



T.C.

MARMARA UNIVERSITY

FACULTY of ENGINEERING

CSE4062 Introduction to Data Science and Analytics

Spring 2025

Group #3

Delivery #3: Predictive Analysis

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

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:

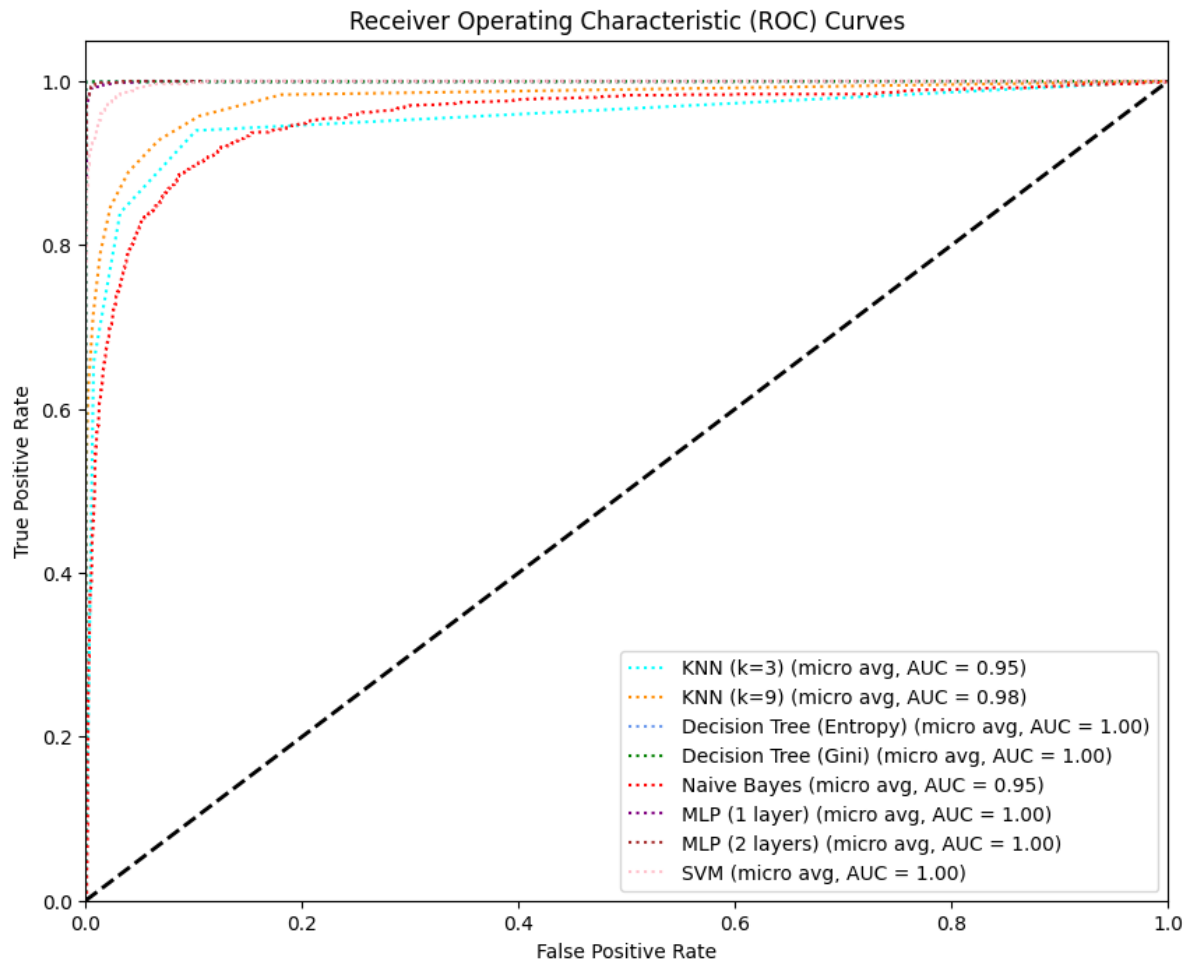
- **Accuracy:** The overall proportion of correct predictions.
- **F1 Score (Macro & Micro):**
 - *Macro*: Average F1 score across all classes, treating them equally.
 - *Micro*: F1 score calculated globally by counting total true positives, false negatives, and false positives.
- **AUC (Area Under the 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.

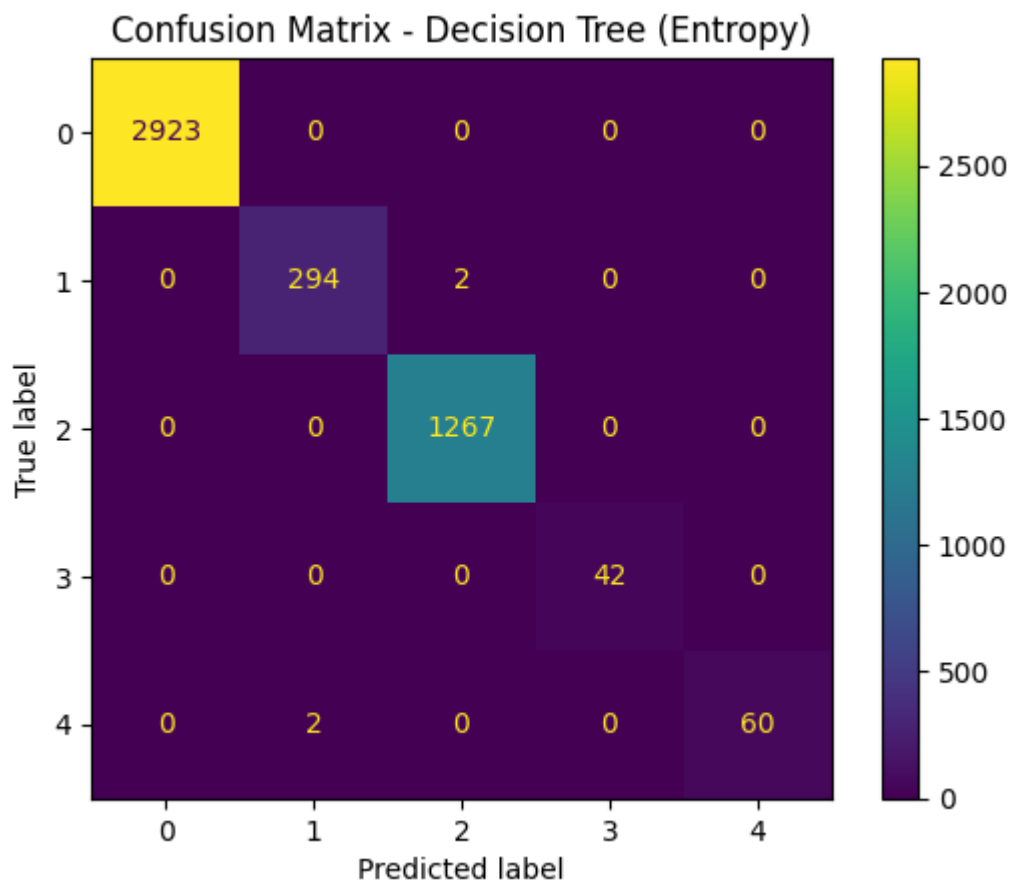


Fig.2 Confusion Matrix for Decision Tree (Entropy)

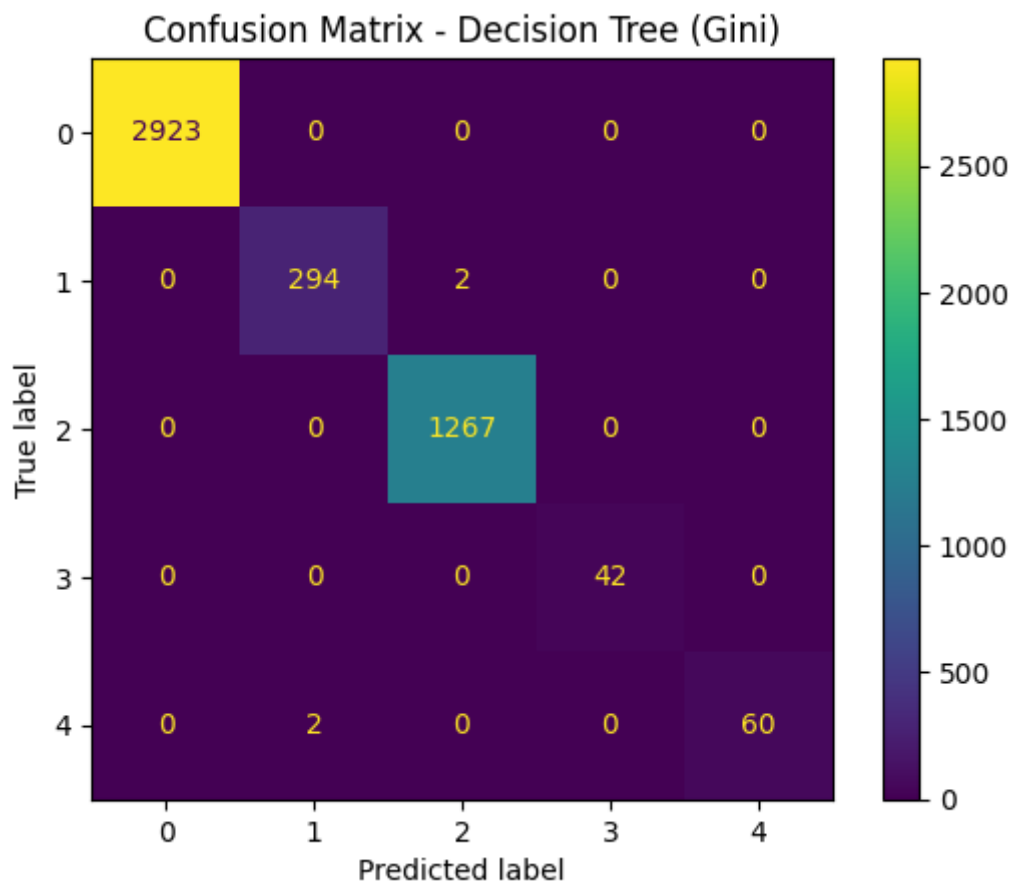


Fig.3 Confusion Matrix for Decision Tree (Gini Index)

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:

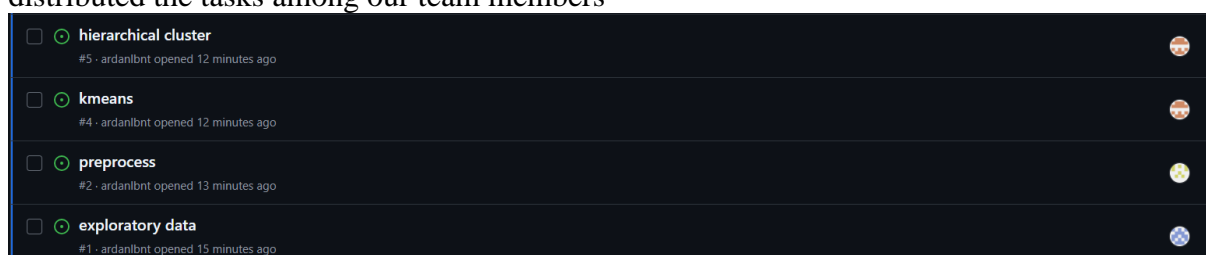
The P-value is = 0.2853

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"/>	<input checked="" type="checkbox"/>	apriori	#9 · by ardanlbnt was closed 1 minute ago							
<input type="checkbox"/>	<input checked="" type="checkbox"/>	dbscan	#7 · by ardanlbnt was closed 5 minutes ago							
<input type="checkbox"/>	<input checked="" type="checkbox"/>	final report	#6 · by ardanlbnt was closed 1 minute ago							
<input type="checkbox"/>	<input checked="" 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

Planning

4

Estimate: 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 progress

0 / 3

Estimate: 0

...

This is actively being worked on

+ Add item

In review

3 / 5

Estimate: 0

...

This item is in review

CSE4062S25_Grp3 #6

final report

CSE4062S25_Grp3 #14

svm

CSE4062S25_Grp3 #15

Naive Bayes

+ Add item

Done

7

Estimate: 0

...

This has been completed

CSE4062S25_Grp3 #8

apriori

CSE4062S25_Grp3 #1

exploratory data

CSE4062S25_Grp3 #2

preprocess

CSE4062S25_Grp3 #7

+ Add item