Universidad Andrés Bello

Facultad de Ingeniería

Professor: Pablo Schwarzenverg

April 24, 2025

Predicting Clinical Severity Using Machine Learning

A Case Study from Hospital El Pino

Authors:

Nicolás Godoy

Marcos Lazo

Miguel Muñoz

1. Introduction:

In this project, we aim to predict the **severity level** of hospitalized patients by extracting and analyzing the last digit of the DRG (Diagnosis Related Group, *GRD* in Spanish) codes from real data collected at Hospital El Pino. Rather than predicting the entire DRG code or all three severity levels, we focus on a **binary classification** task: identifying whether a case is **high severity ("Alta")** or **not high severity ("No Alta")**, derived from the last digit of the DRG code. This distinction supports clinical prioritization and resource allocation, particularly in contexts where identifying high-risk patients is critical.

2. Related Work:

Machine learning techniques have shown promising results in healthcare classification problems, including disease severity estimation. Prior research highlights the potential of predictive modeling for improving clinical outcomes and supporting decision-making in hospital settings.

For instance, Nistal-Nuño (2022) developed machine learning models for predicting mortality in intensive care units, emphasizing the clinical value of early risk stratification using structured EHR data. Similarly, Shamout et al. (2021) reviewed a wide range of machine learning approaches applied to clinical outcome prediction, demonstrating the relevance of models such as logistic regression and ensemble methods in high-stakes medical environments. Aljameel et al. (2021) proposed a model to predict COVID-19 severity using patient data, reinforcing the feasibility of using classical algorithms on structured datasets for severity assessment.

While many studies focus on predicting full DRG codes or diagnostic categories, other research has emphasized the importance of identifying severity levels to improve triage and cost estimation. Traditional models such as logistic regression and random forests have been applied to structured hospital data with good performance, particularly when working with encoded diagnosis and procedure codes. We draw on this foundation to apply classical models for the specific task of severity classification in the context of a real Chilean hospital.

3. Objective:

To develop and evaluate machine learning models capable of predicting whether a patient case is of **high severity (Alta)** or **not high severity (No Alta)** based on features extracted from structured hospital data. The goal is to simplify the original multi-class severity scale into a clinically meaningful binary classification problem, improving focus on critical cases.

4. Methodology

4.1. Dataset Description

The dataset contains hospitalization records from Hospital El Pino, each row representing a single patient stay. Relevant variables include:

- Up to 35 diagnostic codes
- Up to 30 procedure codes
- Patient's age (in years)
- Sex (as categorical)
- DRG (used to derive severity)

From the DRG code, the last digit was extracted to represent severity. Originally, the last digit could take values from 0 to 3, with the following meanings:

- 0 → no severity (typically for outpatient or minor cases)
- 1 → baja (low)
- 2 → media (medium)
- 3 → alta (high)

For this study, we excluded cases labeled as severity 0. This decision was made for the following reason:

From an operational perspective, excluding these cases allows the model to focus exclusively on decisions that impact hospital management, prioritization, and resource allocation.

After filtering, the remaining severities were grouped and binarized as follows:

- **Alta (high)** → class 1
- **No Alta (low or medium)** → class 0

Only entries with valid severities (1, 2, 3) were retained for analysis.

4.2. Development Process

Data Cleaning:

- Extracted numerical codes from DRG labels.
- Filtered out entries with invalid or missing severity values.

Feature Selection:

- Used diagnosis and procedure codes as features.
- Included patient age and sex.

Feature Transformation:

- Categorical variables were encoded using One-Hot Encoding.
- Missing values were imputed.

4.3. Chosen Machine Learning Techniques

The dataset was split into training and testing subsets using a 70/30 ratio. Two classical machine learning models were trained and evaluated: Random Forest and Logistic Regression.

These models were chosen for the following reasons:

- **Random Forest Classifier**: This ensemble method is known for its robustness to overfitting, its ability to handle high-dimensional data, and its flexibility with both categorical and numerical features. It is especially effective when dealing with complex interactions between variables, as is often the case in clinical data.
- **Logistic Regression**: Used as a baseline model, it offers high interpretability and serves as a useful benchmark for comparison. Despite its simplicity, logistic regression can yield solid results in structured healthcare datasets.

4.4. Evaluation Metrics

Given the class imbalance in our dataset—where the majority of cases fall under the "No Alta" (low or medium severity) category—**accuracy alone is not an appropriate performance metric**, as a model could achieve high accuracy by simply predicting the majority class. Therefore, our evaluation focused on metrics better suited to imbalanced classification problems:

- **Precision**: Indicates how many of the predicted high-severity cases were actually correct.
- **Recall (Sensitivity)**: Crucial in our context, it measures how many of the actual high-severity cases were correctly identified by the model.
- **F1-Score**: Provides a balance between precision and recall, and is particularly relevant for imbalanced datasets.
- **Confusion Matrix**: Offers a clear view of class-wise performance, allowing inspection of false negatives, which are especially critical in clinical settings.

These metrics allowed us to assess the models not only on overall performance but more importantly, on their ability to detect high-severity ("Alta") cases, which are the primary focus of this study.

5. Experiments

5.1. Data Analysis

5.1.1. Data Quality Study

- **Completeness:** Most patients have at least a few diagnoses and procedures, though the dataset is sparse.
- **Correctness:** Some codes were malformed or empty and were excluded.
- **Outliers:** Very rare severity labels were filtered during preprocessing.

5.1.2. Descriptive Statistics

Before Binarization:

- **Average age:** ~39 years
- **Sex distribution:**
  - Male: ~34%
  - Female: ~66%
- **Severity distribution:**
  - Baja (low): ~50%
  - Media (medium): ~24%
  - Alta (high): ~26%

After Binarization:

- **Average age:** ~50 years
- **Sex distribution:**
  - Male: ~47%
  - Female: ~53%
- **Severity distribution:**
  - No Alta (low or medium): ~74%
  - Alta (high): ~27%

5.2. Feature Selection Justification

The input features used for all models include diagnosis codes, procedure codes, patient age, and sex. These variables were selected based on their clinical relevance and alignment with the DRG classification system. The target variable is a **binary indicator of severity**: *Alta* (high) vs *No Alta* (medium or low), derived from the last digit of the DRG code. This binary simplification enables more focused modeling on critical care identification.

5.3. Model Training and Results

Model	Threshold
Random Forest	0.5
Random Forest (Adjusted Threshold)	0.35
Logistic Regression	0.5

The Random Forest classifier, using its default decision threshold, achieved high precision for the Alta class (0.87) but suffered from a low recall (0.68), indicating it failed to identify a significant number of high-severity cases. Given the clinical importance of not overlooking severe cases, we adjusted the decision threshold to favor sensitivity. This increased the recall for Alta to 0.84, but at the expense of a drop in precision to 0.73.

In contrast, the Logistic Regression model offered a more balanced performance, with both precision and recall at 0.86 for the Alta class, resulting in a superior F1-score of 0.86. These consistent and balanced results across all key metrics—especially in identifying high-severity cases—led us to select Logistic Regression as the preferred model.

This choice is further supported by the model's better macro average F1-score (0.90) and its ability to handle class imbalance more effectively, making it more suitable for a healthcare-oriented context where both false positives and false negatives carry weight.

Confusion matrices and ROC curves were generated for both models to visualize their performance. The Logistic Regression model ultimately demonstrated a more favorable trade-off between sensitivity and specificity, making it the preferred choice in a healthcare setting where missing severe cases can have serious consequences.

5.3.1. Random Forest

Achieved high precision, but suffered from a low recall.

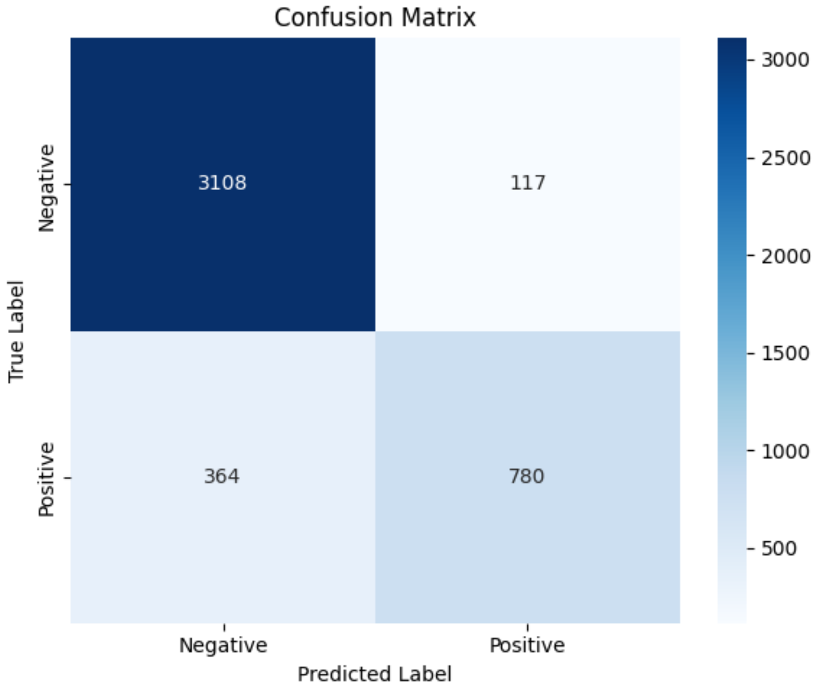

Figure 1: Confusion Matrix – Random Forest (default threshold)

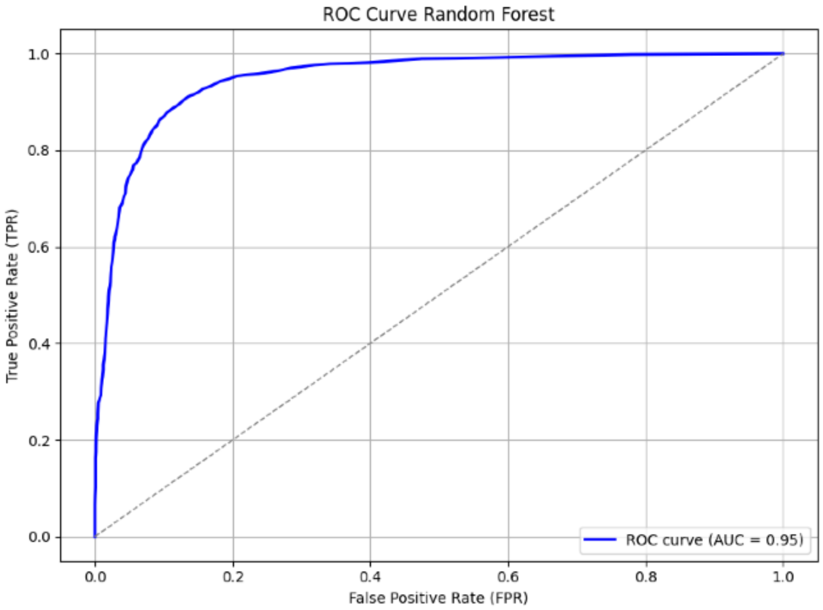

Figure 2: ROC Curve – Random Forest (default threshold)

5.3.2. Random Forest (Adjusted Threshold)

Increased the recall, but at the expense of a drop in precision.

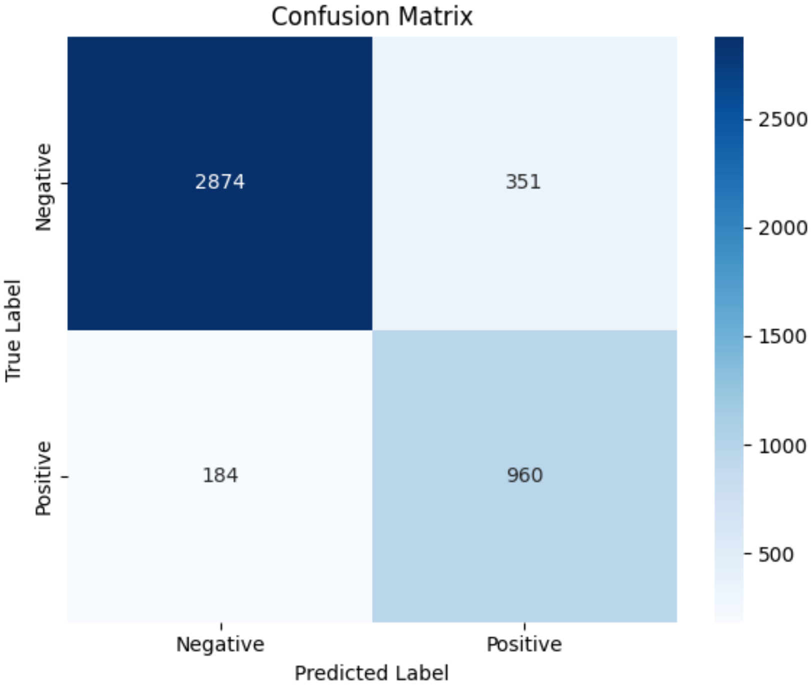

Figure 3: Confusion Matrix – Random Forest (adjusted threshold)

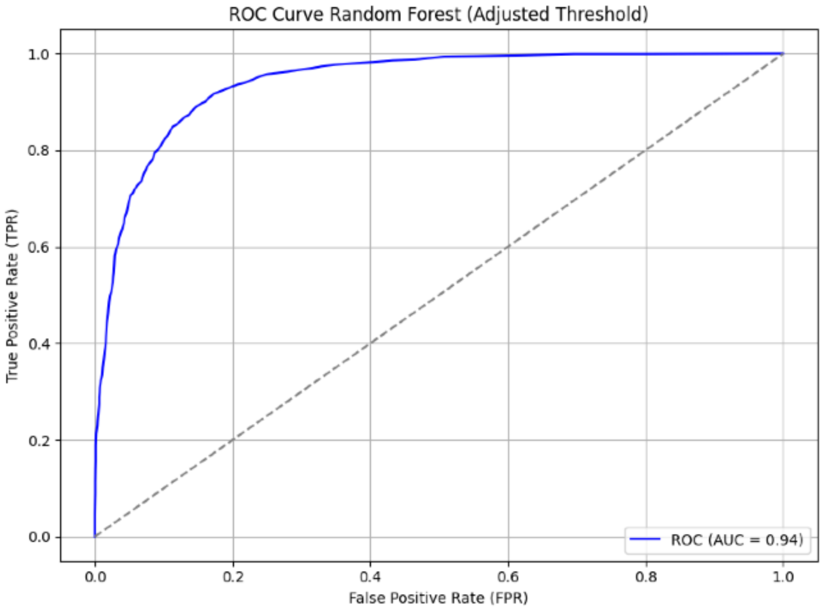

Figure 4: ROC Curve – Random Forest (adjusted threshold)

5.3.3. Logistic Regression

More balanced performance, with both high precision and recall.

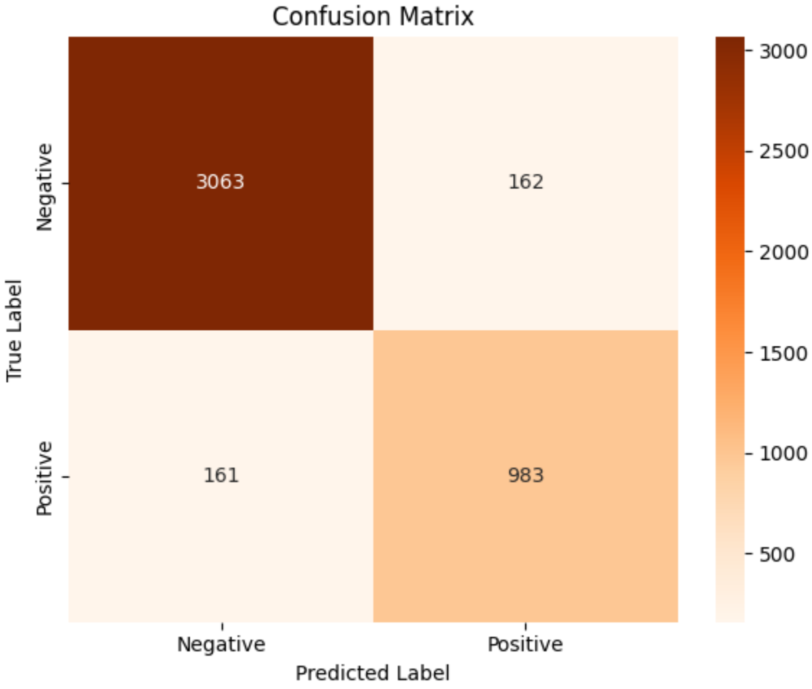

Figure 5: Confusion Matrix – Logistic Regression

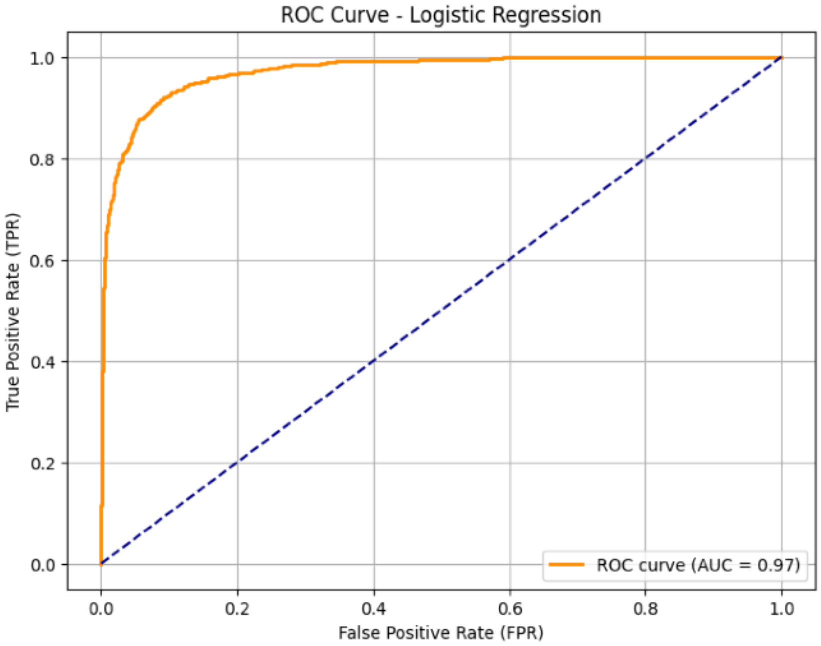

Figure 6: ROC Curve – Logistic Regression

5.4. Model Architecture

Rather than employing neural networks or deep learning techniques, this study focused on classical machine learning models using a structured and reproducible pipeline approach with **scikit-learn**. Each classifier was embedded within a **Pipeline** that included:

- **SimpleImputer**: to handle missing values in numerical and categorical features.
- **OneHotEncoder**: for transforming categorical variables into a numeric format suitable for modeling.
- **ColumnTransformer**: to apply different preprocessing steps to specific column types.
- **Classifier**: either a **RandomForestClassifier** or **LogisticRegression**, depending on the experiment.

This modular architecture facilitated consistent preprocessing and ensured comparability across models. All experiments and visualizations—such as confusion matrices and ROC curves—were performed using libraries including **pandas**, **matplotlib**, and **seaborn** for analysis and plotting.

5.5. Performance Analysis

A closer look at the performance metrics revealed distinct trade-offs between the evaluated models. The Random Forest classifier, when adjusted to improve sensitivity, became more aggressive in detecting high-severity (*Alta*) cases, as reflected in its increased recall. However, this came at the expense of precision, resulting in a higher number of false positives—an important consideration in clinical decision-making where unnecessary alarms can overwhelm healthcare workflows.

Logistic Regression, on the other hand, demonstrated **greater consistency across all performance dimensions**, offering stable behavior without the need for threshold tuning. While its recall was slightly lower than the adjusted Random Forest, the improved **harmonic balance between precision and recall** (as captured by the F1-score) made it a more practical solution in our context.

Importantly, all evaluations were conducted without applying any resampling techniques, allowing us to observe the models' natural behavior on an imbalanced dataset.

5.6. Literature Comparison

Previous research on severity prediction in healthcare often relies on advanced techniques such as deep learning or boosting algorithms, which typically offer strong performance but at the expense of interpretability and computational simplicity (Shamout et al., 2021). While these models can capture complex, non-linear relationships, their "black-box" nature limits their adoption in clinical environments where explainability is crucial.

In contrast, our study demonstrates that classical models like Logistic Regression and Random Forest can achieve competitive results while maintaining transparency—an essential factor in clinical decision-making, especially in contexts such as intensive care unit risk assessment (Nistal-Nuño, 2022) or severity prediction for COVID-19 (Aljameel et al., 2021).

Specifically, Logistic Regression proved effective at balancing precision and recall without the need for threshold adjustments or complex hyperparameter tuning. Random Forest showed strong baseline performance, and with threshold optimization, it became highly sensitive to high-severity cases.

Our approach, rooted in explainable models and reproducible pipelines, contributes to the growing body of evidence that machine learning solutions in clinical contexts must prioritize interpretability—particularly when they are integrated into workflows that affect patient outcomes.

6. Conclusions

By reframing the task into a high-severity (Alta) vs. not high-severity (No Alta) classification, we aligned our modeling with real-world clinical priorities—helping identify patients who may require urgent or intensive care.

Logistic Regression emerged as the most balanced and robust model, offering high precision and recall with minimal tuning.

The Random Forest classifier, although initially biased toward the majority class, demonstrated adaptability with threshold adjustments, enhancing its ability to detect critical cases.

The combination of structured preprocessing, clinically relevant feature selection, and targeted evaluation metrics provides a strong foundation for future deployments or integrations into hospital decision support systems.

7. Limitations and Future Work

Despite promising results, several limitations remain. First, the dataset exhibited class imbalance, which may have influenced the models' performance on underrepresented high-severity cases. While we addressed this partially through threshold adjustments, future work could incorporate resampling strategies such as SMOTE or apply class weighting to further improve minority class detection.

Moreover, our decision to exclude severity 0 cases—often representing outpatients or very minor conditions—streamlined our analysis but also narrowed the model's generalizability. Future studies could explore multi-class classification approaches that include all severity levels to capture a broader spectrum of patient cases.

Other avenues for improvement include:

- Feature engineering strategies, such as incorporating relationships between diagnoses and procedures.
- Exploring neural networks or embedding-based approaches to handle code sequences.
- Validating the model on external datasets from other hospitals to ensure robustness and generalizability.

Ultimately, this work demonstrates that even with simple models and structured data, it is possible to derive clinically meaningful insights—laying the groundwork for scalable machine learning tools in healthcare.

8. References

- Nistal-Nuño, B. (2022). *Developing machine learning models for prediction of mortality in the medical intensive care unit. Computer Methods and Programs in Biomedicine,* 216, 106663. https://doi.org/10.1016/j.cmpb.2022.106663

- Shamout, F., Zhu, T., & Clifton, D. A. (2021). *Machine learning for clinical outcome prediction. IEEE Reviews in Biomedical Engineering,* 14, 116–126. https://doi.org/10.1109/RBME.2020.3007816

- Aljameel, S. S., Khan, I. U., Aslam, N., Aljabri, M., & Alsulmi, E. S. (2021). *Machine learning-based model to predict the disease severity and outcome in COVID-19 patients. Scientific Programming,* 2021, 5587188. https://doi.org/10.1155/2021/5587188