***AI PROJECT(REPORT)***

***Tumor Cancer prediction***

**Supervised by: TA. Alaa Tarek**

**Team Members:**

* **ا**
* **ا**
* **ا**
* **ا**
* **ا**
* **ا**
* **ا**
* First importing all libraries that needed in the project.
* Beginning the preprocessing of the data by:
  1. import the CSV dataset from pandas library by using [read\_csv ()] function.
  2. Dropping all null values by using [dropna ()] function.
  3. Removing all duplicate values by using [drop\_duplicates ()] function.
  4. Changing (diagnosis) column to (0 and 1) instead of (B and M) to be able to normalize the dataset after that by using [LabelEncoder] library.
  5. Normalizing the data by using [MinMaxScalar] library.
  6. splitting the dataset into independent dataset which contain all columns from F1 to F30 columns and dependent dataset which contain [diagnosis] column.
  7. splitting our dataset in 75% training our model and 25% to test our model by using [train\_test\_split] library.
* Then here used four algorithms (logistic regression, decision tree, svm, random forest classifier) and import these libraries from sklearn library.
* train our model by using these four algorithms and print the accuracy score by using our training dataset.
* Then import the metrices from sklearn library to get the confusion matrix from it and get some measurement as accuracy which is calculated by [(tp+tn)/(tp+tn+fp+fn)] and precision which is calculated by [tp/(tp+fp)] and Recall which is calculated by [tp/(tp+fn)] by using our test dataset.
* In SVM algorithm there is three kernels used to train our model (linear, polynomial, radial basis function(rbf)) but linear kernel is the most kernel who print the high accuracy so choose it.
* Then predict by four algorithms the model to see who will get tumor in every algorithm.
* Then make voting in our dataset between the three algorithms (logistic regression, decision tree, svm) to be sure who will get tumor by see every prediction in each algorithm what it predicts and the more prediction we decide and print it.
* Then use [joblib] library to save and load model for the four algorithms and print their accuracy score and we do that to make it easier to use the three algorithms to predict without train the model again.
* Then we make a new dataset to test it by using the loading models and see if we can use algorithms and predict without train it again.
* Then predict by four algorithms the new model to see who will get tumor in every algorithm.
* Then make voting in our new dataset between the three (logistic regression, decision tree, svm) algorithms to be sure who will get tumor by see every prediction in each algorithm what it predicts and the more prediction we decide and print it.
* Then use data scaling to print the normalization of the model which equal [Z = (X - Mean)/Variance] by import [Normalizer] from [sklearn. preprocessing] the fit the data then using distribution plot by using [seaborn] library and we print the standardization also with normalization by make [KDE = True] and print the plot by [matplotlib] library