**Machine learning assignment**

**Hazem medhat**

**ID:211001994**

**Link of the recording:** https://nileuniversity-my.sharepoint.com/:v:/g/personal/h\_medhat2194\_nu\_edu\_eg/ER8AWB4JjdRNiZirbTGNct0BwRNHLhZ8kaU0z\_OY6rZneQ?referrer=Teams.TEAMS-ELECTRON&referrerScenario=MeetingChicletGetLink.view.view

Brief summary about the dataset:   
The dataset contains information about patients including their age, sex, chest pain type, resting blood pressure (Resting), cholesterol levels, fasting blood sugar (FastingBS), resting electrocardiographic results (Restringing), maximum heart rate achieved (MaxHR), exercise-induced angina (Exercise Angina), ST depression induced by exercise relative to rest (Old peak), ST segment slope (ST\_Slope), and whether they have heart disease (Heart Disease).

* **Task1/1.1**: I wrote Python code iterates through each column in a Data Frame and prints the unique value and ser if there any missing values present in that column along with the column name. The output displays various characteristics related to heart health, such as age, gender, chest pain type, blood pressure, cholesterol levels, and electrocardiographic results. Additionally, it provides insights into exercise-induced angina, ST segment slopes, and the presence of heart disease, aiding in the analysis of heart-related data.
* **Task1.3**: I wrote a code to create a histogram to visualize the distribution of different types of chest pain (ChestPainType) in the dataset. It uses Seaborn's histplot function.

A diagram of a column

Description automatically generated

The second code snippet generates a pairplot, which is a grid of scatterplots showing the relationships between different numerical columns in the dataset. It provides a quick overview of the relationships between variables.

A screenshot of a graph

Description automatically generated

The third code generates a correlation heatmap using Seaborn's heatmap function to visualize the correlation between these numerical columns. Each cell in the heatmap represents the correlation coefficient between two variables.

A screenshot of a graph

Description automatically generated

* **Task2**: I wrote a code first to generate a new feature, 'BP\_Cholesterol\_Ratio', by dividing the resting blood pressure ('RestingBP') by the cholesterol level ('Cholesterol'). It then handles missing and infinite values, replacing them with the median of the feature.

Next, it preprocesses the data, converting categorical variables into a binary vector format and scaling numerical features. The data is split into training and testing sets, the model is trained on the training data, and predictions are made on the test data. Finally, the model's accuracy on the test data is printed score of approximately 89.44%.

**Accuracy of logistic regression: 89.44%**

**Task3**: I wrote code segments aims to build and evaluate a predictive model for heart disease classification using a Random Forest Classifier. defines preprocessing steps for numerical and categorical features separately. For numerical features, it imputes missing values with the median and scales the features using StandardScaler.

Next, a model pipeline is created, consisting of preprocessing and a Random Forest Classifier model. The Random Forest Classifier is trained with 100 decision trees. The model is then fitted to the training data, and predictions are made on the test data.

Finally, the model's accuracy and classification report, including precision, recall, and F1-score, are printed. The model achieves an accuracy of approximately 88.89% on the test data.

**Accuracy of classification: 88.89%**

* **Task 3.3.4.5:** In this code I evaluate the performance of three different machine learning models—Logistic Regression, Random Forest, and Support Vector Machine (SVM)—for heart disease classification, calculating the mean accuracy and standard deviation. Finally, it displays the mean accuracy and standard deviation for each model. The output shows that all three models achieve similar mean accuracies, **with Random Forest slightly outperforming the others**.
* **Accuracy of random forest: 83%**
* **Accuracy of logistic regression: 82%**
* **Accuracy of svm: 82%**
* **And this is a barplot showing all the models and their accuracy.**

**Task4:** It first encodes categorical columns using LabelEncoder, then splits the data into training and testing sets, it defines a parameter grid for hyperparameter tuning, performs GridSearchCV to find the best hyperparameters, and evaluates the best model's accuracy on the test set.

**Task5:** This Streamlit application predicts the presence or absence of heart disease based on user input. It loads a dataset containing heart health data and provides a user interface for inputting features such as sex, chest pain type, resting blood pressure, cholesterol level, and fasting blood sugar. Upon clicking the prediction button, the app displays whether the user is predicted to have heart disease or not.

**A screenshot of a computer

Description automatically generated**

**A graph of blue lines

Description automatically generated with medium confidence**

**Limitations:** Addressing these limitations by collecting more data, balancing the dataset, including more relevant features, handling missing data appropriately, and ensuring diversity in the dataset can help improve the accuracy and reliability of the predictive model.

**Conclusion:** In summary, the dataset offers useful information for heart disease prediction based on a variety of patient features. To develop a precise and trustworthy predictive model, however, a few of its limitations must be fixed. A small sample size, an unbalanced class distribution, missing key traits, incomplete data, little diversity, and possibly out-of-date information are some of these problems. By addressing these constraints via feature engineering, data preparation, validation of the model, and data collecting, the predictive model's accuracy and generalizability can be increased. With these limitations, the dataset is a useful place to start creating a heart disease prediction model.