TUMOR CANCER PREDICTION

TA: ALAA TAREK

ARTIFICIAL INTELLIGENCE

TEAM MEMBERS	ID	SECTION
AMIRA YASSER IBRAHIM HUSSIEN	20201700150	SECTION 6
AHMED ISMAIL MAHMOUD SAYED	20201701045	SECTION 1
AHMED MAHMOUD AWAD AHMED	20191700071	SECTION 4
MIRA MICHEAL SAMIR ESAAK	20201701226	SECTION 30
SARA MOHMED ABDELMONEM MOUSSA	20201700334	SECTION 12
HANY MOHMED ALI BADR	20201700952	SECTION 32
MAYAR HESHAM MOHMED GALAL	20201700878	SECTION 30

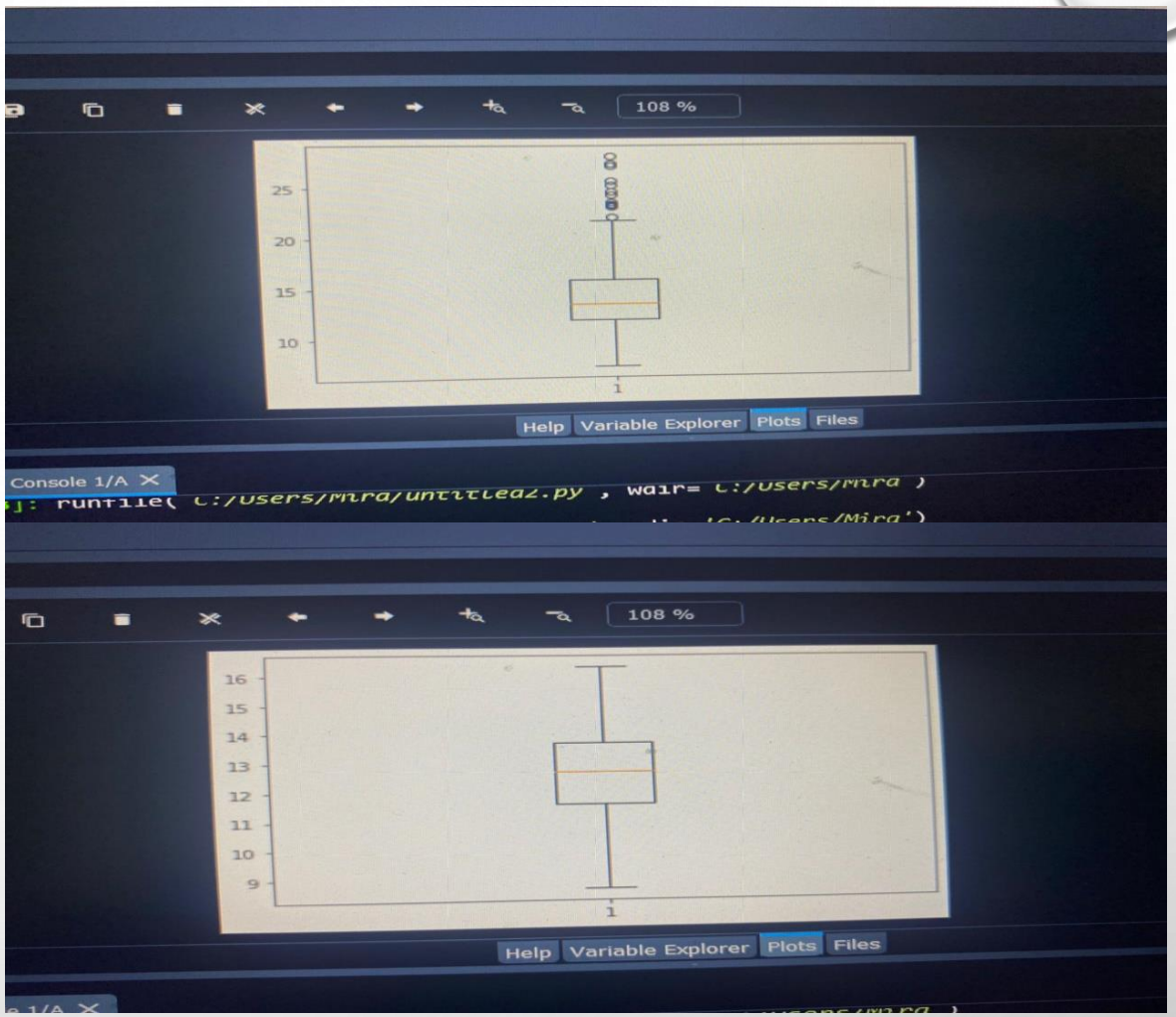
Tumor cancer prediction predict if the patient have cancer or not so in the order of making it happen we made some steps to predict it.

First step is preprocessing the first data set to make the data clean and free of null values and duplicates. We replaced the values in the column diagnosis to zeros and ones to turn it into numerical data (B for 0 and M for 1) if the diagnosis is 0 then the patient doesn't have cancer and if the diagnosis is 1 then the patient has cancer then we created variable to put all the features except the diagnosis Then we found out that there are so many outliers in the dataset so we created a function to calculate the outliers and remove them.

As an example F2 had some outliers

Before removing outliers

After removing outliers



Then we made x equals the input data (the first 30 features) and y equals the output (the last feature)
Then we split the data to 75% train which is used to fit the machine learning model and 25% test which is used to evaluate the machine learning model.

Then the second step is classification:

We used three models to test their accuracy. We made a function for every model

First one: is SVM which is used to solve regression problems. It uses a technique called the kernel trick to transform the data and then based on these transformations it finds an optimal boundary between the possible outputs. We used linear kernel in the SVM because it's used when there are large number of features in the dataset and to separate the data using single line then we fit the data so we can train the data. Then we predict x_train. Then we print the accuracy score of the train data that takes two parameters y_train and y_train_pred. then we print the accuracy score of the test data which takes two parameters x_test and y_test then we printed the classification report of y_test and the prediction of x_test.

SVM train data Accuracy: 0.951515151515151515

SVM test data accuracy: 0.9464285714285714.

We printed the classification report of y_test and the prediction of x_test. We calculated the recall, precision, f1score and support to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report. After that we saved the model SVM to use it to predict the new dataset.

Second model is: Logistic regression we predict x_train. Then we print the accuracy score of the train data that takes two parameters y_train and y_train_pred. then we print the accuracy score of the test data which takes two parameters x_test and y_test then we printed the classification report of y_test and the prediction of x_test.

Logistic regression train data Accuracy: 0.951515151515151515

Logistic regression test data accuracy: 0.9285714285714286.

We printed the classification report of y_test and the prediction of x_test. We calculated the recall, precision, f1score and support to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report. After that we saved the model logistic regression to use it to predict the new dataset.

Third model is: Decision tree is used to create a training model that can use to predict the class or value of the target variable

Then we fit the data so we can train it. . Then we predict x_train. Then we print the accuracy score of the train data that takes two parameters y_train and y_train_pred. then we print the accuracy score of the test data which takes two parameters x_test and y_test .

decision tree train data Accuracy: 1.0

decision tree test data accuracy: 0.9107142857142857.

then we printed the classification report of y_test and the prediction of x_test. We calculated the recall, precision, f1score and support to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report.

After that we saved the model decision tree to use it to predict the new dataset.
Then we created csv file that has all the features to test some data then we did the preprocessing Filled null values in the dataset with zeros and removed the duplicates then we calculated the outliers and removed them. Before the predicting of the new dataset we have to load the models we saved Then we predicted the new dataset by the three models because we can't split the new dataset.
We created the voting module which that trains on the three models and predicts an output based on their highest probability of chosen class as the output then we saved the module so we can use it to predict the new dataset we created list to put the prediction in it then we compared y of the dataset and y_prediction to see the accuracy.