**Heart Disease Prediction**

**CBIO313: Data mining and Machine learning**

**Dr. Muhammed El-Sayeh**

**Made by:**

**Shahd Ebrahim Ahmed**

**ID:**

**211002064**

1. **Introduction**

Heart disease remains one of the leading causes of death globally, yet early detection is often missed due to the tendency of individuals to ignore symptoms or avoid costly medical examinations. As a result, diagnoses frequently occur at advanced stages, reducing the chances of successful intervention.

This project aims to provide a simple, accessible, and cost-effective solution to assess heart health using basic medical information. By leveraging supervised machine learning algorithms, the system can act as an early warning tool that encourages users to seek medical attention when signs of risk are detected.

Designed as a rapid screening aid, it also has potential applications in clinical settings to support physicians with preliminary evaluations. Furthermore, this approach could be commercialized as a user-friendly mobile or web application, promoting broader public awareness and facilitating early intervention in heart disease cases.

1. **Data Exploration & Preprocessing**
   1. This dataset, which includes both numerical and category variables, has 10,000 entries in 21 columns. Age, gender, blood pressure, cholesterol, habits, and medical history are among the health-related information it contains. "Heart Disease Status," which indicates whether or not the person has heart disease, is the target variable.
   2. Searching for unique values in the dataset; a feature is useless if it contains a large number of unique values. The outcome is that all object type features have zero unique values.

A screenshot of a computer

AI-generated content may be incorrect.

* 1. Count Plot Visualization of Categorical Features.[¶](https://www.kaggle.com/code/hossainhedayati/heart-disease-prediction-with-83-8-accuracy#1.3-Count-Plot-Visualization-of-Categorical-Features.)

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

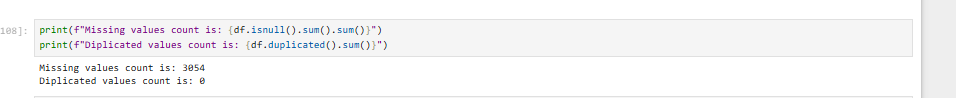
* 1. The distribution of numerical characteristics was examined using a box plot that was created here, exposing remarkable outliers and different feature scales. Certain characteristics, such as fasting blood sugar and triglyceride level, have extreme values that may have an impact on model performance. Before training, the significant variations in feature ranges suggest that normalization or standardization is necessary.

A screenshot of a graph

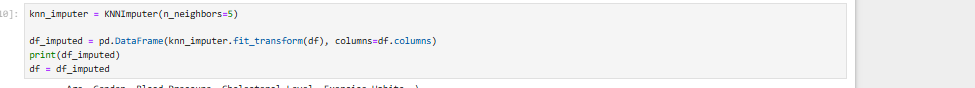
AI-generated content may be incorrect.

1. **Preprocessing**
   1. 12 categorical (object-type) variables must be assigned to numerical variables.

Searching null values and duplicated values:



* 1. KNNImputer from Scikit-learn is used in this section to deal with missing values in a dataset. When a value is missing, the KNNImputer(n\_neighbors=5) substitutes it by calculating the average of the five nearest neighbors for each missing entry.





* 1. Data Distribution Analysis

The previously mentioned histogram matrix shows how several health-related parameters are distributed, such as demographic information (age, gender), lifestyle choices (exercise, smoking, alcohol), and health indicators (blood pressure, cholesterol, diabetes, and more). Visualizing the frequency distribution of each characteristic allows us to see any outliers, skewness, or possible imbalances in the dataset. This preliminary investigation is crucial for comprehending the quality of the data, getting ready for more analysis, and figuring out whether data preprocessing methods like feature scaling or normalization are required.

A screenshot of a computer screen

AI-generated content may be incorrect.

This code creates a heatmap to show how the dataset's numerical features correlate with one another. It aids in determining correlations between variables, although in this instance, the attributes do not significantly correlate with one another.

A screenshot of a computer

AI-generated content may be incorrect.

* 1. This phase uses MinMaxScaler to scale all numerical characteristics to a [0, 1] range such that each one contributes equally to model training. Since the target variable (Heart Disease Status) is a classification label, it is maintained without scaling. Recombining the scaled features with the initial target column allows for additional analysis.

A screenshot of a computer

AI-generated content may be incorrect.

The distribution of the target variable Heart Disease Status is displayed in a bar plot, indicating a class imbalance with the majority of data falling into the no-disease (0) category. A biased model that favors the majority class may result from this imbalance, which would lower the predicted performance for the minority class (heart disease cases).

A screenshot of a graph

AI-generated content may be incorrect.

Risk:  
The model may be biased toward the majority class (0.0) with 80% class 0.0 and 20% class 1.0. This could result in poor prediction performance for the minority class (1.0) and unreliable metrics such as precision, recall, and F1-score.

Solution:   
To balance the dataset, class 1.0 synthetic samples are created using SMOTE. By ensuring that the model learns both classes equally, the minority class's accuracy, recall, and F1-score are all improved.

Why:

By preventing bias, data balancing improves the model's accuracy and predictability for both classes.

A screenshot of a computer program

AI-generated content may be incorrect.

The dataset analysis indicated that no more feature engineering was necessary. During initial testing, no significant changes or artificial features enhanced the model's performance; the preexisting variables already offered adequate predictive power.

1. **model training and hyperparameter tuning**
   1. To ensure each model is optimized, I will use GridSearchCV to find the best hyperparameters for Random Forest, SVM, Logistic Regression, KNN, Naïve Bayes, and Decision Tree Classifier. Grid search will systematically explore different parameter combinations and select the best-performing configuration based on cross-validation scores.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

* 1. This code evaluates six machine learning models using 5-fold cross-validation, reporting weighted precision, recall, and F1-scores to handle class imbalance. The Random Forest performed best (F1: 0.79), while Logistic Regression and Naive Bayes struggled. The zero\_division=0 parameter prevents errors when a model fails to predict certain classes. Warnings for the Decision Tree suggest some folds had imperfect predictions. Results help identify the most reliable algorithm (Random Forest) for deployment.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

* 1. In this evaluation, six models were compared for a binary classification task. RandomForestClassifier achieved the highest accuracy (83%) and the best weighted F1-score (0.81), balancing precision and recall effectively. Other models like SVC and KNeighborsClassifier performed moderately, while LogisticRegression, DecisionTreeClassifier, and GaussianNB showed weaker results, with GaussianNB failing to predict class 1.0 entirely.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a document

AI-generated content may be incorrect.

Why RandomForestClassifier is Best:

1. Highest accuracy and F1-score: Outperformed all other models.
2. Handles class imbalance: Maintained high recall for both classes.
3. No undefined metrics: Reliable and robust performance.

In the end :

The RandomForestClassifier is the best choice due to its superior accuracy, balanced performance, and ability to handle imbalanced data effectively.

1. **Comparison of Models**

A graph of different colored lines

AI-generated content may be incorrect.

Most Important Features for the best model[¶](https://www.kaggle.com/code/hossainhedayati/heart-disease-prediction-with-83-8-accuracy#Most-Important-Features-for-best-model)

A screenshot of a computer

AI-generated content may be incorrect.

This bar chart visualizes the feature importance scores from a Random Forest model, highlighting the most influential factors in the prediction. CRP Level, Homocysteine Level, Cholesterol Level, and BMI emerge as the most significant predictors, suggesting their strong correlation with the target outcome. Lifestyle factors like sleep hours, exercise habits, and alcohol consumption also contribute, though to a lesser extent. Features like gender and high LDL cholesterol have lower importance in the model. This insight helps in understanding key health indicators driving the predictions.

More Exploration of selected model

A screenshot of a computer

AI-generated content may be incorrect.

The ROC Curve for the Random Forest model shows an AUC of 0.83, indicating good performance in distinguishing between the classes. The Confusion Matrix reveals the model's predictions compared to actual labels, highlighting true positives, true negatives, false positives, and false negatives. This analysis helps evaluate the model's effectiveness in handling the imbalanced dataset and its ability to correctly classify instances.

Result:

The heart disease prediction model was developed using a Random Forest Classifier with hyperparameters such as max\_depth=20, min\_samples\_split=2, and n\_estimators=200. The dataset included key health indicators like age, blood pressure, cholesterol levels, and lifestyle habits.

The model's performance was assessed using:

Confusion Matrix: This provided insights into the number of correct and incorrect classifications. The heatmap visualization helped identify the balance between false positives and false negatives.

ROC Curve & AUC Score: The model achieved an AUC score of 0.83, indicating its ability to distinguish between patients with and without heart disease.

1. **Conclusion**

The project successfully demonstrated the application of various machine learning models for predicting heart disease. By addressing class imbalance with SMOTE and evaluating multiple algorithms, the Random Forest model emerged as the best performer. The model was then deployed as a Streamlit web app to enable real-time prediction based on user input. This approach highlights the potential of ML in supporting clinical decision-making.

**GUI using streamlit**

A screenshot of a computer

AI-generated content may be incorrect.