Disease Prediction

FINAL PROJECT

PRESENTED BY

MEHZABEEN GHEESAH

Project Objective

PREDICT A DISEASE BASED ON THE DESCRIBED SYMPTOMS

Dataset Overview

Name: Symptom2Disease

Source: Kaggle

Rows: 1200

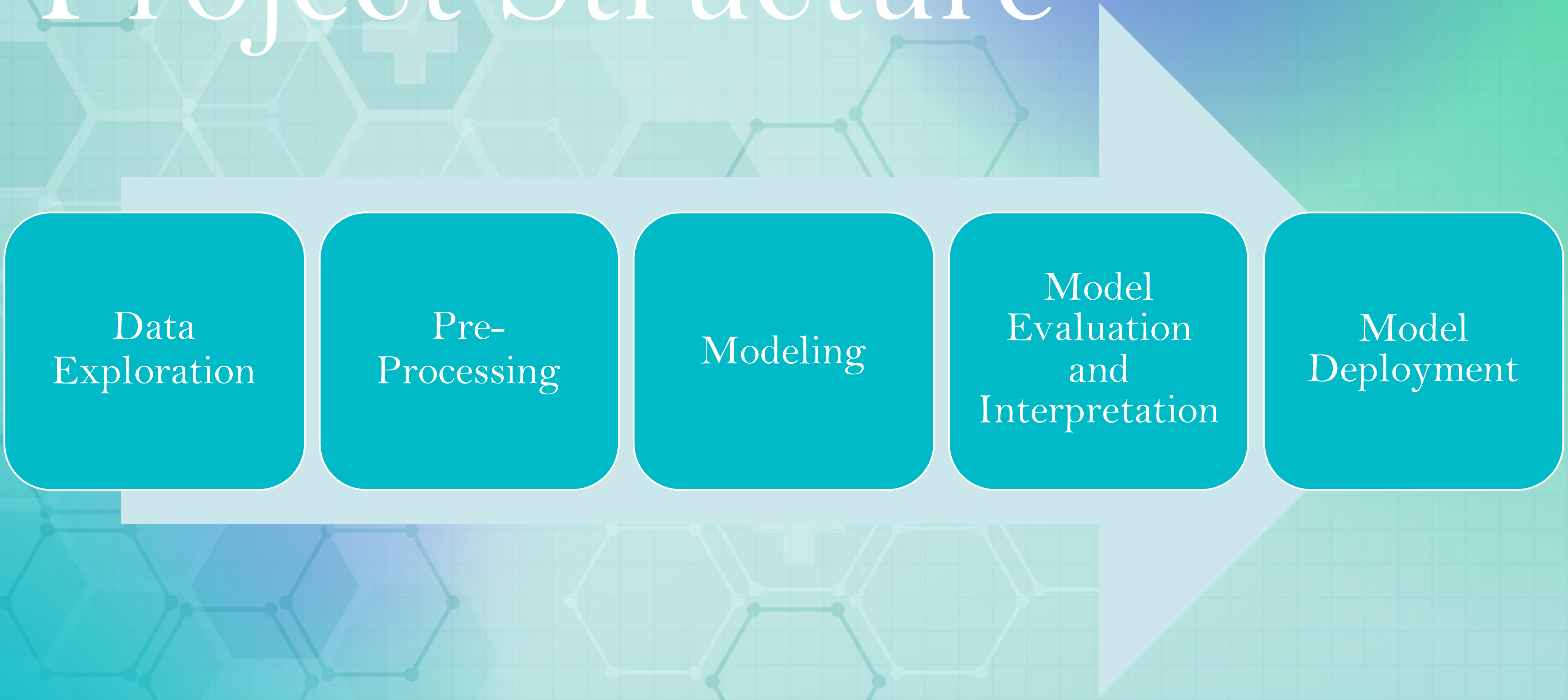
Columns: 3

No missing Values

47 Duplicate Rows

Index	Text	Label
1 - 1200 rows	50 symptom descriptions per label	24 diseases

Project Structure



Data Exploration

Pre-Processing

Modeling

Model Evaluation and Interpretation

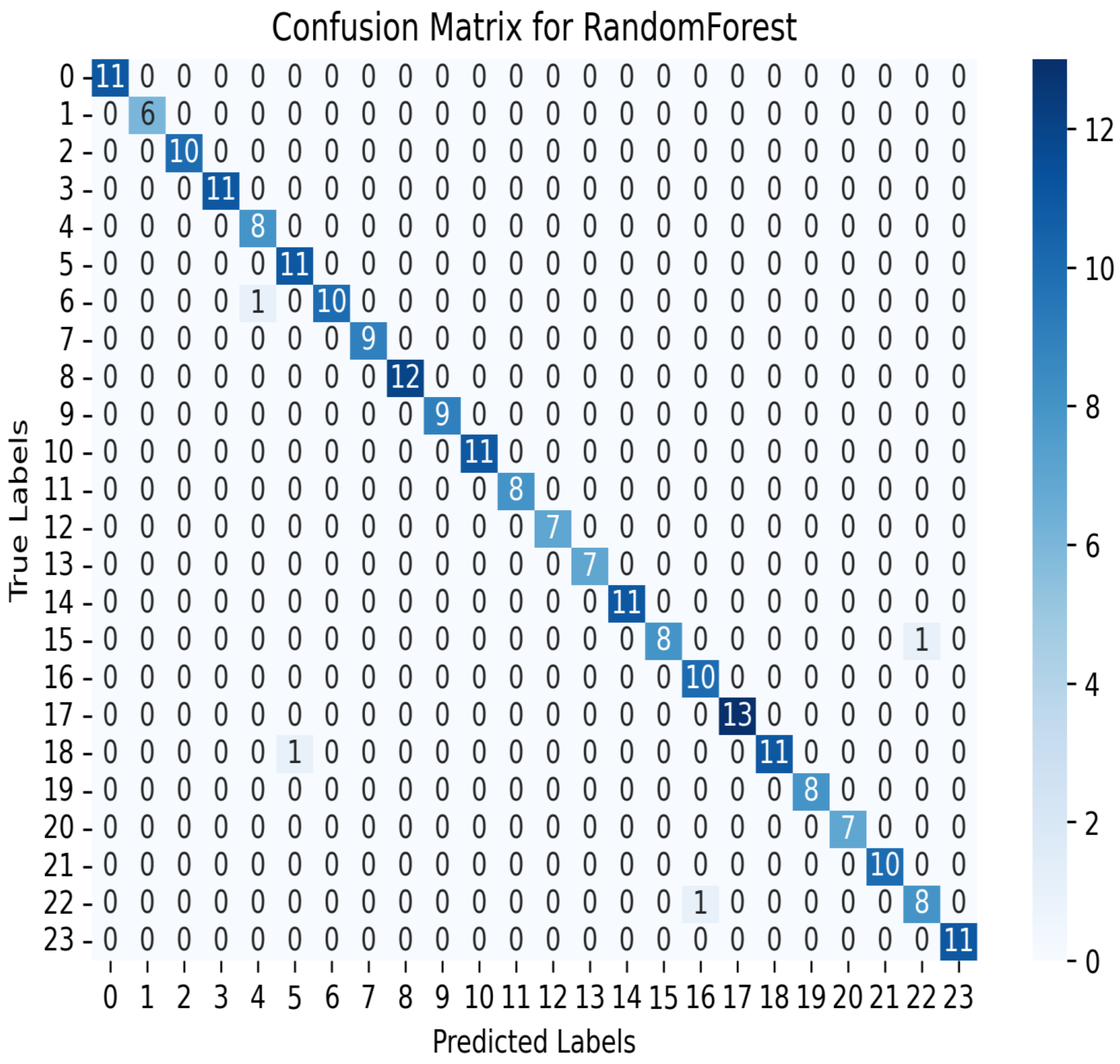
Model Deployment

Models and Results

Models	Accuracy Score before Hyperparameter Tuning		
	CountVectorizer	**Word2Vec**	**TF-IDF**
LinearSVC	94 %	9 %	93 %
RandomForest Classifier	98 %	64 %	96 %

Model with CountVectorizer	Best Parameters	Accuracy Score after Hyperparameter tuning
LinearSVC	`{'C': 0.001, 'max_iter': 1000}`	94 %
RandomForest Classifier	`{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}`	98 %

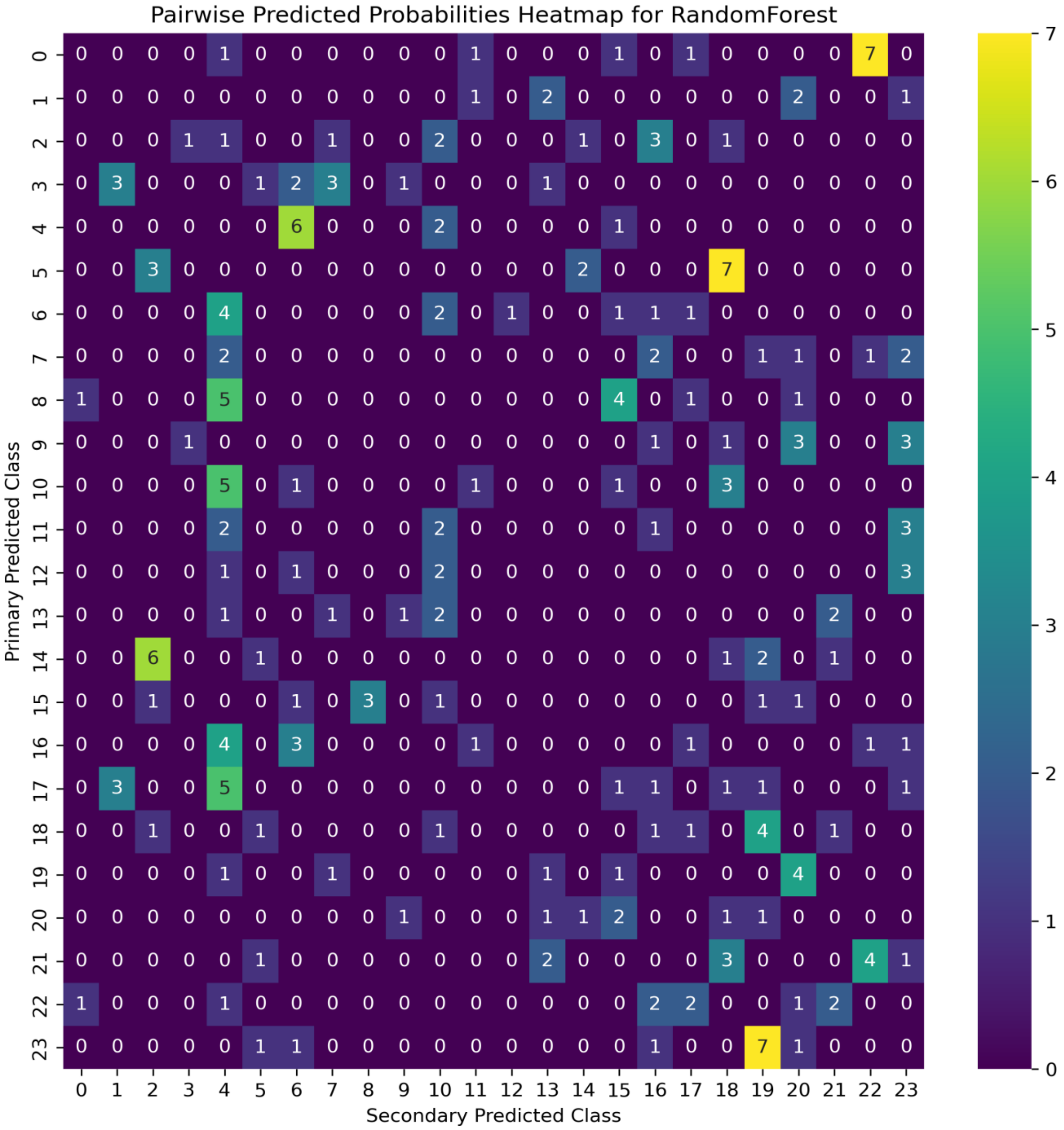
Model Analysis and Evaluation



Confusion Matrix for RandomForest

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	9	0	0	0	0	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	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	1	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0
18	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	8	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11

- Diseases were accurately predicted for a maximum of 13 times
- Confusion matrix shows very few misclassification



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	7	0
1	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	2	0	0	1
2	0	0	0	1	1	0	0	1	0	0	2	0	0	0	1	0	3	0	1	0	0	0	0	0
3	0	3	0	0	0	1	2	3	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	6	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0
5	0	0	3	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	7	0	0	0	0	0
6	0	0	0	0	4	0	0	0	0	0	2	0	1	0	0	1	1	1	0	0	0	0	0	0
7	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	1	0	1	2
8	1	0	0	0	5	0	0	0	0	0	0	0	0	0	0	4	0	1	0	0	1	0	0	0
9	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	3	0	0	3
10	0	0	0	0	5	0	1	0	0	0	0	1	0	0	0	1	0	0	3	0	0	0	0	0
11	0	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	3
12	0	0	0	0	1	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	3
13	0	0	0	0	1	0	0	1	0	1	2	0	0	0	0	0	0	0	0	0	0	2	0	0
14	0	0	6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	1	0	0
15	0	0	1	0	0	0	1	0	3	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
16	0	0	0	0	4	0	3	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1
17	0	3	0	0	5	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	1
18	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	0	4	0	1	0	0
19	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	4	0	0	0
20	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	2	0	0	1	1	0	0	0	0
21	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	3	0	0	0	4	1
22	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	1	2	0	0
23	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	7	1	0	0	0

Model Analysis and Evaluation

- The heatmap highlights some potential confusions between diseases
- Maximum number of confusion 7 times

Conclusion

- Best performing model is the RandomForest Classifier
- Highest accuracy score of 98 %
- Low variability in its predictions
- Some misclassifications

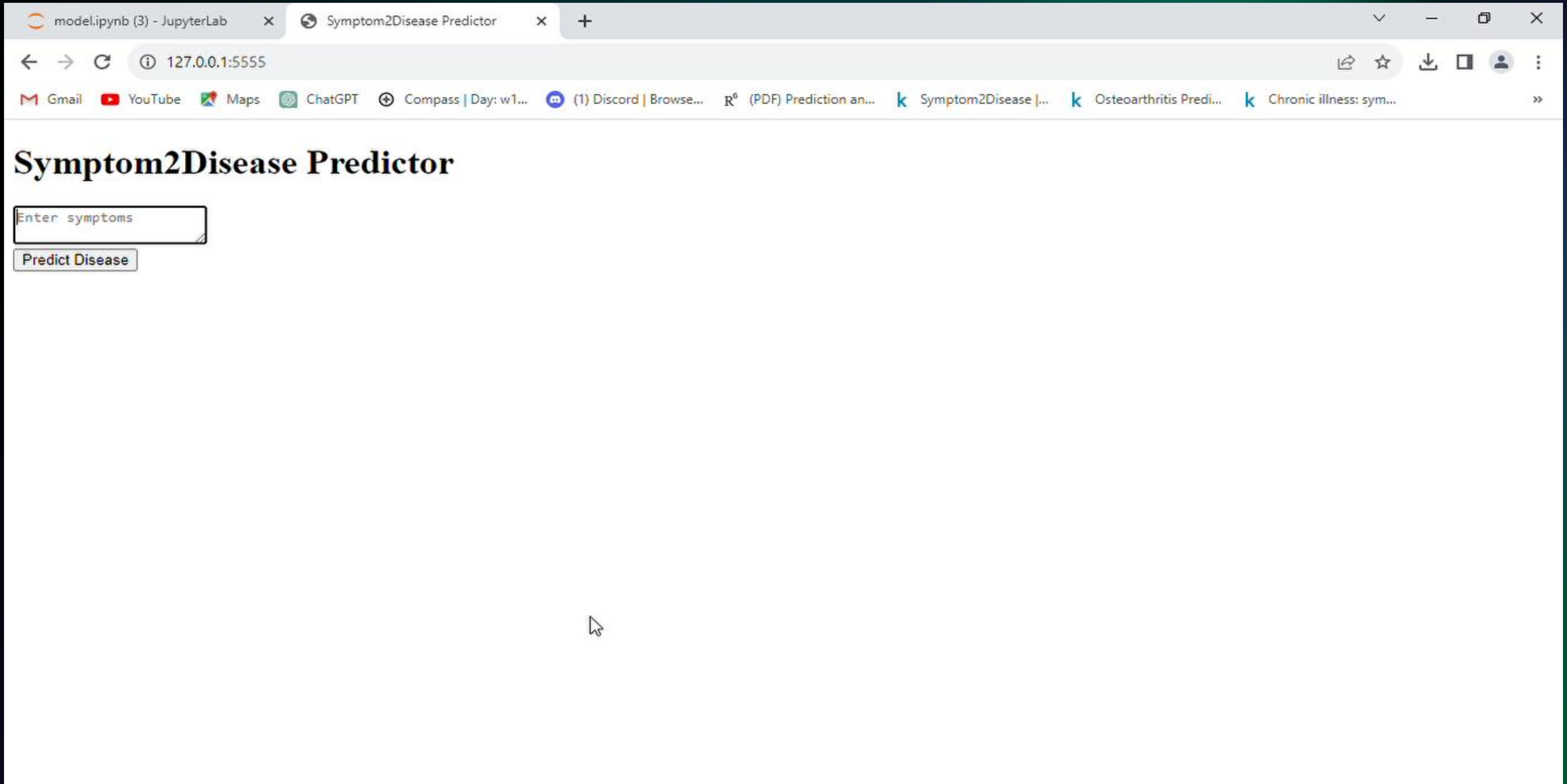
Challenges

- Dealing with text
- Determining the relevance of words
- Hyperparameter Tuning is time consuming

Future Scope

- Refine model performance
- Learn how to effectively remove irrelevant words and noise

Symptom2Disease App Demonstration



Symptom2Disease Predictor

Enter symptoms

Predict Disease

THANK YOU