Group project: predicting mental health risk factors

Team: Kadi Tulver, Rauno Tali, Tõnn Sikk
Github repo: https://github.com/tondza/PAT2023_Menta

1. Introduction

1.1 Motivation

The importance of mental health has significantly escalated in the public consciousness, a trend further accelerated by the challenges of the COVID-19 pandemic. This heightened awareness aligns with observed global trends. Studies have shown a significant increase in depression and anxiety disorders worldwide, affecting a broad spectrum of age groups, from youth to the elderly (Thombs, et al., 2020) and across various professional sectors (Oakman et al., 2020). According to a study on the mental health of the population conducted by the National Institute for Health Development (TAI) and the University of Tartu, nearly a quarter of adults have at least one diagnosed psychiatric disorder, with depression and anxiety

disorders being the most prevalent (Summary press release, including the full report link: https://www.tai.ee/et/uudised/uuring-levinuimad-vaimse-tervise-probleemid-depressioon-ja-arevushaired).

Given the limited accessibility to mental health care, the best approach to addressing this global problem is prevention. To achieve this, it's crucial to identify and understand the factors contributing to mental health disorders. Analyzing data collected from a broad spectrum of geographic and demographic backgrounds allows us to pinpoint key indicators for predicting mental health issues, irrespective of cultural context. This project primarily focuses on measures of depression, anxiety, and stress, which are particularly prevalent mental health struggles globally (World Health Organization, 2017).

Two members of our project team, Tõnn and Rauno, are active in the IT field, where there has been an increasing focus on maintaining employees' mental health through training and compensated services. Remote work, common in the IT sector, especially post-pandemic, often leads to individuals facing mental health issues alone. The analyses conducted in this project could be practically beneficial, not only to our team members but also to everyone with whom this information is shared, thus enhancing awareness of mental health problems and their early prevention. Kadi, the third team member, is a researcher in psychology who has previously shared information on mental health with workgroups. Direct engagement with such data and a clearer understanding of potential risk factors would further improve the illustration and dissemination of this information.

1.2 Goals

Our project aims to predict the likelihood of individuals developing mental health issues based on various parameters such as age, gender, education, and personality type. In data science terms, our goal is to create prediction model(s) with high enough accuracy to be considered relevant. Additionally, we plan to visualize key findings in an engaging and easily comprehensible manner.

More generally, we also seek to raise awareness about mental health among course participants and within our work communities.

2. Data sources and description

2.1 Database

We used an open-source database that comprised data collected through an online survey between 2017-2019. The survey included a questionnaire measuring mental health symptoms descriptive of depression, anxiety, and stress, as well information about demographic background and personality. The online survey was available to the general public, and responders received personalized results at the end. Only answers from people who consented that their data could be used in research and had confirmed that they had

been truthful in their answers, were included in the database. The full database includes responses from 39775 people.

The database is available at the following source: Greenwell, L. (2019). Depression, Anxiety, Stress Scales Responses [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/lucasgreenwell/depression-anxiety-stress-scales-responses.

2.2 Measures

The Depression Anxiety Stress Scales (DASS) were developed by S. H. Lovibond and P. F. Lovibond between 1979 and 1990 as a set of instruments designed to measure the emotional states of depression, anxiety, and stress. One of the objectives of DASS was to differentiate between the symptoms of depression and anxiety and capture their unique aspects. A third factor reflecting general tension (irritability, agitation) emerged from initial testing which was then labeled as "stress". The final version of DASS contains 42 items, which are divided evenly among the three subscales. While initially derived from clinical consensus, the scale has been refined through statistical methods to ensure reliability and validity in measuring these affective states within both clinical and nonclinical populations. Notably, despite the intention to distinguish among the three states, the subscales have shown moderate to high intercorrelations. Since its creation, the DASS has become a widely utilized tool in research due to its robust psychometric properties and its availability at no cost, with both the full and the shortened 21-item version (DASS21) being employed in a variety of research contexts (Yeung, Yuliawati, & Cheung, 2019). The items were measured on a 4-point scale, ranging from 0 = Did not apply to me at all to 3 = Applied to me very much, or most of the time.

In addition to DASS, the Ten Item Personality Inventory (TIPI, Gosling et al., 2003) was administered in the survey. The TIPI includes two items for each of the five dimensions of the Five Factor personality model. The TIPI was designed as a quick method for estimating personality traits as an alternative to longer questionnaires. The TIPI captures the dimensions of Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience. Items were measured on a 7-point scale ranging from 1 = Disagree Strongly to 7 = Agree Strongly.

3. Method and results

3.1 Cleaning and preprocessing

To prepare data for analysis, we explored the data and removed outliers, nonsensical responses, and activity indicative of not paying attention. As a result, we made the following decisions (see “Dataset cleaning and preparation” notebook for more details):

- Removing responses where the same answer was given to all 42 of the DASS scale questions, as this is highly indicative of simply clicking through answers.

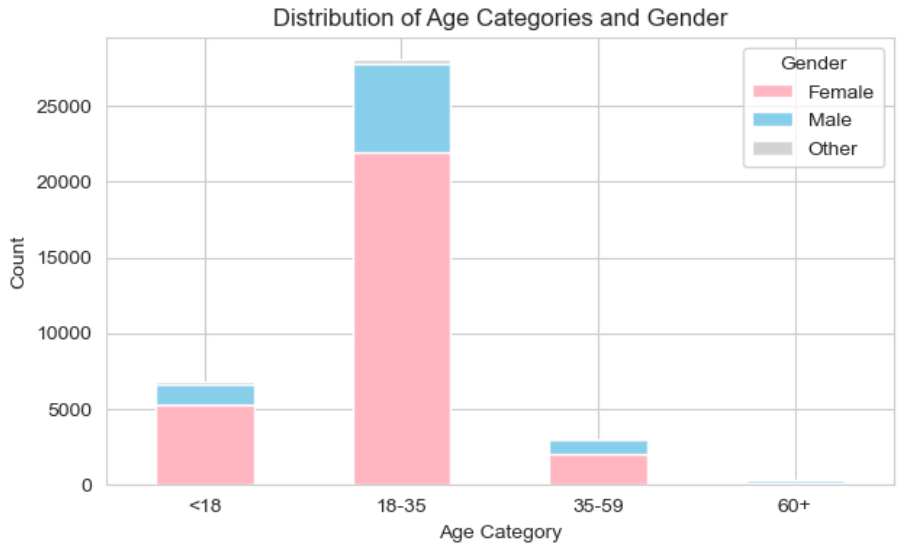
- The survey had included an attention/trustworthiness measure where responders were asked to identify words whose meaning they knew, out of which three items were nonwords. We removed responses where at least two out of the three trick questions were not identified correctly.
- We removed responses where the average response time was less than 1 second, i.e., unrealistically quick, indicating again that the person clicked through the answers without thinking about it.
- We removed outliers based on the continuous variables “age” and “familysize” - removing rows where the age was 90 or higher, and where the family size was above 13 which were in some cases likely to be mistakes or simply out of distribution values. The cut-offs were selected based on plotting the distribution (i.e., visual estimation).

After these steps, 1,442 rows were dropped, which represents approximately 3.63% of the initial dataset.

3.2 Descriptive statistics

3.2.1 Demographic distribution

The study's participants comprised a diverse sample from various countries (N=38333), with the largest groups from Malaysia (55.25%) and the USA (19.85%). 59% of the responders identified as Asian, 26% as White, and 12% as belonging to other racial groups. The predominant religion in the sample was muslim (56.93%). 38.81% of the sample reported having a university degree (undergraduate) and 12.81% a graduate degree. 38.32% had graduated high school, and 10.06% had not. The gender distribution was predominantly female (76.8%), with 21.8% identifying as male. 86.4% had reportedly never been married, 10.9% were married at the time of participation. The average age was 23 years (SD=8.5). The distribution of age categories and gender is illustrated below, highlighting that the majority of responders were young adults (between 18 and 35).

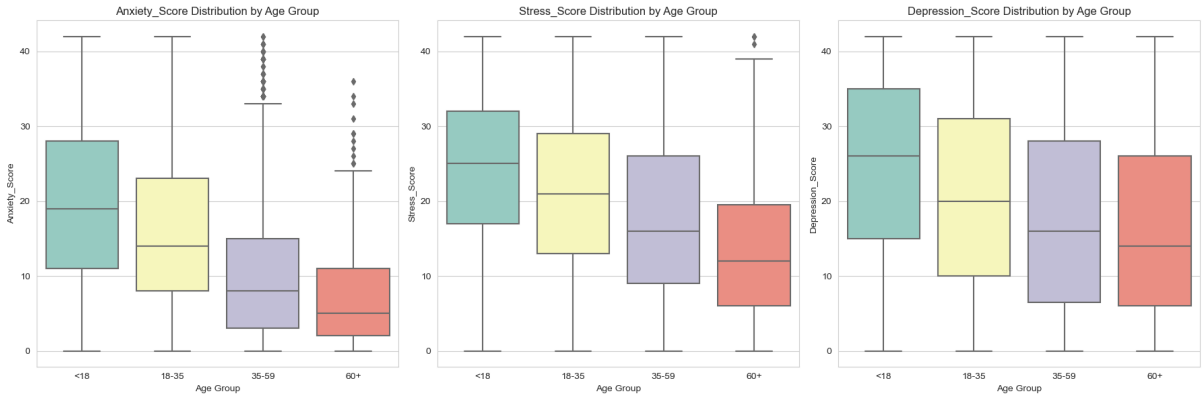


3.2.2 Mental health and personality

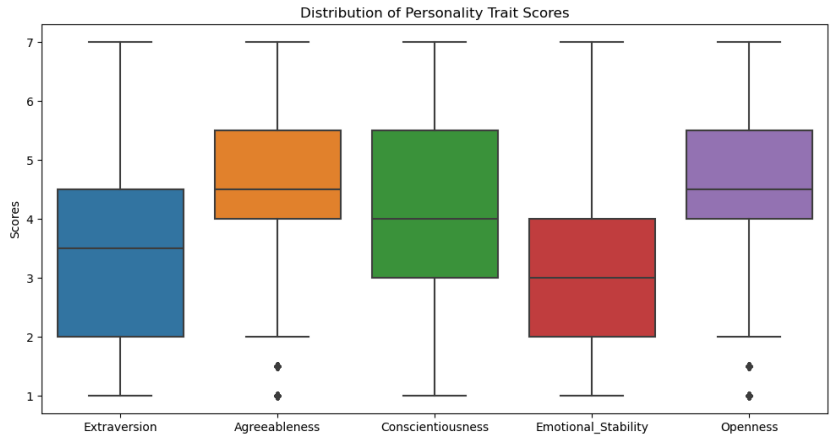
The distribution of DASS scores across sample (N=38333) is illustrated in the histograms below. Furthermore, lines have been added to show the cut-off points of different classes of severity (normal, mild, moderate, severe, extreme).

The average depression score in the sample was 20.99 (SD=12.22), average anxiety scores 15.95 (SD=10.08), and average stress score 21.09 (SD=10.39). Compared to the normative sample scores reported by the authors of the scale (Lovibond & Lovibond, 1995) – Depression: 6.34 (SD=6.97), Anxiety: 4.7 (SD=4.91), Stress: 10.11 (SD=7.91) – it is evident that the scores in this sample are markedly higher. This suggests an elevated prevalence and intensity of mental health problems among the survey responders relative to the general population which indicates that the survey may have been specifically targeted toward risk groups.

The distribution of DASS scores across age groups is plotted below, illustrating that overall, symptoms related to depression, anxiety, and stress tend to decrease with age and that the youngest age groups may be most at risk for developing symptoms.

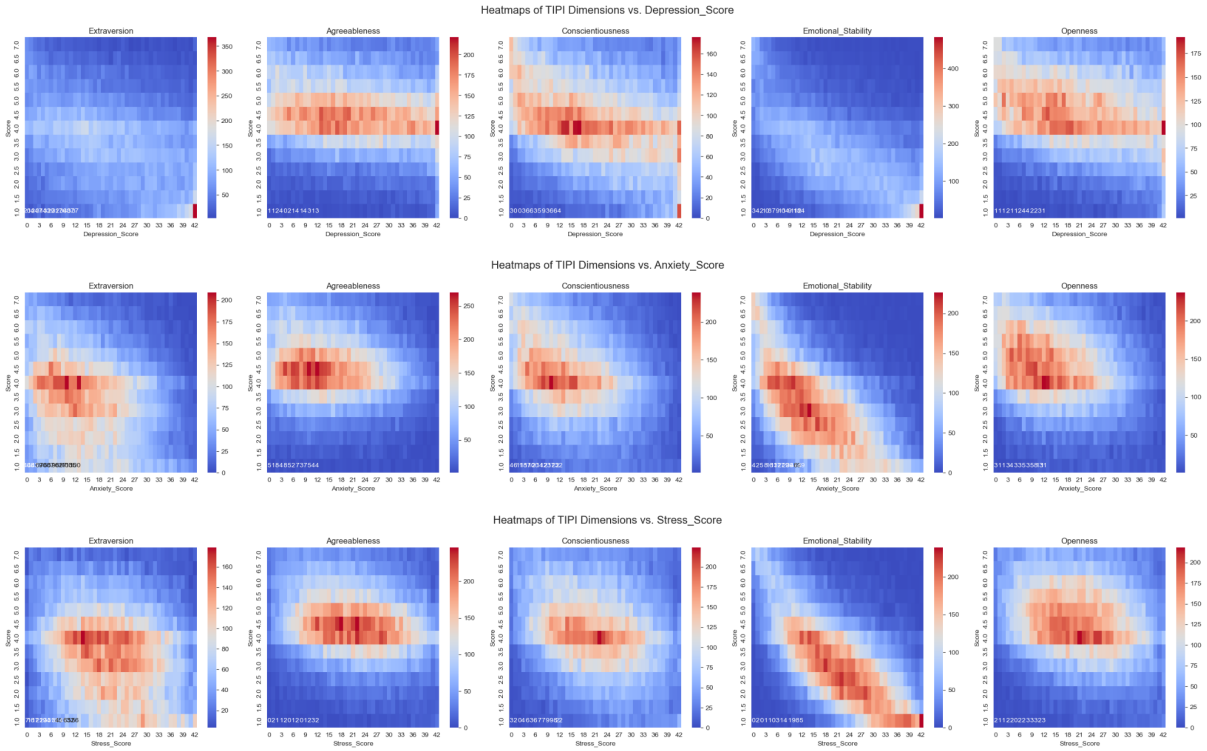


The distribution of the five trait dimensions are depicted in the boxplot figure below.



To explore the potential link between personality traits and mental health, we calculated Spearman's rho correlation coefficients for the combinations of variables as well as plotted heatmaps to illustrate the relationships.

	Anxiety_Score	Stress_Score	Depression_Score
Extraversion	-0.178068	-0.170517	-0.292452
Agreeableness	-0.109462	-0.195180	-0.155523
Conscientiousness	-0.232678	-0.222042	-0.291257
Emotional_Stability	**-0.543459**	**-0.643722**	**-0.523845**
Openness	-0.201843	-0.188473	-0.223576



All the correlations between the DASS and personality dimensions were negative and statistically significant ($p < 0.001$), indicating that higher scores on extraversion, conscientiousness, openness, and agreeableness are all linked with lower scores of depression, anxiety, and stress. However, the highest correlations were with the emotional stability dimension which displayed large effect sizes (>0.5) with all the DASS scales, suggesting that low scores on emotional stability might be a significant risk factor for developing mental health symptoms.

3.3 Binary Classification models

3.3.1 Data Preparation

The initial phase involved meticulously processing the dataset to structure it for binary classification. Variables such as age, gender, education level, and personality traits like Extraversion and Agreeableness were prepared alongside mental health scores. The mental health scores (Depression, Anxiety, and Stress) were converted into binary classes based on predetermined thresholds, classifying them into 'normal/mild' (0) and 'moderate/worse'.

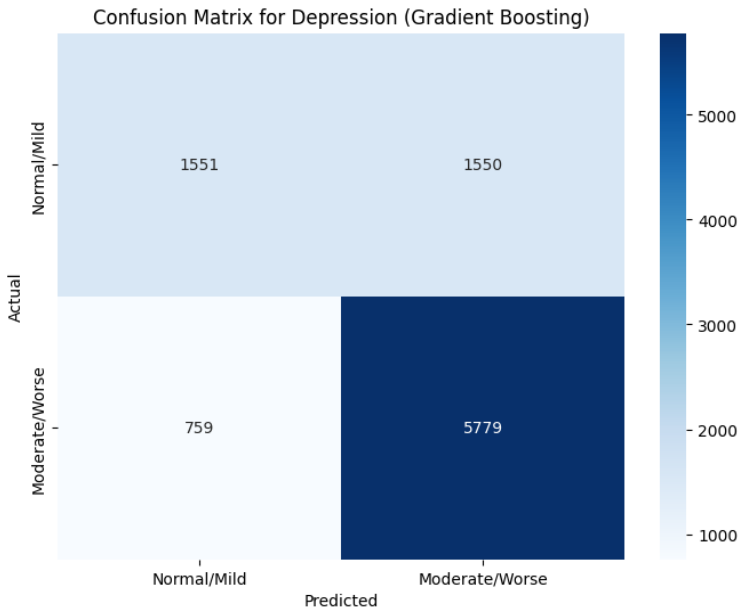
3.3.2 Model Selection

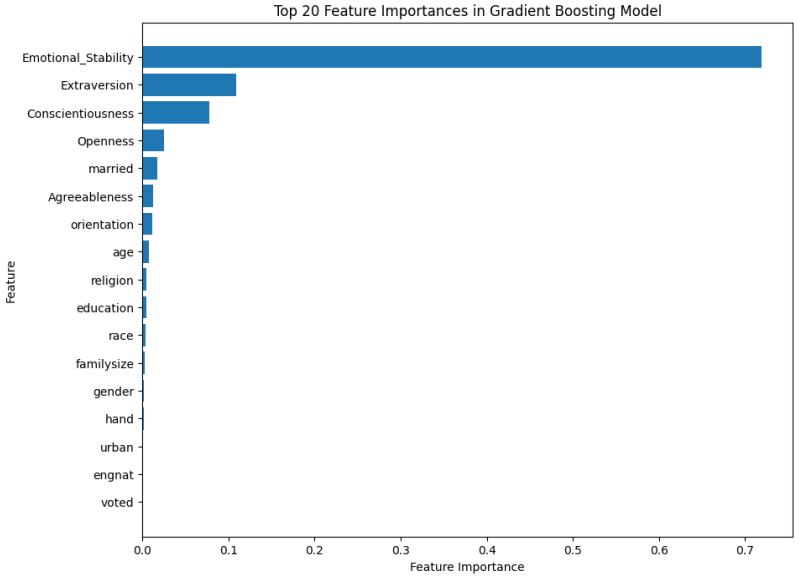
Several classification models were employed to analyze the dataset: Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting Classifier. Each model was trained on three different sets of data targeting Depression, Stress, and Anxiety scores.

3.3.3 Results

Depression Model

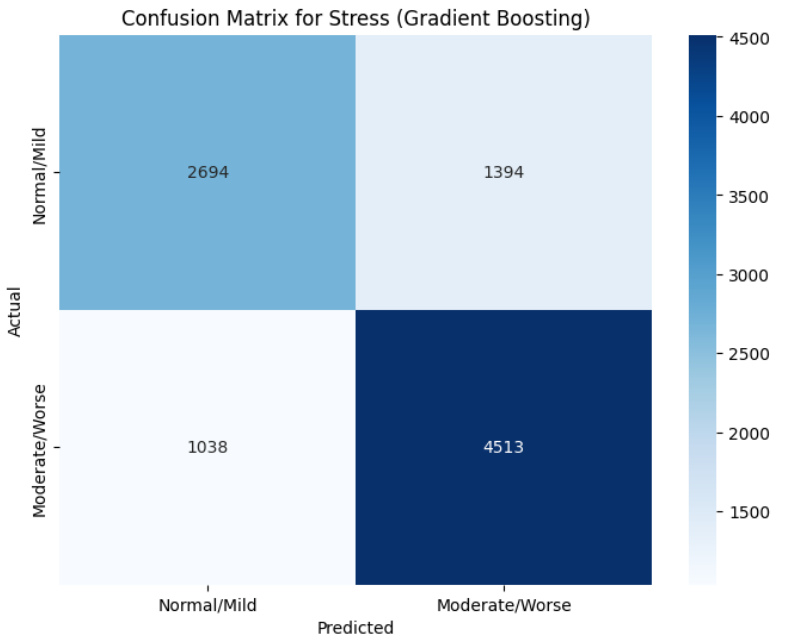
- Logistic Regression: Achieved an accuracy of 76%, with a recall of 0.89 for the moderate/worse class.
- Random Forest: Similar accuracy to Logistic Regression but with a slightly better recall for the moderate/worse class.
- Decision Tree: Slightly lower accuracy at 75% but a balanced performance across classes.
- Gradient Boosting: Best performer with an accuracy of 76%, showing substantial recall for the moderate/worse class.
- Feature Importance based on Gradient Boosting: Emotional Stability (0.71), Extraversion (0.10), Conscientiousness (0.07).

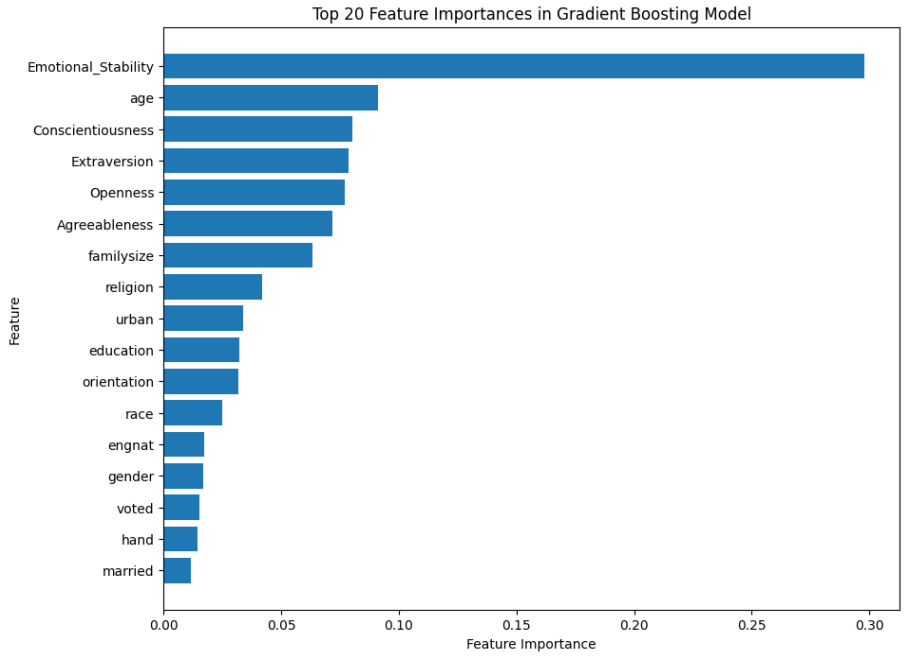




Stress Model:

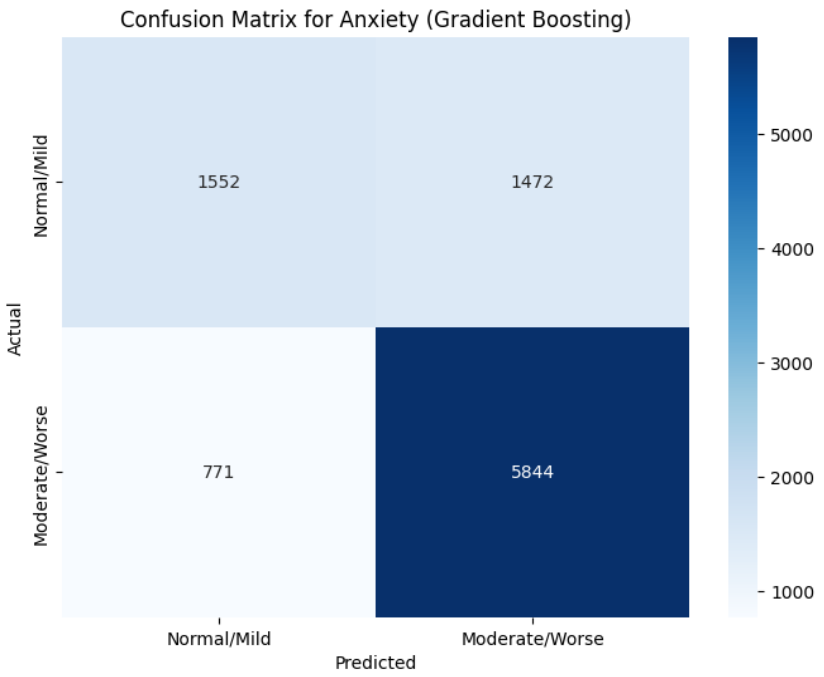
- Logistic Regression: 75% accuracy, better performance in identifying moderate/worse cases.
- Random Forest and Decision Tree: Both showed comparable accuracies around 74%.
- Gradient Boosting: Topped with a 75% accuracy, excelling in classifying the moderate/worse cases.
- Feature Importance based on Gradient Boosting: Emotional Stability (0.28), Age (0.08), Conscientiousness (0.07).

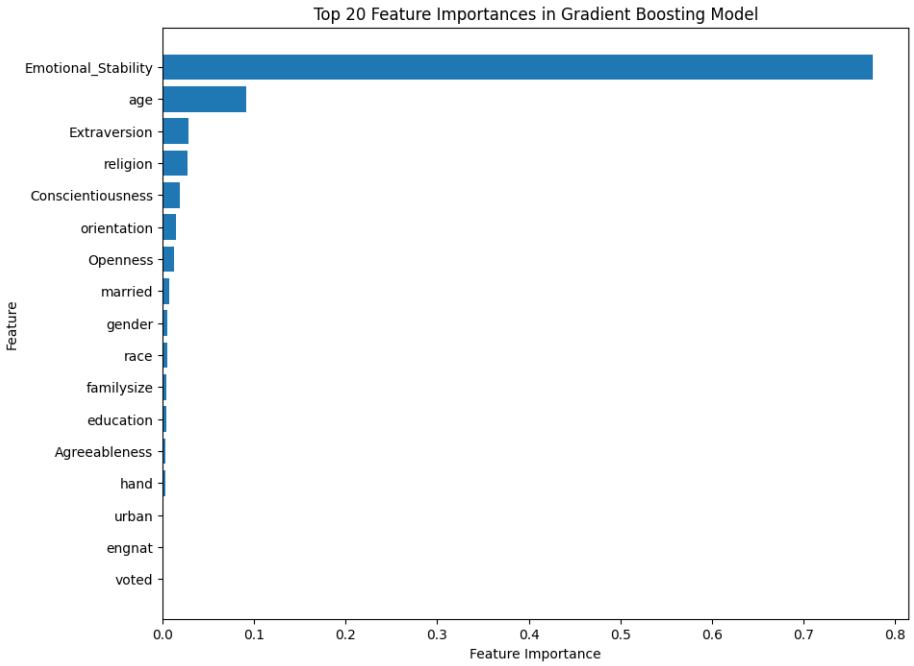




Anxiety Model:

- Logistic Regression: 77% accuracy, favoring the moderate/worse class in prediction.
- Random Forest: Comparable performance to Logistic Regression.
- Decision Tree: Slightly lower performance, with 76% accuracy.
- Gradient Boosting: Emerged as the best model with 77% accuracy.
- Feature Importance based on Gradient Boosting: Emotional Stability (0.77), Age (0.09), Extraversion (0.02).





Confusion Matrix Insights brought out that for all models and conditions, the tendency to predict moderate/worse cases was higher, which is crucial for sensitive applications like mental health assessment.

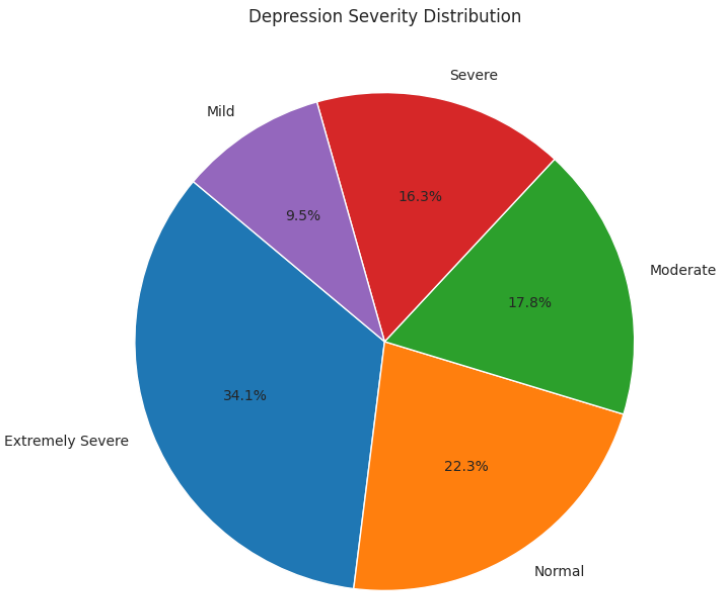
3.4 Category classification models

The goal of this approach is to train a model that would predict the **exact severity class of a mental health condition** (Stress / Anxiety / Depression), possible classes being: Normal, Mild, Moderate, Severe and Extremely Severe.

All the notebooks for these kinds of classification models can be found under the *"Category classification models"* folder.

3.4.1 Data preparation and pre-processing

- One of the key questions was whether those classes would be evenly distributed in the dataset. Taking depression scores as an example, the distribution was not even, so minority oversampling technique (SMOTE was considered when preparing training data):



  However, since models trained with SMOTE data didn't always have higher accuracy, datasets without SMOTE were also tried out
- For missing values treatment, mode imputation was used for categorical values and mean imputation for numerical values. However, as there are models that can handle NaN values as well, version of dataset without mode imputation was tried out
- When starting with just "demographic" question columns in the dataset, "TIPI" columns representing personality traits were later added and proved to increase model performance
- Because of many different combinations to be tried out, a "programmatic approach to dataset preparation" was implemented in *Mental health severity classes prediction - programmatic approach.ipynb*
- Later, a datafile that included most of the preparations was used for predictions, then ordinally encoded severities were used for prediction

3.4.2 Models and parameters used

- Logistic Regression was a fast-performing model with constantly decent results
- Models tried without mode imputation (that could handle NaN values), performed quite well: HistGradientBoosting, XGBoost, LightGBM, CatBoost
- Also tried: Random forest, KNN, SVM
- Main metric to evaluate model performance was its accuracy score (although sometimes it failed to identify a faulty model where classes were seriously imbalanced)

3.4.3 Results

Lessons learned

- Although using SMOTE (Synthetic Minority Over-sampling Technique) technique for balancing prediction classes seemed unnecessary at first, giving lower model accuracy scores, on a closer look it became evident that the non-balanced predictions were totally missing some classes and thus the increased accuracy was fake.
- Hyperparemeter tuning had little effect on improving the accuracy of any model.
- It was assumed that feature-engineering with TIPI questions (combining them into personality traits) would improve model performance, but it didn't.
- It was assumed that ordinally encoding the mental health severity classes to be predicted would improve model performance, but it didn't.
- Incorrect training data can also result in seemingly better predictions. There was a case when our dataset filtering notebook (Dataset cleaning and preparation.ipynb) had a mistake and as a result the dataset had just 3 different severity classes instead of 5. That significantly increased accuracy of trained models (accuracy close to 0.72).

Best models

Unfortunately, the accuracy of the best models doesn't go above 0.5. This is likely because the number of possible severity classes in each condition (5) and greater "chance to miss" than in binary classification.

For depression prediction, best accuracy was achieved with:

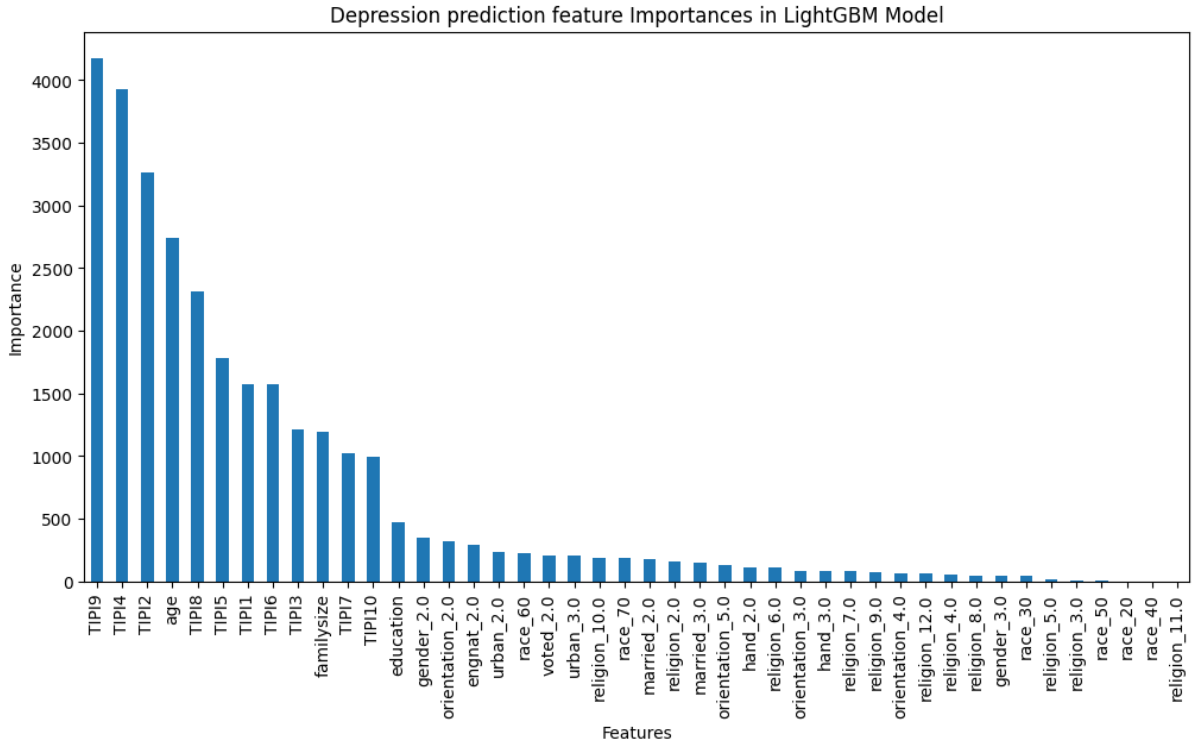
- LightGBM model with parameters `{'learning_rate': 0.01, 'n_estimators': 200, 'num_leaves': 31}`
- TIPI questions included (separately, not combined into personality traits)
- SMOTE not used
- NaN values included in training set

Later, same configuration was used to create prediction models for **Anxiety** and **Stress**

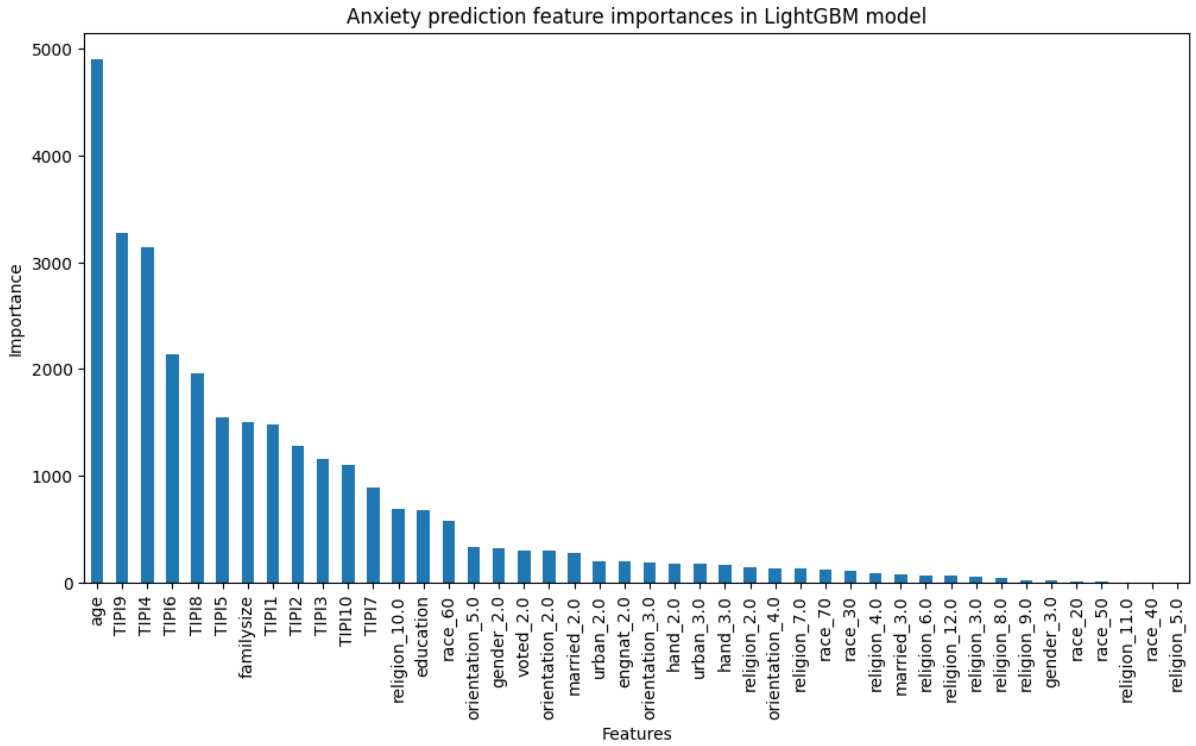
- Anxiety model accuracy = 0.49
- Stress model accuracy = 0.46

Features of best models

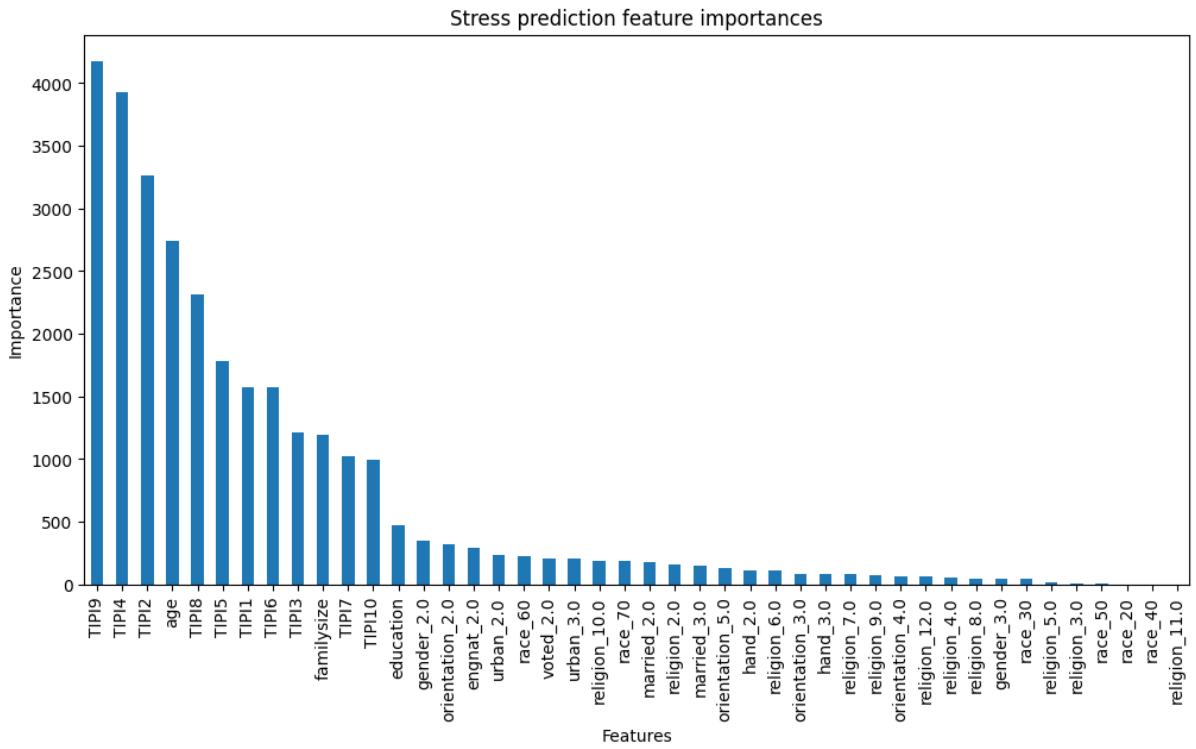
Depression:



Anxiety:



Stress:



Meanings of TOP TIPI columns:
TIPI9 Calm, emotionally stable.
TIPI4 Anxious, easily upset.
TIPI2 Critical, quarrelsome.
TIPI8 Disorganized, careless.

3.5 Continuous score prediction models

3.5.1 Data preparation

For preparing the data, a subset of 20 relevant variables was selected for inclusion in the preliminary model. These included the target measures, as well as the five personality dimensions and several measures for demographic background. After dropping all NAs, 32129 data rows remained. For categorical variables, one-hot encoding was performed, converting each category into a new binary column (numerical input) which facilitates their use in the subsequent modeling process. Continuous variables were standardized to have a mean of zero and a standard deviation of one. This step equalizes the influence of each feature on the model and helps prevent any feature from disproportionately impacting the model's predictions due to its scale. Then, data was split into training and testing sets using scikit-learn and maintaining a standard 80-20 split. This process was repeated for each of the three target variables. The training set size was 25703 and the testing set 6426 rows of data.

3.5.2 Model selection

The dataset includes a mix of continuous and categorical variables that can be used for predicting mental health estimations (Depression, Anxiety, and Stress Scores). To enable the prediction of the continuous scores of DASS scales, three distinct models were chosen: Ridge Regression, Random Forest, and XGBoost.

- As a linear model with L2 regularization, Ridge Regression should allow for capturing linear relationships between the features and the target variables. Its regularization helps prevent overfitting, making it robust, especially when dealing with multicollinearity or datasets with a high number of features.
- Random Forest is a tree-based ensemble method suitable for handling non-linear relationships and complex interactions between features. It works well with a mix of numerical and categorical data and is less likely to overfit than individual decision trees.
- XGBoost is an advanced gradient boosting algorithm known for its performance and speed. It's particularly effective in scenarios where the relationship between the variables is intricate and non-linear. Its flexibility in tuning and handling various types of data makes it a good candidate for achieving high accuracy in predictions.

3.5.3 Results

For this analysis, MSE and R2 were chosen to evaluate model fit due to their relevance in regression analysis and their widespread acceptance in statistical modeling. MSE provides a direct measure of the model's accuracy in terms of prediction errors, while R2 offers an understanding of how well the model explains the variability in the target variable. The results for each of the models, as well as their most important contributing features, are summarized in the table below.

Model	Target Variable	Mean Squared Error (MSE)	R-squared (R^2)	Top Contributing Features
Ridge Regression	Depression	97.148	0.350	Emotional stability (TIPI9), Previously married, Religion Buddhist
	Anxiety	66.079	0.345	Race Indigenous Australian, Emotional stability (TIPI4), Left handed
	Stress	58.213	0.454	Emotional stability (TIPI4, TIPI9), Religion Christian (Mormon)
Random Forest	Depression	97.393	0.349	Emotional stability (TIPI9, TIPI4), Age
	Anxiety	64.551	0.360	Emotional stability (TIPI4, TIPI9), Age

	Stress	56.588	0.469	Emotional stability (TIPI9, TIPI4), TIPI1
XGBoost	Depression	93.011	0.377	Emotional stability, Extraversion, Conscientiousness
	Anxiety	62.235	0.383	Emotional stability (TIPI4, TIPI9), Never married
	Stress	53.984	0.493	Emotional stability (TIPI4, TIPI9), TIPI2

3.5.4 Summary

To summarize, Ridge Regression displayed a moderate performance with MSE scores of around 97, 66, and 58 (for Depression, Anxiety, and Stress Scores, respectively). The R^2 values suggest that the model could explain around 35-45% of the variance in these scores. Random Forest, an ensemble method, showed a similar pattern in MSE and slight improvements in R^2. This suggests an enhanced ability to capture non-linear relationships, with a somewhat limited boost in predictive power. XGBoost offered the best performance among the three, particularly for Stress Scores (with MSE of around 54 and R^2 of almost 0.5) which was to be expected due to its ability to handle a variety of data types and capture complex patterns.

In almost all models, items for emotional stability emerged as the most important feature across all targets. This suggests that higher levels of emotional stability may serve as a protective factor against mental health problems. Individuals with greater emotional stability appear to be less prone to depressive symptoms, indicating that emotional stability could play a crucial role in mitigating the risk of developing depression and anxiety.

4. Discussion

In summary, we found that predicting binary classes of DASS scores (normal/mild vs severe cases) was the easiest task for models, and predicting continuous scores proved most challenging. The classification task of predicting severity categories might have been complicated due to the somewhat unbalanced distribution of severity classes in the sample. Overall, gradient boosting models (XGBoost, LightGBM) exhibited superior performance.

Mental health symptoms are influenced by a complex interplay of factors, most of which were likely not fully captured in the dataset. This complexity inherently limits the precision that such models could attain. The main takeaway from comparing feature importances of the most accurate models was that personality traits (especially emotional stability) is much more impactful in predicting mental health disorders compared to any demographic parameters.

While the models chosen performed well in general, there is potential in further tuning their parameters. Experimenting in depth with more advanced techniques like grid search or random search for hyperparameter optimization was out of the scope of our current project aims but could potentially yield better results in the future.

Although the choice and engineering of features significantly impacts model performance, we also saw that most of our efforts to improve data preparation did not result in notable changes to model evaluation parameters. It is very likely that there are more optimal ways to approach such a task when performed by more experienced data scientists.

Most importantly, however, the performance of the model will reflect the quality and representativeness of the dataset. Although the dataset included responders from many different countries and backgrounds, the categories were not ideally balanced. Also, the sample may have been somewhat skewed in the responses to the DASS questionnaire and hence may not have been entirely representative of the broader population.

References:

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in personality, 37(6), 504-528.

Greenwell, L. (2019). Depression, Anxiety, Stress Scales Responses [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/lucasgreenwell/depression-anxiety-stress-scales-responses.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behaviour research and therapy, 33(3), 335-343.

Oakman, J., Kinsman, N., Stuckey, R., Graham, M., & Weale, V. (2020). A rapid review of mental and physical health effects of working at home: how do we optimise health?. BMC public health, 20, 1-13.

Taquet, M., Holmes, E. A., & Harrison, P. J. (2021). Depression and anxiety disorders during the COVID-19 pandemic: knowns and unknowns. The Lancet, 398(10312), 1665-1666.

World Health Organization. (2017). Depression and Other Common Mental Disorders: Global Health Estimates. World Health Organization.

Yeung, A. Y., Yuliawati, L., & Cheung, S. H. (2020). A systematic review and meta‐analytic factor analysis of the Depression Anxiety Stress Scales. Clinical Psychology: Science and Practice, 27(4), e12362.