

BANA 273: MACHINE LEARNING ANALYTICS

Group Project Final Report

Group B16

Yiwei Lu

Ziqi Zhang

Zaiheng Shen

Wan-Lun Tsai

Chia-Chien Chang

1. Executive summary.....	3
1.1 Introduction.....	3
2. Data Analysis.....	4
2.1 Data Overview.....	4
2.2 Exploratory Data Analysis (EDA).....	5
3. Models without Pre-processing.....	9
3.1 Logistic Regression.....	11
3.2 Gaussian Naive Bayes.....	13
3.3 Decision Trees.....	15
4. Models Improvement with Pre-processing.....	18
4.1 Overfitting.....	18
4.2 Improvement: Standardization and SMOTE.....	20
4.3 Feature Selection.....	22
4.3.1 Feature Selection with the most relevant features(Hemoglobin and Gender).....	22
4.3.2 Feature Selection with the least relevant features(MCH, MCHC, and MCV).....	23
5. Summary and Key Takeaways.....	25
5.1 Models Performance Summary.....	25
5.2 Key Takeaways.....	27

1. Executive summary

This report is focusing on the application of machine learning techniques to predict anemia based on blood test results. By analyzing a dataset of 1,421 entries, we aim to develop accurate and reliable models to assist in early diagnosis and treatment. Our findings indicate that Decision Tree models exhibit superior performance in classifying individuals as anemic or non-anemic. Decision Trees are robust to data imperfections and are capable of capturing complex relationships between features. By leveraging machine learning, we can potentially improve the efficiency and accuracy of anemia diagnosis, leading to better patient outcomes.

1.1 Introduction

Anemia is a health issue that is often overlooked, some can cause serious consequences for individuals and communities. By understanding the risk factors and early indicators of anemia, people adopt proactive measures for prevention and management.

The dataset used in this report was found on Kaggle and it provides important insights into factors associated with anemia. By examining variables such as gender, hemoglobin levels, and red blood cell index, patterns and trends can be identified to inform public health strategies.

The recent enactment of the Iron Deficiency Education and Awareness Act¹ in the United States highlights the growing awareness of the importance of anemia identification. The legislation aims to increase awareness, improve diagnosis, and promote effective treatment of iron deficiency. By adopting a comprehensive approach to anemia treatment, we can contribute to global health improvement and alleviate the burden of this debilitating disease.

¹ AAHFN “Iron Deficiency Education and Awareness Act H. R. 6747

” AAHFN, December 13, 2023. <https://www.aahfn.org/page/IronDeficiencyAct>

2. Data Analysis

2.1 Data Overview

This dataset contains 1,421 entries with 6 columns. It represents medical data related to anemia diagnosis, including blood test results, gender, and a binary classification outcome.

Variables

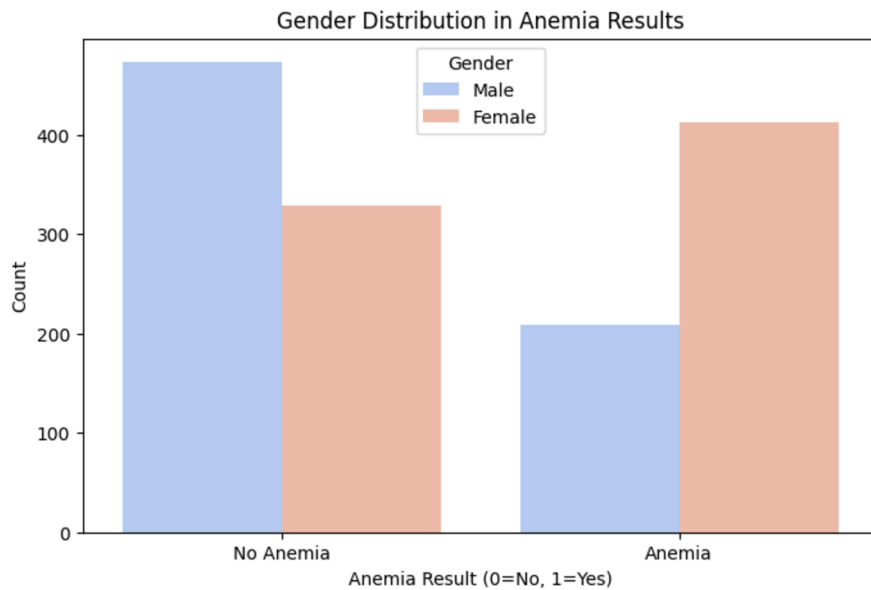
<i>Column Name</i>	<i>Description</i>	<i>Dtype</i>	<i>Non-Null Count</i>	<i>Null Count</i>
Gender	<i>The sex of the individual, with possible values "0 - Male" or "1 - Female."</i>	<i>int64</i>	<i>1421</i>	<i>0</i>
Hemoglobin	<i>Hemoglobin is a protein in your red blood cells that carries oxygen to your body's organs and tissues and transports carbon dioxide from your organs and tissues back to your lungs.</i>	<i>float64</i>	<i>1421</i>	<i>0</i>
MCH	<i>MCH is short for "mean corpuscular hemoglobin." It's the average amount in each of your red blood cells of a protein called hemoglobin, which carries oxygen around your body.</i>	<i>float64</i>	<i>1421</i>	<i>0</i>
MCHC	<i>MCHC stands for mean corpuscular hemoglobin concentration. It's a measure of the average concentration of hemoglobin inside a single red blood cell.</i>	<i>float64</i>	<i>1421</i>	<i>0</i>
MCV	<i>MCV stands for mean corpuscular volume. An MCV blood test measures the average size of your red blood cells.</i>	<i>float64</i>	<i>1421</i>	<i>0</i>
Results	<i>An indicator of anemia status, with possible values "0 - not anemic" or "1 - anemic".</i>	<i>int64</i>	<i>1421</i>	<i>0</i>

dtypes: float64(4), int64(2)

Here is the head of the dataset

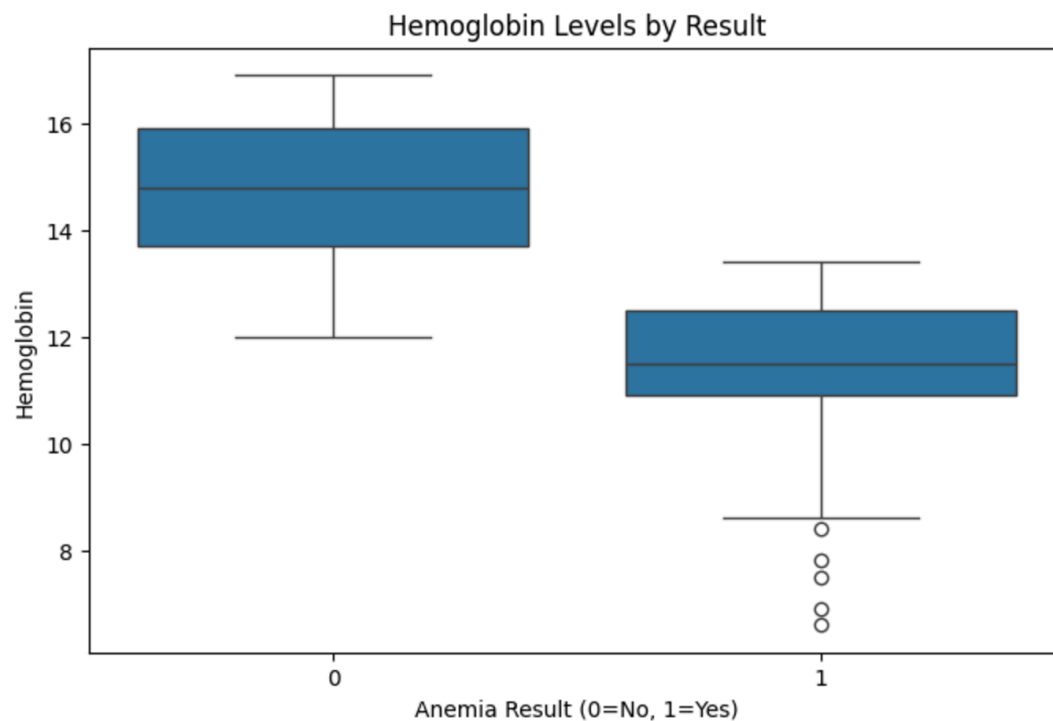
	Gender	Hemoglobin	MCH	MCHC	MCV	Result
0	1	14.9	22.7	29.1	83.7	0
1	0	15.9	25.4	28.3	72.0	0
2	0	9.0	21.5	29.6	71.2	1
3	0	14.9	16.0	31.4	87.5	0
4	1	14.7	22.0	28.2	99.5	0

2.2 Exploratory Data Analysis (EDA)

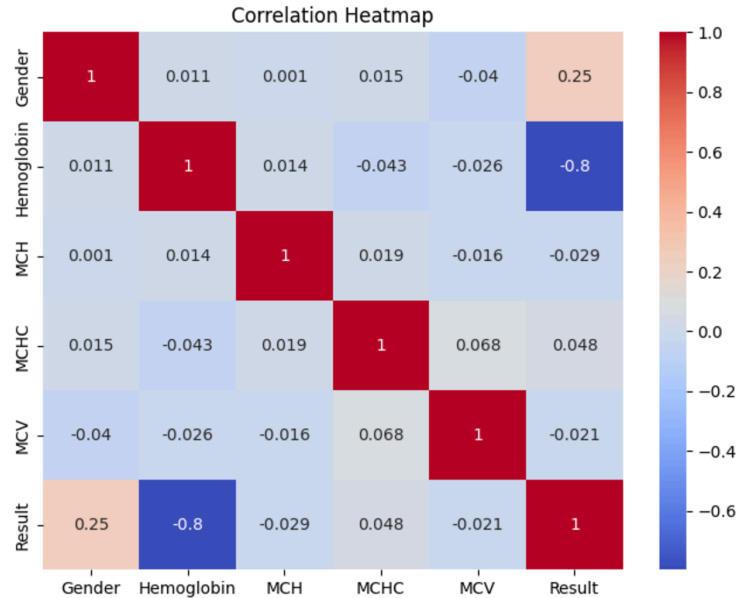


Count plots provide a visual representation of the sex distribution within each anemia category. This visualization enables us to identify gender-based differences in anemia prevalence. By comparing the number of males and females diagnosed with or without anemia, we can gain insight into potential gender-related factors.

This figure illustrates a significantly higher number of anemia cases in females. It suggests that gender may be a factor in anemia or that screening methods may vary by gender. Further investigation is necessary to understand the underlying cause of this discrepancy.



Boxplots offer a visual comparison of hemoglobin levels between individuals with or without anemia. This visualization helps us understand the distribution, central tendency (median), variability (interquartile range), and potential outliers within each group. Hemoglobin levels are a crucial indicator of anemia. Lower hemoglobin levels are characteristic of anemia. Outliers in the boxplot might represent extreme cases or potential data errors. By analyzing the boxplot, we can observe that anemic individuals tend to exhibit lower median hemoglobin levels compared to non-anemic individuals.

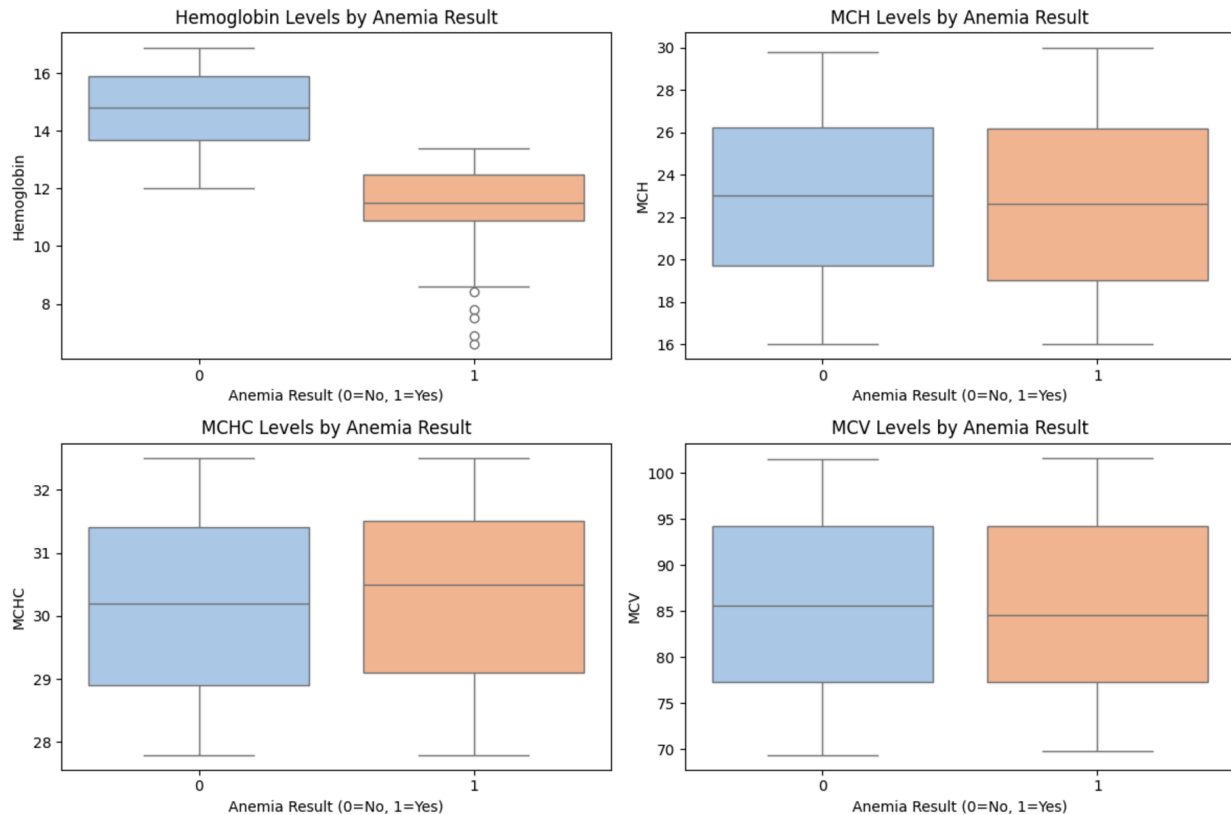


Correlation heatmap provides a visual representation of the strength and direction of relationships among different variables. By examining the patterns in the heatmap, highly correlated predictors can be identified, which can impact model performance and reveal underlying data patterns. Strong correlations may indicate redundancy between variables, which can be taken into account during the feature selection process for predictive modeling.

The heatmap in this case highlights the correlations among various blood parameters and anemia. A notable finding is the weak positive association between gender and anemia, suggesting that females are more likely to experience anemia than males. This can be attributed to biological factors such as menstruation and pregnancy, which can increase iron demands and the risk of iron deficiency anemia.

Another significant observation is the strong negative correlation between hemoglobin levels and anemia, which indicates lower hemoglobin are closely related to anemia. Hemoglobin plays an important role in oxygen transport, and its deficiency is the main cause of anemia.

Furthermore, the moderate negative correlations between MCHC, MCV, and MCH with anemia suggest that abnormalities in red blood cell size, hemoglobin content, and concentration may contribute to the development of anemia.

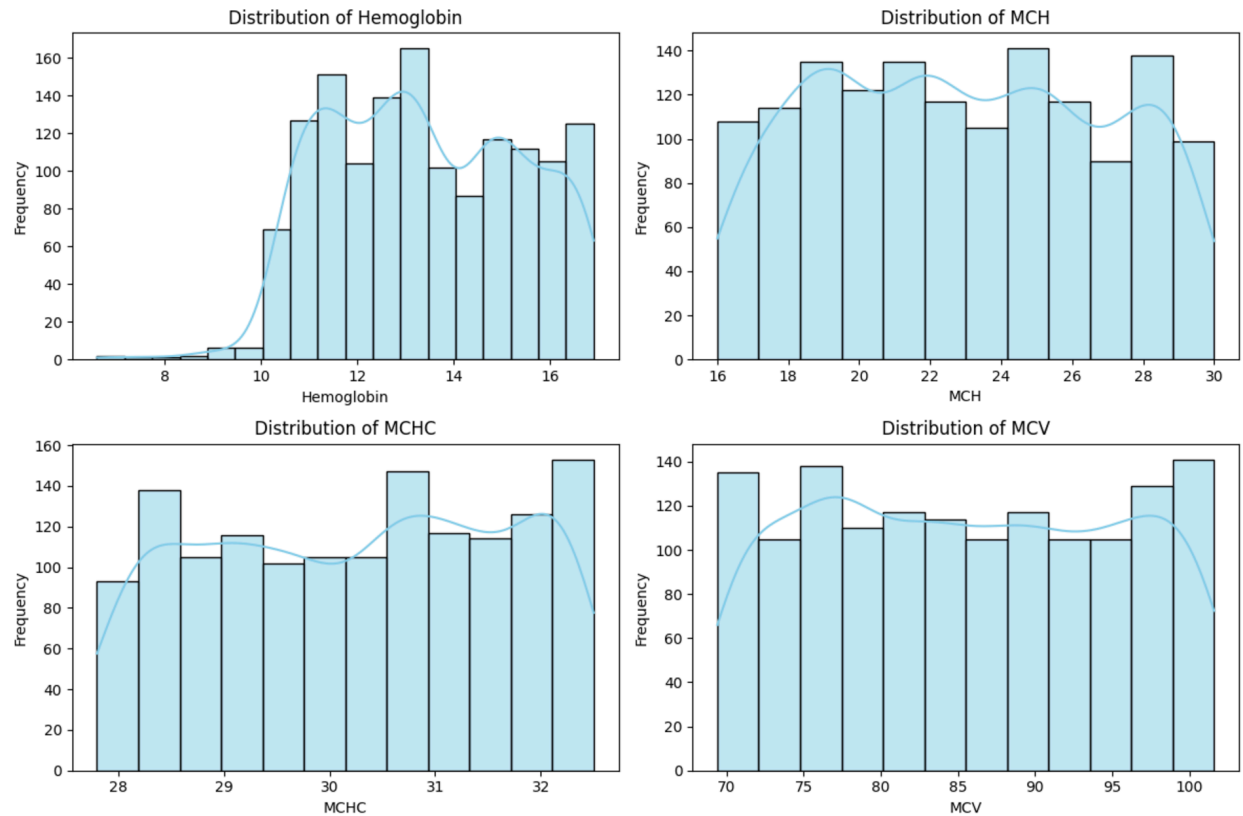


The Boxplots above provide a visual comparison of the distribution of blood parameters in anemic and nonanemic individuals, highlighting differences in central tendency, variability, and potential outliers.

Hemoglobin has become the most important indicator for diagnosing anemia. The non-anemic group shows higher median hemoglobin levels, whereas the anemic group shows significantly lower levels with several outliers. This suggests that lower hemoglobin levels are closely related to anemia.

Differences in MCH, MCHC, and MCV were small between the two groups. Although these parameters may provide additional information on red blood cell characteristics, they may not be independent predictors of anemia in this dataset. They might be more useful when combined with hemoglobin levels and other clinical information to diagnose different types of anemia.

In conclusion, hemoglobin levels are a major factor for initial screening and diagnosis of anemia. Although MCH, MCHC, and MCV can provide complementary information, they should be interpreted in conjunction with other clinical factors.



The histogram above showcased it's not the normal distribution of the indicators.

Hemoglobin: The distribution is skewed, with a peak around 12-14 g/dL. This suggests that most individuals have hemoglobin levels within this range.

MCH, MCHC, and MCV: These blood parameters, MCH, MCHC, and MCV, show a more consistent pattern with less variation in their values. This indicates that red blood cells tend to have a more uniform size, hemoglobin content, and concentration.

After analyzing the histograms, it's evident that the distribution of blood parameters, particularly hemoglobin, does not strictly follow a normal distribution.

3. Models without Pre-processing

Since this is a binary classification problem, we used the following models:

1. Logistic Regression

2. Gaussian Naive Bayes

3. Decision Trees

To train and evaluate our models, we adopt a 70/30 split, using 70% of the data for training and the remaining 30% for validation. When evaluating models for medical applications such as anemia prediction, it's crucial to select metrics that align with the real-world impact of model performance.

In this report, we used a combination of confusion matrix, F1-score, AUC-ROC, and recall to evaluate model performance. It balances precision and recall, and is particularly essential for imbalanced datasets. AUC-ROC provides a comprehensive measure of overall model performance. Recall is crucial for minimizing false negatives, which is extremely important for medical diagnoses. The confusion matrix provides detailed insights into true and false positives and negatives, aiding in model refinement. By leveraging these metrics, we gained a comprehensive understanding of the models' performance and its applicability to real-world anemia diagnosis. Note that all the models were built using a Python package called Sklearn.

Below is the overall benchmark accuracy without pre-processing data, and we will show the details in the following report:

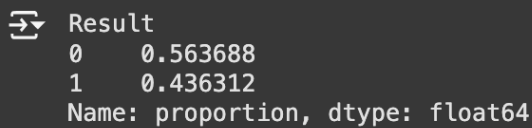
(The following results are based on the testing dataset)

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Logistic Regression (No pre-processing)</i>	<i>0.99</i>	<i>0.99</i>	<i>1</i>	<i>1</i>
<i>Gaussian Naive Bayes (No pre-processing)</i>	<i>0.95</i>	<i>0.95</i>	<i>0.95</i>	<i>0.99</i>
<i>Decision Trees (No pre-processing)</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>

Benchmark accuracy without pre-processing data

Here is the proportions of the class variable:

```
[134] print(data['Result'].value_counts(normalize=True))
```



```
Result
0    0.563688
1    0.436312
Name: proportion, dtype: float64
```

- **0 (not anemic): 56%**
- **1 (anemic): 44%**

3.1 Logistic Regression

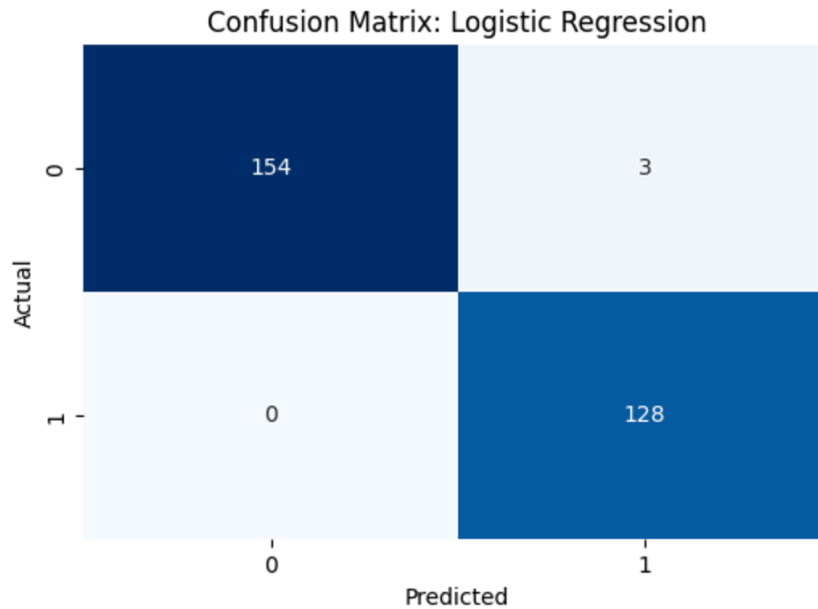
Logistic Regression Metrics:

- Accuracy: 0.99
- F1 Score: 0.99
- Recall: 1.00 (All actual positive cases were correctly predicted.)
- AUC-ROC: 1.00 (Excellent discrimination between positive and negative classes.)

Confusion Matrix:

```
[154  3]
```

```
[ 0 128]
```



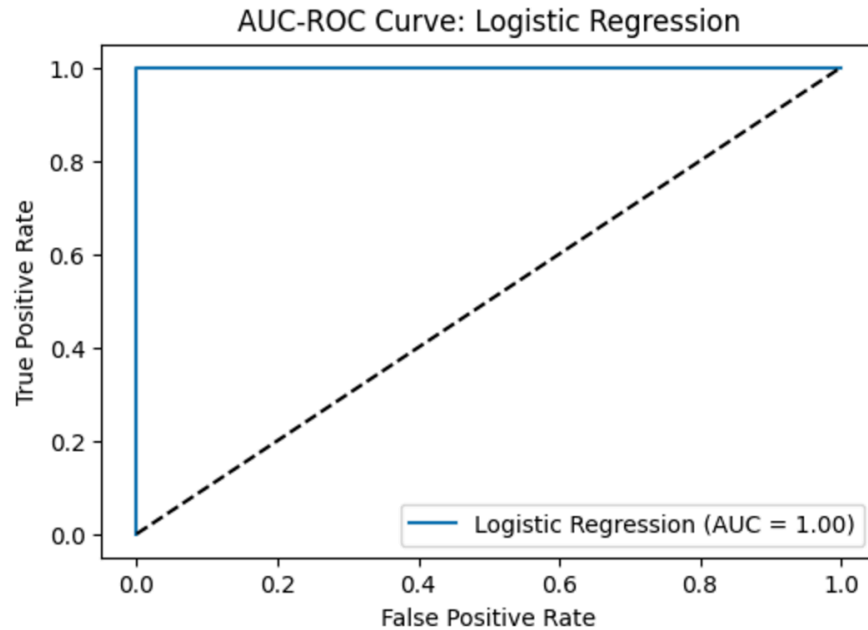
- **TN: 154** (Correctly predicted no anemia).
- **FP: 3** (Incorrectly predicted as anemia but actually not).
- **FN: 0** (No missed anemia cases).
- **TP: 128** (Correctly predicted anemia cases).

The model showed excellent performance in predicting anemia.

- Out of 157 actual cases of no anemia, only 3 were incorrectly classified as positive (false positives).
- Importantly, the model did not miss any actual cases of anemia (no false negatives), which is crucial for accurate medical diagnosis.

This strong performance is reflected in the high accuracy, F1-score, recall, and AUC-ROC scores. The confusion matrix further confirms the model's ability to distinguish between anemic and non-anemic individuals accurately.

While the model's overall performance is impressive, it's worth noting that the few false positives could lead to unnecessary follow-up tests or treatments. Future refinements could focus on reducing these false positives to minimize unnecessary medical interventions.



Logistic Regression: AUC-ROC = 1.00

The AUC-ROC of 1.00 indicates that the Logistic Regression model is a perfect classifier. It can distinguish between anemia and non-anemia cases with 100% accuracy across all thresholds. This means there is no overlap between the positive and negative class predictions.

The model's predictions are highly reliable. Every positive (anemia) case is correctly ranked higher than every negative (no anemia) case.

3.2 Gaussian Naive Bayes

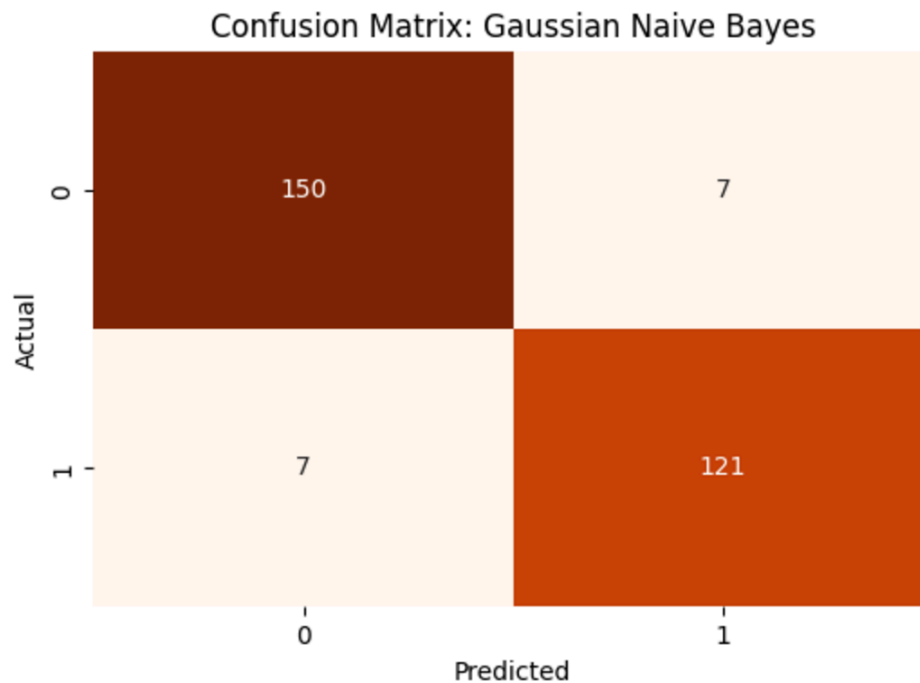
Gaussian Naive Bayes Metrics:

- Accuracy: 0.95
- F1 Score: 0.95
- Recall: 0.95 (Slightly lower compared to other models.)
- AUC-ROC: 0.99 (Great performance, but slightly lower performance.)

Confusion Matrix:

[150 7]

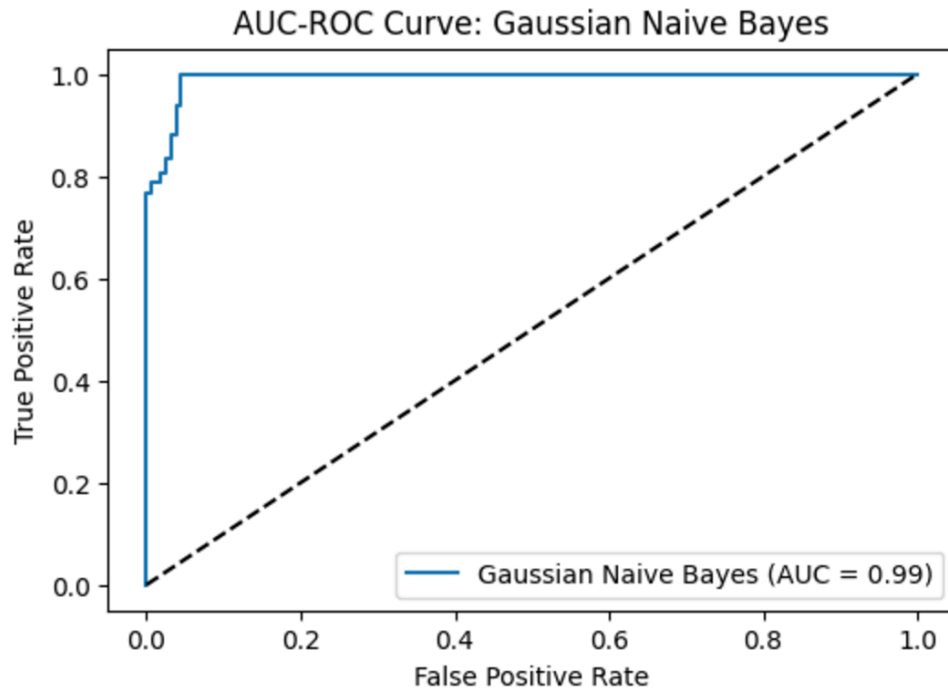
[7 121]



- **TN: 150** (Correctly predicted no anemia)
- **FP: 7** (Incorrectly predicted as anemia but actually not)
- **FN: 7** (Missed anemia cases)
- **TP: 121** (Correctly predicted anemia cases)

While the Gaussian Naive Bayes model exhibited strong performance, it demonstrated a notable drawback: it misclassified 7 actual cases of anemia as non-anemic (false negatives). This is a significant concern in a medical context, as misdiagnosing anemia can have severe consequences. While the model's overall performance is robust, its lower recall indicates a potential for missing some positive cases.

In comparison, Logistic Regression, Decision Tree models demonstrated superior performance, with fewer negatives. These models were more effective in identifying all cases of anemia, making them more reliable for medical diagnosis.



Gaussian Naive Bayes: AUC-ROC = 0.99

An AUC-ROC of 0.99 is excellent, indicating that the model is very close to a perfect classifier. It can distinguish between the two classes with high accuracy but might have slight imperfections.

There is a small chance of overlap between predicted positive and negative cases, which explains the false positives and false negatives seen in the confusion matrix. However, it is still a highly reliable model.

3.3 Decision Trees

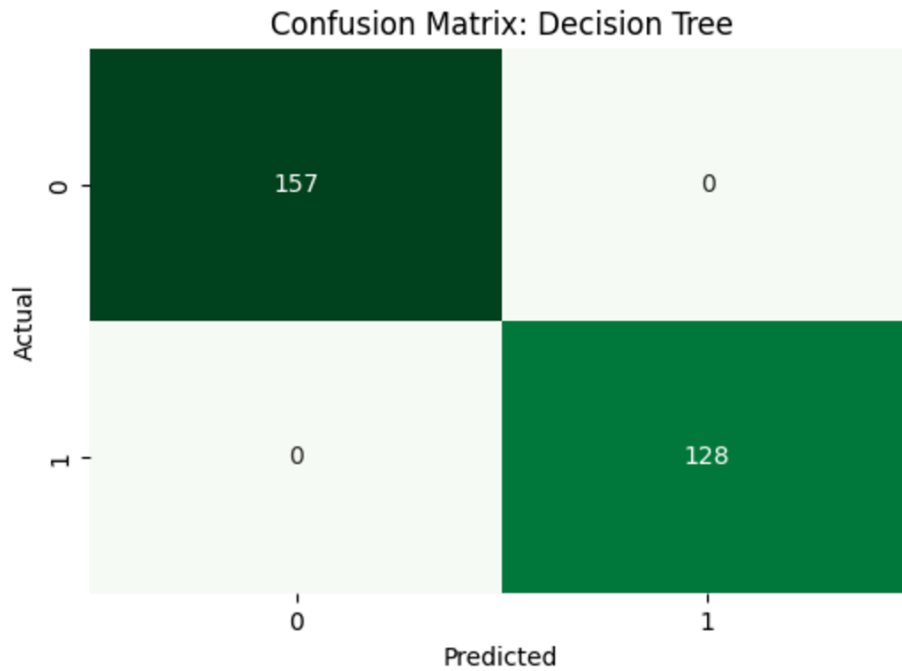
Decision Tree Metrics:

- Accuracy: 1.00
- F1 Score: 1.00
- Recall: 1.00 ((All actual positive cases were correctly predicted.))
- AUC-ROC: 1.00 (Excellent discrimination between positive and negative classes.)

Confusion Matrix:

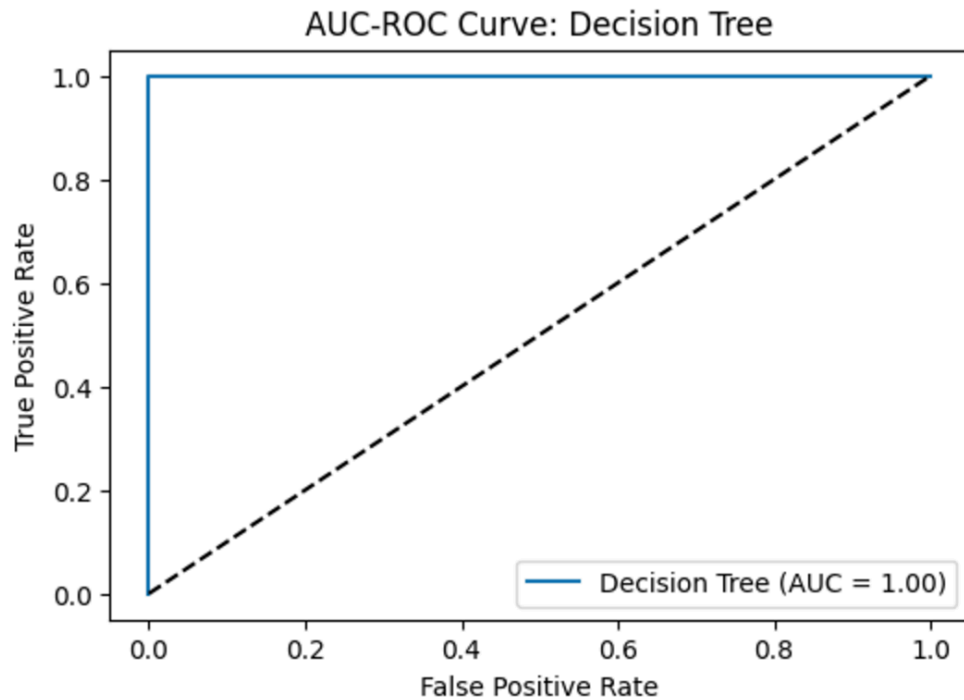
[157 0]

[0 128]



- **TN: 157** (Correctly predicted no anemia).
- **FP: 0** (No incorrect predictions of anemia).
- **FN: 0** (No missed anemia cases).
- **TP: 128** (Correctly predicted anemia cases).

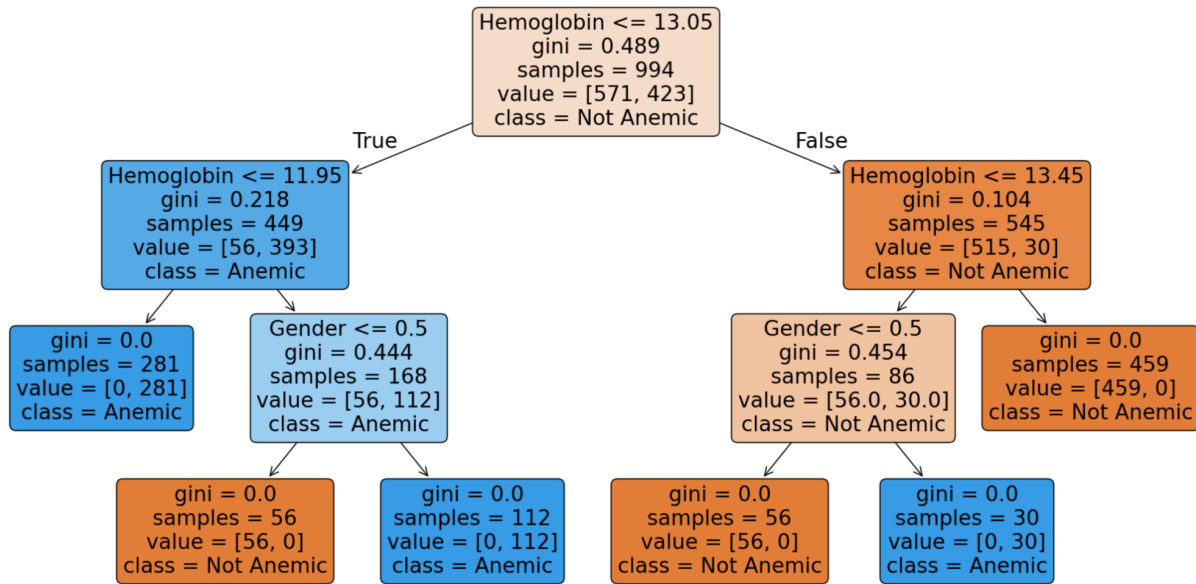
The Decision Tree model exhibited perfect performance, correctly identifying all cases of anemia and non-anemia. This indicates that the model effectively learned the underlying patterns in the data.



Decision Tree: AUC-ROC = 1.00

Similar to Logistic Regression, the AUC-ROC of 1.00 determines the Decision Tree model perfectly distinguishes between positive and negative cases.

There are no false positives or false negatives, confirming the confusion matrix results. The model performs exceptionally well and can be trusted for medical diagnosis.



From the correlation heatmap, Hemoglobin emerged as the most significant attribute, followed by Gender. To validate these findings, we visualized a decision tree. As expected, the tree heavily relied on Hemoglobin as the primary decision node, with Gender playing a secondary role. This confirms that these attributes are crucial for accurate anemia prediction. Notably, less significant attributes were excluded from the tree, highlighting the model's ability to identify and prioritize relevant features.

4. Models Improvement with Pre-processing

4.1 Overfitting

Based on the third section, demonstrated exceptional performance on both the training and testing data. For instance, the Decision Tree model and Logistic Regression models achieved impressive accuracies of 100% and 99%, respectively. While these results are remarkable, they strongly suggest the possibility of overfitting. To address this concern, we implemented the following methods to eliminate the risk of overfitting:

(The following results are based on the testing dataset)

<i>Split</i>	<i>80/20 Split</i>	<i>70/30 Split</i>
<i>Logistic Regression</i>	<i>Accuracy: 0.99</i> <i>F1 Score: 0.99</i> <i>Recall: 1.00</i> <i>AUC-ROC: 1.00</i>	<i>Accuracy: 0.99</i> <i>F1 Score: 0.99</i> <i>Recall: 1.00</i> <i>AUC-ROC: 1.00</i>
<i>Gaussian Naive Bayes</i>	<i>Accuracy: 0.97</i> <i>F1 Score: 0.96</i> <i>Recall: 0.98</i> <i>AUC-ROC: 0.99</i>	<i>Accuracy: 0.95</i> <i>F1 Score: 0.95</i> <i>Recall: 0.95</i> <i>AUC-ROC: 0.99</i>
<i>Decision Trees</i>	<i>Accuracy: 1</i> <i>F1 Score: 1</i> <i>Recall: 1</i> <i>AUC-ROC: 1</i>	<i>Accuracy: 1</i> <i>F1 Score: 1</i> <i>Recall: 1</i> <i>AUC-ROC: 1</i>

Method 1: Adopt an 80/20 split to 70/30: Based on the chart shown above. To train and evaluate our model, we initially adopted an 80/20 split, allocating 80% of the data for training and 20% for validation. The results are performed excellently, prompting us to explore additional possibilities. Using a different train-test split, such as 70/30, and observing consistently strong performance, it suggested that the model might not be overfitting. However, overfitting could still occur if the model memorizes patterns specific to the training data.

(The following results are based on the testing dataset)

Logistic Regression	<p>Cross-Validation Metrics (5-Fold):</p> <p>Accuracy Scores: [0.99497487 0.97487437 0.98492462 1. 0.98989899]</p> <p>Mean Accuracy: 0.99</p> <p>F1 Scores: [0.99435028 0.9726776 0.98342541 1. 0.98876404]</p> <p>Mean F1 Score: 0.99</p> <p>ROC-AUC Scores: [0.99979525 0.99948927 0.99938713 1. 1.]</p> <p>Mean ROC-AUC: 1.00</p>
----------------------------	---

Gaussian Naive Bayes	Cross-validation Accuracy Scores: [0.89949749 0.96482412 0.92462312 0.94974874 0.93939394] Mean Cross-validation Accuracy: 0.9356174813461247 Cross-validation AUC Scores: [0.97583948 0.99775281 0.98508682 0.98947986 0.98904959] Mean Cross-validation AUC: 0.9874415510319494
Decision Trees	Cross-validation Accuracy Scores: [1. 1. 1. 1. 1.] Mean Cross-validation Accuracy: 1.0 Cross-validation AUC Scores: [1. 1. 1. 1. 1.] Mean Cross-validation AUC: 1.0

Method 2: Cross-validation: The second better method (above chart) involves using cross-validation, which splits the data into multiple folds and evaluates the model's performance across all of them. This approach ensures that the model is tested on all parts of the dataset rather than relying on a single fixed train-test split. After applying cross-validation to all three models, the results remained consistently strong, confirming that overfitting is unlikely.

Therefore, we infer that the reason why this set of data performs so well is because:

1. **High-Quality Data:** This set of data is clear and only contains gender, Hemoglobin, MCH, MCHC, MCV, and result data with minimal noise. It appears to have been effectively cleaned and preprocessed.
2. **The Data Has Clear Patterns:** This suggests that the features in the dataset have a strong and well-defined relationship with the target variable.

4.2 Improvement: Standardization and SMOTE

Although the Decision Tree and Logistic Regression models achieved impressive accuracies of 100% and 99%, respectively, we were curious to explore whether the Naive Bayes model could be further improved. To address potential class imbalances in the dataset, we applied SMOTE (Synthetic Minority Oversampling Technique). Additionally, since the data did not follow a normal distribution, we implemented standardization to improve the performance of the model.

Furthermore, according to the correlation heatmap, 'Gender' and 'Hemoglobin' emerged as the most significant features for anemia diagnosis. We focused on scaling these features to improve the performance of Gaussian Naive Bayes, especially considering its sensitivity to feature scales.

(The following results are based on the testing dataset)

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Logistic Regression(No pre-processing)</i>	0.99	0.99	1	1
<i>Logistic Regression (SMOTE)</i>	0.99	0.99	1	1

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Gaussian Naive Bayes (No pre-processing)</i>	0.95	0.95	0.95	0.99
<i>Gaussian Naive Bayes (SMOTE)</i>	0.97	0.97	0.98	0.99

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Decision Trees (No pre-processing)</i>	1	1	1	1
<i>Decision Trees (SMOTE)</i>	1	1	1	1

After applying Standardization and SMOTE, we observed improvements in the Gaussian Naive Bayes model. While the Logistic Regression and Decision Tree models maintained their perfect performance, the Naive Bayes model's accuracy increased from 95% to 97%, F1-score from 95% to 97%, and recall from 95% to 98%.

This improvement can be attributed to the following:

Standardization:

- This ensures that features with larger scales, like Hemoglobin, do not disproportionately influence the model's predictions compared to features with smaller scales, like Gender. The standardization process adjusts the data to have a mean of 0 and a standard deviation of 1, helping to balance their impact on the model.
- Aligns with the Gaussian Naive Bayes assumption of normally distributed features.

SMOTE:

- Our dataset was imbalanced, with more samples in the "not anemic" (56%) class than in the "anemic" (44%) class. This imbalance could lead to a biased model that

underperforms the minority class. To address this, we applied SMOTE, which generates synthetic samples for the minority class, helping to balance the dataset.

- A more balanced dataset allows the model to learn more effectively from both classes, resulting in improved recall for the "anemic" class. This enhances the model's ability to accurately identify true positive cases of anemia.

By addressing these issues, the Gaussian Naive Bayes model experienced a significant improvement in performance, particularly in its ability to identify anemic cases.

4.3 Feature Selection

4.3.1 Feature Selection with the most relevant features(Hemoglobin and Gender)

(The following results are based on the testing dataset)

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Logistic Regression(No pre-processing)</i>	0.99	0.99	1	1
<i>Logistic Regression (Feature Selection)</i>	0.99	0.99	1	1

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Gaussian Naive Bayes (No pre-processing)</i>	0.95	0.95	0.95	0.99
<i>Gaussian Naive Bayes (Feature Selection)</i>	0.97	0.97	0.98	0.99

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Decision Trees (No pre-processing)</i>	1	1	1	1
<i>Decision Trees (Feature Selection)</i>	1	1	1	1

From the previous analysis of correlation heatmaps, we identified Hemoglobin and Gender as the two most significant features correlated with anemia. These features exhibit a strong negative and positive relationship, respectively, with the Results.

To further refine our models, we retrained them using only these two features. Interestingly, while the Logistic Regression and Decision Tree models maintained their perfect performance, the Gaussian Naive Bayes model experienced another improvement. Its accuracy, F1-score, and recall increased from 95% to 97%, 95% to 97%, and 95% to 98%, respectively.

This improvement can be attributed to Fewer Features:

- Features with weak correlations or high noise levels can introduce irrelevant information, potentially confusing the model and leading to incorrect predictions.
- By focusing on the most relevant features, the model can simplify its calculations, reducing computational complexity and enhancing efficiency.
- A simpler model with fewer features is less prone to overfitting, increasing its ability to generalize effectively to new, unseen data.

By focusing on “Hemoglobin” and “Gender”, the model avoids being misled by less relevant features like MCH, MCHC, and MCV. This allows the model to make more accurate predictions, especially for the minority class (anemic).

4.3.2 Feature Selection with the least relevant features(MCH, MCHC, and MCV)

We were curious to assess the impact of less correlated features on our models. We selected "MCH," "MCHC," and "MCV," which exhibit weak correlations with the target variable (Results) of -0.0029, 0.048, and -0.021, respectively.

Upon training our models with these features, we made an interesting observation: the Decision Tree model maintained its performance, while the Logistic Regression and Gaussian Naive Bayes models experienced a significant decline. This shows that anemia can still be accurately identified using these "weaker" attributes when the right model is applied.

Additional Insight:

If we involve medical experts to design and create new derived features (e.g., combining MCH, MCHC, and MCV in clinically meaningful ways), these transformed features can be fed into

models like Logistic Regression and Gaussian Naive Bayes. This approach could significantly improve their performance, allowing them to capture more of the underlying patterns in the data, similar to how the Decision Tree does.

(The following results are based on the testing dataset)

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Logistic Regression(No pre-processing)</i>	0.99	0.99	1	1
<i>Logistic Regression (Feature Selection with the most relevant features)</i>	0.99	0.99	1	1
<i>Logistic Regression (Feature Selection with the least relevant features)</i>	0.55	0.00	0.00	0.51

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Gaussian Naive Bayes (No pre-processing)</i>	0.95	0.95	0.95	0.99
<i>Gaussian Naive Bayes (Feature Selection with the most relevant features)</i>	0.97	0.97	0.98	0.99
<i>Gaussian Naive Bayes (Feature Selection with the least relevant features)</i>	0.57	0.08	0.04	0.56

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Decision Trees (No pre-processing)</i>	1	1	1	1
<i>Decision Trees (Feature Selection with the most relevant features)</i>	1	1	1	1
<i>Decision Trees (Feature Selection with the least relevant features)</i>	0.96	0.96	0.93	0.96

According to the table above, we observed that Logistic Regression and Gaussian Naive Bayes performed poorly, while Decision Trees performed well. The following sections will delve into the details of these performance differences.

1. Why Logistic Regression and Gaussian Naive Bayes Performed Poorly:

- **Logistic Regression:**

Logistic Regression models a linear relationship between the features and the target variable. When using weakly correlated features like MCH, MCHC, and MCV, the model cannot find a clear linear boundary to separate the classes. Therefore, these features provide minimal information to the model, hindering its ability to make accurate predictions.

- **Gaussian Naive Bayes:**

Naive Bayes assumes that each feature follows a normal distribution and makes predictions based on probability calculations. However, when the selected features (MCH, MCHC, and MCV) offer minimal useful information, the model struggles to estimate probabilities accurately. This leads to a performance that is slightly better than Logistic Regression but still unsatisfactory, as the input features do not provide clear distinctions between the classes.

2. Why Decision Trees Performed Well:

- **Decision Trees** are capable of capturing complex, non-linear relationships between features. This allows them to identify subtle patterns even in weakly correlated features. Decision Trees can consider interactions between features. For example, a combination of low MCH and high MCV might be a strong indicator of anemia, even though these features individually may not be highly informative. What is more, Decision Trees can adapt to different data distributions and feature relationships, making them robust to noise and weak correlations.

5. Summary and Key Takeaways

5.1 Models Performance Summary

(The following results are based on the testing dataset)

The summary of our Logistic Regression model performance:

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Logistic Regression(No pre-processing)</i>	0.99	0.99	1	1
<i>Logistic Regression (SMOTE)</i>	0.99	0.99	1	1
<i>Logistic Regression (Feature Selection with the most relevant features)</i>	0.99	0.99	1	1
<i>Logistic Regression (Feature Selection with the least relevant features)</i>	0.55	0.00	0.00	0.51

The summary of our Gaussian Naive Bayes model performance:

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Gaussian Naive Bayes (No pre-processing)</i>	0.95	0.95	0.95	0.99
<i>Gaussian Naive Bayes (SMOTE)</i>	0.97	0.97	0.98	0.99
<i>Gaussian Naive Bayes (Feature Selection with the most relevant features)</i>	0.97	0.97	0.98	0.99
<i>Gaussian Naive Bayes (Feature Selection with the least relevant features)</i>	0.57	0.08	0.04	0.56

The summary of our Decision Tree's' model performance:

<i>Method</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Recall</i>	<i>AUC--ROC</i>
<i>Decision Trees (No pre-processing)</i>	1	1	1	1
<i>Decision Trees (SMOTE)</i>	1	1	1	1
<i>Decision Trees (Feature Selection with the most relevant features)</i>	1	1	1	1
<i>Decision Trees (Feature Selection with the least relevant features)</i>	0.96	0.96	0.93	0.96

Real-world applications:

- The **Decision Tree model** is highly recommended. Its perfect accuracy and recall ensure minimal false negatives, a critical factor in accurate anemia diagnosis. This model's ability to handle various data conditions makes it a robust choice.
- **Logistic Regression** is also a strong alternative, particularly for simpler datasets. However, it's important to be aware of the risk of false positives, which may require additional medical tests or consultations.
- Data preprocessing techniques such as **SMOTE and feature selection** played a key role in improving the performance of some models, especially Gaussian Naive Bayes. On the other hand, the decision tree model required minimal preprocessing to perform at its best.

5.2 Key Takeaways

Our experiments highlight the varying degrees of sensitivity that machine learning models exhibit towards feature quality and correlation.

- **Decision Trees:** Demonstrated outstanding performance, effectively capturing complex and non-linear relationships between features which makes them a reliable choice for anemia prediction.
- **Logistic Regression and Naive Bayes:** These models are very sensitive to feature quality.
 - Logistic Regression performs best when features have a strong linear relationship with the target variable. Features that lack a clear linear relationship make a minimal contribution to the model's predictive performance.
 - Naive Bayes assumes that features are independent and perform well when each feature contributes independently to the classification process. However, weakly correlated features may lead to inaccurate probability estimation, reducing model accuracy.

Understanding these nuances is crucial for selecting the most appropriate model for a given dataset and problem.