# Predicting Prevalence of Mental Health Disorders Using Supervised Learning

Maelice Yamdjieu, Ting-Yu Lin, & Vy Tran

## Abstract

This study explores the relationship between the common mental health disorders and socioeconomic factors using a dataset from the Worl Data repository. The predictor variables primarily focus on economic indicators, such as GDP per Capita and Unemployment Rate, and the percentage of behavioral responses to mental health challenges, including seeking professional psychological help, taking prescribed medications, and making lifestyle adjustments. The response variables are the prevalence rates of anxiety disorders, depressive disorders, and eating disorders. The study employs both regression and classification techniques to analyze the data, utilizing machine learning models such as Random Forest, Support Vector Machines (SVM), and logistic regression. Random Forest models initially explained 48% to 71% of the variance in mental health disorders, with improved R-squared values after hyperparameter tuning. Notably, the tuned models achieved an R-squared of 0.51 for anxiety disorders, 0.65 for depressive disorders, and 0.78 for eating disorders. SVM models demonstrate the highest accuracy for anxiety disorders with linear and polynomial kernels (67.74%), depressive disorders with a linear kernel (54.84%), and eating disorders with a polynomial kernel (90.32%). Logistic regression, after standardizing the data, shows improved prediction accuracy for anxiety (76.19%) and eating disorders (nearly 100%), but a decrease in accuracy for depressive disorders (47.62%). The findings highlight the strong relationship between GDP, unemployment rates, and prevalence of mental health disorders, providing valuable insights for policymakers, healthcare providers, and researchers to develop targeted interventions and strategies for improving mental health outcomes globally.

## Introduction

Mental health disorders represent a significant public health challenge worldwide, affecting millions of individuals across various demographic and socio-economic backgrounds. Understanding the factors that influence the prevalence of these disorders is crucial for developing effective interventions and policies. In this study, an exploration of the intricate interplay between mental health disorders and socioeconomic indicators across diverse countries worldwide is conducted using the “Mental Health” dataset from Our World in Data Repository (IHME, Global Burden of Disease Study (2019) – processed by Our World in Data). This comprehensive dataset comprises four main tables: Mental Health Prevalence, GDP, Behavioral Response, and Unemployment Rate allowing for a global analysis of the relationship between mental health, economic indicators, and behavioral factors. The main goal is to determine which predictor variables have the most significant impact on the prevalence of anxiety disorders and to assess the performance of various machine learning models in predicting these outcomes. Specifically, the study evaluates the Random Forest, Support Vector Machine (SVM), and Logistic Regression models. The initial analysis involves fitting these models to the data and then optimizing their performance through hyperparameter tuning. This study provides valuable insights into the relationship

between mental health disorders and socio-economic factors, demonstrating the effectiveness of various machine learning models in predicting the prevalence of anxiety disorders.

## Theoretical Background

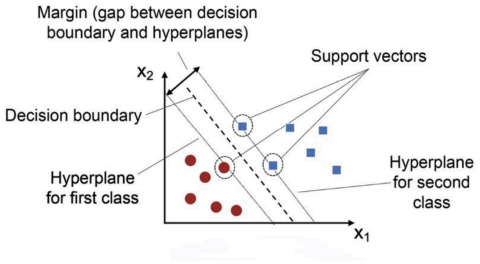
### *Random Forest*

Random forests are an advanced ensemble learning method that extends the principles of bagging by incorporating a technique that decorrelates individual trees, thereby improving overall model performance. Both bagging and random forests involve constructing multiple decision trees using bootstrapped samples from the training data. However, random forests introduce a key modification during the tree-building process: at each split, instead of considering all predictors, only a random subset of predictors is selected as candidates for the split. This subset is chosen anew at each split point. Random forests excel in scenarios with high-dimensional data and when there are many predictors, especially if these predictors are correlated. They offer several advantages, including the reduction of overfitting by averaging multiple trees and decorrelating them, mitigating the risk of overfitting that plagues single decision trees. Additionally, random forests provide insights into the importance of each predictor, aiding in feature selection and model interpretability. Furthermore, the method can maintain accuracy even with missing data points, making it robust for practical applications. Key parameters in random forests include the number of trees and the number of predictors at each split. The number of trees in the forest is typically large to ensure that the error rate stabilizes, as the model will not overfit with an increasing number of trees. The number of predictors randomly chosen at each split is crucial for decorrelating the trees. In conclusion, random forests enhance the predictive power and robustness of ensemble learning methods by introducing randomness in the predictor selection process at each split. This decorrelation of trees leads to lower variance and improved generalization performance, making random forests a powerful tool for both classification and regression tasks, particularly in high-dimensional settings.

### Support Vector Machines (SVM)

SVMs are a collection of techniques used to classify data in classes. The fundamental idea is to identify the hyperplane with the biggest margin of separation between classes. They are employed in regression and classification applications. They encompass different concept such as the maximal margin classifier which requires data to be linearly separable, the support vector classifier, an extension of the margin classifier design to work with broader datasets, and the support vector machine which build upon the support vector classifier can handle data with non-linear boundaries (figure1) [1]. Finding the best hyperplane with the biggest margin in a multidimensional space for splitting classes. Because of its maximum margin approach, SVM is less likely to overfit compared to certain other machine learning algorithms, which makes it perfect for a range of classification applications. It can define the decision boundary it forms using different kernel types. When a straight line or a plane can be used to segregate the data, the linear kernel is the most basic. The radial kernel provides additional flexibility when dealing with complicated patterns in the data since it can

accommodate curved or non-linear boundaries. By adding polynomial equations to create the hyperplane and allowing for an adjustable degree value that controls the boundary's complexity, the polynomial kernel increases versatility. Because of its adaptability to various data and its flexibility, SVM is a good fit for high-dimensional datasets. SVM can, however, be computationally demanding, particularly when using more complex kernels like Radial and polynomial, and training can require a significant amount of processing speed. This is particularly relevant for big datasets or when there are many features in the data. There is a strong correlation between the amount of support vectors and points in SVM and its computing complexity. The quantity of support vectors, that is, the points in the dataset that determine the hyperplane's position has a direct bearing on the computational complexity of SVM. SVM hyperparameters are important because they influence how the algorithm maintains a balance between minimizing misclassification and maximizing the margin. This trade-off is controlled by the cost parameter. Higher costs lead to strict boundaries but lower margins. The gamma parameter in the radial and polynomial kernels controls how much influence singles data points have on the hyperplane. Higher gamma values result in more complex boundaries. The degree of boundary complexity is determined by the degree parameter, which is unique to the polynomial kernel. Cross-validation methods are often employed to ensure that SVM models are stable and do not overfit. The training dataset is divided into subsets by cross-validation, especially 10-fold cross-validation, which enables the model to be validated on several subsets of the data. By ensuring that the model performs well when applied to unseen data, this approach improves the model's reliability as an indicator of performance. By selecting the appropriate kernel and fine-tuning hyperparameters, SVM can provide accurate and reliable results for a wide range of applications.




[1]

Figure 1: Support Vector Machine

***Logistic Regression***

Logistic regression is a statistical method widely used for binary classification problems, where the goal is to model the probability of a binary outcome based on one or more predictor variables. Unlike linear regression, which predicts a continuous outcome, logistic regression predicts the probability that a given input point belongs to a particular class. This makes logistic regression particularly suitable for problems where the response variable is categorical. At the core of logistic regression is the logistic function, also known as the sigmoid function, which maps any real-valued number into the interval (0, 1). The logistic function is defined as σ(z)=1/1+e^−z, where z is a linear combination of the input features X. The output of the logistic function is interpreted as the probability that the response variable Y equals 1, given the predictor variables X.

The probability that Y equals 0 is simply $1-\sigma(z)$. Logistic regression makes several key assumptions: linearity of the logit (the log-odds of the outcome is a linear function of the predictor variables), independence of observations, and no multicollinearity among the predictor variables. Despite its simplicity and interpretability, logistic regression has limitations. It assumes a linear relationship between the log-odds of the outcome and the predictor variables, which may not always be appropriate. It also may not perform well with very large datasets with complex relationships among variables, where more sophisticated models like SVM might be more suitable. It works well for small to moderately sized datasets and provides clear insights into the relationships between predictors and the probability of the outcome. Regularization techniques such as L1 (Lasso) and L2 (Ridge) can be incorporated to handle multicollinearity and prevent overfitting, making logistic regression a robust choice for binary classification problems. While it has limitations, particularly in handling complex, non-linear relationships, its strengths in interpretability and simplicity make it an essential method for a wide range of applications.

## Methodology

### *Data preprocessing*

Data processing and cleaning are crucial steps in preparing the dataset for analysis. In this study, the relevant data on mental health disorders, GDP per capita, and behavioral responses to mental health challenges were obtained from separate datasets for the year 2019. To ensure consistency and facilitate merging, the country identifiers were reformatted to match across all data sources as the 'Entity' column in each file was renamed to 'Country'. Then the rows and columns with missing or incomplete information were removed as mostly in country code and continents columns. The data was then combined using an inner join operation based on the country column, which resulted in a single comprehensive dataset containing information on various aspects of mental health. It includes data on the prevalence of mental health disorders such as anxiety, depression, bipolar, and schizophrenia. Additionally, the dataset features economic indicators like GDP per capita and unemployment rates. The behavioral responses to mental health challenges captured in the dataset include seeking professional psychological help, taking prescribed medications, and lifestyle adjustments in response to mental health conditions.

Following the merging, the dataset still contained some missing values. To address this issue effectively while maintaining the integrity of the dataset, a matrix completion technique using Singular Value Decomposition (SVD) was applied. This method was selected for its ability to leverage the existing data structure to accurately estimate missing entries, thus preserving overall data quality. The application of SVD provided highly accurate imputations as helping to maintain the relationships among variables.

### *Random Forest for regression*

The study employs Random Forest regression models to identify the most influential predictors of the prevalence rates of mental health disorders across different countries. The objective is to analyze the impact of various economic and behavioral factors and determine which factors are most predictive of specific mental disorders in each country.

The predictors are economic indicators (GDP per capita and unemployment rates) and behavioral factors (frequency of seeking psychological help, medication adherence, and participation in mental health-enhancing activities). The response variables are continuous and represent the prevalence rates of mental disorders like anxiety, depression, and eating disorders.

For each mental health disorder, a separate Random Forest regression model is trained on the preprocessed dataset by excluding the country column. Hyperparameter tuning is performed to optimize the model's performance by considering the number of trees, maximum depth of the trees, and the minimum number of samples required to split an internal node. The best hyperparameters are selected based on evaluation metrics such as mean squared error and R squared for the optimal model. Then, to identify the most influential predictors of the specific mental health disorder, the feature importances are extracted from the optimized model. A comparison plots are created to visualize the relative importance of each feature, providing insights into which economic and behavioral factors have the greatest impact on the prevalence rates of the disorder being studied.

By using the metrics to evaluate the model, MSE measures the average squared difference between the predicted and actual values, providing an absolute measure of the model's predictive accuracy, with lower values indicating better performance. On the other hand, R-squared represents the proportion of variance in the target variable explained by the model, ranging from 0 to 1, with higher values indicating a better fit. A model with a low MSE and a high R-squared is generally preferred, as it suggests accurate predictions and explains a substantial amount of the variance in the target variable of mental health disorders.

***SVM***

The study uses SVM models to perform classification analysis. The predictors used are economic indicators, specifically GDP and the unemployment rate. The response variables chosen for this analysis are three specific mental health issues: anxiety disorders, depressive disorders, and eating disorders. These variables are initially continuous, representing the percentage of cases relative to a country's population. For the purpose of classification, these variables are transformed into categorical variables: a prevalence above the average is labeled as '1' (high prevalence), while prevalence below the average is labeled as '0' (low prevalence). The dataset is divided into 80% for training, and the remaining 20% is reserved for testing the accuracy of the model.

In the linear SVM model, the parameter C is adjusted across different values (0.01, 0.1, 1, 5, 10). Cross-validation is employed to identify the optimal C value. The model with the best parameters is subsequently applied to the test data to evaluate its accuracy.

For the SVM model utilizing the RBF kernel, the parameter C is varied among several values (0.001, 0.01, 1, 10) and the gamma parameter is adjusted (0.00001, 0.0001, 0.1, 1, 10, 1000). Cross-validation is used to find the optimal combination of C and gamma. The best-parameter model was then used to fit the test data to determine its accuracy.

For the SVM model method with polynomial kernel, the parameter C is explored across a range of values (0.01, 0.1, 1, 5, 10), and the degree of the polynomial is adjusted between 1, 2, and 3. Similarly, cross-validation was used to identify the optimal parameters, and the best-parameter model was applied to the test data to determine its accuracy.

The accuracy of each model is compared to determine which SVM configuration most effectively predicts each of the mental health issues, based on the optimal parameters.

### *Logistic Regression*

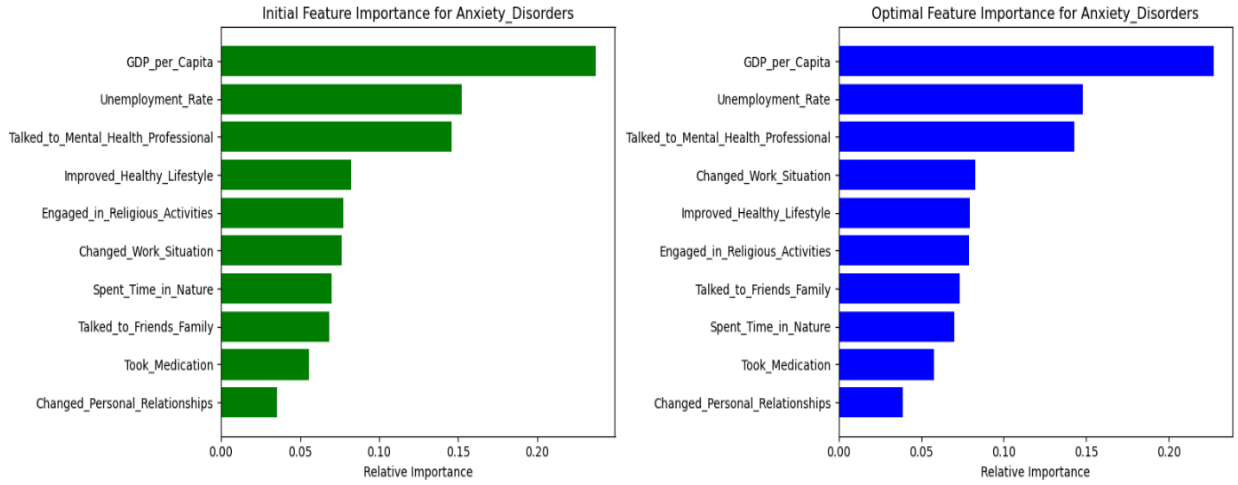
Standardize the dataset, and use logistic regression models to perform classification analysis. The results are then compared with those from the SVM models to determine which method provides better predictive accuracy for the specified mental health issues.

## Computation Results

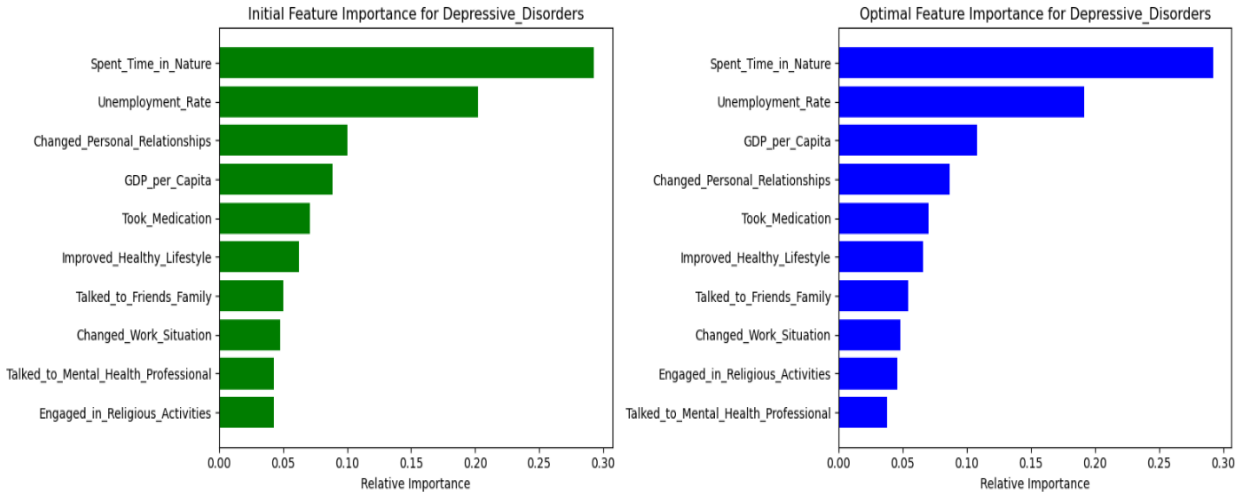
### *Random Forest for regression*

Figure 2 presents the feature importance plots from the Random Forest models for predicting the factors that impact the prevalence of anxiety disorders, eating disorders, and depressive disorders. The plots reveal the relative importance of economic and behavioral factors in influencing these mental health issues. For anxiety disorders, GDP per capita, unemployment rate, and talking to mental health professionals are the most influential predictors. For eating disorders, GDP per capita, talking to mental health professionals, and changed personal relationships are crucial factors. In the case of depressive disorders, spent time in nature, unemployment rate, and GDP per capita are the leading predictors. The initial feature importance plots show the rankings of the predictors before hyperparameter tuning, while the optimal feature importance plots demonstrate the impact of tuning hyperparameters such as the number of trees, maximum depth, and minimum samples split on the relative importance of the predictors, leading to improved model performance and more accurate identification of the key factors impact for each mental health disorder. Notably, economic factors such as GDP per capita and unemployment rate consistently appear as top predictors across all three mental health disorders, highlighting the significant influence of a country's economic conditions on the prevalence of these issues.

```
Best parameters: {'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 100}
```




```
Best parameters: {'max_depth': None, 'min_samples_split': 4, 'n_estimators': 50}
```




```
Best parameters: {'max_depth': 5, 'min_samples_split': 3, 'n_estimators': 50}
```

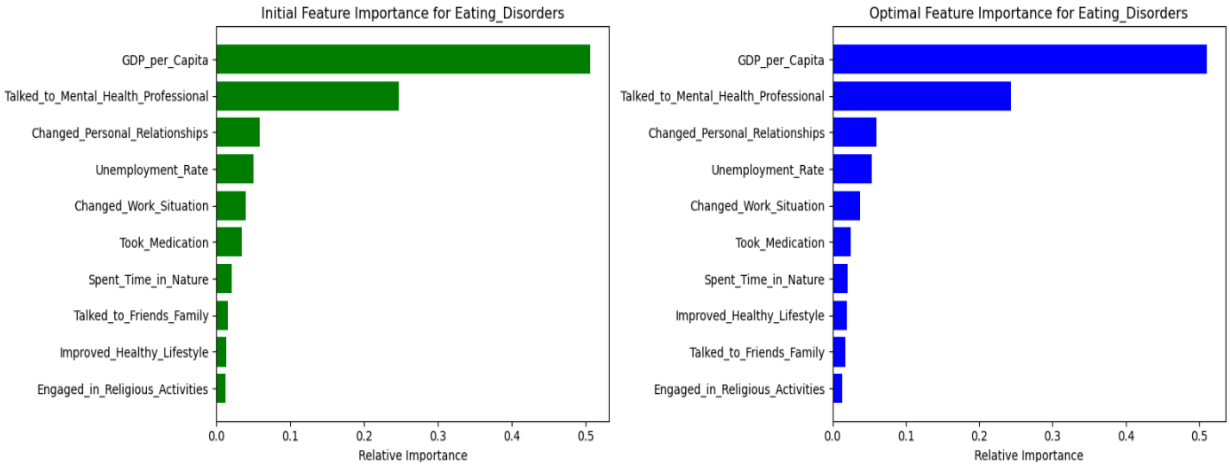



Figure 2: Important Features Impact to Prevalence of Mental Health Issues

Table 1 presents the R-squared values for the Random Forest models in predicting the prevalence of anxiety disorders, depressive disorders, and eating disorders, both before and after hyperparameter tuning. The initial R-squared values range from 0.48 for anxiety disorders to 0.71 for eating disorders, indicating that the models explain a considerable portion of the variance in the prevalence rates. After hyperparameter tuning, the R-squared values improve, with anxiety disorders reaching 0.51, depressive disorders achieving 0.65, and eating disorders attaining 0.78. These higher R-squared values suggest that the optimized models capture an even greater proportion of the variability in the prevalence rates.

| | Anxiety_Disorders | Depressive_Disorders | Eating_Disorders |
|---|---|---|---|
| Initial | 0.48 | 0.61 | 0.71 |
| After Tunning | 0.51 | 0.65 | 0.78 |

Table 1: R-squared for Random Forest

Table 2 shows the corresponding Mean Squared Error (MSE) values for the models. The initial MSE values range from 0.00557 for eating disorders to 1.14807 for anxiety disorders, reflecting the average squared difference between the predicted and actual prevalence rates. After hyperparameter tuning, the MSE values decrease for all three mental health disorders, indicating improved predictive accuracy. The reduction in MSE is particularly notable for anxiety disorders, which drops from 1.14807 to 1.09401. These results demonstrate that hyperparameter tuning enhances the performance of the Random Forest models in predicting the prevalence of mental health disorders across countries.

| | Anxiety_Disorders | Depressive_Disorders | Eating_Disorders |
|---|---|---|---|
| Initial | 1.14807 | 0.23 | 0.00557 |
| After Tunning | 1.09401 | 0.18 | 0.00511 |

Table 2: Mean Squared Error for Random Forest

***SVM***

Figure 3 shows the prevalence rates of three mental health disorders (anxiety disorders, depressive disorders, and eating disorders) in relation to GDP and unemployment rates. The data points for anxiety and depressive disorders do not show a very clear trend, but some observations can still be made. In the graph for anxiety, the green points, which indicate a higher prevalence, are more concentrated in countries with high GDP and low unemployment rates. This may be because a high GDP often correlates with high productivity and competitive work environments, which could lead to increased work stress, a potential factor contributing to anxiety. On the other hand, the graph for eating disorders shows a distinct trend where higher GDP correlates with a higher proportion of eating disorder cases. This could be because economically developed areas usually

have fast-paced work and high pressure. Long hours of work and stress may lead to irregular eating habits and eating disorders.

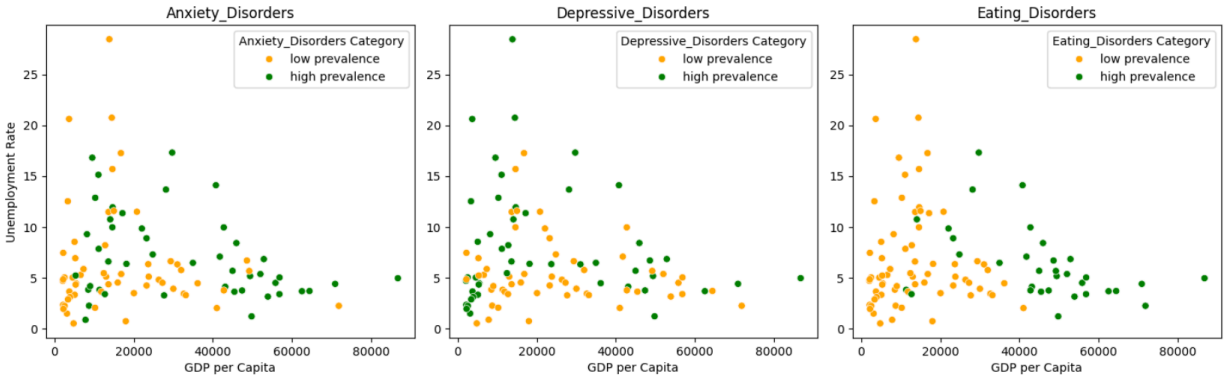



Figure 3 Mental Health Issues Classification by GDP and Unemployment Rate

Table 3 displays the accuracy of different SVM kernels, optimized with the best parameters, for various mental health issues. As the result, for anxiety disorders, both the linear kernel (C value of 0.1) and the polynomial kernel (C value of 0.1 and a degree of 2) emerge as top performers, each achieving the highest accuracy rate of 0.6774. In the case of depressive disorders, the linear kernel (C value of 5) achieves the highest accuracy at 0.5484. For eating disorders, the polynomial kernel (C value of 0.1 and a degree of 2) excels significantly, outperforming others with an exceptionally high accuracy of 0.9032.

| Kernel \ Mental issue | Anxiety_Disorders | Depressive_Disorders | Eating_Disorders |
|---|---|---|---|
| linear | 0.6774 | 0.5484 | 0.8387 |
| RBF | 0.4516 | 0.4594 | 0.6129 |
| polynomial | 0.6774 | 0.4516 | 0.9032 |

Table 3 Testing Accuracy with Optimized SVM Model

***Logistic Regression***

Table 4 presents the accuracy of logistic regression models in predicting three mental health disorders, comparing results between original and standardized data. After standardizing the data, the accuracy for both anxiety disorders and eating disorders improved. In contrast, the accuracy for depressive disorders decreased. This suggests that standardization may have diminished the predictive power of some important features in the case of depressive disorders. Logistic regression model can perform better than SVM model in predicting anxiety disorders and eating disorders.

| | Anxiety_Disorders | Depressive_Disorders | Eating_Disorders |
|---|---|---|---|
| Original data | 0.4286 | 0.6190 | 0.8571 |
| Standardized data | 0.7619 | 0.4762 | 1.0000 |

Table 4 Testing Accuracy with Logistic Model

## Discussion

The Random Forest regression models reveal that economic factors, such as GDP per capita and unemployment rate, and behavioral factors, like talking to mental health professionals and spending time in nature, significantly influence the prevalence of anxiety disorders, depressive disorders, and eating disorders across different countries. Based on the insights gained from the Random Forest analysis, this study focuses on two key economic indicators, GDP per capita and unemployment rate, as the main predictors for the classification models as Support Vector Machines (SVM) and logistic regression, to further investigate the classification of mental health prevalence across countries. It provides valuable insights for policymakers, healthcare providers, and researchers to develop targeted interventions and strategies for improving mental health outcomes globally.

Comparing the logistic regression model with the SVM model, the logistic regression model performs better in predicting mental health issues. This might be due to the presence of outliers in the dataset. SVMs are often more sensitive to outliers than logistic regression. SVMs strive to maximize the margin between classes and can be heavily influenced by points that are difficult to classify. In contrast, logistic regression minimizes logistic loss, which can be less influenced by outliers.

The analysis shows that using GDP and unemployment rate as economic indicators can effectively predict anxiety and eating disorders, with accuracies of 0.76 and 1, respectively. Notably, the testing accuracy of the logistic model to predict eating disorders reaches as high as 1, indicating that the model is highly successful. However, given that the dataset is small, the model might only be fitting these specific samples well. Thus, a larger dataset is needed to increase the credibility of the model. For depressive disorders, the prediction in this study shows a low correlation with GDP and unemployment rates. Future studies could benefit from increasing the data samples and including more economic-related analytical factors, such as inflation rate and interest rates, to better determine whether a country's economic condition can predict mental health.

The prediction of eating disorders shows strong performance across all models, indicating a strong relationship between GDP, unemployment rates, and eating disorders. Therefore, these models could be used to predict regions with high incidences of eating disorders and to enhance public awareness and prevention education in these areas. Providing more specialized medical resources and psychological support may help reduce the incidence of these disorders.

## Conclusion

In conclusion, this study explores the relationship between various mental health disorders and socioeconomic factors using machine learning models, including Random Forest, Support Vector Machines (SVM), and logistic regression. The Random Forest models reveal that economic factors, such as GDP per capita and unemployment rate, along with behavioral factors, like talking to mental health professionals and spending time in nature, significantly influence the prevalence of anxiety disorders, depressive disorders, and eating disorders across different countries. The optimized Random Forest models capture a substantial portion of the variability in the prevalence rates, with R-squared values ranging from 0.51 for anxiety disorders to 0.78 for eating disorders. The SVM models, particularly with linear and polynomial kernels, demonstrate high accuracy in predicting anxiety disorders (67.74%) and eating disorders (90.32%), while the logistic regression models outperform SVM in predicting anxiety and eating disorders after standardizing the data.

The findings of this study provide valuable insights into exploring the relationship between socioeconomic factors and mental health disorders on a global scale. The consistent influence of economic indicators, such as GDP per capita and unemployment rate, across all three mental health issues highlights the importance of considering a country's economic conditions when addressing mental health challenges. The strong performance of the models in predicting eating disorders suggests that these models could be used to identify regions with high incidences of eating disorders and inform targeted interventions, such as public awareness campaigns and the allocation of specialized medical resources. Future research should focus on expanding the dataset, incorporating additional economic and behavioral factors, and exploring the specific mechanisms through which these factors impact mental health outcomes. By understanding these relationships, policymakers, healthcare providers, and researchers can develop more effective strategies to promote mental well-being and address the global burden of mental health disorders.

## References


1. Vitalflux. (n.d.). Classification model- SVM classifier python example. Vitalflux. https://vitalflux.com/classification-model-svm-classifier-python-example/
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Retrieved from https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html
3. Saloni Dattani, Lucas Rodés-Guirao, Hannah Ritchie and Max Roser (2023) - “Mental Health” Published online at OurWorldInData.org. Retrieved from https://ourworldindata.org/mental-health