Data Collection and Preprocessing Phase

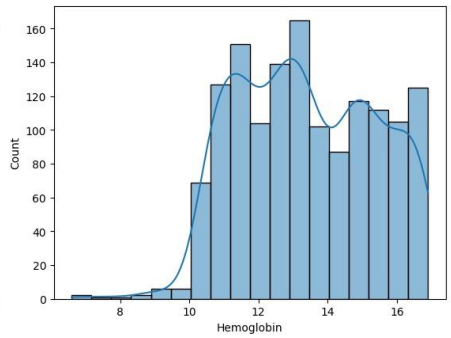
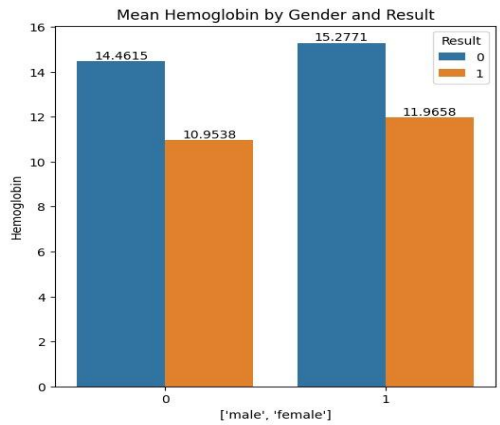
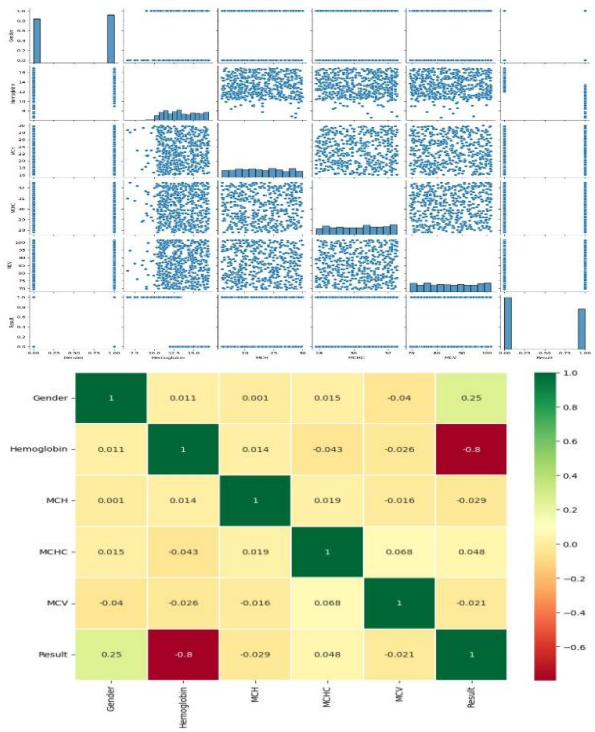
Date	03 August 2025
Project Title	Anemia Sense - Machine Learning for Precise Anemia Recognition
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

Dataset variables will be statistically analysed 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 modelling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	Dimension: 1421 rows × 6 columns Descriptive statistics: `df.describe()` (see table below)
Univariate Analysis	

	Gender	Hemoglobin	MCH	MCHC	MCV	Result
count	1421.000000	1421.000000	1421.000000	1421.000000	1421.000000	1421.000000
mean	0.520760	13.412738	22.905630	30.251232	85.523786	0.436312
std	0.499745	1.974546	3.969375	1.400898	9.636701	0.496102
min	0.000000	6.600000	16.000000	27.800000	69.400000	0.000000
25%	0.000000	11.700000	19.400000	29.000000	77.300000	0.000000
50%	1.000000	13.200000	22.700000	30.400000	85.300000	0.000000
75%	1.000000	15.000000	26.200000	31.400000	94.200000	1.000000
max	1.000000	16.900000	30.000000	32.500000	101.600000	1.000000

	 
Bivariate Analysis	
Multivariate Analysis	

Outliers and Anomalies	-
Data Preprocessing Code Screenshots	
Loading Data	`df.head()` 		Gender	Hemoglobin	MCH	MCHC	MCV	Result
---	---	---	---	---	---	---
0	1	14.9	22.7	29.1	83.7	0
1	0	15.9	25.4	28.3	72.0	0
2	0	9.0	21.5	29.6	71.2	1
3	0	14.9	16.0	31.4	87.5	0

4	1	14.7	22.0	28.2	99.5	0 |
						
						`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	
				`dtypes: float64(4), int64(2)`
				`memory usage: 66.7 KB`
				
				`df.isnull().sum()`
				
						0

Gender	0
Hemoglobin	0
MCH	0
MCHC	0
MCV	0
Result	0 | `dtype: int64`
Handling Missing Data	

Data Transformation	`x = df.drop('Result', axis = 1)`
	`x`
	`0.0s`
	
			Gender	Hemoglobin	MCH	MCHC	MCV

---	---	---	---	---	---
1234	1	16.6	18.8	28.1	70.9
1188	0	15.3	18.3	30.4	93.4
106	0	14.8	20.4	28.5	91.1
954	0	14.6	16.9	31.9	78.1
112	0	15.9	28.7	31.0	81.6
...	...	...	...	...	...
1415	1	13.2	20.1	28.8	91.2
1416	0	10.6	25.4	28.2	82.9
1417	1	12.1	28.3	30.4	86.9
1418	1	13.1	17.7	28.1	80.7
1420	0	11.8	21.2	28.4	98.1
Feature Engineering	Attached the codes in final submission.
Save Processed Data	-