SAT 5114 – First Large Project Final Report

Project Title - Predicting Sepsis Onset: A Machine Learning Approach

Group Member – B. Suma Durga

Introduction

Sepsis is a severe medical condition characterized by the body's extreme response to an infection, leading to organ dysfunction and potentially life-threatening complications. Early detection and prompt intervention are crucial for improving patient outcomes and reducing mortality rates associated with sepsis. In this report, a comprehensive analysis of a dataset containing information related to sepsis cases, aiming to develop a predictive model using statistical ML classifiers capable of accurately identifying sepsis cases based on various patient variables and clinical data.

Data Loading and Overview

- I started by loading the dataset named \`undersampled_sepsis_data_wtarget.csv.zip\`, which contains structured data pertaining to sepsis cases.
- After loading the dataset, I conducted an initial overview to understand its structure and content. This includes checking the number of rows and columns in the dataset, displaying the first few rows to inspect the data format, and listing the column names to identify the available features.
- Additionally, I performed a summary statistics analysis to gain insights into the numerical features' distributions, which helps us understand the data's characteristics and identify potential outliers.

Data Preprocessing

- The next step involves preprocessing the dataset to ensure its suitability for modeling.
- I started by checking and handled missing values using an appropriate imputation strategy, such as median imputation, to fill in any null values present in the dataset.

- Additionally, I checked for class imbalance to ensure that the distribution of sepsis and non-sepsis cases is balanced, which is essential for training unbiased predictive models and found out there is no class imbalance .
- Visualizations are created to explore the distributions of both numerical and categorical features, providing valuable insights into the data's underlying patterns and distributions.
- Furthermore, I calculated and visualize the correlation matrix between features to identify any significant correlations or relationships among them.

Splitting the dataset

To ensure unbiased model evaluation and generalizability, we split the dataset into separate training and test sets using a random split. The random split ensures that both the training and test sets contain representative samples of the overall dataset, enabling the model to learn from diverse patterns and generalize well to new data instances.

Feature Selection and Scaling

- Feature selection is a critical step in building models, as it helps identify the most relevant features that contribute to the model's predictive performance.
- In this report, I employed a Recursive Feature Elimination (RFE) with a Random Forest classifier to select the top features that are most informative for predicting sepsis cases.
- These selected features are then scaled using StandardScaler to standardize their scales and magnitudes, ensuring that they have similar ranges and distributions, which can improve the performance of certain machine learning algorithms.

Feature Importance Analysis

- To gain a deeper understanding of the factors influencing sepsis detection, I conducted a feature importance analysis using a Random Forest classifier.
- After that , I calculated and visualized the correlation matrix between the selected features to identify any significant correlations or relationships among them.

Models

1. Random Forest Model :

- **Introduction:**

The Random Forest model is a powerful ensemble learning technique that combines multiple decision trees to make predictions. It is widely used for classification and regression tasks due to its robustness and ability to handle complex datasets.

- **Parameter Tuning:**

The Random Forest model's hyperparameters were tuned using GridSearchCV to find the optimal combination that maximizes model performance.

The parameter grid defined includes:

- \`n_estimators\`: Number of trees in the forest, with values 100 and 200.
- \`max_depth\`: Maximum depth of the trees, allowing them to grow without restrictions or limiting them to a depth of 10.
- \`min_samples_split\`: Minimum number of samples required to split an internal node, with values 2 and 5.
- \`min_samples_leaf\`: Minimum number of samples required to be at a leaf node, with values 1 and 2.

- **Model Training:**
  The Random Forest classifier was initialized with a \`random_state\` parameter set to 42 for reproducibility.GridSearchCV was employed to fit the Random Forest classifier to the training data, exploring the specified parameter grid and using cross-validation with 3 folds.The best model obtained from GridSearchCV was stored in \`best_rf_model\`.

- **Best Hyperparameters:**

The best hyperparameters determined by GridSearchCV are as follows:

\`max_depth\`: None (allowing trees to grow without restriction)

\`min_samples_leaf\`: 1

\`min_samples_split\`: 2

`n_estimators`: 200

- **Model Evaluation:**

The best Random Forest model was evaluated on the test set to assess its performance.

Evaluation metrics computed include:

- **Accuracy: 90.47%**
- **Precision: 87.98%**
- **Recall: 93.49%**
- **F1-score: 90.65%**

- **Conclusion:**

The Random Forest classifier demonstrated excellent performance on the preprocessed dataset.With an accuracy of 90.47%, the model showcases its ability to correctly classify sepsis cases.The precision of 87.98% indicates the model's capability to minimize false positives, while the recall of 93.49% highlights its ability to capture true positives effectively.The F1-score of 90.65% provides a balanced measure of the model's accuracy and robustness in predicting sepsis cases.

2. Ensemble Model (Decision Tree And XGBoost)

- **Introduction:**

  The ensemble model combines the predictions of multiple base classifiers, Decision Tree (DT), and XGBoost (XGB), to improve overall performance. The VotingClassifier aggregates the predictions using a "hard" voting strategy, where the class with the majority of votes is selected.

- **Parameter Tuning:**

Hyperparameters of the ensemble model and the base classifiers were tuned using GridSearchCV to find the optimal combination.

The parameter grid includes:

- `dt__max_depth`: Maximum depth of the Decision Tree, allowing it to grow without restrictions or limiting it to a depth of 5.
- `xgb__n_estimators`: Number of boosting rounds for XGBoost, with values 50 and 100.
- `xgb__learning_rate`: Learning rate for XGBoost, with values 0.01 and 0.1.

Model Training

The ensemble model was initialized with the specified base classifiers and parameter grid. GridSearchCV was employed to fit the ensemble model to the training data, exploring the parameter grid and using cross-validation with 3 folds.The best performing model obtained from GridSearchCV was stored in `best_ensemble_model`.

- **Best Hyperparameters:**

The best hyperparameters determined by GridSearchCV are as follows:

- `dt__max_depth`: None (allowing the Decision Tree to grow without restriction)
- `xgb__n_estimators`: 100

- `xgb__learning_rate`: 0.1

Model Evaluation:

The best ensemble model was evaluated on the test set to assess its performance.

Evaluation metrics computed include:

- **Accuracy: 78.57%**
- **Precision: 86.89%**
- **Recall: 66.74%**
- **F1-score: 75.49%**

Conclusion:

The ensemble method achieved an accuracy of 78.57%, indicating the percentage of correctly classified instances out of the total instances.The precision of 86.89% suggests that when the ensemble method predicts a positive class (sepsis, in this case), With a recall of 66.74%, the ensemble method successfully identifies all actual positive instances in the dataset.
The F1-score, which is the harmonic mean of precision and recall, is 75.49%. It provides a balance between precision and recall and is useful for assessing the overall performance of the model.

Insights:

- Among two models , the Random Forest model with optimized hyperparameters achieved high performance on the test set.
- It demonstrates strong predictive capability in identifying sepsis cases based on the selected features.
- The model's high accuracy, precision, recall, and F1-score indicate its effectiveness in real-world applications.
- Further optimization and fine-tuning may lead to even better results, ensuring effective sepsis detection in clinical scenarios.

Conclusion

In conclusion, this report presents a detailed analysis of building a predictive model for sepsis detection using machine learning techniques. By following a structured approach to data preprocessing, feature selection, model building, and evaluation, we can develop robust and effective predictive models for critical healthcare applications like sepsis detection. The insights gained from the analysis, including feature importance and model evaluation metrics, provide valuable information for healthcare practitioners and researchers working towards improving sepsis management and patient outcomes. Further research and development efforts can focus on refining the model and exploring additional avenues for enhancing its performance and generalizability in real-world clinical settings.

Dataset reference -
https://www.kaggle.com/datasets/durgalakshmip/undersampled-wtarget-sepsis/data

Github code link - https://github.com/Suma1236/Prediction-Of-Sepsis-using-ML.git