1. Introduction

Mental health presents a concern for an estimated 25% of people in both the United States and England in any given year (Johns Hopkins Medicine; NHS Digital, 2009). Even controlling for accessibility of mental health diagnosis and treatment possibly more pervasive in developed western societies, Our World in Data reports that more than 10.7%, or 792 million people, are estimated to have suffered from a mental disorder globally in 2017 (Dattani, Ritchie and Roser, 2021).

Depression and anxiety disorders account for two of the most ubiquitous mental disorders. Approximately 3.4% and 3.8% of the global population are estimated to have suffered from depression and anxiety in 2017, respectively (Dattani, Ritchie and Roser, 2021). Both disorders are clinically treatable; furthermore, the clinical sciences have established potentially relevant predictors of depression and anxiety. For instance, socioeconomic status, personality and gender have all been shown to predict mental health (Lorant et al, 2003; Burešová et al., 2020; Rosenfield and Mouzon, 2012).

From a data analytic perspective, the established literature on predictors of depression and anxiety presents a currently under-explored opportunity to predict demographics that are likely to disproportionately experience depression and anxiety. The potential utility of such models is immense: first, they can provide mental health providers and policymakers a method to identify susceptible populations *in advance* of a formal diagnosis. Consequently, clinical practice and policy interventions can positively shift from the currently treatment-focused interventions to a preventative approach that addresses vulnerable populations before the development and exacerbation of depression and anxiety. Second, models can be used to formally assess the validity of proposed predictors from a theoretical perspective.

Motivated by this opportunity to enrich psychological science with data analytic principles, our group aimed to examine **whether we can use supervised machine learning techniques to predict an individual's propensity to experience depression and anxiety**. Our final analysis utilised a large dataset of 30434 individuals who provided online responses to the Depression Anxiety Stress Scale 42 (DASS-42) between 2017-2019. The scale comprised 42 questions with possible responses ranging from 1 to 4 and examined the extent respondents were experiencing depression, anxiety and stress/tension symptoms at time of completion. We used a set of accompanying personality and demographic variables as predictors of DASS outcomes. The precise explanations and theoretical justifications for the chosen data are discussed in detail in Section 2 (see 2. Data Description and Preparation).

To address our research problem, we built three classification models in R using decision tree, random forest and logistic regression techniques for both depression and anxiety. Given that depression and anxiety are often comorbid and closely associated, we built distinct models for each disorder to ascertain whether they share common predictors and if not, identify the most significant predictors for each (Beuke, Fischer and McDowall, 2003). The models are presented and evaluated in detail in Section 3 (see 3. Methodology and Analysis).

2. Data Description and Preparation

I. Data description

The analysis used anonymised, publicly available individual-level data obtained from Kaggle's *Predicting Depression, Anxiety and Stress* dataset. The dataset contains 39775 responses to 42 questions corresponding to the Depression Anxiety Stress Scale 42 (DASS-42), 10 questions corresponding to the Ten Item Personality Inventory (TIPI) and 13 questions corresponding to a series of demographic variables. Further, the dataset contains a validity check which requests participants to examine whether a series of 16 words are real words or not. Finally, the dataset contains technical information including the duration spent on the survey. The data was collected between 2017 and 2019 from an online survey available to anyone interested in receiving personalised results on the subject matter. The dataset contains participants who provided consent for their responses to be used for research.

This dataset is highly appropriate for our goal, as it effectively provides outcome measurements of depression and anxiety in addition to potential predictors including personality and demographic variables from the same participant.

II. Data preparation

The outcome variables of depression and anxiety were constructed based on the sum scores for the relevant items in the DASS-42. Consistent with the severity-rating index presented by (Lovibond and Lovibond, 1995), we generated a variable for both depression and anxiety to determine whether the sum score should be considered as high or low for their respective measure, as follows:

Depression, anxiety and stress variables:

- **depression_label**: 'High' 'Low' - Sum scores of depression items which were equal to or exceeded 35 labeled as 'High', otherwise 'Low'.

- **anxiety_label**: 'High' 'Low' - Sum scores of anxiety items which were equal to or exceeded 29 labeled as 'High', otherwise 'Low'.

While we did not aim to evaluate predictors for stress/tension, stress/tension was also examined by the DASS-42 and we generated a corresponding variable for completeness, as follows:

- **stress_label**: 'High' 'Low' - Sum scores of stress/tension items which were equal to or exceeded 40 labeled as 'High', otherwise 'Low'.

It is relevant to note that the severity-rating index considers severities of Normal, Mild, Moderate, Severe and Extremely Severe. However, for our purposes, we grouped all scores above moderate as 'High' and all scores below moderate as 'Low'. We note the implications of this decision in Section 5 (see 5. Conclusion). Further, the original DASS consists of possible responses ranging from 0 to 3, while the DASS administered in our dataset uses a response range from 1 to 4. Therefore, we adjusted the severity-rating index accordingly. Given that the response interval remains the same (4), this has no impact on how the scores are rated.

Personality variables:

The personality variables were constructed based on the scores for the relevant items in the TIPI. The TIPI is a psychometric measurement of the Five-Factor model of personality that consists of traits Extraversion, Agreeableness, Emotional Stability, Conscientiousness and Openness (Gosling, Rentfrow and Swann Jr., 2003). The TIPI contains ten questions, with two questions for each trait. One question for each trait is positively associated (e.g., Extraversion: I see myself as extraverted, enthusiastic) while another question is negatively associated (e.g., Extraversion: I see myself as reserved, quiet). Therefore, we constructed a personality variable for each trait by reverse scoring the negatively associated question to form a composite score for the trait, as follows. All personality descriptions are consistent with the Five-Factor model (Gray and Bjorklund, 2018).

- **extraversion**: Measures the extent one enjoys social interactions and is energised. Composite measure of TIPI1 (extraverted, enthusiastic) and TIPI6 reversed (reserved, quiet).
- **agreeableness**: Measures the extent one acts sympathetically and unselfishly. Composite measure of TIPI7 (sympathetic, warm) and TIPI2 reversed (critical, quarrelsome).
- **emotional_stability**: Measures the extent one does not get upset easily and is not prone to depression. Composite measure of TIPI9 (calm, emotionally stable) and TIPI4 reversed (anxious, easily upset).

- **conscientiousness**: Measures the extent one is industrious towards their goals. Composite measure of TIPI3 (dependable, self-disciplined) and TIPI8 reversed (disorganised, careless).
- **openness**: Measures the extent one seeks out new experiences. Composite measure of TIPI5 (open to new experiences, complex) and TIPI10 reversed (conventional, uncreative).

Demographic variables:

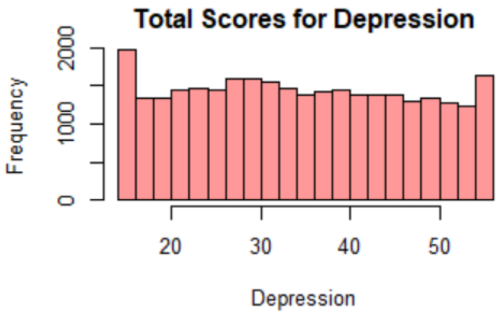
12 of 13 demographic variables were used as the second set of predictors for depression and anxiety, as follows:

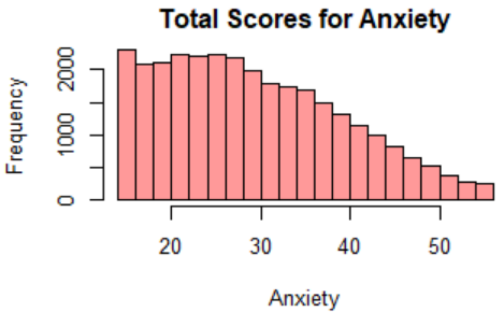
- **education:** Elicits the highest level of education obtained. 1=Less than high school, 2=High school, 3=University degree, 4=Graduate degree. Higher scores indicate greater educational obtainment.
- **urban:** Elicits the type of area the participant lived in as a child. 1=Rural (country side), 2=Suburban, 3=Urban (town, city).
- **gender:** Elicits the gender of the participant. 1=Male, 2=Female, 3=Other.
- **engnat:** Elicits whether the participant's native language is English. 1=Yes, 2=No.
- **age:** Elicits the age of the participant.
- **hand:** Elicits the hand the participant writes with. 1=Right, 2=Left, 3=Both.
- **religion:** Elicits the participant's 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:** Elicits the participant's sexual orientation. 1=Heterosexual, 2=Bisexual, 3=Homosexual, 4=Asexual, 5=Other.
- **race:** Elicits the participant's race. 10=Asian, 20=Arab, 30=Black, 40=Indigenous Australian, 50=Native American, 60=White, 70=Other
- **voted:** Elicits whether the participant has voted in a national election in the past year. 1=Yes, 2=No.
- **married:** Elicits the participant's marital status. 1=Never married, 2=Currently married, 3=Previously married.
- **familysize:** Elicits the number of children the participant's mother has, including the participant.

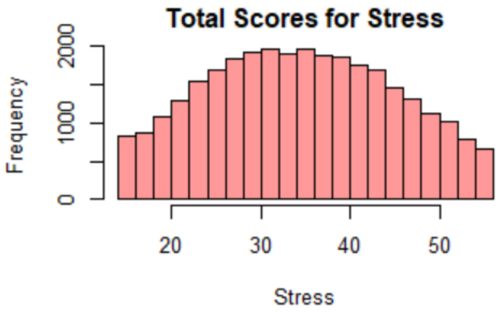
The data was cleaned for missing values, outliers, transformations and validity. The variable **major** was a free-response question that contained a significant number of missing values and is therefore removed. Thereafter, participants with missing responses were removed. Outliers for **age** were handled by removing responses greater than 110. We performed log transformation on **age** and **familysize** given that the two variables demonstrated a highly skewed distribution. Finally, we removed all participants who failed the validity check (i.e., participants who identified any of the not real words as real words). This effectively accounts for acquiescence, whereby participants provide responses without considering the question. The final dataset contained responses from 30434 individuals, among which 15193 were classified as high in depression, 15241 were classified as low in depression, 15098 were classified as high in anxiety and 15336 were classified as low in anxiety. The data was very balanced between high and low scorers in depression and anxiety, providing an approximately equal chance for the models to detect positive and negative signals.

Distribution of scores for depression, anxiety and stress:

The initial data analysis showed that the distributions of scores for depression, anxiety and stress levels exhibit different characteristics such that total scores for depression indicate a roughly uniform distribution, total scores for anxiety being skewed towards the lower end of the scale and total scores for stress following a normal distribution. For all three subscales, the calculated total sums ranged from 14 to 56 which was in accordance with our adjusted scale. No outliers have been observed. The mean for depression level was 34.89 and the median 34.00, for anxiety level the mean was 29.75 and the median 28.00, and for stress level, the mean was 34.96 and the median 35.00.

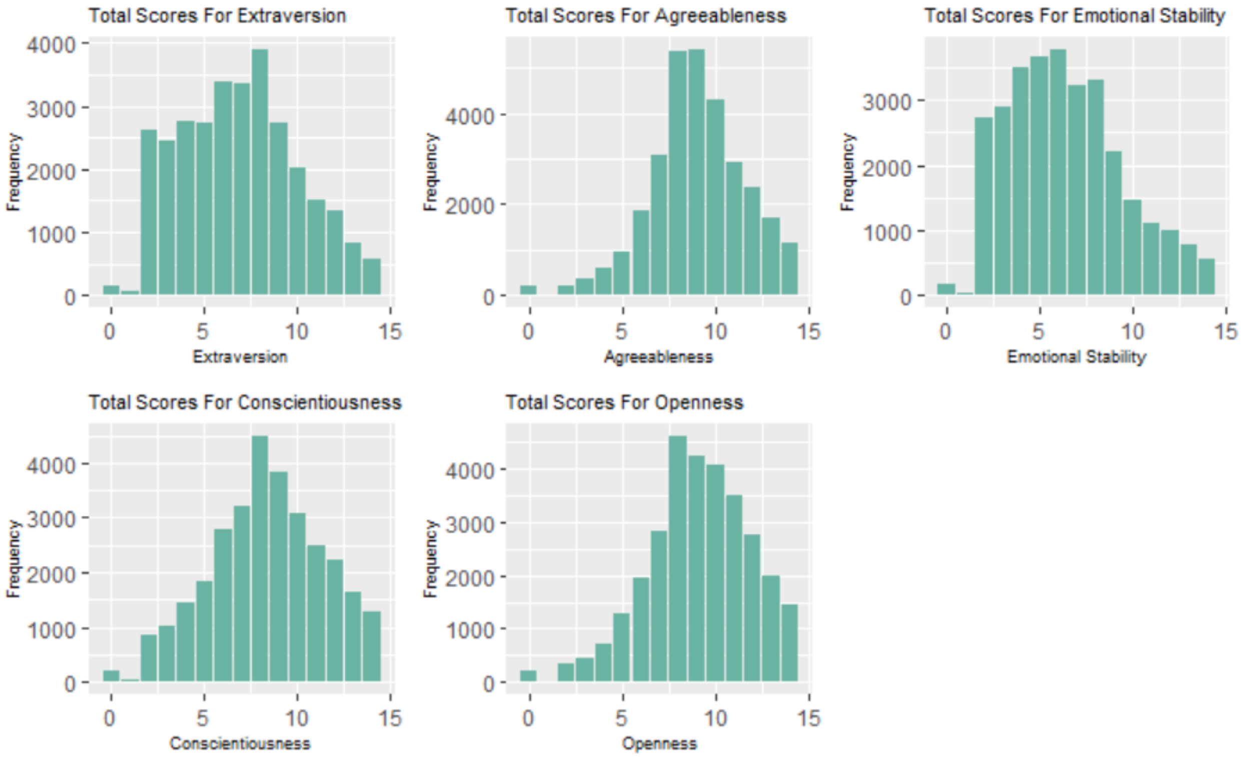






Distribution of scores for personality traits:

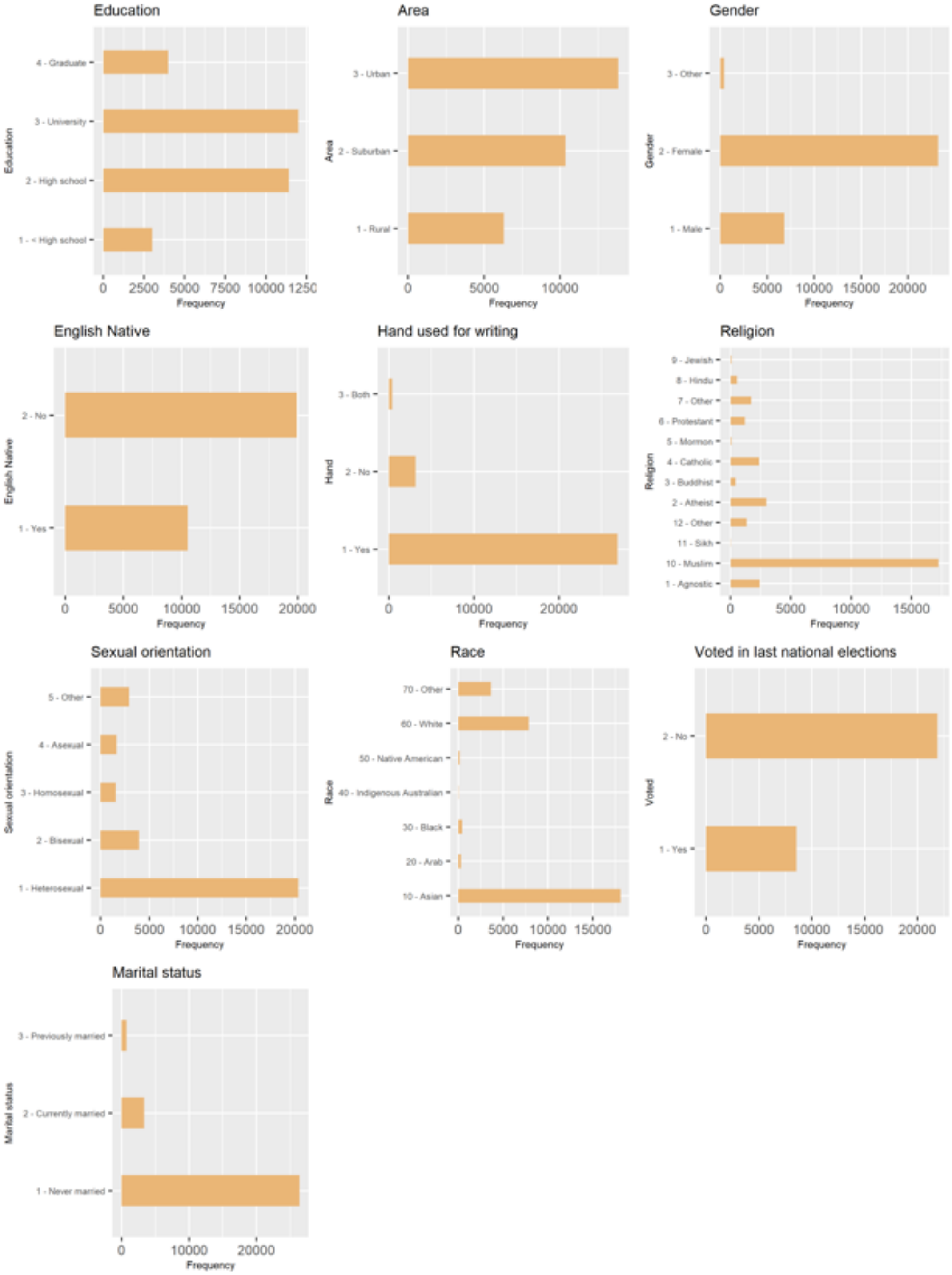
As expected, the analysis that we have run on five personality items showed that to a large extent all distributions followed the shape of a normal distribution. However, the data also deviated from this trend in some important manners. All calculated total scores were within the range of 0-14 in accordance with the design of the test. We observed that the data is slightly skewed towards the right for agreeableness (-0.36), openness (-0.42), and conscientiousness (-0.18) and towards the left for emotional stability (0.47) and extraversion (0.21). This could be in part explained by the specificity of the surveyed group which consisted mostly of young and female respondents as indicated later in this report. However, this skewness is quite small, and the means align quite well with the medians for all variables (extraversion: M = 6.86, Mdn = 7; agreeableness: M = 9.02, Mdn = 9; emotional stability: M = 6.40, Mdn = 6; conscientiousness: M = 8.34, Mdn = 8; openness: M = 9.09, Mdn = 9).



Distributions of categorical demographic variables:

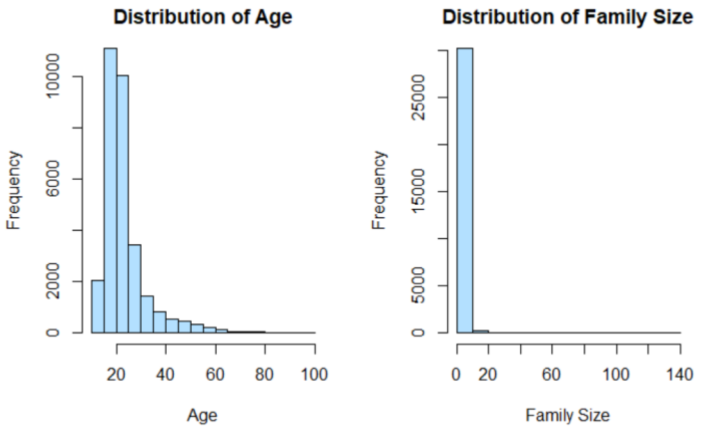
Exploring the categorical variables in our demographic group have revealed interesting characteristics about the respondents that took part in the survey. Most of the respondents indicated to have either a high school diploma (38%) or university degree (39%), and they were living in the urban areas (45%). Interestingly, the majority of the group was female (76%), and their native language was not English

(65%). 88% of people reported to use their right hand for writing. When it comes to religion and sexual orientation, 56% respondents identified as Muslim and 67% reported to be Heterosexual. Surprisingly, when we looked at the reported race, we found out that most of the respondents were Asian (59%) with the second most frequently reported race being White (26%). Lastly, it turned out that most people reported not to have voted in the last national election (72%), and never have been married (87%).



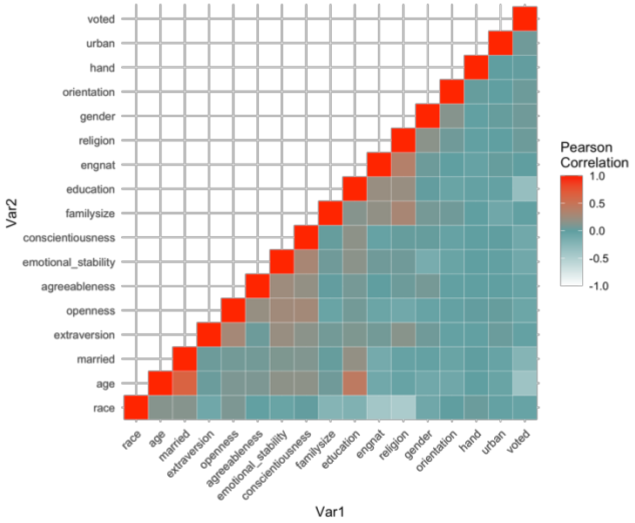
Distributions of numerical demographic variables:

Examining age and family size, two numerical variables from the demographic group, has revealed a significant skewness of scores towards the lower ones. The skewness of age and family size distributions were around 2.32 and 11.56 respectively which suggests that the scores are highly skewed. Moreover, the kurtosis of around 10 for the age variable and around 600 for the family size variable indicate that the scores cluster around the peaks of the distributions closely to the means. The mean age of survey respondents was 23.45 years showing that the data is not strictly representative of a larger population and might mainly portray the situation of young people.



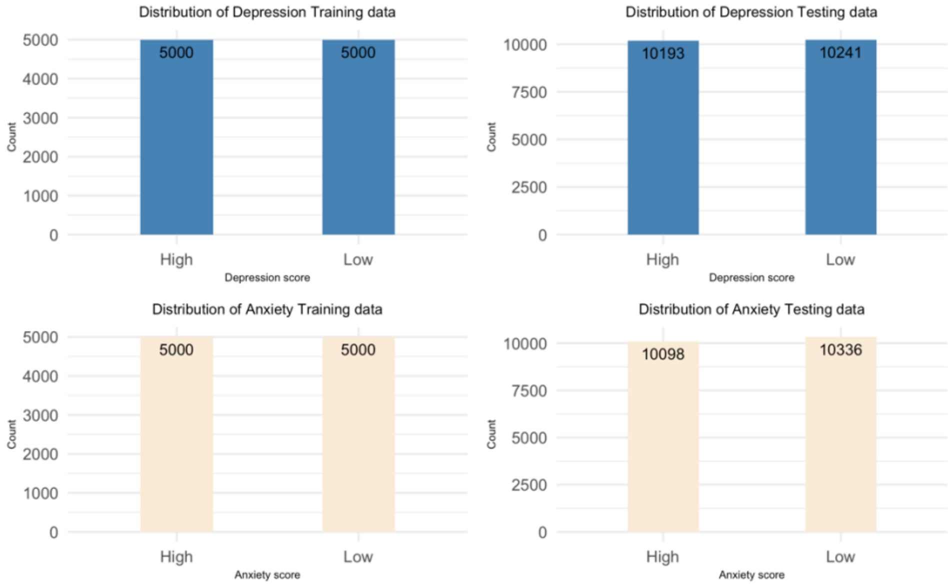
Correlation matrix heatmap:

The heatmap below illustrates the Pearson correlations between the predictor variables.



Splitting training and testing data:

We randomly split the dataset into training and testing data in an approximately 1:3 ratio for both depression and anxiety models. The training dataset for both depression and anxiety contained 10,000 observations, 5,000 of which included high scorers and 5,000 of which included low scorers. The remaining 20434 observations were assigned to the testing datasets. The testing dataset for depression contained 10193 high scorers and 10241 low scorers. The testing dataset for anxiety contained 10098 high scorers and 10366 low scorers. The distributions of our training and testing datasets are visualised below.



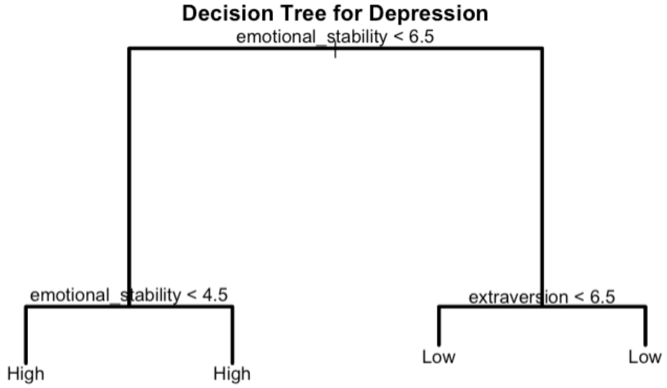
3. Methodology and Analysis

Three classification models were built to predict outcomes of the DASS-42: decision tree, random forest and logistic regression. Given that we aim to ascertain whether depression and anxiety share common predictors, we modelled the two outcomes separately while maintaining the same method to enable comparability.

I. Decision tree

(a) Depression

A decision tree is a supervised technique that identifies variables that contain information on the outcome of interest, rendering it an appropriate method to determine the strongest predictors of depression and anxiety scores. We first fitted a decision tree to predict depression scores using the training dataset and all predictor variables, excluding **depression_label** itself. The depression model below indicates that emotional stability and extraversion are strong predictors for depression outcomes. Low emotional stability is expected to yield high depression scores, while low extraversion is expected to yield low depression scores. To avoid overfitting and enable generalisation beyond the training data, we pruned the tree using cross validation and 10-fold cross validation (Provost and Fawcett, 2013). Both methods confirm that the tree with 4 terminal nodes obtains the minimum entropy (i.e., the minimum deviance).

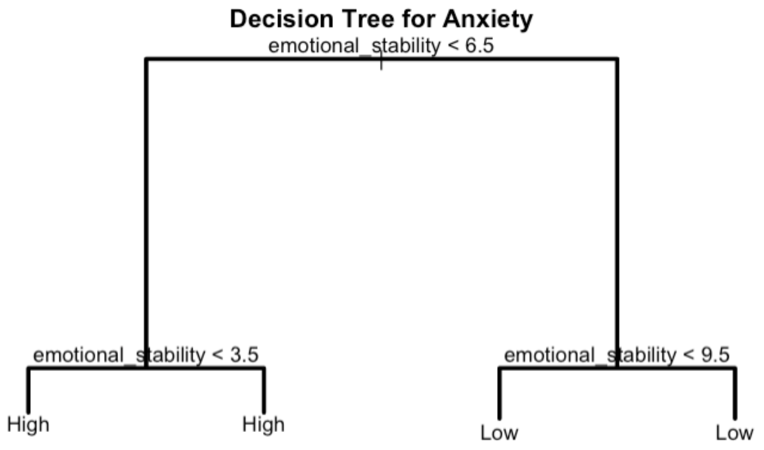

Decision Tree for Depression
emotional_stability < 6.5
emotional_stability < 4.5
extraversion < 6.5
High
High
Low
Low

We subsequently tested the performance of the model using the testing dataset that generated a misclassification rate of 31.8% and an accuracy rate of 68.2%. The corresponding confusion matrix is visualised below.

Prediction/Actual	**High**	**Low**
High	7510	3814
Low	2683	6427

(b) Anxiety

We applied the same method to predict anxiety scores. The decision tree below indicates that emotional stability is the strongest predictor for anxiety outcomes. Emotional stability scores lower than 3.5 are expected to yield high anxiety scores while scores greater than 3.5 and lower than 9.5 are expected to yield low anxiety scores. Cross validation and 10-fold cross validation both affirm that the tree with 4 terminal nodes obtains the minimum entropy (i.e., the minimum deviance).


Decision Tree for Anxiety
emotional_stability < 6.5
emotional_stability < 3.5
emotional_stability < 9.5
High
High
Low
Low

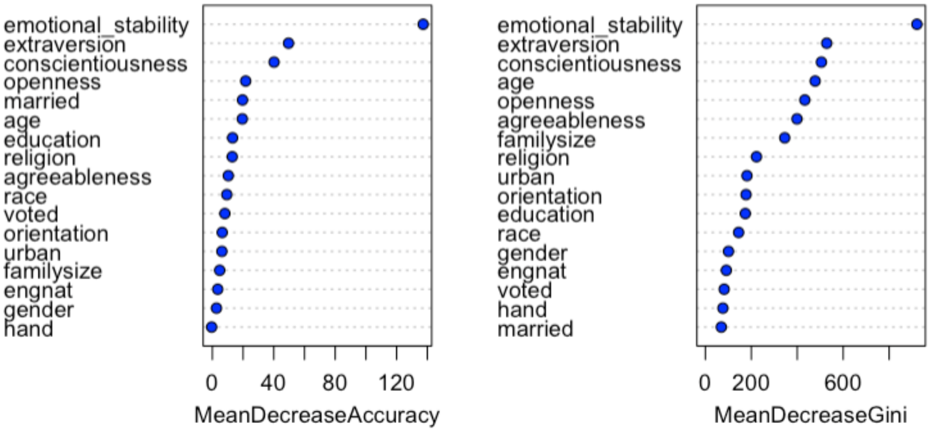
Further, the model achieves a misclassification rate of 30.7% and an accuracy rate of 69.3%. The corresponding confusion matrix is visualised below.

Prediction/Actual	**High**	**Low**
High	7547	3725
Low	2551	6611

II. Random forest

(a) Depression

We subsequently generated a random forest model for depression and anxiety. The random forest classifier utilises multiple individual decision trees to compose an uncorrelated series of trees that are expected to outperform any individual tree, providing high accuracy (Yiu, 2019). Another advantage of the random forest technique is that it provides a clear indication of the relative importance of each predictor (Donges, 2021). This is important for our aim, as depression and anxiety are shown to develop from joint effects of more than one cause (Gray and Bjorklund, 2018).

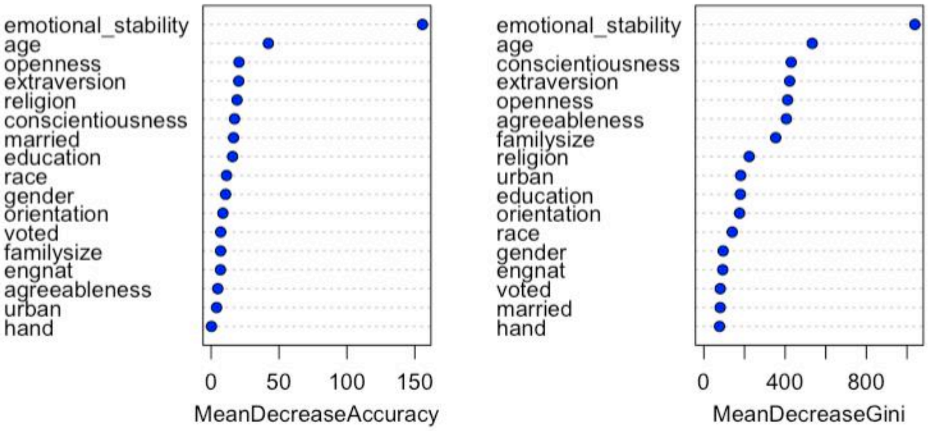


The relative importance of each predictor can be assessed using MeanDecreaseAccuracy and MeanDecreaseGini, as visualised above (Oliveira, 2017). The MeanDecreaseAccurracy plot indicates that the omission of emotional stability, extraversion and conscientiousness from the training set is expected to produce the greatest loss in prediction performance. Consistent with the decision tree, the random forest affirms that emotional stability is by far the most significant predictor of depression scores. Further, the model achieves a misclassification rate of 29.9% and an accuracy rate of 70.1% when tested using the testing dataset. As expected from random forest classifiers, the model is more accurate and demonstrates an incremental improvement over the 68% accuracy obtained by the decision tree. The corresponding confusion matrix is visualised below.

Prediction/Actual	**High**	**Low**
High	7236	3153
Low	2957	7088

(b) Anxiety

The random forest for anxiety demonstrated that emotional stability, age and openness are the strongest predictors of anxiety scores. This supports the decision tree's inclusion of emotional stability and introduces the variables of age and openness as novel predictors. Further, the model attained a misclassification rate of 29.4% and an accuracy rate of 70.6% when tested using the testing dataset. Like the depression models, the random forest for anxiety achieved an incremental improvement over the 69% accuracy rate of the decision tree. The corresponding confusion matrix is visualised below.



Prediction/Actual	**High**	**Low**
High	7337	3254
Low	2761	7091

III. Logistic regression

(a) Depression

Finally, we constructed a logistic regression model for depression and anxiety. Logistic regression estimates the probability an observation belongs to a categorical class via log-odds (Provost and Fawcett, 2013). In comparison to decision trees which bisect predictor space into increasingly smaller regions, logistic regression applies a single line which divides the space into two (Cheesinglee, 2016). Further, logistic regression models yield coefficients that are advantageous as we aim to ascertain the effect of individual predictors on depression and anxiety. This technique is appropriate given that we are interested in a binary outcome variable (high/low scores in depression and anxiety), the observations are all sampled independently from each other and the sample size is sufficiently large (Leung, 2021).

The model was first constructed using all predictor variables and was subsequently improved via stepwise, backward elimination of insignificant variables one by one. The final fitted model contains 7 predictors which are all significant at the p < .05 level standard in psychological research, as follows:

depression_label ~ education + married + extraversion + agreeableness + emotional_stability + conscientiousness + openness

	Estimated Coefficient	**Standard Error**
Intercept	-4.469691	0.154031
Education	0.121332	0.027995
Married	0.201185	0.057262
Extraversion	0.125293	0.007747
Agreeableness	0.022164	0.009667
Emotional stability	0.272116	0.008962
Conscientiousness	0.102879	0.008267
Openness	0.031680	0.009126
Null deviance: 13863 on 9999 degrees of freedom Residual deviance: 11379 on 9992 degrees of freedom AIC: 11395		

The fitted model indicates that all 7 predictor variables share a positive relationship with low scores in depression and a negative relationship with high scores in depression. The magnitudes of the coefficients reveal that emotional stability has the highest effect on depression scores while openness has the lowest. Interestingly, educational attainment and marital status also predict depression scores, whereby individuals with lower educational attainment and individuals who were never married are more likely to score high on depression. Interpreted together with the decision tree and random forest models, the logistic regression affirms that personality variables exert the strongest effect, but demographic variables also demonstrate statistically significant effects. Indicatively, the model illustrates that the variables **urban**, **gender**, **engnat**, **age**, **hand**, **religion**, **orientation, race**, **voted**, and **familysize** do not have a statistically significant effect on depression scores. This suggests that an individual's area of residence as a child, gender, affinity with English, age, hand of writing, religion, orientation, race, recent voting activity and family size are not significant predictors of depression scores. Low standard errors for all predictors suggest that the model generates reliable estimates.

We tested the performance of the model using the testing dataset that generated a misclassification rate of 29.6% and an accuracy rate of 70.4%, rendering the logistic regression model as the most accurate of the three classifiers. The corresponding confusion matrix is illustrated below.

Prediction/Actual	**High**	**Low**
High	7346	3204
Low	2847	7037

A potential limitation of logistic regression is that it assumes the absence of multicollinearity, whereby predictor variables are highly correlated (Leung, 2021). To account for this, we generated the variance inflation factor (VIF) to assess the ratio of model variance to variance of a model with a single predictor. The corresponding values below demonstrate that no VIF value exceeds 5, confirming that multicollinearity is minimal.

Variables	**VIF**
Education	1.048514
Married	1.032545
Extraversion	1.046980
Agreeableness	1.040901
Emotional stability	1.052795
Conscientiousness	1.057320
Openness	1.085351

(b) Anxiety

We applied the same method to predict anxiety scores. The logistic regression model fitted for anxiety contains 11 predictor variables significant at the p < .05 level, as follows:

anxiety_label ~ education + gender + age + religion + orientation + familysize + extraversion + agreeableness + emotional_stability + conscientiousness + openness

	Estimated Coefficient	**Standard Error**
Intercept	-2.833150	0.177317
Education	0.087408	0.032098
Gender	-0.138803	0.055591

Age	0.046592	0.003633
Religion	-0.048771	0.007532
Orientation	-0.086295	0.017630
Family size	-0.026755	0.011788
Extraversion	0.045519	0.008035
Agreeableness	-0.051335	0.009791
Emotional stability	0.335252	0.009584
Conscientiousness	0.022544	0.008182
Openness	0.027934	0.009344
Null deviance: 13863 on 9999 degrees of freedom Residual deviance: 11211 on 9988 degrees of freedom AIC: 11235		

The fitted model indicates that an increase in variables **gender**, **religion**, **orientation**, **familysize** and **agreeableness** is associated with an increase in anxiety scores. Conversely, an increase in variables **education**, **age**, **extraversion**, **emotional_stability, conscientiousness** and **openness** is associated with a decrease in anxiety scores. The magnitudes of the coefficients demonstrate that individuals who are emotionally unstable, low in openness and who identify as a female or other are most likely to score high on anxiety. Further, the model considers the variables **urban**, **engnat**, **hand**, **race**, **voted** and **married** as insignificant predictors. This suggests that an individual's area of residence as a child, affinity with English, hand of writing, race, recent voting activity and marital status are not significant predictors of anxiety scores. Like the depression model, the model indicates low standard errors for all predictors.

The anxiety model also differs from the depression model in several important ways, including the inclusion of demographic predictors of gender, age, religion, orientation and family size established insignificant in the depression model. This suggests that while depression and anxiety share common predictors including personality, some variables may exert an effect on anxiety but not depression. A comprehensive comparison between depression and anxiety models is presented in detail in Section 4 (see 4. Conclusion); important theoretical and practical implications are discussed.

When tested on the testing dataset, the anxiety model achieved a misclassification rate of 29.2% and an accuracy rate of 71.8%, attaining the highest accuracy rate of all three anxiety models. The corresponding confusion matrix is illustrated below.

Prediction/Actual	**High**	**Low**
High	7499	3365
Low	2599	6971

Finally, test of multicollinearity confirms that the predictors share minimal correlation with each other, satisfying the assumption of the logistic regression technique. The VIF values for each predictor are presented below.

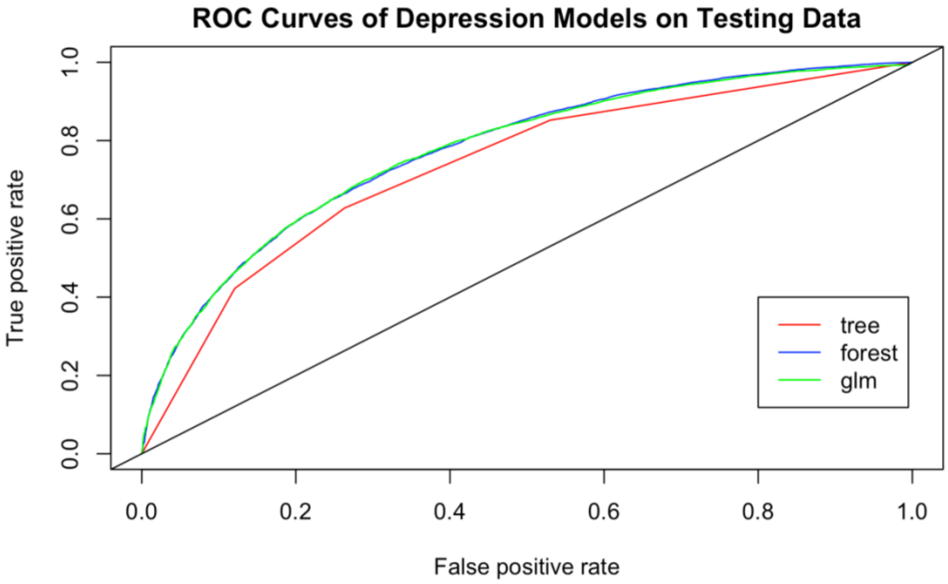
Variables	**VIF**
Education	1.344193
Gender	1.067800
Age	1.277223
Religion	1.264647
Orientation	1.018340
Family size	1.131144
Extraversion	1.115075
Agreeableness	1.098031
Emotional stability	1.159956
Conscientiousness	1.113234
Openness	1.152270

IV. Model evaluation and discussion

(a) Depression

The depression models were evaluated based on classification accuracy on testing data, Receiver Operating Characteristics (ROC) graph and their corresponding area under curve (AUC) visualised below. The ROC graph evaluates model performance by plotting a model's false positive rate against its true positive rate (Provost and Fawcett, 2013).

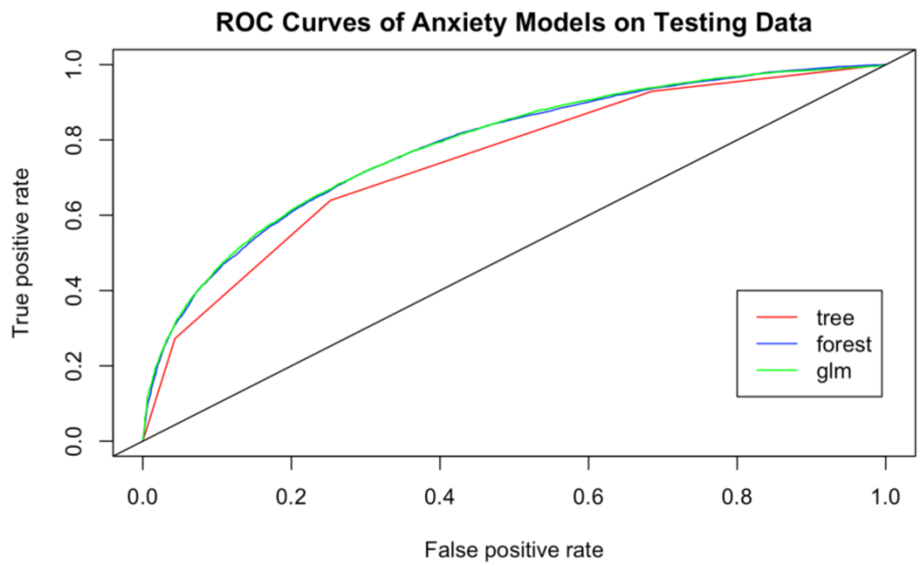
Of the three models constructed, the logistic regression model demonstrated the highest classification accuracy (70.4%) and the second greatest AUC (0.772616). The random forest demonstrated the second highest classification accuracy (70.1%) and greatest AUC (0.7741209). The decision tree demonstrated the lowest classification accuracy (68.2%) and corresponding AUV (0.7328779). However, given that the accuracy range between the most accurate and least accurate models is only 2.2%, all models have approximately comparable performance. Nonetheless, the logistic regression model was most predictive of depression scores.



Depression	**Accuracy**	**AUC**
Decision Tree	68.2%	0.7328779
Random Forest	70.1%	0.7741209
Logistic Regression	70.4%	0.772616

(b) Anxiety

The same method was used to evaluate the performance of the anxiety models and is visualised below. The regression model achieved the highest classification accuracy on testing data (71.8%) and the greatest AUC (0.7841349). This is followed by the random forest, achieving 70.6% accuracy and 0.7818934 AUC. The decision tree was the least accurate model (69.3%), which is reinforced by its relatively low AUC value (0.7443952). The performance of the anxiety models demonstrate similarity with the depression models, with the logistic regression being the most predictive of anxiety scores, followed by random forest and decision tree models.



Anxiety	**Accuracy**	**AUC**
Decision Tree	69.3%	0.7443952
Random Forest	70.6%	0.7818934
Logistic Regression	71.8%	0.7841349

4. Conclusion

Our project implemented three classification techniques (decision tree, random forest and logistic regression) to predict depression and anxiety scores on the DASS-42 using personality and demographic variables within a dataset of 30434 participants. Logistic regression was the most predictive model, followed by random forest and decision tree for both depression and anxiety.

A comparison between the decision tree models demonstrates that depression scores are predicted by emotional stability and extraversion, while anxiety scores are predicted by emotional stability alone. Further, the random forest models indicate that emotional stability, extraversion and conscientiousness are the most significant predictors of depression, while emotional stability, age and openness are the most significant predictors of anxiety. The logistic regression model reveals that emotional stability, marital status and extraversion are most significant predictors of depression, while emotional stability, gender and openness are the most significant predictors of anxiety.

Overall, the models demonstrate the relative importance of personality variables over demographic variables in predicting depression and anxiety. The logistic regression models, for instance, included all 5 personality variables within its 7 predictors for depression and 11 predictors for anxiety. From a theoretical perspective, this observation affirms the predictive validity of the Five-Factor model of personality and the TIPI more specifically, for depression and anxiety outcomes. Researchers are encouraged to replicate the findings of this report, and further, examine the validity of the Five-Factor model for other mental illnesses. From a practical perspective, the models affirm that clinicians and policymakers will find disproportionate success in modelling depression and anxiety using personality predictors than demographic ones. Specifically, emotional stability is an overwhelmingly strong predictor of both depression and anxiety across all models tested.

Further, gender, age, religion, orientation and family size appear important predictors of anxiety but not depression. This suggests that while depression and anxiety may share common predictors, anxiety appears to be influenced by a greater range of demographic variables beyond the personality variables that predict depression. One implication is that anxiety may be more susceptible to the influence of environmental factors than depression, which is strongly predicted by personality traits genetically influenced between 40-50% (Bouchard, 2004). Of course, associating personality with genetics independent from environment is far too reductionist given the nature of gene-environment interaction (Dick, 2011). However, the differential effect of personality and demographic predictors on depression and anxiety can be usefully explored in future research. The second implication is that variables predictive of anxiety but not depression can be used to delineate between depression and anxiety disorders to provide more accurate diagnoses and effective treatments. This is particularly useful given that depression and anxiety are often comorbid and closely associated (Beuke, Fischer and McDowall, 2003).

From a data analytics perspective, depression and anxiety may be most accurately predicted by logistic regression techniques, followed by random forests and decision trees. However, the difference in performance is not significant and should be further investigated. Nonetheless, given the complex nature of psychopathology in both their causes and symptoms, the relative high accuracy of the models demonstrate that data analytic principles can and should be integrated within the psychological sciences to enhance both methodology and practice.

However, there are several limitations to our report. First, the classification models predict a binary, categorical outcome variable (high or low) which may not accurately reflect the continuous distribution of DASS-42 scores. Therefore, the models make broad generalisations regarding whether one experiences depression or anxiety but cannot delineate the severity. Future research can delineate the scores more precisely to ascertain severity of illness.

Second, both the DASS-42 and TIPI are only one measurement of depression, anxiety and personality which may present issues of validity, including construct and external validity. The tests are psychometric measures that may not validly reflect the latent construct of interest (e.g., personality) nor be generalisable to populations beyond the samples observed. For instance, the dataset is saturated with female participants from Asia and may not be representative of other populations and cultures which are shown to differ in emotional expression (Kitayama, Mesquita and Karasawa, 2006). Third, the modelled effects can only be interpreted as correlational and may be subject to confounding variables. Moreover, the models may be limited by reverse causality whereby the direction of effects remain uncertain.

An important extension is to integrate a cost-benefit matrix to identify the optimal cut-off probabilities of the models. Estimated costs and benefits of false negatives and false positives of misdiagnosis vary significantly and should therefore be incorporated on a case-by-case basis by clinicians and governments that have access to local data.

Bibliography

Beuke, C.J., Fischer, R. and McDowall, J. (2003). Anxiety and depression: Why and how to measure their separate effects. *Clinical Psychology Review*, [online] 23(6), pp.831–848. Available at: https://psycnet.apa.org/record/2003-09392-004 [Accessed 10 Feb. 2022].

Bouchard, T.J., Jr. (2004). Genetic Influence on Human Psychological Traits. *Current Directions in Psychological Science*, [online] 13(4), pp.148–151. Available at: https://psycnet.apa.org/record/2004-16045-005 [Accessed 10 Feb. 2022].

Burešová, I., Jelínek, M., Dosedlová, J. and Klimusová, H. (2020). Predictors of Mental Health in Adolescence: The Role of Personality, Dispositional Optimism, and Social Support. *SAGE Open*, [online] 10(2), p.215824402091796. Available at: https://journals.sagepub.com/doi/10.1177/2158244020917963 [Accessed 10 Feb. 2022].

Cheesinglee (2016). *Logistic Regression versus Decision Trees*. [online] BigML Blog. Available at: https://blog.bigml.com/2016/09/28/logistic-regression-versus-decision-trees/ [Accessed 10 Feb. 2022].

Dick, D.M. (2011). Gene-Environment Interaction in Psychological Traits and Disorders. *Annual Review of Clinical Psychology*, [online] 7(1), pp.383–409. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3647367/ [Accessed 10 Feb. 2022].

Donges, N. (2021). *A Complete Guide to the Random Forest Algorithm*. [online] Built In. Available at: https://builtin.com/data-science/random-forest-algorithm [Accessed 10 Feb. 2022].

Gosling, S.D., Rentfrow, P.J. and Swann, W.B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, [online] 37(6), pp.504–528. Available at: https://psycnet.apa.org/record/2003-09807-003 [Accessed 10 Feb. 2022].

Gray, P. and Bjorklund, D.F. (2018). *Psychology*. 8th ed. United States of America: Worth Publishers.

Johns Hopkins Medicine (n.d.). *Mental Health Disorder Statistics*. [online] www.hopkinsmedicine.org. Available at: https://www.hopkinsmedicine.org/health/wellness-and-prevention/mental-health-disorder-statistics#:~:text=An%20estimated%2026%25%20of%20Americans [Accessed 10 Feb. 2022].

Kitayama, S., Mesquita, B. and Karasawa, M. (2006). Cultural affordances and emotional experience: Socially engaging and disengaging emotions in Japan and the United States. *Journal of Personality and Social Psychology*, [online] 91(5), pp.890–903. Available at: https://psycnet.apa.org/record/2006-20034-007 [Accessed 10 Feb. 2022].

Leung, K. (2021). *Assumptions of Logistic Regression, Clearly Explained*. [online] Medium. Available at: https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290 [Accessed 10 Feb. 2022].

Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P. and Ansseau, M. (2003). Socioeconomic Inequalities in Depression: A Meta-Analysis. *American Journal of Epidemiology*, [online] 157(2), pp.98–112. Available at: https://academic.oup.com/aje/article/157/2/98/90059?login=false [Accessed 10 Feb. 2022].

Lovibond, S.H. and Lovibond, P.F. (1995). *Manual for the depression anxiety stress scales*. 2nd ed. Sydney, N.S.W.: Psychology Foundation Of Australia.

NHS Digital (2009). *Adult Psychiatric Morbidity in England - 2007, Results of a household survey - NHS Digital*. [online] NHS Digital. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/adult-psychiatric-morbidity-survey/adult-psychiatric-morbidity-in-england-2007-results-of-a-household-survey [Accessed 10 Feb. 2022].

Oliveira, S.P. de (2017). *A very basic introduction to Random Forests using R*. [online] Oxford Protein Informatics Group. Available at: https://www.blopig.com/blog/2017/04/a-very-basic-introduction-to-random-forests-using-r/ [Accessed 10 Feb. 2022].

Provost, F. and Fawcett, T. (2013). *Data Science for Business*. 1st ed. United States of America: O'Reilly Media.

Rosenfield, S. and Mouzon, D. (2013). Gender and Mental Health. In: C.S. Aneshensel, J.C. Phelan and A. Bierman, eds., *Handbook of the Sociology of Mental Health*. Netherlands: Springer.

Yiu, T. (2019). *Understanding Random Forest*. [online] Medium. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [Accessed 10 Feb. 2022].