Date :
Page No.

1) Data Preprocessing

- Load dataset using pandas

→
```
import pandas as pd
df = pd.read.csv
print (df.head())
```

- check for missing or duplicate values.

checking for missing values —
```
missing values = df.isnull().sum()
print ("Missing values in each column: ")
print (missing_values)
```

- Convert categorical columns using Label Encoding or OneHot Encoding.

converting categorical columns using label encoding —
```
from Sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
df ['category_column'] = label_encoder.
                    fit_transform (df ['category_column'])
print (df.head())
```

Date :
Page No.

2) Exploratory Data Analysis (EDA)

- Plot score distributions using histograms.

→
```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv
plt.figure (figsize = (10, 6))
plt.hist (df['score_column'], bins = 20,
    color = 'blue', alpha = 0.7, edgecolor = 'black')
plt.title ('Score Distribution')
plt.xlabel ('scores')
plt.ylabel ('Frequency')
plt.grid (axis = 'y', alpha = 0.75)
plt.show ()
```

- Compare average scores by gender, lunch type and test preparation status.

```
import pandas as pd
df = pd.read.$csv
average_scores_gender = df.groupby ('gender')
    ['math score', 'reading sorce', 'writing
     score'].mean()

average_scores_lunch = df.groupby ('lunch')
    ['math score', 'reading score', 'writing
     score'].mean()

average_scores_test_prep = df.groupby ('test
    preparation course') ['math score', 'reading
    score', writing score'].mean()
```

Date :
Page No.

```
print ("Average scores by Gender: ")
print (average_scores_gender)

print ("\n Average scores by Lunch Type: ")
print (average_scores_lunch)

print ("\n Average scores by Test Preparation
      Status: ")
print (average_scores_test_prep)
```

- Create a correlation heatmap.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
df = pd.read_csv ('student_performance.csv')

df_encoded = pd.get_dummies (df, columns =
        ['gender', 'lunch', 'test preparation course'],
         drop_first = True)

numeric_cols = ['math score', 'reading score',
        'writing score', 'parental level of
         education']
df_numeric = df_encoded.select_dtypes
        (include = ['int64', 'float64'])

corr_matrix = df_numeric.corr()

plt.figure (figsize = (12,8))
sns.heatmap (corr_matrix,
             annot = True,
```

Date :
Page No.

```
    cmap = 'coolwarm',
    center = 0,
    fmt = '.2f',
    linewidths = 0.5,
    annot_kws = {'size': 10})

plt.title('Student Performance Correlation
          Heatmap', pad=20, fontsize=16)
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.tight_layout()
plt.show()
```

3) Feature Engineering

- Create a new column average ~~sor~~ score.

```
import pandas as pd
df = pd.read_csv
df['average_scorce'] = df[['math score',
        'reading score', 'writing score']].mean(axis=1)
print(df.head())
```

- Convert result to binary labels for machine learning (1 = Pass, 0 = Fail)

```
import pandas as pd
import numpy as np
df = pd.read_csv

df['result_binary'] = np.where(df['average
        _scorce'] >= 50, 1, 0)
print(df[['average_score', 'result_binary']]
        .head())
```