EXPERIMENT ASSESSMENT

ACADEMIC YEAR 2025-26

Course: AIH Lab

Course code: CSDOL 7012

Year: BE SEM: VII

Experiment No. 1
Title: To collect, Clean, Integrate and Transform Healthcare Data based on specific disease.
Name: Ayush Gupta
Roll Number: 11
Date of Performance:15/7/25
Date of Submission:21/7/25

Evaluation

Performance Indicator	Max. Marks	Marks Obtained
Performance	5	
Understanding	5	
Journal work and timely submission.	10	
Total	**20**	

Performance Indicator	Exceed Expectations (EE)	Meet Expectations (ME)	Below Expectations (BE)
Performance	5	3	2
Understanding	5	3	2
Journal work and timely submission.	10	8	4

Checked by

Name of Faculty : Mrs.Kranti Gule

Signature :

Date :

Aim: To collect, clean, integrate and transform Healthcare data

Objective: To develop a robust and efficient data pipeline that collects, cleans, integrates, and transforms healthcare data from diverse sources, ensuring data accuracy, privacy compliance, and usability. This pipeline will facilitate comprehensive and reliable analyses, enabling informed decision-making and insights to drive improvements in healthcare delivery, patient outcomes, and research endeavors.

Theory: The Disease Symptoms and Patient Profile Dataset serves as a pivotal gateway to unraveling the intricate web of diseases. By meticulously intertwining symptomatology, patient demographics, and health metrics, this dataset presents an unprecedented opportunity to discern the concealed correlations within medical conditions. With symptoms such as fever, cough, fatigue, and difficulty breathing interwoven alongside crucial variables like age, gender, blood pressure, and cholesterol levels, this dataset holds the promise of unearthing latent patterns. A transformative resource for medical researchers, healthcare professionals, and data enthusiasts alike, its exploration promises to unveil distinctive symptom profiles and initiate an enthralling expedition into the realm of ailments. As users navigate this treasure trove, a profound revolution in healthcare comprehension beckons, destined to reshape our understanding of medical intricacies and pave the way for enhanced diagnostics and treatment strategies.

At the heart of the Disease Symptoms and Patient Profile Dataset lies a reservoir of invaluable insights waiting to reshape the landscape of healthcare knowledge. This meticulously curated compilation of symptoms, patient characteristics, and health indicators offers an unprecedented vantage point into the complex interplay of factors underlying various diseases. As medical researchers delve into the depths of this dataset, they embark on a journey of discovery, unveiling hidden relationships that have the potential to redefine diagnostic paradigms and treatment approaches. By deciphering the intricate tapestry woven from fever, cough, fatigue, and difficulty breathing, intricately interwoven with age, gender, blood pressure, and cholesterol levels, a new era of personalized and targeted healthcare strategies is on the horizon. This dataset not only promises to revolutionize medical research but also empowers healthcare professionals to make informed decisions that can lead to improved patient outcomes and a brighter future for the field of medicine.

Program and output:

```
import pandas  as pd

df = pd.read_csv("diabetes.csv")

df.head()

print(df.info())
```

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 768 entries, 0 to 767

Data columns (total 9 columns):

#	Column	Non-Null Count	Dtype
0	Pregnancies	768 non-null	int64
1	Glucose	768 non-null	int64
2	BloodPressure	768 non-null	int64
3	SkinThickness	768 non-null	int64
4	Insulin	768 non-null	int64
5	BMI	768 non-null	float64
6	DiabetesPedigreeFunction	768 non-null	float64
7	Age	768 non-null	int64
8	Outcome	768 non-null	int64

dtypes: float64(2), int64(7)

memory usage: 54.1 KB

None

```
cols_with_invalid_zeros = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']

df[cols_with_invalid_zeros] = df[cols_with_invalid_zeros].replace(0, np.nan)

print(df.isnull().sum())
```

Pregnancies 0

Glucose 5

BloodPressure 35

SkinThickness 227

Insulin 374

BMI 11

DiabetesPedigreeFunction 0

Age 0

Outcome 0

dtype: int64

```
from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy='median')

df[cols_with_invalid_zeros] = imputer.fit_transform(df[cols_with_invalid_zeros])

print("\nAfter Imputation:")

print(df.isnull().sum())
```

After Imputation:

Pregnancies 0

Glucose 0

BloodPressure 0

SkinThickness 0

Insulin 0

BMI 0

DiabetesPedigreeFunction 0

Age 0

Outcome 0

dtype: int64

Conclusion- (Write in your own words)

Using Seaborn, we explored the diabetes dataset through three powerful visualizations that revealed insightful patterns. The time series plot illustrated how average glucose levels change over time, helping us identify trends that could signal worsening health conditions or shifts in patient behavior. The correlation heatmap highlighted strong positive relationships between glucose, BMI, and diabetes outcomes, confirming that these features are key indicators of diabetic risk. Lastly, the violin plot offered a deep look into the BMI distribution between diabetic and non-diabetic groups, showing that diabetics tend to have a higher and more variable BMI. Together, these plots provided a clearer, data-driven view of the factors associated with diabetes, supporting more informed healthcare decisions and deeper analysis.