Public Research Module 1 Report

Authors

- **Patricia Honorato Moreira** - Estudante [Inteli]
- **Jefferson Silva** - Coorientador [Inteli]
- **Luciana Rodrigues Carvalho Barros** - Orientadora [Fundação Faculdade de Medicina]
- **Roger Chammas** - Coorientador [Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP)]

1. Research Area

This project is situated within the field of clinical predictive modeling in oncology. In this module, our work focused on integrating diverse clinical and laboratory datasets, advanced data transformation, and the implementation and evaluation of several machine learning and deep learning models. The ultimate goal is to develop tools that can eventually support clinical decision-making, while ensuring that all sensitive data and specific outcomes are kept confidential.

2. Overview of Sprints

Sprint 1: Data Integration and Research Documentation

- Establish the research framework.
- Define the research motivation, problem statement, objectives, and scope.
- Prepare a comprehensive internal research document outlining the rationale, current challenges, and expected impact of integrating clinical and laboratory data for predictive modeling.

Sprint 2: Data Transformation and Preliminary Modeling

a. Data Transformation

- Integrated multiple clinical and laboratory datasets through extensive cleaning, resolving duplicate columns, and standardizing key identifiers.
- Created a unified, anonymized dataframe (internally named df_model) that serves as the basis for subsequent experiments.

b. Preliminary Modeling

- Performed stratified splitting of the dataset into training (60%), validation (20%), and test (20%) sets.
- Developed initial machine learning models using Random Forest, XGBoost, and SVM.
- Experimented with oversampling techniques (SMOTE) for addressing class imbalance.

Sprint 3: Advanced Feature Engineering and Model Exploration

a. Feature Engineering

- Conducted exploratory data analysis on an expanded dataset.
- Reduced feature dimensionality by removing redundant and low-variance variables.
- Developed new features based on composite indices and regression-based feature selection.

b. Model Exploration

- Evaluated shallow learning models and regularized logistic regression to assess the impact of the newly engineered features.

c. Insights

- Identified challenges related to limited sample size and high feature dimensionality.
- Gathered insights that guided future refinements of the modeling approach.

Sprint 4: Data Expansion, Stakeholder Engagement, and Feature Recovery

a. Data Expansion

- Recovered additional clinical exam results and integrated these with the existing dataset.
- Conducted data quality checks, retaining only those exam parameters with reliable non-null values.

b. Stakeholder Engagement

- Presented the project progress at a scientific forum, receiving positive feedback and constructive suggestions from clinicians, biologists, and postgraduate researchers.
- Initiated future collaborations for methodological refinement and advanced feature extraction.

Sprint 5: Comprehensive Modeling and Evaluation

a. Modeling Approaches:

- **T- raditional Models (Non-Oversampled):** Implemented individual models such as Random Forest, XGBoost, and SVM.
- **Stacking Ensemble:** Developed a stacking classifier that combined Random Forest and XGBoost as base models with Logistic Regression (in a pipeline with StandardScaler) as the final estimator.
- **Keras Neural Network:** Designed and trained a deep learning model (multi-layer perceptron) for binary classification.

b. Evaluation

- Models were evaluated using stratified splits into training, validation, and test sets.
- Evaluation metrics included Accuracy, Precision, Recall, F1-Score, and ROC-AUC.
- Experiments were conducted both with and without SMOTE to address class imbalance.

3. Conclusions

The detailed evaluation metric tables originally produced for this module have been withheld in this public report. Instead, a narrative summary is provided below.

a. Non-Oversampled Models:

The individual traditional models (Random Forest, XGBoost, and SVM) showed strong performance in overall accuracy and precision but demonstrated lower recall levels, indicating potential challenges in identifying all positive cases.

b. Stacking Ensemble with SMOTE:

Incorporating SMOTE into the stacking ensemble improved recall, better capturing positive instances, yet this improvement was coupled with a reduction in precision and overall accuracy. These findings highlight a crucial trade-off that must be carefully balanced based on the intended application.

c. Keras Neural Network:

The deep learning model provided competitive performance with similar overall discrimination; however, improvements in recall are needed. Fine-tuning the network architecture and hyperparameters is anticipated to further enhance its sensitivity.

In summary, the research indicates that careful attention must be paid to the balance between false negatives and false positives. The chosen modeling strategy will depend on the operational priorities of the clinical setting, with current efforts emphasizing the importance of improving recall when the cost of missed positive cases is high.

4. Conclusions

The modeling work in this module has demonstrated several important insights:

a. Modeling Approaches:

The experiments with traditional machine learning models and deep learning techniques underscore both the strengths and limitations inherent in each approach.

b. Trade-Offs Observed:

The non-oversampled models yielded higher precision and accuracy; however, they experienced lower recall, which could result in missing some high-risk cases. Conversely, the stacking ensemble with SMOTE improved recall at the expense of a slight reduction in overall accuracy and precision.

These trade-offs are critical in clinical predictive modeling, where the cost of false negatives and false positives must be carefully balanced.

c. Future Directions:

Based on these findings, further work will focus on additional hyperparameter tuning, the exploration of alternative model architectures, and potentially integrating additional clinically relevant features to improve performance.

5. Next Steps

Moving forward, the next phase will concentrate on preparing a scientific article for publication. The article will include:

- A detailed methodology of data integration and transformation.
- An in-depth explanation of our modeling approaches, including the evaluation of strategies with and without oversampling.
- A critical discussion of the observed trade-offs between sensitivity and specificity.
- Future plans for enhanced model tuning and the incorporation of additional clinical variables.

This report represents the cumulative progress achieved across multiple sprints and serves as a foundation for the forthcoming publication.

This public report summarizes our work within the field of clinical predictive modeling in oncology, while ensuring that all sensitive data and internal details remain confidential.