" بسم الله الرحمن الرحيم"

Electrical and Computer Engineering.

Artificial intelligence - ENCS3340.

1st Semester| 2023-2024

Course Project: COVID-19 Diagnosis.

Group members (Prepared by):

Razan Abdalrahman 1200531

Lana Hamayel 1200209

Instructors: Dr. Adnan H. Yahya.

Section: 2.

Date: January | 2024.

❖ Specifications.

I. Tools used.

- Weka 3.8.6 was used for preprocessing, classifying, and simulating the testing results.
- “scikit-learn”, often abbreviated as sklearn, is a popular machine learning library in Python, was used in the python code.

II. Test set used.

Despite several attempts, a dataset could not be obtained from the medical center in Ramallah due to patient privacy concerns. Inquiries were made regarding the procedures followed by them to determine whether a patient has COVID or not. Subsequently, a dataset was created, incorporating information acquired from the medical center. This dataset includes details such as Patient_ID, Age, Gender, Location, Fever, Cough, Shortness_of_Breath, Fatigue, Loss_of_Taste_or_Smell, Diabetes, Hypertension, and COVID_Test_Result.

III. Algorithms utilized

Two machine learning algorithms, Decision Tree and Artificial Neural Network (ANN), are employed for COVID-19 analysis and prediction. The Decision Tree classifier is a tree-like model that makes decisions based on feature conditions at each node, creating a hierarchical structure for classification. It is interpretable and well-suited for categorical data. On the other hand, the ANN, specifically a Multi-Layer Perceptron (MLP), is a type of neural network with multiple layers of interconnected nodes, allowing it to capture complex patterns in data. ANN excels in handling non-linear relationships and is particularly effective for tasks involving numerical and continuous features. Both algorithms are trained on labeled data to learn patterns, enabling the system to make predictions on new, unseen instances. The script evaluates and displays the performance metrics of these models for COVID-19 analysis and provides a user interface for predictions.

❖ Procedure and discussion.

The Python code utilizes the “scikit-learn” library for machine learning and tkinter for creating a graphical user interface (GUI). It begins by importing necessary libraries and defining a function (train_and_evaluate_classifier) to preprocess data and evaluate machine learning models. The data, read from a CSV file (covid_data.csv), is preprocessed using label encoding. Features and target variables are defined, and the data is split into training and testing sets. Two classifiers are created: a Decision Tree classifier (decision_tree_classifier) and a Multi-Layer Perceptron (MLP) classifier (ann_classifier). These classifiers are trained on the respective datasets.

The script also defines functions for analyzing and displaying results using both classifiers. Additionally, there is a function (open_covid_prediction_window) for user input of COVID-related information, which is then used to predict COVID results using the trained models. The GUI is created using tkinter, including buttons for Decision Tree analysis, ANN analysis, and COVID prediction.

In summary, the script combines data preprocessing, machine learning model training, and GUI development to provide a user-friendly interface for COVID analysis and prediction using Decision Tree and ANN techniques.

❖ Results.

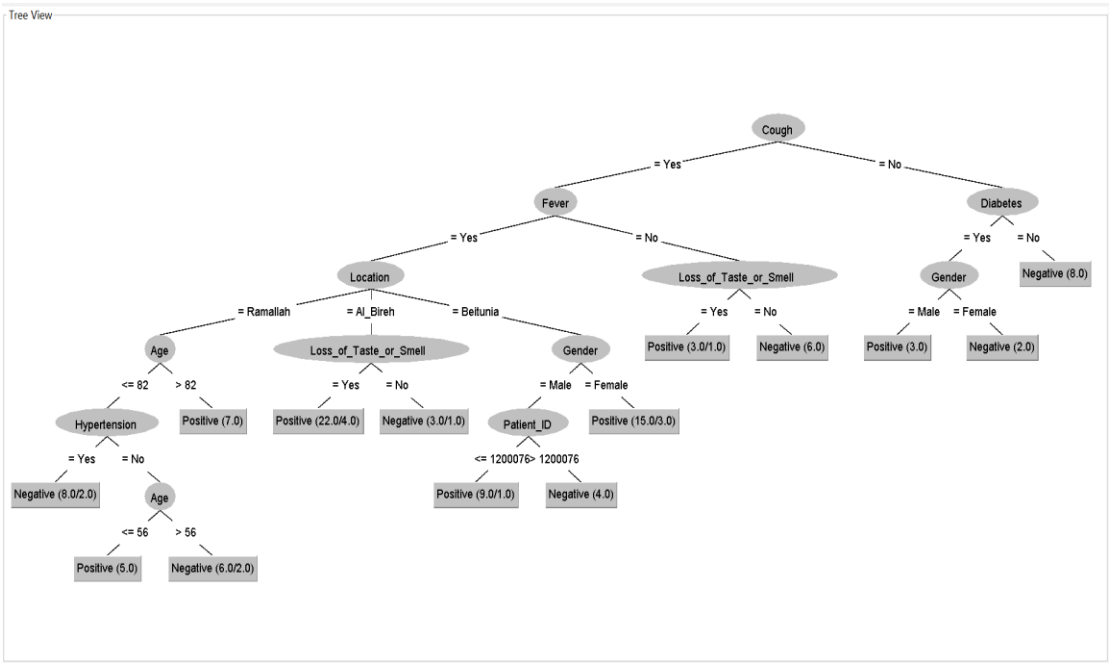
- Results from WEKA:

Results from Decision tree algorithm:

Classifier output

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          66               65.3465 %
Incorrectly Classified Instances        35               34.6535 %
Kappa statistic                          0.2246
Mean absolute error                      0.3966
Root mean squared error                  0.4936
Relative absolute error                 82.1657 %
Root relative squared error            100.4985 %
Total Number of Instances              101

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.867     0.659     0.658      0.867     0.748      0.248    0.635     0.678     Positive
               0.341     0.133     0.636      0.341     0.444      0.248    0.635     0.608     Negative
Weighted Avg.  0.653     0.445     0.649      0.653     0.625      0.248    0.635     0.650

=== Confusion Matrix ===

  a  b   <-- classified as
 52  8 |  a = Positive
 27 14 |  b = Negative
```

Decision tree shape:



Results from ANN algorithm:

Classifier output

```
Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          60               59.4059 %
Incorrectly Classified Instances        41               40.5941 %
Kappa statistic                          0.168
Mean absolute error                      0.4147
Root mean squared error                  0.5833
Relative absolute error                 85.91   %
Root relative squared error            118.7602 %
Total Number of Instances              101

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.633    0.463    0.667      0.633    0.650      0.168    0.582     0.620     Positive
                 0.537    0.367    0.500      0.537    0.518      0.168    0.582     0.519     Negative
Weighted Avg.    0.594    0.424    0.599      0.594    0.596      0.168    0.582     0.579

=== Confusion Matrix ===

  a  b   <-- classified as
 38 22 |  a = Positive
 19 22 |  b = Negative
```

- Results from " scikit-learn " library: is shown by details in the video provided in the folder named "testCases".

❖ Conclusion.

The observation that increasing the size of the training set leads to more positive results in the testing set suggests that our machine learning models, particularly the Decision Tree and Artificial Neural Network (ANN) classifiers, benefit from a larger and more diverse set of training examples. A larger training set allows the models to better generalize patterns and relationships within the data, potentially improving their ability to accurately predict COVID-19 outcomes. This finding underscores the importance of robust training data in enhancing the performance and reliability of our predictive models, reinforcing our commitment to obtaining and utilizing comprehensive datasets for more effective COVID-19 analysis and prediction within our project.