Project Report

On

Disease Symptoms and Patient Profile

Submitted in partial fulfilment of the requirements for the award of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

(Artificial Intelligence & Machine Learning)

by

Ms.D.Meghana (22WH1A6603)

Ms. K.Vijaya Rajasree (22WH1A6648)

Ms. B.Anusha(22WH1A6655)

Ms. K.Pavani Reddy (22WH1A6664)

Under the esteemed guidance of

Ms. A Naga Kalyani

Assistant Professor, CSE(AI&ML)

Department of Computer Science & Engineering

(Artificial Intelligence & Machine Learning)

BVRIT HYDERABAD COLLEGE OF ENGINEERING FOR WOMEN

(Approved by AICTE, New Delhi and Affiliated to JNTUH, Hyderabad)

Accredited by NBA and NAAC with A Grade

Bachupally, Hyderabad – 500090

2023-24

Department of Computer Science & Engineering

(Artificial Intelligence & Machine Learning)

BVRIT HYDERABAD COLLEGE OF ENGINEERING FOR WOMEN

(Approved by AICTE, New Delhi and Affiliated to JNTUH, Hyderabad)

Accredited by NBA and NAAC with A Grade

Bachupally, Hyderabad – 500090

2023-24

CERTIFICATE

This is to certify that the major project entitled **"Disease Symptoms and Patient Profile"** is a bonafide work carried out by **Ms. D.Meghana(22WH1A6603), Ms. K.Vijaya Rajasree (22WH1A6648)**, **Ms. B.Anusha (22WH1A6655), Ms. K.Pavani Reddy (22WH1A6664)** in partial fulfilment for the award of B. Tech degree in **Computer Science & Engineering (AI&ML)**, **BVRIT HYDERABAD College of Engineering for Women, Bachupally, Hyderabad**, affiliated to Jawaharlal Nehru Technological University Hyderabad, Hyderabad under my guidance and supervision. The results embodied in the project work have not been submitted to any other University or Institute for the award of any degree or diploma.

Supervisor

Ms. A Naga Kalyani
Assistant Professor
Dept of CSE(AI&ML)

Head of the Department

Dr. B. Lakshmi Praveena
HOD & Professor
Dept of CSE(AI&ML)

External Examiner

DECLARATION

We hereby declare that the work presented in this project entitled **"Disease Symptoms and Patient Profile"** submitted towards completion of Project work in III Year of B.Tech of CSE(AI&ML) at **BVRIT HYDERABAD College of Engineering for Women,** Hyderabad is an authentic record of our original work carried out under the guidance of **Ms. A Naga Kalyani, Assistant Professor, Department of CSE(AI&ML).**

Sign with Date:
D.Meghana
(22WH1A6603)

Sign with Date:
K.Vijaya Rajasree
(22WH1A6648)

Sign with Date:
B.Anusha
(22WH1A6655)

Sign with Date:
K.Pavani Reddy
(22WH1A6664)

ACKNOWLEDGEMENT

We would like to express our sincere thanks to **Dr. K. V. N. Sunitha, Principal, BVRIT HYDERABAD College of Engineering for Women**, for her support by providing the working facilities in the college.

Our sincere thanks and gratitude to **Dr. B. Lakshmi Praveena, Head of the Department, Department of CSE(AI&ML), BVRIT HYDERABAD College of Engineering for Women,** for all timely support and valuable suggestions during the period of our project.

We are extremely thankful to our Internal Guide, **Ms. A Naga Kalyani , Assistant Professor, CSE(AI&ML), BVRIT HYDERABAD College of Engineering for Women,** for her constant guidance and encouragement throughout the project.

Finally, we would like to thank our Major Project Coordinator, all Faculty and Staff of CSE(AI&ML) department who helped us directly or indirectly. Last but not least, we wish to acknowledge our **Parents** and **Friends** for giving moral strength and constant encouragement.

D.Meghana (22WH1A6603)

K.Rajasree(22WH1A6648)

B. Anusha(22WH1A6655)

K.Pavani Reddy(22WH1A6664)

ABSTRACT:

This project conducts an exploratory data analysis (EDA) on a dataset containing patient profiles and their associated symptoms to better understand the relationships between different diseases, symptoms, and patient characteristics. The dataset includes information on various diseases, their symptoms, and demographic details of patients, such as age and gender. The project involves a thorough analysis of the data through visualizations, statistical summaries, and pattern recognition to uncover insights about symptom distribution, disease prevalence, and potential correlations between patient attributes and health conditions. Key techniques such as data cleaning, missing value imputation, feature encoding, and descriptive statistics are employed to prepare the data for deeper insights. The findings from this analysis aim to provide useful perspectives for healthcare professionals in identifying trends, improving diagnostic accuracy, and understanding patient symptomatology.

Problem Statement:

Healthcare professionals often face challenges in understanding complex relationships between diseases, symptoms, and patient demographics, which can hinder accurate diagnosis and effective treatment planning. With the increasing availability of patient data, there is a need for systematic analysis to uncover hidden patterns, trends, and correlations within the data. However, raw datasets frequently contain issues such as missing values, unstructured information, and a lack of clear visual or statistical summaries, making it difficult to extract meaningful insights.

The problem lies in developing a comprehensive approach to explore and analyze patient datasets, focusing on relationships between diseases, symptoms, and demographic characteristics. This requires employing advanced techniques such as data cleaning, imputation, and feature encoding to prepare the data, along with exploratory data analysis methods to visualize patterns and derive actionable insights. The solution must aim to assist healthcare professionals in improving diagnostic accuracy, understanding symptomatology, and identifying key trends to support informed decision-making in medical practice.

The Data Set that is used :

"C:\Users\cse\Downloads\Disease_symptom_and_patient_profile_dataset.csv"

df.head(20)

OUTPUT:

	Disease	Fever	Cough	Fatigue	Difficulty_Breathing	Age	Gender	Blood_Pressure	Cholesterol_Level	Outcome_Variable	Age_Group
0	Influenza	0	0	0	0	0	Female	Low	Normal	Positive	kid
1	Common Cold	0	0	0	0	0	Female	Normal	Normal	Negative	kid
2	Eczema	0	0	0	0	0	Female	Normal	Normal	Negative	kid
3	Asthma	0	0	0	0	0	Male	Normal	Normal	Positive	kid
4	Asthma	0	0	0	0	0	Male	Normal	Normal	Positive	kid
5	Eczema	0	0	0	0	0	Female	Normal	Normal	Positive	kid
6	Influenza	0	0	0	0	0	Female	Normal	Normal	Positive	kid
7	Influenza	0	0	0	0	0	Female	Normal	Normal	Positive	kid
8	Hyperthyroidism	0	0	0	0	0	Female	Normal	Normal	Negative	kid
9	Hyperthyroidism	0	0	0	0	0	Female	Normal	Normal	Negative	kid
10	Asthma	0	0	0	0	0	Male	High	Normal	Positive	kid
11	Allergic Rhinitis	0	0	0	0	0	Female	Normal	Low	Negative	kid
12	Anxiety Disorders	0	0	0	0	0	Female	Normal	High	Negative	kid
13	Common Cold	0	0	0	0	0	Female	Low	Normal	Negative	kid
14	Diabetes	0	0	0	0	0	Male	Low	Normal	Negative	kid
15	Gastroenteritis	0	0	0	0	0	Female	Normal	Normal	Negative	kid
16	Pancreatitis	0	0	0	0	0	Female	High	Normal	Negative	kid
17	Rheumatoid Arthritis	0	0	0	0	0	Female	High	High	Negative	kid
18	Depression	0	0	0	0	0	Male	High	Normal	Positive	kid
19	Liver Cancer	0	0	0	0	0	Female	Normal	Normal	Positive	kid

Code :

```
import pandas as pd
try:
  df = pd.read_csv('Disease_symptom_and_patient_profile_dataset.csv')
  print("Shape of the DataFrame:", df.shape)
except FileNotFoundError:
  print("Error: 'Disease_symptom_and_patient_profile_dataset.csv' not found. Please ensure the file exists in the current directory or provide the correct path.")
except pd.errors.ParserError:
   print("Error: Unable to parse the CSV file. Please check the file format.")
except Exception as e:
   print(f"An unexpected error occurred: {e}")
```

OUTPUT:

```
Shape of the DataFrame: (349, 10)
```

```
def rename_col(s):
   s = s.replace(' ', '_')
   s = s.replace('-', '_')
   a = s.split('_')
   for i in range(len(a)):
      if not all([c == c.upper() for c in a[i]]):
         a[i] = a[i].capitalize()
   return '_'.join(a)

df.columns = [rename_col(c) for c in df.columns]
df.columns
```

OUTPUT:

```
Index(['Disease', 'Fever', 'Cough', 'Fatigue', 'Difficulty_Breathing', 'Age',
       'Gender', 'Blood_Pressure', 'Cholesterol_Level', 'Outcome_Variable'],
      dtype='object')
```

df.head()

OUTPUT:

Disease	Fever	Cough	Fatigue	Difficulty_Breathing	Age	Gender	Blood_Pressure	Cholesterol_Level	Outcome_Variable
Influenza	Yes	No	Yes	Yes	19	Female	Low	Normal	Positive
Common Cold	No	Yes	Yes	No	25	Female	Normal	Normal	Negative
Eczema	No	Yes	Yes	No	25	Female	Normal	Normal	Negative
Asthma	Yes	Yes	No	Yes	25	Male	Normal	Normal	Positive
Asthma	Yes	Yes	No	Yes	25	Male	Normal	Normal	Positive

What are the most common symptoms for each disease?

```
cols = ['Fever', 'Cough', 'Fatigue', 'Difficulty_Breathing']
for c in cols:
    df[c] = df[c].apply(lambda x : 1 if x == 'Yes' else 0)
p = df[df.Outcome_Variable == 'Positive'].copy()
(
    p[p.Disease.isin(p.Disease.value_counts().nlargest(10).index)]
    .groupby('Disease')[cols]
    .sum()
    .style
    .highlight_max(axis=1, color='red')
)
```

OUTPUT:

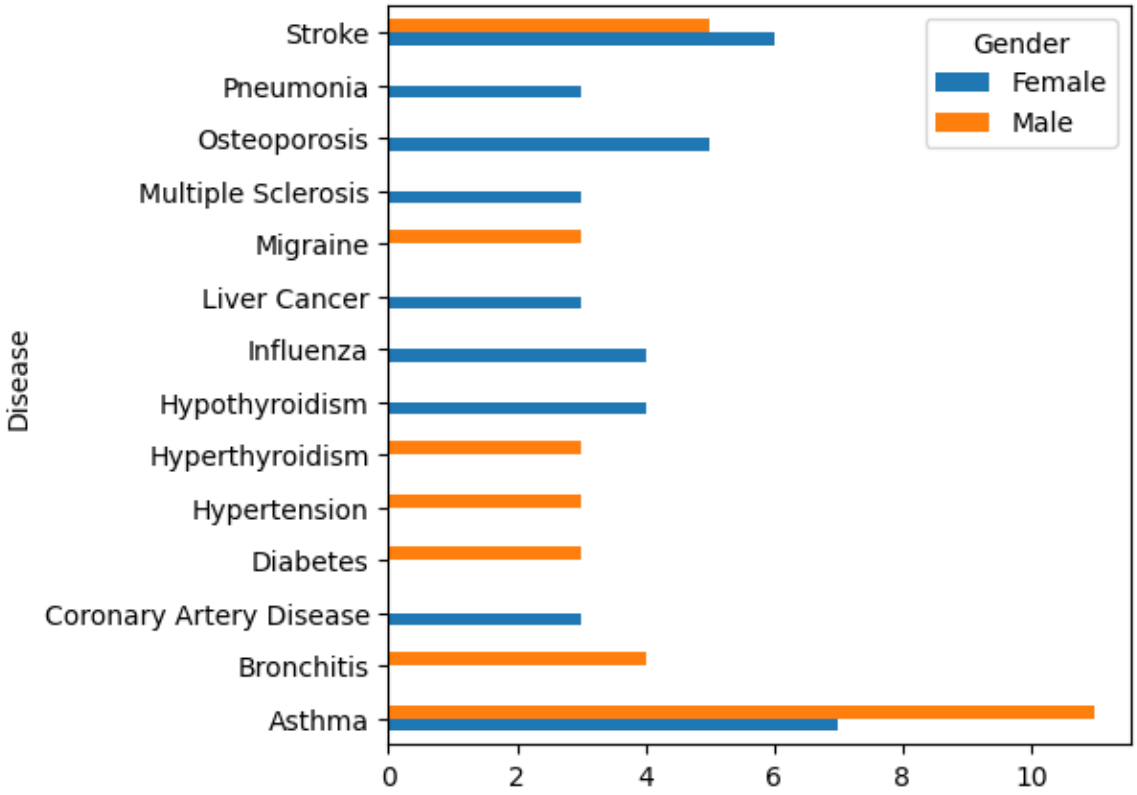
	Fever	Cough	Fatigue	Difficulty_Breathing
Disease				
Asthma	17	13	7	17
Bronchitis	5	4	4	5
Diabetes	4	3	5	1
Hypertension	3	0	4	0
Influenza	6	5	6	5
Liver Cancer	2	3	4	2
Migraine	3	2	2	0
Osteoporosis	5	3	7	1
Pneumonia	4	4	5	5
Stroke	8	5	9	2

How does the distribution of diseases vary with age and gender?

```
t = p.groupby(['Disease','Gender']).size()
t[t > 2].unstack().plot(kind='barh', figsize=(5,5))
```

OUTPUT:

```
<Axes: ylabel='Disease'>
```



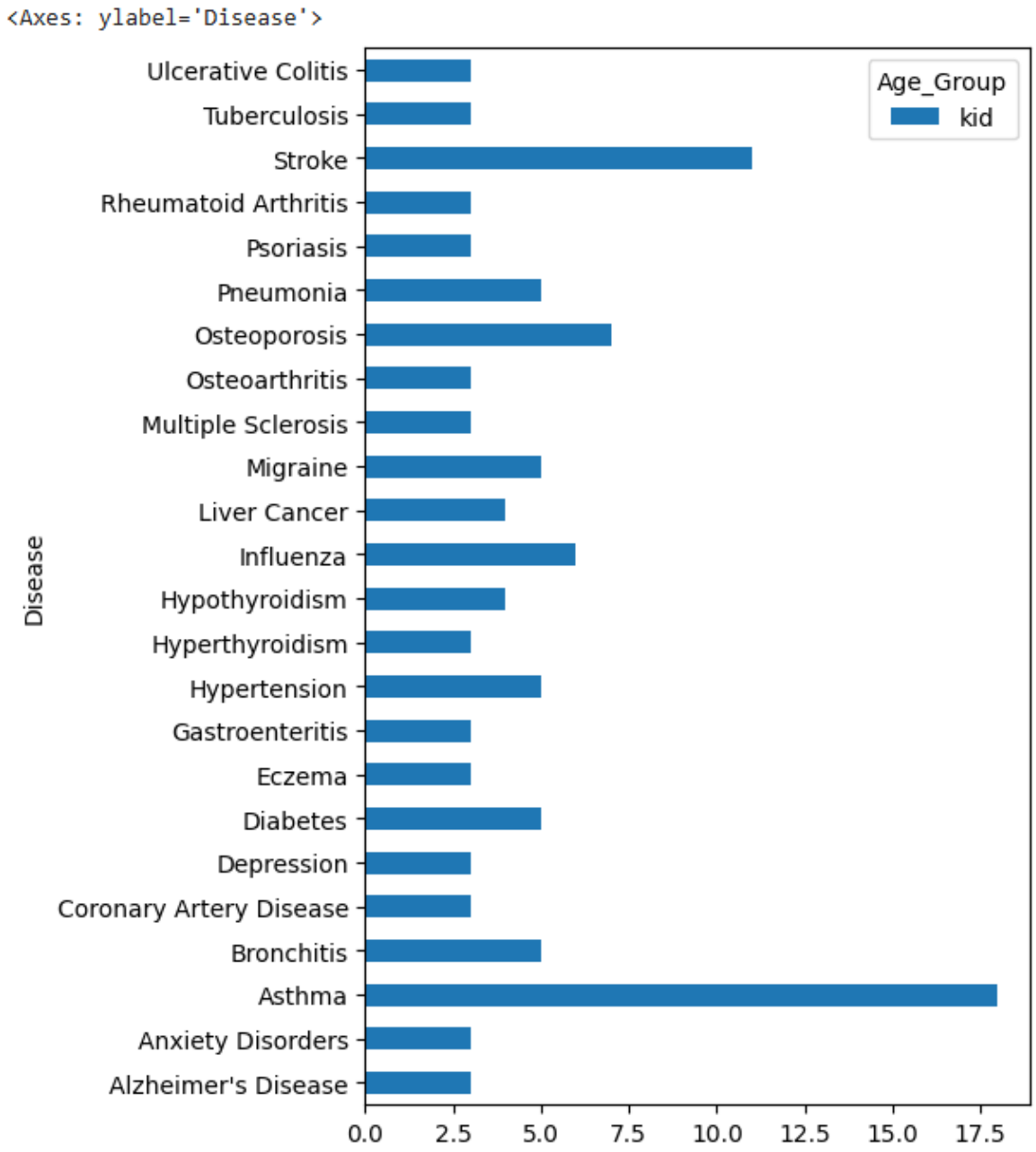

```
def age_group(x):
  if x < 13:
    return 'kid'
  elif x < 20:
    return 'teen'
  elif x <= 60:
    return 'adult'
  else:
    return 'senior'

df['Age_Group'] = df.Age.apply(age_group)
p = df[df.Outcome_Variable == 'Positive'].copy()
```

t = p.groupby(['Disease','Age_Group']).size()

t[t > 2].unstack().plot(kind='barh', figsize=(5,8))

OUTPUT:



Is there a relationship between blood pressure or cholesterol level and specific diseases?

p.Blood_Pressure.value_counts()

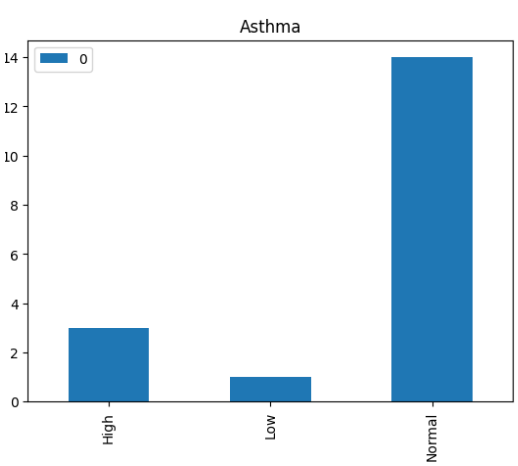
OUTPUT:

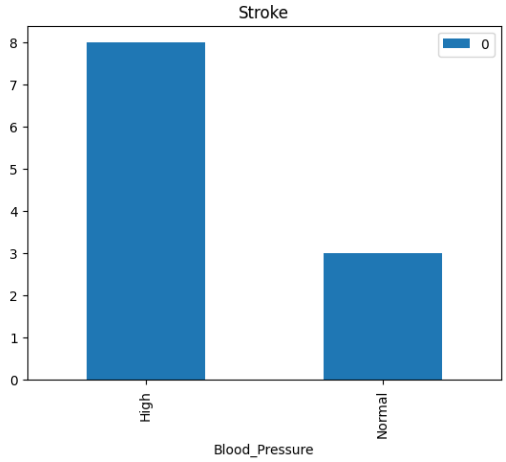
Blood_Pressure	count
High	104
Normal	78
Low	4

dtype: int64

```
for d in p.Disease.unique():
    s = p[p.Disease == d].groupby('Blood_Pressure').size()
    if (s >= 5).any():
        s.to_frame().plot(kind='bar', title=d)
```

OUTPUT:





p.Cholesterol_Level.value_counts()**OUTPUT:**

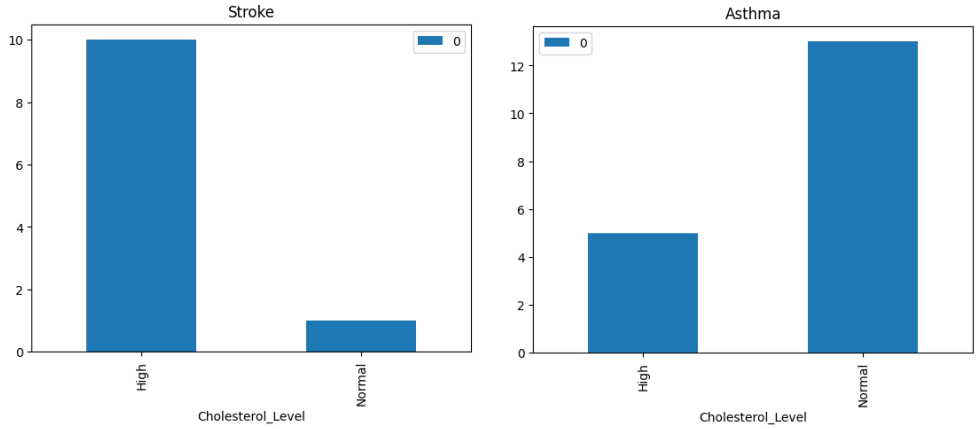
Cholesterol_Level	count
High	115
Normal	61
Low	10

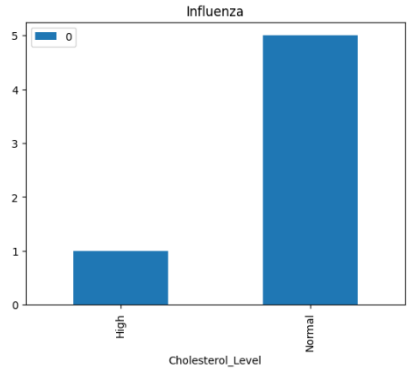
dtype: int64

```
for d in p.Disease.unique():
    s = p[p.Disease == d].groupby('Cholesterol_Level').size()
    if (s >= 5).any():
```

```
    s.to_frame().plot(kind='bar', title=d)
```

OUTPUT:





Which symptoms are the least useful in determining a disease?

```
cols = ['Fever', 'Cough', 'Fatigue', 'Difficulty_Breathing']

for c in cols:

    print(f'{c} is associated with {df[df[c] == 1].Disease.unique().size} diseases')
```

OUTPUT:

```
Fever is associated with 0 diseases
Cough is associated with 0 diseases
Fatigue is associated with 0 diseases
Difficulty_Breathing is associated with 0 diseases
```

GITHUB REPOSITORY LINK: