Predictive Modeling and Statistical Analysis for Anemia Diagnosis Using Complete Blood Count Data

Akash Ahmed Student ID: 2021-3-60-242
Tasnova Haque Mazumder Student ID: 2021-3-60-235
Dilruba Akter Student ID: 2021-3-60-077
Amin Ocin Student ID: 2021-3-60-135

Submission Date: May 30, 2024

Abstract

In this study, a dataset of Complete Blood Count (CBC) values was prepossessed and analyzed using ANOVA hypothesis tests and paired t-tests to select relevant independent variables for predicting anemia. Six machine learning models were then applied, including regression and classification techniques, to detect anemia in patients. The models' performance was evaluated using metrics such as MSE, R-squared, accuracy, precision, recall, and F1 score, and SHAP values were employed to provide interpret ability and insights into the models' predictions.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Dataset Overview

The dataset uses Complete Blood Count (CBC) values to show the prevalence of various anemia forms, along with their severity and correlation with age and gender. The CBC tests that were conducted at the Eureka Diagnostic Center in Lucknow, India, between September 2020 and December 2020 produced the data. A total of 400 random samples were chosen from 1000 CBC studies, and after removing newborns, young children under the age of ten, and pregnant women, the final dataset consisted of 364 individuals. There are eleven attributes in the dataset:. The dataset's goal is to identify cases of anemia and quantify its severity using recommendations from the World Health Organization (WHO), classifying anemia as mild, moderate, or severe. Table:

Name	**Type**	**Min**	**Max**	**Mean**
Serial Number	Numeric	1.00	364.00	182.50
Age	Numeric	11.00	89.00	44.92
Sex	Numeric	0.00	1.00	0.44
RBC(Red Blood Cell Count)	Numeric	1.36	6.90	4.28
PCV(Packed Cell Volume)	Numeric	13.10	56.90	36.76
MCV(Mean Cell Volume)	Numeric	55.70	124.10	87.51
MCH(Mean Cell Hemoglobin)	Numeric	14.70	41.40	28.23
MCHC(Mean Cell Hemoglobin Concentration)	Numeric	23.60	50.20	32.05
RDW(Red cell Distribution Width)	Numeric	10.60	29.20	15.12
TLC(Total Leucocyte Count)	Numeric	2.00	42.42	8.86
PLT(Platelet Count)	Numeric	10.00	660.00	223.75
HGB(Hemoglobin)	Numeric	4.20	19.60	11.91

1.2 Domain Interest

A complete blood count (CBC) is a blood test. It's used to look at overall health and find a wide range of conditions, including anemia, infection and leukemia. A complete

blood count test measures Red blood cells, which carry oxygen, White blood cells, which fight infection, Hemoglobin, the oxygen-carrying protein in red blood cells.[1] Anemia is a condition in which the body does not have enough healthy red blood cells. Red blood cells provide oxygen to body tissues. [2] Often, the first test used to diagnose anemia is a complete blood count (CBC). The test checks hemoglobin levels of human body. A low level of hemoglobin is a sign of anemia.[3]. Recent studies have demonstrated the efficacy of machine learning (ML) algorithms in diagnosing anemia using Complete Blood Count (CBC) data. Various ML models, including DecisionTreeClassifier, RandomForestClassifier, XGBoost, LightGBM, and CatBoost, have been evaluated for their performance in classifying anemia types. Machine learning techniques such as Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine (SVM) have been employed to predict anemia, with Random Forest and Multilayer Perceptron (MLP) achieving high accuracy rates.[4] These findings underscore the potential of ML in enhancing diagnostic accuracy and efficiency in clinical settings, suggesting that ML models can significantly aid in the early detection and management of anemia.

2 Hypothesis Setting

H1:

We want to investigate whether there is any relationship between the gender of the patient and their hemoglobin (hgb) levels. Additionally, we will consider the effect of gender on other status values.

H2:

We want to know the effects of RBC (Red Blood Cell count) and PCV (Packed Cell Volume) on HGB (Hemoglobin level), while considering the interaction between RBC and PCV. Also considering the effect of these independent variables on status.

H3:

We are interested to determine the impact of MCV (Mean Corpuscular Volume) and MCH (Mean Corpuscular Hemoglobin) on HGB (Hemoglobin level),while considering the interaction between MCV and MCH. Also we are considering the impact of these independent variables on status.

H4:

We want to determine how HGB levels vary across different age groups and also considering how status vary across different age.

H5:

We are interested to know the effects of RDW (Red Cell Distribution Width) and MCHC (Mean Corpuscular Hemoglobin Concentration) on HGB (Hemoglobin level), while considering the interaction between MCV and MCH. Also considering the effect of these independent variables on status.

3 Statistical Testing

3.1 Statistical Test

T test: For hypothesis one we applied paired t test to compare the means of two groups male and female based on the 'Sex' column . The t-tests are performed for two dependent variables 'HGB' (Hemoglobin level) and 'Status'.

For HGB,

- **Null Hypothesis:** There is no significant difference in the mean hemoglobin levels (HGB) between males (Sex = 1.0) and females (Sex = 0.0).

Interpretation of Results

- As the p-value for the HGB t-test is less than the chosen significance level (0.05), we reject the null hypothesis, indicating a significant difference in mean HGB levels between males and females.

For 'Status',

There is no significant difference in the mean status values between males (Sex = 1.0) and females (Sex = 0.0)

- **Null hypothesis:** There is no significant difference in the mean status values between males (Sex = 1.0) and females (Sex = 0.0).

- Mathematically: $H_0 : \mu_{\text{Status Male}} = \mu_{\text{Status Female}}$

Interpretation of Results

- As the p-value for the Status t-test is less than the chosen significance level (0.05), we reject the null hypothesis, indicating a significant difference in mean 'Status' values between males and females.

3.2 ANOVA

For hypothesis two(H2), we applied two way Anova to know the effect of two independent variables RBC(Red Blood Cell count) and PVC(Packed Cell Volume). In this case, the dependent variables are HGB (Hemoglobin level) and Status.

*** For HGB,**

Effect of RBC

- **Null Hypothesis:** There is no significant difference in the mean hemoglobin levels (HGB) across different levels of RBC.
  Interpretation of Results - As the p-value for RBC is less than the significance level ,we reject the null hypothesis, indicating a significant effect of RBC on HGB levels.

Effect of PCV

- **Null Hypothesis:** There is no significant difference in the mean hemoglobin levels (HGB) across different levels of PCV. Interpretation of Results.
  Interpretation of Results - As the p-value for PCV is less than the significance level, we reject the null hypothesis, indicating a significant effect of PCV on HGB levels.

Combine Effect

- **Null Hypothesis:** There is no significant interaction effect between RBC and PCV on HGB values.
  Interpretation of Results - As the p-value is greater than the significance level, we do not reject the null hypothesis, indicating no significant interaction effect on status values.

Source	**sum_sq**	**df**	**F**	$PR(>F)$
A	0.112855	1.0	14.257302	1.885881e-04
B	1.940667	1.0	245.169255	7.481446e-42
A:B	0.020817	1.0	2.629825	1.058140e-01
Residual	2.651732	335.0	NaN	NaN

Table 1: ANOVA Table(RBC(A), PVC(B), HGB(C))

*** For Status,**

Effect of RBC

- **Null Hypothesis:** There is no significant difference in the mean status values across different levels of RBC.
  Interpretation of Results - As the p-value for RBC is less than the significance level, we reject the null hypothesis, indicating a significant effect of RBC on status values.

Effect of PCV

- **Null Hypothesis:** There is no significant difference in the mean hemoglobin levels (HGB) across different levels of PCV.
  Interpretation of Results - As the p-value for PCV is less than the significance level, we reject the null hypothesis, indicating a significant effect of PCV on status value.

Combine Effect

- **Null Hypothesis:** There is no significant interaction effect between RBC and PCV on status values.
  Interpretation of Results - As the p-value is less than the significance level, we reject the null hypothesis, indicating significant interaction effect.

For hypothesis three(H3), to determine the impact of two independent variables, MCV and MCH, on a dependent variable, in this case HGB (hemoglobin level) and Status, are examined using the two-way ANOVA. The interaction between MCV and MCH is also consider.

*** For HGB,**

Effect of MCV

- **Null Hypothesis:** There is no significant difference in the mean hemoglobin levels (HGB) across different levels of MCV.
  Interpretation of Results - As the p-value for MCV is greater than the significance level, we do not reject the null hypothesis, indicating no significant effect of PCV on HGB level.

Effect of MCH

- **Null Hypothesis:** There is no significant difference in the mean hemoglobin levels (HGB) across different levels of MCH.
  Interpretation of Results - As the p-value for MCH is less than the significance level, we reject the null hypothesis, indicating a significant effect of MCH on HGB level.

Combine Effect

- **Null Hypothesis:** There is no significant interaction effect between MCV and MCH on HGB level.
  Interpretation of Results - As the p-value is greater than the significance level, we do not reject the null hypothesis, indicating no significant interaction effect on HGB level.

Source	**sum_sq**	**df**	**F**	$PR(>F)$
A	0.017652	1.0	0.692280	0.405983
B	0.357277	1.0	14.011809	0.000214
A:B	0.443904	1.0	17.409185	0.000038
Residual	8.541920	335.0	NaN	NaN

Table 2: ANOVA Table (MCV(A), MCH(B), HGB(C))

*** For status,**

Effect of MCV

- **Null Hypothesis:** There is no significant difference in the mean status value across different levels of MCV.
  Interpretation of Results - As the p-value for MCV is greater than the significance level, we do not reject the null hypothesis, indicating no significant effect of MCV on status values.

Effect of MCH

- **Null Hypothesis:** There is no significant difference in the mean status values across different levels of MCH.
  Interpretation of Results - As the p-value for MCH is less than the significance level, we reject the null hypothesis, indicating a significant effect of MCH on status values.

Combine Effect

- **Null Hypothesis:** There is no significant interaction effect between MCV and MCH on status values.
  Interpretation of Results - As the p-value is less than the significance level, we reject the null hypothesis, indicating significant interaction effect on status values.

For hypothesis four(H4),To find out how one independent variable (AgeGroup) affects a dependent variable, we applied one-way ANOVA. Status and HGB (hemoglobin level) are the dependent variables in this instance.

*** For HGB,**

- **Null Hypothesis:** There is no significant difference in the mean hemoglobin levels (HGB) across different age groups.
  Interpretation of Results - As the p-value is greater than the significance level, we do not reject the null hypothesis, indicating no significant effect of AgeGroup on HGB levels.

Source	**sum_sq**	**df**	**F**	**PR(¿F)**
C	0.242567	4.0	2.046215	0.08762
Residual	9.898426	334.0	NaN	NaN

Table 3: ANOVA Table (HGB, AgeGroup(C))

*** For Status,**

- **Null Hypothesis:** There is no significant difference in the mean status values across different age groups.
  Interpretation of Results - As the p-value is greater than the significance level, we do not reject the null hypothesis, indicating no significant effect of AgeGroup on status value.

Source	**sum_sq**	**df**	**F**	$PR(> F)$
C	2.791699	4.0	2.417232	0.048528
Residual	96.435440	334.0	NaN	NaN

Table 4: ANOVA Table (Status, AgeGroup(C)

For hypothesis five(H5), The two-way ANOVA is used to know the effects of two independent variables, RDW and MCHC, on a dependent variable, which in this case is HGB (Hemoglobin level) and status.The interaction between RDW and MCHC is also considered.

*** For HGB,**

Effect of RDW

- **Null Hypothesis:** There is no significant difference in the mean hemoglobin levels (HGB) across different levels of RDW.
  Interpretation of Results - As the p-value for RDW is less than the significance level, we reject the null hypothesis, indicating a significant effect of RDW on HGB level.

Effect of MCHC

- **Null Hypothesis:** There is no significant difference in the mean hemoglobin levels (HGB) across different levels of MCHC.
  Interpretation of Results - As the p-value for MCHC is less than the significance level, we reject the null hypothesis, indicating a significant effect of MCHC on HGB level.

Combine Effect

- **Null Hypothesis:** There is no significant interaction effect between RDW and MCHC on HGB level.
  Interpretation of Results - As the p-value is greater than the significance level, we do not reject the null hypothesis, indicating no significant interaction effect on HGB level.

Source	**sum_sq**	**df**	**F**	$PR(>F)$
A	1.437881	1.0	59.311341	1.532406e-13
B	0.056101	1.0	2.314103	1.291485e-01
A:B	0.095087	1.0	3.922257	4.846860e-02
Residual	8.121381	335.0	NaN	NaN

Table 5: ANOVA Table(RDW(A), MCHC(B), HGB(C))

*** For status,**

Effect of RDW

- **Null Hypothesis:** There is no significant difference in the mean status values across different levels of RDW.
  Interpretation of Results - As the p-value for RDW is less than the significance level, we reject the null hypothesis, indicating a significant effect of RD on status values.

Effect of MCHC

- **Null Hypothesis:** There is no significant difference in the mean status values across different levels of MCHC.
  Interpretation of Results - As the p-value for MCHC is less than the significance level, we reject the null hypothesis, indicating a significant effect of MCHC on status values.

Combine Effect

- **Null Hypothesis:** There is no significant interaction effect between RDW and MCHC on status values.
  Interpretation of Results - As the p-value is greater than the significance level, we do not reject the null hypothesis, indicating no significant interaction effect on status values.

3.3 Regression Analysis:

Linear regression: Linear regression is an important algorithm in machine learning to predict numerical values based on input features by assuming a linear relationship between the target value and the features. The linear regression model learns the coefficients that best fit the data. As the target variable is HGB in our data set, which is continuous, we have done a linear regression model to fit the data. We have shown the mse and r2 score values to determine how well our model fits.

4 Machine Learning Task

4.1 Data Preparation

- The process involves importing libraries like pandas, numpy, matplotlib, seaborn, and sklearn for data manipulation, visualization, and machine learning tasks.
- The dataset is loaded from a CSV file and renamed to make column names more readable.
- The dataset's shape and initial data are checked, and descriptive statistics are calculated to provide insights into the distribution and central tendency.
- Missing values are handled, and outliers are removed using Interquartile range.
- The Serial No column is dropped to simplify the dataset.
- Age groups are created, and hemoglobin levels are categorized based on gender.
- Categorical variables are encoded using LabelEncoder, and the data is scaled using MinMaxScaler to improve model performance.
- Data is visualized using histograms and a heatmap to understand the data's distribution and correlations.

These steps ensure that the data is clean, well-structured, and ready for further analysis and machine learning model application.

4.2 Model Selection

Regression

Regression analysis is done to investigate the relationship between the independent and dependent variables. An independent variable is an input, driver, or factor that impacts a dependent variable (also called an outcome or response variable). Regression analysis can be useful for predicting the outcomes and changes in dependent variables based on the relationships of dependent and independent variables[5].

- **Linear regression:** Liner regression is an important algorithm in machine learning to predict numerical values based on input features by assuming a linear relationship between the target value and the features. The linear regression model learns the coefficients that best fit the data. As the target variable is HGB in our data set, which is continuous, we have done a linear regression model to fit the data. We have shown the mse and r2 score values to determine how well our model fits.

- **Polynomial regression:** Polynomial regression is a type of regression analyzed in the nth degree polynomial modeling of the relationship between independent and dependent variables. Polynomial regression is a special case of MLR in which the polynomial equation of data blends in with the curvilinear interplay of the dependent and independent variables [6]. We have created and fitted our data in polynomial regression model. In our polynomial regression model, the value of our polynomial degree is 2. The value of mse and r2 score shows how well the model fits..

- **Xgboost regression:** Xgboost is a gradient boosting-based scalable ensemble technique which is a reliable and efficient machine learning challenge solver. We have used the Xgboost regression model to define the relationship between the independent and dependent variables.

Classifier

- **Random Forest:** Random Forest is a widely-used machine learning algorithm developed by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. For our 'Status' classification we use random forest classifier. Reason behind choosing this algorithm is it can handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning.

- **Support Vector Machine:** Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. SVM is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks. SVMs can be used for a variety of tasks,

such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. We used SVC (Support Vector Classifier) for our classification task. We selected this algorithm for it's robustness while dealing with the data.

- **Xgboost Classifier:** XGBoost is a machine learning algorithm that belongs to the ensemble learning category, specifically the gradient boosting framework. It utilizes decision trees as base learners and employs regularization techniques to enhance model generalization. Known for its computational efficiency, feature importance analysis, and handling of missing values, XGBoost is widely used for tasks such as regression, classification, and ranking. We chose this XGBClassifier because It supports a variety of data types and objectives and It incorporates regularization techniques to avoid overfitting and improve generalization performance.

4.3 Training and Testing

To fit a model collected data has to be prepared and split into two parts, one is training and other is testing. Then training and testing data are again split into two parts, features and target value.. Features should be normalized or standardized to ensure consistent scaling before testing. Secondly, we should choose an appropriate machine learning algorithm according to the problem. Such as, linear regression, decision tree, neural networks and so more. 20% of the data is used for testing in our project and the remaining 80% of the data is used to train the regression and classificassion model (test size = 0.2). For the regression model, we have taken SEX, RBC, PCV, MCH, MCHC, RDW columns as feature and HGB column as target value. In the regression model, we applied linear regression, polynomial regression and XGB regression. The feature or target values are same for the classification model but in classification model, we have trained the model based on the status. We included the Random Forest Classifier, XGB Classifier, and SVM in the classification model. It can be understood whether the model train is good or bad by looking at R-squared value, MSE, RMSE value, accuracy.

4.4 Explainable AI Techniques

The shape value assigns a numerical value to each feature, indicating how much the feature contributes to the model's predictions for a particular instance. This is visualized using various plots, such as summary plots, force plots, and dependency plots, which provide a more intuitive understanding of feature contributions. In our project, we use SHAP in polynomial regression and random forest.

In the chart, the x-axis is labelled "mean(SHAP value)", each bar's length represents the average impact of the corresponding variable on the model's output magnitude. The chart shows that the larger bars indicate a greater influence on the model's predictions. For example:

- A longer RBC bar suggests that the Red Blood Cell Count has a significant impact.
- A shorter Sex bar implies that gender has less influence.

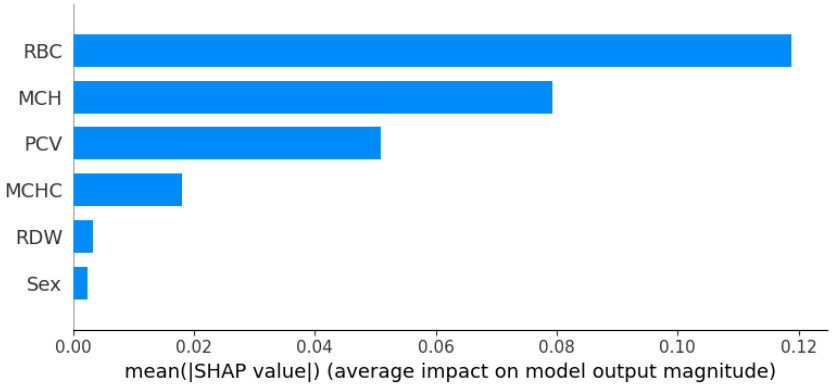

Figure 1: Summary bar plot

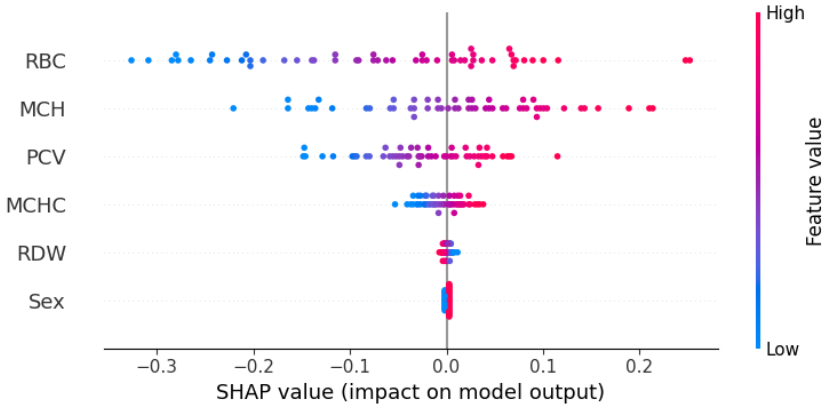

Figure 2: Summary bee-swarm plot

The image shows several horizontal bars, each representing a feature (like RBC, MCH, PCV, etc.). The length of each bar shows how important that feature is for the model's prediction. Each bar has dots (coloured from blue to red) spread across it. These dots represent individual data points. Blue dots mean low impact, and red dots mean high impact. Here, for RDW it is seen that if feature value is high then shap value is negative then we can say that there is negative relation for RDW.

4.5 Insights and Model Interpretation

Model	MSE	R2 Score
Linear Regression	0.002074788681998784	0.9244170890683823
Polynomial Regression	0.0010759968332690346	0.9608022862678582
XGB Regression	0.047588345729096766	0.9175005677458983

Table 6: Regression Summery

Model	Accuracy	Precision	Recall	F1 Score
XBGClassifier	0.9265	0.9350	0.9265	0.9282
SVC	0.8971	0.8526	0.8971	0.8736
RANDOM FOREST	0.9412	0.9480	0.9412	0.9363

Table 7: Classification Summery

5 Conclusion

In this study, we employed a comprehensive approach combining statistical analysis techniques, machine learning models, and explainable AI methods to accurately detect anemia using Complete Blood Count (CBC) values. Our rigorous hypothesis testing identified the most significant independent variables impacting hemoglobin levels and anemia status. We then applied diverse machine learning techniques, which demonstrated strong predictive capabilities in classifying anemia, as validated by various performance metrics. Notably, we leveraged SHAP values to provide valuable insights into the models' decision-making process, enhancing the interpretability of our findings. The results underscore the effectiveness of this multifaceted methodology in accurately identifying anemia and shed light on the key factors influencing hemoglobin levels. This approach holds promise for assisting healthcare professionals in diagnosing anemia more efficiently and accurately.

References

[1] "Complete blood count (CBC) - Mayo Clinic", Mayo Clinic, https://www.mayoclinic.org/tests-procedures/complete-blood-count/about/pac-20384919, Accessed 30 May 2024.

[2] "Anemia - Symptoms and Causes", Penn Medicine, https://www.pennmedicine.org/for-patients-and-visitors/patient-information/conditions-treated-a-to-z/anemia#:~:text=Anemia%20is%20a%20condition%20in,to%20folate%20(folic%20acid)%20deficiency, Accessed 30 May 2024.

[3] "How is Anemia Diagnosed?", Hematology-Oncology Associates of CNY, https://www.hoacny.com/patient-resources/blood-disorders/anemia/how-anemia-diagnosed, Accessed 30 May 2024.

[4] "Anemia", National Center for Biotechnology Information, U.S. National Library of Medicine, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11032364/, Accessed 30 May 2024.

[5] "Anaemia", Evidence-Based Nursing, https://ebn.bmj.com/content/24/4/116, Accessed 30 May 2024.

[6] Soni, R.K., Sharma, D. Linear Regression Comprehensive in Machine Learning: A Review. *Global Transitions Proceedings*, Volume 2, Issue 1, 2021, Pages 81-88, ISSN 2666-285X, https://www.researchgate.net/publication/348111996_A_Review_on_Linear_Regression_Comprehensive_in_Machine_Learning.