Indian Institute of Technology Jodhpur

MLOps

Assignment 2

ML Ops

Submitted by: Jojo Joseph (G23AI2100)

Google Colab :
https://colab.research.google.com/drive/12_uCB1524aOU9-Bp2GtblleoMkBkHAqL?usp=sharing

gitHub : https://github.com/IITJPGD/MLOpsAssignment2

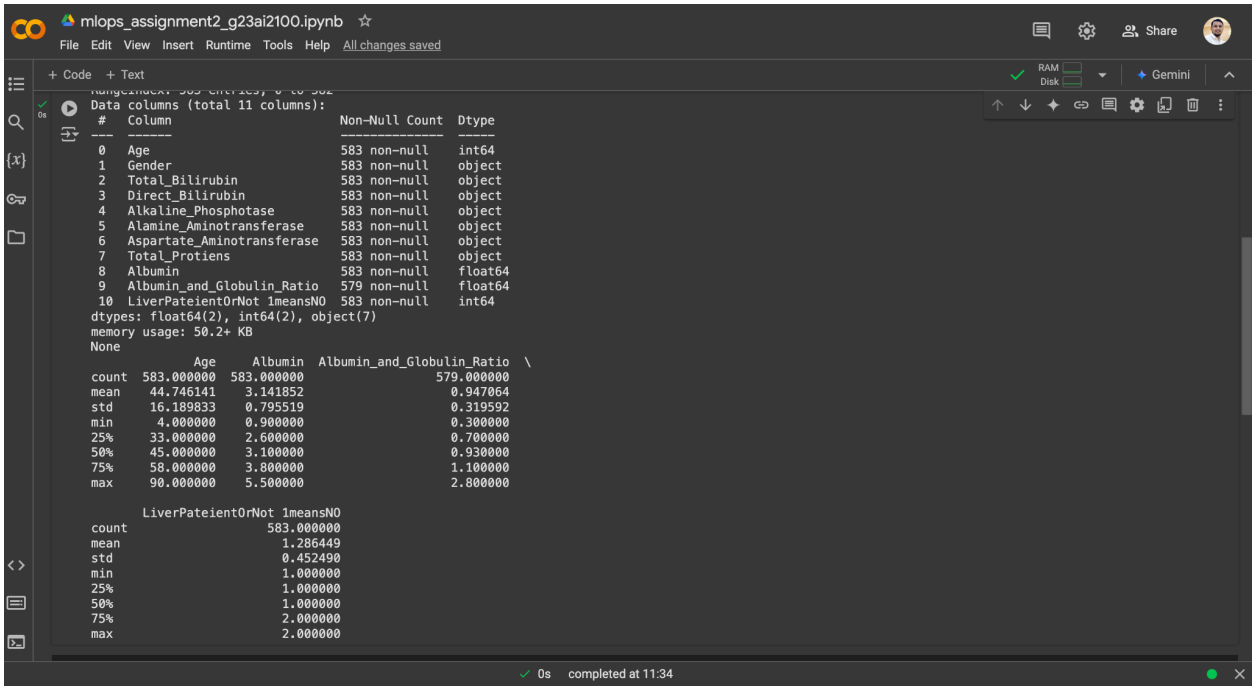
MLOps Assignment

Q1 Use the attached dataset which is a classification problem with Indianliver pateints whether liver patent or not as a output variable

Ans

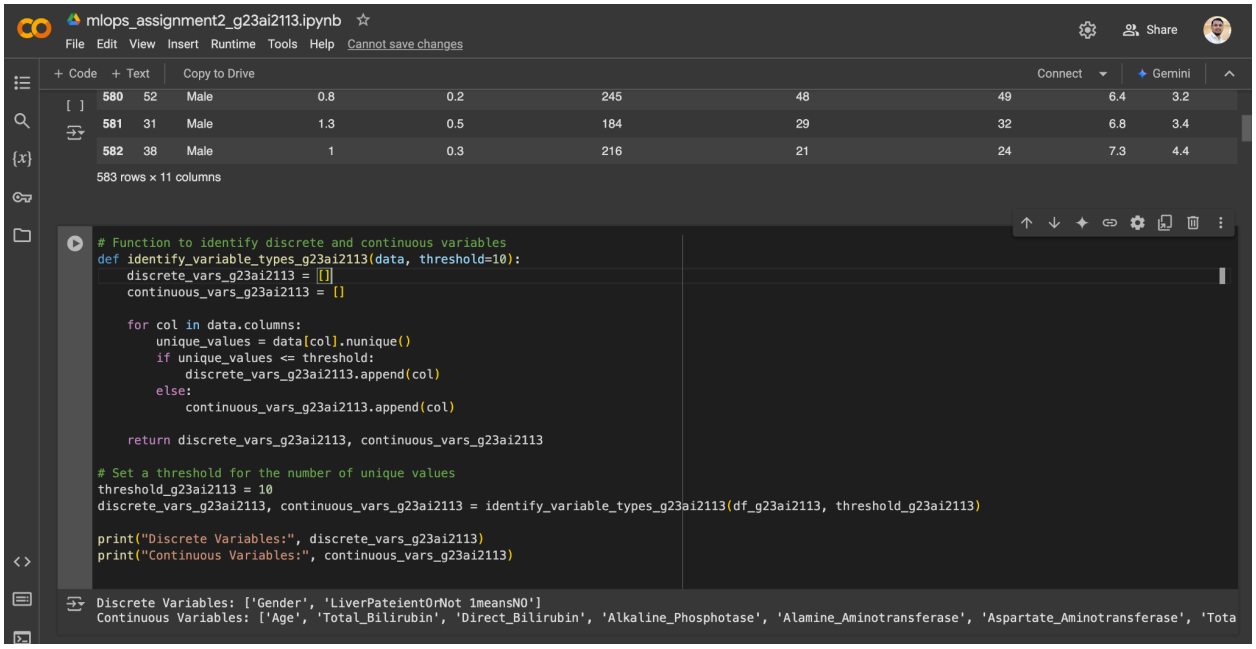
```
import pandas as pd
url = "https://raw.githubusercontent.com/IITJPGD/MLOpsAssignment2/main/indian_liver_patientmissing.csv"
df = pd.read_csv(url)
print(df.head())
print(df.info())
print(df.describe())
```

```
   Age  Gender Total_Bilirubin Direct_Bilirubin Alkaline_Phosphotase  \
0   65  Female             0.7              0.1                  187
1   62    Male            10.9              5.5                  699
2   62    Male             7.3              4.1                  490
3   58    Male               1              0.4                  182
4   72    Male             3.9                2                  195

  Alamine_Aminotransferase Aspartate_Aminotransferase Total_Protiens  Albumin  \
0                       16                         18            6.8      3.3
1                       64                        100            7.5      3.2
2                       60                         68              7      3.3
3                       14                         20            6.8      3.4
4                       27                         59            7.3      2.4

   Albumin_and_Globulin_Ratio  LiverPateientOrNot 1meansNO
0                        0.90                           1
1                        0.74                           1
2                        0.89                           1
3                        1.00                           1
4                        0.40                           1
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 583 entries, 0 to 582
Data columns (total 11 columns):
 #   Column                      Non-Null Count  Dtype
```



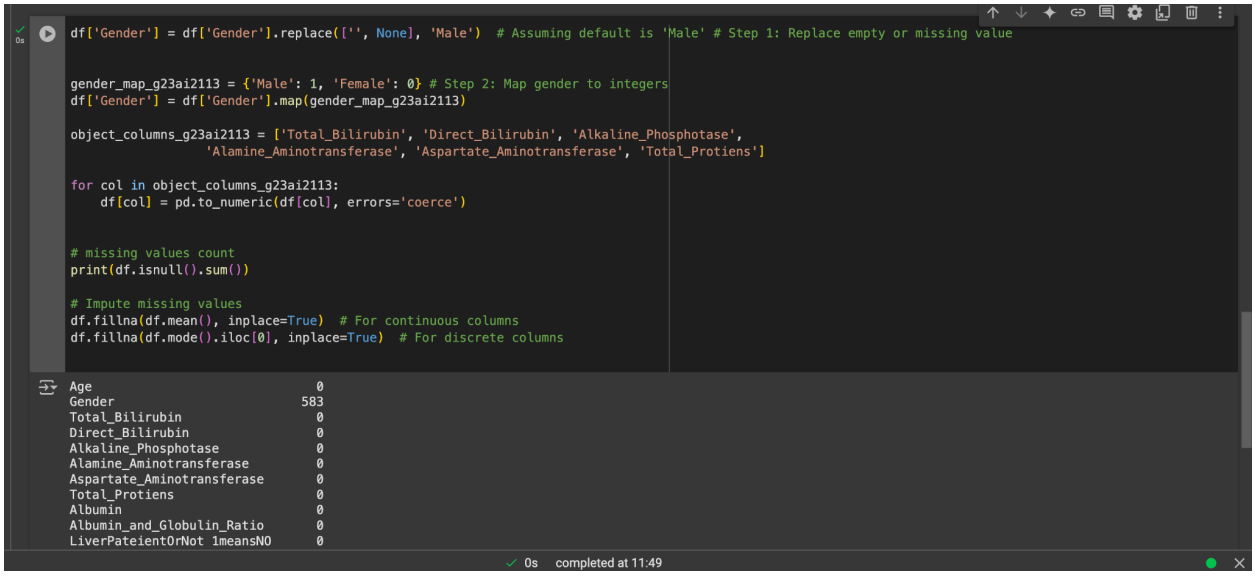
Q2 Do descriptive analytics of all columns by first finding discrete and continous columns

Ans



Q3 Do imputation of relevant missing values

Ans



Q4 Do exploratory analysis by different visualizations

Ans

mlops_assignment2_g23ai2100.ipynb

File Edit View Insert Runtime Tools Help

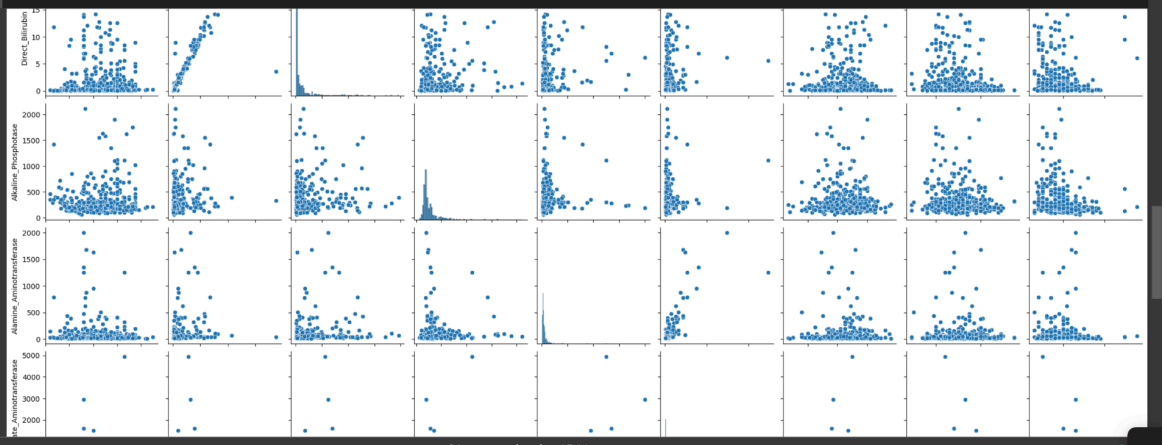
Share

+ Code + Text

RAM Disk

Gemini

```
import matplotlib.pyplot as plt
import seaborn as sns

# Pairplot for continuous variables
sns.pairplot(df[continuous_vars_g23ai2100])
plt.show()
```



26s completed at 15:11

mlops_assignment2_g23ai2100.ipynb

File Edit View Insert Runtime Tools Help All changes saved

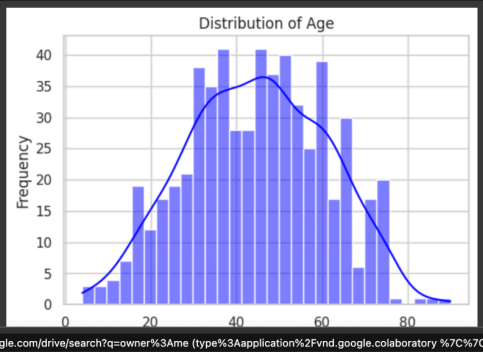
Share

+ Code + Text

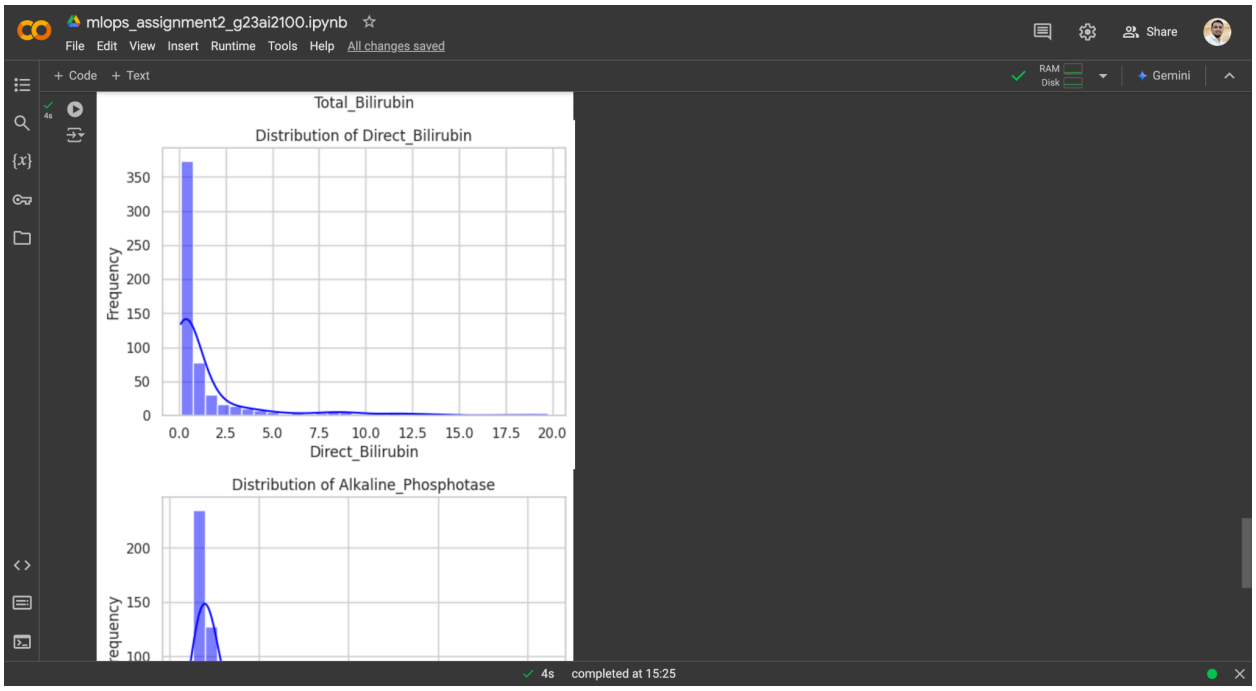
RAM Disk

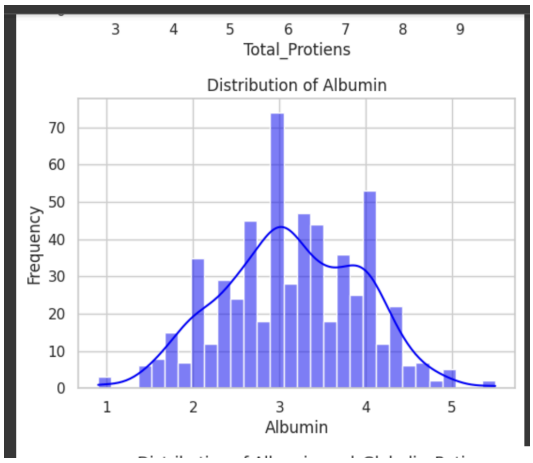
Gemini

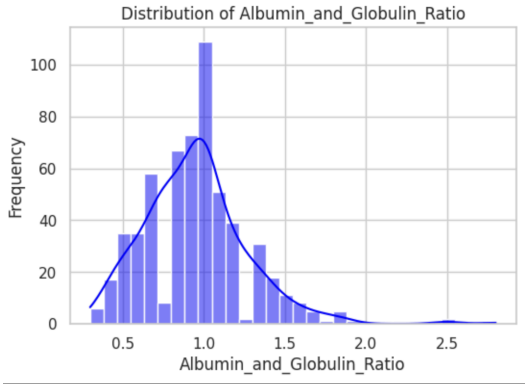
```
[7]
    # Set visualization style
    sns.set(style="whitegrid")
    #  Visualize the distribution of continuous variables

    for col in continuous_vars_g23ai2100:
      plt.figure(figsize=(6, 4))
      sns.histplot(df[col], kde=True, bins=30, color='blue')
      plt.title(f"Distribution of {col}")
      plt.xlabel(col)
      plt.ylabel("Frequency")
      plt.show()
```

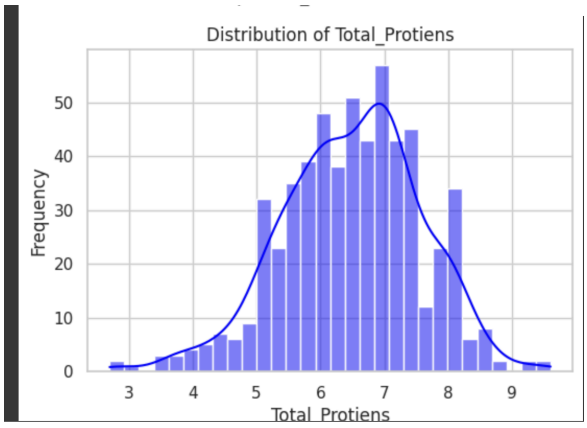


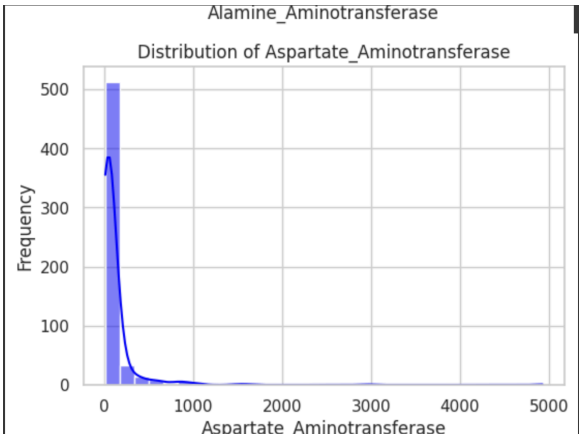
https://drive.google.com/drive/search?q=owner%3Ame (type%3Aapplication%2Fvnd.google.colaboratory %7C%7C type%3Aapplication%2Fvnd.google.colab)&authuser=3


mlops_assignment2_g23ai2100.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Share
+ Code + Text
RAM
Disk
Gemini
Total_Bilirubin
Distribution of Direct_Bilirubin
Frequency
Direct_Bilirubin
Distribution of Alkaline_Phosphotase
4s completed at 15:25

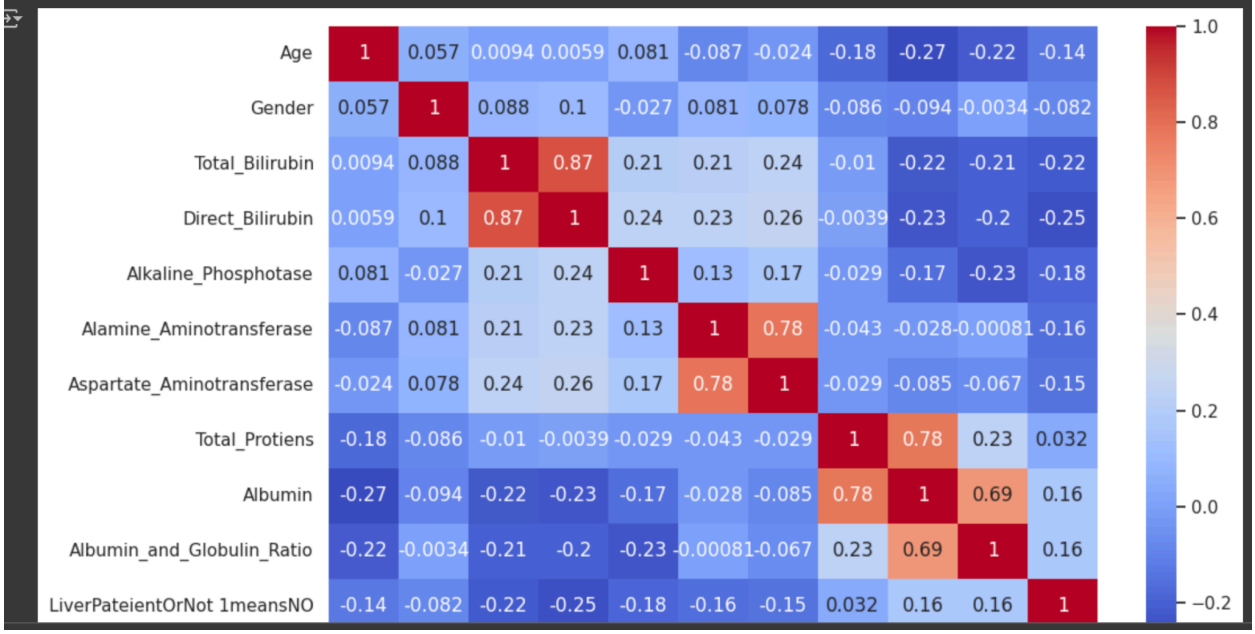

Total_Protiens
Distribution of Albumin
Frequency
Albumin

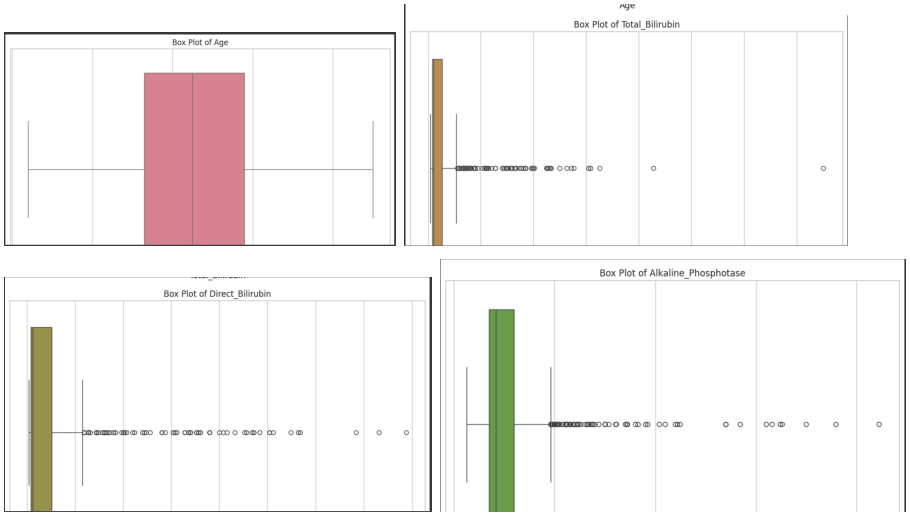

Distribution of Albumin_and_Globulin_Ratio
Frequency
Albumin_and_Globulin_Ratio

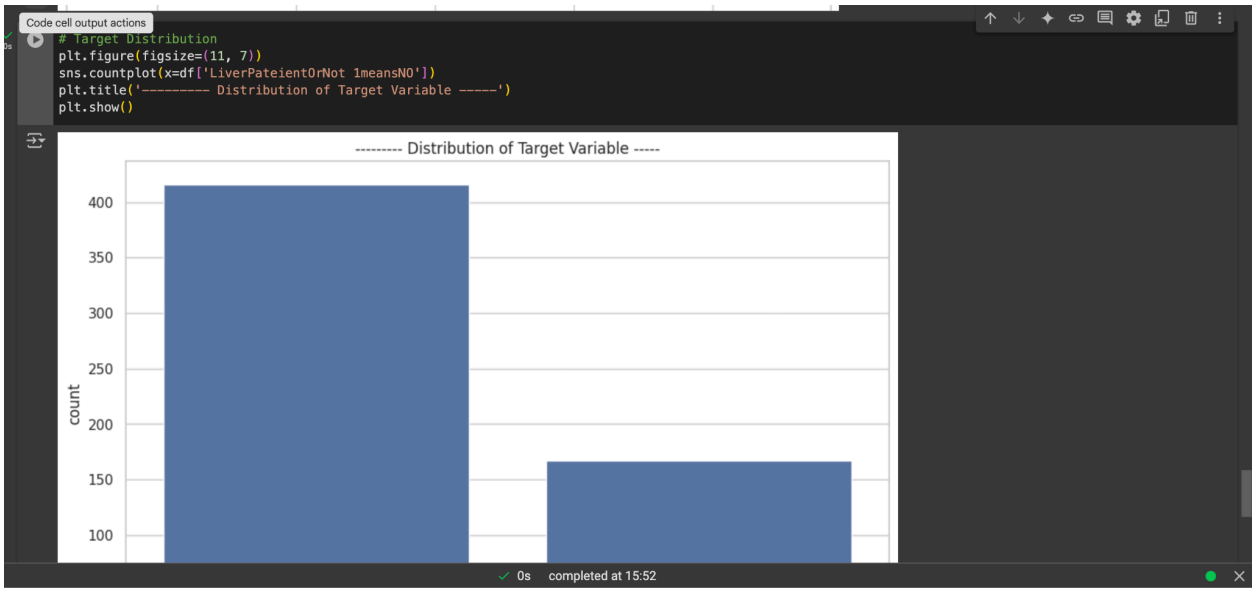

Distribution of Total_Protiens
Frequency
Total_Protiens


Alamine_Aminotransferase
Distribution of Aspartate_Aminotransferase
Frequency
Aspartate_Aminotransferase

```
# ---------- Correlation heatmap ------
plt.figure(figsize=(11, 7))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.show()
```

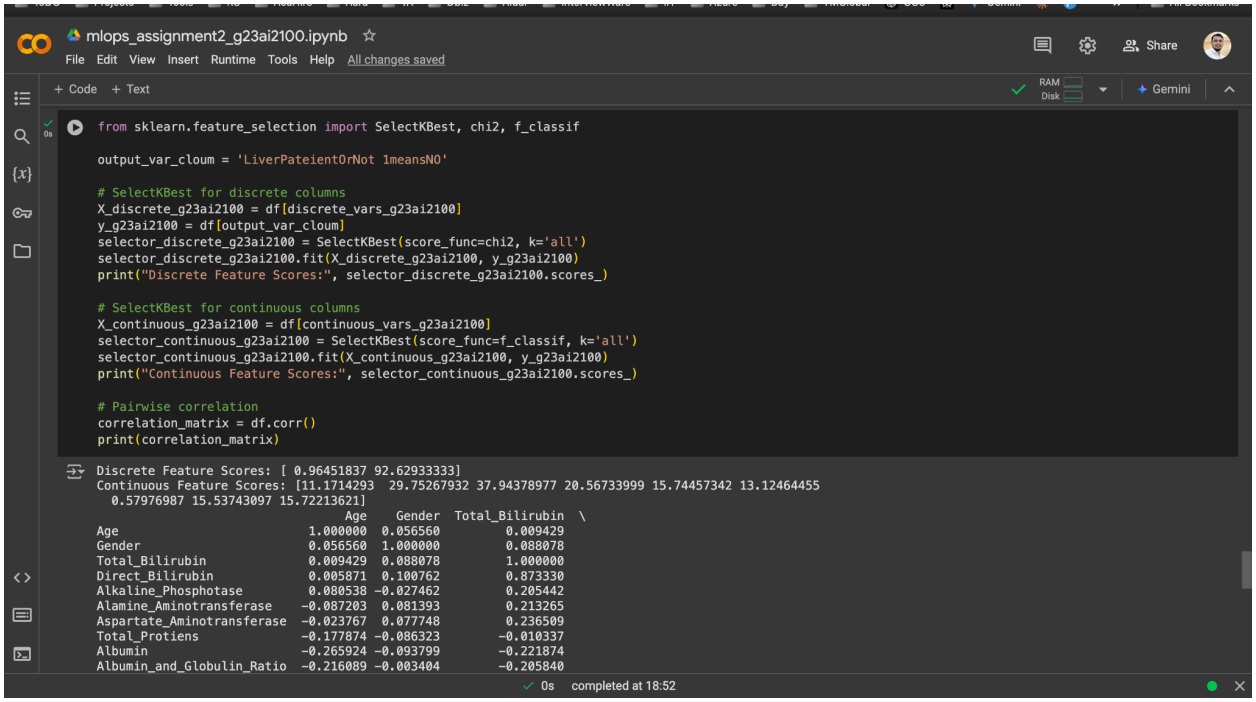






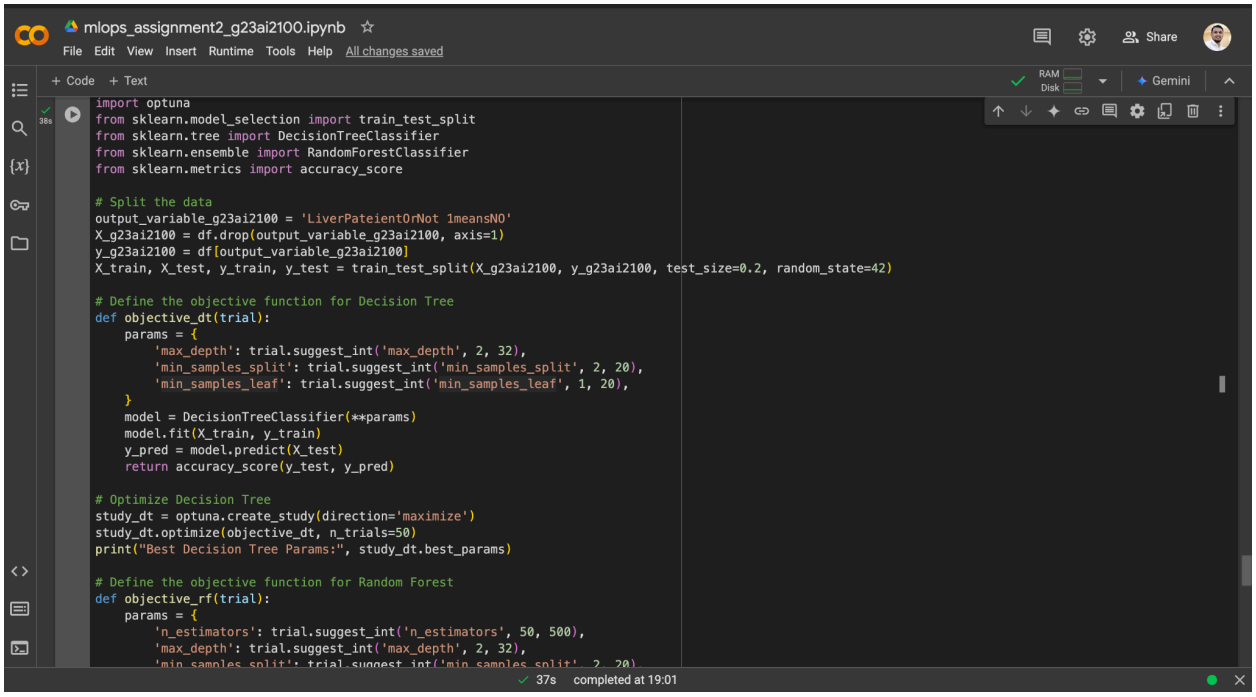
Q5 Perform different feature selection algorithms including pairwise correlation

Ans



Q6 Use optuna to fine tune decision tree, and random forest classification algorithms to find the best hyperparameter combination

Ans

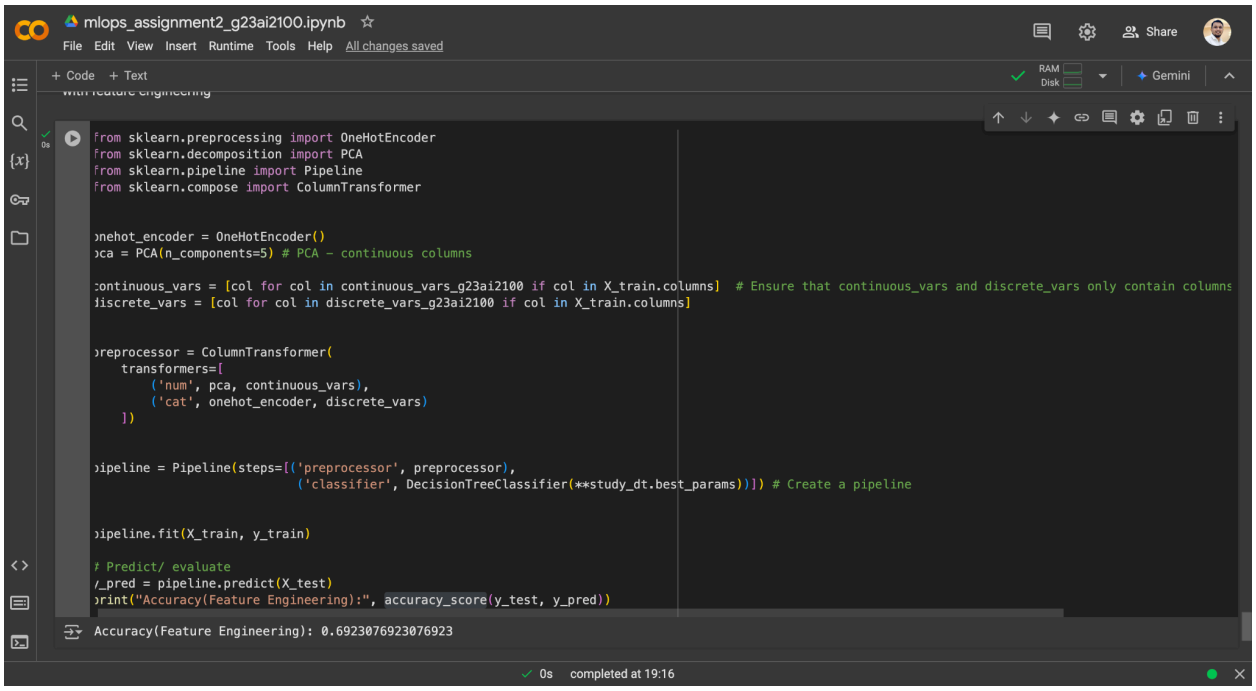


OutPut :

```
Best Random Forest Params: {'n_estimators': 458, 'max_depth': 3,
'min_samples_split': 8, 'min_samples_leaf': 9}
```

Q7 Do different feature engineering steps of PCA and one hot encoding and then do classification and compare without feature engineering and with feature engineering

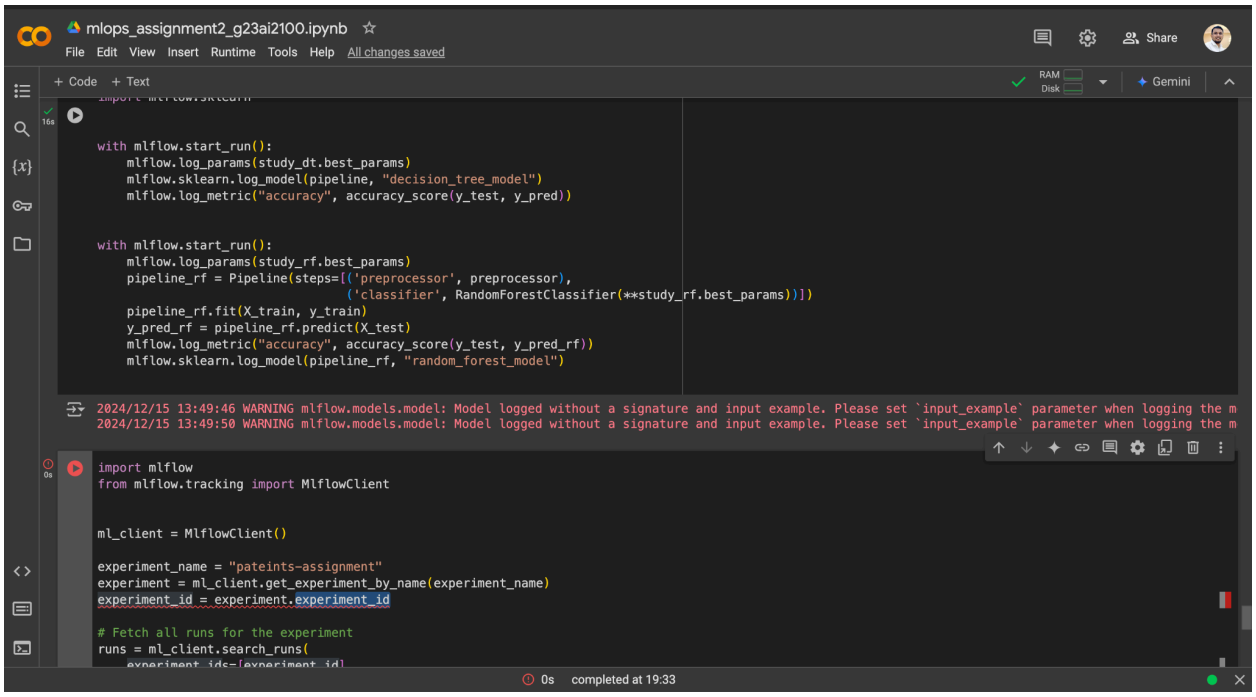
Ans



Output : `Accuracy(Feature Engineering): 0.6923076923076923`

Q8 Use MLFlow to run different runs of the same algorithms Decision Tree , and Random Forest and capture the input hyperparameters and output metrics of the same

Ans



Output

Run ID: 9fa6362996274af7ab6a9db912cab030

Parameters: {'n_estimators': '137', 'min_samples_leaf': '12', 'max_depth': '19', 'min_samples_split': '7'}

Metrics: {'accuracy': 0.7692307692307693}

Artifacts: [<FileInfo: file_size=None, is_dir=True, path='random_forest_model'>]

Run ID: e9ee433632254247889d5ce41f135a56

Parameters: {'min_samples_leaf': '19', 'max_depth': '23', 'min_samples_split': '2'}

Metrics: {'accuracy': 0.7008547008547008}

Artifacts: [<FileInfo: file_size=None, is_dir=True, path='decision_tree_model'>]