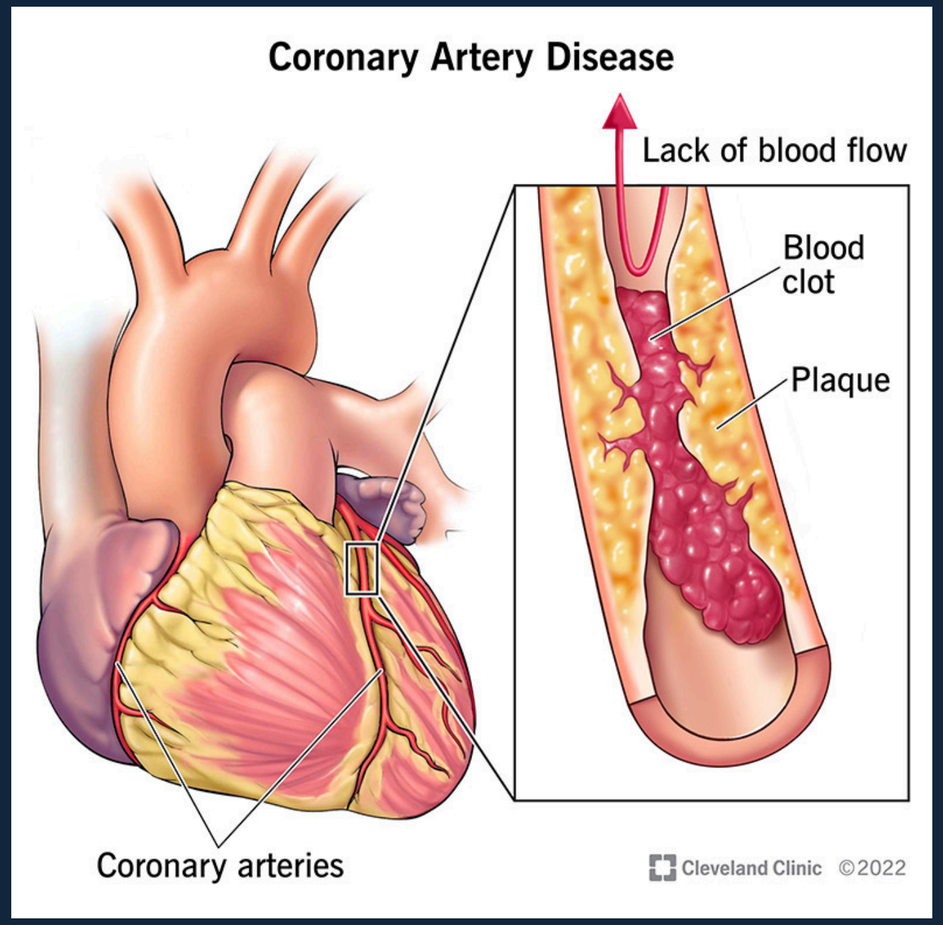
The heart is a vital organ that pumps blood throughout the body, delivering oxygen and nutrients to tissues and removing waste products.

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

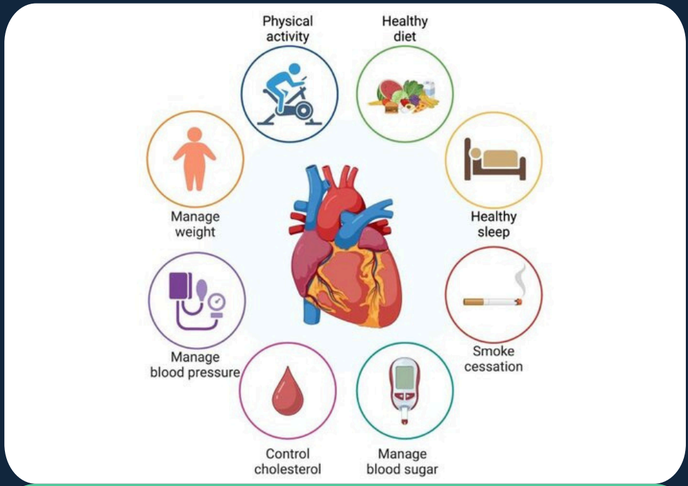
2

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

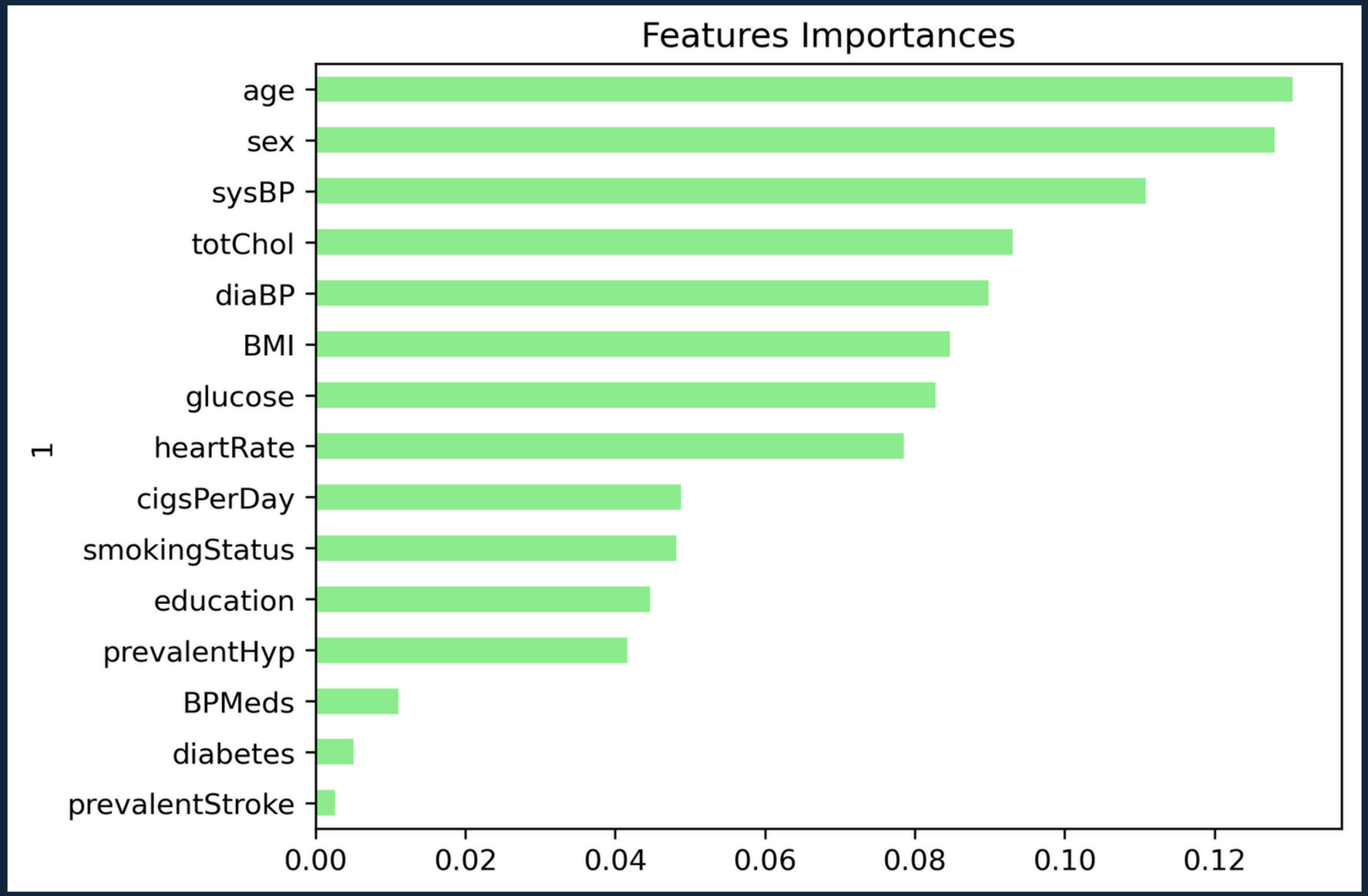
- Top features: age, sex, sysBP, totChol, diaBP

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 reduce feature dimensionality while retaining critical information.

Key Findings:

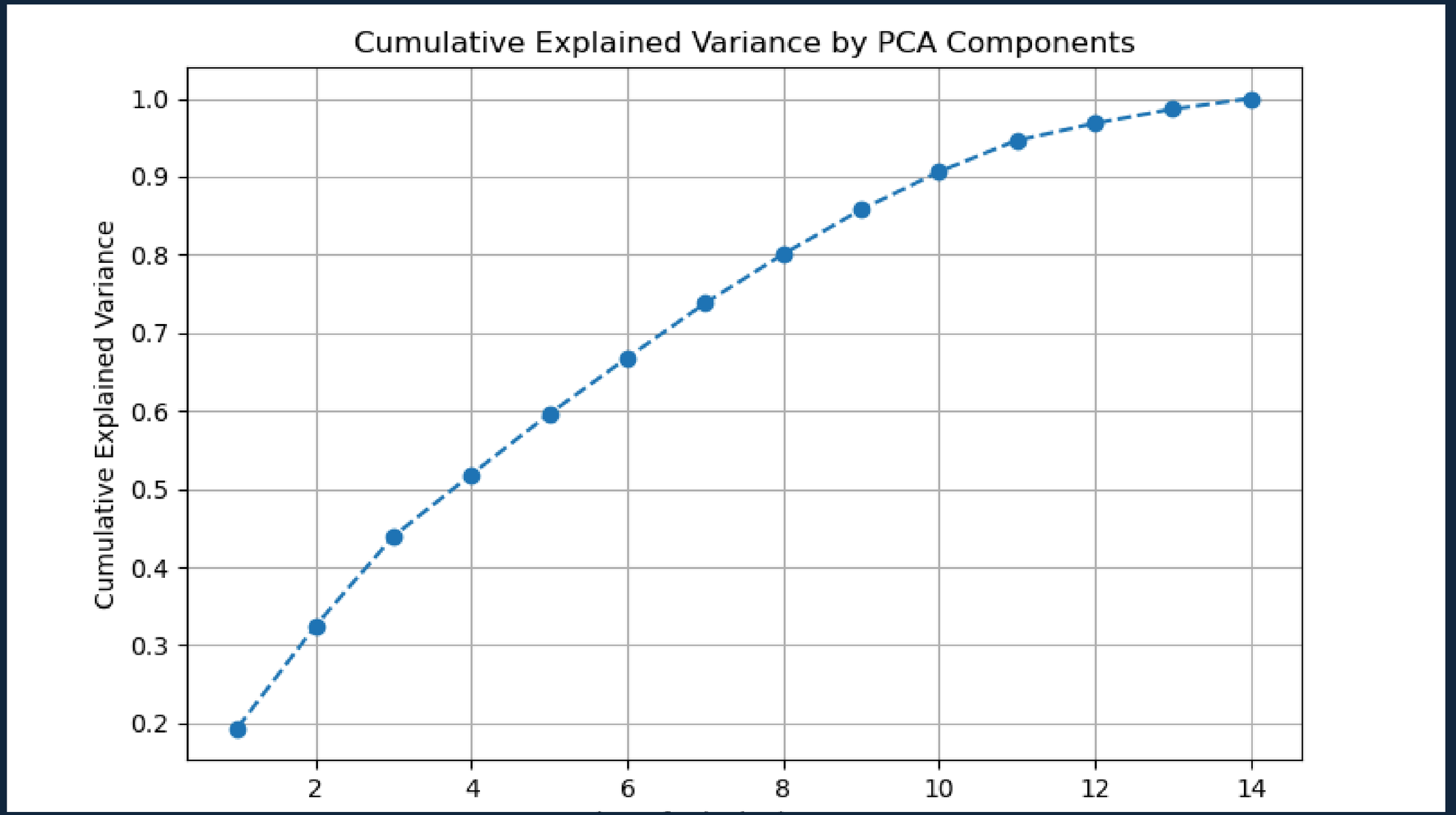
- **Explained Variance:** PC1 explains 19.39% of variance, with the top 5 components capturing 59.61%.
- **Feature Loadings**: Age exhibits strong loadings across multiple components, while Diabetes and glucose contribute notably to specific components.
- **Cumulative Variance:** About 95% of variance 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 on 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 on 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 # 3

Objective

- To optimize the Random Forest model by replacing systolic and diastolic blood pressure with Mean Arterial Pressure (MAP).

STEPS:

- MAP was used as a substitute for systolic and diastolic blood pressure.
- 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.
- 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 # 3

Feature Importances:

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

Cross-validation Results:

- Mean cross-validation accuracy: 84.91%
- Standard deviation of cross-validation accuracy: 0.31%

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

Random Forest Model - Results Comparison

Model Optimizations

Neural Network Model

Model Optimizations

Neural Network Model

Model Optimizations

Neural Networks

Key Findings:

Random Forest Model

1

Explore & Clean Data

2

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

3

Adjust ML algorithms, try different Optimizations & Analyze

Key Findings:

Neural Networks

1

Explore & Clean Data

2

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

3

Adjust ML algorithms, try different Optimizations & Analyze

Recommendations:

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

Conclusion

Heart Disease Dataset

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

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

Consult us for Questions

Thank you..!!