5/3/25

Lab - 1

```
import pandas as pd
df = pd.read_csv("housing.csv")
```

(i)
```
df.info()
```

(ii)
```
df.describe()
```

(iii)
```
df["Ocean Proximity"].value_counts()
```

(iv)
```
mis_val = df.isnull().sum()
val = mis_val[mis_val>0]
print(val)
```

Diabetes

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder

df = pd.read_csv('/content/Dataset of Diabetes - CSV')

print(df.head())
#missing values
print(df.isnull().sum())

# Impute
nc = df.select_dtypes(include=['float64','int64']).
                                    columns

imputer = SimpleImputer(strategy='mean')
df[num_columns] = imputer.fit_transform(df[num_columns])
cat = df.select_dtypes(include=['object']).columns
imputer_cat = SimpleImputer(strategy='most_frequent')
df[cat] = imputer_cat.fit_transform(df[cat_columns])

Handling categorical data

label_encoder = LabelEncoder()
df['Gender'] = label_encoder.fit_transform(df['Gender'])
df['CLASS'] = label_encoder.fit_transform(df['CLASS'])

#Handling outliers
Q1 = df[num_columns].quantile(0.25)
Q3 = df[num_columns].quantile(0.75)
IQR = Q3 - Q1

df_clean = df[~((df[num_columns] < (Q1 - 1.5 * IQR)) |
        (df[num_columns] > (Q3 + 1.5 * IQR))).any
          (axis = 1)]

#Data Transformation
#Apply MinMax or standard scaler
scaler_choice = 'minmax'
if scaler_choice == 'minmax':
    scaler = MinMaxScaler()
else:
    scaler = StandardScaler()
df_scaled = pd.DataFrame(scaler.fit_transform(df_clean
        [num_columns]), columns = num_columns)
```

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

(1) Which columns in the dataset had missing values? How did you handle them?

(A) No columns had missing values.
We handled them by writing the code:
`df.isnull().sum()`

(2) Which categorical columns did you identify in the dataset? How did you encode them?

(A) Gender and Class
We encoded them by using label encoding

(3) What is the difference between MinMax Scaling & Standardization? When would you use one over the other?

(A) Min-Max Scaling :-
Transforms values to fixed range (usually 0 to 1) and useful if there is a fixed range for features bounded range, and also used when no outliers are present
Formula : $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$

Standardization (z-score normalization) :-
Transforms data to have zero mean and unit variance.
outliers exist.
Gaussian distribution, such as in regression generally used in the normal distribution
Formula :- $x' = \frac{x - \mu}{\sigma}$

26/3/25