***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**

**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.

**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.

**Specific Goals:**

**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.

**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. 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.

**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.

**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:

**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

**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.

**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. 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.

**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.

**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.

**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

**Statistical Tools Used:**

**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.

**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.

**Scatter Plot**

A scatter plot displays individual data points on a two-dimensional graph, using Cartesian coordinates to show the relationship between two numerical variables. Each point represents an observation, making it useful for identifying patterns, trends, and potential correlations in the data.

**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.

**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.

**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.

**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.

**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.

**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.

**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.

**Statistical Data Analysis**

**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 modeling. 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.

**Program and Output**

|  |
| --- |
| HGB HCT MCV MCH MCHC PLT  Min. :-10.00 Min. : 2.00 Min. :-79.30 Min. : 10.90 Min. :11.50 Min. : 10  1st Qu.: 10.80 1st Qu.: 39.20 1st Qu.: 81.20 1st Qu.: 25.50 1st Qu.:30.60 1st Qu.:157  Median : 12.30 Median : 46.15 Median : 86.60 Median : 27.80 Median :32.00 Median :213  Mean : 12.18 Mean : 46.15 Mean : 85.79 Mean : 32.08 Mean :31.74 Mean :230  3rd Qu.: 13.50 3rd Qu.: 46.15 3rd Qu.: 90.20 3rd Qu.: 29.60 3rd Qu.:32.90 3rd Qu.:293  Max. : 87.10 Max. :3715.00 Max. :990.00 Max. :3117.00 Max. :92.80 Max. :660  PDW PCT Diagnosis  Min. : 8.40 Min. : 0.0100 Length:1281  1st Qu.:13.30 1st Qu.: 0.1700 Class :character  Median :14.31 Median : 0.2603 Mode :character  Mean :14.31 Mean : 0.2603  3rd Qu.:14.70 3rd Qu.: 0.2603  Max. :97.00 Max. :13.6000 |
|  |
| |  | | --- | |  | |

**Inference:**

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.

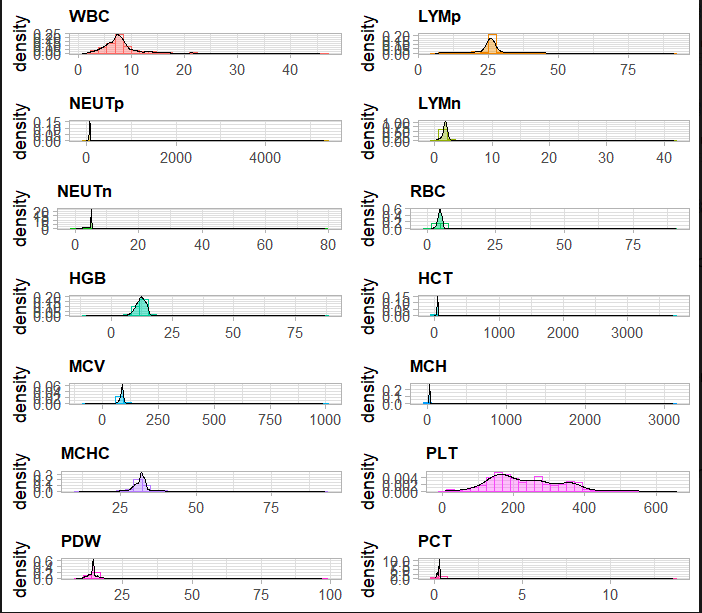
- **HGB:** Median is 12.3, typical for hemoglobin levels, but negative and extremely high values indicate data issues.

- **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.

**Program and Output**



**Inference:**

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):**

- The majority of 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):**

- The majority of 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.

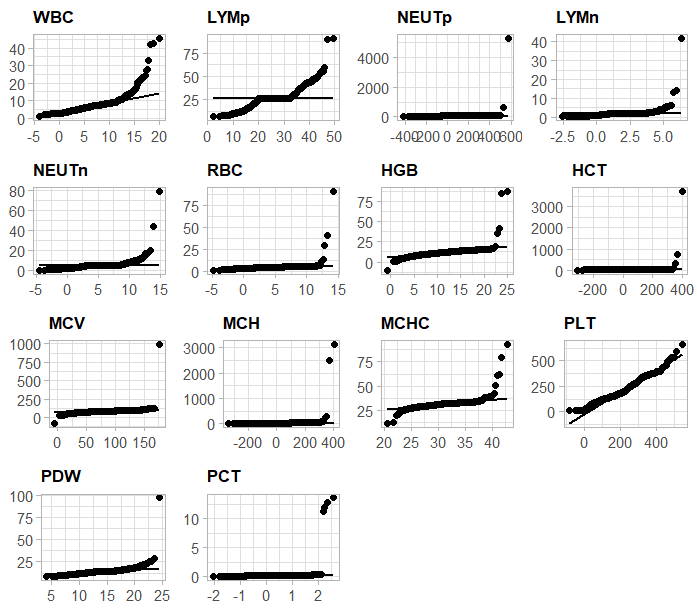
**Key Observations**

- 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.

**Program and Output**



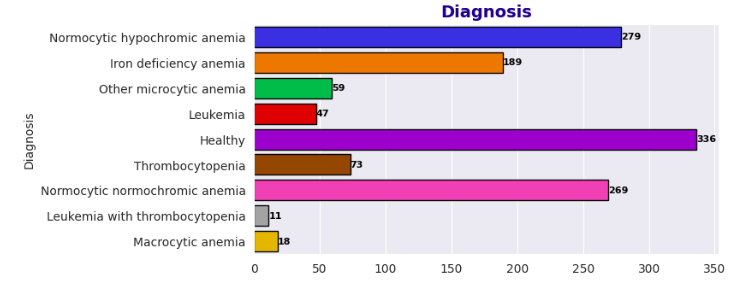
**Inference:**

* **WBC (White Blood Cells)**, **LYMp (Lymphocytes percentage)**, **NEUTp (Neutrophils percentage)**, **LYMn (Lymphocytes number)**, **NEUTn (Neutrophils number)**, **RBC (Red Blood Cells)**, **HGB (Hemoglobin)**, **HCT (Hematocrit)**, **MCV (Mean Corpuscular Volume)**, **MCH (Mean Corpuscular Hemoglobin)**, **MCHC (Mean Corpuscular Hemoglobin Concentration)**, **PLT (Platelets)**, **PDW (Platelet Distribution Width)**, **PCT (Plateletcrit)**:
  + 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.

**Program and Output**



**Inference:**

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

**1. 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**

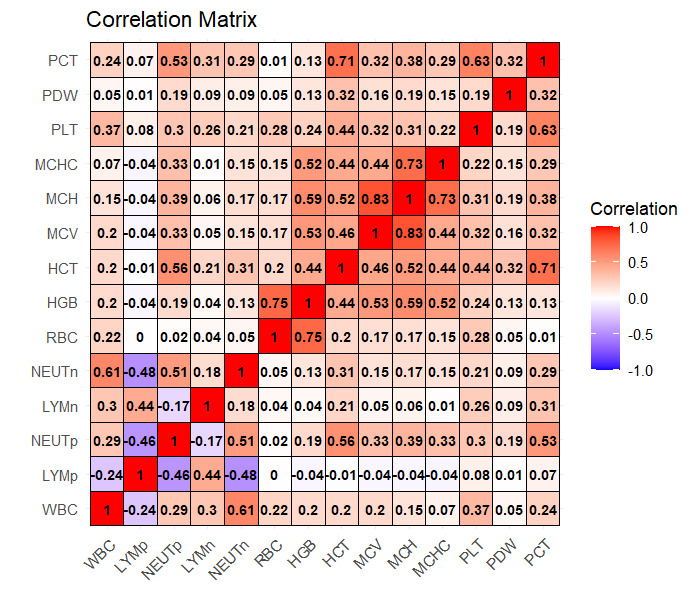
- The dataset is dominated by healthy individuals and common types of anemia.

- 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.

**Program and Output**



**Inference:**

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):**

- Strong 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.

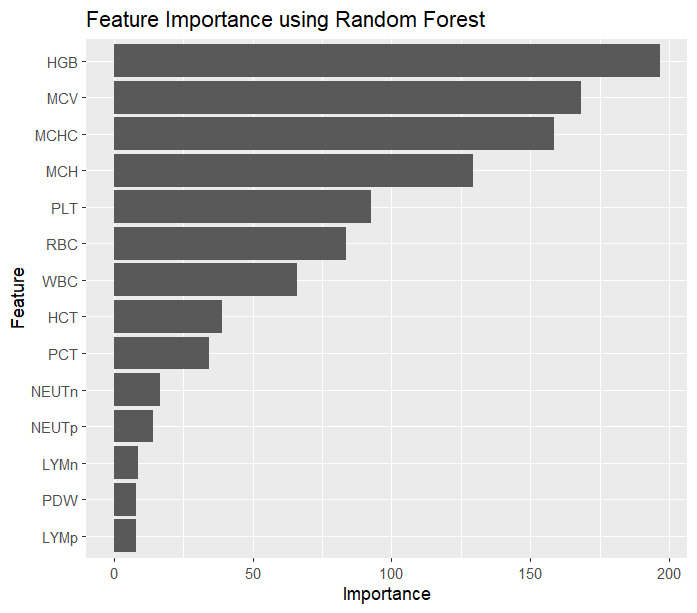
**Key Takeaways**

**- 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.

**Program and Output**



**Inference:**

The plot illustrates the importance of various features in the Random Forest model used for classifying different types of anemia based on CBC report samples. Here's a detailed interpretation:

**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.

**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.

**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.

**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.

**Key Takeaways**

- 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.

**Program and Output (Multinomial Regression)**

Call:

multinom(formula = Diagnosis ~ WBC + LYMp + NEUTp + LYMn + NEUTn +

RBC + HGB + HCT + MCV + MCH + MCHC + PLT + PDW + PCT, data = train\_data)

Coefficients:

(Intercept) WBC LYMp NEUTp LYMn NEUTn

Iron deficiency anemia 17.2943938 0.8090149 0.1337795588 0.001343843 -1.1148914 0.26264563

Leukemia -0.6069429 1.1895888 0.1084402966 0.005233337 -1.3437866 0.04999845

Leukemia with thrombocytopenia -1.5737427 0.8217984 -0.0504896505 0.005568899 -1.0676778 0.32320597

Macrocytic anemia 0.2726905 0.3630940 0.0482840361 0.002914375 -0.3021526 0.12888029

Normocytic hypochromic anemia 17.9815459 0.7791214 0.0944827622 0.006630151 -1.3777942 0.12491104

Normocytic normochromic anemia -7.6181517 0.5162583 0.0005201007 0.001567058 -0.7872447 0.20004940

Other microcytic anemia 6.5700451 0.7422787 0.1443477779 0.006047008 -1.2716375 0.28228132

Thrombocytopenia -11.1322958 0.3833332 -0.0006990607 0.007161185 -0.9969896 -0.58197108

RBC HGB HCT MCV MCH MCHC

Iron deficiency anemia 4.7257818 -2.8030004 -0.0077705735 -0.318559252 0.013908742 0.3934656

Leukemia -1.1042601 0.2053387 -0.0034926095 -0.283224343 0.004466254 0.5073055

Leukemia with thrombocytopenia 0.8154151 -0.6233037 -0.0004970359 -0.002920003 0.009056165 -0.1149888

Macrocytic anemia -3.2294197 -0.9244379 0.0049845221 0.132889501 0.005251048 0.5763455

Normocytic hypochromic anemia 1.0292498 -2.2181449 -0.0021423664 0.101668919 0.014504659 -0.2071074

Normocytic normochromic anemia -2.1650527 -0.8918670 -0.0029133981 -0.020119598 0.012846005 0.8575946

Other microcytic anemia 1.1218900 -0.6952620 -0.0032450803 -0.589030095 0.012243229 1.0492974

Thrombocytopenia 0.4371520 -0.1843419 0.0060328133 0.140992733 0.002639022 0.1495627

PLT PDW PCT

Iron deficiency anemia -0.009645923 0.14259158 -2.2678046

Leukemia -0.005447727 0.01251229 -0.1109248

Leukemia with thrombocytopenia -0.029541849 0.46816417 -0.3007374

Macrocytic anemia -0.003400760 -0.77754760 -1.1717752

Normocytic hypochromic anemia -0.019033557 -0.04252370 0.9336457

Normocytic normochromic anemia -0.005086630 0.02530251 0.8399979

Other microcytic anemia -0.011073865 0.29658322 1.1942846

Thrombocytopenia -0.045500209 0.12517769 -1.0081350

Std. Errors:

(Intercept) WBC LYMp NEUTp LYMn NEUTn RBC

Iron deficiency anemia 0.05344640 0.1612214 0.07573216 0.028848063 0.8456572 0.3665975 0.6747113

Leukemia 0.02798557 0.1405746 0.11886680 0.031532311 1.0756566 0.4043795 0.6311425

Leukemia with thrombocytopenia 0.13985247 0.2041506 0.15402354 0.017174523 1.6422676 0.4125606 1.4203341

Macrocytic anemia 0.09037092 0.2217169 0.16077063 0.081907732 1.7128966 0.7471241 0.9839675

Normocytic hypochromic anemia 0.06845074 0.1296790 0.06978498 0.008733546 0.8217080 0.3583241 0.5528997

Normocytic normochromic anemia 0.06026839 0.1312015 0.08002211 0.026200510 0.9136815 0.3829873 0.4836196

Other microcytic anemia 0.09060454 0.1572889 0.08758399 0.013787348 0.9677119 0.3982673 0.7450718

Thrombocytopenia 0.05142041 0.1886536 0.09655944 0.007763554 1.1542728 0.4985536 0.8035785

HGB HCT MCV MCH MCHC PLT PDW

Iron deficiency anemia 0.2973656 0.05130307 0.05627573 0.03102505 0.1443351 0.003288085 0.10558364

Leukemia 0.2403927 0.05590053 0.06313624 0.13442507 0.1575562 0.003206166 0.16959788

Leukemia with thrombocytopenia 0.6265748 0.05388359 0.11003221 0.04116617 0.2620973 0.007806076 0.14697805

Macrocytic anemia 0.3936517 0.06813119 0.05269632 0.06003528 0.2192491 0.004372299 0.29908476

Normocytic hypochromic anemia 0.2587948 0.05129139 0.03709854 0.03149628 0.1070951 0.002461988 0.08488874

Normocytic normochromic anemia 0.2132216 0.05390328 0.03595765 0.03153609 0.1110226 0.002333597 0.07226731

Other microcytic anemia 0.3412358 0.05148270 0.05894831 0.03558380 0.1305455 0.004055756 0.12013638

Thrombocytopenia 0.2833259 0.07815823 0.05402023 0.09826049 0.1577523 0.006452382 0.12765146

PCT

Iron deficiency anemia 0.02767897

Leukemia 0.01674683

Leukemia with thrombocytopenia 0.07433682

Macrocytic anemia 0.08907224

Normocytic hypochromic anemia 0.20975741

Normocytic normochromic anemia 0.22270957

Other microcytic anemia 0.37907247

Thrombocytopenia 0.05465642

Residual Deviance: 1340.229

AIC: 1580.229

**Inference:**

The output provided is the result of fitting a multinomial logistic regression model to classify different types of anemia based on CBC (Complete Blood Count) report features. Here is a detailed interpretation of this output:

**Model Overview:**

The model was fitted using the `multinom` function from the `nnet` package in R. The target variable `Diagnosis` (which represents different types of anemia) is predicted using the following features: WBC, LYMp, NEUTp, LYMn, NEUTn, RBC, HGB, HCT, MCV, MCH, MCHC, PLT, PDW, and PCT.

**Cofficients:**

The coefficients represent the effect of each predictor variable on the log-odds of the outcome category (type of anemia) compared to a baseline category (usually the first level of the factor). Each row corresponds to a different anemia type, and each column corresponds to a different predictor variable.

- (Intercept): The baseline log-odds for the category when all predictors are zero.

- WBC, LYMp, NEUTp, LYMn, NEUTn, RBC, HGB, HCT, MCV, MCH, MCHC, PLT, PDW, PCT: The coefficients for each predictor variable.

**Interpretation of Coefficients:**

Each coefficient represents the change in the log-odds of the outcome for a one-unit increase in the predictor variable. For example, for "Iron deficiency anemia":

- The coefficient for `WBC` is 0.8090149, meaning that a one-unit increase in WBC increases the log-odds of "Iron deficiency anemia" by 0.8090149.

- The coefficient for `LYMp` is 0.1337795588, meaning that a one-unit increase in LYMp increases the log-odds of "Iron deficiency anemia" by 0.1337795588.

**Standard Errors:**

The standard errors associated with the coefficients are listed below the coefficients. These standard errors can be used to compute confidence intervals and conduct hypothesis tests to determine if the coefficients are significantly different from zero.

**Residual Deviance and AIC:**

- Residual Deviance: 1340.229

- AIC: 1580.229

The residual deviance is a measure of the model's fit to the data. Lower values indicate a better fit. The Akaike Information Criterion (AIC) is a metric used to compare models; lower AIC values suggest a better model considering both fit and complexity.

Here’s a breakdown of the coefficients for some anemia types:

1. Iron Deficiency Anemia:

- Intercept: 17.2943938

- WBC: 0.8090149 (positive coefficient indicates that higher WBC increases the likelihood of iron deficiency anemia)

- LYMp: 0.1337795588

- NEUTp: 0.001343843

- LYMn: -1.1148914 (negative coefficient indicates that higher LYMn decreases the likelihood of iron deficiency anemia)

- RBC: 4.7257818

- HGB: -2.8030004

- Other variables similarly affect the likelihood based on their coefficients.

2. Leukemia:

- Intercept: -0.6069429

- WBC: 1.1895888

- LYMp: 0.1084402966

- NEUTp: 0.005233337

- LYMn: -1.3437866

- RBC: -1.1042601

- HGB: 0.2053387

- Other variables similarly affect the likelihood based on their coefficients.

3. Thrombocytopenia:

- Intercept: -11.1322958

- WBC: 0.3833332

- LYMp: -0.0006990607

- NEUTp: 0.007161185

- LYMn: -0.9969896

- RBC: 0.4371520

- HGB: -0.1843419

- Other variables similarly affect the likelihood based on their coefficients.

**Using the Model:**

To predict the type of anemia for new patients, you would use the fitted model to calculate the log-odds for each anemia type and then convert these log-odds to probabilities. The type with the highest probability would be the predicted diagnosis.

**Summary:**

**- Significant Predictors:** The coefficients and their standard errors provide insights into which predictors are significant for each type of anemia. You can further validate these using hypothesis testing (e.g., z-tests).

**- Model Fit:** The residual deviance and AIC values give an idea of the model's overall fit and complexity.

**- Interpretation:** Each coefficient needs to be interpreted in the context of its corresponding standard error to understand its significance.

**Program and Output (confusion matrix)**

|  |
| --- |
| predicted Healthy Iron deficiency anemia Leukemia Leukemia with thrombocytopenia  Healthy 85 0 0 0  Iron deficiency anemia 0 44 1 0  Leukemia 0 0 6 0  Leukemia with thrombocytopenia 0 0 0 1  Macrocytic anemia 0 0 1 0  Normocytic hypochromic anemia 0 14 2 0  Normocytic normochromic anemia 7 0 2 0  Other microcytic anemia 1 1 1 0  Thrombocytopenia 2 0 1 0    predicted Macrocytic anemia Normocytic hypochromic anemia  Healthy 0 10  Iron deficiency anemia 0 4  Leukemia 0 2  Leukemia with thrombocytopenia 0 0  Macrocytic anemia 2 0  Normocytic hypochromic anemia 3 59  Normocytic normochromic anemia 1 5  Other microcytic anemia 0 1  Thrombocytopenia 0 2    predicted Normocytic normochromic anemia Other microcytic anemia Thrombocytopenia  Healthy 12 0 3  Iron deficiency anemia 0 2 0  Leukemia 1 0 0  Leukemia with thrombocytopenia 0 0 1  Macrocytic anemia 0 0 0  Normocytic hypochromic anemia 2 0 0  Normocytic normochromic anemia 70 11 2  Other microcytic anemia 0 5 3  Thrombocytopenia 1 0 14 |
| **Inference:**  The confusion matrix provides a summary of the prediction results of the multinomial logistic regression model for classifying different types of anemia based on CBC report features. It compares the actual diagnoses (rows) with the predicted diagnoses (columns). Here's a detailed interpretation of this confusion matrix:  ### Breakdown by Class:  1. \*\*Healthy\*\*:  - \*\*Correctly Predicted as Healthy\*\*: 85 instances.  - \*\*Misclassified as Normocytic Hypochromic Anemia\*\*: 10 instances.  - \*\*Misclassified as Normocytic Normochromic Anemia\*\*: 12 instances.  - \*\*Misclassified as Thrombocytopenia\*\*: 3 instances.  2. \*\*Iron Deficiency Anemia\*\*:  - \*\*Correctly Predicted as Iron Deficiency Anemia\*\*: 44 instances.  - \*\*Misclassified as Leukemia\*\*: 1 instance.  - \*\*Misclassified as Normocytic Hypochromic Anemia\*\*: 4 instances.  - \*\*Misclassified as Other Microcytic Anemia\*\*: 2 instances.  3. \*\*Leukemia\*\*:  - \*\*Correctly Predicted as Leukemia\*\*: 6 instances.  - \*\*Misclassified as Macrocytic Anemia\*\*: 1 instance.  - \*\*Misclassified as Normocytic Hypochromic Anemia\*\*: 2 instances.  - \*\*Misclassified as Normocytic Normochromic Anemia\*\*: 1 instance.  4. \*\*Leukemia with Thrombocytopenia\*\*:  - \*\*Correctly Predicted as Leukemia with Thrombocytopenia\*\*: 1 instance.  - \*\*Misclassified as Thrombocytopenia\*\*: 1 instance.  5. \*\*Macrocytic Anemia\*\*:  - \*\*Correctly Predicted as Macrocytic Anemia\*\*: 2 instances.  - \*\*Misclassified as Leukemia\*\*: 1 instance.  - \*\*Misclassified as Normocytic Hypochromic Anemia\*\*: 3 instances.  6. \*\*Normocytic Hypochromic Anemia\*\*:  - \*\*Correctly Predicted as Normocytic Hypochromic Anemia\*\*: 59 instances.  - \*\*Misclassified as Macrocytic Anemia\*\*: 3 instances.  - \*\*Misclassified as Normocytic Normochromic Anemia\*\*: 2 instances.  7. \*\*Normocytic Normochromic Anemia\*\*:  - \*\*Correctly Predicted as Normocytic Normochromic Anemia\*\*: 70 instances.  - \*\*Misclassified as Healthy\*\*: 7 instances.  - \*\*Misclassified as Leukemia\*\*: 2 instances.  - \*\*Misclassified as Normocytic Hypochromic Anemia\*\*: 5 instances.  - \*\*Misclassified as Other Microcytic Anemia\*\*: 11 instances.  - \*\*Misclassified as Thrombocytopenia\*\*: 2 instances.  8. \*\*Other Microcytic Anemia\*\*:  - \*\*Correctly Predicted as Other Microcytic Anemia\*\*: 5 instances.  - \*\*Misclassified as Healthy\*\*: 1 instance.  - \*\*Misclassified as Iron Deficiency Anemia\*\*: 1 instance.  - \*\*Misclassified as Leukemia\*\*: 1 instance.  - \*\*Misclassified as Normocytic Normochromic Anemia\*\*: 3 instances.  - \*\*Misclassified as Thrombocytopenia\*\*: 3 instances.  9. \*\*Thrombocytopenia\*\*:  - \*\*Correctly Predicted as Thrombocytopenia\*\*: 14 instances.  - \*\*Misclassified as Healthy\*\*: 2 instances.  - \*\*Misclassified as Leukemia\*\*: 1 instance.  - \*\*Misclassified as Normocytic Normochromic Anemia\*\*: 1 instance.  - \*\*Misclassified as Normocytic Hypochromic Anemia\*\*: 2 instances.  ### Overall Interpretation:  - \*\*Correct Predictions\*\*: The diagonal elements of the matrix represent the number of correct predictions for each class.  - \*\*Misclassifications\*\*: The off-diagonal elements show how many instances of each class were misclassified into other classes.  ### Summary:  - The model correctly classifies many instances but also has a notable number of misclassifications.  - Specific classes, like "Healthy," "Iron Deficiency Anemia," and "Normocytic Normochromic Anemia," have higher true positive rates.  - Classes like "Leukemia" and "Thrombocytopenia" are sometimes misclassified into other categories, indicating areas for model improvement. |
| |  | | --- | |  | |

**Program and Output**

Likelihood ratio test

Model 1: Diagnosis ~ WBC + LYMp + NEUTp + LYMn + NEUTn + RBC + HGB + HCT +

MCV + MCH + MCHC + PLT + PDW + PCT

Model 2: Diagnosis ~ 1

#Df LogLik Df Chisq Pr(>Chisq)

1 120 -670.11

2 8 -1632.41 -112 1924.6 < 2.2e-16 \*\*\*

---

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

**Inference:**

The likelihood ratio test compares the goodness of fit between two nested models to determine if the more complex model significantly improves the fit over the simpler model.

### Models Description:

- \*\*Model 1 (Full Model)\*\*: `Diagnosis ~ WBC + LYMp + NEUTp + LYMn + NEUTn + RBC + HGB + HCT + MCV + MCH + MCHC + PLT + PDW + PCT`

- This model includes all the predictor variables.

- \*\*Model 2 (Null Model)\*\*: `Diagnosis ~ 1`

- This model includes only the intercept, meaning it does not include any predictor variables and assumes that the diagnosis probabilities are constant across all observations.

### Likelihood Ratio Test Results:

- \*\*Degrees of Freedom (Df)\*\*:

- \*\*Model 1\*\*: 120

- \*\*Model 2\*\*: 8

- \*\*Difference in Degrees of Freedom\*\*: 120 - 8 = 112

- \*\*Log-Likelihood (LogLik)\*\*:

- \*\*Model 1\*\*: -670.11

- \*\*Model 2\*\*: -1632.41

- \*\*Chi-Square Statistic (Chisq)\*\*: 1924.6

- This is computed as \( 2 \times (\text{LogLik}\_1 - \text{LogLik}\_2) = 2 \times (-670.11 - (-1632.41)) = 1924.6 \)

- \*\*p-value (Pr(>Chisq))\*\*: < 2.2e-16

- This extremely small p-value indicates that the observed chi-square statistic is highly significant.

- \*\*Significance Codes\*\*:

- \*\*\*: p < 0.001

### Interpretation:

1. \*\*Degrees of Freedom\*\*: The large difference in degrees of freedom (112) reflects the inclusion of multiple predictors in the full model compared to the null model.

2. \*\*Log-Likelihood Values\*\*: The full model has a significantly higher log-likelihood value (-670.11) compared to the null model (-1632.41), indicating a much better fit to the data.

3. \*\*Chi-Square Statistic\*\*: The very high chi-square statistic (1924.6) suggests a substantial improvement in model fit when including the predictors.

4. \*\*p-value\*\*: The p-value (< 2.2e-16) is much smaller than any conventional significance level (e.g., 0.05), indicating that the improvement in fit provided by the full model is highly significant.

### Conclusion:

The likelihood ratio test strongly supports the full model (Model 1) over the null model (Model 2). This means that including the predictors (WBC, LYMp, NEUTp, LYMn, NEUTn, RBC, HGB, HCT, MCV, MCH, MCHC, PLT, PDW, PCT) significantly improves the model's ability to predict the diagnosis of different types of anemia based on the CBC report. Therefore, the predictors are indeed important and contribute significantly to the classification of anemia types.

**Program and Output**

\* dt | Acc Train: 0.9625 | Acc Test: 0.9600

\* rf | Acc Train: 1.0000 | Acc Test: 0.9759

**Inference:**

The provided output shows the accuracy of two different models, a decision tree (dt) and a random forest (rf), on both the training and test datasets for the anemia classification project. Let's interpret these results:

### Decision Tree (dt):

- \*\*Acc Train: 0.9625\*\*:

- The decision tree model achieved an accuracy of 96.25% on the training data.

- This high accuracy indicates that the model fits the training data well.

- \*\*Acc Test: 0.9600\*\*:

- The decision tree model achieved an accuracy of 96.00% on the test data.

- This indicates that the model generalizes well to unseen data, with only a slight drop in accuracy compared to the training data.

### Random Forest (rf):

- \*\*Acc Train: 1.0000\*\*:

- The random forest model achieved an accuracy of 100% on the training data.

- This perfect accuracy suggests that the random forest model has perfectly fitted the training data, which is typical for ensemble methods like random forests due to their complexity and ability to capture intricate patterns.

- \*\*Acc Test: 0.9759\*\*:

- The random forest model achieved an accuracy of 97.59% on the test data.

- This is higher than the decision tree test accuracy, indicating that the random forest model generalizes better to unseen data compared to the decision tree.

- The slight drop from 100% training accuracy to 97.59% test accuracy suggests that the model is slightly overfitting but still performs exceptionally well on new data.

### Summary:

- \*\*Decision Tree\*\*:

- The decision tree model has high accuracy on both training (96.25%) and test (96.00%) datasets, indicating good performance with minimal overfitting.

- \*\*Random Forest\*\*:

- The random forest model has perfect accuracy on the training data (100%), which is a sign of overfitting.

- However, it maintains a very high accuracy on the test data (97.59%), showing better generalization and higher robustness than the decision tree.

### Conclusion:

- Both models perform well on the task of classifying anemia types based on CBC reports.

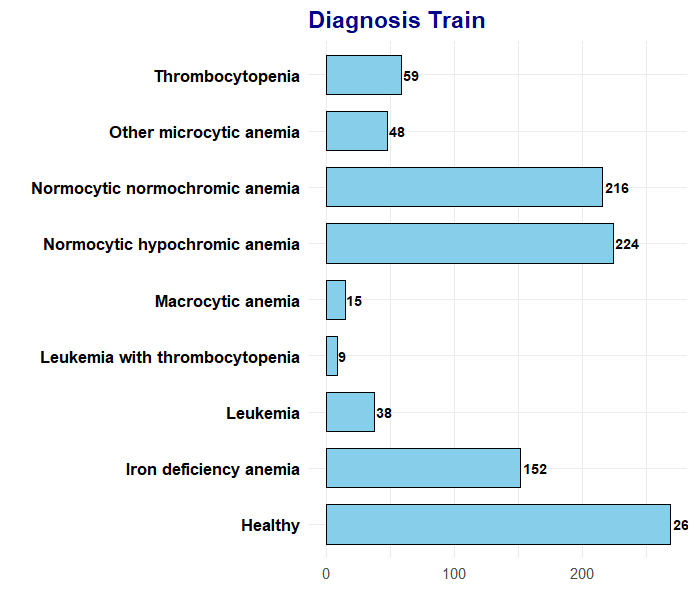
- The decision tree model provides high accuracy with simplicity, making it easier to interpret.

- The random forest model provides even higher test accuracy, demonstrating its power and robustness, although it may be slightly overfitting the training data due to its complexity.

- For practical purposes, if interpretability is crucial, the decision tree might be preferred. If accuracy is the primary goal, the random forest would be the better choice given its superior performance on the test data.

**Program and Output**

A graph with red squares

Description automatically generated with medium confidence

**Inference:**

**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

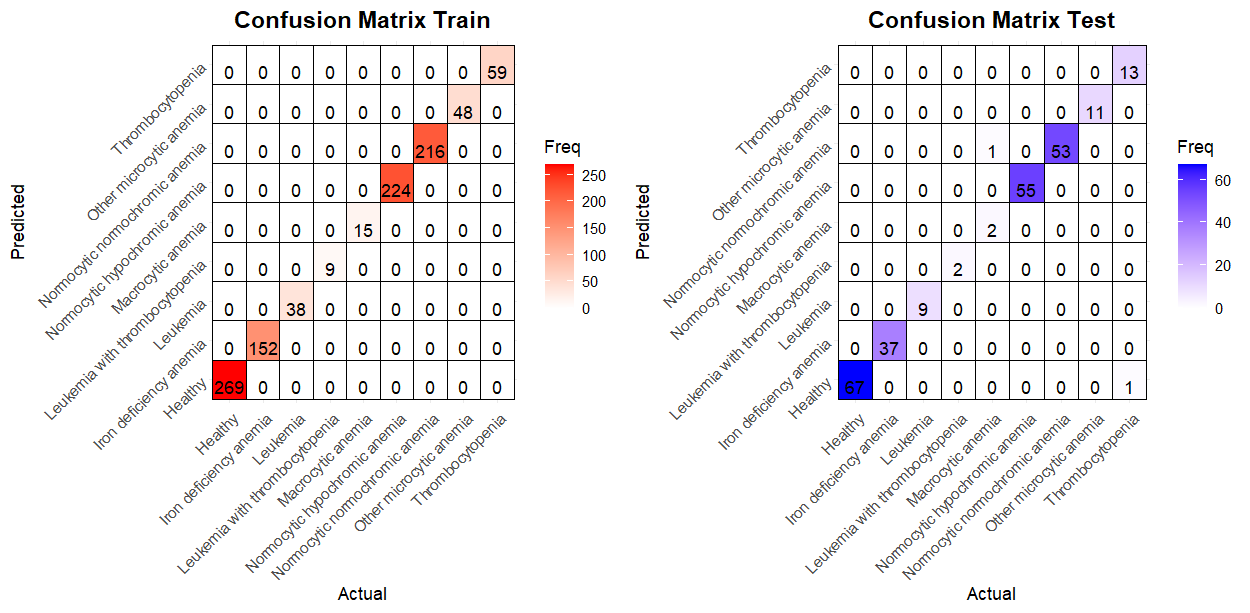
**Testing Data (Diagnosis Test):**

* **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

**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.

**Program and Output**



**References**

* Fundamentals of Mathematical Statistics - SC Gupta, VK Kapoor
* Linear Models in Statistics **–** Alvin C. Rencher
* Kaggle Dataset.<https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification>