**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

Introduction1

Research Methods and Results2

*Task 1. Data Cleaning and Dictionary2*

*Task 2. Data Visualisation2*

*Task 2. Data Preprocessing* 5

*Task 3. SMART Question 1* 5

*Task 3.1 EDA*

*Task 3.2 Statistical Testing*

*Task 3.3 Modelling and Evaluation*

*Task 3.4 Feature Importance*

*Task 4. SMART Question 1* 5

*Task 4.1 EDA*

*Task 4.2 Statistical Testing*

*Task 4.3 Modelling and Evaluation*

*Task 4.4 Feature Importance*

*Task 5. SMART Question 1* 5

*Task 5.1 EDA*

*Task 5.2 Statistical Testing*

*Task 5.3 Modelling and Evaluation*

*Task 5.4 Feature Importance*

Conclusions12

Recommendations12

References

Introduction

According to Johns Hopkins, 1 in 4 American adults suffers from a diagnosable mental disorder in a given 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?

***Additional Research Questions:***

Question 1: What are the top five factors contributing to the escalation of stress levels in individuals?

Question 2: Does a family history of mental health issues influence whether individuals seek treatment for mental stress?

Question 3: What specific factors impact the growing stress for students?

**Research Methods**

To achieve these objectives, we will perform the following tasks:

***Task 1. Data Cleaning and Data Dictionary***

A screenshot of a computer

Description automatically generated

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. The data’s original source is Ourworldindata.org. The method of collection is via survey.

**Data Dictionary:**

Timestamp: ['8/27/2014 11:37' ,'8/27/2014 11:43' ,'8/27/2014 11:49' , ..., '2/1/2016 23:04']

Gender: ['Female', 'Male']

Country: ['United States' ,'Poland' ,'Australia' ,'Canada' ,'United Kingdom' ',South Africa' , ..., 'Philippines']

Occupation: ['Corporate' ,'Student' ,'Business' ,'Housewife' ,'Others']

Business: Business owned by an individual

Corporate: An individual who works for corporate

self\_employed: ['No' ,'Yes']

family\_history: ['Yes' ,'No']

treatment: ['Yes' ,'No']

Days\_Indoors: ['1-14 days' ,'Go out Every day' ,'More than 2 months' ,'15-30 days','31-60 days']

Growing\_Stress: ['Yes' ,'No' ,'Maybe']

Changes\_Habits: ['No' ,'Yes' ,'Maybe']

Mental\_Health\_History: ['Yes', 'No' ,'Maybe']

Mood\_Swings: ['Medium' ,'Low' ,'High']

Coping\_Struggles: ['No' ,'Yes']

Work\_Interest: ['No' ,'Maybe' ,'Yes']

Social\_Weakness: ['Yes','No' ,'Maybe']

mental\_health\_interview: ['Maybe', 'No' ,'Yes']

care\_options: ['Yes', 'Not sure' ,'No']

***Task 2. Data Visualization***

We will perform exploratory data analysis (EDA) to uncover key trends, relationships, and distributions within the dataset. The analysis reveals that the distribution of individuals across most feature categories is fairly even, with the exception of those related to self-employment, mental health interviews, and country of origin.A group of pie charts

Description automatically generated

***Task 3: SMART Question 1***

**What are the top five factors contributing to the escalation of stress levels in individuals?**

***Task 3.1: EDA***

***Significant observations with growing stress in all occupations:***

***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***

The figure illustrates that 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.

***Task 3.2: Statistical Testing***

**H₀: {The feature} and Growing\_Stress are independent.**

**H₁: {The feature} and Growing\_Stress are not independent.**

**Self\_Employed:** p-value = 0.15, **Timestamp:** p-value = 0.93, **Country:** p-value = 0.99

The p-value > 0.05 for all the 3 variables suggest that they are independent of growing stress, hence removing these variables for the modelling.

***Task 3.3: Modeling***

Various classification models will be employed, including Logistic Regression, Decision Trees, and Random Forest, to assess the impact of identified factors on mental health outcomes.

**Question 1: What are the top five factors contributing to the escalation of stress levels in individuals?**

**Model 1: Logistic Regression**

**Reason for Selection:** Logistic Regression was chosen as it is a fundamental classification algorithm. Despite the correlation matrix indicating a lack of clear linear relationships between variables, the model was used to validate its performance.

**Modeling Details:**

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

***Task 3.3: Evaluation***

The model performed poorly, achieving an accuracy of only 54%.

**Confusion matrix:**

A screenshot of a color chart

Description automatically generated

**High misclassification rates:**

* **9866 instances** where the model predicted **1** for actual **0** (false positives).
* **7486 instances** where the model predicted **0** for actual **1** (false negatives).

**Poor recall and precision:**

* High **false negatives** (7486) indicate the model struggles to correctly identify class 1 (actual positives).
* High **false positives** (9866) suggest the model frequently predicts class 1 incorrectly.

**Imbalanced performance:**

* The model misclassifies a large number of actual class 0 and class 1 instances, leading to unreliable predictions.

**ROC/AUC:**

A graph of a curve

Description automatically generated

The model performed poorly with an accuracy of 54% and an AUC score of 0.58, indicating that the model has limited ability to distinguish between individuals with escalating stress levels and those without.

**Model 2: Decision Tree**

**Reason for Selection:**

The decision tree was chosen due to the underperformance of the logistic regression model. Tree-based models excel with categorical variables and given the presence of numerous categorical predictors in this case, the decision tree was deemed the most suitable option**.**

**Modeling Details:**

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

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

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

**Cross-Validation Accuracies for Each Fold:** [0.84604697, 0.8528614, 0.8406219, 0.83122726,0.83493219, 0.82712537, 0.83969567, 0.83162421, 0.83652001, 0.85080058]

**Mean Cross-Validation Accuracy**: 0.8391455560739811

**Standard Deviation of Cross-Validation Accuracy:** 0.008139265793375025

**Final Training Accuracy**: 0.8372466903519045

**Final Test Accuracy:** 0.8346829681380332

**Confusion matrix:**

A blue and white grid with numbers and labels

Description automatically generated

**High Classification Rates:**

* The model demonstrated strong classification rates with a total of 15,913 correct predictions for class 1 and 15,628 correct predictions for class 0, indicating effective overall performance.

**Good Recall and Precision:**

* The model showed good recall by accurately identifying a significant number of true positives, and it also maintained high precision with many correct predictions for class 0.

**Balanced Performance:**

* The performance reflects a balanced approach, with the model effectively managing to predict both classes, although further tuning could enhance its accuracy in handling misclassifications.

**ROC/AUC:**

A graph with a red line

Description automatically generated

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

**Model 2: Random Forest Tree**

**Reason for Selection:** The Random Forest algorithm was selected due to the slight overfitting observed in the decision tree model. Random Forests help mitigate this issue by combining multiple decision trees, which enhances generalization and reduces the risk of overfitting. This ensemble approach leverages the strengths of individual trees while maintaining robust performance across various data subsets.

**Modeling Details:**

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

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

**Evaluation:**

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

C**ross-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 Cross-Validation Accuracy:** 0.0061

**Final Training Accuracy:** 0.9112

**Final Test Accuracy:** 0.9124

**Confusion matrix:**

A screenshot of a graph

Description automatically generated

**High Classification Rates:**

* The model demonstrated strong classification rates with a total of 18,379 correct predictions for class 1 and 16,09 correct predictions for class 0, indicating effective overall performance.

**Good Recall and Precision:**

* The model showed good recall by accurately identifying a significant number of true positives, and it also maintained high precision with many correct predictions for class 0.

**Balanced Performance:**

* The performance reflects a balanced approach, with the model effectively managing to predict both classes, and performs much better than decision tree.

**ROC/AUC:**

A graph with a red line

Description automatically generated

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

**Feature Importance:**

**A graph with blue rectangular bars

Description automatically generated with medium confidence**

Using the importance function from the random forest, we can evaluate how the variables influence growing stress levels as it is the best performing model of all the others. The following image highlights the top five factors affecting growing stress, which can be interpreted as:

**Mental Health History**: A history of mental health issues or uncertainty about mental health (responses of "yes" or "maybe") indicates a greater likelihood of stress escalation

**Days Indoors**: Spending more days indoors (7.5 to 365 days) is positively correlated with increased stress levels.

**Work Interest**: Changes in work interest (e.g., "yes" or "maybe") are associated with higher stress, reflecting varying engagement levels.

**Mood swings**: Mood swings—whether high, medium, or low—appear to significantly contribute to growing stress levels.

**Changes in Habits:** Changes in habits are associated with high stress levels, where significant changes (e.g., "yes" or "maybe") contribute to growing stress.

***Task 4: SMART Question 3***

**What specific factors impact the growing stress among students?**

**Task 4.1 EDA:**

**A graph of stress distribution

Description automatically generated**

In our dataset, students had the second-highest proportion of survey responses that indicated experiencing feelings of growing stress. To add to this, we found this would be an interesting question to gain insight into since we’re students.

***Task 4.2: Statistical Testing***

**H₀: {Occupation} and Growing\_Stress are independent.**

**H₁: {Occupation} and Growing\_Stress are not independent.**

**P-value = 0.00**

The p-value < 0.00 for all the occupation and our target variable Growing Stress suggest that they not independent.

***Task 4.3: Modeling***

For this question, we will deploy logistic regression modeling techniques.

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

**Model: Logistic Regression**

**Reason for Selection:** Logistic Regression was chosen as it is a fundamental classification algorithm. Despite the correlation matrix indicating a lack of clear linear relationships between variables, the model was used to validate its performance.

**Modeling Details:**

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

***Task 4.3: Evaluation***

The model performed poorly, achieving an accuracy of only 61%.

**Confusion Matrix:**

**A blue squares with numbers

Description automatically generated**

**High misclassification rates:**

* **2063 instances** where the model predicted **1** for actual **0** (false positives).

**Poor recall and precision:**

* High **false positives** (3469) suggest the model frequently predicts class 1 incorrectly.

**Imbalanced performance:**

* The model misclassifies a large number of actual class 0 instances, leading to unreliable predictions.
* The model was quick to predict a student experiencing growing stress when in actuality, they did not report so

**ROC/AUC:**

A graph with a line going up

Description automatically generated

The model didn’t perform very well. With an accuracy of 61% and a AUC of 0.62, it left much to be desired.

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.

**Coefficient evaluation:**

A graph with red squares and green squares

Description automatically generated

From observing the coefficients, we concluded the following:

**Changes\_Habits**: Changes in habits had the largest effect on whether or not students were more likely to report growing stress. Students who reported a change in their habits were more likely not to report growing stress.

**Days Indoors**: Students spending more days indoors are more likely to report increased stress levels.

**Gender**: Men, more than women, were more likely to report growing stress.

**Mood Swings**: Mood swings—whether high, medium, or low—also impact growing stress among students, but not as much as the other variables.

**Social Weakness:** Students who reported having social weakness were only slightly more likely to report growing stress.

**Results**

**Conclusions**

In conclusion, the top five factors contributing to growing stress across all occupations include Mental Health History, Days Indoors, Work Interest, Mood Swings, and Changes in Habits. For students, the primary stressors are Changes in Habits, Days Indoors, Gender, Mood Swings, and Social Weakness. Individuals with a family history of mental health issues may be more likely to recognize their symptoms or seek treatment. To mitigate stress, it is important across all sectors to spend more time outdoors, acknowledge stress when experiencing changes in habits or frequent mood swings, and seek treatment when symptoms arise. The negative relationship between growing stress and treatment underscores the critical need to address stress proactively.

**Recommendations**

**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)