Project Abstract: -

Topic: - Personalized Cancer Diagnosis using Machine Learning

Team Members: -

1. Karun (203350008) 203350008@iitb.ac.in

2. Sachin (203180019) 203180019@iitb.ac.in

3. Yashu Chaudhary (203350010) 203180010@iitb.ac.in

4. Brijal (203180010) 203180010@iitb.ac.in

Problem Statement: - Classify genetic variations based on evidence from the text-based clinical literature or research papers using multi class classification among the 9 classes and classify which class it belongs to.

Domain: - Text Processing, Healthcare

Introduction- Cancer is the second leading cause of death globally and accounted for 8.8 million deaths in 2015. It has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. For better clinical decisions, it is important to accurately distinguish between benign and malignant tumors. Conventionally, statistical methods have been used for classification of high risk and low risk cancer, despite the complex interactions of high-dimensional medical data. To overcome the drawbacks of conventional statistical methods, machine learning has emerged as a promising technique for handling high-dimensional data, with increasing application in clinical decision support. This paper highlights new research directions and discusses main challenges related to machine learning approaches in cancer detection. Basically we are trying to reduce the time consumption in classifying the type of tumor using Multiple machine learning models such as Naïve Bayes, Linear SVM, KNN, logistic Regression, Random Forest.

Libraries: We are using multi class classification problem. We are using libraries such as Pandas, matplotlib, nltk, scikitlearn, NumPy etc. for this project. We are getting the real world data of cancer patients from a US website. Based on that data, we are going to get the probability for the cancer diagnosis. The early diagnosis of cancer can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. So,by using the data, we will try to get the probability of cancer diagnosis.

MOTIVATION: -

This kind of project can help in reduction in time and effort consumption in analyzing the evidence related to each type of variation, classifying them and predicting the probability of each data point.

Methodology: -

We will go through the basic literature out there for implementation of multi-class classification and understand different techniques that one can employ. There can be multiple Performance metric like Area under curve, Precision, Recall, log loss, F1score, confusion metric. Performance metric are also known as KPI (Key Performance Indicator). It is very important to choose the right metric. So, we have chosen multi-class log loss because of our business constraints, our aim was to predict probabilities and for the same, the best model is multi-class Log loss and also for the reason because we want to penalize the model for error since errors are very costly in this scenario. For Better visual interpretation result, we have used confusion matrix. Then we will be choosing the best and most interpretable model i.e., Linear Regression, Linear SVM, Random Forest,

Reference:

1. We are using the data from www.kaggle.com ,
Research paper: Applications of Machine Learning in Cancer Prediction and Prognosis.

2. Noor, M. M., & Narwal, V. (2017). Machine learning approaches in cancer detection and diagnosis: mini review. *IJ Mutil Re App St*, *1*(1), 1-8.

3. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17.