Data Collection and Preprocessing Phase

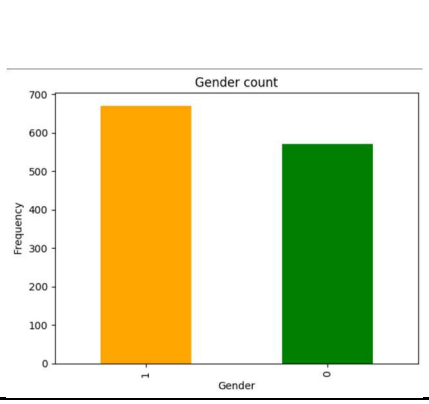
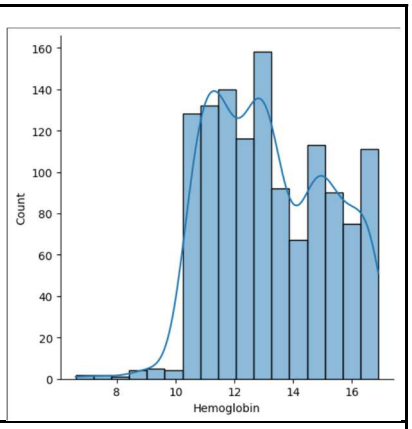
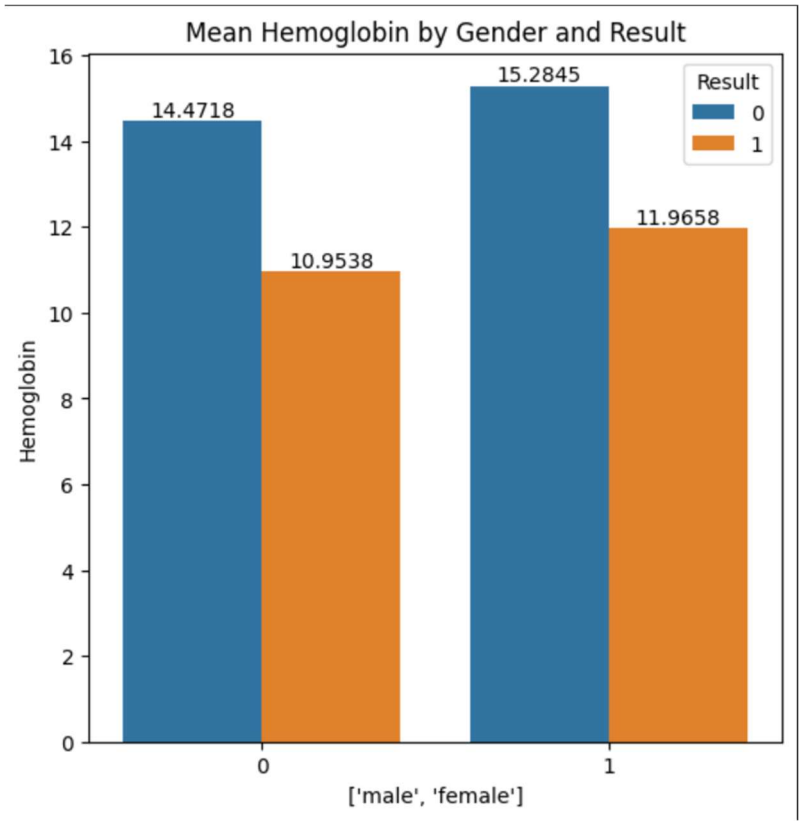
Date	24 July 2025
Team ID/ Skill Wallet ID	**SWUID20250195143**
Project Title	Anemia Sense: Leveraging Machine Learning For Precise Anemia Recognitions
Maximum Marks	6 Marks

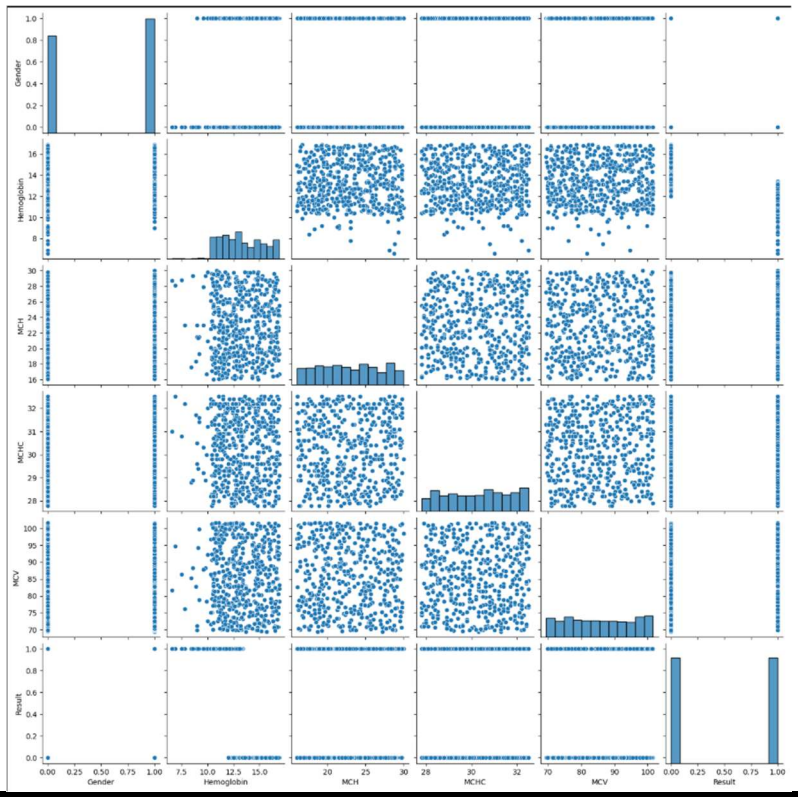
Data Exploration and Preprocessing Report

The dataset variables were statistically analyzed to understand distributions, identify patterns, and detect outliers. Python was used for preprocessing tasks, including normalization and feature selection. Data cleaning involved handling missing values and removing anomalies to ensure high-quality input for the machine learning model. These steps established a reliable foundation for accurate anemia prediction.

Section	Description
Data Overview	Dimension: 1421 rows × 6 columns Descriptive statistics: 
Univariate Analysis	

	Gender	Hemoglobin	MCH	MCHC	MCV	Resul
count	1240.000000	1240.000000	1240.000000	1240.000000	1240.000000	1240.000000
mean	0.540323	13.218145	22.903952	30.277984	85.620968	0.500000
std	0.498573	1.976190	3.993624	1.394515	9.673794	0.500202
min	0.000000	6.600000	16.000000	27.800000	69.400000	0.000000
25%	0.000000	11.500000	19.400000	29.100000	77.300000	0.000000
50%	1.000000	13.000000	22.700000	30.400000	85.300000	0.500000
75%	1.000000	14.900000	26.200000	31.500000	94.225000	1.000000
max	1.000000	16.900000	30.000000	32.500000	101.600000	1.000000

	 
Bivariate Analysis	

Multivariate Analysis	
Outliers and Anomalies	NONE

Data Preprocessing Code Screenshots

Loading Data

```
#importing the dataset which is in csv file
data = pd.read_csv('/content/Dataset/loan_prediction.csv')
data
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome
0	LP001002	Male	No	0	Graduate	No	5849	0.0
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0
4	LP001008	Male	No	0	Graduate	No	6000	0.0

<table>
<tr>
<td>Handling Missing Data</td>
<td>

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1421 entries, 0 to 1420
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Gender      1421 non-null   int64
 1   Hemoglobin  1421 non-null   float64
 2   MCH         1421 non-null   float64
 3   MCHC        1421 non-null   float64
 4   MCV         1421 non-null   float64
 5   Result      1421 non-null   int64
```

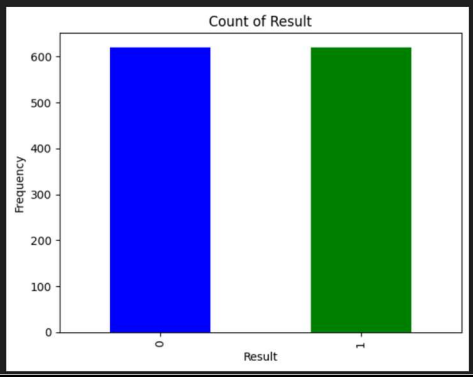
None found
</td>
</tr>
<tr>
<td>Data Transformation</td>
<td>

```
# to balance anemia count with not-anemia count

from sklearn.utils import resample

majorclass = df[df['Result'] == 0]
minorclass = df[df['Result'] == 1]

major_downsample = resample(majorclass, replace=False, n_samples=len(minorclass), random_state=42)

df = pd.concat([major_downsample, minorclass])

print(df['Result'].value_counts())
```

Python

```
Result
0    620
1    620
Name: count, dtype: int64
```

```
# to plot the new and balanced data

results = df['Result'].value_counts()
results.plot(kind='bar', color=['blue', 'green'])
plt.xlabel('Result')
plt.ylabel('Frequency')
plt.title('Count of Result')
plt.show()
```



</td>
</tr>
<tr>
<td>Feature Engineering</td>
<td>Not needed.</td>
</tr>
</table>

Save Processed Data

```
df.describe()
```

	Gender	Hemoglobin	MCH	MCHC	MCV	Result
count	1240.000000	1240.000000	1240.000000	1240.000000	1240.000000	1240.000000
mean	0.540323	13.218145	22.903952	30.277984	85.620968	0.500000
std	0.498573	1.976190	3.993624	1.394515	9.673794	0.500202
min	0.000000	6.600000	16.000000	27.800000	69.400000	0.000000
25%	0.000000	11.500000	19.400000	29.100000	77.300000	0.000000
50%	1.000000	13.000000	22.700000	30.400000	85.300000	0.500000
75%	1.000000	14.900000	26.200000	31.500000	94.225000	1.000000
max	1.000000	16.900000	30.000000	32.500000	101.600000	1.000000