

Data Collection and Preprocessing Phase

Date	28 June 2024
Team ID	739897
Project Title	Predictive Pulse: Harnessing Machine Learning For Blood Pressure Analysis.
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

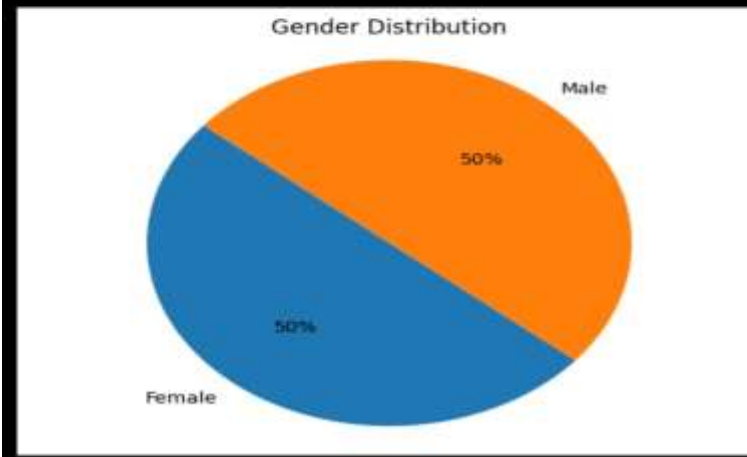
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description																																																																									
Data Overview	Descriptive Analysis:- <div><pre>df.describe()</pre><table><tr><th></th><th>Gender</th><th>Age</th><th>History</th><th>Patient</th><th>TakeMedication</th><th>Severity</th><th>BreathShortness</th><th>VisualChanges</th><th>NoseBleeding</th><th>WhenDiagnosed</th><th>Systolic</th><th>Diastolic</th><th>ControlledDiet</th><th>Stages</th></tr><tr><td>count</td><td>1825</td><td>1825</td><td>1825</td><td>1825</td><td>1825</td><td>1825</td><td>1825</td><td>1825</td><td>1825</td><td>1825</td><td>1825</td><td>1825</td><td>1825</td></tr><tr><td>unique</td><td>2</td><td>4</td><td>2</td><td>2</td><td>3</td><td>3</td><td>2</td><td>2</td><td>3</td><td>3</td><td>5</td><td>5</td><td>2</td></tr><tr><td>top</td><td>Female</td><td>51-64</td><td>Yes</td><td>No</td><td>No</td><td>Moderate</td><td>No</td><td>No</td><td>No</td><td><1 Year</td><td>111 - 120</td><td>81 - 90</td><td>No</td><td>HYPERTENSION (Stage-1)</td></tr><tr><td>freq</td><td>913</td><td>475</td><td>1657</td><td>984</td><td>744</td><td>697</td><td>976</td><td>940</td><td>984</td><td>625</td><td>1008</td><td>708</td><td>984</td><td>648</td></tr></table></div>		Gender	Age	History	Patient	TakeMedication	Severity	BreathShortness	VisualChanges	NoseBleeding	WhenDiagnosed	Systolic	Diastolic	ControlledDiet	Stages	count	1825	1825	1825	1825	1825	1825	1825	1825	1825	1825	1825	1825	1825	unique	2	4	2	2	3	3	2	2	3	3	5	5	2	top	Female	51-64	Yes	No	No	Moderate	No	No	No	<1 Year	111 - 120	81 - 90	No	HYPERTENSION (Stage-1)	freq	913	475	1657	984	744	697	976	940	984	625	1008	708	984	648
		Gender	Age	History	Patient	TakeMedication	Severity	BreathShortness	VisualChanges	NoseBleeding	WhenDiagnosed	Systolic	Diastolic	ControlledDiet	Stages																																																											
	count	1825	1825	1825	1825	1825	1825	1825	1825	1825	1825	1825	1825	1825																																																												
	unique	2	4	2	2	3	3	2	2	3	3	5	5	2																																																												
	top	Female	51-64	Yes	No	No	Moderate	No	No	No	<1 Year	111 - 120	81 - 90	No	HYPERTENSION (Stage-1)																																																											
	freq	913	475	1657	984	744	697	976	940	984	625	1008	708	984	648																																																											

Univariate
Analysis

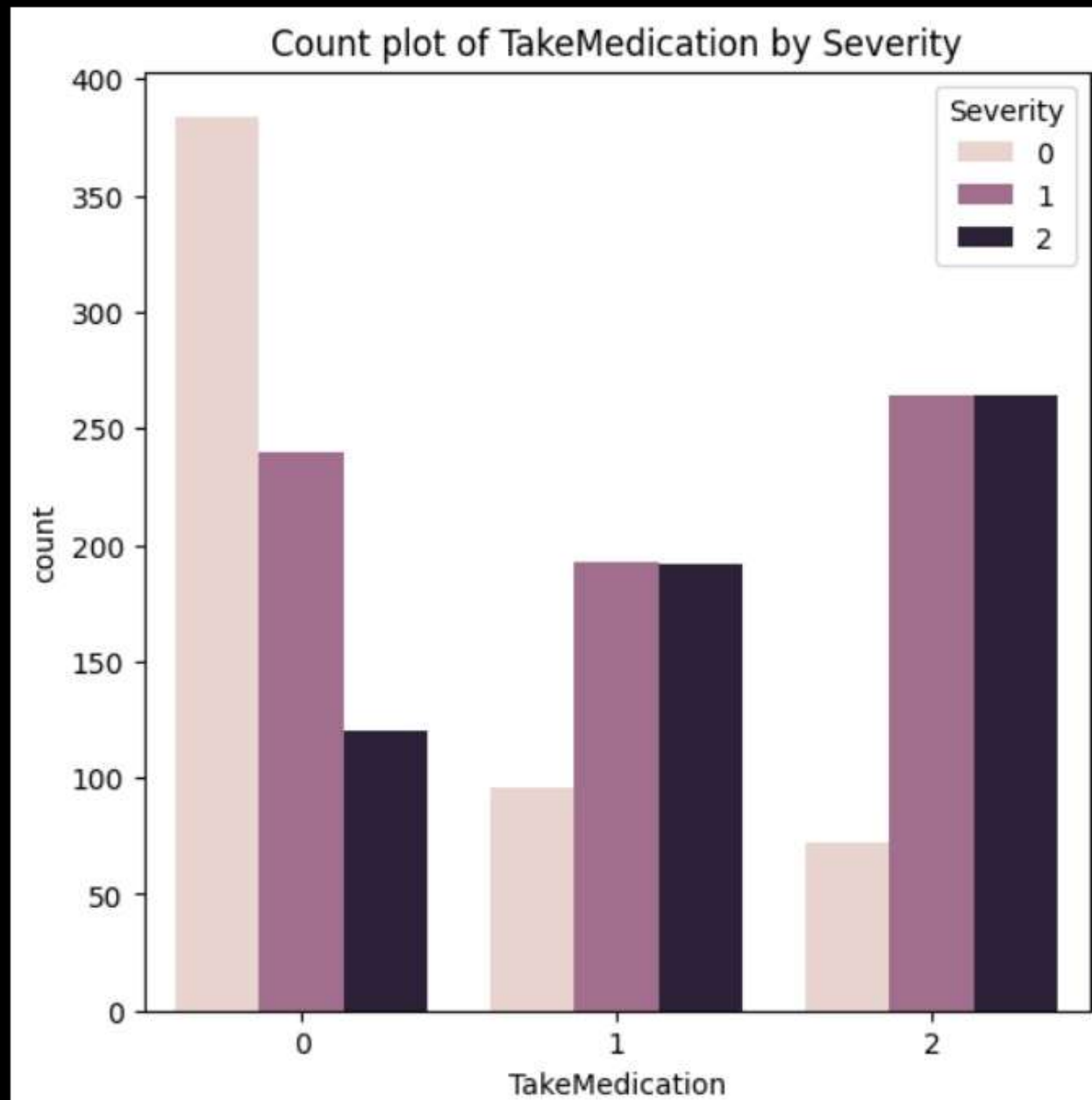
```
gender_counts = df['Gender'].value_counts()

# Plotting the pie chart
plt.pie(gender_counts, labels=gender_counts.index, autopct='%1.0f%%', startangle=140)
plt.title('Gender Distribution')
plt.axis('equal')
plt.show()
```

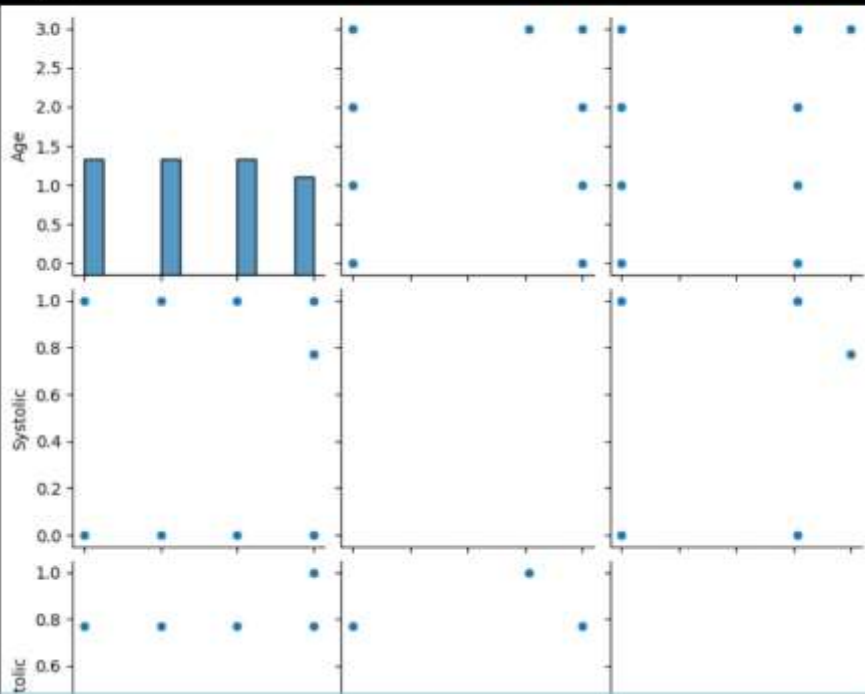


```
sns.countplot(x='TakeMedication', hue='Severity', data=df)  
plt.title('Count plot of TakeMedication by Severity')  
plt.show()
```

✓ 0.1s



Bivariate
Analysis

Multivariate Analysis	<div data-bbox="397 247 1291 1039"> <pre>sns.pairplot(df[['Age', 'Systolic', 'Diastolic']]) plt.show()</pre>  </div>
Outliers and Anomalies	-
Data Preprocessing Code Screenshots	

Loading Data

```
#Importing data
df = pd.read_csv('patient_data.csv')
```

Python

```
df.head()
```

Python

	C	Age	History	Patient	TakeMedication	Severity	BreathShortness	VisualChanges	NoseBleeding	Whendiagnosed	Systolic	Diastolic	ControlledDiet	Stages
0	Male	18-34	Yes	No	No	Mild	No	No	No	<1 Year	111 - 120	81 - 90	No	HYPERTENSION (Stage-1)
1	Female	18-34	Yes	No	No	Mild	No	No	No	<1 Year	111 - 120	81 - 90	No	HYPERTENSION (Stage-1)
2	Male	35-50	Yes	No	No	Mild	No	No	No	<1 Year	111 - 120	81 - 90	No	HYPERTENSION (Stage-1)
3	Female	35-50	Yes	No	No	Mild	No	No	No	<1 Year	111 - 120	81 - 90	No	HYPERTENSION (Stage-1)
4	Male	51-64	Yes	No	No	Mild	No	No	No	<1 Year	111 - 120	81 - 90	No	HYPERTENSION (Stage-1)

Handling
Missing Data

```
#checking for null values
df.isnull().sum()
```

```
Gender      0
Age         0
History     0
Patient     0
TakeMedication  0
Severity    0
BreathShortness  0
VisualChanges  0
NoseBleeding  0
Whendiagnosed  0
Systolic    0
Diastolic   0
ControlledDiet  0
Stages      0
dtype: int64
```

Data
Transformation

```
#converting categorical into numerical value
from sklearn.preprocessing import LabelEncoder

columns = ['Gender' , 'Severity' , 'History' , 'Patient', 'TakeMedication', 'BreathShortness',
| | | 'VisualChanges', 'NoseBleeding', 'ControlledDiet', 'Stages']

label_encoder = LabelEncoder()
for col in columns:
| df[col] = label_encoder.fit_transform(df[col])
```

Feature
Engineering

Attached the codes in final submission.

Save
Processed Data

-