Detection of Mental Health Condition (Depression,Stress & Anxiety) using different machine learningmodels

Drashti Raval, Viraj Parkhe *Department of Technology Savitribai Phule Pune University***Pune, India**

***Abstract*— This research paper presents an in-depth exploration of data-driven mental health assessment, employing data science methodologies to develop predictive models for identifying depression, stress, and anxiety. Through meticulous data preprocessing, exploratory data analysis, and feature engineering, we transformed raw questionnaire data into a structured format conducive to analysis. Our study involved the evaluation of various machine learning models, including Random Forest, Decision Tree, Support Vector Machines, Logistic Regression, and k-Nearest Neighbors, through rigorous cross-validation. The results showcase the potential of these models for accurate mental health prediction. Additionally, ethical considerations surrounding privacy, consent, and potential misclassification are highlighted, emphasizing the importance of responsible model deployment in sensitive domains. This research contributes to the field by showcasing the power of data science in mental health assessment while acknowledging the ethical complexities associated with its implementation.**

Keywords*— *Data Science, Machine Learning, Predictive modeling, Depression, Stress, Anxiety, Logistic regression model, K-nearest neighbor classifier model, Support vector machine classifier, Decision tree model Random forest model, Mean accuracy, Precision, Recall, F-1 Score, Model evaluation, Cross validation, K-fold cross validation

INTRODUCTION

Mental health assessment stands as a pivotal aspect of healthcare, offering insights into individuals' psychological well-being and facilitating timely interventions. With the proliferation of digital platforms and the increasing demand for accessible and accurate assessment methods, traditional approaches are being reimagined. This research paper embarks on a journey to harness the potential of data science and machine learning techniques in redefining mental health assessment. The prevalent challenges associated with conventional methods, marked by subjectivity and resource constraints, underscore the pressing need for innovative, data-driven approaches that can enhance accuracy and accessibility. Use the enter key to start a new paragraph. The appropriate spacing and indent are automatically applied.

In this context, our research endeavors to develop predictive models capable of identifying key mental health conditions—depression, stress, and anxiety—based on a multidimensional dataset collected through Google Forms. Leveraging a diverse set of features encompassing demographic, behavioral, and psychological aspects, our study aims to construct robust classifiers that can predict mental health conditions with high accuracy. By employing advanced techniques such as feature engineering, exploratory data analysis, and model evaluation, we seek to enhance the quality and efficiency of mental health assessment.

The dataset, meticulously curated and cleaned, forms the foundation of our investigation. Through meticulous preprocessing, we've transformed raw questionnaire responses into structured data amenable to analysis. Our research methodology entails the utilization of various machine learning models, including Random Forest, Decision Tree, Support Vector Machines, Logistic Regression, and k-Nearest Neighbors. By rigorously subjecting these models to Cross-Validation, we endeavor to identify the most effective approach for accurate classification.

Ethical considerations hold paramount significance in this pursuit. While the promise of data-driven assessment is evident, it brings forth ethical complexities related to privacy, consent, and the potential for misclassification. Thus, alongside model development, we emphasize the need for responsible deployment and vigilant monitoring of these models in real-world applications.

In essence, this research contributes to the evolving landscape of mental health assessment by showcasing the power of data science to transform healthcare practices. Through a multidisciplinary approach that combines data analysis, machine learning, and ethical awareness, our study underscores the potential for innovation in an area that profoundly impacts individual well-being.

1. METHODOLOGY

1.1 DATA SOURCES AND COLLECTION:

The primary source of data for this study is a carefully designed survey administered through a Google Form. This survey is specifically tailored to capture information related to symptoms, demographic characteristics, and lifestyle patterns of participants. By directly collecting responses from individuals, we ensure the authenticity and relevance of the data used for model development and evaluation.

Description of the features that were added in thequestionnaire made for the data collection is as follows:

1. Gender: Categorical variable indicating the gender of the participant (e.g., Male, Female, Non-binary).
2. Age: Continuous variable representing the age of the participant in years.

3. Education: Categorical variable indicating the highest level of education attained by the participant (e.g., High School, Bachelor's Degree, Master's Degree).
4. Medical Condition: Binary variable indicating whether the participant has a pre-existing medical condition (Yes/No).
5. Traumatic Experience: Binary variable indicating whether the participant has experienced a traumatic event (Yes/No).
6. Smoking/Drinking Frequency: Numeric variable representing the frequency of smoking or drinking on a scale from 1 to 10.
7. Childhood Neglect/Abuse: Binary variable indicating whether the participant experienced neglect or abuse during childhood (Yes/No).
8. Personality: Categorical variable representing the personality type of the participant (e.g., Introverted, Extroverted, Ambivert).
9. Depression Symptoms: Numeric variable indicating the severity of depression symptoms experienced by the participant.
10. Depression Symptom Frequency Scale: Numeric variable indicating the frequency of specific depression symptoms.
11. Feeling of Worthlessness Frequency: Numeric variable indicating the frequency of feelings of worthlessness.
12. Excessive Guilt Frequency: Numeric variable indicating the frequency of experiencing excessive guilt.
13. Persistent Sadness/Low Mood Frequency: Numeric variable indicating the frequency of persistent sadness or low mood.
14. Loss of Interest Scale: Numeric variable indicating the level of interest or engagement in activities.
15. Stress Cause: Categorical variable indicating the primary cause of stress for the participant.
16. Stress Symptom Frequency Scale: Numeric variable indicating the frequency of stress-related symptoms.
17. Stress-Related Physical Symptom: Binary variable indicating whether the participant experiences physical symptoms due to stress (Yes/No).
18. Stress-Induced Irritability Scale: Numeric variable indicating the level of irritability induced by stress.
19. Anxiety Symptom Check: Binary variable indicating whether the participant experiences symptoms of anxiety (Yes/No).
20. Anxiety Frequency Scale: Numeric variable indicating the frequency of anxiety symptoms.
21. Anxiety Avoidance Scale: Numeric variable indicating the extent of avoidance behaviors due to anxiety.
22. Sleep Pattern Changes: Binary variable indicating whether the participant experiences changes in sleep patterns (Yes/No).
23. Mental Health Condition: Categorical variable indicating the reported mental health condition (e.g., Depression, Anxiety, Stress).

1.2. EXPLORATORY DATA ANALYSIS:

. In this phase of the project, we leveraged a range of techniques and commands to gain a comprehensive understanding of our dataset.

1.2.1 Data summary and Information

We initiated our EDA by obtaining a quick overview of the dataset using commands such as head() and tail(). These commands provided us with a glimpse of the initial and final records, respectively, giving us a snapshot of the data's format and content.

To delve deeper into the dataset's structure, we utilized the info() command. This command not only informed us about the data types and non-null counts of each column but also alerted us to potential missing values.

1.2.2 Descriptive Statistics

To gain insights into the central tendencies and dispersions of our data, we employed the **describe()** command. This command provided summary statistics such as mean, standard deviation, minimum, maximum, and quartiles for the numeric columns. These statistics aided us in understanding the distribution of our variables and identifying potential outliers.

1.2.3 Understanding the Relationship between Mental Health and Demographic Factors:

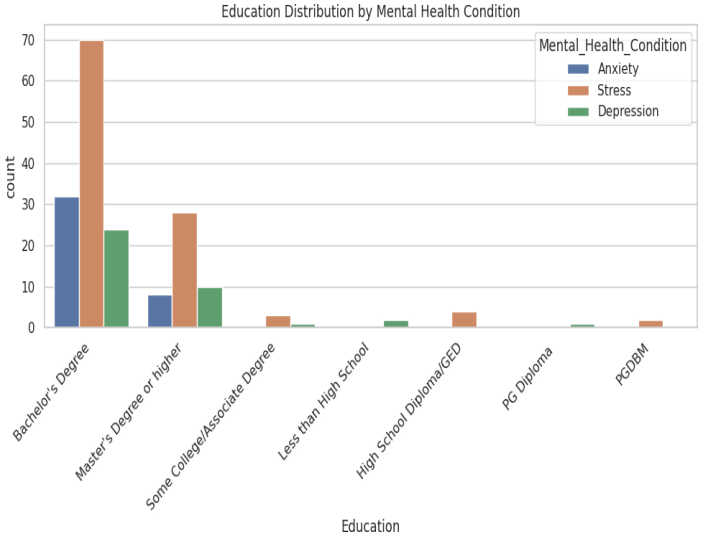

Fig 1.2.1: Education Distribution by Mental Health Condition

The above figurewe examine the education level distribution across mental health conditions. It is evident that individuals with Bachelor's degree and Master's degree or higher have higher occurrences of mental health conditions, indicating a possible correlation between education and mental well-being. These visualizations provide valuable initial insights, prompting us to explore these patterns further in subsequent analyses.

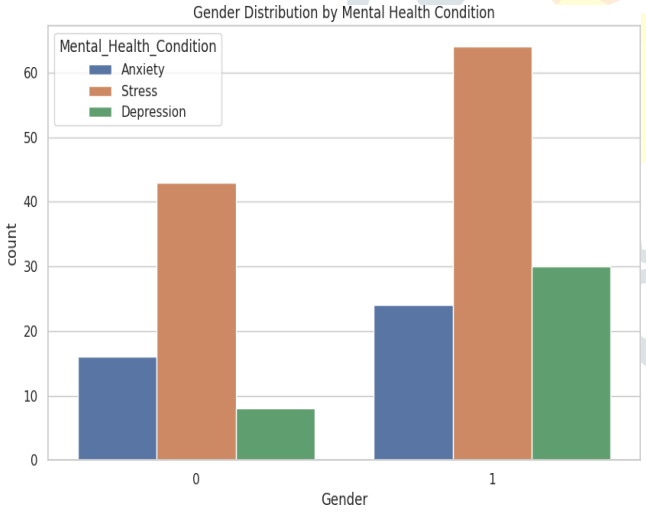

Fig 1.2.2: Gender Distribution by Mental Health Condition

In the above figure the gender distribution by mental health condition demonstrates potential gender-based differences (0 – Female, 1- Male) in the prevalence of mental health conditions. This observation suggests that gender could be a contributing factor to the manifestation of certain mental health conditions.

From the graph we can conclude that there are higher chances of occurrence of some mental health condition in the males as compared to females.

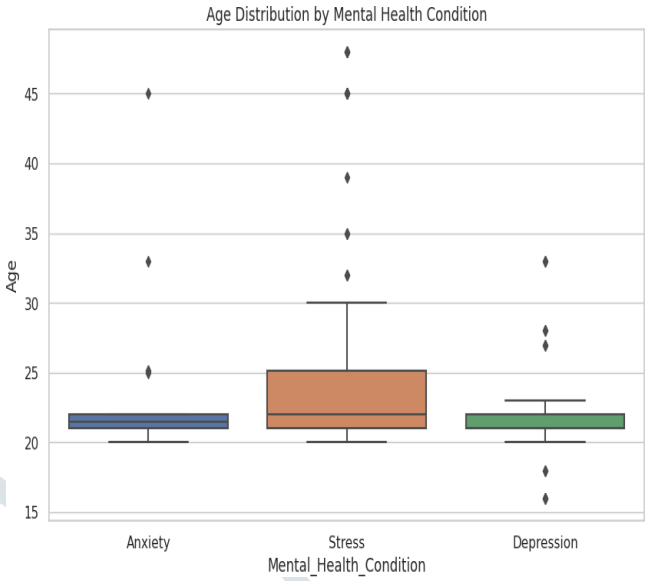

Fig 1.2.3: Age Distribution by Mental Health Condition

An insightful analysis of the age distribution across different mental health conditions was conducted using a box plot visualization, as demonstrated in Figure 1. The box plot presents a clear depiction of the central tendency, variability, and potential outliers within each mental health category. Notably, the box plot reveals interesting patterns in the age distribution across the three mental health conditions. While the median age for individuals with Anxiety and Depression seems relatively consistent, there is a distinctive characteristic for the Stress category. A larger interquartile range and a visibly extended upper whisker in the Stress box plot indicate greater variability and the presence of potential outliers in the age distribution for this condition. This observation suggests that stress-related symptoms may affect a wider range of age groups, potentially spanning both younger and older individuals. The presented box plot thus offers a valuable visual representation of the age distribution and its variability within different mental health conditions, underscoring the importance of further investigation into the factors driving these patterns.

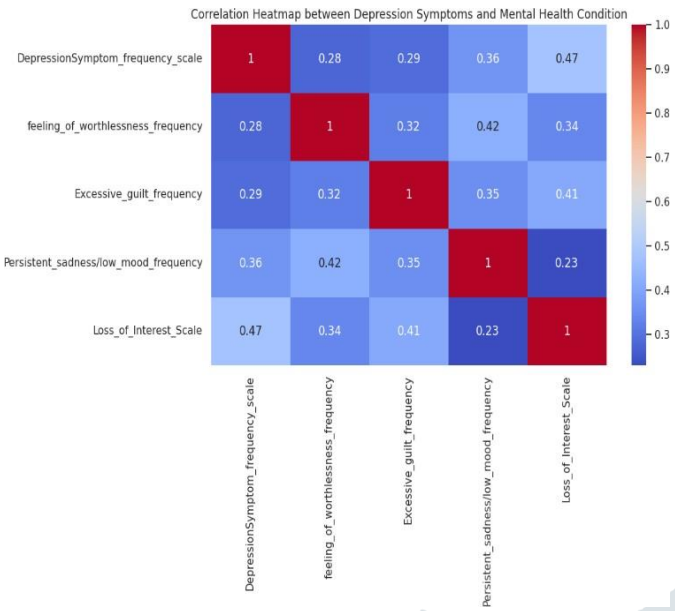

Fig 1.2.4: Correlation Heatmap between Depression Symptoms and Mental_Health_Condition

The above figure shows heatmap of the correlation between the symptoms of depression and mental health condition through which the individual is going through.

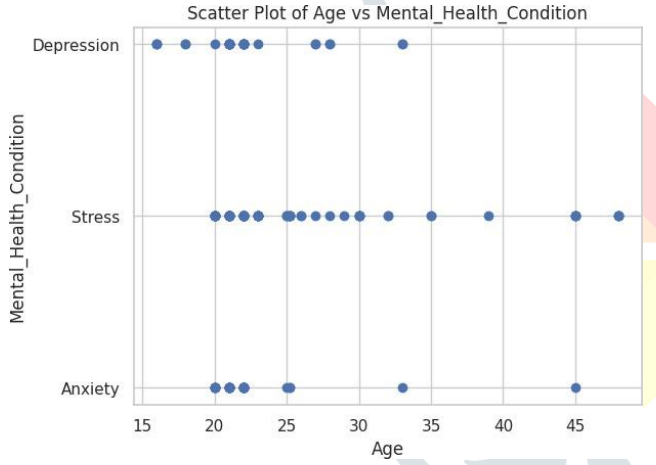

Fig: 1.2.5: Scatter plot of Age Vs Mental_Health_Condition

The above figure shows what age shows which mental health condition. From the figure we can see that stress is the most frequent mental health condition found in the data and the age group of 20 to 35 show most of the mental heath conditions.

1.3 DATA PRE-PROCESSING:

1.3.1 Handling missing values:

We also utilized the **isnull()** command to identify missing values within the dataset.

We found that there were no missing values in the data. This step was essential for addressing gaps in the data and making informed decisions about handling missing values, such as imputation or exclusion.

1.3.2 Columns renaming for enhanced clarity:

Our data collection process involved gathering responses through a Google Form, resulting in column names that corresponded to the survey questions. To enhance the clarity and readability of our dataset, we undertook a strategic column renaming process. This step was pivotal in ensuring that our dataset's attributes align with the research context and facilitate effective communication of our findings.

The renaming process was conducted using the rename() function, allowing us to map the original column names to more descriptive titles. The renaming of columns was guided by the principle of conciseness while preserving the core information captured by the questions. This endeavour empowers us to engage with the dataset more effectively and fosters clarity when presenting our insights and analysis.

1.3.3 Standardization of categorical responses:

As part of our data preparation phase, we recognized the need to standardize certain categorical responses within specific columns. This standardization was essential to consolidate various synonymous variations and streamline our data for consistency and accuracy. We employed a systematic approach using Python code to achieve this standardization. For instance, consider the column "Anxiety Symptom Check." We observed that respondents used multiple variations such as "None," "No," "na," "Nothing," and "-," among others, to indicate absence of anxiety symptoms. To ensure uniformity, we created a variation mapping dictionary. This dictionary associated all synonymous variations with the standardized response "None." Through an iterative process, we replaced these variations with the standardized option, enhancing the homogeneity and interpretability of the data.

Similarly, the column "Stress Cause" exhibited similar diversities in response terms like "N/A," "No," "None," "Nothing," and "-." By utilizing another variation mapping dictionary, we systematically replaced these variations with the standardized response "None."

1.3.4 Data cleanup and column removal:

In addition to standardization, we performed data cleanup by removing certain columns that did not contribute to our analytical objectives. Columns such as "Timestamp" and "Email" were identified for removal since they held no substantive relevance to our analysis.

The cleaned dataset, referred to as "data_cleaned," was generated by dropping the specified columns using the drop() function. This streamlined version of the dataset is now devoid of redundant or irrelevant information, enhancing its clarity and relevance to our research.

As a supplementary step, we had the option to save the cleaned dataset to a new CSV file named "cleaned_dataset.csv." This safeguarded our cleaned data for subsequent analysis, ensuring that the data remained consistent and aligned with our data preparation efforts.

1.3.5 Encoding the categorical data:

Since the data was collected by the means of Google forms and questionnaire being the questionnaire for mental health assessment, most of the data collected was in the form of categorical data. There was a need to convert this categorical data into numerical data for further analysis, which is why we encoded the categorical data into numerical data.

The encoding techniques used for the conversion are as mentioned below:

1. Label Encoding:

   Label encoding is a technique used to convert categorical variables into numerical values. It assigns a unique integer to each category in a categorical variable. This technique is particularly useful for variables that have an inherent order or ranking.

   How Label Encoding Works:

   - Assigning Labels: Each unique category in a categorical variable is assigned a unique integer label, starting from 0. The order of assignment is typically based on the alphabetical order of the categories.
   - Conversion: The categorical values in the variable are replaced with their corresponding integer labels.

2. Ordinal Encoding:

   Ordinal encoding is a data transformation technique used to convert categorical variables with ordinal relationships into numerical values, ensuring that the order among categories is preserved. It is a valuable tool in data preprocessing for machine learning tasks, helping algorithms understand and utilize the inherent order within the data while minimizing computational complexity.

   The process of ordinal encoding involves two main steps:

   - Establishing Order: First, an order is defined among the categorical variables. This order can be based on factors such as hierarchy, ranking, or levels of a specific characteristic within the data. It is essential to ensure that the order assigned accurately reflects the relationship between categories.
   - Assigning Integer Values: Once the order is established, each category is assigned a corresponding integer value based on its position in the defined order. The assignment of integer values is sequential, with lower integers assigned to categories with lower positions in the order and higher integers assigned to categories with higher positions.

3. Custom Encoding:

   Custom encoding, also known as manual encoding, refers to the process of transforming categorical variables into numerical values using a customized mapping that you define based on the specific characteristics of your data. Unlike standard encoding techniques like label encoding or one-hot encoding, which follow predefined rules, custom encoding allows you to tailor the transformation to the unique requirements of your dataset.

   The process of custom encoding involves the following steps:

   - Mapping Categories: Identify the categorical variable you want to encode and examine its unique categories. For each category, decide on the corresponding numerical value you want to assign. This mapping is created based on your domain knowledge, data understanding, or specific objectives of your analysis.
   - Create Mapping Dictionary: Create a dictionary or mapping table that associates each category with its corresponding numerical value. This dictionary serves as a reference for the encoding process.
   - Apply Custom Encoding: Replace the categorical values in your dataset with the numerical values according to the mapping dictionary you created. This process converts the categorical variable into a numerical format that can be used for analysis or modelling.

1.3.6 Identifying columns having outliers:

To ensure the quality and reliability of our dataset, we conducted an analysis to identify potential outliers within the variables. Outliers, which are data points that deviate significantly from the overall distribution, can have a considerable impact on statistical analyses and model performance. Identifying and addressing outliers is crucial to maintain the integrity of our findings.

To accomplish this, we employed the z-score method, which measures the number of standard deviations a data point is from the mean. We set a threshold for outliers at 3 standard deviations from the mean. Any data points exceeding this threshold were flagged as potential outliers within the dataset.

Upon applying this method, we identified the following columns that contained potential outliers:

- Age
- feeling_of_worthlessness_frequency
- medical_condition
- stress_related_physical_symptom

1.3.7 Outlier replacement using mean:

As part of our data preprocessing efforts, we recognized the importance of addressing potential outliers that could impact the quality of our analyses and model building. Outliers, which are data points that deviate significantly from the general trend, can disproportionately affect statistical metrics and model performance.

To address this, we focused on specific columns where outliers were likely to have the greatest influence. These columns, namely 'Age', 'feeling_of_worthlessness_frequency', 'medical_condition', and 'stress_related_physical_symptom', were identified based on their significance to our research objectives.

We adopted a method to replace outliers with the mean of the respective column. This approach ensures that extreme values are moderated without distorting the overall distribution of the data. The process involved the following steps:

- Calculation of Mean and Standard Deviation: For each column of interest, we calculated the mean and standard deviation. These metrics are essential for establishing a threshold beyond which data points are considered outliers.
- Threshold Definition: We defined a threshold for outliers, which was set at 3 standard deviations from the mean. Data points exceeding this threshold were deemed potential outliers warranting replacement.
- Outlier Replacement: Using the calculated mean and standard deviation, we replaced outlier values with the mean of the column. This adjustment mitigates the impact of outliers on subsequent analyses and modelling.

This approach ensures that extreme values do not disproportionately influence our findings while maintaining the integrity of our data. By replacing outliers with the mean, we aim to strike a balance between addressing data anomalies and preserving the genuine characteristics of the dataset.

1.4 FEATURE IMPORTANCE USING L1 REGULARIZATION AND MODEL LIKE RANDOM FOREST:

1.4.1 USING RANDOM FOREST:

To optimize the predictive power of our models, we embarked on the critical process of feature selection. The objective was to identify the most influential attributes within our dataset that contribute significantly to predicting mental health conditions. Feature selection is crucial for mitigating the impact of irrelevant or redundant variables, enhancing model performance, and reducing overfitting.

We began by structuring our dataset, segregating the features and the target variable. After splitting the data into training and testing sets, we fit the Random Forest model to the training data. This model learned the intricate patterns within the features and their influence on predicting the mental health condition outcomes.

A significant advantage of the Random Forest model lies in its ability to provide insights into feature importance. By analyzing the accumulated decisions of multiple decision trees, we were able to extract the relative significance of each feature in predicting the target variable.

The feature importance scores were extracted from the model and presented graphically using a horizontal bar plot. Each feature's contribution was depicted as a bar, with longer bars indicating higher importance. The feature names were displayed along the y-axis, enabling an easy comparison of their relative impacts.

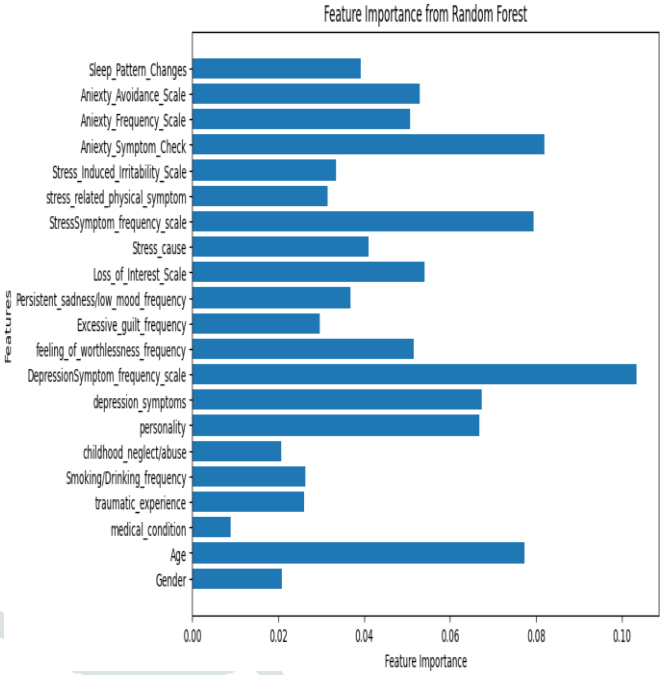

Fig: 1.4.1: Feature importance using random forest

As it is seen in the graph, the columns: “DepressionSymptom_frequency_scale”, “age”, Anxiety_Symptom_Check” and “StressSymptom_frequency_scale” are one of the most target variable affecting coulmns.

1.4.2 USING L1 REGULARIZATION:

Utilizing the Logistic Regression model with L1 regularization, we aimed to ascertain which features held significant predictive power for mental health conditions. The regularization process adjusted the coefficients associated with each feature, with some coefficients potentially shrinking to zero. Features with non-zero coefficients after regularization were deemed essential for the predictive task.

Upon fitting the Logistic Regression model to the training data, we identified the selected features by extracting the non-zero coefficients. These selected features collectively contribute to the model's predictive capabilities, ensuring that only the most relevant attributes influence the prediction of mental health conditions. The analysis yielded a set of selected features that demonstrate their importance in predicting mental health conditions. These features encompass aspects such as age, personality traits, symptom frequencies, and behavioural patterns. The selection process underscores the significance of these attributes in the context of mental health assessment.

Below are the features which were affecting the target variable the most:

['Age', 'traumatic_experience', 'Smoking/Drinking_frequency', 'personality', 'depression_symptoms', 'DepressionSymptom_frequency_scale', 'feeling_of_worthlessness_frequency', 'Excessive_guilt_frequency', 'Loss_of_Interest_Scale', 'Stress_cause', 'StressSymptom_frequency_scale',

'Stress_Induced_Irritability_Scale',
'Aniexty_Symptom_Check', 'Aniexty_Frequency_Scale',
'Aniexty_Avoidance_Scale', 'Sleep_Pattern_Changes']

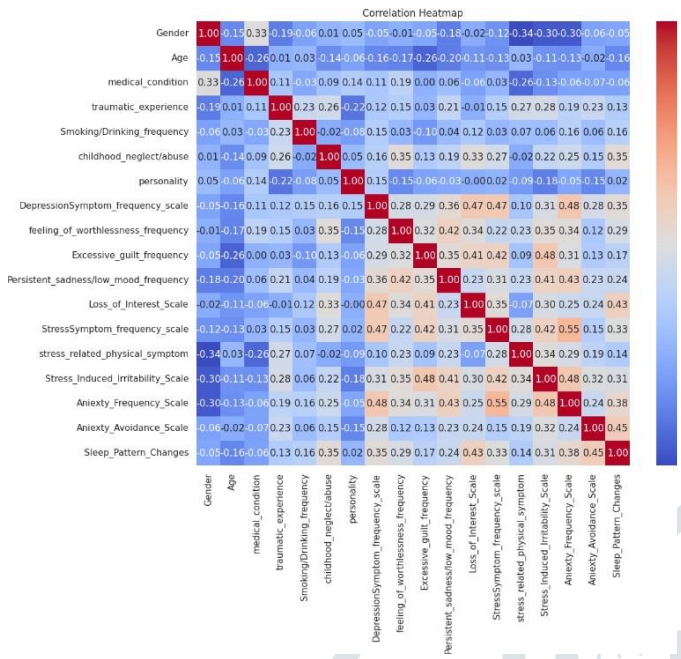

Fig 1.4.2: Correlation Heatmap

The above figure shows the correlation of different features.

1.5 MODEL FITTING AND EVALUATION:

After executing data preprocessing steps, including the replacement of outliers with the mean for specific columns and identifying the important features affecting the target variable, we proceeded to build predictive models using a range of machine learning algorithms using the above selected features. The goal was to assess the performance of these models in predicting mental health conditions based on the collected features.

1.5.1 Models that were fitted and evaluated were as follows:

- LOGISTICS REGRESSION MODEL:

We employed logistic regression, a fundamental classification technique, to predict mental health conditions. Logistic regression is well-suited for binary classification tasks and provides insights into the relationship between features and the target variable. By fitting the logistic regression model on our cleaned dataset, we aimed to leverage the relationships among various attributes to predict the likelihood of different mental health conditions.

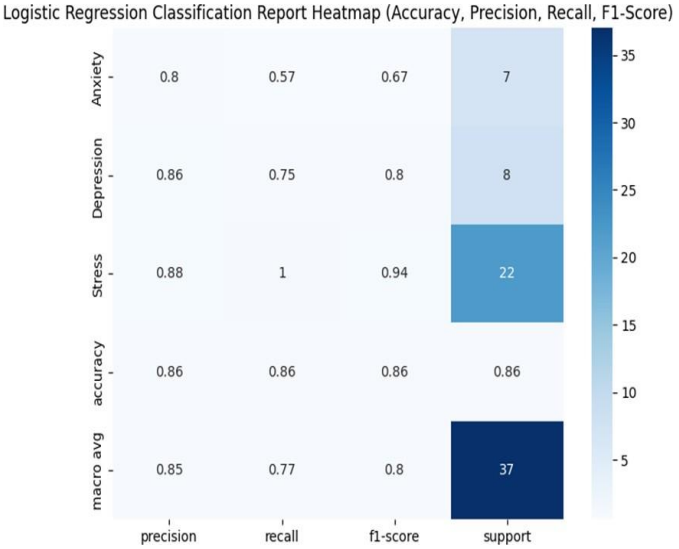

Fig 1.5.1: Logistic Regression Classification Report Heatmap (Accuracy, Precision, Recall and F1-Score)

- SUPPORT VECTOR MACHINE (SVM) CLASSIFIER:

The SVM classifier is another powerful algorithm we employed for predictive modelling. With its ability to capture complex relationships and create optimal hyperplane boundaries, SVM is particularly effective in handling non-linear separable data. Our SVM model was trained on the dataset to classify instances into distinct mental health conditions, leveraging features that were carefully prepared and standardized.

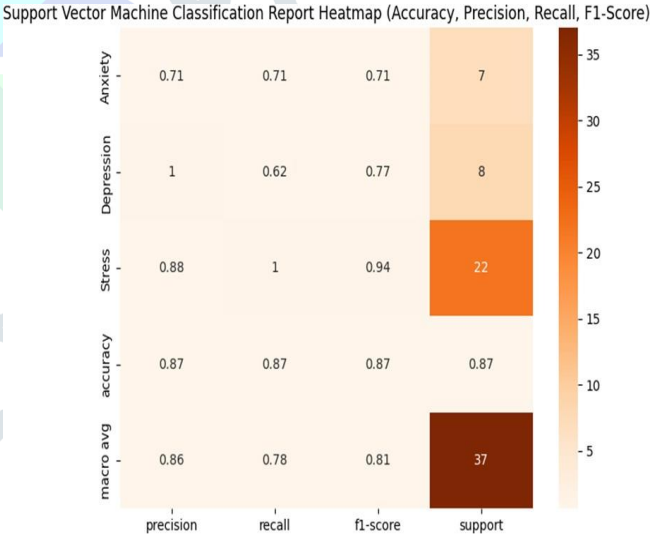

Fig 1.5.2: Support Vector Machine Classification Report Heatmap (Accuracy, Precision, Recall, F1-Score)

- RANDOM FOREST AND DECISION TREE CLASSIFIER MODELS:

We explored ensemble techniques like random forest and decision trees to harness the collective power of multiple trees for prediction. Random forest constructs numerous decision trees and aggregates their predictions to enhance accuracy and

reduce overfitting. Decision trees, on the other hand, provide interpretable rules for classification. By fitting these models, we sought to capitalize on their ability to capture intricate relationships within the data, while also assessing their performance against other algorithms.

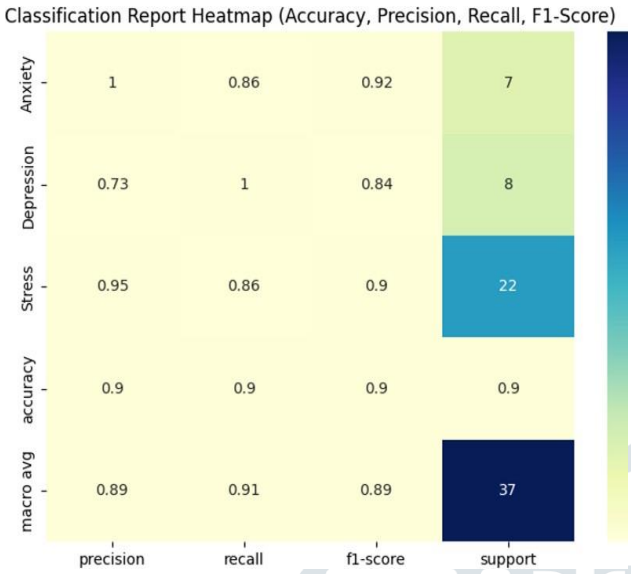

Fig 1.5.3: Decision Tree Classification Report Heatmap (Accuracy, Precision, Recall and F1-Score)

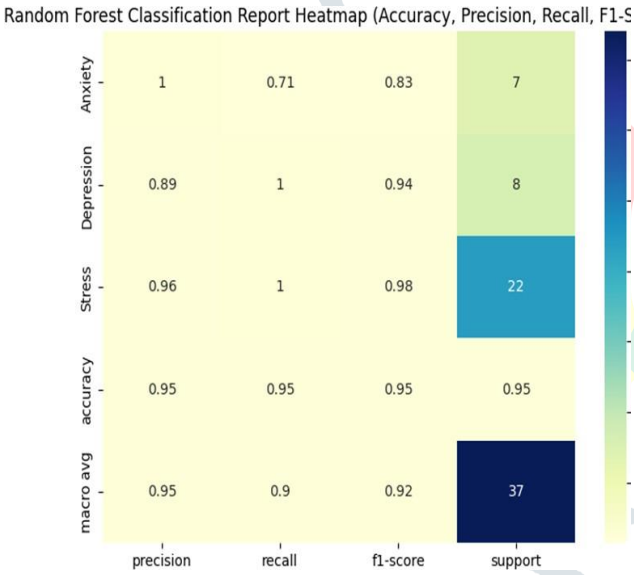

Fig 1.5.4: Random Forest Classification Report Heatmap (Accuracy, Precision, Recall and F1-Score)

- K-NEAREST NEIGHBOUR (KNN) CLASSIFIER MODEL:

KNN is a proximity-based classification algorithm that relies on the similarities between instances. Our KNN model leveraged the dataset's features to classify instances by identifying the nearest neighbours. With its simplicity and intuitive approach, KNN provided an additional perspective on the data's patterns and separability.

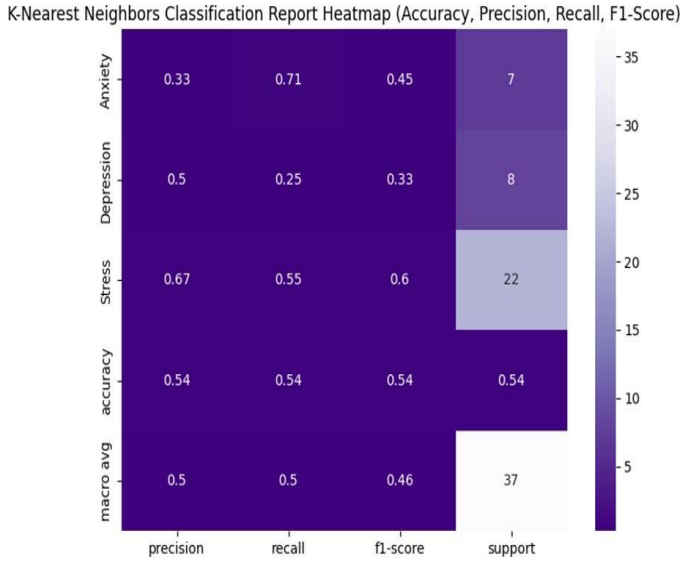

Fig 1.5.5: K-Nearest Neighbours Classification Report Heatmap (Accuracy, Precision, Recall and F1-Score)

We utilized the various algorithm to predict mental health conditions and evaluated its effectiveness through a Classification Report Heatmap. This heatmap visualizes crucial metrics including accuracy, precision, recall, and F1-score across different conditions. The shades represent metric values, with warmer shades indicating higher values. The heatmap offers insights into the model's strengths and areas of improvement, guiding our understanding of mental health prediction.

Therefore, by looking at the heatmaps we can show that the accuracies of the models were:

Random Forest: 0.95 i.e., 95%

Decision Tree: 0.90 i.e., 90%

SVM Classifier: 0.87 i.e., 87%

Logistic Regression: 0.86 i.e., 86%

K-Nearest Neighbours: 0.54 i.e., 54%

1.5.2 Model evaluation using K-fold cross validation:

In our pursuit of creating accurate and reliable predictive models for mental health assessment, we adopted a rigorous validation technique known as K-Fold Cross Validation. This method goes beyond a simple train-test split and provides a comprehensive evaluation of model performance, offering a more realistic estimate of how the models would generalize to unseen data.

K-Fold Cross Validation involves partitioning the dataset into 'K' subsets or folds. The model is trained on 'K-1' folds and evaluated on the remaining fold. This process is repeated 'K' times, with each fold serving as the validation set once. The performance metrics obtained from each fold are then averaged to derive an overall assessment of the model's predictive capabilities.

For each of our models – including Logistic Regression, Support Vector Machines, Random Forest, Decision Tree, and K-Nearest Neighbours – we applied K-Fold Cross Validation. This approach enabled us to validate the models on multiple subsets of the data and obtain a robust understanding of their performance.

1.5.3 Performance metrices across different models:

The mean accuracy scores obtained from the cross-validation process are as follows:

- Random Forest Model: 0.93
- Decision Tree Model: 0.91
- Logistic Regression Model: 0.81
- Support Vector Machines (SVM) Model: 0.78
- K-Nearest Neighbours (KNN) Model: 0.62

Through this comprehensive modelling and evaluation process, we aimed to uncover the optimal approach for mental health condition prediction. The results of this analysis provide valuable insights into the applicability of different machine learning algorithms in the context of mental health assessment and contribute to our broader research objectives.

Below are the graphs showing the accuracies of all the models at each fold:

1. Random Forest:

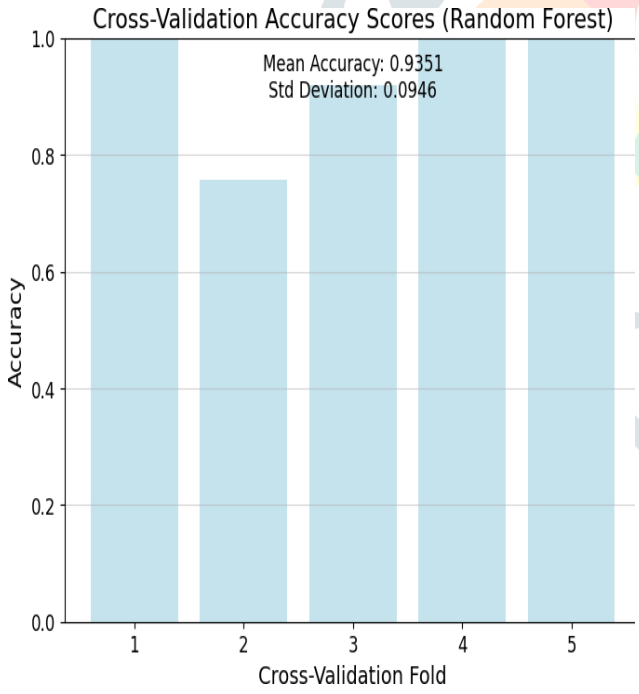

Fig: 1.5.6: Cross-Validation Accuracy Scores (Random Forest)

2. Decision Tree

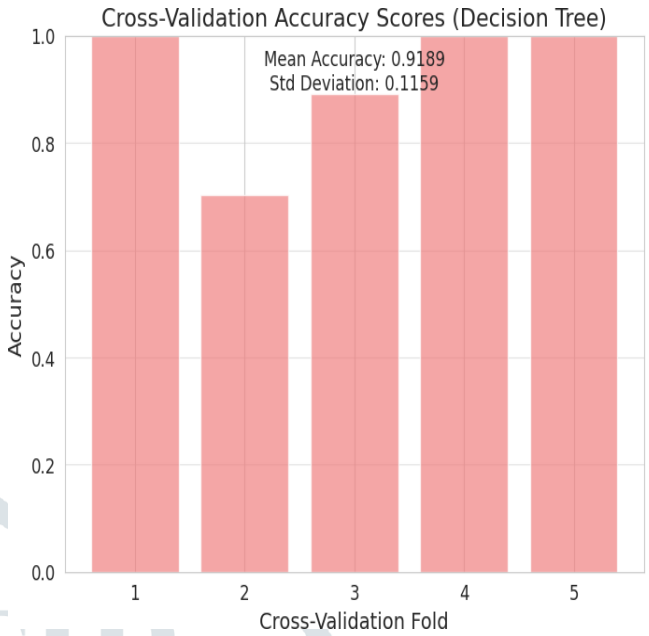

Fig: 1.5.7: Cross-Validation Accuracy Scores (Decision Tree)

3. Logistic Regression:

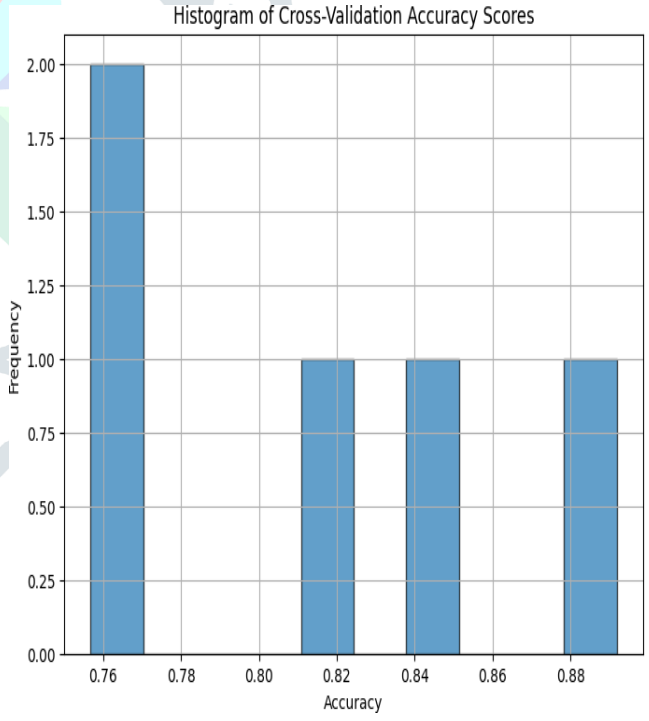

Fig: 1.5.8: Cross-Validation Accuracy Scores (Logistic Regression)

4. Support Vector Machine:

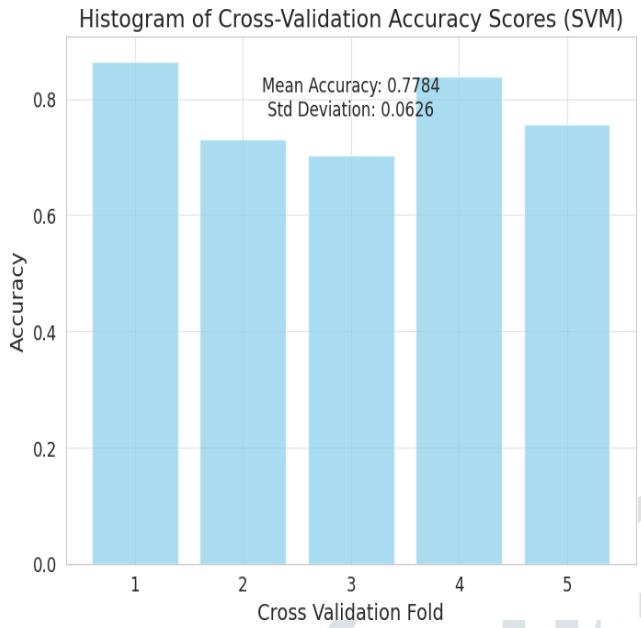

Fig: 1.5.9: Cross-Validation Accuracy Scores (SVM)

5. K-Nearest Neighbours:

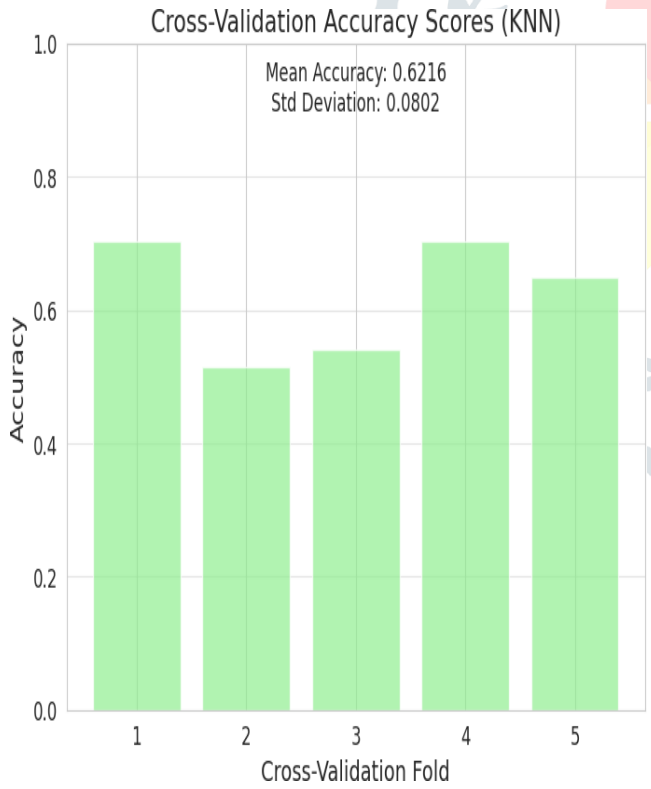

Fig: 1.5.10: Cross-Validation Accuracy Scores (KNN)

2. CONCLUSION

This research aimed to assess mental health conditions and their associated factors using data collected through an online survey. Through meticulous data preprocessing, exploratory data analysis, and predictive modeling, we have gained valuable insights into the relationships between various demographic factors and mental health conditions. Our findings revealed significant associations between age, personality traits, depression symptoms, stress-related factors, and mental health outcomes. Moreover, the predictive models, including Decision Tree, KNN, SVM, Logistic Regression, and Random Forest, yielded promising results in classifying mental health conditions. Out of the models used, Random Forest, Decision Tree and Logistic Regression models showed more promising results. The application of feature selection techniques further enhanced the accuracy and interpretability of our models. Our study contributes to the growing field of mental health research and offers actionable insights for healthcare practitioners, policymakers, and researchers. However, we acknowledge the limitations of our study, including potential biases in survey data and the need for further validation. Moving forward, future research endeavors can build upon our findings to develop targeted interventions and strategies for mental health support.

3. ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to the participants who generously dedicated their time and shared their valuable insights for our mental health assessment project. Their willingness to open up and provide us with essential data has played a pivotal role in the accomplishment of our research.

Furthermore, we acknowledge the unwavering support and guidance offered by Mrs. Sunita Bangal and Mrs. Poonam Bhawke of Department of Technology of Savitribai Phule Pune University. Their expertise has been vital in shaping the trajectory of our research and refining our methodologies.

We would like to extend our warm thanks to the academic community for providing an environment conducive to exploration and learning. The diverse array of resources and seminars have significantly enriched my understanding of both the subject matter and the research process.

We would also like to extend our gratitude to our mentors, colleagues, and friends for their invaluable feedback, thought-provoking discussions, and continuous encouragement. Their insights have significantly influenced our project's development and enhanced the quality of our outcomes.

We wish to acknowledge the collective efforts of all these individuals and resources, without which this project would not have come to fruition. We are sincerely thankful for their contributions and support.

4.REFERENCES

1. StressNet: Detecting Stress in Thermal Videos By Satish Kumar, A S M Iftekhar, Michael Goebel, Tom Bullock, Mary H. MacLean, Michael B. Miller, Tyler Santander, Barry Giesbrecht, Scott T. Grafton, B.S. Manjunath, University of California Santa Barbara ,Dept. of Electrical & Computer Engineering ,Dept. of Psychological and Brain Sciences and Institute for Collaborative Biotechnologies.
2. The heterogeneity of mental health assessment by Jennifer J. Newson, Daniel Hunter, Tara C. Thiagarajan.
3. On the State of Social Media Data for Mental Health Research By Keith Harrigian, Carlos Aguirre, Mark Dredze of Johns Hopkins University
4. Shared depressive symptomatology across mental disorders: implications for mental health assessments By Adam Horvath of University of Sydney, Department of Psychology.
5. Do Models of Mental Health Based on Social Media Data Generalize? By Keith Harrigian, Carlos Aguirre, Mark Dredze of Johns Hopkins University kharrigian@jhu.edu , caguirr4@jhu.edu , mdredze@cs.jhu.edu
6. ATTENTION-BASED LSTM FOR PSYCHOLOGICAL STRESS DETECTION FROM SPOKEN LANGUAGE USING DISTANT SUPERVISION by Genta Indra Winata, Onno Pepijn Kampman, Pascale Fung Human Language Technology Center Department of Electronic and Computer Engineering Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong {giwinata , opkampman}@connect.ust.hk , pascale@ece.ust.hk
7. Kaiser Permanente Research Brief Mental health. This brief summarizes the contributions of Kaiser Permanente Research since 2007 on the topic of mental health, including depression, anxiety, and other affective and stress disorders.
8. The mental health assessment study [MeHAS] – PROTOCOL, version 1.1
9. Mobile technology for mental health assessment By Patricia A. Areàn, Kien Hoa Ly & Gerhard Andersson.
10. Detection of Maternal and Fetal Stress from the Electrocardiogram with Self-Supervised Representation Learning By Pritam Sarkar1 , Silvia Lobmaier2,* Bibiana Fabre 5 , Diego Gonz´alez 5 , Alexander Mueller 6 , Martin G. Frasch3,4,*, Marta C. Antonelli2,7,* Ali Etemad1,*
11. Towards Emotional Support Dialog Systems By Siyang Liu1,2∗ , Chujie Zheng1∗ , Orianna Demasi3 , Sahand Sabour1 , Yu Li3 , Zhou Yu4 , Yong Jiang2 , Minlie Huang1
12. The prevalence of mental illness in refugees and asylum seekers: A systematic review and meta-analysis: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7505461/
13. Fatigue and cognitive impairment in Post-COVID-19 Syndrome: A systematic review and meta-analysis: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8715665/
14. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8960395/
15. Childhood trauma and adult mental disorder: A systematic review and meta-analysis of longitudinal cohort studies: Michael T. McKay, Mary Cannon, Derek Chambers, Ronán M. Conroy, Helen Coughlan, Philip Dodd, Colm Healy, Laurie O'Donnell, Mary C. Clarke. https://onlinelibrary.wiley.com/doi/10.1111/acps.13268
16. The Long-Term Effectiveness of Interventions Addressing Mental Health Literacy and Stigma of Mental Illness in Children and Adolescents: Systematic Review and Meta-Analysis https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8714636/
17. Effects of parental mental illness on children's physical health: systematic review and meta-analysis https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/effects-of-parental-mental-illness-on-childrens-physical-health-systematic-review-and-metaanalysis/25B5458F1686C6FB8FB4830D862DD247
18. Effectiveness of nutrition and dietary interventions for people with serious mental illness: systematic review and meta-analysis: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9828433/
19. Laughter therapy: A humor-induced hormonal intervention to reduce stress and anxiety:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8496883/
20. Stress and Coping Strategies Among Malawian Undergraduate Nursing Students:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8169817/
21. Association Between Physical Activity and Risk of Depression: A Systematic Review and Meta-analysis: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9008579/
22. Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis: https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bjc.12333