HACKATHON LINK - https://www.kaggle.com/t/a7bf10b9003b4bbca94a00243c48359f

HERE I HAVE WRITTEN A SUMMARY OF WHAT I HAVE DONE THROUGHOUT THIS PROJECT.

1). Import the dataset.

2). Split the dataset into 2 sets of first 150 and next 250 elements.

3). Now, we do **EDA (Exploratory Data Analysis)** which includes determining :-

i). Shape of dataset, head and tail elements.

ii). No. of columns containing missing values.

iii). The Value Count of Class and plotting it against the Number of Patients of each class.

iv). Replace all the null values with the mean value of that respective column

v). Splitting the Dataset into Features (X) and Target (y).

4). Split the dataset into 20% testing and 80% training dataset.

5). Now we normalize our training and testing dataset so as to decrease bias and increase performance.

6). Pick the top 300 most relevant features by Mutual Information Score. I have also plotted the bar graph between top 50 features V/S their MI Score.

7). Now, our data is ready to fit into various ML models. Here, I have used :-

i). **RANDOM FOREST CLASSIFIER :-**

- The One V/S Rest Classifier is used to simplify the Multidimensional Classification problem into multiple binary decisions, each comparing one class to all others.
- We calculate the accuracy of our model and also print the Classification Report which includes Precision, Recall, F1 – Score and number of samples which support that particular class.
- We draw a confusion matrix to see our model’s prediction V/S Actual Value. I have also visualized it using seaborn.
- To plot ROC curves for each class, we binarize the target labels to match the One V/S Rest format of the predicted class probabilities.
- Now, we load our test data (401 samples for prediction) and pre – process it which includes handling missing values, normalizing our testing dataset features and predict on behalf of those 300 most relevant features which we chosen during training our model.
- We predict the class of the testing data using our RF model.
- Finally, we create a submission file with Columns - Id and Class (Predicted).

ii). **LOGISTIC REGRESSION :-**

- We again use One V/S Rest Classifier.
- Check model's accuracy and report to see precision, recall, F1 – Score and support samples for different classes.
- Plot confusion matrix and visualize it using seaborn.
- Plot ROC curve for each class.
- The LR model is applied to the pre - processed test data previously prepared during the Random Forest classification to generate class predictions.
- Finally, we create a submission file with Columns - Id and Class (Predicted).

REASONS FOR SELECTING ONLY LOGISTIC REGRESSION AND RANDOM FOREST CLASSIFIER :-

1). Both Random Forest and Logistic Regression handle high-dimensional data efficiently. After selecting the top 300 most relevant features using Mutual Information, Random Forest uses feature bagging internally ensuring performance without overfitting.

2). Random Forest helps us find out which genes are most important for predicting cancer. Logistic Regression, on the other hand, helps us understand how each gene affects the prediction, whether it increases or decreases the chances of a certain cancer type, allowing clear understanding of each feature's impact on class prediction.

3). Models such as KNN and Naive Bayes often perform poorly in handling high-dimensional data due to the curse of dimensionality where all points start to look equally distant, irrelevant features overpower useful ones and model could not learn effectively.

4). SVM and neural networks need careful tuning of things like kernel type, learning rate, and layers while models like RF and LR are easier and quicker to use in real world problems.

--- THE END ---