**Project Name:** Disease-Prediction

**Github Link:** https://github.com/utkrisht2000/Disease-Prediction.git

**Why was this project created?**

With the fast advancement of technology and data, the healthcare sector is currently one of the most important study areas. Managing the vast volume of patient data is challenging. Big Data Analytics makes handling this data simpler. There are several methods used all around the world to cure various ailments. A new method that aids in illness detection and prediction is machine learning.

**What problem is it solving?**

This system's goal is to forecast more widespread, often occurring illnesses that, if left untreated, can become deadly. The system uses decision tree algorithms and data mining approaches. Based on the symptoms provided and the precautions needed to prevent disease aggressiveness, this system will forecast the most likely sickness. It will also assist clinicians in analysing the distribution of diseases in society. The illness prediction system in this project will do data mining in its early phases and be taught utilising both machine learning and data mining.

**Entire explanation of project**

* **PROPOSED APPROACH**

Any machine learning task must start with data preparation. An imported dataset consists of two CSV files; a training file and a testing file. The dataset has 133 columns in total, 132 of which are devoted to symptoms, with the last column serving as the prognosis. In a machine learning project, cleaning is the most crucial phase. Our machine learning model's quality is based on the quality of our data. The data must thus always be cleaned before being fed to the model for training. All of the columns in our dataset are numerical, except for the goal column, prognosis, which is a string type and is converted to a numerical form using a label encoder.

The data may be used to train a machine learning model once it has been collected and cleaned. The Support Vector Classifier, Naive Bayes Classifier, and Random Forest Classifier will all be trained using this cleaned data. To assess the models' quality, a confusion matrix will be used. Once the three models have been trained, we will combine their predictions to forecast the illness from the input symptoms. This strengthens and improves the accuracy of our total forecast. Finally, we will define a function that accepts a list of symptoms separated by commas, uses the trained models to forecast the illness based on the symptoms, and delivers the predictions in a JSON format.

Algorithm for creating next word prediction model :

**Step 1:** Dataset is imported

**Step 2:** Data Cleaning

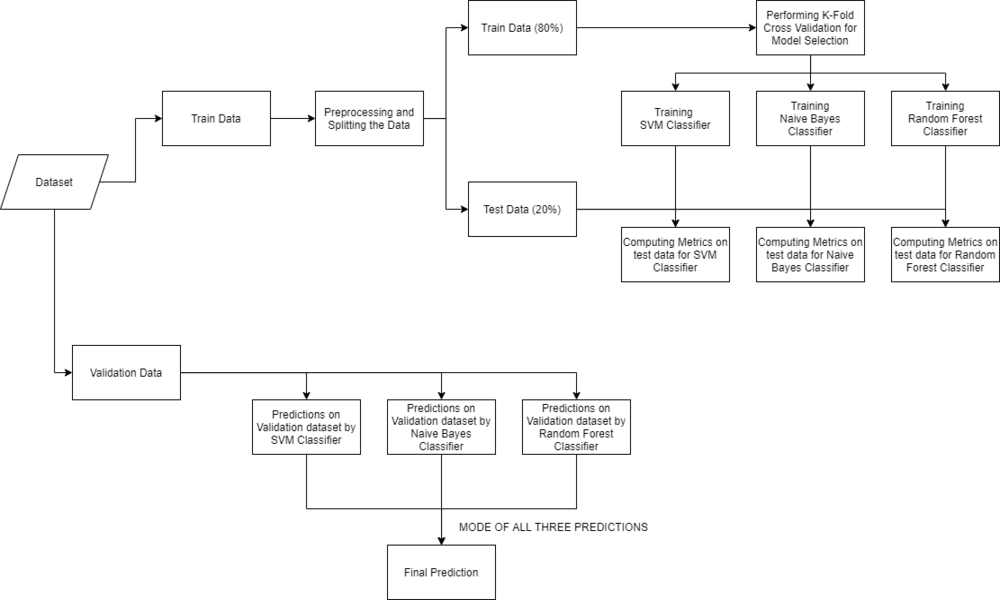
**Step 3:** Label Encoding

**Step 4:** Model Building

**Step 5:** Choose Best Model {Random Forest/ Decision Tree/ Naive Bayes}

**Step 6:** Prediction

* **DATA FLOW DIAGRAM**



* **RESULT**



* **CONCLUSION**

We can see that all of the data items were correctly categorised by our combined model. The last step in the implementation process will include building a function that accepts a list of symptoms, separated by commas, as input and returns the illness predicted by the combined model based on the input symptoms. The 132 symptoms in the dataset should all perfectly match the symptoms provided as input to the method.

Decision Tree 84.5%

Random Forest 98.95%

Naïve Bayes 89.4%