***Statistical Study & Data Analysis on***

Classification of Different Types of Anemia

*Project work Submitted to*

The Department of Statistics, Pondicherry University

*By*

**Ms. Annanya Singh**

Regd. No.: 23375010

*&*

**Mr. Hrithik Dineshan**

Regd. No.: 20384308

As an Assignment for the Partial Fulfilment of the Course work

*Under the Guidance of*

**Dr. Sudesh Pundir**

Professor & Course Teacher

**STAT- 428 Biostatistics**

**Table of Contents**

1. **Introduction**
2. **Literature Review**
   1. **Etiological Classification**
   2. **Morphological Classification**
3. **Objective**
   1. **Specific Goals**
4. **Research Methodology**
   1. **Secondary Data Analysis**
   2. **Data Collection**
   3. **Data Processing**
   4. **Data Analysis**
   5. **Data Interpretation**
   6. **Documentation and Reporting**
5. **Statistical Tools Used**
   1. **Mean, Median, Mode**
   2. **Bar Plot**
   3. **Q-Q Plot**
   4. **Correlation Matrix**
   5. **Density Plots**
   6. **Multinomial Regression**
   7. **Likelihood Ratio Test**
   8. **Model Building**
   9. **Decision Trees**
   10. **Random Forest**
   11. **Confusion Matrix**
   12. **Feature Importance Analysis**
6. **Descriptive Data Analysis**
   1. **Exploratory Data Analysis**
7. **Multinomial Regression Analysis**
   1. **Multinomial Regression**
   2. **Likelihood Ratio Test (LRT)**
8. **Model Building**
   1. **Decision Tree**
   2. **Random Forest**
   3. **Confusion Matrix**
   4. **Training and Testing Model**
   5. **Feature Importance Analysis**
9. **Conclusion**
   1. **Key Findings**
   2. **Implications and Future Directions**

**Annexure I**

**Annexure II**

**References**

1. **Introduction**

Anemia is a common blood disorder characterized by a deficiency in the number or quality of red blood cells (RBCs) or hemoglobin, leading to reduced oxygen transport capacity in the blood. It affects millions of people worldwide and can result from various underlying causes, including nutritional deficiencies, chronic diseases, genetic disorders, and bone marrow problems. Accurate classification of anemia types is crucial for effective diagnosis, treatment, and management of the condition.

Complete Blood Count (CBC) tests are one of the most frequently used diagnostic tools in medicine, providing a comprehensive overview of the hematological parameters of an individual. CBC reports typically include measurements such as hemoglobin concentration, hematocrit, RBC count, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), and several other indices that can provide insights into a person's overall health and help identify different types of anemia.

The primary objective of this project is to leverage machine learning techniques to classify different types of anemia based on CBC report samples. Additionally, the project aims to identify key hematological factors that indicate various anemia types, enhancing our understanding of the condition's underlying mechanisms and improving diagnostic accuracy. By integrating advanced data analysis and machine learning models, we aim to develop a robust, accurate, and interpretable tool for anemia classification, which can be utilized by healthcare professionals for better patient outcomes.

This project involves several key steps, including data exploration and visualization, data preprocessing, feature engineering, model development and model interpretation. We will employ various classification algorithms, evaluate their performance using appropriate metrics, and select the best-performing model for deployment. Furthermore, we will interpret the results to determine which CBC parameters are most indicative of different anemia types, providing valuable insights for clinical practice.

Ultimately, this project aims to contribute to the field of hematology by developing an effective tool for anemia classification and by providing deeper insights into the factors that drive different types of anemia. Through careful analysis and rigorous model development, we hope to enhance the diagnostic process and support healthcare professionals in delivering more precise and personalized care to patients with anemia.

1. **Literature Review**

Anemia is a prevalent condition characterized by a decrease in the number of red blood cells (RBCs) or the hemoglobin concentration within the blood, leading to reduced oxygen transport capacity. This condition can result from various underlying causes, necessitating a comprehensive classification system to aid in diagnosis and management. The classification of anemia can be approached from multiple perspectives, including etiology, morphology, and pathophysiology. This review synthesizes key literature on the classification of anemia, focusing on these primary frameworks.

**2.1 Etiological Classification**

The etiological classification of anemia categorizes the condition based on its underlying causes. These can be broadly divided into three main groups: decreased RBC production, increased RBC destruction, and blood loss.

**2.1.1 Decreased RBC Production:** This category includes anemia resulting from deficiencies in essential nutrients, such as iron, vitamin B12, and folate. Iron deficiency anemia (IDA) is the most common form globally, often due to dietary insufficiency, malabsorption, or chronic blood loss (e.g., from gastrointestinal bleeding) (Camaschella, 2015). Anemias due to vitamin B12 or folate deficiency often result in megaloblastic anemia, characterized by the presence of abnormally large RBCs (megaloblasts) (Green & Allen, 2015).

**2.1.2 Increased RBC Destruction:** Hemolytic anemias occur when the lifespan of RBCs is significantly shortened due to intrinsic defects (e.g., hereditary spherocytosis, sickle cell anemia) or extrinsic factors (e.g., autoimmune disorders, infections) (Kato et al., 2018).

**2.1.3 Blood Loss:** Acute or chronic blood loss can lead to anemia. Acute blood loss typically results from trauma or surgery, while chronic blood loss may arise from gastrointestinal lesions, menstruation, or parasitic infections (Munoz et al., 2009).

**2.2 Morphological Classification**

The morphological classification is based on the size (mean corpuscular volume, MCV) and hemoglobin content (mean corpuscular hemoglobin concentration, MCHC) of RBCs. This approach divides anemia into three main categories:

**2.2.1 Microcytic Hypochromic Anemia:** Characterized by small (low MCV) and pale (low MCHC) RBCs. The most common cause is iron deficiency anemia, but other causes include thalassemia and anemia of chronic disease (Auerbach et al., 2020).

**2.2.2 Normocytic Normochromic Anemia:** RBCs are of normal size and hemoglobin content, but their number is reduced. This type includes anemia of chronic disease, acute blood loss, and hemolytic anemia (Weiss & Goodnough, 2005).

**2.2.3 Macrocytic Anemia:** Characterized by large RBCs (high MCV). This category includes megaloblastic anemias due to vitamin B12 or folate deficiency and non-megaloblastic anemias caused by liver disease, alcoholism, and hypothyroidism (Aslinia et al., 2006).

**2.3 Pathophysiological Classification**

The pathophysiological approach considers the underlying mechanisms affecting RBC production and survival:

**2.3.1 Hypoproliferative Anemias:** Result from inadequate RBC production due to bone marrow disorders (e.g., aplastic anemia, myelodysplastic syndromes) or suppression by chronic diseases (e.g., chronic kidney disease) (Adamson & Longo, 2012).

**2.3.2 Maturation Disorders:** Involve defects in DNA synthesis (e.g., megaloblastic anemia) or hemoglobin synthesis (e.g., thalassemia, sideroblastic anemia) (Hoffbrand & Moss, 2011).

**2.3.3 Hemolytic Anemias:** Caused by increased destruction of RBCs due to intrinsic factors (e.g., membrane defects, enzymopathies) or extrinsic factors (e.g., autoimmune hemolytic anemia, microangiopathic processes) (Gomes & Berger, 2015).

The classification of anemia is essential for accurate diagnosis, management, and treatment. By considering etiological, morphological, and pathophysiological perspectives, healthcare providers can better understand the diverse causes and manifestations of anemia. Future research should continue to refine these classifications and explore novel diagnostic markers and therapeutic approaches

1. **Objective**

The primary objective of this project is to develop a robust machine learning model to classify different types of anemia based on Complete Blood Count (CBC) report samples. Additionally, the project aims to identify key hematological factors that are indicative of various anemia types. This will be achieved through comprehensive data analysis, model development, and interpretation of feature importance, ultimately providing insights that can aid in the accurate diagnosis and understanding of anemia.

* 1. **Specific Goals:**

**3.1.1 Data Exploration and Understanding**:

- Conduct an exploratory data analysis (EDA) to understand the distribution and characteristics of CBC report samples.

- Visualize the data to identify patterns and correlations between different blood parameters and anemia types.

**3.1.2 Data Preprocessing**

- Handle missing values, outliers, and noise in the dataset.

- Normalize and standardize numerical features to ensure consistency and improve model performance.

- Encode categorical variables appropriately for inclusion in machine learning models.

**3.1.3 Model Development:**

- Train multiple classification models (e.g., Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, Support Vector Machines, and Neural Networks) to classify different types of anemia.

- Evaluate the models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

**3.1.4 Model Interpretation and Feature Importance:**

- Interpret the results of the best-performing model to understand which CBC parameters are most indicative of different types of anemia.

- Use feature importance scores, SHAP values, or LIME to provide insights into model decisions.

By achieving these goals, the project will provide valuable tools and insights for the classification and understanding of anemia types based on CBC report samples, contributing to better diagnostic practices and patient outcomes.

1. **Research Methodology**

Data methodology refers to the systematic approach and procedures used to collect, process, analyze, and interpret data in research or business contexts. It is crucial for ensuring the accuracy, reliability, and validity of the information gathered. Here are key aspects of data methodology:

**4.1 Secondary Data Analysis**: Secondary data analysis involves the use of existing data that was collected by someone else for a different purpose. Researchers or analysts utilize this data for their own investigations, avoiding the need to collect new data

**4.2 Data Collection**: This involves the process of gathering raw data from various sources such as surveys, sensors, databases, or observations. Methods can include online surveys, interviews, experiments, or scraping data from websites.

**4.3 Data Processing**: Once collected, the data often needs cleaning and preprocessing to remove errors, inconsistencies, or missing values. This step involves transforming the data into a format suitable for analysis, which may include normalization, scaling, or encoding categorical variables.

**4.4 Data Analysis**: This step involves using statistical or computational methods to derive insights, identify patterns, or test hypotheses within the data. Techniques range from basic descriptive statistics to advanced machine learning algorithms.

**4.5 Data Interpretation**: After analysis, the results need to be interpreted in the context of the research question or business problem. This involves drawing conclusions, making recommendations, or creating visualizations to communicate findings effectively.

**4.6 Documentation and Reporting**: Proper documentation of the entire data methodology process is crucial for transparency and reproducibility. This includes documenting data sources, processing steps, analysis methods, and assumptions made. Clear reporting ensures that others can understand, replicate, or build upon the work.

1. **Statistical Tools Used:**

**5.1** **Mean, Median, Mode**

- Mean: The average of a set of numbers, calculated by summing all the values and dividing by the count of values. It provides a central value but can be affected by outliers.

- Median: The middle value of a dataset when ordered. It is robust to outliers and represents the 50th percentile.

- Mode: The most frequently occurring value in a dataset. It is useful for identifying the most common value.

**5.2 Bar Plot**

A bar plot is a graphical representation that uses rectangular bars to display and compare the frequency, count, or other measures (like mean) of different categories or groups. The length of each bar is proportional to the value it represents, making it easy to visualize and compare categorical data.

**5.3 Q-Q Plot**

A Q-Q (quantile-quantile) plot is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, typically the normal distribution. The plot displays the quantiles of the dataset on the y-axis against the quantiles of the theoretical distribution on the x-axis. If the data follows the theoretical distribution, the points on the Q-Q plot will lie approximately along a straight line. Deviations from this line indicate departures from the theoretical distribution, such as skewness, kurtosis, or other distributional anomalies. Q-Q plots are widely used in statistical analysis to assess the normality of data, identify outliers, and understand the underlying distributional characteristics, thereby aiding in the selection of appropriate statistical tests and models.

**5.4 Correlation Matrix**

A correlation matrix displays the correlation coefficients between pairs of variables in a dataset. The coefficients range from -1 to 1, indicating the strength and direction of linear relationships:

- 1: Perfect positive correlation.

- 0: No correlation.

- -1: Perfect negative correlation.

**5.5 Density Curves**

Density curves are smoothed representations of a dataset's distribution, similar to histograms but continuous. They provide insights into the distribution shape, central tendency, and spread of data.

**5.6 Multinomial Regression**

Multinomial regression is a type of regression analysis used for predicting outcomes of a categorical dependent variable with more than two levels. It extends logistic regression to handle multiple classes by modelling the probability of each class as a linear combination of predictor variables. This method is particularly useful for classification problems where the target variable can take on three or more distinct categories. By estimating the relative risk of each outcome, it provides a comprehensive understanding of how predictor variables influence different categories.

**5.7 Likelihood Ratio Test**

The likelihood ratio test is a statistical method used to compare the goodness of fit between two models: a full model and a reduced model. It evaluates whether the more complex model (full model) significantly improves the fit of the data compared to the simpler model (reduced model). This is done by comparing the log-likelihoods of the two models and calculating the test statistic, which follows a chi-squared distribution. A significant result indicates that the full model provides a better fit, justifying the inclusion of additional predictors.

**5.8 Model Building**

Model building involves selecting and training algorithms on data to predict outcomes or understand relationships. Key steps include:

- Data Preprocessing: Cleaning and preparing data for analysis.

- Feature Selection: Choosing relevant variables.

- Model Training: Using algorithms to learn from data.

- Model Evaluation: Assessing performance using metrics like accuracy, precision, recall, and F1-score.

**5.9 Decision Trees**

Decision trees are a type of supervised learning algorithm used for classification and regression tasks. They split data into subsets based on feature values, creating a tree-like model of decisions:

- Root Node: The top node representing the entire dataset.

- Internal Nodes: Decision points based on feature values.

- Leaf Nodes: Final output labels or values.

**5.10 Random Forest**

Random forests are an ensemble learning method that combines multiple decision trees to improve prediction accuracy and control overfitting. Each tree is trained on a random subset of the data, and the final prediction is made by averaging (regression) or majority voting (classification):

- Ensemble Method: Combines predictions from several models to enhance performance.

- Robustness: Less prone to overfitting compared to individual decision trees.

- Feature Importance: Provides insights into the importance of variables in making predictions.

**5.11 Confusion Matrix**

A confusion matrix is a powerful tool used to evaluate the performance of a classification model. It is a table that allows visualization of the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model. Here's a breakdown:

- True Positives (TP): Correctly predicted positive instances.

- True Negatives (TN): Correctly predicted negative instances.

- False Positives (FP): Incorrectly predicted positive instances (also known as Type I errors).

- False Negatives (FN): Incorrectly predicted negative instances (also known as Type II errors).

The confusion matrix provides detailed insight into how well a classification model performs, highlighting specific areas where the model may be making mistakes, and is essential for improving model accuracy and reliability.

**5.12 Feature Importance Analysis**

Feature importance analysis identifies and quantifies the contribution of each feature to a machine learning model's predictive power. Methods include coefficients in linear models, impurity reduction in tree-based methods, permutation importance, and advanced techniques like SHAP and LIME. Understanding feature importance improves model transparency, performance, and provides valuable domain insights. This analysis helps in selecting relevant features and interpreting the model's decision-making process effectively.

1. **Descriptive Data Analysis**

**6.1 Exploratory Data Analysis (EDA)** is a crucial initial step in understanding a dataset's characteristics. It involves summarizing main features, often with visual methods like histograms or scatter plots, to uncover patterns, spot anomalies, and test assumptions. EDA helps identify relationships between variables and provides insights for further analysis or modelling. By exploring data distributions, central tendencies, and outliers, EDA aids in formulating hypotheses and refining data preprocessing strategies. This iterative process sets the stage for more sophisticated analyses, ensuring a comprehensive understanding of the dataset's nuances and informing the direction of subsequent statistical or machine learning techniques.

**Table 1.**

**Summary Statistics**

| **Statistic** | **WBC** | **LYMp** | **NEUTp** | **LYMn** | **NEUTn** | **RBC** | **HGB** | **HCT** | **MCV** | **MCH** | **MCHC** | **PLT** | **PDW** | **PCT** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Min.** | 0.800 | 6.20 | 0.70 | 0.200 | 0.500 | 1.360 | -10.00 | 2.00 | -79.30 | 10.90 | 11.50 | 10 | 8.40 | 0.0100 |
| **1st Qu.** | 6.000 | 25.84 | 71.10 | 1.881 | 5.100 | 4.190 | 10.80 | 39.20 | 81.20 | 25.50 | 30.60 | 157 | 13.30 | 0.1700 |
| **Median** | 7.400 | 25.84 | 77.51 | 1.881 | 5.141 | 4.600 | 12.30 | 46.15 | 86.60 | 27.80 | 32.00 | 213 | 14.31 | 0.2603 |
| **Mean** | 7.863 | 25.84 | 77.51 | 1.881 | 5.141 | 4.708 | 12.18 | 46.15 | 85.79 | 32.08 | 31.74 | 230 | 14.31 | 0.2603 |
| **3rd Qu.** | 8.680 | 25.84 | 77.51 | 1.881 | 5.141 | 5.100 | 13.50 | 46.15 | 90.20 | 29.60 | 32.90 | 293 | 14.70 | 0.2603 |
| **Max.** | 45.700 | 91.40 | 5317.00 | 41.800 | 79.000 | 90.800 | 87.10 | 3715.00 | 990.00 | 3117.00 | 92.80 | 660 | 97.00 | 13.6000 |

The dataset contains hematological parameters for 1281 samples, along with their diagnoses. Here's a brief interpretation of key statistics:

- **WBC:** Median is 7.4, indicating a typical range of white blood cells.

- **LYMp and NEUTp:** Median percentages of lymphocytes (25.84%) and neutrophils (77.51%) show typical distributions.

- **LYMn and NEUTn:** Medians are 1.881 and 5.141, respectively, reflecting common lymphocyte and neutrophil counts.

- **RBC:** Median of 4.6 indicates normal red blood cell count, but extreme values suggest outliers.

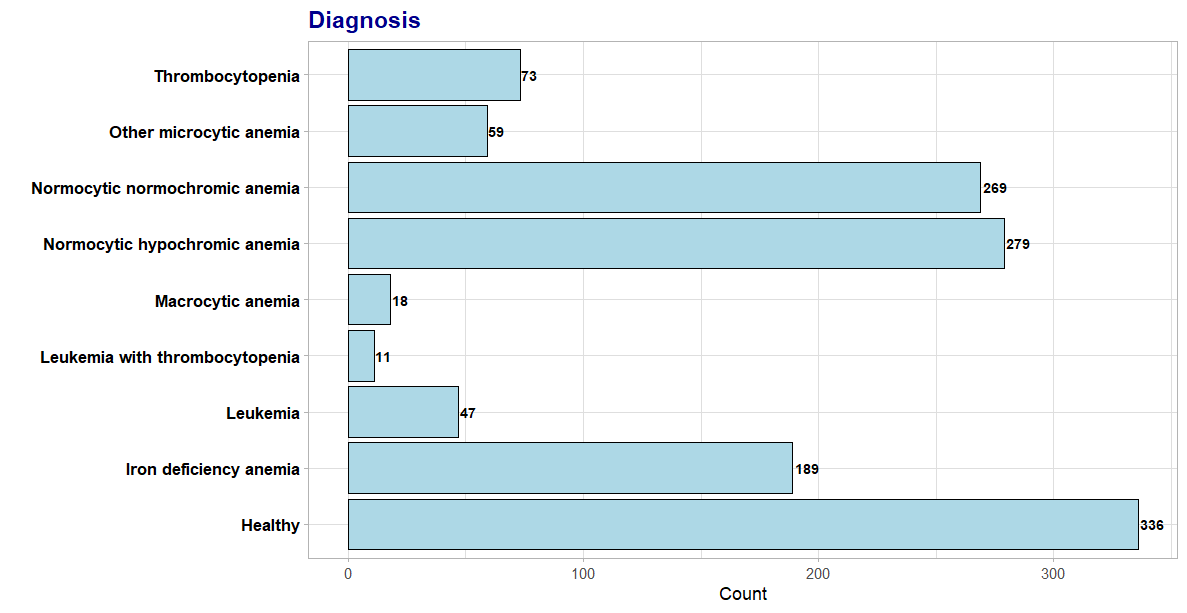
- **HGB:** Median is 12.3, typical for hemoglobin levels, but negative and extremely high values indicate data issues.

- **HCT, MCV, MCH, MCHC:** Medians are within normal ranges, but extreme values again indicate potential outliers.

- **PLT:** Median is 213, showing normal platelet count, with a wide range suggesting variability.

- **PDW and PCT:** Medians are 14.31 and 0.2603, respectively, within expected ranges for platelet distribution and plateletcrit.

**Bar Plot indicating the count of each diagnosis.**



**Figure 1.**

The bar plot illustrates the distribution of different diagnoses within the dataset. Here's a breakdown of the findings:

**a. Healthy:**

- Count: 336

- Observation: The largest group in the dataset, indicating that a significant number of samples are from healthy individuals.

**2. Normocytic Hypochromic Anemia:**

- Count: 279

- Observation: This is the second-largest group, showing a common type of anemia characterized by normal cell size but low hemoglobin content.

**3. Normocytic Normochromic Anemia:**

- Count: 269

- Observation: Another common type of anemia with normal-sized red blood cells and normal hemoglobin concentration.

**4. Iron Deficiency Anemia:**

- Count: 189

- Observation: A significant number of samples have this common type of anemia, typically caused by insufficient iron.

**5. Thrombocytopenia:**

- Count: 73

- Observation: A moderate number of cases with low platelet counts.

**6. Other Microcytic Anemia:**

- Count: 59

- Observation: A smaller group indicating other types of anemia with small red blood cells.

**7. Leukemia:**

- Count: 47

- Observation: A notable number of cases with this serious blood disorder.

**8. Macrocytic Anemia:**

- Count: 18

- Observation: Few cases of anemia characterized by larger than normal red blood cells.

**9. Leukemia with Thrombocytopenia:**

- Count: 11

- Observation: The smallest group, indicating a combined condition of leukemia and low platelet count.

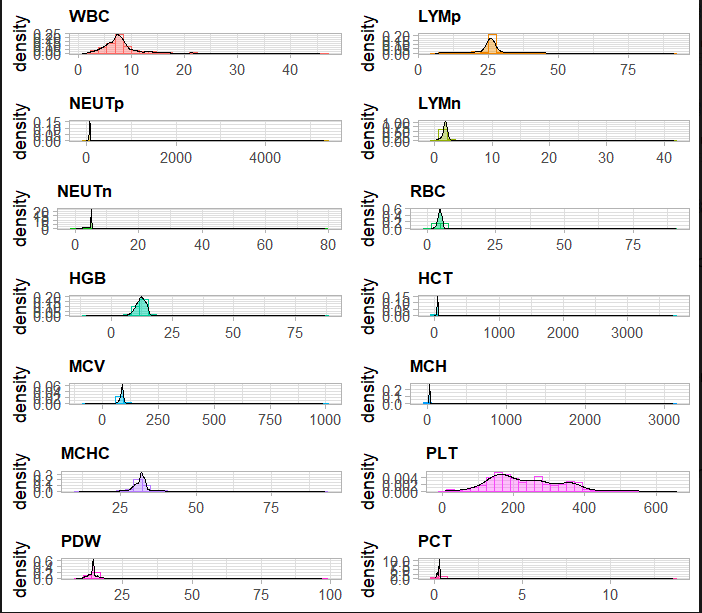
**Key Takeaways**

- The dataset is dominated by healthy individuals and common types of anemia.

- Normocytic hypochromic and normocytic normochromic anemia are prevalent, suggesting these conditions are frequently diagnosed.

- Iron deficiency anemia is also a major concern, highlighting the importance of monitoring iron levels.

- There are fewer cases of serious conditions like leukemia and combined disorders such as leukemia with thrombocytopenia.

**Density Plots illustrating the spread of each variable.**

**Figure 2.**

These density plots show the distribution of various hematological parameters in the dataset:

**1. WBC (White Blood Cell count)** - Most values cluster around 5-10, indicating a normal range, but there are a few extreme values up to 45.7.

**2. LYMp (Lymphocyte percentage)** - Most values are around 25%, with a long tail extending to higher percentages, indicating a skewed distribution.

**3. NEUTp (Neutrophil percentage)** - Most values are around 70-80%, but there are extreme outliers, significantly higher than typical values.

**4. LYMn (Lymphocyte number)** - Most values are around 0-5, showing a typical distribution with some high outliers.

**5. NEUTn (Neutrophil number)** - Values are mostly around 0-10, with some extreme outliers up to 79.

**6. RBC (Red Blood Cell count)** - The distribution peaks around 4-5, indicating a normal RBC count, but has extreme values up to 90.8.

**7. HGB (Hemoglobin)** - The majority of values are around 12-15, but there are some extremely high outliers.

**8. HCT (Hematocrit)** - The main cluster is around 30-50, with extreme outliers indicating potential data entry errors.

**9. MCV (Mean Corpuscular Volume)** - Most values are between 80-100, typical for MCV, but there are very high outliers.

**10. MCH (Mean Corpuscular Hemoglobin)** - Many values are around 20-30, with significant outliers.

**11. MCHC (Mean Corpuscular Hemoglobin Concentration)** - Most values are around 30-35, indicating a normal range, but some outliers are very high.

**12. PLT (Platelet count)** - Values cluster around 200-300, with a long tail and some very high outliers.

**13. PDW (Platelet Distribution Width)** - Most values are around 10-20, with a few extreme outliers.

**14. PCT (Plateletcrit)** - The distribution is heavily skewed with most values close to 0.1-0.3, but some extreme values.

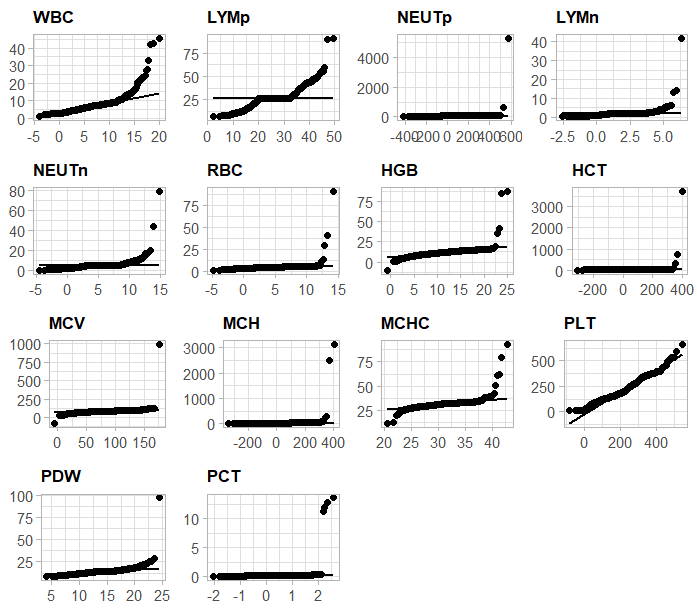
**Key Observations**

- Outliers and Extreme Values: Several parameters (e.g., NEUTp, HCT, MCV) have extreme outliers, which may indicate data entry errors or rare cases.

- Skewed Distributions: Some parameters like LYM% and PCT show skewed distributions, suggesting non-normality.

- Normal Ranges: Parameters like WBC, RBC, HGB, and PLT mostly fall within expected clinical ranges but still include outliers.

**Q-Q Plots to understand normality of the values of the variables.**



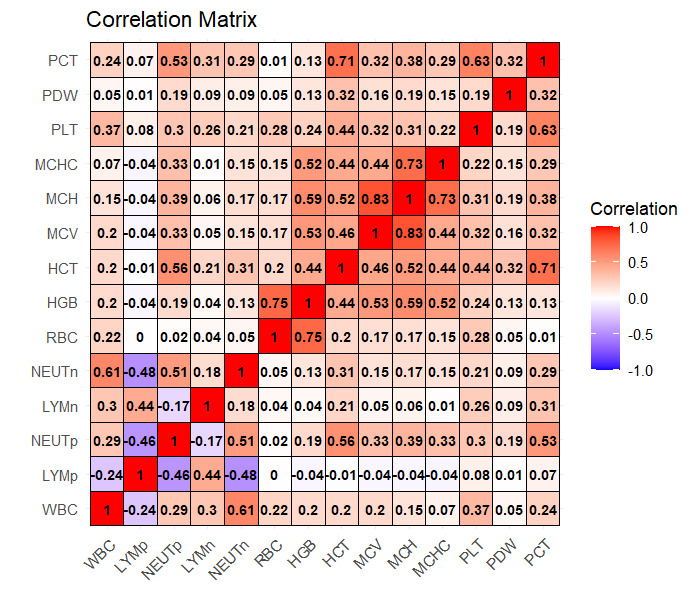
**Figure 3.**

Most of these plots show points deviating significantly from the straight line, especially at the tails. This suggests that these variables do not follow a normal distribution. Extreme outliers can be observed in several plots, indicating that some values are far from the expected normal range.

**Detailed Observations:**

* **WBC, LYM, and PLT** show a curved pattern with significant deviation at the tails, indicating heavy-tailed distributions.
* **NEUTp** shows extreme outliers and a distinct non-linear pattern, suggesting significant deviation from normality.
* **MCH and HCT** have points scattered widely, indicating a high degree of non-normality and the presence of outliers.
* **PCT and PDW** show some deviations but not as extreme as other variables.

**Correlation Matrix illustrating the correlation between the variables.**



**Figure 4.**

The correlation matrix visualizes the relationships between various Complete Blood Count (CBC) parameters. Here's a detailed interpretation of the key correlations:

**1. WBC (White Blood Cell count)** - Positive correlation with NEUTn (0.61) and NEUTp (0.29), indicating higher WBC counts are associated with higher neutrophil counts and percentages. Negative correlation with LYMp (-0.24), suggesting that higher WBC counts tend to be associated with lower lymphocyte percentages.

**2. LYMp (Lymphocyte percentage)** - Strong negative correlation with NEUTp (-0.46), meaning higher lymphocyte percentages are associated with lower neutrophil percentages. Negative correlation with NEUTn (-0.48), reinforcing the inverse relationship between lymphocytes and neutrophils.

**3. NEUTp (Neutrophil percentage)** - Strong positive correlation with NEUTn (0.51), showing that higher neutrophil percentages correspond to higher neutrophil counts. Negative correlation with LYMp (-0.46), as discussed.

**4. LYMn (Lymphocyte number)** - Moderate positive correlation with WBC (0.30) and NEUTn (0.18), indicating a relationship between lymphocyte count and overall WBC/neutrophil count.

**5. NEUTn (Neutrophil number)** - Strong positive correlations with WBC (0.61) and NEUTp (0.51), highlighting the key role of neutrophils in total WBC counts.

**6. RBC (Red Blood Cell count)** - Weak positive correlations with WBC (0.22) and other red cell parameters like HGB (0.44) and HCT (0.20), indicating some interdependence among these blood components.

**7. HGB (Hemoglobin)** - Strong positive correlation with HCT (0.75), suggesting that higher hemoglobin levels are closely associated with higher hematocrit.Positive correlation with MCH (0.59), MCHC (0.52), and MCV (0.53), showing interrelatedness among red cell indices.

**8. HCT (Hematocrit)** - Strong correlations with HGB (0.75), MCV (0.44), and MCH (0.52), indicating that hematocrit is a key measure related to the size and hemoglobin content of red blood cells.

**9. MCV (Mean Corpuscular Volume) - S**trong correlations with MCH (0.83) and MCHC (0.44), reflecting the relationships between cell volume, hemoglobin content, and concentration.

**10. PLT (Platelet count)** - Moderate positive correlation with WBC (0.37), suggesting a relationship between platelet count and overall WBC count.

**11. PDW (Platelet Distribution Width)** - Weak correlations with most other parameters, indicating a relatively independent variability in platelet width distribution.

**12. PCT (Plateletcrit)** - Strong positive correlation with PLT (0.63) and moderate correlations with HCT (0.32), MCV (0.38), and MCH (0.29), highlighting its dependence on platelet count and red cell parameters.

**Key Takeaways**

**- Strong Positive Correlations:** Between parameters like HGB and HCT, MCV and MCH, indicating these measures are often jointly affected in various blood conditions.

**- Strong Negative Correlations:** Notably between LYMp and NEUTp/NEUTn, reflecting the inverse relationship between lymphocytes and neutrophils.

**- Moderate Correlations:** RBC has moderate positive correlations with HGB, HCT, and WBC, indicating interconnectedness but with independent variability.

1. **Multinomial Regression**

**7.1 Multinomial Regression**

Multinomial regression is a type of regression analysis used when the dependent variable is categorical with more than two levels (i.e., more than two categories or classes). It is an extension of logistic regression that generalizes to multiclass problems. This method models the probability of each possible outcome of a categorical dependent variable based on one or more independent variables.

* **Dependent Variable:** Must be categorical with three or more levels.
* **Independent Variables:** Can be continuous or categorical.
* **Modeling:** Estimates the probability of each class by comparing it to a reference class.
* **Applications:** Commonly used in scenarios where the outcome can fall into multiple categories, such as predicting the type of anemia based on various blood parameters.

The model estimates the log-odds of each category compared to a reference category, and the coefficients represent the change in the log-odds for a one-unit change in the predictor variables.

**Table 2.**

|  | **(Intercept)** | **WBC** | **RBC** | **HGB** | **MCV** | **MCHC** | **PLT** | **PCT** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Iron deficiency anemia** | 234.7090699 | 0.4701796 | -1.7558460 | -1.616174563 | -1.44377086 | -2.8371930 | 0.008778187 | -29.5791235 |
| **Leukemia** | 53.5703944 | 0.9601132 | -2.6171937 | 0.356841466 | -0.48714276 | -0.4322590 | 0.001374433 | 0.9126915 |
| **Leukemia with thrombocytopenia** | 0.6071842 | 0.8501513 | 0.7280117 | -0.418268842 | 0.11158898 | -0.2268335 | -0.04728360 | -21.2120773 |
| **Macrocytic anemia** | 58.8600208 | 0.1535578 | -6.5629050 | -0.653933563 | 0.18544050 | -1.2848364 | 0.005215775 | -15.7475178 |
| **Normocytic hypochromic anemia** | 121.9543089 | 0.4646056 | -2.8298678 | -1.759989967 | -0.15997588 | -2.3462129 | -0.01006954 | 0.7142998 |
| **Normocytic normochromic anemia** | 23.0198692 | 0.3269029 | -2.3026174 | -1.578542342 | -0.03894518 | 0.2757199 | 0.000095064 | 0.4436780 |
| **Other microcytic anemia** | 117.8475497 | 0.4619715 | -0.1766043 | -0.840498733 | -1.41141352 | 0.2339772 | -0.01459189 | 1.2125521 |
| **Thrombocytopenia** | -15.2678732 | -0.1763268 | 0.2231883 | 0.008920833 | 0.18282179 | 0.2193595 | -0.04465769 | -10.9523218 |
|  |  |  |  |  |  |  |  |  |

Residual Deviance: 964.223

AIC: 1092.223

**7.1.1 Model Overview**

The model is used to predict the diagnosis category based on the blood test parameters: WBC (White Blood Cell count), RBC (Red Blood Cell count), HGB (Hemoglobin), MCV (Mean Corpuscular Volume), MCHC (Mean Corpuscular Hemoglobin Concentration), PLT (Platelet count), and PCT (Plateletcrit).

**7.1.2 Coefficients**

The coefficients represent the log odds of being in a particular diagnosis category relative to the baseline category (not specified here, but often the first or most common category). Positive coefficients indicate higher likelihood, while negative coefficients indicate lower likelihood of the diagnosis as the predictor variable increases.

**7.1.3 Standard Errors**

Standard errors provide a measure of the variability of the coefficient estimates. Smaller standard errors indicate more precise estimates of the coefficients. For example:

The standard error for the intercept of “Iron deficiency anemia” is 0.0066, which is very small, indicating a precise estimate.

**7.1.4 Interpretation**

**Iron deficiency anemia:** An increase in WBC slightly increases the log odds of being diagnosed with iron deficiency anemia. Higher RBC and HGB levels decrease the log odds of this diagnosis significantly, as do MCV and MCHC.

**Leukemia:** Higher WBC levels increase the log odds of being diagnosed with leukemia. RBC has a strong negative effect, while HGB has a slight positive effect.

**Leukemia with thrombocytopenia:** WBC again has a positive effect, while RBC and HGB have mixed effects.

**Macrocytic anemia:** Higher WBC increases the log odds slightly, but RBC decreases it significantly.

**Normocytic hypochromic anemia**: Higher WBC and RBC have a similar impact as for other anemias, with significant negative effects from RBC and HGB.

**Normocytic normochromic anemia:** The effects are similar to those of normocytic hypochromic anemia but with slight variations in magnitude.

**Other microcytic anemia:** WBC has a positive effect, but RBC and MCV have significant negative impacts.

**Thrombocytopenia:** The effects vary, with WBC having a negative effect and RBC having a slight positive effect.

**7.1.5** **Residual Deviance and AIC**

**Residual Deviance:** 964.223 - Lower deviance indicates a better fit of the model to the data.

**AIC (Akaike Information Criterion):** 1092.223 - A lower AIC indicates a better model fit when comparing multiple models.

**7.1.6 Model Summary**

The logistic regression model is designed to classify types of anemia based on blood test parameters: WBC, RBC, HGB, MCV, MCHC, PLT, and PCT. The coefficients indicate the log odds of a diagnosis category relative to a baseline category, with positive values suggesting higher likelihood and negative values indicating lower likelihood as predictor variables increase. The model highlights how different blood parameters influence the probability of various anemias and related conditions such as leukemia and thrombocytopenia. For instance, higher RBC and HGB levels are strongly associated with lower odds of iron deficiency anemia, whereas higher WBC levels increase the likelihood of leukemia. The model's fit is assessed using Residual Deviance (964.223) and AIC (1092.223), with lower values indicating better model performance. The initial logistic regression model had an Akaike Information Criterion (AIC) close to 1500. After applying stepwise regression to refine the model, the AIC significantly decreased to 1092.223. This substantial reduction in AIC indicates an improved model fit, demonstrating the effectiveness of stepwise regression in enhancing the model's predictive accuracy.

**7.2 Likelihood Ratio Test (LRT)**

The Likelihood Ratio Test (LRT) is a statistical test used to compare the goodness-of-fit of two nested models. The test determines whether a more complex model provides a significantly better fit to the data than a simpler model.

* **Nested Models:** The simpler model (null model) is a special case of the more complex model (alternative model).
* **Log-Likelihood Values:** The test compares the log-likelihoods of the two models.
* **Chi-Square Distribution:** The test statistic follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the models.
* **Significance:** A significant p-value indicates that the more complex model provides a significantly better fit to the data

The Likelihood Ratio Test (LRT) is used to compare two nested models to determine if the more complex model significantly improves the fit to the data compared to the simpler model. We have two models:

Likelihood ratio test

Model 1: Diagnosis ~ WBC + RBC + HGB + MCV + MCHC + PLT + PCT

Model 2: Diagnosis ~ 1

#Df LogLik Df Chisq Pr(>Chisq)

* + 1. 64 -482.11

2 8 -1632.41 -56 2300.6 < 2.2e-16 \*\*\*

Signif. Codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**7.2.1 Log-Likelihood Values –** Model 1 has a log-likelihood of -482.11. Model 2 has a log-likelihood of -1632.41. Higher (less negative) log-likelihood values indicate a better fit to the data. Thus, Model 1 fits the data much better than Model 2.

**7.2.2 Degrees of Freedom** – Model 1 has 64 degrees of freedom, while Model 2 has only 8 degrees of freedom. The difference in degrees of freedom between the two models is 56.

**7.2.3 Chi-Square Statistic-** The LRT statistic (Chisq) is 2300.6, which is very large.

**7.2.4 p-value** – The p-value is less than 2.2e-16, which is exceedingly small.

The very small p-value (< 2.2e-16) indicates that the improvement in model fit from Model 2 to Model 1 is statistically significant. In other words, including the predictors (WBC, RBC, HGB, MCV, MCHC, PLT, PCT) in Model 1 significantly improves the model’s ability to explain the variation in the diagnosis of anemia compared to the intercept-only Model 2.Thus, we reject the null hypothesis that the simpler model (Model 2) is sufficient, and conclude that the more complex model (Model 1) provides a significantly better fit to the data.

**7.2.5 Testing summary**

The very small p-value suggests that Model 1, which includes the predictors, significantly improves the model fit over Model 2, the null model. Thus, we reject the null hypothesis that the simpler model is sufficient, concluding that the more complex model provides a significantly better fit.

1. **Model Building**

**8.1 Decision Tree**

A decision tree is a supervised learning algorithm used for both classification and regression tasks. It splits the data into subsets based on the value of input features, creating branches that lead to decision nodes and leaf nodes. Each node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome.

* **Interpretability:** Easy to interpret and visualize.
* **Splitting Criteria:** Uses criteria like Gini impurity, information gain, or variance reduction to split nodes.
* **Prone to Overfitting:** Can overfit the training data if not pruned or controlled by parameters like maximum depth.

**8.2 Random Forest**

A random forest is an ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction. It addresses the overfitting problem of individual decision trees.

* **Bootstrap Aggregation (Bagging):** Each tree is trained on a random subset of the data.
* **Feature Randomness:** Randomly selects a subset of features for splitting nodes, adding more diversity, and reducing correlation between trees.
* **Robustness:** More robust to overfitting compared to individual decision trees.

\* dt | Acc Train: 0.9680 | Acc Test: 0.9479

\* rf | Acc Train: 1.0000 | Acc Test: 0.9505

The output provides valuable insights into the performance of two machine learning models, Decision Tree (dt) and Random Forest (rf), in predicting anemia diagnoses.

**Accuracy on Training Set -** The high accuracy scores on the training set indicate that both models were able to learn from the training data and make accurate predictions. The Random Forest model achieved a perfect accuracy of 100% on the training set, suggesting that it effectively captured the complex relationships within the training data.

**Accuracy on Test Set -** The accuracy scores on the test set are also high, indicating that both models generalize well to new, unseen data. The Random Forest model slightly outperforms the Decision Tree model on the test set, achieving a slightly higher accuracy score.

**Model Effectiveness:** The high accuracy scores on both the training and test sets suggest that the machine learning models are effective in predicting anemia diagnoses based on the input features (WBC, RBC, HGB, MCV, MCHC, PLT, PCT).

**Model Comparison:** While both models perform well, the Random Forest model shows a slight improvement in predictive accuracy compared to the Decision Tree model. This suggests that the ensemble approach used in Random Forest, which combines multiple decision trees, is beneficial in this context.

**Generalization:** The fact that both models perform well on unseen test data indicates that they have learned meaningful patterns from the training data and can generalize to new patient data effectively.

**Potential Application:** These well-performing models could be deployed in clinical settings to assist healthcare professionals in diagnosing anemia based on patient blood test results. They could help identify different types of anemia (such as iron deficiency anemia, macrocytic anemia, etc.) accurately and efficiently, aiding in timely and appropriate patient management.

**8.3 Confusion Matrix**

A confusion matrix is a performance measurement tool for classification problems. It is a table that describes the performance of a classification model by comparing predicted labels with true labels.

* **True Positives (TP):** Correctly predicted positive observations.
* **True Negatives (TN):** Correctly predicted negative observations.
* **False Positives (FP):** Incorrectly predicted positive observations.
* **False Negatives (FN):** Incorrectly predicted negative observations.

**8.3.1 Metrics Derived:**

* **Accuracy:** (TP+TN)/(TP+TN+FP+FN)(TP+TN)/(TP+TN+FP+FN)
* **Precision:** TP/(TP+FP)TP/(TP+FP)
* **Recall (Sensitivity):** TP/(TP+FN)TP/(TP+FN)
* **F1 Score:** Harmonic mean of precision and recall.

**Confusion Matrix for Diagnoses**

**Table 3.**

| **Predicted/Actual** | **Healthy** | **Iron deficiency anemia** | **Leukemia** | **Leukemia with thrombocytopenia** | **Macrocytic anemia** | **Normocytic hypochromic anemia** | **Normocytic normochromic anemia** | **Other microcytic anemia** | **Thrombocytopenia** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Healthy** | 84 | 0 | 2 | 0 | 0 | 2 | 2 | 0 | 4 |
| **Iron deficiency anemia** | 2 | 51 | 1 | 0 | 0 | 3 | 0 | 1 | 0 |
| **Leukemia** | 0 | 0 | 7 | 0 | 0 | 2 | 1 | 0 | 0 |
| **Leukemia with thrombocytopenia** | 0 | 0 | 2 | 1 | 0 | 1 | 1 | 0 | 0 |
| **Macrocytic anemia** | 0 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 1 |
| **Normocytic hypochromic anemia** | 1 | 5 | 1 | 0 | 2 | 69 | 2 | 4 | 0 |
| **Normocytic normochromic anemia** | 6 | 0 | 0 | 0 | 1 | 3 | 79 | 0 | 0 |
| **Other microcytic anemia** | 0 | 3 | 0 | 0 | 0 | 1 | 0 | 13 | 3 |
| **Thrombocytopenia** | 2 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 15 |

**Healthy**

* + Predicted as healthy 84 times (True Positives).
  + Misclassified as leukemia 2 times.
  + Few misclassifications into other categories, indicating good model performance for this class.

**Iron Deficiency Anemia**

* + Correctly identified 51 times.
  + Misclassified as healthy (2), leukemia (1), normocytic hypochromic anemia (3), other microcytic anemia (1).

**Leukemia**

* + Correctly identified 7 times.
  + Misclassified into thrombocytopenia, macrocytic anemia, normocytic hypochromic anemia, and normocytic normochromic anemia occasionally, indicating some overlap.

**Leukemia with Thrombocytopenia**

* + Correctly identified only once.
  + Frequently misclassified into other leukemia types and normocytic normochromic anemia.

**Macrocytic Anemia**

* + Correctly identified 3 times.
  + Misclassified into leukemia, thrombocytopenia, normocytic hypochromic anemia.

**Normocytic Hypochromic Anemia**

* + Correctly identified 69 times.
  + Misclassified into iron deficiency anemia (5), healthy (1), and few others.

**Normocytic Normochromic Anemia**

* + Correctly identified 79 times.
  + Misclassified as healthy (2), macrocytic anemia (1), normocytic hypochromic anemia (3), and thrombocytopenia (1).

**Other Microcytic Anemia**

* + Correctly identified 13 times.
  + Misclassified into iron deficiency anemia (3), normocytic hypochromic anemia (4), thrombocytopenia (3).

**Thrombocytopenia**

* + Correctly identified 15 times.
  + Misclassified into healthy (4), normocytic hypochromic anemia (2), normocytic normochromic anemia (1), macrocytic anemia (1).

**8.3.1 Summary**

* Healthy, Iron deficiency anemia, Normocytic hypochromic anemia, and Normocytic normochromic anemia are predicted with high accuracy.
* Leukemia with thrombocytopenia and macrocytic anemia show higher misclassification rates, indicating difficulties in distinguishing these conditions with the current model.
* Some diagnoses are more frequently confused with specific others, suggesting potential overlap in symptoms or test results. For example, normocytic hypochromic anemia is often confused with iron deficiency anemia.

**8.4 Training and Testing Model**

**Bar Plots showing the count of the Diagnosis through training and testing set.**

A graph with red squares

Description automatically generated with medium confidenceA graph with blue bars

Description automatically generated with medium confidence

**Figure 5(a) Figure 5(b)**

**8.4.1 Training Data (Diagnosis Train):**

* **Thrombocytopenia**: 59 instances
* **Other microcytic anemia**: 48 instances
* **Normocytic normochromic anemia**: 216 instances
* **Normocytic hypochromic anemia**: 224 instances
* **Macrocytic anemia**: 15 instances
* **Leukemia with thrombocytopenia**: 9 instances
* **Leukemia**: 38 instances
* **Iron deficiency anemia**: 152 instances
* **Healthy**: 26 instances

**8.4.2 Testing Data (Diagnosis Test):**

* **Thrombocytopenia**: 14 instances
* **Other microcytic anemia**: 11 instances
* **Normocytic normochromic anemia**: 53 instances
* **Normocytic hypochromic anemia**: 55 instances
* **Macrocytic anemia**: 3 instances
* **Leukemia with thrombocytopenia**: 2 instances
* **Leukemia**: 9 instances
* **Iron deficiency anemia**: 37 instances
* **Healthy**: 67 instances
  + 1. **Comparison and Interpretation:**

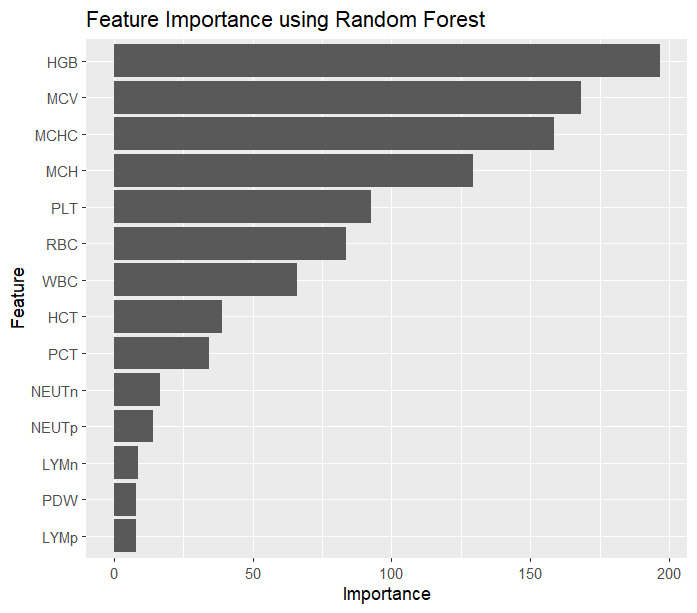
1. **Class Distribution**:
   * Both the training and testing datasets have similar distributions, which is crucial for ensuring that the machine learning model generalizes well to new data.
   * The most common diagnoses in both sets are **Normocytic hypochromic anemia** and **Normocytic normochromic anemia**.
   * **Healthy** instances are significantly more in the test set (67) compared to the train set (26).
2. **Imbalance**:
   * There is an imbalance in both datasets, with certain classes (like **Macrocytic anemia** and **Leukemia with thrombocytopenia**) having very few instances compared to others. This could affect the model’s performance on these less represented classes.
   * Techniques such as oversampling, undersampling, or using stratified sampling could be considered to address this imbalance.
3. **Model Evaluation**:
   * The presence of a similar distribution in both training and testing sets suggests that the evaluation metrics derived from the test set would be reliable and reflective of the model’s performance in a real-world scenario.
   * However, the model’s ability to correctly identify underrepresented classes might still be limited due to the inherent class imbalance.

**8.5 Feature Importance Analysis**

Feature importance analysis ranks the features of a dataset based on their contribution to the predictive power of a model. In random forests, feature importance is commonly derived from measures such as the mean decrease in impurity or the mean decrease in accuracy.

* **Mean Decrease in Impurity (Gini Importance):** Measures the total decrease in node impurity (like Gini impurity) brought by a feature, averaged over all trees in the forest.
* **Mean Decrease in Accuracy (Permutation Importance):** Measures the decrease in model accuracy when the feature values are randomly shuffled, averaged over all trees in the forest.

**Feature Importance Analysis using Random Forest**



**Figure 6.**

The plot illustrates the importance of various features in the Random Forest model

**8.5.1 Top Features**

* + **HGB (Hemoglobin):** Most important feature. Hemoglobin levels are critical in diagnosing different types of anemia.
  + **MCV (Mean Corpuscular Volume):** Second most important, indicating cell size is crucial for anemia classification.
  + **MCHC (Mean Corpuscular Hemoglobin Concentration):** High importance, showing concentration of hemoglobin in cells is key.

**8.5.2 Moderately Important Features**

* + **MCH (Mean Corpuscular Hemoglobin)**: Indicates the average amount of hemoglobin per red cell.
  + **PLT (Platelet count):** Highlights the role of platelets in differentiating conditions like thrombocytopenia.
  + **RBC (Red Blood Cell count):** RBC count is fundamental in the overall assessment of anemia.
  + **WBC (White Blood Cell count):** Important for differentiating conditions like leukemia.

**8.5.3 Lesser Important Features**

* + **HCT (Hematocrit):** Measures the proportion of blood volume occupied by red blood cells.
  + **PCT (Plateletcrit):** Reflects the total platelet mass, important but less so compared to other factors.
  + **NEUTn (Neutrophil number) and NEUTp (Neutrophil percentage):** Useful in specific conditions like infections and inflammatory diseases.

**8.5.4 Least Important Features**

* + **LYMn (Lymphocyte number) and LYMp (Lymphocyte percentage):** Although useful, they are less crucial compared to other parameters in anemia classification.
  + **PDW (Platelet Distribution Width):** Indicates variability in platelet size, which is less significant for anemia types classification.

**8.5.5 Summary of Feature Importance Analysis**

* + Hemoglobin (HGB) is the most critical feature for classifying anemia, followed by MCV and MCHC.
  + Features related to red blood cell characteristics (HGB, MCV, MCH, MCHC) dominate in importance, reflecting their central role in diagnosing anemia.
  + Platelet-related measures (PLT, PCT) and white blood cell counts (WBC) also play significant roles but are secondary to red blood cell parameters.
  + Understanding the importance of these features helps prioritize them in diagnostic models and clinical assessments.

1. **Conclusion**

Anemia is a widespread blood disorder that can have significant impacts on individuals' health and quality of life. Throughout this project, we've delved into various aspects of anemia, including its types, causes, symptoms, diagnostic criteria, and potential treatments. Our main aim being to develop effective predictive models to assist in diagnosing anemia based on blood test results, facilitating early detection and appropriate management.

**9.1 Key Findings**

**Strong Positive Correlations:** Between parameters like HGB and HCT, MCV and MCH, indicating these measures are often jointly affected in various blood conditions.

**Strong Negative Correlations:** Notably between LYMp and NEUTp/NEUTn, reflecting the inverse relationship between lymphocytes and neutrophils.

**Moderate Correlations:** RBC has moderate positive correlations with HGB, HCT, and WBC, indicating interconnectedness but with independent variability.

**Residual Deviance & AIC (Akaike Information Criterion):** The residual deviance decreased from 1340.229 to 964.223, and the Akaike Information Criterion (AIC) improved from 1580.229 to 1092.223 after applying stepwise regression. These reductions indicate that the refined model has significantly enhanced its ability to predict the factors involved in classifying the type of anemia.

**Likelihood Ratio Test:** The very small p-value (< 2.2e-16) indicates that the improvement in model fit from Model 2 to Model 1 is statistically significant. In other words, including the predictors (WBC, RBC, HGB, MCV, MCHC, PLT, PCT) in Model 1 significantly improves the model's ability to explain the variation in the diagnosis of anemia compared to the intercept-only Model 2.Thus, we reject the null hypothesis that the simpler model (Model 2) is sufficient, and conclude that the more complex model (Model 1) provides a significantly better fit to the data.

**Accuracy on Training Set:** The high accuracy scores on the training set indicate that both models were able to learn from the training data and make accurate predictions. The Random Forest model achieved a perfect accuracy of 100% on the training set, suggesting that it effectively captured the complex relationships within the training data.

**Accuracy on Test Set:** The accuracy scores on the test set are also high, indicating that both models generalize well to new, unseen data. The Random Forest model slightly outperforms the Decision Tree model on the test set, achieving a slightly higher accuracy score.

**Confusion Matrix:** Healthy, Iron deficiency anemia, Normocytic hypochromic anemia, and Normocytic normochromic anemia are predicted with high accuracy. Leukemia with thrombocytopenia and macrocytic anemia show higher misclassification rates, indicating difficulties in distinguishing these conditions with the current model. Some diagnoses are more frequently confused with specific others, suggesting potential overlap in symptoms or test results.

**Feature Importance Analysis:** - Hemoglobin (HGB) is the most critical feature for classifying anemia, followed by MCV and MCHC. Features related to red blood cell characteristics (HGB, MCV, MCH, MCHC) dominate in importance, reflecting their central role in diagnosing anemia. Platelet-related measures (PLT, PCT) and white blood cell counts (WBC) also play significant roles but are secondary to red blood cell parameters. Understanding the importance of these features helps prioritize them in diagnostic models and clinical assessments.

**9.2 Implications and Future Directions**

- **Clinical Application:** The developed models have the potential to assist healthcare professionals in diagnosing anemia accurately and efficiently, leading to timely interventions and improved patient outcomes.

- **Continued Refinement:** Further refinement and fine-tuning of the models can enhance their predictive accuracy and reliability. This may involve incorporating additional relevant features or exploring advanced machine learning techniques.

- **Integration with Healthcare Systems:** Integration of these predictive models into existing healthcare systems can streamline the diagnostic process and provide valuable decision support to clinicians.

**- Research Opportunities:** Continued research into the underlying mechanisms and risk factors of anemia can further enrich our understanding and inform the development of more sophisticated predictive models.

In conclusion, our project represents a significant step towards leveraging machine learning techniques for the early detection and management of anemia, ultimately contributing to improved healthcare outcomes and patient well-being.

**Annexure I**

**Terminologies**

* HGB: The amount of hemoglobin in the blood, crucial for oxygen transport.
* PLT: The number of platelets in the blood, involved in blood clotting.
* WBC: The count of white blood cells, vital for immune response.
* RBC: The count of red blood cells, responsible for oxygen transport.
* MCV (Mean Corpuscular Volume): Average volume of a single red blood cell.
* MCH (Mean Corpuscular Hemoglobin): Average amount of hemoglobin per red blood cell.
* MCHC (Mean Corpuscular Hemoglobin Concentration): Average concentration of hemoglobin in red blood cells.
* PDW: a measurement of the variability in platelet size distribution in the blood
* PCT: A procalcitonin test can help your health care provider diagnose if you have sepsis from a bacterial infection or if you have a high risk of developing sepsis
* LYMp: Lymphocytes percentage
* NEUTp: Neutrophils percentage
* LYMn: Lymphocytes number
* NEUTn: Neutrophils number
* Diagnosis: Anemia type based on the CBC parameters – Iron Deficiency Anemia, Thrombocytopenia, Normocytic Normochromic Anemia, Macrocytic Anemia, Normocytic Hypochromic Anemia, Leukemia, Leukemia with Thrombocytopenia, Other Microcytic Anemia and Healthy.

**Annexure II**

data = read.csv(choose.files())

head(data)

summary(data)

str(data)

# Libraries

library(ggplot2)

library(gridExtra)

library(dplyr)

library(qqplotr)

library(nnet)

library(MASS)

library(caret)

library(reshape2)

library(gridExtra)

library(lmtest)

library(rpart)

library(randomForest)

# Model Learning

# Define the partition

trainIndex <- createDataPartition(y, p = 0.8, list = FALSE, times = 1)

# Split the data

X\_train <- X[trainIndex, ]

X\_test <- X[-trainIndex, ]

y\_train <- y[trainIndex]

y\_test <- y[-trainIndex]

# Print the number of samples in the training and testing sets

cat("X\_train samples:", length(y\_train), "\n")

cat("X\_test samples:", length(y\_test), "\n")

# Convert "Diagnosis" into a factor

data$Diagnosis <- as.factor(data$Diagnosis)

# Split the data into training and test sets

set.seed(123) # For reproducibility

sample <- sample(1:nrow(data), 0.7 \* nrow(data))

train\_data <- data[sample, ]

test\_data <- data[-sample, ]

# Fit a multinomial logistic regression model

model <- multinom(Diagnosis ~ WBC + LYMp + NEUTp + LYMn + NEUTn + RBC + HGB + HCT + MCV + MCH + MCHC + PLT + PDW + PCT, data = train\_data)

model

# Check the summary of the model

summary(model)

# Perform stepwise selection using AIC as the criterion

stepwise\_model <- stepAIC(model, direction = "both")

stepwise\_model

summary(stepwise\_model)

# Make predictions on the test set

predicted <- predict(model, newdata = test\_data)

predicted <- predict(stepwise\_model, newdata = test\_data)

# Evaluate the model

confusion\_matrix <- table(predicted, test\_data$Diagnosis)

print(confusion\_matrix)

X <- data[, !(names(data) %in% "Diagnosis")]

y <- data$Diagnosis

# Likelihood Ratio Test

lr\_test <- lrtest(stepwise\_model)

# Print test results

print(lr\_test)

# Train the Random Forest classification model

model2 <- randomForest(Diagnosis ~ ., data = data, importance = TRUE)

# Get feature importances

importance\_scores <- importance(model2)

# Convert to a data frame for easier plotting

importance\_df <- data.frame(Feature = rownames(importance\_scores), Importance = importance\_scores[, "MeanDecreaseGini"])

# Plot feature importances

ggplot(importance\_df, aes(x = reorder(Feature, Importance), y = Importance)) +

geom\_bar(stat = "identity") +

coord\_flip() +

xlab("Feature") +

ylab("Importance") +

ggtitle("Feature Importance using Random Forest")

#MODEL BUILDING

# Ensure y\_pred and y\_true have the same levels

ensure\_levels <- function(y\_pred, y\_true) {

levels(y\_pred) <- levels(y\_true)

return(y\_pred)

}

# Function to calculate balanced accuracy

balanced\_accuracy <- function(y\_true, y\_pred) {

y\_pred <- ensure\_levels(y\_pred, y\_true)

cm <- confusionMatrix(y\_pred, y\_true)

return(mean(cm$byClass[,"Balanced Accuracy"]))

}

# Initialize models

SEED <- 42

models <- list(

dt = rpart,

rf = randomForest

)

# Loop through models

for (name in names(models)) {

cat(paste0("\* ", name, " | "))

# Fit model

if (name == "dt") {

model <- models[[name]](y\_train ~ ., data = X\_train)

y\_pred\_train <- predict(model, X\_train, type = "class")

y\_pred\_test <- predict(model, X\_test, type = "class")

} else if (name == "rf") {

model <- models[[name]](X\_train, y\_train, ntree = 100)

y\_pred\_train <- predict(model, X\_train)

y\_pred\_test <- predict(model, X\_test)

}

# Calculate balanced accuracy

acc\_train <- balanced\_accuracy(y\_train, y\_pred\_train)

acc\_test <- balanced\_accuracy(y\_test, y\_pred\_test)

cat(sprintf("Acc Train: %.4f | Acc Test: %.4f\n", acc\_train, acc\_test))

}

# Plots

# Select numeric columns

features <- data %>% select\_if(is.numeric) %>% names()

# Create color palette

colors <- scales::hue\_pal()(length(features))

# Create individual plots and store in a list

plots <- list()

for (i in seq\_along(features)) {

p <- ggplot(data, aes\_string(x = features[i])) +

geom\_histogram(aes(y = ..density..), fill = NA, color = colors[i], bins = 30) +

geom\_density(fill = colors[i], alpha = 0.4) +

labs(title = features[i]) +

theme\_light() +

theme(plot.title = element\_text(size = 10, face = "bold"),

axis.title.x = element\_blank())

plots[[i]] <- p

}

# Arrange plots in a 7x2 grid

grid.arrange(grobs = plots, ncol = 2, nrow = 7)

# Create individual Q-Q plots and store in a list

plots <- list()

for (i in seq\_along(features)) {

p <- ggplot(data, aes(sample = !!sym(features[i]))) +

stat\_qq\_line() +

stat\_qq\_point() +

labs(title = features[i]) +

theme\_light() +

theme(plot.title = element\_text(size = 10, face = "bold"),

axis.title.x = element\_blank(),

axis.title.y = element\_blank())

plots[[i]] <- p

}

# Arrange plots in a 7x2 grid

grid.arrange(grobs = plots, ncol = 4, nrow = 4)

# Create the count plot with horizontal bars and adjusted text

plot <- ggplot(data, aes(x = Diagnosis)) +

geom\_bar(fill = "lightblue", color = "black") +

geom\_text(stat = "count", aes(label = ..count..), hjust = -0.1, size = 3, fontface = "bold") +

labs(x = "", y = "Count", title = "Diagnosis") +

theme\_light() +

theme(

plot.title = element\_text(size = 14, face = "bold", color = "darkblue"),

axis.text.y = element\_text(size = 10, face = "bold", color = "black")

) +

coord\_flip() # Flip the coordinates to make the bar plot horizontal

# Show the plot

print(plot)

# Select only numeric columns

numeric\_data <- data %>% select\_if(is.numeric)

# Calculate the correlation matrix using the Spearman method

corr\_matrix <- cor(numeric\_data, method = "spearman")

# Convert the correlation matrix to long format

corr\_matrix\_long <- melt(corr\_matrix)

# Create the heatmap

plot <- ggplot(corr\_matrix\_long, aes(Var1, Var2, fill = value)) +

geom\_tile(color = "black") +

scale\_fill\_gradient2(low = "blue", mid = "white", high = "red",

midpoint = 0, limits = c(-1, 1), name = "Correlation") +

geom\_text(aes(label = round(value, 2)), size = 3, fontface = "bold") +

theme\_minimal() +

labs(x = "", y = "", title = "Correlation Matrix", color = "black",

title.fontface = "bold", fill = "Correlation") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

# Show the plot

print(plot)

# Convert y\_train to a data frame for ggplot

df\_y\_train <- data.frame(Diagnosis = y\_train)

# Create the plot

ggplot(data = df\_y\_train, aes(x = Diagnosis)) +

geom\_bar(aes(y = after\_stat(count)), fill = "skyblue", color = "black", width = 0.7) +

geom\_text(stat = 'count', aes(label = after\_stat(count)), hjust = -0.1, size = 3, fontface = "bold", color = "black") +

labs(title = "Diagnosis Train", y = "", x = "") +

theme\_minimal() +

theme(

plot.title = element\_text(size = 14, face = "bold", color = "darkblue"),

axis.text.y = element\_text(size = 10, face = "bold", color = "black")

) +

coord\_flip() # Flip the coordinates to match the horizontal bar plot

df\_y\_test <- data.frame(Diagnosis = y\_test)

# Create the plot for y\_test

ggplot(data = df\_y\_test, aes(x = Diagnosis)) +

geom\_bar(aes(y = after\_stat(count)), fill = "lightcoral", color = "black", width = 0.7) +

geom\_text(stat = 'count', aes(label = after\_stat(count)), hjust = -0.1, size = 3, fontface = "bold", color = "black") +

labs(title = "Diagnosis Test", y = "", x = "") +

theme\_minimal() +

theme(

plot.title = element\_text(size = 14, face = "bold", color = "darkblue"),

axis.text.y = element\_text(size = 10, face = "bold", color = "black")

) +

coord\_flip() # Flip the coordinates to match the horizontal bar plot

# Manual Testing

# Create a new sample data frame

new\_sample <- data.frame(WBC = 8.0, LYMp = 30.0, NEUTp = 60.0, LYMn = 1.5, NEUTn = 4.0, RBC = 4.5,

HGB = 12.0, HCT = 36.0, MCV = 80.0, MCH = 27.0, MCHC = 33.0, PLT = 200, PDW = 15.0, PCT = 0.25)

# Make a prediction

new\_sample\_prediction <- predict(model2, new\_sample)

# Print the prediction

print(new\_sample\_prediction)

**References**

* Adamson, J. W., & Longo, D. L. (2012). Anemia and polycythemia. In D. L. Longo (Ed.), *Harrison's Principles of Internal Medicine* (18th ed.). McGraw-Hill.
* Aslinia, F., Mazza, J. J., & Yale, S. H. (2006). Megaloblastic anemia and other causes of macrocytosis. *Clinical Medicine & Research*, 4(3), 236-241.
* Auerbach, M., Adamson, J. W., & Goodnough, L. T. (2020). Iron deficiency anemia: Evaluation and management. *Hematology/Oncology Clinics of North America*, 34(1), 177-195.
* Camaschella, C. (2015). Iron-deficiency anemia. *New England Journal of Medicine*, 372(19), 1832-1843.
* Gomes, C., & Berger, E. (2015). Hemolytic anemia. *Pediatric Clinics of North America*, 62(4), 853-866.
* Green, R., & Allen, L. H. (2015). Anemia and vitamin B12 deficiency. *Annual Review of Nutrition*, 35, 357-382.
* Hoffbrand, A. V., & Moss, P. A. H. (2011). *Essential Haematology* (6th ed.). Wiley-Blackwell.
* Kato, G. J., Piel, F. B., Reid, C. D., Gaston, M. H., Ohene-Frempong, K., Krishnamurti, L., ... & Vichinsky, E. P. (2018). Sickle cell disease. *Nature Reviews Disease Primers*, 4, 18010.
* Munoz, M., Garcia-Erce, J. A., & Remacha, A. F. (2009). Disorders of iron metabolism. Part II: Iron deficiency and iron overload. *Journal of Clinical Pathology*, 62(4), 318-326.
* Weiss, G., & Goodnough, L. T. (2005). Anemia of chronic disease. *New England Journal of Medicine*, 352(10), 1011-1023.
* Kaggle Dataset.<https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification>