Weekly Report - 4

This week I focused on building other models for disease prediction. I mainly built 3 models: Support Vector Classifier, Gaussian Naive Bayes Classifier and Random Forest. So, for my final analysis I will compare the 4 models: Decision Tree, Support Vector Classifier, Gaussian Naive Bayes Classifier and Random Forest. We had seen how Decision Tree had performed last week. In this report I will give a sample of how other models work. I had divided my dataset to extract symptoms for 40 unique diseases and it contained 120 samples for each disease as below:

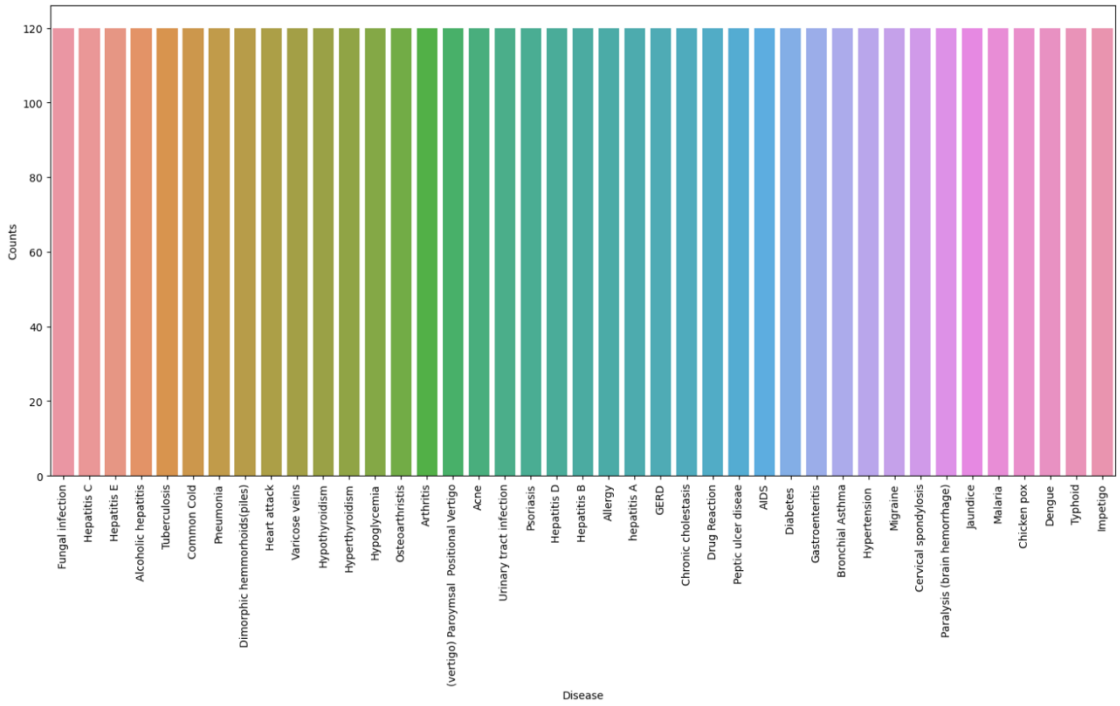

Fig. 1: Spread of diseases in the Dataset

I have used Support Vector Classifier, Gaussian Naive Bayes Classifier, and Random Forest Classifier for cross-validation, to obtain the results as below:

```
============================================================
SVC
Scores: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 1.0
============================================================
Gaussian NB
Scores: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 1.0
============================================================
Random Forest
Scores: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 1.0
```

Fig. 2: K-Fold Cross-Validation for model selection

In the final code, I took a clinical note as an input. Then, I extracted the symptoms separated by commas and encoded them into a numerical format suitable for model prediction. Then, I initialized a list representing the input data with zeros and set entries corresponding to the input symptoms to 1. This encoded input data was then reshaped into a format compatible with model prediction. Using trained models, I generated individual predictions for each model based on the input data. These predictions were then combined by computing the mode (most frequent prediction) across all models, resulting in the final prediction of the disease. The function returned a dictionary containing the predictions of each model as well as the final prediction. Overall, it served as an ensemble predictor, leveraging multiple models to provide a more robust prediction based on the input symptoms.

An example of the same is shown below:

```
The note is: The patient is complaining of Itching. There is Skin Rash that is observed in the patient's hand and there are Nodal Skin Eruptions seen on the right hand.
The keywords extracted are: Itching,Skin Rash,Nodal Skin Eruptions

Disease Predicted:
 {'Random Forest Model': 'Fungal infection', 'Naive Bayes': 'Fungal infection', 'Support Vector Classifier': 'Fungal infection', 'Final Prediction': 'Fungal infection'}
```

Fig.3: Disease Prediction Based on Sample Note

As we can see, in the above example, the clinical note I had inputted was “The patient is complaining of Itching. There is Skin Rash that is observed in the patient's hand and there are Nodal Skin Eruptions seen on the right hand.”. My code extracted the symptoms: “Itching, Skin Rash, Nodal Skin Eruptions” and then predicted “Fungal infection” as the disease using all the models.

In my next steps, I plan to take clinical note as a user input and then predict the disease by extracting the symptoms.