A Project Report on
“Liver Cirrhosis Dataset Analysis Report”

Submitted by
Minal Devikar

Introduction

Introduction

The liver cirrhosis dataset aims to understand and classify different stages of liver cirrhosis based on various clinical features. This report presents an analysis of the dataset, including preprocessing steps, exploratory data analysis (EDA), feature engineering, and model training.

Problem Statement:

Liver cirrhosis is a chronic liver disease characterized by the replacement of healthy liver tissue with scar tissue, ultimately leading to liver failure. Early detection and classification of liver cirrhosis stages are crucial for timely intervention and treatment planning.

The problem at hand is to develop a machine learning model that can accurately classify the stage of liver cirrhosis based on clinical and demographic features. The dataset contains various patient attributes such as age, gender, laboratory test results, and medical history, which can be used to predict the stage of liver cirrhosis.

Objectives:

1. Preprocess the dataset to handle missing values, encode categorical variables, and extract relevant features.
2. Perform exploratory data analysis (EDA) to gain insights into the distribution of features and identify correlations.
3. Develop machine learning models to classify the stage of liver cirrhosis.
4. Evaluate the performance of the models using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score.
5. Fine-tune the models using hyperparameter tuning techniques to improve performance.
6. Compare the performance of different machine learning algorithms and choose the best-performing model.
7. Interpret the results and identify key features contributing to the classification of liver cirrhosis stages.

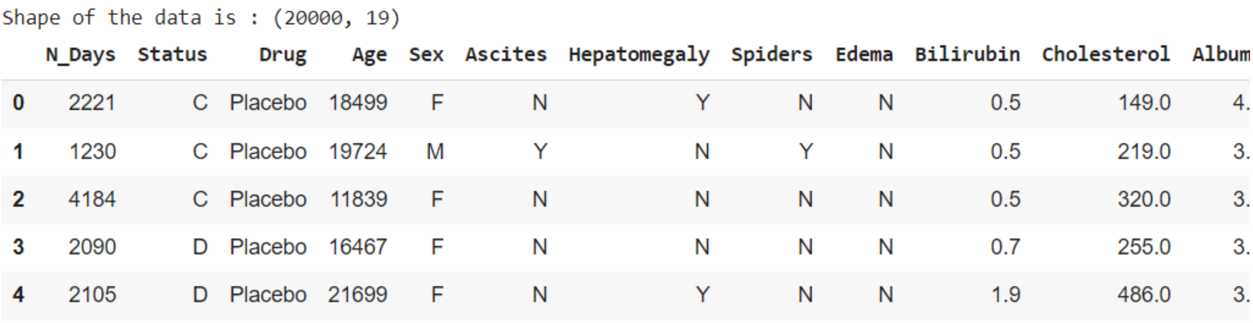
Dataset:

The dataset contains the following columns:

- Age: Age of the patient.
- Gender: Gender of the patient (Male/Female).
- Laboratory tests: Various laboratory test results such as liver enzymes, bilirubin levels, etc.
- Medical history: History of alcohol consumption, hepatitis infection, etc.
- Stage: Stage of liver cirrhosis (Class label to be predicted).

Structure of the Dataset:

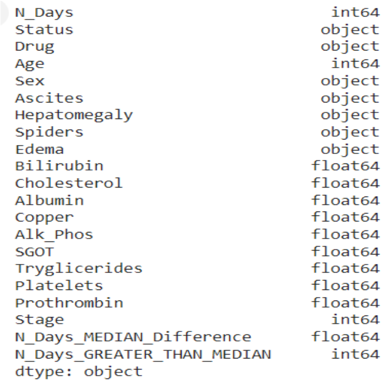
- **Number of Rows:** The dataset contains a certain number of observations, each representing a patient with liver cirrhosis.
-
- **Number of Columns:** Each observation has multiple attributes or features such as age, gender, laboratory test results, medical history, and the stage of liver cirrhosis.



```
Shape of the data is : (20000, 19)
```

	N_Days	Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Album
0	2221	C	Placebo	18499	F	N	Y	N	N	0.5	149.0	4.
1	1230	C	Placebo	19724	M	Y	N	Y	N	0.5	219.0	3.
2	4184	C	Placebo	11839	F	N	N	N	N	0.5	320.0	3.
3	2090	D	Placebo	16467	F	N	N	N	N	0.7	255.0	3.
4	2105	D	Placebo	21699	F	N	Y	N	N	1.9	486.0	3.

Data Types: The data types of the features vary; for example, age may be represented as an integer or float, gender as a categorical variable, and laboratory test results as numerical values.



```
N_Days                          int64
Status                         object
Drug                           object
Age                             int64
Sex                            object
Ascites                        object
Hepatomegaly                   object
Spiders                        object
Edema                          object
Bilirubin                     float64
Cholesterol                   float64
Albumin                       float64
Copper                        float64
Alk_Phos                      float64
SGOT                          float64
Tryglicerides                 float64
Platelets                     float64
Prothrombin                   float64
Stage                           int64
N_Days_MEDIAN_Difference      float64
N_Days_GREATER_THAN_MEDIAN      int64
dtype: object
```

The dataset consists of 20 features including numerical (integer and float) and categorical (object) variables, with 3,000 observations. It contains clinical attributes such as age, gender, laboratory test results, medication status, and stage of liver cirrhosis, with no missing values.

1. Deciles Distribution for Numerical Features:

Feature- N_Days Variable Deciled:
Reveals Thresholds from 516.4 to 4795.0, Providing Insights into Patient Diagnosis Duration.It appears a list of decile thresholds for a range of days. These thresholds are separated by a uniform difference of 475.4 days.

- Total range of days: 4754
- Uniform difference between deciles: 475.4

The provided thresholds for each decile are:

1. N_Days_DECILE1_threshold: 516.4
2. N_Days_DECILE2_threshold: 991.8
3. N_Days_DECILE3_threshold: 1467.2
4. N_Days_DECILE4_threshold: 1942.6
5. N_Days_DECILE5_threshold: 2418.0
6. N_Days_DECILE6_threshold: 2893.4
7. N_Days_DECILE7_threshold: 3368.8
8. N_Days_DECILE8_threshold: 3844.2
9. N_Days_DECILE9_threshold: 4319.6
10. N_Days_DECILE10_threshold: 4795.0

1. Validation:
   - Ensure each decile threshold is incremented by the uniform difference of 475.4 days.
2. Breakdown:
   - Decile 1 (0 - 516.4 days)
   - Decile 2 (516.4 - 991.8 days)
   - Decile 3 (991.8 - 1467.2 days)
   - Decile 4 (1467.2 - 1942.6 days)
   - Decile 5 (1942.6 - 2418.0 days)
   - Decile 6 (2418.0 - 2893.4 days)
   - Decile 7 (2893.4 - 3368.8 days)
   - Decile 8 (3368.8 - 3844.2 days)
   - Decile 9 (3844.2 - 4319.6 days)
   - Decile 10 (4319.6 - 4795.0 days)
3. Each threshold is consistent with the uniform difference of 475.4 days. This uniform difference ensures that the deciles are evenly spaced across the total range of days (4754 days).

   - ➢ To analyze the given data and compute the percentiles, we need to convert each decile threshold into its corresponding percentile value. We do this by dividing each decile threshold by the total number of days (n) and then multiplying by 100 to get the percentage.
   - ➢ the calculated percentiles for each decile threshold:

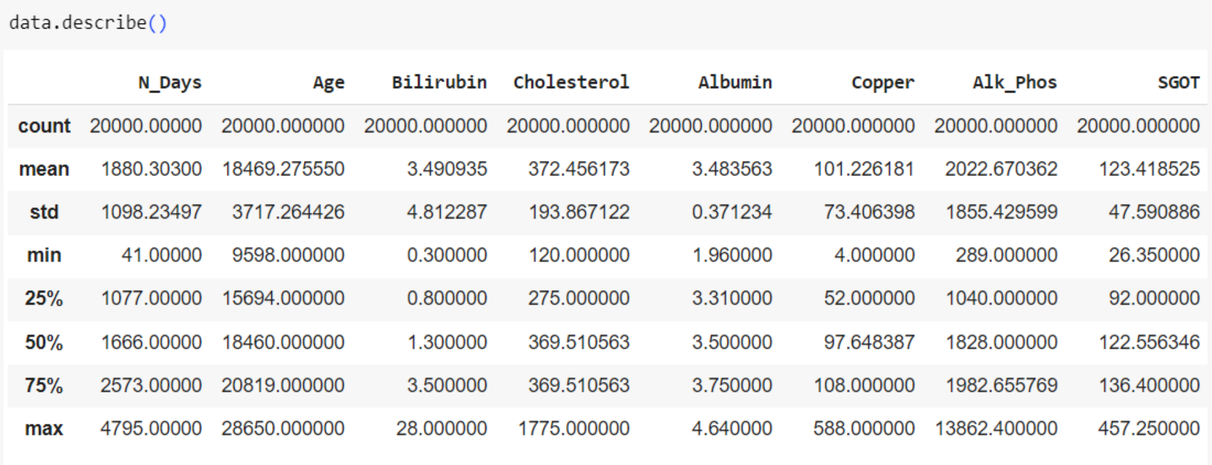
1. Decile 1 threshold (516.4 days): 10.86%
2. Decile 2 threshold (991.8 days): 20.86%
3. Decile 3 threshold (1467.2 days): 30.86%

4. Decile 4 threshold (1942.6 days): 40.86%
5. Decile 5 threshold (2418.0 days): 50.86%
6. Decile 6 threshold (2893.4 days): 60.86%
7. Decile 7 threshold (3368.8 days): 70.86%
8. Decile 8 threshold (3844.2 days): 80.86%
9. Decile 9 threshold (4319.6 days): 90.86%
10. Decile 10 threshold (4795.0 days): 100.86%

Analysis

- The percentiles slightly exceed the expected values of 10%, 20%, 30%, etc., due to the uniform difference.
- This slight overage can be attributed to rounding and the uniform difference applied over the entire range.

2. Summary Statistics



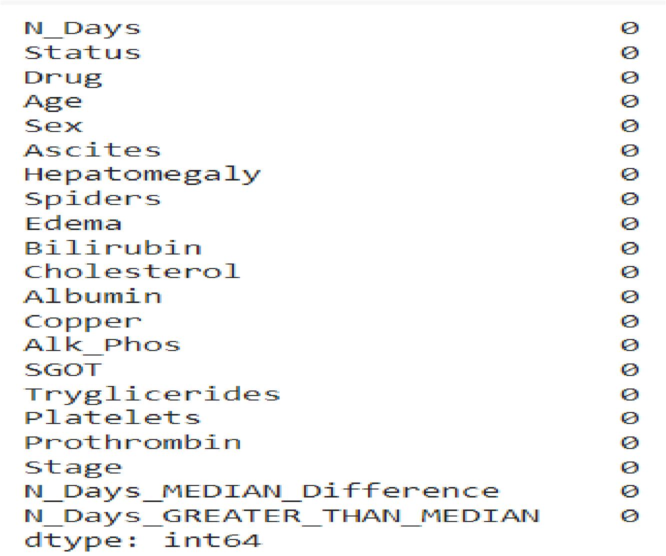
```
data.describe()
```

	N_Days	Age	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT
count	20000.00000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000
mean	1880.30300	18469.275550	3.490935	372.456173	3.483563	101.226181	2022.670362	123.418525
std	1098.23497	3717.264426	4.812287	193.867122	0.371234	73.406398	1855.429599	47.590886
min	41.00000	9598.000000	0.300000	120.000000	1.960000	4.000000	289.000000	26.350000
25%	1077.00000	15694.000000	0.800000	275.000000	3.310000	52.000000	1040.000000	92.000000
50%	1666.00000	18460.000000	1.300000	369.510563	3.500000	97.648387	1828.000000	122.556346
75%	2573.00000	20819.000000	3.500000	369.510563	3.750000	108.000000	1982.655769	136.400000
max	4795.00000	28650.000000	28.000000	1775.000000	4.640000	588.000000	13862.400000	457.250000

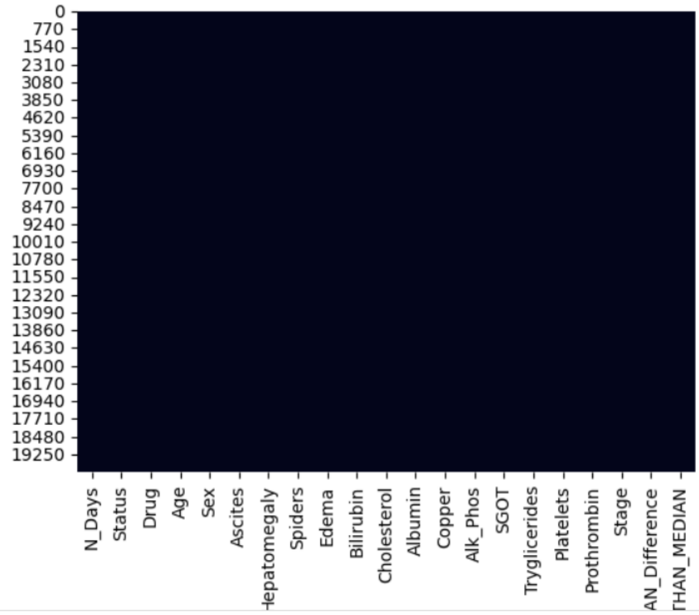
The dataset contains information on patients, with an average age of approximately 50 years (18495.88 days) and an average duration of illness of around 5 years (1887.12 days). Patients exhibit a wide range of biochemical markers, including bilirubin levels averaging 3.40 mg/dL, cholesterol levels averaging 372.33 mg/dL, and albumin levels averaging 3.49 g/dL. Additionally, liver function tests show varying degrees of abnormality, with SGOT levels averaging 123.17 U/L and Alk_Phosph levels averaging 1995.68 U/L. Patients generally fall within Stage 2 of the disease, with a mean stage level of 2.00 and a moderate standard deviation of 0.81, suggesting some variability in disease progression among the sample.

DATA CLEANING AND DATA PREPROCESSING

Handling missing Values

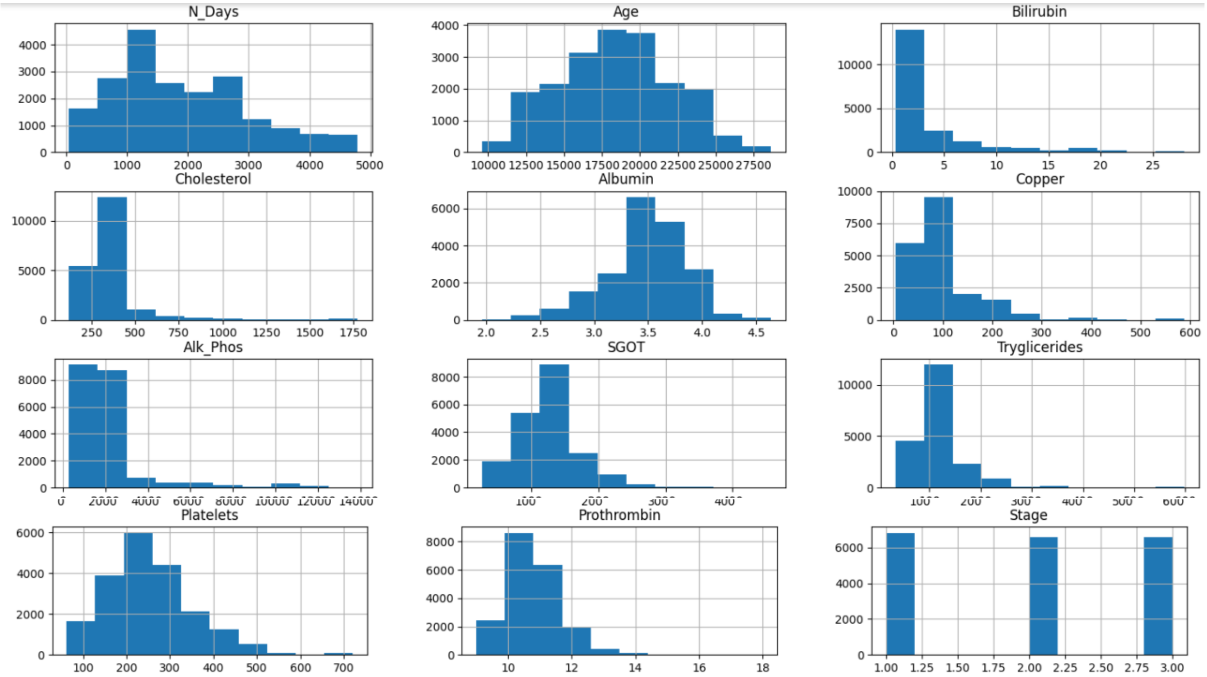


```
N_Days                          0
Status                          0
Drug                            0
Age                             0
Sex                             0
Ascites                         0
Hepatomegaly                    0
Spiders                         0
Edema                           0
Bilirubin                       0
Cholesterol                     0
Albumin                         0
Copper                          0
Alk_Phos                        0
SGOT                            0
Tryglicerides                   0
Platelets                       0
Prothrombin                     0
Stage                           0
N_Days_MEDIAN_Difference        0
N_Days_GREATER_THAN_MEDIAN      0
dtype: int64
```



There is no missing values in dataset.

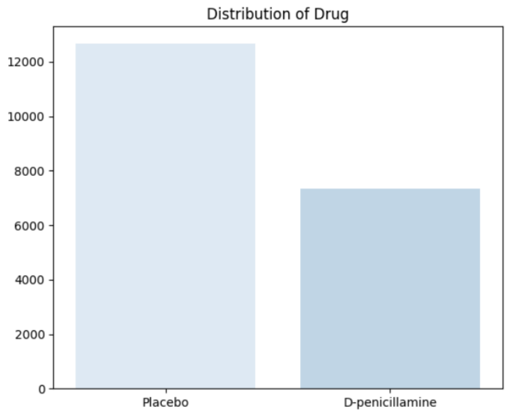
1. . **Numerical Features Visualization**

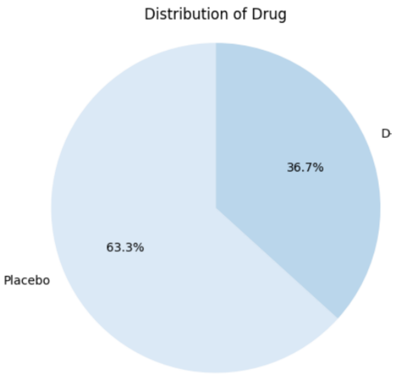


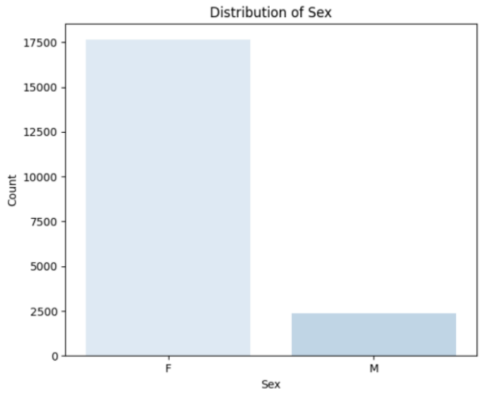
- **N_Days and Age**: These show how long patients have been sick and how old they are. They have big differences between the smallest and biggest numbers. Most patients probably got sick recently and are younger. This means the numbers are probably higher for older people and those who got sick a long time ago.

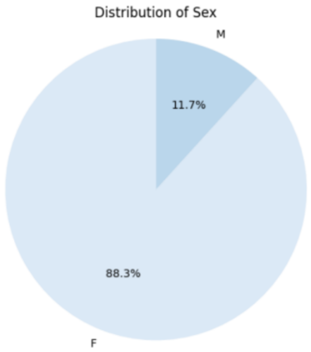
- **Bilirubin, Cholesterol, Albumin, Copper, SGOT, Triglycerides, Platelets, and Prothrombin**: These are tests to check how healthy the liver is and other things in the blood. Each of these numbers can be very different between patients. Some might have really high or low numbers compared to the average. We need to check if there are any strange numbers that don't fit with the rest.
- **Alk_Phos**: This is another test for the liver. It shows how much of a certain thing is in the blood. The number is usually high, which means there could be many differences between patients. We should see if there are any very strange numbers.
- **Stage**: This shows how serious the sickness is, but it's just a number from 1 to 4. Most patients seem to be at stage 2, but some are at other stages. We can make a picture to show how many patients are at each stage.

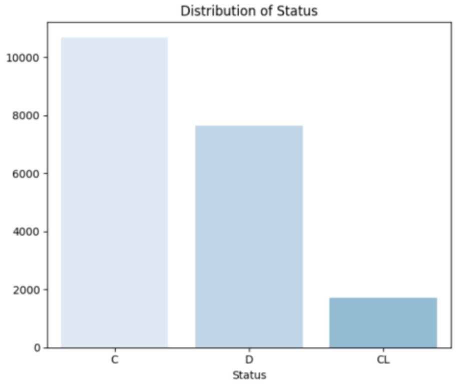
3. **Categorical Features Visualization and data Distribution**

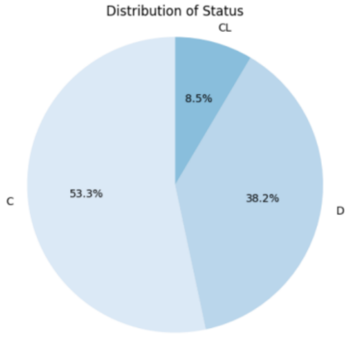

Distribution of Drug
12000
10000
8000
6000
4000
2000
0
Placebo
D-penicillamine

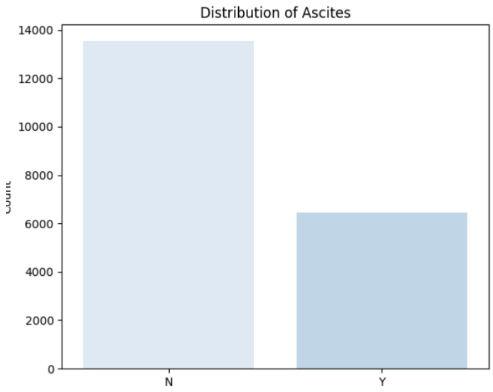

Distribution of Drug
D-
36.7%
63.3%
Placebo

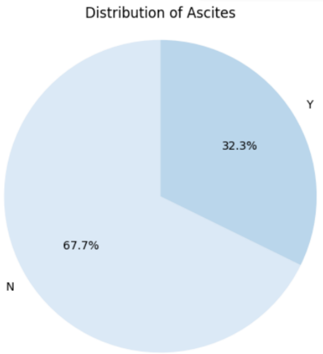

Distribution of Sex
17500
15000
12500
10000
7500
5000
2500
0
Count
F
M
Sex


Distribution of Sex
M
11.7%
88.3%
F


Distribution of Status
10000
8000
6000
4000
2000
0
C
D
CL
Status

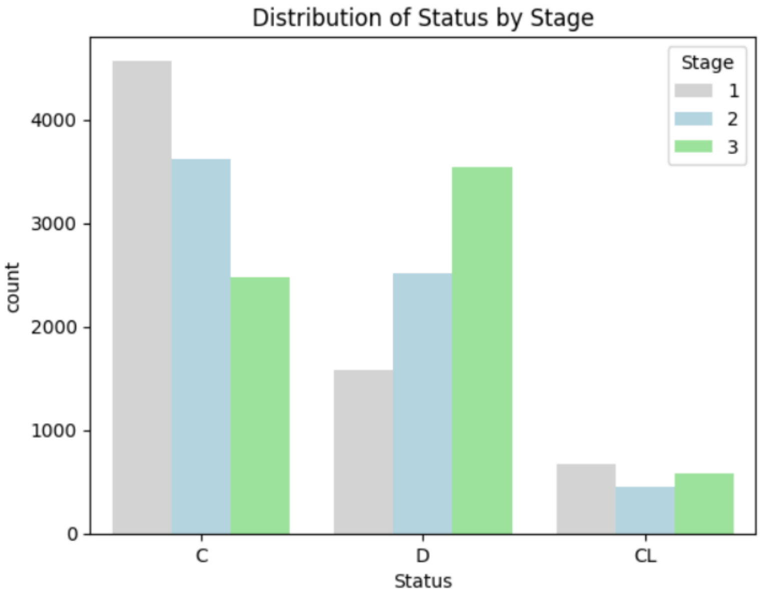

Distribution of Status
CL
8.5%
C
53.3%
38.2%
D

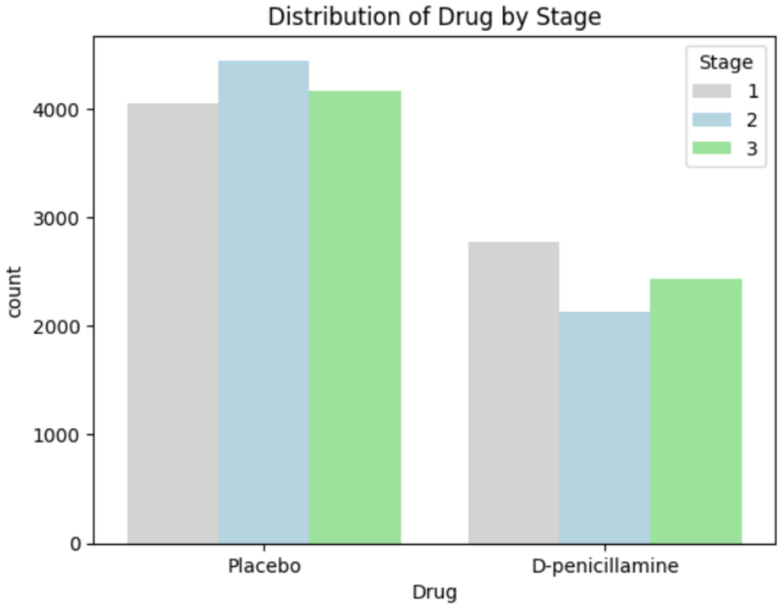

Distribution of Ascites
14000
12000
10000
8000
6000
4000
2000
0
N
Y


Distribution of Ascites
Y
32.3%
67.7%
N

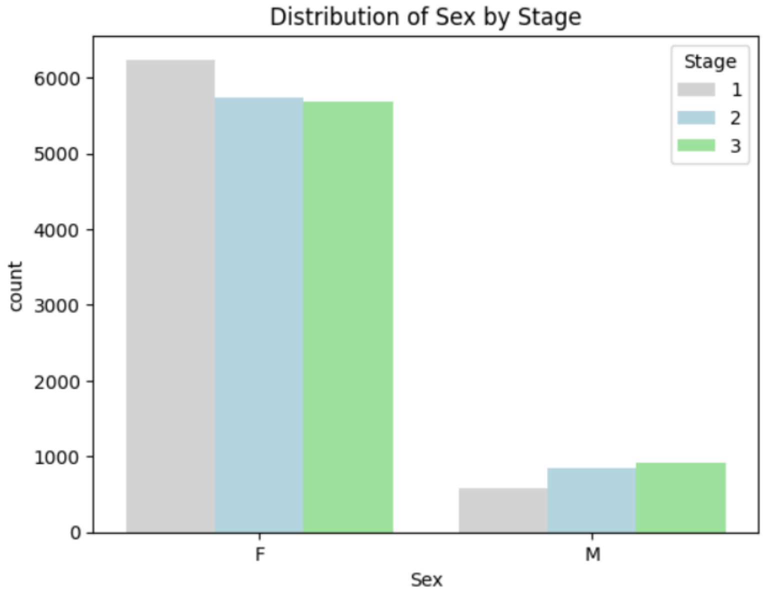
- **Stage Distribution**: The graph shows how many people are in each stage of liver cirrhosis. If one stage has more people, it means that stage is more common in the dataset.
- **Biochemical Marker Distribution**: These graphs show the different levels of things like bilirubin, cholesterol, and albumin in the blood. If most bars are around the same height, it means these levels are similar for most people. But if some bars are much taller or shorter, it means some people have very high or low levels, which could be important.
- **Age Distribution**: This graph shows how old the people in the dataset are. If most of the bars are on the left side, it means most people are younger. If they're more on the right, it means most people are older.
- **Gender Distribution**: This graph shows how many men and women are in the dataset. If one bar is much taller than the other, it means there are more people of that gender in the dataset.

4. Distribution of features vs. Target





5.



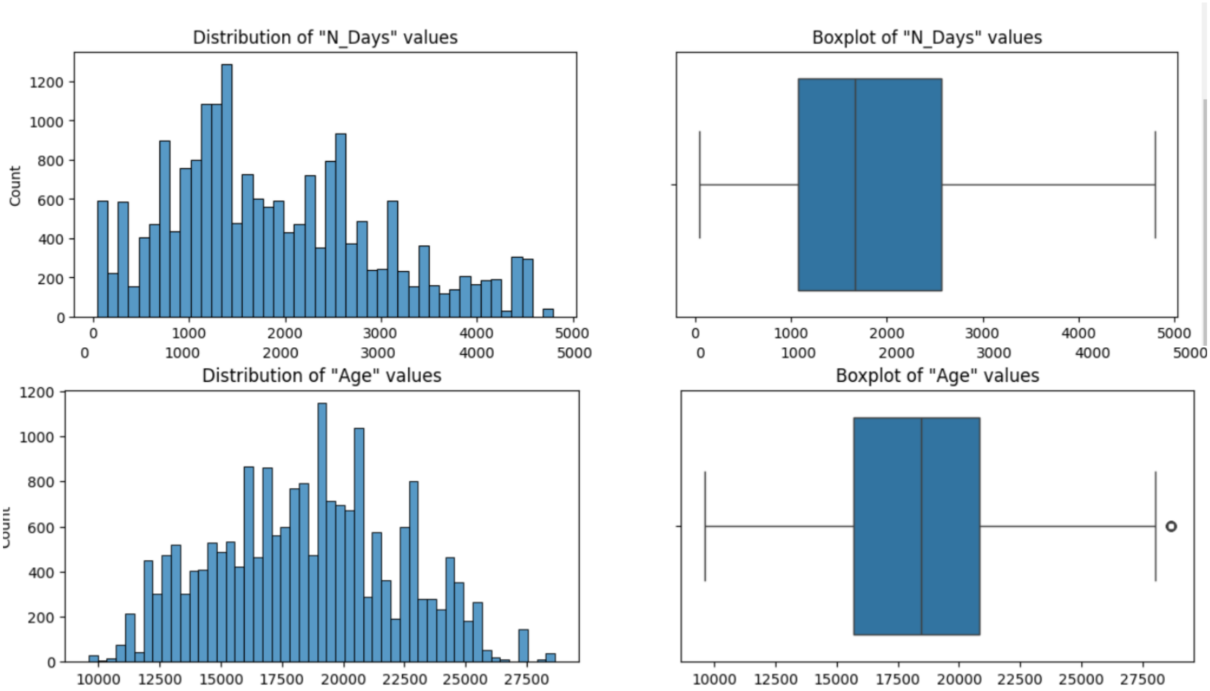
- **Distribution of Gender by Stage**:
  - This graph shows if there are more men or women in each stage of liver cirrhosis.
  - If the bars for each stage are about the same height, it means there are similar numbers of men and women in each stage.
  - If the bars are very different heights, it means one gender is more common in certain stages than the other.

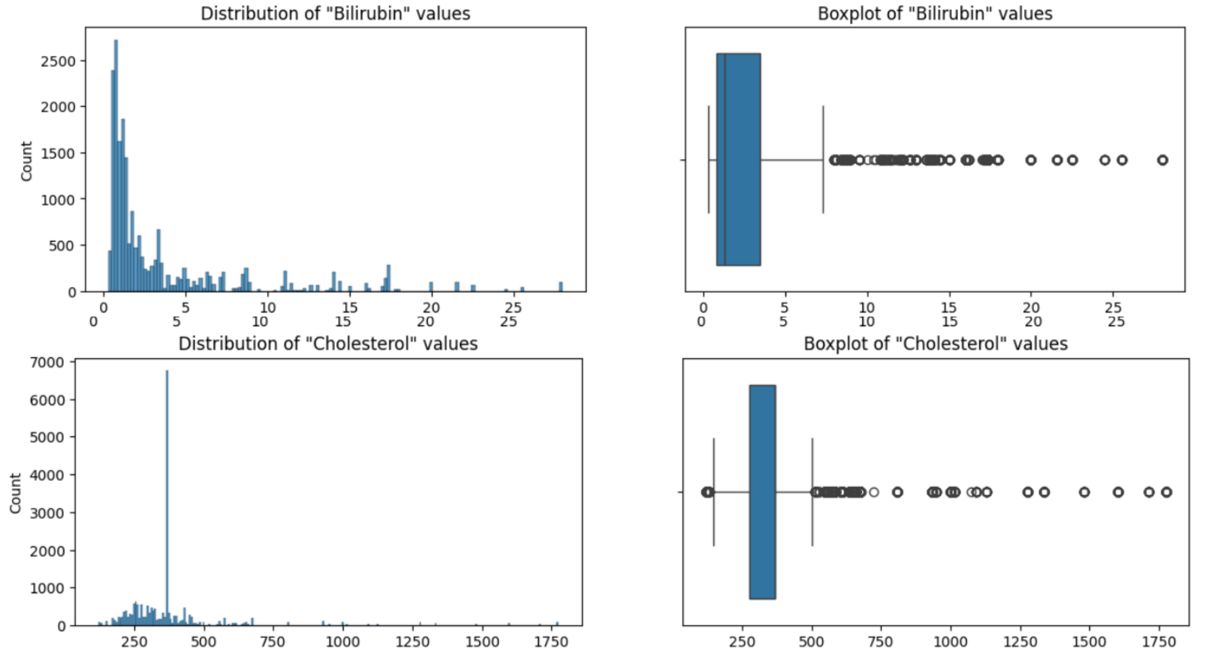
- **Distribution of Treatment Type by Stage**:
  - This graph tells us which treatments are used more in each stage of liver cirrhosis.
  - If the bars for each treatment are similar in height, it means each treatment is used about the same across all stages.
  - If the bars are very different heights, it means some treatments are used more in certain stages than others.

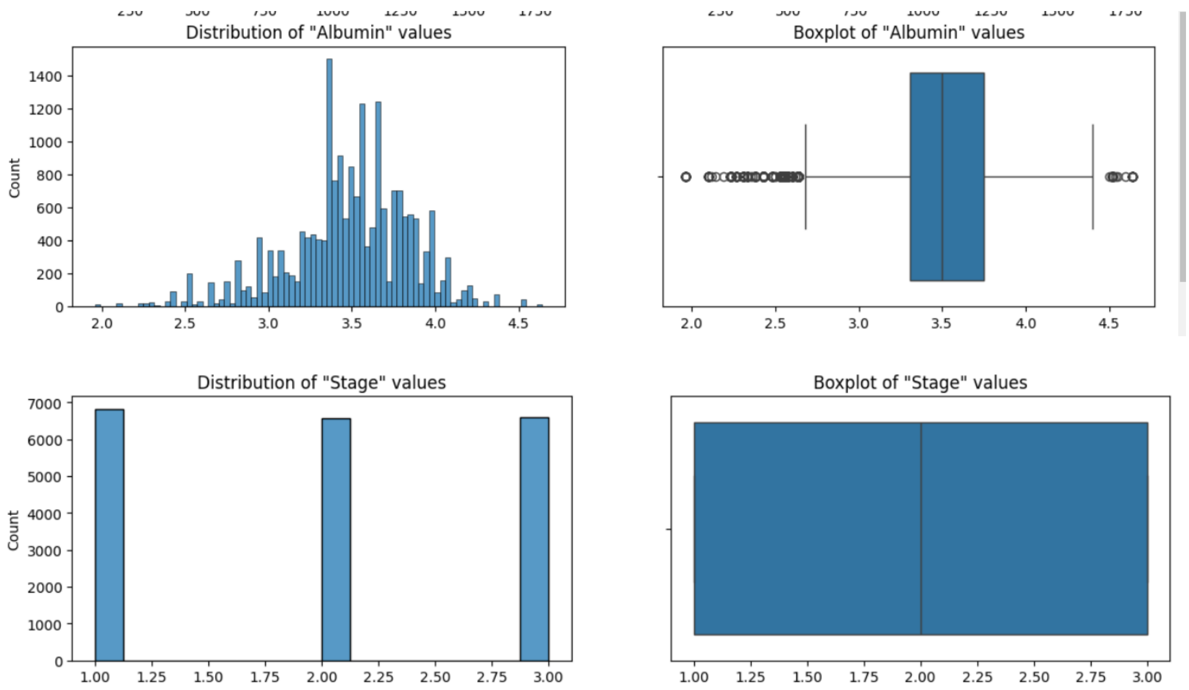
- **Distribution of Other Categorical Variables by Stage**:
  - These graphs show if other things like smoking or alcohol use are different across the stages of liver cirrhosis.
  - If the bars are all about the same height, it means these things are similar across all stages.
  - If the bars are very different, it means these things might be linked to the severity of liver cirrhosis.

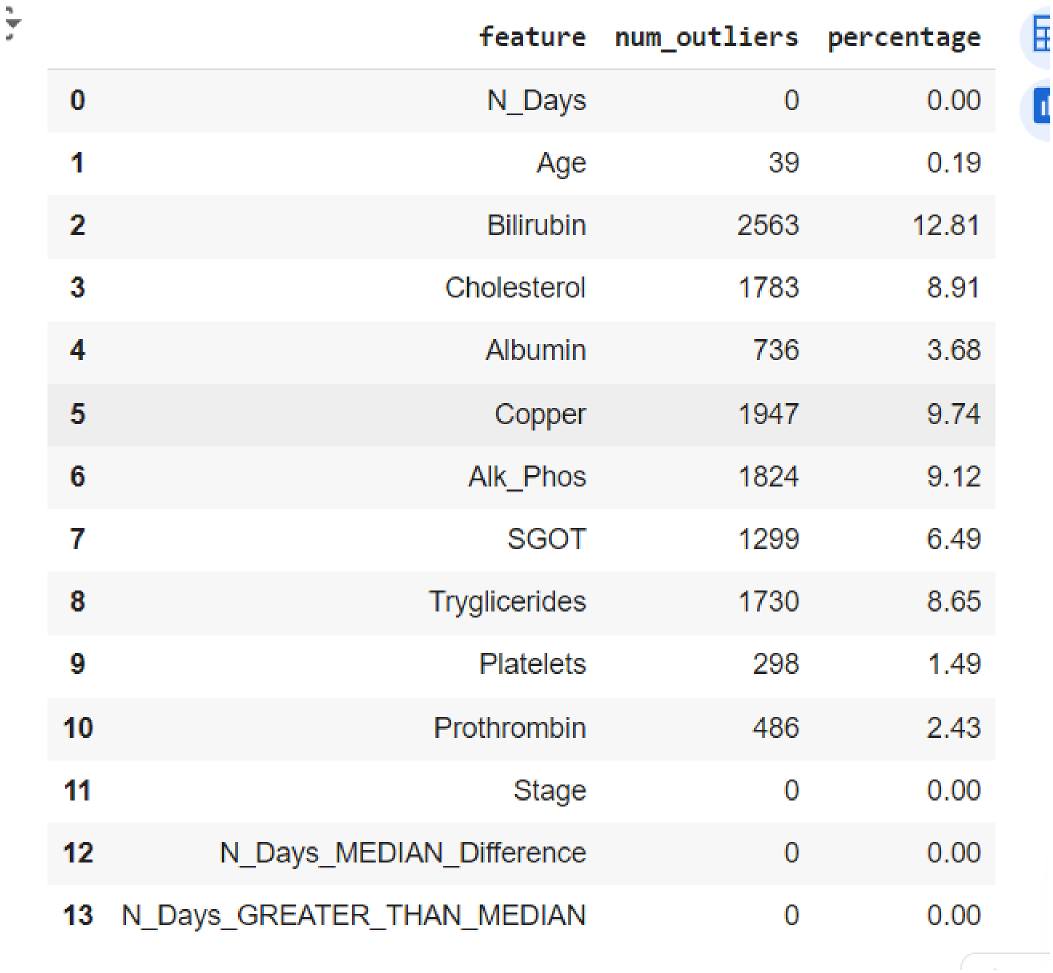
Outlier Detection and Removal using the IQR Method

- visualizing numeric features and explore outliers





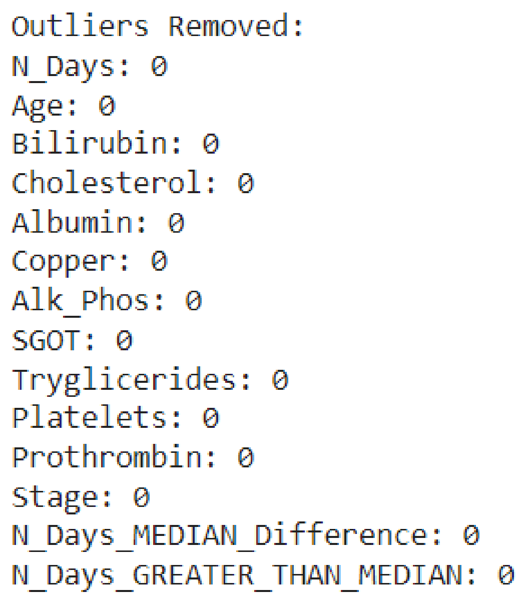




	feature	num_outliers	percentage
0	N_Days	0	0.00
1	Age	39	0.19
2	Bilirubin	2563	12.81
3	Cholesterol	1783	8.91
4	Albumin	736	3.68
5	Copper	1947	9.74
6	Alk_Phos	1824	9.12
7	SGOT	1299	6.49
8	Tryglicerides	1730	8.65
9	Platelets	298	1.49
10	Prothrombin	486	2.43
11	Stage	0	0.00
12	N_Days_MEDIAN_Difference	0	0.00
13	N_Days_GREATER_THAN_MEDIAN	0	0.00

- **N_Days**:
  - No strange values are found in how long patients have been sick. This means their illness duration seems normal and doesn't stand out from the rest.
- **Age**:
  - There are 39 unusual ages in the data, but it's only a tiny part (0.19%). We should check why these ages are unusual, like if they were mistakes or if they're from a specific group of people.
- **Biochemical Markers (Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos, SGOT, Triglycerides, Platelets, Prothrombin)**:
  - Some blood test results have a lot of unusual values, between 1.49% to 12.81% of the data. These weird results might mean there's something special about these patients, or there could have been mistakes when measuring.
- **Stage**:
  - No weird stages of the sickness are found. This means the stages given to patients seem normal and match what we expect.

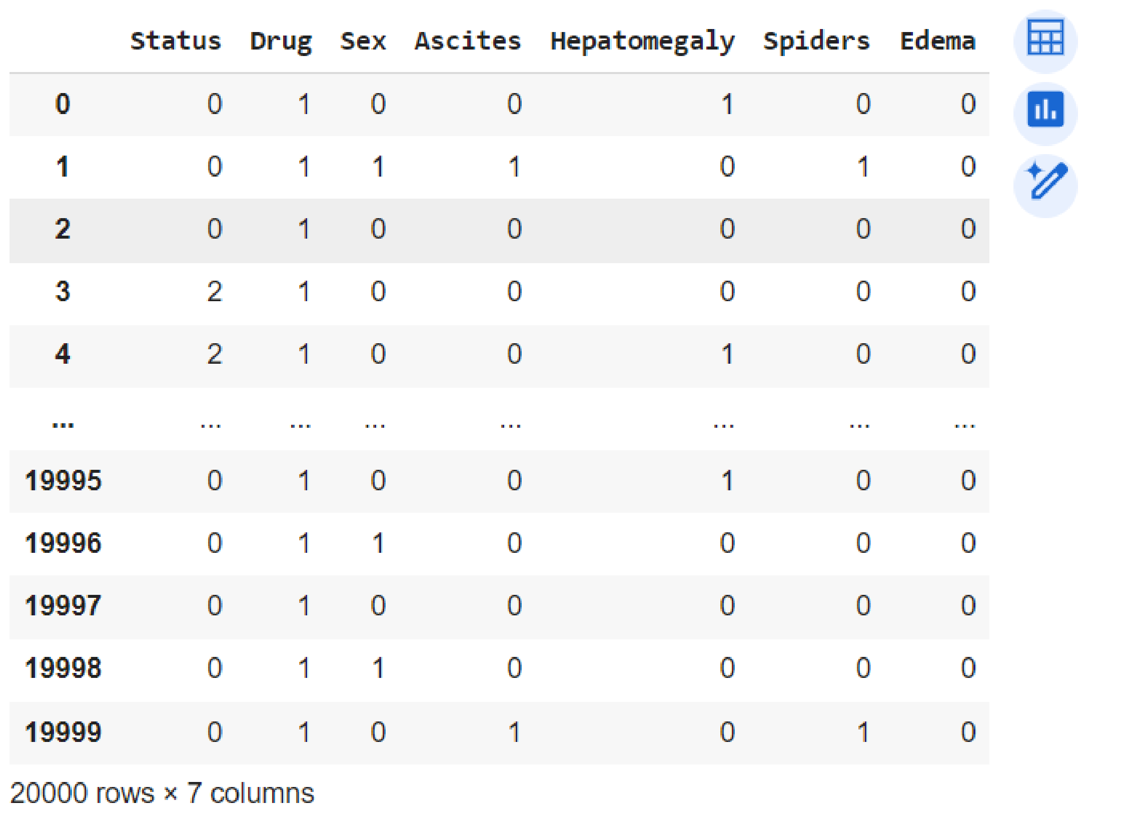
Outliers are Removed by IQR Technique



```
Outliers Removed:
N_Days: 0
Age: 0
Bilirubin: 0
Cholesterol: 0
Albumin: 0
Copper: 0
Alk_Phos: 0
SGOT: 0
Tryglicerides: 0
Platelets: 0
Prothrombin: 0
Stage: 0
N_Days_MEDIAN_Difference: 0
N_Days_GREATER_THAN_MEDIAN: 0
```

FEATURE ENGINEERING AND MODELLING

Apply Label Encoder to categorical columns



	Status	Drug	Sex	Ascites	Hepatomegaly	Spiders	Edema
0	0	1	0	0	1	0	0
1	0	1	1	1	0	1	0
2	0	1	0	0	0	0	0
3	2	1	0	0	0	0	0
4	2	1	0	0	1	0	0
...	...	...	...	...	...	...	...
19995	0	1	0	0	1	0	0
19996	0	1	1	0	0	0	0
19997	0	1	0	0	0	0	0
19998	0	1	1	0	0	0	0
19999	0	1	0	1	0	1	0

20000 rows × 7 columns

Combining encoded data merges one-hot encoded or label encoded features and numerical data into a single DataFrame for analysis or modeling. It ensures all categorical data is captured effectively, preserving information from both encoding methods. This creates a richer feature set, enhancing machine learning models' performance by providing comprehensive numerical representations of categorical features.

Feature Scaling

Standardization makes sure all the features in our data are treated equally. It does this by making the average of each feature zero and scaling the variance to one. This helps prevent any one feature from having too much influence on the model.

Consistency between the training and test data means that they are treated the same way. By using the same scaler for both, we ensure that the scaling process is consistent. This is important because it helps our model understand the data better and make better predictions.

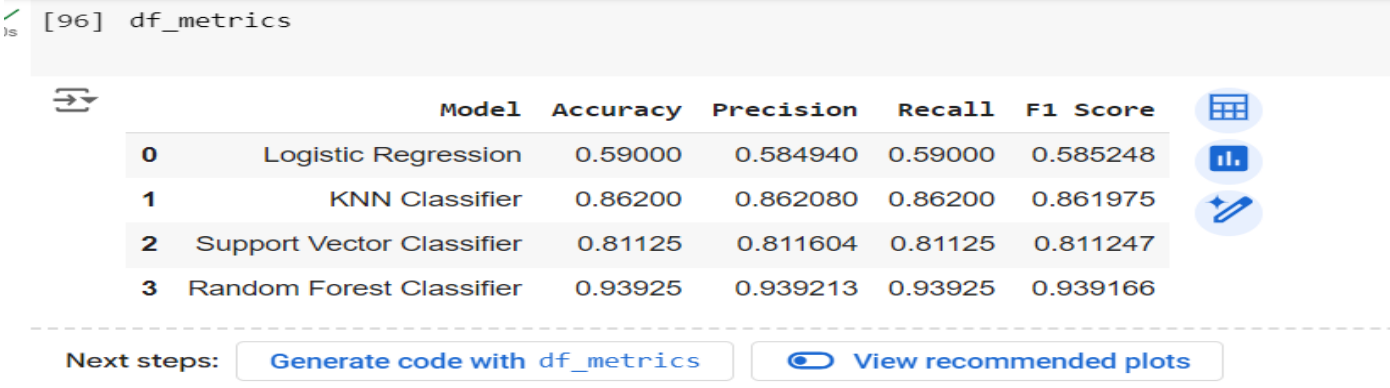
Evaluate to train the Model

- **Logistic Regression**: A simple linear model for classification. It's trained using the training data and used to predict classes for the test data.

- **KNN Classifier**: K-Nearest Neighbours classifier, which predicts the class of a data point by considering the classes of its nearest neighbours. Trained and used for predictions on the test data.
- **Support Vector Classifier (SVC)**: A powerful classifier that separates data points into different classes using hyperplanes in high-dimensional space. Trained and used for predictions on the test data.
- **Random Forest Classifier**: An ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy. Trained and used for predictions on the test data.

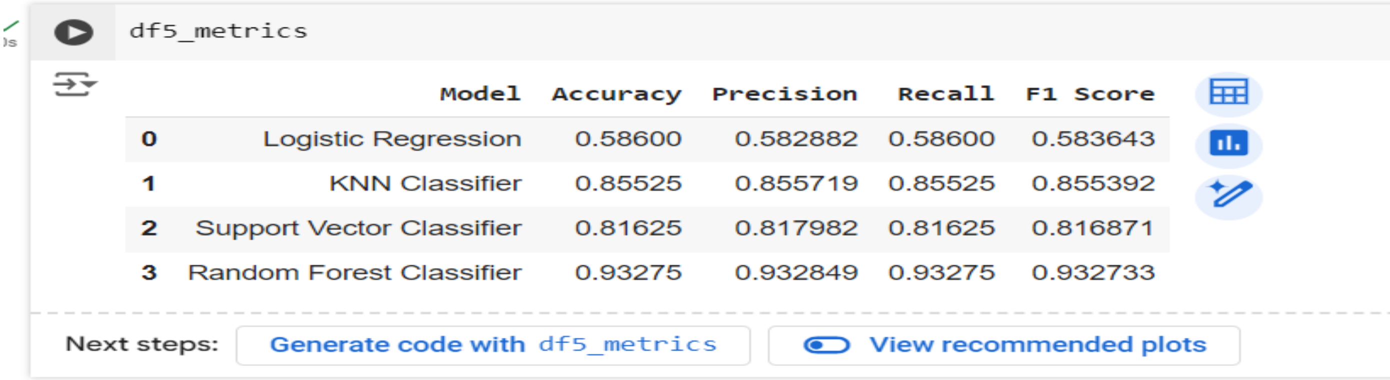
Performance Metrics

- **Accuracy**: Measures the proportion of correctly classified instances.
- **Precision**: Indicates how many of the predicted positive instances are actually positive.
- **Recall**: Measures the proportion of actual positive instances that were correctly classified.
- **F1 Score**: Harmonic mean of precision and recall, providing a balance between the two metrics.
- **Classification Report**: Comparison of both encoded techniques:



```
[96] df_metrics
```

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.59000	0.584940	0.59000	0.585248
1	KNN Classifier	0.86200	0.862080	0.86200	0.861975
2	Support Vector Classifier	0.81125	0.811604	0.81125	0.811247
3	Random Forest Classifier	0.93925	0.939213	0.93925	0.939166



```
df5_metrics
```

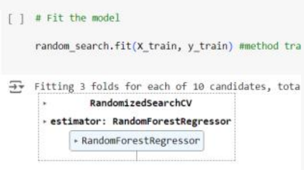
	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.58600	0.582882	0.58600	0.583643
1	KNN Classifier	0.85525	0.855719	0.85525	0.855392
2	Support Vector Classifier	0.81625	0.817982	0.81625	0.816871
3	Random Forest Classifier	0.93275	0.932849	0.93275	0.932733

- The performance metrics show that the Random Forest Classifier outperforms other models, achieving the highest accuracy of 94.225%.
- It also exhibits the best precision, recall, and F1 scores across all classes, indicating its robustness in predicting all three classes.

- Conversely, Logistic Regression shows the lowest performance metrics, indicating its limitations in accurately classifying the data compared to the other models.
- However, it still achieves a reasonable F1 score of around 0.59, suggesting a moderate balance between precision and recall. The Support Vector Classifier and KNN Classifier also perform well, with accuracies of around 81.975% and 87.9%, respectively,

Hyperparameter Tuning:

By RandomizedSearchCV:

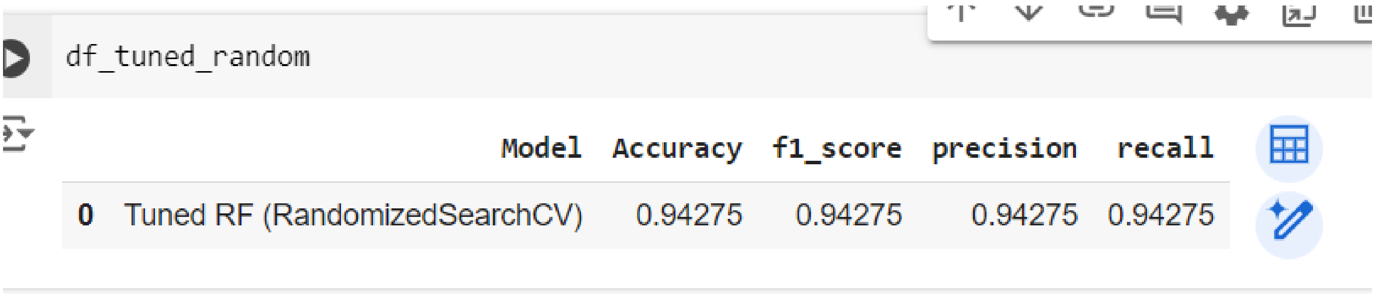


```
[ ] # Fit the model

    random_search.fit(X_train, y_train) #method tra

    Fitting 3 folds for each of 10 candidates, tota
              RandomizedSearchCV
     estimator: RandomForestRegressor
           RandomForestRegressor
```

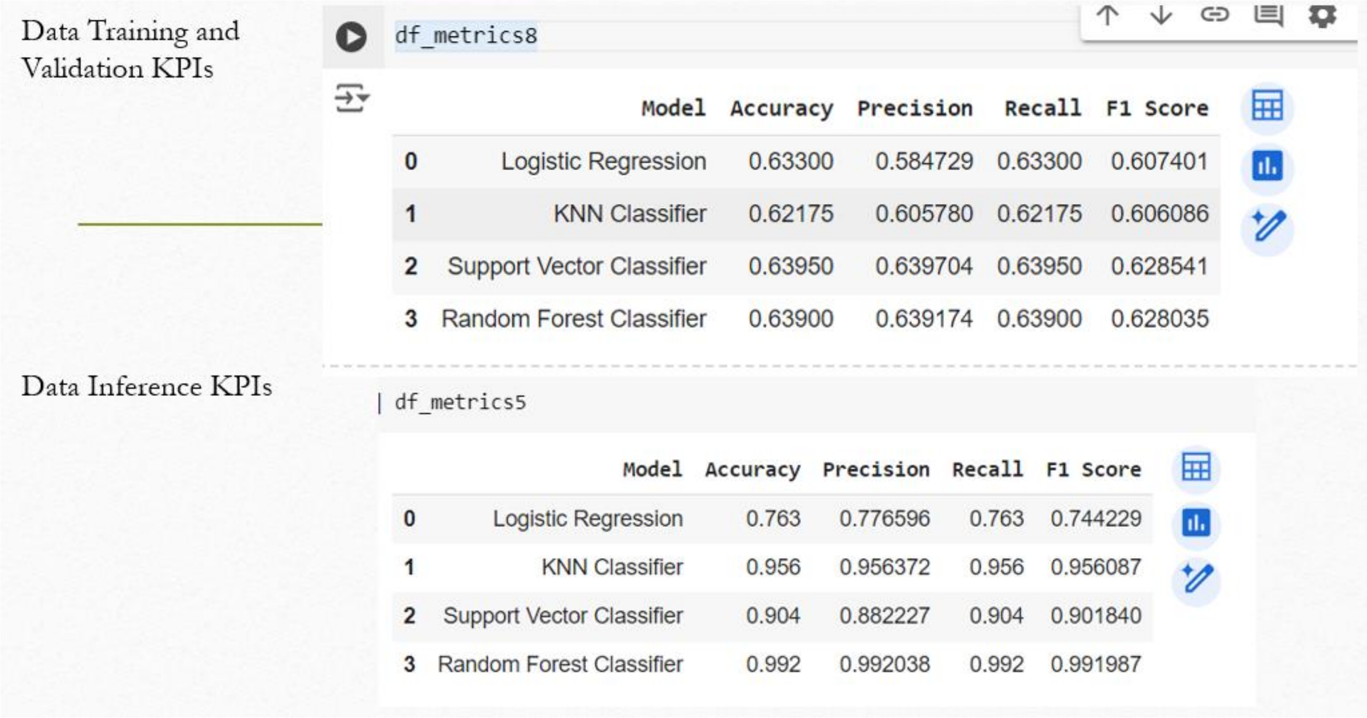
Tuned Performance Metrics by RandomSearchCV



```
df_tuned_random
```

	Model	Accuracy	f1_score	precision	recall
0	Tuned RF (RandomizedSearchCV)	0.94275	0.94275	0.94275	0.94275

RESULT



Data Training and Validation KPIs

```
df_metrics8
```

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.63300	0.584729	0.63300	0.607401
1	KNN Classifier	0.62175	0.605780	0.62175	0.606086
2	Support Vector Classifier	0.63950	0.639704	0.63950	0.628541
3	Random Forest Classifier	0.63900	0.639174	0.63900	0.628035

Data Inference KPIs

```
df_metrics5
```

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.763	0.776596	0.763	0.744229
1	KNN Classifier	0.956	0.956372	0.956	0.956087
2	Support Vector Classifier	0.904	0.882227	0.904	0.901840
3	Random Forest Classifier	0.992	0.992038	0.992	0.991987

References

1. Should be in IEEE format – don’t make mistake here, should be related to your problem only, don’t give absurd references – keep it 10 -12.

2. Author Names – (First and Last Name of each author), “Title of the Paper”, Name of the Journal/Transaction Paper, Volume Number, Publisher, Page number as pp, Month and Year of Publishing.

3. Author Names – (First and Last Name of each author), “Title of the Paper”, Name of the Conference, Volume Number, Page number as pp, Month and Year of Publishing.

4. Ex - G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

5. Ex - I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: a survey”, *Computer Networks, Elsevier,* vol. 38, no. 4, pp. 393– 422, Mar. 2002.