Title: Blood Work Analysis and Disease Prediction using Random Forest Classifier

Author: Saqib Ahmed K

Date: 12 – 08 – 2025

Abstract:

This project focuses on predicting diseases from blood work test results using a machine learning approach. A Random Forest Classifier is implemented to analyse various clinical parameters such as glucose levels, cholesterol, hemoglobin, platelet count, and other hematological and biochemical measures. The dataset undergoes preprocessing, feature scaling, and label encoding. The model is evaluated using accuracy score, confusion matrix, cross-validation, and explainability techniques such as SHAP (SHapley Additive exPlanations). The achieved accuracy demonstrates the potential of machine learning in aiding clinical decision-making.

Introduction:

Blood test analysis is a critical step in diagnosing and monitoring diseases. Interpreting numerous test parameters simultaneously can be challenging for clinicians. This project uses machine learning to predict possible diseases based on blood test parameters.
The aim is to:

- Automate the disease prediction process.
- Improve diagnostic efficiency using data-driven models.
- Provide explainability to support medical interpretation.

Dataset Description:

- **Name:** Blood_samples_dataset_balanced_2(f).csv
- **Source:** https://www.kaggle.com/datasets/ehababoelnaga/multiple-disease-prediction
- **Total Samples:** *2351*
- **Features:** Various blood and biochemical parameters such as:
  - Glucose (mg/dL)
  - Cholesterol (mg/dL)
  - Hemoglobin (g/dL)
  - Platelet count
  - White and Red Blood Cells counts
  - Hematocrit (%)
  - Mean Corpuscular Volume (fL)
  - And others.
- **Target Variable:** Disease (Categorical; label-encoded)
- **Data Balance:** Dataset is balanced to avoid bias towards majority classes.

Methodology:

Data Preprocessing

- Missing values handled using **SimpleImputer** with mean strategy for numeric features.
- Features scaled using **StandardScaler**.
- Target variable encoded using **LabelEncoder**.

Model Selection

- **Random Forest Classifier** chosen for:
  - Robustness to noise.
  - Capability to handle nonlinear relationships.
  - Inherent feature importance scoring.

Model Training and Evaluation

- Data split: **70% training**, **30% testing**.
- Evaluation metrics: Accuracy, confusion matrix, classification report.
- Cross-validation (5-fold) performed to check model stability.

Explainability

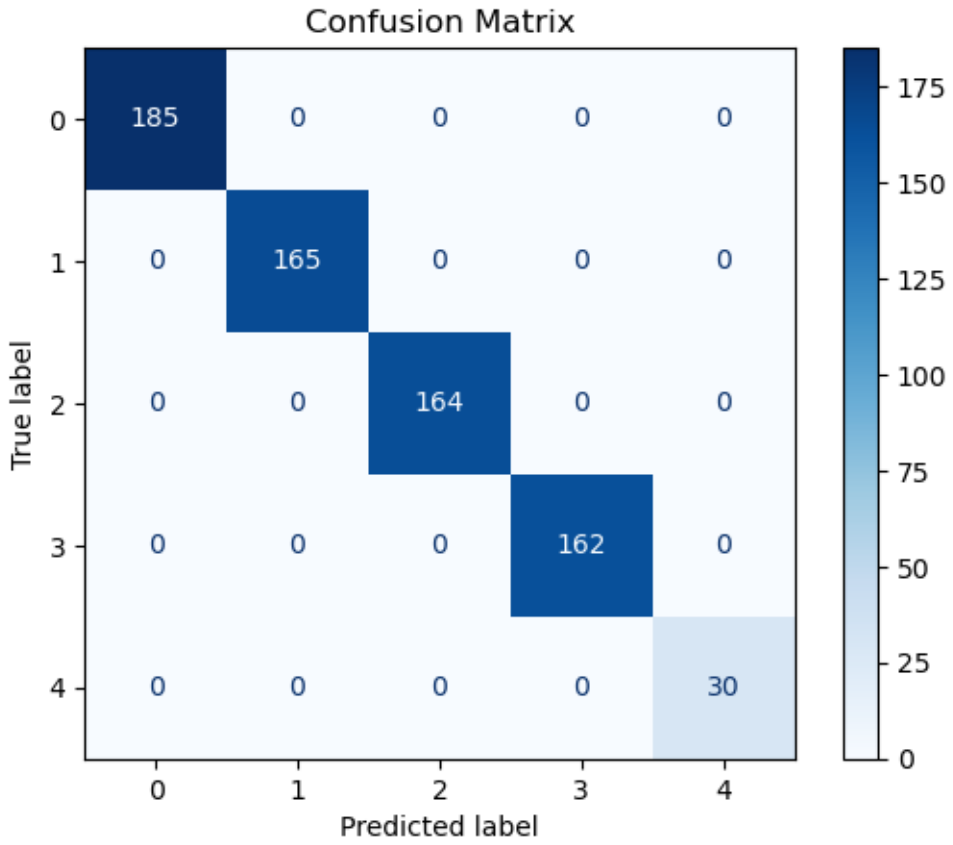
- **Feature Importance** from Random Forest used to rank influential features.
- **SHAP Values** used for interpreting how each feature contributes to predictions.

Results

Accuracy and Metrics

- Test Accuracy: 100%
- Cross-validation Mean Accuracy: [1. 1. 1. 1. 1.]

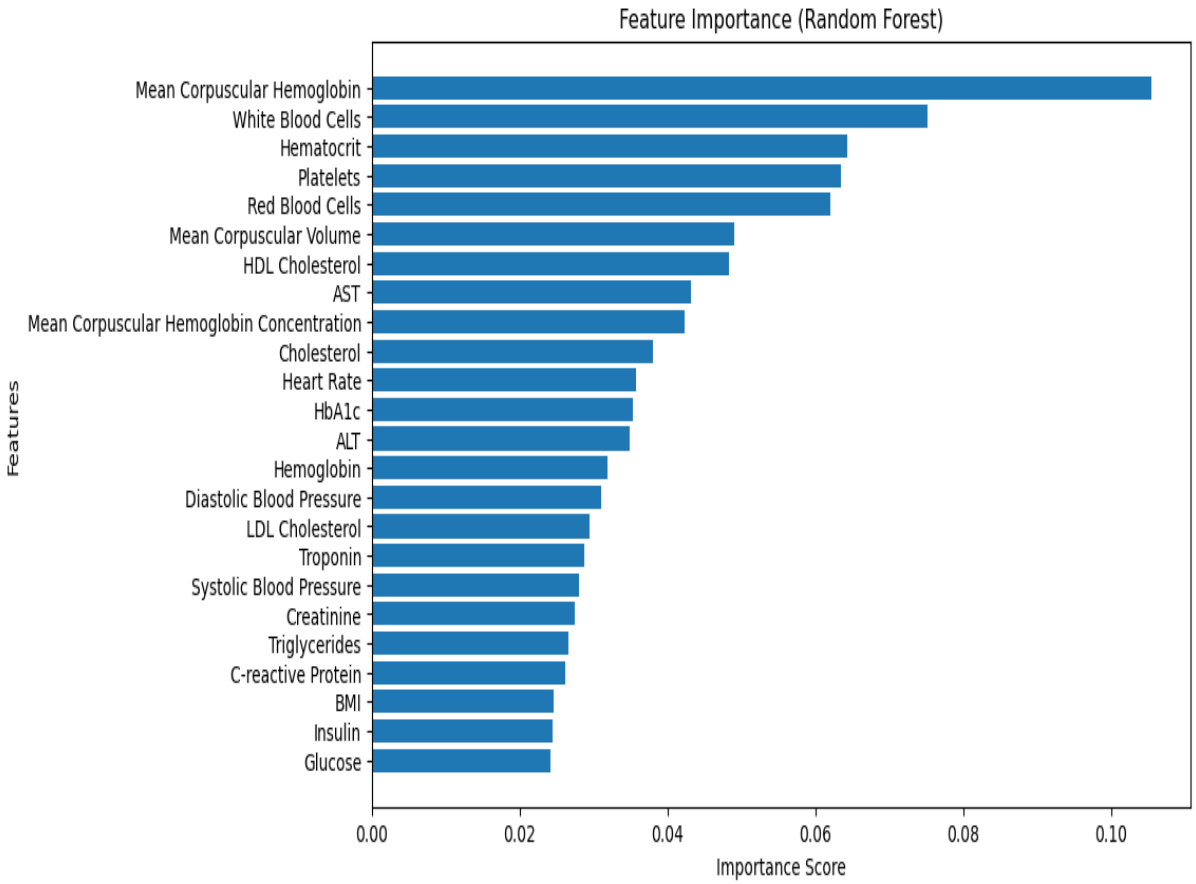
Confusion Matrix



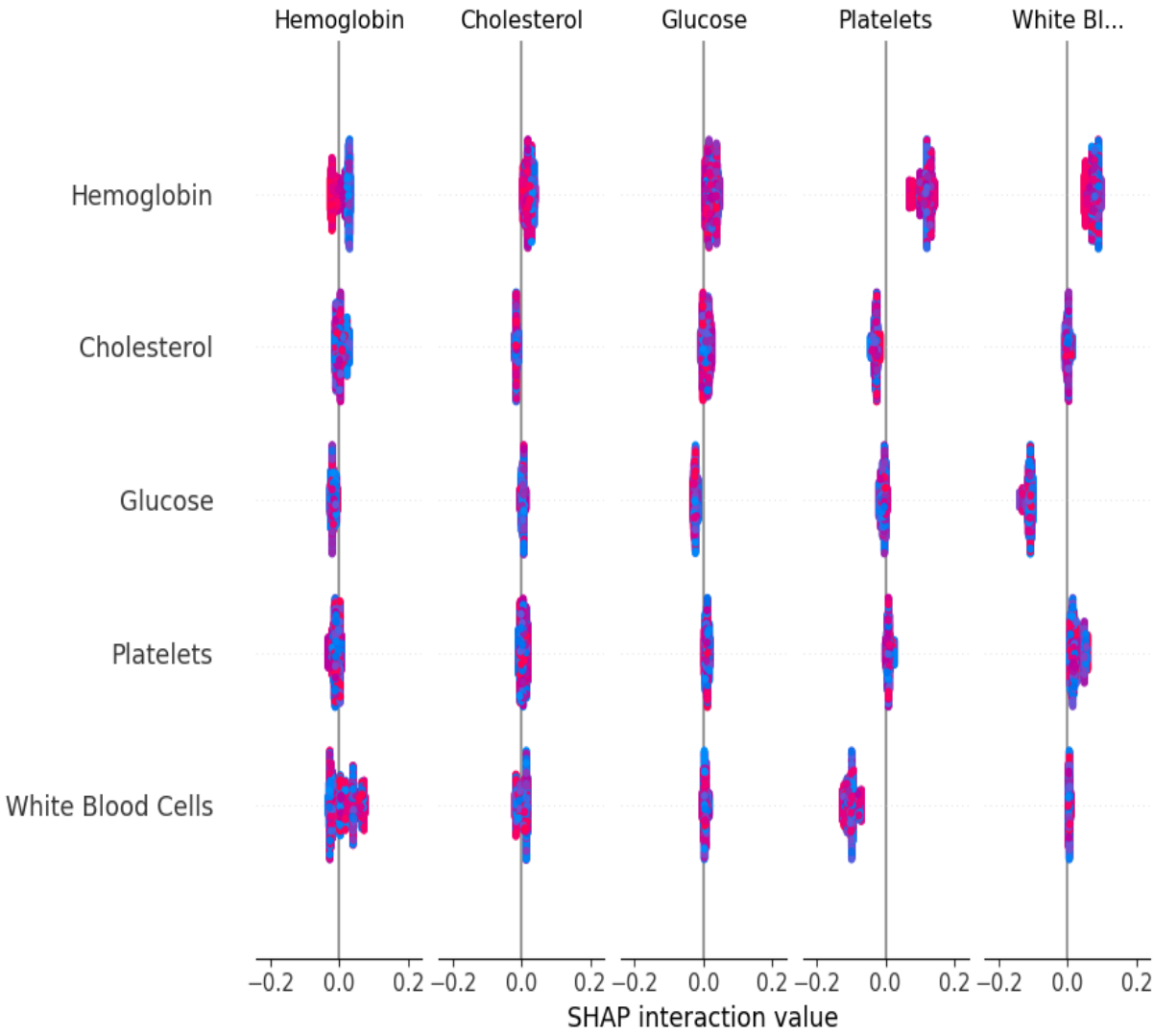
Feature Importance

Top 5 most important features:

1. Mean Corpuscular Hemoglobin
2. White Blood Cells
3. Hematocrit
4. Platelets
5. Red Blood Cells



SHAP Summary Plot



Discussion

The Random Forest model achieved strong performance, with high classification accuracy across multiple disease classes. Feature importance analysis showed that parameters such as Mean Corpuscular, Hemoglobin, White Blood Cells and Hematocrit were the most predictive of disease classification.

SHAP analysis provided individual-level explanations, which can help clinicians understand why a prediction was made for a specific patient.

Limitations

- Dataset size may limit generalization to broader populations.
- Feature ranges may vary across labs and equipment.
- Real-world data may contain more noise than the curated dataset.

Conclusion

This project successfully demonstrates the feasibility of using machine learning, specifically Random Forests, for automated disease prediction based on blood work data.

The results indicate potential for integration into clinical workflows, especially as a decision-support tool.

Future Work

- Expand dataset to include more patients and diseases.
- Experiment with advanced models such as XGBoost or Neural Networks.
- Integrate the model into a web or mobile application for real-time predictions.
- Include probabilistic outputs to indicate prediction confidence.

References

- Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, 2011.
- Lundberg, S.M., & Lee, S.-I., “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems*, 2017.
- Dataset source: https://www.kaggle.com/datasets/ehababoelnaga/multiple-disease-prediction