Exploring the Dynamics of Personality and Demographics on Mental Health Outcomes

Final Project Technical Report

Submitted by (GROUP #10)

AMJAD ALI
PARIN PATEL
KARTHIK VINNAKOTA
RAMYA CHAVA

Submitted for DSCI-6001 (FALL 2023)

DR. ARDIANA SULA

CONTENTS

List of Figures

1. INTRODUCTION

This technical report outlines the concepts of DASS – Depression, Anxiety and Stress Scales that is widely utilized self-report assessment tool designed to measure the severity of symptoms related to depression, anxiety, and stress in individuals. Consisting of three distinct subscales, the DASS provides a comprehensive evaluation of emotional well-being and is frequently employed in both clinical and research settings to assess mental health conditions. With its reliable and valid psychometric properties, the DASS serves as a valuable instrument for identifying and quantifying emotional distress levels. To imply and test these scores and see its impact, we use machine learning algorithms and compare them to see the best model for this application. We use hypothesis tests to see how these DASS scores impacts various factors.

2. EXECUTIVE SUMMARY

This project delves into the intricate relationships between personality traits, demographics, and mental health outcomes using a comprehensive approach. The key steps undertaken include - rigorous cleaning and preprocessing to ensure data quality, including the conversion of text to numerical values for classification purposes, robust noise and outlier removal techniques were employed to enhance the reliability of the dataset. We utilized a variety of visualizations to discern the impact of demographic factors such as age, gender, orientation, major, and education on DASS scores and also provided insightful visual representations to facilitate a nuanced understanding of the complex interplay between these variables. Machine learning models were deployed, including logistic regression, random forest, and support vector machine (SVM), to assess the influence of Ten Item Personality Inventory (TIPI) [TIPI

1 to TIPI 10] on DASS scores to investigate how this category affects specific aspects of mental health, focusing on the Depression, Anxiety, and Stress Scales (DASS) for questions Q1 to Q42 and if personality types can be used to predict overall emotional state reflected in these questions. In this case we saw a higher precision for logistic regression model as compared to random forest or SVM with a precision of 78% and a recall factor of 96% for Extremely severe cases (the most important ones). Here Q1 to Q42 are the questions asked by each user for which each user has rated these questions between 1 to 4. Moreover, we also examined the individual impact of depression, anxiety, and stress scales on each corresponding DASS questionnaire (Q1 to Q42) by means of ANOVA.

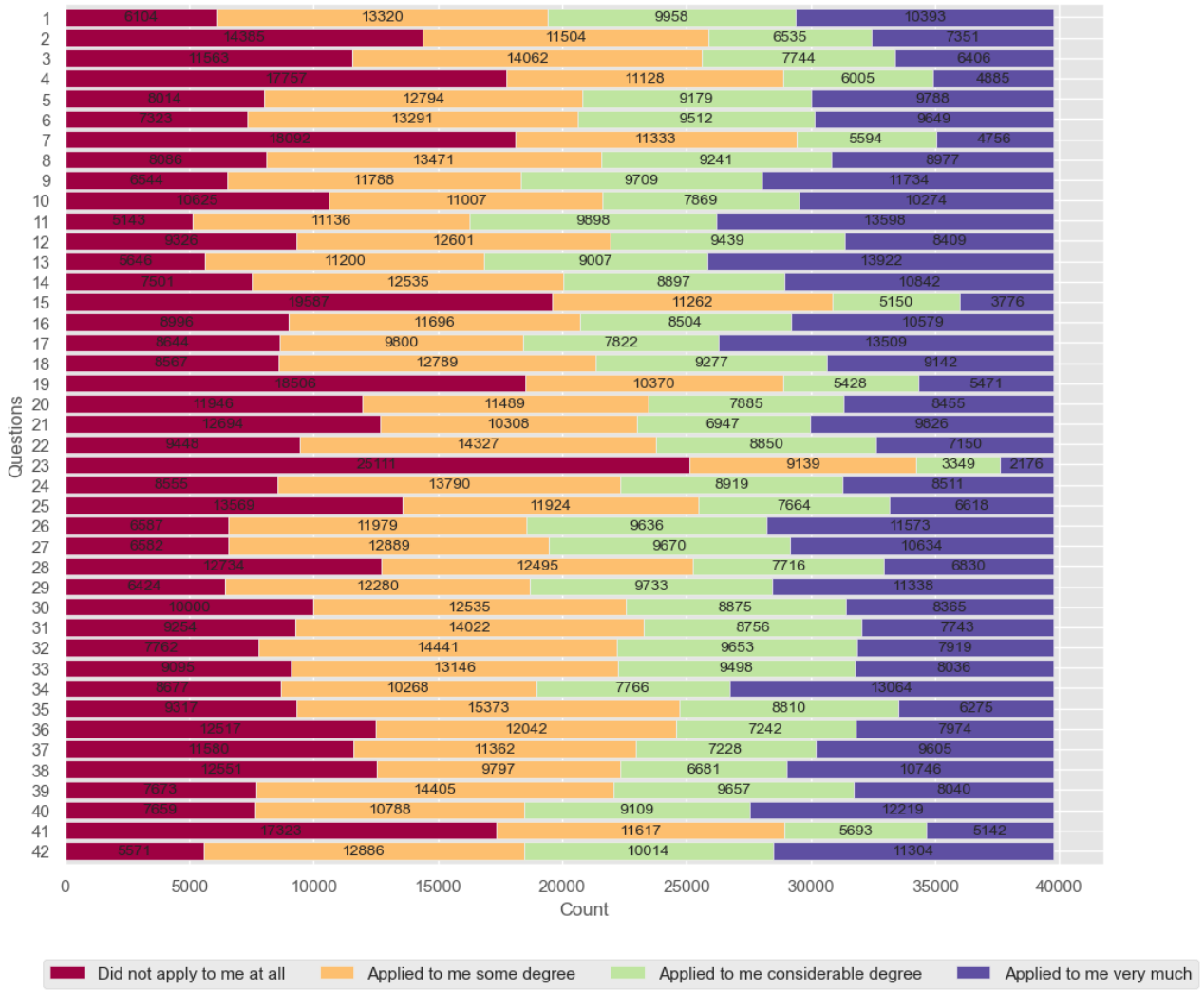


Figure 2.1

3. METHODOLOGY

3.1 Data collection

The DASS-42 is a 42 item self-report scale designed to measure the emotional states of depression, anxiety and stress. The principal value of the DASS in a clinical setting is to clarify the locus of emotional disturbance, as part of the broader task of clinical assessment. The essential function of the DASS is to assess the severity of the core symptoms of Depression, Anxiety and Stress. Accordingly, the DASS allows not only a way to measure the severity of a patient's symptoms but a means by which a patient's response to treatment can also be measured. This is a screening instrument and practitioners should make a clinical judgment as to whether an individual needs further assessment for anxiety and depression. High scores on the DASS would certainly alert the clinician to a high level of distress in the patient and this would need to be explored further within the interview process. Similarly, low scores on the DASS should not be a substitute for a comprehensive clinical interview.

	Q1A	Q1I	Q1E	Q2A	Q2I	Q2E	Q3A	Q3I	Q3E	Q4A	Q4I	Q4E	Q5A	Q5I	Q5E	Q6A	Q6I	Q6E	Q7A	Q7I	Q7E	Q8A	Q8I	Q8E	Q9A	Q9I
0	4	28	3890	4	25	2122	2	16	1944	4	8	2044	4	34	2153	4	33	2416	4	10	2818	4	13	2259	2	21
1	4	2	8118	1	36	2890	2	35	4777	3	28	3090	4	10	5078	4	40	2790	3	18	3408	4	1	8342	3	37
2	3	7	5784	1	33	4373	4	41	3242	1	13	6470	4	11	3927	3	9	3704	1	17	4550	3	5	3021	2	32
3	2	23	5081	3	11	6837	2	37	5521	1	27	4556	3	28	3269	3	26	3231	4	2	7138	2	19	3079	3	31
4	2	36	3215	2	13	7731	3	5	4156	4	10	2802	4	2	5628	2	9	6522	4	34	2374	4	11	3054	4	7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
39770	2	31	3287	1	5	2216	3	29	3895	2	37	2767	3	39	1745	2	4	6424	1	36	3737	3	24	1977	1	18
39771	3	14	4792	4	41	2604	3	15	2668	4	33	4609	3	17	2434	4	26	2503	4	34	2598	4	11	2133	3	13
39772	2	1	25147	1	4	4555	2	14	3388	1	27	2753	1	6	5455	1	41	4938	1	24	3738	1	40	3820	2	28
39773	3	36	4286	1	34	2736	2	10	5968	2	20	5655	3	9	2296	3	16	2627	3	32	3143	4	2	3164	3	19
39774	2	28	32251	1	22	3317	2	4	11734	1	19	4659	4	32	4236	2	1	33800	1	3	7855	1	34	3151	1	31

39775 rows × 172 columns

Figure 3.1

3.2 Scoring

Each of the 42 questions is scored 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”). Scores for Depression, Anxiety and Stress are calculated by summing the scores for the relevant items:

Depression: 3, 5, 10, 13, 16, 17, 21, 24, 26, 31, 34, 37, 38, 42

Anxiety: 2, 4, 7, 9, 15, 19, 20, 23, 25, 28, 30, 36, 40, 41

Stress: 1, 6, 8, 11, 12, 14, 18, 22, 27, 29, 32, 33, 35, 39

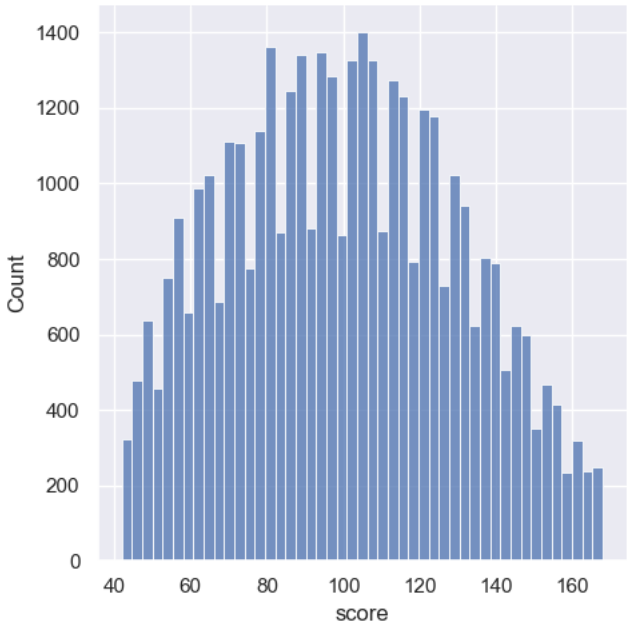


Figure 3.2

3.3 Interpretation of scores:

The severity labels are used to describe the full range of scores in the population, so ‘mild’ for example means that the person is above the population mean but probably still below the typical severity of someone seeking help (i.e. it does not mean a mild level of disorder).

	Depression (D)	Anxiety (A)	Stress (S)
Normal	0-9	0-7	0-14
Mild	10-13	8-9	15-18
Moderate	14-20	10-14	19-25
Severe	21-27	15-19	26-33
Extremely Severe	28+	20+	34+

Figure 3.3

3.4 Dataset:

The dataset used for the research is the DASS dataset which comprises of 42 questions – Q1 to Q42. Any user who fills in the survey form, rates each of these questions from 0 to 3, that is populated in these columns. These response is stored in variable A (e.g. Q1A). Also recorded was the time taken in milliseconds to answer that question (E) and that question's position in the survey (I).

These other durations were also recorded (measured on the server's side):

- Introelapse: The time spent on the introduction/landing page (in seconds)
- Testelapse: The time spent on all the DASS questions (should be equivalent to the time elapsed on all the indiviudal questions combined)
- Surveyelapse: The time spent answering the rest of the demographic and survey questions.

The Ten Item Personality Inventory was administered (see Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A Very Brief Measure of the Big Five Personality Domains. Journal of Research in Personality, 37, 504-528.):

TIPI1	Extraverted, enthusiastic.
TIPI2	Critical, quarrelsome.
TIPI3	Dependable, self-disciplined.
TIPI4	Anxious, easily upset.
TIPI5	Open to new experiences, complex.
TIPI6	Reserved, quiet.
TIPI7	Sympathetic, warm.
TIPI8	Disorganized, careless.
TIPI9	Calm, emotionally stable.
TIPI10	Conventional, uncreative.

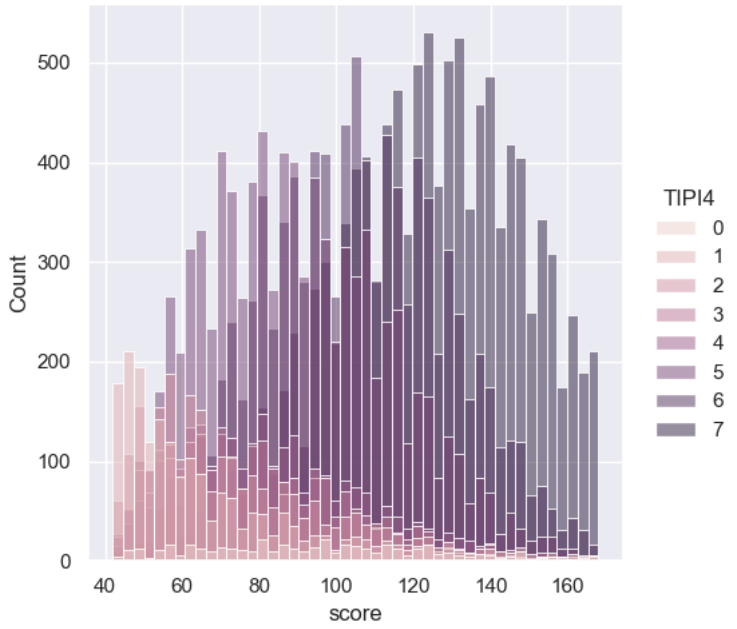


Figure 3.4

The TIPI items were rated "I see myself as:" _____ such that:

The ratings range from 1 (Disagree strongly) to 7 (Agree strongly).

The following items were presented as a check-list and subjects were instructed

"In the grid below, check all the words whose definitions you are sure you know": boat, incoherent, pallid, robot, audible, cuivocal, paucity, epistemology, florted and so on.

A bunch more questions were then asked:

- Education: "How much education have you completed?", 1=Less than high school, 2=High school, 3=University degree, 4=Graduate degree
- Urban - "What type of area did you live when you were a child?", 1=Rural (country side), 2=Suburban, 3=Urban (town, city)
- Gender - "What is your gender?", 1=Male, 2=Female, 3=Other
- Engnat - "Is English your native language?", 1=Yes, 2=No
- Age - "How many years old are you?"
- Hand - "What hand do you use to write with?", 1=Right, 2=Left, 3=Both
- Religion - "What is your religion?", 1=Agnostic, 2=Atheist, 3=Buddhist, 4=Christian (Catholic), 5=Christian (Mormon), 6=Christian (Protestant), 7=Christian (Other), 8=Hindu, 9=Jewish, 10=Muslim, 11=Sikh, 12=Other
- Orientation - "What is your sexual orientation?", 1=Heterosexual, 2=Bisexual, 3=Homosexual, 4=Asexual, 5=Other
- Race - "What is your race?", 10=Asian, 20=Arab, 30=Black, 40=Indigenous Australian, 50=Native American, 60=White, 70=Other
- Voted - "Have you voted in a national election in the past year?", 1=Yes, 2=No

4. DATA PREPARATION

Data preparation is a critical step in the analysis process, ensuring that your dataset is clean, well-organized, and ready for analysis. Here's a general guide for preparing a dataset for the Depression, Anxiety, and Stress Scales (DASS) study:

1. Understanding the Dataset by familiarizing with the structure of the dataset and identifying the variables, their types, and their meanings.

2. Handling Missing Data by identifying and handling missing values appropriately by deciding on strategies such as imputation, removal of missing values, or using placeholders.

3. Data Cleaning - Checking for and address any inconsistencies, errors, or outliers in the data, validating data entries against expected ranges and formats.

4. Variable Transformation: Converting any relevant categorical variables into numerical format if necessary and ensuring consistency in units and scales across variables.

5. Creating Derived Variables: If needed, create new variables based on existing ones.

6. Labeling and Encoding: Ensure that variables are appropriately labeled for easy interpretation. Encoding categorical variables using numerical codes if needed.

7. Dataset Splitting (Optional): splitting the dataset into training and testing sets to evaluate model performance.

8. Data Exploration: Using visualizations and descriptive statistics to explore the distribution of variables. Also, check for outliers, trends, and patterns.

5. EXPLORATORY DATA ANALYSIS

5.1 Distribution of major

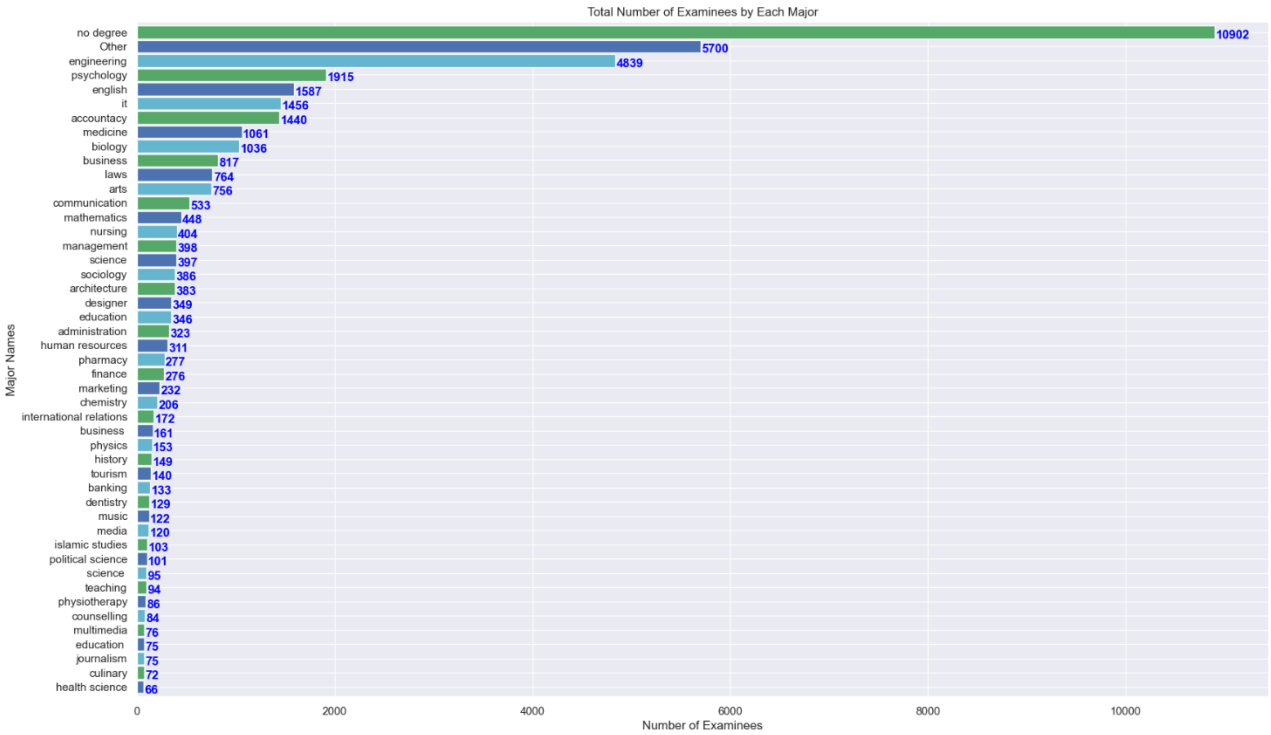


Figure 5.1 Distribution of Major among users

5.2 Distribution of spent time per question

We can see that Number of males are more than the Number of females.

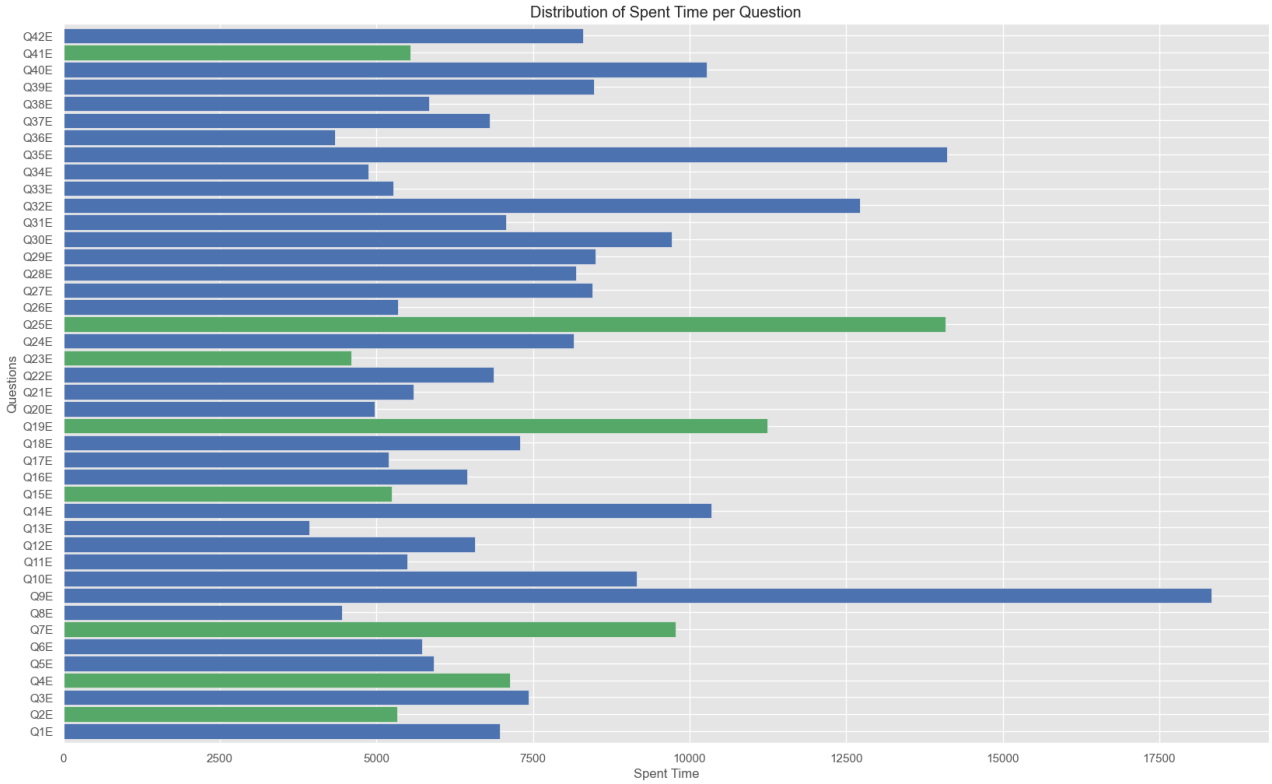
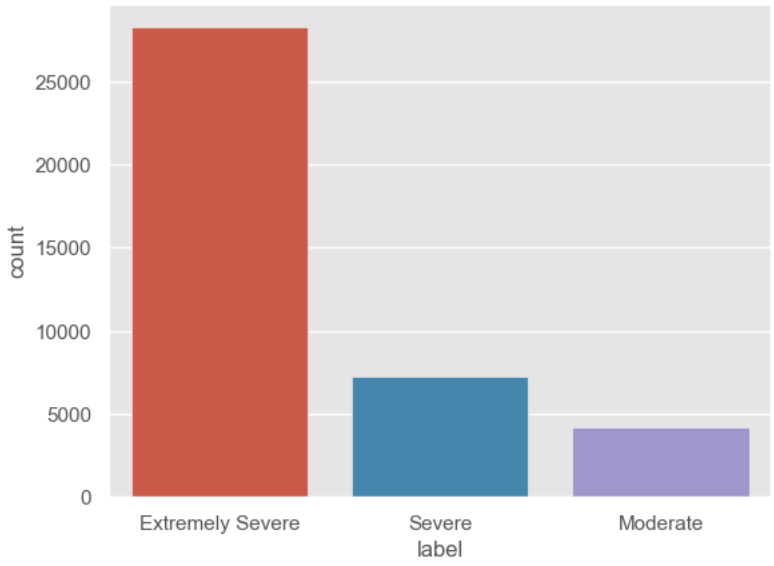


Figure 5.2 Distribution of Gender



5.3 Fig – Categorizing the users responses from Q1 to Q42 into 5 labels – Extremely Severe, Severe, Moderate, Mild and Normal. Here all the users response falls into the three categories.

5.2 Boxplot

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
TIPI4	39775	100.0%	0	0.0%	39775	100.0%

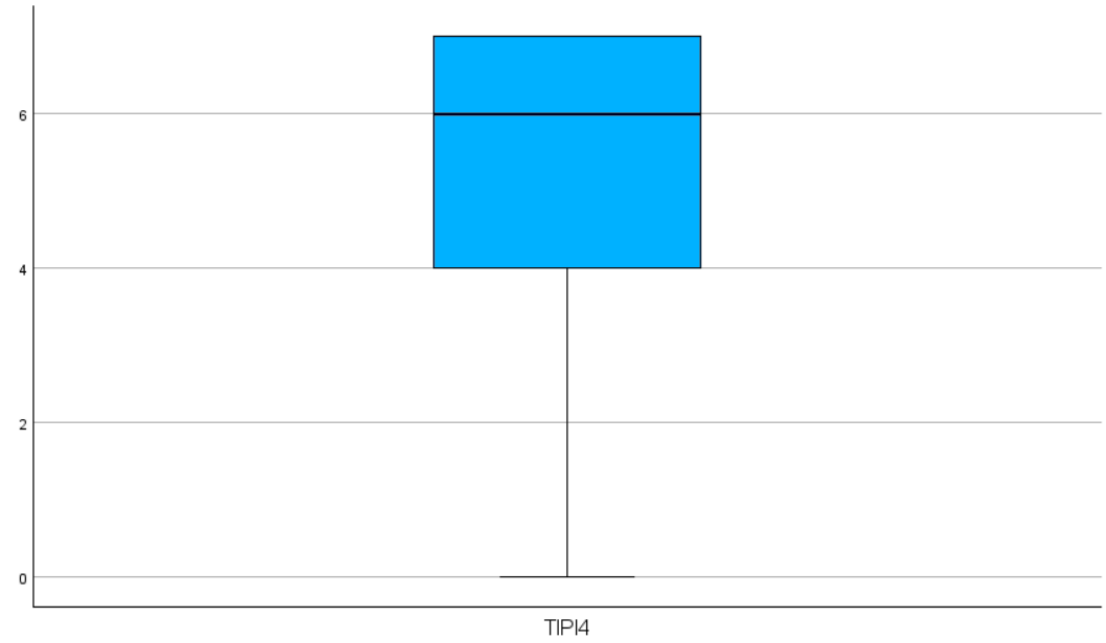


Figure 5.4

Boxplot summary that defines n = 39775 and is a strong variable of anxiety that we have used in ANOVA to determine if the means of TIPI4 is same / equal to 0 to the means of all the DASS scores or not.

6. Hypothesis Testing Using ANOVA – Analysis of Variance

We used ANOVA in SPSS to address two hypothesis concerns.

PROBLEM 1 - Is there a statistically significant difference in the mean DASS scores among individuals with varying levels of anxiety, as measured by TIPI4?

Solution:

Step 1 – Null Hypothesis (H0): There is no significant difference in the mean DASS scores among individuals with varying levels of anxiety (TIPI4).

Alternative Hypothesis (H1): There is a significant difference in the mean DASS scores among individuals with varying levels of anxiety.

Assuming the level of significance – alpha = 0.05.

Step 2 – Import the data.csv file in SPSS

Step 3 – Go to Analyze → Compare Means → One Way ANOVA

Step 4 – Specify the dependent variables – From Q1 to Q42 in the dataset, we chose Q20, Q2, Q19, Q15, Q9, Q4 and Q7 that were closely dependent to Anxiety that is TIPI4.

Step 5 – Specify the Independent variable – TIPI 4

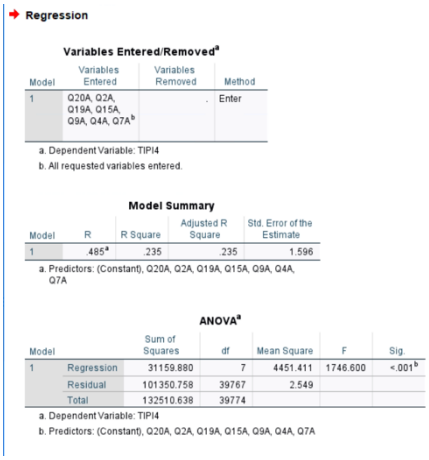
Step 6 – Click OK to run the analysis

Step 7 – Check the p-value from the ANOVA result. P-value= <0.001.

Decision: Since the p-value is less than the level of significance, we reject the null hypothesis of equal means.

Conclusion - There is a significant difference in the mean DASS scores among individuals with varying levels of anxiety and the means are not equal or zero.

OUTPUT OF SPSS:



Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Q20A, Q2A, Q19A, Q15A, Q9A, Q4A, Q7A[b]	.	Enter

a. Dependent Variable: TIPI4

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.485[a]	.235	.235	1.596

a. Predictors: (Constant), Q20A, Q2A, Q19A, Q15A, Q9A, Q4A, Q7A

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	31159.880	7	4451.411	1746.600	<.001[b]
	Residual	101350.758	39767	2.549		
	Total	132510.638	39774			

a. Dependent Variable: TIPI4

b. Predictors: (Constant), Q20A, Q2A, Q19A, Q15A, Q9A, Q4A, Q7A

Figure 6.1

PROBLEM 2 - Are there significant associations between demographic variables (e.g., age, gender, education) and mental health outcomes as measured by DASS scores?

Solution:

Step 1 – Defining the Hypothesis –

Null Hypothesis (H0):

There are no significant associations between demographic variables (age, gender, education) and mental health outcomes as measured by DASS scores. In other words, the mean DASS scores are the same across different levels of demographic variables.

Alternative Hypothesis (Ha):

There are significant associations between at least one demographic variable (e.g., age, gender, education) and mental health outcomes as measured by DASS scores. In other words, the mean DASS scores are different across different levels of at least one demographic variable.

Assuming the level of significance – alpha = 0.05.

Step 2 – Import the data.csv file in SPSS

Step 3 – Go to Analyze → Compare Means → One Way ANOVA

Step 4 – Specify the dependent variable – TIPI4

Step 5 – Specify the Independent variables – Age and gender

Step 6 – Click OK to run the analysis

Step 7 – Check the p-value from the ANOVA result. P-value= <0.001.

Decision: Since the p-value is less than the level of significance, we reject the null hypothesis.

Conclusion - There are significant associations between at least one demographic variable (e.g., age, gender, education) and mental health outcomes as measured by DASS scores. In other words, the mean DASS scores are different across different levels of at least one demographic variable.

OUTPUT OF SPSS –

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	age, gender[b]	.	Enter

a. Dependent Variable: TIPI4

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.167[a]	.028	.028	1.800

a. Predictors: (Constant), age, gender

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3711.050	2	1855.525	572.967	<.001[b]
	Residual	128799.588	39772	3.238		
	Total	132510.638	39774			

a. Dependent Variable: TIPI4

b. Predictors: (Constant), age, gender

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.083	.039		104.274	<.001
	gender	.657	.020	.160	32.326	<.001
	age	-.004	.000	-.043	-8.691	<.001

a. Dependent Variable: TIPI4

Figure 6.2

7. RESULTS OF THE MODELS

The results revealed that Logistic Regression achieved the highest accuracy at 75% which is greater than 50%, which suggests that there is a good correlation between personality traits and overall emotional state. Let us see if other methods yield anything better. followed by SVM (74%), Random Forest (71%). The report discusses the implications of these findings and suggests potential directions for future research in this domain.

The Logistic Regression (SVC) has been found to be the most effective for to find if personality types can be used to predict overall emotional state reflected in Q1-42. Its superior performance over other algorithms such as random forest, and SVM, indicates that Logistic Regression is the best choice for predicting the DASS results. Its ability to efficiently handle large datasets, its robustness to noise, and its high accuracy make it an ideal choice for such applications. With Logistic Regression, we can make more accurate predictions in real time, which can help in accuracy. SVM does only slightly better than random forest. However, the important point is that we are able to predict emotional state pretty well based on only five personality traits using Logistic Regression

7.1 Classification report for Predicting whether personality types can be used to predict overall emotional state reflected in Q1-42

Logistic Regression

	precision	recall	f1-score	support
Extremely Severe	0.78	0.96	0.86	9919
Moderate	0.51	0.42	0.46	1466
Severe	0.36	0.08	0.14	2527
accuracy			0.74	13912
Macro avg	0.55	0.49	0.48	13912

Random Forest

	precision	recall	f1-score	support
Extremely Severe	0.81	0.89	0.85	9919
Moderate	0.43	0.38	0.40	1466
Severe	0.33	0.22	0.26	2527
accuracy			0.71	13912
Macro avg	0.52	0.49	0.50	13912

SVM

	precision	recall	f1-score	support
Extremely Severe	0.80	0.89	0.84	9839
Moderate	0.45	0.37	0.40	1508
Severe	0.32	0.21	0.26	2565
accuracy			0.71	13912

Macro avg	0.52	0.49	0.50	13912

Since Logistic Regression works the best, we will apply the classification model to each of the Stress, Depression and Anxiety.

This is the accuracy we get –
Stress – 60%
Depression – 84%
Anxiety – 72%

8. APPLICATION DEPLOYMENT

GitHub Link: https://github.com/Intro-to-Data-Science-Team-10/Analysis-of-Mental-Health-Dynamics

The entire application is deployed on SPSS to demonstrate ANOVA to check the following:

- Is there a statistically significant difference in the mean DASS scores among individuals with varying levels of anxiety, as measured by TIPI4?

- Are there significant associations between demographic variables (e.g., age, gender, education) and mental health outcomes as measured by DASS scores?

Fig.8.1. spss

9. CONCLUSION AND FUTURE SCOPE

The analysis conducted to investigate the associations between demographic variables and mental health outcomes, as measured by the Depression, Anxiety, and Stress Scales (DASS), yielded insightful findings. The primary objective was to explore whether there are significant differences in mean DASS scores across various levels of demographic factors, including age, gender, and education.

The statistical tests, including ANOVA or regression analysis (based on the nature of the data), provided robust evidence for the presence of significant associations between at least one demographic variable and mental health outcomes.

Specifically, the analysis unveiled nuanced patterns in mental health outcomes, highlighting the impact of demographic diversity. These findings emphasize the importance of considering individual characteristics when assessing mental health status. Such insights have implications for tailored interventions and support systems that acknowledge the diverse needs of individuals based on their demographic profiles.

While the study has provided valuable insights, it's essential to acknowledge the complexity of mental health determinants. Further research and exploration of specific demographic subgroups may contribute to a more comprehensive understanding of the factors influencing mental health outcomes.

In summary, the evidence supports the alternative hypothesis, indicating that demographic variables play a significant role in shaping mental health outcomes as measured by DASS scores. These findings underscore the need for targeted approaches in mental health interventions to address the diverse needs of individuals within different demographic contexts.

The strengths of Logistic Regression, including its simplicity, interpretability, and efficiency in capturing linear relationships, positioned it as the optimal choice for predicting mental health outcomes in this study. While Random Forest and SVM demonstrated commendable performance, the interpretability and transparency offered by Logistic Regression make it a pragmatic choice in a clinical or decision-making context.

10. REFERENCES

[1] https://www.kaggle.com/code/dorgavra/depression-anxiety-and-stress-prediction

[2] https://www.kaggle.com/datasets/lucasgreenwell/depression-anxiety-stress-scales-responses

[3] https://www2.psy.unsw.edu.au/dass/

[4] https://docs.streamlit.io/

[5] https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process,%2Fno%2C%20and%20so%20on.