Received: September 2025, Accepted: November 2025, Published: December 2025
Digital Object Identifier: https://doi.org/10.34302/CJEE/BUEP2247

REGRESSION-BASED PREDICTION OF ANXIETY SEVERITY

Bogdan **CHIȘ**, Diana **DULF**, Mădălina **MORAR,** Eleonora **POP**
Technical University of Cluj-Napoca
chis.do.paul@student.utcluj.ro, dulf.da.diana@student.utcluj.ro,
jula.an.madalina@student.utcluj.ro, eleonora.pop@ieec.utcluj.ro

Keywords: Mental health, Anxiety disorders, Machine learning, Regression analysis

Abstract*:* *The present study explores the use of machine learning to predict self-reported anxiety levels based on demographic, behavioral, and physiological data. To this end, we used a dataset comprising 11,000 survey responses and applied multiple regression models, including Linear Regression, Gradient Boosting Regression, Extreme Gradient Boosting Regression Light Gradient-Boosting Machine Regression after data preprocessing. The performance of the models was evaluated using the mean absolute error, the root mean square error, and the coefficient of determination. Among the models evaluated, Gradient Boosting Regression achieved the best results, with a mean cross-validated R² of 0.759 after five-fold cross-validation.*

1. INTRODUCTION

Anxiety is among the most prevalent mental health concerns on a global scale, often evading detection in nonclinical populations due to its subjective and intricate nature [1]. In this context, Machine learning (ML) offers new opportunities to detect patterns associated with anxiety. These patterns are detected based on accessible, structured data.

The objective of this study is to predict anxiety levels using characteristics related to demographics, lifestyle, physiology, and psychological well-being. A series of multiple regression models were trained and evaluated.

A second objective of this study is the identification of the most significant predictive factors. The results of this study may support the development of data-driven screening tools for the early identification of people at risk of anxiety disorders.

2. RELATED WORK

This section reviews previous studies on mental health prediction using ML, organized first by general mental health disorders followed by anxiety, depression and stress.

2.1. General mental health

Chung and Teo [2] conducted an evaluation of classifiers for automated mental health screening, using response data from the Open Sourcing Mental Illness tech survey, https://osmhhelp.org/research.html. They compared five base model approaches with Extreme Gradient Boosting and a Deep Neural Network and a voting ensemble built from the same base learners. Under repeated cross-validation, Gradient Boosting Machine led all methods with an 0.88 accuracy. Extreme Gradient Boosting and Deep Neural Network achieved 0.87 and 0.86, respectively, while the voting ensemble reached 0.85 and the remaining base classifiers scored between 0.82 and 0.84.

Jain et al. [3] focused on predicting depression as a first step toward preventing suicide, using lifestyle and demographic attributes collected from 1,429 individuals. The study leveraged 76 features, ranging from income, marital status, and substance use to medical and financial information, to build predictive models using eight main ML algorithms. The Support Vector Machine classifier achieved the highest accuracy of 0.83.

Tate et al. [4] investigated the feasibility of predicting general mental health problems in adolescence using ML techniques. Using data from 7,638 participants in the Child and Adolescent Twin Study in Sweden [5], the authors trained models on 474 predictors derived from parental reports and national register data. Several common models were compared. Random Forest achieved the highest area under the receiver operating characteristic curve (AUROC) of 0.73.

Garriga et al. [6] developed a ML model to predict the probability of mental health crises over a 28-day horizon using electronic health records of 17,122 participants. Aimed at enabling proactive intervention, the model was designed for continuous patient monitoring and achieved an AUROC of 0.79, with 0.58 sensitivity and 0.85 specificity. Although the area under the precision-recall curve was only 0.15, the system demonstrated clinical utility. In a six-month prospective study, predictions were found helpful in 64% of cases, either by informing caseload management or mitigating crisis risk. This study is notable as one of the first to implement continuous, real-time risk prediction across a broad spectrum of mental health crises in a clinical setting.

2.2. Depression, anxiety and stress

Bhatnagar et al. [7] investigated the prevalence and impact of anxiety among Indian university students, focusing specifically on a sample of 127 engineering students. Using a Likert questionnaire, the study quantified anxiety levels and examined associated causes and effects. The authors applied multiple classifiers to predict anxiety severity. Among the tested algorithms, Random Forest achieved the highest accuracy of 0.78, followed by Support Vector Machine with accuracy of 0.75, and Naïve Bayes and Decision Tree both at 0.71.

Richter et al. [8] tackled the challenge of distinguishing anxiety from depression, using objective behavioral measures rather than self-report alone. In their study, 125 subclinical participants were stratified into four groups: high symptoms of depression, anxiety, or both and the non-symptomatic control group. Participants completed cognitive–emotional bias tasks. Models were trained to classify individuals based on their performance across these tasks. The analysis achieved 0.71 sensitivity and 0.70 specificity in distinguishing symptomatic from non-symptomatic participants. In a two-group model, the classifiers reached 0.68 accuracy for high-depression and 0.74 for high-anxiety groups.

Nemesure et al. [9] developed models to predict generalized anxiety disorder (GAD) and major depressive disorder (MDD) in a non-clinical setting using electronic health records. Drawing on a sample of 4,184 undergraduate students, the study excluded all direct psychiatric inputs and instead trained an ensemble ML pipeline on 59 biomedical and demographic features. On a test set, the model achieved an AUROC of 0.73 for GAD (sensitivity: 0.66, specificity: 0.70) and 0.67 for MDD (sensitivity: 0.55, specificity: 0.70). Feature attribution analysis using SHAP values revealed that variables such as satisfaction with living conditions and public health insurance were predictive of MDD, while up-to-date vaccinations and marijuana use were most predictive of GAD.

The Depression Anxiety Stress Scales (DASS) is a well-established self-report instrument designed to measure the emotional states of depression, anxiety, and stress. DASS-42 consists of 42 items and DASS-21 being its shorter 21-item version. Kumar et al. [10] investigated the prediction of anxiety, depression, and stress severity using two datasets: the DASS-42, sourced online, and the DASS-21, collected by the authors. The study employed eight ML algorithms along with a hybrid classification approach. Each model aimed to classify symptoms into five severity levels. While the hybrid method generally outperformed individual classifiers, the highest accuracy was achieved by the Radial Basis Function Network, which reached 0.97 for anxiety, 0.96 for depression, 0.96 for stress on DASS-42, and 0.82 for anxiety, 0.96 for depression, and 0.89 for stress on DASS-21.

Priya and Garg [11] applied ML algorithms to classify the severity of anxiety, depression, and stress into five levels using responses from 348 participants of the standardized DASS-21 questionnaire. The dataset included individuals from varied professional and cultural backgrounds. Five common methods were tested. Although Naïve

Bayes achieved the highest raw accuracy, Random Forest was selected as the best performing model based on the F1 score, which was prioritized due to class imbalance in the data. The Random Forest model had 0.79 accuracy for depression, 0.71 for anxiety and 0.72 for stress. The study also analyzed feature importance, identifying "scared without any good reason", "life was meaningless", and "difficult to relax" as the most predictive items for anxiety, depression, and stress, respectively.

Chavanne et al. [12] examined the potential of early prediction of clinical anxiety in adolescents using a combination of magnetic resonance imaging derived brain features and psychometric data. In a longitudinal study of 580 participants, non-anxious individuals at age 14 were followed up to assess anxiety diagnoses between ages 18 and 23. A voting classifier combining Random Forest, Support Vector Machine, and Linear Regression was trained on baseline gray matter volumes and psychological measures. The model achieved moderate predictive performance for pooled anxiety disorders with AUROC of 0.68. Psychometric features such as neuroticism, hopelessness, and emotional symptoms were the primary contributors for prediction. While brain imaging data did not improve prediction for pooled anxiety outcomes, it did enhance the detection of GAD, particularly with contributions from the caudate and pallidum volumes.

3. METHODOLOGY

The workflow adopted begins with an exploratory data analysis phase comprising univariate, bivariate, and multivariate analyses to understand the distribution, relationships, and interactions within the dataset. This step provides insights that inform subsequent preprocessing decisions. Afterwards, ML experiments are conducted using various regression models. The performance of these models is then evaluated based on appropriate metrics to assess predictive accuracy and generalization capability.

3.1. Dataset description

The dataset used in this study was obtained from Kaggle, https://www.kaggle.com/datasets/natezhang123/social-anxiety-dataset, and is based on survey responses from a diverse population on factors believed to be associated with social anxiety.

The dataset consists of 11,000 rows and 19 columns in total. The target attribute is Anxiety Level, which is a self-reported numerical value that indicates the respondent's perceived level of anxiety. The remaining features are independent variables. An overview of the characteristics of the dataset, accompanied by their classifications and descriptions, is provided in *table 1*.

Table 1. Features description

Name	Description	Type
Age	Age of the respondent	Numerical
Gender	Biological sex of the respondent	Categorical
Occupation	Professional field or job role	Categorical
Sleep Hours	Average sleep hours per night	Numerical
Physical Activity (hrs/week)	Weekly physical activity duration	Numerical
Caffeine Intake (mg/day)	Average daily caffeine intake	Numerical
Alcohol Consumption (drinks/week)	Alcohol consumption per week	Numerical
Smoking	Whether the individual smokes	Boolean
Family History of Anxiety	Presence of anxiety in family history	Boolean
Stress Level (1-10)	Self-reported stress level	Numerical
Heart Rate (bpm)	Resting heart rate	Numerical
Breathing Rate (breaths/min)	Respiratory rate	Numerical
Sweating Level (1-5)	Degree of sweating	Numerical
Dizziness	Experience of dizziness	Boolean
Medication	Use of medication for anxiety	Boolean
Therapy Sessions (per month)	Monthly therapy sessions	Numerical
Recent Major Life Event	Recent stressful life event	Boolean
Diet Quality (1-10)	Nutritional quality of diet	Numerical
Anxiety Level (1-10)	Self-reported anxiety level	Numerical

3.2. Data analysis

In this section univariate, bivariate and multivariate analysis is carried out on the dataset to uncover underlying relationships between attributes. By conducting this analysis, integrity of the features is ensured before they are used as inputs for the predictive models.

3.2.1. Univariate analysis

An initial analysis revealed information on the distribution of data considering numerical attributes. The mean and median are nearly equal for most dataset columns as described in *table 2*. Such a distribution implies that the data is not heavily influenced by outliers and approximates a normal distribution, which provides a statistically sound basis for the subsequent application of regression algorithms. By examining the consistency of each attribute, we mitigate the risk of introducing bias. The symmetry in the data distribution is a key indicator of the reliability of the data collection process.

Table 2. Descriptive statistics of dataset columns

Feature	Mean	Median	Std	Min	25%	50%	75%	Max
Age	40.24	40.00	13.23	18.00	29.00	40.00	51.00	64.00
Sleep Hours	6.65	6.70	1.22	2.30	5.90	6.70	7.50	11.30
Physical Activity	2.94	2.80	1.82	0.00	1.50	2.80	4.20	10.10
Caffeine Intake	286.09	273.00	144.81	0.00	172.00	273.00	382.00	599.00
Alcohol Consumption	9.70	10.00	5.68	0.00	5.00	10.00	15.00	19.00
Stress Level	5.86	6.00	2.92	1.00	3.00	6.00	8.00	10.00
Heart Rate	90.92	92.00	17.32	60.00	76.00	92.00	106.00	119.00
Breathing Rate	20.96	21.00	5.16	12.00	17.00	21.00	25.00	29.00
Sweating Level	3.08	3.00	1.39	1.00	2.00	3.00	4.00	5.00
Therapy Sessions	2.43	2.00	2.18	0.00	1.00	2.00	4.00	12.00
Diet Quality	5.18	5.00	2.89	1.00	3.00	5.00	8.00	10.00
Anxiety Level	3.93	4.00	2.12	1.00	2.00	4.00	5.00	10.00

3.2.2. Bivariate analysis

A Pearson correlation matrix was computed to assess the relationships between numerical variables. The strongest positive correlation (r = 0.67) was observed between Anxiety Level and Stress Level, suggesting that stress is a significant contributing factor. A negative correlation (r = -0.49) was identified between Sleep Hours and Anxiety Level, suggesting that reduced sleep may exacerbate symptoms. A negative correlation (r = -0.41) between Diet Quality and Anxiety Level was also identified. The results of the therapeutic sessions indicated a slight mitigating effect (r = -0.22), while Caffeine Intake exhibited a weak positive correlation (r = 0.35) with Anxiety Level.

3.2.3. Multivariate analysis

Multivariate analysis indicates that people experiencing the highest levels of anxiety tend to have elevated stress levels, shorter sleep duration, and lower levels of physical activity as shown in *figure 1*.

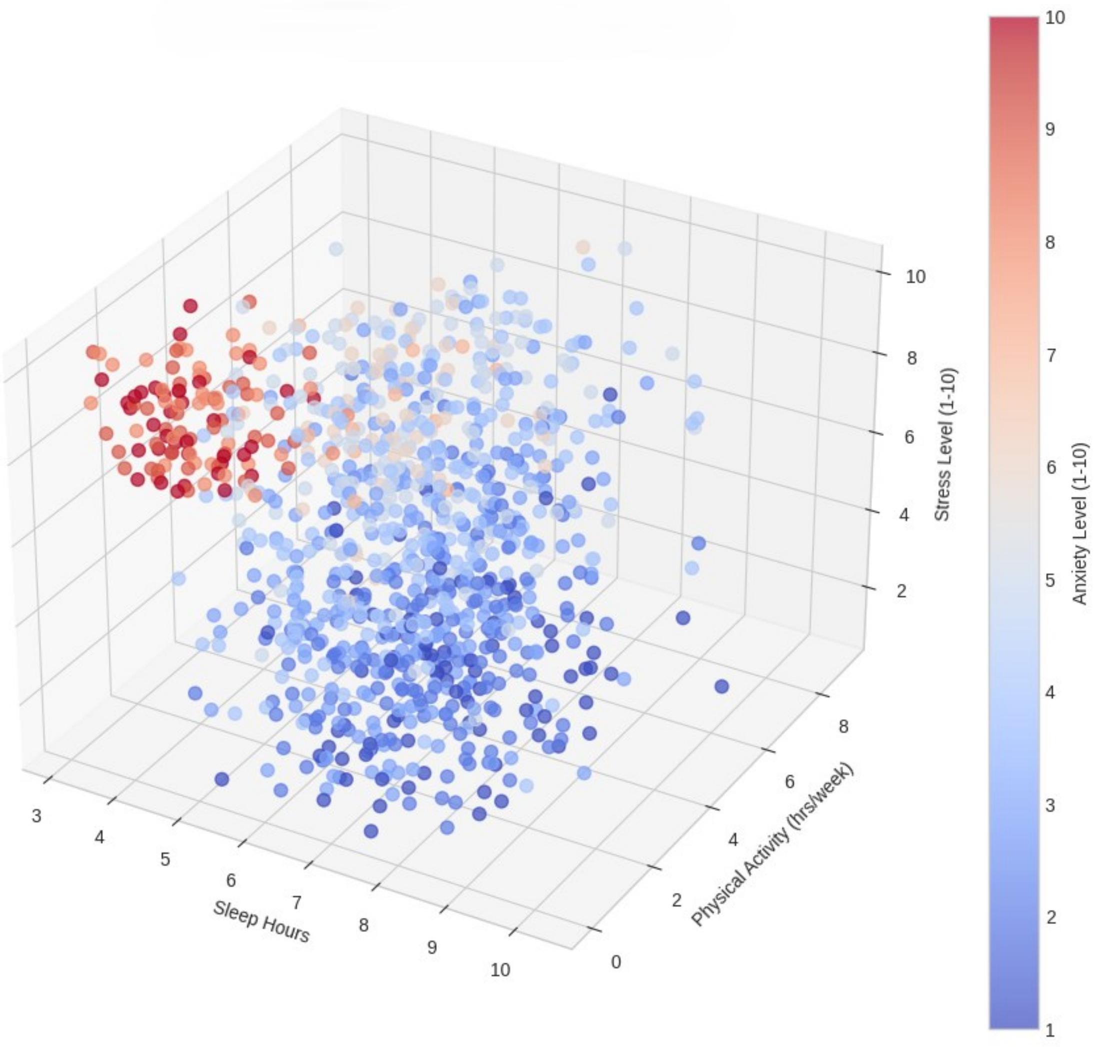

Fig. 1. Relationship between sleep, activity, stress and anxiety

3.3. Data preprocessing

To prepare the data for ML models, several preprocessing steps were implemented. Columns with a weak correlation with the target variable were eliminated, leaving five features, excluding the target variable. The remaining features are Sleep Hours, Caffeine Intake, Stress Level, Physical Activity and Therapy Sessions. The Physical Activity feature and the Therapy Sessions feature were transformed using log1p transformation to remove outlier bias. The values of all columns were scaled using min-max scaling. The data set was partitioned into a training set and a testing set, with the training set constituting 80% of the total data and the testing set comprising the remaining 20%.

4. EXPERIMENTS

Experimental procedures were conducted on the Google Colab platform using Python version 3.11. Data processing and visualization were supported by standard libraries, including Scikit-learn, Pandas, and Matplotlib.

The performance of all models is shown in *table 3*. Five-fold cross-validation was implemented for all models to obtain an average cross-validated R² score. Gradient Boosting Regression delivered the best performance, having a cross-validated R² of 0.759. Support Vector Regression and Light Gradient-Boosting Machine Regression attained comparable performance, with mean R² values of 0.758 and 0.756, respectively. Random Forest Regression demonstrated a slightly lower mean R² value of 0.749. Extreme Gradient Boosting Regression obtained a performance of 0.733. Linear Regression demonstrated significantly lower performance than previous models, with an R² value of 0.688. Decision Tree Regression exhibited the lowest performance, with a mean R² of 0.528.

Table 3. Performance comparison of the regression models

Model	MAE	RMSE	R^2	CV R^2 (mean)
Gradient Boosting Regression	0.822	1.032	0.770	0.759
Support Vector Regression	0.828	1.031	0.770	0.758
Light Gradient-Boosting Machine Regression	0.829	1.030	0.771	0.756
Random Forest Regression	0.851	1.058	0.759	0.749
Extreme Gradient Boosting Regression	0.869	1.079	0.749	0.733
Linear Regression	0.942	1.189	0.695	0.688
Decision Tree Regression	1.093	1.439	0.553	0.528

5. CONCLUSIONS

This study explored the predictive modeling of anxiety severity using a combination of demographic, behavioral, and physiological data derived from a large-scale survey dataset. After correlation analysis, the research identified stress level, number of sleep hours, physical activity, caffeine intake, and number of therapy sessions as factors significantly correlated with anxiety.

Among the suite of regression models evaluated, Gradient Boosting Regression emerged as the most effective, achieving the highest performance after cross-validation. Support Vector Regression and Light Gradient-Boosting Machine Regression also delivered comparable results, demonstrating the utility of ensemble and kernel-based models in mental health prediction tasks.

Key insights from the analysis reaffirm the central role of stress, sleep, and physical activity in influencing anxiety levels. These findings are consistent with existing psychological research and underline the potential for ML to support data-driven mental health assessments and early intervention tools.

However, certain limitations must be acknowledged. The reliance on self-reported survey data introduces potential response bias, and the cross-sectional nature of the dataset prevents observation of temporal trends. These factors should be addressed in future research for more comprehensive mental health modeling.

REFERENCES

[1] World Health Organization, *World Mental Health Report: Transforming Mental Health for All*, World Health Organization, 2022.

[2] J. Chung, J. Teo, *Single classifier vs. ensemble machine learning approaches for mental health prediction*, Brain informatics, vol. 10, no. 1, p. 1, 2023.

[3] T. Jain, A. Jain, P.S. Hada, H. Kumar, V.K. Verma, A. Patni, *Machine learning techniques for prediction of mental health*, 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1606-1613, 2021.

[4] A.E. Tate, R.C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, R. Kuja-Halkola, *Predicting mental health problems in adolescence using machine learning techniques*, PloS one, vol. 15, no. 4, 2020.

[5] H. Anckarsäter, S. Lundström, L. Kollberg, N. Kerekes, C. Palm, E. Carlström, N. Långström, P.K. Magnusson, L. Halldner, S. Bölte, C. Gillberg, C. Gumpert, M. Råstam, P. Lichtenstein, *The child and adolescent twin in Sweeden (CATSS)*, Twin Research and Human Genetics, vol. 14, no. 6, pp. 495-508, 2011.

[6] R. Garriga, J. Mas, S. Abraha, J. Nolan, O. Harrison, G. Tadros, A. Matic, *Machine learning model to predict mental health crises from electronic health records*, Nature medicine, vol. 28, no. 6, pp. 1240-1248, 2022.

[7] S. Bhatnagar, J. Agarwal, O.R. Sharma, *Detection and classification of anxiety in university students through the application of machine learning*, Procedia Computer Science, vol. 218, pp. 1542-1550, 2023.

[8] T. Richter, B. Fishbain, A. Markus, G. Richter-Levin, H. Okon-Singer, *Using machine learning-based analysis for behavioral differentiation between anxiety and depression*, Scientific reports, vol. 10, no. 1, p. 16381, 2020.

[9] M.D. Nemesure, M.V. Heinz, R. Huang, N.C. Jacobson, *Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence*, Scientific reports, vol. 11, no. 1, p. 1980, 2021.

[10] P. Kumar, S. Garg, A. Garg, *Assessment of anxiety, depression and stress using machine*

learning models, Procedia Computer Science, vol. 171, pp. 1989-1998, 2020.

[11] A. Priya, S. Garg, N.P. Tigga, *Predicting anxiety, depression and stress in modern life using machine learning algorithms*, Procedia Computer Science, vol. 167, pp. 1258-1267, 2020.

[12] A.V. Chavanne, M.L. Paillere Martinot, J. Penttilä, Y. Grimmer, P. Conrod, A. Stringaris, B. Van Noort, C. Isensee, A. Becker, T. Banaschewski, et al., *Anxiety onset in adolescents: a machine-learning prediction*, Molecular psychiatry, vol. 28, no. 2, pp. 639-646, 2023.