STSCI5111 Final Project Report
Patient Administration Prediction Using Blood Test Result

Zixiao Wang (zw699), Suphakrit Lertkitcharoenvong (sl3355)

1 Introduction

In the modern healthcare system, efficient patient management is crucial for ensuring optimal health outcomes and effective allocation of resources within hospitals. A key aspect of patient assessment involves the analysis of blood test results, which can provide vital insights into a variety of health conditions and help determine the appropriate level of care required. This report presents a study that aimed to develop a predictive model capable of accurately classifying whether patients should be admitted to the hospital as inpatients or treated within the outpatient department based on their blood test results from a private hospital in Indonesia. The study employed exploratory data analysis, preprocessing techniques, and evaluated several machine learning algorithms including logistic regression, support vector machines, k-nearest neighbors, and random forest models. The random forest model demonstrated the best performance with 77% accuracy, 76% precision, and 75% recall, though this still falls short of the desired 85%+ accuracy for clinical applications. While promising, the findings suggest additional data sources and more advanced modeling techniques may be needed to further improve predictive performance and ensure safe, reliable integration into healthcare workflows for making critical patient management decisions.

2 Background

In the healthcare system, efficient patient management is crucial to ensuring optimal health outcomes and resource allocation within hospitals. An integral part of patient assessment involves the analysis of blood test results, which can indicate a variety of health conditions and dictate the required level of care. For example, in the United States, blood tests are routinely used to make critical decisions regarding patient care. A real-world instance of this can be observed in the management of patients with suspected anemia. Hematocrit and hemoglobin levels are assessed to diagnose the condition and its severity. If a patient's hemoglobin level is significantly low, indicating severe anemia, they may require inpatient care for blood transfusions and close monitoring. Conversely, a mild case might be managed on an outpatient basis with oral supplements and follow-up tests.

2.1 Problem Statement

Determining whether a patient needs inpatient care based on blood test results is a significant challenge in modern healthcare. Hospitals are constantly seeking efficient ways to decide if a patient should be admitted as an inpatient or treated in the outpatient department. Blood tests, which provide vital information about a patient's health status, can be instrumental in making these decisions. In the dataset used from a private hospital in Indonesia, the focus is on laboratory test results including hematocrit, hemoglobin, erythrocytes, leukocytes, and thrombocytes. These measures are crucial for diagnosing various conditions and determining patient care requirements. For instance, an elevated leukocyte count may suggest an infection that requires immediate and intensive treatment, potentially justifying inpatient care. Similarly, abnormalities in thrombocyte counts can indicate bleeding disorders or other hematological issues that necessitate close medical supervision.

The challenge lies in developing a system or model that can accurately interpret these lab results and recommend the most appropriate care setting—whether inpatient or outpatient. Such a system must be reliable, given the potential consequences of under-treatment (such as worsening of the patient's condition) or over-treatment (such as unnecessary hospitalization).

2.2 Importance of Predicting Inpatient Care Needs

1. Enhanced Patient Care: The ability to predict whether a patient requires inpatient care from blood test results can drastically improve the quality of care provided. By identifying patients who need immediate and intensive treatment, healthcare providers can prioritize care delivery, ensuring that those in urgent.
2. Reduced Hospital Congestion: Accurately classifying patients based on their need for inpatient care helps in managing hospital congestion effectively. With better prediction, hospitals can optimize the usage of limited resources such as beds and medical kits.
3. Cost Savings: Hospital stays are considerably more expensive than outpatient care due to the intensive use of resources, including specialized medical staff, equipment, and facilities. By correctly determining which patients can be effectively treated in an outpatient setting, healthcare providers can significantly cut costs.

2.3 Objective

The primary objective of the study detailed in the document is to develop a predictive model that can accurately classify whether patients undergoing basic laboratory tests should be admitted to the hospital (incorporated as inpatients) or can be treated within the outpatient department of the hospital. This classification is crucial as it aids in optimizing patient management, ensuring that hospital resources are allocated efficiently, and improving patient outcomes by quickly identifying those who need intensive care.

Given the critical nature of decisions in the medical industry, the predictive model needs to achieve a high level of accuracy. In medical settings, especially where patient care and resource allocation depend heavily on the outcomes predicted by the model, an accuracy of more than 85% is considered necessary to ensure reliability and trustworthiness. Models with lower accuracy levels might lead to misclassifications, which can have serious implications, such as unnecessary hospitalizations or inadequate care for those who need immediate inpatient treatment. By utilizing advanced machine learning algorithms, the project aims to automate and enhance the decision-making process in hospital settings. Achieving and exceeding the 85% accuracy threshold is critical for the adoption of such models in live clinical environments.

3 Exploratory Data Analysis, Statistical Testing, and Data Cleaning

3.1 Data Introduction

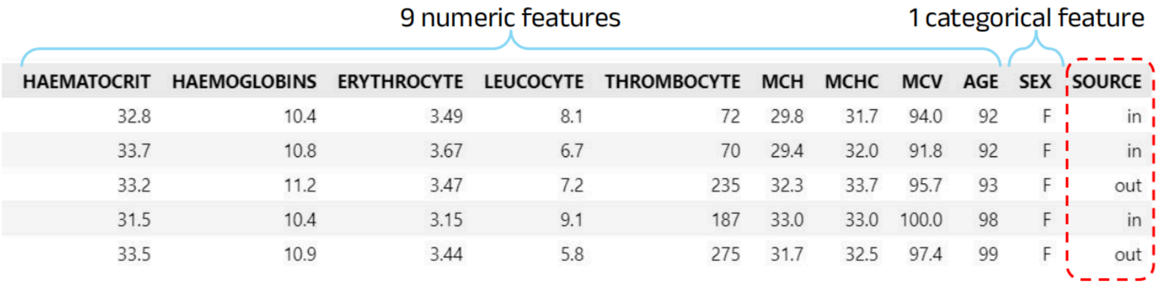
3.1.1 Data Collection

The dataset used in the study consists of Electronic Health Records (EHR) from a private hospital in Indonesia, focusing on laboratory test results. These tests include measures like hematocrit, hemoglobins, erythrocytes, leukocytes, and thrombocytes, which are crucial for diagnosing various conditions and determining patient care requirements.

In the study presented in the EHC classification report, data handling is a crucial process that prepares the dataset for effective model training and validation. The dataset primarily consists of laboratory test results which are fundamental to the study's objective of predicting hospital admission.

It's noted that the dataset contains 4512 rows and a mix of numeric and categorical features. The target labels are encoded as '0' for outpatient and '1' for inpatient, indicating whether a patient should be admitted.

Access: data is attached in the zip file with the name ‘data-ori.csv’. It can be accessed directly as is shown in the Python code.


9 numeric features
1 categorical feature

HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	AGE	SEX	SOURCE
32.8	10.4	3.49	8.1	72	29.8	31.7	94.0	92	F	in
33.7	10.8	3.67	6.7	70	29.4	32.0	91.8	92	F	in
33.2	11.2	3.47	7.2	235	32.3	33.7	95.7	93	F	out
31.5	10.4	3.15	9.1	187	33.0	33.0	100.0	98	F	in
33.5	10.9	3.44	5.8	275	31.7	32.5	97.4	99	F	out

Figure 1. Data set example

3.1.2 Variable Definition

The dataset includes several key laboratory measurements that are significant indicators of a patient's health status. From Figure 1 showing an example of a data set, Each variable was selected based on its medical relevance and predictive value in assessing the need for inpatient care:

1. Haematocrit (HCT): This measures the percentage of red blood cells in the blood. A critical measure as it can indicate conditions like anemia or polycythemia, affecting the decision on the level of care needed.
2. Haemoglobins (HGB): This involves the level of hemoglobin in the blood, crucial for carrying oxygen throughout the body. Abnormal levels can signify serious health issues that might require inpatient care.
3. Erythrocyte (RBC Count): The count of red blood cells, which is vital for diagnosing anemias and other blood disorders.

4. Leucocyte (WBC Count): White blood cell count, indicative of the body's immune response. High or low levels can indicate infections or immune disorders, potentially requiring more intensive hospital care.
5. Thrombocyte (Platelet Count): Platelet count is important for blood clotting. Low levels can lead to bleeding disorders, whereas high levels might indicate thrombotic risks, both possibly necessitating inpatient care.
6. Mean Corpuscular Hemoglobin (MCH): This reflects the average amount of hemoglobin per red blood cell, useful in diagnosing different types of anemia.
7. Mean Corpuscular Hemoglobin Concentration (MCHC): This measures the concentration of hemoglobin in a given volume of red blood cells, providing insights into the hemoglobin status and possible anemias.
8. Mean Corpuscular Volume (MCV): The average size of red blood cells, which helps in classifying anemias.
9. Age: Patient's age
10. Sex: Patient's gender

3.1.3 Variable Implication

The implications for each blood result are described below:

1. Haematocrit (HCT)
    - Role: Haematocrit measures the percentage of red blood cells in the blood, reflecting blood viscosity and oxygen-carrying capacity.
    - Lower Levels: Low HCT values can indicate anemia, potentially caused by nutritional deficiencies, chronic disease, or blood loss. Low levels can lead to fatigue, weakness, and poor oxygenation of tissues.
    - Higher Levels: Elevated HCT can occur in conditions like polycythemia vera or as a response to chronic hypoxia. High levels increase blood viscosity, raising the risk of thrombosis and cardiovascular events, necessitating careful monitoring and potential therapeutic intervention.
2. Haemoglobins (HGB)
    - Role: Hemoglobin is a protein in red blood cells responsible for oxygen transport.
    - Lower Levels: Lower hemoglobin levels are indicative of anemia, leading to symptoms such as pallor, shortness of breath, and tachycardia. Treatment may involve iron supplementation, blood transfusions, or treatment of the underlying cause.
    - Higher Levels: High hemoglobin levels might be due to dehydration, smoking, or bone marrow disorders, increasing the risk for blood clots and strokes. Management might include hydration, phlebotomy, or addressing the underlying causes.
3. Erythrocyte (RBC Count)
    - Role: Red blood cells carry oxygen from the lungs to the body's tissues and bring carbon dioxide back to the lungs.
    - Lower Levels: Decreased RBC count is a primary indicator of anemia and can lead to reduced oxygen transport, requiring treatments ranging from dietary adjustments to more complex medical interventions.
    - Higher Levels: High RBC counts can lead to polycythemia, which can increase blood thickness and slow blood flow, potentially leading to complications such as thrombosis.
4. Leucocyte (WBC Count)
    - Role: White blood cells are part of the body's immune system, helping fight infections.
    - Lower Levels: Low WBC counts (leukopenia) can make a person more prone to infections and might be a side effect of treatments like chemotherapy or a sign of bone marrow suppression.

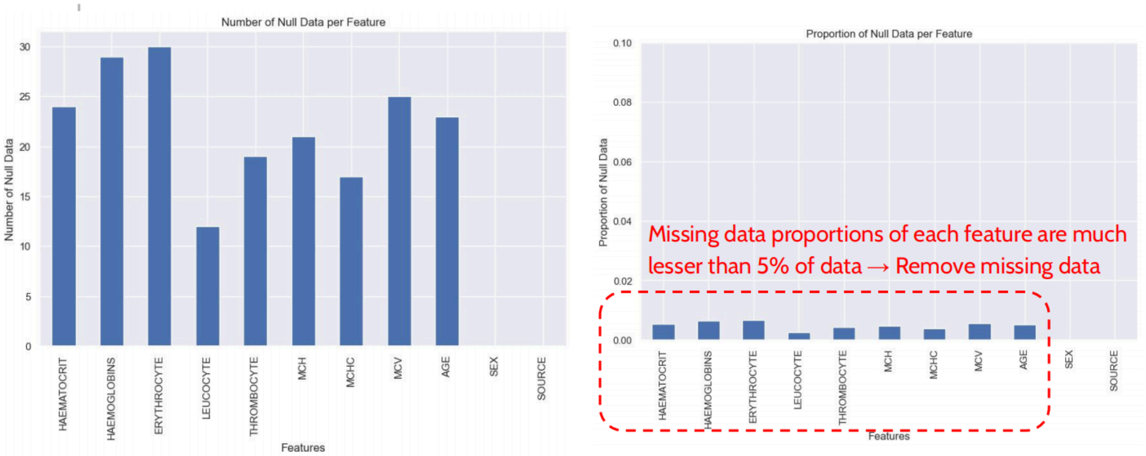
- Higher Levels: Elevated WBC counts may indicate an infection, inflammatory disease, stress, or more serious conditions like leukemia, requiring specific treatments based on the underlying cause.

5. Thrombocyte (Platelet Count)
    - Role: Platelets are crucial for blood clotting and wound repair.
    - Lower Levels: Thrombocytopenia can result in easy bruising and bleeding, serious cases of which might need treatments like platelet transfusion or steroids.
    - Higher Levels: High platelet counts, or thrombocytosis, can lead to unnecessary clot formation, which may require antiplatelet medications or treatment of the underlying condition.
6. Mean Corpuscular Hemoglobin (MCH)
    - Role: MCH measures the average amount of hemoglobin per red blood cell.
    - Lower Levels: Low MCH values can indicate hypochromic anemias, where red blood cells do not have enough hemoglobin, often treated with iron supplements.
    - Higher Levels: Elevated MCH values can indicate hyperchromic anemias, such as vitamin B12 deficiency or macrocytic anemia, requiring specific vitamin supplementation or dietary changes.
7. Mean Corpuscular Hemoglobin Concentration (MCHC)
    - Role: MCHC measures the concentration of hemoglobin in a given volume of red blood cells, indicating the hemoglobin content in RBCs.
    - Lower Levels: Low MCHC values can suggest hypochromic anemias, where cells lack sufficient hemoglobin, addressed typically by treating the underlying cause.
    - Higher Levels: High MCHC can occur in hereditary spherocytosis or other conditions leading to hyperdense cells, which may require interventions such as folic acid supplementation or even splenectomy in severe cases.
8. Mean Corpuscular Volume (MCV)
    - Role: MCV measures the average volume of red blood cells, useful in classifying anemias.
    - Lower Levels: Low MCV indicates microcytic anemia, commonly due to iron deficiency, and treatment involves iron supplementation.
    - Higher Levels: High MCV indicates macrocytic anemia, often due to vitamin B12 or folate deficiency, treated with nutritional supplementation and dietary management.

3.2 Exploratory Data Analysis (EDA) and Statistical Testing

3.2.1 Data Completeness Analysis

The presentation in Figure 2 identifies that the dataset includes rows with missing data (NA values). A total of 100 rows with missing data are noted, and the decision is made to remove these rows to ensure the quality and reliability of the models as the total missing data are less than 2% of the total data and missing values in each feature are less than 5% of the total data.

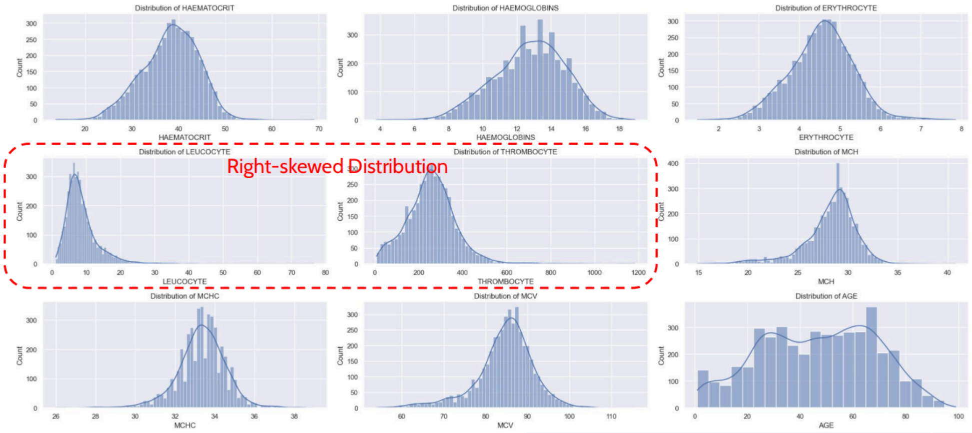


Missing data proportions of each feature are much lesser than 5% of data → Remove missing data

Figure 2. Data missing count (left) and percentage of data missing by each feature (right)

3.2.2 Numerical Analysis

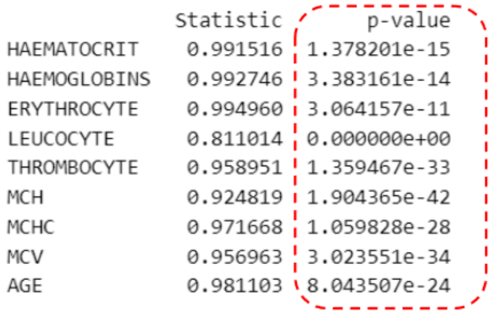
1. Distribution Analysis: Histograms and Kernel Density Estimate (KDE) plots are used to assess the distribution of each numerical feature. It's noted that several features exhibit right-skewed distributions, including LEUCOCYTE and THROMBOCYTE, which might require transformation to normalize the data.



Right-skewed Distribution

Figure 3. Numerical value distribution. Highlighted is the right-skewed distribution in LEUCOCYTE and THROMBOCYTE variables

2. Normality Tests: The Shapiro-Wilk test is conducted to test the normality of the data distributions. The results lead to the rejection of the null hypothesis for all numeric variables, indicating that they do not follow a normal distribution. This insight prompts considerations for data transformations to achieve normality.



```
             Statistic       p-value
HAEMATOCRIT   0.991516  1.378201e-15
HAEMOGLOBINS  0.992746  3.383161e-14
ERYTHROCYTE   0.994960  3.064157e-11
LEUCOCYTE     0.811014  0.000000e+00
THROMBOCYTE   0.958951  1.359467e-33
MCH           0.924819  1.904365e-42
MCHC          0.971668  1.059828e-28
MCV           0.956963  3.023551e-34
AGE           0.981103  8.043507e-24
```

Figure 4. Shapiro-Wilk normality test suggests all numeric variables are not normally distributed

3. QQ Plots: These plots are used to further analyze the normality of the data. The QQ plots confirm the results from the Shapiro-Wilk test, showing heavy-tailed distributions that deviate from normality.

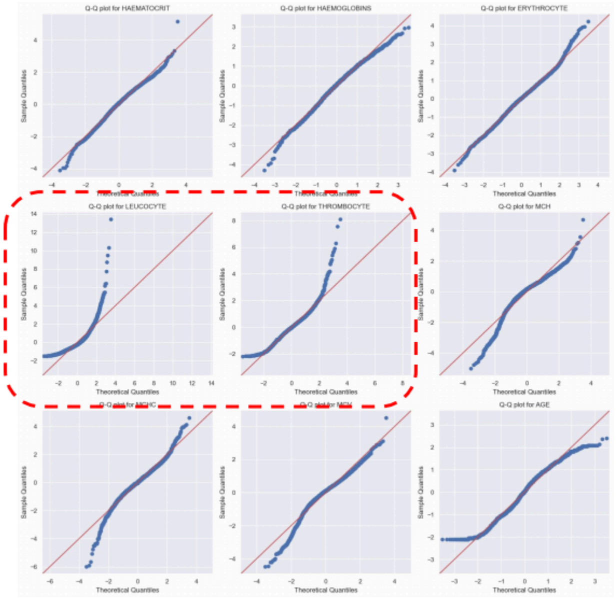
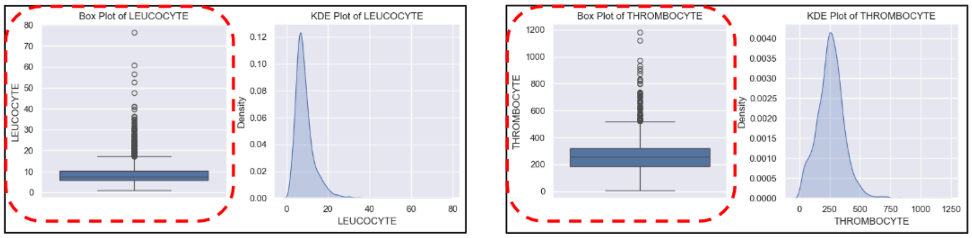


Figure 5. Q-Q plots showing evidence of right skewness as highlighted. Also, notice the heavy tail effect in other plots

4. Outlier Analysis: The presentation discusses the identification of outliers through KDE plots, noting that the max values for some features are significantly higher than the mean, suggesting potential outliers or extreme values that could impact model performance.



Max value is much higher than the mean

Figure 6. Box plots showing the existence of outliers in LEUCOCYTE and THROMBOCYTE

3.2.2 Correlation Analysis

In the correlation analysis conducted on the dataset (see Figure 7), various relationships between features were examined using correlation heatmaps and bivariate correlations.

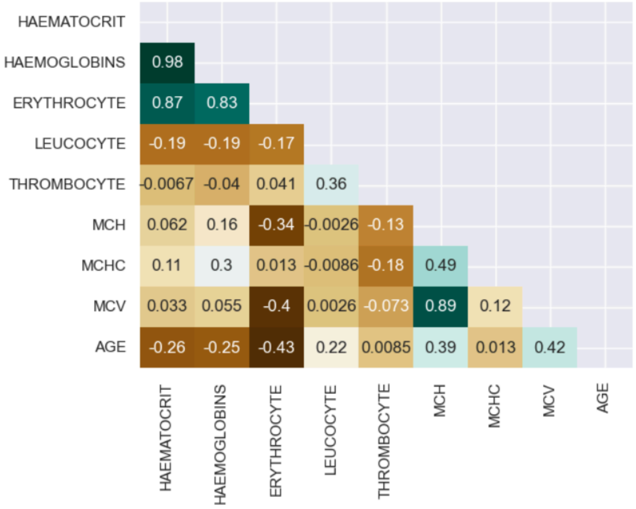


Figure 7. Heatmap of correlation

Here's a breakdown of the findings and decisions made based on positive, negative, and other correlations:

Positive Correlation:

1. HAEMATOCRIT (RBC volume) with HAEMOGLOBINS (RBC counts): These two features highly correlate, which is expected as they are both indicators of red blood cell (RBC) health and functionality. Given their high correlation and similar definitions, it was decided to remove one of the correlated features, in this case, HAEMATOCRIT, to simplify the model and potentially enhance its performance.
2. MCV (Avg vol of RBC) with MCH (Avg counts of hemoglobin): Similarly, MCV and MCH highly correlate, which is logical as both metrics provide insights into the size and content of RBCs. As with HAEMATOCRIT and HAEMOGLOBINS, one of these correlated features, MCH, was removed to streamline the model.

Negative Correlation:

1. ERYTHROCYTE and MCV: A negative correlation exists between erythrocyte count and mean corpuscular volume (MCV). This implies that as the erythrocyte count increases, MCV tends to decrease. This relationship is indicative of microcytic anemias, where higher RBC counts are associated with smaller cell sizes.
2. AGE and ERYTHROCYTE: Another negative correlation is observed between age and erythrocyte counts. This suggests that erythrocyte counts tend to decrease with age, which could have clinical relevance, especially in the context of age-related health conditions.

Other Correlations:

1. LEUKOCYTE and THROMBOCYTE: A moderate positive association exists between leukocyte and thrombocyte counts, indicating simultaneous responses to certain conditions such as infections or inflammations.
2. AGE and MCV: A positive correlation is observed between age and mean corpuscular volume (MCV), suggesting potential changes in red blood cell size with aging.

In the correlation analysis, it was decided to remove the HAEMATOCRIT and MCH columns due to their high correlation with other features and to simplify the model. However, it's noteworthy that despite the high correlation between HAEMOGLOBINS and ERYTHROCYTE (RBC), these features were retained in the analysis. This decision was based on the understanding that they represent different particles and serve distinct physiological functions.

Although HAEMOGLOBINS and ERYTHROCYTE (RBC) are closely related as components of blood, they represent different aspects of blood composition. HAEMOGLOBINS measure the concentration of hemoglobin in the blood, which is responsible for carrying oxygen to tissues and organs. ERYTHROCYTE (RBC) counts, on the other hand, indicate the number of red blood cells present in the blood, which is crucial for oxygen transport and overall blood volume. They can provide complementary information about a patient's health status. For instance, while high levels of both HAEMOGLOBINS and ERYTHROCYTE (RBC) may suggest a state of polycythemia, there are cases where the relationship between these two features can be altered due to various physiological factors or underlying health conditions. For example, there may be instances where a patient exhibits high levels of HAEMOGLOBINS but low ERYTHROCYTE (RBC) counts or vice versa.

3.2.3 Categorical Analysis

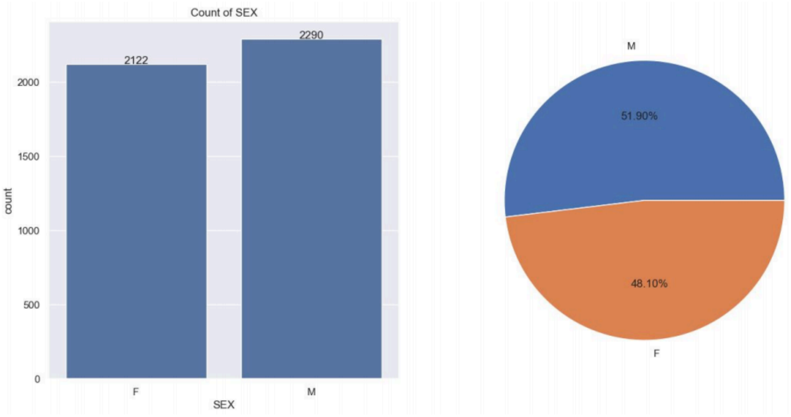


Figure 8. SEX variable analysis

The distribution of the categorical variable (sex) is examined to ensure there's no significant imbalance that could bias the classification models. The presentation notes a balanced distribution between male and female categories.

3.2.4 Label Analysis

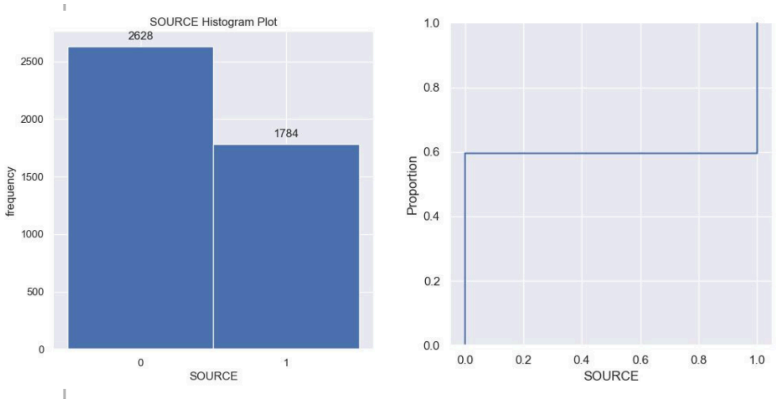


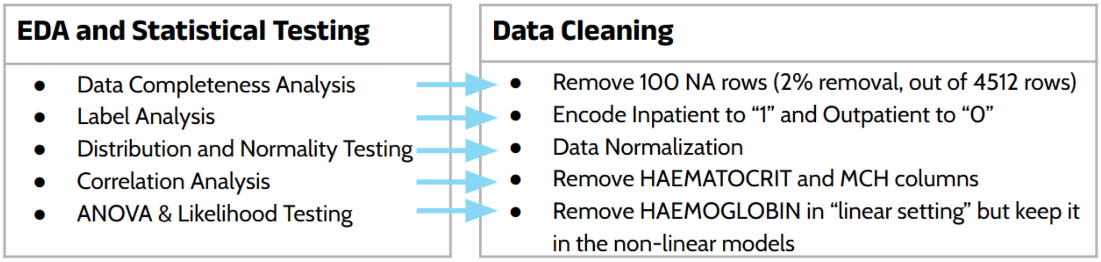
Figure 9. Label distribution

The distribution of the target variable ('SOURCE') is analyzed to check for class imbalance, which could influence model performance and evaluation metrics. A slight imbalance is noted with a 60:40 distribution but is considered mild enough not to require remedial measures.

3.3 Data Cleaning

From the EDA performed previous parts, we clean the data accordingly. Figure 10 shows the summary of data cleaning from the EDA process.

1. Handling Missing Values:Techniques such as imputation or removing rows with missing data to ensure the model's performance isn't adversely affected by gaps in the data.
2. Encoding Categorical Variables: Since machine learning models require numerical input, categorical variables like sex (male, female) were transformed into numerical codes.
3. Feature Scaling: To ensure that no variable dominates the others due to its scale, methods like Min-Max scaling, Z-score normalization, or Robust Scaling were applied. This standardization helps in optimizing the performance of algorithms that are sensitive to the scale of the data, such as SVM and K-nearest neighbors.
4. Column Removal: Remove HAEMATOCRIT and MCH columns due to their correlation



EDA and Statistical Testing	**Data Cleaning**
● Data Completeness Analysis	● Remove 100 NA rows (2% removal, out of 4512 rows)
● Label Analysis	● Encode Inpatient to “1” and Outpatient to “0”
● Distribution and Normality Testing	● Data Normalization
● Correlation Analysis	● Remove HAEMATOCRIT and MCH columns
● ANOVA & Likelihood Testing	● Remove HAEMOGLOBIN in “linear setting” but keep it in the non-linear models

Figure 10. Summarization of data cleaning from actions obtained in the EDA and statistical testing. Note: ANOVA and likelihood testing will be described in next section

4 Modeling

4.1 Variable Selection

When predicting patient admission based on blood test results, a detailed statistical analysis was conducted to determine the significance and impact of including the hemoglobin variable in the predictive model. This analysis focused on assessing whether hemoglobin levels vary significantly by categories such as sex and source (inpatient vs. outpatient), and whether including this variable improves the model's predictive power. Here's how these analyses were conducted and interpreted:

4.1.1 ANOVA Testing

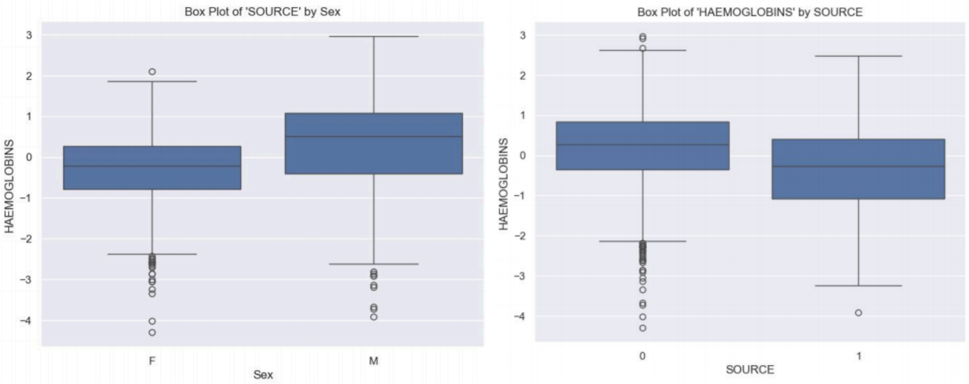
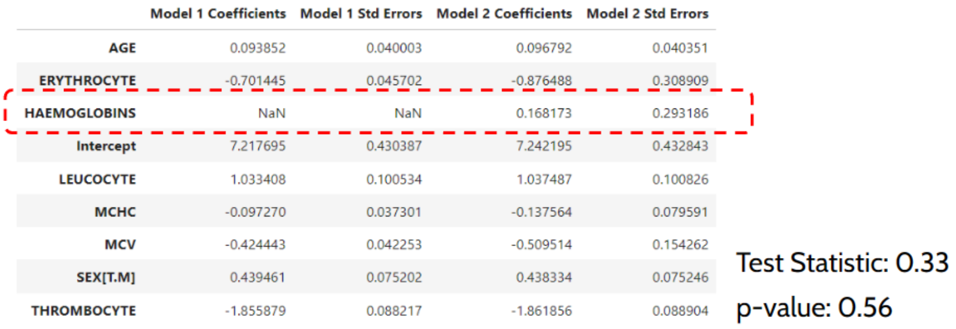


Figure 11. ANOVA testing result

- HAEMOGLOBINS vs. SEX:
  Hypothesis: The null hypothesis (H0) posited that there is no difference in mean hemoglobin levels across different sex categories.
  Results: The F-value was extremely high at 462.23 with a P-value of 1.37e-97, strongly rejecting the null hypothesis. This indicates significant variations in hemoglobin levels between different sex categories.
- HAEMOGLOBINS vs. SOURCE:
  Hypothesis: The null hypothesis (H0) stated that there is no difference in mean hemoglobin levels between patients categorized as inpatients and outpatients.
  Results: The F-value was 308.75 with a P-value of 7.40e-67, which also led to the rejection of the null hypothesis, showing significant differences in hemoglobin levels based on patient admission status.

4.1.2 Likelihood Ratio Test

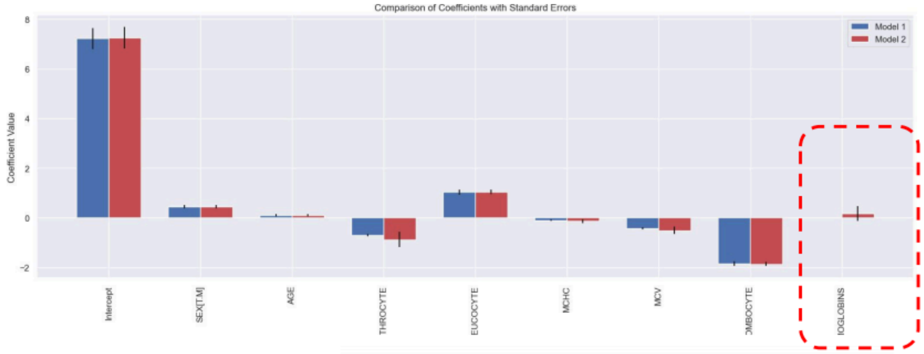


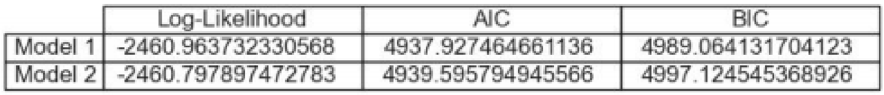
	Model 1 Coefficients	Model 1 Std Errors	Model 2 Coefficients	Model 2 Std Errors
AGE	0.093852	0.040003	0.096792	0.040351
ERYTHROCYTE	-0.701445	0.045702	-0.876488	0.308909
HAEMOGLOBINS	NaN	NaN	0.168173	0.293186
Intercept	7.217695	0.430387	7.242195	0.432843
LEUCOCYTE	1.033408	0.100534	1.037487	0.100826
MCHC	-0.097270	0.037301	-0.137564	0.079591
MCV	-0.424443	0.042253	-0.509514	0.154262
SEX[T.M]	0.439461	0.075202	0.438334	0.075246
THROMBOCYTE	-1.855879	0.088217	-1.861856	0.088904

Test Statistic: 0.33

p-value: 0.56

Figure 12. Likelihood ratio test coefficients





	Log-Likelihood	AIC	BIC
Model 1	-2460.963732330568	4937.927464661136	4989.064131704123
Model 2	-2460.797897472783	4939.595794945566	4997.124545368926

Figure 13. Likelihood ratio test coefficient plots and information criteria

This test assessed whether adding hemoglobins to the predictive model significantly improved its ability to explain variations in patient admission status (SOURCE). From Figure 12, the test statistic was 0.33 with a P-value of 0.56, and with higher AIC and BIC values than the original model as seen in Figure 13, indicating that the inclusion of hemoglobins does not significantly improve the model's explanatory power.

Interpretation and Conclusion

The findings from the ANOVA tests indicate that hemoglobin levels are indeed an important factor in differentiating between categories of sex and source, suggesting that they could be a critical variable in understanding patient conditions. However, the likelihood ratio test suggests that its inclusion in the logistic regression model does not significantly enhance the model's predictive accuracy.

Given these results, several conclusions and actions were considered:

- Statistical Significance vs. Predictive Power: While hemoglobin levels show statistical significance across categories, their contribution to improving the predictive power of the model for patient admission is limited. This distinction is crucial in model building where the focus is on variables that offer predictive utility.
- Model Simplification: Removing hemoglobins from the model might be considered to simplify the model without sacrificing accuracy. This can lead to a more streamlined model that is easier to interpret and manage.
- Clinical Relevance: Despite its limited impact on predictive performance, the clinical significance of hemoglobin levels cannot be overlooked. It may still be used as a

supplementary variable for clinical assessments and decisions outside of the predictive modeling framework.

4.2 Binary Classification Algorithms

In this section, we describe the algorithm used for this binary classification problem of “Given features X (blood particle counts, blood concentration & volume, and patient factors), can we correctly classify their label y into either 0 (outpatient) or 1 (inpatient)” and also the motivation for selecting these algorithms

4.2.1 Logistic Regression

Advantages:

- Simplicity and Interpretability: Logistic Regression is straightforward to implement, interpret, and explain, making it a good baseline for binary classification problems.
- Efficiency: It is computationally less intensive, making it suitable for situations where speed is a factor.

Disadvantages:

- Assumption of Linearity: Assumes a linear relationship between the independent variables and the logit of the dependent variable, which might not hold in complex scenarios like medical diagnostics.
- Performance Limitations: Can underperform when relationships are non-linear and in cases where feature interactions are important.

4.2.2 Support Vector Machine (SVM)

Advantages:

- Effectiveness in High-Dimensional Spaces: SVM works well in complex decision spaces where the number of dimensions exceeds the number of samples.
- Versatility: The kernel trick helps in solving any complex problem by transforming the linear inseparable data to linearly separable ones in higher dimensions.

Disadvantages:

- Scalability and Speed: SVMs can be computationally intensive, especially with large datasets, making them less scalable and slower in training.
- Kernel Dependency: The choice of kernel and its parameters can have a significant impact on the performance of the SVM model, requiring careful tuning.

4.2.3 K-Nearest Neighbors (KNN)

Advantages:

- Simplicity: KNN is very simple and easy to implement.
- No Assumptions on Data: Non-parametric, which means it does not make any assumptions about the underlying data distribution.

Disadvantages:

- Computationally Intensive: As the dataset grows, the efficiency or speed of the algorithm declines as it needs to compute the distance of each instance to all the training samples.

- Sensitive to Noisy Data: Performance can be severely impacted by the presence of noise or irrelevant features.

4.2.4 Random Forest

Advantages:

- Handling Overfitting: Random Forest reduces the risk of overfitting by averaging multiple trees, making it robust against noise and variance in data.
- Performance with Non-linear Data: Performs well on datasets with complex relationships thanks to the ensemble of decision trees that can model non-linear interactions.
- Feature Importance: Provides insightful outputs regarding the importance of each feature in the decision-making process, crucial for medical diagnostic applications.

Disadvantages:

- Model Complexity and Interpretability: More complex than a single decision tree, making it harder to interpret.
- Computation Demands: Although faster than some algorithms like SVM, Random Forest can be slower compared to others like Logistic Regression when it comes to training on very large datasets.

4.2.5 Decision to Choose Random Forest

Random Forest was chosen due to its ability to handle the intricacies and complexities of medical data effectively. Its robustness against overfitting and its superior performance on non-linear and multifaceted data make it particularly suited for predicting patient outcomes based on a variety of medical tests and features. While Logistic Regression provided a good baseline, and SVM and KNN offered valuable insights, Random Forest's balance of accuracy, interpretability of feature importance, and robustness across various metrics (accuracy, precision, recall, F1-score, and ROC-AUC) made it the most suitable choice for the task of classifying patients into inpatient and outpatient categories based on blood test results.

4.3 Model Evaluation

To assess the effectiveness of each model, the following metrics were used:

- Precision and Recall: These metrics are crucial in medical applications where both the accuracy of the positive predictions and the ability to capture all potential positive cases are important.
- F1-Score: The harmonic mean of precision and recall, providing a single metric to assess model performance considering both precision and recall.
- ROC-AUC Score: This score represents a model’s ability to discriminate between the classes at various threshold settings, which is particularly useful for imbalanced datasets.

4.4 Hyperparameter Tuning

Grid Search CV was likely employed to optimize model parameters to improve model performance. This involves searching through a specified parameter space for each algorithm, determining which configurations yield the best performance based on cross-validation results. By applying these methods, the study aims to establish a robust predictive model capable of accurately classifying patient admission necessity, thereby supporting clinical decision-making processes.

4.5 Model Training and Evaluation

Models were trained using the preprocessed data, and their performance was evaluated based on precision, recall, F1-score, and ROC-AUC scores. Hyperparameter tuning was carried out using techniques like Grid Search to optimize model parameters.

5. Results

5.1 Metric and Algorithm Comparison

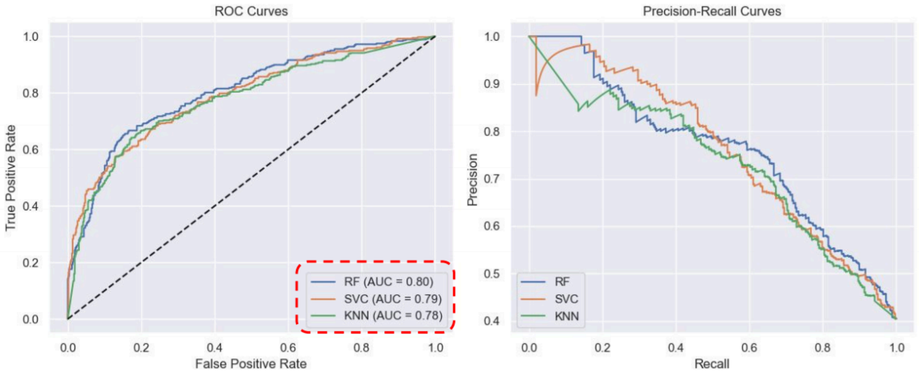


Figure 14. ROC and precision-recall curves

In the comparative analysis of the machine learning models employed to predict patient admission necessity, the ROC and Precision-Recall curves reveal closely competing performances among the models tested. Random Forest (RF), with an AUC of 0.80, demonstrates a slight edge over Support Vector Machine (SVC) and K-Nearest Neighbors (KNN), which scored AUCs of 0.79 and 0.78, respectively. This indicates a marginally better discriminative power of the Random Forest model in distinguishing between inpatient and outpatient cases across varying thresholds. The Precision-Recall analysis further supports these findings, showing a similar pattern of precision and recall trade-offs among the models, with none distinctly outperforming the others across the full range of recall values. This consistency underscores the competitive nature of the models, guiding the decision to favor Random Forest due to its slight advantage in overall classification efficacy as reflected in its ROC performance.

1. ROC and Precision-Recall Curves: The models are compared based on their Receiver Operating Characteristic (ROC) curves and Precision-Recall plots as seen in Figure 14. These metrics are crucial for assessing the trade-offs between true positive rates (sensitivity) and false positive rates, as well as the trade-offs between precision and recall, respectively. The presentation notes that all three algorithms (SVM, KNN, RF) have similar ROC scores, indicating comparable performance in distinguishing between the classes.

2. F2 Score: The F2 score, which weights recall higher than precision, is particularly important in medical settings where missing a positive case (failing to detect a true inpatient) can have serious consequences. The Random Forest model is highlighted for its balance of precision, recall, and F2 score, suggesting it is effective at maintaining a good balance between sensitivity and precision. The formula for F2 score is described below:

$$F - \text{Score} = \left(1+\beta^2\right) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

3. Recall and False Negative Analysis: With the importance of minimizing false negatives in a medical context, Random Forest has the least number of false negatives among the models tested, making it particularly suitable for medical applications where failing to identify an inpatient could result in inadequate care.

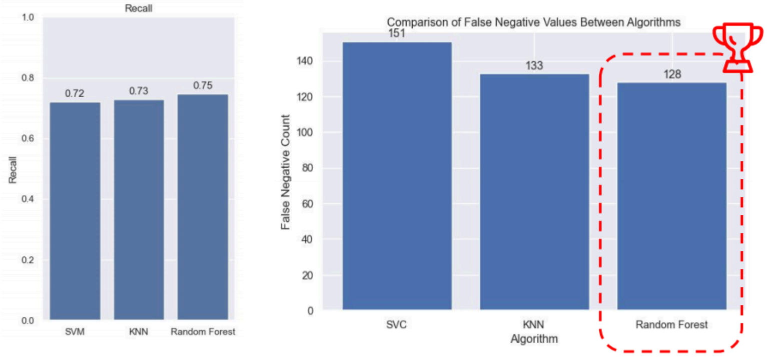


Figure 15. Recall and false negative analysis

In our comprehensive evaluation (see Figure 16.) of machine learning models designed to predict patient admission needs, the Random Forest model exhibited superior performance across multiple critical metrics. With an accuracy of 77%, precision of 76%, and recall as well as an F2 score of 75%, Random Forest not only most accurately classified patients but also maintained the highest sensitivity in identifying those requiring inpatient care. This model's ability to balance accuracy with a strong emphasis on recall makes it particularly suited for deployment in healthcare settings, where accurately detecting patients in need of immediate care is paramount. These results suggest that while all tested models performed robustly, Random Forest stands out for its ability to deliver reliable predictions crucial for effective patient management and resource allocation in hospital settings

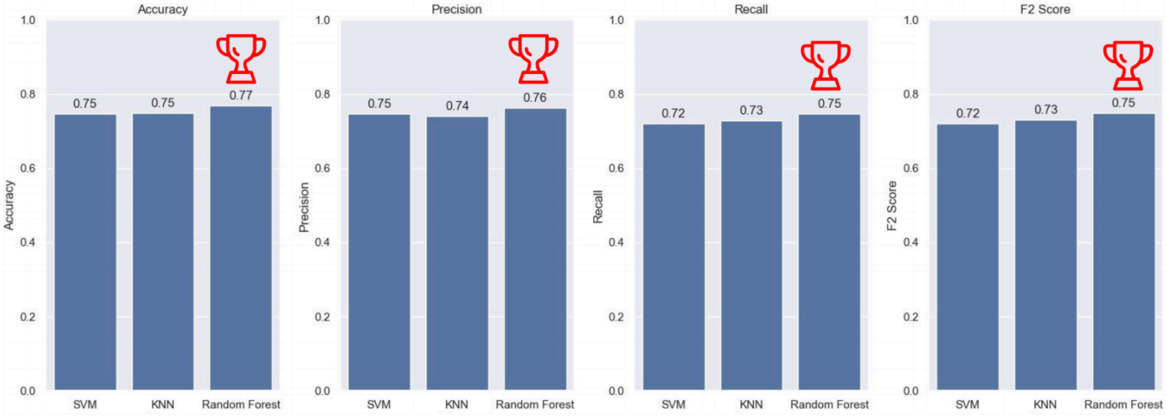


Figure 16. Algorithm comparison

5.2 Random Forest Performance Details

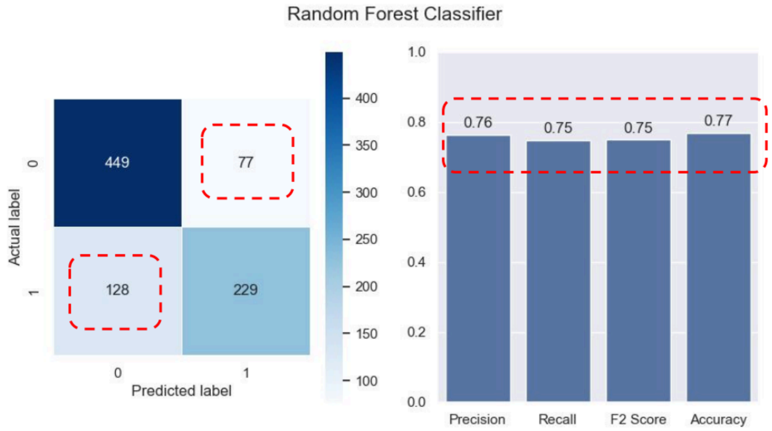


Figure 17. Random Forest Confusion Matrix and score summary

In our evaluation of the Random Forest model's capability to predict the necessity for patient admission, we utilized a confusion matrix and observed key performance metrics. The model achieved an accuracy of 77%, with a precision of 76% and a recall of 75%. The F2 score, which emphasizes recall over precision, stood at 0.75, highlighting the model's ability to prioritize the detection of patients who genuinely require inpatient care.

Despite these promising results, the accuracy of 77% still falls short of the higher standards typically demanded in medical applications, where more precise diagnostic capabilities are crucial. Medical standards often require accuracies well above 85% to ensure optimal patient safety and care. Although the Random Forest model outperforms other tested models in our study, indicating a robust base model, there remains a significant need for improvement to meet the stringent accuracy requirements of medical diagnostics.

The confusion matrix also points out specific areas for improvement, particularly in minimizing false negatives (128) and false positives (77). These figures are concerning because false negatives could lead to missed inpatient care, and false positives could result in unnecessary hospital admissions, both of which have serious implications for patient safety and hospital resource management.

5.3 Feature Importance in Random Forest

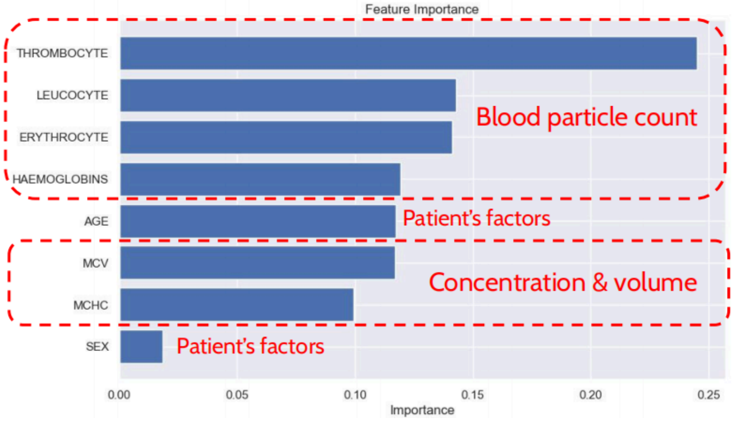


Figure 18. Random forest feature importance

1. Blood Particle Counts: Features related to blood particle counts (e.g., Thrombocyte, Leucocyte, Erythrocyte, and Haemoglobins) are identified as top contributors to the predictive power of the model.
2. Other Factors: The model also considers patient factors like age, which is shown to be more important than some blood measurement specifics like MCHC and MCV. However, sex has the least predictive power.

5.4 Medical Insights from Blood Parameters

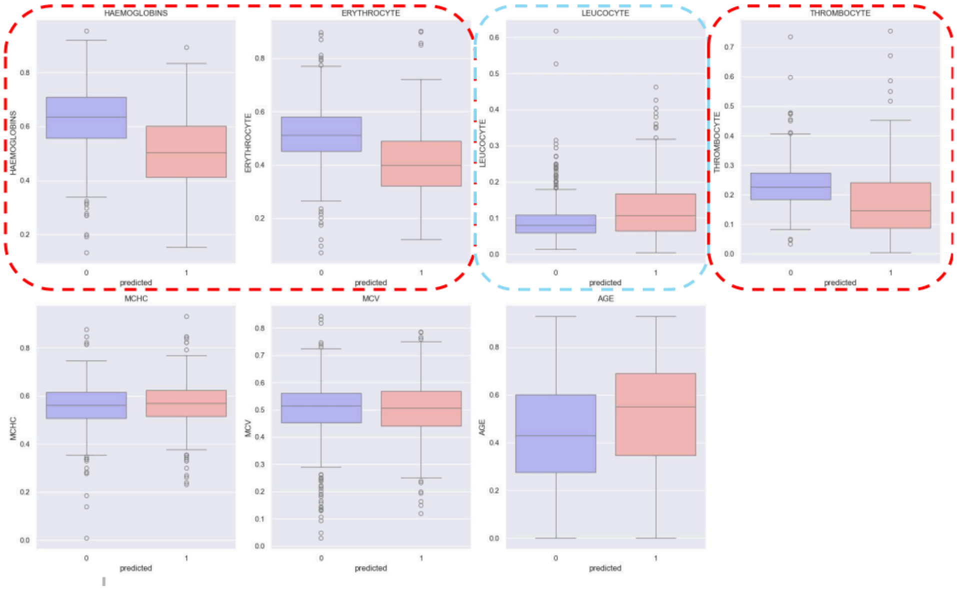


Figure 19. Medical insight. Patients with lower HAEMOGLOBINS, ERYTHROCYTE (RBC), and THROMBOCYTE (Platelets) are more likely to be admitted as inpatient as shown in the red boxes. However, patients with higher LEUCOCYTES (WBC) are more likely to be admitted as inpatient as shown in the blue boxes.

The figure above describes the insight gained from the prediction. Below is a detailed explanation of the impact of each covariate on the prediction result:

1. Haemoglobins, Erythrocytes, and Thrombocytes:
    - Observation: Patients predicted as inpatients generally show lower levels of Haemoglobins, Erythrocytes (RBCs), and Thrombocytes (platelets).
    - Medical Implication: Such conditions are indicative of anemia or severe bleeding. Anemia can be caused by nutritional deficiencies, chronic diseases, or bone marrow problems, while low platelet levels might suggest bleeding disorders or other hematological issues.
    - Action Required: Immediate medical intervention might be necessary to address severe anemia or active bleeding, which can be life-threatening if untreated. Treatment may involve transfusions, medication, or other urgent interventions to stabilize the patient's condition.
2. Leucocytes (WBC):
    - Observation: Higher Leucocyte counts are noted in patients predicted to be inpatients.
    - Medical Implication: Elevated white blood cell counts are typically associated with infections or inflammatory responses. This indicates that the immune system is actively fighting an infection.
    - Action Required: Such patients may require immediate isolation and treatment with antibiotics or other modalities to manage the infection and prevent its spread within hospital settings, especially considering the high-risk environments.
3. Age:
    - Observation: Older age appears to correlate with a higher likelihood of being admitted as an inpatient.
    - Medical Implication: Older patients often have more complex health profiles with multiple comorbidities, making them more susceptible to severe complications and necessitating inpatient care.
    - Action Required: This demographic insight underscores the need for targeted care plans and possibly more intensive monitoring for older patients, aligning resources such as ICU beds or specialized geriatric care as needed.

6. Conclusion

6.1 Model Performance and Limitations

While the Random Forest model demonstrated the best performance among the models tested, with an accuracy of around 77% and reasonably good precision and recall, these metrics still fall short of the ideal in a high-stakes field like healthcare. Particularly concerning is the high rate of false positives, which in a medical context could lead to unnecessary treatments, increased patient anxiety, and higher healthcare costs. This indicates that relying solely on blood test data might not be sufficient for making critical patient management decisions.

6.2 Key Predictive Features and Their Implications

The analysis identified key features such as hematocrit and leukocytes as significant predictors for hospital admissions. These findings underscore the importance of these blood components in clinical assessments and highlight how data-driven insights can augment traditional medical decision-making processes. However, the variability in model accuracy suggests that additional variables, possibly from patient medical histories or other diagnostic tests, might be needed to enhance predictive accuracy.

6.3 Practical Implications and Integration Challenges

Integrating predictive models into hospital information systems could indeed streamline operations and aid medical staff in making more informed decisions. However, the integration of such systems must be handled with care to ensure they support, rather than complicate, the decision-making process for healthcare providers. Systems need to be intuitive and must seamlessly fit into the clinical workflow to be truly effective. Moreover, Patient care requires a holistic approach that considers not just laboratory results but also symptoms, medical history, vital signs, and physical examinations. Blood tests provide quantitative data which must be interpreted within the broader context of the patient's overall health condition.

6.4 Future Work and Model Improvement

Given the limitations observed in the current models, future research should focus on:

- Incorporating More Data: Adding more comprehensive datasets, including a wider range of clinical parameters and patient demographics, could help in developing more robust models.
- Advanced Modeling Techniques: Exploring more complex machine learning models or deep learning approaches that can capture non-linear relationships and interactions between a larger set of variables may yield better predictive performance.
- Real-World Validation: It is crucial to test these models in real-world clinical settings to understand their practical utility and refine them based on feedback from healthcare professionals.

6.5 Broader Impact and Ethical Considerations

The potential of machine learning to transform healthcare is immense, but it also comes with responsibilities. It is vital to address ethical concerns such as patient privacy, data security, and the risk of bias in algorithmic decisions. Ensuring transparency in how models make predictions and allowing for human oversight can help mitigate some of these issues.

Appendix

Labor Division

Zixiao Wang primarily focuses on data collection and analysis, EDA, initial modeling, and report writing. On the other hand, Suphakrit Lertkitcharoenvong takes on the responsibilities related to modeling, classification, results, and PowerPoint presentation design. Additionally, Suphakrit's role in PowerPoint presentation design ensures that the findings of the study are effectively communicated to the audience.