Probability & Statistics (ENG 024)

Final Project Report

Heart Disease Diagnosis using ML

Team:

1- عبدالعزيز ناصر عبدالعزيز محمد (211600183)
2- عمر أحمد إسماعيل محمد (211600202)
3- عمر محمد فاروق درويش (211600208)
4- محمد فريد عبده محمد (211600307)
5- ياسر نبيل ياسر ربيع (211600397)
6- مهند مؤمن ابراهيم محمد (211600457)
7- عمار عمر حسن رفعت (211600200)
8- اسلام حسن جمال الدين السيد (211600070)
9- عمر محمد حنفي صبحي (231600171)
10- أحمد محمد عبدالرؤوف عبدالله شراب (211600042)

Under the Supervision of:

Dr. Ahmed Magdy Dr. Sherif Fathi

2024-2025

Presentation Video Link:

https://drive.google.com/file/d/1FKiNwJ-6UCA79aQm96qc9nxJTJi1bR-h/view?usp=sharing

Code Link:

https://colab.research.google.com/drive/1g4RSozjO8vwz4c37rFYI6l11JTO_YAGS?usp=sharing

Objective:

To develop a machine learning model that can predict the likelihood of a person having heart disease based on various health features like age, blood pressure, cholesterol levels, and more.

Tools and Technologies:

- **Python:** For scripting, data processing, and model implementation.
- **scikit-learn:** For implementing the machine learning algorithms.
- **pandas:** For data manipulation and preprocessing.
- **matplotlib / seaborn:** For data visualization and exploratory analysis.

Techniques:

1. **Naïve Bayes Classifier:**
    - A probabilistic model that will be used to predict the presence of heart disease based on the symptoms and medical data of the patient.
    - This model assumes that features (health attributes) are independent and will predict the probability of disease.
2. **Confusion Matrix:**
    - A performance evaluation tool that provides insights into the model's accuracy, precision, recall, and F1 score.
    - It will help assess the number of true positives, true negatives, false positives, and false negatives.

Steps:

1. Data Collection and Preprocessing

- **Data Loading:**
  - Load the **Heart Disease UCI Dataset** into a pandas Data Frame.
  - Handle missing values and encode categorical variables (e.g., sex, chest pain type).
- **Data Exploration:**
  - Perform exploratory data analysis (EDA) to understand the relationships between features and the target variable (`target` indicating presence of heart disease).
  - Visualize the distribution of key features like age, cholesterol, blood pressure, and target (disease/no disease).
- **Feature Engineering:**
  - Encode categorical features using one-hot encoding or label encoding.
  - Normalize numerical features to improve model performance.
- **Train-Test Split:**
  - Split the data into training and testing sets (e.g., 80% training, 20% testing).

2. Model Implementation

- **Naïve Bayes:**
  - Use the `GaussianNB` implementation from scikit-learn to train the model on the training set.
  - Train the model using the features and the target variable (disease vs no disease).

3. Model Evaluation

- **Confusion Matrix:**
  - Generate a confusion matrix to evaluate model performance.
  - Use the confusion matrix to calculate performance metrics like accuracy, precision, recall, and F1 score.
- **Cross-Validation:**
  - Optionally perform cross-validation to get a better estimate of the model's performance.

4. Model Tuning

- **Hyperparameter Tuning:**
  - Tune the hyperparameters (if any) of the Naïve Bayes model to enhance performance.

5. Prediction and Interpretation

- **Predict Disease:**

    - Use the trained model to predict the likelihood of heart disease for a given patient based on input features.
    - Display the predicted probability for each class (disease vs no disease).
- **Visualization:**
    - Visualize the confusion matrix to see the performance of the model.
    - Plot ROC curves or precision-recall curves to evaluate model effectiveness.

Expected Outcomes:

- A trained Naïve Bayes model capable of predicting the presence or absence of heart disease based on health features.
- A performance evaluation using a confusion matrix and metrics like accuracy, precision, recall, and F1 score.
- Visualizations of model results and prediction probabilities for better understanding.

Dataset Overview:

The **Heart Disease UCI dataset** contains several attributes related to a patient's medical record and symptoms. The dataset has the following features:

- **age**: Age of the patient
- **sex**: Gender (1 = male, 0 = female)
- **cp**: Chest pain type (4 values)
- **trestbps**: Resting blood pressure (in mm Hg)
- **chol**: Serum cholesterol (in mg/dl)
- **fbs**: Fasting blood sugar (> 120 mg/dl) (1 = true, 0 = false)
- **restecg**: Resting electrocardiographic results (values 0, 1, 2)
- **thalach**: Maximum heart rate achieved
- **exang**: Exercise induced angina (1 = yes, 0 = no)
- **oldpeak**: Depression induced by exercise relative to rest
- **slope**: Slope of the peak exercise ST segment
- **ca**: Number of major vessels colored by fluoroscopy (0-3)
- **thal**: Thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect)
- **target**: Presence of heart disease (1 = disease, 0 = no disease)

Potential Challenges and Mitigations:

- **Imbalanced Data:** If the dataset has an imbalance in the classes (e.g., more samples with "no disease"), this may affect model performance. This can be mitigated by techniques like oversampling the minority class or using class weights.

- **Missing Values:** Handle any missing data through imputation (e.g., using the median or mode for numerical/categorical columns).