

A Generalizable Normalization Pipeline for Explainable Machine Learning Analysis in Mental Health Datasets

Martín Di Felice*, Ariel Deroche, Ilan Trupkin, Parag Chatterjee and
María Florencia Pollo-Cattaneo

Universidad Tecnológica Nacional, Buenos Aires, Argentina

Abstract

Depression is among the most prevalent mental health disorders worldwide, yet its diagnosis still relies largely on subjective clinical assessment, and the datasets available for computational analysis are markedly heterogeneous in structure, variable types, missing-value patterns, and target definitions. This heterogeneity forces analytical decisions that are difficult to standardize and difficult for healthcare professionals to audit. This work addresses that problem through a metadata-driven normalization and explainable Machine Learning pipeline for mental health dataset analysis. The pipeline organizes the analysis into six stages — data collection, preprocessing, feature engineering, algorithm selection, results analysis, and results explanation — of which the sixth is introduced here to translate technical outputs into physician-oriented interpretations. Configuration is externalized to dataset-level metadata rather than dataset-specific code, so heterogeneous sources follow a common and traceable execution path. The pipeline was evaluated on four mental health datasets, two supervised and two unsupervised. LightGBM offered the best overall balance across metrics, reaching an F1-score of 0.7338 and AUC of 0.9439 on the first supervised dataset and 0.9766 and 0.9955 on the second, while Random Forest achieved the highest AUC (0.9971). In the unsupervised experiments, K-Means produced the most stable two-cluster partitions, with post-clustering statistical profiling identifying the variables that differentiate the resulting groups. The results indicate improved traceability, comparability, and interpretability across heterogeneous mental health data, supporting decision-support rather than automated diagnosis.

Keywords

Machine Learning, Explainable Artificial Intelligence, Normalization Pipeline, Mental Health, Depression, Clinical Decision Support

1. Introduction

Depression is one of the most prevalent mental health disorders worldwide and a leading contributor to global disability [1]. Its diagnosis continues to depend largely on subjective clinical assessment and self-reported symptoms, which contributes to underdiagnosis, delayed intervention, and variability across clinicians. Artificial Intelligence (AI) and Machine Learning (ML) techniques have therefore been increasingly applied to mental health analysis, particularly in depression detection, patient profiling and clinical decision support, where they offer a means of identifying patterns across questionnaire, behavioral and physiological data.

Research in this area has followed several complementary directions. Supervised prediction uses labeled datasets, typically derived from validated instruments such as the PHQ-9 [2] or the BDI-II [3], to train classifiers distinguishing depressed from non-depressed individuals; tree-based ensembles such as Random Forest [4], XGBoost [5] and LightGBM [6] are frequently reported as strong performers on structured mental health data, and this paradigm has previously been applied to depression detection [7]. Unsupervised analysis addresses the fact that mental health labels are costly to obtain and often clinically inconsistent or unavailable: clustering is used not to predict a diagnosis but to reveal latent behavioral

ICAIW 2026: Workshops at the 9th International Conference on Applied Informatics 2026, October 13–17, 2026, Orlando, United States

*Corresponding author.

✉ mdifelice@frba.utn.edu.ar (M. D. Felice); arielderoche@gmail.com (A. Deroche); ilantrup@gmail.com (I. Trupkin);
parag@frba.utn.edu.ar (P. Chatterjee); flo.pollo@gmail.com (M. F. Pollo-Cattaneo)

ORCID 0000-0003-1388-3220 (M. D. Felice); 0009-0000-0354-8803 (A. Deroche); 0009-0006-6594-643X (I. Trupkin);
0000-0001-6760-4704 (P. Chatterjee); 0000-0003-4197-3880 (M. F. Pollo-Cattaneo)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

profiles, with quality assessed through internal validation indices such as the Silhouette coefficient [8], the Davies–Bouldin Index [9] and the Calinski–Harabasz score [10], an approach previously applied to depression and anxiety data [11]. Systematic reviews of the field consistently report the same structural weaknesses across studies: small and heterogeneous cohorts, inconsistent preprocessing, limited external validation, and limited interpretability [12]. More recently, these two paradigms have been combined within a single modular framework that applies supervised classification and unsupervised clustering to heterogeneous mental health datasets through a shared preprocessing stage [13], establishing the analytical structure that the present work extends. A more recent direction concerns explainability. As models are proposed for clinical use, predictive performance alone has proven insufficient, since clinicians require an account of how a result was produced. Model-agnostic methods have become central to this discussion, including SHAP for globally consistent feature attribution [14, 15], LIME for locally faithful surrogate explanations [16] and Anchor for high-precision rule-based conditions [17]. These methods are, however, typically evaluated in isolation and on a single modeling task, rather than embedded as a standard stage within a reproducible analytical workflow.

A persistent gap follows from this. Studies tend to address a single dataset, a single learning paradigm or a single explanation technique, and the analytical process itself — how data were cleaned, encoded, scaled and filtered before modeling — is often reported only summarily. In practice, each dataset presents a different structure, variable types, missing-value pattern, target definition, and class distribution, so the analysis depends on technical decisions that are difficult to standardize and difficult to interpret for healthcare professionals without programming knowledge. In clinical contexts, model performance alone is not enough: physicians also need to understand how the data were processed, which variables influenced the results, and what limitations should be considered.

To address this, the present work proposes a metadata-driven normalization and explainable Machine Learning pipeline for mental health dataset analysis. Its contributions are: (i) a metadata-driven configuration mechanism that externalizes preprocessing and encoding decisions from the analytical code; (ii) a unified workflow accommodating both supervised and unsupervised tasks without structural modification; (iii) a sixth methodological stage, results explanation, which extends the five-stage structure of [13] by translating model outputs, relevant variables, and data-quality warnings into physician-oriented interpretations; and (iv) an empirical evaluation across four mental health datasets, two supervised and two unsupervised. The remainder of the paper is organized as follows. Section 2 describes the pipeline and its six stages, including the explainability framework. Section 3 presents the experimental evaluation. Section 4 discusses the findings and limitations, and Section 5 concludes.

2. Methodology: Proposed Pipeline

The proposed pipeline organizes the analysis into six stages: data collection, preprocessing, feature engineering, algorithm selection, results analysis, and results explanation. The first five follow the analytical structure adopted in previous studies [11, 7], while the sixth is introduced in this work to support clinical interpretation of the results. Figure 1 summarizes the workflow and Table 1 details the operations and outputs of each stage.

Metadata-driven configuration: The defining characteristic of the pipeline is that dataset-specific decisions are externalized into a declarative metadata definition rather than embedded in the analytical code. For each dataset, the metadata specifies the file format and location, the target variable when one exists, attributes to be removed, categorical encoding rules, the treatment of multiple-response and date fields, missing-value thresholds, correlation-filtering criteria, and the scaling strategy to apply. The execution code is therefore identical across datasets, and adapting the pipeline to a new source requires only a new metadata definition. This design serves two purposes. Technically, it makes runs reproducible and comparable, since every transformation applied to a dataset is declared explicitly and can be inspected without reading code. Clinically, it makes the analytical process auditable: a physician or domain reviewer can examine the metadata to see which variables were discarded, how the target was defined, and what preprocessing was applied, without programming knowledge.

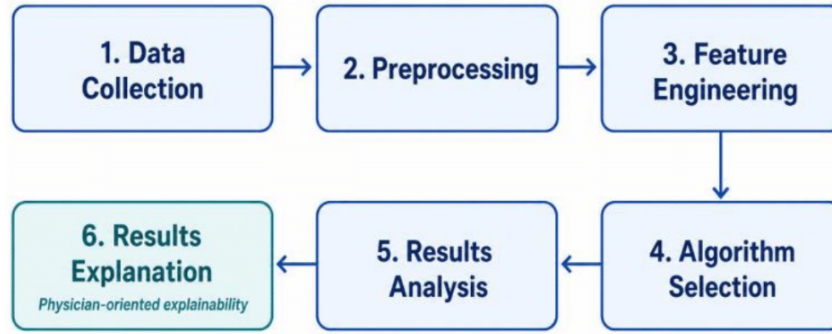


Figure 1: Methodological workflow of the proposed pipeline. The first five stages follow the analytical structure used in previous works, while the sixth stage introduces physician-oriented explainability for interpreting the results.

Data collection: Structured datasets are loaded according to their metadata definitions, and an initial profile is produced describing the number of records and attributes, variable types, and the presence of a target variable. This profile establishes whether the dataset will follow the supervised or the unsupervised branch of the workflow.

Preprocessing: The pipeline standardizes column names, detects and handles empty or invalid values, removes duplicate records, and discards rows and columns exceeding predefined missing-value thresholds. Thresholds are metadata-defined rather than fixed, since the acceptable level of incompleteness varies with dataset size and the clinical importance of individual variables.

Feature engineering: The dataset is transformed into a representation suitable for modeling. This includes removing metadata-defined irrelevant variables, binarizing the target when required, encoding categorical attributes according to their cardinality, handling multiple-response variables typical of survey instruments, transforming date fields, filtering highly correlated variables to limit redundancy, applying the configured scaling strategy, and generating model-compatible feature names.

Algorithm selection: The pipeline supports both learning paradigms without structural modification. For supervised datasets, classification models are trained and evaluated using cross-validation, reducing dependence on a single train–test partition — a relevant consideration given the modest size of many mental health datasets. For unsupervised datasets, clustering algorithms are applied to identify latent groups, with the number of clusters selected through internal validation rather than fixed in advance.

Results analysis: Model behavior is compared using task-specific metrics. Supervised models are evaluated using F1-score, AUC, accuracy, precision, recall, specificity, the KS statistic, and Lift@10, together with execution time, so that predictive quality and computational cost can be weighed together. Unsupervised models are assessed using the Silhouette coefficient [8], the Davies–Bouldin Index [9] and the Calinski–Harabasz score [10], the resulting number of clusters, and statistical tests between the identified groups.

Post-clustering statistical profiling: Because no target variable is used during clustering, the objective is not to predict depression but to identify latent groups and then characterize them. After cluster assignment, independent T-tests are applied between clusters for each variable to identify those with statistically significant differences ($\alpha = 0.05$). Two outputs follow: the individual variables that differentiate the groups, which constitute the interpretable description of each cluster, and an aggregate indicator reported in the results tables as T-Test, corresponding to the proportion of evaluated variables showing statistically significant differences between clusters. Higher values therefore indicate partitions separated along a larger share of the available variables, complementing the geometric information provided by the internal validation indices.

Results explanation: The final stage generates interpretation-oriented outputs, including model comparison, visualizations, relevant-variable analysis, data-quality warnings, and explicit limitations. To ensure transparency, the pipeline integrates a multi-layered Explainable AI (XAI) framework adapted

to the learning paradigm.

For supervised tasks, three complementary methods are applied. SHAP [14] provides globally consistent feature attribution based on Shapley values and is used for model auditing: identifying which variables drive predictions overall and whether the model relies on clinically plausible signals. LIME [16] produces locally faithful surrogate explanations for individual cases, addressing the question of why a particular subject received a particular prediction. Anchor [17] yields high-precision rule-based explanations expressed as sufficient conditions, that is, sets of variable values under which the model’s prediction holds with high probability. The three operate at different granularities – global attribution, local attribution, and local rules – and are reported together rather than in isolation, since each answers a different question a reviewer may raise.

For unsupervised tasks, no ground truth is available to anchor feature importance, so explainability is addressed through statistical and geometric profiling of the identified groups. Cluster structure is characterized using the internal validation indices together with the Elbow method for assessing the number of clusters, while clinical relevance is approached through the post-clustering T-test profiling described above, which converts cluster assignments into an interpretable statement about which variables distinguish each group. These outputs are intended to help physicians review the analytical process and interpret the results as decision-support information rather than automatic diagnostic conclusions.

Table 1

Table of Variable Rationale.

Stage	Main operations	Output
Data collection	Load CSV/XLSX per metadata, identify variables and target	Initial dataset profile
Preprocessing	Missing-value thresholds, empty/invalid value handling, duplicate removal, column standardization	Clean dataset
Feature engineering	Encoding, target binarization, multiple-response and date handling, correlation filtering, scaling	ML-ready feature set
Algorithm selection	Supervised classification with cross-validation, or clustering with internal validation	Trained models
Results analysis	Task-specific metric computation and model comparison	Performance metrics
Results explanation	SHAP / LIME / Anchor or cluster profiling, relevant variables, warnings, limitations	Interpretable results

3. Results

The experimental evaluation was designed to assess the proposed pipeline across four mental health datasets: two supervised datasets with target variables and two unsupervised datasets without predefined labels. Each dataset was processed using the same six-stage methodology described in Section 2, with dataset-specific rules defined through metadata. The evaluation focused on the pipeline’s ability to normalize heterogeneous datasets, preserve a common execution structure, generate comparable outputs, and provide interpretation-oriented information for reviewing the results. Table 2 summarizes the datasets used in the study. All four datasets are publicly available and fully anonymized [18, 19, 20, 21]. They were selected to span a range of population types, sample sizes, feature counts, and labeling conditions, so that the pipeline could be assessed under heterogeneous input conditions rather than on a single data profile.

3.1. Dataset 1 - Supervised

For the first supervised dataset, the evaluated models achieved consistent performance, with AUC and accuracy values above 0.90. The best overall result was obtained by LGBM, reaching an F1-score of

Table 2

Summary of datasets used in the experimental evaluation.

ID	Dataset	Features	Analysis type
1	Argentina Post-Pandemic Survey	n=1,100, 13 features	Supervised
2	Bangladeshi Depression Survey	n=4,464, 13 features	Supervised
3	Japanese College Students	n=268, 50 features	Unsupervised
4	Bangladeshi College Students	n=823, 36 features	Unsupervised

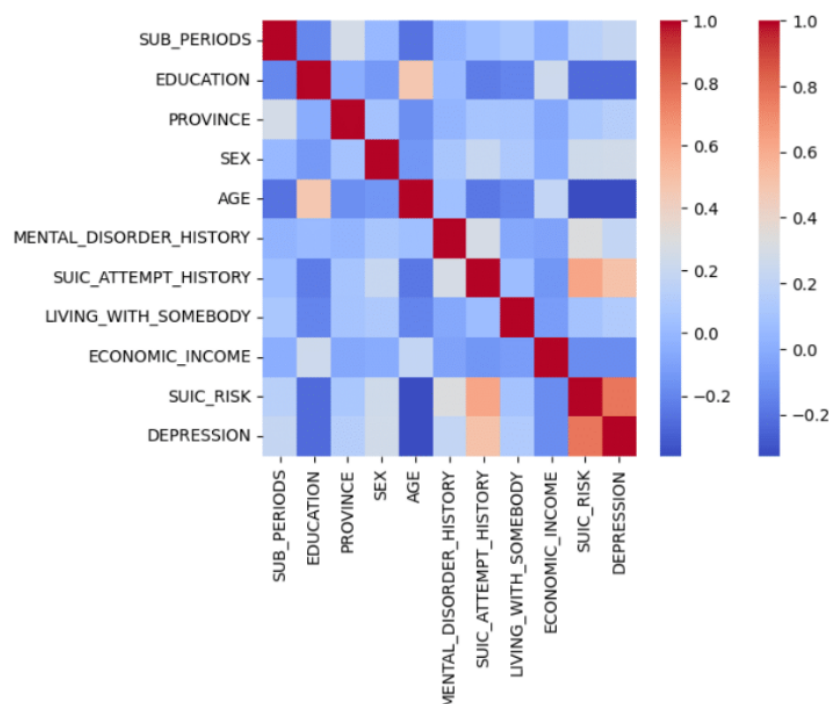
0.7338, AUC of 0.9439, accuracy of 0.9273, precision of 0.7450, recall of 0.7333, and specificity of 0.9579. These results indicate a balanced behavior between the identification of positive and negative cases.

Table 3

Performance comparison of supervised models for Dataset 1.

Model	F1	AUC	Accuracy	Precision	Recall	Specificity	Time (s)
LGBM	0.7338	0.9439	0.9273	0.7450	0.7333	0.9579	0.1528
RF	0.6864	0.9526	0.9182	0.7933	0.6333	0.9632	0.6876
XGB	0.6462	0.9535	0.9045	0.6686	0.6333	0.9474	0.1490

Although RF achieved the highest precision and specificity among the evaluated models, LGBM showed the best overall balance across the main metrics. XGB obtained lower performance but required the shortest execution time, suggesting a favorable trade-off between computational cost and predictive performance. Therefore, LGBM was selected as the best-performing model for this dataset, while XGB may be considered when execution time is prioritized. Figure 4 shows the SHAP explanations for the selected LGBM model. The beeswarm plot (left) summarizes global model behavior, indicating that SUIC_RISK and SUIC_ATTEMPT_HISTORY have the widest influence on the model's output across the cohort. The bar plot (right) shows a local explanation for a single case, decomposing that individual prediction into the signed contribution of each variable, with MENTAL_DISORDER_HISTORY pushing

**Figure 2:** Heatmap of dataset 1 features.

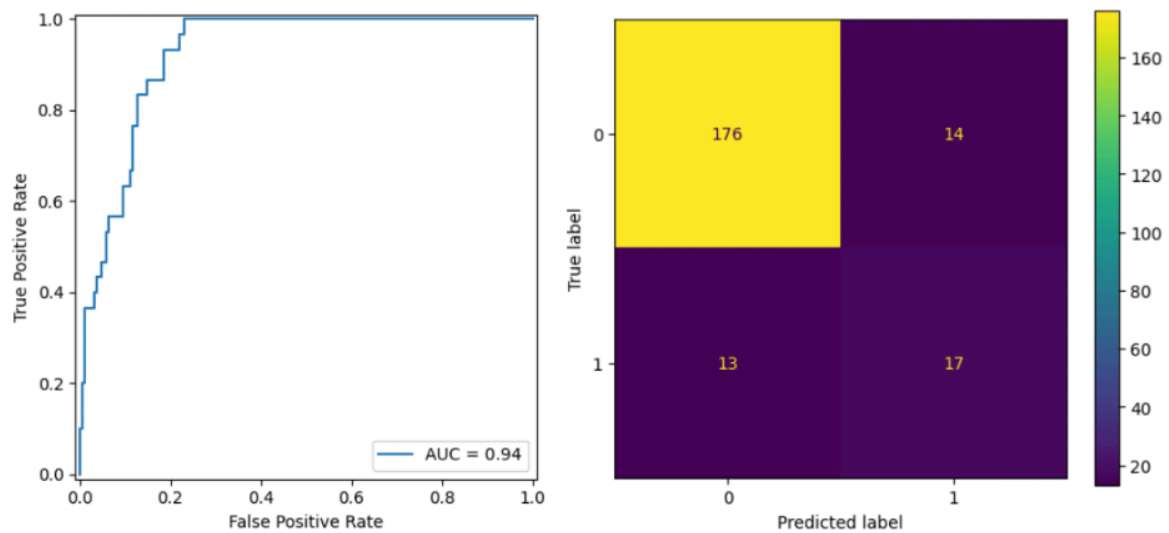


Figure 3: AUC for LGBM (left) and CM for LGBM (right).

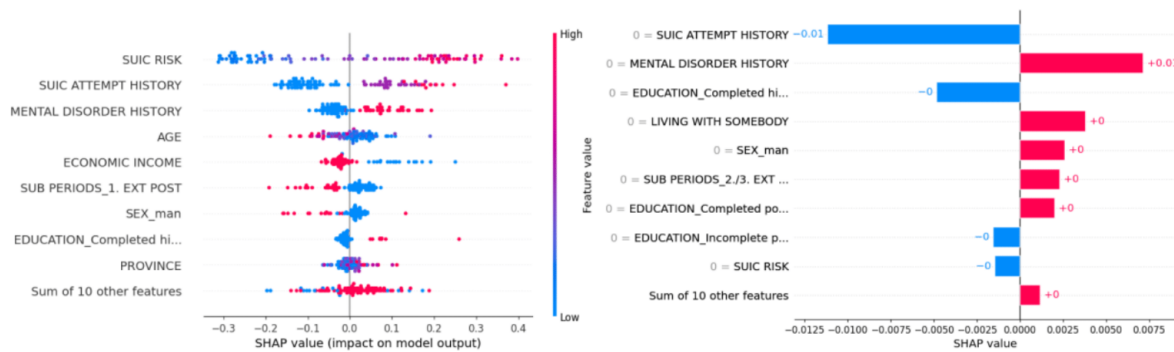


Figure 4: Global (beeswarm, left) and local single-case (right) SHAP explanations for Dataset 1.

the prediction toward the positive class and SUIC_ATTEMPT_HISTORY pushing it in the opposite direction.

3.2. Dataset 2 - Supervised

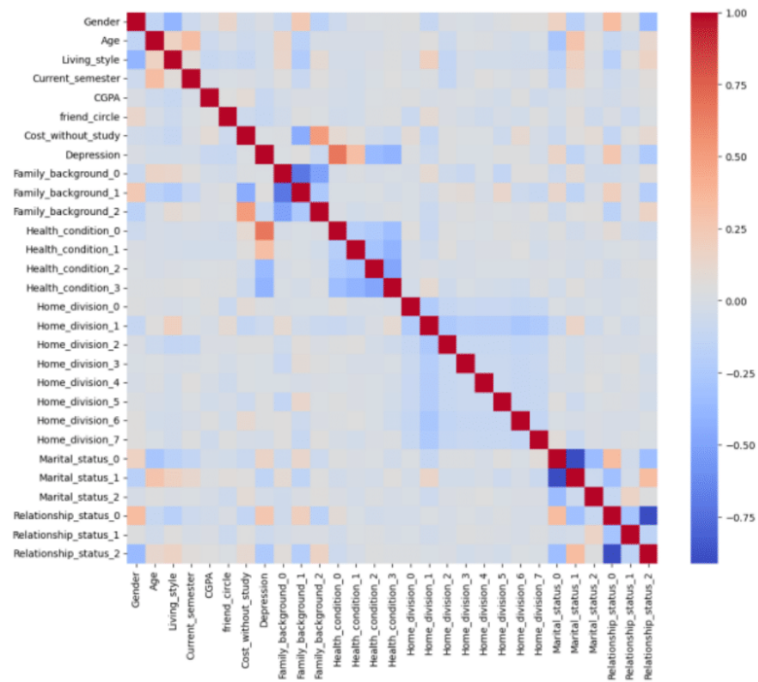
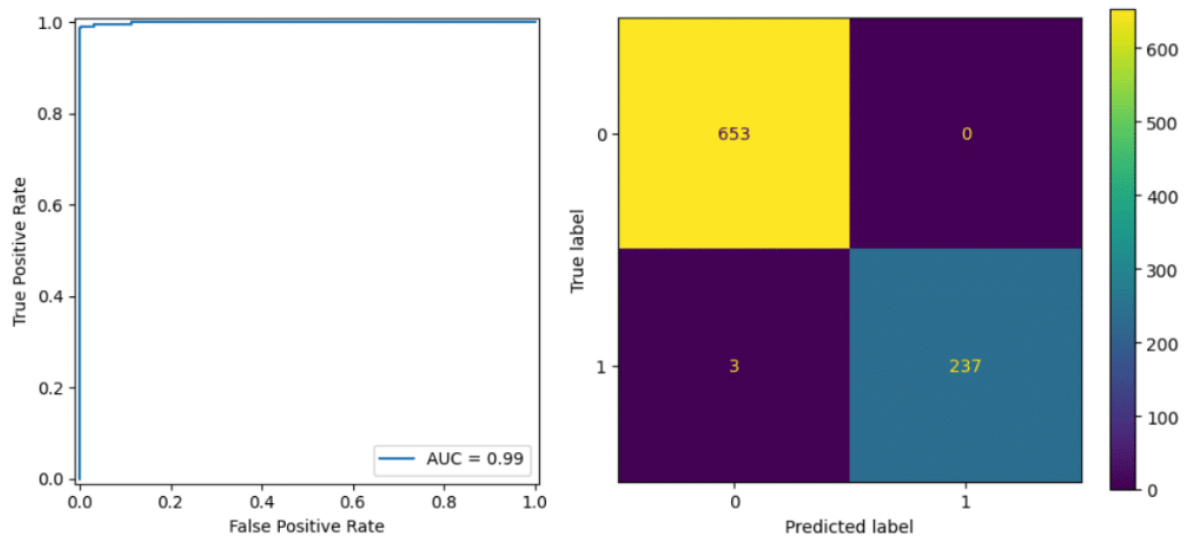
For the second supervised dataset, all models achieved high predictive performance, with AUC and accuracy values above 0.97 and F1-scores above 0.95. RF obtained the highest AUC value, reaching 0.9971, which indicates strong discriminative capacity. However, LGBM achieved the best overall balance across the main classification metrics, with an F1-score of 0.9766, accuracy of 0.9877, precision of 0.9916, recall of 0.9625, specificity of 0.9969, and the lowest execution time among the evaluated models. Therefore, LGBM was considered the most balanced and efficient model for this dataset, while RF stood out in terms of AUC. Figure 7 shows the corresponding SHAP explanations. In the beeswarm plot, the encoded values of Health_condition dominate the global attribution. The local explanation for a single case shows Health_condition_Normal and Health_condition_Good as the largest positive contributions to that prediction, while Health_condition_Critical, relationship status, and friend_circle contribute in the opposite direction.

In the supervised experiments, both datasets showed strong model performance. For Dataset 1, LGBM achieved the best overall balance, with the highest F1-score and accuracy among the evaluated models. For Dataset 2, LGBM obtained the best global classification performance, while RF achieved the highest AUC. These results suggest that the pipeline can compare different algorithms across datasets

Table 4

Performance comparison of supervised models for Dataset 2.

Model	F1	AUC	Accuracy	Precision	Recall	Specificity	Time (s)
LGBM	0.9766	0.9955	0.9877	0.9916	0.9625	0.9969	0.2642
XGB	0.9676	0.9932	0.9832	0.9833	0.9542	0.9939	0.3052
RF	0.9570	0.9971	0.9776	0.9840	0.9333	0.9938	0.9435

**Figure 5:** Heatmap of dataset 2 features.**Figure 6:** AUC for LGBM (left) and CM for LGBM (right).

and identify the most suitable model according to multiple performance criteria rather than relying on a single metric.

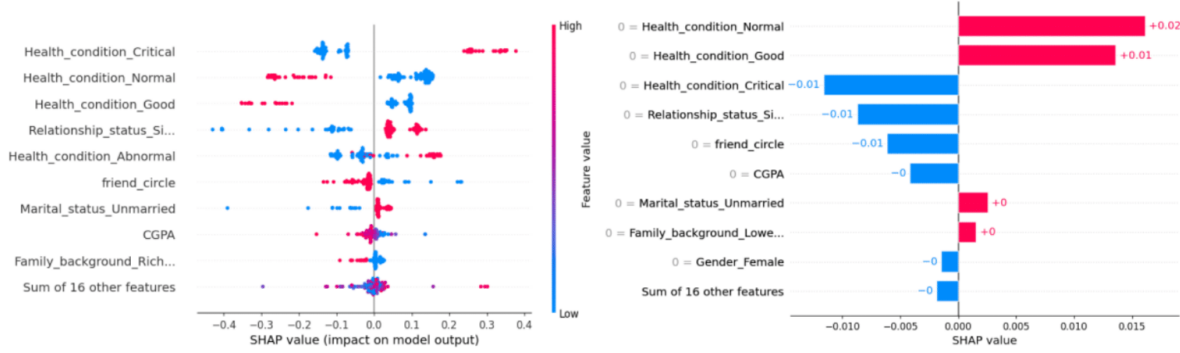


Figure 7: Global (beeswarm, left) and local single-case (right) SHAP explanations for Dataset 2.

The results show that model selection cannot be reduced to a single metric, since different algorithms may optimize different aspects of performance, such as discrimination capacity, recall, specificity, or execution time.

3.3. Dataset 3 - Unsupervised

In Dataset 3, both Gaussian Mixture and K-Means produced two-cluster partitions. The T-Test post-analysis showed that *inter_dom* was significant in both methods, suggesting that it consistently contributed to cluster separation. Additional variables related to guilt, loneliness, professional support, and religion-related dimensions also showed significant differences between clusters. These results should be interpreted as exploratory profiles rather than diagnostic categories. Thus, the unsupervised module supports physician-oriented explainability by transforming cluster assignments into interpretable group differences.

In both unsupervised datasets, the optimal number of clusters was not fixed in advance but determined by the pipeline through internal validation, which identified $k = 2$ as the preferred partition in each case. This coincidence is therefore a result of the analysis rather than a configuration parameter, and it indicates that both datasets are most stably separated into two broad profiles.

Table 5

Cluster post-analysis for Dataset 3.

Model	Silhouette	DBI	Calinski-Harabasz	T-Test
KM	0.09258	2.956	25.66	0.06173
GMM	0.09236	2.968	25.49	0.07407

3.4. Dataset 4 - Unsupervised

For the last, unsupervised dataset, K-Means showed the most acceptable clustering behavior among the evaluated configurations. In this case, the model also proposes 2 clusters as the optimal division. Although the Silhouette values were low, indicating limited cluster separation, K-Means obtained the best Davies-Bouldin scores and the highest Calinski-Harabasz values. In particular, the StandardScaler + K-Means configuration achieved the highest CHI value and the lowest DBI, suggesting a comparatively better-defined cluster structure than the Gaussian Mixture alternatives.

The Gaussian Mixture configurations showed weaker internal clustering performance, especially with StandardScaler, where the Silhouette value was close to zero, and the DBI was substantially higher. Although RobustScaler + Gaussian Mixture obtained the highest global T-Test score, its internal validation metrics were weaker than K-Means. Therefore, K-Means can be considered the most promising

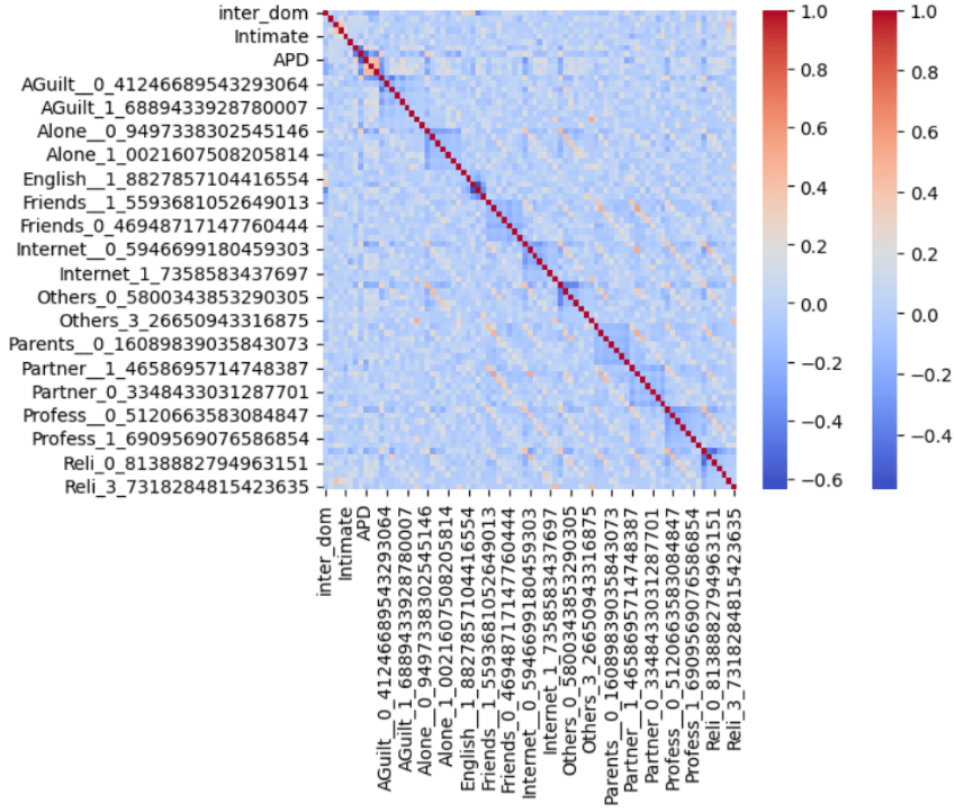


Figure 8: Heatmap of dataset 3 features.

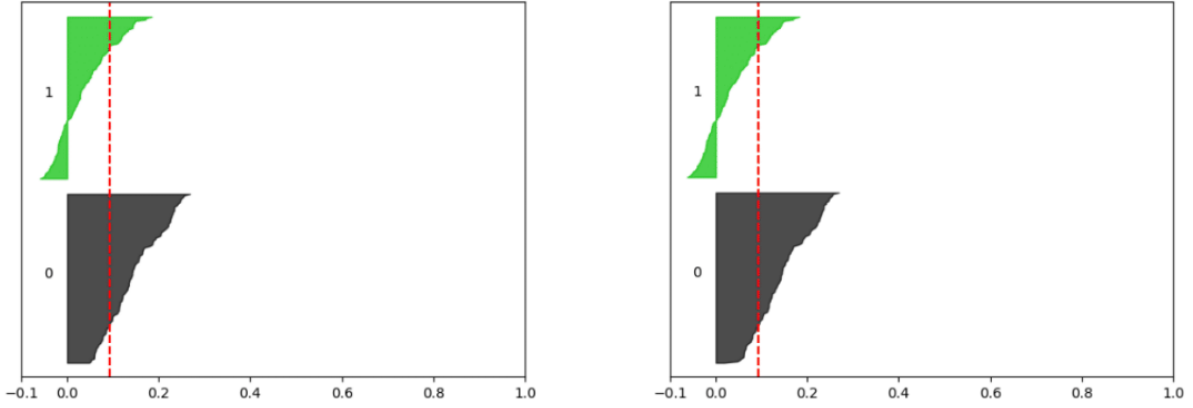


Figure 9: Silhouette scores for dataset 3 (KM left, GMM right).

configuration for this dataset, while the overall clustering quality remains moderate and should be interpreted cautiously.

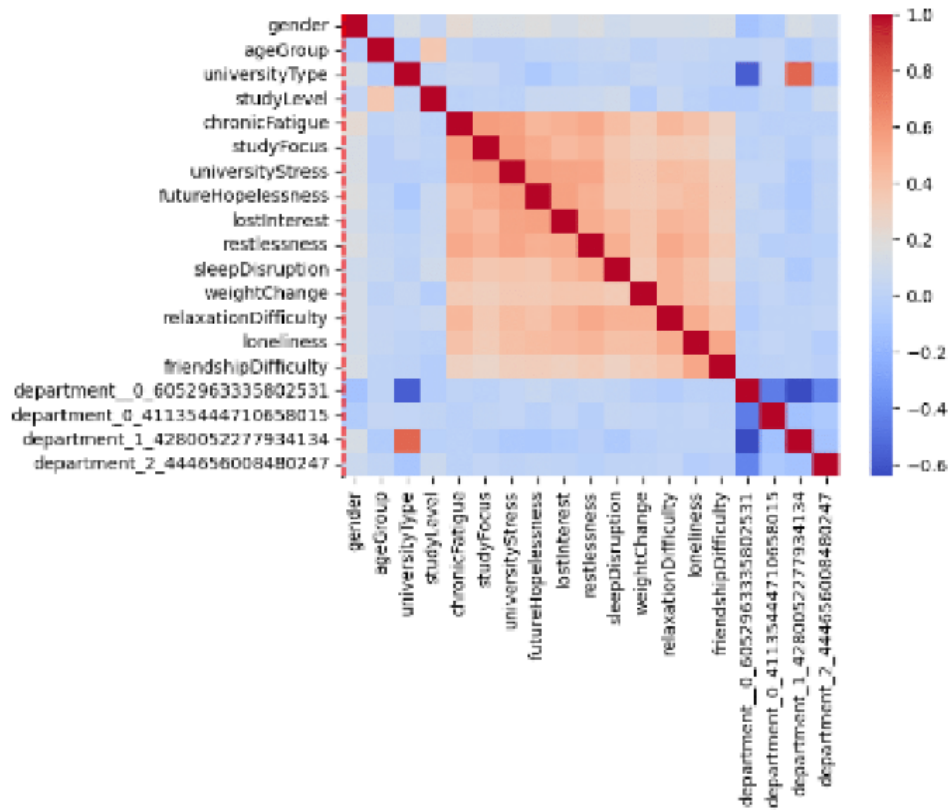
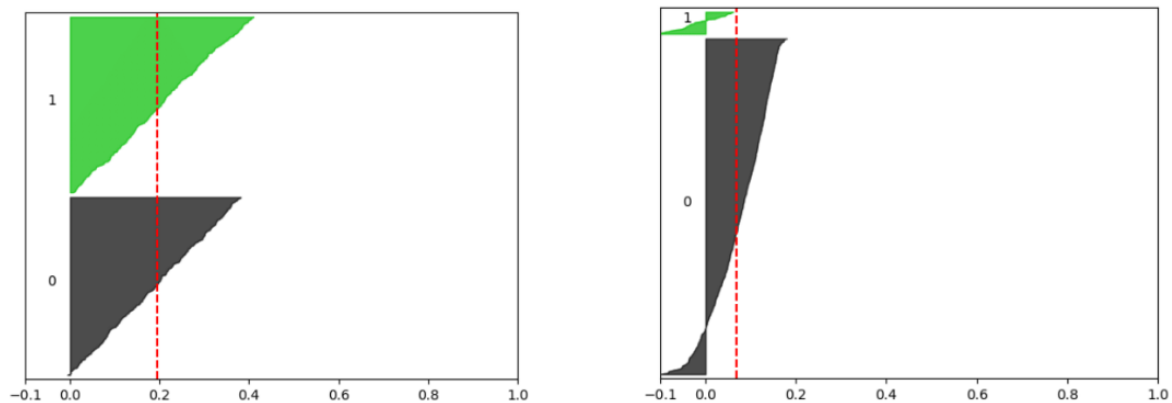
4. Discussion

The application of Explainable AI varies significantly depending on the underlying learning paradigm. In supervised learning, the existence of a definitive target variable allows feature impact to be attributed directly to specific predictions, which makes these models comparatively straightforward to explain. The pipeline leverages a threefold approach. SHAP [14] remains the reference method for comprehensive model auditing and for understanding global feature interactions, although its exhaustive output can be

Table 6

Cluster post-analysis for Dataset 4.

Model	Silhouette	DBI	Calinski-Harabasz	T-Test
KM	0.1956	1.774	242.5	0.6316
GMM	0.0689	5.587	6.148	0.2632

**Figure 10:** Heatmap of dataset 4 features.**Figure 11:** Silhouette scores for dataset 4 (KM left, GMM right).

cognitively demanding at the clinical point of care. LIME [16] and Anchor [17] are more pragmatic for real-time use: Anchor in particular provides deterministic, rule-based explanations expressed as sufficient conditions, which closely mirror clinical reasoning and allow a physician to validate a decision and communicate it to a patient. Reporting the three together is deliberate, since each answers a different

question, and agreement among them is itself informative about the stability of an explanation. SHAP was selected for this illustration because a single additive attribution framework yields both the global summary and the per-case decomposition reported above. LIME and Anchor are local by construction and are implemented in the pipeline, but their value lies in explaining individual predictions reviewed by a clinician, which the present evaluation does not include.

Explaining unsupervised models presents a greater challenge, because no ground truth exists to anchor feature importance. Explainability therefore shifts from tracing individual predictions to structural profiling of the discovered groups, relying on internal validation indices and statistical testing to interpret the geometric boundaries and possible clinical relevance of each cluster. The experimental results illustrate both the value and the limits of this approach: the pipeline consistently identified two-cluster partitions and isolated the variables differentiating them, but the low Silhouette values obtained for Datasets 3 and 4 indicate weak geometric separation. The resulting profiles should accordingly be read as exploratory hypotheses about population structure, not as diagnostic categories. A further observation concerns model selection. Across both supervised datasets, no single algorithm dominated on every metric: LightGBM achieved the best overall balance, Random Forest the highest AUC and precision, and XGBoost the shortest execution time. This supports the design decision to report a panel of metrics rather than optimize a single figure of merit, since the appropriate trade-off between discrimination, recall, specificity, and computational cost depends on the clinical context in which a model would be deployed.

Several limitations should be considered when interpreting these results. First, the evaluation covers four datasets drawn from questionnaire-based instruments in specific populations, so the reported performance figures are not intended as generalizable estimates of diagnostic accuracy. Second, the datasets vary considerably in size and feature count, so performance is not directly comparable across them. Third, the unsupervised results show weak cluster separation, and the aggregate statistical indicator used to summarize between-cluster differences is a descriptive measure without correction for multiple comparisons. Fourth, class imbalance is present in several of the datasets and was not addressed through resampling, which affects the interpretation of F1-scores in particular. Finally, and most importantly, the physician-oriented explanation layer has not yet been evaluated with clinicians: its usefulness is argued on design grounds rather than demonstrated empirically, and validation with healthcare professionals remains necessary before any claim of clinical utility can be made.

5. Conclusion

This work presented a metadata-driven normalization and explainable Machine Learning pipeline for mental health dataset analysis. Building on prior supervised, unsupervised, and systematic review studies, the pipeline extends the analytical workflow with a sixth stage focused on results explanation and externalizes dataset-specific decisions into declarative metadata so that heterogeneous sources follow a common and traceable execution path. The pipeline was evaluated on four datasets covering two supervised and two unsupervised scenarios. In the supervised experiments, it achieved strong predictive performance on both datasets and enabled comparison across algorithms using multiple evaluation criteria rather than a single metric, with LightGBM offering the best overall balance in each case. In the unsupervised experiments, the results were exploratory: K-Means showed the most promising behavior, particularly on Dataset 4, where it obtained better internal validation and post-analysis results than Gaussian Mixture, while overall cluster separation remained moderate. Taken together, the results suggest that the proposed pipeline improves traceability, comparability, and interpretability in mental health dataset analysis. Its principal contribution lies not in model execution but in the translation of technical outputs into physician-oriented explanations covering model behavior, relevant variables, data-quality warnings, and limitations. Future work will focus on refining the clinical explanation layer, validating it with healthcare professionals, addressing class imbalance explicitly, and extending the evaluation to additional and more diverse mental health datasets.

Declaration on Generative AI

During the preparation of this work, the authors used Claude to improve the writing style of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] World Health Organization, Depressive disorder (depression), 2023. URL: <https://www.who.int/news-room/fact-sheets/detail/depression>.
- [2] K. Kroenke, R. L. Spitzer, The phq-9: a new depression diagnostic and severity measure, 2002.
- [3] A. T. Beck, Manual for the beck depression inventory-ii, (No Title) (1996).
- [4] L. Breiman, Random forests, *Machine learning* 45 (2001) 5–32.
- [5] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [6] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, *Advances in neural information processing systems* 30 (2017).
- [7] M. Di Felice, A. Deroche, I. Trupkin, P. Chatterjee, M. F. Pollo-Cattaneo, Predictive modeling for detection of depression using machine learning, in: *International Conference on Applied Informatics*, Springer, 2023, pp. 47–57.
- [8] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics* 20 (1987) 53–65.
- [9] D. L. Davies, D. W. Bouldin, A cluster separation measure, *IEEE transactions on pattern analysis and machine intelligence* (1979) 224–227.
- [10] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, *Communications in Statistics-theory and Methods* 3 (1974) 1–27.
- [11] M. Di Felice, A. Deroche, I. Trupkin, P. Chatterjee, M. F. P. Cattaneo, Depression and anxiety diagnosis using unsupervised learning approach., in: *ICAI Workshops*, 2023, pp. 12–24.
- [12] M. Di Felice, I. Trupkin, A. Deroche, M. F. Pollo Cattaneo, P. Chatterjee, Depression diagnosis using artificial intelligence: a systematic review, *F1000Research* 13 (2024) 1549.
- [13] M. Di Felice, A. Deroche, I. Trupkin, M. F. Pollo-Cattaneo, P. Chatterjee, A modular dual-pipeline machine learning framework for supervised and unsupervised depression analysis, 2026.
- [14] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *Advances in neural information processing systems* 30 (2017).
- [15] J. Hernandez, M. F. Pollo-Cattaneo, D. H. Peluffo-Ordóñez, H. Florez, Cognitive architecture for learning in autonomous agents: A study in discrete navigation, in: *Communications in Computer and Information Science*, Springer, 2025, pp. 3–19. URL: https://link.springer.com/10.1007/978-3-032-07175-0_1. doi:10.1007/978-3-032-07175-0_1.
- [16] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?" explaining the predictions of any classifier, in: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [17] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, 2018.
- [18] L. C. López Steinmetz, M. A. Dutto Florio, C. A. Leyes, S. B. Fong, A. Rigalli, J. C. Godoy, Levels and predictors of depression, anxiety, and suicidal risk during covid-19 pandemic in argentina: the impacts of quarantine extensions on mental health state, *Psychology, Health & Medicine* 27 (2022) 13–29.
- [19] O. Haque, S. Alom, Student depression detection dataset, 2026.
- [20] M.-H. Nguyen, M.-T. Ho, Q.-Y. T. Nguyen, Q.-H. Vuong, A dataset of students' mental health and help-seeking behaviors in a multicultural environment, *Data* 4 (2019) 124.
- [21] J. Hossain, M. R. A. Rahi, M. S. I. Ovi, Minimal symptom assessment of depression severity in university students, 2026.