IBM 312- BUSINESS ANALYTICS

TOPIC- DISEASE PREDICTION FROM SYMPTOM DESCRIPTION

Group No.- 20

LAKSHYA (20115059)
LAKSHIT SHARMA(20115058)
SURYANSH SINGH BISHT(20115150)
TANMAY GUPTA(20115154)
RAHUL KUMAR MEENA(18119028)

ABSTRACT-

Our project involves building a disease predictor bot which can effectively diagnose an illness based on a user's description of his symptoms. The bot reads the user's description of their symptoms and uses NLP to retrieve relevant information. Then, logistic regression, Support Vector Machine and Artificial Neural network models which have been trained on a dataset of diseases and their associated symptom description are used to predict the disease that most closely matches the user's symptoms. The result is returned along with some suggested steps. A GUI built using the Tkinter library is used to display the results.

MOTIVATION-

The idea for this work was inspired by the fact that identifying illnesses is frequently a difficult and extended procedure. Confusion about the actual illness that may be producing symptoms may raise tension and anxiety in those who are experiencing them.

In addition, going to the doctor can be expensive and time-consuming, which discourages some people from getting care until their symptoms get worse.
We attempted to create a simple and intuitive disease predictor bot that uses machine learning and natural language processing to precisely diagnose an illness based on a user's symptoms in order to overcome these difficulties.

By using the possible benefits of these technologies, we aim to develop a tool that is simple to use, accessible, and capable of quickly and accurately diagnosing illnesses for people who might not have access to medical professionals or who might be reluctant to seek treatment.

Although its scope is currently limited, it can be improved upon using web development technologies to increase its accessibility and make it user-friendly. With further improvements, it has the potential to serve as a foundation for the development of more complex healthcare systems in the future.

SNAPSHOT OF DATA

The dataset we used for this project consists of 1200 data points and has two columns: “label” and “text”.

1. Label- consists of disease labels
2. Text- consists of natural language symptoms description

The dataset has 24 **different diseases**, and each disease has **50 symptom descriptions** that results in a total of **1200 data points**.

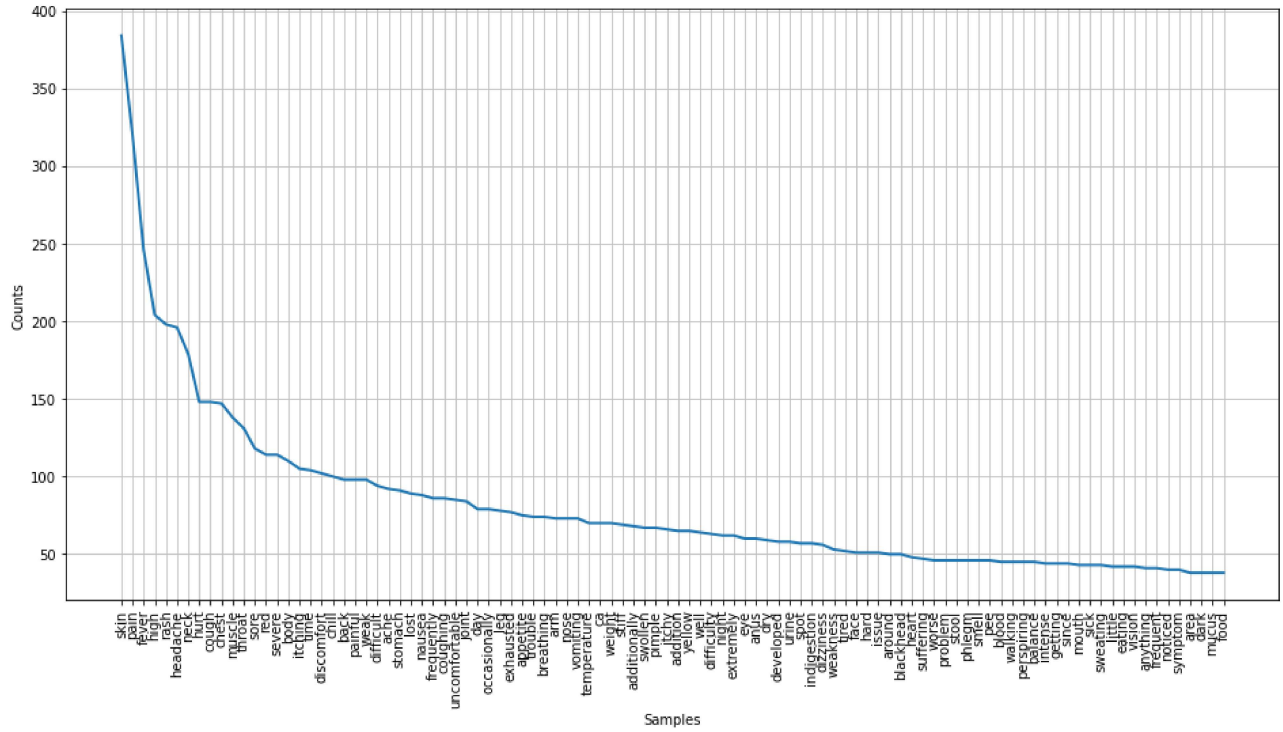
The following 24 diseases have been covered in the dataset: Psoriasis, Varicose Veins, Typhoid, Chicken pox, Impetigo, Dengue, Fungal infection, Common Cold, Pneumonia, Dimorphic Hemorrhoids, Arthritis, Acne, Bronchial Asthma, Hypertension, Migraine, Cervical spondylosis, Jaundice, Malaria, urinary tract infection, allergy, gastroesophageal reflux disease, drug reaction, peptic ulcer disease, diabetes.

	label	text
0	Psoriasis	I have been experiencing a skin rash on my arm...
1	Psoriasis	My skin has been peeling, especially on my kne...
2	Psoriasis	I have been experiencing joint pain in my fing...
3	Psoriasis	There is a silver like dusting on my skin, esp...
4	Psoriasis	My nails have small dents or pits in them, and...

METHODOLOGY-

1. Text cleaning and Conversion to numerical array

- First the dataset is loaded from the csv file using pandas.read_csv.
- To analyze the given data, all the description strings are converted to lowercase, concatenated together and tokenized.
- Stop words for the English language and punctuation symbols are removed. The total number of tokens reduces from 43418 to 18524.
- The tokens are lemmatized using WordNetLemmatizer.
- Some custom stops are identified and also removed. These words are very common in the dataset but do not carry meaning or inclination towards any class.
- The following bar plot shows the frequency of 100 most common words after removing stop words and lemmatization.



- Thus the ‘clean_text’ function is defined which includes converting to lowercase, tokenization, removal of stop words including custom stopwords and then lemmatization.

- This function is applied to each row of the dataset.

	label	text	clean_text
0	Psoriasis	I have been experiencing a skin rash on my arm...	skin rash arm leg torso past week red itchy co...
1	Psoriasis	My skin has been peeling, especially on my kne...	skin peeling especially knee elbow scalp peeli...
2	Psoriasis	I have been experiencing joint pain in my fing...	joint pain finger wrist knee pain achy throbbi...

- This clean text is converted into a TF-IDF matrix using TF-IDF vectorizer. Only unigrams are considered and max features are set to 500.

2. Machine learning models

- The disease classes are encoded using LabelEncoder.
- Various machine learning techniques are applied to this dataset. The results obtained are in the following table:-

Machine learning technique	Training Accuracy	Testing Accuracy
Logistic regression	99.16 %	97.5 %
Support Vector Machine (rbf kernel)	100 %	97.08 %
Multilayer Perceptron (one hidden layer of 150 neurons)	100 %	97.08%

- As all three models have high testing accuracy, we combine the results of the three (ensemble methods). The final answer is obtained by majority voting of the three classifiers.
- A function is defined which takes a string as input, cleans its text, predicts the disease. It returns a string giving the disease and also some suggested steps to take. The suggestions are taken from an excel file created by our group. Example output-

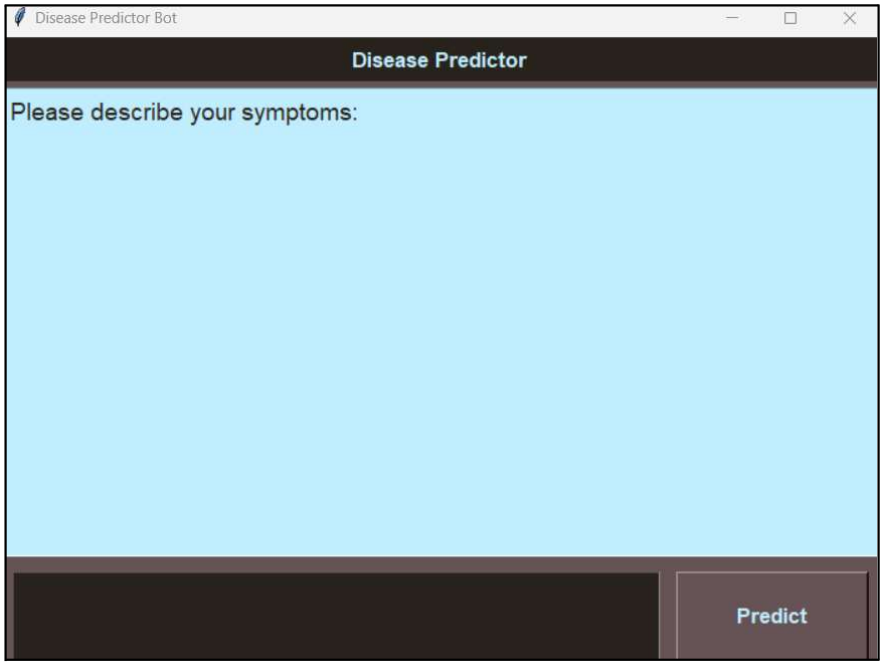
```
predict("I have a dry mouth and i tend to urinate frequently")
```

```
'According to the disease prediction model, you are suffering from diabetes. The suggested steps are - Take care of your food h
abits; Eat non-starchy vegetables, leafy greens, fatty fish, etc. Consult your doctor if your condition worsens.'
```

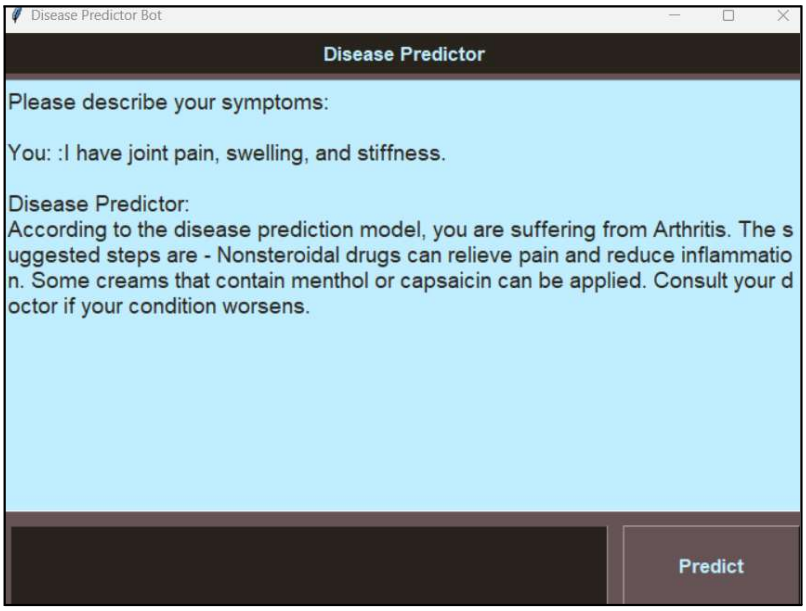
3. Disease-Predictor-Bot using Tkinter

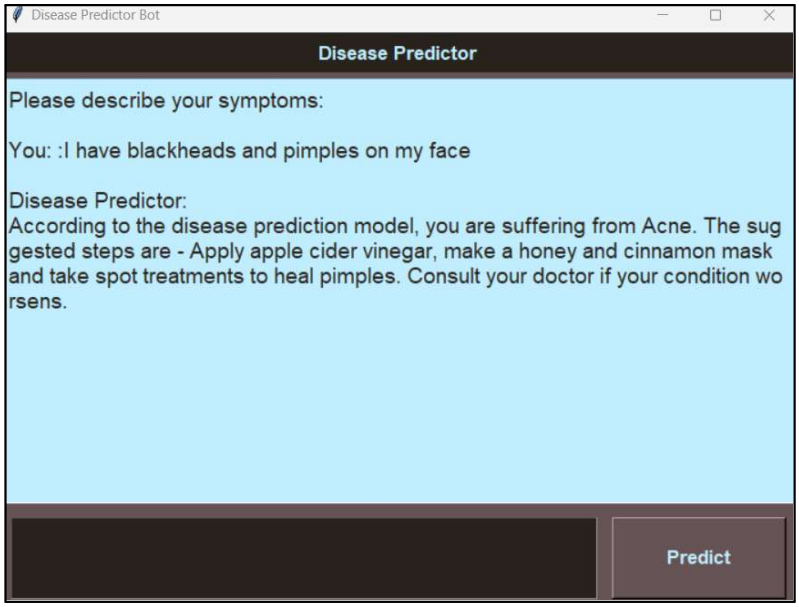
A simple GUI is created using the Tkinter library in python. It has various components like message box, predict button, heading and display panel.

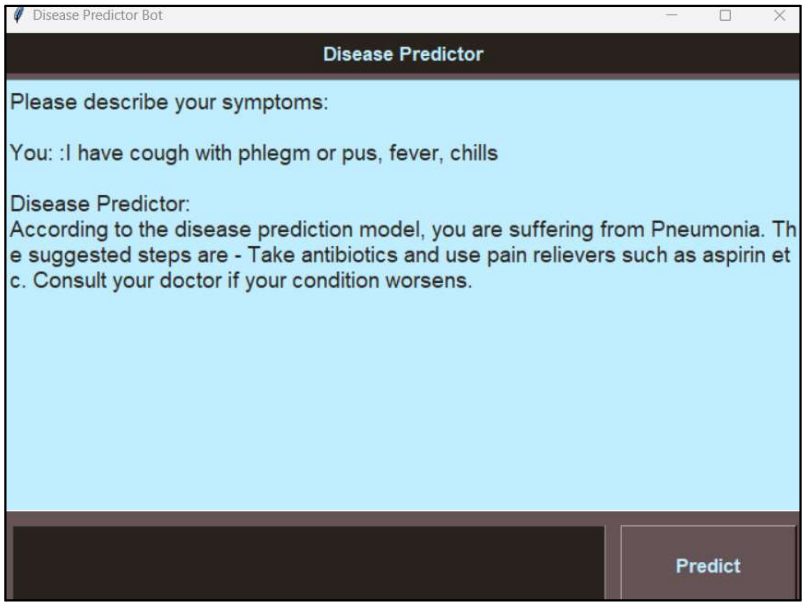
To run it, all the files need to be downloaded and then the app.py file needs to be executed.



RESULTS AND ANALYSIS







We successfully implemented a disease predictor bot which takes some symptoms as input and detects the most probable disease that is present, moreover it also suggests some important treatment measures to ease the disease. For disease prediction, we have used the concept of ensemble learning, and combined these predictions from Logistic Regression, Support Vector Machines, and Multi-Layer Perceptron.

REFERENCES-

https://www.kaggle.com/datasets/niyarrbarman/symptom2disease
https://scikit-learn.org/stable/user_guide.html
https://www.geeksforgeeks.org/gui-chat-application-using-tkinter-in-python/