

Lab -1

Data Processing

PAGE NO :

DATE : 5/3/25

Lab -1

Demonstrate various data pre-processing technique for a given dataset

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

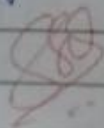
```
df = pd.read_csv('Diabetes.csv')
print(df.head())
```

```
from sklearn.preprocessing import
MinMaxScaler, StandardScaler, LabelEncoder
from sklearn.preprocessing import
SimpleImputer
```

```
df = pd.
```

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

```
num_column = df.select_dtypes(include
= ['float', 'int64']).column
imputer = SimpleImputer(strategy='mean')
df[num_column] = imputer.fit_transform
(df[num_column])
```



```
cat_col = df.select_dtypes(include=[object]).columns
```

```
imputer_cat = SimpleImputer(strategy='most_frequent')
df[cat_col] = imputer_cat.fit_transform(df[cat_columns])
```

(iii) handling outliers

```
Q1 = df[num_col].quantile(0.25)
```

```
Q3 = df[num_col].quantile(0.75)
```

```
IQR = Q3 - Q1
```

```
df_clean = df[(df[num_col] < (Q1 - 1.5 * IQR)) |
               (df[num_col] > (Q3 + 1.5 * IQR))]
```

```
df_scaled = pd.DataFrame(Scaler.fit_transform(df_clean[num_col]), columns=num_col)
```

```
df_final = pd.concat([df_clean[cat_col],
                      df_scaled], axis=1)
```

```
print("\ncleaned & scaled data:")
```

```
print(df_final.head())
```

① load .csv file into dataframe

```
df = pd.read_csv("housing.csv")
```

② display informat of all column

```
print(df.info())
```


- ③ To display Statistical info of all numerical
`print(df.describe())`
- ④ To display the count of unique labels for
 "Ocean proxy" column
`print(df["ocean proxy"].count())`
- ⑤ To display which attributes (col) in a dataset
 have missing values count greater than
 zero
`Miss = df.isnull().sum()`
`col_miss = miss[miss > 0]`
`print(col_miss)`

Diabetes

- ① which column in the Dataset had missing
 values? how did you handle them?
 None, but if had then num column's
NAN can be replaced with median &
 categorical col null can be replaced by
mode
- ② which categorical col did you identify in
 the dataset? how did you encode them?
 gender, class
 using ordinal Encoder we can encode
 categorical column. \rightarrow low-0, medium-1,
 High-2.
- ③ What is the difference b/w Min-Max Scaling
 and Standardizat? when would you
 use one over the other?

Min-Max scaling is also called normalization. It transforms data to fit within a specific range [0 to 1] by scaling based on min-max values in dataset.

Standardization scales data by subtracting mean and dividing.