

Andrea Barroso Gil

Behavioral Risk Prediction in Acute Mental Health Inpatient Unit

MASTER'S THESIS

supervised by Dr. Joan T. Matamalas (URV) and advised

Laura Miró Mezquita (Hospital Universitari Germans Trias i Pujol)

Master's Degree in Health Data Science (MHEDAS)



Barcelona, 2025

Acknowledgments

As we come to the end of this journey, I want to thank all the people who have accompanied me along the way. First and foremost, the team at the Acute Mental Health Inpatient Unit at Hospital Universitari Germans Trias i Pujol, for opening their doors to me (in every sense) and showing me the reality of their daily lives, first during my internship and later during the preparation of my master's thesis. To the promoters of this project, Dr. Cuevas, Dr. de Pablo, and Anna Moreno, because without them, this solution would not exist and they have given me the knowledge and motivation necessary to achieve my goals.

In particular, I would like to thank my advisor Laura, who, together with the psychiatry team, does a magnificent job providing innovative solutions to healthcare practice. She was the instigator of this idea, laid the groundwork, and has been a key player in this project, contributing creativity, rigor, and method, a great traveling companion. I also want to thank Guillem, the hospital's data scientist, who has helped advance this project and, with his experience, has enriched the work.

Finally, I would like to thank my academic advisor, Joan Matamalas, because, despite the distance, the time changes, and the daily workload, he has always been available to answer my questions and offer his support.

To my fellow Master's students, for having accompanied me during these two years culminating in this final project: thank you for making me feel understood and supported. And, finally, to my friends, my family, and my boyfriend, for never doubting me and for always believing in me. That energy you transmit to me has helped me achieve all my goals.

And finally, I remind myself: "If you believe it, you create it." A lot of effort and work, but the result is always rewarding.

Abstract

The transition to open-doors, patient-centered models in acute mental health inpatient units has introduced new demands for effective risk management of self-harm, aggression, and absconding behaviors. Existing risk assessment tools, largely based on static or retrospective information, are often inadequate for anticipating critical events in these dynamic care settings.

This study examines the ability of several machine learning models to predict short-term risky behaviors using systematically collected clinical data from an open-door psychiatric unit. Analysis of this real-world dataset reveals that predictive performance is generally moderate, with ensemble methods such as XGBoost and Random Forest showing slightly better results when using daily data aggregation. However, the advantage over other algorithms is limited, and the model's accuracy relies heavily on a small set of variables, notably schizophrenia diagnosis, gender, age, and certain risk indicators.

The restricted set of influential predictors highlights both the strengths and limitations of current routine data for this purpose. Difficulties persist in accurately identifying less frequent events due to class imbalance and limited diversity of variables. These findings suggest that while machine learning models can provide valuable support for risk assessment in open psychiatric settings, significant improvement will depend on expanding the scope and quality of clinical data, as well as refining modeling strategies to address the complexities of real-world practice.

Resumen

La transición a modelos de puertas abiertas centrados en el paciente en unidades de hospitalización de salud mental aguda ha generado nuevas demandas para una gestión eficaz del riesgo de autolesiones, agresión y conductas de fuga. Las herramientas de evaluación de riesgos existentes, basadas principalmente en información estática o retrospectiva, suelen ser inadecuadas para anticipar eventos críticos en estos entornos de atención dinámicos.

Este estudio examina la capacidad de varios modelos de aprendizaje automático para predecir conductas de riesgo a corto plazo utilizando datos clínicos recopilados sistemáticamente en una unidad psiquiátrica de puertas abiertas. El análisis de este conjunto de datos reales revela que el rendimiento predictivo es generalmente moderado, y que métodos de conjunto como XGBoost y Random Forest muestran resultados ligeramente mejores al utilizar la agregación diaria de datos. Sin embargo, la ventaja sobre otros algoritmos es limitada, y la precisión del modelo depende en gran medida de un conjunto reducido de variables, en particular el diagnóstico de esquizofrenia, el sexo, la edad y ciertos indicadores de riesgo.

El conjunto limitado de predictores influyentes destaca tanto las fortalezas como las limitaciones de los datos rutinarios actuales para este propósito. Persisten las dificultades para identificar con precisión los eventos menos frecuentes debido al desequilibrio de clases y a la limitada diversidad de variables. Estos hallazgos sugieren que, si bien los modelos de aprendizaje automático pueden brindar un valioso apoyo para la evaluación de riesgos en entornos psiquiátricos abiertos, una mejora significativa dependerá de la ampliación del alcance y la calidad de los datos clínicos, así como del perfeccionamiento de las estrategias de modelado para abordar las complejidades de la práctica clínica real.

Resum

La transició a models de portes obertes i centrats en el pacient en unitats d'hospitalització de salut mental aguda ha introduït noves demandes per a una gestió eficaç del risc d'autolesió, agressió i conductes de fuga. Les eines d'avaluació de riscos existents, basades en gran part en informació estàtica o retrospectiva, sovint són inadequades per anticipar esdeveniments crítics en aquests entorns d'atenció dinàmica.

Aquest estudi examina la capacitat de diversos models d'aprenentatge automàtic per predir comportaments de risc a curt termini utilitzant dades clíniques recollides sistemàticament d'una unitat psiquiàtrica de portes obertes. L'anàlisi d'aquest conjunt de dades del món real revela que el rendiment predictiu és generalment moderat, amb mètodes de conjunt com XGBoost i Random Forest que mostren resultats lleugerament millors quan s'utilitza l'agregació de dades diàries. Tanmateix, l'avantatge sobre altres algoritmes és limitat, i la precisió del model depèn en gran mesura d'un petit conjunt de variables, en particular el diagnòstic d'esquizofrènia, el sexe, l'edat i certs indicadors de risc.

El conjunt restringit de predictors influents destaca tant els punts forts com les limitacions de les dades rutinàries actuals per a aquest propòsit. Persisteixen les dificultats per identificar amb precisió esdeveniments menys freqüents a causa del desequilibri de classes i la diversitat limitada de variables. Aquestes troballes suggereixen que, si bé els models d'aprenentatge automàtic poden proporcionar com un suport valuós per a l'avaluació de riscos en entorns psiquiàtrics oberts, una millora significativa dependrà d'ampliar l'abast i la qualitat de les dades clíniques, així com del refinament de les estratègies de modelització per abordar les complexitats de la pràctica del món real.

List of Figures

Figure 1. RutiSafeNet DashBoard	23
Figure 2. RutiSafeNet main screen. Overview of the digital panel with risk traffic lights and patient load. The door status and overall risk level are displayed at the top. Personal data has been hidden for confidentiality reasons.	24
Figure 3. Progress graphs in RutiSafeNet. Visualization of changes in risk and workload over time (by shift, day, or week), useful for clinical monitoring and planning.....	25
Figure 4. Monthly distribution of the total number of clinical observations recorded during the analysis period.	42
Figure 5. Total activation frequency of clinical risk indicators extracted from the free text in RutiSafeNet.....	43
Figure 6. Activation frequency of care burden indicators, reflecting the functional complexity of the patients.	43
Figure 7. Distribution of key sociodemographic characteristics among unique patients included in the study. Each subplot shows the frequency of patients by category for the following variables: (A) Presence of a schizophrenia diagnosis, (B) Age over 35 years, (C) Involuntary admission status, (D) Suicide risk as a reason for admission, and (E) Risk of harming others as a reason for admission, ...	44
Figure 8. Total observations by risk category documented in Geslin: aggression, self-harm, absconding, and no risk.....	45
Figure 9. Distribution of the three main clinical risk categories by gender, in total number of observations per group.....	46
Figure 10. Complete correlation matrix between all clinical indicators, care burden factors, and risk categories.	47
Figure 11. Correlation submatrix centered on the target variables Aggressive, Self-Harm, and Absconding, showing their relationship with the various clinical and care indicators.	48
Figure 12. Visual representation of the experimental configurations: combinations of analysis unit, classification type, and input features.....	51
Figure 13. Workflow for data preprocessing, variable encoding, and model validation using leave-one-group-out cross-validation in all experimental configurations.	52
Figure 14. ROC curves for the top 3 models, binary risk prediction by shift (turn)	56

Figure 15. ROC curves for the top 3 models, binary risk prediction by day	56
Figure 16. Confusion matrices for the top binary models (Random Forest by shift, XGBoost by day) using binary features.	57
Figure 17. Confusion matrices for neural network (binary and weighted features) in multiclass risk prediction.....	59
Figure 18. Confusion matrix for XGBoost (binary or weighted features) in multiclass risk prediction	59
Figure 19. Project Gantt Chart.....	68

List of Tables

Table 1. Behavioral risk assessment tools and predictive models in mental health, classified according to their technique and level of automation.	20
Table 2. Example of original Geslin register structure, where each row represents a shift and risk behaviors are assigned through the patient identifier in the corresponding column.	29
Table 3. Example of reordered structure where each row represents a single patient in a given shift, with binary coding for each risk category.	30
Table 4. Weighting assigned to each risk factor	37
Table 5. Risk behavior dataset (by shift)	38
Table 6. Binary risk factor dataset (by patient-shift)	39
Table 7. Weighted risk factor dataset (per patient/shift)	39
Table 8. Sociodemographic dataset (per patient/shift)	40
Table 9. Summary of experimental configurations	50
Table 10. Comparative performance of all models for binary risk prediction (best values in green).	54
Table 11. Comparative performance of all models for multiclass risk prediction (best values in green).	54
Table 12. Pairwise McNemar test results for binary classification scenarios	57
Table 13. Wilcoxon signed-rank test results comparing top multiclass models per scenario.	60
Table 14. Top 10 Most Influential Features (All Models, All temporal grouping)	61

SUMMARY

1. GLOSSARY	11
2. INTRODUCTION	12
3. OBJECTIVES	13
4. STATE-OF-THE-ART	14
4.1. Behavioral risk in psychiatric hospitalization: clinical context and traditional management	14
4.2. Risk assessment tools	16
4.3. Hospital Universitari Germans Trias i Pujol case	22
5. DESIGN AND DEVELOPMENT	27
5.1. Data collection instruments	27
5.2. Preprocessing steps	28
5.3. Final dataset structure	38
6. EXPERIMENTS AND RESULTS	41
6.1. Dataset overview and variable description	41
6.2. Exploratory data analysis	46
6.3. Experimental setup	48
6.4. Results	53
7. DISCUSSION	63
8. PROJECT IMPACT AND SCHEDULE	66
8.1. Ethical-social impact	66
8.2. Sustainability	66
8.3. Diversity	67
8.4. Project Timeline	68
9. CONCLUSIONS AND FUTURE WORK	69
REFERENCES	70
ANNEXES	74

1. Glossary

AI: Artificial Intelligence

AUC: Area Under the Curve

BVC: Brøset Violence Checklist

CEIC: Clinical Research Ethics Committee

GDPR: General Data Protection Regulation

Geslin: Name of the structured digital form for incident recording at the hospital; not a traditional acronym, but capitalized as a system/tool name in the thesis

HCR-20: Historical Clinical Risk Management-20

HUGTiP: Hospital Universitari Germans Trias i Pujol

KNN: k-Nearest Neighbors (machine learning algorithm)

LOPOCV: Leave-One-Patient-Out Cross-Validation

ML: Machine Learning

MLP: Multilayer Perceptron

NHC: Número de Historia Clínica (Patient Identifier Number)

NLP: Natural Language Processing

OxMIV: Oxford Mental Illness and Violence

OxMIS: Oxford Mental Illness and Suicide

RutiSafeNet: Name of the digital risk monitoring platform; not a traditional acronym, but a proprietary tool

SHAP: SHapley Additive exPlanations

SVM: Support Vector Machine

START: Short-Term Assessment of Risk and Treatability

XGBoost: Extreme Gradient Boosting

2. Introduction

Healthcare in psychiatric inpatient units has evolved toward more open and patient-centered models, moving away from closed and restrictive structures [1,2]. This transformation, aligned with international human rights frameworks, aims to promote autonomy, recovery, and reduce stigma in people with mental disorders.

However, open units present significant clinical challenges, particularly regarding risk management. The presence of patients exhibiting potentially dangerous behaviors (such as self-harm, aggression, or absconding) requires effective mechanisms to anticipate these events without systematically resorting to restrictive measures [3–5].

In clinical practice, care teams use structured tools such as the HCR-20 [6], START [7], or the Brøset Violence Checklist [8] to assess the risk of violence. However, these tools are largely based on static variables, lack the ability to adapt to the changing context of an acute care unit, and are not designed to be integrated into daily care [9–11].

Faced with these limitations, digital technologies and artificial intelligence (AI)-based approaches offer new opportunities. The ability to analyze structured clinical data and systematically collected observations allows for the construction of personalized predictive models that are sensitive to daily changes and specific to the care environment [12–14].

The HUGTiP has initiated this transformation by digitizing clinical incidents (Geslin) and implementing the RutIsSafeNet system, which visually and updatably captures individual and collective risk in the unit [15]. This digital environment generates a considerable volume of data that can be analyzed using machine learning techniques to anticipate short-term risk behaviors with greater accuracy than traditional tools [16,17].

This work is part of this line of innovation, aiming to develop a clinically useful predictive model adapted to the real-life context of an acute psychiatric unit with an open-door policy. This tool aims to strengthen the care team's capacity to prevent critical events, reduce the use of coercive interventions, and improve the quality of care through risk management based on real, local, and dynamic data [18,19].

3. Objectives

This work focuses on the analysis of a unique set of clinical data obtained from patients in acute mental health inpatient unit, to understand the occurrence of conflictual behaviors. The main objective is to maximize the value of this data by identifying patterns and factors associated with risky behaviors (aggressiveness, self-harm, or absconding attempts). Complementarily, the development of a predictive model is proposed to estimate the probability of these events, always prioritizing the analysis and interpretation of the information.

Main objective

Analyze the data collected in a structured manner during hospitalization, identifying patterns and factors related to the occurrence of conflictive behaviors, and develop a predictive model that helps estimate individual risk.

Secondary objectives

- Build the longitudinal dataset by patient and shift.
- Evaluate different ways to represent temporal information and clinical variables.
- Analyze the impact of different forms of coding clinical indicators.
- Explore different modeling approaches for risk classification.
- Identify combinations of risk factors associated with conflictive behaviors.

4. State-of-the-Art

This section is structured into several subsections that progressively address the main approaches and tools available for the assessment and prediction of behavioral risk in psychiatric hospitalization units. First, the clinical context and traditional risk management models are described (4.1), followed by a review of structured assessment tools, both conventional and based on artificial intelligence (4.2). Subsequently, the current limitations of automated predictive models are analyzed (4.2.3). Finally, the specific case of the HUGTiP (4.3) is presented, contextualizing the local evolution and development of the model proposed in this study.

4.1. Behavioral risk in psychiatric hospitalization: clinical context and traditional management

Acute psychiatric hospitalization poses a highly complex clinical environment, characterized by the coexistence of severe conditions, high patient turnover and a significant frequency of disruptive behaviors. In this context, events such as self-harm, aggression towards others or absconding represent a constant risk for both patients and care teams. Managing these risks effectively and ethically remains one of the major challenges in the field of mental health.

For years, the predominant model has been that of closed units, where patient mobility is restricted as a preventive measure. This approach is based on the assumption that environmental control reduces the probability of critical incidents. However, different studies have questioned this premise, showing that closed units do not always manage to reduce absconding or suicide rates, and can, instead, generate greater subjective discomfort, distrust of staff and increased relational conflicts [1,20,21].

In response, a paradigm shift towards less restrictive models, centered on the autonomy and dignity of the user, has been promoted at the international level [22]. Open-door units are presented as a possible alternative, allowing greater freedom of movement within the hospital space. The available evidence suggests that, when implemented with adequate protocols and sufficient resources, these models do not imply a significant increase in clinical risk. Reductions in certain indicators of conflict have even been documented, especially in settings where good supervision and appropriate staffing ratios are ensured [3,4].

However, these approaches present significant operational challenges, especially in high-intensity care units. Risk management in patients with severe behavioral disturbances requires continuous observation and individualized intervention plans.

Some institutions have opted for hybrid approaches that combine openness measures with enhanced risk management protocols, tailored to the needs of each shift and available resources. This is the case of HUGTiP where the present project is framed, and where initiatives have been launched to reduce coercive measures through digitalization strategies, structured monitoring and early risk assessment [15].

In parallel to structural changes, care models have also been developed aimed at reducing the frequency of critical incidents by improving the therapeutic climate. Within this framework, structured interventions aimed at preventing the escalation of disruptive behavior through communication, anticipation and the active involvement of the care teams have been promoted [2]. In some centers, these strategies have been complemented with digital tools that allow real-time monitoring of relevant clinical variables, thus facilitating proactive decision-making and dynamic adjustment of supervision levels (9,10).

Safewards is a structured care model that proposes ten evidence-based interventions designed to reduce conflict and restraint in psychiatric inpatient units. Its focus is on improving unit climate and communication between patients and professionals, based on the premise that many disruptive behaviors are not a direct consequence of individual patient characteristics, but are influenced by environmental and interpersonal factors that can be modified. The implementation of Safewards has proven effective in decreasing critical incidents and reducing the use of coercive measures, although the results depend on the context and fidelity in their application [1].

At HUGTiP, the Safewards model has been progressively integrated since 2019, as part of a broader strategy to transform care environments. Its implementation has been accompanied by digital tools such as RutiSafeNet, a real-time risk monitoring system that makes it easier for clinical teams to identify early signs of behavioral escalation, enabling proactive interventions and better organization of care resources [15].

Taken together, psychiatric risk management is increasingly oriented toward minimizing coercion while ensuring safety through evidence-based strategies. Current practices combine open-door policies, structured models like Safewards, and real-time digital assessment tools. These approaches reflect a shift toward proactive, patient-centered care grounded in clinical risk analysis and continuous monitoring.

4.2. Risk assessment tools

4.2.1. Traditional risk assessment tools

Risk assessment in acute psychiatric inpatient units has traditionally focused on identifying the likelihood of certain adverse events occurring during the stay. This assessment is usually organized around three main categories: aggression toward others, selfharm behaviors and absconding. Each represents a high-impact clinical concern, both in terms of its implications for the safety of the environment and for therapeutic planning and care burden [18].

Violence or aggressiveness includes physical aggression, verbal threats and destruction of objects, and is one of the most critical manifestations in daily clinical practice due to its direct impact on the health care team. Although these episodes are usually concentrated in a minority of patients, their occurrence is frequently associated with active psychosis or substance use [5]. Self-injurious behaviors include suicidal ideation, suicide attempts and non-suicidal self-injury, and their early identification is a constant challenge due to the poor predictive capacity of many classic tools [19]. Absconding, understood as unauthorized departure of the patient, also represents a significant risk factor, with potential consequences such as relapse, suicidal risk and loss of the therapeutic bond.

Several structured tools are currently used to assess these types of risk:

- HCR-20 (Historical Clinical Risk Management-20) was first developed by in 1997. Now in its third version, it is one of the most widely used tools for violence risk assessment. It includes 20 items across three domains: historical (e.g., prior violence), clinical (e.g., current symptoms), and risk management factors (e.g., treatment compliance). The HCR-20 has been applied in over 35 countries and is considered a reference standard in forensic psychiatry. However, it relies on clinical judgment and has shown only moderate inter-rater reliability and predictive accuracy in some studies [6].
- START (Short-Term Assessment of Risk and Treatability), introduced in the early 2000s, was designed to assess multiple types of short-term risk, including violence, self-harm, suicide, and victimization. The tool focuses on dynamic and protective factors, allowing for an assessment sensitive to rapid clinical changes. Although it offers a comprehensive approach tailored to diverse contexts, its application is subject to some variability in scoring, and its direct impact on improving clinical outcomes has not yet been fully established [7].
- Brøset Violence Checklist (BVC), developed in 2000, is a brief clinical observation tool that allows for the prediction of violent behavior within a very short time horizon,

typically between 24 and 48 hours. It assesses six key behavioral indicators: confusion, irritability, boisterous behavior, verbal threats, physical agitation, and recent aggression. Its use has been shown to be effective in predicting immediate risk in acute inpatient units, although its usefulness diminishes when a longer assessment is required or in more complex clinical settings [8].

- OxMIV (Oxford Mental Illness and Violence), developed in 2016, estimates the probability that a person with schizophrenia or bipolar disorder will commit a violent crime within a year. This model, based on logistic regression, incorporates 16 clinical and sociodemographic variables routinely available in clinical practice, including criminal record, substance use history, age, sex, and socioeconomic status. Its digital and accessible design allows for simple implementation in clinical settings, with high discriminatory capacity in initial validation studies [9,10].
- OxMIS (Oxford Mental Illness and Suicide), developed in 2020, aims to estimate the 12-month suicide risk in patients with serious mental illness. It is based on a set of 17 variables, including previous suicide attempts, primary psychiatric diagnosis, criminal history, and recent hospitalization data. Its structure as a clinical screening tool has been assessed as adequate, and although its performance has been acceptable in recent external validations, its evaluation in other clinical settings is still needed to confirm its general applicability [11,23].

The latter two belong to a collection of online calculators developed by the Oxford group, being among the most validated in clinical settings. Their transparency and accessibility make them suitable for routine use in clinical practice, although their performance in non-European populations is still under investigation [23,24]. Other tools from this group include OxRec2, a model for assessing the risk of violent recidivism among released prisoners, based on routinely collected data and intended for application in the criminal justice setting; FOxWeb, a structured framework for long-term risk assessment in forensic psychiatry services, commonly used for leave or discharge planning; and FoVox, a brief tool designed to assess short-term risk of violence at the time of admission. All are designed to complement clinical judgment and prioritize ease of use [25].

4.2.2. AI-based risk assessment studies

Although Artificial Intelligence (AI) technology has been widely deployed in medicine for applications related to physical health, its adoption in the field of mental health has been slower. Mental health professionals work daily with the more human dimension of patients, building therapeutic bonds and directly observing behaviors and emotions, which makes it difficult to automate these processes. In addition, clinical data in mental health are often qualitative and subjective, based on the perceptions of patients and professionals, which adds complexity to their treatment by AI algorithms.

Much of the AI algorithms developed for mental health have been trained using very specific databases, limiting the applicability of their results to specific contexts or populations [12]. Even in studies that apply machine learning techniques, sample sizes are often small, making it difficult to generalize findings [26]. Reasons for the delay in the adoption of AI in routine psychiatric practice include the sensitive nature of the data generated in patient-professional interaction and the complexity of diagnostic criteria. Unlike more objective tasks where current AI performs well, in psychiatry both clinical observations and therapeutic decisions are subjective, incomplete, and variable between evaluators, making model training difficult [26].

Despite these challenges, AI offers great potential to redefine mental illness more objectively, personalize treatments, and encourage patients' active participation in their own care. However, its development must address critical issues such as bias, privacy, transparency, and other ethical issues. Ideally, advances in AI will contribute to providing more compassionate and ethical care for diverse populations.

On the other hand, AI algorithms can be used to design new clinical risk prediction models, provided they have rigorous internal and external validation processes to ensure their clinical utility [27]. A particular challenge in psychiatry is that the performance of supervised algorithms is highly dependent on the quality of the diagnostic labels used for their training. Given the subjective nature of many diagnoses in mental health, disease labels may lack the specificity needed to achieve algorithms with high sensitivity and specificity. An emerging alternative is to use machine learning to predict specific symptoms or specific functional consequences rather than global diagnoses [27].

In addition to the tools developed in Oxford, small-scale algorithms have been developed to specifically predict different types of risk. One of the most widely used approaches has been the comparative combination of different algorithms within the same clinical sample, such as penalized logistic regression, support vector machines (SVMs), multilayer neural networks and ensemble models such as random forests.

In studies applied to patients diagnosed with schizophrenia admitted to acute units, random forest-based models have shown very high levels of accuracy in predicting aggressive behaviors, with areas under the curve (AUC) close to 0.96 [16]. This type of algorithm, by combining multiple decision trees and reducing the risk of overfitting, has demonstrated a remarkable ability to handle heterogeneous data and to detect nonlinear relationships between clinical and behavioral variables. Moreover, its ability to provide probabilistic estimates and to assess the relative importance of variables makes it a particularly useful option in clinical settings where risk factors need to be interpreted and prioritized.

In the field of suicide risk, specific models have been developed using classical logistic regression applied to routine data collected in medical records. These tools allow for estimating suicide risk within a one-year period based on predefined sets of variables, such as previous suicide attempts, current diagnoses, legal status, and sociodemographic characteristics.

Although these models have more moderate accuracy (with AUCs around 0.70) they have been positively evaluated for their ease of implementation and for offering a structured framework that allows for automated alerts within clinical information systems [11,23].

Models focused on predicting interpersonal violence in the medium term have also been developed, integrating clinical variables, criminal history, substance use, and social conditions. These models have been validated in large population-based databases and have shown high discriminatory capacity, with AUCs exceeding 0.85. Their usefulness lies in their applicability as screening tools, especially for planning secondary prevention measures and allocating resources for intensive surveillance or intervention [25].

It is worth noting that some of these tools, such as OxMIS and OxMIV, were already mentioned in the previous section due to their clear structure and widespread clinical use. However, their inclusion in this section is due to the fact that they represent a middle ground between traditional instruments and more advanced predictive models. Although they use classical statistical techniques, their design is geared toward individualized risk prediction, they integrate multiple routine variables, and are available as automated digital platforms. This hybrid nature justifies their presence in both contexts: as part of the current structured assessment arsenal and as direct predecessors of artificial intelligence-based models.

More recently, the development of deep learning models has made it possible to explore previously underutilized sources of information, such as clinical notes written in natural language. Using natural language processing (NLP) techniques and deep neural networks, models have been trained to identify the risk of aggression from texts generated by clinical staff during daily care. Some of these models have outperformed human evaluators and traditional structured models in short-term prediction tasks (e.g., 72 hours), achieving F1 scores close to 0.75. This approach allows for the incorporation of nuances of clinical discourse and the detection of latent patterns in language that could anticipate relevant behavioral changes, representing a promising development in the field of prediction based on unstructured data [17].

A summary of the main AI-based tools and models for risk assessment is presented in Table 1.

Table 1. Behavioral risk assessment tools and predictive models in mental health, classified according to their technique and level of automation.

Model or Tool	Technique	Predicted Risk	Target Population	Key Variables	Classification
HCR-20 (v3)	Structured clinical judgment	Future violence	General and forensic psychiatry	Violence history, current symptoms, treatment compliance	Clinical support tool
START	Structured dynamic assessment	Violence, suicide, self-harm, victimization	Acute psychiatric inpatients	Dynamic risk and protective factors	Clinical support tool
BVC (Brøset)	Behavioral checklist	Imminent aggression (24–48h)	Psychiatric inpatients	Irritability, confusion, verbal threats, agitation	Clinical support tool
OxMIV (2016)	Logistic regression (16 factors)	12-month violence risk	Schizophrenia or bipolar disorder	Criminal history, substance use, age, sex, SES	Hybrid prediction tool
OxMIS (2020)	Logistic regression (17 factors)	12-month suicide risk	Severe mental illness	Suicide attempts, diagnosis, legal history	Hybrid prediction tool
OxRec2	Advanced logistic regression	Violent reoffending after release	Post-prison population	Judicial data, clinical history, social factors	Hybrid prediction tool
FoVOx / FOxWeb	Structured digital tool	Short/long-term violence risk	Forensic psychiatry	Structured risk factor scales	Clinical support tool
Random Forest model	Machine learning - Random Forest	Aggression during hospitalization	Inpatients with schizophrenia	Clinical data, family support, insight	AI-based predictive model
NLP-Deep Learning model	Deep learning + NLP	Violence toward staff in 72h	General psychiatric inpatients	Free text notes and structured data	AI-based predictive model

The table above summarizes the main tools currently used for behavioral risk assessment and prediction in the field of mental health, ranging from structured clinical support tools to advanced artificial intelligence-based models. This classification allows us to observe the progressive evolution from approaches focused on clinical judgment to automated systems capable of processing large volumes of data and generating individualized predictions. Logistic regression models, such as OxMIS or OxMIV, represent a transition point between the two approaches, while algorithms such as random forests or deep neural networks already operate as fully automated predictive models.

4.2.3. AI tools limitations

However, the application of artificial intelligence in psychiatric clinical settings poses a number of ethical, technical, and organizational challenges that must be addressed before considering its widespread implementation. Although machine learning models have demonstrated improvements in predictive accuracy, their integration into inpatient units requires considering fundamental aspects such as algorithmic bias, transparency, the protection of sensitive data, and their actual clinical utility.

One of the main risks identified is the reproduction of pre-existing biases. Some studies have shown that models trained with historical data can perpetuate structural inequalities, associating demographic or socioeconomic variables with greater risk without considering the social or clinical context. This can lead to disproportionate healthcare decisions, especially in vulnerable populations [28].

Another critical aspect is the interpretability of the models. Many of the most accurate algorithms, especially those based on deep neural networks, act as black boxes. This lack of explainability can hinder their acceptance by clinical staff, who need to understand and justify the decisions derived from a prediction in order to integrate them into clinical practice. In response to this limitation, explainable artificial intelligence approaches are being developed that seek to balance performance and understandability [29].

Privacy protection is also a key factor. Most models require access to large volumes of clinical information, including electronic records, narrative notes, and biometric data. Without adequate security measures, there is a risk of unauthorized secondary uses, such as applications in insurance or employment processes, especially when data is shared between institutions or stored on commercial platforms [30].

From an operational perspective, the usefulness of these models depends on their effective integration into clinical workflows. Predictions must be available in real time, in a clear format aligned with local protocols, enabling concrete actions such as increased observation, staff redeployment, or medication review. Without these links between prediction and action, the practical value of the tool is significantly reduced [31].

Finally, there are significant limitations in terms of validation and generalizability. Most current models have been developed and evaluated in specific institutional or national contexts. There are few cases of rigorous external validation and even fewer prospective studies in real-life clinical settings. This lack of evidence limits their universal applicability and reinforces the need to develop models tailored to each healthcare setting, validated with local data, and designed in collaboration with the professionals who will use them [32].

The review in this section has shown how behavioral risk assessment in mental health has shifted from relying primarily on structured tools based on clinical judgment to incorporating automated models using artificial intelligence techniques. The former remains useful due to their simplicity and applicability, but have limitations in accuracy and adaptability to rapid changes. Data-driven predictive models offer greater anticipatory capacity, although they require validation, clinical integration, and ethical safeguards. Both approaches perform complementary functions and mark a starting point for the development of solutions more tailored to the healthcare context, such as the one proposed in this study.

4.3. Hospital Universitari Germans Trias i Pujol case

The acute psychiatric inpatient unit at HUGTiP was opened in 2018 and is located on the 12th floor of the hospital. It serves adults with psychiatric decompensations, eating disorders with medical or comorbid risks, and patients undergoing detoxification from substance use. It has 14 beds distributed across six double rooms and two single rooms, the latter for patients with more complex clinical conditions. The unit has video surveillance and separate spaces for clinical care, leisure time, and family contact.

Since its opening, the unit has operated under a closed-door model, although with a flexible policy regarding accompanied departures. Beginning in 2019, a shift began toward a blended model, combining the structured Safewards approach with progressive opening criteria, regulated by risk assessment. This model's main objectives are promoting patient autonomy, reducing stigma, and improving the quality of care. Opening doors is considered an ethical and therapeutic component, especially for voluntarily hospitalized patients, and is subject to clinical, organizational, and safety criteria. This requires a systematic risk assessment that takes into account both patient behavior and staff care burden.

In this context, the hospital has developed two complementary tools. The first is Geslin, a digital form implemented in 2012 that allows for the structured recording of relevant clinical incidents such as assaults, absconding attempts, and self-harm. This tool acts as a retrospective database that facilitates the analysis of events occurring during hospitalization.

The second tool, and the focus of this section, is the risk categorization system initially tracked on a handwritten board at the nursing station. This was later digitized into the RutiSafeNet platform (Figure 1), a comprehensive and interactive web-based system developed in collaboration with a software company [15]. This transition from analog to digital was crucial to ensure consistency, efficiency, and ease of updating risk data in real time.

RutiSafeNet was designed to replicate the visual clarity and intuitive functionality of the original board to ensure ease of use and high staff adherence. It captures both static patient data (such as name, admission date, and medical record number) and dynamic information—namely behavioral risk, care burden, permissions, and type of admission. One of the main requirements during the platform's development was to maintain a user-friendly and visually intuitive interface, with a particular emphasis on privacy—personal identifiers are concealed to preserve confidentiality in case patients view the screen.

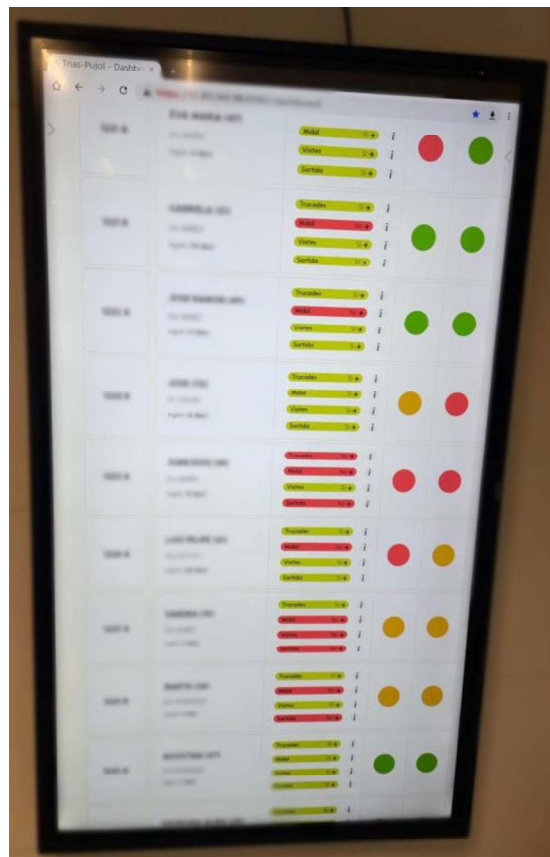


Figure 1. RutiSafeNet DashBoard

The platform interface (see Figure 1) features a main dashboard where each patient is represented by two traffic light indicators. The first light denotes overall behavioral risk, and the second corresponds to the care burden. The traffic light system uses color coding—green (0) for low risk, yellow (1) for moderate, and red (2) for high—linked to specific predefined clinical criteria. For instance, risk evaluations may reflect observed frustration, self-harm ideation, absconding attempts, or aggressive behavior. The care burden assessment takes into account the patient's dependence for daily activities, the need for observation, and spontaneous demand for staff attention.

The top bar of the dashboard displays the overall unit risk score and door status. If the cumulative risk score across all patients surpasses a defined threshold (e.g., 20), it signals that it may be unsafe to open the doors. However, final decisions are made collaboratively by the clinical team. The platform also includes a manual override button that allows staff to indicate the actual door status if it differs from what the calculated score suggests.

As seen in Figure 6, the system not only enhances visibility of individual risk but also provides real-time visual summaries of the entire unit's status. Moreover, it integrates an internal algorithm that automates score calculation and color coding, supporting timely clinical decision-making.

Beyond real-time monitoring, the platform includes data visualization tools (Figure 7), which chart the evolution of patient risk and care burden across different time frames (daily, weekly, or throughout hospitalization). This enables trend tracking and supports more informed planning.

Importantly, RutiSafeNet is fully customizable, meaning that fields and functionalities can be adjusted to match the unique requirements of the unit. This adaptability ensures that the platform remains a dynamic and evolving support system for clinical care.



Figure 2. RutiSafeNet main screen. Overview of the digital panel with risk traffic lights and patient load. The door status and overall risk level are displayed at the top. Personal data has been hidden for confidentiality reasons.

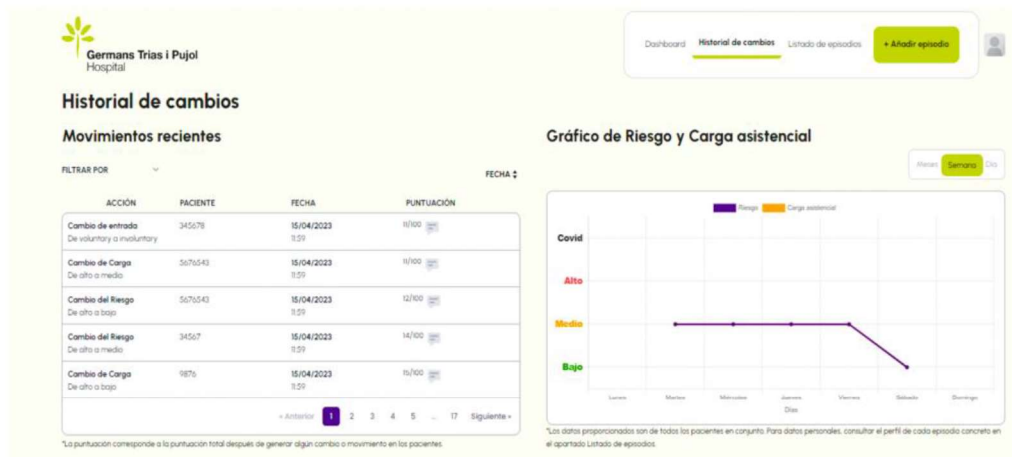


Figure 3. Progress graphs in RutisSafeNet. Visualization of changes in risk and workload over time (by shift, day, or week), useful for clinical monitoring and planning.

While Geslin allows for retrospective analysis of relevant clinical incidents and RutisSafeNet facilitates immediate and visual management of care and behavioral risk, neither of these tools alone can accurately anticipate the likelihood of critical events occurring in upcoming shifts. This lack represents a significant limitation, especially in a clinical environment transitioning to an open-door model, where risk anticipation is key to ensuring safety without resorting to coercive measures.

In this context, the RutisSafeNet project emerged, conceived within the framework of this Master's Thesis, with the aim of developing a local predictive model based on machine learning techniques. This model uses routine data already available in the unit, systematically collected through existing tools, to estimate the short-term risk of disruptive behaviors such as aggression, self-harm, or attempted absconding.

Unlike more general predictive models used in psychiatry, which typically focus on static variables (such as diagnoses, medical history, or sociodemographic data) and long-term outcomes (such as readmissions or six-month outcomes), RutisSafeNet focuses on immediate and dynamic prediction, integrated into daily clinical practice. This approach allows for the incorporation of variables highly sensitive to changes in the patient's clinical status, such as observed behavioral indicators and the care burden perceived by the team; critical factors for decision-making in acute hospitalization units.

The originality and added value of the RutisSafeNet model lies in its ability to seamlessly combine two dimensions typically excluded from standard models: on the one hand, dynamic behavioral indicators, recorded by the care team in real time, and on the other, subjective care burden, which reflects the patient's functional impact on the clinical team.

Both factors are especially relevant in units seeking to move toward less restrictive models without compromising the safety of the therapeutic environment

Furthermore, the model is specifically designed to fit the operational and organizational characteristics of the HUGTiP unit, making it more reliable and applicable than other generic developments that do not consider local specificities. This local adaptation allows the model to be continuously updated with real-world data, ensuring its validity and clinical utility.

In a complex and changing environment such as acute psychiatric hospitalization, where workload, patient turnover, and decisions must be made quickly and by consensus, having a tool that supports clinical judgment and strengthens the preventive capacity of the care team represents a significant advance. This is not only about improving safety, but also about contributing to more ethical, personalized, and data-driven care, aligned with the principles of reducing the use of coercive measures and promoting more open and respectful therapeutic environments.

5. Design and development

This section describes the methodological design and technical development of the system created to transform heterogeneous clinical records into a unified, coherent and analytically exploitable data structure. The objective has been to build a set of data that accurately represents the clinical status and risk factors of patients admitted to a psychiatric unit, so that it is compatible with advanced predictive models.

The system is based on the integration of two main sources of information: the Geslin structured form, completed by the nursing staff at the end of each shift, and the RutiSafeNet digital platform, used by the clinical team to record risk factors and care burden in real time. Based on these sources, multiple stages of processing have been applied, including cleaning, temporal standardization, longitudinal restructuring and imputation of sociodemographic variables.

In addition, the set of clinical and care indicators has been codified in binary and weighted formats, taking into account both the presence and the intensity of the risk. The final result is a longitudinal dataset per patient and shift, which allows detailed monitoring of the evolution of risk during admission and applying supervised learning techniques for the prediction of critical events.

It is important to make clear that, although the original sources contain the unique patient identifier (NHC), this data has not been used directly in the analysis. In the methodology, all the data have been pseudo-anonymized before being processed, replacing the NHC with an internal code specific to the study. In this way, it is not possible to re-identify patients at any point in the analysis or in the results. All the work has been carried out following data protection criteria and the codes used do not allow tracing the identity of the patients.

5.1. Data collection instruments

The system developed for the construction of the dataset starts from two main sources of clinical information: the Geslin form and the RutiSafeNet digital platform. Both instruments are used by the healthcare staff of the psychiatric hospitalization unit, although they fulfill different functions and complement each other in terms of the type of information recorded and the time of collection.

Geslin is a structured form that the nursing team completes manually at the end of each shift (morning, afternoon or night). Its purpose is to record, in a standardized way, the occurrence of risky behaviors observed during the shift. These behaviors include verbal or physical aggression, attempts at self-harm or suicide, absconding behaviors, among others.

The information is coded using unique patient identifiers (NHC), located in the columns corresponding to each type of behavior. In this way, it is possible to know what type of risk each patient presented and in what specific time it occurred. This instrument constitutes a structured and retrospective source of clinically relevant events.

On the other hand, **RutiSafeNet** is a digital tool used by the clinical team to record in real time both clinical risk factors and situations that involve a greater care burden. Unlike Geslin, which is completed at the end of the shift, RutiSafeNet allows you to enter observations during the care activity itself, through an interface with drop-down lists.

The entries include brief descriptions of detected risks and specific clinical needs of the patient (for example, verbal aggressiveness, accumulation of medication, need for constant accompaniment, etc.), which allows to capture clinical dynamics continuously and with a high level of structuring.

Both sources record the date and time of the registration, which has allowed the data to be synchronized and assigned to a specific shift. In addition, both contain explicit references to the patient by means of an identifier code (NHC), which makes their integration possible. The objective of the development has been to transform these two sources into a single data structure, where each row represents a patient in a specific shift, and includes the presence or absence of risk behaviors, dynamic clinical indicators and stable socio-demographic variables. This structure has been designed to be directly usable in supervised analyzes and in the construction of predictive models.

5.2. Preprocessing steps

Once the sources of information were defined, an exhaustive preprocessing process was carried out to transform the original data into a structured and analytically useful format. This process included cleaning tasks, standardization of formats, temporal filtering, normalization of identifiers and restructuring of records. In addition, pattern-based information extraction techniques, integration of socio-demographic variables and binary and weighted coding of risk factors will be applied. Each of these stages is detailed below for both data sources: Geslin and RutiSafeNet.

5.2.1. Geslin

The processing of the Geslin fountain constituted the first stage of technical development. This source contains structured records that the nursing team completes at the end of each shift, recording the appearance of relevant clinical risk behaviors during that period. The main objective was to clean, standardize and restructure these records to build a database oriented to analysis per patient and shift.

Initially, the source file was loaded and the Date column was converted to the datetime data type, forcing the conversion with `errors='coerce'` to detect invalid entries. Rows with non-convertible dates (NaT) were deleted. Next, a temporal filter was applied to select only the records between January 22, 2024 and February 21, 2025, an interval defined as the study's analysis frame. Finally, the data were ordered chronologically according to the date of the shift.

Of the total variables available in the Geslin form, only those related to risk behaviors observed during the shift will be selected. These are distributed in eight columns, each corresponding to a different risk category.

In the original table, each row represents a single shift (morning, afternoon or night), and the columns contain the identifiers of the patients who have presented this behavior. The selected columns were the following:

- 1) NHC_AV: verbal aggressiveness,
- 2) NHC_AFO: physical aggression against objects,
- 3) NHC_AFP: physical aggression against persons,
- 4) NHC_IA: self-harm attempts,
- 5) NHC_IS: suicide attempts,
- 6) NHC_IF: absconding attempts,
- 7) NHC_PF: family exit without return,
- 8) NHC_F: absconding from controlled environment.

The blank spaces or empty cells were replaced by null values (NaN) and the rows where there was no identifier registered in the eight columns mentioned were eliminated. Duplicates will also be eliminated by combination of Date and Turn, keeping only the first record in case of repetition. As shown in Table 2, patient identifiers appear within the columns corresponding to the different risk categories only when that type of behavior has been observed during the shift and day.

Table 2. Example of original Geslin register structure, where each row represents a shift and risk behaviors are assigned through the patient identifier in the corresponding column.

NHC_AV	NHC_AFO	NHC_AFP	NHC_IA	NHC_IS	NHC_IF	NHC_PF	NHC_F
patient1							

NHC_AV	NHC_AFO	NHC_AFP	NHC_IA	NHC_IS	NHC_IF	NHC_PF	NHC_F
	patient2					patient3	
			patient4				

Given that the original format is not directly exploitable for longitudinal analysis (since it groups multiple patients and risk categories in a single row), a restructuring process was applied. This consists of iterating through each row of the register and, for each non-empty cell in the risk columns, generating a new row in an auxiliary table. Each new row contains a unique patient identifier (NHC), the date and turn of the register, and a binary coding where the detected risk category is marked with 1 and the rest with 0. Thus, if in the same shift several patients present different types of behavior, multiple independent rows are created, one for each patient and type of risk.

The final result is a table (as showed in example below Table 3) in which each row represents an individual patient in a specific shift, with binary coding by type of risk. This structure allows for a clear analysis of the temporal evolution of risk per patient, and facilitates its subsequent integration with variables from other sources.

Table 3. Example of reordered structure where each row represents a single patient in a given shift, with binary coding for each risk category.

NHC	Date	NHC_ AV	NHC_ AFO	NHC_ AFP	NHC_ IA	NHC_ IS	NHC_ IF	NHC_ PF	NHC_ F
patient1	28-01-2024	1	0	0	0	0	0	0	0
patient2	20-08-2024	0	1	0	0	0	0	0	0
patient3	20-08-2024	0	0	0	0	0	0	1	0
patient4	19-12-2024	0	0	0	1	0	0	0	0

5.2.1.1. Integration of Sociodemographic Variables

Once the table of risk behaviors per patient and shift was restructured, it was enriched with socio-demographic and stable clinical information, that is, variables that do not vary throughout the hospital admission. These variables are found in specific blocks of the Geslin file and are not present in the daily records of shifts, so it was necessary to integrate them manually from the patient identifier (NHC).

The Geslin form contains three different blocks that group information per patient: blocks P, Y and AH. Each one of them contains a column that stores the patient identifier (NHC) and several adjacent columns with relevant sociodemographic and clinical variables.

These columns are organized by fixed positions, not by numbers, which required accessing them by column index. The variables included were:

- 1) Gender (H/M)
- 2) Diagnosis of schizophrenia? (S/N)
- 3) Over 35 years of age? (S/N)
- 4) Involuntary admission (S/N)
- 5) Admission due to the presence of suicide risk (S/N)
- 6) Admission for risk of harming others (S/N)

A sequential search procedure was constructed for each row of the main dataset (already structured by patient-shift). For each patient (NHC), we looked for its presence in the three Geslin blocks in the following order: first the P block, then the Y block, and finally the AH block.

If an exact match was found with the identifier in any of the blocks, the six associated variables were extracted directly from the corresponding columns.

Once the values were extracted, they were assigned to all the rows of the patient in the main table. In this way, each patient keeps a copy of their socio-demographic variables in each shift register, maintaining their stability throughout the entire time sequence. The extracted values were normalized to guarantee consistency in the format (for example, eliminating spaces or dashes, and converting text responses to binary categorical format).

In cases where the identifier was not found in any of the three blocks, null values or empty markers were assigned. Records were not eliminated due to lack of socio-demographic information, since the objective was to preserve all observations in turn, even if the demographic variable was not available.

This process made it possible to build a base integrated by patient and shift, which combines dynamic variables (risks per shift) with fixed variables (clinical and demographic profile), thus establishing the basic structure for use in supervised models.

5.2.2. RutiSafeNet

RutiSafeNet is a digital platform used by the clinical team to record real-time risk factors and care burden indicators during hospital admission. Unlike the Geslin form, which is completed retrospectively at the end of each shift, RutiSafeNet allows observations to be recorded as they occur, through a structured interface with drop-down lists and standardized fields. This source provides complementary information of a dynamic nature, enriching the clinical analysis with events collected within the development of the shift.

The preprocessing process began with the combination of multiple platform exports, unified in a single data structure. Once the records were consolidated, the format of the date and time column (Date) was standardized by converting it to the datetime type, which allowed the day (Date) and time (Hour) of each observation to be extracted separately.

From the recorded time, a time classification was applied to assign each observation a care shift. The morning shift (M) was assigned to records made between 08:00 and 14:59, the afternoon shift (T) between 15:00 and 21:59, and the night shift (N) to observations made between 22:00 and 07:59 of the following day. This temporary normalization made it possible to align the RutiSafeNet records with the shift structure used in Geslin.

Subsequently, the records were filtered to retain only those corresponding to the analysis period defined between January 22, 2024 and February 21, 2025. Observations without a valid patient identifier were also eliminated, after converting the corresponding column to numerical type and excluding non-convertible values. The field was renamed and standardized to ensure its compatibility with the identifiers of the rest of the system.

In addition to the date, the shift and the patient's identifier, two key fields are preserved: Item and Care burden. Both fields contain structured text selected through drop-down menus, which facilitates the subsequent extraction of clinical variables through pattern searches. The Item field collects observations related to clinical risk factors (such as verbalizations, self-injurious behavior or threats), while Care Burden describes situations that involve a greater demand or complexity of care (for example, the need for accompaniment, restraints or isolation).

This preprocessing phase left the RutiSafeNet data ready for structured exploitation, allowing the systematic detection of clinical indicators by means of regular expressions and their integration with the data from the Geslin form.

5.2.2.1. Detection of indicators via keyword matching

Once the Item and Care Charge columns of RutiSafeNet were preprocessed and normalized, the automatic extraction of clinical variables was carried out by searching for keywords. The objective was to transform the textual content of these columns into structured indicators, represented as binary variables, which will reflect the presence or absence of clinical risk factors and care burden in each shift.

For this, a set of clinical acronyms was defined, each one associated with one or several key terms. These terms were defined based on frequent input patterns in RutiSafeNet and validated in collaboration with the clinical team. Regular expressions capable of identifying both exact matches and morphological or semantic variants of the terms were implemented.

Two groups of variables were distinguished according to the column of origin:

- Variables extracted from Item, associated with dynamic clinical risk
- Variables extracted from Care Burden, associated with care overload

Each indicator was coded as a new binary column, where 1 indicates that the associated pattern was detected in the text of the corresponding turn, and 0 indicates its absence.

1. Clinical indicators extracted from the Item column:

Autolytic risk:

- VIA: verbalization of autolytic ideation
- VIPM: verbalization of passive ideas of death
- AMG: moderate or severe self-harm
- ALT: minor self-harm
- CAU: autolytic behaviors
- AOPAH: concealment of potentially harmful objects
- AMED: accumulation of medication
- SRAMF: suspicions expressed by the family
- OPCC: Occultation in blind spots
- PAP: psychomotor preagitation

Risk of absconding:

- MVE: explicit verbalization of absconding
- HVA: hypervigilance with access
- DCAT: registration requirements during the shift
- PIDT: insistent request for documentation
- NPMT: refusal to take medication
- AFPI: antecedents of involuntary income absconding
- DCMSI: cognitive impairment with ambulation
- PII: involuntary entry
- CCA: acute cognitive change
- SPS: suspicion towards him staff

Heteroaggressive risk:

- CHP: physical assaults on persons
- CHO: physical attacks on objects
- HVT: verbal aggressiveness
- CA: threatening behavior
- DR: recent withdrawal of containment

2. Care Burden indicators extracted from the Care Burden column:

- PH: hyperdemand patient
- PD-ABVD: dependency for basic activities of daily life
- PAH: need for assistance in hygiene
- PCA: patient in environmental containment
- AC: isolation due to contagion
- PTCA: presence of eating disorder
- PCM: mechanical containment
- RAC: need for constant accompaniment
- RAI: need for intermittent accompaniment
- RCE-PSA: care for acute somatic pathology
- RCO: renewal of oppositional conduct

This procedure generated a total of 36 additional binary variables, which were incorporated into the main table, aligned by patient, date and shift. Its standardized format allows a systematic analysis and its use as predictors in supervised classification models.

5.2.3. Merging clinical sources

Once both sources of information (Geslin and RutiSafeNet) were preprocessed and codified, they were integrated into a single consolidated structure. The objective of this fusion was to unify, in turn, the information recorded retrospectively on risk behaviors (from Geslin), with the dynamic risk factors and the healthcare burden situations collected in real time (from RutiSafeNet), together with the stable socio-demographic variables.

To carry out this integration, a composite key formed by three columns was used: NHC (patient identifier), Date and Shift. These three variables uniquely define each unit of analysis, that is, each individual observation of a patient in a given shift.

Before performing the fusion, the formats of the keys in both sets of data were normalized. The patient identifiers (NHC) were converted to text, eliminating possible spaces or inconsistent symbols. The date column will be standardized to the datetime format, and the shift values will be limited to valid categories (M, T, N).

The fusion operation was implemented as a left join, using as a base table the set structured from Geslin (which includes risk behaviors and socio-demographic variables). In this way, it was guaranteed that all the records of the base system would be maintained, adding the columns from RutiSafeNet when there was an exact match in the three keys. In the cases where no information from RutiSafeNet was found for that patient, date and shift, the variables corresponding to the dynamic risk factors and care burden were imputed with zeros or null values, according to the type of coding.

Additionally, a specific case was considered: the presence of patients registered only in RutiSafeNet but absent in Geslin. To avoid the loss of information, these patients and unrepresented shifts will be identified, and additional records will be generated with the available data. In these new records, the missing columns (such as risk behaviors or socio-demographic variables) will be filled with zeros or empty values, as appropriate. This strategy made it possible to capture events detected in real time even if they had not been documented at the end of the shift.

As a final result, a unified table was obtained in which each row represents a patient in a specific shift, with all the relevant variables for the analysis: observed behaviors, dynamic risk factors, care burden and clinical and demographic characteristics. This consolidated structure constitutes the analytical basis on which the following phases of the system will be carried out.

5.2.4. Longitudinal structuring of hospitalization

Although the system does not explicitly reconstruct the complete sequence of hospitalization per patient (from the date of admission to discharge), a temporal structuring phase was applied aimed at guaranteeing a homogeneous representation of the clinical shifts per day. The objective was to ensure that, for each day with records, the three care shifts (morning, afternoon and night) were represented, thus allowing a more consistent analysis of clinical risk over time.

This process was carried out based on the data structured by patient, date and shift. For each date with existing records, undocumented shifts will be identified, and additional rows will be generated to complete them. These completed turns were assigned default values ('NA') in the risk variables, indicating that no specific clinical information had been recorded for that interval.

It is important to note that the registration logic differs between the two main sources. In Geslin, each shift must be documented by regulation, regardless of whether relevant clinical events have occurred. Therefore, it is expected to have continuous records in turn. Instead, RutiSafeNet is a dynamic use tool that is activated only when there is a clinical event that justifies its use.

This means that if no significant clinical change or overload situation occurs in a given shift, nothing is recorded in the platform. Therefore, it is normal that there are absent shifts in RutiSafeNet and that not all patients are represented every day.

Once the matrix of dates and shifts was completed, the information of patients with risk detected in each time slot was added. For each combination of date and shift, three columns were constructed that list the identifiers (NHC) of the patients who will present the following risk categories:

- Aggressive: aggressive behaviors (verbal or physical),
- Self-Harm: self-harm or suicidal risk,
- Absconding: escapes or attempts to leave the therapeutic environment.

In cases where no patient was registered in a category, 'NA' was assigned. On this basis, two additional binary variables will be generated:

- Risk: takes value 1 if there is at least one patient with risk in that turn.
- No_risk: takes value 1 if no type of risk was detected.

This structure allows for a more robust daily representation by turns, even in the absence of explicit events, maintaining temporal coherence and facilitating longitudinal analysis aggregated at the level of the care unit.

5.2.5. Data coding

Once the data structure per patient, date and shift had been consolidated, the clinical variables were codified in order to adapt them to the requirements of the supervised models. For this, two parallel versions of the dataset will be generated: one with standard binary coding and another with weighted coding, based on the relative clinical impact of each factor.

- **Binary encoding**

The binary version of the dataset contains exclusively values 0 or 1 for each variable of clinical risk and care burden. A value of 1 indicates the presence of the corresponding factor in a given turn, while a 0 indicates its absence.

This coding was applied both to the risk behaviors extracted from Geslin and to the indicators derived from RutiSafeNet (obtained by searching for patterns in the Item and Care Burden fields). This representation is the basis on which all the previous stages of the system were structured and constitutes the main version for exploratory analyzes and supervised modeling tasks.

- **Weighted coding**

In parallel, a weighted version of the dataset was generated. In this case, each binary variable was replaced by a numerical value that reflects its relative weight based on the clinical severity or the level of alertness it represents. The weights were defined in collaboration with the clinical team of the unit, based on care and operational criteria previously used in similar projects.

The allocation of weightings is shown in the following table, where each clinical variable (risk indicator or care burden) is associated with a value between 1 and 2, depending on its estimated impact:

Table 4. Weighting assigned to each risk factor

Risk factor	Weight	Risk factor	Weight
VIA	2	VIPM	1
AMG	2	ALT	1
CAU	2	AOPAH	2
AMED	2	SRAMF	1
OPCC	1	PAP	2
MVE	2	HVA	2
DCAT	1	PIDT	1
NPMT	1	AFPI	2
DCMSI	1	PII	1
CCA	1	SPS	1
CHP	2	CHO	2
HVT	1	CA	1
DR	2	PH	1
PD-ABVD	2	PAH	1
PCA	2	AC	2
PTCA	2	PCM	1
RAC	1	RAI	1
RCE-PSA	1	RCO	1

The transformation logic consisted of going through each binarily coded column and, if the value was 1, replacing it with the corresponding weight; if the value was 0, it was kept. This procedure was applied only to the clinical variables derived from the text, not to the socio-demographic variables or to the objective variables.

Both versions of the dataset (binary and weighted) share the same structure: one row per patient and turn, sociodemographic variables replicated per observation, and predictor variables structured by columns. This duplicity makes it possible to evaluate the impact of the representation of the information on the performance of the models.

5.3. Final dataset structure

At the end of the integration, codification and temporal standardization phases, four different versions of the dataset were obtained, each with a structure oriented to a concrete analytical objective. Although they share a common basis of organization by patient, date and turn, they differ by the level of aggregation, the representation of risk factors and the way of coding the clinical variables.

Next, the four generated structures are described, together with a representative example of each one.

1. Risk behavior dataset (by shift)

This version is organized by date and shift, not by patient. For each day and time slot combination, the identifiers of patients who presented with one of the three risk categories are listed. In addition, a global binary coding is included indicating whether or not clinical risk was recorded during that shift.

Example:

Table 5. Risk behavior dataset (by shift)

Date	Shift	Aggressive	Self-Harm	Absconding	Risk	No_risk
2024-02-01	M	P00123, P00312	NA	NA	1	0
2024-02-01	T	NA	NA	NA	0	1
2024-02-01	N	NA	P00456	NA	1	0

- The Aggressive, Self-Harm, and Absconding columns contain lists of NHC identifiers.
- Risk = 1 indicates that at least one category has patients assigned; No_risk = 1 if no risk is present.

2. Binary risk factor dataset (by patient-shift)

This structure collects information by patient, shift, and day. Clinical variables derived from textual records are coded binary (0 = absence, 1 = presence). This version is the most straightforward for supervised models that work with binary inputs.

Example:

Table 6. Binary risk factor dataset (by patient-shift)

NHC	Date	Shift	VIA	AMG	AFPI	PCA	PCM	CHP
P00123	2024-02-01	M	1	0	1	1	0	1
P00123	2024-02-01	T	0	0	1	1	0	0
P00456	2024-02-03	N	0	1	0	1	1	0

- Columns such as VIA, AMG, and CHP indicate the presence (1) or absence (0) of each factor.

3. Weighted risk factor dataset (per patient/shift)

Structurally identical to the binary version, but with weighted values instead of binary ones. The assigned weights reflect the relative impact of each clinical indicator, defined by clinical consensus.

Example:

Table 7. Weighted risk factor dataset (per patient/shift)

NHC	Date	Shift	VIA	AMG	AFPI	PCA	PCM	CHP
P00123	2024-02-01	M	2	0	2	2	0	2
P00123	2024-02-01	T	0	0	2	2	0	0
P00456	2024-02-03	N	0	2	0	2	1	0

- Values of 2 indicate factors with greater clinical significance.
- The target variables and the structure by patient and shift are preserved.

4. Sociodemographic dataset (per single patient)

This version includes a single row per patient. The variables collected were extracted from the sociodemographic blocks of the Geslin form. They were used as fixed variables across all patient shifts.

Example:

Table 8. Sociodemographic dataset (per patient/shift)

NHC	Gender	Schizophrenia	>35 years	Involuntary	Suicide Risk	Risk to others
P00123	H	1	0	1	1	0
P00456	M	0	1	0	0	1

- All variables are coded as binary (1 = Yes, 0 = No).
- These variables are replicated across shift versions to maintain structural consistency.

6. Experiments and results

This section presents the exploratory analysis, the experiments conducted, and the main results obtained from the previously structured dataset. First, the different versions of the dataset, the available variables, and their overall distribution are described. The procedures applied to prepare the dataset for supervised modeling are then detailed, including correlation dimensionality reduction and variable analysis by risk type. Finally, the results obtained from the evaluated models are presented, along with a discussion of their performance and clinical relevance.

6.1. Dataset overview and variable description

The analysis is based on a longitudinal dataset structured by patient and shift, with a total of 1,825 observations corresponding to 390 unique patients. Each row represents a unique combination of patient, date, and time slot (morning, afternoon, or evening) and includes stable sociodemographic variables, dynamic clinical factors, workload indicators, and the risk coding observed at the end of the shift.

The available variables are grouped into four blocks: temporal identification data (NHC, Date, Shift), sociodemographic variables extracted from Geslin, clinical and care indicators coded from RutiSafeNet, and risk labels observed at the end of the shift. In total, the dataset contains 49 columns and covers 357 different dates within the study period.

The variables are grouped into the following categories:

- Basic identification variables:
 - NHC: unique patient identifier.
 - Date: day of observation.
 - Shift: corresponding time slot (M: morning, T: afternoon, N: evening).
- Sociodemographic variables (constant per patient)
 - Gender
 - Diagnosis of schizophrenia
 - Over 35 years old?
 - Involuntary Admission
 - Admission due to the presence of a suicide risk
 - Admission due to a risk of harming others

- Clinical risk indicators (25 variables): extracted from structured text using patterns (e.g., VIA, AMG, AFPI, CHP, etc.).
- Care burden indicators (11 variables): related to containment, accompaniment, hygiene, or care complexity (e.g., PH, PCM, RCO, etc.).
- Target variables (risk categories):
 - Aggressive, Self-Harm, Absconding, No_risk (binary, extracted from Geslin).

To contextualize the temporal density of the record, the monthly frequency of observations was analyzed (Figure 4). A reasonably constant distribution was observed across the months, without abrupt interruptions, confirming the continuity of the care record during the period covered. This guarantees an adequate time base for longitudinal or period-segmented analyses.

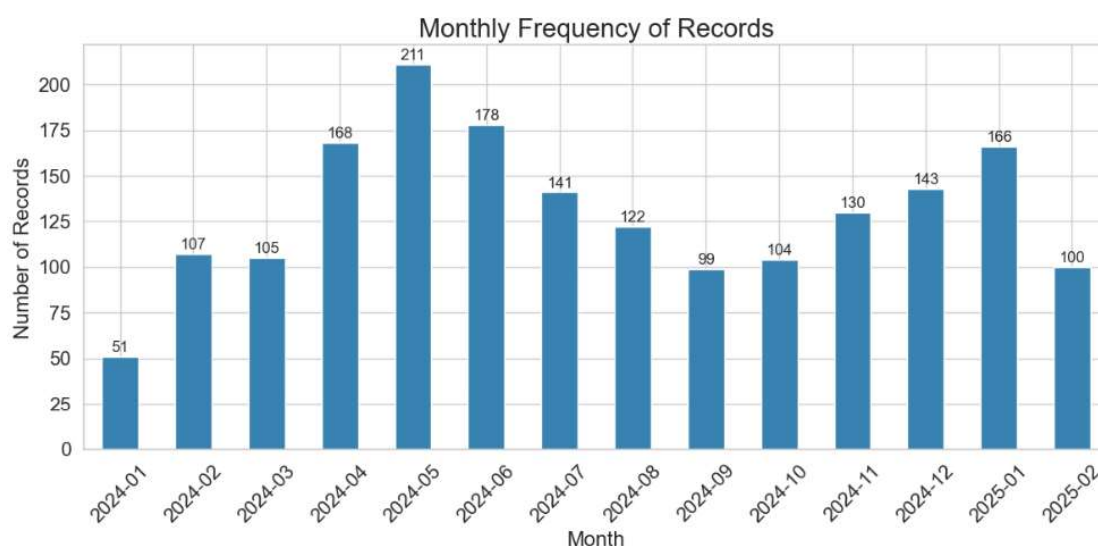


Figure 4. Monthly distribution of the total number of clinical observations recorded during the analysis period.

Regarding the clinical profile, the total activation frequency of risk indicators extracted from free text was examined (Figure 5). The most prevalent were PII (involuntary admission), followed by suicidal verbalizations (VIPM, VIA), behaviors demanding discharge (DCAT), and signs of suspicion (SPS). Indicators of aggression (CA, HVT), medication hoarding (AMED), and history of elopement (AFPI) also appeared. This distribution allows for the identification of the most recurrent foci in the unit, both from the perspective of autolytic risk and heteroaggression or elopement risk.

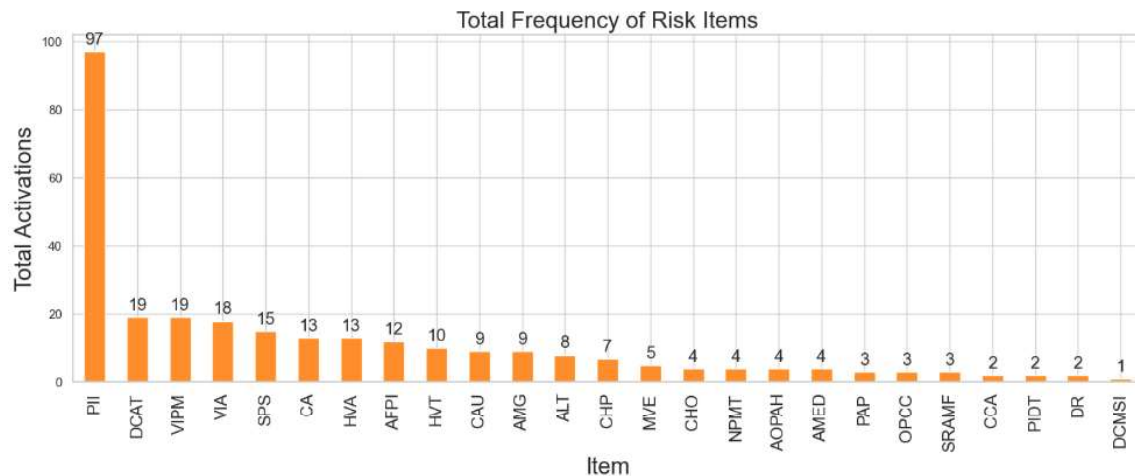


Figure 5. Total activation frequency of clinical risk indicators extracted from the free text in RutiSafeNet.

The frequency of indicators related to care burden was also analyzed (Figure 6). Particularly notable among these were cases requiring intermittent support (ISC), hygiene assistance (HAA), or high-demand patients (HDP). These factors do not necessarily represent risk behaviors, but they do imply a greater operational burden for the clinical team and, together, define the level of functional complexity of the admitted patients.

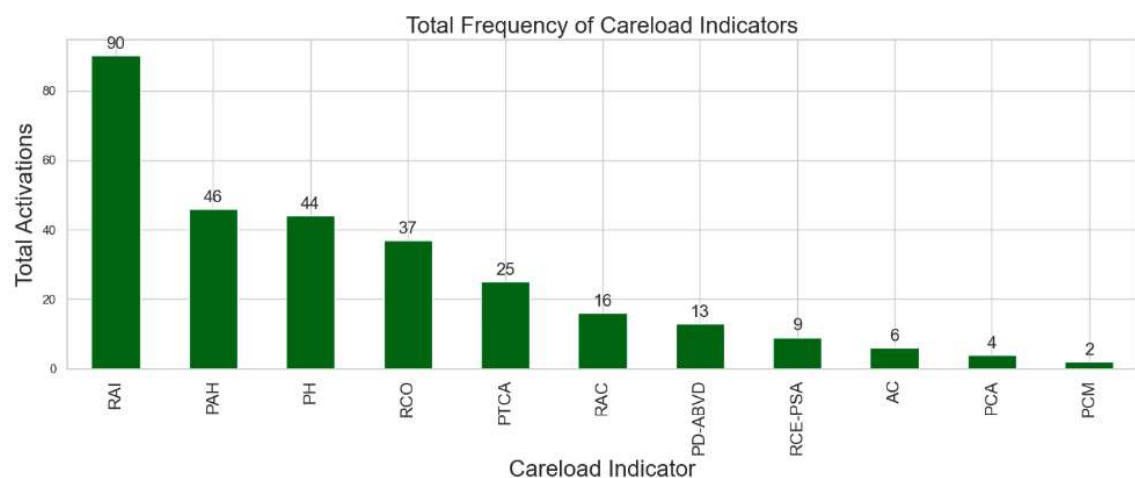


Figure 6. Activation frequency of care burden indicators, reflecting the functional complexity of the patients.

From a demographic perspective, the Gender variable was standardized to eliminate coding inconsistencies, and the analysis was restricted to valid values: H (male) and M (female). The resulting distribution (Figure 7) shows a balanced proportion, with a slight female predominance. The other sociodemographic variables—schizophrenia diagnosis, age over 35, involuntary admission, admission due to suicide risk, and admission due to risk of harming others—were binary-coded (Y/N) and analyzed at the individual patient

level. The results, presented together in Figure 7, indicate that a considerable portion of the sample consists of patients over 35 years of age, with a predominance of voluntary admissions and no diagnosis of schizophrenia.

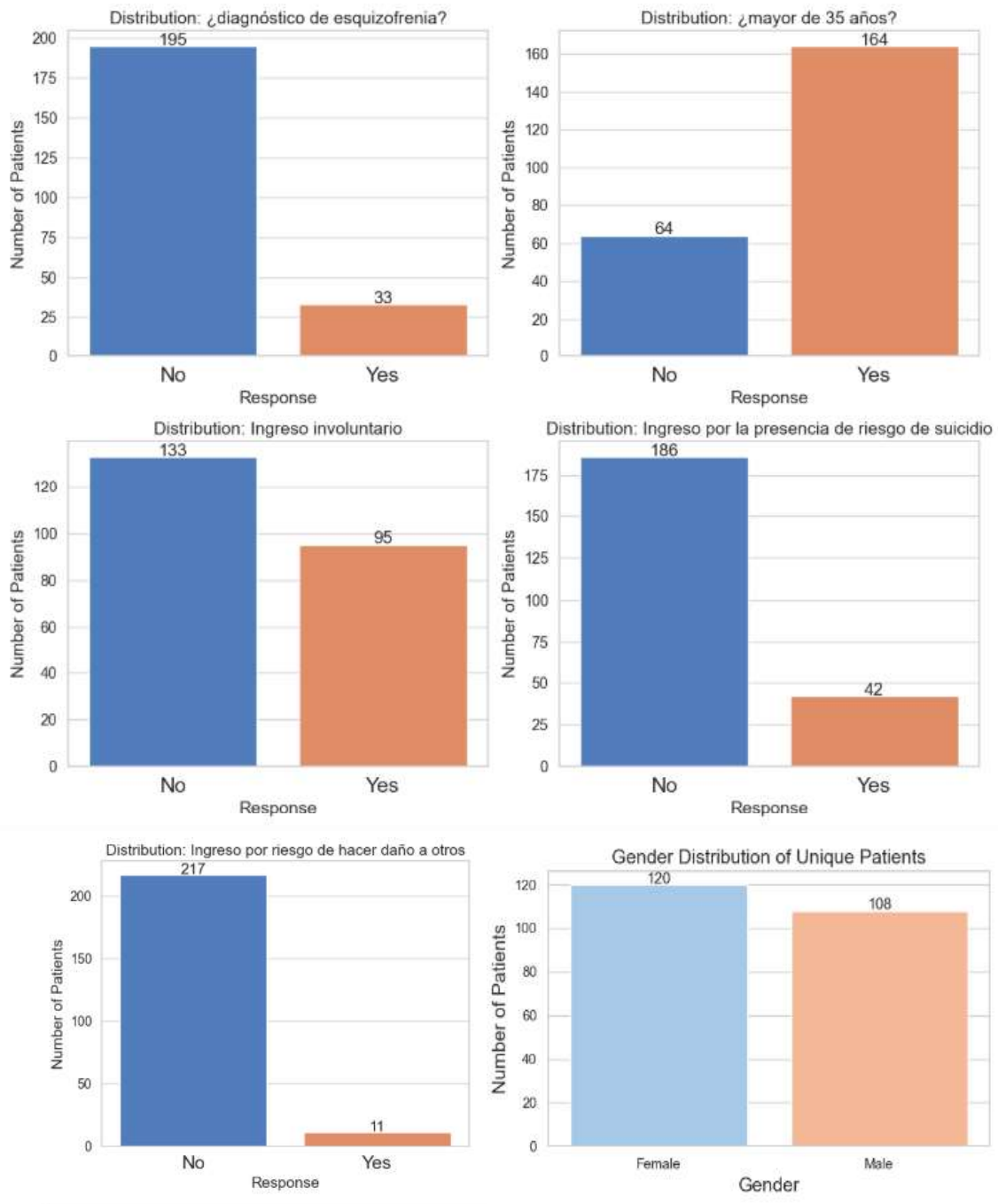


Figure 7. Distribution of key sociodemographic characteristics among unique patients included in the study.

Each subplot shows the frequency of patients by category for the following variables:

(A) Presence of a schizophrenia diagnosis,

(B) Age over 35 years,

(C) Involuntary admission status,

(D) Suicide risk as a reason for admission, and

(E) Risk of harming others as a reason for admission,

(F) Gender (Female, Male).

Categories are coded as "S" (Yes) and "N" (No), and reflect binary clinical or demographic attributes relevant to risk assessment and care planning.

Regarding the target variables, the overall distribution of the risk categories documented in Geslin was quantified. Eighty-six percent of appointments were recorded as risk-free, while the remaining 14% included some type of conflictive behavior. Among these, aggression (verbal or physical) was the most frequent pattern, followed by self-harm and behaviors such as running away from or abandoning the therapeutic setting. This distribution justifies the clinical interest in developing models capable of anticipating the emergence of this type of behavior, given its low frequency but high relevance (Figure 8).

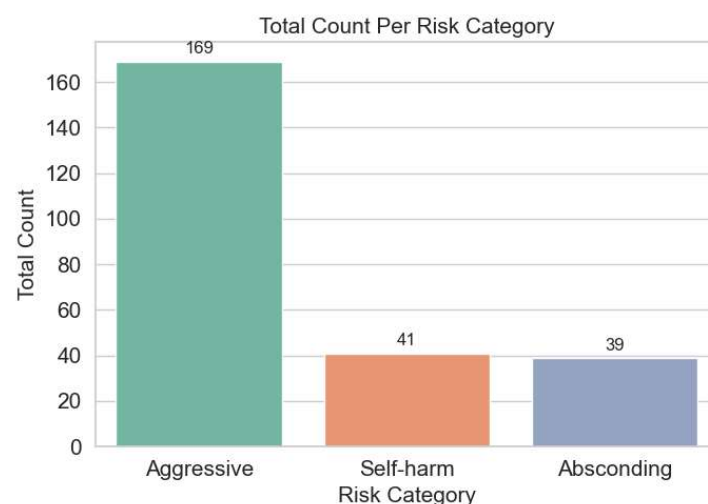


Figure 8. Total observations by risk category documented in Geslin: aggression, self-harm, absconding, and no risk.

When disaggregating the risk categories by gender, a differential distribution in the occurrence of conflict behaviors is observed (Figure 9). In general, men have a higher absolute frequency in the three main risk categories (Aggressive, Self-Harm, and Absconding), although this difference may be influenced by their slight overrepresentation in the total sample. This preliminary analysis allows us to explore whether there are differentiated risk patterns by demographic profile, which could be relevant in the development of stratified predictive models or in the clinical interpretation of the results.

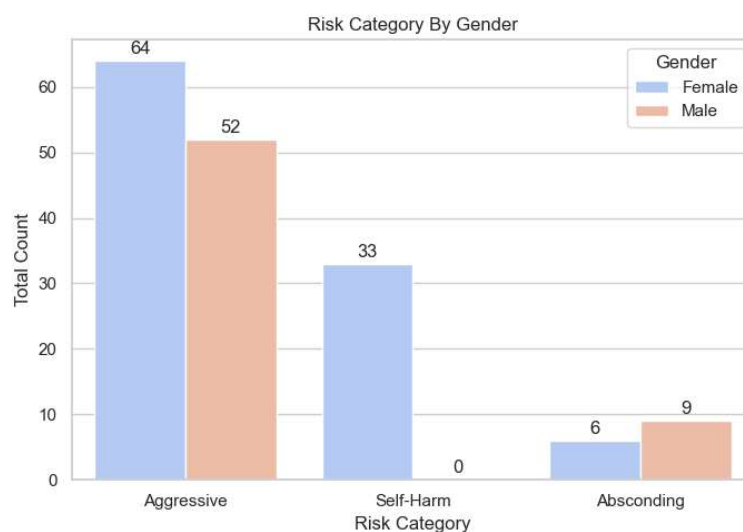


Figure 9. Distribution of the three main clinical risk categories by gender, in total number of observations per group.

The dataset includes complete temporal information and a diverse set of clinical and demographic variables. This structure allows for the analysis of both the emergence and progression of risk behaviors throughout hospitalization. It should be noted, however, that the main technical challenge of this study is the imbalance between the different risk classes: aggression (169 cases), self-harm (41 cases), and absconding (39 cases). The following sections examine these relationships and their clinical relevance in detail.

6.2. Exploratory data analysis

Following the general description of the variables, an exploratory analysis was performed to identify patterns of relationships between clinical factors, care burden indicators, and the observed risk categories. The objective was to understand the internal structure of the dataset, evaluate potential redundancies, and detect predictively relevant associations.

The exploratory analysis began with an examination of the overall correlation matrix between all clinical indicators (item_cols), care burden factors (workload_cols), and the four risk categories (Aggressive, Self-Harm, Absconding, No_risk). The matrix (Figure 10) shows the Pearson correlation between pairs of variables, coded by strength and sign. The strongest correlations are observed between certain pairs of indicators that tend to be activated together, such as AMG and CAU, or AOPAH and CHO, suggesting potential semantic redundancies or clinical overlap. Moderate positive correlations are also observed between some clinical variables and risk categories, especially between aggression indicators (CHP, CA, HVT) and the variable Aggressive.

To clarify the specific relationship between the indicators and the three main risk categories, a directed correlation submatrix was generated, focusing exclusively on Aggressive, Self-Harm, and Absconding as dependent variables (Figure 11). This more compact matrix allows for a clearer view of which factors are associated with each type of disruptive behavior. For example, HVT, CHP, and CA show significant correlations with Aggressive, while VIA, VIPM, and AMG show significant correlations with Self-Harm, and AFPI, PII, and DCAT are associated with Absconding. Although the correlations are not very high in absolute terms, they do reflect consistent trends that can inform the construction of composite variables or weights in subsequent models.

This analysis also identified clinical indicators with little or no correlation with any risk category, suggesting a low predictive value in isolation. Such variables could be candidates for exclusion in dimensionality reduction processes or regrouped into more robust clinical constructs.

Taking together, this exploratory analysis provides a first approximation of the dataset's structural relationships, allowing variables to be prioritized, redundancies to be detected, and predictor selection to be guided for the modeling phases.

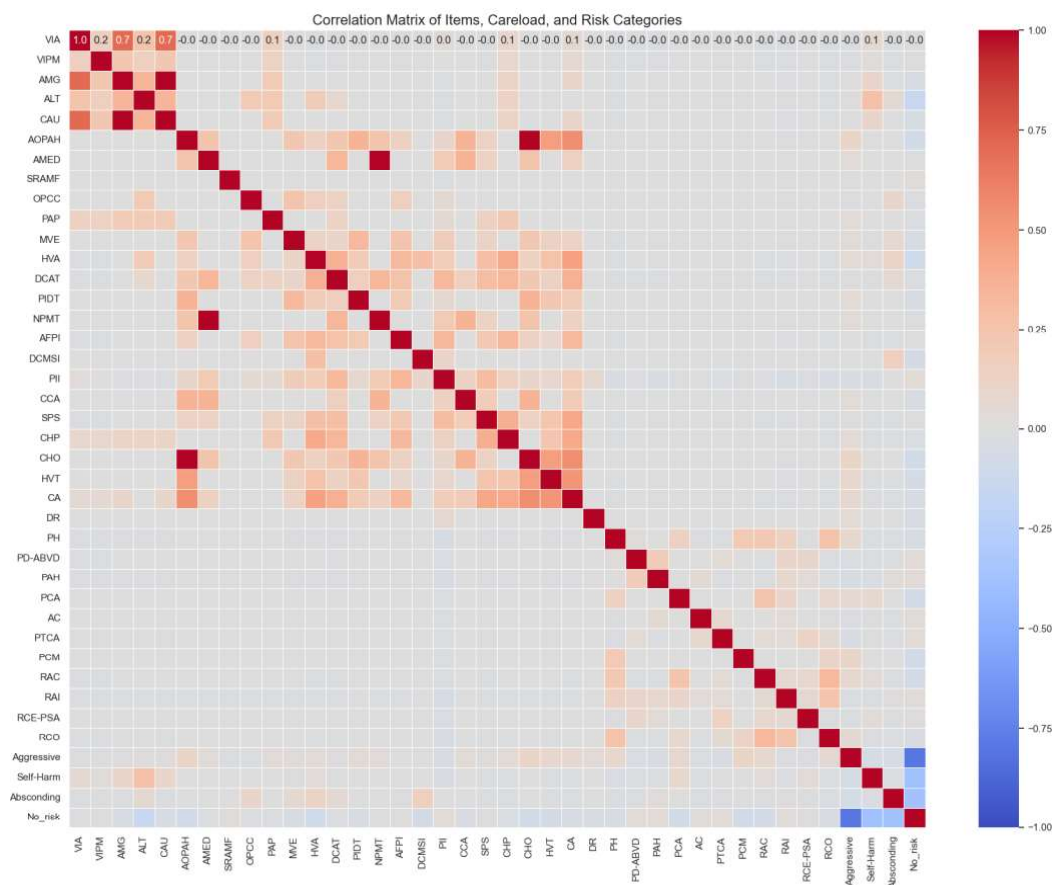


Figure 10. Complete correlation matrix between all clinical indicators, care burden factors, and risk categories.

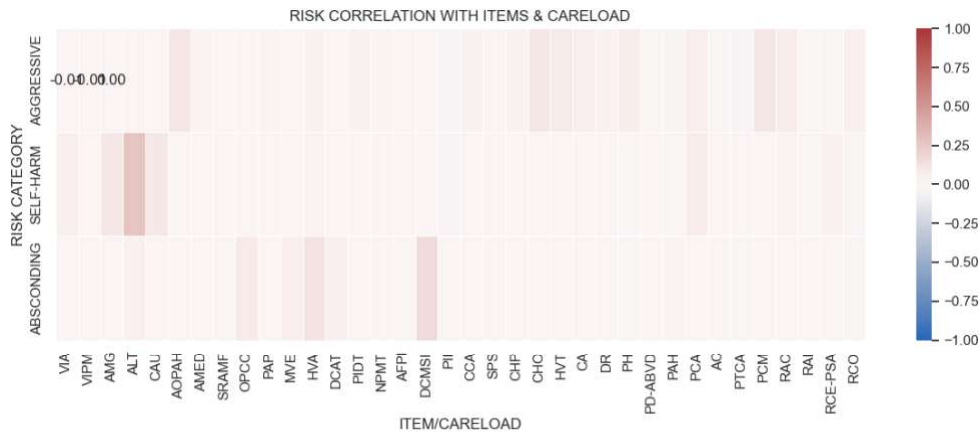


Figure 11. Correlation submatrix centered on the target variables Aggressive, Self-Harm, and Absconding, showing their relationship with the various clinical and care indicators.

6.3. Experimental setup

The objective of this phase was to evaluate the performance of different supervised classification algorithms in predicting risk behaviors observed at the end of each clinical shift. To this end, an experimental framework was defined that includes different variable coding strategies, different levels of temporal aggregation, and two classification approaches: multiclass and binary.

6.3.1. Algorithms Evaluated

To compare the prediction of risk behaviors observed at the end of each clinical shift, five widely used supervised learning models were selected, each representing a distinct family of classification algorithms. This diversity facilitates robust performance comparison and is consistent with standard practice in applied clinical machine learning research [16,26,33]. All models were trained and evaluated under identical preprocessing, imputation, and cross-validation strategies. Hyperparameters were set according to established recommendations in the literature and, when appropriate, adapted to the characteristics of the dataset.

1. Logistic Regression (Baseline):

Logistic regression is a classical linear classification algorithm commonly used as a baseline in predictive modeling. L2 regularization was employed to prevent overfitting, and class weights were balanced to address class imbalance (`penalty='l2'`, `class_weight='balanced'`). The model was fitted using the 'lbfgs' solver with a maximum of 1000 iterations. For multiclass classification, the multinomial variant was applied (`multi_class='multinomial'`) [34].

2. Random Forest:

Random Forest is an ensemble method based on the aggregation of multiple decision trees, which improves predictive accuracy and reduces overfitting. The configuration included 100 trees (`n_estimators=100`) without depth restriction (`max_depth=None`) and class weighting (`class_weight='balanced'`). Feature importance was assessed using Gini coefficients. Both binary and multiclass configurations were evaluated [35].

3. XGBoost:

XGBoost is a gradient boosting algorithm that constructs decision trees sequentially to minimize classification errors. For the binary classification, the logloss objective was used, with `scale_pos_weight=1.87` to address class imbalance. The configuration included 100 trees (`n_estimators=100`), a maximum tree depth of 4 (`max_depth=4`), and a learning rate of 0.1 (`learning_rate=0.1`). For multiclass classification, the 'multi:softmax' objective was used with the same hyperparameters. Hyperparameter choices were based on common usage in the literature [36].

4. Deep Neural Network (MLP):

A multilayer perceptron (MLP) was implemented in Keras with two hidden layers (128 and 64 neurons, ReLU activation), each followed by batch normalization and dropout (rate 0.3). The output layer used sigmoid activation for binary and softmax for multiclass tasks. Models were trained with the Adam optimizer (learning rate 0.001), batch size 32, up to 8 epochs, early stopping, and a 10% validation split. Model performance was evaluated using leave-one-patient-out cross-validation [37]. The architecture was selected to balance learning capacity with regularization, as recommended for small to moderate clinical datasets.

The selection of algorithms was motivated by their methodological diversity (linear models, ensembles, boosting, and neural networks), their documented performance in the context of class imbalance, and their adaptability to both binary and multiclass classification.

6.3.2. Experimental Structure

An experimental structure was defined to evaluate how different combinations of variable coding (binary or weighted), levels of temporal aggregation (by shift or by day), and classification approaches (binary and multiclass) affect predictive performance. For each experiment, multiple ML models, including logistic regression (baseline), random forest, XGBoost, and neural network, were trained and evaluated to compare how the coding of variables and the aggregation level influence prediction accuracy and variability.

- **Variable coding:**

The predictor variables were processed in two alternative ways. On one hand, binary coding was used, where each clinical indicator was represented as 0 (absence) or 1 (presence). On the other hand, a weighted coding was applied, assigning values of 0, 1, or 2 based on the estimated clinical impact or severity, according to expert criteria. The performance of all models was systematically compared using both coding schemes, allowing us to determine whether the choice of coding method led to significant differences in model behavior.

- **Type of classification:**

Two classification tasks were considered:

- Multiclass: Prediction of the specific category of risk behavior manifested at the end of the shift (aggression, self-harm, absconding, or no risk).
- Binary: Prediction of the presence or absence of any risky behavior, regardless of type.

- **Temporal analysis unit:**

To assess the effect of temporal granularity, two levels were evaluated:

- Clinical shift: Each observation corresponds to a unique combination of patient, day, and shift, capturing intra-daily variations and facilitating real-time prediction.
- Full day: The data from all shifts of the same day were grouped to obtain a daily temporal resolution, allowing for performance comparison against shift-by-shift analysis.

The systematic combination of classification type (binary or multiclass), temporal unit (shift or day), and variable coding (binary or weighted) resulted in 8 main experimental configurations. This factorial design enabled the evaluation of both the individual and combined effects of each methodological choice on model performance, providing a comprehensive assessment of the best predictive strategies for acute psychiatric risk behaviors.

Below is a table of the combinations made (Table 9):

Table 9. Summary of experimental configurations

No.	Unit of analysis	Classification type	Datasets used	Models evaluated
1	Shift	Multiclass (4 categories)	Binary, Weighted	Logistic Regression, Random Forest, KNN, SVM, XGBoost, Deep Learning
2	Shift	Binary (Risk / No Risk)	Binary, Weighted	Logistic Regression, Random Forest, KNN, SVM, XGBoost, Deep Learning
3	Day	Binary	Binary, Weighted	Logistic Regression, Random Forest, KNN, SVM, XGBoost, Deep Learning

No.	Unit of analysis	Classification type	Datasets used	Models evaluated
4	Day	Multiclass	Binary, Weighted	Logistic Regression, Random Forest, KNN, SVM, XGBoost, Deep Learning

To complement the tabular overview, the following diagram (Figure 12) provides a visual synthesis of the experimental setup, highlighting how classification types, temporal units, and variable encodings intersect with the selected models.

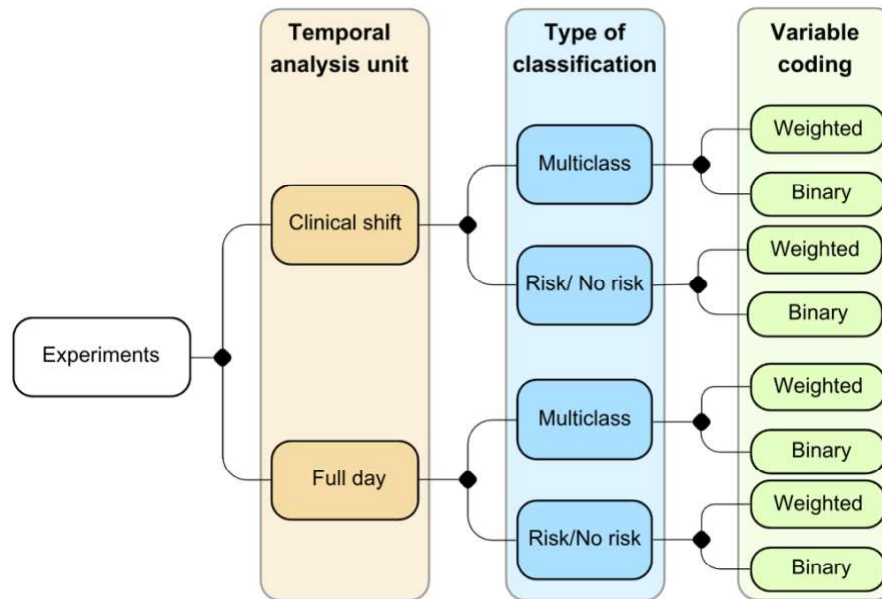


Figure 12. Visual representation of the experimental configurations: combinations of analysis unit, classification type, and input features.

6.3.3. Data Preparation and Partitioning Pipeline

The process of data preparation and partitioning followed a structured sequence in seven stages, aimed at ensuring rigorous and consistent model training and validation.

- 1- First, clinical and care burden indicators were loaded, available in both binary (0/1) and weighted (0/1/2) versions, forming the basis of the predictor variables.
- 2- Fixed sociodemographic variables were then integrated using the unique patient code (NHC), so that each record combined clinical information with the individual patient context.
- 3- The target or output variable was defined according to the experiment type. For binary classification, a variable was generated to encode the presence or absence of risk ("Risk" vs. "No Risk"). For multiclass classification, each record was labeled as "Aggressiveness," "Self-harm," "Absconding," or "No Risk," based on observed behavior.

- 4- All non-predictor columns, such as identifiers, dates, and any output labels, were then removed to prevent information leakage during model training.
- 5- All categorical variables were transformed using one-hot encoding, and numerical variables were standardized with StandardScaler, ensuring that all variables had a comparable scale-a crucial step for algorithms sensitive to data magnitude.
- 6- In cases with missing values, mean imputation was performed using SimpleImputer to maintain the integrity and completeness of the data matrix.
- 7- Finally, to robustly validate the generalization capability of the algorithms, a leave-one-group-out cross-validation (LOPOCV) strategy was implemented. In this approach, the model is trained by leaving out all the data from one group and then evaluated on that excluded group. This simulates real scenarios where the goal is to predict risk in patients or time periods not previously seen by the model.

The data preparation and partitioning pipeline is the same regardless of whether the validation is performed per patient (shift) or per day. The only difference is the group used for the leave-one-group-out cross-validation strategy: in one case it is grouped by patient, and in the other, by date.

Figure 13 shows the pipeline followed for data processing, feature engineering, and model validation using leave-one-group-out cross-validation.

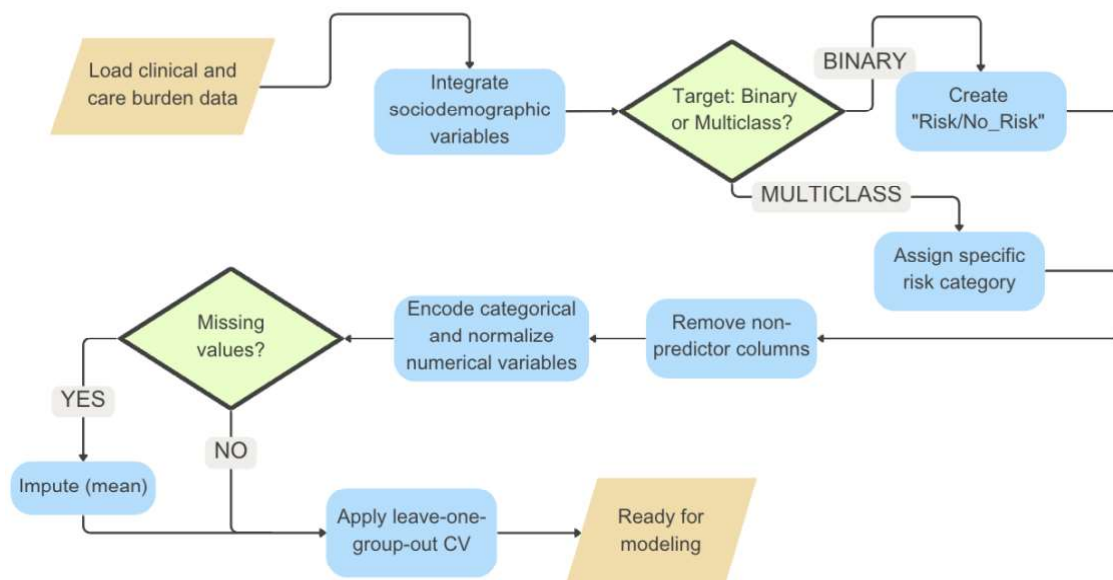


Figure 13. Workflow for data preprocessing, variable encoding, and model validation using leave-one-group-out cross-validation in all experimental configurations.

6.4. Results

This section presented a comparative summary of the results obtained with the different classification models across all experimental scenarios. Key performance metrics (including accuracy, precision, recall, and F1-score) were reported for each model and configuration, grouped by classification type (binary or multiclass), unit of analysis (shift or day), and variable coding method (binary or weighted). For binary classification tasks, receiver operating characteristic (ROC) curves and area under the curve (AUC) values were also provided, allowing for direct comparison of discrimination performance between models.

Where appropriate, confidence intervals for performance metrics and statistical tests for the comparison of model results were included to assess the significance of observed differences. The following sections detailed the results for each classification task and highlighted the main findings from these experiments.

The most essential results have been included in this section to facilitate direct interpretation and comparison. All complementary tables, figures, and detailed outputs are provided in the Annexes, where the full set of metrics, SHAP plots, Grid-Search analysis, and additional model explainability analyses can be consulted for reference.

6.4.1. Comparative Model Performance

The comparative results of all evaluated models are summarized below in two separate tables. Table 10 presents the performance metrics for binary risk prediction (risk/no risk), while Table 11 summarizes the results for multiclass risk prediction (aggressiveness, self-harm, absconding, no risk). Each table covers all experimental configurations: binary and weighted features, cross-validation by patient (shift) and by day, and all four classification algorithms.

For each experiment and metric (accuracy, precision, recall, F1-score, and AUC for binary classification), the best value among the models is highlighted in green. This allows for easy identification of the optimal approach for each scenario.

The highest accuracy metric among all models was achieved by XGBoost, when grouping by day, with both the binary and weighted feature codings yielding the same value (Accuracy = 0.761). In any case, the differences between the accuracy values of the various models are minimal, which is why in the following section, additional statistical metrics such as confidence intervals will be explored to further assess these differences.

In the multiclass prediction task, the same pattern is observed: XGBoost, grouped by day, achieved the highest accuracy (Accuracy = 0.727) for both binary and weighted feature codings.

It is also observed that the binary classification setup achieves higher accuracy values compared to the multiclass setup. However, this is to be expected, as multiclass classification involves a greater number of classes, and thus overall results may be comparable in relative terms.

Table 10. Comparative performance of all models for binary risk prediction (best values in green).

Model	Grouping	Dataset	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	Turn	Binary	0.653	0.502	0.459	0.480	0.611
Logistic Regression	Turn	Weighted	0.653	0.502	0.459	0.480	0.611
Logistic Regression	Day	Binary	0.706	0.578	0.586	0.582	0.679
Logistic Regression	Day	Weighted	0.706	0.578	0.586	0.582	0.679
Random Forest	Turn	Binary	0.669	0.533	0.396	0.457	0.658
Random Forest	Turn	Weighted	0.669	0.533	0.396	0.457	0.658
Random Forest	Day	Binary	0.757	0.674	0.586	0.627	0.747
Random Forest	Day	Weighted	0.757	0.674	0.586	0.627	0.747
XGBoost	Turn	Binary	0.680	0.551	0.437	0.487	0.687
XGBoost	Turn	Weighted	0.680	0.551	0.437	0.487	0.687
XGBoost	Day	Binary	0.761	0.667	0.631	0.648	0.766
XGBoost	Day	Weighted	0.761	0.667	0.631	0.648	0.766
Neural Net	Turn	Binary	0.611	0.367	0.162	0.225	0.591
Neural Net	Turn	Weighted	0.644	0.478	0.248	0.319	0.646
Neural Net	Day	Binary	0.711	0.628	0.419	0.503	0.736
Neural Net	Day	Weighted	0.721	0.649	0.432	0.519	0.726

Table 11. Comparative performance of all models for multiclass risk prediction (best values in green).

Model	Grouping	Dataset	Accuracy	Precision	Recall	F1-score
Logistic Regression	Turn	Binary	0.386	0.606	0.386	0.449
Logistic Regression	Turn	Weighted	0.381	0.601	0.381	0.444
Logistic Regression	Day	Binary	0.487	0.741	0.487	0.544
Logistic Regression	Day	Weighted	0.480	0.738	0.480	0.538
Random Forest	Turn	Binary	0.504	0.698	0.504	0.565

Model	Grouping	Dataset	Accuracy	Precision	Recall	F1-score
Random Forest	Turn	Weighted	0.504	0.698	0.504	0.565
Random Forest	Day	Binary	0.562	0.747	0.562	0.627
Random Forest	Day	Weighted	0.562	0.747	0.562	0.627
XGBoost	Turn	Binary	0.626	0.559	0.626	0.589
XGBoost	Turn	Weighted	0.626	0.559	0.626	0.589
XGBoost	Day	Binary	0.727	0.687	0.727	0.702
XGBoost	Day	Weighted	0.727	0.687	0.727	0.702
Neural Net	Turn	Binary	0.641	0.526	0.641	0.543
Neural Net	Turn	Weighted	0.633	0.515	0.633	0.544
Neural Net	Day	Binary	0.669	0.587	0.669	0.591
Neural Net	Day	Weighted	0.673	0.588	0.673	0.599

6.4.2. Results by Classification Task

This section presents the main results obtained for each classification task. The performance of all evaluated models is reported separately for binary risk prediction and multiclass risk prediction, under all experimental configurations.

6.4.2.1. Binary Risk Prediction

As presented in Table 10 (see Section 6.4.1), the main performance metrics for all binary risk prediction experiments were compared across all models and scenarios. Based on these results, the best-performing models were selected according to their highest AUC and F1-score values.

Subsequent analyses in this section (including the ROC curves, confusion matrices, and statistical tests) are focused on these top-performing models.

Figure 14 presents the ROC curves for the top three models (logistic regression, random forest, and XGBoost) in the binary risk prediction task, grouped by shift (turn). The best AUC values observed were 0.61 for logistic regression, 0.67 for random forest, and 0.64 for XGBoost. Among these, random forest achieved the highest discriminative performance in this grouping, as indicated by the larger area under its ROC curve.

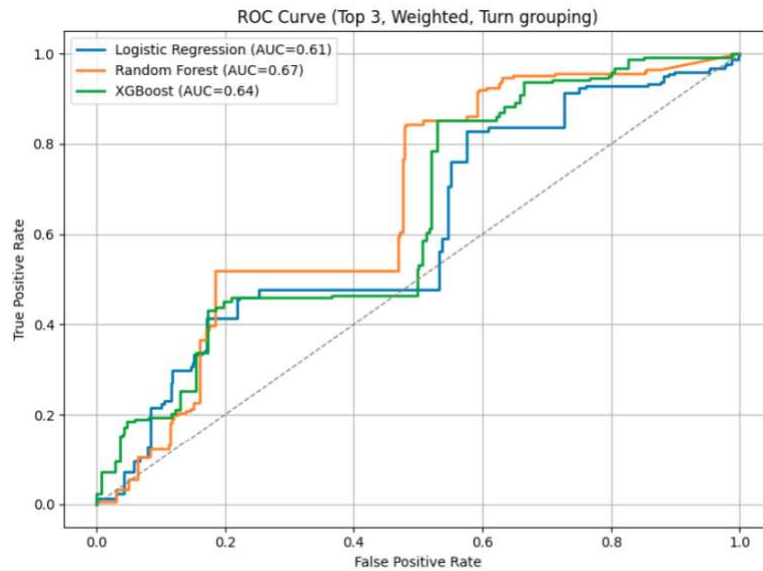


Figure 14. ROC curves for the top 3 models, binary risk prediction by shift (turn)

Figure 15 displays the ROC curves for the same models when grouping by day. In this case, the best AUC values were higher: 0.68 for logistic regression, 0.76 for random forest, and 0.77 for XGBoost. Notably, both ensemble models outperformed logistic regression, with XGBoost obtaining the highest AUC overall in this scenario.

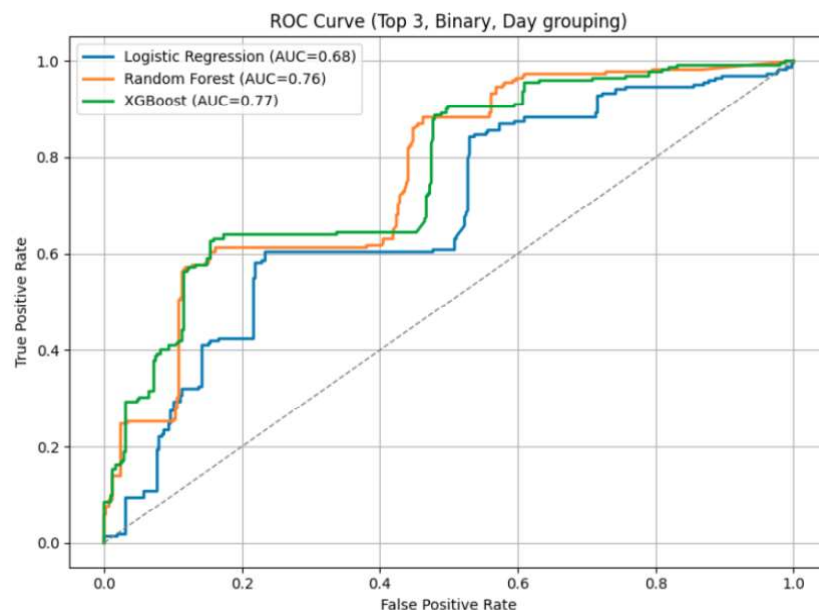


Figure 15. ROC curves for the top 3 models, binary risk prediction by day

These plots highlight the improvement in model performance when grouping by day rather than by shift, with ensemble methods showing particular benefit in this setting.

Figure 16 presents the confusion matrices for the top-performing models in binary risk prediction: Random Forest (by shift) and XGBoost (by day), using binary feature coding. For Random Forest grouped by shift, the maximum number of correct predictions was observed for the "No Risk" class (338 true negatives). For XGBoost grouped by day, the maximum correct classification was also for "No Risk" (345 true negatives). The corresponding results with weighted feature coding were identical, and are therefore not shown.

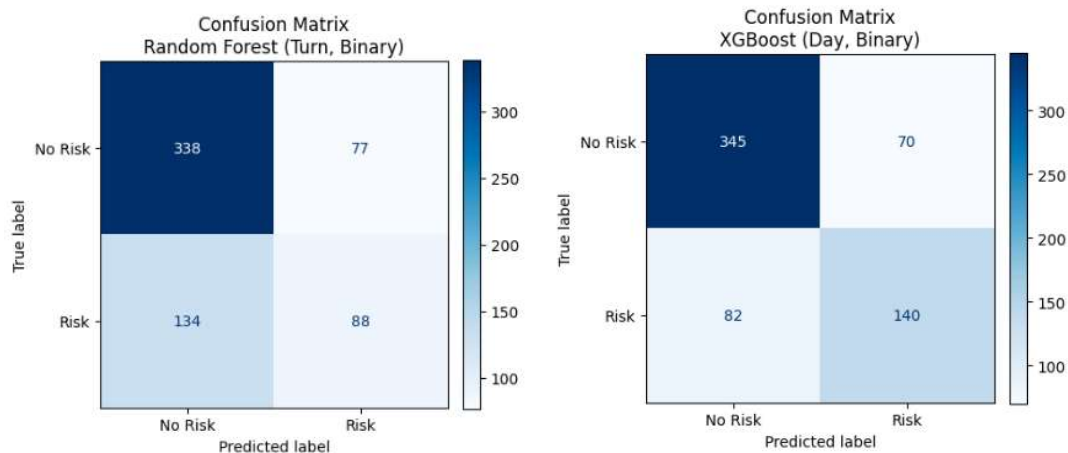


Figure 16. Confusion matrices for the top binary models (Random Forest by shift, XGBoost by day) using binary features.

Table 12 shows the results of the pairwise McNemar test applied to all binary classification scenarios. This statistical test was used to compare the predictions of each pair of models for each scenario (by shift and by day, with both binary and weighted features). The table reports the p-value obtained for each comparison, as well as whether the difference is statistically significant ($p < 0.05$).

Significant differences were found between several model pairs, especially in the groupings by day. For example, in the "Day/Binary" scenario, the comparison between logistic regression and both ensemble models (random forest and XGBoost) resulted in p-values < 0.001 , indicating statistically significant differences in their predictions. In contrast, some model comparisons, such as random forest vs. XGBoost in both binary and weighted feature settings, did not show significant differences.

Table 12. Pairwise McNemar test results for binary classification scenarios

Scenario	Model 1	Model 2	McNemar p-value	Significant ($p < 0.05$)
Turn/Binary	Logistic Regression	Random Forest	0.3374	No
Turn/Binary	Logistic Regression	XGBoost	0.0396	Yes

Scenario	Model 1	Model 2	McNemar p-value	Significant (p < 0.05)
Turn/Binary	Logistic Regression	Neural Net	0.0207	Yes
Turn/Binary	Random Forest	XGBoost	0.3817	No
Turn/Binary	Random Forest	Neural Net	0.0004	Yes
Turn/Binary	XGBoost	Neural Net	0.0000	Yes
Turn/Weighted	Logistic Regression	Random Forest	0.3374	No
Turn/Weighted	Logistic Regression	XGBoost	0.0396	Yes
Turn/Weighted	Logistic Regression	Neural Net	0.6427	No
Turn/Weighted	Random Forest	XGBoost	0.3817	No
Turn/Weighted	Random Forest	Neural Net	0.1214	No
Turn/Weighted	XGBoost	Neural Net	0.0297	Yes
Day/Binary	Logistic Regression	Random Forest	0.0002	Yes
Day/Binary	Logistic Regression	XGBoost	0.0000	Yes
Day/Binary	Logistic Regression	Neural Net	0.8408	No
Day/Binary	Random Forest	XGBoost	0.6900	No
Day/Binary	Random Forest	Neural Net	0.0038	Yes
Day/Binary	XGBoost	Neural Net	0.0018	Yes
Day/Weighted	Logistic Regression	Random Forest	0.0002	Yes
Day/Weighted	Logistic Regression	XGBoost	0.0000	Yes
Day/Weighted	Logistic Regression	Neural Net	0.4069	No
Day/Weighted	Random Forest	XGBoost	0.6900	No
Day/Weighted	Random Forest	Neural Net	0.0178	Yes
Day/Weighted	XGBoost	Neural Net	0.0112	Yes

The weighted feature scenarios generally produced similar significance patterns to the binary feature scenarios, and most model comparisons showed consistent results across both coding methods. The statistical results for each pairwise comparison largely held regardless of whether binary or weighted features were used.

6.4.2.2. Multiclass Risk Prediction

For multiclass classification, ROC curves and AUC values are not included, as these metrics are designed for binary problems and do not extend directly to multiclass scenarios without additional transformations. Instead, model performance is summarized using confusion matrices, which clearly show how predictions are distributed across all classes.

Figure 17 presents the confusion matrices for the neural network model in multiclass risk prediction. Results are shown separately for the binary and weighted feature codings, as performance differs between these two settings. These observed differences will be further examined to determine whether they are statistically significant.

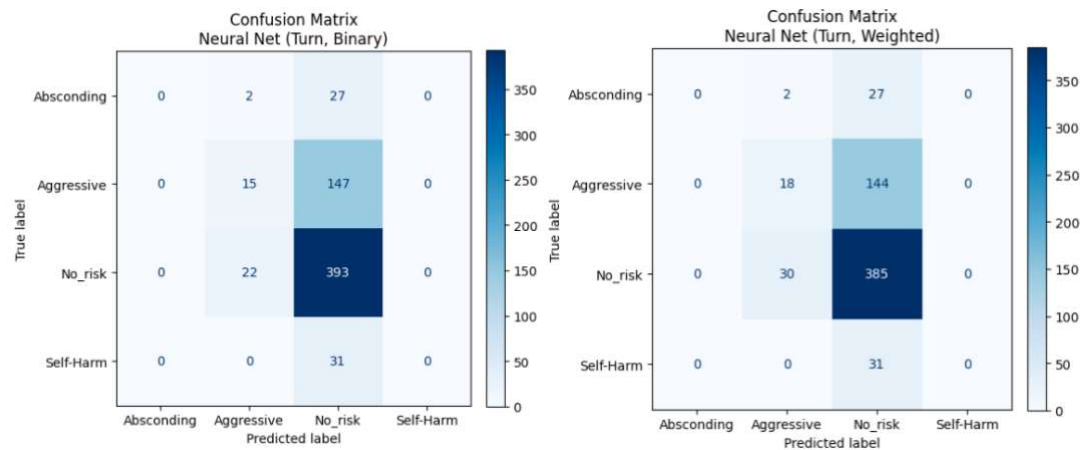


Figure 17. Confusion matrices for neural network (binary and weighted features) in multiclass risk prediction

Figure 18 displays the confusion matrix for the XGBoost model in multiclass risk prediction. In this case, both binary and weighted feature codings yielded identical results, so only a single confusion matrix is provided.

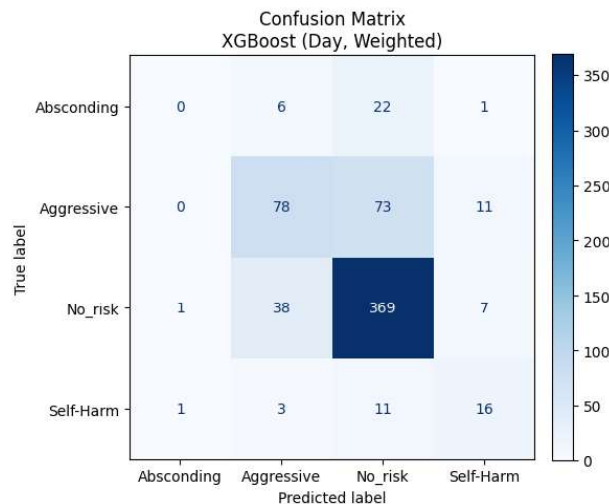


Figure 18. Confusion matrix for XGBoost (binary or weighted features) in multiclass risk prediction

Table 13 shows the results of the Wilcoxon signed-rank test, conducted to statistically compare the accuracy of the two best-performing models in each multiclass risk prediction scenario. For grouping by shift (turn), differences between neural network and XGBoost

were not statistically significant for either binary or weighted feature codings ($p > 0.05$). In contrast, when grouping by day, the accuracy of XGBoost was significantly higher than that of the neural network for both binary and weighted features ($p < 0.05$).

Scenario	Best Model	Second Best Model	Accuracy 1	Accuracy 2	Wilcoxon p-value	Significant ($p < 0.05$)
Turn/Binary	Neural Net	XGBoost	0.641	0.626	0.1269	No
Turn/Weighted	Neural Net	XGBoost	0.633	0.626	0.6042	No
Day/Binary	XGBoost	Neural Net	0.727	0.669	0.0026	Yes
Day/Weighted	XGBoost	Neural Net	0.727	0.673	0.0137	Yes

Table 13. Wilcoxon signed-rank test results comparing top multiclass models per scenario.

6.4.3. Variable Importance and Model Explainability

The SHAP values were used to identify which variables were most important in each model and experiment carried out. This allowed us to clearly see which indicators or variables had the greatest weight in predicting the risk. All related figures and detailed plots are provided in the appendix; here, only the main observed results are summarized to support the understanding of the document.

In the logistic regression, for the binary classification in turn and with binary coding, the variables that most contributed to the prediction were the diagnosis of schizophrenia (present and absent), the male gender and some risk items such as PPII and PTCA. When weighted variables were used, the diagnosis of schizophrenia and gender were again among the most relevant, along with the PTCA item.

In the multiclass analysis, the important variables were similar. For the absconding class, they highlighted the male gender, the absence of suicide risk and several risk items. In the aggressive and non-risk classes, the sociodemographic indicators also had a lot of weight, and in self-harm the item PAH appeared as a relevant variable. This pattern was repeated with both binary and weighted variables.

In the binary random forest model in turn, the sociodemographic variables and the risk items ALT, PII, VIA and RAI were the most influential. In multiclass, PAH stood out as the main predictor for the absconding class, while in the aggressive and self-harm classes the risk items and sociodemographic data continued to be relevant. The results were almost identical using the weighted variables.

In the neural network (binary, in turn) the same patterns were also repeated: the sociodemographic variables and the PAH risk item were the most influential in the prediction.

All the SHAP graphs that show these results are included in the Annexes. These graphics were added to give transparency to the models and to help visualize in a simple way which variables will carry the most weight in the model's decisions.

An adaptive table (Table 14) was constructed to clearly display the most influential features for prediction across the different models. As observed, sociodemographic variables consistently appear as the most important predictors, followed by the specific clinical risk and care burden indicators.

Rank	Feature Name	Type	Description
1	Schizophrenia Diagnosis (Yes)	Sociodemographic	Diagnosis of schizophrenia
2	Schizophrenia Diagnosis (No)	Sociodemographic	No diagnosis of schizophrenia
3	Gender (Male)	Sociodemographic	Patient gender: male
4	Gender (Female)	Sociodemographic	Patient gender: female
5	Age > 35 years (Yes)	Sociodemographic	Age above 35
6	PPII	Clinical risk indicator	Involuntary psychiatric admission
7	PAH	Care burden indicator	Need for assistance in hygiene
8	PTCA	Care burden indicator	Presence of eating disorder
9	ALT	Clinical risk indicator	Minor self-harm
10	VIA	Clinical risk indicator	Verbalization of autolytic ideation

Table 14. Top 10 Most Influential Features (All Models, All temporal grouping)

6.4.4. Main Findings

The comparative analysis of the different machine learning models demonstrated that all algorithms achieved similar performance metrics across the various experimental setups. XGBoost consistently achieved the highest accuracy in both binary and multiclass classification tasks when grouping by day, with identical results observed for binary and weighted feature codings. However, the differences in performance between the models were generally small, and statistical significance was only confirmed in some scenarios, as demonstrated by the results of the McNemar and Wilcoxon tests.

Binary classification models achieved higher accuracy than multiclass models, as expected given the increased complexity of multiclass prediction. For the binary risk prediction task, ensemble models (Random Forest and XGBoost) outperformed logistic regression and neural network models in terms of AUC and overall accuracy, particularly when using day-level aggregation. In multiclass prediction, both XGBoost and neural networks showed the best results, although the differences between these models were not always statistically significant.

Model explainability analyses using SHAP values identified a consistent set of key predictors across all models. Sociodemographic features (such as schizophrenia diagnosis, gender, and age) were the most influential in the prediction of risk, followed by specific clinical risk and care burden indicators (e.g., PAH, PPII, PTCA, ALT, and VIA). These patterns were robust to changes in feature coding (binary vs. weighted) and the classification algorithm used.

A grid search procedure was performed to optimize the hyperparameters of the XGBoost models in both binary and multiclass tasks (Results in Annexes). This process enabled the identification of parameter settings that led to the best F1-score and accuracy for each configuration. The results confirmed that even with hyperparameter optimization, the relative performance rankings between models and scenarios remained stable, reinforcing the robustness of the main findings.

The main findings indicate that ensemble models, particularly XGBoost, provided the best performance overall, especially when the data were aggregated by day. However, all models demonstrated comparable results, and the most important variables for risk prediction remained stable across all algorithms and experimental conditions.

7. Discussion

In this section, the results presented above are interpreted in depth, contrasting the findings with the literature, discussing the clinical and methodological implications, and pointing out in detail the limitations detected. It reflects critically on the validity, usefulness and possible biases of the models, as well as the areas where the work could be expanded or improved.

One of the first aspects that deserves to be highlighted is the relative homogeneity of the results between the different classification algorithms evaluated, especially in terms of metrics such as accuracy, precision, recall and F1-score. Although XGBoost consistently showed the best results (for example, accuracy of 0.761 in binary classification grouped by day and 0.727 in multiclass classification), the absolute difference with the other models was, in general, modest. This invites us to ask to what extent the choice of the algorithm is decisive in this context, or if the performance is conditioned mainly by the nature and quality of the data rather than by the model itself.

However, the superiority of assembly methods (Random Forest and XGBoost) over linear models and neural networks in scenarios aggregated by day suggests that the ability to capture non-linear relationships and possible interactions between variables is relevant to the problem addressed. The improvement of AUC in daily groupings (from 0.61 to 0.77) points in that direction, which could be related to greater statistical stability and noise reduction when working with temporal windows longer than the shift.

It is remarkable that the performance was systematically higher in the binary classification tasks compared to the multiclass ones. This is to be expected, given that the multiclass classification, by increasing the number of categories to predict, increases the complexity of the problem and the risk of overlap between classes. However, it is worth asking if part of that loss of performance in the multiclass approach is due not only to the inherent difficulty but also to the possible imbalance of classes, something that could especially affect less frequent categories such as "self-harm" or "absconding".

Another question that arises is whether the type of coding of variables (binary vs weighted) provides relevant information or simply introduces redundancy. The results show that, in many cases, both codings yielded identical metrics, which could be due to a lack of sensitivity of the model to small variations in weight, or to the fact that the main predictors are already sufficiently informative in their dichotomous version.

The use of statistical tests such as McNemar and Wilcoxon allowed to go beyond the superficial comparison of metrics. In fact, although the absolute differences between models were small, in some scenarios (especially when grouping by day) statistically significant differences were detected, which validates the importance of applying these tests so as not to fall into simplistic conclusions.

It is interesting to note that the comparison between assembly methods (Random Forest vs XGBoost) did not show significant differences, which suggests that, although XGBoost has slight advantages, in practice any of the two could be a valid choice, depending on other factors such as interpretability, training time or ease of adjustment.

On the other hand, the fact that in some cases there are no significant differences (for example, between the two best multiclass models using turns) may mean that all models have already reached their maximum possible performance. This may be due to the quality of the data available or to the fact that predicting this type of behavior is very difficult.

Although the maximum accuracy reported was 0.761, the dispersion of accuracy values between models and scenarios was relatively low. This could be interpreted in several ways: on the one hand, it can indicate that the problem is simple enough for most models to achieve a similar performance (possibly due to the predictive power of a few key variables). On the other hand, it can indicate the existence of an upper limit in the predictive capacity due to information not captured in the data (for example, clinical variables not registered, quality of the information, etc.). In addition, the consistency of the results when optimizing hyperparameters of XGBoost suggests that the margin of improvement in this way is limited.

The analysis with SHAP values showed that the sociodemographic variables (schizophrenia diagnosis, gender and age) and some specific clinical indicators (PPII, PAH, PTCA, ALT, VIA) were systematically the most relevant for the prediction of risk, both in binary and multiclass models and in all the configurations analyzed. This consistency suggests that most of the predictive capacity of the models depends on a small group of main variables.

The highest risk profile identified by the models is a male patient, over 35 years old, with a diagnosis of schizophrenia, admitted involuntarily and who presents clinical indicators such as the need for help with hygiene, the presence of an eating disorder, a history of minor self-injury and the verbalization of autolytic ideation.

The limited variability in the importance of the variables indicates that, without the incorporation of new data or sources of information, it is unlikely that significant improvements will be achieved in the predictive performance. In addition, the dependence on socio-demographic variables poses risks of bias and must be considered in the interpretation and application of the models.

One of the main limitations of this study is the quantity and quality of the data with which it was worked. The size of the sample and the fact that some classes, especially in the multiclass classification, have very few cases make the model have difficulties in correctly detecting less common events.

This makes the models tend to predict more the majority class, which is a problem that could be solved using techniques to balance the classes or by applying specific penalties. In addition, although binary and weighted coding were used to represent the variables, these forms may not capture all the relevant clinical information. There is also the choice of using data grouped by turn or by day: grouping by day seems to give more stable and accurate results, but it can cause important details that occur in short periods to be lost, which could affect the practical use in making quick decisions.

Another important aspect is the generalization of the models. These were developed and validated with data from a single center or population, so it is not clear if they would work equally well in other hospitals or regions. Therefore, it would be necessary to test with external data to see how well they adapt. Although methods like SHAP were used to understand which variables have the most influence, assembly models like Random Forest and XGBoost are still difficult to interpret, especially for professionals who make clinical decisions, which could limit their acceptance. As for the statistical analyses, the number of repetitions and the independence of the samples also influence the robustness of the results.

For future research, it would be interesting to test more advanced models, such as recurrent neural networks (LSTM), which can better capture temporal information. It could also be useful to incorporate additional variables, such as environmental factors, service events or medication information, and even use natural language processing techniques to take advantage of unstructured data, such as clinical notes. Another important line would be to investigate how to improve prediction in minority classes using balancing techniques or specific adjustments and evaluate how these models can be integrated into real clinical practice, considering the risk of false positives or negatives and their impact on patient care.

The models showed a good performance, especially XGBoost with data grouped by day, but the most relevant was to show the strong influence of the socio-demographic variables and certain clinical indicators in the prediction of the risk. For each risk category, the socio-demographic variables had a significant weight, together with some specific clinical indicators. This finding is important because it facilitates the creation or mapping of profiles of patients at risk, which can contribute to earlier detection. However, it will be essential to validate and analyze with clinical professionals the real impact of these variables to ensure that the models are useful and reliable in clinical practice.

8. Project Impact and Schedule

This section presents the main aspects related to the project's ethical, social, environmental, and diversity impact. It also details the planned timeline, including key milestones and phases for its development and implementation. The objective is to provide an overview of the key considerations and project planning.

8.1. Ethical-social impact

The development of a predictive model of risk behaviors in mental health units involves fundamental ethical and social considerations, which have been carefully evaluated in the design of the research protocol presented to the Clinical Research Ethics Committee (CEIC) of the HUGTiP, obtaining a favorable opinion (PI-25-072).

From an ethical point of view, the study is based on the analysis of secondary data from pre-existing clinical registries (RutiSafeNet and Geslin), without direct contact or intervention on patients. The data is pseudo-anonymized in origin, guaranteeing the protection of the subject's identity in accordance with the General Data Protection Regulation (RGPD, EU 2016/679) and the Organic Law 3/2018 on the Protection of Personal Data. Access to information is restricted exclusively to authorized researchers, through secure corporate systems, and international data transfers are not foreseen. Given the retrospective nature of the study, the minimal invasiveness and the practical impossibility of obtaining the consent of all subjects, the exemption of informed consent has been requested and adequately justified.

From the social perspective, the objective of the model is to anticipate conflictive behaviors (such as self-harm attempts, episodes of aggression or absconding risk) in hospitalized patients, thus allowing preventive action. This contributes to reducing the use of coercive measures, improving healthcare security and promoting more respectful and humanized hospital environments. During this phase of the study, the developed algorithm will not make clinical decisions, ensuring that there is no direct impact on patient care. In addition, as an additional measure of protection, an impact assessment on the protection of personal data will be carried out to identify possible risks and establish adequate safeguards for future applications.

8.2. Sustainability

The project uses existing resources in the hospital environment, such as the RutiSafeNet and Geslin information systems, without the need to develop new data collection tools or modify the current technological infrastructures.

The database created for the analysis will be hosted on the hospital's corporate servers, managed under internal secure access protocols and without requiring additional investment.

In terms of medium and long-term sustainability, the developed model can be integrated into the clinical management systems already in use, allowing its possible updating or maintenance without generating a significant burden of resources. Moreover, since it is a retrospective analysis project, it does not involve additional consumption of materials, physical resources or waste generation, thus minimizing its environmental impact.

The continuity and future updating of the model will depend on the availability of updated data in the existing systems and the human resources necessary for its supervision and validation in subsequent phases.

8.3. Diversity

The population included in the study is made up of patients hospitalized in the adult mental health unit of the HUGTiP during the period between January 22, 2024 and February 22, 2025. No gender restrictions were applied and both men and women over 18 years of age were included. The final sample is made up of 390 unique patients, corresponding to the annual number of discharges from the unit, and is considered representative of the population served. The gender distribution in the sample is approximately balanced, with 120 women and 108 men.

In addition, the dataset incorporates relevant sociodemographic variables, such as gender, age, ethnic group (classified as black/African American, Asian Indian, Chinese, Japanese, Caucasian/white, gypsy or Muslim) and type of admission (voluntary or involuntary), which facilitates an analysis that takes into account the diversity of profiles present in the unit.

It should be noted that the ethnic group variable has not been addressed in this study due to lack of data and inconsistencies in its recording.

The consideration of these variables will allow us to evaluate the behavior of the predictive model in different subgroups of the population, guaranteeing its applicability within the specific care context of the hospital.

8.4. Project Timeline

Prior to the start of the Master's Final Thesis (MFT), curricular internships were carried out at the hospital, during which progress was made in structuring the database. For the MFT, the datasets were updated with more recent information; however, the methodology and organization of these data had been under development since September 23, 2024, spanning a period of approximately 24 weeks until the formal start of the MFT on March 7, 2025. During this period, work was carried out on defining variables, integrating sources (RutiSafeNet, Geslin), and standardizing time and clinical data, which allowed for the construction of a solid longitudinal database. The MFT focused on the development and validation of predictive models based on this structured database.

The Gantt Chart presented in Figure 19 corresponds only to the development and planning of the MFT, showing the different phases and the estimated duration of each of them over a period of approximately 12 weeks. In the schedule, the activities are organized with a biweekly review and follow-up, thus allowing a regular control of the progress of the project.

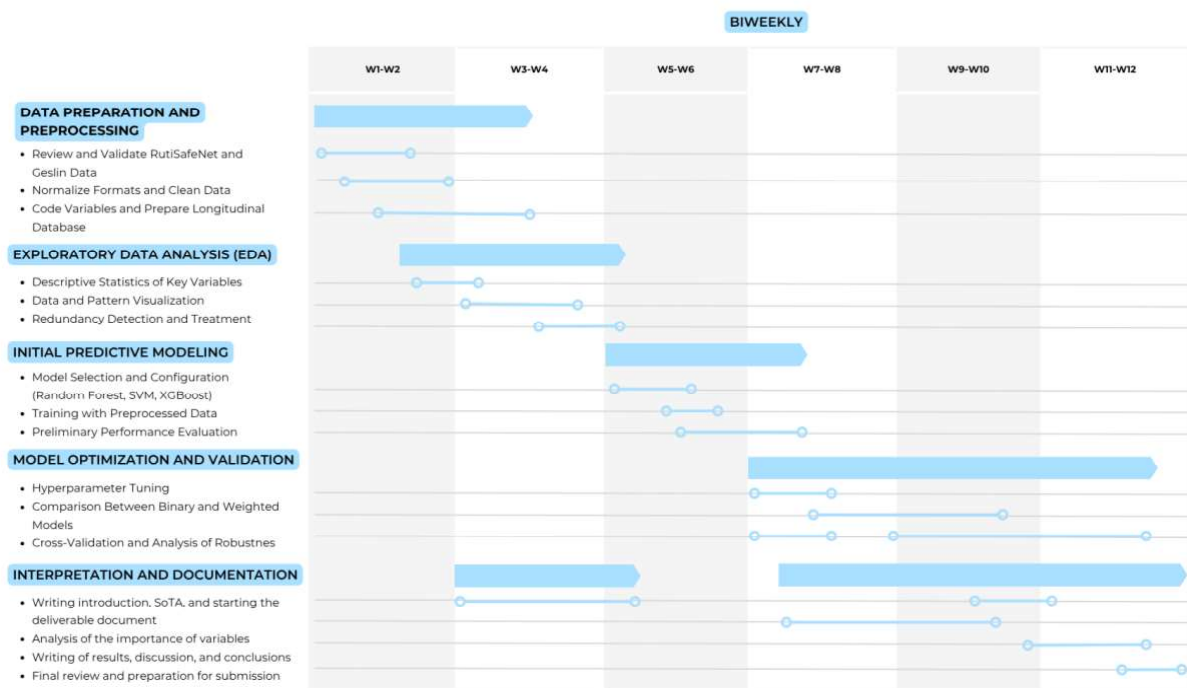


Figure 19. Project Gantt Chart

9. Conclusions and future work

This study evaluated the applicability of predictive models based on routine clinical data to anticipate conflictive behaviors in an adult mental health inpatient unit. The results show that the predictive ability is moderate and relatively consistent across algorithms, with XGBoost and Random Forest performing slightly better in most scenarios, although without substantial differences compared to linear models or neural networks. Grouping data by day improves the stability of predictions, but the difficulty of correctly identifying less frequent events due to class imbalance persists.

The importance of variables such as schizophrenia diagnosis, gender, age, and certain health indicators is a constant across models, reflecting that much of the predictive ability is concentrated in a small set of easily accessible factors. However, this dependence also limits the potential for improvement of the models without the incorporation of new sources of clinical or contextual information. The work demonstrates that, in the current context of open and less restrictive units, the integration of predictive tools based on existing data can provide added value for risk management, although their practical impact will depend on improving the representativeness of the data, refining the management of class imbalances and adapting the models to the specific needs of each healthcare setting.

Furthermore, it is important to highlight that the implementation of such a predictive tool could have a significant clinical impact, especially in open mental health inpatient units. In these settings, risk prediction models could help guide the degree of openness and management of involuntary patients, minimizing both legal risks for professionals and potential harms for patients. An involuntary patient absconding not only poses a substantial legal problem for staff but also represents a serious clinical risk. Therefore, this work should be considered as a first step towards the development of clinically useful prediction models, which will require larger data sets and thorough validation before they can be safely and reliably integrated into daily clinical practice. Future development should also focus on validation in other centers and on close collaboration with professionals to adapt the predictive tools to the reality of clinical practice.

To promote transparency and allow others to review the work done, all Python scripts (.py files) developed during this project have been temporarily uploaded to the public GitHub repository on behavioral risk and mental health [39]. This allows interested practitioners and researchers to access the analyses, replicate the results, and contribute to future developments. Please note that these files will remain available for evaluation purposes only and will be deleted one month after the thesis defense.

References

- [1] Bowers L. Safewards: a new model of conflict and containment on psychiatric wards. *J Psychiatr Ment Health Nurs* 2014;21:499.
<https://doi.org/10.1111/JPM.12129>.
- [2] Bowers L, James K, Quirk A, Simpson A, Stewart D, Hodsoll J. Reducing conflict and containment rates on acute psychiatric wards: The Safewards cluster randomised controlled trial. *Int J Nurs Stud* 2015;52:1412–22.
<https://doi.org/10.1016/J.IJNURSTU.2015.05.001>.
- [3] Huber CG, Schneeberger AR, Kowalinski E, Fröhlich D, von Felten S, Walter M, et al. Suicide risk and absconding in psychiatric hospitals with and without open door policies: a 15 year, observational study. *Lancet Psychiatry* 2016;3:842–9.
[https://doi.org/10.1016/S2215-0366\(16\)30168-7](https://doi.org/10.1016/S2215-0366(16)30168-7).
- [4] Schneeberger AR, Kowalinski E, Fröhlich D, Schröder K, von Felten S, Zinkler M, et al. Aggression and violence in psychiatric hospitals with and without open door policies: A 15-year naturalistic observational study. *J Psychiatr Res* 2017;95:189–95. <https://doi.org/10.1016/J.JPSYCHIRES.2017.08.017>.
- [5] Cornaggia CM, Beghi M, Pavone F, Barale F. Aggression in psychiatry wards: a systematic review. *Psychiatry Res* 2011;189:10–20.
<https://doi.org/10.1016/J.PSYCHRES.2010.12.024>.
- [6] Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). HCR-20. Assessing the risk of violence. Version 2. Burnaby Simon Fraser University and Forensic Psychiatric Services Commission of British Columbia. - References - Scientific Research Publishing n.d.
<https://www.scirp.org/reference/referencespapers?referenceid=673579> (accessed March 17, 2025).
- [7] Gunenc C, O'Shea LE, Dickens GL. Structured risk assessment for reduction of multiple risk outcomes in a secure mental health setting: Use of the START. *Criminal Behaviour and Mental Health* 2018;28:61–71.
<https://doi.org/10.1002/CBM.2036>.
- [8] Woods P, Almvik R. The Brøset violence checklist (BVC). *Acta Psychiatr Scand Suppl* 2002;106:103–5. <https://doi.org/10.1034/J.1600-0447.106.S412.22.X>.
- [9] Coid JW, Ullrich S, Kallis C. Predicting future violence among individuals with psychopathy. *Br J Psychiatry* 2013;203:387–8.
<https://doi.org/10.1192/BJP.BP.112.118471>.



- [10] OxMIV (TVP) | OxRisk n.d. <https://oxrisk.com/oxmiv-7/> (accessed March 17, 2025).
- [11] OxMIS | OxRisk n.d. <https://oxrisk.com/oxmis/> (accessed March 17, 2025).
- [12] Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Curr Psychiatry Rep* 2019;21:116. <https://doi.org/10.1007/S11920-019-1094-0>.
- [13] Passos IC, Ballester PL, Barros RC, Librenza-Garcia D, Mwangi B, Birmaher B, et al. Machine learning and big data analytics in bipolar disorder: A position paper from the International Society for Bipolar Disorders Big Data Task Force. *Bipolar Disord* 2019;21:582–94. <https://doi.org/10.1111/BDI.12828>.
- [14] Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, et al. Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2021;6:856–64. <https://doi.org/10.1016/J.BPSC.2021.02.001>.
- [15] Mezquita LM, Francesc J, López A. Digitalización del proceso de control de riesgos en pacientes con trastornos de salud mental que precisan hospitalización 2023.
- [16] Cheng N, Guo M, Yan F, Guo Z, Meng J, Ning K, et al. Application of machine learning in predicting aggressive behaviors from hospitalized patients with schizophrenia. *Front Psychiatry* 2023;14:1016586. <https://doi.org/10.3389/FPSYT.2023.1016586/BIBTEX>.
- [17] Dobbins NJ, Chipkin J, Byrne T, Ghabra O, Siar J, Sauder M, et al. Deep Learning Models Can Predict Violence and Threats Against Healthcare Providers Using Clinical Notes. *MedRxiv* 2024:2024.05.27.24308001. <https://doi.org/10.1101/2024.05.27.24308001>.
- [18] Cuomo A, Koukouna D, Macchiarini L, Fagiolini A. Patient Safety and Risk Management in Mental Health. *Textbook of Patient Safety and Clinical Risk Management* 2020:287–98. https://doi.org/10.1007/978-3-030-59403-9_20.
- [19] Large M, Sharma S, Cannon E, Ryan C, Nielssen O. Risk factors for suicide within a year of discharge from psychiatric hospital: a systematic meta-analysis. *Aust N Z J Psychiatry* 2011;45:619–28. <https://doi.org/10.3109/00048674.2011.590465>.
- [20] Haglund K, Von Knorring L, Von Essen L. Psychiatric wards with locked doors--advantages and disadvantages according to nurses and mental health nurse assistants. *J Clin Nurs* 2006;15:387–94. <https://doi.org/10.1111/J.1365-2702.2006.01489.X>.

- [21] Lang UE, Hartmann S, Schulz-Hartmann S, Gudlowski Y, Ricken R, Munk I, et al. Do locked doors in psychiatric hospitals prevent patients from absconding? *Eur J Psychiatry* 2010;24:199–204.
- [22] QualityRights: instrumento de calidad y derechos de la OMS: evaluando y mejorando la calidad y los derechos humanos en los establecimientos de salud mental y de apoyo social n.d. <https://iris.who.int/handle/10665/150398> (accessed March 17, 2025).
- [23] Fazel S, Wolf A, Larsson H, Mallett S, Fanshawe TR. The prediction of suicide in severe mental illness: development and validation of a clinical prediction rule (OxMIS). *Transl Psychiatry* 2019;9. <https://doi.org/10.1038/S41398-019-0428-3>.
- [24] Fazel S, Wolf A, Larsson H, Lichtenstein P, Mallett S, Fanshawe TR. Identification of low risk of violent crime in severe mental illness with a clinical prediction tool (Oxford Mental Illness and Violence tool [OxMIV]): a derivation and validation study. *Lancet Psychiatry* 2017;4:461. [https://doi.org/10.1016/S2215-0366\(17\)30109-8](https://doi.org/10.1016/S2215-0366(17)30109-8).
- [25] Wolf A, Fanshawe TR, Sariaslan A, Cornish R, Larsson H, Fazel S. Prediction of violent crime on discharge from secure psychiatric hospitals: A clinical prediction rule (FoVOx). *Eur Psychiatry* 2018;47:88–93. <https://doi.org/10.1016/J.EURPSY.2017.07.011>.
- [26] Passos IC, Ballester PL, Barros RC, Librenza-Garcia D, Mwangi B, Birmaher B, et al. Machine learning and big data analytics in bipolar disorder: A position paper from the International Society for Bipolar Disorders Big Data Task Force. *Bipolar Disord* 2019;21:582–94. <https://doi.org/10.1111/BDI.12828>.
- [27] Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, et al. Artificial Intelligence for Mental Healthcare: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2021;6:856. <https://doi.org/10.1016/J.BPSC.2021.02.001>.
- [28] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* (1979) 2019;366:447–53. https://doi.org/10.1126/SCIENCE.AAX2342/SUPPL_FILE/AAX2342_OBERMEYER_SM.PDF.
- [29] Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *N Engl J Med* 2018;378:981. <https://doi.org/10.1056/NEJMP1714229>.

- [30] Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019;25:37–43. <https://doi.org/10.1038/S41591-018-0272-7>.
- [31] Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *Npj Digital Medicine* 2023 6:1 2023;6:1–7. <https://doi.org/10.1038/s41746-023-00751-9>.
- [32] Velasquez VT, Chang J, Waddell A. The development of early warning scores or alerting systems for the prediction of adverse events in psychiatric patients: a scoping review. *BMC Psychiatry* 2024;24:742. <https://doi.org/10.1186/S12888-024-06052-Z>.
- [33] O'Reilly LM, Fazel S, Rickert ME, Kuja-Halkola R, Cederlöf M, Hellner C, et al. Evaluating Machine Learning for Predicting Youth Suicidal Behavior up to 1 Year After Contact With Mental-Health Specialty Care. *Clinical Psychological Science* 2024. <https://doi.org/10.1177/21677026241301298>.
- [34] LogisticRegression — scikit-learn 1.6.1 documentation n.d. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed June 3, 2025).
- [35] RandomForestClassifier — scikit-learn 1.6.1 documentation n.d. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> (accessed June 3, 2025).
- [36] GradientBoostingClassifier — scikit-learn 1.6.1 documentation n.d. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html> (accessed June 3, 2025).
- [37] 1.17. Neural network models (supervised) — scikit-learn 1.6.1 documentation n.d. https://scikit-learn.org/stable/modules/neural_networks_supervised.html (accessed June 3, 2025).
- [38] Fazel S, Wolf A, Larsson H, Mallett S, Fanshawe TR. The prediction of suicide in severe mental illness: development and validation of a clinical prediction rule (OxMIS). *Translational Psychiatry* 2019 9:1 2019;9:1–10. <https://doi.org/10.1038/s41398-019-0428-3>.
- [39] anbarrosog/behavioural-risk-mental-health: Master thesis of “behavioural risk prediction in acute mental health inpatient unit” n.d. <https://github.com/anbarrosog/behavioural-risk-mental-health> (accessed June 2, 2025).

Annexes

Models' metrics

Comparative Model Performance (Binary Classification)

	Model	Grouping	Dataset	Accuracy	Precision	Recall	F1-score	AUC
0	Logistic Regression	Turn	Binary	0.653	0.502	0.459	0.480	0.611
1	Logistic Regression	Turn	Weighted	0.653	0.502	0.459	0.480	0.611
2	Logistic Regression	Day	Binary	0.706	0.578	0.586	0.582	0.679
3	Logistic Regression	Day	Weighted	0.706	0.578	0.586	0.582	0.679
4	Random Forest	Turn	Binary	0.669	0.533	0.396	0.455	0.667
5	Random Forest	Turn	Weighted	0.669	0.533	0.396	0.455	0.667
6	Random Forest	Day	Binary	0.757	0.674	0.586	0.627	0.763
7	Random Forest	Day	Weighted	0.757	0.674	0.586	0.627	0.763
8	XGBoost	Turn	Binary	0.680	0.551	0.437	0.487	0.643
9	XGBoost	Turn	Weighted	0.680	0.551	0.437	0.487	0.643
10	XGBoost	Day	Binary	0.761	0.667	0.631	0.648	0.766
11	XGBoost	Day	Weighted	0.761	0.667	0.631	0.648	0.766
12	Neural Net	Turn	Binary	0.611	0.367	0.162	0.225	0.591
13	Neural Net	Turn	Weighted	0.644	0.478	0.248	0.326	0.612
14	Neural Net	Day	Binary	0.711	0.628	0.419	0.503	0.715
15	Neural Net	Day	Weighted	0.721	0.649	0.432	0.519	0.726

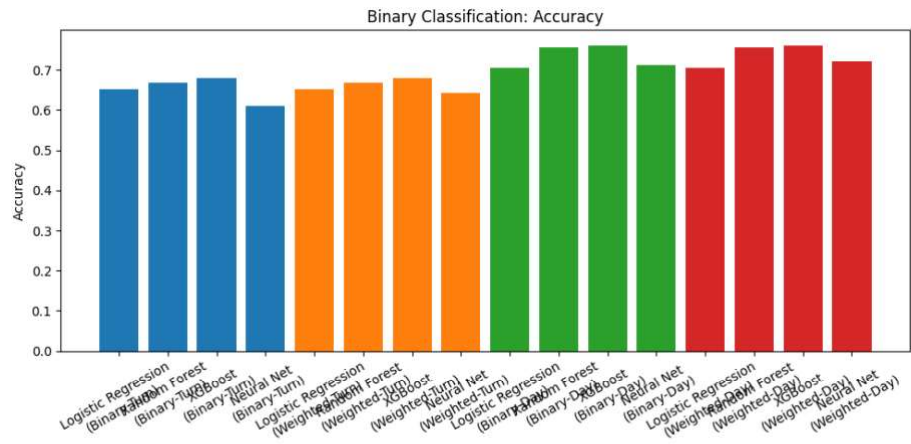
	Model	Grouping	Dataset	Accuracy	Precision	Recall	F1-score
0	Logistic Regression	Turn	Binary	0.386	0.606	0.386	0.449
1	Logistic Regression	Turn	Weighted	0.381	0.601	0.381	0.443
2	Logistic Regression	Day	Binary	0.487	0.741	0.487	0.544
3	Logistic Regression	Day	Weighted	0.480	0.738	0.480	0.538
4	Random Forest	Turn	Binary	0.504	0.698	0.504	0.565
5	Random Forest	Turn	Weighted	0.504	0.698	0.504	0.565
6	Random Forest	Day	Binary	0.562	0.747	0.562	0.622
7	Random Forest	Day	Weighted	0.562	0.747	0.562	0.622
8	XGBoost	Turn	Binary	0.626	0.559	0.626	0.578
9	XGBoost	Turn	Weighted	0.626	0.559	0.626	0.578
10	XGBoost	Day	Binary	0.727	0.687	0.727	0.702
11	XGBoost	Day	Weighted	0.727	0.687	0.727	0.702
12	Neural Net	Turn	Binary	0.641	0.526	0.641	0.543
13	Neural Net	Turn	Weighted	0.633	0.519	0.633	0.544
14	Neural Net	Day	Binary	0.669	0.587	0.669	0.591
15	Neural Net	Day	Weighted	0.673	0.588	0.673	0.599

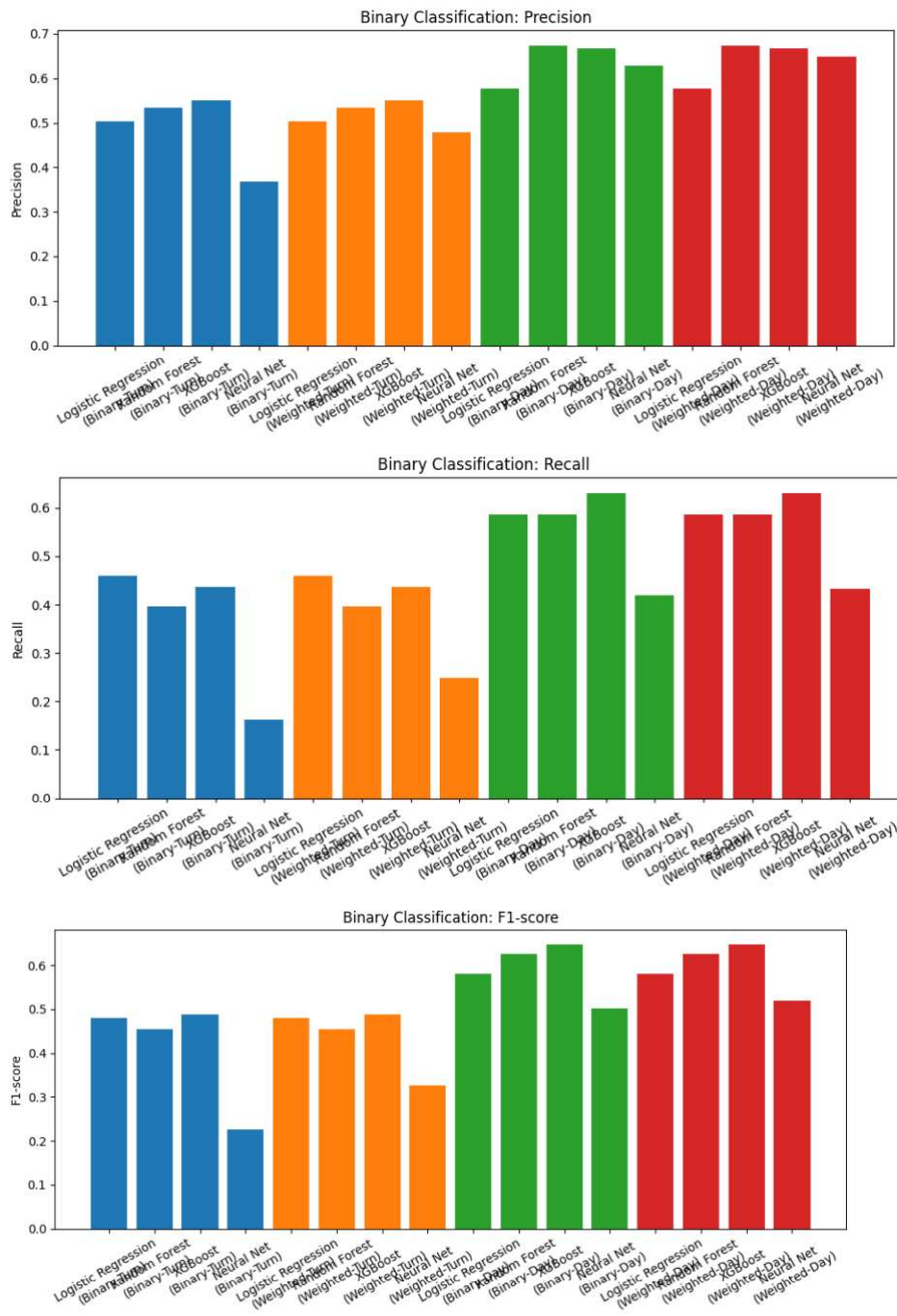
Pairwise McNemar Test (Binary Classification Scenarios)

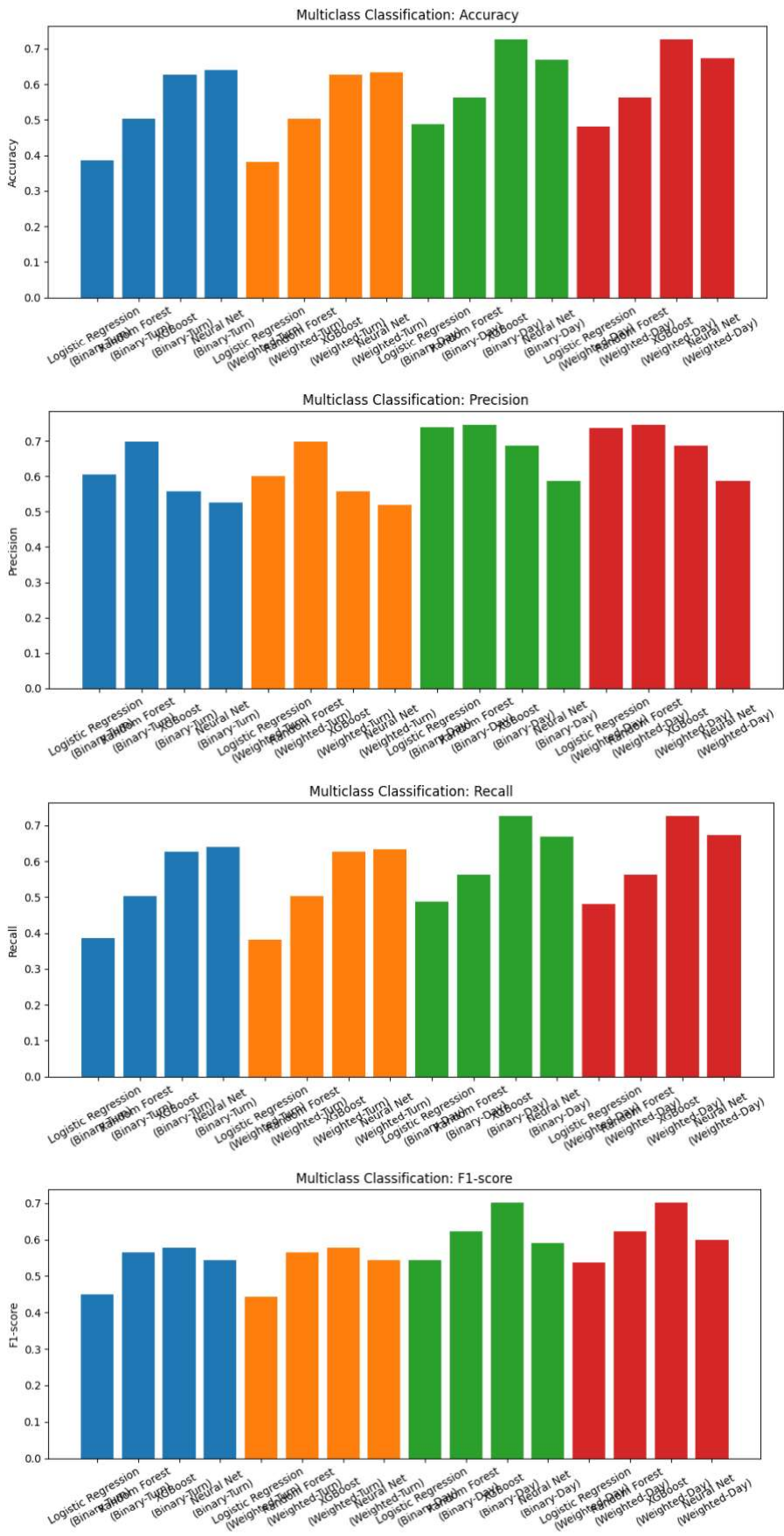
Scenario	Model 1	Model 2	McNemar p-value	Significant (p < 0.05)
0 Turn/Binary	Logistic Regression	Random Forest	0.3374	No
1 Turn/Binary	Logistic Regression	XGBoost	0.0396	Yes
2 Turn/Binary	Logistic Regression	Neural Net	0.0207	Yes
3 Turn/Binary	Random Forest	XGBoost	0.3817	No
4 Turn/Binary	Random Forest	Neural Net	0.0004	Yes
5 Turn/Binary	XGBoost	Neural Net	0.0000	Yes
6 Turn/Weighted	Logistic Regression	Random Forest	0.3374	No
7 Turn/Weighted	Logistic Regression	XGBoost	0.0396	Yes
8 Turn/Weighted	Logistic Regression	Neural Net	0.6427	No
9 Turn/Weighted	Random Forest	XGBoost	0.3817	No
10 Turn/Weighted	Random Forest	Neural Net	0.1214	No
11 Turn/Weighted	XGBoost	Neural Net	0.0297	Yes
12 Day/Binary	Logistic Regression	Random Forest	0.0002	Yes
13 Day/Binary	Logistic Regression	XGBoost	0.0000	Yes
14 Day/Binary	Logistic Regression	Neural Net	0.8408	No
15 Day/Binary	Random Forest	XGBoost	0.6900	No
16 Day/Binary	Random Forest	Neural Net	0.0038	Yes
17 Day/Binary	XGBoost	Neural Net	0.0018	Yes
18 Day/Weighted	Logistic Regression	Random Forest	0.0002	Yes
19 Day/Weighted	Logistic Regression	XGBoost	0.0000	Yes
20 Day/Weighted	Logistic Regression	Neural Net	0.4069	No
21 Day/Weighted	Random Forest	XGBoost	0.6900	No
22 Day/Weighted	Random Forest	Neural Net	0.0178	Yes
23 Day/Weighted	XGBoost	Neural Net	0.0112	Yes

Wilcoxon Test Between Best Multiclass Models per Scenario

Scenario	Best Model	Second Best Model	Accuracy 1	Accuracy 2	Wilcoxon p-value	Significant (p < 0.05)
0 Turn/Binary	Neural Net	XGBoost	0.641	0.626	0.1269	No
1 Turn/Weighted	Neural Net	XGBoost	0.633	0.626	0.6042	No
2 Day/Binary	XGBoost	Neural Net	0.727	0.669	0.0026	Yes
3 Day/Weighted	XGBoost	Neural Net	0.727	0.673	0.0137	Yes







Best Accuracy (Binary):

Model XGBoost
Grouping Day
Dataset Binary
Accuracy 0.761381
Precision 0.666667
Recall 0.630631
F1-score 0.648148
AUC 0.765636
Name: 10, dtype: object

Best F1-score (Binary):

Model XGBoost
Grouping Day
Dataset Binary
Accuracy 0.761381
Precision 0.666667
Recall 0.630631
F1-score 0.648148
AUC 0.765636
Name: 10, dtype: object

Best Accuracy (Multiclass):

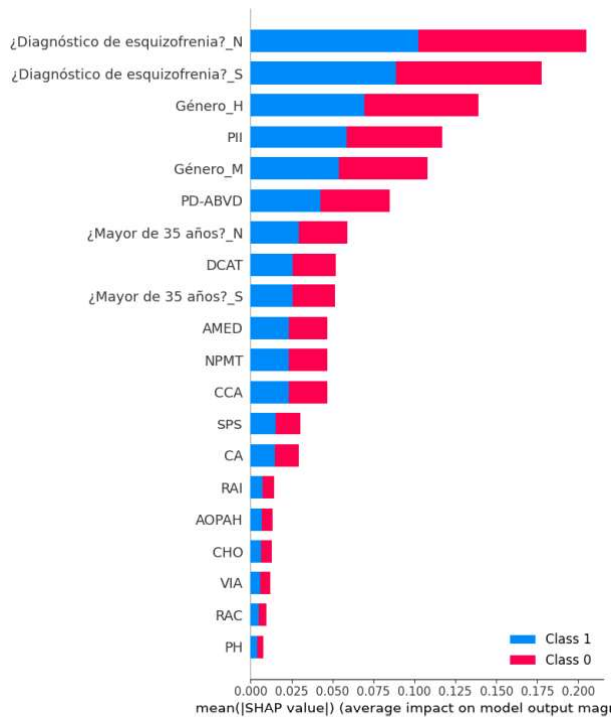
Model XGBoost
Grouping Day
Dataset Binary
Accuracy 0.726845
Precision 0.687047
Recall 0.726845
F1-score 0.702056
Name: 10, dtype: object

Best F1-score (Multiclass):

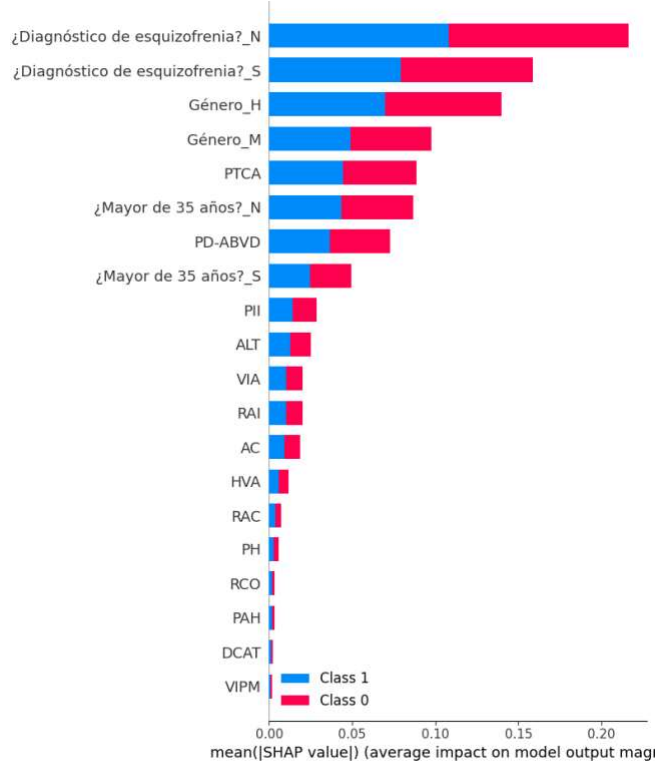
Model XGBoost
Grouping Day
Dataset Binary
Accuracy 0.726845
Precision 0.687047
Recall 0.726845
F1-score 0.702056
Name: 10, dtype: object

SHAP

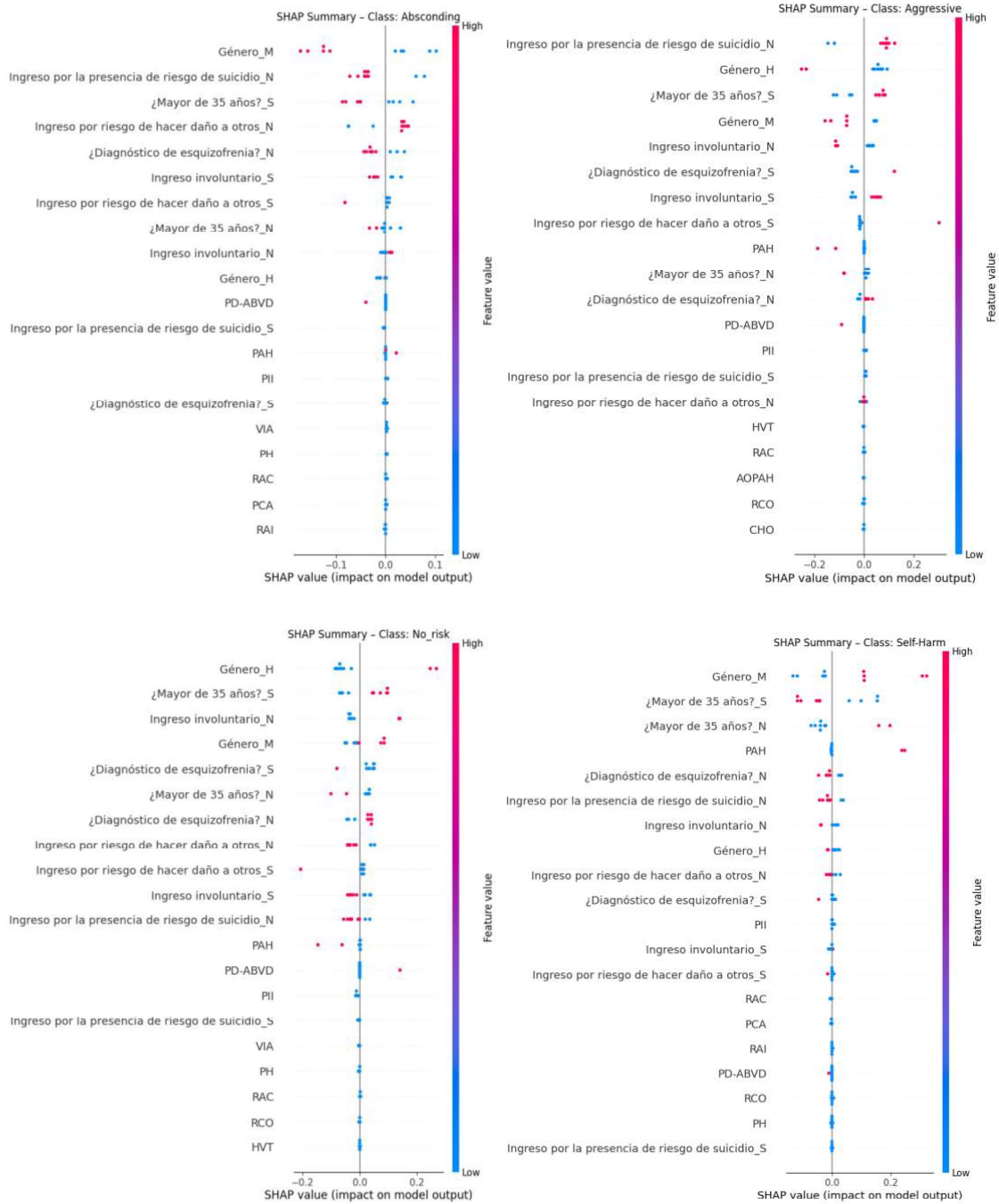
SHAP - Logistic Regression (Binary, Binary Features, by TURN)



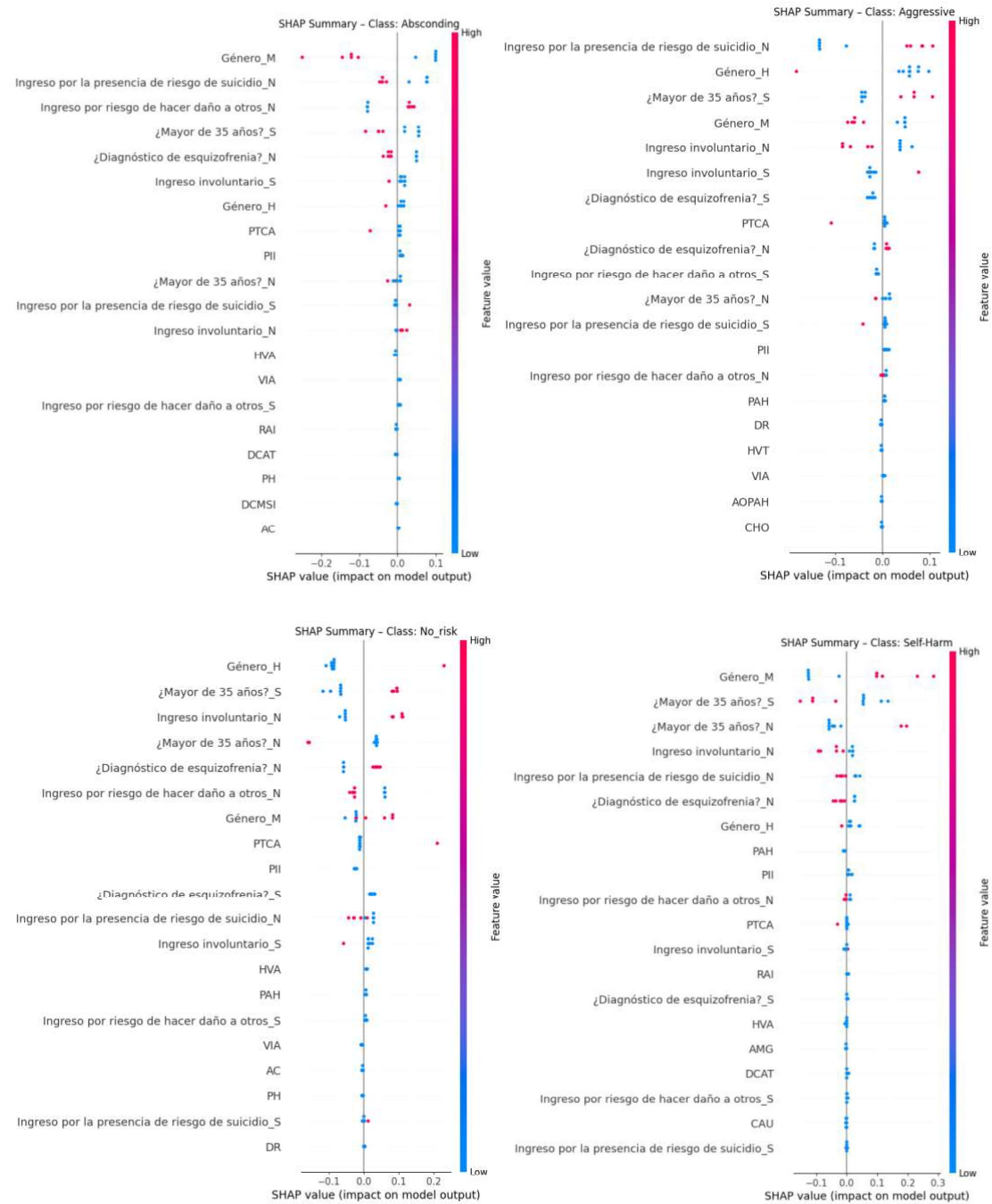
SHAP - Logistic Regression (Binary, Weighted Features, by TURN)



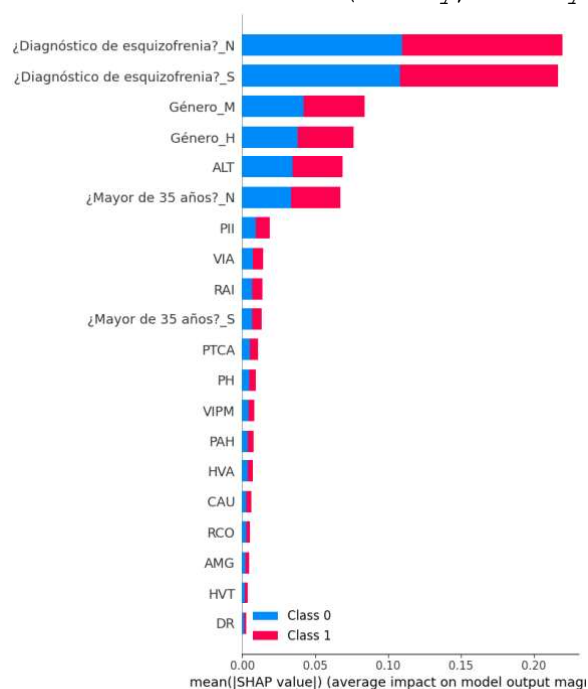
SHAP - Logistic Regression (Multiclass, Binary Features, by TURN)



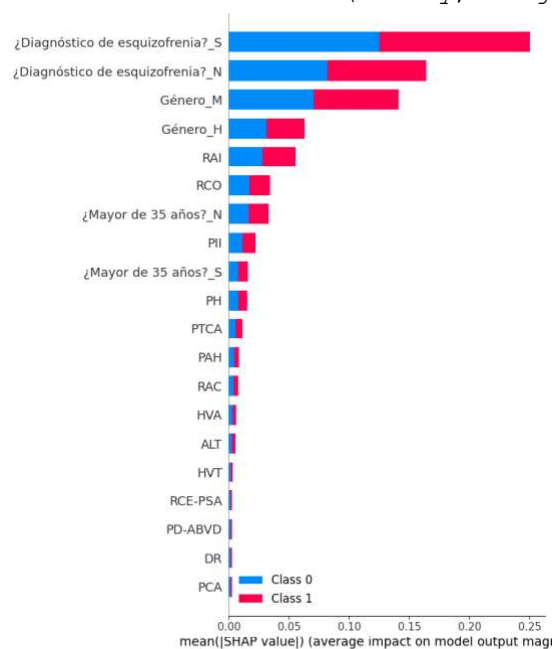
SHAP - Logistic Regression (Multiclass, Weighted Features, by TURN)



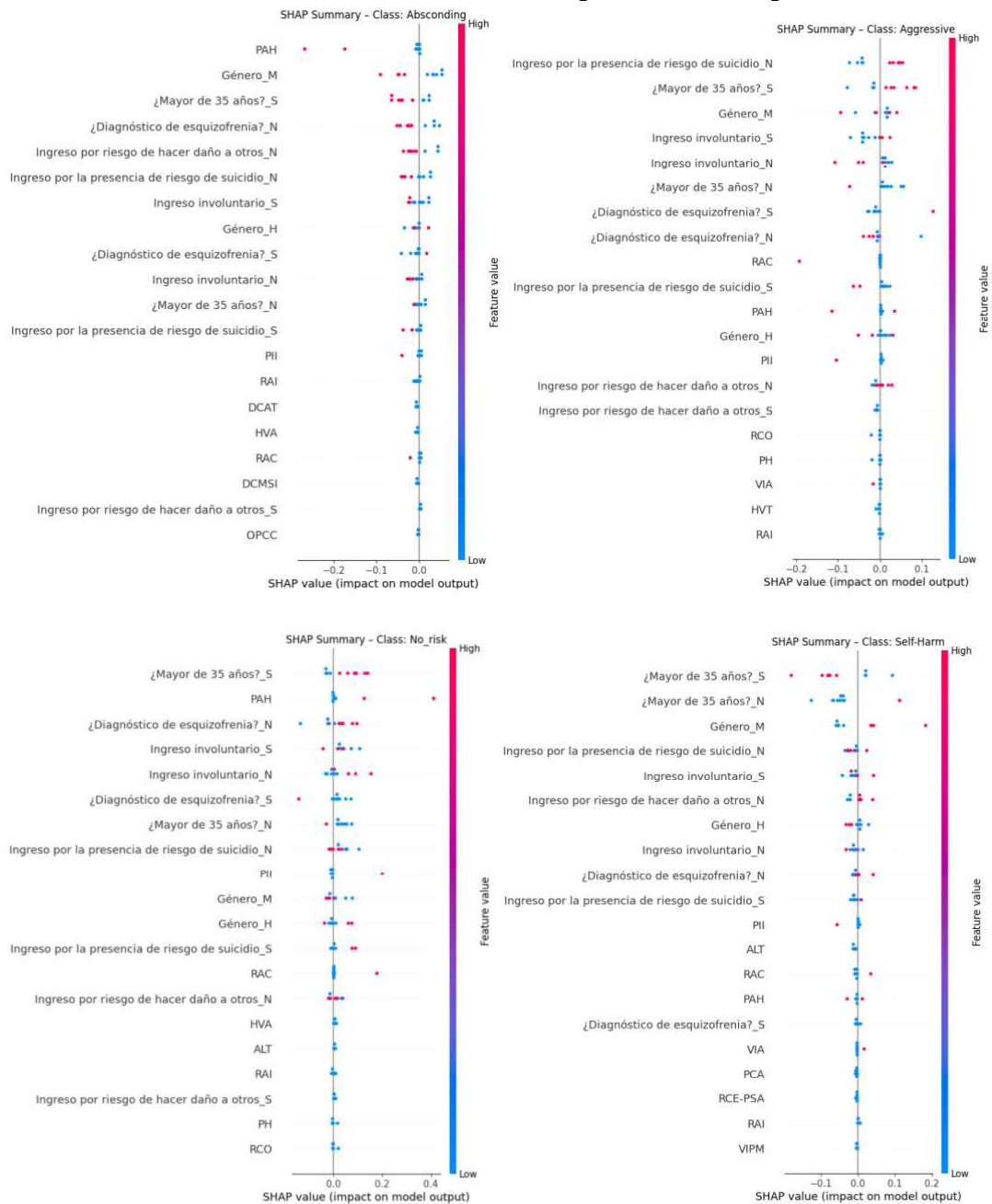
SHAP - Random Forest (Binary, Binary Features, by TURN)



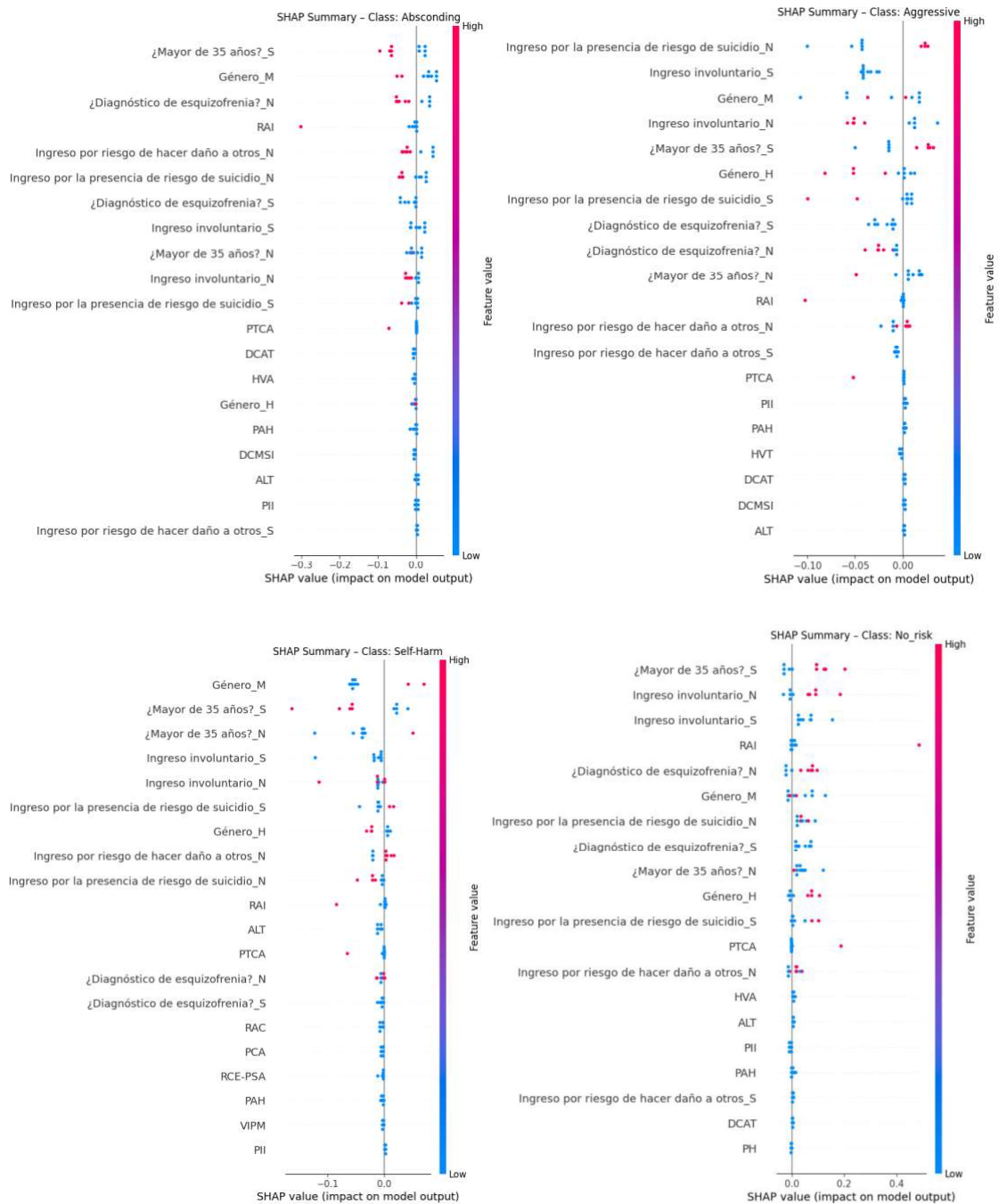
SHAP - Random Forest (Binary, Weighted Features, by TURN)



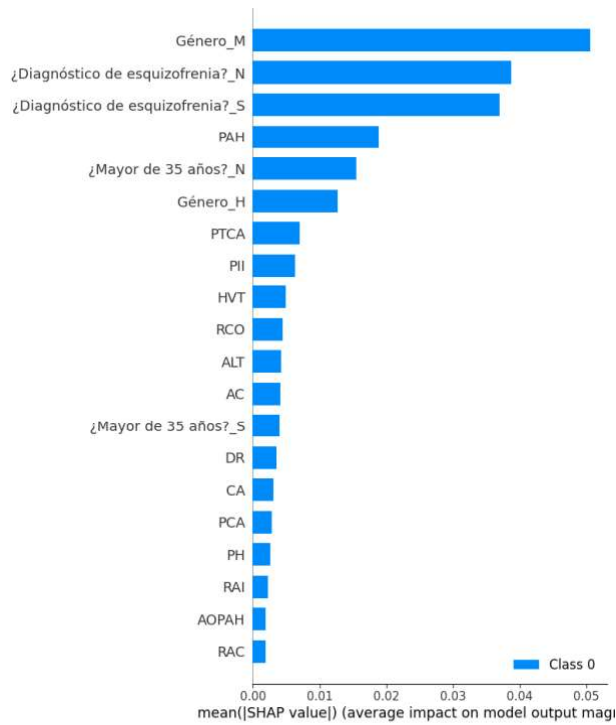
SHAP - Random Forest (Multiclass, Binary Features, by TURN)



SHAP - Random Forest (Multiclass, Weighted Features, by TURN)



SHAP - Neural Network (Binary, Binary Features, by TURN)



Grid Search

Best parameters for Binary: {'learning_rate': 0.01, 'max_depth': 4, 'n_estimators': 50, 'subsample': 1.0}
Best F1 score for Binary: 0.0864

Best parameters for Weighted: {'learning_rate': 0.01, 'max_depth': 4, 'n_estimators': 50, 'subsample': 1.0}
Best F1 score for Weighted: 0.0864

Grid Search XGBoost (Binary features, multiclass classification, LOPOC V by patient)
Best parameters: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100, 'subsample': 0.8}
Best F1 score: 0.7789630211058782

Grid Search XGBoost (Weighted features, multiclass classification, LOP OCV by patient)
Best parameters: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100, 'subsample': 0.8}
Best F1 score: 0.7789630211058782

--- Optimized XGBoost (Binary features) ---				
	precision	recall	f1-score	support
0	0.81	0.73	0.76	415
1	0.57	0.68	0.62	222
accuracy			0.71	637
macro avg	0.69	0.70	0.69	637
weighted avg	0.73	0.71	0.71	637
AUC: 0.7698686638445675				

```
--- Optimized XGBoost (Weighted features) ---
      precision    recall  f1-score   support

     0         0.81      0.73      0.76         415
     1         0.57      0.68      0.62         222

 accuracy          0.71         637
  macro avg         0.69      0.70      0.69         637
 weighted avg         0.73      0.71      0.71         637

AUC: 0.7698686638445675
```