IDS 561 Analytics for Big Data (45604) 2022 Spring

Final Project

Disease Prediction Using Symptoms

Group Members:
Mehul Gupta(UIN:677991579)
Saraschandra Addanki(UIN:658694881)
Shivani Erigineni(UIN:665751065)

Problem Setting:

The healthcare area is one of the most important research subjects in the modern period, thanks to rapid advancements in technology and data. It's difficult to keep track of such a large amount of patient data.

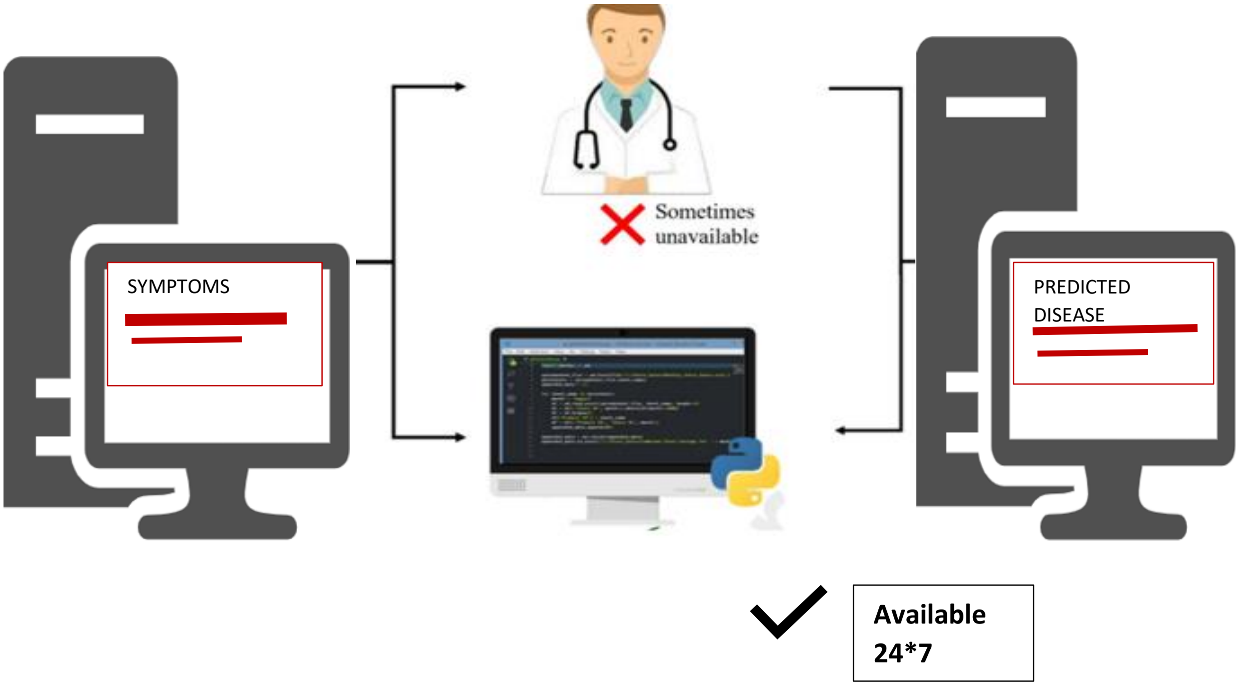
Big Data Analytics makes it simpler to handle Electronic Health Records data which is one of the biggest examples of the application of big data in healthcare.

Machine Learning and Big Data are two innovative methods for predicting and diagnosing diseases and THE PROJECT aims to implement a robust machine learning model that can efficiently predict the disease of a human, based on the symptoms that he/she possesses

Project Description

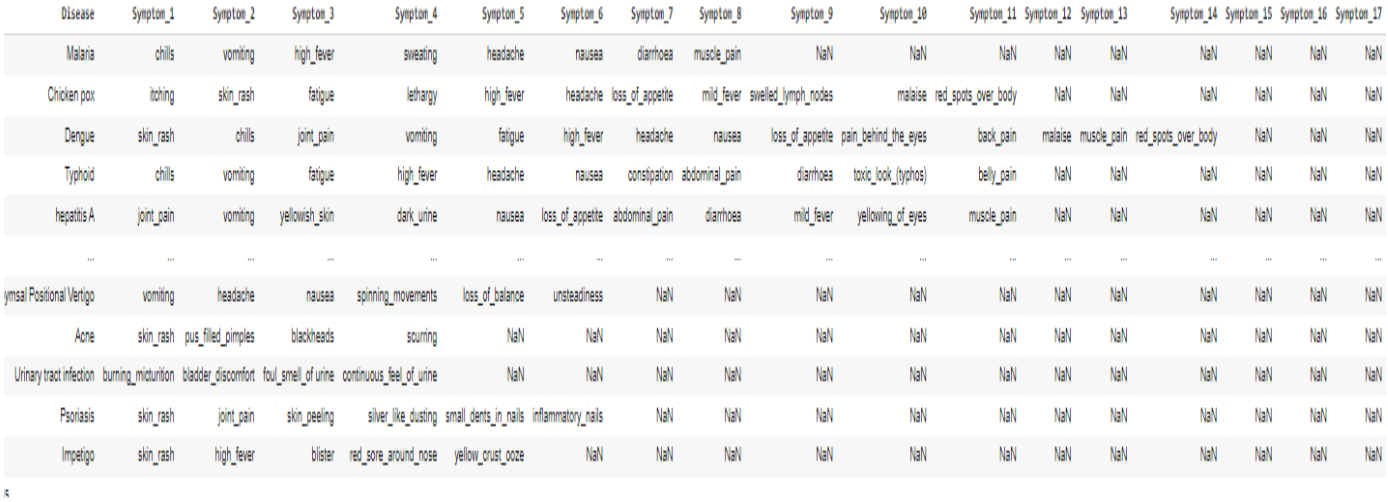
The Main Motivation of the Project is inspired by an online Chatbot we encountered by chance on a medical advisory website. Basically, it was the first step we need to go through in order to get assigned to a specific medical department for further detailed diagnosis. The chatbot asks the user to enter the symptoms the user had been facing and then gives a rough diagnosis.

With the fast advancement of technology and data, the healthcare sector is one of the most significant study topics in the contemporary era. It is challenging to manage the vast volume of patient data. Big Data Analytics makes it easier to manage this data. Around the world, there are several ways for treating various ailments. Machine Learning is a new method that aids in disease prediction and diagnosis. This study illustrates the use of machine learning to predict illness based on symptoms. On the presented dataset, machine learning methods such as Naive Bayes, Decision Tree, and Random Forest are used to forecast the illness. The python programming language is used to implement it. The research demonstrates the best algorithm based on their accuracy. The accuracy of an algorithm is determined by the performance of the given dataset.

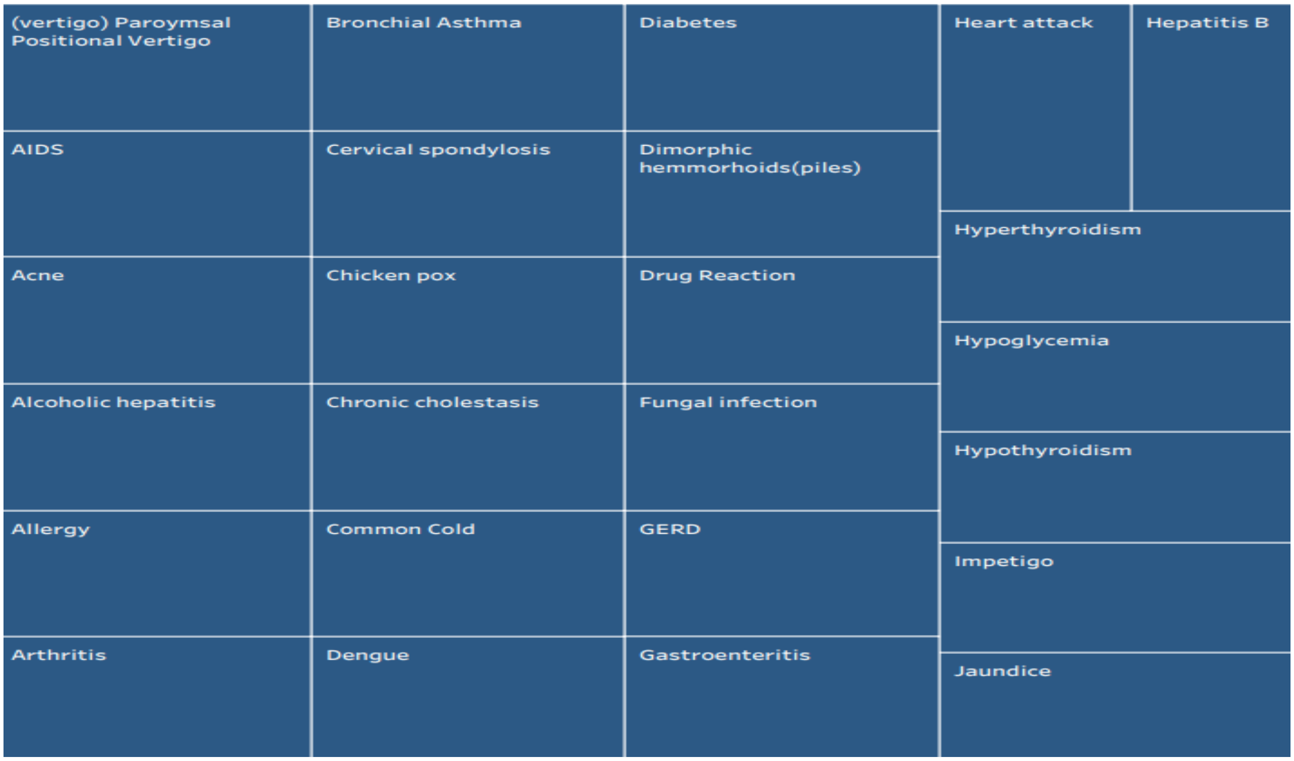


Data Description:

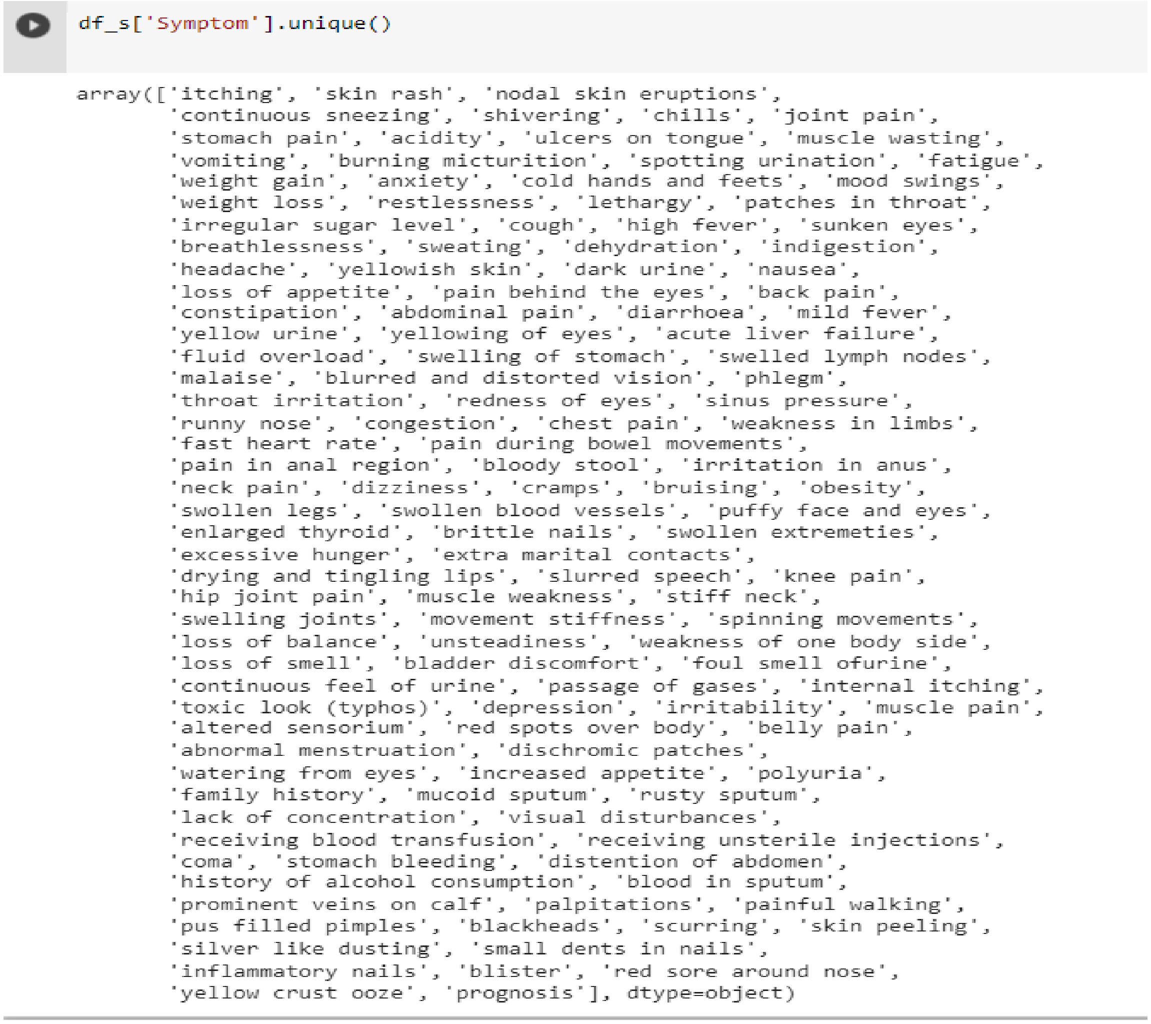
- The dataset was taken from the Kaggle.
- It comprises the diseases and their symptoms**.** It has information of the diseases and what might be the symptoms of these diseases.
- There are 4920 observations with Maximum of 17 symptoms. For example, for chicken pox- symptoms experienced by one person are itching, skin rash, fatigue.etc and it varied person to person.



Disease	Sympton_1	Sympton_2	Sympton_3	Sympton_4	Sympton_5	Sympton_6	Sympton_7	Sympton_8	Sympton_9	Sympton_10	Sympton_11	Sympton_12	Sympton_13	Sympton_14	Sympton_15	Sympton_16	Sympton_17
Malaria	chills	vomiting	high_fever	sweating	headache	nausea	diarrhoea	muscle_pain	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Chicken pox	itching	skin_rash	fatigue	lethargy	high_fever	headache	loss_of_appetite	mild_fever	swelled_lymph_nodes	malaise	red_spots_over_body	NaN	NaN	NaN	NaN	NaN	NaN
Dengue	skin_rash	chills	joint_pain	vomiting	fatigue	high_fever	headache	nausea	loss_of_appetite	pain_behind_the_eyes	back_pain	malaise	muscle_pain	red_spots_over_body	NaN	NaN	NaN
Typhoid	chills	vomiting	fatigue	high_fever	headache	nausea	constipation	abdominal_pain	diarrhoea	toxic_look_(typhos)	belly_pain	NaN	NaN	NaN	NaN	NaN	NaN
hepatitis A	joint_pain	vomiting	yellowish_skin	dark_urine	nausea	loss_of_appetite	abdominal_pain	diarrhoea	mild_fever	yellowing_of_eyes	muscle_pain	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
ymsal Positional Vertigo	vomiting	headache	nausea	spinning_movements	loss_of_balance	unsteadiness	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Acne	skin_rash	pus_filled_pimples	blackheads	scurring	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Urinary tract infection	burning_micturition	bladder_discomfort	foul_smell_of urine	continuous_feel_of_urine	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Psoriasis	skin_rash	joint_pain	skin_peeling	silver_like_dusting	small_dents_in_nails	inflammatory_nails	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Impetigo	skin_rash	high_fever	blister	red_sore_around_nose	yellow_crust_ooze	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN



The above tree map shows the different number of diseases in the dataset.



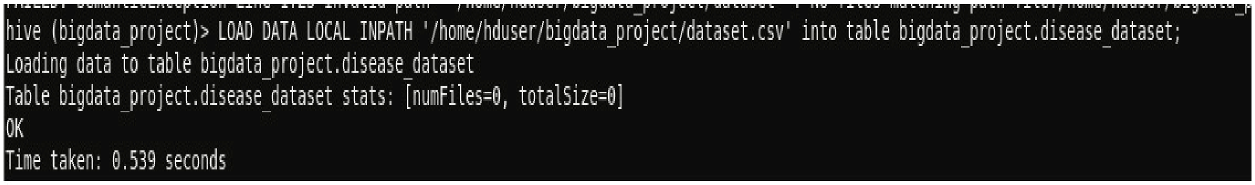
```
df_s['Symptom'].unique()
```

```
array(['itching', 'skin rash', 'nodal skin eruptions',
       'continuous sneezing', 'shivering', 'chills', 'joint pain',
       'stomach pain', 'acidity', 'ulcers on tongue', 'muscle wasting',
       'vomiting', 'burning micturition', 'spotting urination', 'fatigue',
       'weight gain', 'anxiety', 'cold hands and feets', 'mood swings',
       'weight loss', 'restlessness', 'lethargy', 'patches in throat',
       'irregular sugar level', 'cough', 'high fever', 'sunken eyes',
       'breathlessness', 'sweating', 'dehydration', 'indigestion',
       'headache', 'yellowish skin', 'dark urine', 'nausea',
       'loss of appetite', 'pain behind the eyes', 'back pain',
       'constipation', 'abdominal pain', 'diarrhoea', 'mild fever',
       'yellow urine', 'yellowing of eyes', 'acute liver failure',
       'fluid overload', 'swelling of stomach', 'swelled lymph nodes',
       'malaise', 'blurred and distorted vision', 'phlegm',
       'throat irritation', 'redness of eyes', 'sinus pressure',
       'runny nose', 'congestion', 'chest pain', 'weakness in limbs',
       'fast heart rate', 'pain during bowel movements',
       'pain in anal region', 'bloody stool', 'irritation in anus',
       'neck pain', 'dizziness', 'cramps', 'bruising', 'obesity',
       'swollen legs', 'swollen blood vessels', 'puffy face and eyes',
       'enlarged thyroid', 'brittle nails', 'swollen extremeties',
       'excessive hunger', 'extra marital contacts',
       'drying and tingling lips', 'slurred speech', 'knee pain',
       'hip joint pain', 'muscle weakness', 'stiff neck',
       'swelling joints', 'movement stiffness', 'spinning movements',
       'loss of balance', 'unsteadiness', 'weakness of one body side',
       'loss of smell', 'bladder discomfort', 'foul smell ofurine',
       'continuous feel of urine', 'passage of gases', 'internal itching',
       'toxic look (typhos)', 'depression', 'irritability', 'muscle pain',
       'altered sensorium', 'red spots over body', 'belly pain',
       'abnormal menstruation', 'dischromic patches',
       'watering from eyes', 'increased appetite', 'polyuria',
       'family history', 'mucoid sputum', 'rusty sputum',
       'lack of concentration', 'visual disturbances',
       'receiving blood transfusion', 'receiving unsterile injections',
       'coma', 'stomach bleeding', 'distention of abdomen',
       'history of alcohol consumption', 'blood in sputum',
       'prominent veins on calf', 'palpitations', 'painful walking',
       'pus filled pimples', 'blackheads', 'scurring', 'skin peeling',
       'silver like dusting', 'small dents in nails',
       'inflammatory nails', 'blister', 'red sore around nose',
       'yellow crust ooze', 'prognosis'], dtype=object)
```

The above output shows different symptoms that could lead to the diseases mentioned above.

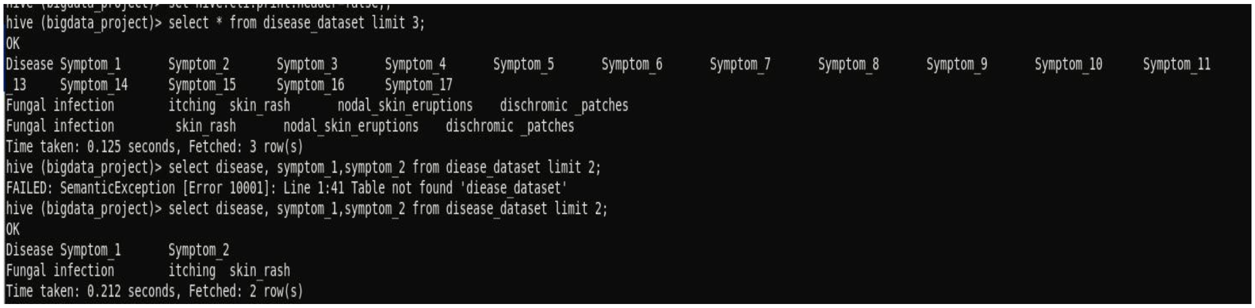
Using HIVE To Import the Data:

Setting the server to hive and copying the path of dataset where it is stored.



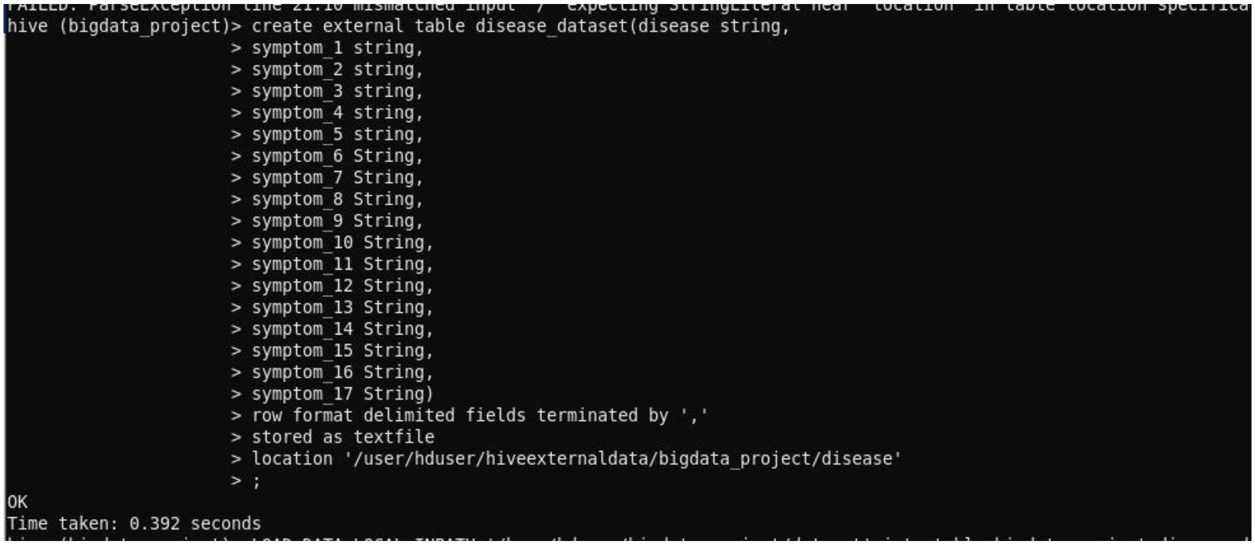
```
FAILED: SemanticException Line 1:23 Invalid path ''/home/hduser/bigdata_project/dataset'': No files matching path file:/home/hduser/bigdata_p
hive (bigdata_project)> LOAD DATA LOCAL INPATH '/home/hduser/bigdata_project/dataset.csv' into table bigdata_project.disease_dataset;
Loading data to table bigdata_project.disease_dataset
Table bigdata_project.disease_dataset stats: [numFiles=0, totalSize=0]
OK
Time taken: 0.539 seconds
```

Running a query to see the dataset details:



```
hive (bigdata_project)> set hive.cli.print.header=false;;
hive (bigdata_project)> select * from disease_dataset limit 3;
OK
Disease Symptom_1       Symptom_2       Symptom_3       Symptom_4       Symptom_5       Symptom_6       Symptom_7       Symptom_8       Symptom_9       Symptom_10      Symptom_11
_13     Symptom_14      Symptom_15      Symptom_16      Symptom_17
Fungal infection        itching  skin_rash       nodal_skin_eruptions    dischromic _patches
Fungal infection         skin_rash       nodal_skin_eruptions    dischromic _patches
Time taken: 0.125 seconds, Fetched: 3 row(s)
hive (bigdata_project)> select disease, symptom_1,symptom_2 from diease_dataset limit 2;
FAILED: SemanticException [Error 10001]: Line 1:41 Table not found 'diease_dataset'
hive (bigdata_project)> select disease, symptom_1,symptom_2 from disease_dataset limit 2;
OK
Disease Symptom_1       Symptom_2
Fungal infection        itching  skin_rash
Time taken: 0.212 seconds, Fetched: 2 row(s)
```

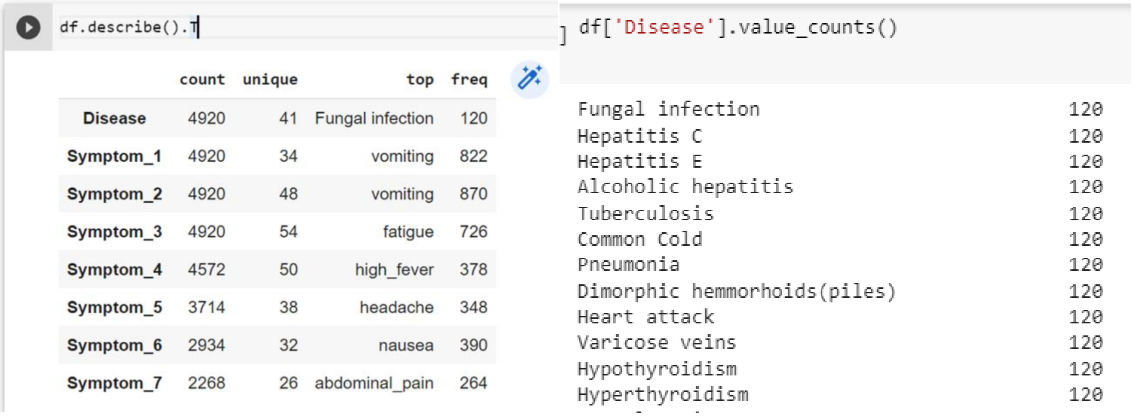
Describing the dataset



```
FAILED: ParseException line 21:10 mismatched input '/' expecting StringLiteral near 'location' in table location specifica
hive (bigdata_project)> create external table disease_dataset(disease string,
                      > symptom_1 string,
                      > symptom_2 string,
                      > symptom_3 string,
                      > symptom_4 string,
                      > symptom_5 string,
                      > symptom_6 String,
                      > symptom_7 String,
                      > symptom_8 String,
                      > symptom_9 String,
                      > symptom_10 String,
                      > symptom_11 String,
                      > symptom_12 String,
                      > symptom_13 String,
                      > symptom_14 String,
                      > symptom_15 String,
                      > symptom_16 String,
                      > symptom_17 String)
                      > row format delimited fields terminated by ','
                      > stored as textfile
                      > location '/user/hduser/hiveexternaldata/bigdata_project/disease'
                      > ;
OK
Time taken: 0.392 seconds
```

Data Pre-Processing and Data Cleaning:

For the better understanding of the dataset, we wanted to see how many unique diseases and symptoms we are dealing with in the data set. The following are the clear results: We found out that we are dealing with 41 unique diseases. And each of the 41 diseases has 120 set of symptoms which we found to be very fascinating and a well-balanced data.



```
df.describe().T
```

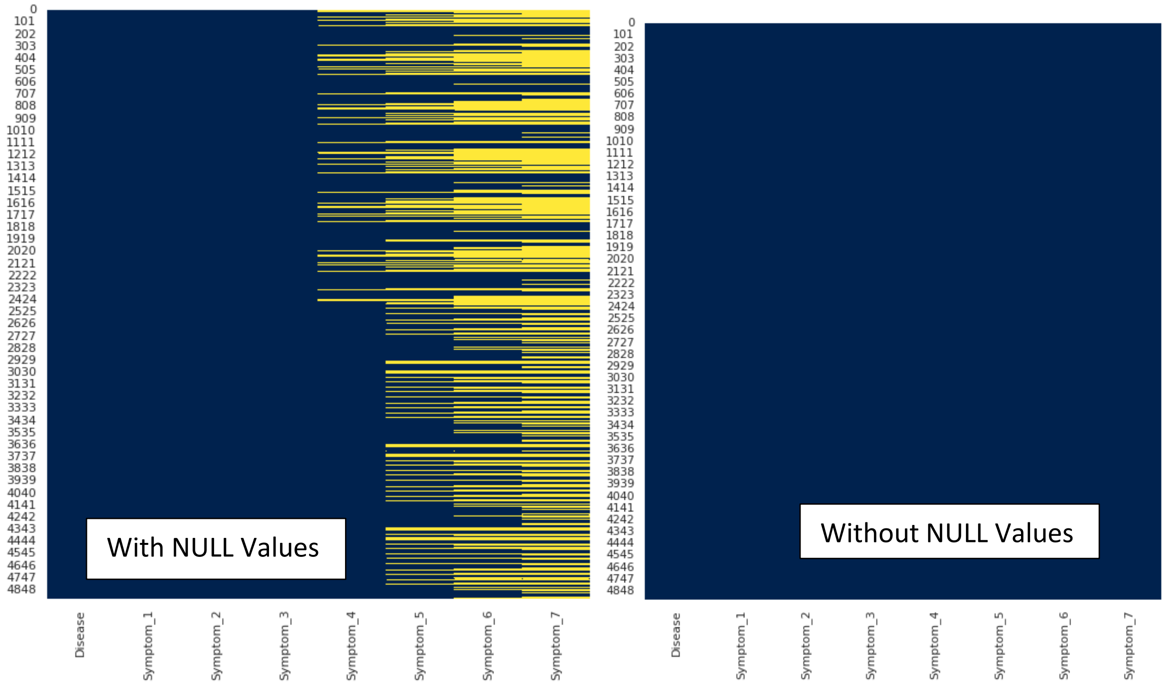
	count	unique	top	freq
Disease	4920	41	Fungal infection	120
Symptom_1	4920	34	vomiting	822
Symptom_2	4920	48	vomiting	870
Symptom_3	4920	54	fatigue	726
Symptom_4	4572	50	high_fever	378
Symptom_5	3714	38	headache	348
Symptom_6	2934	32	nausea	390
Symptom_7	2268	26	abdominal_pain	264

```
df['Disease'].value_counts()
```

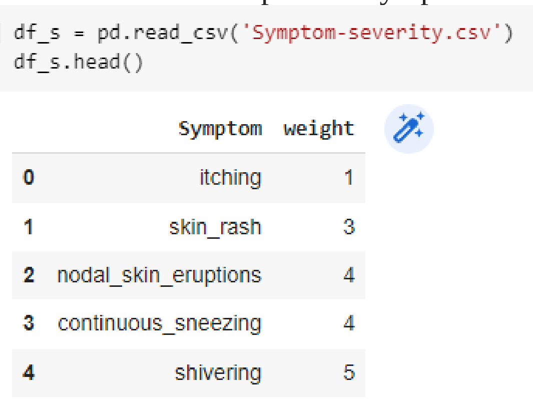
```
Fungal infection                                 120
Hepatitis C                                      120
Hepatitis E                                      120
Alcoholic hepatitis                              120
Tuberculosis                                     120
Common Cold                                      120
Pneumonia                                        120
Dimorphic hemmorhoids(piles)                     120
Heart attack                                     120
Varicose veins                                   120
Hypothyroidism                                   120
Hyperthyroidism                                  120
```

- We did some cleaning and replaced the ‘NaN’ values with Zeros.

We have removed any null values, hyphens, insignificant columns, and rows that had many null values because they provided no meaningful information and have given weights and done normalization for various symptoms.



Downloaded and Imported ‘Sympton-severity.csv’ to get severity scores.

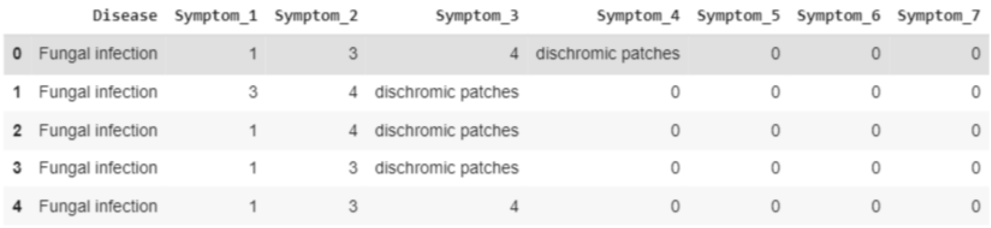


```
df_s = pd.read_csv('Symptom-severity.csv')
df_s.head()
```

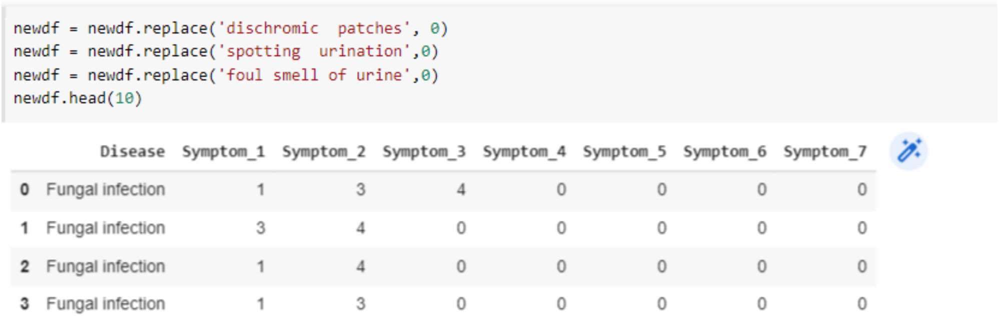
	Symptom	weight
0	itching	1
1	skin_rash	3
2	nodal_skin_eruptions	4
3	continuous_sneezing	4
4	shivering	5

Here, in this dataset, each of the symptom is given weights as per their severity and we want to plug in these weights in the dataset in the corresponding Symptoms. We have replaced the symptoms text data into numerical weights and noticed that three symptoms i.e., dyschromic patches, spotting urination and foul smell of urine are not given any weights.

So, we have assigned 0 weights to those, and this is how our new data and final data looks like:



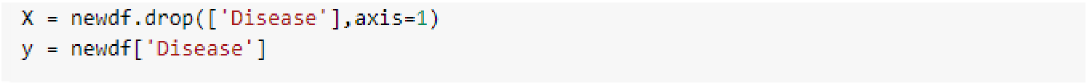
	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7
0	Fungal infection	1	3	4	dischromic patches	0	0	0
1	Fungal infection	3	4	dischromic patches	0	0	0	0
2	Fungal infection	1	4	dischromic patches	0	0	0	0
3	Fungal infection	1	3	dischromic patches	0	0	0	0
4	Fungal infection	1	3	4	0	0	0	0



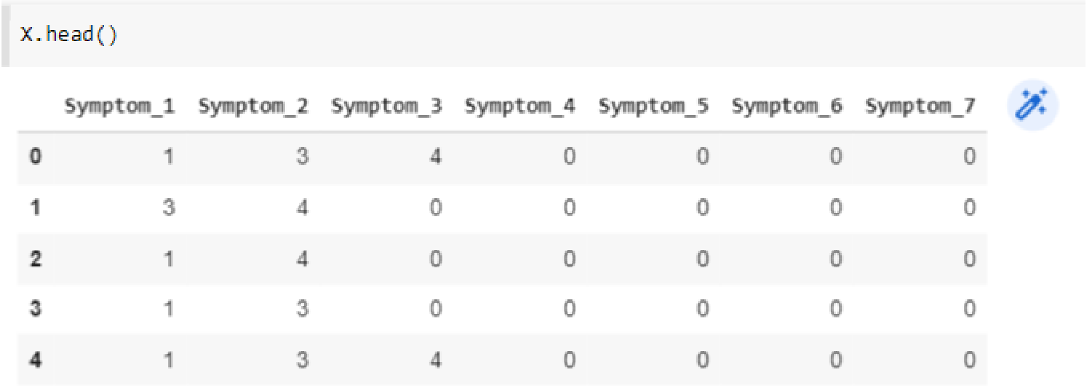
```
newdf = newdf.replace('dischromic  patches', 0)
newdf = newdf.replace('spotting  urination',0)
newdf = newdf.replace('foul smell of urine',0)
newdf.head(10)
```

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7
0	Fungal infection	1	3	4	0	0	0	0
1	Fungal infection	3	4	0	0	0	0	0
2	Fungal infection	1	4	0	0	0	0	0
3	Fungal infection	1	3	0	0	0	0	0

Selection of features for Training Purpose: Deleting the disease column



```
X = newdf.drop(['Disease'],axis=1)
y = newdf['Disease']
```

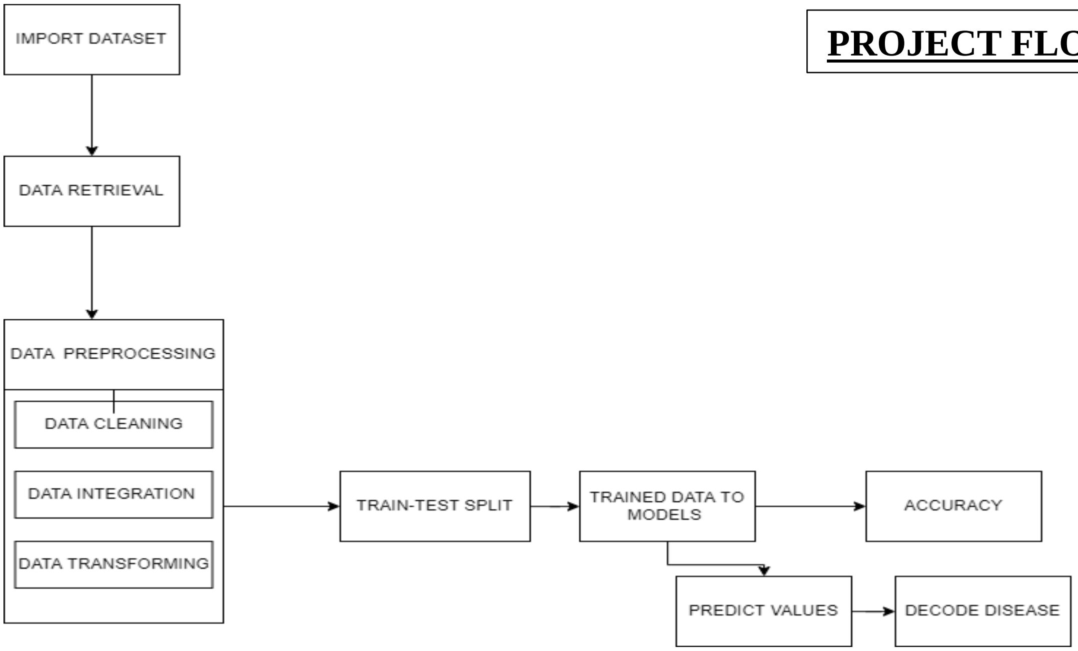


```
X.head()
```

	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7
0	1	3	4	0	0	0	0
1	3	4	0	0	0	0	0
2	1	4	0	0	0	0	0
3	1	3	0	0	0	0	0
4	1	3	4	0	0	0	0

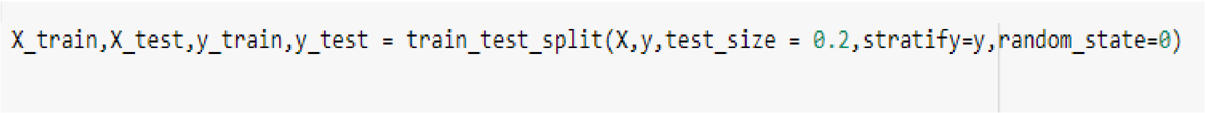
Final dataset looks like above.

PROJECT FLOW DIAGRAM



Splitting the data:

The data set is divided into test and train with 80% and 20% probability.



```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size = 0.2,stratify=y,random_state=0)
```

Models:

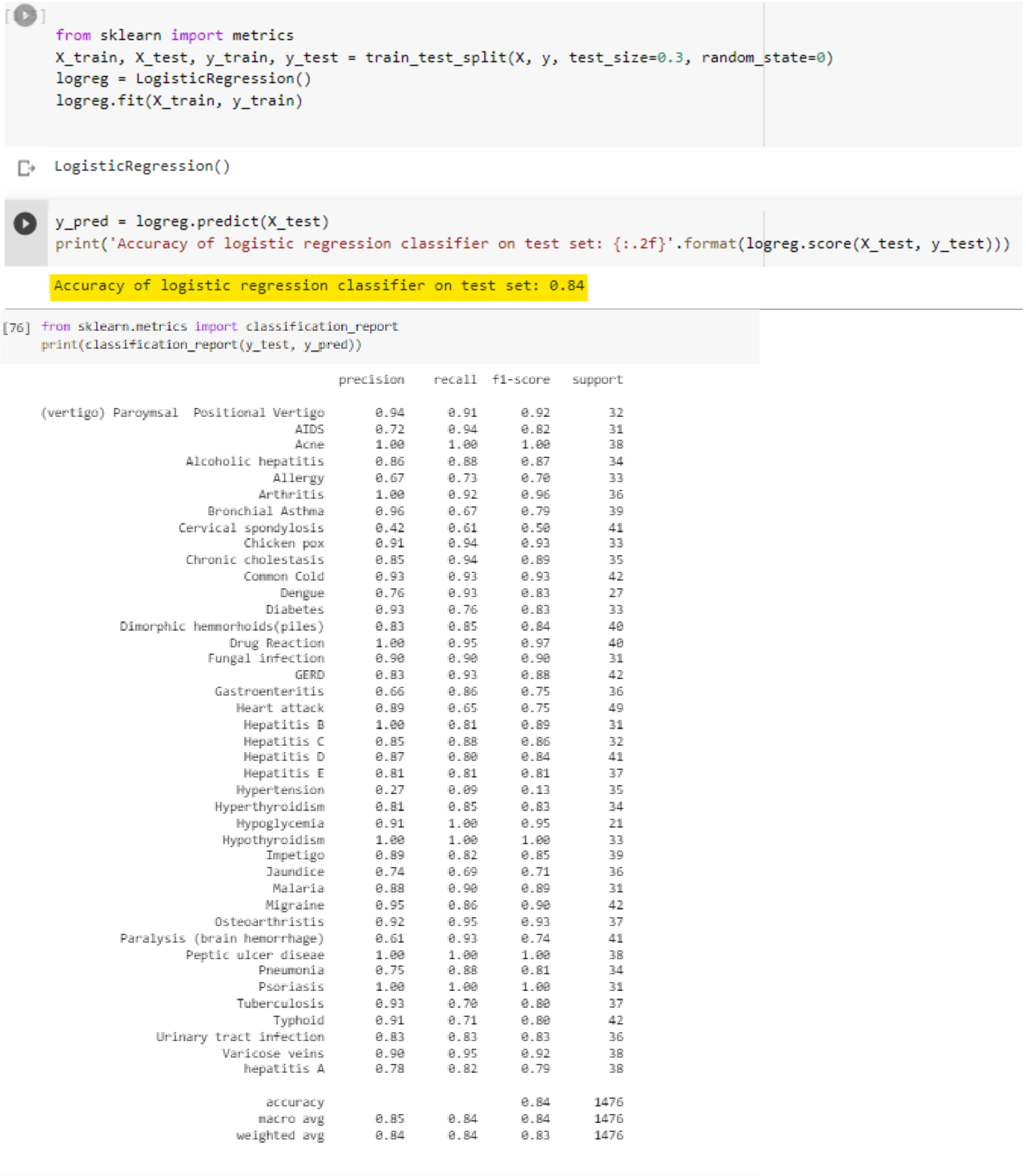
We used the following three Machine Learning Models for our dataset:

1. Logestic Regression
2. Random Forest Classifier
3. SVM Model
4. KNN

Logestic Regression:

Logistic Regression is a statistical and machine-learning technique classifying records of a dataset based on the values of the input fields. It predicts a dependent variable based on one or more set of independent variables to predict outcomes.

Screenshot of the logistic model with classification report



```
from sklearn import metrics
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
```

```
LogisticRegression()
```

```
y_pred = logreg.predict(X_test)
print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_test, y_test)))
```

```
Accuracy of logistic regression classifier on test set: 0.84
```

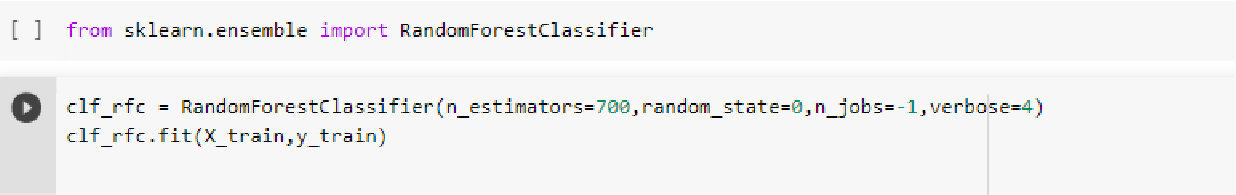
```
[76] from sklearn.metrics import classification_report
     print(classification_report(y_test, y_pred))
```

```
                                          precision    recall  f1-score   support

(vertigo) Paroymsal  Positional Vertigo       0.94      0.91      0.92        32
                                   AIDS       0.72      0.94      0.82        31
                                   Acne       1.00      1.00      1.00        38
                    Alcoholic hepatitis       0.86      0.88      0.87        34
                                Allergy       0.67      0.73      0.70        33
                              Arthritis       1.00      0.92      0.96        36
                       Bronchial Asthma       0.96      0.67      0.79        39
                   Cervical spondylosis       0.42      0.61      0.50        41
                            Chicken pox       0.91      0.94      0.93        33
                    Chronic cholestasis       0.85      0.94      0.89        35
                            Common Cold       0.93      0.93      0.93        42
                                 Dengue       0.76      0.93      0.83        27
                               Diabetes       0.93      0.76      0.83        33
           Dimorphic hemmorhoids(piles)       0.83      0.85      0.84        40
                          Drug Reaction       1.00      0.95      0.97        40
                       Fungal infection       0.90      0.90      0.90        31
                                   GERD       0.83      0.93      0.88        42
                       Gastroenteritis       0.66      0.86      0.75        36
                           Heart attack       0.89      0.65      0.75        49
                            Hepatitis B       1.00      0.81      0.89        31
                            Hepatitis C       0.85      0.88      0.86        32
                            Hepatitis D       0.87      0.80      0.84        41
                            Hepatitis E       0.81      0.81      0.81        37
                           Hypertension       0.27      0.09      0.13        35
                        Hyperthyroidism       0.81      0.85      0.83        34
                           Hypoglycemia       0.91      1.00      0.95        21
                         Hypothyroidism       1.00      1.00      1.00        33
                               Impetigo       0.89      0.82      0.85        39
                               Jaundice       0.74      0.69      0.71        36
                                Malaria       0.88      0.90      0.89        31
                               Migraine       0.95      0.86      0.90        42
                       Osteoarthristis       0.92      0.95      0.93        37
           Paralysis (brain hemorrhage)       0.61      0.93      0.74        41
                    Peptic ulcer diseae       1.00      1.00      1.00        38
                              Pneumonia       0.75      0.88      0.81        34
                              Psoriasis       1.00      1.00      1.00        31
                           Tuberculosis       0.93      0.70      0.80        37
                                Typhoid       0.91      0.71      0.80        42
                Urinary tract infection       0.83      0.83      0.83        36
                         Varicose veins       0.90      0.95      0.92        38
                            hepatitis A       0.78      0.82      0.79        38

                               accuracy                           0.84      1476
                              macro avg       0.85      0.84      0.84      1476
                           weighted avg       0.84      0.84      0.83      1476
```

Random Forest:

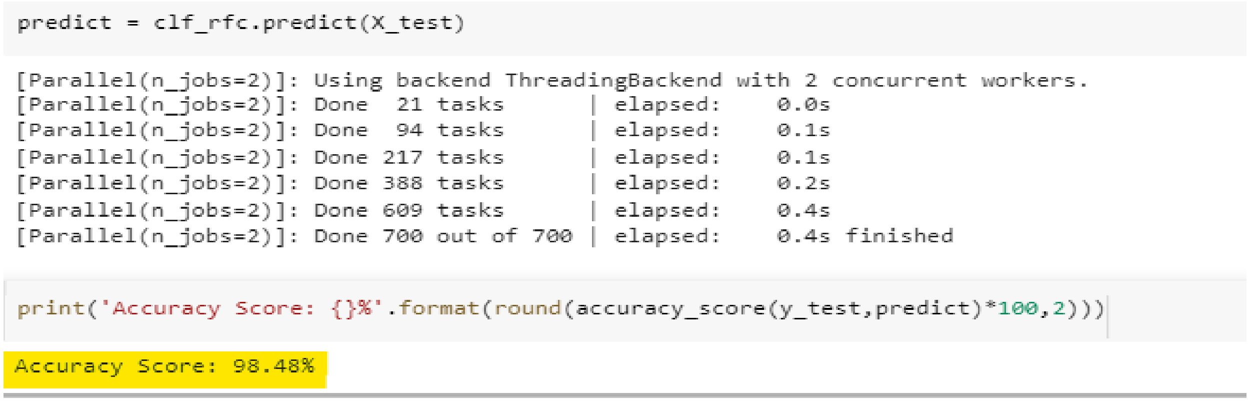
Random Forest is a supervised learning algorithm used for both classification and regression. It chooses random data samples from dataset and constructs decision trees for every sample dataset chosen, most voted prediction will be selected and be presented as result of classification.

Screenshot of the Random Forest Classifier



```
[ ]  from sklearn.ensemble import RandomForestClassifier
```

```
clf_rfc = RandomForestClassifier(n_estimators=700,random_state=0,n_jobs=-1,verbose=4)
clf_rfc.fit(X_train,y_train)
```

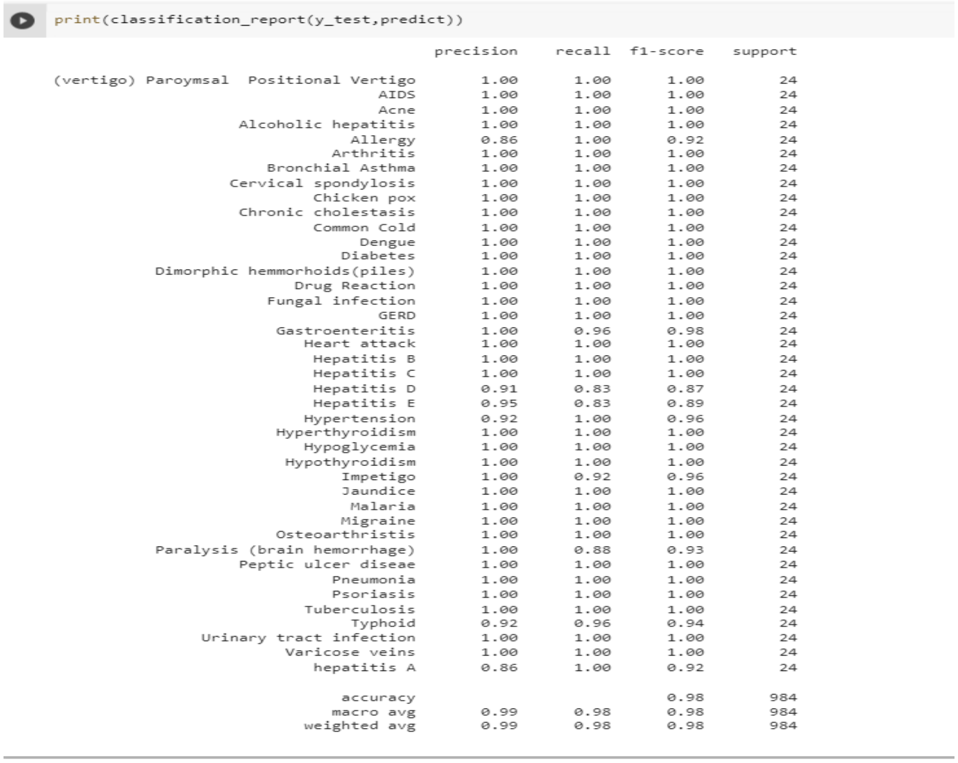


```
predict = clf_rfc.predict(X_test)
```

```
[Parallel(n_jobs=2)]: Using backend ThreadingBackend with 2 concurrent workers.
[Parallel(n_jobs=2)]: Done  21 tasks      | elapsed:    0.0s
[Parallel(n_jobs=2)]: Done  94 tasks      | elapsed:    0.1s
[Parallel(n_jobs=2)]: Done 217 tasks      | elapsed:    0.1s
[Parallel(n_jobs=2)]: Done 388 tasks      | elapsed:    0.2s
[Parallel(n_jobs=2)]: Done 609 tasks      | elapsed:    0.4s
[Parallel(n_jobs=2)]: Done 700 out of 700 | elapsed:    0.4s finished
```

```
print('Accuracy Score: {}%'.format(round(accuracy_score(y_test,predict)*100,2)))
```

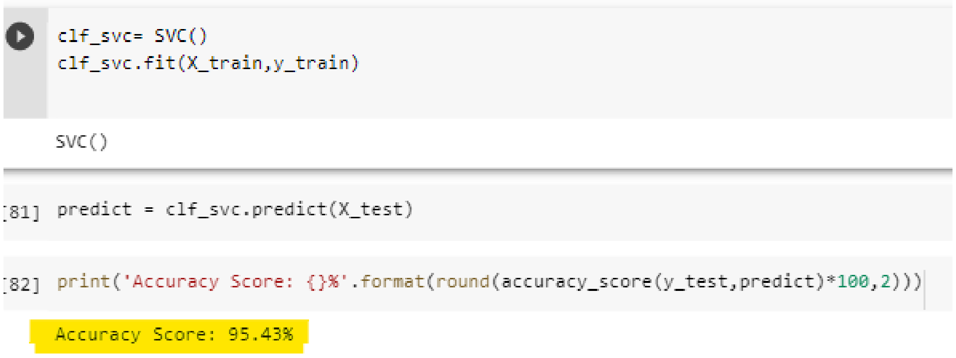
```
Accuracy Score: 98.48%
```



```
print(classification_report(y_test,predict))
```

```
                                             precision    recall  f1-score   support

    (vertigo) Paroymsal  Positional Vertigo       1.00      1.00      1.00        24
                                       AIDS       1.00      1.00      1.00        24
                                       Acne       1.00      1.00      1.00        24
                        Alcoholic hepatitis       1.00      1.00      1.00        24
                                    Allergy       0.86      1.00      0.92        24
                                  Arthritis       1.00      1.00      1.00        24
                           Bronchial Asthma       1.00      1.00      1.00        24
                       Cervical spondylosis       1.00      1.00      1.00        24
                                Chicken pox       1.00      1.00      1.00        24
                        Chronic cholestasis       1.00      1.00      1.00        24
                                Common Cold       1.00      1.00      1.00        24
                                     Dengue       1.00      1.00      1.00        24
                                   Diabetes       1.00      1.00      1.00        24
               Dimorphic hemmorhoids(piles)       1.00      1.00      1.00        24
                              Drug Reaction       1.00      1.00      1.00        24
                           Fungal infection       1.00      1.00      1.00        24
                                       GERD       1.00      1.00      1.00        24
                            Gastroenteritis       1.00      0.96      0.98        24
                               Heart attack       1.00      1.00      1.00        24
                                Hepatitis B       1.00      1.00      1.00        24
                                Hepatitis C       1.00      1.00      1.00        24
                                Hepatitis D       0.91      0.83      0.87        24
                                Hepatitis E       0.95      0.83      0.89        24
                               Hypertension       0.92      1.00      0.96        24
                            Hyperthyroidism       1.00      1.00      1.00        24
                               Hypoglycemia       1.00      1.00      1.00        24
                             Hypothyroidism       1.00      1.00      1.00        24
                                   Impetigo       1.00      0.92      0.96        24
                                   Jaundice       1.00      1.00      1.00        24
                                    Malaria       1.00      1.00      1.00        24
                                   Migraine       1.00      1.00      1.00        24
                            Osteoarthristis       1.00      1.00      1.00        24
               Paralysis (brain hemorrhage)       1.00      0.88      0.93        24
                        Peptic ulcer diseae       1.00      1.00      1.00        24
                                  Pneumonia       1.00      1.00      1.00        24
                                  Psoriasis       1.00      1.00      1.00        24
                               Tuberculosis       1.00      1.00      1.00        24
                                    Typhoid       0.92      0.96      0.94        24
                    Urinary tract infection       1.00      1.00      1.00        24
                             Varicose veins       1.00      1.00      1.00        24
                                hepatitis A       0.86      1.00      0.92        24

                                   accuracy                           0.98       984
                                  macro avg       0.99      0.98      0.98       984
                               weighted avg       0.99      0.98      0.98       984
```

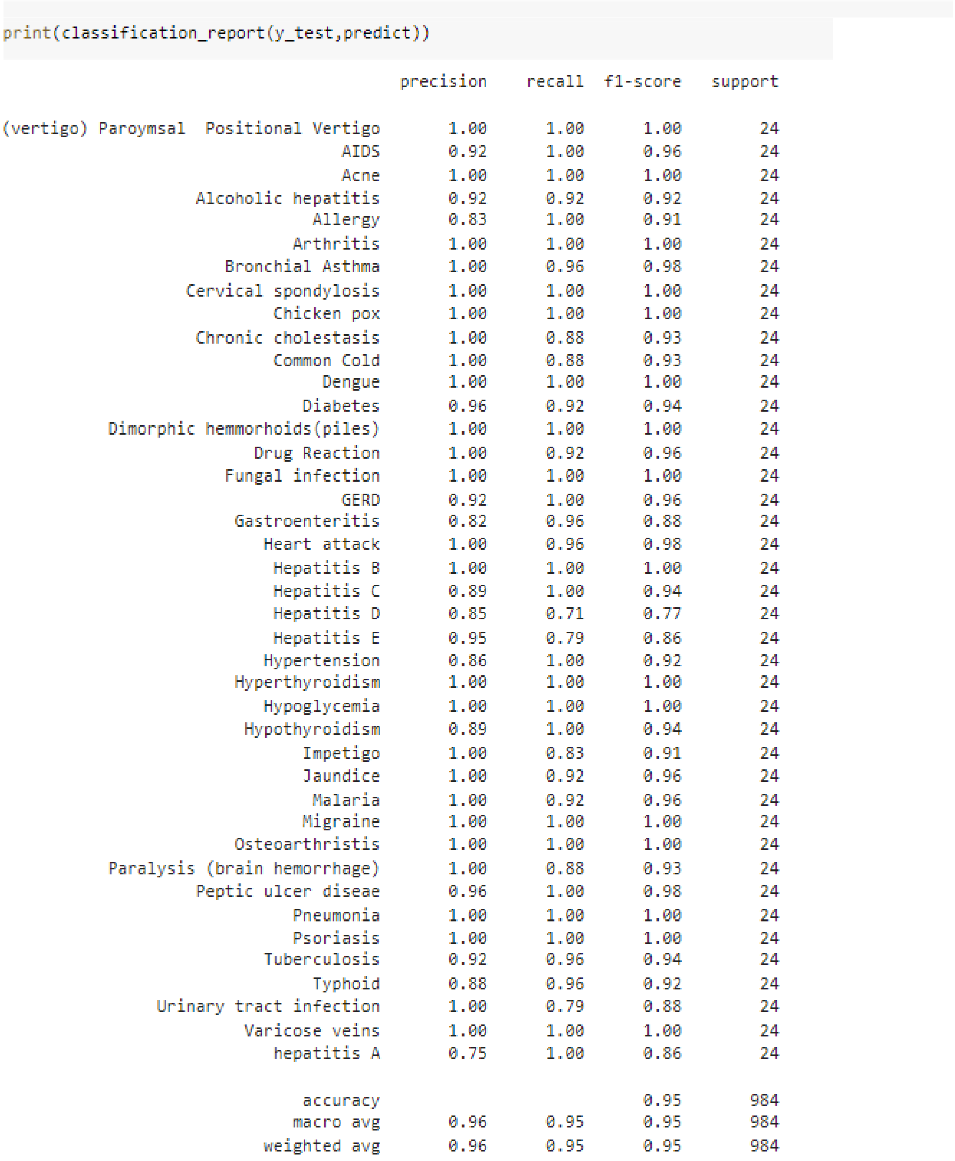
SVM Model:



```
clf_svc= SVC()
clf_svc.fit(X_train,y_train)

SVC()
```

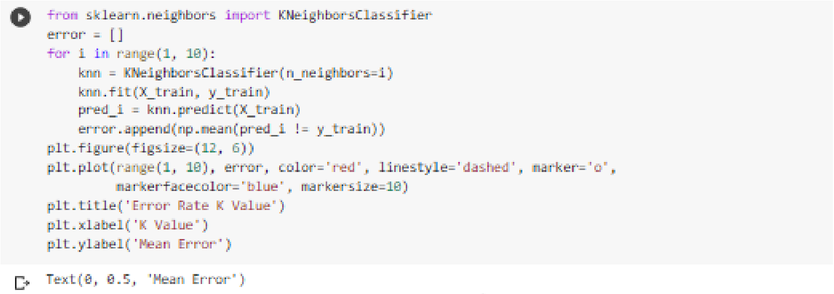
```
[81] predict = clf_svc.predict(X_test)
```

```
[82] print('Accuracy Score: {}%'.format(round(accuracy_score(y_test,predict)*100,2)))

Accuracy Score: 95.43%
```



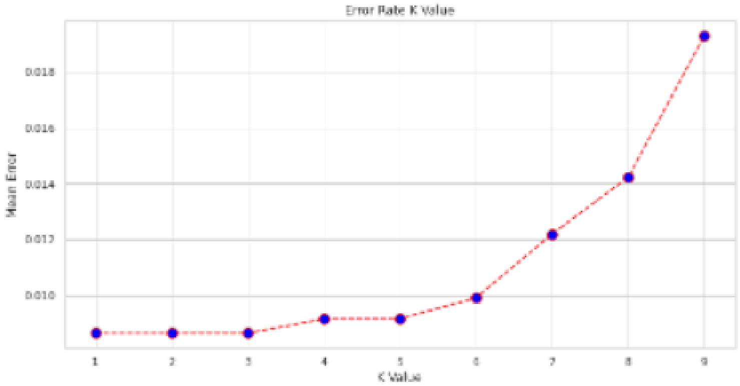
```
print(classification_report(y_test,predict))

                                         precision    recall  f1-score   support

(vertigo) Paroymsal  Positional Vertigo       1.00      1.00      1.00        24
                                   AIDS       0.92      1.00      0.96        24
                                   Acne       1.00      1.00      1.00        24
                    Alcoholic hepatitis       0.92      0.92      0.92        24
                                Allergy       0.83      1.00      0.91        24
                              Arthritis       1.00      1.00      1.00        24
                       Bronchial Asthma       1.00      0.96      0.98        24
                   Cervical spondylosis       1.00      1.00      1.00        24
                            Chicken pox       1.00      1.00      1.00        24
                    Chronic cholestasis       1.00      0.88      0.93        24
                            Common Cold       1.00      0.88      0.93        24
                                 Dengue       1.00      1.00      1.00        24
                               Diabetes       0.96      0.92      0.94        24
           Dimorphic hemmorhoids(piles)       1.00      1.00      1.00        24
                          Drug Reaction       1.00      0.92      0.96        24
                        Fungal infection       1.00      1.00      1.00        24
                                   GERD       0.92      1.00      0.96        24
                        Gastroenteritis       0.82      0.96      0.88        24
                           Heart attack       1.00      0.96      0.98        24
                            Hepatitis B       1.00      1.00      1.00        24
                            Hepatitis C       0.89      1.00      0.94        24
                            Hepatitis D       0.85      0.71      0.77        24
                            Hepatitis E       0.95      0.79      0.86        24
                           Hypertension       0.86      1.00      0.92        24
                        Hyperthyroidism       1.00      1.00      1.00        24
                           Hypoglycemia       1.00      1.00      1.00        24
                         Hypothyroidism       0.89      1.00      0.94        24
                               Impetigo       1.00      0.83      0.91        24
                               Jaundice       1.00      0.92      0.96        24
                                Malaria       1.00      0.92      0.96        24
                               Migraine       1.00      1.00      1.00        24
                        Osteoarthristis       1.00      1.00      1.00        24
           Paralysis (brain hemorrhage)       1.00      0.88      0.93        24
                    Peptic ulcer diseae       0.96      1.00      0.98        24
                              Pneumonia       1.00      1.00      1.00        24
                              Psoriasis       1.00      1.00      1.00        24
                           Tuberculosis       0.92      0.96      0.94        24
                                Typhoid       0.88      0.96      0.92        24
                Urinary tract infection       1.00      0.79      0.88        24
                         Varicose veins       1.00      1.00      1.00        24
                            hepatitis A       0.75      1.00      0.86        24

                               accuracy                           0.95       984
                              macro avg       0.96      0.95      0.95       984
                           weighted avg       0.96      0.95      0.95       984
```

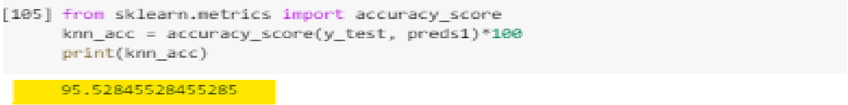
KNN Model:



```
from sklearn.neighbors import KNeighborsClassifier
error = []
for i in range(1, 10):
    knn = KNeighborsClassifier(n_neighbors=i)
    knn.fit(X_train, y_train)
    pred_i = knn.predict(X_train)
    error.append(np.mean(pred_i != y_train))
plt.figure(figsize=(12, 6))
plt.plot(range(1, 10), error, color='red', linestyle='dashed', marker='o',
         markerfacecolor='blue', markersize=10)
plt.title('Error Rate K Value')
plt.xlabel('K Value')
plt.ylabel('Mean Error')
```

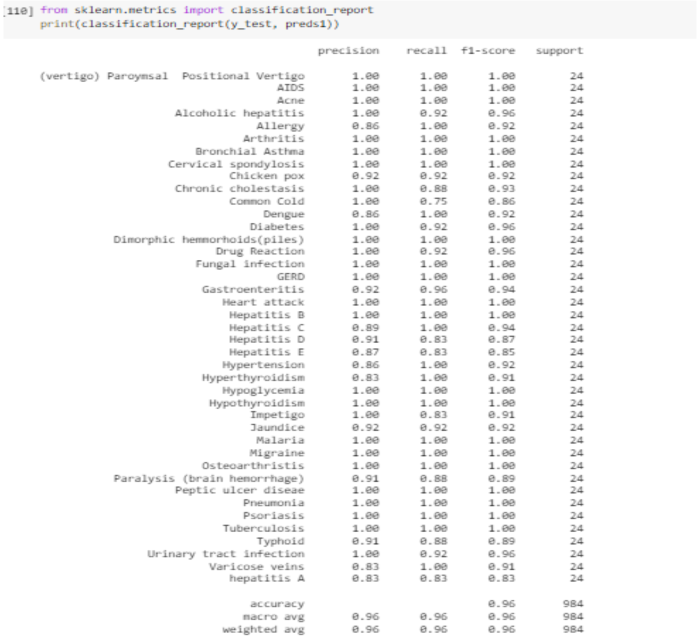
```
Text(0, 0.5, 'Mean Error')
```





```
[105] from sklearn.metrics import accuracy_score
      knn_acc = accuracy_score(y_test, preds1)*100
      print(knn_acc)
```

```
      95.52845528455285
```



```
[110] from sklearn.metrics import classification_report
      print(classification_report(y_test, preds1))
```

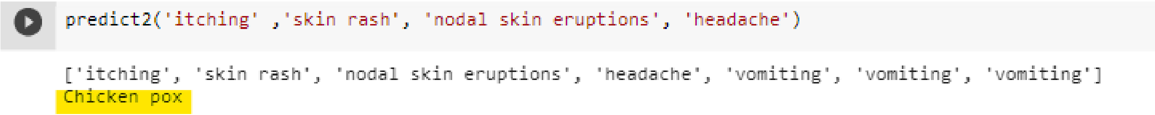
```
                                          precision    recall  f1-score   support

(vertigo) Paroymsal  Positional Vertigo       1.00      1.00      1.00        24
                                   AIDS       1.00      1.00      1.00        24
                                   Acne       1.00      1.00      1.00        24
                    Alcoholic hepatitis       1.00      0.92      0.96        24
                                Allergy       0.86      1.00      0.92        24
                              Arthritis       1.00      1.00      1.00        24
                       Bronchial Asthma       1.00      1.00      1.00        24
                   Cervical spondylosis       1.00      1.00      1.00        24
                            Chicken pox       0.92      0.92      0.92        24
                    Chronic cholestasis       1.00      0.88      0.93        24
                            Common Cold       1.00      0.75      0.86        24
                                 Dengue       0.86      1.00      0.92        24
                               Diabetes       1.00      0.92      0.96        24
           Dimorphic hemmorhoids(piles)       1.00      1.00      1.00        24
                          Drug Reaction       1.00      0.92      0.96        24
                       Fungal infection       1.00      1.00      1.00        24
                                   GERD       1.00      1.00      1.00        24
                        Gastroenteritis       0.92      0.96      0.94        24
                           Heart attack       1.00      1.00      1.00        24
                            Hepatitis B       1.00      1.00      1.00        24
                            Hepatitis C       0.89      1.00      0.94        24
                            Hepatitis D       0.91      0.83      0.87        24
                            Hepatitis E       0.87      0.83      0.85        24
                           Hypertension       0.86      1.00      0.92        24
                        Hyperthyroidism       0.83      1.00      0.91        24
                           Hypoglycemia       1.00      1.00      1.00        24
                         Hypothyroidism       1.00      1.00      1.00        24
                               Impetigo       1.00      0.83      0.91        24
                               Jaundice       0.92      0.92      0.92        24
                                Malaria       1.00      1.00      1.00        24
                               Migraine       1.00      1.00      1.00        24
                        Osteoarthristis       1.00      1.00      1.00        24
           Paralysis (brain hemorrhage)       0.91      0.88      0.89        24
                     Peptic ulcer diseae       1.00      1.00      1.00        24
                              Pneumonia       1.00      1.00      1.00        24
                              Psoriasis       1.00      1.00      1.00        24
                           Tuberculosis       1.00      1.00      1.00        24
                                Typhoid       0.91      0.88      0.89        24
                Urinary tract infection       1.00      0.92      0.96        24
                         Varicose veins       0.83      1.00      0.91        24
                            hepatitis A       0.83      0.83      0.83        24

                               accuracy                           0.96       984
                              macro avg       0.96      0.96      0.96       984
                           weighted avg       0.96      0.96      0.96       984
```

Results:

We found the results of the accuracy and F1 score and predicted for unknown symptoms for each model. For example, here we have given random symptoms, and we got the prediction for Logistic Regression as chicken pox, for Random Forest also it is chicken pox, for KNN also it is chicken pox but for support vector machine it has predicted Hepatitis B.

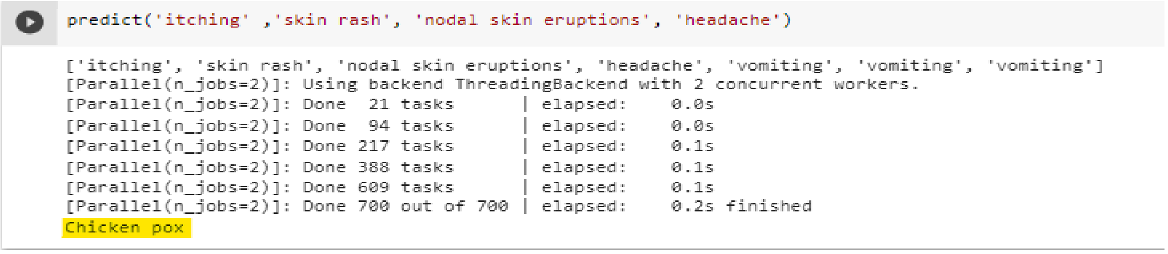
	Logistic regression	Random Forest	Support Vector Machine	KNN
Accuracy	0.84	0.94	0.96	0.95
F-1 Score	0.84	0.98	0.96	0.96
Prediction	Chicken Pox	Chicken Pox	Hepatitis B	Chicken Pox

Logestic Regression Prediction :



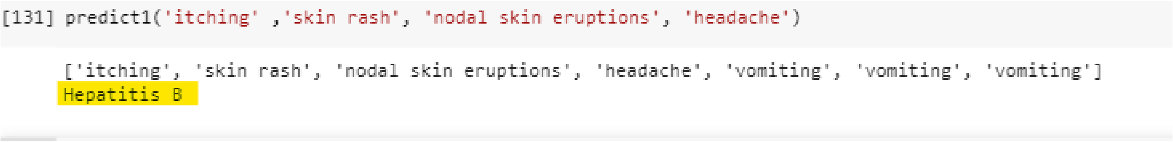
```
predict2('itching' ,'skin rash', 'nodal skin eruptions', 'headache')

['itching', 'skin rash', 'nodal skin eruptions', 'headache', 'vomiting', 'vomiting', 'vomiting']
Chicken pox
```

Random Forest Classifier Prediction:



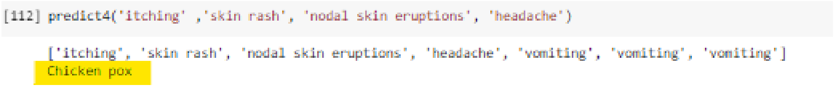
```
predict('itching' ,'skin rash', 'nodal skin eruptions', 'headache')

['itching', 'skin rash', 'nodal skin eruptions', 'headache', 'vomiting', 'vomiting', 'vomiting']
[Parallel(n_jobs=2)]: Using backend ThreadingBackend with 2 concurrent workers.
[Parallel(n_jobs=2)]: Done  21 tasks      | elapsed:    0.0s
[Parallel(n_jobs=2)]: Done  94 tasks      | elapsed:    0.0s
[Parallel(n_jobs=2)]: Done 217 tasks      | elapsed:    0.1s
[Parallel(n_jobs=2)]: Done 388 tasks      | elapsed:    0.1s
[Parallel(n_jobs=2)]: Done 609 tasks      | elapsed:    0.1s
[Parallel(n_jobs=2)]: Done 700 out of 700 | elapsed:    0.2s finished
Chicken pox
```

SVM Model:



```
[131] predict1('itching' ,'skin rash', 'nodal skin eruptions', 'headache')

['itching', 'skin rash', 'nodal skin eruptions', 'headache', 'vomiting', 'vomiting', 'vomiting']
Hepatitis B
```

KNN:



```
[112] predict4('itching' ,'skin rash', 'nodal skin eruptions', 'headache')

['itching', 'skin rash', 'nodal skin eruptions', 'headache', 'vomiting', 'vomiting', 'vomiting']
Chicken pox
```

Conclusion:

We have taken four machine learning models which predicts result based on the symptoms given. Except for SVM model, rest three models give same result for above symptoms. When other symptoms are given SVM is giving the right result whereas one of the other three models are predicting different result. There is a scope of improvement in the project and as for now we are taking the majority.

Role of Each Team Member:

Mehul Gupta: Made all the four Models and Parameter Tuning code and made the related Report.

Saraschandrika Addanki: Did Data Cleaning/preprocessing code, Report and Presented during the class.

Shivani Erigineni: Made four Models and predictions code, Report and made the PowerPoint Presentation