JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
http://jurnal.iaii.or.id/index.php/JOSEIT

Quantifying Target Leakage in Rule-Derived Mental Health Screening Labels

Rozi Meri[1*], Ikhsan[2], Dian Eka Putra[3], Sofia Yosse[4], Ismael[5]
[12345]Department of Information Technology, Politeknik Negeri Padang, Padang, Indonesia

[*]*Corresponding author email: rozimeri@pnp.ac.id*

Abstract

Publicly hosted tabular datasets increasingly package screening-questionnaire items alongside computed, not clinically observed, risk and severity labels. This creates a risk of target leakage: if a label is a deterministic function of the same features given to a classifier, reported predictive performance measures rule reconstruction, not screening validity. We audit a recent IEEE DataPort dataset (n = 2,005) scoring 27 Likert items across five conditions – ADHD, Autism Spectrum Disorder (ASD), Social Pragmatic Communication Disorder (SPCD), Depression, and Anxiety – each with a binary risk flag and a three-level severity label. The dataset's documentation discloses that every label is generated from a threshold or tertile rule on the summed items of its own condition, yet its usage instructions recommend fitting Random Forest or XGBoost with SHAP directly on the 27 raw items. We verify the documented rule empirically, then compare a trivial single-feature decision stump (the summed score) against a 300-tree Random Forest trained on all 27 items, under identical 5-fold stratified cross-validation, across all ten condition-by-task combinations (5 conditions × {risk, severity}). The single-feature stump matches or exceeds the 27-feature ensemble in every combination, reaching 100% cross-validated accuracy on nine of ten targets. Random Forest feature importance further shows 72–88% of predictive weight concentrated on each condition's own items. These results indicate the pipeline the dataset recommends reconstructs a disclosed arithmetic rule rather than learning a clinically meaningful pattern; any predictive-validity study on this dataset should exclude a condition's own summed items from that condition's feature set.

Keywords: target leakage, label circularity, mental health screening, construct validity, random forest

Cite as: Rozi Meri, Ikhsan, Dian Eka Putra, Yosse, S., & Ismael. (2026). Quantifying Target Leakage in Rule-Derived Mental Health Screening Labels. Journal of Systems Engineering and Information Technology (JOSEIT), 3(3). https://doi.org/10.29207/joseit.v3i3.8470

Received by the Editor: 8-10-2024
***Final Revision*:** 15-12-2024
***Published*:** 30-12-2024

This is an open-access article under the CC BY-4.0 license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Machine learning for mental-health screening increasingly relies on tabular, questionnaire-derived datasets rather than clinician-adjudicated cohorts. This lowers the barrier to benchmarking, but it also reintroduces a longstanding problem in applied predictive modeling: target leakage, defined as the unintentional inclusion of information about the label in the feature representation available to the model [6]. Leakage inflates reported performance without producing a model that generalizes to any setting where the leaking measurement is unavailable [6], [8]. In clinical and biomedical machine learning specifically, leakage has been documented for tasks ranging from antihypertensive-drug-based hypertension prediction to sepsis detection from the very vital-sign thresholds used to define sepsis [8], [9]. A related but more precise failure mode is circularity: when a target label is deterministically computable from a subset of the model's own input features, a sufficiently flexible classifier can reconstruct the label-generating function itself rather than learn any generalizable structure, producing what has been termed "false perfection" [7].

A recent IEEE DataPort dataset packages exactly this configuration for college-student mental-health screening [10]. It provides 2,005 respondents' answers to a 27-item, 0–4 Likert instrument spanning five conditions – ADHD, ASD, SPCD, Depression, and Anxiety – with item sections stated to align with recognized screening constructs: the Adult ADHD Self-Report Scale, ASRS-v1.1 [1]; the Autism-Spectrum Quotient, AQ [2]; pragmatic-language markers; the Patient Health Questionnaire-9, PHQ-9 [3]; and the Generalized Anxiety Disorder-7, GAD-7 [4]. For every condition, the dataset also supplies a binary risk flag and a three-level severity label. Critically, the dataset's documentation states that these labels are derived from threshold and tertile rules applied to each condition's own summed item scores, while its usage guidance simultaneously recommends fitting Random Forest, XGBoost, or a compact neural network on the 27 raw items to predict these same labels, with SHAP explanation of the result [10], [17]. No leakage caveat accompanies this recommendation – precisely the configuration flagged as high-risk by recent clinical machine-learning leakage frameworks [8], [9], here applied to a screening-questionnaire release rather than an electronic health record.

We audit this dataset empirically rather than relying on its documentation alone. We ask: (1) do the class-conditional summed-score ranges for risk and severity actually separate as cleanly as the documented rule implies; (2) does a Random Forest trained on all 27 raw items – the pipeline the dataset itself recommends – outperform a trivial single-feature rule built only from the relevant summed score; and (3) where a Random Forest is fit, does its feature importance concentrate on a condition's own items, consistent with rule reconstruction, or spread across conditions, consistent with genuine comorbidity signal. Because the item counts used here (four to seven items per condition) are abbreviated relative to the full instruments they are "aligned to" – the ASRS-v1.1 screener alone uses six to eighteen items [1], the AQ uses fifty [2], the PHQ-9 uses nine [3], and the GAD-7 uses seven [4] – we treat the released dataset as a synthetic, clinically inspired screening simulacrum rather than a validated instrument, and confine our claims to the internal statistical structure of its released labels.

The clinical stakes behind this audit are not abstract. Common mental disorders are highly prevalent among the population this dataset targets: the WHO World Mental Health Surveys International College Student Project found that a substantial share of first-year university students across eight countries met criteria for at least one common mental disorder in the past year [11], and ADHD and ASD are themselves frequently comorbid, sharing both phenotypic overlap and, per recent clinical reviews, partially shared etiological pathways [12] -- which is precisely the kind of cross-condition signal a legitimate predictive model would need to recover, and precisely what Section 3.4 shows the audited pipeline does not do. A parallel scoping review of machine-learning applications in mental health has catalogued hundreds of published studies applying classifiers of exactly this kind -- Random Forest, XGBoost, and related tree ensembles -- to screening and severity-prediction tasks [13], the application context in which a circularity failure like the one documented here would propagate silently into a published literature.

The broader reproducibility literature suggests this is not an isolated incident. A systematic survey of leakage across seventeen scientific fields that have adopted machine learning found leakage-driven overoptimism in hundreds of published studies, arising from causes ranging from textbook train/test contamination to subtler circularity of the kind audited here [14]. Structured dataset documentation, of the kind proposed under the "datasheets for datasets" framework, is intended to surface precisely this class of provenance risk before a dataset reaches downstream users [15]; the present dataset's own "Derivation of Labels" disclosure is, in effect, an unusually candid instance of such documentation, which is what makes an empirical audit against it tractable in the first place. The dataset's blend of synthetic and real respondents (1,975 and 30, respectively) further situates it within a growing synthetic-health-data literature that stresses evaluating synthetic releases against the same validity criteria as real clinical data, rather than assuming synthetic construction alone guarantees downstream usability [16].

2. Method

All analysis was performed in Python 3 (pandas, scikit-learn) on the released CSV. The protocol below is fully reproducible from the dataset's public DOI [10].

2.1 Dataset

The dataset (DOI: 10.21227/krk9-qk90) contains 2,005 respondents aged 18–25: 1,975 synthetically generated respondents plus 30 anonymized real student responses scored with the same instrument [10]. Each respondent answers 27 Likert items (0 = never to 4 = very often) partitioned across five conditions: ADHD (7 items, sum range 0–28), ASD (6 items, 0–24), SPCD (4 items, 0–16), Depression (5 items, 0–20), and Anxiety (5 items, 0–20). For each condition the release includes a binary risk flag and a three-level severity label (Low / Medium / High). No missing values are present in any of the 2,005 × 40 columns.

2.2 Documented Label-Generation Rule

The dataset's accompanying documentation discloses the exact label-generation formula: risk = 1 when a condition's summed item score reaches a fixed threshold (ADHD ≥19/28, ASD ≥16/24, SPCD ≥11/16, Depression ≥13/20, Anxiety ≥14/20), and severity is set by tertile of the maximum possible score (Low < 33%, Medium 33–66%, High > 66%) [10]. Both rules take only the condition's own items as input; no cross-condition information enters either formula. The same documentation recommends Random Forest, XGBoost [17], or a compact neural network on the full 27-item vector, explained with SHAP, as the modeling approach, and does not flag the resulting configuration as leakage-prone.

2.3 Empirical Verification Protocol

We verified the documented rule directly against the released labels rather than assuming it holds exactly, since the same documentation states that noise was injected to reproduce realistic symptom correlations and comorbidity overlap. For each condition we computed the summed score of its own items for all 2,005 respondents and examined the class-conditional minimum and maximum of that sum for every level of the risk flag and the severity label. We then fit two classifier families under identical stratified 5-fold cross-validation (scikit-learn, random_state = 42): (i) a depth-2 decision-tree stump using only the condition's summed score as input – the

simplest model capable of reconstructing a two- or three-way threshold rule; and (ii) a 300-tree Random Forest [5] using all 27 raw items as input – the pipeline the dataset's documentation recommends. Both were evaluated identically for the risk flag (binary) and the severity label (three-class), and compared against a majority-class baseline. We additionally fit a Random Forest on the full item set (no cross-validation split, for feature-importance estimation only) and computed, for each condition's risk model, the share of total impurity-based feature importance attributable to that condition's own items versus the remaining 20–23 items belonging to the other four conditions.

2.4 Evaluation Metrics

For the risk flag we report cross-validated accuracy, F1 on the positive class, and area under the ROC curve (AUC). For severity we report cross-validated accuracy and Cohen's kappa [20], which corrects for class imbalance (majority-class share ranges from 41% to 57% across conditions) and is interpreted against the conventional benchmark scale in which $\kappa > 0.81$ is considered "almost perfect" agreement [19]. All folds preserve the class distribution of the target.

2.5 Overview of the Audit Workflow

Fig. 1 summarizes Sections 2.1-2.4 as a six-stage protocol organized into three verification layers: a provenance layer that extracts the dataset's own disclosed label-generation rule (Stage 1), a structural layer that empirically confirms class-conditional separability against that rule (Stage 2), and a target-validity layer that builds the trivial and recommended-pipeline comparator models, evaluates both under identical cross-validation, computes the metrics of Section 2.4, and attributes Random Forest feature importance (Stages 3-6). The three-layer structure mirrors the provenance / structural / target-validity distinction used throughout this audit series and separates claims about what the dataset documentation states from claims verified empirically against the released data.



Figure 1. Modeling Pipeline

3. Results and Discussion

3.1 Class-Conditional Separability of the Risk Label

Table 1 compares the documented risk threshold against the empirical class-conditional sum ranges. For four of the five conditions – ADHD, ASD, Depression, and Anxiety – the risk = 0 and risk = 1 sum ranges do not overlap at all: a single number, the item sum, perfectly separates the two classes. SPCD is the only condition with any overlap: 138 of 2,005 respondents (6.9%) fall in the boundary zone (sum 11–16) where both risk = 0 and risk = 1 occur, consistent with the dataset's own note that noise was injected into the label-generation process. Even within this noisy zone, the best single threshold on the sum recovers 99.55% of SPCD risk labels. Severity was separable without exception: for all five conditions, the tertile cut-points implied by the documentation produced zero overlap between adjacent severity levels.

Table 1. Documented and Empirically Verified Risk-Label Rule

Condition	Items	Sum range	Threshold	Risk = 0 range	Risk = 1 range	Overlap (n)
ADHD	7	0–28	≥ 19	0–18	19–28	0
ASD	6	0–24	≥ 16	0–15	16–24	0
SPCD	4	0–16	≥ 11	0–16	11–16	138
Depression	5	0–20	≥ 13	0–12	13–20	0
Anxiety	5	0–20	≥ 14	0–13	14–20	0

3.2 A Trivial Single-Feature Rule Reconstructs Both Labels

Table 2 and Figure 2 report cross-validated risk-classification performance. The sum-only decision stump reaches 100.00% accuracy, F1, and AUC for ADHD, ASD, Depression, and Anxiety, and 99.55% accuracy / 0.997 AUC for SPCD – essentially reproducing Table 1's class-conditional ranges under cross-validation rather than in-sample fitting. Table 3 and Figure 3 report severity classification: the sum-only stump reaches 100.00% cross-validated accuracy and a Cohen's kappa of 1.000 for all five conditions -- the ceiling of the scale, and well above the $\kappa > 0.81$ threshold conventionally read as "almost perfect" agreement [19].

Table 2. Risk Classification: 5-Fold Cross-Validated Performance

Condition	Majority (%)	Stump Acc (%)	Stump F1	Stump AUC	RF-27 Acc (%)	RF-27 F1	RF-27 AUC
ADHD	89.18	100.00	1.000	1.000	98.25	0.914	0.999
ASD	93.42	100.00	1.000	1.000	98.15	0.838	0.999
SPCD	93.57	99.55	0.966	0.997	98.35	0.857	0.998
Depression	81.10	100.00	1.000	1.000	99.30	0.981	1.000
Anxiety	81.40	100.00	1.000	1.000	98.70	0.964	0.999

Table 3. Severity Classification: 5-Fold Cross-Validated Performance

Condition	Majority (%)	Stump Acc (%)	Stump κ	RF-27 Acc (%)	RF-27 κ
ADHD	51.22	100.00	1.000	95.76	0.927
ASD	54.41	100.00	1.000	94.66	0.902
SPCD	57.46	100.00	1.000	96.41	0.933
Depression	47.88	100.00	1.000	97.01	0.951
Anxiety	41.00	100.00	1.000	96.96	0.952

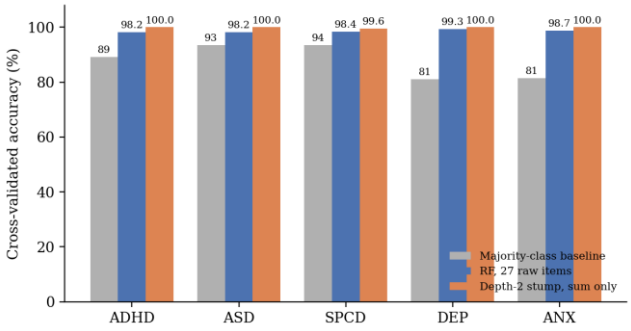


Figure 2. Cross-Validated Risk-Classification Accuracy

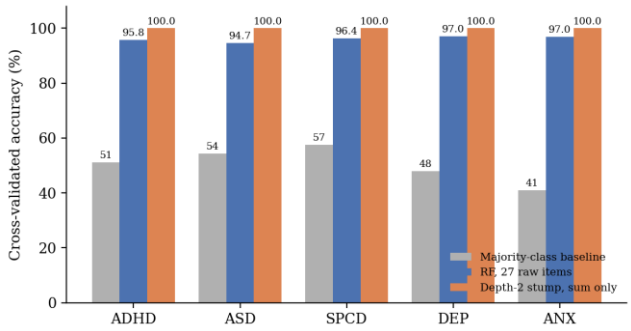


Figure 3. Cross-Validated Severity-Classification Accuracy

3.3 The 27-Feature Ensemble Never Outperforms the Trivial Rule

The comparison in Tables 2 and 3 yields a consistent, and initially counter-intuitive, pattern: in all ten condition-by-task combinations, the 300-tree Random Forest trained on the full 27-item vector performs at or below the one-feature decision stump – it never wins. For ADHD risk, for example, the Random Forest reaches 98.25% cross-validated accuracy against the stump's 100.00%; for ADHD severity, 95.76% against 100.00% with kappa falling from 1.000 to 0.927.

This is not a failure of the Random Forest so much as a property of the task: because each split in a Random Forest considers only a random subset of the 27 items, individual trees rarely see the exact linear combination that defines each condition's summed score, so bagging can only approximate – not exactly reconstruct – a deterministic threshold function that a two-node decision stump recovers exactly. The practical implication is that the 98–99% accuracy such an ensemble reports on this dataset is not evidence of predictive signal beyond the summed score; it is a noisier reconstruction of the same signal, obtained at far greater model complexity.

3.4 Where the Ensemble's Importance Concentrates

Impurity-based feature importance from the full-data Random Forest confirms that its residual accuracy gap relative to the stump is not being spent on genuine cross-condition (comorbidity) signal. Each condition's own items account for the large majority of total importance in that condition's risk model: 85.7% for ADHD, 78.5% for ASD, 72.1% for SPCD, 87.9% for Depression, and 87.6% for Anxiety. SPCD shows the lowest own-domain share and is also the only condition with a documented noisy risk boundary (Section 3.1); this is consistent with the Random Forest partially compensating for that noise using items from comorbid conditions, rather than discovering new predictive structure. Across all five conditions, the 20–23 “extra” items available to the ensemble but absent from the trivial rule contribute only secondary importance.

3.5 Implication for Downstream Use of This Dataset

Any study that follows the dataset's own recommended pipeline – Random Forest or XGBoost on the 27 items, explained with SHAP – to predict risk or severity will report near-perfect metrics. Sections 3.2–3.4 show that these metrics measure how well the model reconstructs a two-parameter rule already published in the dataset's documentation, not clinical predictive validity in the sense evaluated by the original screening instruments' own validation studies, where, for comparison, the ASRS Screener's concordance with clinician diagnosis reached an AUC of 0.90 against an independent clinical interview [1].

A SHAP explanation fitted to such a model would attribute importance to the same own-domain items that define the label, which looks clinically plausible on inspection but is tautological under the circularity framework of Hagmann and Riezler [7]; this is consistent with independent work showing that post hoc explanation methods including SHAP can be made to produce plausible-looking attributions that misrepresent what a model actually relies on [18]. We therefore treat the dataset's “Recommended Analysis Tools & Models” guidance and its “Derivation of Labels” disclosure as, in effect, a single document: the latter functions as a mandatory feature-exclusion list for the former, not as independent background reading.

4. Conclusion

This audit shows that the risk and severity labels released with a recent college-student mental-health screening dataset are, with one minor and documented exception, exact deterministic functions of the same 27 Likert items the dataset's own instructions recommend as classifier input. A single-feature decision stump built from each condition's summed score reconstructs the risk flag with 99.6–100% cross-validated accuracy and the severity label with 100% accuracy and a perfect Cohen's kappa across all five conditions, matching or exceeding a 300-tree Random Forest trained on the full item set in every one of ten condition-by-task comparisons. Random Forest feature importance further shows that 72–88% of predictive weight concentrates on each condition's own items, indicating that even the additional cross-condition information available to the ensemble contributes little beyond the disclosed rule. Reported predictive performance from the pipeline the dataset itself recommends therefore measures reconstruction of a published arithmetic rule rather than screening validity. We recommend that any predictive-validity study built on this dataset exclude a condition's own summed items from that condition's feature set, or predict against a criterion external to the release – for instance, one condition's severity from only the other four conditions' items – which would test the dataset's injected comorbidity structure rather than its label-generation arithmetic.

References

[1] R. C. Kessler et al., “The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short screening scale for use in the general population,” Psychological Medicine, vol. 35, no. 2, pp. 245–256, 2005, doi: 10.1017/S0033291704002892.

[2] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, “The Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians,” Journal of Autism and Developmental Disorders, vol. 31, no. 1, pp. 5–17, 2001, doi: 10.1023/A:1005653411471.

[3] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: validity of a brief depression severity measure," Journal of General Internal Medicine, vol. 16, no. 9, pp. 606–613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.
[4] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the GAD-7," Archives of Internal Medicine, vol. 166, no. 10, pp. 1092–1097, 2006, doi: 10.1001/archinte.166.10.1092.
[5] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[6] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: formulation, detection, and avoidance," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, Art. 15, 2012, doi: 10.1145/2382577.2382579.
[7] M. Hagmann and S. Riezler, "False perfection in machine prediction: detecting and assessing circularity problems in machine learning," arXiv:2106.12417, 2021, doi: 10.48550/arXiv.2106.12417.
[8] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A review of challenges and opportunities in machine learning for health," arXiv:1806.00388, 2018 (also in AMIA Joint Summits on Translational Science Proceedings, vol. 2020, pp. 191–200, 2020), doi: 10.48550/arXiv.1806.00388.
[9] S. E. Davis, M. E. Matheny, S. Balu, and M. P. Sendak, "A framework for understanding label leakage in machine learning for health care," Journal of the American Medical Informatics Association, vol. 31, no. 1, pp. 274–280, 2024, doi: 10.1093/jamia/ocad178.
[10] S. Indish and M. Dangore, "College-Student Mental Health Screening Dataset: ADHD, ASD, SPCD, Depression & Anxiety (Likert, Multi-Label)," IEEE DataPort, 2026, doi: 10.21227/krk9-qk90.
[11] R. P. Auerbach et al., "WHO World Mental Health Surveys International College Student Project: prevalence and distribution of mental disorders," Journal of Abnormal Psychology, vol. 127, no. 7, pp. 623–638, 2018, doi: 10.1037/abn0000362.
[12] K. M. Antshel and N. Russo, "Autism spectrum disorders and ADHD: overlapping phenomenology, diagnostic issues, and treatment considerations," Current Psychiatry Reports, vol. 21, no. 5, Art. 34, 2019, doi: 10.1007/s11920-019-1020-5.
[13] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," Psychological Medicine, vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151.
[14] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, Art. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
[15] T. Gebru et al., "Datasheets for datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.
[16] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," BMC Medical Research Methodology, vol. 20, Art. 108, 2020, doi: 10.1186/s12874-020-00977-1.
[17] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[18] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods," in Proc. AAAI/ACM Conf. on AI, Ethics, and Society, 2020, pp. 180–186, doi: 10.1145/3375627.3375830.
[19] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159–174, 1977, doi: 10.2307/2529310.
[20] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.