UNIVERSITAS
GADJAH MADA

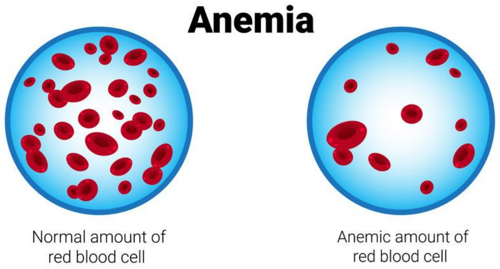
Thalassemia and Anemia Blood Test Classification

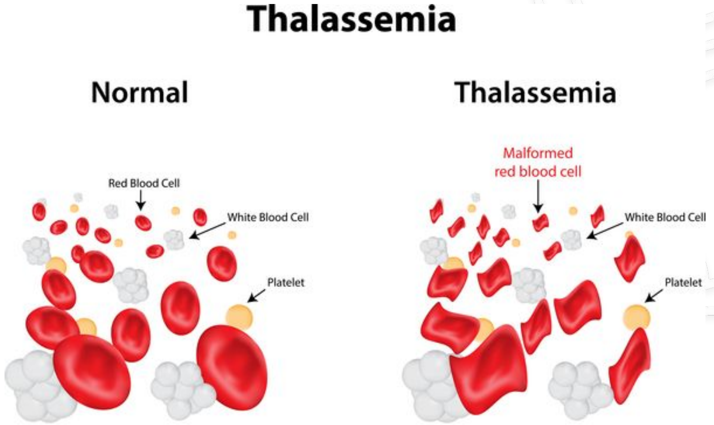
Project Members:

Reynard Pradhitya	(21/472680/PA/20321)
Muhammad Alfi Ramadhan	(21/472839/PA/20345)
Reza Aurelio Brilliansah	(21/475039/PA/20515)
Muhammad Zaky Firdaus	(21/477171/PA/20637)

Introduction

Anemia and thalassemia are blood disorders characterized by reduced oxygen-carrying capacity due to low red blood cell levels or abnormal hemoglobin production. Accurate differentiation is crucial for effective treatment and minimal complications. Traditional methods are time-consuming, subjective, and prone to errors. Machine learning algorithms can automate and enhance diagnosis, providing efficient and reliable decision support tools for clinicians.


Anemia
Normal amount of red blood cell
Anemic amount of red blood cell



Dataset

Patients' blood test data from Dr. Sardjito General Hospital, Yogyakarta comprising:

- 129 training data
- 65 testing data
- 1 = Thalassemia
- 2 = Anemia
- Proportion of Anemia and Thalassemia classes = 1 : 1.04
- 19 attributes & 1 label column (diagnosis)

Class	Training Data	Testing Data
1	43	22
2	70	34

Methods: Data Preprocessing

Data Cleaning

- Convert values with comma decimal separators to use periods
- Drop rows with a diagnosis value of 3
- Drop identifier column 'No'

Dataset Splitting

The original training data was split into training and validation sets using an 80:20 split. This resulted in three sets of data: training set, validation set, and testing set

Handling Outliers

- The standard deviation for each column is calculated.
- A threshold, twice the standard deviation, is determined for each column
- Values exceeding this threshold are replaced with the column's average

Handling Missing Values

Missing values are replaced by their respective column's average value

Methods: The Model

Algorithms used

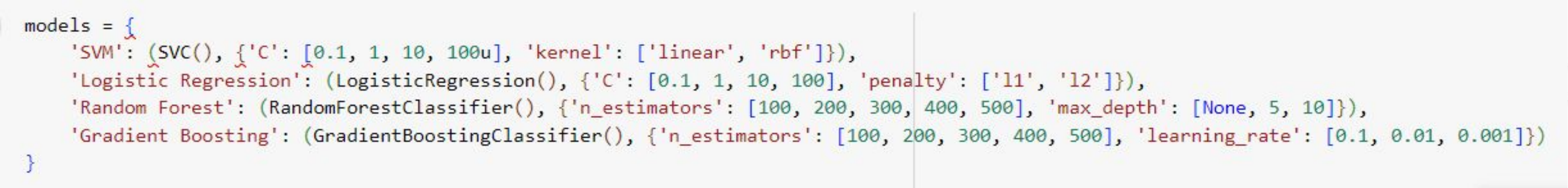
- SVM (Support Vector Machines)
- Logistic Regression
- Random Forest
- Gradient Boosting

GridsearchCV

- Utilized to compare and evaluate the performance of different models
- Enables testing multiple algorithms simultaneously
- Allows for hyperparameter tuning to achieve optimal results

Through this process,

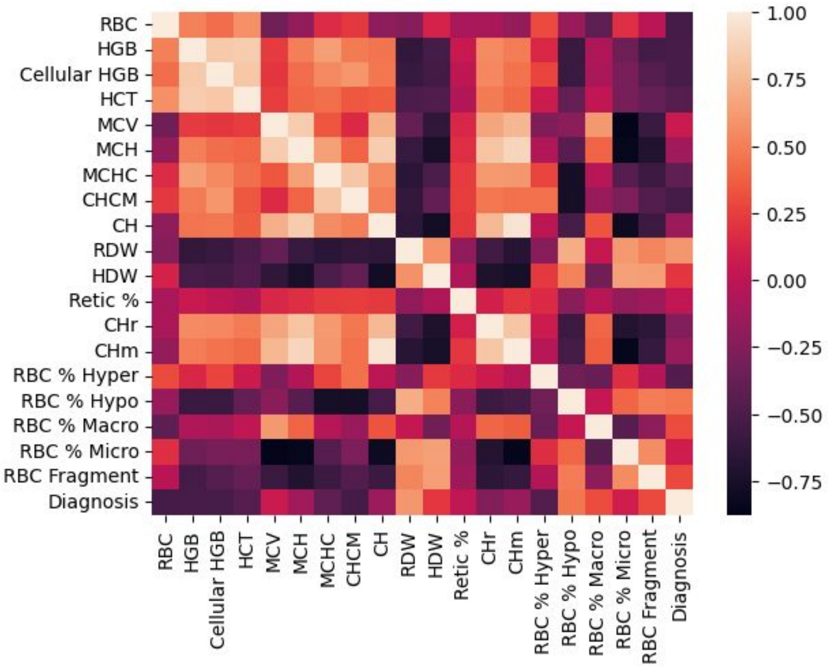
- The best and most accurate model can be identified
- Optimal hyperparameters can be determined
- Ensures the selection of the most appropriate model for testing



```
models = {
    'SVM': (SVC(), {'C': [0.1, 1, 10, 100u], 'kernel': ['linear', 'rbf']}),
    'Logistic Regression': (LogisticRegression(), {'C': [0.1, 1, 10, 100], 'penalty': ['l1', 'l2']}),
    'Random Forest': (RandomForestClassifier(), {'n_estimators': [100, 200, 300, 400, 500], 'max_depth': [None, 5, 10]}),
    'Gradient Boosting': (GradientBoostingClassifier(), {'n_estimators': [100, 200, 300, 400, 500], 'learning_rate': [0.1, 0.01, 0.001]})
}
```

Exploratory Data Analysis

Heatmap of variables



Model Comparison

Gridsearch Results Summary

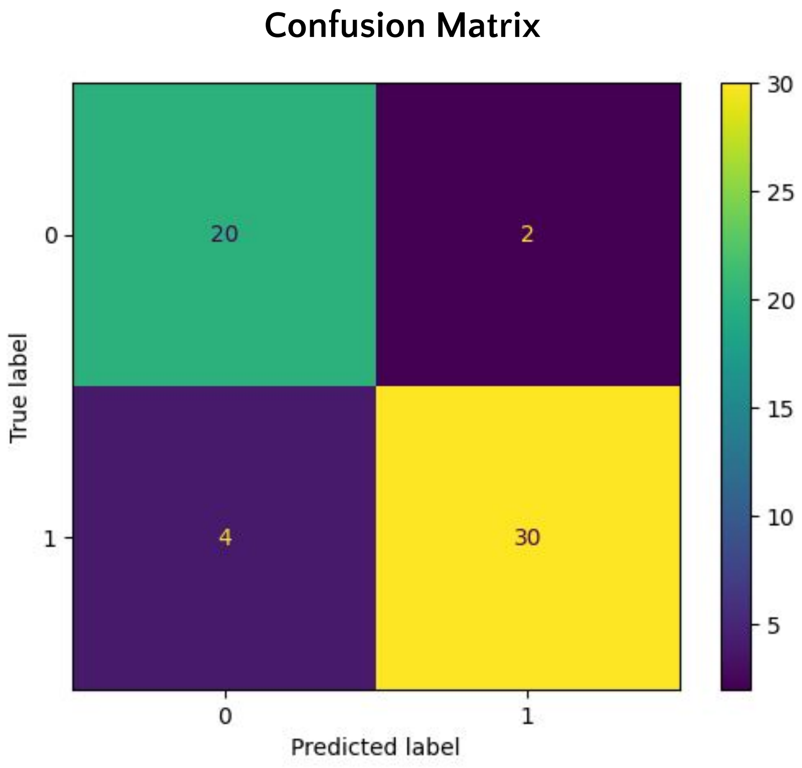
Model	Training Score	Validation Score	Best Params
SVM	0.86	0.95	{'C': 100, 'kernel': 'rbf'}
Logistic Regression	0.85	0.86	{'C': 1, 'penalty': 'l2'}
Random Forest	0.87	0.82	{'max_depth': None, 'n_estimators': 200}
Gradient Boosting	0.86	0.73	{'learning_rate': 0.001, 'n_estimators': 400}

Model Testing - SVM Accuracy

SVM Accuracy

Class	Precision	Recall	F1-Score
1	0.83	0.91	0.87
2	0.94	0.88	0.91
Macro Avg	0.89	0.90	0.89
Weighted Avg	0.90	0.89	0.90
Accuracy			0.89

Model Testing - Confusion Matrix



Accurate Classifications

- 20 instances correctly classified as Thalassemia
- 30 instances correctly classified as Anemia

Misclassifications

- 2 instances misclassified as Anemia
- 4 instances misclassified as Thalassemia

Model Accuracy: 0.89

Analysis & Conclusion

- Best performing model: SVM
- Reasons for SVM's high performance:
  - Works well with a high number of features (19 in this case)
  - Better generalization ability even with limited training data
- Although the training scores of all four models were similar, SVM achieved a significantly higher validation score compared to other models
- Final model accuracy: 89%
- For medical purposes, there is still room for improvement in achieving higher accuracy
  - Further fine-tuning and improvement of the model to achieve an accuracy closer to 100%
  - Implementing cross-validation to address limited data
  - Exploring more parameters in grid search for potential better models