Data Analytics and Visualization

DS250

Assignment 1C

Animesh Raj : 11940120

Puja Bansal : 11940910

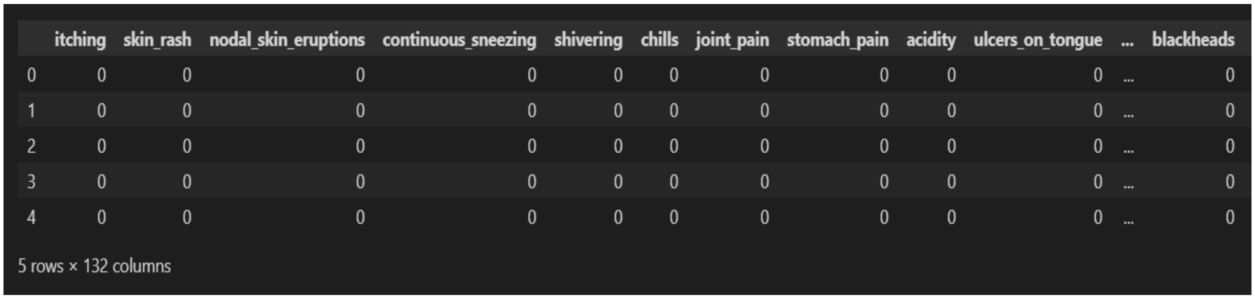
INTRODUCTION :

In this assignment we have made a medical diagnosis app in which the user can enter the symptoms they are facing and our model can predict the disease as per the information given. It would also show the description and the precautions that one must take for the predicted disease.

Step 1: Preparation of the data in a suitable format

We have imported dataset.csv and symptoms-severity.csv files from kaggle (Kaggle link) and extracted unique symptoms and diseases from the dataset.

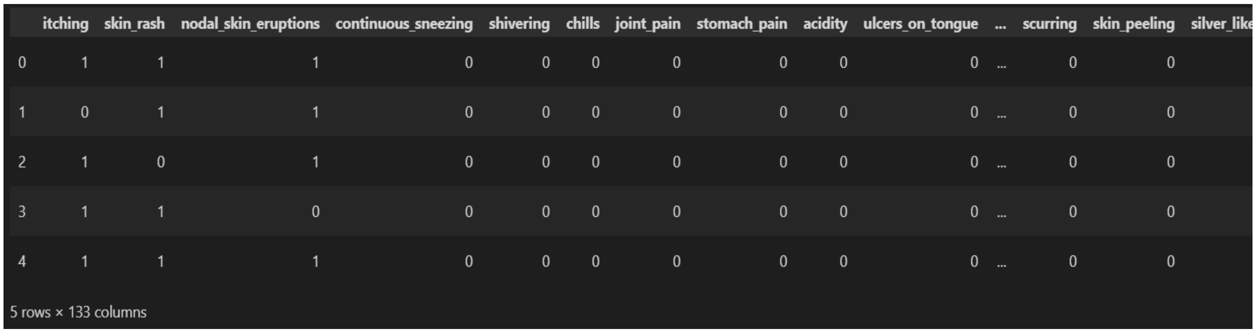
We have removed trailing white spaces from all symptom columns and also replaced NaN with integer ‘0’. We have created a table with all symptoms as the columns attribute and initialize it with 0.



	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	blackheads
0	0	0	0	0	0	0	0	0	0	0	...	0
1	0	0	0	0	0	0	0	0	0	0	...	0
2	0	0	0	0	0	0	0	0	0	0	...	0
3	0	0	0	0	0	0	0	0	0	0	...	0
4	0	0	0	0	0	0	0	0	0	0	...	0

5 rows × 132 columns

We have compared each row of symptoms of the dataset.csv file with this symptoms_table and if the symptoms match, then we have put integer ‘1’ in that place. In a nutshell, we have created a binary nominal attribute table of symptoms.



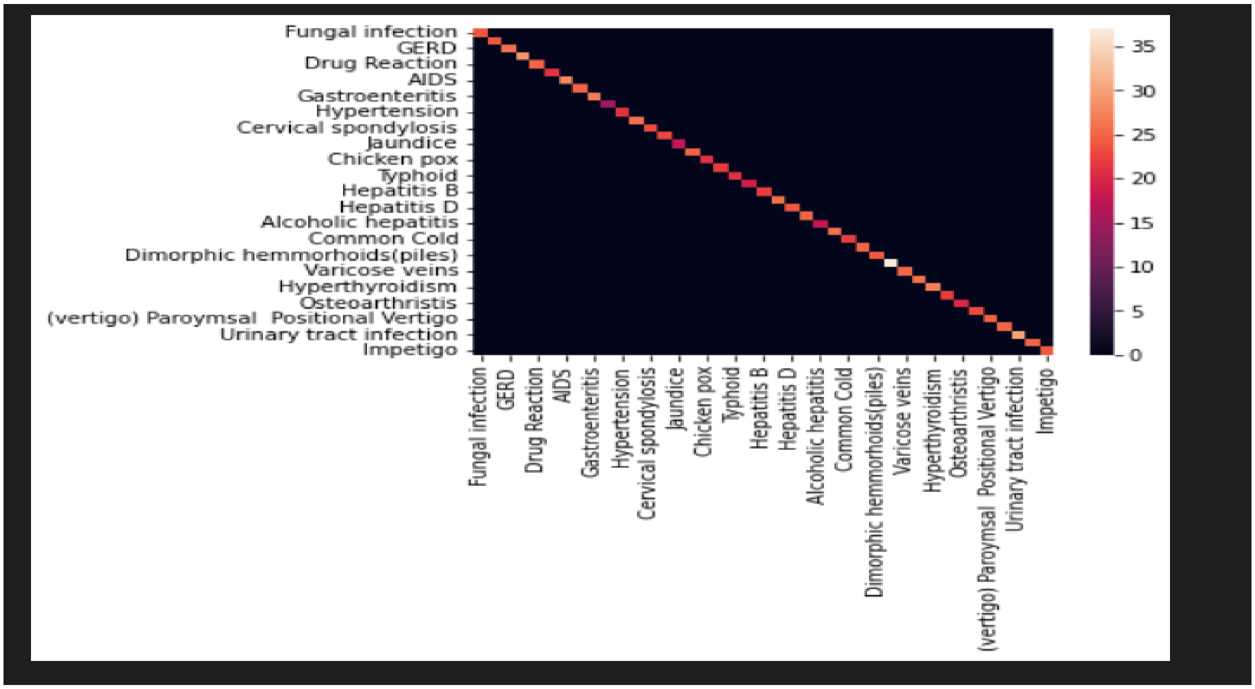
	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	scurring	skin_peeling	silver_like
0	1	1	1	0	0	0	0	0	0	0	...	0	0	
1	0	1	1	0	0	0	0	0	0	0	...	0	0	
2	1	0	1	0	0	0	0	0	0	0	...	0	0	
3	1	1	0	0	0	0	0	0	0	0	...	0	0	
4	1	1	1	0	0	0	0	0	0	0	...	0	0	

5 rows × 133 columns

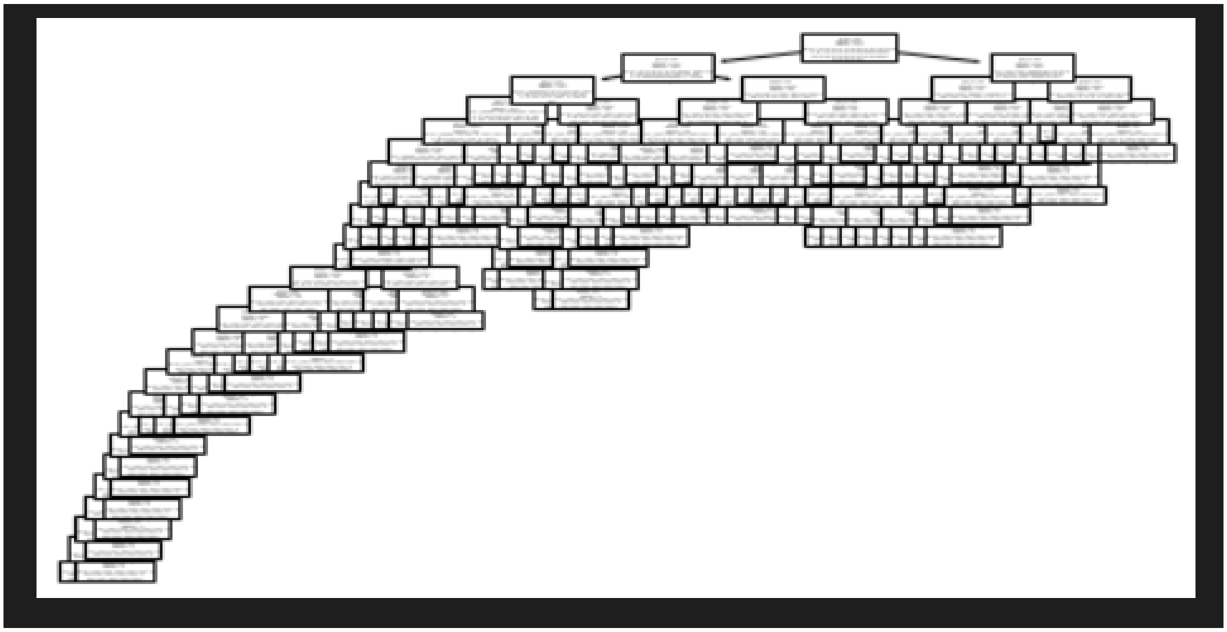
Step 2: Training a decision tree

A) We have used the decision tree classifier of scikit-learn directly.

We have plotted a heat map and decision tree with max depth.

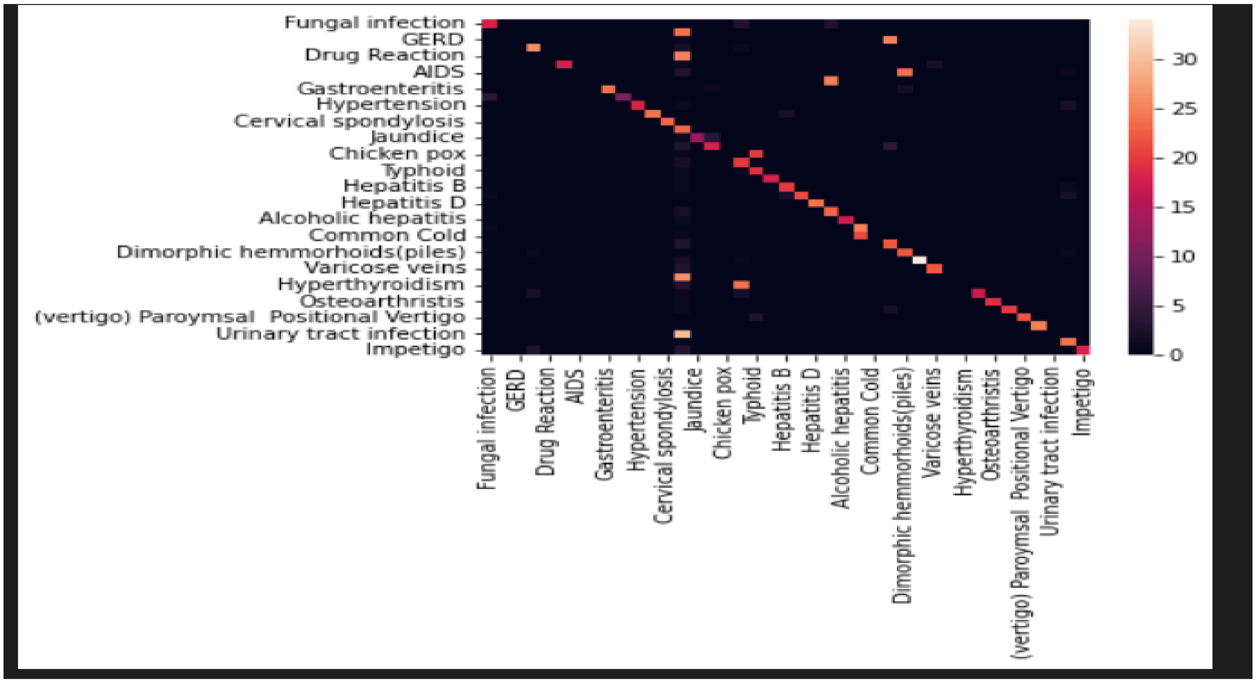


Heat Map for max_depth

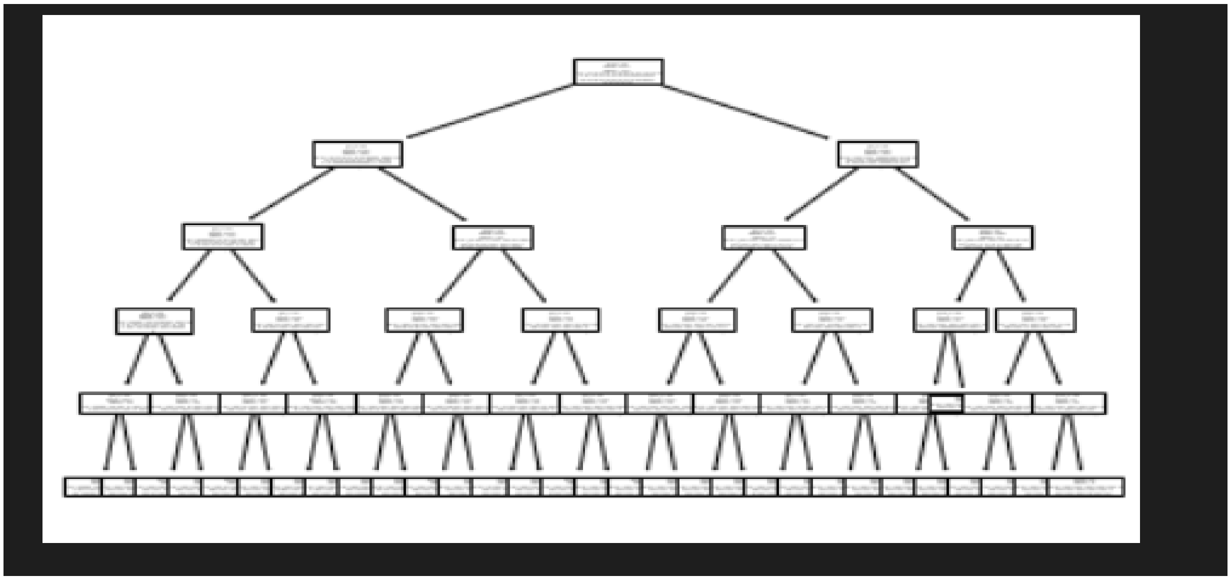


Decision Tree for max_depth

We have plotted a heat map and decision tree with max depth = 5. Now we could see the outliers in the heat map and the accuracy also decreases a bit.



Heat Map for max_depth = 5



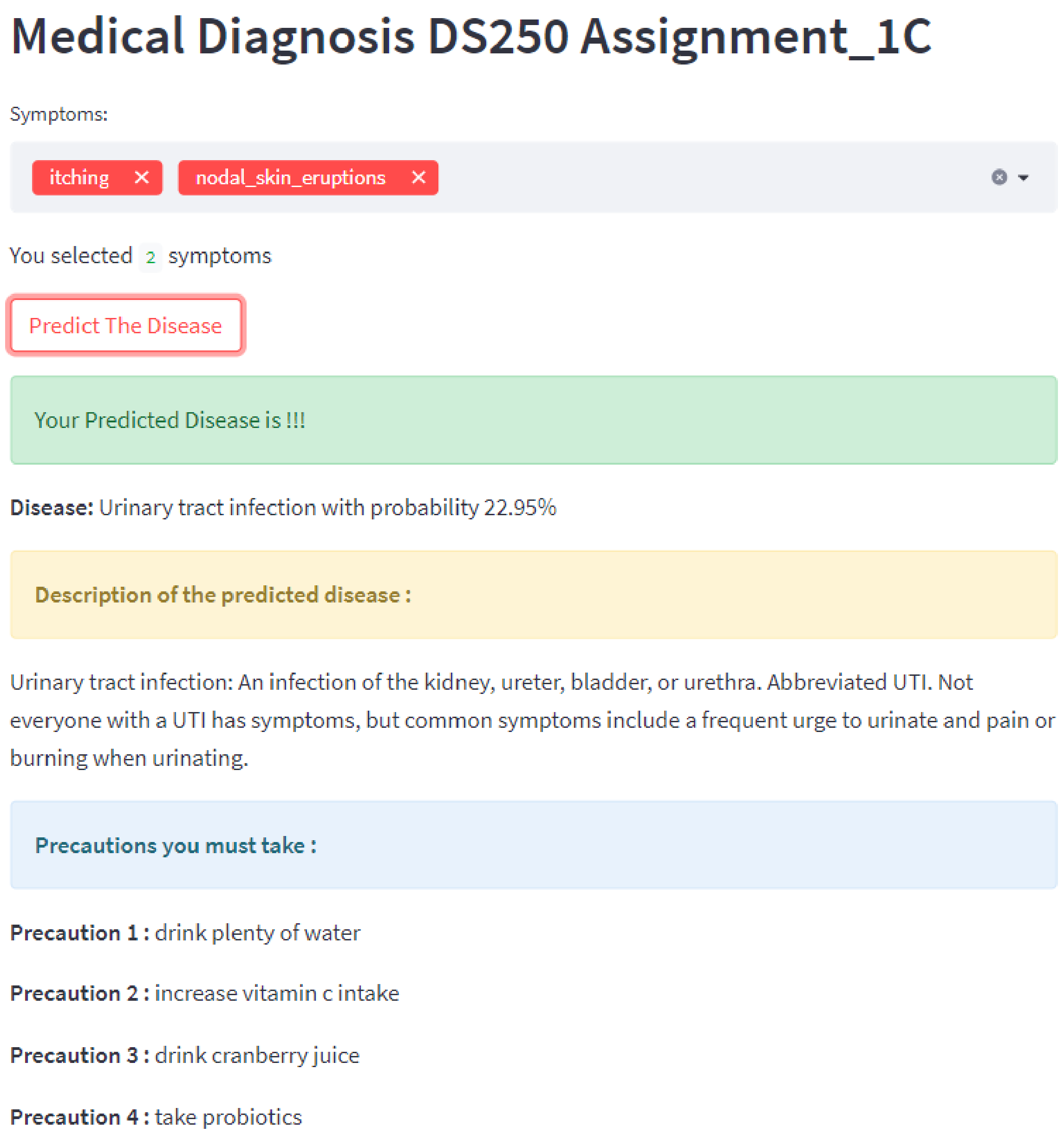
Decision Tree for max_depth = 5

B) We have implemented our own decision tree classifier with the help of entropy and information gain function. If information gain is greater than the previous one, update the max_information_gain and max_information_gain_index and splitting the decision tree further in this manner.

Step 3: Interactive Web app

We have implemented our model in streamlit. We have made a multi-selector so that the user can input more than 1 symptom that the user is facing.

As a result, it shows the predicted disease, description of the predicted disease and the precautions that must be taken for it.



Medical Diagnosis DS250 Assignment_1C

Symptoms:

itching ✕ nodal_skin_eruptions ✕

You selected 2 symptoms

Your Predicted Disease is !!!

Disease: Urinary tract infection with probability 22.95%

Description of the predicted disease :

Urinary tract infection: An infection of the kidney, ureter, bladder, or urethra. Abbreviated UTI. Not everyone with a UTI has symptoms, but common symptoms include a frequent urge to urinate and pain or burning when urinating.

Precautions you must take :

Precaution 1 : drink plenty of water

Precaution 2 : increase vitamin c intake

Precaution 3 : drink cranberry juice

Precaution 4 : take probiotics

Medical Diagnosis Web App(Streamlit)