CSE4062

S25 Spring Semester

Group 3

Delivery #2 - Descriptive Analytics

ML-Powered Anemia Detection

NAME	NUMBER	DEPARTMENT	E-MAIL
UMUT BAYAR	150120043	COMPUTER ENG	umutbayar@marun.edu.tr
AHMET ARDA NALBANT	150121004	COMPUTER ENG	ardanalbant@marun.edu.tr
BURÇE PEKER	150619006	CHEMICAL ENG	burcepeker@marun.edu.tr
MUHARREM KEREM PAÇACI	150820053	BIOENGINEERING	kerempacaci@marun.edu.tr

1- Data Preprocessing Steps

1.1 First of all, we add necessary libraries.

```python
# 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".

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

```python
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()
```

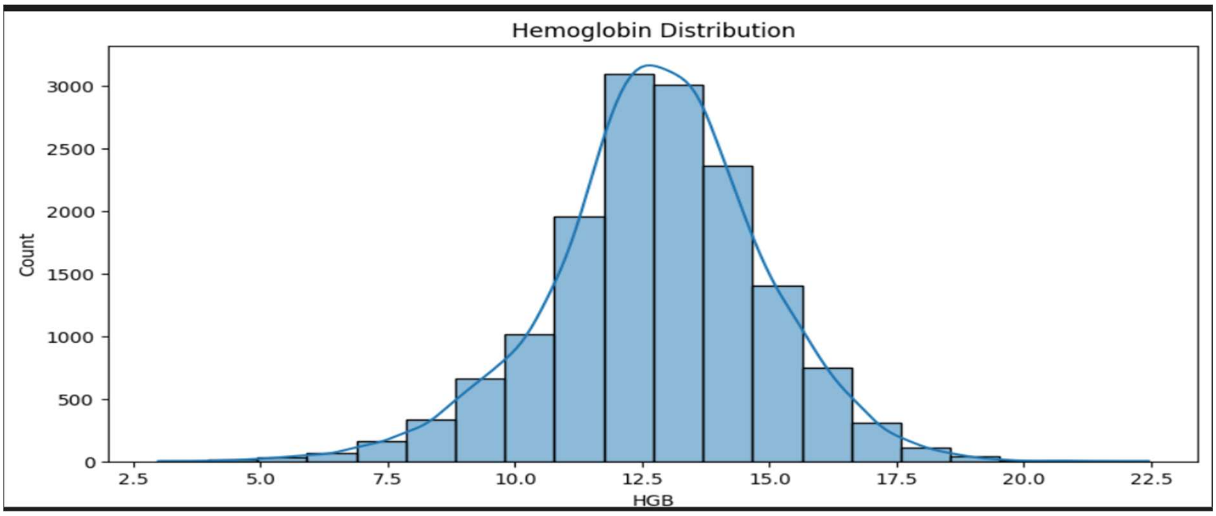
0.0s

```
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)

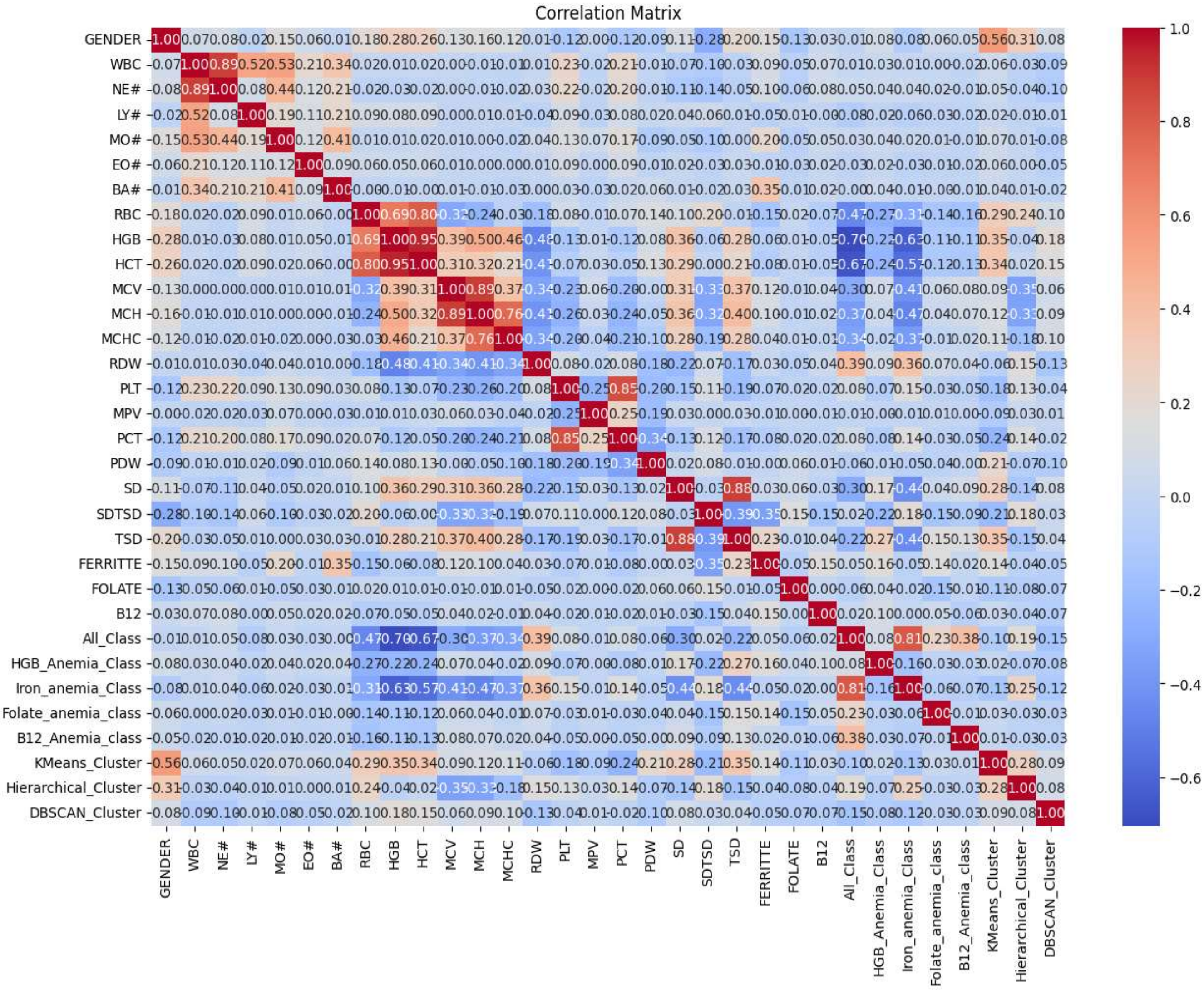
2.1 Hemoglobin Distribution Histogram

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

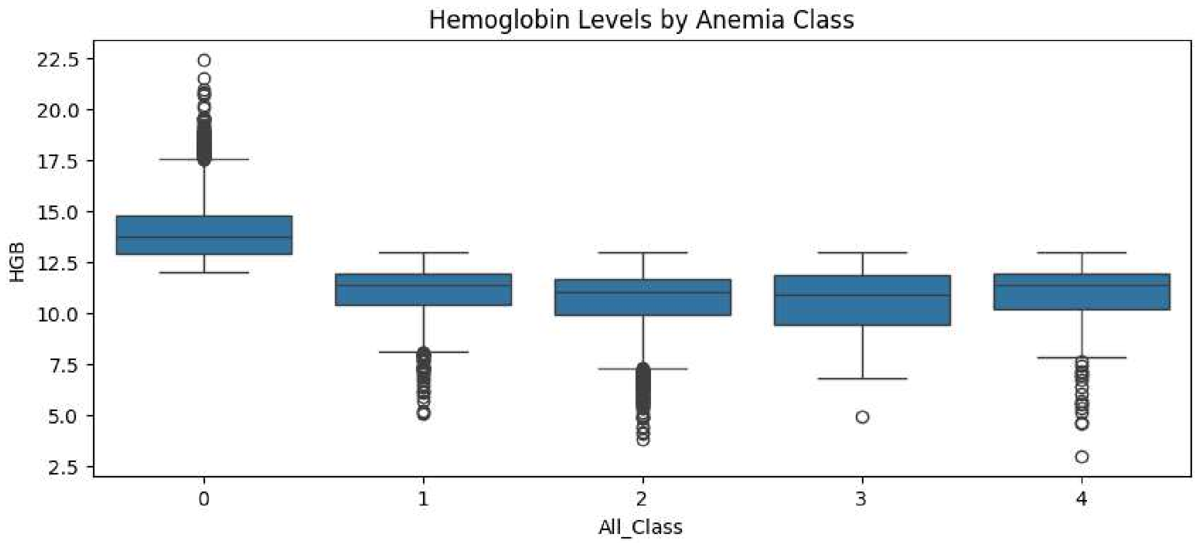
Image shows a correlation matrix heatmap, which visualizes the pairwise correlation coefficients between multiple variables in a dataset. This matrix helps identify which features are redundant (highly correlated), potential predictors (strongly related to target classes), and insights into relationships 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

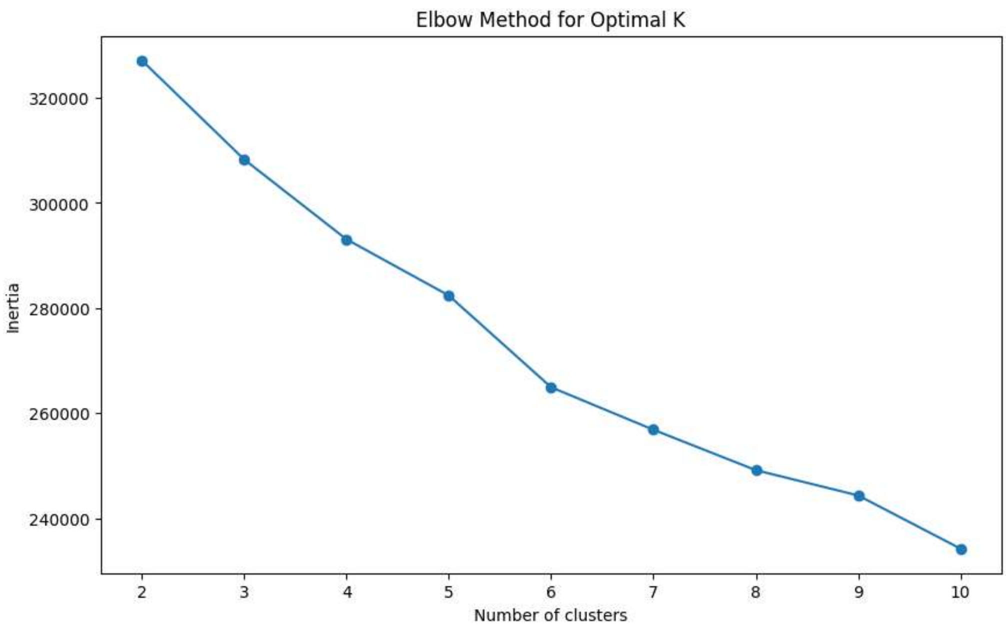
This boxplot shows the distribution of hemoglobin (HGB) levels across different anemia classes (`All_Class`, ranging from 0 to 4).

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)

1. **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:

1. Evidence-based (ferritin-HGB correlation in Cluster 2)
2. Scalable EHR integration
3. 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
```

Business Implications

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:
    1. Lock the patient file in RED
    2. Send SMS alert to the hematology team
    3. Prioritize appointment scheduling

IRON DEFICIENCY (Cluster 1):

- Example Patient: Woman with HGB 12.9 (mildly low) and RDW 48 (abnormal cell size)
- Action:
    1. Tag file for "Iron Protocol"
    2. Auto-order iron level tests every 3 months
    3. 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
                GENDER       WBC       NE#       LY#       MO#       EO#  \
DBSCAN_Cluster
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

                     BA#       RBC        HGB        HCT  ...   FERRITTE  \
DBSCAN_Cluster                                            ...
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**.

---

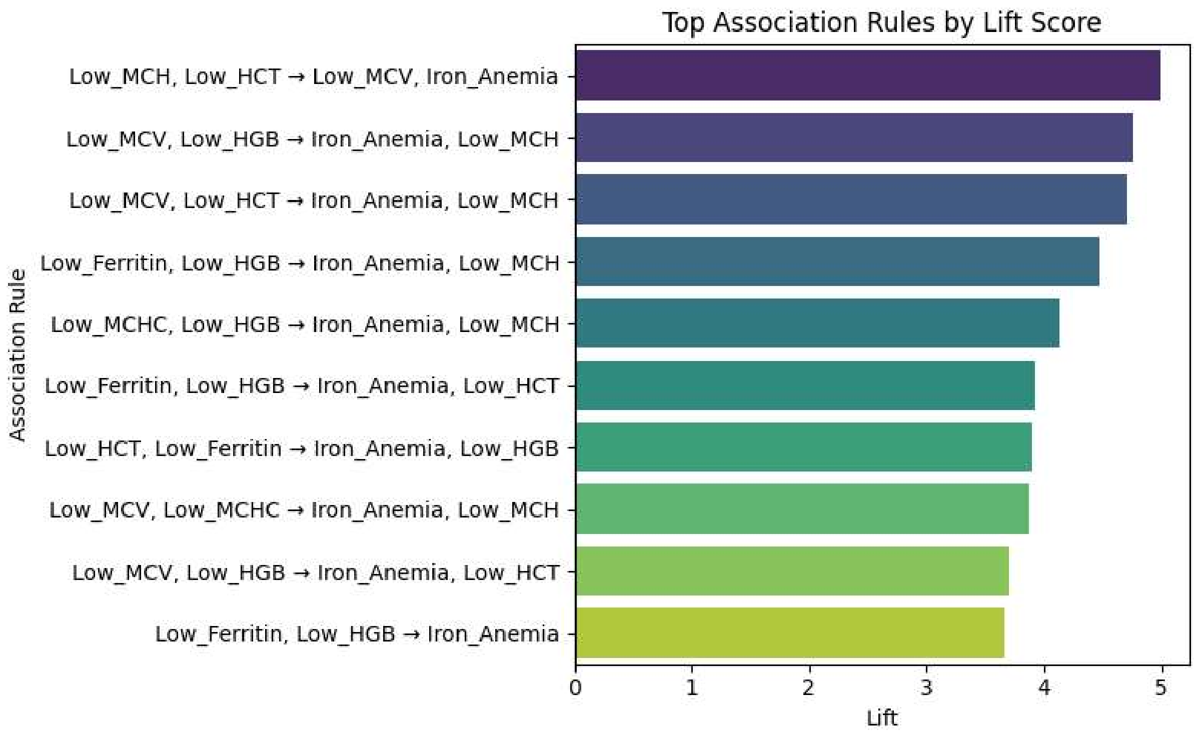
5. Triple Marker Confirmation

- **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:

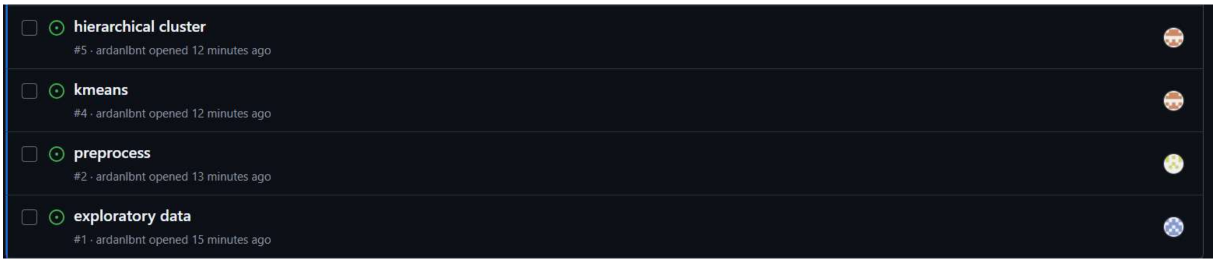
✓ **Reduce unnecessary tests**

✓ **Accelerate treatment decisions**

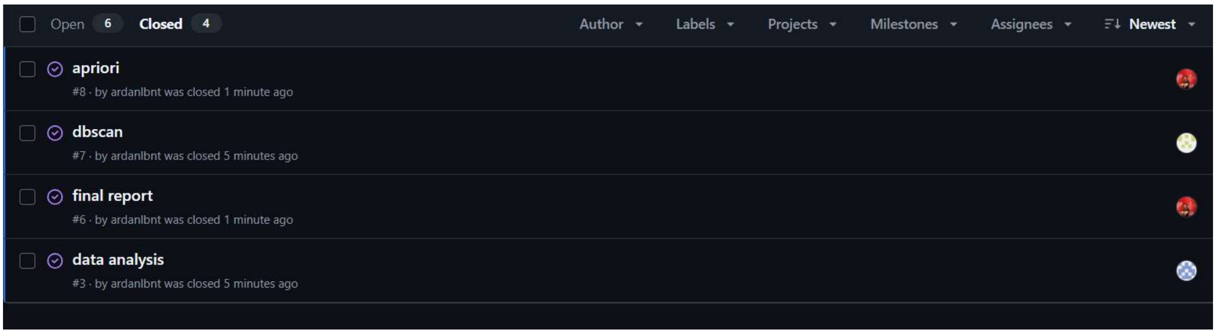
✓ **Improve accuracy** in anemia classification

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.



This is CANBAN version of our task assignment