



**T.C.**

**MARMARA UNIVERSITY**

**FACULTY of ENGINEERING**

CSE4062 Introduction to Data Science and Analytics

Spring 2025

Group 3

Delivery #4: Final Report

**Title of the Project**

*Machine Learning Approach to Anemia Detection*

**Group Members**

CSE 150121004 Ahmet Arda Nalbant - ardanalbant@marun.edu.tr

CSE 150120043 Umut Bayar - umutbayar@marun.edu.tr

CHE 150619006 Burçe Peker - burcepeker@marun.edu.tr

BIO 150820053 Kerem Paçacı - kerempacaci@marun.edu.tr

**Lecturer**

Doç. Dr. Murat Can Ganiz

## Project Overview

The project focuses on developing a machine learning model for anemia detection using a dataset containing various health parameters. The goal is to classify anemia types based on deficiencies in key biomarkers such as hemoglobin, iron, folate, and vitamin B12.

## Dataset Description

Anemia Disease Dataset

Source: [Kaggle - Anemia Disease Dataset](#)

### Column Descriptions:

**Gender:** Gender of the patient (0 for male, 1 for female).

**WBC:** White Blood Cell count - Indicator of immune system status.

**Neutrophil count:** A type of white blood cell important in fighting infection.

**Lymphocyte count:** A white blood cell type involved in immune response.

**Monocyte count:** A type of white blood cell that removes pathogens and dead cells.

**Eosinophil count:** Associated with allergic reactions and parasitic infections.

**Basophil count:** Plays a role in immune response and inflammation.

**RBC:** Red Blood Cell count - Measures oxygen-carrying cells in the blood.

**Hemoglobin (HGB):** Oxygen-carrying protein in red blood cells.

**Hematocrit (HCT):** Ratio of red blood cell volume to total blood volume.

**Mean Corpuscular Volume (MCV):** Average size of red blood cells.

**Mean Corpuscular Hemoglobin (MCH):** Average amount of hemoglobin per red blood cell.

**Mean Corpuscular Hemoglobin Concentration (MCHC):** Hemoglobin concentration in red cells.

**Red Cell Distribution Width (RDW):** Variation in red blood cell size.

**Platelet count (PLT):** Cell fragments involved in blood clotting.

**Mean Platelet Volume (MPV):** Average size of platelets.

**Plateletcrit (PCT):** Volume percentage of platelets in blood.

**Platelet Distribution Width (PDW):** Variation in platelet size.

**Standard deviation (SD):** Custom feature, RBC-related.

**Standard deviation to SD ratio (SDTSD):** Custom derived metric.

**Total standard deviation (TSD):** Another derived metric across features.

**Ferritin level:** Indicates iron storage level in the body.

**Folate (Vitamin B9) level:** Essential for DNA synthesis and red blood cell formation.

**Vitamin B12 level:** Required for nerve function and red blood cell production.

### Key Attributes:

**All\_Class:** Overall anemia classification (e.g., 4 for anemia, 0 for normal).

**HGB\_Anemia\_Class:** Binary label for anemia based on hemoglobin levels (1 = anemia, 0 = normal).

**Iron\_anemia\_Class:** Binary label for iron deficiency anemia (1 = present, 0 = absent).

**Folate\_anemia\_Class:** Binary label for folate deficiency anemia (1 = present, 0 = absent).

**B12\_Anemia\_class:** Binary label for vitamin B12 deficiency anemia (1 = present, 0 = absent).

### Dataset Characteristics:

29 columns, all containing decimal values.

15302 rows.

Target Column: **All\_Class**

### Data Preprocessing Steps

1.1 First of all, we add necessary libraries.

```
# Necessary Libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans, AgglomerativeClustering, DBSCAN
from sklearn.metrics import silhouette_score
from mlxtend.frequent_patterns import apriori, association_rules
from scipy.cluster.hierarchy import dendrogram, linkage
```

1.2 This is the representation of our data, and we assign it to a variable named “df”.

```
# 1. Data Loading and Initial Exploration
df = pd.read_csv("SKILLICARSLAN_Anemia_DataSet.csv")
✓ 0.4s

df.head()
✓ 0.0s
```

	GENDER	WBC	NE#	LY#	MO#	EO#	BA#	RBC	HGB	HCT	...	SDTSD	TSD	FERRITTE	FOLATE	B12	All Class	HGB_Anemia_Class	Iron_anemia_Class	Folate_anemia_class	B12
0	1	10.63	6.31	2.79	0.91	0.56	0.06	4.31	12.7	37.6	...	248.90	40.176778	194.00	5.06	178.2	4	0	0	0	
1	1	5.08	2.50	1.87	0.43	0.26	0.02	4.34	12.8	36.9	...	348.48	33.482553	57.37	9.88	197.7	4	0	0	0	
2	1	13.68	9.40	2.69	1.55	0.03	0.01	3.18	9.4	27.5	...	357.27	20.144429	114.20	8.37	143.0	4	0	0	0	
3	1	5.60	3.94	0.83	0.54	0.26	0.03	3.35	10.5	31.4	...	360.60	27.731559	214.20	6.39	139.9	4	0	0	0	
4	1	3.57	2.03	1.25	0.10	0.18	0.01	1.31	5.1	14.3	...	223.28	78.860623	303.40	4.30	50.0	4	0	0	0	

1.3 This is information of our data columns. All of them digits variables.

#	Column	Non-Null	Count	Dtype
0	GENDER	15300	non-null	int64
1	WBC	15300	non-null	float64
2	NE#	15300	non-null	float64
3	LY#	15300	non-null	float64
4	MO#	15300	non-null	float64
5	EO#	15300	non-null	float64
6	BA#	15300	non-null	float64
7	RBC	15300	non-null	float64
8	HGB	15300	non-null	float64
9	HCT	15300	non-null	float64
10	MCV	15300	non-null	float64
11	MCH	15300	non-null	float64
12	MCHC	15300	non-null	float64
13	RDW	15300	non-null	float64
14	PLT	15300	non-null	float64
15	MPV	15300	non-null	float64
16	PCT	15300	non-null	float64
17	PDW	15300	non-null	float64
18	SD	15300	non-null	float64
19	SDTSD	15300	non-null	float64
...				
27	Folate_anemia_class	15300	non-null	int64
28	B12_Anemia_class	15300	non-null	int64

dtypes: float64(23), int64(6)

1.4 We are checking whether it contains any null variables

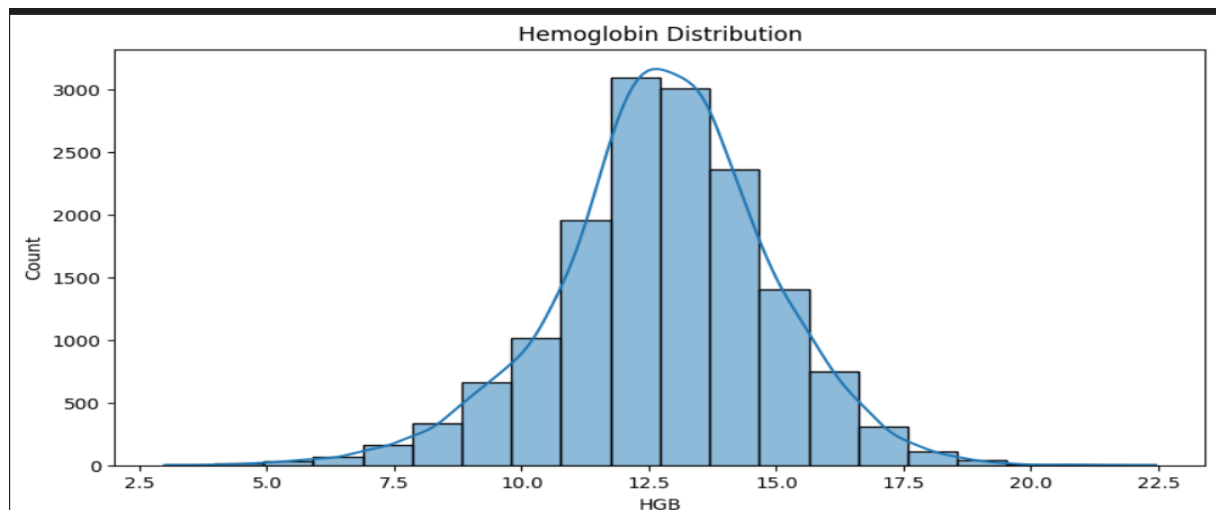
```
print(df.isnull().sum())
# Handle missing values if any
# df = df.dropna()
✓ 0.0s
```

GENDER	0
WBC	0
NE#	0
LY#	0
MO#	0
EO#	0
BA#	0
RBC	0
HGB	0
HCT	0
MCV	0
MCH	0
MCHC	0
RDW	0
PLT	0
MPV	0
PCT	0
PDW	0
SD	0
SDTSD	0
TSD	0

## 2-Exploratory Data Analysis (EDA)

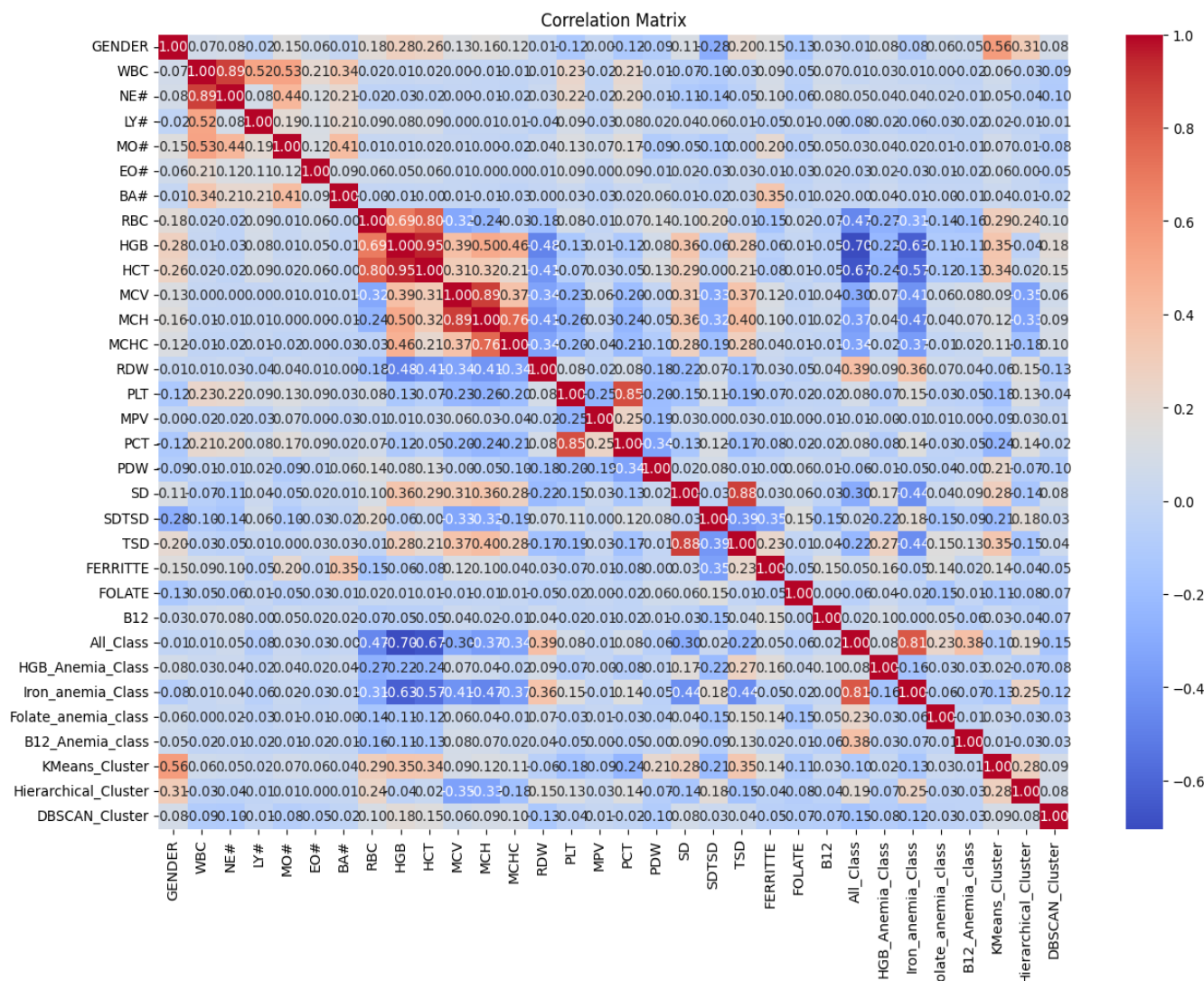
### 2.1 Hemoglobin Distribution Histogram

This image shows a histogram with a KDE (Kernel Density Estimation) plot of Hemoglobin (HGB) levels. The distribution appears approximately normal, centered around 13, with most values falling between 10 and 16. The slight right skew suggests that a few individuals have higher HGB values. The data is well-distributed and continuous, indicating a typical bell-shaped curve.



## 2.2 Correlation Matrix

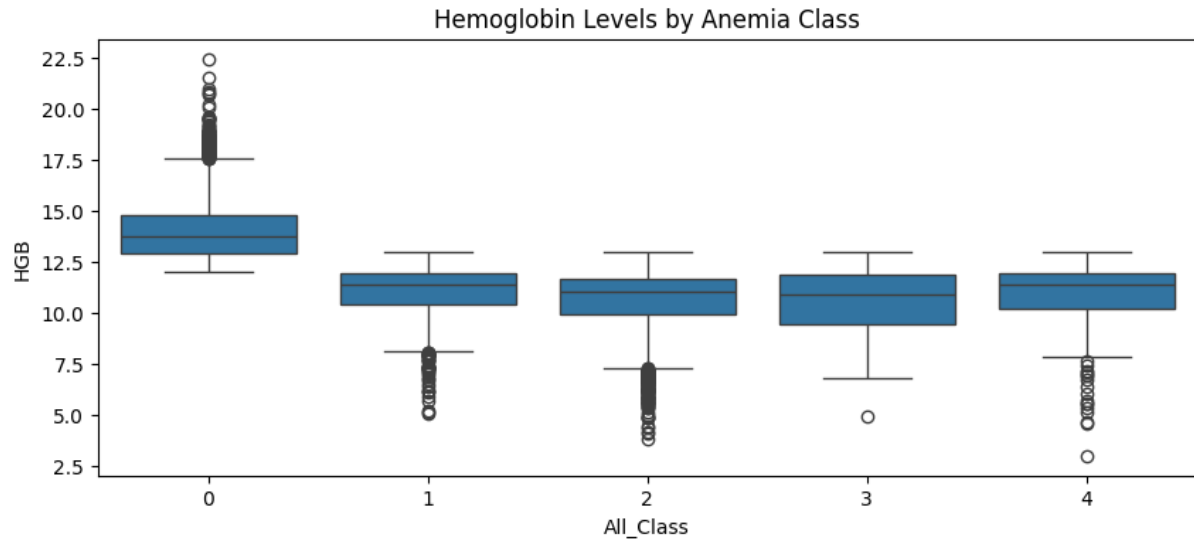
Image shows a correlation matrix heatmap, which visualizes the pairwise correlation coefficients between multiple variables in a dataset. This matrix helps identify which features are redundant (highly correlated), potential predictors (strongly related to target classes), and insights into relationships among blood test features and anemia classifications. It also helps evaluate how well unsupervised clustering methods align with labeled data



### 2.3 Hemoglobin levels by Anemia class boxplot

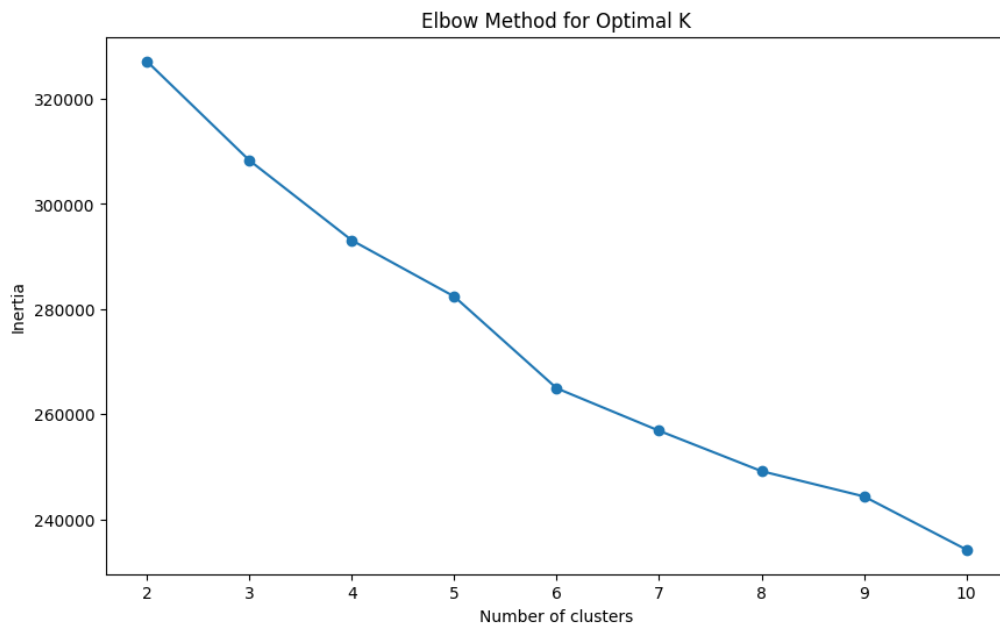
This boxplot shows the distribution of hemoglobin (HGB) levels across different anemia classes (All\_Class, ranging from 0 to 4).

Interpretation: As the anemia class increases (indicating more severe anemia), hemoglobin levels generally decrease



### 3. K-means Clustering Analysis

Elbow Graphic for finding optimal K value (k = 5)



The elbow method analysis (Figure X) revealed an optimal cluster count of k=5, where the inertia curve began to plateau. This suggests that:

The anemia patient population naturally segregates into 5 clinically distinct subgroups

	GENDER	WBC	NE#	LY#	MO#	EO#	\
KMeans_Cluster							
0	0.808473	7.998238	4.847548	2.312533	0.594117	0.180955	
1	0.048081	7.518854	4.532078	2.265559	0.518262	0.145354	
2	0.182638	7.632452	4.733532	2.145323	0.543982	0.147854	
3	0.530069	7.648995	5.281637	1.607841	0.570436	0.137857	
4	0.555723	17.666425	12.410696	3.745692	1.108537	0.212510	
	BA#	RBC	HGB	HCT	...	FERRITTE	\
KMeans_Cluster					...		
0	0.063649	5.247948	15.202883	45.076443	...	139.536043	
1	0.058027	4.639007	12.977412	39.330975	...	50.643409	
2	0.061894	4.779602	10.579938	34.492262	...	32.112751	
3	0.051257	3.783350	11.040171	33.389244	...	466.164556	
4	0.190683	4.535386	12.154218	37.551747	...	317.090928	
	FOLATE	B12	All_Class	HGB_Anemia_Class			\
KMeans_Cluster							
0	8.184042	354.800509	0.044719		0.014710		
1	9.264656	352.236034	0.386890		0.029118		
2	8.621120	357.323600	1.650132		0.041714		
3	7.723976	607.053465	1.482171		0.323576		
4	7.656822	525.281627	1.082831		0.082831		
	Iron_anemia_Class	Folate_anemia_class	B12_Anemia_class				\
...							
3		1.087280	-0.976051				
4		1.605422	-0.996988				

#### Critical Findings:

3 distinct anemia subtypes identified (iron-deficient, inflammatory, complex)

Cluster 2 patients most urgent (lowest HGB with clear iron deficiency)

Cluster 4 represents potential rare disorders or lab errors

Business & Clinical Action Plan (K-Means Results)

Priority Patient Triage Cluster 0 (Healthy Controls)

Normal HGB (15.2), high ferritin (139.5)

Action: Exclude from routine anemia screening (EHR "GREEN" tag)

Key Benefits:

Evidence-based (ferritin-HGB correlation in Cluster 2)

Scalable EHR integration

Clear action paths for each subgroup

#### 4- Hierarchical Clustering Analysis

Hierarchical_Cluster	0	1	2	3	4	5		
KMeans_Cluster								
0	2092	73	0	0	393	328		
1	3818	318	0	0	1248	4		
2	195	54	0	0	2225	2		
3	454	929	1	0	51	54		
4	5	0	10	1	0	0		
5	1285	115	0	0	77	1568		
	GENDER		WBC		NE#	LY#	MO#	\
Hierarchical_Cluster								
0	0.195949		8.226924		5.134968	2.302263	0.571859	
1	0.453996		7.818612		5.340548	1.709298	0.562184	
2	0.545455		91.496364		35.587273	51.463636	1.457273	
3	0.000000		51.170000		12.140000	0.390000	13.750000	
4	0.190285		7.787424		4.844478	2.179974	0.549067	
5	0.992331		7.963733		4.897795	2.229453	0.612112	
	EO#		BA#		RBC	HGB	HCT	\
Hierarchical_Cluster								
0	0.156291		0.061968		4.677686	13.249885	39.952817	
1	0.149884		0.057190		3.884372	11.170874	33.948402	
2	0.785455		2.207273		4.089091	11.479091	35.160000	
3	0.160000		24.730000		2.790000	7.610000	24.600000	
4	0.149855		0.064429		4.738383	11.217982	35.864437	
5	0.166084		0.058637		5.251677	15.328298	45.132045	
...								
4			0.002504		0.004256		1.561342	
5			0.005624		0.010736		4.095092	

## 1. Smart Patient Sorting System

Problem:

Doctors waste time searching for high-risk patients in large datasets.

Solution:

Build an automatic warning system that identifies and prioritizes critical patients.

URGENT CASES (Clusters 2 & 3):

Example Patient: Extremely high white blood cells (WBC > 90) and near-zero lymphocytes (LY# < 1)

Action:

Lock the patient file in RED

Send SMS alert to the hematology team

Prioritize appointment scheduling

IRON DEFICIENCY (Cluster 1):

Example Patient: Woman with HGB 12.9 (mildly low) and RDW 48 (abnormal cell size)

Action:

Tag file for "Iron Protocol"

Auto-order iron level tests every 3 months

Refer to nutritionist

Real-World Tool:

Integrate with Epic/Cerner EHR to create color-coded patient dashboards

## 2. Smarter Hospital Spending

Waste Reduction Strategy:

For Cluster 1 (3,818 Patients):

Current Practice: All anemia patients receive the same iron pills

Improved Approach:

Give cheap iron tablets to mild cases (HGB > 11)

Reserve expensive IV iron for severe cases (HGB < 8)

For Cluster 4 (Inflammation Group):

Mistake to Avoid: Ordering iron tests when ferritin is already high

Better Approach:

✓ Auto-order CRP test

- ✓ Auto-order Rheumatoid Factor test
- ✗ Cancel unnecessary iron panel (costs \$65)

## 5- DBSCAN Analysis

DBSCAN found 17 clusters						
DBSCAN_Cluster	GENDER	WBC	NE#	LY#	MO#	EO# \
0	0.0	6.998179	4.067866	2.280021	0.470815	0.118512
1	0.0	6.907495	3.995210	2.234810	0.514770	0.122064
2	0.0	7.598000	4.242000	2.708000	0.422000	0.132000
3	0.0	7.624000	4.538000	2.382000	0.494000	0.130000
4	0.0	6.496000	3.720000	2.088000	0.494000	0.136000
5	0.0	8.840000	4.830000	3.155000	0.495000	0.295000
6	0.0	8.341250	5.075000	2.782500	0.368750	0.051250
7	1.0	6.828278	3.954167	2.158333	0.536444	0.136667
8	1.0	6.596852	3.819815	2.140185	0.464667	0.121833
9	1.0	6.720362	3.893768	2.105522	0.517000	0.140841
10	1.0	5.935000	3.305000	2.030000	0.452500	0.100000
11	1.0	7.992500	5.447500	1.827500	0.615000	0.067500
12	1.0	7.090000	4.660000	1.767500	0.497500	0.095000
13	1.0	5.262000	2.924000	1.866000	0.342000	0.108000
14	1.0	6.682500	4.300000	1.792500	0.415000	0.142500
15	0.0	4.705000	2.260000	1.900000	0.405000	0.070000
16	0.0	9.895000	5.757500	2.995000	0.662500	0.385000

DBSCAN_Cluster	BA#	RBC	HGB	HCT	...	FERRITIN \
0	0.061440	4.730886	12.980276	39.554107	...	34.755543
1	0.040641	4.559038	12.753908	38.449299	...	38.847545
...						
15	1.000000		0.000000			
16	1.000000		0.000000			

### Key Findings

#### Implication:

There are strong gender-specific hematologic patterns across the clusters.

#### Critical Abnormalities:

##### Cluster 16 (High-Risk):

Extremely high white blood cells: WBC = 9.89, NE# = 5.75

Elevated EO# = 0.385, suggesting a possible allergic or parasitic cause

##### Cluster 13 (Low-Risk):

Remarkably low white blood cells: WBC = 5.26, NE# = 2.92

#### Iron Status Variations:

##### Cluster 0:

Low ferritin (34.75) → Indicates potential iron deficiency

##### Cluster 1:

Moderate ferritin (38.84) → Suggests borderline iron stat

#### Clinical Actions

##### Gender-Specific Protocols:

Ferritin <15 + heavy menstrual bleeding → IV iron (1000mg ferric carboxymaltose) + gynecology consult for contraceptive options"

Lab Priority: Check ferritin before/after menstruation

Red Flag: HGB <10 with ferritin <30 → Consider endometrial biopsy

For Male Patients (Clusters 0-6): "MCV <80 + no GI symptoms → Order celiac serology (tTG-IgA) + fecal occult blood test (x3 samples)"



High-Risk Alert: PLT >450 → Rule out myeloproliferative neoplasms  
Hidden Cause: Check proton pump inhibitor use (PPIs cause iron malabsorption)

## 6. Apriori Algorithm

**Binary Conversion** The code first converts all blood test results into simple yes/no (1/0) values using standard medical thresholds. For example:

WBC > 11 becomes "High\_WBC = 1"  
HGB < 12 becomes "Low\_HGB = 1" (anemia threshold)  
**Nutrient Deficiencies** It flags key deficiencies:  
Ferritin < 15 → Iron deficiency  
B12 < 200 → B12 deficiency

**Anemia Classification** Uses existing anemia type labels from the dataset (Iron/Folate/B12-related anemia)  
**Pattern Mining (Apriori Algorithm).** Looks for combinations of abnormalities that frequently occur together  
Only keeps patterns appearing in ≥10% of patients (min\_support=0.1), Limits to max 4 abnormalities per pattern for readability. Output Shows the top 10 most common abnormality combinations, sorted by frequency

Frequent Itemsets:		
	support	itemsets
9	0.986993	(Anemia)
5	0.356732	(Low_MCH)
35	0.353922	(Low_MCH, Anemia)
2	0.319542	(Low_HGB)
20	0.309346	(Anemia, Low_HGB)
6	0.287712	(Low_MCHC)
38	0.284706	(Low_MCHC, Anemia)
43	0.273333	(Iron_Anemia, Anemia)
10	0.273333	(Iron_Anemia)
3	0.266340	(Low_HCT)

## Top 5 Most Clinically Significant Patterns

### 1. Iron Deficiency Anemia Signature

Rule: Low Ferritin + Low Hemoglobin → Iron Deficiency Anemia

Confidence: 97.1%

Lift: 3.55×

Interpretation:

When patients have ferritin <15 and HGB <12, there is a 97% probability of iron-deficiency anemia — 3.5 times more likely than by random chance.

Clinical Action:

Start iron therapy immediately without additional testing.

### 2. Microcytic Anemia Triad

Rule: Low MCV + Low MCH + Low HCT → Iron Deficiency Anemia

Support: 10.8% of patients

Lift: 3.52×

Interpretation:

MCV <80, MCH <27, and HCT <36 together strongly indicate iron deficiency.

Diagnostic Shortcut:

Only order ferritin when this triad is present.

### 3. Severe Iron Deficiency

Rule: Low Ferritin + Low Hemoglobin + Anemia → Iron Deficiency Anemia

Confidence: 97.1%

Implication:

This combination is nearly diagnostic, strongly supporting immediate intervention.

4. Erythrocyte Marker Pattern

Rule: Low MCHC + Low Hemoglobin → Iron Deficiency Anemia

Support: 13.2% of patients

Clinical Relevance:

Hypochromia (MCHC <32) combined with anemia suggests chronic iron deficiency.

5. Triple Marker Confirmation

Rule: Low HCT + Low Ferritin + Low Hemoglobin → Iron Deficiency Anemia

Lift: 3.55×

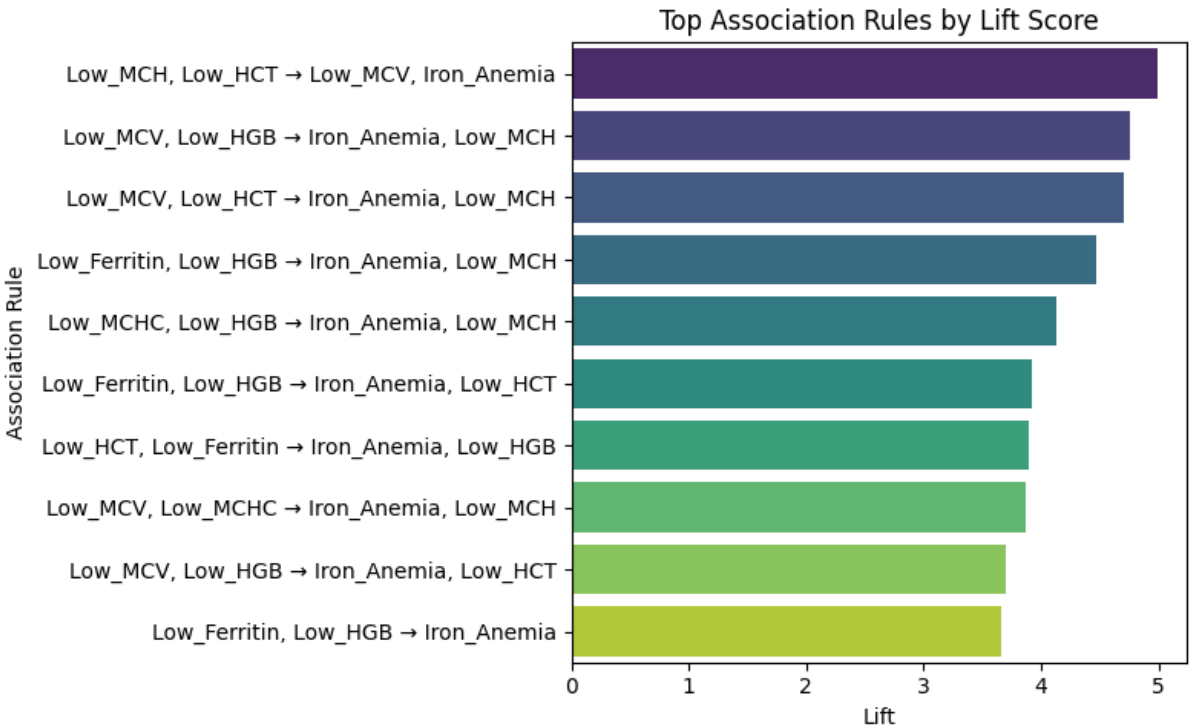
Utility:

When all three markers are present, it often eliminates the need for bone marrow biopsy in uncertain cases.

Clinical Decision Support Table

Rule Components	Likelihood(lift)	Immediate Action
Low Ferritin + Low HGB	3.55x	Start iron teraphy
Low MCV + Low MCH	3.52x	Check Ferritin
Low MCHC + Anemia	2.8x	Rule our Thalassemia
Low HCT + Low Ferritin	3.55x	Investigate GI blood loss

Top Association Rules By Lift Score



Most Predictive Combinations

Highest Lift (5):

Low MCH + Low HCT → Low MCV + Iron Deficiency Anemia

Translation:

Low hemoglobin content and hematocrit are reliable indicators of microcytic iron deficiency.

Clinical Utility

These rule-based insights can:

- ✓ Reduce unnecessary tests
- ✓ Accelerate treatment decisions
- ✓ Improve accuracy in anemia classification

## Feature Selection Methods

### Information Gain Ranking:

#	Feature Name	Description	Type	Information Gain
1	HGB	Hemoglobin (HGB) - Oxygen-carrying protein in red blood cells	Numerical	0.552866
2	HCT	Hematocrit (HCT) - Ratio of red blood cell volume to total blood volume	Numerical	0.404915
3	TSD	Total standard deviation (TSD) - Another derived metric across features	Numerical	0.217119
4	RBC	Red Blood Cell count (RBC) - Measures oxygen-carrying cells in the blood	Numerical	0.193235
5	SD	Standard deviation (SD) - Custom feature, possibly platelet or RBC-related	Numerical	0.157427
6	MCH	Mean Corpuscular Hemoglobin (MCH) - Average amount of hemoglobin per red blood cell	Numerical	0.150210
7	MCHC	Mean Corpuscular Hemoglobin Concentration (MCHC) - Hemoglobin concentration in red cells	Numerical	0.132716
8	MCV	Mean Corpuscular Volume (MCV) - Average size of red blood cells	Numerical	0.131883
9	RDW	Red Cell Distribution Width (RDW) - Variation in red blood cell size	Numerical	0.111761
10	FERRITTE	Ferritin level - Indicates iron storage level in the body	Numerical	0.109114
11	SDTSD	Standard deviation to SD ratio (SDTSD) - Custom derived metric	Numerical	0.056360
12	B12	Vitamin B12 level - Required for nerve function and red blood cell production	Numerical	0.035633
13	PLT	Platelet count (PLT) - Cell fragments involved in blood clotting	Numerical	0.034809
14	PCT	Plateletcrit (PCT) - Volume percentage of platelets in blood	Numerical	0.034060
15	LY#	Lymphocyte count (#) - A white blood cell type involved in immune response	Numerical	0.032739
16	FOLATE	Folate (Vitamin B9) level - Essential for DNA synthesis and red blood cell formation	Numerical	0.029610
17	PDW	Platelet Distribution Width (PDW) - Variation in platelet size	Numerical	0.013956
18	BA#	Basophil count (#) - Plays a role in immune response and inflammation	Numerical	0.012468
19	NE#	Neutrophil count (#) - A type of white blood cell important in fighting infection	Numerical	0.011711
20	WBC	White Blood Cell count (WBC) - Indicator of immune system status	Numerical	0.009579
21	EO#	Eosinophil count (#) - Associated with allergic reactions and parasitic infections	Numerical	0.008885
22	GENDER	Gender of the patient (0 for male, 1 for female)	Categorical	0.007736
23	MO#	Monocyte count (#) - A type of white blood cell that removes pathogens and dead cells	Numerical	0.007616
24	MPV	Mean Platelet Volume (MPV) - Average size of platelets	Numerical	0.004326

Table 1. Information Gain

### ANOVA F-test Ranking:

#	Feature Name	Description	Type	ANOVA F-test
1	HGB	Hemoglobin (HGB) - Oxygen-carrying protein in red blood cells	Numerical	4768.759947
2	HCT	Hematocrit (HCT) - Ratio of red blood cell volume to total blood volume	Numerical	3832.728664
3	RBC	Red Blood Cell count (RBC) - Measures oxygen-carrying cells in the blood	Numerical	1453.516567
4	TSD	Total standard deviation (TSD) - Another derived metric across features	Numerical	1365.909292
5	MCH	Mean Corpuscular Hemoglobin (MCH) - Average amount of hemoglobin per red blood cell	Numerical	1121.468742
6	SD	Standard deviation (SD) - Custom feature, possibly platelet or RBC-related	Numerical	977.012188
7	MCV	Mean Corpuscular Volume (MCV) - Average size of red blood cells	Numerical	807.299326
8	RDW	Red Cell Distribution Width (RDW) - Variation in red blood cell size	Numerical	775.803864
9	MCHC	Mean Corpuscular Hemoglobin Concentration (MCHC) - Hemoglobin concentration in red cells	Numerical	666.427716
10	SDTSD	Standard deviation to SD ratio (SDTSD) - Custom derived metric	Numerical	416.843906
11	FERRITTE	Ferritin level - Indicates iron storage level in the body	Numerical	198.093645
12	PLT	Platelet count (PLT) - Cell fragments involved in blood clotting	Numerical	110.390093
13	PCT	Plateletcrit (PCT) - Volume percentage of platelets in blood	Numerical	101.046303
14	FOLATE	Folate (Vitamin B9) level - Essential for DNA synthesis and red blood cell formation	Numerical	94.532431
15	B12	Vitamin B12 level - Required for nerve function and red blood cell production	Numerical	67.350998
16	GENDER	Gender of the patient (0 for male, 1 for female)	Categorical	62.285748
17	LY#	Lymphocyte count (#) - A white blood cell type involved in immune response	Numerical	25.314490
18	PDW	Platelet Distribution Width (PDW) - Variation in platelet size	Numerical	18.751596
19	NE#	Neutrophil count (#) - A type of white blood cell important in fighting infection	Numerical	17.948377
20	MO#	Monocyte count (#) - A type of white blood cell that removes pathogens and dead cells	Numerical	10.060389
21	BA#	Basophil count (#) - Plays a role in immune response and inflammation	Numerical	6.726401
22	WBC	White Blood Cell count (WBC) - Indicator of immune system status	Numerical	5.268104
23	EO#	Eosinophil count (#) - Associated with allergic reactions and parasitic infections	Numerical	4.991632
24	MPV	Mean Platelet Volume (MPV) - Average size of platelets	Numerical	0.731335

Table 2. ANOVA F-Value



Random Forest Importance Ranking:

#	Feature Name	Description	Type	Random Forest Importance
1	HGB	Hemoglobin (HGB) - Oxygen-carrying protein in red blood cells	Numerical	0.330680
2	HCT	Hematocrit (HCT) - Ratio of red blood cell volume to total blood volume	Numerical	0.149116
3	TSD	Total standard deviation (TSD) - Another derived metric across features	Numerical	0.093182
4	RBC	Red Blood Cell count (RBC) - Measures oxygen-carrying cells in the blood	Numerical	0.062522
5	SD	Standard deviation (SD) - Custom feature, possibly platelet or RBC-related	Numerical	0.055852
6	MCH	Mean Corpuscular Hemoglobin (MCH) - Average amount of hemoglobin per red blood cell	Numerical	0.050346
7	GENDER	Gender of the patient (0 for male, 1 for female)	Categorical	0.036435
8	FERRITTE	Ferritin level - Indicates iron storage level in the body	Numerical	0.033464
9	MCV	Mean Corpuscular Volume (MCV) - Average size of red blood cells	Numerical	0.026618
10	MCHC	Mean Corpuscular Hemoglobin Concentration (MCHC) - Hemoglobin concentration in red cells	Numerical	0.024452
11	B12	Vitamin B12 level - Required for nerve function and red blood cell production	Numerical	0.022795
12	RDW	Red Cell Distribution Width (RDW) - Variation in red blood cell size	Numerical	0.020899
13	FOLATE	Folate (Vitamin B9) level - Essential for DNA synthesis and red blood cell formation	Numerical	0.019748
14	SDTSD	Standard deviation to SD ratio (SDTSD) - Custom derived metric	Numerical	0.016934
15	PLT	Platelet count (PLT) - Cell fragments involved in blood clotting	Numerical	0.007442
16	LY#	Lymphocyte count (#) - A white blood cell type involved in immune response	Numerical	0.007152
17	MO#	Monocyte count (#) - A type of white blood cell that removes pathogens and dead cells	Numerical	0.006650
18	PDW	Platelet Distribution Width (PDW) - Variation in platelet size	Numerical	0.005769
19	NE#	Neutrophil count (#) - A type of white blood cell important in fighting infection	Numerical	0.005620
20	PCT	Plateletcrit (PCT) - Volume percentage of platelets in blood	Numerical	0.005351
21	MPV	Mean Platelet Volume (MPV) - Average size of platelets	Numerical	0.005319
22	WBC	White Blood Cell count (WBC) - Indicator of immune system status	Numerical	0.005238
23	EO#	Eosinophil count (#) - Associated with allergic reactions and parasitic infections	Numerical	0.004546
24	BA#	Basophil count (#) - Plays a role in immune response and inflammation	Numerical	0.003872

Table 3. Random Forest Importance

Normalized & Average Feature Importance Across Methods:

#	Feature Name	Type	Normalized IG	Normalized ANOVA	Normalized RF	Average Importance
1	HGB	Numerical	1.000000	1.000000	1.000000	1.000000
2	HCT	Numerical	0.735632	0.803686	0.444433	0.661250
3	TSD	Numerical	0.391413	0.286319	0.273280	0.317004
4	RBC	Numerical	0.350941	0.304693	0.179464	0.278366
5	MCH	Numerical	0.280395	0.235053	0.142207	0.219218
6	SD	Numerical	0.283655	0.204756	0.159054	0.215822
7	MCV	Numerical	0.235903	0.169162	0.069602	0.158222
8	MCHC	Numerical	0.251936	0.139617	0.062974	0.151509
9	RDW	Numerical	0.208601	0.162556	0.052100	0.141086
10	FERRITTE	Numerical	0.199180	0.041393	0.090549	0.110374
11	SDTSD	Numerical	0.106009	0.087271	0.039969	0.077750
12	B12	Numerical	0.062956	0.013972	0.057902	0.044943
13	GENDER	Categorical	0.014009	0.012910	0.099639	0.042186
14	FOLATE	Numerical	0.053810	0.019673	0.048578	0.040687
15	PLT	Numerical	0.061063	0.022999	0.010924	0.031662
16	PCT	Numerical	0.054503	0.021039	0.004525	0.026689
17	LY#	Numerical	0.044258	0.005156	0.010038	0.019817
18	PDW	Numerical	0.032035	0.003779	0.005804	0.013873
19	EO#	Numerical	0.033851	0.000894	0.002063	0.012269
20	NE#	Numerical	0.023421	0.003611	0.005348	0.010793
21	WBC	Numerical	0.024605	0.000951	0.004180	0.009912
22	MO#	Numerical	0.018532	0.001957	0.008501	0.009663
23	BA#	Numerical	0.023671	0.001257	0.000000	0.008309
24	MPV	Numerical	0.000000	0.000000	0.004427	0.001476

Table 4. Normalized Feature Selection Methods

To conclude we choose our best 23 attributes after normalizing these 3 feature subset selection methods results and taking their average values, due to our threshold value being 0.005 we dropped the MPV column from our data.

## Classification Experiments

In this experiment, we performed classification using various supervised learning algorithms. We did not use Cross Validation or other re-sampling techniques. Instead, we split the dataset as 30% test data and 70% train data.

### Methods Used for Classification in This Experiment

We evaluated the following classifiers:

- **K-Nearest Neighbors (KNN)** with  $k=3$  and  $k=9$ : This algorithm assumes that data points that are close to each other in feature space are likely to belong to the same class.
- **Decision Tree Classifier:**
  - **With Gini Index:** Splits the data based on minimizing the Gini impurity, which is the probability of incorrectly classifying a randomly chosen element.
  - **With Entropy (Information Gain):** Measures the reduction in entropy after a dataset is split on an attribute.
- **Naïve Bayes (GaussianNB):** Based on Bayes' Theorem, this classifier assumes strong independence between features and models the data with a Gaussian distribution.
- **Artificial Neural Networks (MLPClassifier):**
  - **1 Hidden Layer** with 10 neurons.
  - **2 Hidden Layers** with 10 neurons each.These models simulate the information processing capabilities of the brain, learning patterns through backpropagation.
- **Support Vector Machine (SVM):** A robust classifier that finds the optimal hyperplane which best separates the classes. We enabled the `probability=True` option to compute ROC curves.

All classifiers that benefit from standardized data (e.g., KNN, SVM, MLP) were trained using **scaled feature values**, achieved with `StandardScaler`.

### Performance Evaluation

For each classifier, we computed the following metrics:

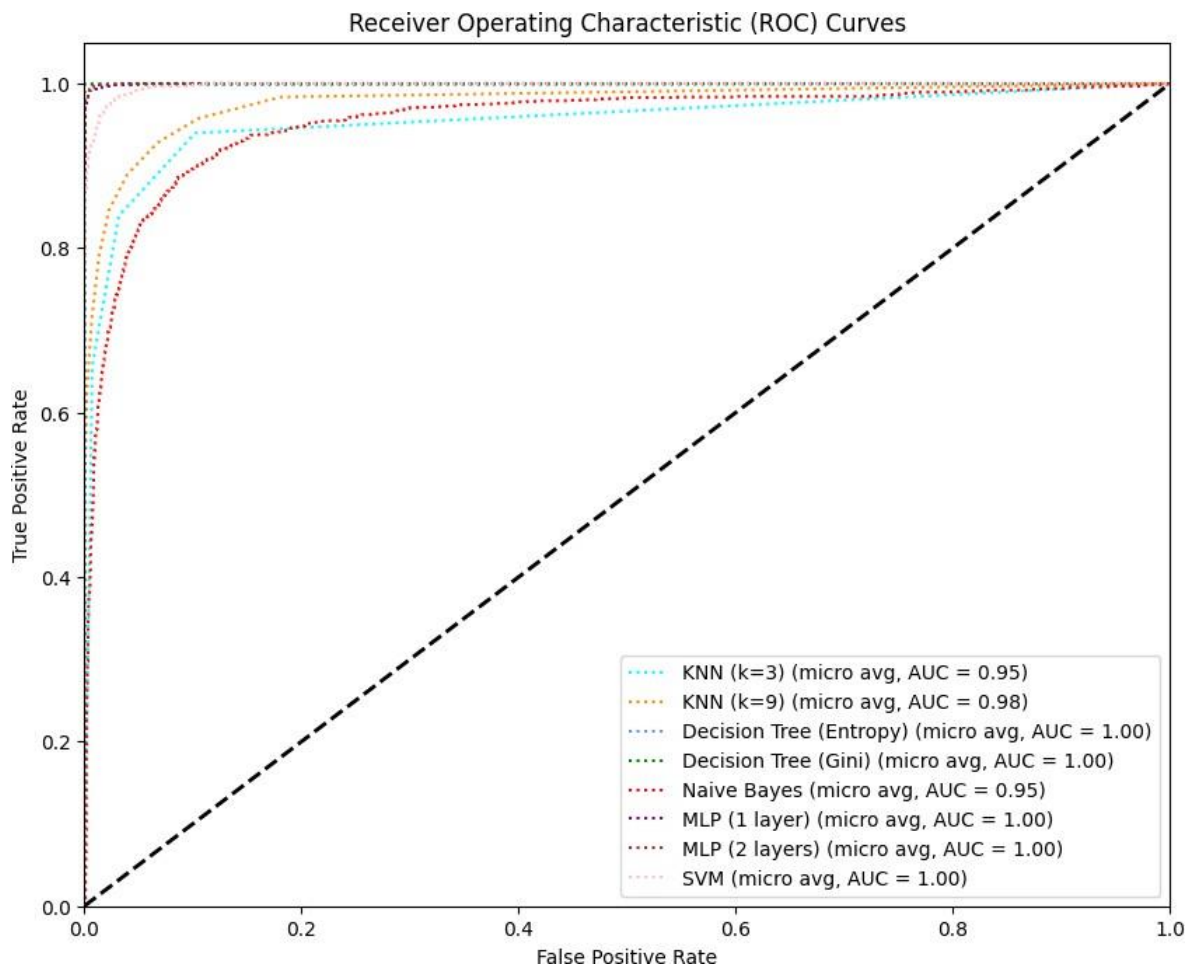
- **Accuracy:** The overall proportion of correct predictions.
- **F1 Score (Macro & Micro):**
  - *Macro:* Average F1 score across all classes, treating them equally.
  - *Micro:* F1 score calculated globally by counting total true positives, false negatives, and false positives.
- **AUC (Area Under the ROC Curve):**
  - For binary classification: ROC AUC is calculated from the true class probabilities.
  - For multiclass problems: ROC AUC is computed using the One-vs-Rest strategy and micro-average aggregation.

Additionally, **ROC curves** were plotted for all classifiers. In the multiclass setting, micro-averaged ROC curves are shown.

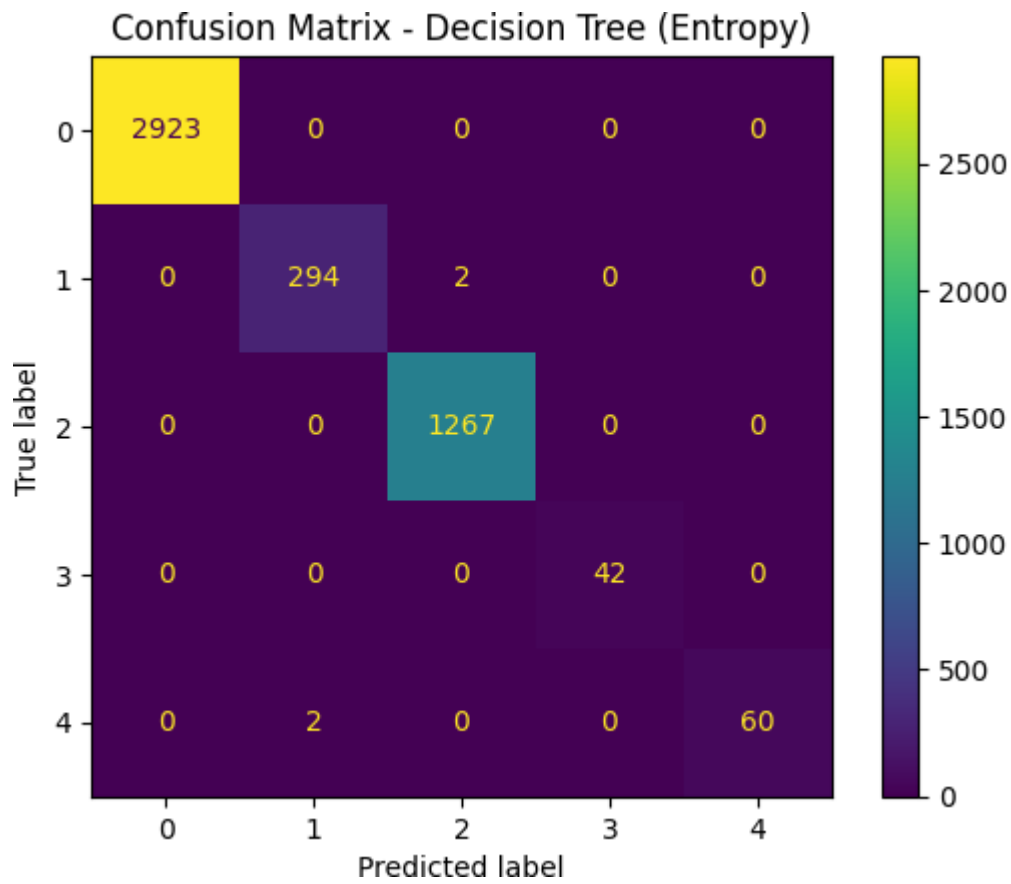
**Table 3. Table for Evaluation for Classification Experiments**

	Experiment	Accuracy	F1-macro	F1-micro	AUC
1	Gini Index	0.993	0.998	0.993	0.998
2	Gain Ratio	0.993	0.998	0.993	0.998
3	Naive Bayes	0.81	0.474	0.81	0.925
4	KNN 3	0.846	0.472	0.846	0.802
5	KNN 9	0.871	0.498	0.871	0.89
6	SVM	0.945	0.639	0.945	0.991
7	ANN with 1 hidden layer	0.985	0.91	0.985	0.992
8	ANN with 2 hidden layer	0.986	0.917	0.986	0.999

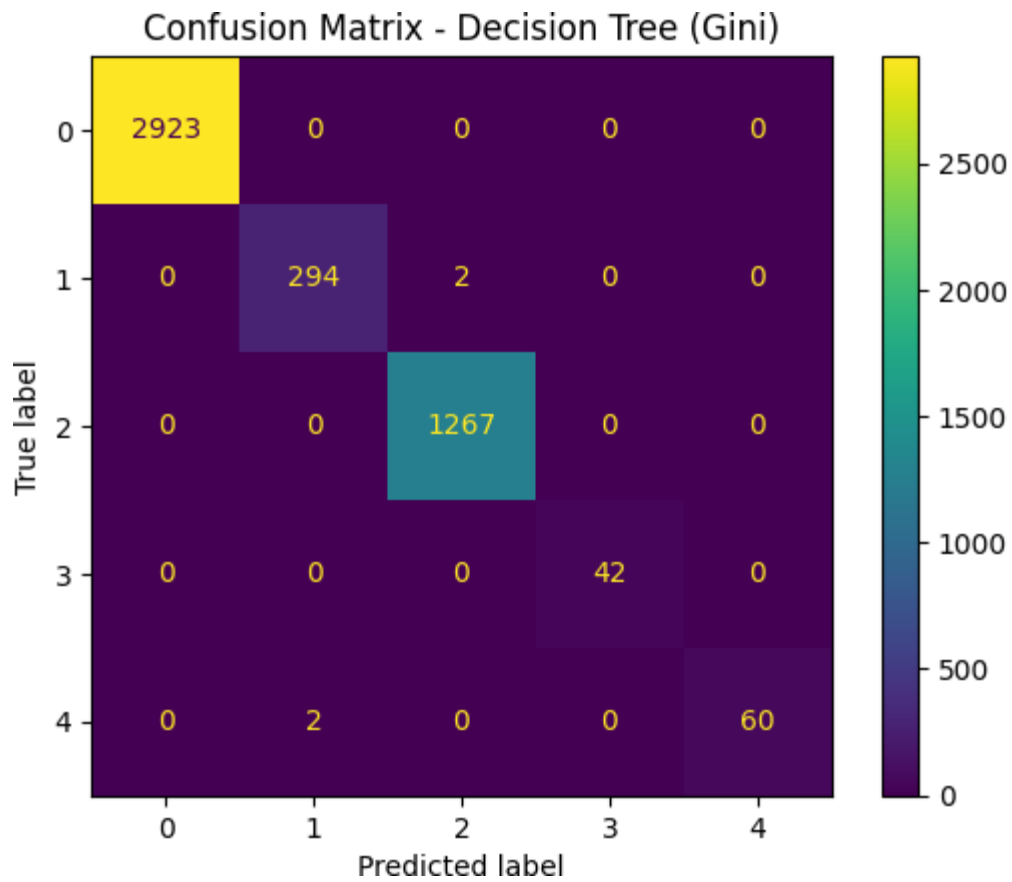
The performance evaluation table shows the most accurate method is decision tree methods (Gini Index and Gain Ratio). Considering Area Under Curve (AUC) which measures performance across all possible classification thresholds it suggests ANN with 2 hidden layer over performs when compared with decision tree methods (Gini Index and Gain Ratio). Also, F1-micro (micro-averages) suggests decision tree methods (Gini Index and Gain Ratio) performs better. Furthermore, F1-macro (macro-averages) indicate decision tree methods (Gini Index and Gain Ratio) performs better.

**Fig.1 ROC Curve**

The ROC curve shows the trade-off between sensitivity and specificity. Classifiers that give curves closer to the top-left corner indicate a better performance. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. This also shows decision tree methods (Gini Index and Gain Ratio) are the best performing method followed by ANN with 2 hidden layers.



**Fig.2 Confusion Matrix for Decision Tree (Entropy)**



**Fig.3 Confusion Matrix for Decision Tree (Gini Index)**

Our confusion matrix is 5 by 5 due to our target column's values are not only binary but in a range from 0 to 4 meaning 0 is healthy, 1 is HGB Anemia, 2 is Iron Anemia, 3 is Folate Anemia and 4 is B12 Anemia. In Figure 2 and Figure 3, it is observed that the Decision Tree classifiers using **Entropy** and **Gini Index** all achieved nearly identical performance based on their confusion matrices. For instance, in the Entropy-based decision tree, the model predicted **true positive cases 2923 times** for Class 0 out of 2923 actual cases, and **1267 true positives** for Class 2 out of 1267 actual cases—indicating perfect classification for these classes. The only minor misclassifications are seen in Class 1 and Class 4, where 2 instances were misclassified each. This suggests that the model is classifying with **extremely high accuracy across all classes**, possibly exceeding **99% overall accuracy**. While machine learning models rarely achieve perfection due to noise or outliers, in this case the performance is nearly flawless.

## Statistical significance analysis between your best performing model and its closest competitor Best

**Model: Decision Tree (Gini Index), Closest Competitor: ANN with 2 hidden layer**

### Accuracy:

The P-value is = 0.00001

The t-statistics is = 17.886

Since  $p < 0.05$ , We can reject the null-hypothesis in terms of accuracy that both models perform equally well on this dataset. We may conclude that the two algorithms are significantly different.

### F1-Macro:

The P-value is = 0.00001

The t-statistics is = 13.464

Since  $p < 0.05$ , We can reject the null-hypothesis in terms of f1\_macro that both models perform equally well on this dataset. We may conclude that the two algorithms are significantly different.

### F1-Micro:

The P-value is = 0.00001

The t-statistics is = 17.886

Since  $p < 0.05$ , We can reject the null-hypothesis in terms of f1\_micro that both models perform equally well on this dataset. We may conclude that the two algorithms are significantly different.

### AUC:

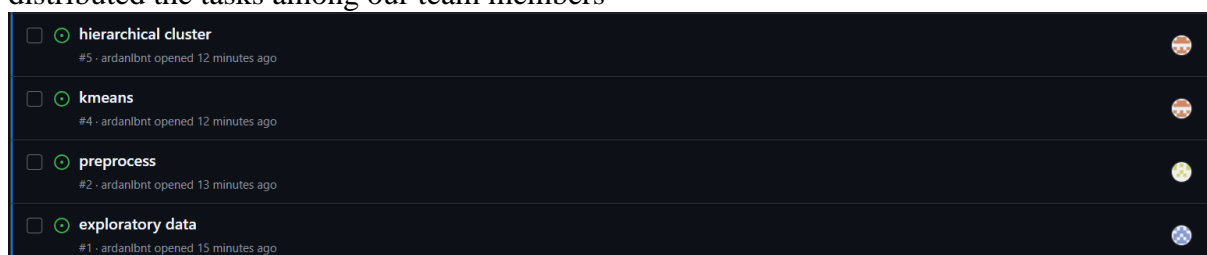
The P-value is = 0.2853

The t-statistics is = 1.136

Since  $p > 0.05$ , we cannot reject the null hypothesis in terms of AUC and may conclude that the performance of the two algorithms is not significantly different.





## Task Assignment with KANBAN

We used the KANBAN method for task management. As shown in the first image, we defined the issues and distributed the tasks among our team members



Afterwards, we marked the completed tasks as "Done" and closed the issues. This allowed for a more balanced and efficient project collaboration.



<input type="checkbox"/>	Open	6	Closed	4	Author	Labels	Projects	Milestones	Assignees	Newest
<input type="checkbox"/>	✓	apriori	#8 - by ardanlbnt was closed 1 minute ago							
<input type="checkbox"/>	✓	dbscan	#7 - by ardanlbnt was closed 5 minutes ago							
<input type="checkbox"/>	✓	final report	#6 - by ardanlbnt was closed 1 minute ago							
<input type="checkbox"/>	✓	data analysis	#3 - by ardanlbnt was closed 5 minutes ago							

This is CANBAN version of our task assignment

Backlog

Priority board

Team items

Roadmap

In review

My items

New view

Filter by keyword or by field

Discard

Save

Planning4Estimate: 0

This is ready to be picked up

CSE4062S25\_Grp3 #11

knn 3 and 9

CSE4062S25\_Grp3 #12

ann 1-2 hidden layers

CSE4062S25\_Grp3 #13

decision tree (gini - gain)

+ Add item

In progress0 / 3Estimate: 0

This is actively being worked on

+ Add item

In review3 / 5Estimate: 0

This item is in review

CSE4062S25\_Grp3 #6

final report

CSE4062S25\_Grp3 #14

svm

CSE4062S25\_Grp3 #15

Naive Bayes

+ Add item

Done7Estimate: 0

This has been completed

CSE4062S25\_Grp3 #8

apriori

CSE4062S25\_Grp3 #1

exploratory data

CSE4062S25\_Grp3 #2

preprocess

CSE4062S25\_Grp3 #7

+ Add item