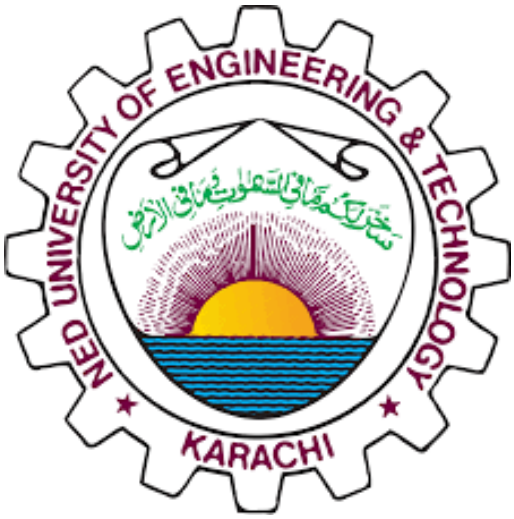


Machine Learning (CS-324)
OEL

"Disease Prediction from Symptoms"

Submitted to: Miss Mahnoor Malik

Prepared by

GROUP MEMBERS	ROLL NO.
Rooman	CS-21008
Syed Maarij	CS-21022
Usama Khalid	CS-21028

TABLE OF CONTENTS

1. INTRODUCTION

This report explores the Disease Prediction dataset from Kaggle, performs data preprocessing and exploratory data analysis (EDA), and builds predictive models using various machine learning algorithms. The primary objective is to predict the disease that a person has based on the features available in the dataset. There is an other data 'Disease-Symptom Knowledge Database' which is related to this Disease Prediction dataset, it gives more information. This dataset has 3 columns Disease, Count of Disease Occurrence, Symptom. Disease Prediction dataset used for model training and 'Disease-Symptom Knowledge Database' used for exploratory data analysis (EDA).

2. DATA COLLECTION

Applying Knowledge to field of Medical Science and making the task of Physician easy is the main purpose of this Disease Prediction dataset. The disease prediction dataset was downloaded from Kaggle. It contains data for 42 diseases. This dataset has 132 symptoms on which 42 different types of diseases can be predicted. It contains 4920 records. 132 features are; itching,skin_rash,nodal_skin_eruptions,continuous_sneezing,shivering,chills,joint_pain,stomach_pain,acidity,ulcers_on_tongue,muscle_wasting,vomiting,burning_micturition,spotting_urination,fatigue,weight_gain,anxiety,cold_hands_and_feets,mood_swings,weight_loss,restlessness,lethargy,patches_in_throat,irregular_sugar_level,cough,high_fever,sunken_eyes,breathlessness,sweating,dehydration,indigestion,headache,yellowish_skin,dark_urine,nausea,loss_of_appetite,pain_behind_the_eyes,back_pain,constipation,abdominal_pain,diarrhoea,mild_fever,yellow_urine,yellowing_of_eyes,acute_liver_failure,fluid_overload,swelling_of_stomach,swelled_lymph_nodes,malaise,blurred_and_distorted_vision,phlegm,throat_irritation,redness_of_eyes,sinus_pressure,runny_nose,congestion,chest_pain,weakness_in_limbs,fast_heart_rate,pain_during_bowel_movements,pain_in_anal_region,bloody_stool,irritation_in_anus,neck_pain,dizziness,cramps,bruising,obesity,swollen_legs,swollen_blood_vessels,puffy_face_and_eyes,enlarged_thyroid,brittle_nails,swollen_extremeties,excessive_hunger,extra_marital_contacts,drying_and_tingling_lips,slurred_speech,knee_pain,hip_joint_pain,muscle_weakness,stiff_neck,swelling_joints,movement_stiffness,spinning_movements,loss_of_balance,unsteadiness,weakness_of_one_body_side,loss_of_smell,bladder_discomfort,foul_smell_ofurine,continuous_feel_of_urine,passage_of_gases,internal_itching,toxic_look_(typhos),depression,irritability,muscle_pain,altered_sensorium,red_spots_over_body,belly_pain,abnormal_menstruation,dischromic_patches,watering_from_eyes,increased_appetite,polyuria,family_history,mucoid_sputum,rusty_sputum,lack_of_concentration,visual_disturbances,receiving_blood_transfusion,receiving_unsterile_injections,coma,stomach_bleeding,distention_of_abdomen,history_of_alcohol_consumption,fluid_overload,blood_in_sputum,prominent_veins_on_calf,palpitations,painful_walking,pus_filled_pimples,blackheads,scurring,skin_peeling,silver_like_dusting,small_dents_in_nails,inflammatory_nails,blister,red_sore_around_nose,yellow_crust_ooze. The outcome variable is 'prognosis'. The diseases are 'Fungal infection', 'Allergy', 'GERD', 'Chronic cholestasis', 'Drug Reaction', 'Peptic ulcer diseae', 'AIDS', 'Diabetes ', 'Gastroenteritis', 'Bronchial Asthma', 'Hypertension ', 'Migraine', 'Cervical spondylosis', 'Paralysis (brain hemorrhage)', 'Jaundice', 'Malaria', 'Chicken pox', 'Dengue', 'Typhoid', 'hepatitis A', 'Hepatitis B', 'Hepatitis C', 'Hepatitis D', 'Hepatitis E', 'Alcoholic hepatitis', 'Tuberculosis', 'Common Cold', 'Pneumonia', 'Dimorphic hemmorhoids(piles)', 'Heart attack', 'Varicose veins', 'Hypothyroidism', 'Hyperthyroidism', 'Hypoglycemia', 'Osteoarthristis', 'Arthritis', '(vertigo) Paroymsal Positional Vertigo', 'Acne', 'Urinary tract infection', 'Psoriasis', 'Impetigo'

3. DATA PREPROCESSING

Disease Prediction dataset doesn't need preprocessing but 'Disease-Symptom Knowledge Database' need preprocessing in order to make it easy to visualize.

Data Cleaning

- **Missing Values**: The Disease and Count of nDisease Occurrence columns had missing values. ffill function was used.
- Remove UMLS codes for diseases and symptoms.
- After doing some processing get the cleaned data that is generated from raw data.

4. EXPLORATORY DATA ANALYSIS (EDA)

Plotly, matplotlib and seaborn was used in order to explore data and getting insights from it.

1-Among 42 diseases hypertensive disease has the largest number of occurrences (8.66%).

2- Among 42 diseases decubitus ulcer has the smallest number of occurrences (0.108%).

3- Subplots shows the symptoms responsible for the disease and also shows the count of these symptoms. For example the Gerd disease have symptoms stomach pain, acidity, ulcers on tongue, vomiting, cough and chest pain, among these symptoms stomach pain, cough and chest pain are major symptoms of Gerd disease.

4- Symptoms and associated sub plots shows the all those diseases that have this specific symptom and also shows how much is the importance of that symptom in the disease. For example migraine and gerd have symptom acidity and acidity is the major symptom of migraine disease.

5- Count of each symptom shows fatigue is the major symptom among 42 diseases.

5. FEATURE ENGINEERING

Created 3 feature Disease, Count of Disease Occurrence, Symptom by doing processing on 'Disease-Symptom Knowledge Database'.

6. MODEL BUILDING

Machine Learning Algorithms

- **Decision Trees**: Implemented using scikit-learn and from scratch.
- **Random Forest**: Implemented using scikit-learn and from scratch.

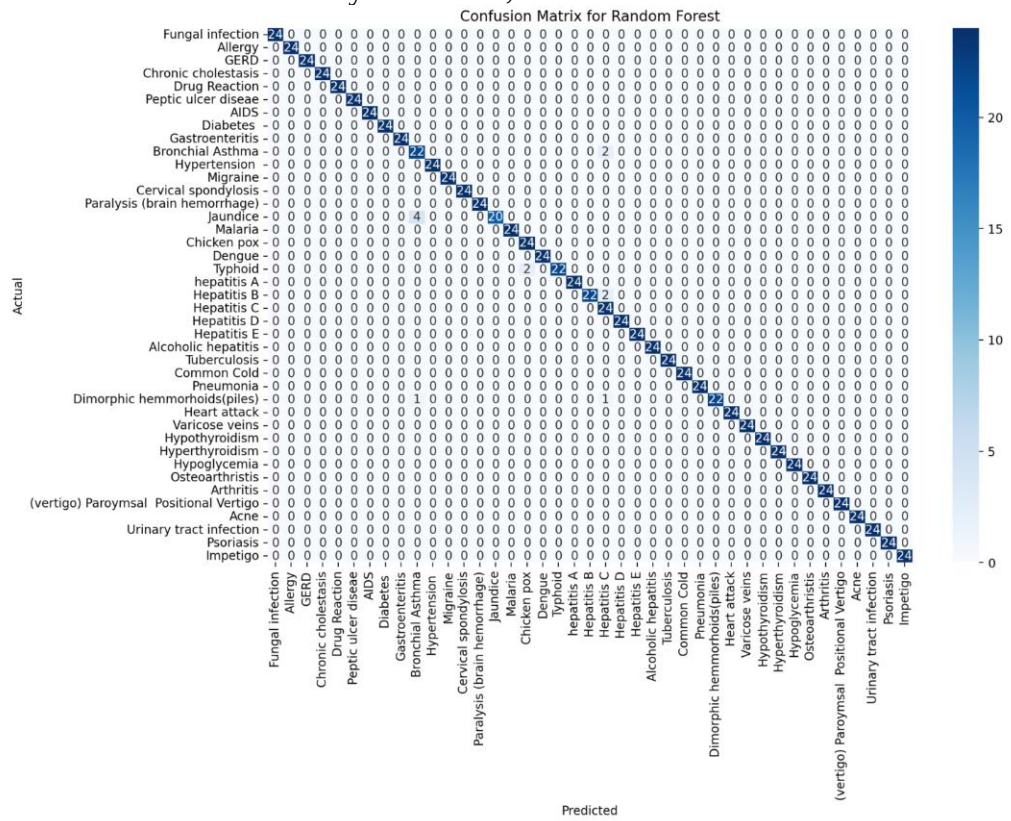
7. MODEL EVALUATION

Evaluation Metrics

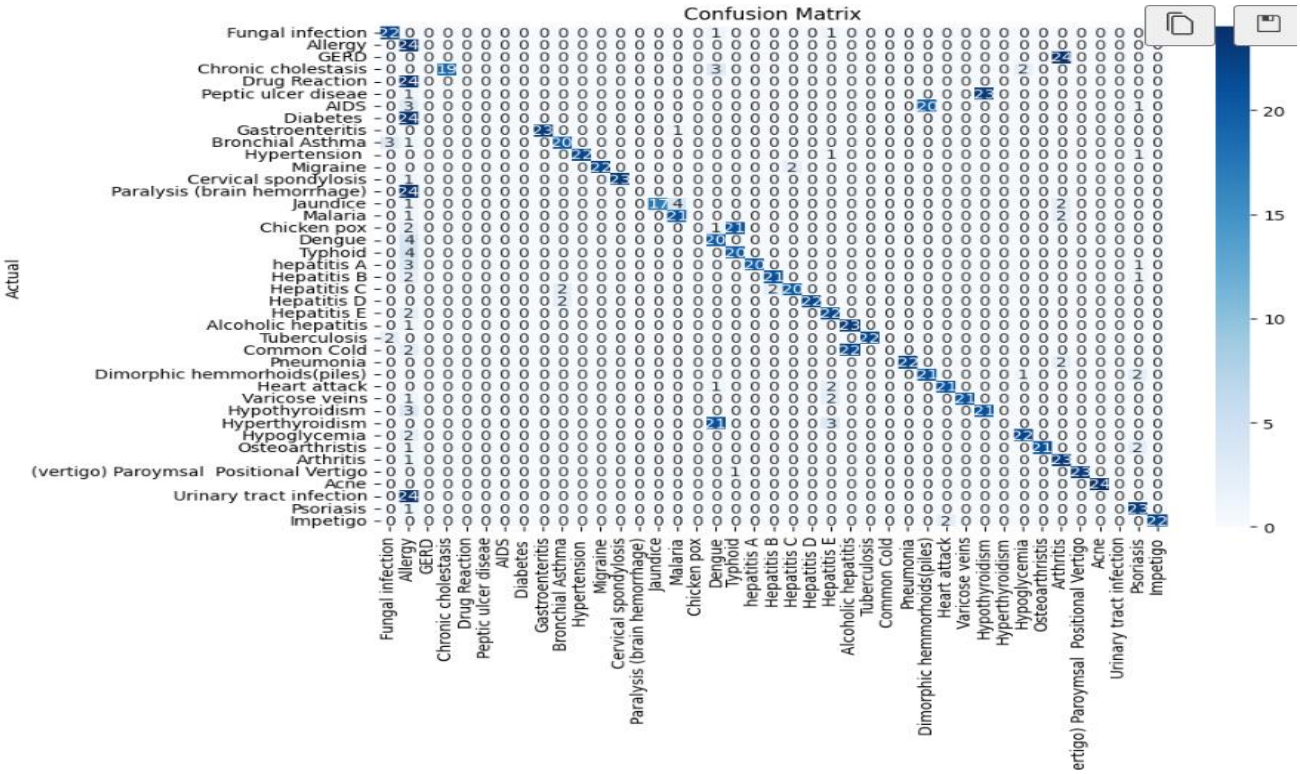
- Accuracy: Measured the overall accuracy of each model.
- Confusion Matrix: Evaluated the performance by understanding the true positives, false positives, true negatives, and false negatives.
- Classification Report: Provided precision, recall, and F1-score for each class.

Model Comparison

- **Random Forest**: Achieved an accuracy of 97.62%,



- **Decision Tree**: Achieved an accuracy of 67.78%



Confusion Matrix for Decision Tree

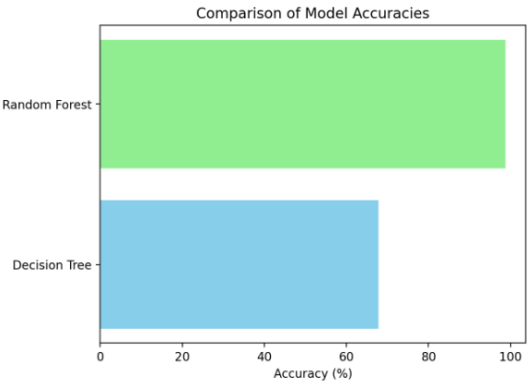
- **CLASSIFICATION REPORT**

Classification Report:

	precision	recall	f1-score	support
Fungal infection	0.81	0.92	0.86	24
Allergy	0.15	1.00	0.27	24
GERD	0.00	0.00	0.00	24
Chronic cholestasis	1.00	0.79	0.88	24
Drug Reaction	0.00	0.00	0.00	24
Peptic ulcer diseae	0.00	0.00	0.00	24
AIDS	0.00	0.00	0.00	24
Diabetes	0.00	0.00	0.00	24
Gastroenteritis	1.00	0.96	0.98	24
Bronchial Asthma	0.83	0.83	0.83	24
Hypertension	1.00	0.92	0.96	24
Migraine	1.00	0.92	0.96	24
Cervical spondylosis	1.00	0.96	0.98	24
Paralysis (brain hemorrhage)	0.00	0.00	0.00	24
Jaundice	1.00	0.71	0.83	24
Malaria	0.81	0.88	0.84	24
Chicken pox	0.00	0.00	0.00	24
Dengue	0.43	0.83	0.56	24
Typhoid	0.48	0.83	0.61	24
hepatitis A	1.00	0.83	0.91	24
Hepatitis B	0.91	0.88	0.89	24
Hepatitis C	0.91	0.83	0.87	24
Hepatitis D	1.00	0.92	0.96	24
Hepatitis E	0.71	0.92	0.80	24
Alcoholic hepatitis	0.51	0.96	0.67	24
Tuberculosis	1.00	0.92	0.96	24
Common Cold	0.00	0.00	0.00	24
Pneumonia	1.00	0.92	0.96	24
Dimorphic hemmorhoids(piles)	0.51	0.88	0.65	24
Heart attack	0.91	0.88	0.89	24
Varicose veins	1.00	0.88	0.93	24
Hypothyroidism	0.48	0.88	0.62	24
Hyperthyroidism	0.00	0.00	0.00	24
Hypoglycemia	0.88	0.92	0.90	24
Osteoarthristis	1.00	0.88	0.93	24
Arthritis	0.43	0.96	0.60	24
(vertigo) Paroymsal Positional Vertigo	1.00	0.96	0.98	24
Acne	1.00	1.00	1.00	24
Urinary tract infection	0.00	0.00	0.00	24
Psoriasis	0.74	0.96	0.84	24
Impetigo	1.00	0.92	0.96	24
accuracy			0.68	984
macro avg	0.62	0.68	0.63	984
weighted avg	0.62	0.68	0.63	984

Classification Report on Test Data for Random Forest:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	1.00	1.00	1.00	1
2	1.00	1.00	1.00	1
3	1.00	1.00	1.00	1
4	1.00	1.00	1.00	1
5	1.00	1.00	1.00	1
6	1.00	1.00	1.00	1
7	1.00	1.00	1.00	1
8	0.50	1.00	0.67	1
9	1.00	1.00	1.00	1
10	1.00	1.00	1.00	1
11	1.00	1.00	1.00	1
12	1.00	1.00	1.00	1
13	1.00	1.00	1.00	1
14	1.00	1.00	1.00	1
15	1.00	0.50	0.67	2
16	1.00	1.00	1.00	1
17	1.00	1.00	1.00	1
18	1.00	1.00	1.00	1
19	1.00	1.00	1.00	1
20	1.00	1.00	1.00	1
21	1.00	1.00	1.00	1
...				
accuracy			0.98	42
macro avg	0.99	0.99	0.98	42
weighted avg	0.99	0.98	0.98	42



8. CONCLUSION

For implementing disease prediction model, use random forest because it has high accuracy 98%. The insights come out from EDA are hypertensive disease has the largest number of occurrences, decubitus ulcer has the smallest number of occurrences and fatigue is the major symptom among 42 diseases.

9. LIMITATIONS AND FUTURE WORK

Limitations:

1. Dataset Size: Limited to 4920 records and 42 diseases, which may not capture full variability.
2. Symptom Overlap: Common symptoms among diseases can lead to misclassification.
3. Symptom Details: Severity and duration of symptoms are not considered.
4. Static Data: The model does not adapt to new data over time.
5. Data Quality: Potential remaining noise or inaccuracies in the dataset.

Future Work:

1. Larger Datasets: Incorporate more comprehensive datasets with more records and diseases.
2. Temporal Analysis: Include symptom progression and duration.
3. Real-time Integration: Enable real-time data updates.
4. Advanced Features: Explore sophisticated feature engineering techniques.
5. Model Interpretability: Improve model explanations for healthcare professionals.
6. Multi-modal Data: Integrate genetic, imaging, and patient history data.