AIML PROJECT

HEART DISEASES PREDICTION

- **Student name:**Gangisetty Dimple Aishwarya
- **Roll number:** 2366250173
- **Institution:** Sri Aditya Degree College
- **Department name:** BSc Data Science

- **Submission date:** 5/02/202

1.Introduction:

Heart disease is one of the leading causes of death worldwide. As medical technology advances, the ability to predict heart diseases and intervene early has become increasingly important. Heart disease prediction involves using various data sources, including medical history, lifestyle factors, and diagnostic tests, to estimate the likelihood of an individual developing heart-related issues such as heart

attacks, coronary artery disease, arrhythmias, and more. This predictive process helps doctors provide personalized care and allows for earlier, more effective prevention strategies.
The goal of heart disease prediction is to prevent serious outcomes by detecting risk factors and warning signs early on, allowing for timely interventions such as lifestyle changes, medications, or further diagnostic testing.

2.Data collection:

Data collection is a crucial step in the development of prediction models for heart disease. The quality, diversity, and accuracy of the data directly affect the reliability of any model used to predict the likelihood of heart-related conditions. Collecting comprehensive and relevant data allows for more accurate risk assessments, early diagnosis, and targeted treatment plans for individuals at risk of heart disease.

- **Age**: The risk of heart disease increases as a person gets older, particularly after the age of 45 for men and 55 for women.
- **Gender**: Men generally have a higher risk of heart disease at a younger age compared to women. However, after menopause, women's risk increases, and they often experience different symptoms and outcomes.
- **Ethnicity**: Certain ethnic groups, such as African Americans, Hispanics, and South Asians, are more prone to certain heart conditions, like hypertension and diabetes, which increase the risk of heart disease.

- **Family History**: A family history of heart disease, especially if a first-degree relative (parent or sibling) developed heart problems early, increases a person's risk of heart disease.
- **Blood Pressure**: High blood pressure (hypertension) is a leading risk factor for heart disease. Regular blood pressure readings help in assessing the risk.
- **Cholesterol Levels**: Blood cholesterol levels, specifically low-density lipoprotein (LDL, the "bad" cholesterol) and high-density lipoprotein (HDL, the "good" cholesterol), provide key information about the risk of plaque buildup in arteries, a major contributor to heart disease.
- **Blood Glucose (Diabetes)**: Diabetes is a major risk factor for heart disease, as high blood sugar can lead to the development of artery-blocking plaque. Monitoring blood glucose levels helps detect potential diabetes-related heart risks.
- **Electrocardiogram (ECG)**: An ECG measures the electrical activity of the heart and helps detect abnormalities in heart rhythm (arrhythmias), heart attacks, or other conditions that may signal heart disease.
- **Echo and Ultrasound**: Echocardiograms (ultrasounds of the heart) help evaluate heart function, including pumping ability and valve function. They can reveal conditions like heart failure or valve disease.
- **Chest X-rays or CT Scans**: These imaging techniques are used to detect structural issues in the heart and lungs that might contribute to cardiovascular problems.

3.Data preprocessing:

Data Preprocessing for Heart Disease Prediction involves preparing raw data for analysis by performing several essential steps to improve model accuracy:

1. **Handling Missing Data**: Address missing values by imputation, removal, or prediction using statistical methods or machine learning models.
2. **Outlier Detection**: Identify and handle extreme values that may distort predictions, either by removing or capping them.
3. **Normalization and Scaling**: Rescale data features to a standard range (e.g., [0,1] or mean 0, standard deviation 1) to ensure consistent input for models.
4. **Categorical Encoding**: Convert non-numeric categorical data (like "Yes"/"No" or gender) into numerical format using label encoding or one-hot encoding.
5. **Feature Engineering**: Create new variables (e.g., BMI from height and weight) or derive meaningful interactions between existing features to improve model performance.
6. **Feature Selection**: Identify and retain the most relevant features, removing irrelevant or redundant ones to enhance predictive accuracy and reduce complexity.
7. **Data Splitting**: Divide the dataset into training and testing sets to evaluate model performance effectively.

4. Machine Learning Model Selection:

involves preparing raw data for analysis by performing several essential steps to improve model accuracy:

1. **Handling Missing Data**: Address missing values by imputation, removal, or prediction using statistical methods or machine learning models.
2. **Outlier Detection**: Identify and handle extreme values that may distort predictions, either by removing or capping them.
3. **Normalization and Scaling**: Rescale data features to a standard range (e.g., [0,1] or mean 0, standard deviation 1) to ensure consistent input for models.
4. **Categorical Encoding**: Convert non-numeric categorical data (like "Yes"/"No" or gender) into numerical format using label encoding or one-hot encoding.

5. **Feature Engineering**: Create new variables (e.g., BMI from height and weight) or derive meaningful interactions between existing features to improve model performance.
6. **Feature Selection**: Identify and retain the most relevant features, removing irrelevant or redundant ones to enhance predictive accuracy and reduce complexity.
7. **Data Splitting**: Divide the dataset into training and testing sets to evaluate model performance effectively.

These steps ensure that the data is clean, consistent, and ready for machine learning models to make accurate heart disease predictions.

You said:

Machine Learning Model Selection in heart diseases prediction in short defnetion

ChatGPT said:

ChatGPT

Machine Learning Model Selection for Heart Disease Prediction involves choosing the most suitable algorithm based on the data characteristics, problem complexity, and desired outcomes. Commonly used models include:

1. **Logistic Regression**: Ideal for binary classification (e.g., predicting whether a patient has heart disease or not).
2. **Decision Trees**: Useful for visualizing decision-making processes and handling both categorical and continuous data.
3. **Random Forest**: An ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting.
4. **Support Vector Machines (SVM)**: Effective for high-dimensional spaces and when there is a clear margin of separation between classes.

5. **K-Nearest Neighbors (KNN)**: Simple, instance-based learning that classifies a new instance based on the majority label of its nearest neighbors.
6. **Naive Bayes**: A probabilistic classifier based on Bayes' theorem, often used when features are independent.

---

5. Model Training and Evaluation:

1. **Model Training**: Using the training dataset to teach the selected machine learning model patterns in the data. During training, the model learns the relationships between input features (e.g., age, cholesterol levels, lifestyle) and the target variable (e.g., heart disease presence).
2. **Model Evaluation**: After training, the model is evaluated on a separate testing dataset to assess its performance. Key evaluation metrics include:
   - **Accuracy**: The proportion of correct predictions (true positives + true negatives) over total predictions.
   - **Precision**: The proportion of true positives among all predicted positives (important when false positives are costly).
   - **Recall**: The proportion of true positives among all actual positives (important when false negatives are risky).
   - **F1-Score**: The harmonic mean of precision and recall, balancing both.
   - **AUC-ROC**: The area under the receiver operating characteristic curve, measuring the model's ability to distinguish between classes.

Training and evaluation help refine the model for better prediction accuracy and generalization to new data.

---

6. Model Optimization:

Model optimization for heart disease prediction involves fine-tuning machine learning models to improve their accuracy, efficiency, and reliability in predicting heart disease outcomes. This can be achieved through various techniques such as:

1. **Feature Engineering**: Selecting the most relevant features (e.g., age, blood pressure, cholesterol levels) and transforming the data to improve model performance.
2. **Hyperparameter Tuning**: Adjusting model parameters (e.g., learning rate, tree depth) to find the best configuration for optimal prediction.
3. **Cross-validation**: Using techniques like k-fold cross-validation to reduce overfitting and improve generalizability.
4. **Model Selection**: Testing various algorithms like Logistic Regression, Random Forest, SVM, and Neural Networks to find the most effective one.
5. **Ensemble Methods**: Combining predictions from multiple models (e.g., boosting or bagging) to enhance accuracy.
6. **Data Preprocessing**: Addressing imbalanced datasets (through oversampling/undersampling), handling missing data, and normalizing features.

7.Deployment and Monitoring:

Deployment and monitoring for heart disease prediction involve putting the trained model into a production environment and ensuring its ongoing performance:

1. **Deployment**: The model is integrated into an application or system where it can receive real-time or batch input data (e.g., patient health records) and provide predictions. This can be done through APIs, web services, or cloud platforms.
2. **Monitoring**: Continuously tracking the model's performance in the real world. This includes checking for accuracy, detecting data drift (changes in input data patterns), and ensuring the model remains reliable. Monitoring helps identify when the model needs retraining or adjustments.

Execution:

pip install pandas numpy scikit-learn

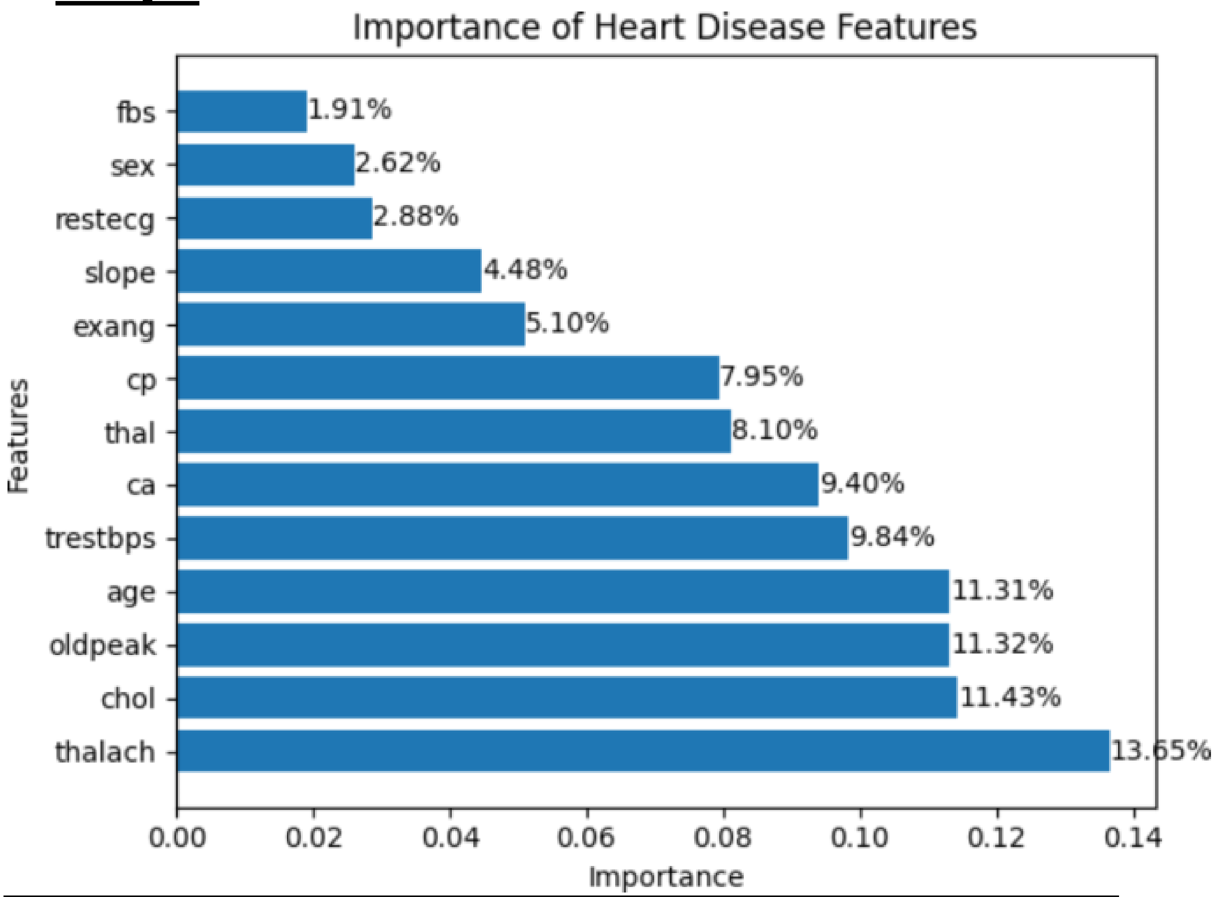
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score

# Load the Cleveland Heart Disease dataset
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/heart-
disease/processed.cleveland.data'
columns = ['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang',
'oldpeak', 'slope', 'ca', 'thal', 'target']

# Load dataset into pandas DataFrame
data = pd.read_csv(url, header=None, names=columns)

# Preprocess: Handle missing values
# Replace '?' with NaN and drop rows with missing values
data.replace('?', pd.NA, inplace=True)
data.dropna(inplace=True)

# Convert 'ca' and 'thal' columns to numeric (as they were strings) after replacing
'?' with NaN
data['ca'] = pd.to_numeric(data['ca'], errors='coerce')
data['thal'] = pd.to_numeric(data['thal'], errors='coerce')

# Ensure the target variable and other columns are of the correct data type
data['target'] = data['target'].astype(int)

# Features (X) and target (y)
X = data.drop('target', axis=1)  # Features
y = data['target']  # Target variable

# Split the dataset into training (80%) and testing (20%) sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Initialize the RandomForestClassifier
model = RandomForestClassifier(n_estimators=100, random_state=42)

# Train the model on the training set
```

```
model.fit(X_train, y_train)

# Predict the target values for the test set
y_pred = model.predict(X_test)

# Evaluate the model's performance
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))

print("\nClassification Report:")
print(classification_report(y_test, y_pred))

print("\nAccuracy Score:")
print(accuracy_score(y_test, y_pred))
```

Output:

```
Confusion Matrix:
[[35  1  0  0  0]
 [ 5  0  3  1  0]
 [ 2  1  1  1  0]
 [ 1  2  3  0  1]
 [ 1  2  0  0  0]]

Classification Report:
          precision   recall  f1-score   support

       0      0.80     0.97     0.88       36
       1      0.00     0.00     0.00        9
       2      0.14     0.20     0.17        5
       3      0.00     0.00     0.00        7
       4      0.00     0.00     0.00        3

   accuracy                      0.60       60
  macro avg      0.19     0.23     0.21       60
weighted avg      0.49     0.60     0.54       60


Accuracy Score:
0.6
```

8.Graph:



9.Conclusion:

Heart disease prediction models are powerful tools that leverage machine learning techniques to identify individuals at risk of heart disease based on various health-related features. By utilizing factors such as age, cholesterol levels, blood pressure, and other medical parameters, these models can predict the likelihood of heart disease with high accuracy. This enables earlier interventions, personalized treatment plans, and better resource allocation in healthcare systems.