MAIS 202

Team #3

Anastasiia Nemyrovska, Shirley Ding, Michelle Sayeh, Lucie Shi

MAIS 202 - PROJECT DELIVERABLE 1

Choice of dataset

We selected the Heart Disease Prediction dataset for its comprehensive nature and ease of use. It includes essential patient health metrics such as age, cholesterol levels, blood pressure, and other indicators relevant to heart disease prediction which are used as the model's features. One of the main advantages of this dataset is that it contains no missing or mismatched data, making preprocessing straightforward. This allows us to focus on feature engineering and model development rather than data cleaning. The numerical format of the dataset's features (either continuous or binary) is also ideal for machine learning models, reducing the need for additional transformations. The target variable in this dataset is the "Heart Disease" column, which indicates whether a patient has heart disease (labeled as "Presence") or not (labeled as "Absence").

Methodology

We aim to build a machine learning model that predicts the likelihood of heart disease based on patient data. The dataset contains multiple features that are critical indicators of heart health. Our project will involve preprocessing the data, training a machine learning model, and evaluating its performance.

Data Preprocessing

Given that the dataset is already in a well-structured format, and contains no missing data, the preprocessing steps will focus on optimizing the dataset for model training rather than cleaning it. This includes:

- Label Conversion: The "Heart Disease" column, which contains "Presence" or "Absence" as values, will be converted into a binary format ("Presence" to 1 and "Absence" to 0). All other features are already in numerical format. This step will make the dataset ready for classification algorithms.
- Feature Scaling: Since the dataset contains continuous variables like age, cholesterol, and blood pressure, feature scaling will be applied. We will use techniques such as min-max normalization or standardization to ensure that all features are on the same scale. This is

crucial for machine learning algorithms, particularly those like logistic regression or neural networks, where the magnitude of the features can affect model performance. Without scaling, features with larger values might disproportionately influence the predictions, leading to biased results.

- Class Balancing: One potential issue we will address is class imbalance. In medical datasets, it is common to have more examples of healthy patients than those with a disease, and this can lead to biased models. If we observe an imbalance between the classes (i.e., more patients without heart disease than those with it), we will implement techniques like SMOTE (Synthetic Minority Over-sampling Technique) to ensure that the model does not favor the majority class. This step is important because class imbalance can lead to misleading accuracy results, where the model appears to perform well by simply predicting the majority class.

Machine Learning Model

From the chosen dataset, we want to predict the probability of heart disease given the features of the dataset. Since the desired output is the probability score for a binary classification, the suitable machine learning model is Logistic Regression.

Logistic Regression is a good model for small to medium-sized datasets where the features have a mostly linear relationship with the outcome. This model is fast to train and does not require extensive memory or hyperparameter tuning. As for the results, it is good at interpreting how certain features will influence the outcome, which is especially useful in healthcare for identifying risk-factors. However, the downsides of logistic regression include the need for clean data, meaning that outliers and missing values must be handled before training the dataset. Also, the features must have a linear relationship with the outcome of heart disease, as the model assumes that the relationships are all linear.

Another machine learning algorithm that was taken into consideration is the Random Forest model, which is known to provide accurate predictions, a critical factor in medical diagnosis. However, since Random Forest tends to work better on large datasets, it can still overfit smaller datasets if note tuned properly. Additionally, this model is more complex to train and requires more memory.

The model that aligns best with the objectives of this project is Logistic Regression. While accuracy is important, it is also crucial to interpret how the coefficients influence the prediction.

Since our dataset is small, Logistic Regression would be more suitable than random forest as it is less prone to overfitting in smaller datasets. Additionally, Logistic Regression is more time and memory-efficient. Therefore, Logistic Regression better meets our needs than Random Forest.

Evaluation Metric

Confusion Matrix

The confusion matrix provides a breakdown of predicted vs. actual values. It allows us to calculate other metrics like precision, recall, and accuracy. There are four values in the matrix. At the top left corner, we have the True Positives (TP), which indicates cases where the model correctly predicted heart disease. To its right, we have the False Positives (FP), which indicates the cases where the model incorrectly predicted the disease. At the bottom right, we have True Negatives (TN), which indicates cases where the model correctly predicted no heart disease. Finally, on its left, we have the False Negatives (FN), which indicates the cases where the model incorrectly predicted no heart disease.

Precision

Precision focuses on the proportion of the true positive predictions among all positive predictions. A precision close to or above 0.9 is desirable to reduce the number of false positives (incorrectly diagnosing heart disease).

Recall

Recall measures the proportion of actual positives that the model correctly identifies. A high recall (above 0.85) is crucial in heart disease detection because we want to minimize false negatives (cases where the disease is missed).

Specificity

Specificity measures the proportion of actual negatives that the model correctly identifies. High specificity (above 0.85) is also critical to avoid over-diagnosing patients who don't have heart disease.

F1 Score

The F1 score is the harmonic mean of precision (accuracy of positive predictions) and recall (representation of how well a model can identify actual positive cases). This metric is useful when both false positives and false negatives need to be minimized. A score above 0.85 would be acceptable for medical contexts.

Application

1. User Input

   The user will provide several medical and lifestyle-related inputs in the form of text fields. These inputs might include:

   - Age
   - Sex
   - Resting blood pressure
   - Cholesterol level
   - Maximum heart rate achieved

   This can be done via a simple web form with labeled input fields where users type or select their information.

2. Output

   The output will be a prediction indicating the likelihood of heart disease. After the user submits their data, the model will process the inputs and return a probability score (e.g., 85%) or a classification (e.g., "High risk" or "Low risk" for heart disease).

   The result will be displayed directly on the web page in an easy-to-understand manner, such as:

   - A color-coded response (e.g., green for low risk, red for high risk).
   - A percentage likelihood with explanatory text (e.g., "Based on your inputs, you have a 75% chance of developing heart disease").

   Additionally, the web app can provide some basic health advice or suggest next steps, such as seeing a doctor for further examination.