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

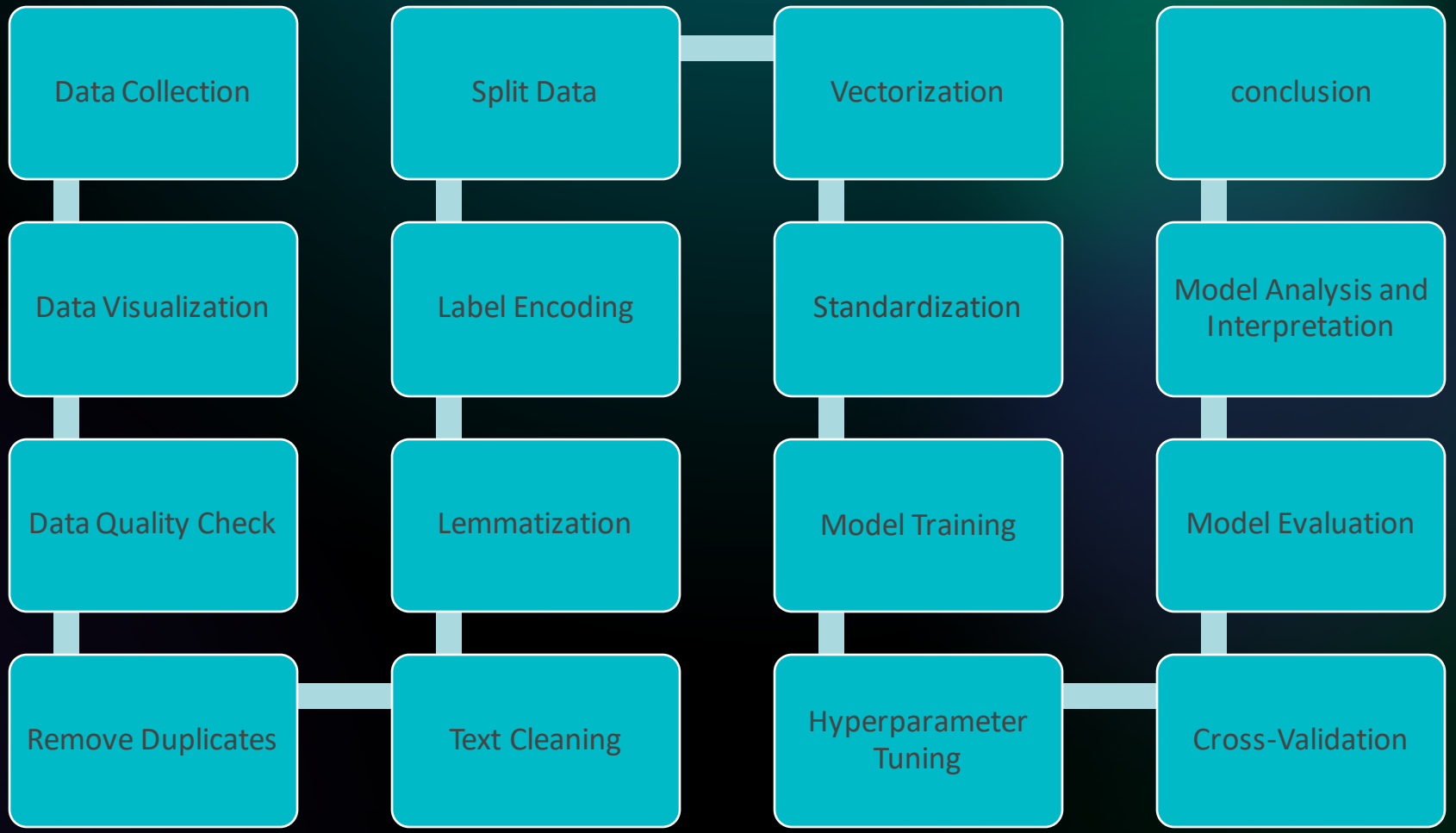
Columns: 3

Unnamed (Index)	Text (Symptoms)	Labels (Diseases)
1 - 1200 rows	50 symptom descriptions per label	24 diseases

No missing Values

47 Duplicate Rows

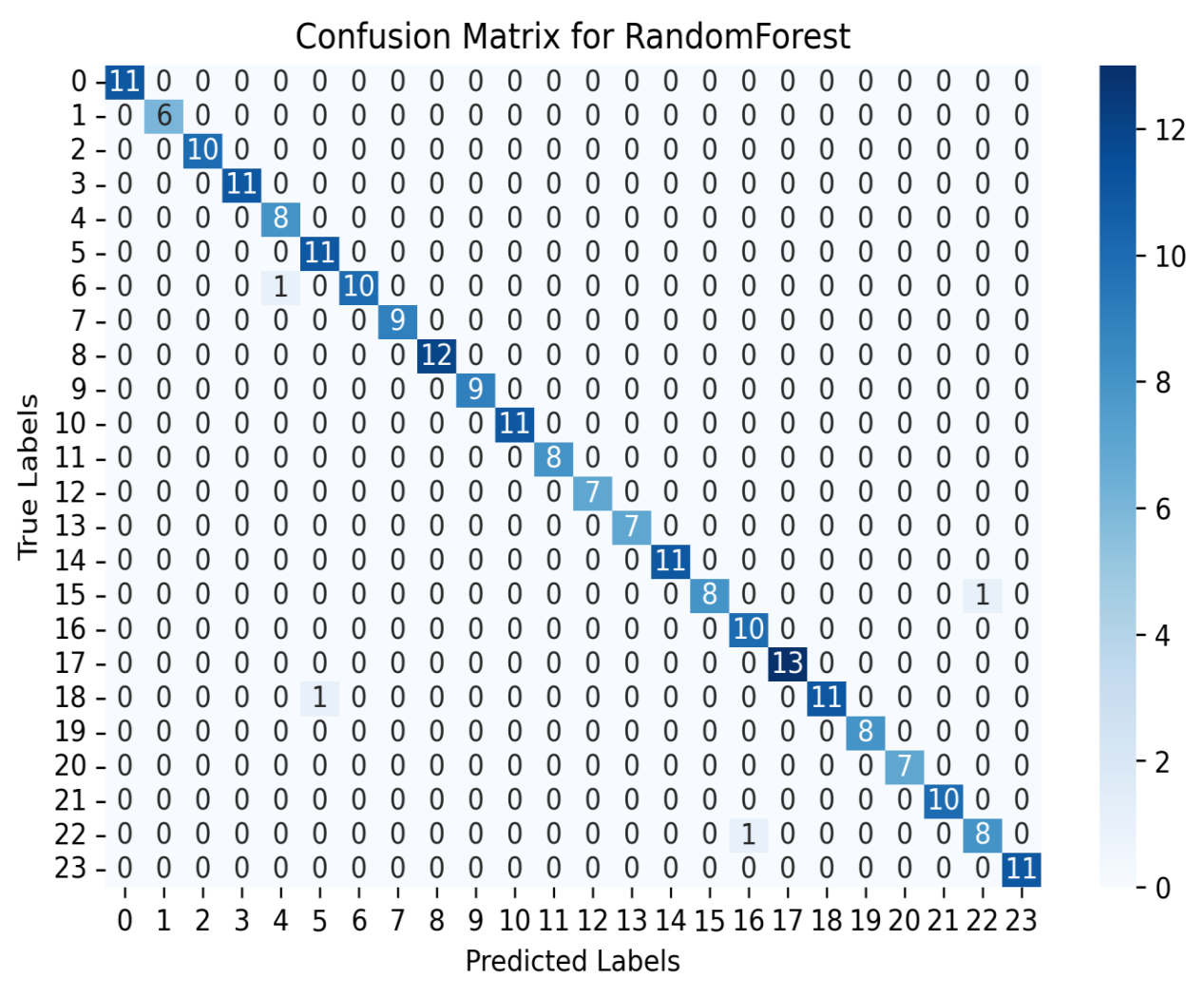
Project Structure



Models and Results

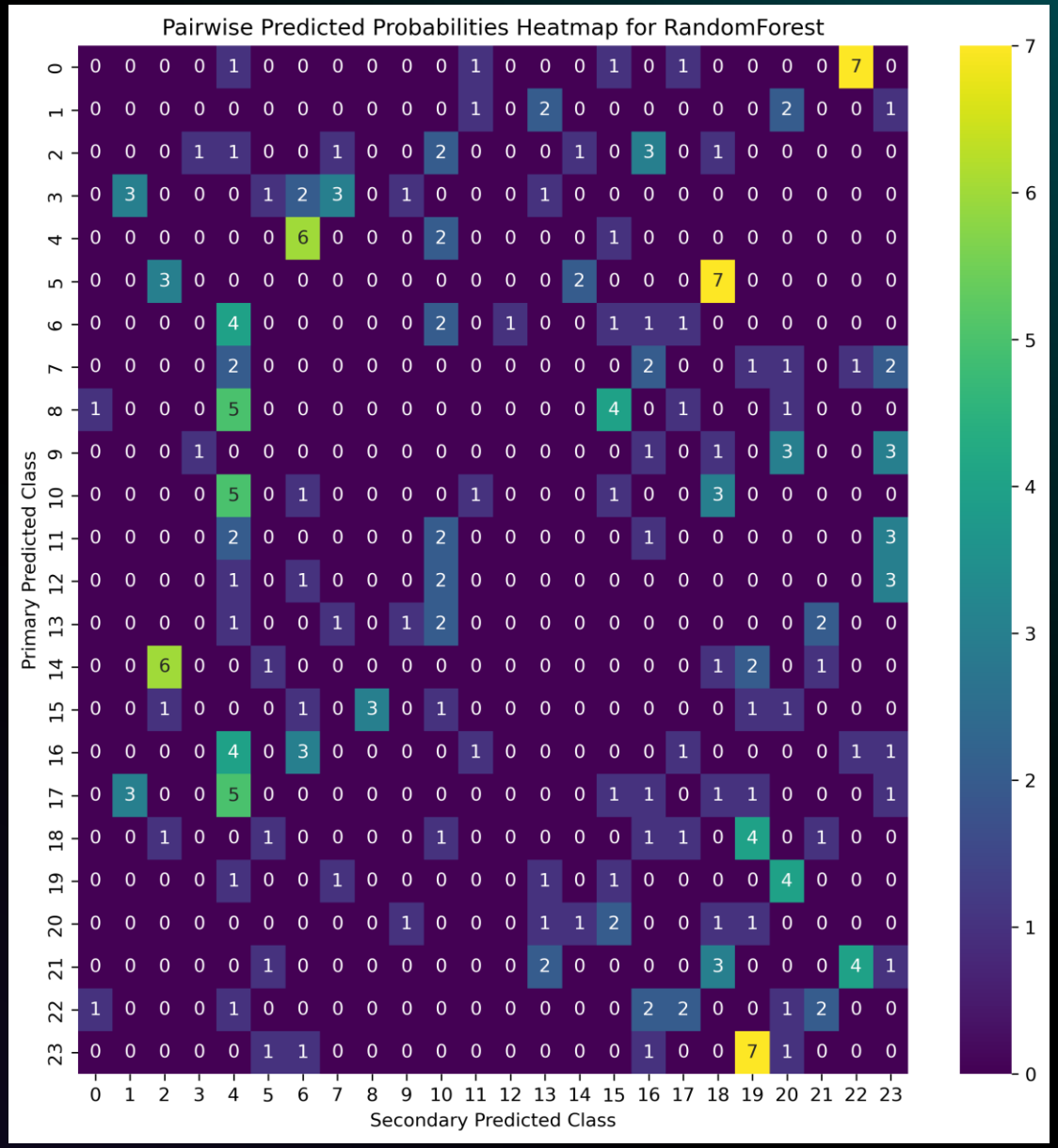
Models	Accuracy Score before Hyperparameter Tuning		
	CountVectorizer	**Word2Vec**	**TF-IDF**
LinearSVC	94 %	9 %	93 %
RandomForest Classifier	98 %	64 %	96 %

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 %



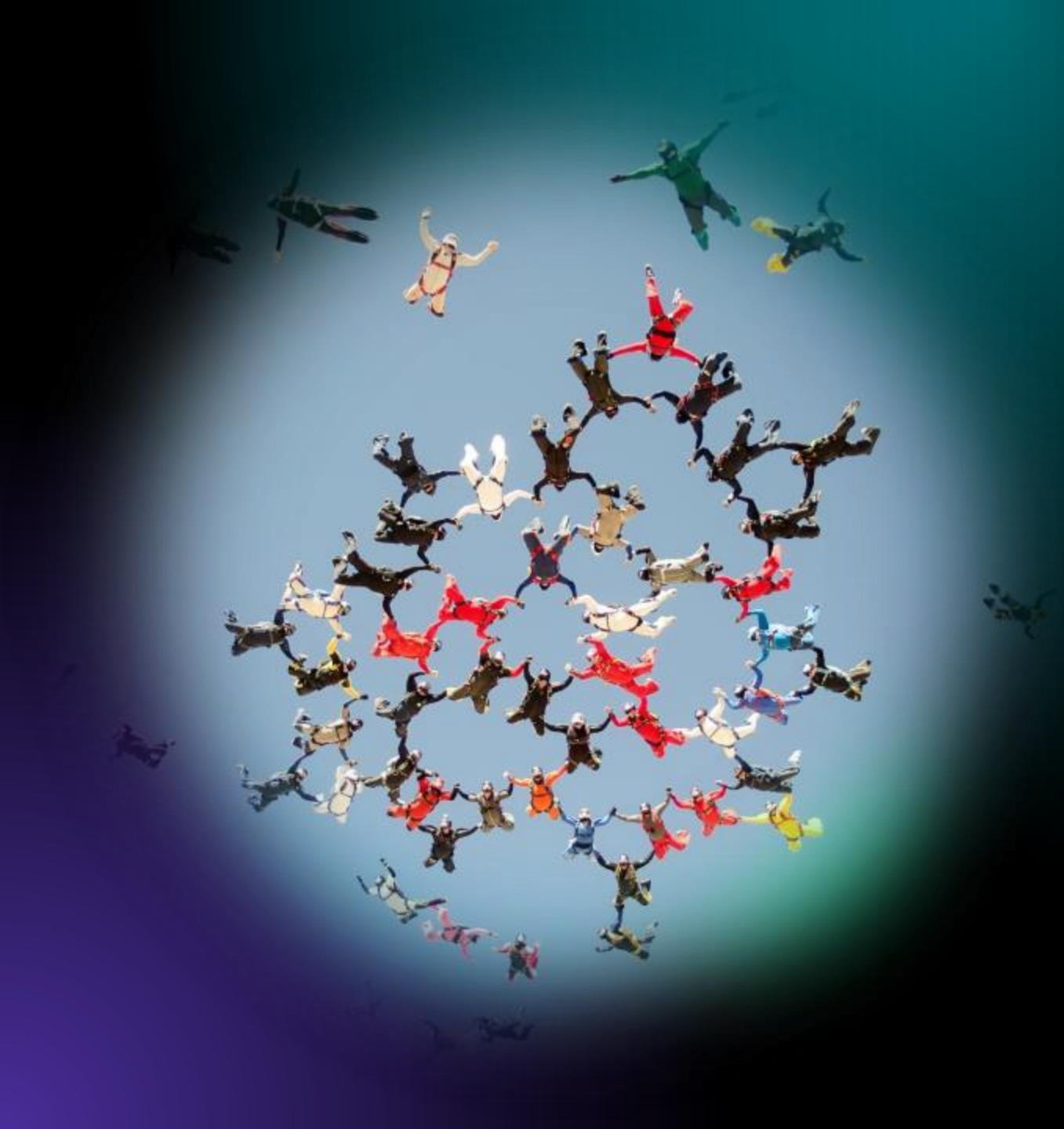
Model Analysis and Evaluation

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



Model Analysis and Evaluation

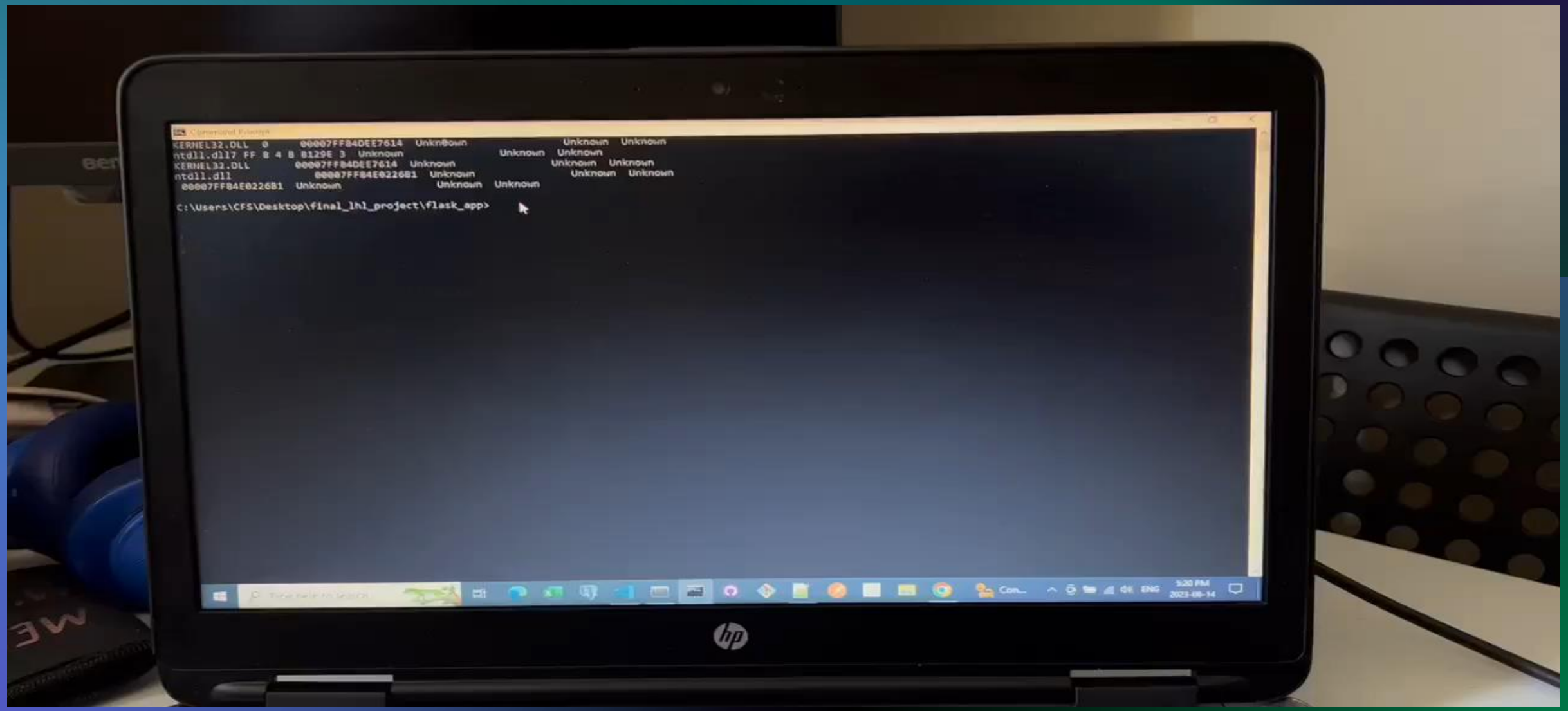
- The heatmap highlights some potential confusions between diseases
- Maximum number of confusion 7 times



Conclusion

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

App Demonstration



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


THANK YOU