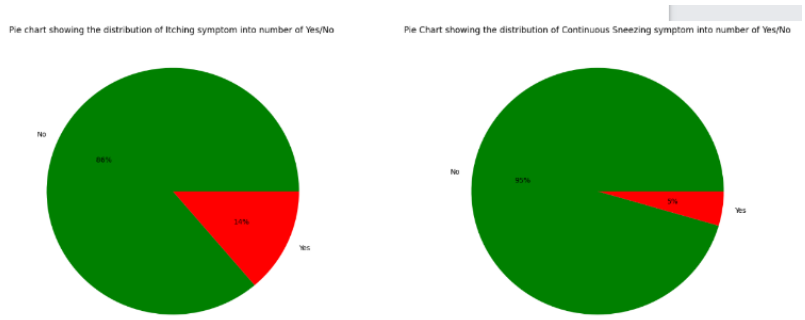
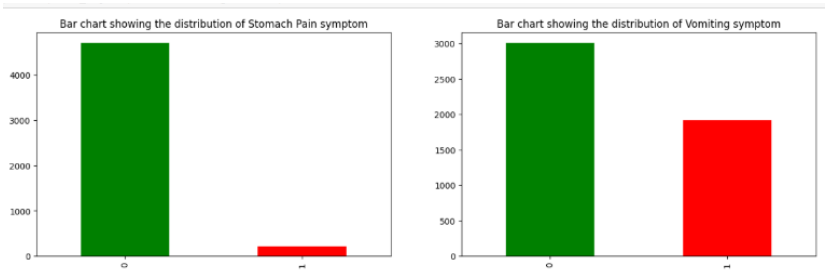
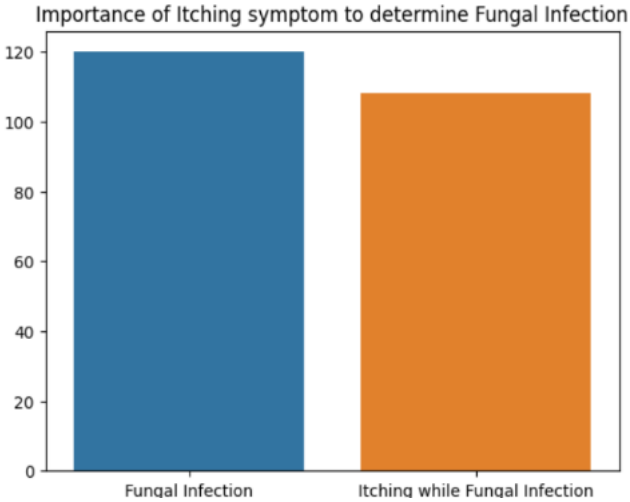
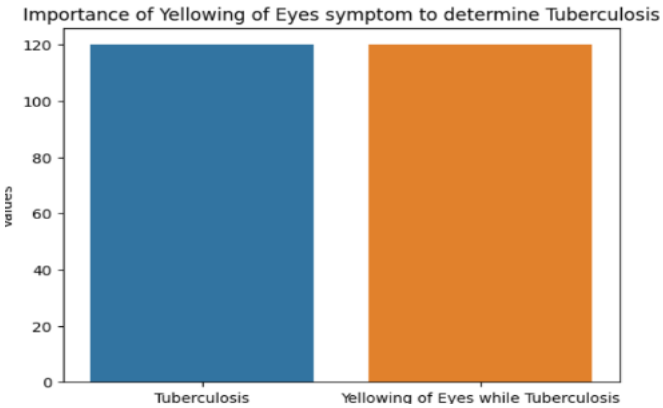
Data Collection and Preprocessing Phase

Date	15 June 2024
Team ID	740003
Project Title	Disease prediction using Machine Learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	**Description**
Data Overview	Dimension: 8 rows x 131 columns Descriptive stastistics: train_data.describe() 8 rows × 131 columns
Univariate Analysis	

	
Bivariate Analysis	
	

Multivariate Analysis	

	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ul
itching	1.000000	0.318158	0.326439	-0.086906	-0.059893	-0.175905	-0.160650	0.202850	-0.086906	
skin_rash	0.318158	1.000000	0.298143	-0.094786	-0.065324	-0.029324	0.171134	0.161784	-0.094786	
nodal_skin_eruptions	0.326439	0.298143	1.000000	-0.032566	-0.022444	-0.065917	-0.060200	-0.032566	-0.032566	
continuous_sneezing	-0.086906	-0.094786	-0.032566	1.000000	0.608981	0.446238	-0.087351	-0.047254	-0.047254	
shivering	-0.059893	-0.065324	-0.022444	0.608981	1.000000	0.295332	-0.060200	-0.032566	-0.032566	
chills	-0.175905	-0.029324	-0.065917	0.446238	0.295332	1.000000	-0.004688	-0.095646	-0.095646	
joint_pain	-0.160650	0.171134	-0.060200	-0.087351	-0.060200	-0.004688	1.000000	-0.087351	-0.087351	
stomach_pain	0.202850	0.161784	-0.032566	-0.047254	-0.032566	-0.095646	-0.087351	1.000000	0.433917	
acidity	-0.086906	-0.094786	-0.032566	-0.047254	-0.032566	-0.095646	-0.087351	0.433917	1.000000	
ulcers_on_tongue	-0.059893	-0.065324	-0.022444	-0.032566	-0.022444	-0.065917	-0.060200	0.649078	0.608981	
muscle_wasting	-0.059893	-0.065324	-0.022444	-0.032566	-0.022444	-0.065917	-0.060200	-0.032566	-0.032566	
vomiting	-0.057763	-0.225046	-0.119543	-0.173459	-0.119543	0.144263	0.199921	0.031406	0.019355	
burning_micturition	0.207896	0.166507	-0.032103	-0.046581	-0.032103	-0.094285	-0.086108	0.412239	-0.046581	
spotting_ urination	0.350585	0.298143	-0.022444	-0.032566	-0.022444	-0.065917	-0.060200	0.608981	-0.032566	
fatigue	0.069744	-0.105248	-0.120465	0.041755	-0.120465	0.269437	0.066652	-0.174797	-0.174797	

Outliers and Anomalies	-

Data Preprocessing Code Screenshots

Loading train Data	

```
train_data=pd.read_csv('/content/Training.csv')
train_data
```

	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	scurr
0	1	1	1	0	0	0	0	0	0	0	...	
1	0	1	1	0	0	0	0	0	0	0	...	
2	1	0	1	0	0	0	0	0	0	0	...	
3	1	1	0	0	0	0	0	0	0	0	...	
4	1	1	1	0	0	0	0	0	0	0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
4915	0	0	0	0	0	0	0	0	0	0	...	

Loading test Data	

```
test_data=pd.read_csv('/content/Testing.csv')
test_data
```

	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	black
0	1	1	1	0	0	0	0	0	0	0	...	
1	0	0	0	1	1	1	0	0	0	0	...	
2	0	0	0	0	0	0	0	1	1	1	...	
3	1	0	0	0	0	0	0	0	0	0	...	
4	1	1	0	0	0	0	0	1	0	0	...	
5	0	0	0	0	0	0	0	0	0	0	...	
6	0	0	0	0	0	0	0	0	0	0	...	
7	0	0	0	0	0	0	0	0	0	0	...	
8	0	0	0	0	0	0	0	0	0	0	...	
9	0	0	0	0	0	0	0	0	0	0	...	
10	0	0	0	0	0	0	0	0	0	0	...	
11	0	0	0	0	0	0	0	0	1	0	...	
12	0	0	0	0	0	0	0	0	0	0	...	

Handling Missing Data
In train and test

```
[ ] train_data.isnull().sum()
```

```
itching                    0
skin_rash                  0
nodal_skin_eruptions       0
continuous_sneezing        0
shivering                  0
                        ...
blister                    0
red_sore_around_nose       0
yellow_crust_ooze          0
prognosis                  0
Unnamed: 133            4920
Length: 134, dtype: int64
```

```
[ ] train_data.isna().sum().sum()
```

```
4920
```

REMOVING NULL COLUMNS IN TRAINING DATA

```
[ ] train_data['Unnamed: 133'].value_counts()
```

```
Series([], Name: count, dtype: int64)
```

```
[ ] train_data.drop("Unnamed: 133",axis = 1,inplace=True)
    train_data.drop("fluid_overload",axis = 1,inplace=True)
```

```
[ ] train_data.shape
```

```
(4920, 132)
```

```
[ ] test_data.isnull().sum()
```

```
itching                0
skin_rash              0
nodal_skin_eruptions   0
continuous_sneezing    0
shivering              0
                      ..
inflammatory_nails     0
blister                0
red_sore_around_nose   0
yellow_crust_ooze      0
prognosis              0
Length: 133, dtype: int64
```

```
test_data.drop("fluid_overload",axis = 1,inplace=True)
```

Data Transformation	
	``python
	from sklearn.preprocessing import LabelEncoder
	label_encoder =LabelEncoder()
	train_data['prognosis']= label_encoder.fit_transform(train_data['prognosis'])
	train_data['prognosis'].unique()
	``
	``
	array([15, 4, 16, 9, 14, 33, 1, 12, 17, 6, 23, 30, 7, 32, 28, 29, 8,
	       11, 37, 40, 19, 20, 21, 22, 3, 36, 10, 34, 13, 18, 39, 26, 24, 25,
	       31, 5, 0, 2, 38, 35, 27])
	``
	``python
	label_encoder =LabelEncoder()
	test_data['prognosis']= label_encoder.fit_transform(test_data['prognosis'])
	test_data['prognosis'].unique()
	``
	``
	array([15, 4, 16, 9, 14, 33, 1, 12, 17, 6, 23, 30, 7, 32, 28, 29, 8,
	       11, 37, 40, 19, 20, 21, 22, 3, 36, 10, 34, 13, 18, 39, 26, 24, 25,
	       31, 5, 0, 2, 38, 35, 27])
	``

Feature Engineering	-
Save Processed Data	-