**Data Science Project:**

**Introduction on Gastric cancer and its severity:**

* Stomach cancer, also called gastric cancer, begins when cells in the stomach start to grow out of control.
* Gastric carcinogenesis is a multistep and multifactorial process.
* Age, diet, and stomach disease can affect the risk of developing gastric cancer.

**Symptoms of Gastric cancer:**

* Difficulty swallowing
* Feeling bloated after eating
* Feeling full after eating small amounts of food
* Heartburn

**Problem statement:**

The problem statement is to identify if a particular individual has Gastric cancer or not. In the case, they are found to have contracted Gastric cancer, they will be classified into one of the two types of Gastric cancer,

The three classes in the dataset are,

1. **DIFFUSE**
2. **INTESTINAL**
3. **NORMAL**

**Methods:**

Below mentioned are the steps involved in this data science project,

1. Explore the data and find patterns and insights,
2. Perform Feature selection, to find the most important features
3. Split the prepared dataset into training set and Test set,
4. Build the machine learning algorithms onto the dataset,
5. Evaluate the performance of machine learning models.

As specified, all the three types of classes can be classified using various classification machine learning techniques, and the four classification algorithms used in this project are,

* LOGISTIC REGRESSSION
* DECISION TREE
* NAÏVE BAYES
* K-NEAREST NEIGHBOR

**NOTE:** *Since, the medical related data are collected on the assurance that non-authorized persons won’t be able to access patient's data.*

* *The independent features of the dataset are coded in a way to uphold this assurance made by the authority collecting the data.*
* *Due to this, we won't be able to fully grasp each feature's characteristic.*

**DATASET:**

* The dataset comprises of gene expression profiles of various individuals spread across 4524 features.
* The dataset consists of 30 rows.
* “class” feature in the dataset is the dependent variable.
* The dependent variable consists of three classes

1. Normal
2. Diffuse
3. Intestinal

* Since, the dependent variable consists of three classes, this is understood to be a multi-class classification problem.

**Analysis:**

Before building the machine learning models onto the dataset, it is necessary to understand the data, so I’ve applied basic preprocessing techniques and Feature selection is done using ANOVA test. Since, the dataset contains above 4524 features.

Therefore, by building the classification models using sklearn library in python, I have created four classification models, their performance is evaluated by using performance metrics called Confusion matrix, Classification report and Accuracy.

* The Accuracy score is used to determine the performance of the model.
* The Classification Report consist of several components such as Precision, Recall, F1-score etc., which also determines the quality of the model.
* The Confusion Matrix is a matrix used to determine the performance of the classification models for a given set of test data.

**Results:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison of performance metrics** | | | | |
| **Classification Algorithms** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Logistic Regression | 62% | 0.55 | 0.62 | 0.58 |
| Decision Tree Classifier | 87.5 % | 0.94 | 0.88 | 0.89 |
| Naïve Bayes | 75 % | 0.88 | 0.75 | 0.80 |
| K Nearest Neighbors | 50 % | 0.40 | 0.50 | 0.44 |

**Conclusion:**

The goal of classification is to accurately predict the target class for each case in the data. In order to believe any predictive model, the accuracy of the model must be estimated. We noticed that Decision Tree Classifier model showed the highest accuracy.

As Decision Tree Classifier gives the overall accuracy of 87.5%, it is more significant than comparing all the other 3 classification models.