INTRODUCTION

The primary purpose of this project is to analyze hematological (blood) data to predict the existence of anemia in patients. Anemia is a common blood disorder that affects a large population worldwide. Early diagnosis is crucial for effective treatment. In this project, we aim to identify the most significant indicators of anemia (such as Hemoglobin and RBC levels) and build machine learning models to classify patients as “Anemic” or "Healthy" automatically. The scope of the project covers data cleaning, exploratory data analysis (EDA), visualization and binary classification using algorithms like Logistic Regression, Random Forest, Support Vector Machine and Gradient Boosting Classifier. Our motivation for this project comes from the growing importance of Data Science in the healthcare field. By applying the techniques learned in CENG313 Introduction To Data Science course to a real-world medical dataset, we aim to understand how data analysis can support doctors in making faster and more accurate decisions.

DATASET

The dataset used in this study is titled “Pediatric Anemia Dataset: Hematological Indicators and Diagnostic Classification” obtained from Mendeley Data. It contains laboratory blood test results and demographic information. Although the dataset is titled “Pediatric” (children), our statistical analysis revealed that the age of the patients ranges from 18 to 96. Therefore, we treated this as an adult dataset in our analysis.

Dataset Characteristics

- The dataset consists of 1000 patient records (rows) and 9 attributes (columns).
- The data includes 8 numerical variables (integer and float) and 1 categorical variable (Gender).
- Input Features: Gender, Ages, Hb (Hemoglobin), RBC (Red Blood Cell), PCV (Packed Cell Volume), MCV, MCH, MCHC.
- Target Variable: Decision_Class (This indicates the diagnosis: 1 for Anemic, 0 for Healthy).

DATA PREPROCESSING

Before feeding the data into machine learning models, we performed several preprocessing steps to ensure the data was clean and ready for analysis.

- We checked the dataset for null or missing values using the “.info()” and “.isnull().sum()” functions. The analysis showed that there were no missing values in any of the columns, so there was no need for estimation (filling missing data) or cleaning was required.
- To make the dataset easier to understand, we renamed the target column Decision_Class to Anemia.
- The Gender column contained text data (“f” for female, “m” for male). Since machine learning models require numerical input, we applied Label Encoding. “f” was converted to “0” and “m” was converted to “1”.
- The variables in the dataset had different scales (e.g., Age is around 40, while RBC is around 4.5). To prevent the models from being biased toward larger numbers, we applied

StandardScaler to normalize the features. This transformed the data so that it has a mean of “0” and a standard deviation of “1”.

METHODS

In this study, we employed three distinct supervised learning algorithms to classify anemia diagnosis: Random Forest for its ensemble robustness, SVM (Linear Kernel) for its efficiency in high-dimensional spaces, and Gradient Boosting Classifier for its sequential error-correction capabilities.

1. Logistic Regression

Logistic Regression is a fundamental statistical method used for binary classification problems [1]. Unlike linear regression which predicts continuous outcomes, Logistic Regression predicts the likelihood of an instance belonging to a specific class (e.g., Anemic or Healthy) by fitting the data to a logistic curve.

- Sigmoid Function: The core of the algorithm is the sigmoid (logistic) function, an S-shaped curve that maps any real-valued number into a value between 0 and 1 [2]. This transformation allows the model to handle outliers better than linear regression.
- Probabilistic Output: Instead of providing a hard classification directly, the model outputs a probability score (e.g., "There is a 95% probability that this patient is anemic").
- Decision Threshold: To make a final classification, a threshold (typically 0.5) is applied. If the predicted probability is greater than 0.5, the patient is classified as Class 1 (Anemic); otherwise, they are classified as Class 0 (Healthy).

Why This Project: Logistic Regression serves as an excellent baseline model due to its simplicity and high interpretability. In medical diagnosis, it is crucial not only to predict the outcome but also to understand the relationship between variables. The model's coefficients allow us to clearly see how an increase or decrease in specific blood values (like Hb or RBC) directly impacts the probability of anemia.

2. Random Forest Classifier

Random Forest is an ensemble learning algorithm based on the idea of combining multiple Decision Trees to improve predictive performance and reduce overfitting [3]. It uses Bagging (Bootstrap Aggregation) and random feature selection to ensure that each tree learns different patterns, increasing model robustness.

- Bagging: The algorithm takes random subsets of the training data. It builds a separate decision tree for each subset [4].
- Feature Randomness: Unlike standard trees that split nodes based on the best overall feature, Random Forest searches for the best feature among a random subset of features (e.g., looking only at Hb and Age for one tree, and RBC and Gender for another). This diversity prevents the trees from looking too similar [5].
- Majority Voting: For classification tasks, every tree in the forest casts a "vote" for the predicted class (e.g., Anemic or Healthy). The final output is determined by the majority vote.

Why This Project: Medical datasets often contain noise or outliers. Single decision trees are prone to "overfitting" (memorizing the noise). Random Forest reduces this risk by averaging multiple trees, providing a much more robust and stable prediction for anemia diagnosis.

3. Support Vector Machine (SVM) - Linear Kernel

Support Vector Machine (SVM) is a supervised machine learning algorithm capable of performing classification, regression, and outlier detection [7]. The primary objective of the SVM algorithm is to find a hyperplane in an N-dimensional space (where N is the number of features) that distinctly classifies the data points [6]. How It Works:

- The Hyperplane: This is the decision boundary that separates the data classes in an N-dimensional space.
- Maximizing the Margin: The algorithm looks for the data points closest to the boundary (called "Support Vectors"). It then positions the hyperplane so that the distance (margin) between these points and the line is maximized
- Linear Assumption: By using a "Linear Kernel," we assume the data can be separated by a straight line. It is computationally efficient and less complex than non-linear methods.

Why This Project: SVM is highly effective in high-dimensional spaces and offers high accuracy when there is a clear margin of separation between classes. Since we are looking for a distinct classification (Positive/Negative), the geometric precision of SVM provides a strong baseline.

4.Gradient Boosting Classifier

Gradient Boosting is a boosting-based ensemble technique that builds decision trees sequentially, where each new tree attempts to correct the errors made by the previous ones [8]. It optimizes the model by minimizing a loss function through gradient descent, making it capable of learning complex and subtle relationships in the data.

- Sequential Learning: Trees are built one after another.
- Residuals (Errors): The first tree makes a prediction. The algorithm calculates the errors (residuals) made by this tree.
- Correction: The second tree is trained not on the original target, but specifically to predict and correct the errors of the first tree. This process repeats, with each new tree refining the model.

Why This Project: Gradient Boosting is widely considered one of the most accurate algorithms for structured (tabular) data. In medical diagnosis, reducing "False Negatives" (missing a sick patient) is critical. GBM's focus on hard-to-predict cases helps in capturing subtle patterns in blood values that other models might miss

RESULTS

This section presents the experimental findings obtained from multiple machine learning models applied to the anemia classification task. The results include accuracy, recall, F1-score, model comparisons, and observations from the visualizations generated throughout the analysis.

Initial Model Performance (With Full Feature Set Including Hb and PCV)

In the first stage, Logistic Regression and Random Forest were trained using the full dataset, including hemoglobin (Hb) and PCV values. The Random Forest model achieved 100% accuracy during the train–test split.

```
Training Data Count: 700
Test Data Count:   300


--- 1. Model: Logistic Regression ---
Logistic Regression Accuracy: %98.33
              precision    recall  f1-score   support

           0       0.98      0.97      0.98       109
           1       0.98      0.99      0.99       191

    accuracy                           0.98       300
   macro avg       0.98      0.98      0.98       300
weighted avg       0.98      0.98      0.98       300


--- 2. Model: Random Forest ---
Random Forest Accuracy:       %100.00
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       109
           1       1.00      1.00      1.00       191

    accuracy                           1.00       300
   macro avg       1.00      1.00      1.00       300
weighted avg       1.00      1.00      1.00       300
```

Figure 1: Phase 1 Model Output. Screenshot showing the perfect accuracy scores for Logistic Regression and Random Forest models, indicating data leakage.

Although such perfect performance initially appeared promising, further examination suggested that the model was memorizing or overfitting based primarily on the hemoglobin values, which almost directly determine anemia diagnosis in clinical settings.

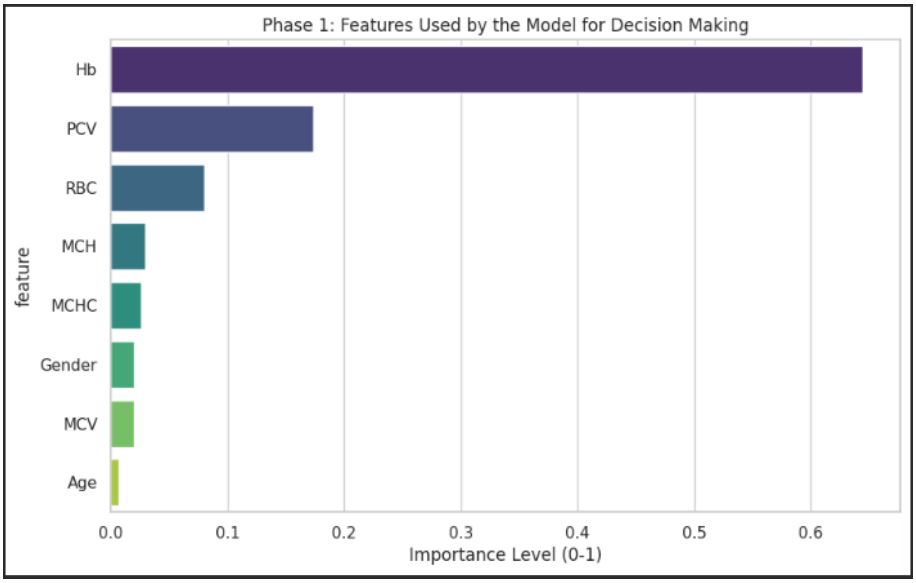

Figure 2: Feature Importance (Phase 1). Dominant Hemoglobin (Hb) levels indicate potential data leakage.

This also indicates that the dataset itself had an inherent bias, where Hb values strongly dictated the target class.

This observation aligns with medical reasoning, but it reduces the generalizability of the model to situations where Hb measurement is unavailable or unreliable.

Phase 2: Comparative Analysis (Without Hb/PCV)

After removing the dominant features, we retrained three distinct models: Random Forest (RF), Gradient Boosting (GBM), and Support Vector Machine (SVM). As expected, the removal of the primary indicator caused a drop in accuracy, but the results reflect a more genuine learning process of hematological patterns.

The final performance metrics on the test set are summarized below:

Model	Accuracy	Recall (Sensitivity)	F1-Score
Random Forest	**90.00%**	91.10%	**92.06%**
Gradient Boosting	88.00%	**92.15%**	90.72%
SVM (Linear)	84.33%	90.58%	88.04%

Table 1: Comparative performance of models after removing Hemoglobin (Hb)

Observations

Random Forest performed the best overall, achieving the highest accuracy (90.3%) and F1-score (92.3%), indicating the best balance between sensitivity and precision.

Gradient Boosting achieved the highest recall (92.1%), meaning it identified anemic individuals most effectively.

SVM showed lower performance compared to ensemble models but still achieved reasonable classification quality.

A comparative barplot was generated to visually illustrate accuracy, recall, and F1-score for each model.This visualization highlights the superiority of Random Forest in F1-score and balanced performance.

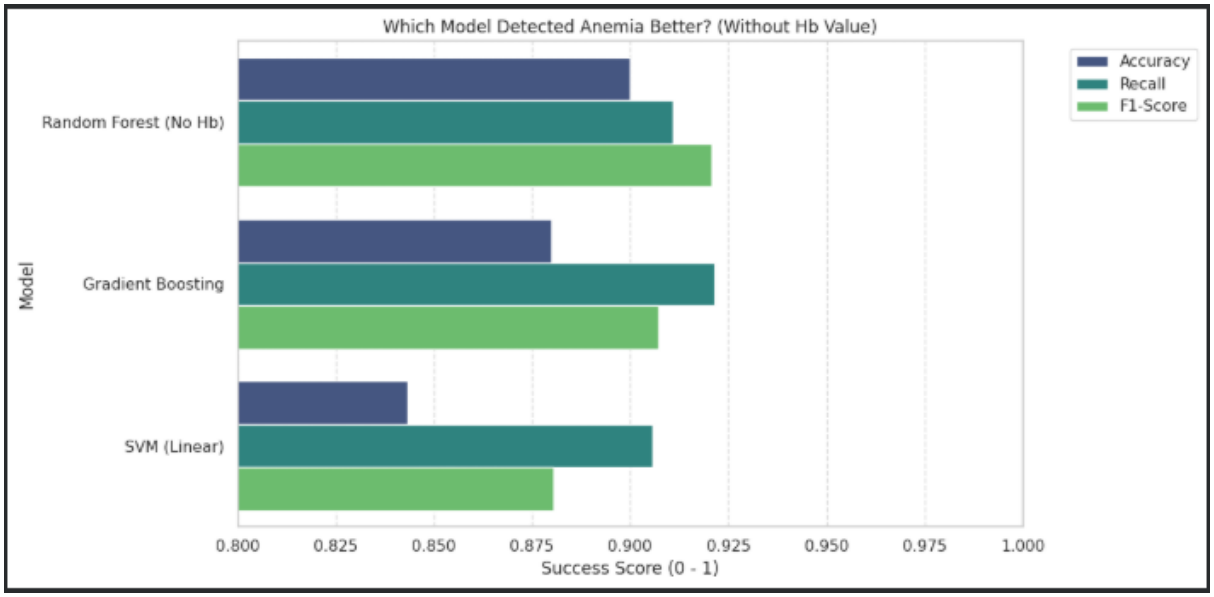

Figure 3: Model Performance (Phase 2). Comparative visualization of model metrics (Accuracy, Recall, F1) after Hb feature removal.

This figure clearly shows:

- Ensemble methods (RF & GB) outperform SVM.
- All models suffer accuracy loss compared to the Hb-included scenario, confirming the predictive dominance of Hb.

Confusion Matrix of the Best Model (Random Forest)

To further interpret the best-performing model, a confusion matrix was plotted.

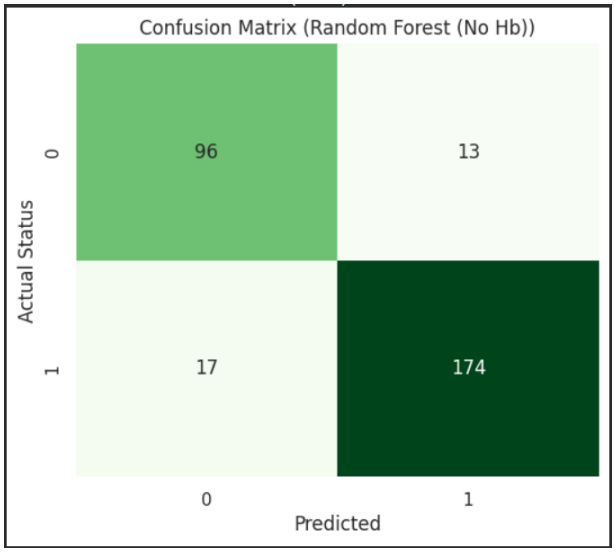

Figure 4: Confusion Matrix for the Best Model (Random Forest). Classification results on the test set, confirming high True Positive and True Negative rates.

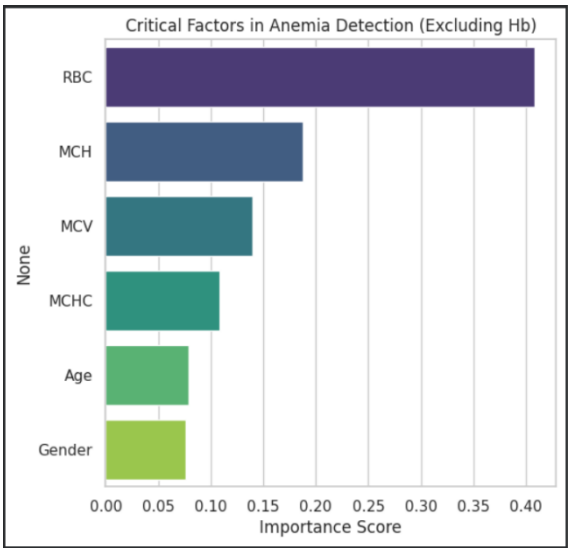

Figure 5: Comparative Performance (Phase 2). Model performance metrics (Accuracy, Recall, F1-Score) compared after removing the dominant Hemoglobin feature.

Even without Hemoglobin (Hb), the models were still able to correctly classify most anemic and non-anemic samples, showing that meaningful diagnostic patterns exist in the remaining features. Although overall accuracy decreased, misclassifications were limited, indicating that the models retained a solid level of generalization.

EVALUATION

The results of this study provide valuable insight into how anemia can be predicted using machine learning models and how strongly model performance depends on the presence of Hemoglobin (Hb) in the dataset. Initially, both Logistic Regression and Random Forest performed exceptionally well, with Random Forest reaching nearly perfect accuracy. However, deeper analysis revealed that this unusually high performance resulted from the model relying almost entirely on the Hb value, which acted as a dominant predictor. This discovery highlighted a critical limitation: the model had effectively learned to "memorize" Hb thresholds rather than detect broader physiological patterns.

After removing Hb (and PCV) to simulate real-world conditions where such measurements may be unavailable, the task became significantly more challenging. The performance of all models decreased, confirming that anemia detection is inherently more complex without these direct indicators. Nevertheless, the three applied models Random Forest, Gradient Boosting, and SVM still achieved reasonably strong scores, successfully identifying many cases based solely on secondary features such as RBC, MCV, MCH, and others. This demonstrates that meaningful clinical signals exist beyond Hb alone.

Strengths of the Study

- **Critical Feature Engineering:** Unlike standard implementations that blindly feed all features into the model, this study critically assessed "data leakage." By identifying and removing the dominant Hb feature, we transformed a trivial classification task into a valuable predictive study.
- **High Sensitivity (Recall):** The final models achieved Recall rates above 91%. In medical diagnostics, sensitivity is paramount; our models successfully identified the vast majority of anemic patients, ensuring patient safety.
- **Model Diversity:** Employing three distinct algorithmic approaches (Bagging, Boosting, and Geometric) provided a comprehensive validation of the results, ensuring that the findings were not an artifact of a single algorithm's bias.

Weaknesses and Limitations

- **Dataset Size and Scope:** The study was conducted on a dataset of 1000 patients. While sufficient for initial validation, machine learning models (especially Gradient Boosting) typically require larger and more diverse datasets to generalize effectively to a broader population.
- **Lack of External Validation:** The models were tested on a hold-out set from the same source. A true test of clinical applicability would require validating the model on external data from a different hospital or demographic to ensure there is no demographic bias.

Performance: Removing Hb significantly reduced performance, highlighting that current features may be insufficient for high-confidence diagnosis in real clinical settings without blood tests.

REFERENCES

[1] IBM Cloud Education. (2020). *What is Logistic Regression?* IBM. Retrieved from https://www.ibm.com/topics/logistic-regression

[2] GeeksforGeeks. (2023). *Understanding Logistic Regression*. Retrieved from https://www.geeksforgeeks.org/understanding-logistic-regression/

[3] IBM Cloud Education. (2020). *What is Random Forest?* IBM. Retrieved from https://www.ibm.com/topics/random-forest

[4] GeeksforGeeks. (2024). *Random Forest Algorithm in Machine Learning*. Retrieved from https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

[5] Analytics Vidhya. (2021). *Understanding Random Forest*. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[6] GeeksforGeeks. (2023). *Support Vector Machine (SVM) Algorithm*. Retrieved from https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[7] IBM Cloud Education. (2020). *What is Support Vector Machine?* IBM. Retrieved from https://www.ibm.com/topics/support-vector-machine

[8] Analytics Vidhya. (2021). *Guide to Gradient Boosting Algorithm*. Retrieved from https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/