Public Research Module 2 Report

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

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 report summarizes the activities and deliverables completed during Module 2 of the breast cancer recurrence risk prediction project. The project aims to develop interpretable and clinically applicable machine learning models to predict 10-year recurrence or death using clinical and laboratory data, following TRIPOD reporting standards.

2. Overview of Sprints

Sprint 1

Objective: Initiate scientific communication materials and establish manuscript structure.

Activities:

- Designed a scientific poster summarizing project objectives, methodology, and preliminary results for conference presentation.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

- Drafted the Introduction section of the manuscript, including clinical background, project motivation, and study aims.

- Defined the full manuscript structure, detailing sections and subsections to align with academic publication requirements.

- Generated initial visualizations for poster and manuscript figures, including distributions of laboratory markers and prognostic ratios.

Sprint 2

Objective: Develop the general predictive model and document technical methodology.

Activities:

- Completed the Methods and Results sections describing the complete analytical pipeline, from cohort selection to model interpretation.

- Developed a general predictive model based on pre-treatment laboratory data using Random Forest, XGBoost, SVM, and stacked ensemble methods.

- Applied class imbalance correction using SMOTE and evaluated model performance with standard metrics (AUC, precision, recall, SHAP interpretability).

- Performed preliminary survival analysis (Kaplan–Meier) stratifying patients by predicted risk categories.

Note: The development of subtype-specific models was postponed due to pending access to tumor subtype data.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

Sprint 3

Objective: Refine manuscript drafts and improve modeling reproducibility.

Activities:

- Completed the first draft of the Results section, integrating performance metrics, confidence intervals, SHAP analysis, and survival analysis.
- Organized and formatted project notebooks for reproducibility and clarity:
    - Notebook 1: Data preparation, exploratory data analysis, feature engineering.
    - Notebook 2: Model training, feature selection, hyperparameter tuning, evaluation structure.

Note: The modeling pipeline was aligned with TRIPOD recommendations to ensure transparent and reproducible reporting.

Sprint 4

Objective: Analyze model results in context and draft the Discussion section.

Activities:

- Completed the first version of the Discussion section, covering:
    - Performance and stability of the stacked ensemble model (ROC-AUC approximately 0.82–0.83).
    - Comparisons with existing prognostic tools.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

- Alignment of model interpretability results with clinical expertise, emphasizing tumor staging and inflammatory biomarkers.
- Potential for clinical integration in resource-limited healthcare environments.
- Study limitations, including single-center data and the need for external validation.

Sprint 5

Objective: Consolidate model evaluation, interpretability, and survival analyses in a single, reproducible notebook and complete the article abstract.

Activities:

- Combined Notebook 3 and Notebook 4 into one comprehensive evaluation and interpretation notebook, including:
  - Performance evaluation of all trained models (Random Forest, XGBoost, SVM, Stacking) with AUC, accuracy, precision, and recall.
  - ROC curves for individual models and combined performance comparison.
  - Global feature importance analysis using Random Forest and XGBoost.
  - SHAP interpretability analysis for feature contribution understanding.
  - Kaplan–Meier survival analysis by predicted risk groups.
- Drafted the scientific abstract for the article, summarizing methods, key results, and clinical relevance.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

Status:

- The full article manuscript is not yet finalized. Major content sections (Introduction, Methods, Results, Discussion, Abstract) have been prepared.

- Completing and refining the manuscript for submission remains a priority for the next project module.

4. Conclusions

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

Module 2 delivered significant progress toward the project objectives, including general model development, interpretability analyses, survival stratification, and manuscript drafting. With these foundations established, the next module will focus on completing the scientific article, expanding modeling to cancer subtypes, and advancing toward external validation and dissemination.

5. Next Steps

- Finalize the complete article manuscript and prepare for submission to a peer-reviewed journal.

- Develop and evaluate subtype-specific prognostic models incorporating tumor subtype data. These models were not finalized during the current module.

- Perform additional exploratory analyses, including PCA and UMAP, to characterize the data structure if required for publication.

- Continue improving reproducibility, technical documentation, and figure quality to meet scientific publication standards.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

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.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br