

# Disease Prediction from Symptoms – Project (FPU)

## Project Contents

- **Files except .py files and .ipynb files**
  - **training\_dataset.csv** : Contains Diseases and its Symptoms.  
Path: "*\Disease-Prediction-from-Symptoms-master\dataset\training\_data.csv*"
  - **cleaned\_data.csv** : Removed incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within the dataset.  
Path : */content/dataset/cleaned\_data.csv*  
[Notebook or Colab]

- **raw\_data.xlsx** : Contains Disease, Count of Disease Occurrence and its Symptoms  
Path : */content/dataset/raw\_data.xlsx*  
[Notebook or Colab]
- **test\_data.csv** : File that provides a final, real-world check of an unseen dataset to confirm that the ML algorithm was trained effectively.  
Path : *" \Disease-Prediction-from-Symptoms-master\dataset\test\_data.csv"*

- **.py files and .ipynb files**

- **main.py** : code for loading kaggle dataset, training & saving the model for Standalone Program

Path : "*\Disease-Prediction-from-Symptoms-master\main.py*"

- **infer.py** : For running the inference on test set or on custom inputs for Standalone Program

Path : "*\Disease-Prediction-from-Symptoms-master\infer.py*"

- **Disease-Prediction-from-Symptoms-checkpoint.ipynb** : For running an interactive demo for Colab or Jupyter Notebook

Path : "*\Disease-Prediction-from-Symptoms-master\notebook\Disease-Prediction-from-Symptoms-checkpoint.ipynb*"

- **.joblib files**

Joblib is a set of tools to provide lightweight pipelining in Python.

Built by main.py file when executed in case of our project files.

- **decision\_tree.joblib** : Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter.

This is used for disease prediction in our project

Path : "*\Disease-Prediction-from-Symptoms-*

*master\saved\_model\decision\_tree.joblib"*

- **gradient\_boosting.joblib** : It is a technique of producing an additive predictive model by combining various weak predictors, typically Decision Trees.

Uses decision\_tree model in our project and produces an additive predictive model.

Path : "*\Disease-Prediction-from-Symptoms-master\saved\_model\gradient\_boost.joblib*"

- **mnb.joblib** : Multinomial Naive Bayes algorithm is a probabilistic learning method. This works well for data which can easily be turned into counts.

Path : "*\Disease-Prediction-from-Symptoms-master\saved\_model\mnb.joblib*"

- **random\_forest.joblib** : Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. Used in infer.py as in our project.

Path : "*\Disease-Prediction-from-Symptoms-master\saved\_model\random\_forest.joblib*"

- **.yaml files**

- YAML – Yet Another Markup Language**

- This file contains the parameters for a specific type for the project.

- **config.yaml** : This file contains the Project Configuration Parameters

- Path : " \Disease-Prediction-from-Symptoms-master\config.yaml"

- **environment.yml** : This file contains the project environment parameters and dependencies

- Path : " \Disease-Prediction-from-Symptoms-master\environment.yml"