A Machine Learning Approach of Liver Cirrhosis Prediction

Introduction

This report documents the process of building a machine learning model to predict liver cirrhosis stages using a dataset of patient characteristics. The objective is to develop a model with high accuracy for early detection and management of liver cirrhosis.

Methodology

The methodology involved the following steps:

1. **Data Loading and Preprocessing:**
    - The dataset named "Liver Cirrhosis prediction_Cleaned.csv" was loaded into a Pandas Data Frame.
    - The dataset's structure was explored using shape, head, dtypes, and value_counts to understand the data types and distribution of the target variable ("Stage").
2. **Data Encoding:**
    - Numerical and categorical features were identified.
    - Categorical features were encoded using Label Encoder to transform them into numerical representations suitable for machine learning models.
    - Correlation analysis was performed using a heatmap to visualize the relationships between features and the target variable. Features with low correlation ("Copper", "Spiders", "SGOT", etc.) were removed.
3. **Model Building and Evaluation:**
    - The dataset was split into training and testing sets using train_test_split with stratification to maintain the class proportions in both sets.
    - Three classification models were chosen: Random Forest, XGBoost, and Gradient Boosting.
    - These models were trained within an ImbPipeline that included StandardScaler for feature scaling and SMOTE for handling class imbalance.
    - GridSearchCV was used to find the best hyperparameters for each model, optimizing for weighted F1-score.

- The performance of each model was evaluated using accuracy, weighted F1-score, confusion matrix, and classification report.

4. **Feature Engineering:**
    - Interaction term between "Bilirubin" and "Albumin" was created: Bilirubin_Albumin_Interaction.
    - Log transformation applied to "Prothrombin": Prothrombin_log.
    - These engineered features were added to the dataset to potentially improve model performance.
5. **Ensemble Model with Voting Classifier:**
    - An ensemble model was created using Voting Classifier with the best individual models (Random Forest, XGBoost, and Gradient Boosting) to combine their predictions.
    - The ensemble model's performance was evaluated similarly to the individual models.
6. **Model Saving:**
    - Each individual model (Random Forest, XGBoost, Gradient Boosting) was saved to a separate file using joblib.dump() for future use.

Results

- The performance of each model was assessed using appropriate metrics, revealing the most effective model.
- The confusion matrices provided insights into model performance across different classes.
- The ensemble model demonstrated the potential for further performance improvement by combining the individual models' strengths.

Conclusion

This notebook successfully demonstrated a comprehensive workflow for building and evaluating machine learning models to predict liver cirrhosis stages. The ensemble method and extensive hyperparameter tuning led to the development of models with promising accuracy.

Recommendations

- Further exploration of feature engineering and selection could enhance model performance.
- Investigation of alternative algorithms or advanced ensemble techniques might yield further improvements.
- Deployment of the model within a user-friendly interface for real-world applications could make this system valuable in clinical settings.