3.1 Data Collection and Preprocessing Phase

Date	11 July 2024
Team ID	SWTID1720099206
Project Title	Anemia Sense: Leveraging Machine Learning For Precise Anemia Recognitions
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	Anemiasense leverages machine learning algorithms to provide precise recognition and management of anemia, a condition characterized by a deficiency of red blood cells or hemoglobin.
Data Collection Plan	The dataset was taken from the SmartInternz platfrom.
Raw Data Sources Identified	The data set comprised values of **Gender Hemoglobin MCH MCHC MCV Result** Of each patient.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset 1: Anemia.csv	Contains all the primary readings required for detecting anemia.	Link of Dataset 1	CSV	33.8kb	Public

3.2 Data Collection and Preprocessing Phase

Date	11 July 2024
Team ID	SWTID1720099206
Project Title	Anemia Sense: Leveraging Machine Learning For Precise Anemia Recognitions
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

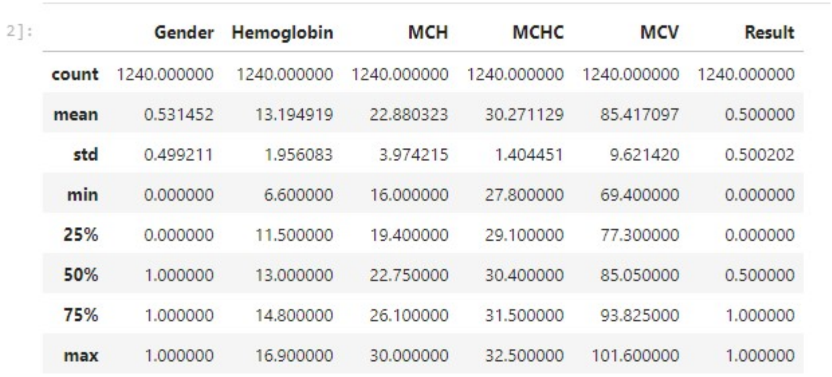
Data Source	Data Quality Issue	Severity	Resolution Plan
Dataset	Mention the issues faced in the selected dataset.	Low/ Moderate / High	Give the solution for that issue technically.
Anemia.csv	No issue	Nil	Nill

3.3 Data Collection and Preprocessing Phase

Date	11 July 2024
Team ID	SWTID1720099206
Project Title	Anemia Sense: Leveraging Machine Learning For Precise Anemia Recognitions
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

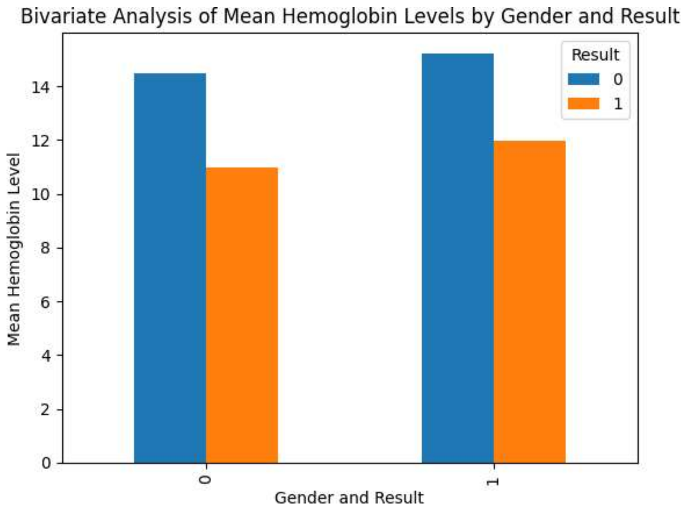
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	**Description**
Data Overview	Shape: (1421, 6)
Univariate Analysis	

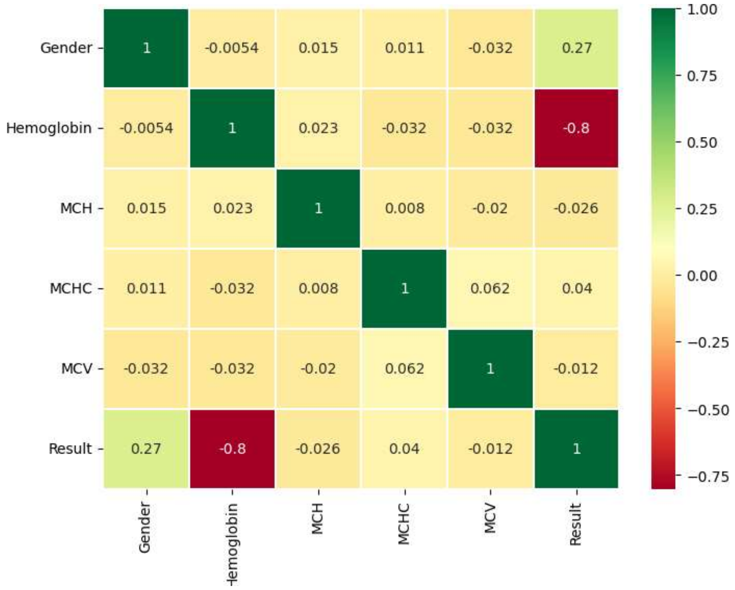
2]:

	Gender	Hemoglobin	MCH	MCHC	MCV	Result
count	1240.000000	1240.000000	1240.000000	1240.000000	1240.000000	1240.000000
mean	0.531452	13.194919	22.880323	30.271129	85.417097	0.500000
std	0.499211	1.956083	3.974215	1.404451	9.621420	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	22.750000	30.400000	85.050000	0.500000
75%	1.000000	14.800000	26.100000	31.500000	93.825000	1.000000
max	1.000000	16.900000	30.000000	32.500000	101.600000	1.000000

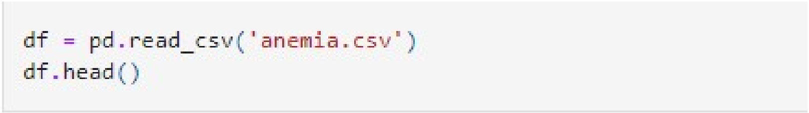
<table>
<tr>
<td>Bivariate Analysis</td>
<td>



</td>
</tr>
<tr>
<td>Multivariate Analysis</td>
<td>

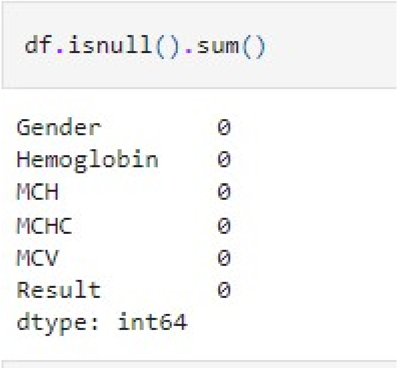


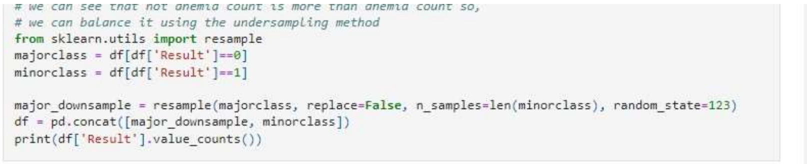
</td>
</tr>
<tr>
<td>Outliers and Anomalies</td>
<td>Identification and treatment of outliers.</td>
</tr>
<tr>
<td colspan="2"><b>Data Preprocessing Code Screenshots</b></td>
</tr>
<tr>
<td>Loading Data</td>
<td>

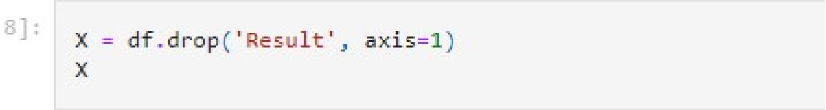


```
df = pd.read_csv('anemia.csv')
df.head()
```

</td>
</tr>
</table>

Handling Missing Data	
	`df.isnull().sum()`
	`Gender        0`
	`Hemoglobin    0`
	`MCH           0`
	`MCHC          0`
	`MCV           0`
	`Result        0`
	`dtype: int64`

Data Transformation	
	`# we can see that not anemia count is more than anemia count so,`
	`# we can balance it using the undersampling method`
	`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=123)`
	`df = pd.concat([major_downsample, minorclass])`
	`print(df['Result'].value_counts())`

Feature Engineering	
	`8]:`
	`X = df.drop('Result', axis=1)`
	`X`

Save Processed Data	Code to save the cleaned and processed data for future use.