As part of this project, I want to create a classification system that can predict the possibility of contracting a heart disease. The dataset used for classification is known as the heart disease dataset. The reason why this problem was chosen: heart diseases are expensive to treat leading to financial burden to patients and family members. Heart diseases can be easily prevented and maintained by identifying risks and adjusting lifestyle. The goal of this project is to provide data driven ideas to improve health care and human life. The link to the dataset: <https://www.kaggle.com/datasets/mexwell/heart-disease-dataset>

This heart disease dataset is curated by combining 5 popular heart disease datasets already available independently but not combined before. The 5 datasets used for its curation are: Cleveland, Hungarian, Switzerland, Long Beach Virginia and Statlog Heart Data Set. The dataset has 1190 and 12 columns. The columns include: cholesterol (value of serum cholesterol concentration), fasting blood sugar (determines if a patient's blood sugar level is low to be considered fasting) and chest pain type(type of chest pains). It doesn't contain any missing values. The dataset has nominal data in variables such as sex, chest pain type, ST slope and exercise angina and integer data in variables such as age, resting bp s, cholesterol and max heart rate.

For data visualization, I transformed nominal data variables, target and chest pain, into character values. Normal hearts had a higher mean than deceased hearts. Majority of normal hearts had a maximum heart rate of around 150 and 175 while deceased hearts had 2 peaks at 125 and 138. In a normal heart, when the body demands more oxygen during exercise or stress, the heart responds by increasing its rate and pumping more blood to deliver oxygen to the tissues. However, in hearts with disease, such as coronary artery disease or heart failure, there may be structural or functional abnormalities which cause issues with how efficiently they pump blood. This reduces the ability to increase heart rate. Count of deceased hearts increases with age until the age group 55 to 60 then decreases. The decline of count is caused by decreased sample size. People of the age group 55 to 60 and above had a higher count of deceased hearts than normal hearts. Factors that contribute to increased risk of heart disease with age: blood vessels become less elastic and more rigid, risk factors such as high blood pressure, high cholesterol and obesity tend to worsen with age and older adults have reduced physical activity. Asymptomatic chest pains had the highest count compared to other classes. Asymptomatic value refers to absence of symptoms. My assumption was also used to categorize patients who did not define their chest pains or describe chest pains that can't be categorized into other classes. Typical angina had the highest ratio of normal heart to deceased heart. Typical angina describes pressure, squeezing or tightness in the chest. It is often associated with underlying coronary artery disease.

The heart disease dataset has features with different scales and units. Standardization was performed to ensure all features are on a similar scale. The mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. This also helps mitigate the impact of outliers by scaling the data.

The heart disease dataset was divided into two subsets: a training dataset and a test dataset by the ratio 70:30. The training dataset is used to train the machine learning model, while the test dataset was kept separate and used for evaluating the model's performance. The machine learning algorithms used for training include:

a. K-Nearest Neighbors (KNN):

In KNN, the model does not learn explicit parameters during training. Instead, it memorizes the training dataset. To fit the training dataset, the KNN algorithm simply stores the feature vectors and corresponding class labels.

b. Support Vector Machines (SVM):

SVM aims to find the hyperplane that best separates different classes in the feature space. During training, the SVM algorithm optimizes the hyperplane's parameters based on the training dataset to maximize the margin between classes. The fitted SVM model consists of the support vectors and associated weights.

c. Naive Bayes:

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of feature independence. During training, Naive Bayes calculates class-conditional probabilities and prior probabilities based on the training dataset. The fitted Naive Bayes model stores these probabilities for each class and feature.

d. Decision Trees:

Decision trees recursively split the feature space into regions based on feature values to minimize impurity (e.g., Gini impurity or entropy). During training, the decision tree algorithm selects optimal split points based on feature values to separate the classes. The fitted decision tree model consists of a tree structure with decision nodes and leaf nodes, representing the learned decision rules.

e. Random Forest:

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance. During training, multiple decision trees are trained on random subsets of the training dataset (bootstrap samples) and random subsets of features. The fitted Random Forest model consists of the collection of decision trees, each trained on a subset of the data.

A 5 fold cross validation was used to assess how well a model generalizes to unseen data by repeatedly splitting the dataset. It is a more reliable estimate of a model's performance by averaging performance across multiple splits. It helps identify models that generalize well rather than a model that memorizes the training data. The cross validation accuracy ranges from 76% to 89%. Linear and sigmoid support vector machines had an accuracy below 80%. Random forest classification has the optimal accuracy of 89%.

Test dataset was used for prediction. Evaluation metrics such as accuracy, precision, recall and F1 score were computed to assess the model's performance on unseen data. Random Forest Classification was the model with the highest accuracy of 94%. It had a 93% precision and 95% recall. The weighted average and macro average were similar indicating a balanced dataset. Decision trees and random forest exhibit a higher accuracy compared to linear and sigmoid support vector machines. Decision trees and random forests are capable of capturing complex, non-linear relationships between features and target variable. Random forest naturally accounts for interaction between features allowing them to capture complex patterns and interaction in the data while linear and sigmoid support vector machines assume that features contribute linearly to the decision boundary.

Machine learning algorithms analyzes various types of medical data, such as electronic health records, medical imaging (e.g., X-rays, MRIs), and genetic data, to identify patterns and risk factors associated with heart disease. Machine learning models can help in the early detection and diagnosis of heart disease. They identify subtle abnormalities or predicting the likelihood of developing cardiovascular diseases based on individual risk factors. These factors include being overweight, taking drugs and lack of exercise. Machine learning techniques can analyze multiple risk factors and biomarkers to group individuals into different risk categories for heart disease.

Predictive machine learning algorithms can estimate the risk of developing specific cardiovascular activities by monitoring heart rate, cholesterol level and blood pressure.

They can send alerts to healthcare providers identifying high risk individuals who may benefit from early intervention or lifestyle modification to prevent heart disease. Machine learning algorithms can analyze large-scale clinical data to identify optimal treatment strategies and novel treatment regimens for individual patients. This is based on data-driven insights from demographics data, medical history data, genetic data, and response to treatment data. Personalized risk assessment and treatment recommendations can improve patient outcomes and reduce the burden of heart disease by specifying interventions to each patient's specific needs. Machine learning techniques, such as deep learning, have shown promising results in analyzing medical imaging data for the detection and characterization of cardiovascular abnormalities.

This has helped reduce test time and increased accuracy in results. The image analysis algorithms can assist radiologists, doctors, and medical doctors in interpreting cardiac imaging studies (e.g., echocardiograms, cardiac MRI) by identifying anatomical incorrect structures, and detecting abnormalities such as thick blood vessels around the heart, blood clots in the heart or structural heart defects. Machine learning holds great promise for improving the prevention, diagnosis, treatment, and management of heart disease by leveraging the power of data-driven insights and machine learning.

**APPENDIX**

****

****

****

****