Heart Disease Prediction Using Machine Learning Models
---by Janani Sri S(125018035)

1. Introduction

Coronary heart diseases or CHDs are some of the leading killer diseases worldwide. Thus, the identification of CHD risk level based on routine medical data is crucial for early effective prevention. This study utilizes the Framingham Heart Study dataset to predict the 10-year risk of CHD by applying and comparing several machine learning models: First, it will use Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree algorithms. From the previously mentioned models, such as Logistic Regression, Naive Bayes, KNN, and Decision Tree, these were chosen for their differences in the model complexity where ranging from simple and easy to explain (Logistic Regression and Naive Bayes) to the models which work with parameters (KNN and Decision Tree). The goal is to define what kind of model is the best using Accuracy and ROC_AUC coefficients.

2. Methodology

1. **Dataset Overview and Preprocessing**:
    - The Framingham dataset provides data on risk factors such as age, gender, cholesterol levels, and blood pressure, essential for studying heart disease.
    - Preprocessing steps included:
        - Dropping the education column and renaming the male column to Sex_male.
        - Removing rows with missing values to ensure data integrity.
        - Standardizing features (age, cigsPerDay, totChol, sysBP, glucose) using StandardScaler for consistent scaling across all models.
2. **Train-Test Split**:
    - The dataset was split into training (70%) and testing (30%) sets using a random seed (random_state=4) to allow reproducibility.
3. **Model Training and Evaluation**:

- Four models were trained using the preprocessed data:
  - **Logistic Regression**: Suitable for binary classification problems due to its simplicity and interpretability.
  - **Naive Bayes**: A probabilistic model that assumes independence between features, offering speed and efficiency.
  - **K-Nearest Neighbors (KNN)**: An instance-based model, with K=5 used initially for this study.
  - **Decision Tree**: Splits data based on feature values, creating a tree structure for decision-making.
- Evaluation Metrics:
  - **Accuracy**: Proportion of correctly predicted cases out of the total cases.
  - **ROC-AUC Score**: Measures the model's ability to distinguish between classes, with higher values indicating better discriminatory power.

3. Results

The models were evaluated based on their performance metrics, and the results are as follows:

Model	Accuracy (%)	ROC-AUC Score
Logistic Regression	84.0	0.72
Naive Bayes	83.0	0.72
K-Nearest Neighbors	83.0	0.61
Decision Tree	74.0	0.54

1.Logistic Regression:

- **Accuracy**: 84.0%
- **ROC-AUC**: 0.72
- Logistic Regression demonstrated the highest accuracy, indicating its effectiveness in predicting CHD risk. The model's simplicity and interpretability also make it ideal for healthcare settings.

2.Naive Bayes:

- **Accuracy**: 83.0%
- **ROC-AUC**: 0.72
- The Naive Bayes model performed similarly to Logistic Regression in terms of accuracy and ROC-AUC. Its fast computation makes it suitable for real-time applications, but its assumptions about feature independence could limit performance.

3.K-Nearest Neighbors (KNN):

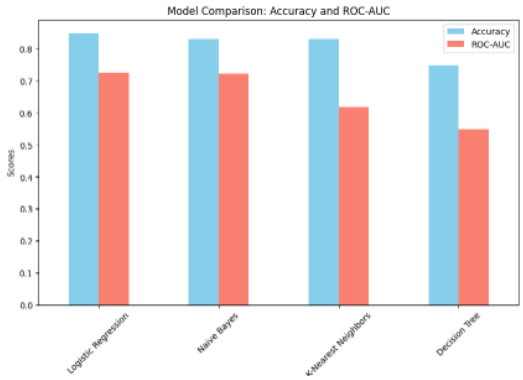
- **Accuracy**: 83.0%
- **ROC-AUC**: 0.61
- While KNN achieved a comparable accuracy to Logistic Regression and Naive Bayes, its lower ROC-AUC score suggests weaker class separation ability. This is likely due to the sensitivity of KNN to the chosen value of K (set at 5) and the distance metric. Future work should explore tuning these parameters to enhance the model's effectiveness.

4.Decision Tree:

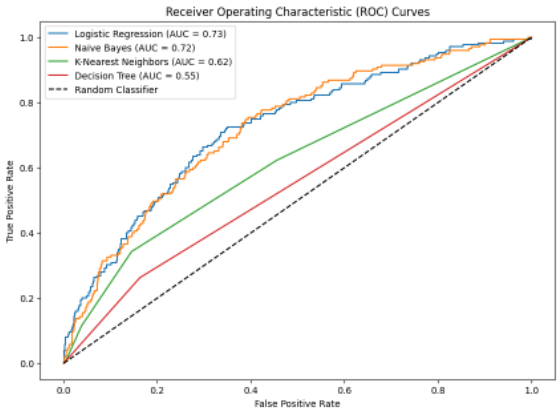
- **Accuracy**: 74.0%
- **ROC-AUC**: 0.54
- The Decision Tree model had the lowest ROC-AUC score, indicating its difficulty in distinguishing between classes effectively. This suggests that the model may be overfitting, capturing specific patterns that do not generalize well to unseen data. Techniques such as pruning or using ensemble methods like Random Forest could potentially improve this model's performance.

4.Visual Analysis:

- **Bar Chart** (Figure 1): Illustrates the accuracy and ROC-AUC scores for each model.



ROC Curves (Figure 2): Displays the trade-off between true positive and false positive rates for each model, showing how Logistic Regression and Naive Bayes maintain a balanced performance.



The results highlight Logistic Regression as the most effective model for predicting the 10-year risk of CHD based on the selected features. Despite a moderate ROC-AUC score (0.72), its high accuracy (84.0%) suggests that it is capable of correctly classifying a substantial proportion of cases. Its interpretability is also beneficial for healthcare professionals who require clear, understandable results.

5. Discussion

The results highlight **Logistic Regression** as the most effective model for predicting the 10-year risk of CHD based on the selected features. Despite a moderate ROC-AUC score (0.72), its high accuracy (84.0%) suggests that it is capable of correctly classifying a substantial proportion of cases. Its interpretability is also beneficial for healthcare professionals who require clear, understandable results.

6. Model Comparisons:

- Naive Bayes offers a viable alternative, performing similarly to Logistic Regression with the added advantage of faster computation.

- KNN's performance suggests the need for parameter tuning; the ROC-AUC score of 0.61 indicates that the model struggles with class separation, possibly due to the chosen K value or distance metric. Exploring higher values of K and different distance measures could improve the model.
- The Decision Tree model's low ROC-AUC score (0.54) indicates poor generalization. Implementing techniques such as pruning or switching to ensemble methods like **Random Forest** could enhance its performance by reducing variance and improving generalization.

7. Limitations:

- The dataset's class imbalance could influence the models' ability to generalize. Techniques like **SMOTE** (Synthetic Minority Over-sampling Technique) or adjusting class weights could be employed to address this issue.
- The model evaluation relied on only six features. Incorporating additional features, such as physical activity levels or dietary habits, could provide a more comprehensive picture of each individual's risk profile.

8. Conclusion

This study confirms that Logistic Regression is the most effective model for predicting the 10-year risk of CHD, with an accuracy of 84.0% and a ROC-AUC score of 0.72. For future work:

- **Hyperparameter Optimization**: Further tuning of parameters in KNN and Decision Trees could refine performance.
- **Ensemble Methods**: Implementing advanced techniques like Random Forest or boosting algorithms may enhance accuracy and robustness.
- **Feature Expansion**: Exploring and engineering additional features could strengthen the model's predictive power and accuracy, potentially leading to a more reliable risk assessment tool.

9. Visualizations:

- Figures, including ROC curves and accuracy bar charts, provide clear insights into model performance differences. They visually demonstrate the effectiveness of Logistic Regression and highlight areas where other models could be improved.

By refining these approaches, a more accurate and reliable prediction system can be developed, contributing to early diagnosis and intervention strategies for individuals at risk of heart disease.