Data Mining & Warehousing

Heart Disease Prediction

Eisha Fatima - 345554
Shahab Ali - 353303

Outline

- Introduction
- Dataset Description
- Exploratory Data Analysis
- Data Preprocessing
- Algorithms Used
- Explanation of Algorithms
- Project Timeline

Motivation

Significance

Human heart beats around 100,000 times, pumping 2,000 gallons of blood through the body.

Gender-Specific Symptoms

Subtlety of heart attack symptoms in women compared to men, underscoring the necessity to differentiate these symptoms for accurate diagnosis

Public Health Impact

Addressing the global prevalence of heart disease, one of the leading causes of death worldwide

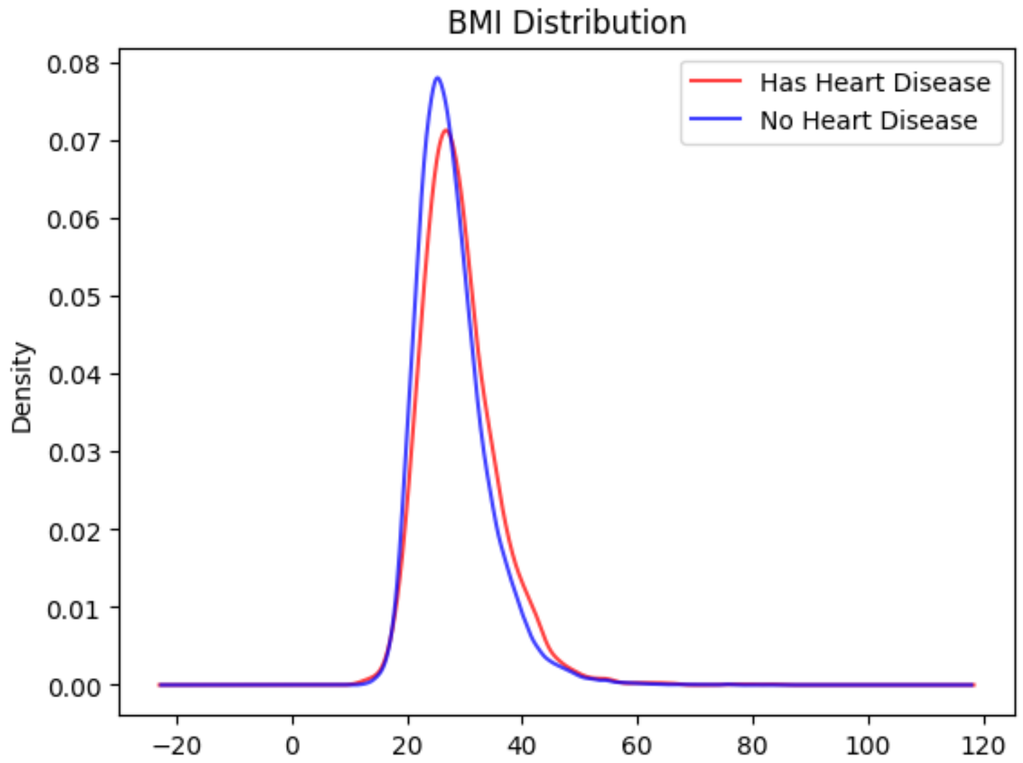
Dataset Description

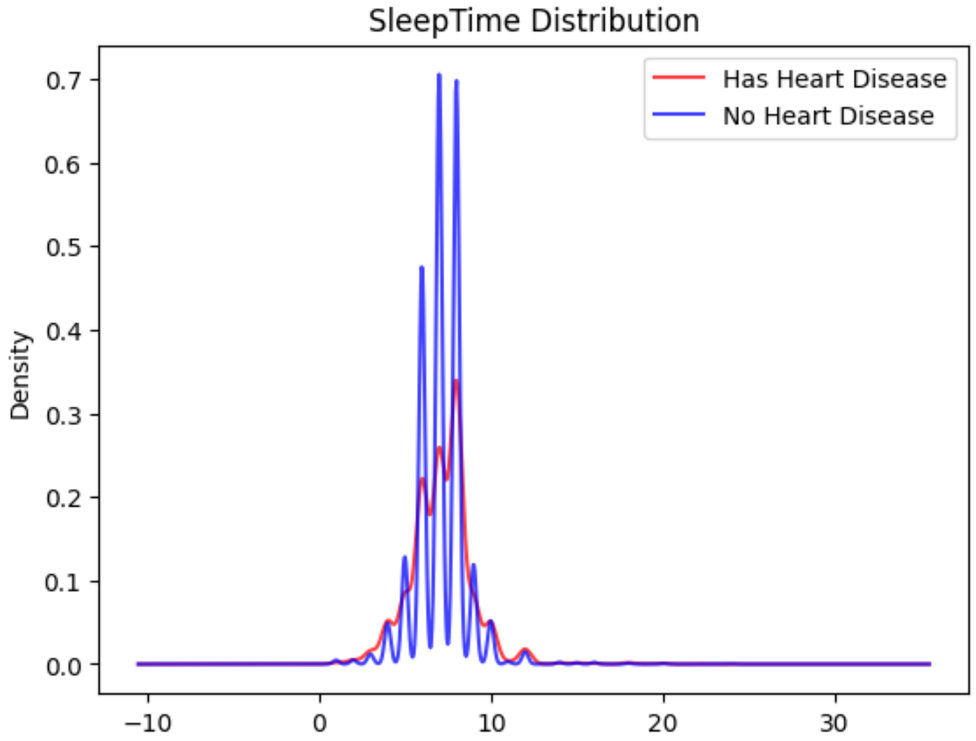
- **Source:** Personal Key Indicators of Heart Disease Dataset
- **Dataset Size:** 2020 annual CDC survey data of about 400k adults related to their health status.

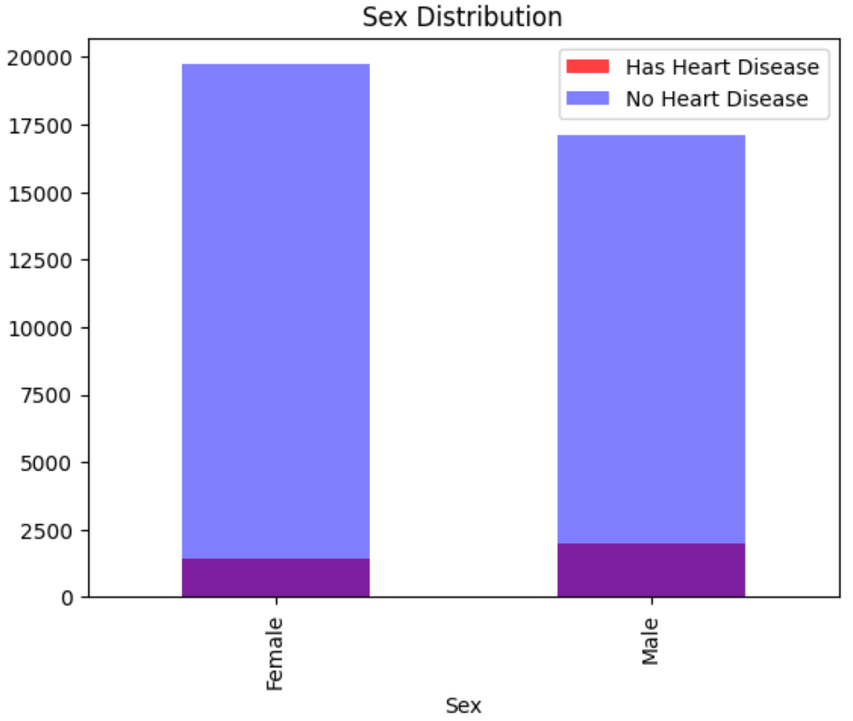
- It consists of **319,795** rows
- It consists of **18** columns.
- Originally, the dataset had about 300 variables.
- The variables were reduced to **18.**

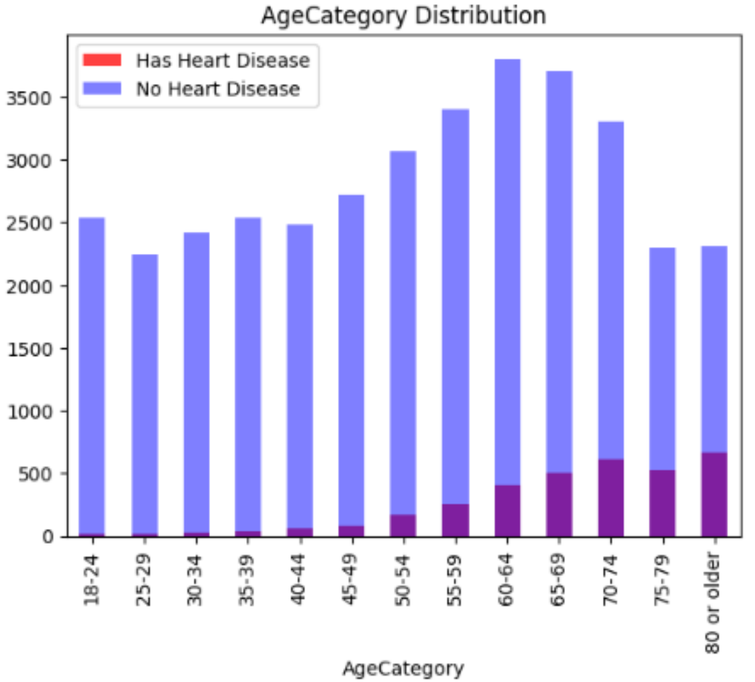
Exploratory Data Analysis

We visualized numerical information using kernel density estimation curves and categorical information using histograms. We did this separately for individuals with and without heart disease. This helped us understand how the attributes are distributed among these two groups.






Sex Distribution
Has Heart Disease
No Heart Disease
20000
17500
15000
12500
10000
7500
5000
2500
0
Female
Male
Sex


AgeCategory Distribution
Has Heart Disease
No Heart Disease
3500
3000
2500
2000
1500
1000
500
0
18-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80 or older
AgeCategory

Data Preprocessing

The data was preprocessed to make it clean and remove any error and noise.

The following steps were taken to preprocess the data:

- Exploring Unique Values and Error Checking in Dataset.
- Check for nan values.
- Replacing Binary Columns (No/Yes) with 0 and 1.
- Column Renaming and Typo Corrections.
- Min-Max Normalization on Numerical Attributes.
- One-Hot Encoding.

Algorithms Used

We used different algorithms to compare their performance so that we can select the best:

- Logistic Regression
- Decision Tree
- Linear SVM
- Categorical Naive Bayes
- Numerical Naive Bayes
- aseline Neural Network

Comparison Table

Model	Accuracy	Precision (No Heart Disease)	Precision (Yes Heart Disease)	Recall (No Heart Disease)	Recall (Yes Heart Disease)	F1-Score (No Heart Disease)	F1-Score (Yes Heart Disease)
Logistic Regression	0.91	0.92	0.52	0.99	0.10	0.96	0.17
Decision Tree	0.86	0.93	0.23	0.92	0.25	0.92	0.24
Linear SVM	0.91	0.91	0	1	0	0.96	0
Categorical Naive Bayes	0.89	0.94	0.35	0.93	0.37	0.94	0.36
Gaussian Naive Bayes	0.87	0.93	0.22	0.93	0.21	0.93	0.22
Baseline Neural Network	0.91	0.93	0.5	0.98	0.16	0.95	0.24

Explanation of Algorithms

Logistic Regression

The Logistic Regression model excels in overall accuracy compared to non-Neural Network models but demonstrates lower precision, recall, and f1-score for individuals with heart disease, indicating a tendency to misclassify positive cases.

Decision Tree

The Decision Tree model, despite having the lowest overall accuracy, outperforms Logistic Regression in f1 score and recall for individuals with heart disease, indicating a higher tendency to classify positives. However, it has reduced precision due to overfitting.

Linear SVM

The Linear SVM, despite achieving higher overall accuracy than the Decision Tree, is the worst-performing model, classifying all individuals as not having heart disease.

Explanation of Algorithms

Categorical Naive Bayes

Categorical Naive Bayes, despite modest overall accuracy, stands out with superior f1-score and recall for individuals with heart disease, surpassing other models. It exhibits high precision and true positives.

Numerical Naive Bayes

Numerical Naive Bayes shares similarities with the Decision Tree model, displaying low overall accuracy but slight improvements in recall and f1-score for individuals with heart disease compared to logistic regression.

Baseline Neural Network

The baseline Neural Network shows accuracy similar to Logistic Regression but excels with higher recall and f1-score for individuals with heart disease. While precision is comparable, the model registers nearly double the false positives.

Conclusion & Discussion

- Heart diseases are a major cause of global deaths, influenced by various factors like lifestyle choices (smoking, alcohol, sleep), health conditions (diabetes, strokes), and personal factors (race, age, sex). This report used machine learning models to understand these influences, including logistic regression, decision tree, linear SVM, Naive Bayes, and Neural Networks.

Imbalanced Data

Occurrence of heart disease cases (positive class) might be significantly lower than non-heart disease cases

Feature Selection and Relevance

Identifying which attributes significantly contribute to predicting heart disease and how to handle less influential or redundant variables is crucial for model performance.

Challenges

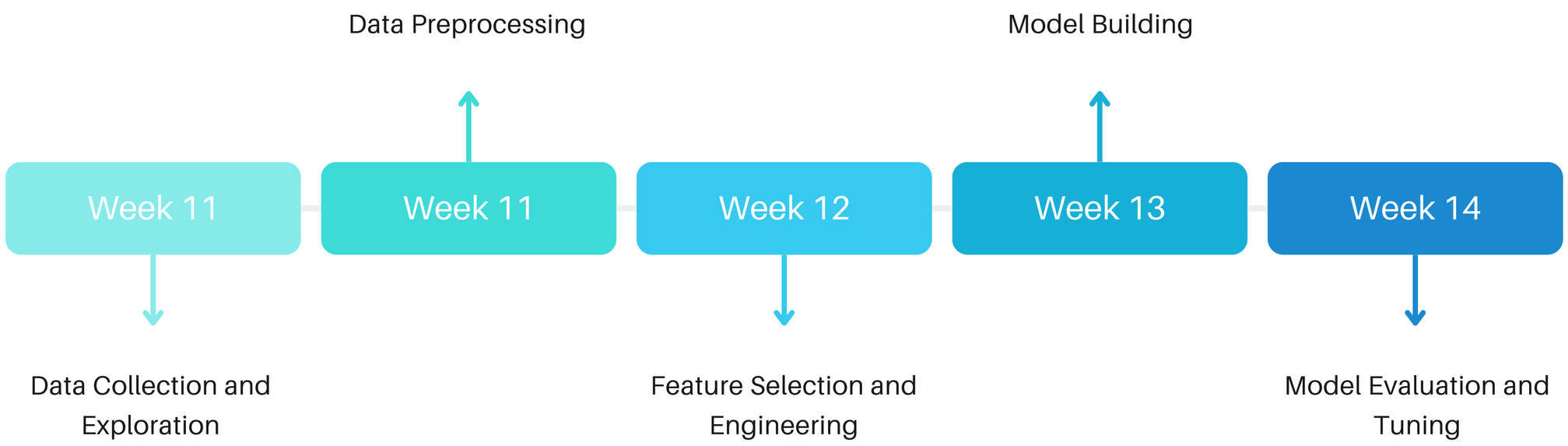
Handling Missing Data and Outliers

These can affect the quality of predictions.

Interactive Dashboards or Reporting

Predictions in a user-friendly format for healthcare professionals or stakeholders.

Project Timeline



Fauget

Thank You

For Watching

End of Pages