STTHK2113 DATA ANALYTICS (B)

Predictive Modelling Task

Blood Donor and Non-Blood Donor

LECTURER: TS. DR. MOHAMED ALI B. SAIP

Group Members

HANA SYAKIRAH
299403

SITI AIN ATHIQAH
297545

NUR FAIZLYANA
300442

NUR ALIAH NADHIRA
300595

ANISA NADIAH
299892

Contents

- Introduction: Problem Statement, Chosen Methodology,
- Dataset Description
- Preprocessing Techniques
- Data Cleaning: Issues Identified, Handing the Issues, Result
- Data Transformation: Preprocessing, Result
- Modelling Process: Model Selected, Model Development
- Performance Evaluation
- Best Model Selection
- Conclusion and Recomendations

Problem Statement

The primary objective:

- to develop a predictive model capable of determining whether a patient is eligible to donate blood or not.
- framed as a binary classification problem
- classifies individuals into 'Blood Donor' or 'Non-Blood Donor' categories based on their demographic and biochemical attribute.

Chosen Methodology

Dataset Selection: The Hepatitis C Virus (HCV) dataset from Kaggle was chosen, containing 615 records and 14 attributes, suitable for binary classification.

Data Preprocessing:

- **Data Cleaning:** Missing values were imputed using the median, outliers were handled by a clipping method, restricting values within the 1st and 99th percentiles, and highly skewed numerical distributions underwent a log transformation (log1p).

- **Data Transformation:** Categorical variables, specifically 'Sex' and 'Category', were encoded. The 'Category' column was transformed into a binary 'Target' variable, 0 for blood donors, 1 for non-blood donors.

Chosen Methodology

Model Development: Logistic Regression was selected as the predictive model due to its suitability for binary classification tasks.

Model Evaluation: The model was trained and evaluated across various train-test split ratios (10:90 to 90:10). Performance was assessed using a suite of metrics including Accuracy, Precision, Recall, F1-Score, ROC-AUC, Confusion Matrix, Sensitivity, and Specificity.

Key Findings

- The model's consistent perfect precision, indicating no false positives.
- The 80:20 train-test split emerged as the best-performing configuration, achieving the **highest accuracy, perfect ROC-AUC, and the highest recall and sensitivity** among the evaluated splits, while maintaining **perfect precision and specificity**. This configuration provides the **optimal balance** for safety and effective screening in blood donation eligibility.

Dataset Description: Number of Features and Records

- contains 615 records (rows) and 14 attributes (columns).

Column	Value Type
category (**target**), sex	Categorical
index, age	Numerical (int)
ALB (Albumin), ALP (Alkaline Phosphatase), ALT (Alanine Aminotransferase), AST (Aspartate Aminotransferase), BIL (Bilirubin), CHE (Cholinesterase), CHOL (Cholesterol), CREA (Creatinine), GGT (Gamma-Glutamyl Transferase), PROT (Total Protein)	Numerical (float)

Preprocessing Techniques

Data Cleaning

conduct the data cleaning to address the missing values, outliers, and skewed distributions.
handle the issues using:

- missing values: impute with median
- outliers: clipping method
- skewed distributions: log transformation (log1p)

Data Transformation

conduct appropriate preprocessing techniques were applied to prepare the dataset for model development.

- standardization (RobustScaler method)
- categorical encoding

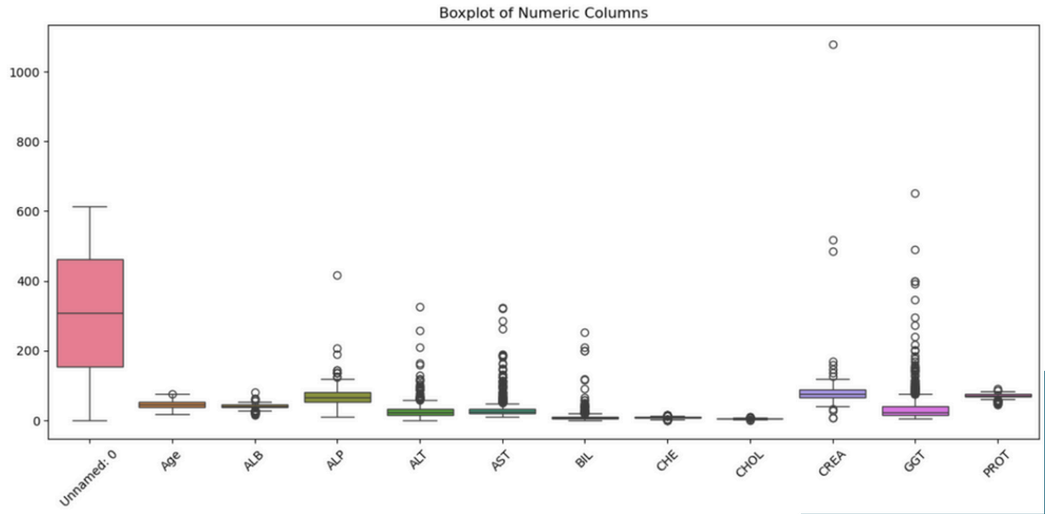
Data Cleaning: Issues Identified

```
Missing data count:
 Unnamed: 0     0
Category        0
Age             0
Sex             0
ALB             1
ALP            18
ALT             1
AST             0
BIL             0
CHE             0
CHOL           10
CREA            0
GGT             0
PROT            1
dtype: int64

Number of outliers in each column (Z-score > 3):
 Unnamed: 0     0
Age             0
ALB            13
ALP             3
ALT            10
AST            14
BIL             7
CHE             9
CHOL            7
CREA            3
GGT            10
PROT            9
dtype: int64
```

```
Skewness of each numeric feature:
        Feature   Skewness                Category
0          CREA  15.169291           Highly Skewed
1           BIL   8.385437           Highly Skewed
2           GGT   5.632734           Highly Skewed
3           ALT   5.506114           Highly Skewed
4           AST   4.940327           Highly Skewed
5           ALP   4.654921           Highly Skewed
6          CHOL   0.375828  Approximately Symmetric
7           Age   0.267134  Approximately Symmetric
8    Unnamed: 0   0.000000  Approximately Symmetric
9           CHE  -0.110233  Approximately Symmetric
10          ALB  -0.176768  Approximately Symmetric
11         PROT  -0.963687       Moderately Skewed

Number of features in each skewness category:
 Category
Highly Skewed              6
Approximately Symmetric    5
Moderately Skewed          1
Name: count, dtype: int64
```



Data Cleaning: Handling the Issues

Handle Missing Values:

- All missing values are numerical → suitable for imputation.
- Median imputation is used to reduce the influence of outliers (robust to extreme values).

Handle Outliers:

- Use clipping method (within 1st and 99th percentiles).
- To reduce the influence of extreme values.

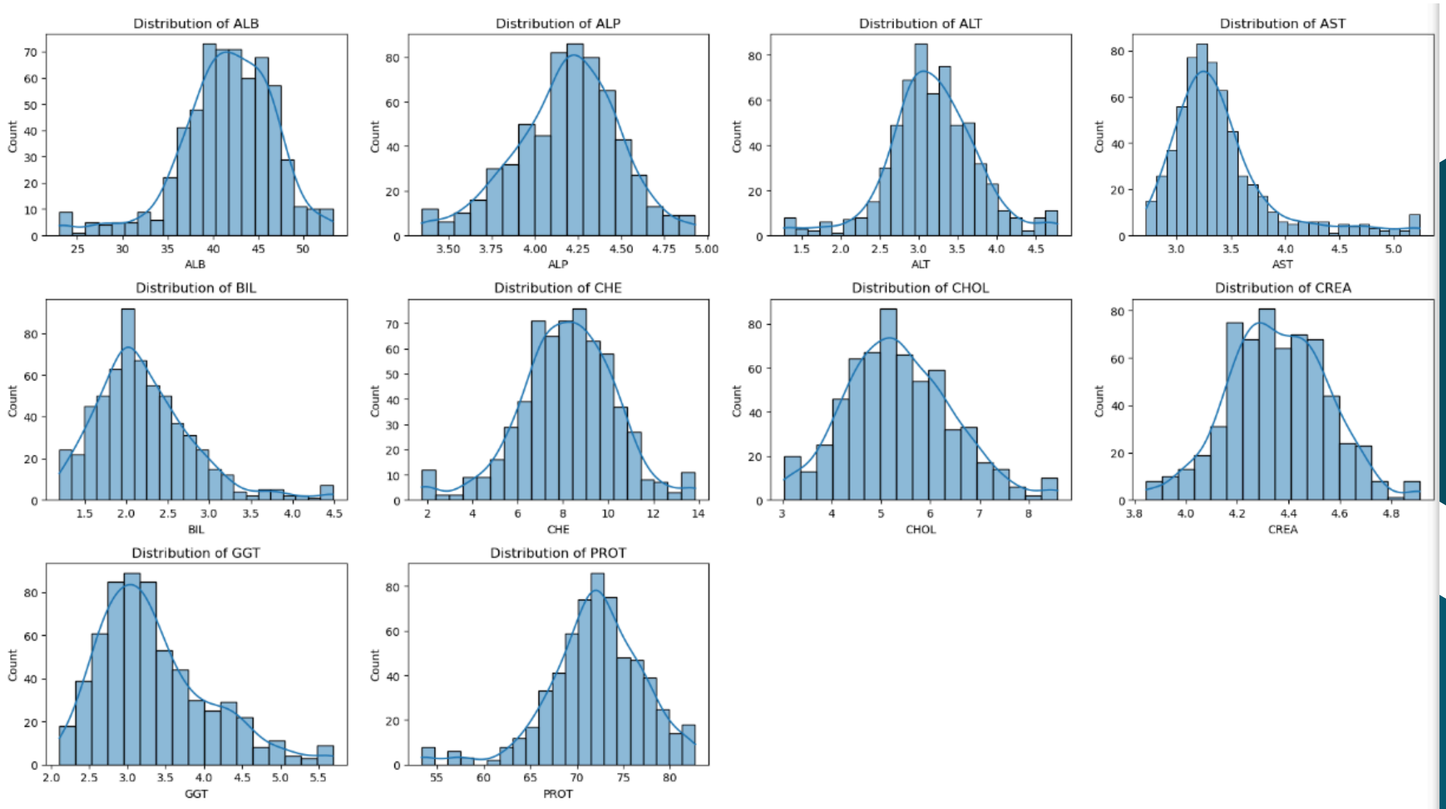
Handle Outliers:

- Applied log1p transformation to features with high skewness (skew > 1).
- Helps normalize the distribution → improves model performance.
- Six skewed features transformed; remaining features left untransformed to retain clinical interpretability.

Data Cleaning (Result)

```
Missing values after imputation:
 Unnamed: 0    0
Category       0
Age            0
Sex            0
ALB            0
ALP            0
ALT            0
AST            0
BIL            0
CHE            0
CHOL           0
CREA           0
GGT            0
PROT           0
dtype: int64

Skewness of numerical columns after transformation:
ALB    -0.817669
ALP    -0.317343
ALT    -0.166088
AST     1.815691
BIL     1.101718
CHE    -0.274983
CHOL    0.359162
CREA    0.042370
GGT     0.931163
PROT   -0.738850
dtype: float64
```



Data Transformation (Preprocessing)

Standardization with RobustScaler:

- Applied to all numerical columns.
- Improves model performance by ensuring consistent feature scaling.
- Its resistance to outliers, common in medical datasets.
- Helps deal with skewed distributions and extreme values.

Categorical Encoding:
Sex column

- Female (f) → 0
- Male (m) → 1

Target column (category):

- Blood Donor (0) → 0
- All others (0s, 1, 2, 3) → 1 (Non-Blood Donor)

Data Transformation (Result)

```
   Age  Sex       ALB       ALP       ALT       AST       BIL       CHE  \
0 -1.0    1 -0.539062 -0.573258 -1.511458 -0.375602  0.036028 -0.500942
1 -1.0    1 -0.539062  0.148909 -0.347973 -0.112550 -0.797445  1.096045
2 -1.0    1  0.773437  0.299475  0.652788  1.700326 -0.236290  0.218456
3 -1.0    1  0.195312 -0.596867  0.409771 -0.322789  1.323169 -0.350282
4 -1.0    1 -0.429687  0.279467  0.501181 -0.102973  0.370107  0.335217

       CHOL      CREA       GGT      PROT  Target
0 -1.442509  1.174605 -0.684213 -0.524590       0
1 -0.348432 -0.145732 -0.421995  0.704918       0
2 -0.069686  0.405751  0.378448  1.163934       0
3 -0.390244  0.140232  0.397707  0.573770       0
4 -0.682927 -0.047945  0.266082 -0.573770       0
Preprocessed dataset saved as 'Preprocessed_Dataset.xlsx'
```

Modelling Process

Model Selected

- Model selected to perform predictive analytics: Logistic Regression.
- Final target: Determine whether a patient can donate blood or not (binary classification).
- Logistic Regression suits this type of problem as it is designed to handle binary outcomes effectively.
- The outcome variable ‘Target’ column was derived by encoding the original ‘Category’ column.
- Values were set to 0 (blood donors), and 1 (non-blood donors).

Model Development

- Train-test ratios used: 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, 90:10, implemented to observe model's performance under various training data amount.
- Train: Fit the model, learn patterns from data.
- Test: Predicts and evaluate how model performs on unseen data.

Target Variable	'Target'
Features Selected	'Age', 'Sex', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE', 'CHOL', 'CREA', 'GGT', 'PROT'

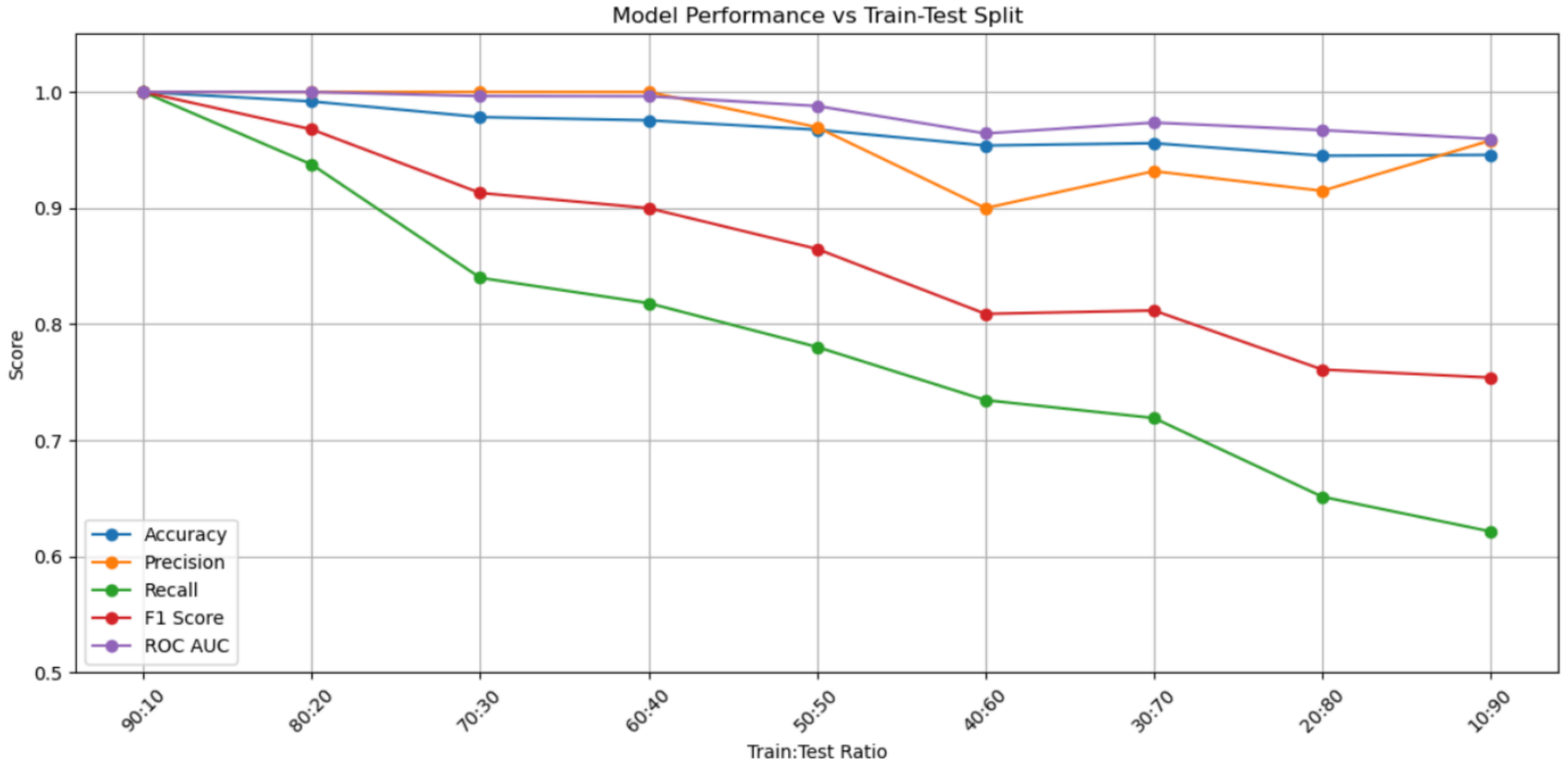
- Features were selected as they were relevant to liver function and blood chemistry (important to determine blood donor eligibility).

Model Development

Top 3 train-test split ratios:

- 80:20
- 70:30
- 60:40

** With Accuracy as main factor



Logistic Regression Performance Summary:

	Train:Test Ratio	Accuracy	Precision	Recall	F1 Score	ROC AUC
0	90:10	1.000000	1.000000	1.000000	1.000000	1.000000
1	80:20	0.991870	1.000000	0.937500	0.967742	1.000000
2	70:30	0.978378	1.000000	0.840000	0.913043	0.996500
3	60:40	0.975610	1.000000	0.818182	0.900000	0.996301
4	50:50	0.967532	0.969697	0.780488	0.864865	0.987942
5	40:60	0.953930	0.900000	0.734694	0.808989	0.964286
6	30:70	0.955916	0.931818	0.719298	0.811881	0.973637
7	20:80	0.945122	0.914894	0.651515	0.761062	0.967172
8	10:90	0.945848	0.958333	0.621622	0.754098	0.959600

Why 90:10 is not considered as top configuration:

- Small test size
- Model's performance overestimated
- Potential misclassification (e.g. ineligible as eligible donor)

Performance Evaluation

Train: Test Ratio	Accuracy	Precision	Recall	F1 Score	ROC-AUC
80:20	0.991870	1.000000	0.937500	0.967742	1.000000
70:30	0.978378	1.000000	0.840000	0.913043	0.996500
60:40	0.975610	1.000000	0.818182	0.900000	0.996301

80:20 Split

- **Accuracy:** The highest, most reliable predictions overall.
- **Recall:** Correctly identified 93.75% of non-blood donors.
- **ROC-AUC:** Flawless discrimination between classes.
- **F1-Score:** Optimal balance between precision and recall.

70:30 Split

- **Recall:** Missed 16% of non-blood donors.
- **F1 Score:** Still a strong performance
- **ROC-AUC:** Near-perfect discrimination

60:40 Split

- **Recall**: Missed 18.18% of non-blood donors.
- **F1 Score**: 90% correct at identifying non-blood donors while avoiding mistakes.
- **ROC-AUC**: Excellent at class seperation.

Precision

- **Perfect Across All Splits =** 1.00000
- **No false positive -** Every predicted "non-blood donor" was correct.
- Critical for healthcare application to avoid wrong classificaition.

Confusion Matrix

True Positive (TP)

- The model correctly predicted the positive class
- The model predicts someone is a blood donor and they actually are.

True Negative(TN)

- The model correctly predicted the negative class
- The model predicts someone is not a blood donor, and they acutally are not.

False Positive (FP)

- The model incorrectly predicted the positive class
- the model predicts someone is a blood donor but they are actually not.

False Negative (FN)

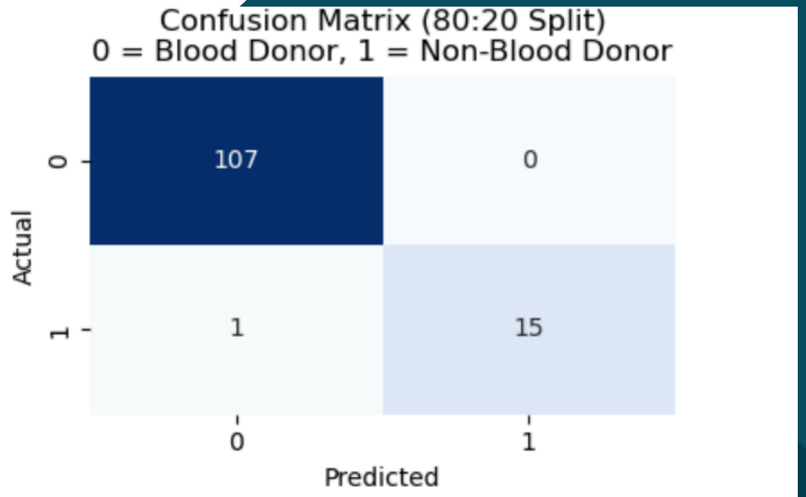
- The model incorrectly predicted the negative class
- The model predicts someone is not a blood donor but they actually are.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix (80:20)

- The model successfully identified 107 real donors. This shows that the model highly accurate in spotting people who donate blood.
- 15 people who do not donate blood were correctly labeled. This is very important for correctly spotting who should not be classified as donor.
- 1 person was misclassified which is very low. This means that very few errors when predicting blood donors.
- False Negative = 0, which is perfect. This means that the model did not miss any real blood donors.

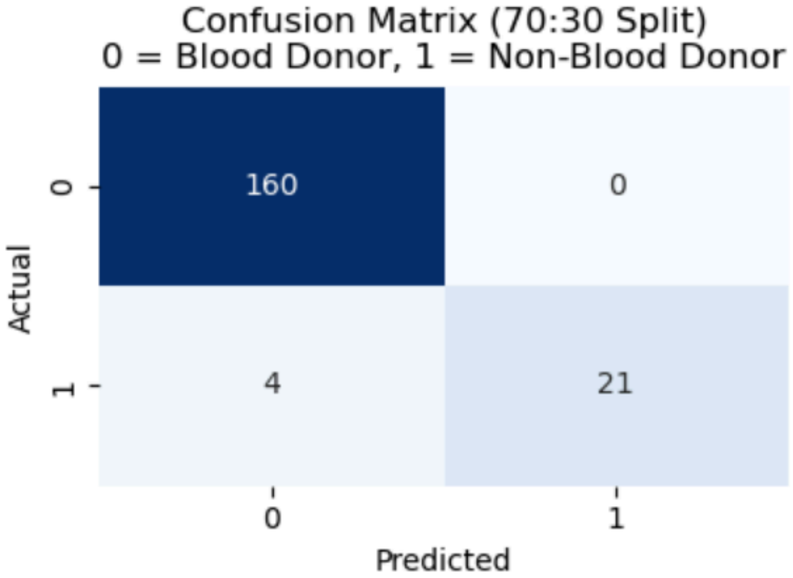
- This split gave the best overall performance and correctly identified all blood donors and almost perfectly identified non-blood donors.



Confusion Matrix (70:30)

- 160 people who donate blood were correctly identified
- The model correctly label 21 people as non-blood donors.
- 4 people who do not donate blood were misclassified as donors. This is a problem when we want to identify all non-blood donors accurately.
- FN = 0 means that all real blood donors were detected correctly.
- its ability to identify all non-donors is a bit weaker here than in the 80:20 split.

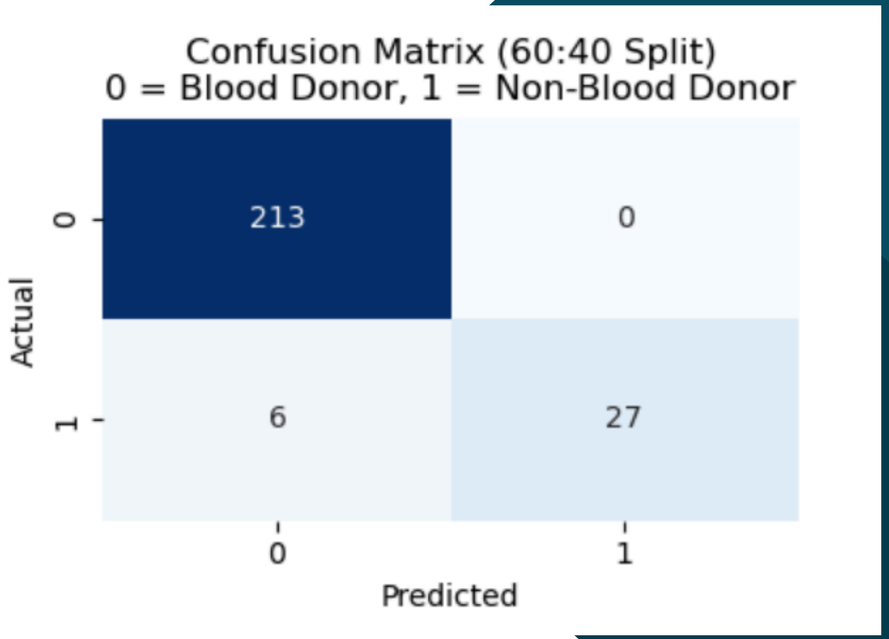
- Overall, it is still a good model, but some non-blood donors are getting missed.



Confusion Matrix (60:40)

- Total of 213 blood donors were correctly identified by the model
- No blood donors were incorrectly predicted as non-blood donors.
- The model correctly identified 27 non-blood donors.
- 6 non-blood donors were wrongly predicted as blood donors

- This split struggled the most at detecting non-blood donors.
- Less Effective among the other splits.



Sensitivity and Specificity

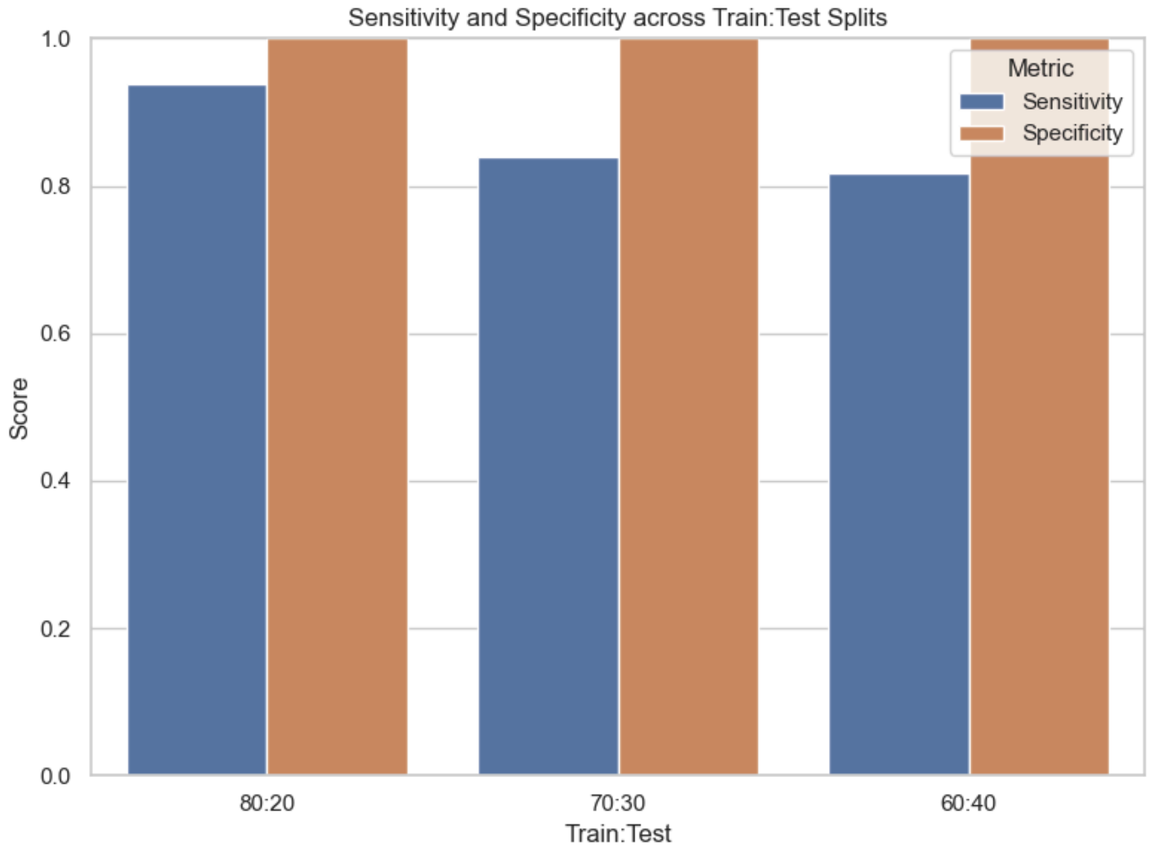
Specificity

- Perfect across all splits = 1.0
- Means: Correctly identifying blood donors.
- The model never wrongly predicted a blood donor as non-donor.
- This also means there were no false positive in any splits.

Sensitivity (Recall)

- Sensitivity shows how well the model finds non-blood donors.
- High Sensitivity = Few missed non-blood donors.
- Decreased slightly as the test size got bigger.

80:20 split gives the best balance, with high accuracy for both groups.



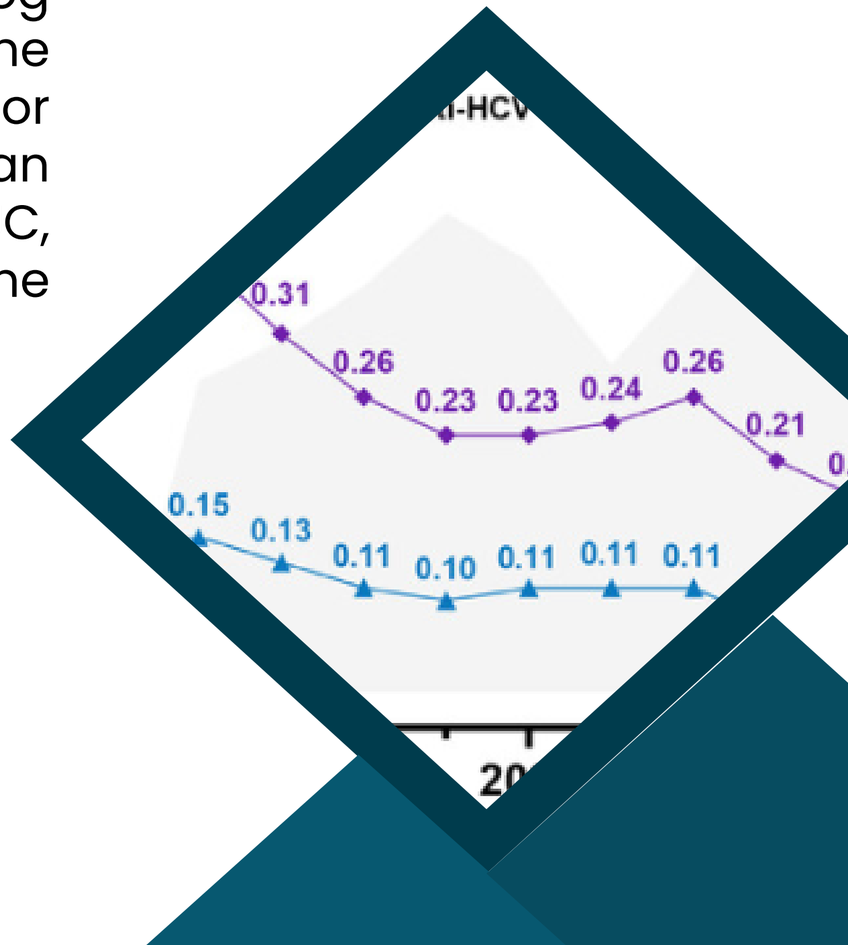
Best Model Selection

The most effective model configuration, based on a comprehensive evaluation of various metrics, is achieved with the **80:20 Split**.

- **Highest Overall Accuracy:** The model was correct 99.2% of the time.
- **Perfect Precision (No False Positives):** Every person the model identified as a 'Blood Donor' was truly a blood donor.
- **Best Recall for Non-Donors:** The model correctly identified nearly 94% of people who were not blood donors.
- **Excellent Discrimination (ROC-AUC):** The model is exceptional at telling the difference between blood donors and non-donors.
- **Perfect Specificity (No False Negatives for Donors):** All actual blood donors were correctly identified.
- **Optimal Balance:** Provided the best balance

Conclusion

A Logistic Regression model was successfully developed to predict blood donation eligibility using the HCV dataset. Key preprocessing steps included median imputation, outlier clipping, log transformation, and encoding of categorical variables. The 'Category' was converted into a binary 'Target' for donor classification. RobustScaler improved resilience to outliers, and an 80:20 train test split yielded excellent accuracy, perfect ROC-AUC, and precision vital for safe donor screening. This highlights the model’s strong performance and practical value.



Few recommendations to further enhance the model's capabilities:

- **Explore advanced models.**
  - Test models like Random Forests, SVMs, and Gradient Boosting to better handle skewed features (AST, BIL) and improve prediction accuracy by capturing complex patterns.
- **Acquire a larger and more diverse dataset.**
  - Increase dataset size and diversity to improve model generalizability and robustness across different populations and medical contexts.
- **Implement a real-world validation and continuous monitoring system.**
  - Validate the model in clinical settings and set up continuous monitoring to maintain accuracy, adapt to new data, and align with evolving medical guidelines.

Thank You