**A close-up of a human heart

Description automatically generatedHeart disease prediction**

**Team members:**

|  |  |
| --- | --- |
| Name | id |
| Taha mohamed | **21-01656** |
| Sama mostafa Ibrahim | **21-01553** |
| Renad Mohamed fawzy | **21-01502** |
| Selwan Ezzat mohamed | **21-01589** |
| Mahmoud adel | **21-00940** |

**Teaching Associate**: Eng... Rana Mohamed

**Dr: Mahmoud Bassioni**

**Project overview(main topics):**

* **Dataset Description**
* **preprocessing steps**
* **choice of machine learning algorithms**
* **evaluation of our model**
* **creation of GUI**

**The idea of the project:**

Predict heart disease based on patient data, by building a predictive model to forecast disease outcomes.

**Let’s talk briefly about project:**

We used suitable machine learning model and trained it on large datasets containing information from patients with and without heart disease to learn patterns and make accurate predictions, These models learn from historical patient data, including medical history, diagnostic tests, and lifestyle factors, to identify patterns and relationships, By analyzing these patterns and use the appropriate algorithms such we use in this project (Decision Tree, Logistic regression and Random Forest).

In evaluation phase, to measure the performance of our model we used some techniques such as (accuracy ,precision, recall and F1 score) on a sperate validation dataset to evaluate the model, and we satisfied with model performance(also on unseen data) and its accuracy !

**Dataset Description:**

The dataset used for training our machine learning model is typically sourced from reputable sources such as healthcare databases, research studies, or public repositories.

It includes a collection of individual patient records, with each record containing various features or attributes that serve as inputs to our predictive model.

The dataset is split into two subsets: a training set used to train the model and a test set used to evaluate its performance.

**Features Included in the Dataset:**

Common features included in the dataset for predicting heart disease

(age ,sex ,cp, trestbps,chol,fbs,restec ,thalach ,exang and oldpeak)lets explain these features briefly:

-**Age:** Age of the patient in years.

**-Sex:** Gender of the patient (binary variable: male or female).

**-Chest Pain Type(CP):** This feature categorizes the type of chest pain experienced by the patient.

**-Resting Blood Pressure (Trestbps):** blood pressure measured in millimeters of mercury (mmHg) at rest

**-Cholesterol Levels (Chol):** Total serum cholesterol levels measured in milligrams

**-Fasting Blood Sugar (FBS):** Blood glucose levels measured after a period of fasting,Elevated fasting blood sugar levels may indicate insulin resistance, diabetes, which are risk factors for heart disease.

**-Resting Electrocardiogram (Restecg):** Electrocardiogram (ECG) results obtained at rest, which provide information about the heart's electrical activity.

**-Maximum Heart Rate Achieved (Thalach):** Maximum heart rate achieved during exercise testing

**-Exercise Induced Angina (Exang):** Presence or absence of angina (chest pain) induced by physical exercise.

-ST Depression Induced by Exercise Relative to Rest (Oldpeak)

These features provide valuable information for assessing an individual's cardiovascular risk profile .Machine learning model can use these features to build predictive models that assist in early detection and personalized management of heart-related conditions.

we do some preprocessing steps performed on the data like ( handling missing values, feature scaling, encoding categorical variables)lets talk briefly about them:

**Handling Missing Values:**

Missing values are a common issue in real-world datasets and can affect the performance of machine learning models. to handling missing values is imputation, where missing values are replaced with estimated values based on the available data. This could involve replacing missing numerical values with the mean, median, or mode of the respective feature.

**Feature Scaling:**

Feature scaling is important for ensuring that numerical features are on a similar scale, preventing features with larger magnitudes from dominating the model training process. Common scaling techniques which we use is StandardScaler (to standardize numerical features by scaling them to have a mean of 0 and a standard deviation of 1) useful when the features in the dataset have different scales or units StandardScaler ensures that they have similar scales and allows machine learning algorithms to converge more quickly during training. The mathematical formula for standardization using the StandardScaler is:

Standardized Feature=**(**feature-mean**) ÷ (**standard deviation**)**Machine learning algorithms typically require numerical inputs, so categorical variables need to be encoded into numerical values. One-hot encoding is a

**Encoding Categorical Variables:**

common technique used for handling categorical variables with multiple categories. It creates binary dummy variables for each category, where a value of 1 indicates the presence of the category and 0 indicates its absence.

**Clearing the duplicated data:**

Clearing duplicated data from a dataset involves identifying and removing duplicate observations or rows to ensure data integrity and prevent redundancy.

By performing these preprocessing steps, the dataset is cleaned, standardized, and transformed into a suitable format for training machine learning models. This enhances the model's ability to learn from the data effectively and make accurate predictions on unseen data.

**choice of machine learning algorithms for our project of predicting heart disease:**

For predicting heart disease, the choice of machine learning algorithms depends on various factors, including the nature of the data, the size of the dataset, the desired interpretability of the model, and the trade-offs between accuracy and complexity, here it is the algorithms we used in our project:

**1-Decision Trees:**

**-**Decision trees are non-linear models that partition the feature space into regions based on the values of the input features.

**-**They are easy to interpret and visualize, making them useful for understanding the decision-making process of the model.

**-**Decision trees can handle both numerical and categorical features and are robust to outliers and irrelevant features.

**2-Logistic Regression:**

**-**Logistic regression is a simple and interpretable classification algorithm that is well-suited for binary classification tasks like predicting the presence or absence of heart disease.

**-**It models the probability of a binary outcome based on one or more predictor variables (features) using a logistic function.

**-**Logistic regression can provide insights into the relationship between individual features and the likelihood of heart disease.

**3-Random Forest:**

**-**Random forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and reduce overfitting.

**-**It builds multiple decision trees using bootstrapped samples of the dataset and random subsets of features at each split.

**-**Random forest is highly effective for handling high-dimensional datasets with complex interactions between features.

The choice of algorithm were be based on empirical evaluation, considering factors such as model performance, computational resources, interpretability, and the specific requirements of the project. And these algorithms were the most suitable for our model for predicting heart disease.

**For evaluation of our model**

we used some techniques such as (Precision, Recall, F1 Score) commonly used for classification tasks, including predicting heart disease.

Let’s talk briefly about these techniques :

**1- Precision:**

**-**Precision is a measure of the accuracy of the positive predictions made by the classifier.

**-**It is calculated as the ratio of true positive predictions to the total number of positive predictions made by the classifier.

**-**The mathematical formula for Precision is :

Precision=**(**true positives**) ÷** **(**true positives **+** false positives**)**

**2-Recall :**

**-**Recall, also known as sensitivity or true positive rate, measures the classifier's ability to correctly identify positive cases from all actual positive cases in the dataset.

**-**It is calculated as the ratio of true positive predictions to the total number of actual positive cases in the dataset.

**-**The mathematical formula for Recall is :

Recall= **(**true positives**) ÷ (**true positives **+** false positives**)**

**3- F1 Score:**

**-**It is calculated as the weighted average of precision and recall, with equal importance given to both metrics.

**-**The F1 score ranges from 0 to 1

**-**The mathematical formula for F1 Score is :

F1 Score = 2**(**Recall **×** Precision**) ÷ (**Recall **×** Precision**)**

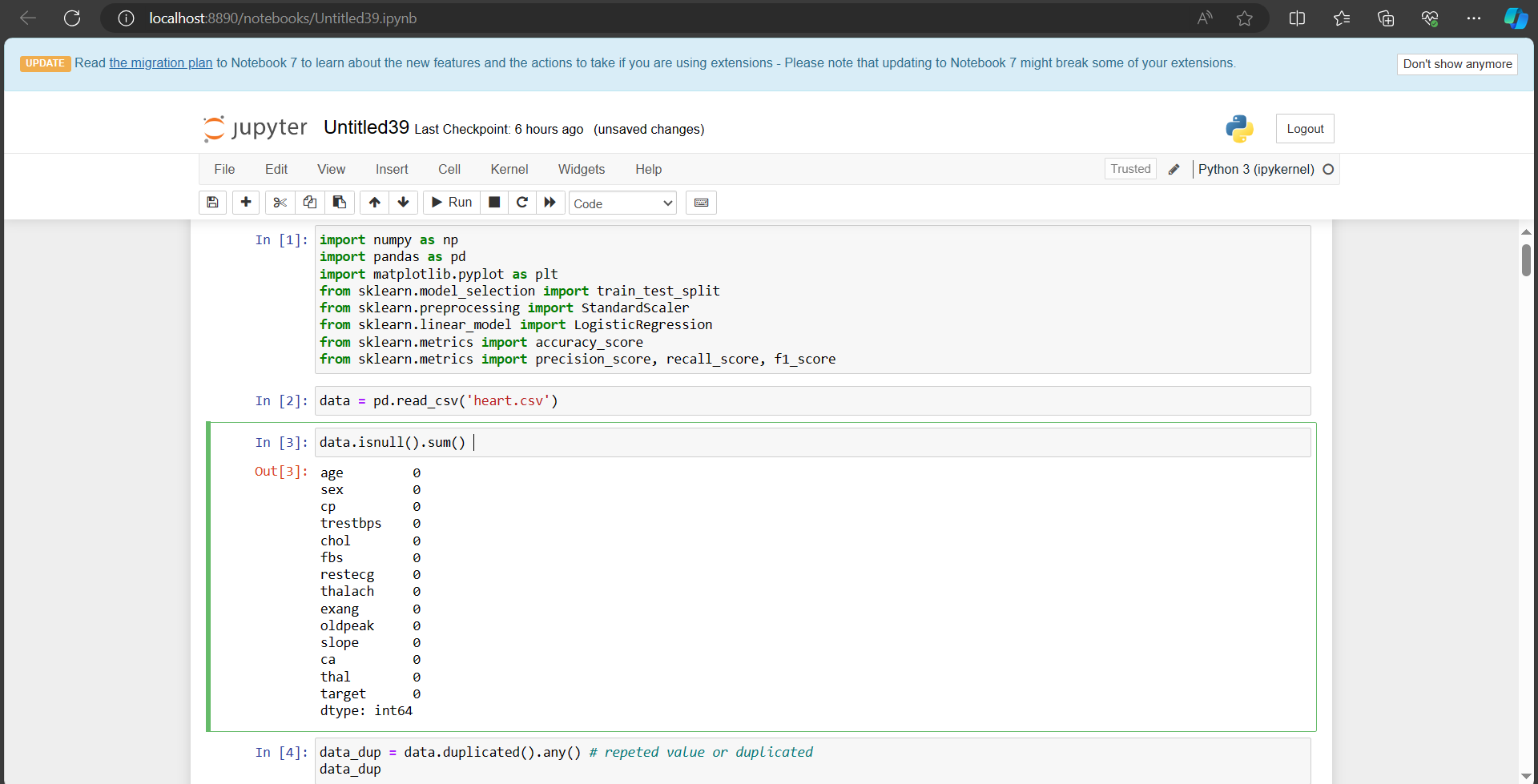
**When we applied these techniques to evaluate our model ,it gives us satisfied results that ensure that our model trained well and his performance in unseen data is great!**

We create **GUI** that allows users to input their clinical data related to heart disease risk factors and obtain a prediction of the likelihood of having heart disease based on a trained machine learning model. After the user enter these data and press the button (predict) the prediction result will display,

If the model predicts no heart disease, it displays "No Heart Diseases"; otherwise, it displays "Possibility of Heart Diseases".

**Link of the code (GitHub):**

THE RUNING OF THE PROJECT:

**-Import the libraries and loading dataset**

**-Preprocessing steps**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**A screenshot of a computer

Description automatically generated** **-describing and get information about dataset**

A screenshot of a computer

Description automatically generated

**A screenshot of a computer

Description automatically generated-splitting the Features and Target (split dataset)**

A screenshot of a computer

Description automatically generated

**-Decision tree algorithm and its accuracy**

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

**-Random forest algorithm and its accuracy**

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

**A screenshot of a computer

Description automatically generated-Logistic Regression algorithm and its accuracy**

A screenshot of a computer

Description automatically generated

**A screenshot of a computer

Description automatically generated-prediction phase**

**A screenshot of a computer

Description automatically generated-Creating GUI**

A screenshot of a computer

Description automatically generated

**-GUI and the predicted result**

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

**Thank You….**