PREDICTING HEART DISEASE

PRESENTATION BY

TENDAI JONHASI

2024

EXECUTIVE SUMMARY

Introduction
The goal of this project is to develop a predictive model that can accurately classify individuals as either having or not having heart disease. By analyzing a comprehensive set of patient attributes, including demographic, clinical, and lifestyle factors, the model aims to identify patterns and features indicative of the presence or absence of heart disease. This will support medical professionals in making informed decisions regarding patient care and prevention strategies.

Objectives
1. Perform exploratory data analysis (EDA) to understand the data and relationships between variables.
2. Preprocess the data, including handling missing values and scaling features.
3. Build multiple predictive models to classify individuals as having or not having heart disease.
4. Evaluate the models using appropriate metrics such as accuracy, precision, recall, and AUC-ROC.
5. Select the best-performing model and interpret its results.

Multiple models were built, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and Convolutional Neural Networks (CNNs).

Model Performance Overview
Logistic Regression -- Accuracy: 71,8%
- Key Strengths: High recall for non-heart disease cases (76%)
- Weaknesses: Lower recall for heart disease cases (68%)
- Age, cholesterol, blood pressure, and weight are significant predictors of heart disease.

Random Forest -- Accuracy: 72.5%
- Key Strengths: Balanced recall for both non-heart disease (75%) and heart disease cases (70%), offering a robust model for identifying at-risk patients.
- Age, systolic blood pressure, and weight are the most influential factors.

Support Vector Machine (SVM) -- Accuracy: 72.3%
- Key Strengths: Effective in high-dimensional spaces and good at identifying non-heart disease cases.
- Weaknesses: Lower recall for heart disease cases (68%), which could lead to underdiagnosis.
- Blood pressure, cholesterol, and age are the most significant factors in the model.

Convolutional Neural Network -- Accuracy: 76%
- Key Strengths: Best performance among all models with the highest accuracy and AUC-ROC, indicating strong predictive power.
- Weaknesses: While effective overall, CNNs are less interpretable compared to traditional models.
- Unlike other models, CNNs automatically learn to extract features, which may include the same key factors (age, blood pressure, weight) but are less transparent.

Recommendations
1. Suggested Model: Convolutional Neural Network (CNN)
   The CNN outperforms the other models, achieving the highest accuracy (76%) and AUC-ROC score (0.80).It provides a strong predictive capability for distinguishing between individuals with and without heart disease.

2. Second-Best Model: Random Forest
   The Random Forest model offers a good balance between recall for both classes, making it a reliable choice when interpretability and feature importance are critical.

3. Feature Importance & Coefficient Impact: Logistic Regression and Random Forest models are recommended for understanding feature importance. The insights from these models can guide targeted interventions and personalized care, focusing on managing age, blood pressure, cholesterol levels, and weight to mitigate the risk of heart disease.

In conclusion, the models showed that individuals who are older, have high blood pressure, high cholesterol, are overweight or obese, and have unhealthy lifestyle habits like smoking, excessive alcohol consumption, and low physical activity are at the highest risk of developing heart disease. These factors should be closely monitored and managed to reduce the risk of heart disease, and targeted interventions should be considered for those at high risk.

AGENDA SLIDE

EXPLORATORY DATA ANALYSIS

LOGISTIC REGRESSION MODEL

RANDOM FOREST MODEL

SUPPORT VECTOR MACHINE MODEL

CONVOLUTIONAL NEURAL NETWORKS

RECOMMENDATION

EXPLORATORY DATA ANALYSIS (EDA)

1

Data Preparation and Preprocessing Overview

In the data preparation and preprocessing phase, we focused on transforming raw data into a format suitable for machine learning models. This process ensures that our models perform optimally by eliminating noise, scaling features, and encoding categorical variables appropriately.

1. Calculating Body Mass Index (BMI) for each individual in the dataset
    - Conversion of Height from centimeters to meters
    - BMI is calculated using the formula $BMI = \frac{weight(kg)}{height(m)^2}$
2. Categorizing BMI into simplified categories: Underweight, Normal weight, Overweight, and Obesity
3. Converting Age from Days to Years
4. Cleaning Up the DataFrame by removing columns like height_m, weight, height, and identifiers (date, id)
5. Checking for Missing Values (there are none)
6. Encoding categorical variables into a numerical format that can be used by machine learning algorithms. One-Hot Encoding was used to convert categories into dummy/indicator (0/1) columns.
7. Feature Scaling to ensure the features have similar ranges, have a mean of 0 and a standard deviation of 1, which is important for many machine learning algorithms.

2

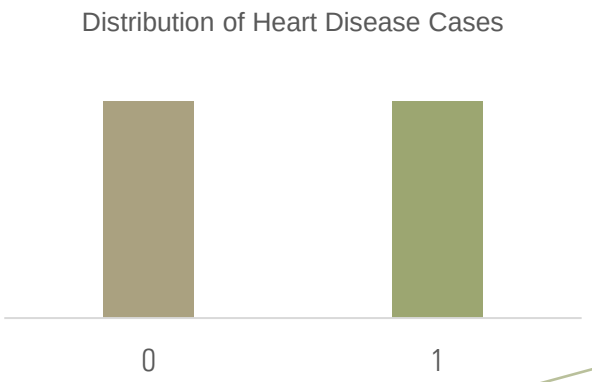
Binary Classification

The target variable 'disease' indicates two classes:

1: Will have heart disease.

0: Does not have heart disease.

Binary classification problems are flexible and straightforward to manipulate, making them suitable for this task. Multiple models that suit binary classification were built, including Logistic Regression, Random Forest, SVM, and Convolutional Neural Networks (CNNs).



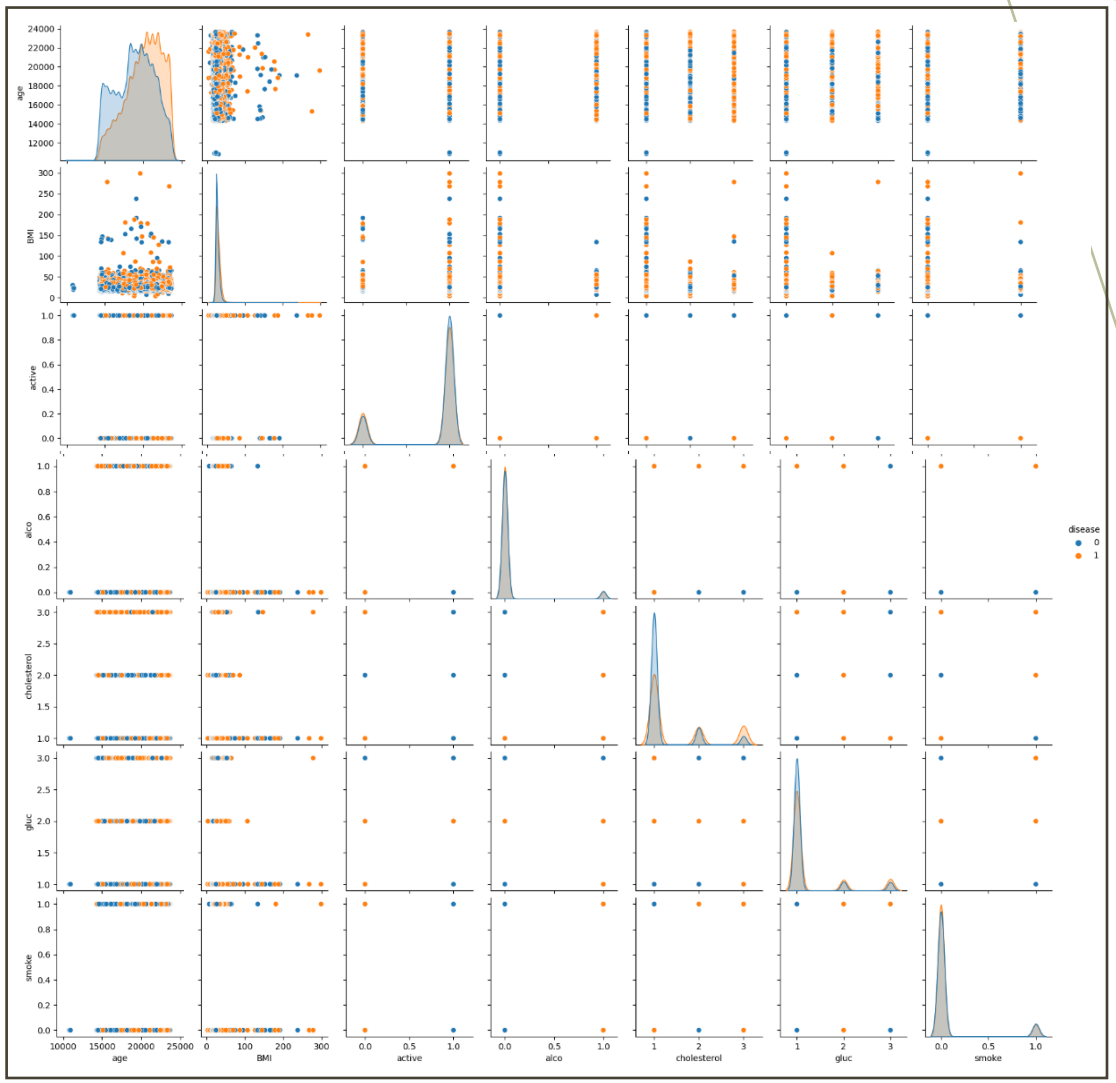
EXPLORATORY DATA ANALYSIS (EDA)

3

Pair Plots

We will explore the dataset to understand the distribution of the data, relationships between features, and key insights that could help in model building.

The scatterplot matrix shows the relationships between several features in the dataset, with points color-coded by the presence (1) or absence (0) of heart disease.



Key Observations:

Each diagonal element shows the distribution of the individual features, separated by the disease variable. For example, in the top-left plot, the age distribution shows some differences between patients with and without heart disease.

Off-diagonal elements are scatter plots showing the relationship between pairs of features, with points colored by the disease label. These plots help visualize how the features interact with each other and whether any patterns emerge that separate those with heart disease from those without.

Age and Cholesterol: There seems to be a slight differentiation in age between those with heart disease (in orange) and those without (in blue). Cholesterol levels appear to have a noticeable difference between patients with and without heart disease, especially as cholesterol increases.

Active, Alco, Gluc, and Smoke: These binary features show clustering where patients with heart disease might tend to be more frequent in certain categories, though the spread is not very pronounced . Further machine learning modeling will confirm if these variables are significant.

LOGISTIC REGRESSION

Interpretation

Age, Gender, Cholesterol, Blood Pressure (both systolic and diastolic), and Obesity are significant positive predictors of heart disease. Being active, consuming alcohol, higher glucose levels, and smoking are significant negative predictors.

Glucose level has a negative coefficient, indicating that higher glucose levels are associated with a lower probability of heart disease in this model, which could be counterintuitive and might require further exploration.

Insignificant Features (p-value > 0.05): Gender Country (Indonesia, Malaysia, Singapore) Occupation (Most occupation categories, except for Teacher).

These results suggest that while certain demographic features like country and occupation may not be significant predictors, physiological metrics like age, gender, obesity, blood pressure, and cholesterol are highly significant in predicting heart disease.

Key Influences

Age, Gender, Cholesterol, Blood Pressure (both systolic and diastolic), and Obesity are significant positive predictors of heart disease. Being active, consuming alcohol, higher glucose levels, and smoking are significant negative predictors.

Logit Regression Results

Dep. Variable:	disease	No. Observations:	49000
Model:	Logit	Df Residuals:	48974
Method:	MLE	Df Model:	25
Date:	Thu, 22 Aug 2024	Pseudo R-squ.:	0.1343
Time:	11:56:11	Log-Likelihood:	-29404.
converged:	True	LL-Null:	-33964.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0929	0.010	9.138	0.000	0.073	0.113
active	-0.0756	0.010	-7.625	0.000	-0.095	-0.056
age	0.7062	0.233	3.031	0.002	0.250	1.163
alco	-0.0445	0.011	-4.190	0.000	-0.065	-0.024
ap_hi	5.4252	0.107	50.553	0.000	5.215	5.636
ap_lo	0.0499	0.013	3.719	0.000	0.024	0.076
cholesterol	0.3643	0.012	30.145	0.000	0.341	0.388
gender	0.0361	0.011	3.389	0.001	0.015	0.057
gluc	-0.0622	0.012	-5.366	0.000	-0.085	-0.039
smoke	-0.0351	0.011	-3.161	0.002	-0.057	-0.013
disease	0.0755	0.019	3.961	0.000	0.038	0.113
BMI	-0.3409	0.233	-1.464	0.143	-0.797	0.116
Age_in_Years	0.0025	0.012	0.206	0.837	-0.021	0.026
country_Indonesia	-0.0005	0.012	-0.043	0.966	-0.024	0.023
country_Malaysia	-0.0062	0.012	-0.510	0.610	-0.030	0.018
country_Singapore	-0.0149	0.013	-1.133	0.257	-0.041	0.011
occupation_Architect	-0.0299	0.013	-2.259	0.024	-0.056	-0.004
occupation_Chef	-0.0307	0.013	-2.329	0.020	-0.056	-0.005
occupation_Doctor	-0.0306	0.013	-2.319	0.020	-0.056	-0.005
occupation_Engineer	-0.0180	0.013	-1.363	0.173	-0.044	0.008
occupation_Lawyer	-0.0277	0.013	-2.087	0.037	-0.054	-0.002
occupation_Nurse	-0.0163	0.013	-1.232	0.218	-0.042	0.010
occupation_Others	-0.0136	0.013	-1.033	0.302	-0.039	0.012
occupation_Teacher	0.1465	0.019	7.584	0.000	0.109	0.184
BMI_Category_Obesity	0.0879	0.013	6.953	0.000	0.063	0.113
BMI_Category_Overweight	-0.0189	0.011	-1.752	0.080	-0.040	0.002

LOGISTIC REGRESSION

Model Evaluation

The Logistic Regression model shows reasonable performance in predicting heart disease with an accuracy of about 71.8%. Strengths: The model is particularly good at identifying non-heart disease cases (high recall for class 0). 76% but falls short for identifying heart disease cases (lower recall for class 1), at 68% which is critical in medical diagnostics where false negatives are costly.

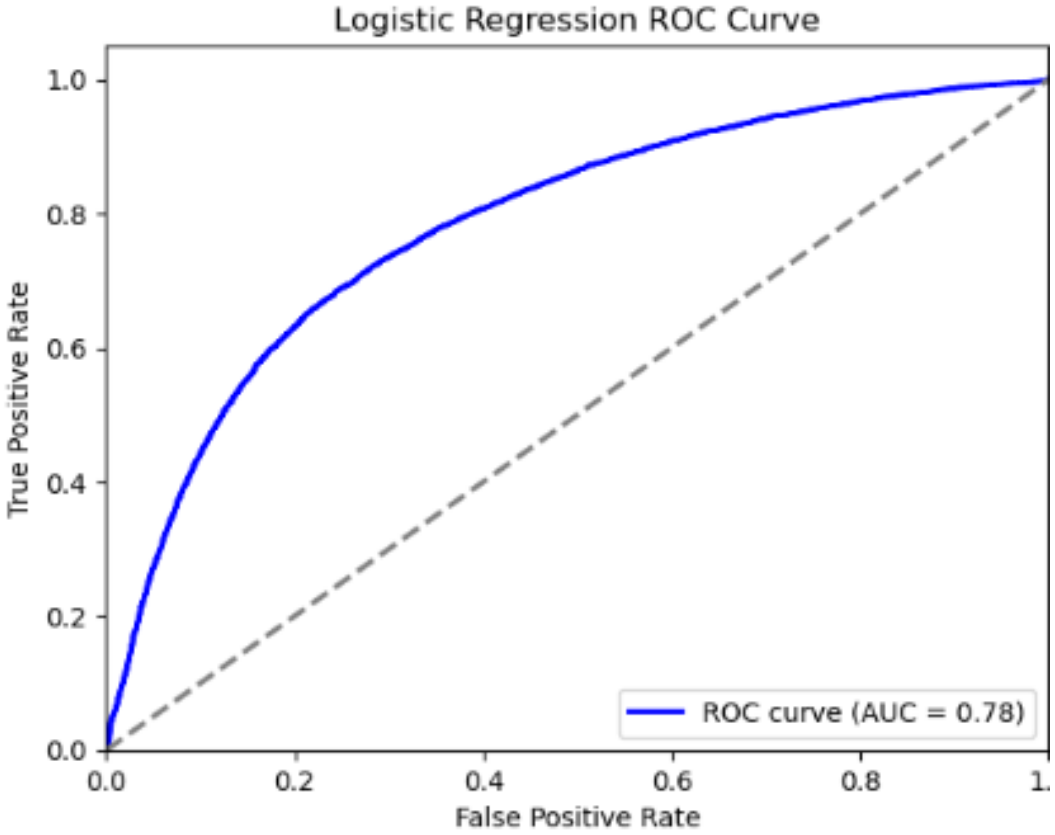
```
Confusion Matrix:
[[7943 2518]
 [3391 7148]]

Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.76      0.73     10461
           1       0.74      0.68      0.71     10539

    accuracy                           0.72     21000
   macro avg       0.72      0.72      0.72     21000
weighted avg       0.72      0.72      0.72     21000


Accuracy Score:
0.7186190476190476
```



RANDOM FOREST MODEL

Interpretation

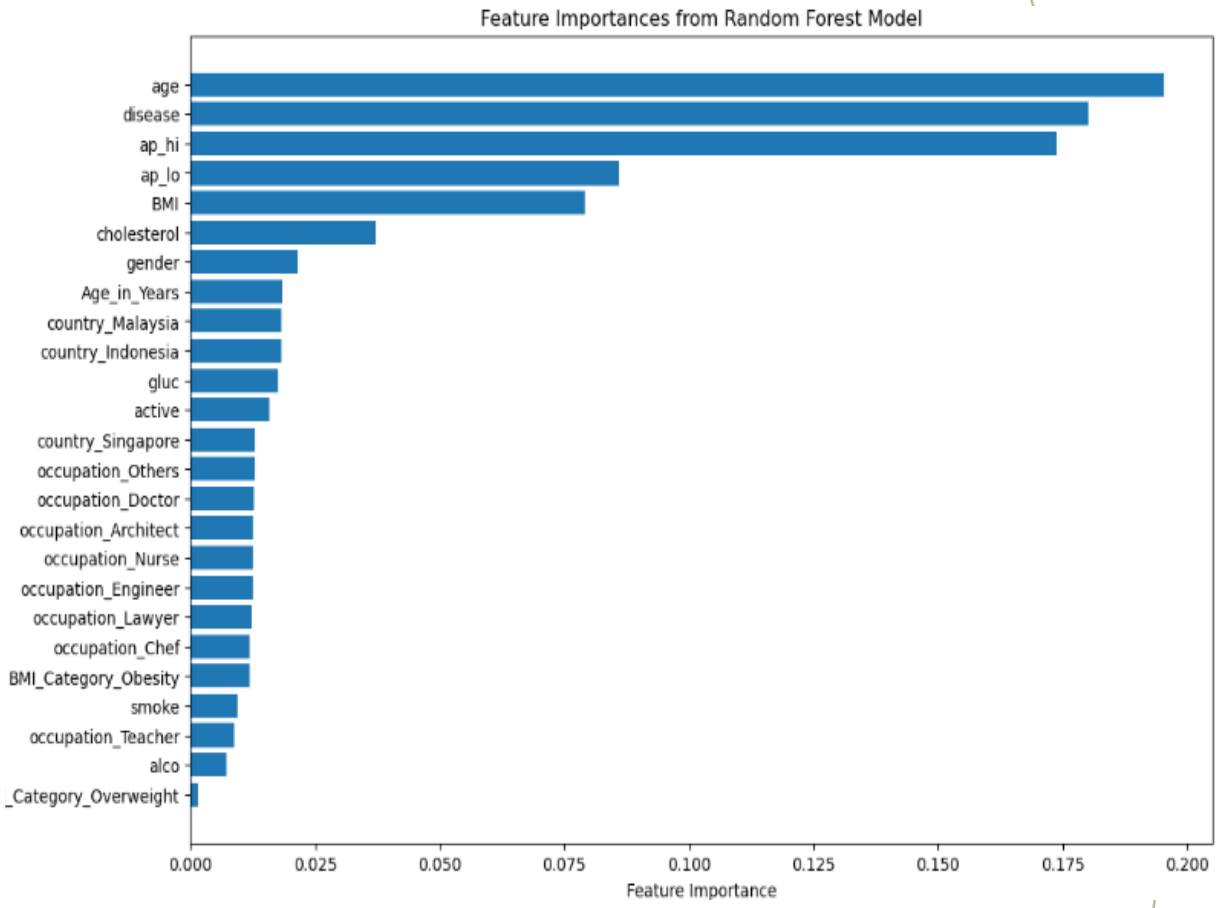
Age: This is the most important feature according to the Random Forest model. Age is often a critical factor in heart disease, with older individuals generally being at higher risk.

Systolic Blood Pressure (ap_hi): The second most important feature. High systolic blood pressure is a significant risk factor for heart disease, as it indicates how much pressure your blood is exerting against artery walls when the heart beats.

Diastolic Blood Pressure (ap_lo): This measures the pressure in your arteries when your heart rests between beats. While it's less impactful than systolic pressure, it's still important in assessing cardiovascular health.

BMI plays a crucial role as excess body weight relative to height is often associated with high risk factors like hypertension, diabetes, and cholesterol levels, which contribute to heart disease.

Cholesterol: Elevated cholesterol levels can lead to the buildup of plaques in arteries, a condition known as atherosclerosis, which increases the risk of heart disease.



Other Features

Gender, Glucose, and Occupation: These features have lower importance scores, meaning they contribute less to the model’s predictions. However, they still provide some predictive power. Gender, for example, is a known factor in heart disease risk, with men typically at higher risk at younger ages compared to women.

Key Influences

Age, Systolic Blood Pressure, Diastolic Blood Pressure, BMI and Cholesterol are the most influential factors in the model's decision-making process are significant positive predictors of heart disease.

RANDOM FOREST MODEL

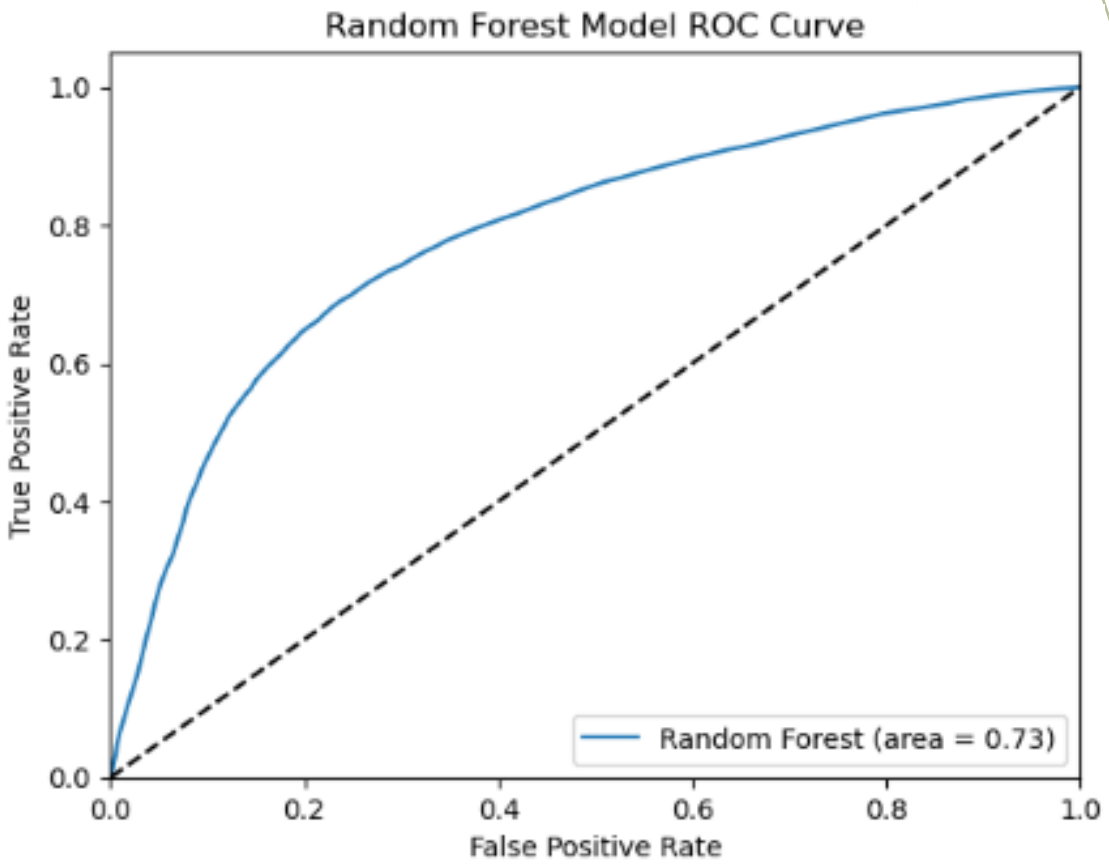
Model Evaluation

The Random Forest model shows reasonable performance in predicting heart disease with an accuracy of about 72.5%. Strengths: The model is particularly good at identifyin non-heart disease cases (high recall for class 0). 74% and for identifying heart disease cases (lower recall for class 1), this is also high at 71% which is critical in medical diagnostics where false negatives are costly.

```
Accuracy Score: 0.7255238095238096
ROC AUC Score: 0.7839073544044086
Confusion Matrix:
 [[7777 2684]
 [3080 7459]]
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.74      0.73     10461
           1       0.74      0.71      0.72     10539

    accuracy                           0.73     21000
   macro avg       0.73      0.73      0.73     21000
weighted avg       0.73      0.73      0.73     21000
```



SUPPORT VECTOR MACHINE MODEL

Interpretation

The top features influencing the prediction are:

ap_hi (**Systolic blood pressure**): This feature has the highest importance, suggesting it plays the most significant role in predicting heart disease.

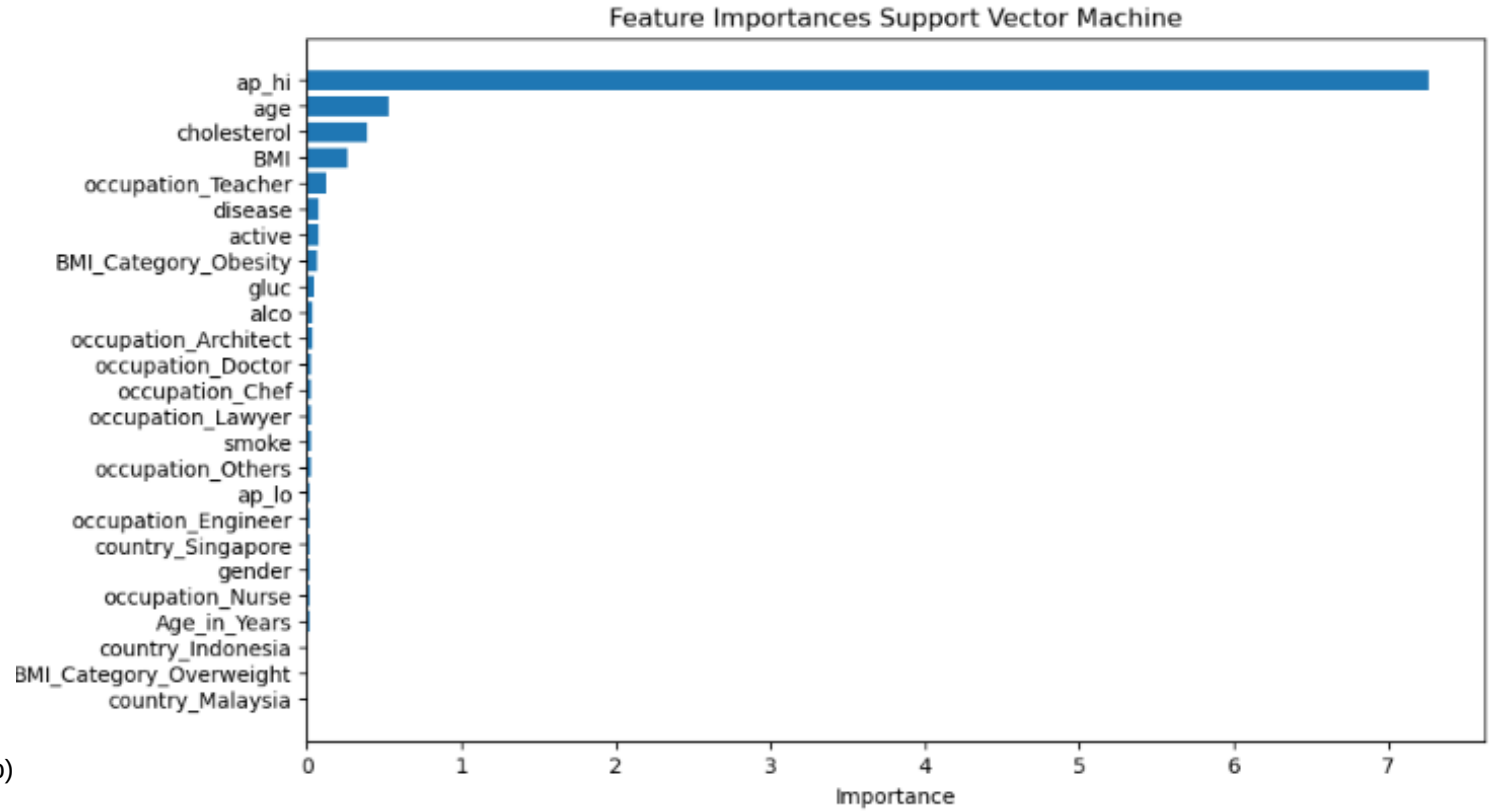
cholesterol: This is also a highly important feature.

age: Another key factor.

BMI and BMI_Category_Obesity have a noticeable impact on prediction

The occupation **Teacher** is also a highly significant feature.
While active, gluc (Glucose levels) and Diastolic Blood Pressure (ap_lo) have moderate importance. Less important features include specific occupations, alcohol consumption (alco), smoking, and gender, among others.

Country: The model might be capturing some region-specific variations in heart disease prevalence, but these features have relatively low importance.



Key Influences

Systolic Blood Pressure, Age, BMI and Cholesterol are the most influential factors in the model's decision-making process are significant positive predictors of heart disease.

SUPPORT VECTOR MACHINE MODEL

Model Evaluation

The SVM model shows reasonable performance in predicting heart disease with an accuracy of about 72.3%. Strengths: The model is particularly good at identifying non-heart disease cases (high recall for class 0) 81%. However, it struggles more with identifying heart disease cases (lower recall for class 1), which is critical in medical diagnostics where false negatives are costly.

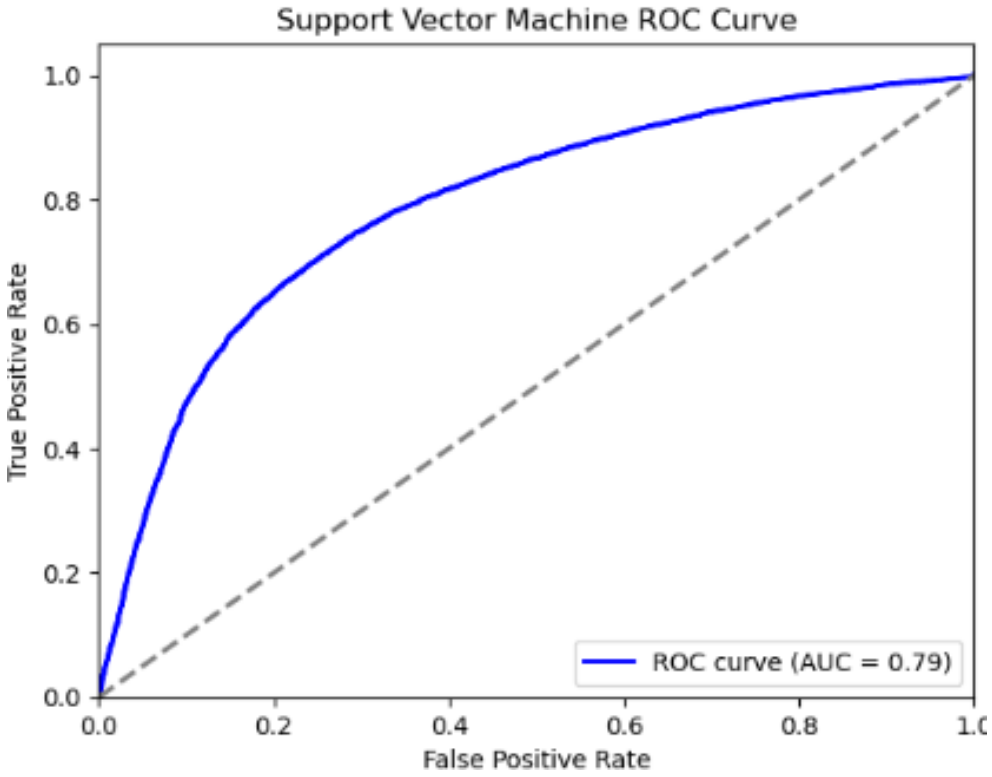
```
Confusion Matrix:
[[8441 2020]
 [3780 6759]]

Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.81      0.74     10461
           1       0.77      0.64      0.70     10539

    accuracy                           0.72     21000
   macro avg       0.73      0.72      0.72     21000
weighted avg       0.73      0.72      0.72     21000


Accuracy Score:
0.7238095238095238
```



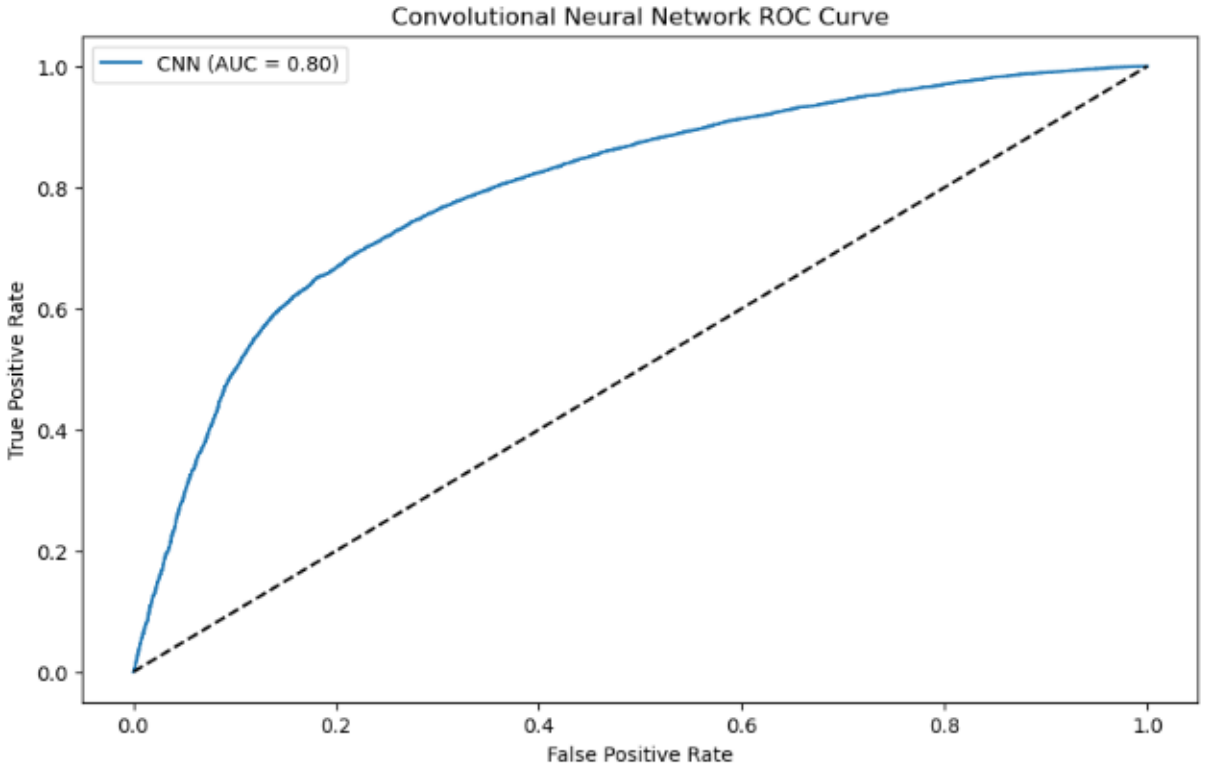
CONVOLUTIONAL NEURAL NETWORK MODEL

Model Evaluation

The CNN model slightly outperforms the previous models, with a higher accuracy of 73,5% and slightly better balance between precision, recall, and F1-scores. The ROC AUC Score of 0.799 is also a good indicator that the CNN model has strong predictive power in distinguishing between the two classes. Strengths: The model is particularly good at identifying non-heart disease cases (high recall for class 0) at 77% but slightly less effective at identifying heart disease cases (lower recall for class 1), at 70%.

Unlike the other machine learning models, CNNs are a type of deep learning model typically used for image data and tasks involving spatial hierarchies. They automatically learn to extract features through convolutional layers and filters, rather than relying on manually defined features.

For better interpretability, consider combining CNN with other techniques or models that allow feature importance extraction.




```
CNN Accuracy Score: 0.7350952380952381
CNN ROC AUC Score: 0.7996109724107849
CNN Confusion Matrix:
 [[8016 2445]
 [3118 7421]]
CNN Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.77      0.74     10461
           1       0.75      0.70      0.73     10539

    accuracy                           0.74     21000
   macro avg       0.74      0.74      0.73     21000
weighted avg       0.74      0.74      0.73     21000
```

RECOMMENDATION

1

The Convolutional Neural Network (CNN) outperforms the other models in both AUC-ROC and accuracy, indicating superior capability in predicting heart disease. The SVM model shows competitive performance and could be preferred in scenarios where neural networks are not feasible. While the Logistic Regression offers good performance with the advantage of interpretability, useful for clinical settings where understanding feature impact is crucial. Random Forest is beneficial when dealing with complex and non-linear relationships but falls slightly behind in performance metrics.

Based on the models and their interpretations, individuals at high risk of developing heart disease typically share the following **characteristics:**

- Age is consistently identified as a strong predictor of heart disease across all models. As people age, the risk of developing heart disease increases significantly.
- High Systolic Blood Pressure: Elevated systolic blood pressure is a key factor in heart disease risk. This is the pressure in your blood vessels when your heart beats. High Diastolic Blood Pressure: Though not highlighted as strongly as systolic pressure, elevated diastolic pressure (the pressure in your blood vessels when your heart rests between beats) also contributes to risk.
- High Cholesterol: Elevated cholesterol levels, particularly LDL (bad cholesterol), are a significant predictor of heart disease. High cholesterol can lead to the buildup of plaques in arteries, increasing the risk of heart attacks and strokes.
- High Body Weight/Obesity/BMI: Excess body weight relative to height, particularly obesity, is another major risk factor. Higher weight is associated with increased blood pressure, cholesterol, and a higher likelihood of developing heart disease.

Lifestyle Factors:

- Lack of regular physical activity can contribute to the development of heart disease by exacerbating other risk factors like obesity, high blood pressure, and high cholesterol.
- Smoking is a well-known risk factor for heart disease, as it damages the lining of the arteries and can lead to the buildup of plaques.
- Excessive alcohol consumption can increase blood pressure and contribute to heart disease risk.
- Individuals with high glucose levels, particularly those with diabetes, are at higher risk for heart disease due to the damage high blood sugar can cause to blood vessels.

To conclude, these factors should be closely monitored and managed to reduce the risk of heart disease, and targeted interventions should be considered for those at high risk. By understanding which factors are most important, personalized treatment plans can be developed that address the highest-risk features for each patient.

THANK YOU.

tendaimjonhasi@gmail.com
LinkedIn profile
Github