Introduction

Problem Statement

Heart disease is the leading cause of death in the United States, claiming approximately 695,000 lives annually, which accounts for one in every five deaths. Coronary artery disease (CAD), the most common type of heart disease, alone was responsible for 375,476 deaths in 2021. The prevalence of heart disease is alarming, affecting about 5% of adults aged 20 and older, and this rate is climbing. With heart disease posing such a significant public health threat, I chose to focus on this issue for my Predictive Analytics project. My goal for this project is to identify key predictors of heart disease based on the data provided.

Why the Problem is Important

The gravity of heart disease is underscored by the fact that it kills one person every 33 seconds in the United States. Beyond the immense loss of life, heart disease incurs a heavy financial burden, costing an estimated $239.9 billion annually in healthcare services, medications, and lost productivity. The widespread impact of this disease is felt across all demographic groups, regardless of age, gender, or ethnicity. Addressing heart disease is not merely a matter of improving health outcomes but also a societal and economic necessity. Given the scale of the problem, the need for effective predictive models that can identify individuals at high risk of heart disease is critical.

Who Would Be Interested in Solving This Problem?

The potential stakeholders in solving the heart disease problem are vast. Healthcare organizations such as the Centers for Disease Control and Prevention (CDC) and the U.S. Department of Health and Human Services (HHS) would be primary beneficiaries of improved predictive models for heart disease. Public and private hospitals, clinics, and healthcare professionals, including doctors and nurses, would also have a vested interest in utilizing these models to enhance patient care and preventive measures. Moreover, policymakers, insurance companies, and public health advocates would likely be interested in these findings, as they directly impact healthcare costs and public health strategies.

Data Source

For this project, I used a comprehensive heart disease dataset obtained from IEEE.org. The dataset was submitted by Manu Siddhartha from Liverpool John Moores University and is a result of merging five popular, previously independent heart disease datasets: Cleveland, Hungary, Switzerland, Long Beach VA, and Statlog (Heart) Data Set. This combined dataset is the largest available for research on heart disease, encompassing 1,190 instances with 11 features each. The dataset was specifically created to facilitate research on CAD-related machine learning and data mining algorithms, making it highly relevant for predictive analytics in this context.

Fig 1

Heart Disease Dataset Attribute Description

S.No.	Attribute	Code given	Unit	Data type
1	age	Age	in years	Numeric
2	sex	Sex	1, 0	Binary
3	chest pain type	chest pain type	1,2,3,4	Nominal
4	resting blood pressure	resting bp s	in mm Hg	Numeric
5	serum cholesterol	cholesterol	in mg/dl	Numeric
6	fasting blood sugar	fasting blood sugar	1,0 > 120 mg/dl	Binary
7	resting electrocardiogram results	resting ecg	0,1,2	Nominal
8	maximum heart rate achieved	max heart rate	71–202	Numeric
9	exercise induced angina	exercise angina	0,1	Binary
10	oldpeak =ST	oldpeak	depression	Numeric
11	the slope of the peak exercise ST segment	ST slope	0,1,2	Nominal
12	class	target	0,1	Binary

Fig 2

Description of Nominal Attributes

Attribute	Description
Sex	1 = male, 0= female;
Chest Pain Type	-- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic
Fasting Blood sugar	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
Resting electrocardiogram results	-- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Exercise induced angina	1 = yes; 0 = no
the slope of the peak exercise ST segment	-- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
class	1 = heart disease, 0 = Normal

Why the Data is Useful

The dataset is particularly useful for solving the heart disease problem because it contains a variety of features that are known to be associated with heart disease, such as age,

cholesterol levels, blood pressure, and more. These features provide a solid foundation for building predictive models that can identify individuals at risk of developing heart disease. Additionally, the dataset's size and the diversity of its sources enhance the robustness of the models that can be developed, making the findings more generalizable across different populations.

Methods and Results

Data Exploration

Visualizations played a crucial role in understanding the relationships between different features and heart disease. For example, scatter plots were effective in showing the relationship between age, resting blood pressure, and cholesterol levels. Histograms and density plots helped in visualizing the distribution of individual features, allowing for a deeper understanding of their impact on the target variable.

Data Preparation

Fortunately, the data required minimal preparation. The individual who uploaded the dataset had already gone through the painstaking process of cleaning and pre-processing it. There were no missing values, and categorical variables had already been converted into numerical values where necessary (e.g., converting gender into binary values). This allowed me to focus on model building and evaluation without the need for extensive data wrangling.

Modeling Approach

I employed two types of models for this project: Logistic Regression and Random Forests. Logistic Regression is a straightforward model that is easy to interpret, making it useful for understanding the impact of each feature on the probability of heart disease. It is particularly well-suited for binary classification problems like this one. On the other hand, Random Forests are more complex models that can capture non-linear relationships between features and the target variable. They are robust to overfitting and can handle both numerical and categorical features, making them a versatile choice for this dataset.

Model Evaluation

To evaluate the performance of the models, I used several metrics: accuracy, precision, F1 score, and cross-validation. Accuracy measures the overall correctness of the model, while precision focuses on the proportion of true positive predictions among all positive predictions. The F1 score is a balanced metric that considers both precision and recall, making it useful for assessing the model's overall effectiveness. Cross-validation was used to ensure that the models generalize well to unseen data by evaluating them on different subsets of the data.

The Logistic Regression model achieved an accuracy of 86.1%, with a precision of 87.1% and an F1 score of 87.5%. The cross-validation results showed consistent performance, with a mean accuracy of 81.9%. However, the Random Forest model outperformed Logistic Regression, achieving an accuracy of 94.54%, precision of 93.38%, and an F1 score of

95.13%. The cross-validation results for Random Forest were similarly strong, with an average accuracy of 92.86% and a precision of 93.13%.

Logistic Regression Model Evaluation

```
Confusion Matrix:
[[ 90  17]
 [ 16 115]]
Accuracy: 0.8613445378151261
Precision: 0.8712121212121212
F1 Score: 0.8745247148288973
Cross-validation Accuracy Scores: [0.80672269 0.89915966 0.86554622 0.76470588 0.84033613 0.86554622
 0.78151261 0.77310924 0.81512605 0.78151261]
Mean Cross-validation Accuracy: 0.819327731092437
```

Random Forest Model Evaluation

```
Confusion Matrix:
 [[ 98   9]
 [  4 127]]
Accuracy: 0.9453781512605042
Precision: 0.9338235294117647
F1 Score: 0.951310861423221
Cross-validated Accuracy: 0.9285714285714286
Cross-validated Precision: 0.9312755672648905
Cross-validated F1 Score: 0.9332100467879512
```

Conclusion

What I Learned

The most critical predictors of heart disease, according to the Random Forest model, are the ST slope, chest pain type, oldpeak, and maximum heart rate. These findings align with established medical knowledge, where these factors are commonly used to assess cardiovascular health. The superior performance of the Random Forest model over Logistic

Regression suggests that complex models that can capture non-linear relationships are more effective for this type of data.

Feature	Importance
ST slope	0.194539
chest pain type	0.134759
oldpeak	0.116515
max heart rate	0.116115
cholesterol	0.111434
age	0.094971
resting bp s	0.078512
exercise angina	0.068288
sex	0.038240
resting ecg	0.027517
fasting blood sugar	0.019110

Recommendations

Based on my analysis, I recommend deploying the Random Forest model for predicting heart disease. Its high accuracy and ability to identify key predictors make it a valuable tool for early diagnosis and prevention. Additionally, healthcare providers should focus on monitoring the most important features identified in this analysis, such as the ST slope and chest pain type, as part of routine cardiovascular assessments.

Model Readiness for Deployment

Seeing as I've never been a part of a project were the model I was working on got deployed, I can't say for sure whether either of the models in this project are ready for deployment or not. However, further validation on a larger and more diverse data set would be beneficial to ensure its generalizability across an even wider population demographic. Additionally,

integrating the model into healthcare systems would require collaboration with medical professionals to ensure its practical applicability.

Future Work

While the current model performs well, future work could focus on incorporating additional features that were not available in the dataset, such as weight, physical activity levels, and dietary habits. These factors are known to influence heart disease risk and could potentially enhance the model's predictive power. Moreover, exploring other machine learning models, such as neural networks, could provide further insights and improve prediction accuracy.

Ethical Considerations

Ethically, it is crucial to ensure that the model is used responsibly. Predictive models in healthcare must be deployed with caution to avoid misdiagnoses or over-reliance on automated systems. Transparency in how the model's predictions are generated is essential to maintain trust among healthcare providers and patients. Additionally, data privacy must be safeguarded, particularly when dealing with sensitive health information.

Mitigating Ethical Concerns

To mitigate potential ethical concerns, it is important to use the model as a supplementary tool rather than a definitive diagnostic tool. Medical professionals should continue to play a central role in interpreting model predictions and making final decisions. Regular audits of the model's performance and updates to its algorithms based on new medical research

will also help ensure that it remains accurate and reliable. Finally, patient consent and data anonymization should be prioritized to protect individual privacy.

References

- Centers for Disease Control and Prevention. (2024, May 15). Heart disease facts. Centers for Disease Control and Prevention. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html#:~:text=About%20695%2C000%20people%20died%20from,lost%20productivity%20due%20to%20death.

- Siddhartha, M. (2020, June 11). Heart disease dataset (comprehensive) | IEEE dataport. IEEEDataPort. https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

- World Health Organization. (2021). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)