Predicting bipolar disorder incidence in young adults using gradient boosting: a 5-year follow-up study

Bruno Braga Montezano[a,b,*], Vanessa Gnielka[b], Augusto Ossamu Shintani[a,b], Kyara Rodrigues de Aguiar[a,b], Thiago Henrique Roza[a], Taiane de Azevedo Cardoso[c], Luciano Dias de Mattos Souza[d], Fernanda Pedrotti Moreira[d], Ricardo Azevedo da Silva[d], Thaíse Campos Mondin[d], Karen Jansen[d], Ives Cavalcante Passos[a,b]

[a]*Graduate Program in Psychiatry and Behavioral Sciences, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil*
[b]*Molecular Psychiatry Laboratory, Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil*
[c]*Mood Disorders Program, Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Canada*
[d]*Graduate Program in Health and Behavior, Catholic University of Pelotas, Pelotas, Rio Grande do Sul, Brazil*

Abstract

This study aimed to develop a classification model predicting incident bipolar disorder (BD) cases in young adults within a 5-year interval, using sociodemographic and clinical features from a large cohort study. We analyzed 1,091 individuals without BD, aged 18 to 24 years at baseline, and used the XGBoost algorithm with feature selection and oversampling methods. Forty-nine individuals (4.49%) received a BD diagnosis five years later. The best model had an acceptable performance (test AUC: 0.786, 95% CI: 0.686, 0.887) and included ten features: feeling of worthlessness, sadness, current depressive episode, self-reported stress, self-confidence, lifetime cocaine use, socioeconomic status, sex frequency, romantic relationship, and tachylalia. We performed a permutation test with 10,000 permutations that showed the AUC from the built model is significantly better than random classifiers. The results provide insights into BD as a latent phenomenon, as depression is its typical initial manifestation. Future studies could monitor subjects during other developmental stages and investigate risk populations to improve BD characterization. Furthermore, the usage of digital health data, biological, and neuropsychological information and also neuroimaging can help in the rise of new predictive models.

Keywords: Bipolar Disorder, Supervised Machine Learning, Incidence, Risk Factors, Young Adult, Cohort Studies

*Corresponding author
Email address: bmontezano@hcpa.edu.br (Bruno Braga Montezano)

Preprint submitted to MedRχiv

1. Introduction

Bipolar disorder is a chronic psychiatric condition associated with high morbidity and mortality (McIntyre et al., 2020). The global lifetime prevalence of bipolar disorder is approximately 2.4%, being 0.6% for bipolar I disorder, 0.4% for bipolar II disorder and 1.4% for individuals with subthreshold presentations (Merikangas et al., 2011). Previous investigations describe that these patients present a significant reduction of life expectancy of about 10 years relative to the general population (Kessing et al., 2015). Cardiovascular disease is the main factor associated with premature mortality in bipolar disorder; nonetheless, deaths by suicide are more commonly reported in bipolar disorder than in other mental health conditions, with these patients presenting a twenty to thirty times higher chance of dying by suicide (McIntyre et al., 2020; Plans et al., 2019; Kessing et al., 2015). In addition, bipolar disorder patients present significant functional and psychosocial impairment, also representing an important economic cost (McIntyre et al., 2020). For instance, evidence from the United States described that the total costs associated with bipolar I disorder exceeded $200 billion in the year of 2015 (Cloutier et al., 2018).

Even though the majority of the patients with BD present clinical symptoms before the age of 25, there is a significant delay of 6-10 years between the onset of the symptoms and the correct diagnosis (Yatham et al., 2018; Scott & Leboyer, 2011). Furthermore, delayed diagnosis is associated with longer duration of untreated illness, which is ultimately linked to a poorer prognosis in terms of hospitalizations, functioning and recurrence of episodes (Altamura et al., 2015). High rates of psychiatric comorbidity, difficulty in the differential diagnosis, usual onset with depressive symptoms, and reduced help-seeking behavior are some of the reasons for the delay in the proper recognition of bipolar disorder (McIntyre & Calabrese, 2019; McIntyre et al., 2020; Yatham et al., 2018). Nevertheless, the diagnosis of bipolar disorder is eminently clinical, with limited evidence to support the use of neuroimaging or laboratory biomarkers during clinical investigation (McIntyre et al., 2020).

Taking into account this context, the rise of the concept of precision psychiatry, with the use of big data and machine learning tools represents a promise, which may ultimately bring a revolution in terms of diagnosis, treatment selection and prognosis in the field of mental health (Fernandes et al., 2017; Passos et al., 2016). To this date several studies have explored the use of these techniques in bipolar disorder, based on distinct data sources (including neuroimaging, clinical and sociodemographic data, peripheral biomarkers, neuropsychological tests, genetics, among others), with the majority of these models being focused on classification tasks that help in the differential diagnosis between bipolar disorder and other psychiatric conditions such as schizophrenia, major depression and healthy individuals (Librenza-Garcia et al., 2017; Passos et al., 2019). Nevertheless, most of these studies present modest classification performances, are based on small and clinical samples, originary from cross-sectional procedures of data collection, or present short periods of follow-up (Librenza-Garcia et al., 2017). All these limitations may compromise the generalizability

and the translation of the results of such investigations to clinical and public health settings (Passos et al., 2019).

Thus, considering these gaps, the present study aims to create a binary classification model capable of predicting incident cases of bipolar disorder in a 5-year interval through sociodemographic and clinical features in a sample of young adults, from a large and population-based cohort study.

2. Methods

2.1. Participants

This was a prospective cohort study that collected sociodemographic and clinical information from a population-based sample of young adults aged between 18 and 24 years, living in the urban area of the city of Pelotas, located in southern Brazil. The first phase took place between 2007 and 2009, and the sample was selected through cluster sampling, considering eighty-nine randomly selected census-based sectors from 448 total sectors (Brazilian Institute of Geography and Statistics, 2010).

The following inclusion criteria were considered at baseline: (1) age between 18 and 24 years old; (2) live in the urban area. Severe cognitive disability (assessed through clinical judgement) that could cause difficulties in understanding study instruments was considered the only exclusion criteria. All eligible subjects (n = 1762) were invited to participate, of which 1560 accepted and consented to participate. Trained interviewers conducted a face-to-face interview at the participants' homes, so that data confidentiality was ensured. Data were collected through printed paper questionnaires with research instruments and diagnostic criteria for mental disorders.

The follow-up occurred from 2012 to 2014, that is, an average interval of five years after the first assessment. The participants from baseline (n = 1560) were invited for a second data collection. All interviewers met weekly to discuss the assessments, focusing on those who were uncertain about the BD diagnosis. In these situations, a psychiatrist was recruited to carry out the reassessment. 1244 individuals were located and consented to be reevaluated (79.7% of retention), and 14 (0.9%) were lost due to death. Since the present study aims to predict BD incidence, subjects who met diagnostic criteria for a lifetime manic or hypomanic episode were excluded. Unlike the baseline, data were collected through tablets using Open Data Kit (ODK), an open-source mobile data collection platform (Hartung et al., 2010). The forms were filled out offline, and the data were later backed up to computers through secure data transfer protocols.

This study was approved by the Research Ethics Committee of Universidade Católica de Pelotas under protocol number 2008/118. The subjects who presented any psychiatric diagnosis in the clinical interview were referred for specialized treatment in the local health system. All participants signed a printed informed consent form and could withdraw from the study at any time.

2.2. Outcome

The BD diagnosis was built with modules A and D from Mini International Neuropsychiatric Interview 5.0 (MINI), in order to assess current or past depressive episodes and current or past manic or hypomanic episodes, respectively. The BD diagnoses were reassessed in those cases where the diagnosis was questionable. MINI is a short-term diagnostic interview designed for clinical assessment of mental disorders according to Diagnostic and Statistical Manual of Mental Disorders — Fourth Edition (DSM-IV) and ICD-10 criteria (American Psychiatric Association, 1994; World Health Organization, 1993). Despite evaluating several disorders, MINI psychometric properties for the diagnosis of lifetime manic episode (sensitivity: 81.0%; specificity: 94.0%; positive predictive value: 76.0%; negative predictive value: 95.0%) and major depressive episode (sensitivity: 96.0%; specificity: 88.0%; positive predictive value: 87.0%; negative predictive value: 97.0%) are reliable when compared to DSM Structured Clinical Interview (Amorim, 2000).

2.3. Predictors

One hundred and ninety features were included in the original dataset before preprocessing steps. These variables include demographic, social, clinical, and environmental characteristics. The following features were included in the modeling pipeline:

a) Sociodemographic and environmental variables: Sex, skin color, age, socioeconomic status (3 levels and 5 levels), current occupation, currently studying, worked for money, has a partner, has a religion, access to psychotherapy, knows someone who attempted suicide or committed suicide, involvement in physical fights, family gun ownership, social support, has divorced parents, has any deceased parents, has someone close by already deceased, individual and family stress problems, lives with parents, family suicide attempts, seat belt wearing, helmet use when riding a motorcycle, suffered an accident that led to an emergency room, drove or took a ride with a drunk driver.
b) Substance use variables: Indicative of substance abuse or dependence (tobacco, alcohol, cannabis, cocaine, crack, amphetamines, inhalants, sedatives, hallucinogens, opioids, illicits, any other substances) assessed by Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) and lifetime use features (same substances cited above), age that first used drugs, injected drugs use, use of medication for stress problems in the last 30 days.
c) Clinical variables: mental disorder diagnoses (anxiety, mood and personality disorders), eating disorders, current suicide risk, serious organic disease, lifetime psychiatrist or psychologist visit, lifetime psychotherapeutic treatment, interrupted treatment.
d) Sex-related variables: age of first sexual intercourse, sexual intercourse in the last week (sex frequency), condom use, alcohol use before sexual intercourse, number of sexual partners, number of pregnancies, lifetime sexual abuse, lifetime sexual intercourse (binary).

e) Psychometric instrument items: Beck Depression Inventory (BDI) [21 items], Hypomania Checklist (HCL-32) [32 items], Social Readjustment Rating Scale (SRRS) [26 items], Beck Scale for Suicide Ideation (BSS) [21 items].

2.4. Machine learning analysis

Aiming to predict new cases of bipolar disorder in young adults, using features previously described, we created an ML pipeline to generate a predictive model using supervised learning. We used a vastly used machine learning algorithm for tabular data called tree gradient boosting, implemented through the *XGBoost* library (Chen & Guestrin, 2016) in the R programming language on version 4.2.1 (R Core Team, 2022).

Tree gradient boosting is part of what is called ensemble algorithms — joining many models to make predictions together — in statistical learning methods. Boosting improves this concept by building a sequence of originally weak models into progressively more powerful models. Additionally, in gradient boosting techniques, the gradient of a loss function is used to choose the best approach to improve a weak learner (James et al., 2021). In the context of gradient-boosted trees, weak learners are decision trees.

The following tree boosting hyperparameters were tuned (Kuhn & Vaughan, 2022a; Chen & Guestrin, 2016):

a) *mtry*: Number of predictors that is randomly sampled at each split.
b) *trees*: Number of trees contained in the ensemble.
c) *min_n*: Minimum number of observations in a node required for the node to be split further.
d) *tree_depth*: Maximum depth (number of splits) of each tree.
e) *loss_reduction*: Reduction in the loss function required to split further.
f) *learn_rate*: Step size at each iteration while moving toward a loss function optimization.
g) *sample_size*: Proportion of the data set used for modeling within an iteration.

When it comes to tabular data, gradient boosting decision trees (GBDT) are seen as the state-of-the-art, reinforced by several competitions in the ML scenario. In addition, a study found that GBDT perform better than deep learning models across multiple tabular datasets, and also requires less hyperparameter tuning (Shwartz-Ziv & Armon, 2021).

The implementation of the data modeling routines was carried out using the *tidymodels* framework (Kuhn & Wickham, 2020). The *tidymodels* is a R metapackage made up of multiple packages that assist in different stages of a machine learning pipeline. In order to split the data and create cross-validation resamples, *rsample* package was used (Silge et al., 2022), *parsnip* was used to access *XGBoost* functions in an unified manner (Kuhn & Vaughan, 2022a), *recipes* for preprocessing functions (Kuhn & Wickham, 2022), *workflows* to bundle the preprocessing, modeling and post-processing routines (Vaughan, 2022), *yardstick* to easily calculate performance measures (Kuhn & Vaughan, 2022b).

In the cross-validation, fifteen hyperparameter combinations were used as candidate parameter sets. The values for each hyperparameter were randomly chosen based on an algorithm that attempts to maximize the determinant of the spatial correlation matrix between coordinates (Santner et al., 2003).

2.5. Preprocessing

Before any preprocessing routine was performed, the data was divided into two subsets. A training set, consisting of 70% of the sample, and a test set with the remaining samples (30% of the total samples). The entire test set was isolated until the end of all tuning and validation procedures to build the model, to then be used to simulate the model performance on new data.

Some preprocessing techniques were adopted to clean and tidy data prior to modeling. The following preprocessing steps were applied:

1) Remove all features with more than 10% of missing values.
2) Impute categorical features with mode.
3) Impute numeric features with median.
4) Remove near-zero variance features (few unique values relative to the number of observations and also a ratio of the frequency of the second most common value is large [ratio of 10]).
5) Create dummy variables with $C - 1$ categories from categorical features.

2.6. Feature selection

The feature selection process aims to automatically filter variables from the data matrix considering their relevance to the predictive modeling problem. Therefore, we can build more accurate and parsimonious models while, at the same time, saving computational resources through the use of less data in the next model fitting steps.

The Boruta system was implemented for feature selection in the present pipeline. It consists of a random forest based algorithm that iteratively removes features that are statistically less important than random synthetic features (artificial noise). For each iteration, removed variables are prevented from being considered for the next iteration (Kursa et al., 2010). Boruta is considered a wrapper method as it takes into account a subset of variables with different combinations in each iteration.

2.7. Class imbalance

Class imbalance is a common problem in classification modeling. It happens when we face a set of examples that presents a given level way more frequently than other. Since most ML classifiers assume data equally distributed, they tend to be more biased towards the majority class, causing bad performance on minority class classification.

This concept is especially important in the context of predicting mental disorders, as subjects who will present the disease will be exposed to greater health

risk. Therefore, it is necessary that the classifiers of such outcomes can adequately predict this portion of the population. BD still has the aggravating factor of having a complex prognosis regarding the neuroprogression, which can be worsened by the length of disease (Librenza-Garcia et al., 2021).

For this paper, an algorithm named ROSE (Random Over-Sampling Examples) was used. ROSE is a smoothed-bootstrap-based technique that creates new artificial observations in data in order to minimize or eliminate class imbalance (Menardi & Torelli, 2014). The *themis* and *ROSE* R packages were adopted to implement the previously described algorithm (Hvitfeldt, 2022; Lunardon et al., 2014).

2.8. Cross-validation

The cross-validation (CV) process was used to tune *XGBoost* hyperparameters described earlier. We used the k-fold cross-validation technique with 5 folds repeated five times. In order to optimize the hyperparameter combinations, we used a racing method proposed by Kuhn (2014). It consists in calculating the area under the receiver operating characteristic (ROC) curve for each parameter set across validation folds. After evaluating the parameter combinations for three resamples, a repeated measure ANOVA model is fitted. The combinations that are statistically different (based on α level for one-sided confidence interval of 5%) from the best setting are excluded from further validation procedures. The ANOVA racing method was implemented via *finetune* R package (Kuhn, 2022).

In Figure 1, the cross-validation procedure can be visualized inside the orange area. For each fold, the Boruta and ROSE algorithms were applied just in training folds, leaving the testing fold untouched in order to properly estimate model error. A maximum of 25 runs (5-fold CV repeated up to five times) was considered for hyperparameter tuning. At the end, the remaining models of the ANOVA racing process were evaluated, and the model with the highest validation AUC was chosen to be tested in testing set from the initial data split.

2.9. Performance measures

Aiming to evaluate the performance of the algorithm, some evaluation metrics were used. Firstly, the area under the receiver operating characteristic (ROC) curve (AUC) was used to diagnose the classifier ability to predict correctly across multiple discrimination thresholds (Fawcett, 2006).

Sensitivity and specificity were also used to assess model ability to correctly detect subjects with BD who have the disorder, and correctly detect subjects that did not present BD who actually does not have BD, respectively (Yerushalmy, 1947). Positive and negative predictive values (PPV and NPV) show the proportion of positive and negative predictions that are truly positive or negative, correspondingly.

Accuracy measures the proportion of correct predictions among the total number of observations evaluated (Metz, 1978). In order to take into account the class imbalance previously described, we also used balanced accuracy as it inputs both sensitivity and specificity into its formula:

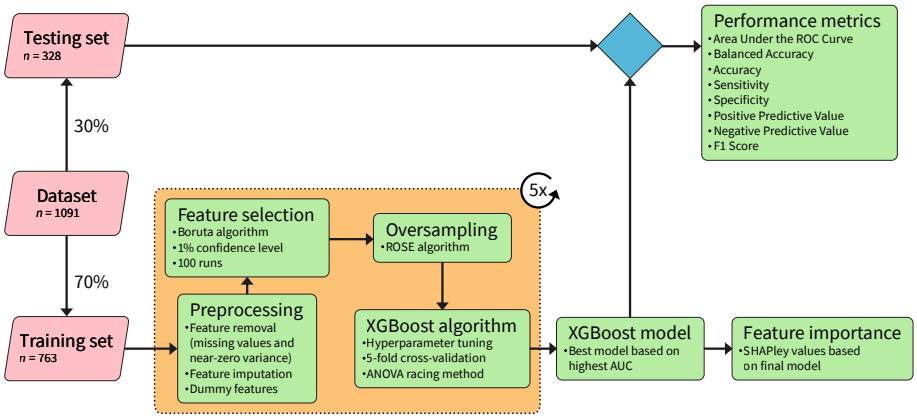


Figure 1: Machine learning pipeline flowchart. The figure shows data splitting, preprocessing routine, feature selection, cross-validation, model fitting, model assessment and feature importance steps using *XGBoost* algorithm.

$$\text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{1}$$

The F1 score is defined as the harmonic mean of PPV and sensitivity (Equation 2). This metric is able to demonstrate another model accuracy measure, being more robust in class imbalance scenarios.

$$F_1 = 2 \cdot \frac{\text{Positive Predictive Value} \cdot \text{Sensitivity}}{\text{Positive Predictive Value} + \text{Sensitivity}} \tag{2}$$

2.10. Model interpretability

In order to make model predictions more interpretable, we used SHAPley values. SHAPley values shows how to fairly distribute the total output among all features. Beyond that, SHAP (SHapley Additive exPlanations) allow for explanation on individual predictions (Lundberg & Lee, 2017). In the present study, the SHAPley values were obtained using the R package *SHAPforxgboost* (Liu & Just, 2021). This package provides functions to create SHAP-related visualizations from a *XGBoost* model object.

In addition to the use of feature importance visualization, partial dependence plots (PDP) were also employed. They are able to show the marginal effect a feature has on the predicted outcome of a machine learning model (Friedman, 2001). The PDP were built with the *pdp* (Greenwell, 2017) and *SHAPforxgboost* (Liu & Just, 2021) R packages, along with the *ggplot2* (Wickham, 2016) and *patchwork* (Pedersen, 2020) packages for plot composition.

3. Results

The present study aimed to create a model to predict bipolar disorder onset on young adults based on 5-year follow-up data. We assessed 1,091 subjects at

follow-up interview who had no current or past episode of mania or hypomania at the first assessment. Of these, 4.49% (n = 49) young adults received a diagnosis of BD five years later. Descriptive tables of demographic features at baseline are presented in Table 1. Absolute and relative frequency of missing values in each feature are described in Table 2. Table 3 shows the selected hyperparameter set from the cross-validation.

XGBoost showed an acceptable performance predicting BD five years before the diagnosis with a test set AUC of 0.786 [95% CI: 0.686, 0.887] (Figure 2). The other performance metrics using a cut-off of 0.5 for class decision boundary[1] can be seen in Table 4.

The six most relevant baseline features in BD prediction were feeling like a failure (BDI item 3), sadness (BDI item 1), current depressive episode, self-reported stress problems, self-confidence (HCL-32 item 3) and lifetime cocaine use. Feature importance can be seen in more detail in Figure 3. Given the importance of interpreting the model trajectory to a given prediction, to visualize the influence of each feature on the prediction of a specific sample, a force plot was built. The SHAPley values for each training sample is shown in Figure 4. Partial dependence plots can be seen in Figure 5.

In addition to the main pipeline, 1,000 different random training and testing splits were sampled in order to fit the final model. The estimates can be visualized in Figure 6. In this way, an adequate AUC can be seen in the model performance — within the estimated confidence intervals — including a robustness in the predictive power shown through the resamples. Along with the random splits, we also performed a permutation test as proposed by Fisher (1935) to compare the distribution of ROC AUC performance of random rearrangements of the outcome with the original test data using 10,000 permutations. We observed statistical difference between the original and permuted models ($p<0.001$). The distribution of the permuted AUCs is available in Figure 7. This result shows that our model predictions for BD incidence after the five years are more accurate than random classifiers.

[1]Class decision boundary separates the data points into classes, where the algorithm switches from one class to another. In the present paper, a threshold of 0.5 was used. If a prediction had a probability ≥ 0.5, it was classified as a positive instance, otherwise, as a negative one.

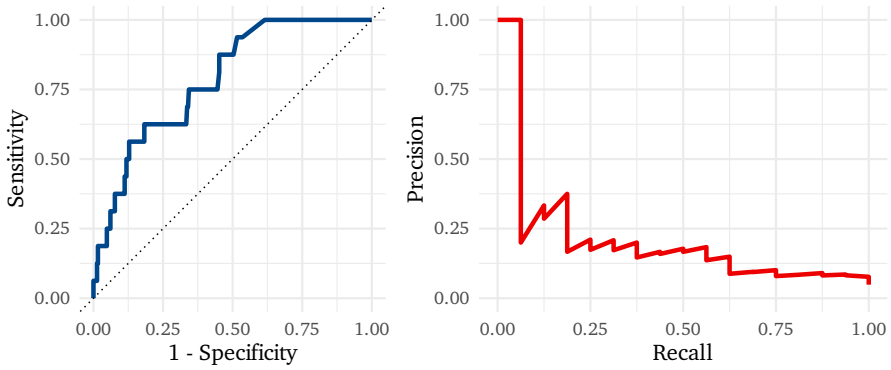


Figure 2: Receiver operating characteristic (ROC) curve and precision recall (PR) curve of the final model fitted on the training set with best parameter combination from cross-validation step, assessed on test set, with an area under the ROC curve of 0.786 and area under the PR curve of 0.208.

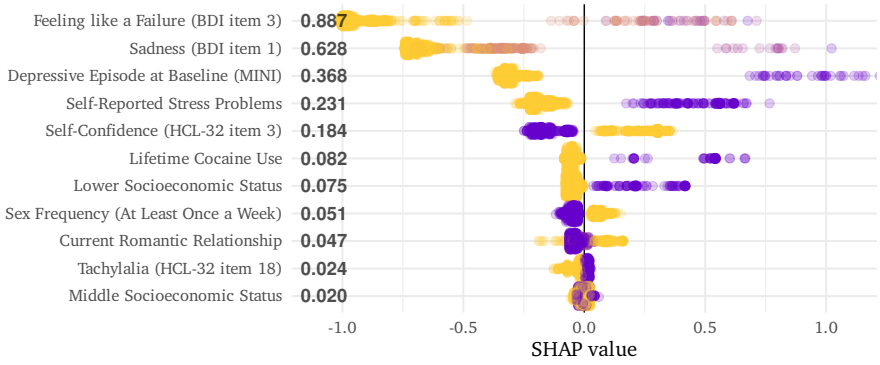


Figure 3: SHAPley values for each feature included in the final model. Each y-axis tick represents a feature, sorted by the highest absolute contribution across all observations, regardless of the direction of the association. Each dot represents a participant in the training set (n = 763). Observations with SHAPley values lower than zero behaved as protective factors, otherwise they were risk factors. The fill color represents the value of the variable for a given individual (purple corresponds to higher values, and yellow corresponds to lower values).

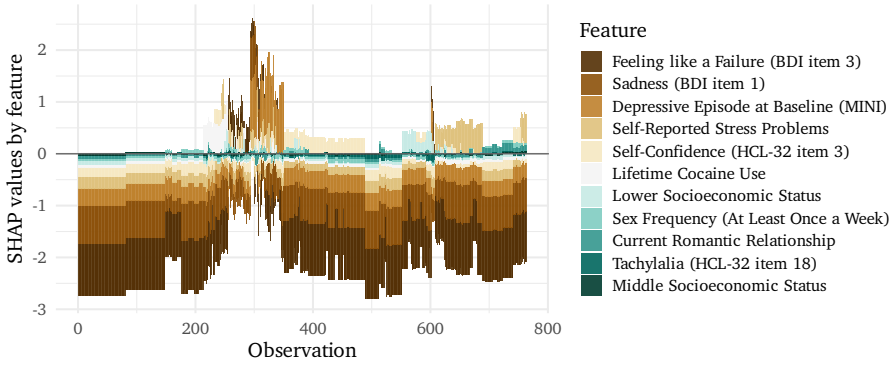


Figure 4: SHAPley force plot. The y-axis demonstrates the influence of each feature on current prediction based on SHAPLey values. The x-axis represents all samples used to train the final model. The whole training set (n = 763) is presented.

Table 1: Sociodemographic features measured at baseline grouped by bipolar disorder diagnosis in the follow-up.

Features	With BD (n = 49)	Without BD (n = 1042)	Overall (n = 1091)	p-value
Sex				0.302
Male	16 (32.7%)	457 (43.9%)	473 (43.4%)	
Female	33 (67.3%)	585 (56.1%)	618 (56.6%)	
Age				0.353
Mean (SD)	20.9 (2.28)	20.5 (2.10)	20.5 (2.11)	
Median [Min, Max]	21.0 [17.0, 25.0]	20.0 [16.0, 26.0]	20.0 [16.0, 26.0]	
Missing	0 (0%)	8 (0.8%)	8 (0.7%)	
Socioeconomic status				0.635
Upper	14 (28.6%)	401 (38.5%)	415 (38.0%)	
Middle	26 (53.1%)	509 (48.8%)	535 (49.0%)	
Lower	9 (18.4%)	132 (12.7%)	141 (12.9%)	
Skin color				0.800
White	34 (69.4%)	765 (73.4%)	799 (73.2%)	
Non-white	15 (30.6%)	273 (26.2%)	288 (26.4%)	
Missing	0 (0%)	4 (0.4%)	4 (0.4%)	
Currently working				0.965
No	8 (16.3%)	245 (23.5%)	253 (23.2%)	
Yes	14 (28.6%)	380 (36.5%)	394 (36.1%)	
Missing	27 (55.1%)	417 (40.0%)	444 (40.7%)	
Has a partner				0.953
No	13 (26.5%)	247 (23.7%)	260 (23.8%)	
Yes	33 (67.3%)	696 (66.8%)	729 (66.8%)	
Missing	3 (6.1%)	99 (9.5%)	102 (9.3%)	

The p-values were calculated from t-tests for numeric features and χ-squared tests for categorical features.

Table 2: Missing values of each feature in the entire dataset (n = 1091), before data splitting procedure. Only features with at least one missing value are present in the table. All features were collected at baseline.

Feature	Number of missing values	% of missing values
Age	8	0.73
Skin color	4	0.37
Currently working (last year)	444	40.70
Deceased parents	1	0.09
Has a religion	1	0.09
Psychiatric hospitalization	2	0.18
Stress problems	2	0.18
Psychiatric medication use	8	0.73
Maternal stress problems	1	0.09
Paternal stress problems	1	0.09
Sibling stress problems	2	0.18
Grandparents stress problems	1	0.09
Partner suicide attempt	753	69.02
Children suicide attempt	785	71.95
Paternal suicide attempt	725	66.45
Maternal suicide attempt	718	65.81
Sibling suicide attempt	725	66.45
Friend suicide attempt	712	65.26
Lifetime sexual intercourse	8	0.73
Worked for money	6	0.55

Table 2 continued from previous page

Feature	Number of missing values	% of missing values
Has a significant disease	2	0.18
Has ever visited a psychiatrist of psychologist	3	0.27
Has ever been treated by a psychiatrist or psychologist	107	9.81
BSS item 1	4	0.37
BSS item 2	2	0.18
BSS item 3	4	0.37
BSS item 4	1	0.09
BSS item 5	1	0.09
BSS item 6	1	0.09
BSS item 7	1	0.09
BSS item 8	1	0.09
BSS item 9	1	0.09
BSS item 10	1	0.09
BSS item 11	1	0.09
BSS item 12	1	0.09
BSS item 13	1	0.09
BSS item 14	1	0.09
BSS item 15	1	0.09
BSS item 16	1	0.09
BSS item 17	1	0.09
BSS item 18	1	0.09

Table 2 continued from previous page

Feature	Number of missing values	% of missing values
BSS item 19	1	0.09
BSS item 20	3	0.27
BSS item 21	1057	96.88
Knows someone who tried suicide	1	0.09
BDI item 1 (sadness)	1	0.09
BDI item 2 (hopelessness about the future)	1	0.09
BDI item 3 (feeling like a failure)	1	0.09
BDI item 4 (anhedonia)	1	0.09
BDI item 5 (sense of guilt)	1	0.09
BDI item 6 (feel that you are being punished)	1	0.09
BDI item 7 (self-hatred)	1	0.09
BDI item 8 (blame yourself for everything)	1	0.09
BDI item 9 (death thoughts)	1	0.09
BDI item 10 (excessive crying)	1	0.09
BDI item 11 (irritability)	1	0.09
BDI item 12 (loss of interest in personal relationships)	1	0.09
BDI item 13 (impaired decision making)	1	0.09
BDI item 14 (self-image)	1	0.09
BDI item 15 (occupational performance)	1	0.09
BDI item 16 (sleep)	1	0.09
BDI item 17 (fatigue)	1	0.09

Table 2 continued from previous page

Feature	Number of missing values	% of missing values
BDI item 18 (loss of appetite)	1	0.09
BDI item 19 (weight loss)	1	0.09
BDI item 20 (physical health concern)	1	0.09
BDI item 21 (loss of sexual interest)	1	0.09
Age that first used drugs	139	12.74
Age of first sexual intercourse	111	10.17
Sex frequency (sexual intercourse in the last week)	106	9.72
Condom use during last sexual intercourse	101	9.26
Alcohol use before the last sexual intercourse	101	9.26
Has a partner	102	9.35
Number of people you had sex with in the last year	138	12.65
Number of times you have gotten or made someone pregnant	114	10.45
Have you experienced forced sex?	405	37.12
Current anorexia nervosa	1	0.09
Antisocial personality disorder	2	0.18
Do you have access to psychotherapy (public or private healthcare)?	964	88.36
Current melancholic depressive episode	1	0.09
Past depressive episode	1	0.09

Table 3: Hyperparameter set chosen for final model based on highest area under the receiver operating characteristic curve in cross-validation routine.

Hyperparameter	**Value on final model**
mtry	9
trees	1429
min_n	10
tree_depth	14
learn_rate	0.0017
loss_reduction	0.6120
sample_size	0.9075

4. Discussion

We proposed to create a model capable of predicting BD in a 5-year interval with acceptable classification performance. Our final model performed with good metrics (AUC: 78.6%), suggesting good predictive capacity. To the best of our knowledge, this is the second Brazilian study to investigate the prediction of bipolar disorder incidence. A previous study investigated the development of a prediction model, with the use of elastic net algorithms, to identify participants who would develop bipolar disorder over the follow-up, in a large community birth cohort, from the city of Pelotas in Brazil (Rabelo-da Ponte et al., 2020). According to the results of this investigation, the model with the best performance (AUC of 0.82) predicted bipolar disorder at the age of 22 years, using clinical and sociodemographic data from the age of 18 years (Rabelo-da Ponte et al., 2020). A recent systematic review on clinical prediction models in psychiatry pointed to several aspects that predictive models could improve, such as overfitting prevention, generalizability and clinical utility (Meehan et al., 2022). The present paper used a larger sample than most studies to predict BD with statistical learning, despite having a low value of events per variable (EPV) of approximately 5.8.

The current study corroborates previous findings in which depressive symptoms would be one of the main predictors for BD conversion (Hafeman et al., 2017; Perich et al., 2015). Notably, the three primary factors found by the prediction model for BD developed in this paper are correlated constructs linked to depression: failure feeling, sadness and current depressive episode. This finding suggests that these factors could be prodromal symptoms of the disorder (Faedda et al., 2019; Van Meter et al., 2016), or even evidence of genetic predisposition to emotional distress (Smeland et al., 2018). It also reinforces the perspective of BD as a worsening trajectory, and the first mood episode as a milestone signaling for a complex disorder onset (Duffy et al., 2014). In the vast majority of cases, the first mood episode of a patient with BD is a depressive one, often years prior to a manic episode (Mesman et al., 2017; Duffy et al.,

Table 4: Performance metrics from the *XGBoost* model applied on testing set using 0.5 as threshold for positive classification. Area under the receiver operating characteristic curve: 0.786.

Performance metrics	***XGBoost* on test set**
Sensitivity (recall)	0.375
Specificity	0.920
PPV	0.194
NPV	0.966
Balanced accuracy	0.647
Accuracy	0.893
F1-score	0.255

Positive predictive value or precision (PPV); Negative predictive value (NPV).

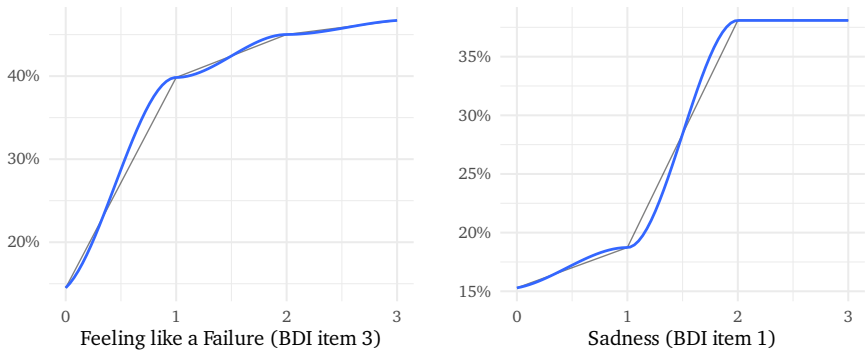


Figure 5: Partial dependence plots for continuous depression-related features. It shows the average trend of each feature. Other variables are held constant. The plots show an upward trend, which indicates that the higher the values of the variables of feeling like a failure and sadness, the higher the predicted probabilities for developing bipolar disorder after five years. The blue line indicates a regression line using the LOESS (locally weighted polynomial regression) method.

Figure 6: Histogram of the area under the receiver operating characteristic curve (AUC) based on 1,000 random training and testing data splits. The AUC mean and 95% CI found were 0.723 [0.719, 0.726]. This analysis is able to demonstrate the predictive performance robustness of the selected boosting model.

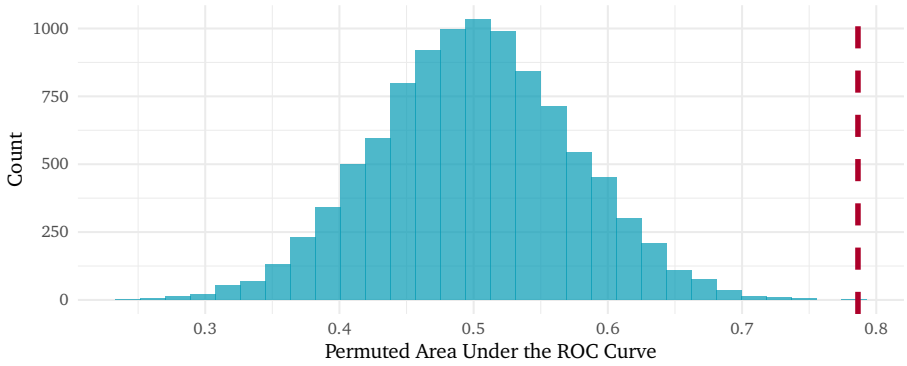


Figure 7: Distribution of the areas under the receiver operating characteristic curves (AUC) from the permutation test with 10,000 rearrangements. The red line indicates the AUC on the original test set (AUC = 0.786).

2014; Mesman et al., 2013), which turns the model also useful to ease the differential diagnosis between unipolar and bipolar depression, because together with other predictors, it is possible to verify whether there is a greater chance that a given current depressive episode is from a unipolar or a bipolar clinical condition.

Lifetime cocaine use is another major predictor evaluated in our study. Several studies have investigated the role of substance use in general and its association with the development of mood disorders. A cross-sectional study conducted in 2013 identified subsequent mood disorders developed in individuals with primary substance use disorder (SUD), and the average time between SUD onset and mood disorder was 11 years (Kenneson et al., 2013). The odds of developing bipolar disorder were particularly high among individuals with drug dependence in this study. A systematic review published in 2021 showed that substance use is a predictor for BD and (hypo)manic symptoms (Lalli et al., 2021). Specific data regarding cocaine use and BD has also been published. A prospective study investigated lifetime cocaine use as a potential predictor for conversion from major depressive disorder to bipolar disorder (de Azevedo Cardoso et al., 2020). The study analysis showed that the risk for conversion from major depressive disorder to BD was 3.41-fold higher in subjects who reported lifetime cocaine use at baseline. A systematic review also found a five-fold increased risk on the development of BD in individuals with lifetime cocaine use (Marangoni et al., 2016). Therefore, we consider this finding as part of the advancement of studies in the area, corroborating the information already established in the literature.

This study has some positive points to be noted. Initially, our sample is composed of young adults between 18 and 24 years old. According to the literature, BD symptoms usually appear before the age of 25, so the population used to build the model allows us to think about an early identification of the disease. This is possible because the population used to build the model was in a critical period of development for the onset of symptoms. Additionally, our team had an external psychiatrist to confirm the diagnosis whenever there were doubts through the standardized diagnostic interview (MINI), which sets a gold standard for characterization of the diagnosis. The average period between the initial interview and the follow-up was an average of five years, higher than in other studies in this field (Ribeiro et al., 2020).

The study has a large sample, collected through a probabilistic sample, obtained from the population of a city in southern Brazil with approximately 343,651 inhabitants. These factors bring robustness to our model. However, the outcome presented is difficult to predict due to: 1) the rarity of the outcome and 2) control participants may develop BD later. Nonetheless, this is a common challenge in studies in this area and we try to address these issues, whenever possible, statistically. Another point that must be taken into account when understanding the results presented here is the generalizability and applicability of the model. Studies in the area of precision psychiatry are on the rise. In this work, we aim and manage to present satisfactory results (test AUC 0.78), however, we understand that the data presented are primarily for scien-

tific purposes and as a basis for future improvements. This study demonstrates that, in the near future, it will be possible to think of a calculator capable of being implemented in basic health systems. The information presented may be useful especially for patients who present characteristics seen here as of potential importance in the face of the diagnosis of BD: current depressive episode, depressive symptoms (mainly related to feelings of failure and sadness) and lifetime use of cocaine. Such a tool has the potential for robust screening, enabling symptomatic treatment, ensuring a better prognosis and preventing more severe clinical conditions.

In summary, we developed a binary model with a state-of-the-art algorithm capable of predicting the diagnosis of BD in approximately five years in a specific population of young adults, through clinical, socio-environmental, substance use, sex-related variables and demographic data, collected through a probabilistic sample. However, aiming for a better characterization of the BD diagnosis, future studies should focus on making systematic follow-ups that seek to follow these subjects during other developmental stages, as well as investing in studies that use specific risk populations, such as depressed patients or children of parents with BD. Furthermore, the inclusion of digital health data, biological and neuropsychological information and the use of neuroimaging can help in the rise of new models with greater applicability for the future.

References

Altamura, A. C., Buoli, M., Caldiroli, A., Caron, L., Melter, C. C., Dobrea, C., Cigliobianco, M., & Quarantini, F. Z. (2015). Misdiagnosis, duration of untreated illness (DUI) and outcome in bipolar patients with psychotic symptoms: A naturalistic study, . *182*, 70–5. doi:10.1016/j.jad.2015.04.024. arXiv:25978716.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*. (4th ed.). Arlington, TX: American Psychiatric Press.

Amorim, P. (2000). Mini International Neuropsychiatric Interview (MINI): validação de entrevista breve para diagnóstico de transtornos mentais, . *22*, 106–115. doi:10.1590/S1516-44462000000300003.

de Azevedo Cardoso, T., Jansen, K., Mondin, T. C., Moreira, F. P., de Lima Bach, S., da Silva, R. A., de Mattos Souza, L. D., Balanzá-Martínez, V., Frey, B. N., & Kapczinski, F. (2020). Lifetime cocaine use is a potential predictor for conversion from major depressive disorder to bipolar disorder: A prospective study, . *74*, 418–423. doi:10.1111/pcn.13012.

Brazilian Institute of Geography and Statistics (2010). Brazilian 2010 census. URL: https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System, . doi:10.1145/2939672.2939785. arXiv:1603.02754.

Cloutier, M., Greene, M., Guerin, A., Touya, M., & Wu, E. (2018). The economic burden of bipolar I disorder in the United States in 2015, . *226*, 45–51. doi:10.1016/j.jad.2017.09.011. arXiv:28961441.

Duffy, A., Horrocks, J., Doucette, S., Keown-Stoneman, C., McCloskey, S., & Grof, P. (2014). The developmental trajectory of bipolar disorder, . *204*, 122–128. doi:10.1192/bjp.bp.113.126706. arXiv:24262817.

Faedda, G. L., Baldessarini, R. J., Marangoni, C., Bechdolf, A., Berk, M., Birmaher, B., Conus, P., DelBello, M. P., Duffy, A. C., Hillegers, M. H. J., Pfennig, A., Post, R. M., Preisig, M., Ratheesh, A., Salvatore, P., Tohen, M., Vázquez, G. H., Vieta, E., Yatham, L. N., Youngstrom, E. A., Van Meter, A., & Correll, C. U. (2019). An International Society of Bipolar Disorders task force report: Precursors and prodromes of bipolar disorder, . *21*, 720–740. doi:10.1111/bdi.12831. arXiv:31479581.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874. URL: https://doi.org/10.1016/j.patrec.2005.10.010. doi:10.1016/j.patrec.2005.10.010.

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of 'precision psychiatry', . *15*, 80. doi:10.1186/s12916-017-0849-x. arXiv:28403846.

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine., . *29*, 1189–1232. doi:10.1214/aos/1013203451.

Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *The R Journal*, *9*, 421–436. URL: https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.

Hafeman, D. M., Merranko, J., Goldstein, T. R., Axelson, D., Goldstein, B. I., Monk, K., Hickey, M. B., Sakolsky, D., Diler, R., Iyengar, S., Brent, D. A., Kupfer, D. J., Kattan, M. W., & Birmaher, B. (2017). Assessment of a Person-Level Risk Calculator to Predict New-Onset Bipolar Spectrum Disorder in Youth at Familial Risk, . *74*, 841–847. doi:10.1001/jamapsychiatry.2017.1763. arXiv:28678992.

Hartung, C., Lerer, A., Anokwa, Y., Tseng, C., Brunette, W., & Borriello, G. (2010). Open data kit: Tools to build information services for developing regions. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development - ICTD '10*. ACM Press. URL: https://doi.org/10.1145/2369220.2369236. doi:10.1145/2369220.2369236.

Hvitfeldt, E. (2022). *themis: Extra Recipes Steps for Dealing with Unbalanced Data*. URL: https://CRAN.R-project.org/package=themis r package version 0.2.2.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer Texts in Statistics (2nd ed.). New York, NY: Springer.

Kenneson, A., Funderburk, J. S., & Maisto, S. A. (2013). Substance use disorders increase the odds of subsequent mood disorders, . *133*, 338–343. doi:10.1016/j.drugalcdep.2013.06.011.

Kessing, L. V., Vradi, E., & Andersen, P. K. (2015). Life expectancy in bipolar disorder, . *17*, 543–548. doi:10.1111/bdi.12296. arXiv:25846854.

Kuhn, M. (2014). Futility Analysis in the Cross-Validation of Machine Learning Models, . doi:10.48550/arXiv.1405.6974. arXiv:1405.6974.

Kuhn, M. (2022). *finetune: Additional Functions for Model Tuning*. URL: https://CRAN.R-project.org/package=finetune r package version 0.2.0.

Kuhn, M., & Vaughan, D. (2022a). *parsnip: A Common API to Modeling and Analysis Functions*. URL: https://CRAN.R-project.org/package=parsnip r package version 1.0.0.

Kuhn, M., & Vaughan, D. (2022b). *yardstick: Tidy Characterizations of Model Performance*. URL: https://CRAN.R-project.org/package=yardstick r package version 1.0.0.

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*. URL: https://www.tidymodels.org.

Kuhn, M., & Wickham, H. (2022). *recipes: Preprocessing and Feature Engineering Steps for Modeling*. URL: https://CRAN.R-project.org/package=recipes r package version 1.0.0.

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta – a system for feature selection. *Fundamenta Informaticae*, *101*, 271–285. URL: https://doi.org/10.3233/fi-2010-288. doi:10.3233/fi-2010-288.

Lalli, M., Brouillette, K., Kapczinski, F., & de Azevedo Cardoso, T. (2021). Substance use as a risk factor for bipolar disorder: A systematic review, . *144*, 285–295. doi:10.1016/j.jpsychires.2021.10.012.

Librenza-Garcia, D., Kotzian, B. J., Yang, J., Mwangi, B., Cao, B., Lima, L. N. P., Bermudez, M. B., Boeira, M. V., Kapczinski, F., & Passos, I. C. (2017). The impact of machine learning techniques in the study of bipolar disorder: A systematic review, . *80*, 538–554. doi:10.1016/j.neubiorev.2017.07.004. arXiv:28728937.

Librenza-Garcia, D., Suh, J. S., Watts, D. P., Ballester, P. L., Minuzzi, L., Kapczinski, F., & Frey, B. N. (2021). Structural and Functional Brain Correlates of Neuroprogression in Bipolar Disorder, . *48*, ;. doi:10.1007/7854_2020_177. arXiv:33040317.

Liu, Y., & Just, A. (2021). *SHAPforxgboost: SHAP Plots for 'XGBoost'*. URL: https://CRAN.R-project.org/package=SHAPforxgboost r package version 0.1.1.

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: a Package for Binary Imbalanced Learning. *The R Journal*, *6*, 79–89. URL: https://doi.org/10.32614/RJ-2014-008. doi:10.32614/RJ-2014-008.

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions, . doi:10.48550/arXiv.1705.07874. arXiv:1705.07874.

Marangoni, C., Hernandez, M., & Faedda, G. L. (2016). The role of environmental exposures as risk factors for bipolar disorder: A systematic review of longitudinal studies, . *193*, 165–174. doi:10.1016/j.jad.2015.12.055.

McIntyre, R. S., Berk, M., Brietzke, E., Goldstein, B. I., López-Jaramillo, C., Kessing, L. V., Malhi, G. S., Nierenberg, A. A., Rosenblat, J. D., Majeed, A., Vieta, E., Vinberg, M., Young, A. H., & Mansur, R. B. (2020). Bipolar disorders, . *396*, 1841–1856. doi:10.1016/S0140-6736(20)31544-0. arXiv:33278937.

McIntyre, R. S., & Calabrese, J. R. (2019). Bipolar depression: the clinical characteristics and unmet needs of a complex disorder, . *35*, 1993–2005. doi:10.1080/03007995.2019.1636017. arXiv:31311335.

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges, . *27*, 2700–2708. doi:10.1038/s41380-022-01528-4.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data, . *28*, 92–122. doi:10.1007/s10618-012-0295-5.

Merikangas, K. R., Jin, R., He, J.-P., Kessler, R. C., Lee, S., Sampson, N. A., Viana, M. C., Andrade, L. H., Hu, C., Karam, E. G., Ladea, M., Medina-Mora, M. E., Ono, Y., Posada-Villa, J., Sagar, R., Wells, J. E., & Zarkov, Z. (2011). Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative, . *68*, 241–251. doi:10.1001/archgenpsychiatry.2011.12. arXiv:21383262.

Mesman, E., Nolen, W. A., Keijsers, L., & Hillegers, M. H. J. (2017). Baseline dimensional psychopathology and future mood disorder onset: findings from the Dutch Bipolar Offspring Study, . *136*, 201–209. doi:10.1111/acps.12739. arXiv:28542780.

Mesman, E., Nolen, W. A., Reichart, C. G., Wals, M., & Hillegers, M. H. J. (2013). The Dutch bipolar offspring study: 12-year follow-up, . *170*, 542–549. doi:10.1176/appi.ajp.2012.12030401. arXiv:23429906.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*, 283–298. URL: https://doi.org/10.1016/s0001-2998(78)80014-2. doi:10.1016/s0001-2998(78)80014-2.

Passos, I. C., Ballester, P. L., Barros, R. C., Librenza-Garcia, D., Mwangi, B., Birmaher, B., Brietzke, E., Hajek, T., Jaramillo, C. L., Mansur, R. B., Alda, M., Haarman, B. C. M., Isometsa, E., Lam, R. W., McIntyre, R. S., Minuzzi, L., Kessing, L. V., Yatham, L. N., Duffy, A., & Kapczinski, F. (2019). Machine learning and big data analytics in bipolar disorder: A position paper from the International Society for Bipolar Disorders Big Data Task Force, . *21*, 582–594. doi:10.1111/bdi.12828. arXiv:31465619.

Passos, I. C., Mwangi, B., & Kapczinski, F. (2016). Big data analytics and machine learning: 2015 and beyond, . *3*, 13–15. doi:10.1016/S2215-0366(15)00549-0. arXiv:26772057.

Pedersen, T. L. (2020). *patchwork: The Composer of Plots*. URL: https://CRAN.R-project.org/package=patchwork r package version 1.1.1.

Perich, T., Lau, P., Hadzi-Pavlovic, D., Roberts, G., Frankland, A., Wright, A., Green, M., Breakspear, M., Corry, J., Radlinska, B., McCormack, C., Joslyn,

C., Levy, F., Lenroot, R., Jnr, J. I. N., & Mitchell, P. B. (2015). What clinical features precede the onset of bipolar disorder?, . *62*, 71–7. doi:10.1016/j.jpsychires.2015.01.017. arXiv:25700556.

Plans, L., Barrot, C., Nieto, E., Rios, J., Schulze, T. G., Papiol, S., Mitjans, M., Vieta, E., & Benabarre, A. (2019). Association between completed suicide and bipolar disorder: A systematic review of the literature, . *242*, 111–122. doi:10.1016/j.jad.2018.08.054. arXiv:30173059.

Rabelo-da Ponte, F. D., Feiten, J. G., Mwangi, B., Barros, F. C., Wehrmeister, F. C., Menezes, A. M., Kapczinski, F., Passos, I. C., & Kunz, M. (2020). Early identification of bipolar disorder among young adults - a 22-year community birth cohort, . *142*, 476–485. doi:10.1111/acps.13233. arXiv:32936930.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Ribeiro, J. S., Pereira, D., Salagre, E., Coroa, M., Oliveira, P. S., Santos, V., Madeira, N., Grande, I., & Vieta, E. (2020). Risk Calculators in Bipolar Disorder: A Systematic Review, . *10*, 525. doi:10.3390/brainsci10080525. arXiv:32781733.

Santner, T. J., Williams, B. J., & Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. New York, NY, USA: Springer. URL: https://link.springer.com/book/10.1007/978-1-4757-3799-8.

Scott, J., & Leboyer, M. (2011). Consequences of delayed diagnosis of bipolar disorders, . *37*, 3. doi:10.1016/S0013-7006(11)70048-3. arXiv:22212870.

Shwartz-Ziv, R., & Armon, A. (2021). Tabular data: Deep learning is not all you need, . URL: https://arxiv.org/abs/2106.03253. doi:10.48550/ARXIV.2106.03253.

Silge, J., Chow, F., Kuhn, M., & Wickham, H. (2022). *rsample: General Resampling Infrastructure*. URL: https://CRAN.R-project.org/package=rsample r package version 1.0.0.

Smeland, O. B., Wang, Y., Frei, O., Li, W., Hibar, D. P., Franke, B., Bettella, F., Witoelar, A., Djurovic, S., Chen, C.-H., Thompson, P. M., Dale, A. M., & Andreassen, O. A. (2018). Genetic Overlap Between Schizophrenia and Volumes of Hippocampus, Putamen, and Intracranial Volume Indicates Shared Molecular Genetic Mechanisms, . *44*, 854–864. doi:10.1093/schbul/sbx148.

Van Meter, A. R., Burke, C., Youngstrom, E. A., Faedda, G. L., & Correll, C. U. (2016). The Bipolar Prodrome: Meta-Analysis of Symptom Prevalence Prior to Initial or Recurrent Mood Episodes, . *55*, 543–555. doi:10.1016/j.jaac.2016.04.017.

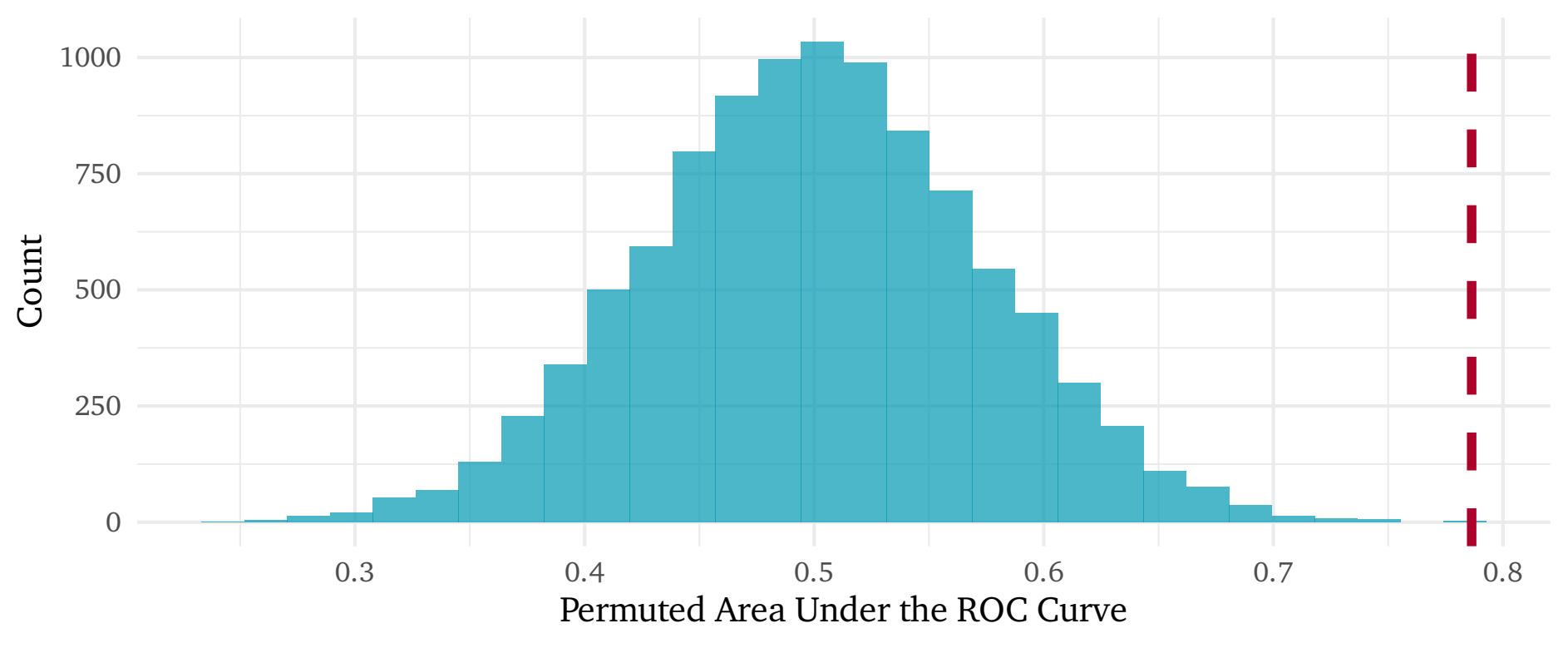
Vaughan, D. (2022). *workflows: Modeling Workflows*. URL: `https://CRAN.R-project.org/package=workflows` r package version 0.2.6.

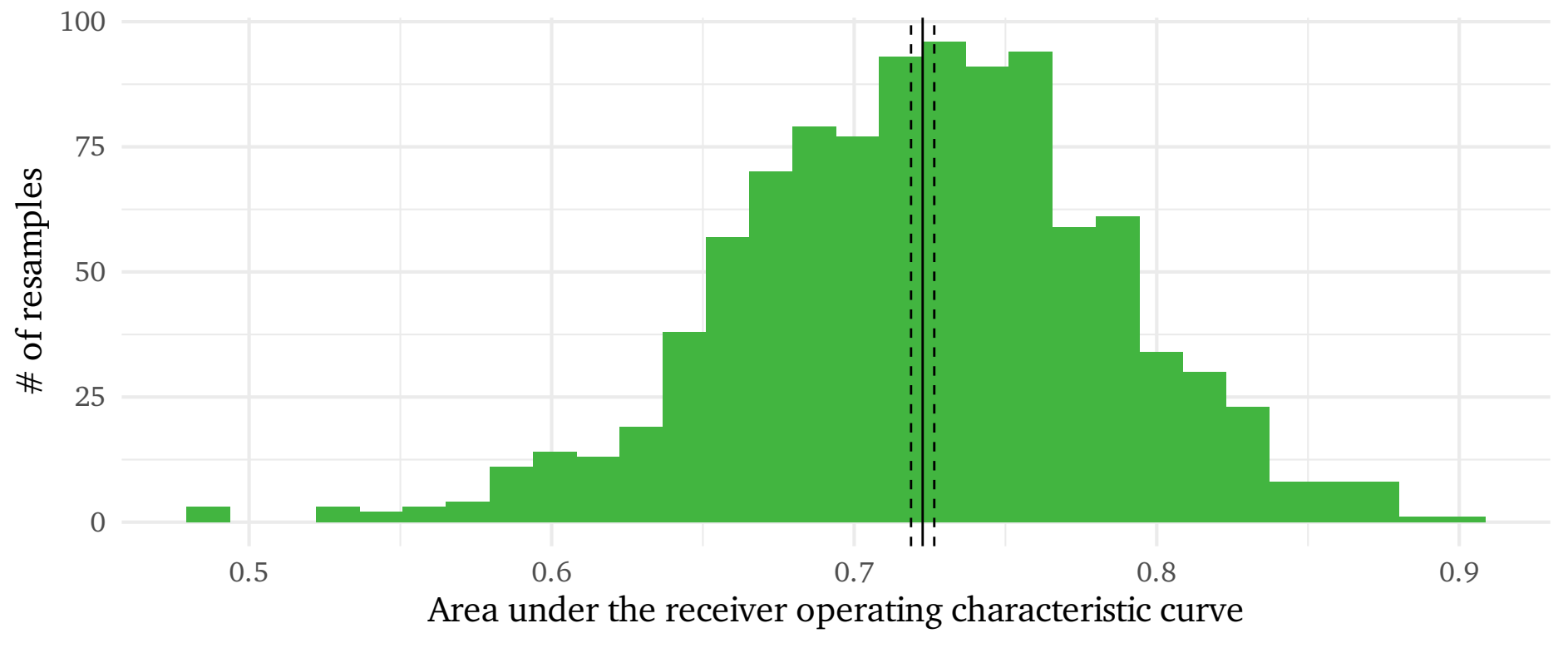
Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL: `https://ggplot2.tidyverse.org`.

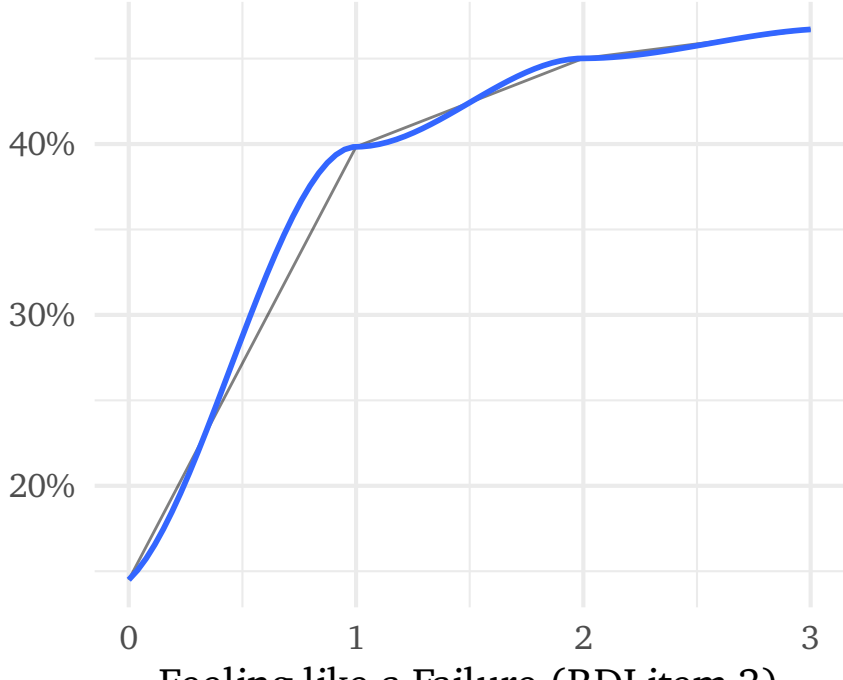
World Health Organization (1993). *The ICD-10 classification of mental and behavioural disorders*. Genève, Switerzland: World Health Organization.

Yatham, L. N., Kennedy, S. H., Parikh, S. V., Schaffer, A., Bond, D. J., Frey, B. N., Sharma, V., Goldstein, B. I., Rej, S., Beaulieu, S., Alda, M., MacQueen, G., Milev, R. V., Ravindran, A., O'Donovan, C., McIntosh, D., Lam, R. W., Vazquez, G., Kapczinski, F., McIntyre, R. S., Kozicky, J., Kanba, S., Lafer, B., Suppes, T., Calabrese, J. R., Vieta, E., Malhi, G., Post, R. M., & Berk, M. (2018). Canadian Network for Mood and Anxiety Treatments (CANMAT) and International Society for Bipolar Disorders (ISBD) 2018 guidelines for the management of patients with bipolar disorder, . *20*, 97–170. doi:`10.1111/bdi.12609`. `arXiv:29536616`.

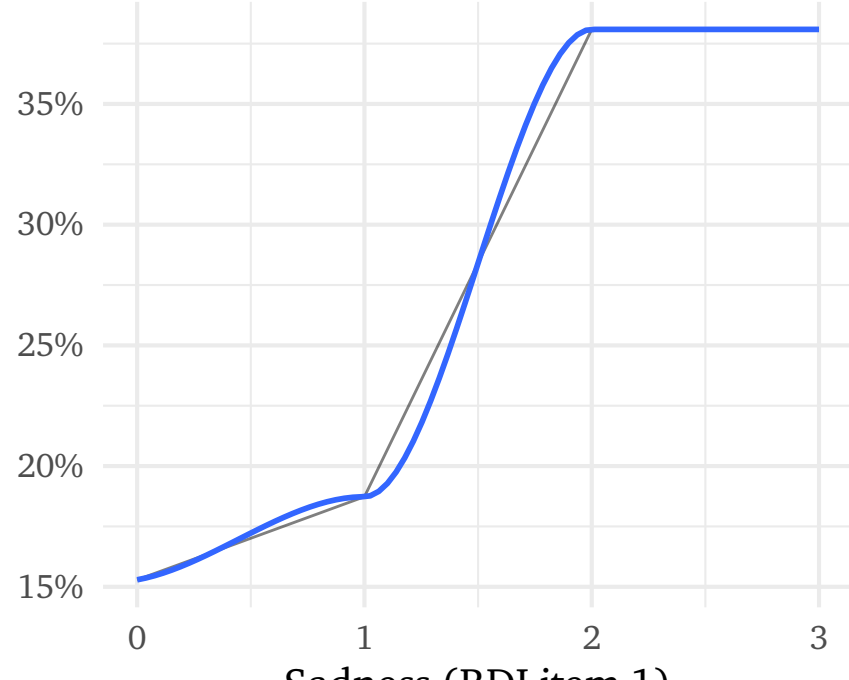
Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. *Public Health Reports (1896-1970)*, *62*, 1432. URL: `https://doi.org/10.2307/4586294`. doi:`10.2307/4586294`.


1000
750
500
250
0
Count
0.3
0.4
0.5
0.6
0.7
0.8
Permuted Area Under the ROC Curve

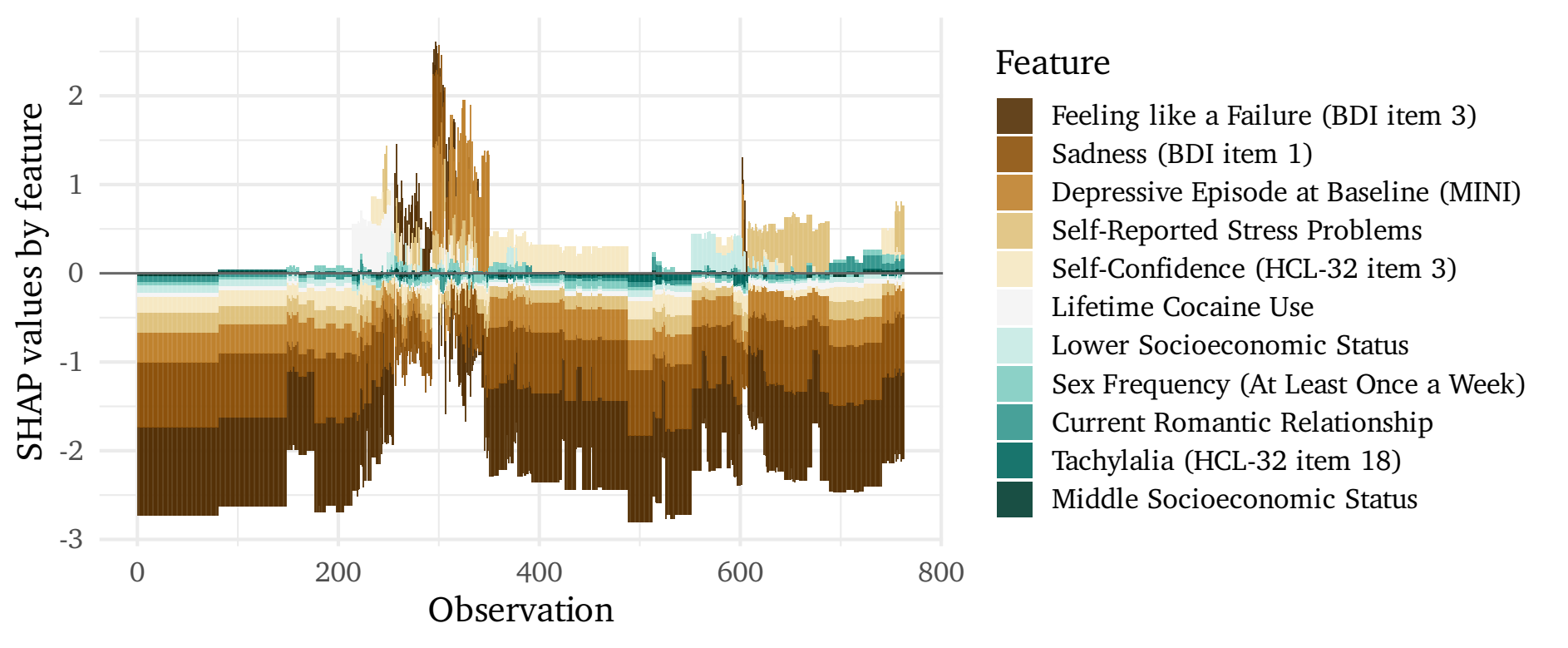

100
75
50
25
0
# of resamples
0.5
0.6
0.7
0.8
0.9
Area under the receiver operating characteristic curve

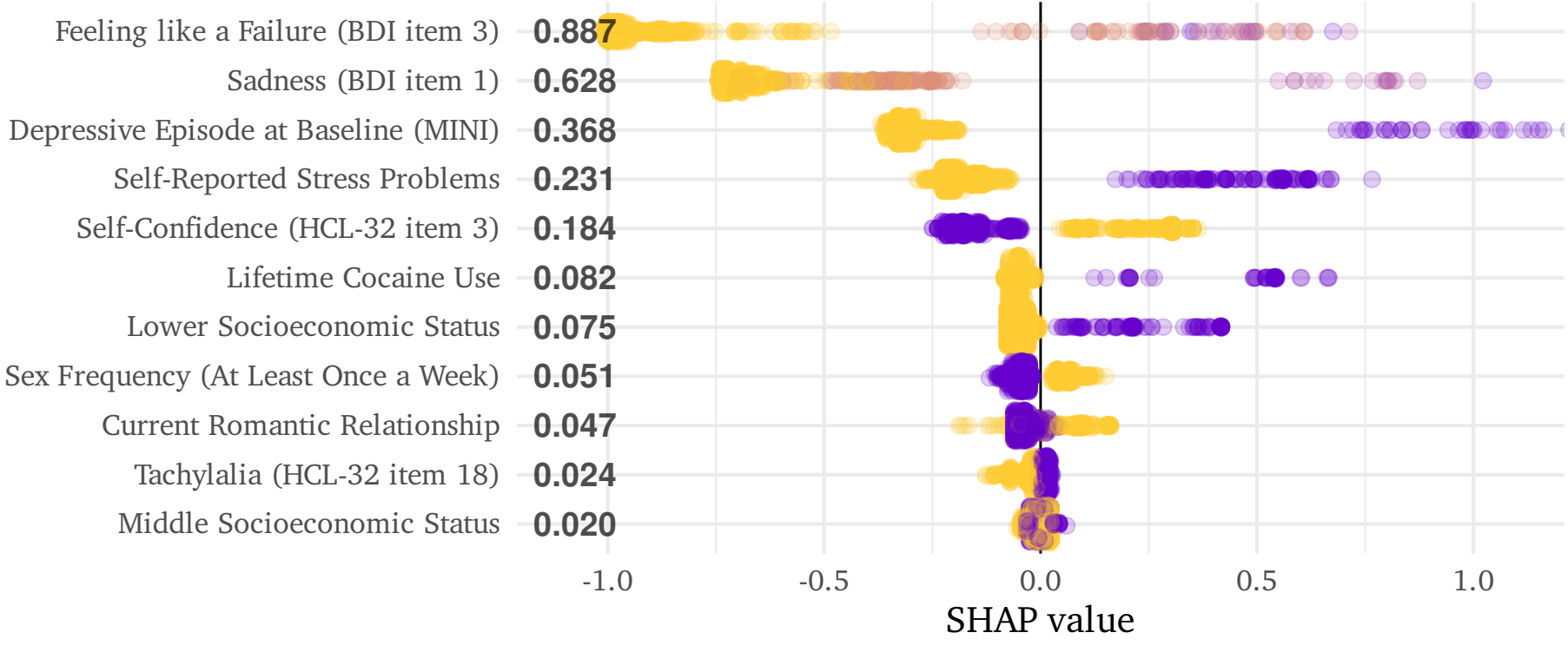


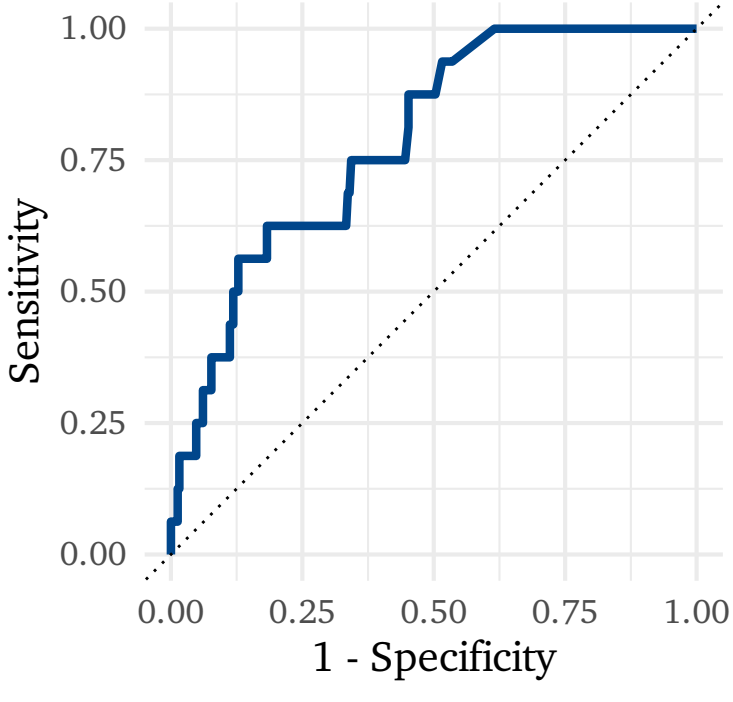
Feeling like a Failure (BDI item 3)

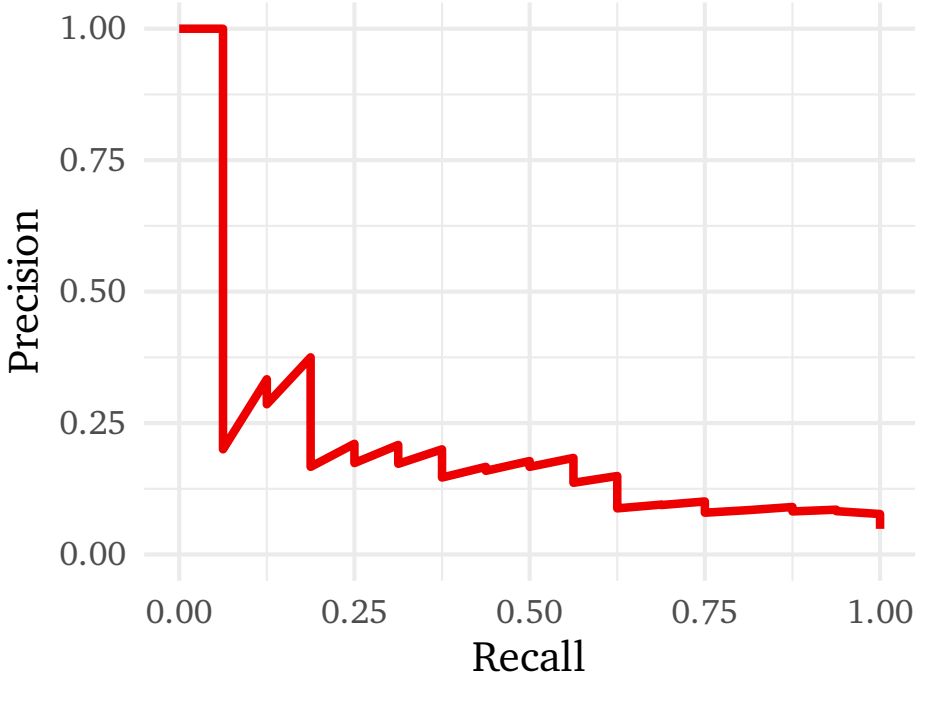


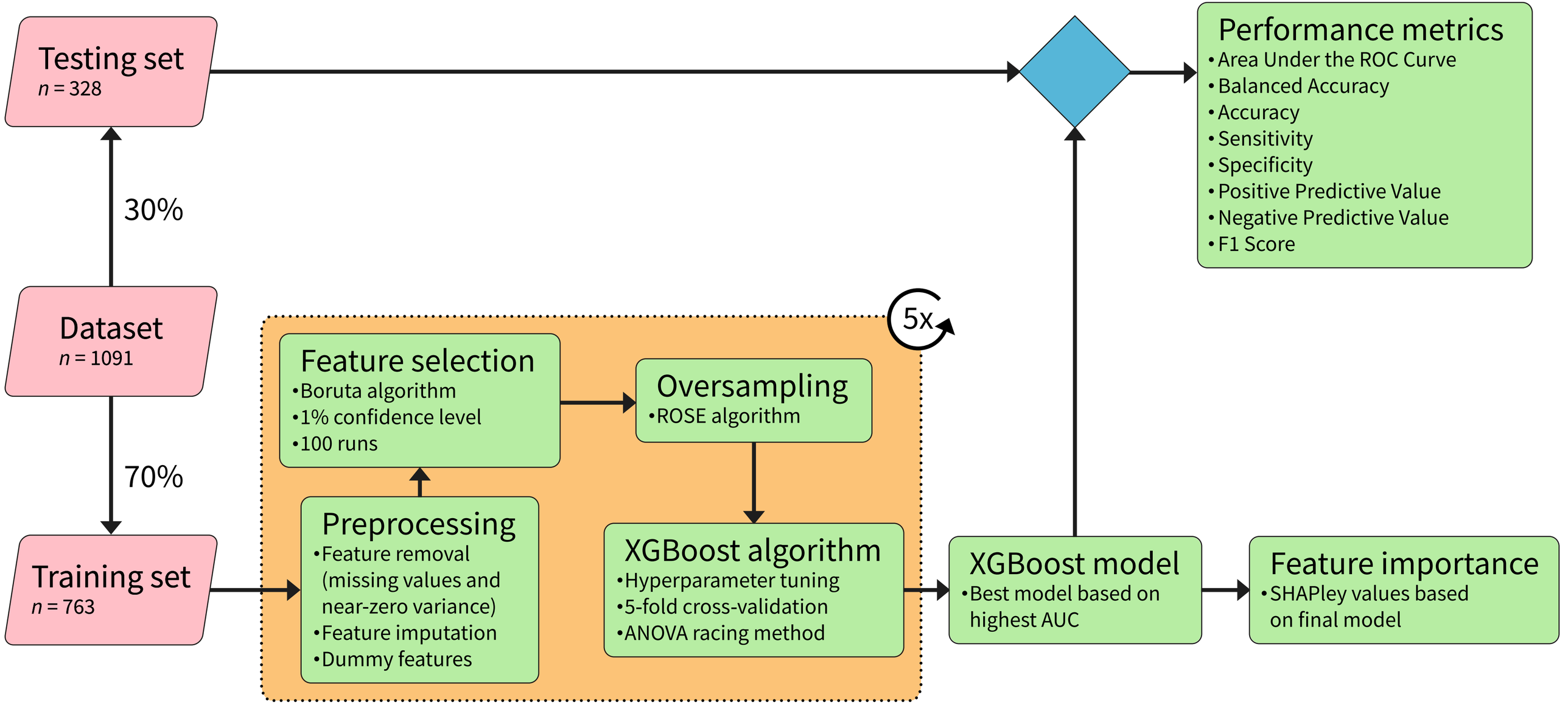
Sadness (BDI item 1)


SHAP values by feature
2
1
0
-1
-2
-3
0
200
400
600
800
Observation
Feature
Feeling like a Failure (BDI item 3)
Sadness (BDI item 1)
Depressive Episode at Baseline (MINI)
Self-Reported Stress Problems
Self-Confidence (HCL-32 item 3)
Lifetime Cocaine Use
Lower Socioeconomic Status
Sex Frequency (At Least Once a Week)
Current Romantic Relationship
Tachylalia (HCL-32 item 18)
Middle Socioeconomic Status


Feeling like a Failure (BDI item 3)
0.887
Sadness (BDI item 1)
0.628
Depressive Episode at Baseline (MINI)
0.368
Self-Reported Stress Problems
0.231
Self-Confidence (HCL-32 item 3)
0.184
Lifetime Cocaine Use
0.082
Lower Socioeconomic Status
0.075
Sex Frequency (At Least Once a Week)
0.051
Current Romantic Relationship
0.047
Tachylalia (HCL-32 item 18)
0.024
Middle Socioeconomic Status
0.020
-1.0
-0.5
0.0
0.5
1.0
SHAP value


1.00
0.75
0.50
0.25
0.00
Sensitivity
0.00
0.25
0.50
0.75
1.00
1 - Specificity


1.00
0.75
0.50
0.25
0.00
Precision
0.00
0.25
0.50
0.75
1.00
Recall


Testing set
n = 328
30%
Dataset
n = 1091
70%
Training set
n = 763
Performance metrics
• Area Under the ROC Curve
• Balanced Accuracy
• Accuracy
• Sensitivity
• Specificity
• Positive Predictive Value
• Negative Predictive Value
• F1 Score
5x
Feature selection
•Boruta algorithm
•1% confidence level
•100 runs
Oversampling
•ROSE algorithm
Preprocessing
•Feature removal (missing values and near-zero variance)
•Feature imputation
•Dummy features
XGBoost algorithm
•Hyperparameter tuning
•5-fold cross-validation
•ANOVA racing method
XGBoost model
•Best model based on highest AUC
Feature importance
•SHAPley values based on final model