**Exploring and Modeling**

**on Mental Health Data**

**: Insights from Kaggle Data Analysis**

Prepared for: Prof. Divya Pandove

(TA: Bhumika Mallikarjunhorapet)

Prepared by: Team 7

Haeyeon Jeong, Sai Rachana Kandikattu

Abirham Getie, Yonathan Shimelis

Dec 17, 2024

Table of Contents

Introduction1

Research Methods and Results2

*Task 1. Data Acquisition*2

*Task 2. Data Preprocessing*4

*Task 3. Exploratory Data Analysis (EDA) and Statistical Testing*6

*Part 1) SMART Question 1*7

*Part 2) SMART Question 29*

*Part 3) SMART Question 3*10

*Task 4. Modeling and Evaluations* 11

*Part 1) SMART Question 1*11

*Part 2) SMART Question 2*11

*Part 3) SMART Question 3*25

Conclusions and Recommendations29

References 30

Introduction

According to Johns Hopkins (2024), 1 in 4 American adults suffers from a diagnosable mental disorder each year. A statistic that proves mental health challenges are becoming more visible across different sectors, affecting individuals in diverse ways. We chose this topic to try and use our data science skills for good. The first step to developing effective support systems is to understand the factors that contribute to mental stress. This research will explore mental health data using various classification models to identify key factors influencing mental health outcomes. Specifically, we will examine how variables such as mood swings, family history, treatment status, occupations, and time spent indoors affect mental health. By analyzing these factors through data models, we hope to uncover patterns that can help improve mental health management and prevention strategies.

This study seeks to address the following key question:

***Initial Research Question:***

How can we develop models to assess mental stress, and which factors are most influential in predicting mental health outcomes?

***Revised Research Questions:***

Based on the insights gathered from Exploratory Data Analysis (EDA) and subsequent statistical testing, we refined and expanded our research focus to address the following additional questions:

**Question 1:** What are the top 5 factors that influence the growing stress?

**Question 2:** Do people with family history receive treatment or not?

**Question 3:** What are the factors that impact the growing stress for students?

Our research includes four main tasks, including data acquisition, data preprocessing, Exploratory Data Analysis (EDA) and statistical testing, and modeling and evaluations.

**Research Methods and Results**

In this section, we will introduce the research methods and results of three tasks of our research.

***Task 1. Data Acquisition***

The dataset used in this study is available on Kaggle at this ([link](https://www.kaggle.com/datasets/bhavikjikadara/mental-health-dataset)). It contains 292,364 observations with 17 variables (**Table 1**). The data’s original source is Ourworldindata.org. The method of collection is via mental health survey. The variables include text-based categorical responses from the survey.

***Table 1*** *Data Dictionary of Mental Health Dataset*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Values/Options** |
| **Timestamp** | The date and time each entry was recorded. | ['8/27/2014 11:37', '8/27/2014 11:49', ..., '2/1/2016 23:04'] |
| **Gender** | representing the gender of the individual. | ['Female', 'Male'] |
| **Country** | representing the country of the individual. | ['United States', 'South Africa', ..., 'Philippines'] |
| **Occupation** | the occupation or employment status of the individual. | ['Corporate', 'Student', 'Business', 'Housewife', 'Others'] |
| **Business** | Describes if the individual owns a business. | 'Business owned by an individual' |
| **Corporate** | Describes if the individual works in a corporate environment. | 'An individual who works for corporate' |
| **Self-Employed** | indicating whether the individual is self-employed. | ['No', 'Yes'] |
| **Family History** | indicating whether the individual has a family history of mental health issues. | ['Yes', 'No'] |
| **Treatment** | indicating whether the individual is currently receiving mental health treatment. | ['Yes', 'No'] |
| **Days Indoors** | describing how often the individual stays indoors. | ['1-14 days', 'Go out every day', 'More than 2 months', '15-30 days', '31-60 days'] |
| **Growing Stress** | indicating whether the individual is experiencing growing stress. | ['Yes', 'No', 'Maybe'] |
| **Changes in Habits** | indicating whether the individual has noticed any changes in their habits. | ['No', 'Yes', 'Maybe'] |
| **Mental Health History** | Categorical data indicating whether the individual has a history of mental health issues. | ['Yes', 'No', 'Maybe'] |
| **Mood Swings** | describing the severity of the individual's mood swings. | ['Medium', 'Low', 'High'] |
| **Coping Struggles** | indicating whether the individual is struggling to cope with stress or mental health challenges. | ['No', 'Yes'] |
| **Work Interest** | indicating whether the individual has an interest in work or is actively seeking work. | ['No', 'Maybe', 'Yes'] |
| **Social Weakness** | indicating whether the individual experiences social weakness or difficulties. | ['Yes', 'No', 'Maybe'] |
| **Mental Health Interview** | indicating whether the individual has had a mental health interview. | ['Maybe', 'No', 'Yes'] |
| **Care Options** | indicating whether the individual is considering or has access to mental health care options. | ['Yes', 'Not sure', 'No'] |

***Task 2. Data Preprocessing***

During the data preprocessing phase, several steps were undertaken to ensure the quality and suitability of the dataset for analysis:

Step 1. Handling Missing Values:  
We identified 5,202 missing values in the 'self\_employed' column. Given the significance of this variable, we opted to drop the rows with missing values to maintain data integrity and ensure the accuracy of our analysis.

Step 2. Removing Irrelevant Columns:

Following an initial round of Exploratory Data Analysis (EDA) and statistical testing, we determined that certain columns, such as 'Timestamp' and 'Country', were not directly relevant to the research objectives. As these variables did not contribute meaningfully to the predictive modeling, they were excluded from the dataset.

Step 3. Numerical Encoding:

For categorical variables with binary or ordinal responses, such as 'No', 'Yes', and 'Maybe', we applied numerical encoding. Specifically, the values were converted as follows: 'No' became 0, 'Yes' became 1, and 'Maybe' was assigned a value of 2. This transformation facilitated the use of these variables in machine learning models.

Step 4: One-Hot Encoding for Categorical Variables:

The 'Occupation' variable was encoded using One-Hot Encoding, creating a separate binary column for each category (e.g., 'Corporate', 'Student', 'Business', etc.). This technique ensured that each category was treated independently without implying any ordinal relationship, which is important for accurate modeling.

Step5: Handling Ambiguous Responses:

For the growing stress variable, we excluded the 'Maybe' category. The presence of 'Maybe' created ambiguity, as it did not provide a clear indication of whether an individual was experiencing growing stress. To maintain the clarity of the target variable, only definitive responses ('Yes' and 'No') were retained.

These preprocessing steps were crucial in preparing the dataset for subsequent modeling and analysis. Further details of the exploratory analysis and statistical tests will be discussed in the following sections.

***Task 3. Exploratory Data Analysis (EDA) and Statistical Testing***

We proceeded with Exploratory Data Analysis (EDA) to uncover key trends, relationships, and distributions within the dataset. Initially, we examined the general distribution of responses across the dataset, providing an overview of how the variables were distributed. As can be seen in **Figure 1**, this initial analysis revealed that the distribution of individuals across most feature categories was fairly balanced, with the exception of variables related to ‘self-employment’, ‘mental health interviews’, and ‘country’. These features exhibited significant imbalances, which warranted further investigation.

After understanding the general distribution of the variables, we moved on to more detailed EDA and statistical testing for each of the research questions. This allowed us to explore specific patterns, relationships, and potential correlations within the data, providing deeper insights into the factors influencing mental stress and mental health outcomes.

***Figure 1*** *Distribution of Each Variables using Pie Chart*

A group of pie charts

Description automatically generated***Part 1) SMART Question 1 (Top Five Factors Contributing Growing Stress)***

The **figure 2** illustrates the overall growing stress distribution across various variables within the dataset. The occupations associated with the highest levels of growing stress include Business, followed closely by Students and Corporate roles. Additionally, the data indicates that spending extended periods indoors, changes in habits, work interest, mood swings, mental health history can have a significant impact on mental health.

***Figure 2*** *Overall Growing Stress Distribution*

***A graph of stress distribution

Description automatically generated A graph of stress distribution

Description automatically generated***

***A graph of a stress distribution

Description automatically generated*** ***A graph of stress distribution

Description automatically generated***

***A graph showing stress distribution

Description automatically generated*** ***A graph of a graph showing a number of different colored squares

Description automatically generated with medium confidence***

We applied statistical tests to examine the independence of various features with respect to the growing\_stress variable. The hypotheses for each feature were as follows:

**Null Hypothesis (H₀):** The feature and Growing\_Stress are independent.

**Alternative Hypothesis (H₁)**: The feature and Growing\_Stress are not independent.

The results of the p-values for each of the three features are summarized in **Table 2**.

***Table 2*** *Statistical Testing Results for Features with Respect to Growing Stress*

|  |  |  |
| --- | --- | --- |
| **Feature** | **p-value** | **Conclusion** |
| **Self\_Employed** | 0.15 | Fail to reject H₀ (independent) |
| **Timestamp** | 0.93 | Fail to reject H₀ (independent) |
| **Country** | 0.99 | Fail to reject H₀ (independent) |

As the p-values for all three features (self\_employed, timestamp, country) were greater than the significance level of 0.05, we failed to reject the null hypothesis, indicating that these features were independent of growing stress and could be excluded from the model.

***Part 2) SMART Question 2 (Family History vs Treatment)***

***Figure 3*** *Correlation Analysis***A diagram of a number of numbers

Description automatically generated**

As can be seen in the **Figure 3**, the correlation analysis reveals that family history and treatment exhibit a moderate positive correlation of 0.37. This suggests a noteworthy relationship between these two variables, indicating that individuals with a family history of mental health issues are somewhat more likely to receive treatment. In contrast, other variables do not show strong linear relationships with growing stress.

To examine the relationship between family history and treatment, statistical testing was performed:

**Null Hypothesis (H₀):** Family history and treatment are independent.

**Alternative Hypothesis (H₁)**: Family history and treatment are not independent.

The p-value was found to be 0.00, which is less than the significance level of 0.05. As a result, the null hypothesis is rejected, confirming that there is a significant relationship between family history and treatment. This means that individuals with a family history of mental health issues are significantly more likely to seek treatment.

***Part 3) SMART Question 3 (Family History vs Treatment)***

In the **Figure 4**, students had the second-highest proportion of survey responses that indicated experiencing feelings of growing stress in our dataset. This finding was significant, especially since the research was being conducted by students. We found it an interesting area to explore, both from a personal and analytical perspective.

To analyze the relationship between Occupation and Growing Stress, we formulated the following hypotheses.

***Figure 4*** *Growing Stress Distribution by Occupation*

**A graph of stress distribution

Description automatically generatedNull Hypothesis (H₀):** Occupation and Growing Stress are independent.

**Alternative Hypothesis (H₁)**: Occupation and Growing Stress are not independent.

The p-value obtained from the statistical test was 0.00. Since the p-value is less than the significance level of 0.05, we reject the null hypothesis, indicating that Occupation and Growing Stress are not independent. This suggests a significant relationship between a person's occupation and the likelihood of experiencing growing stress.

***Task 4. Modeling and Evaluations***

***Part 1) SMART Question 1 (Top Five Factors Contributing Growing Stress)***

For SMART Question 1, we aimed to identify the top five factors contributing to the escalation of growing stress in individuals. To achieve this, we employed three different models: Logistic Regression, Decision Tree, and Random Forest. These models were chosen based on their ability to handle classification tasks and capture the relationships within the data. Below, we describe each model, the parameters used, and their evaluation results.

Model 1: Logistic Regression

Logistic Regression is a fundamental classification algorithm commonly used for binary outcomes. It was selected for its interpretability and because it serves as a benchmark for comparison against more complex models. Although the correlation matrix suggested weak linear relationships between the predictors, Logistic Regression was used to assess its performance on this dataset.

Modeling Details:

* Response Variable (Y): Growing stress
* Predictors (X): All 16 predictors

The Logistic Regression model performed poorly, with an accuracy of 54%. It also exhibited high misclassification rates, as summarized in the confusion matrix below (**Figure 5**).

The model displayed high misclassification rates, as evidenced by 9,866 instances where it incorrectly predicted the target class (1) for actual class 0, resulting in false positives. Additionally, there were 7,486 instances where the model predicted class 0 for actual class 1, leading to false negatives. These misclassifications indicate poor recall and precision in the model's performance.

***Figure 5*** *Logistic Regression Model Confusion Matrix of SMART Question 1*

A blue squares with black text

Description automatically generated

The high number of false negatives (7,486) suggests that the model struggles to correctly identify actual positive cases of growing stress, while the large number of false positives (9,866) shows that the model frequently misidentifies individuals as having growing stress when they do not.

Moreover, the overall performance was imbalanced, as the model misclassified a significant number of instances from both classes (0 and 1). This misclassification led to unreliable predictions, highlighting the model's difficulties in distinguishing between the two classes, ultimately affecting its effectiveness in predicting the escalation of stress levels.

As can be seen in the **Figure 6**, the AUC was 0.58, indicating that the model had a limited ability to distinguish between individuals with and without growing stress.

***Figure 6*** *ROC curve of Logistic Regression Model of SMART Question 1* *A graph of a curve

Description automatically generated*

Model 2: Decision Tree

The Decision Tree model was chosen due to the poor performance of Logistic Regression. Decision Trees are well-suited for handling categorical data and can capture non-linear relationships in the dataset. This makes them effective for datasets with mixed data types.

Modeling Details:

* Response Variable (Y): Growing stress
* Predictors (X): All 16 predictors
* Parameter Tuning **(Table 3):** The parameters selected for tuning include max\_depth, min\_samples\_split, min\_samples\_leaf, and max\_features, which were determined based on training and testing accuracy using 10-fold cross-validation.

***Table 3*** *Parameter Tuning for Decision Tree Model of SMART Question 1*

|  |  |
| --- | --- |
| Max depth | 9 |
| Min samples split | 10 |
| Min sample leaf | 10 |
| Max features | sqrt |

After utilizing the selected features, the following accuracies were obtained through cross-validation (**Table 4**), training, and testing. But as observed, the testing accuracy seems to be a little lower than training accuracy hence, there can be some improvement.

***Table 4***

*Decision Tree: Cross-Validation Accuracies and Model Performance of SMART Question1*

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **Cross-Validation Accuracies for Each Fold** | [0.846, 0.853, 0.841, 0.831, 0.835, 0.827, 0.840, 0.832, 0.837, 0.851] |
| **Mean Cross-Validation Accuracy** | 0.8391 |
| **Standard Deviation of Accuracy** | 0.0081 |
| **Final Training Accuracy** | 0.8372 |
| **Final Test Accuracy** | 0.8347 |

The model demonstrated strong classification rates (**Figure 7**), correctly predicting 15,913 instances for class 1 and 15,628 instances for class 0. This indicates an effective overall performance, with the model successfully identifying both classes.

***Figure 7*** *Decision Tree Model Confusion Matrix for SMART Question 1*

**A blue squares with black text

Description automatically generated**

Additionally, the model exhibited good recall, accurately identifying a significant number of true positives. It also maintained high precision, with many correct predictions for class 0. The performance was balanced, as the model effectively predicted both classes without a noticeable bias toward either. However, there is room for improvement in terms of further tuning the model to enhance its accuracy and better manage any misclassifications.

The model performed well with an accuracy of 84% and an AUC score of 0.94 (**Figure 8**), indicating that the model effectively distinguishes between individuals with escalating stress levels and those without.

***Figure 8*** *ROC curve of Decision Tree Model for SMART Question 1*

A graph showing a curve

Description automatically generated with medium confidence

Model 3: Random Forest Tree

The Random Forest algorithm was selected to mitigate overfitting observed in the Decision Tree model. By combining multiple decision trees, Random Forest improves generalization and reduces the risk of overfitting. It also provides a more robust prediction across various subsets of the data.

Modeling Details:

* Response Variable (Y): Growing stress
* Predictors (X): All 16 predictors
* Parameter Tuning (**Table 5**): The parameters selected for tuning include max\_depth, min\_samples\_split, min\_samples\_leaf, and max\_features, which were determined based on training and testing accuracy using 10-fold cross-validation.

***Table 5*** *Parameter Tuning for Random Forest Tree Model of SMART Question 1*

|  |  |
| --- | --- |
| N\_estimators | 10 |
| Max depth | 7 |
| Min samples split | 20 |
| Min sample leaf | 20 |
| Max features | sqrt |

After utilizing the selected features, the following accuracies were obtained through cross-validation (**Table 6**), training, and testing. But as observed, now the testing accuracy seems to improve and is greater than training accuracy.

***Table 6***

*Random Forest: Cross-Validation Accuracies and Model Performance of SMART Question1*

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **Cross-Validation Accuracies for Each Fold** | 0.8916, 0.8842, 0.8984, 0.9041, 0.8932, 0.9039, 0.8886, 0.9000, 0.8973, 0.8976 |
| **Mean Cross-Validation Accuracy** | 0.8959 |
| **Standard Deviation of Accuracy** | 0.0061 |
| **Final Training Accuracy** | 0.9112 |
| **Final Test Accuracy** | 0.9124 |

As can be seen in the **Figure 9**, the model demonstrated strong classification rates, with a total of 18,379 correct predictions for class 1 and 16,097 correct predictions for class 0, indicating effective overall performance. The model also showed good recall, accurately identifying a significant number of true positives. Additionally, it maintained high precision, with many correct predictions for class 0. The performance was balanced, as the model effectively predicted both classes, and it performed much better than the decision tree model.

***Figure 9*** *Random Forest Tree Model Confusion Matrix for SMART Question 1* A blue squares with black text

Description automatically generated

The model performed well with an accuracy of 91% and an AUC score of 0.97 (**Figure 10**), indicating that the model very effectively distinguishes between individuals with escalating stress levels and those without**.**

***Figure 10*** *ROC curve of Random Forest Tree Model for SMART Question 1*

A graph with a red line

Description automatically generated

Using the importance function from the Random Forest model, we assessed how various variables influence the escalation of growing stress levels, as this model performed the best among all others. The analysis revealed the top five factors contributing to growing stress.

As shown in the **Figure 11**, Mental Health History has the highest importance score, suggesting that individuals with a history of mental health issues (or uncertainty about their mental health) are more likely to experience escalating stress.

Days Indoors and Work Interest are also critical factors, with more days spent indoors and changes in work interest leading to higher stress levels. Mood Swings and Changes in Habits also play significant roles, indicating that emotional fluctuations and changes in behavior contribute to growing stress.

***Figure 11*** *Top 5 Feature Importance using Random Forest for SMART Question 1*

**A graph of blue rectangular bars

Description automatically generated with medium confidence**

As observed in the **Table 7**, the Random Forest model was the most effective in identifying key factors contributing to the escalation of growing stress in individuals. It outperformed both Logistic Regression and Decision Tree in terms of accuracy, AUC, and overall classification performance. The top five factors influencing stress escalation include Mental Health History, Days Indoors, Work Interest, Mood Swings, and Changes in Habits.

***Table 7*** *Comparison of Model Results for SMART Question 1*

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Logistic Regression** | **Decision Tree** | **Random Forest** |
| **Accuracy** | 54% | 83.50% | 91% |
| **AUC Score** | 0.58 | 0.94 | 0.97 |

These findings provide valuable insights into the primary drivers of stress escalation and suggest potential areas for intervention to mitigate stress in individuals.

***Part 2) SMART Question 2 (Family History vs Treatment)***

To further explore the relationship between family history and treatment, two classification models—Logistic Regression and Random Forest—were used.

Model 1: Logistic Regression

Modeling Details:

* Response Variable (Y): Treatment
* Predictors (X): 12 predictors

The Logistic Regression model achieved an accuracy of 68% and AUC score of 0.72 (**Figure 12**).

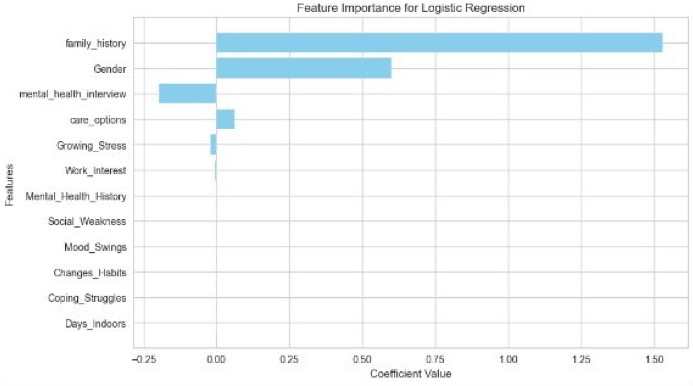
***Figure 12*** *ROC curve of Logistic Regression Model for SMART Question 2*

A graph showing the value of a receiver

Description automatically generated with medium confidence

As shown in the **Figure 13**, The top features identified in the Logistic Regression model were Family History, Gender, and Mental Health Interview. These variables had the most significant impact on predicting whether individuals with a family history of mental health issues would seek treatment.

***Figure 13*** *Feature Importance using Logistic Regression for SMART Question 2*



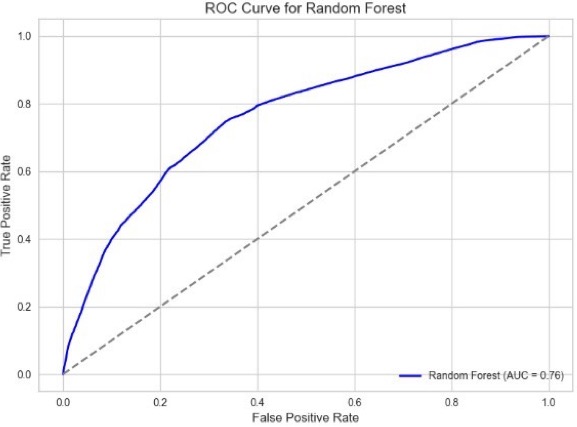
Model 2: Random Forest

Modeling Details:

* Response Variable (Y): Treatment
* Predictors (X): 12 predictors

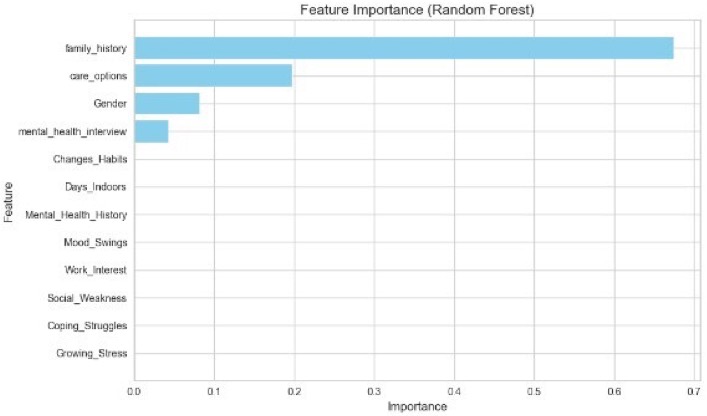
The Random Forest model outperformed Logistic Regression with a higher accuracy of 70% and a better AUC score of 0.76 (**Figure 14**). The higher accuracy and improved AUC score demonstrate the model's superior ability to classify treatment outcomes and distinguish between individuals who are likely to seek treatment and those who are not.

***Figure 14*** *ROC curve of Random Forest Model for SMART Question 2*



As shown in the **Figure 15**, the top features identified in the Random Forest model were Family History, Care Options, Gender, and Mental Health Interview. Similarly, the Logistic Regression model also identified Family History, Gender, and Care Options as key predictors, with Mental Health Interview ranked fourth in both models. This shows strong agreement between the two models on the importance of Family History, Gender, and Care Options.

***Figure 15*** *Feature Importance using Random Forest for SMART Question 2*



The analysis confirms a strong relationship between family history and treatment. Individuals with a family history of mental health issues are more likely to seek treatment. Additionally, care options and gender also play significant roles in determining whether treatment is sought. The Random Forest model performed better than the Logistic Regression model, providing more accurate predictions and better model performance metrics (higher accuracy and AUC). This underscores the importance of family history in predicting treatment outcomes, alongside other influencing factors such as care options and gender.

***Part 3) SMART Question 3 (Family History vs Treatment)***

For this analysis, we used Logistic Regression to investigate the factors that impact growing stress among students. Logistic regression was chosen due to its applicability in predicting binary outcomes, which in this case was whether a student was experiencing growing stress or not.

Modeling Details:

* Response Variable (Y): Growing stress
* Predictors (X): 13 predictors

The logistic regression model performed poorly with an accuracy of just 61%. The model's effectiveness in predicting stress levels among students was limited. The Confusion Matrix showed that the model misclassified a significant number of instances, leading to high rates of false positives (**Figure 16**).

***Figure 16*** *Confusion Matrix for SMART Question 3*

*A blue squares with black text

Description automatically generated*

The model showed high misclassification rates, predicting growing stress (1) in 2,063 instances where there was no actual stress (false positives). This resulted in poor recall and precision, as the model struggled to correctly identify true positives. Additionally, the model demonstrated imbalanced performance, misclassifying many class 0 (no stress) instances, and was overly quick to predict growing stress even when not reported.

***Figure 17*** *ROC curve for SMART Question 3*

A graph with a line going up

Description automatically generated

The ROC curve for this model indicated that it did not perform well, with an accuracy of only 61% and an AUC score of 0.62 (**Figure 17**).

The model’s shortcomings can be proof of the lack of predictive power our current variables give us just for students. It could be possible that if the data included variables more tailored for students, we’d have a logistic regression model with stronger predictive power. Variables like academic performance or financial strife could provide key insight in predicting whether or not a student is experiencing growing stress. The model did, however, leave us with more revealing coefficients.

***Figure 18*** *Feature Importance using Coefficient evaluation for SMART Question 3*

A graph with red squares and green squares

Description automatically generated

Upon examining the coefficients of the logistic regression model, we identified several key factors that influence growing stress among students.

Changes in habits had the largest effect, with students reporting changes in habits less likely to experience growing stress. Days indoors was positively correlated with increased stress, suggesting isolation contributes to stress. Gender showed that men were more likely to report stress than women. Mood swings also impacted stress, though to a lesser extent. Lastly, social weakness had a minor influence on stress levels.

The logistic regression model for predicting growing stress among students showed limited performance, primarily due to the lack of student-specific data. Despite this, the model provided valuable insights, such as the significant influence of habit changes, time spent indoors, and gender on growing stress levels. These findings suggest that tailored features, such as academic performance or financial strain, may improve future models targeting student stress.

**Conclusions and Recommendations**

**Conclusion**

In conclusion, the key factors contributing to growing stress across all occupations are Mental Health History, Days Indoors, Work Interest, Mood Swings, and Changes in Habits. For students specifically, the primary stressors identified are Changes in Habits, Days Indoors, Gender, Mood Swings, and Social Weakness. These findings suggest that both environmental factors (such as time spent indoors) and personal factors (like changes in habits and mood swings) significantly contribute to stress levels.

Furthermore, individuals with a family history of mental health issues appear more likely to recognize their symptoms or seek treatment, which underscores the importance of early identification and intervention. To mitigate growing stress, it is crucial across all sectors to encourage spending more time outdoors, to acknowledge and address stress triggered by changes in habits or mood swings, and to seek treatment when symptoms arise. The negative relationship between growing stress and treatment highlights the importance of addressing stress proactively, emphasizing the need for early mental health support and intervention.

**Recommendations**

1. Stratified Sampling for Employment Groups: To ensure balanced representation, stratified sampling should be implemented to include an equal number of self-employed and non-self-employed individuals in the study. This will enable more accurate comparisons and reduce any biases in analyzing the impact of employment status on stress and mental health outcomes.

2. Inclusion of Self-Employed Individuals: Including self-employed individuals in the model could provide valuable insights into the relationship between employment type, treatment outcomes, and pre-existing mental health conditions. This will enhance the understanding of how mental health dynamics differ across various employment groups, contributing to a more comprehensive understanding of the factors that influence stress and mental well-being.

**References**

Johns Hopkins Medicine. (2024). *Mental health disorder statistics*. John Hopkins Medicine. https://www.hopkinsmedicine.org/health/wellness-and-prevention/mental-health-disorder-statistics

Our World in Data. (2024). *Our World in Data*. Our World in Data. https://ourworldindata.org/

Git Repository: [**github**](https://github.com/sairachanak/6103-Team7.git)