Can Machine Learning Revolutionize Anemia Diagnosis?

Group B16
Yiwei Lu
Ziqi Zhang
Zaiheng Shen
Wan-Lun Tsai
Chia-Chien Chang

Table of contents

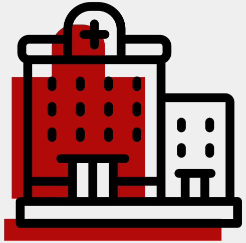
Introduction

Introduction

Anemia is a widespread health challenge that often goes undiagnosed, potentially impacting millions of individuals worldwide. By leveraging data-driven insights and understanding key risk factors, we can develop more effective strategies for early detection and treatment. Our research aims to shed light on this critical health issue and contribute to improved public health outcomes.

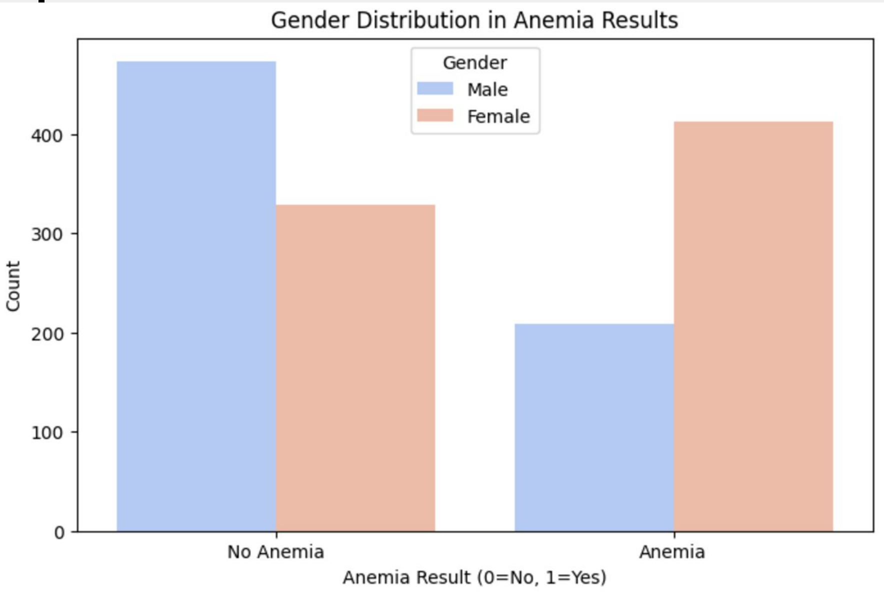
Data Analysis

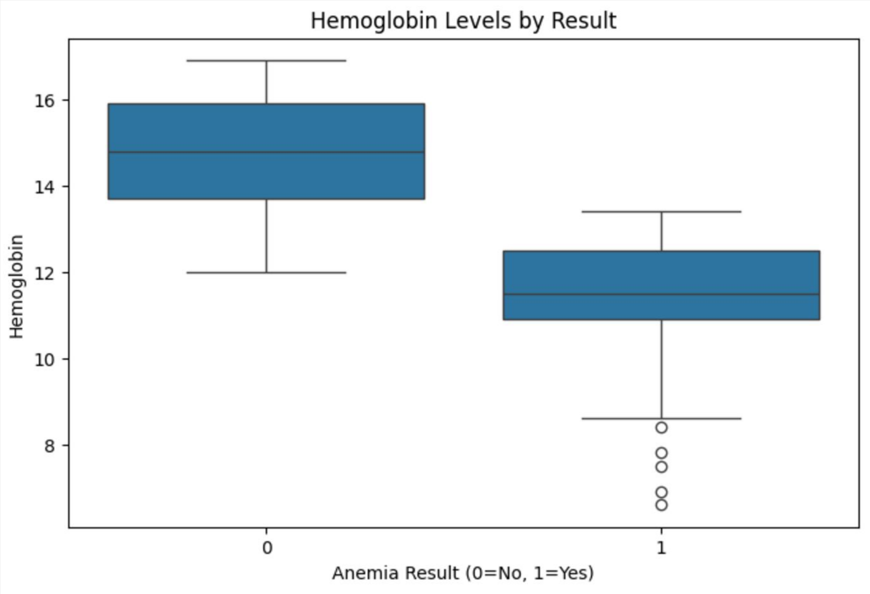
Data Analysis



Data Overview:
This dataset contains 1421 people with categories of Gender Hemoglobin MCH MCHC MCV and Results

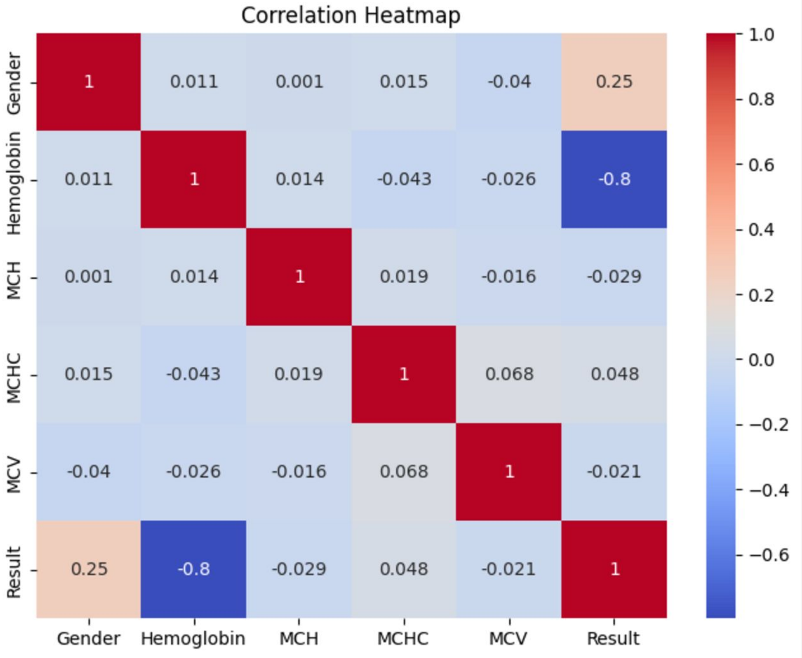
Data Analysis

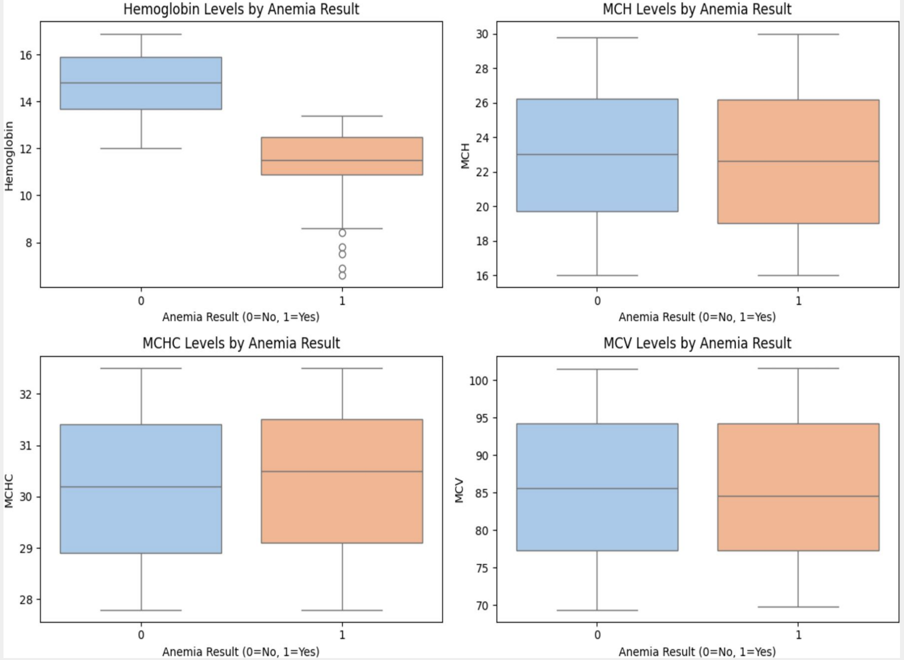




- Females diagnosed with Anemia is greater.
- Patients diagnosed with Anemia appears with a lower medium Hemoglobin

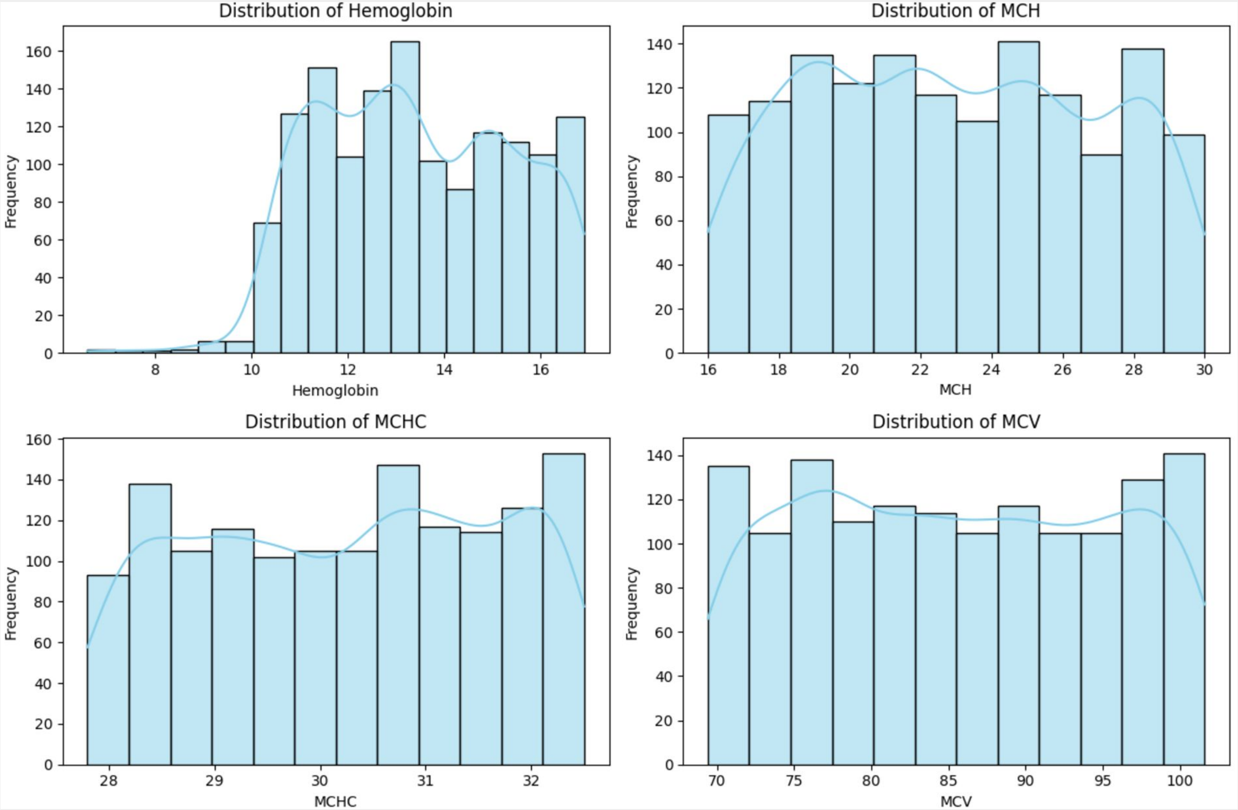
Data Analysis





- Correlation heatmap showcase the correlation between different variables
- Boxplots of all variable

Distribution



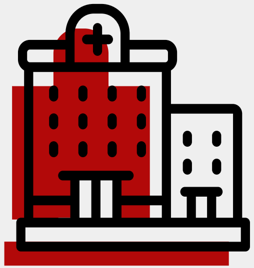
The proportions of the class variable

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

Machine Learning Models

*Without Pre-processing

Machine Learning Models

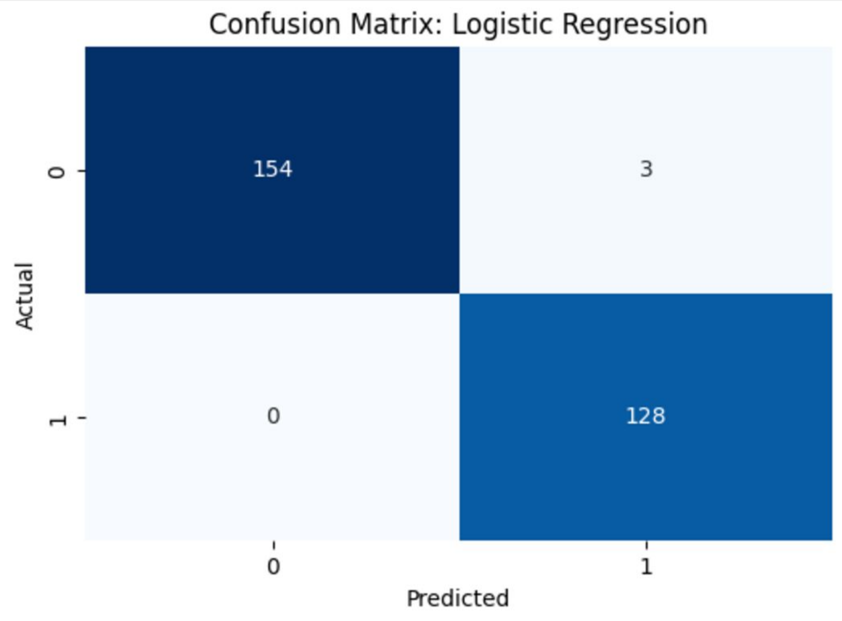


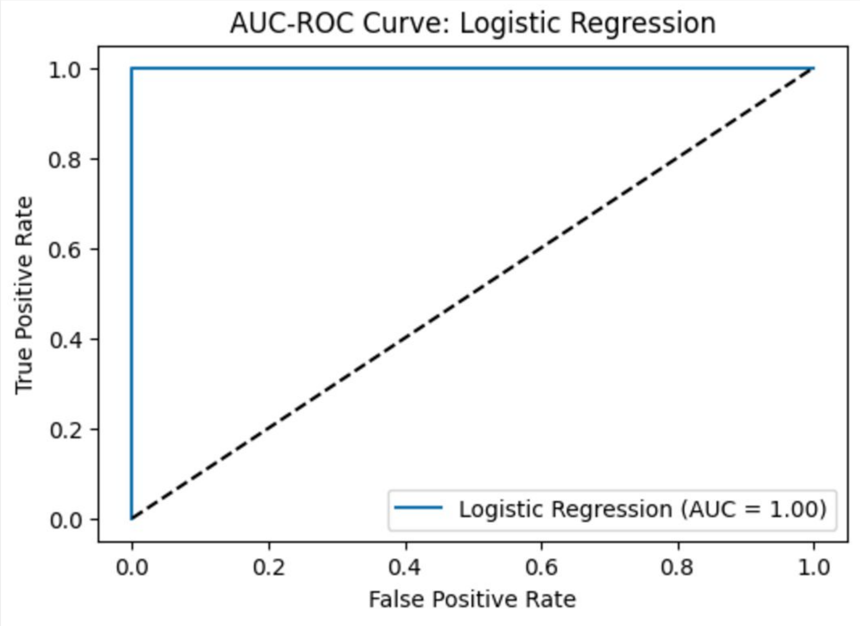
Models used in this projects:

1. Logistic Regression
2. Gaussian Naive Bayes
3. Decision Trees

70/30 split was adopted in this project

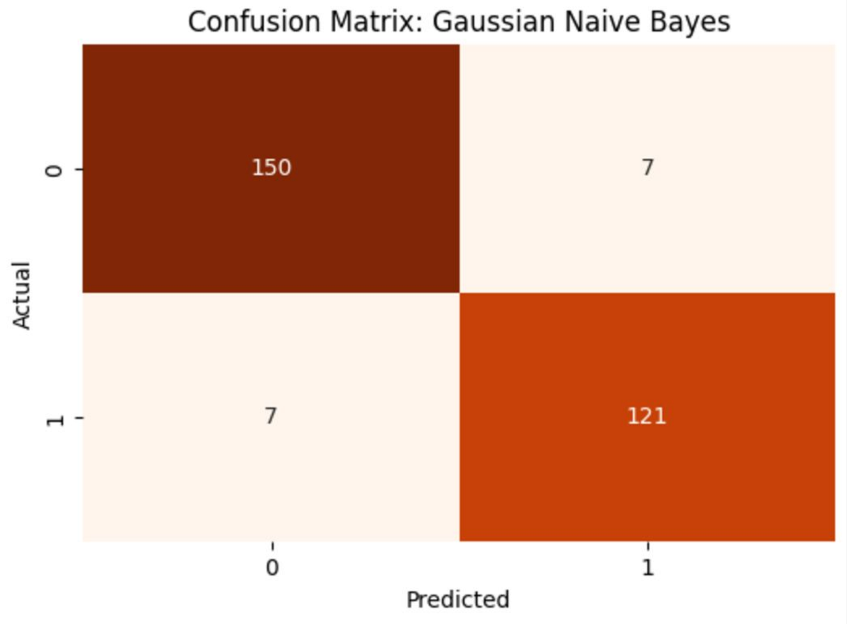
Logistic Regression

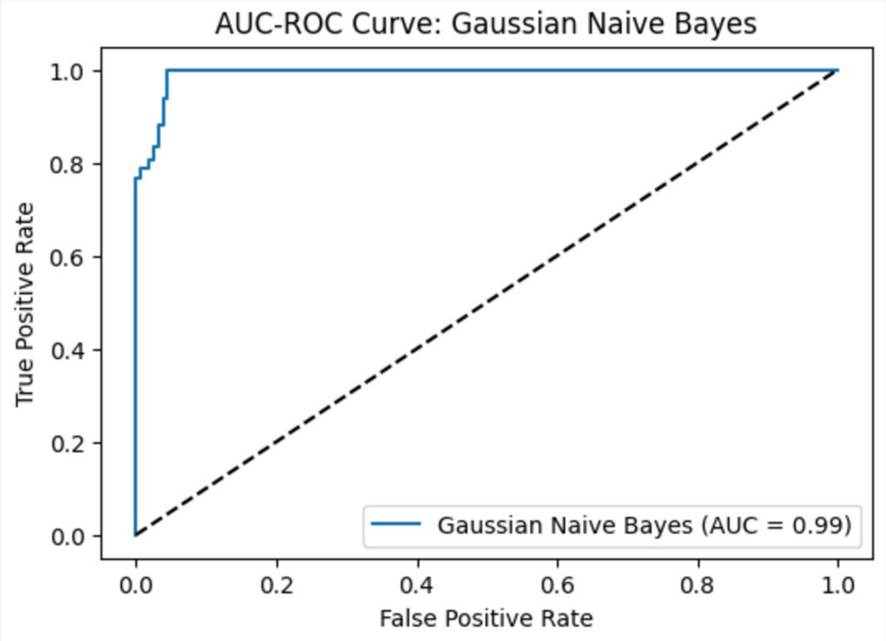




Accuracy: 0.99 F1 Score: 0.99 Recall: 1.00 AUC-ROC; 1.00

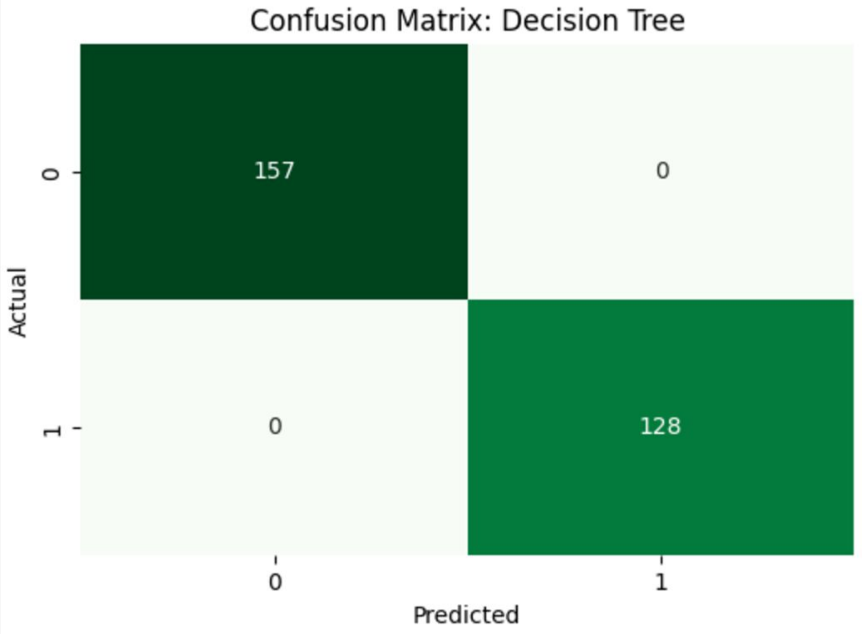
Gaussian Naive Baye

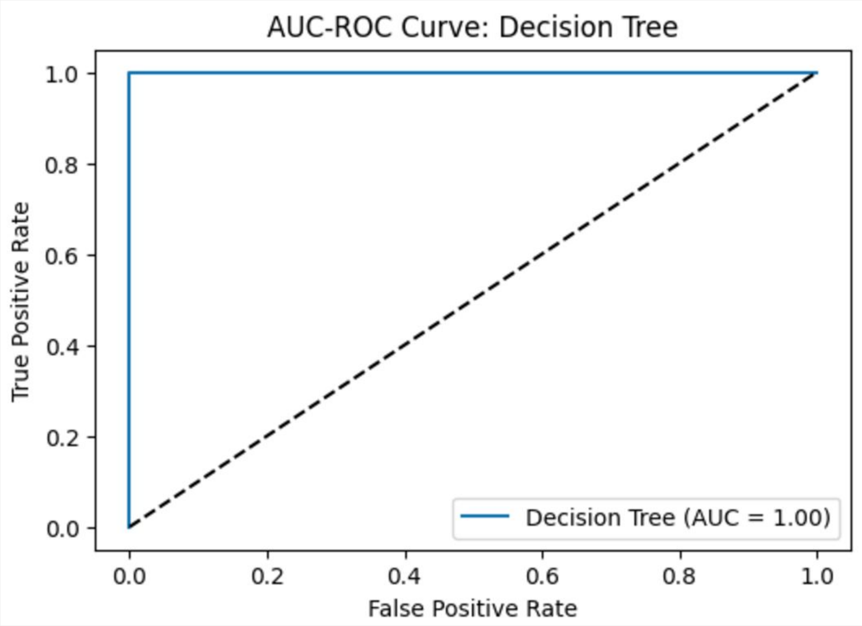




Accuracy: 0.95 F1 Score: 0.95 Recall: 0.95 AUC-ROC;:0.99

Decision Tree





Accuracy: 1.00 F1 Score: 1.00 Recall: 1.00 AUC-ROC: 1.00

Overfitting?

Method 1

Adopt an 80/20 split to 70/30

Method 2

Cross-validation

Method 1:

Adopt an 80/20 split to 70/30

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

Method 2:

Cross-validation

Logistic Regression	Cross-Validation Metrics (5-Fold): Accuracy Scores: [0.99497487 0.97487437 0.98492462 1. 0.98989899] Mean Accuracy: 0.99 F1 Scores: [0.99435028 0.9726776 0.98342541 1. 0.98876404] Mean F1 Score: 0.99 ROC-AUC Scores: [0.99979525 0.99948927 0.99938713 1. 1.] Mean ROC-AUC: 1.00
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 0.98947906 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

If there is no overfitting, why does this set of data perform so well?

Improvement

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

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

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

Standardization

The data did not follow a normal distribution

SMOTE

The potential class imbalances in the dataset

What happens after improvement?

The Naive Bayes model's accuracy increased from 95% to 97%, F1-score from 95% to 97%, and recall from 95% to 98%

Feature Selection

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

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

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

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

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

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

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

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

Summary

Logistic Regression model performance:

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

Gaussian Naive Bayes model performance:

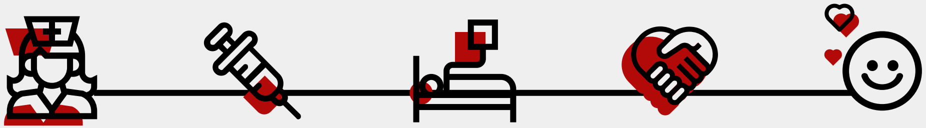
Method	*Accuracy*	*F1-score*	*Recall*	*AUC--ROC*
Gaussian Naive Bayes (No pre-processing)	*0.95*	*0.95*	*0.95*	*0.99*
Gaussian Naive Bayes (SMOTE)	*0.97*	*0.97*	*0.98*	*0.99*
Gaussian Naive Bayes (Feature Selection with the most relevant features)	*0.97*	*0.97*	*0.98*	*0.99*
Gaussian Naive Bayes (Feature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

Decision Tree model performance:

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

Real - World Applications

- Decision Tree model
- Logistic Regression model
- SMOTE and feature selection



Key Takeaways

Decision Trees

Logistic Regression

Naive Bayes

Thanks!

Do you have any questions?