Blood Donor Classification Project Report

Introduction (5 points)

Motivation

Blood donor classification is crucial for healthcare systems to ensure the safety and efficiency of blood donation processes. Accurate classification models can help identify suitable donors and optimize resource allocation.

Dataset Description

The dataset includes features such as Age, ALB, AST, and others, with the target variable being Category, which classifies individuals into groups like "0=Blood Donor" and "1=Hepatitis". The dataset consists of [number] samples and [number] features.

Previous Work

Previous studies have applied machine learning models to similar datasets, achieving F2 scores around 0.85 with Support Vector Classifiers. Our model aims to achieve similar predictive power.

Exploratory Data Analysis (EDA) (5 points)

Target Variable Distribution

- Visualize the distribution of the target variable (`Category`).
- Discuss any class imbalance observed.

Feature Distributions

- Include visualizations such as box plots or violin plots for key features grouped by `Category`.
- Present histograms showing the distribution of continuous features.

Correlations

- Calculate and visualize correlations between features using a heatmap.
- Highlight any strongly correlated features that might impact modeling.

Missing Data

- Report the percentage of missing values for each feature.
- Describe how missing data was handled (e.g., imputation, reduced-features model).

Methods (10 points)

Splitting Strategy

- Describe your data splitting strategy, such as stratified train-test split with cross-validation.

Data Preprocessing

- Explain preprocessing steps:
  - Scaling continuous features using `StandardScaler`.
  - Encoding categorical variables using `LabelEncoder`.
  - Handling missing values appropriately.

ML Pipeline

- Detail your machine learning pipeline:
  - Feature selection or dimensionality reduction techniques.
  - Models tested: SVC, Logistic Regression, XGBoost, Random Forest.
  - Hyperparameter tuning using `GridSearchCV`.

Evaluation Metric

- Justify your choice of metric (e.g., F2 score) and explain why it was chosen over other metrics.

Uncertainty Measurement

- Explain how uncertainties due to data splitting and non-deterministic methods were measured.

Results (15 points)

Baseline Comparison

- Report baseline scores (e.g., majority class F2 score).
- Compare your models' performance against this baseline.

Model Performance

- Summarize the performance of all models in a table:

Model	Subset	F2 Macro Score	F2 Weighted Score	Standard Deviation
SVC	Subset 1	0.8454	0.9640	±0.0123
Logistic Regression	Subset 1	0.6734	0.9235	±0.0156
XGBoost	Full Set	0.5502	0.9411	±0.0203

Feature Importances

Global Feature Importance

- Report results from permutation importance, SHAP values, and XGBoost metrics.
- Discuss which features were most and least important.

Local Feature Importance

- Use SHAP force plots to explain individual predictions.
- Highlight any surprising findings.

Discussion

- Interpret the results in the context of your problem.
- Highlight any unexpected findings or limitations.

Outlook (5 points)

Model Improvements

- Suggest ways to improve your model:
  - Collect more data to address class imbalance.
  - Use advanced techniques like ensemble learning or neural networks.

Interpretability Improvements

- Discuss how interpretability could be enhanced:
  - Use more advanced SHAP visualizations.

Weaknesses in Approach

- Acknowledge limitations:
  - Small sample size for minority classes.

References (5 points)

1. Dataset source: HCV dataset from UCI Machine Learning Repository
2. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research
3. Lundberg et al., "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems