Technical Objective:

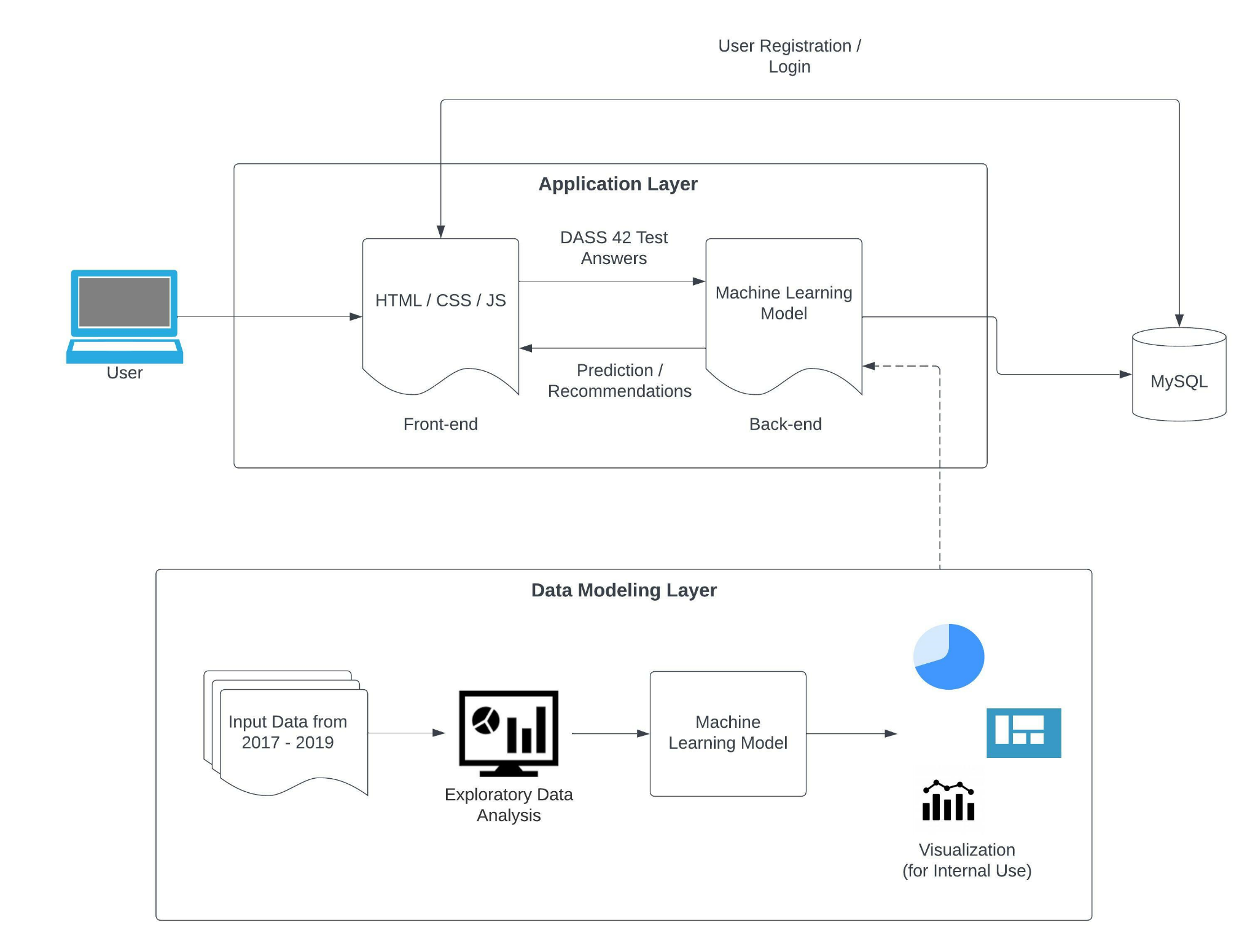
To create a scalable and interactive application to analyze the mental health of customers and provide them with a rating showing the severity of their mental health disorder. All the users should be able to take the test, but the registered users will be able to view the final results. All non-registered users can view their results once they register. All users can view their history by logging in to the application. The application should allow the user to book an appointment with the specialist and provide valuable resources like motivational videos, recommendations, suggested books, and information about support groups to boost their mental health.

The DASS42 questionnaire, when answered, generates the dataset for modeling. Using Machine learning models, mainly classification models such as decision tree, random forest, Naive Bayes, and support vector machine, to predict the severity of mental health. After comparing evaluation measures such as accuracy, f1-score, precision, and recall, the support vector machine is selected as the best model.

Technical Challenges:

The questions analyze multiple parameters (more than 100 variables) of mental health, and each of these parameters plays a different part in assessing mental health capabilities, deciding between these attributes. The challenge arises to find the best combination of the parameters to analyze the answers. We faced a challenge in adding the labels which show the range of severity to the output graphs for stress, anxiety, and depression. Finding each feature’s importance and deciding which features to use for model prediction.

**TECHNICAL ARCHITECTURE**

****

**About Data**

We took the dataset from: <https://openpsychometrics.org/_rawdata/DASS_data_21.02.19.zip>,this Data was collected from 2017 - 2019 with 5000 online participants who participated in a questionnaire available through the Open Source Psychometrics Project (https://openpsychometrics.org/). This organization maintains an open website for the public to take psychometric tests for educational and entertainment purposes.

We worked on the online questionnaire to assess personality traits and whether the user shows certain emotional states (Stress, Anxiety, and Depression). The DASS assesses negative emotional symptoms using a 4-point Likert scale, ranging from 0 to 3. It has three subscales (Stress, Anxiety, and Depression), with 14 items in each subscale for the DASS 42. We scored these subscales by adding the total item scores, ranging from 0 to 42. The higher score on the DASS indicates greater severity or frequency of these negative emotional symptoms.

**Data Preprocessing**

The data has nearly 40k rows and 172 features. However, most data is about DASS - item, response time, and estimated random number of items. It would be better to calculate the total and sub-domain scores of DASS and other questionnaires. Depression : 0-14, Anxiety: 15-28, Stress: 29-42, and we evaluated these parts in 3 parts - Physical Symptoms, Mental Symptoms, and Total Score.

We calculated the DASS score with Questions 2, 4, 7, 15, 19, 23, 25, and 41, categorized as Physical symptoms, and other questions as mental symptoms.

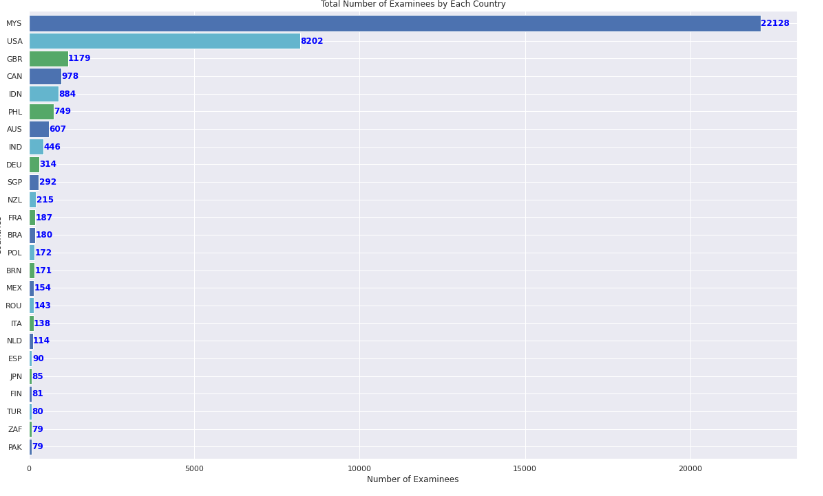
We cleaned the ***major*** column and simplified it using a method called Simplifier. We replaced the major columns filled with 0’s, 1’s, 2’s. The education column had some errors; it ideally should have had 1, 2, 3, and 4; however, it had 0’s. So we fixed that and replaced it with the ideal values.

We changed the birth year inputs and altered it to ***age***. Some years are unrealistic. We handled missing variables in ***TIPI*** by KNN imputer. We filled the country column with mode values and converted the ISO code 2 to 3.

Vocabulary Assessment(VCL), there were 16 items in ***VCL***. It asked examinees whether they knew or didn't know the given vocabulary. The given words were not actual words that were used as reliability items. Hence we created a summation of VCL; each correct word is +1, and reliability items are -1 each.

**Exploratory Data Analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations. From our data, we have taken a total count for examinees by each major in which no degree has the highest entries of 10902. As our data is diverse from all the countries, the highest participation is from Malesia, with a count of 22128 & the sample is not equally distributed; therefore, we cannot assume this sample size reflects the entire population.



In the below graph, the following items were answered as the examinee's felt that way most of the time:

Question 11: I found myself getting upset rather easily.

Question 13: I felt sad and depressed.

Question 18: I felt that I was rather touchy.

Question 34: I felt I was pretty worthless.

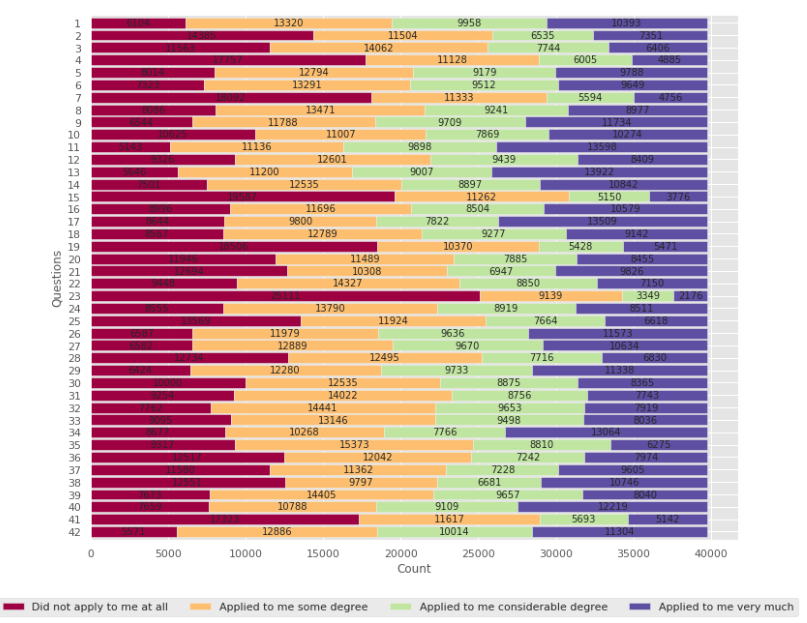
Question 40: I was worried about situations in which I might panic and make a fool of myself.

On the other hand, the following items were answered as the examinee's felt that way nearly never:

Question 15: I had a feeling of faintness.

Question 23: I had difficulty in swallowing.

It is obvious that most of the participants had mental/emotional symptoms of SAD, and did not have physical symptoms.

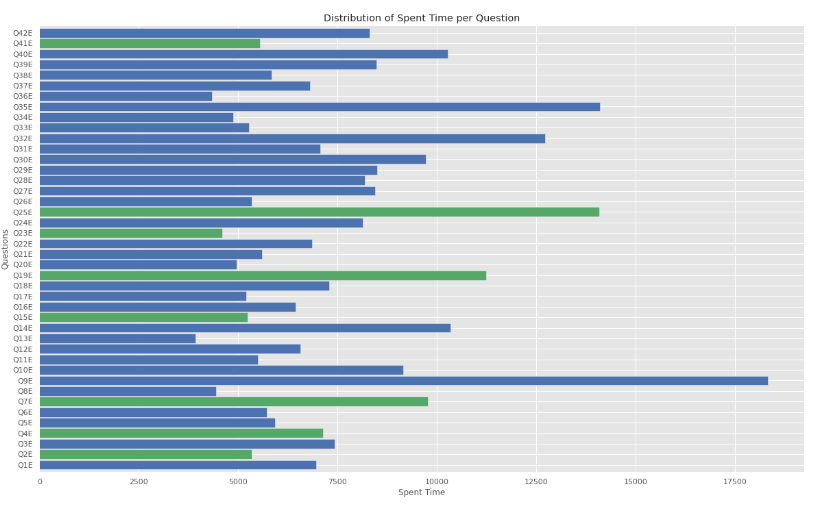


One of the important factors that need to be taken under consideration while evaluating a mental disorder is Time. The more time the participant takes to answer the question, the high chances that the participant is thinking & giving answers. There is a possibility that the answer is not correct or manipulated. For the same as per the statistics, the highest time is spent on Question No 9 followed by Question No 35 & Question No 25. The question says that:

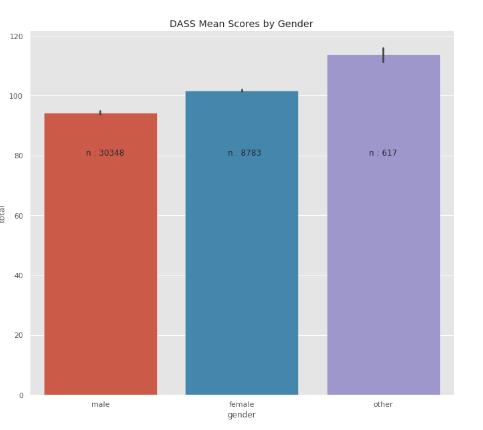
Q9: I found myself in situations that made me so anxious I was most relieved when they ended.

Q25: I was aware of the action of my heart in the absence of physical exertion (eg, sense of heart rate increase, heart missing a beat).

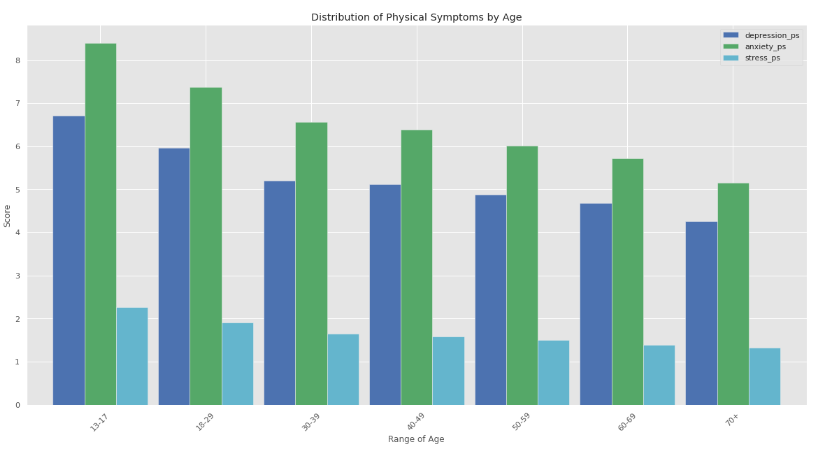
Q35: I was intolerant of anything that kept me from getting on with what I was doing.



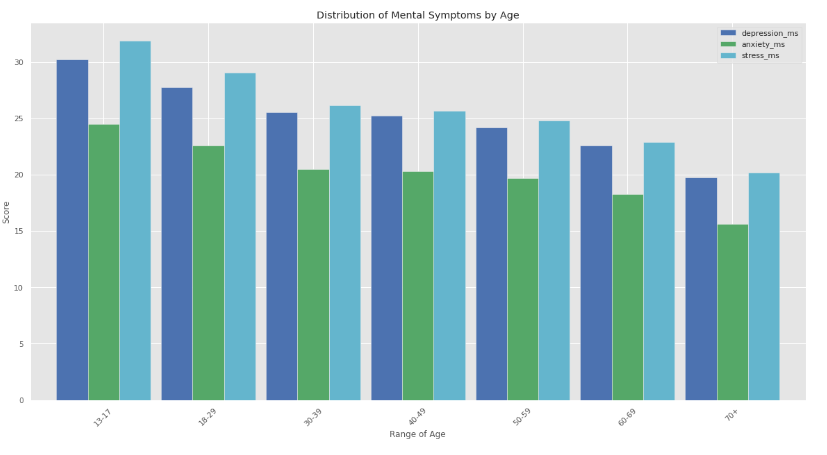
Similarly, I have calculated the DASS score as per the gender in which Male gender has the lower DASS score.



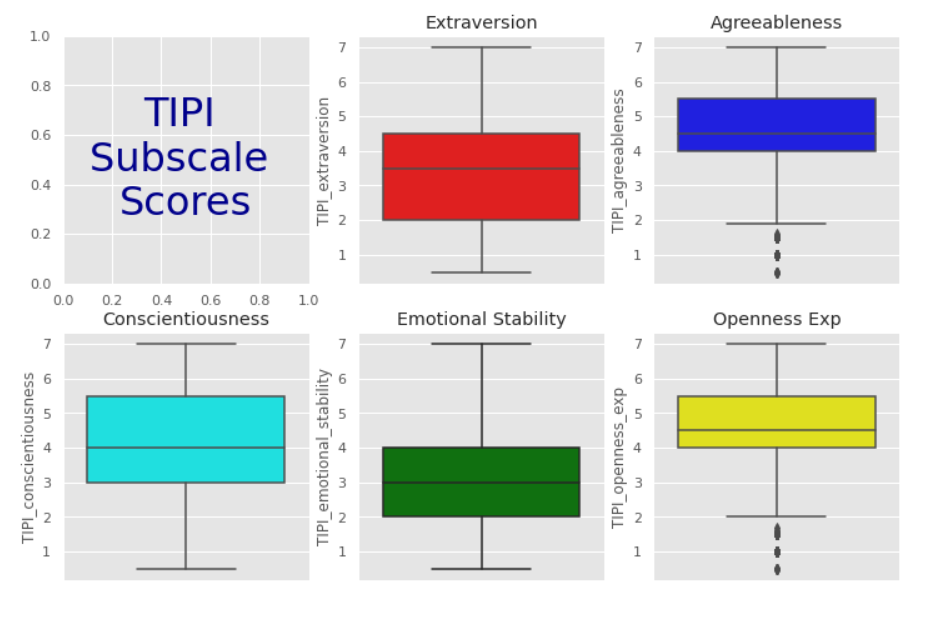
In our data, one of the factors is Age which plays an important role in having a mental disorder. I have divided it into physical symptoms & mental symptoms by age.



From the Physical Symptoms graph, we can say that Anxiety is more in the age group 13-17. Whereas the Mental Symptoms graph says that the amount of Stress is more in every age group as compared with depression & Anxiety. Depression, anxiety, and stress levels are getting lower with age. While anxiety causes physical symptoms, stress causes mental symptoms at all ages. Apart from that, depression causes both symptoms.

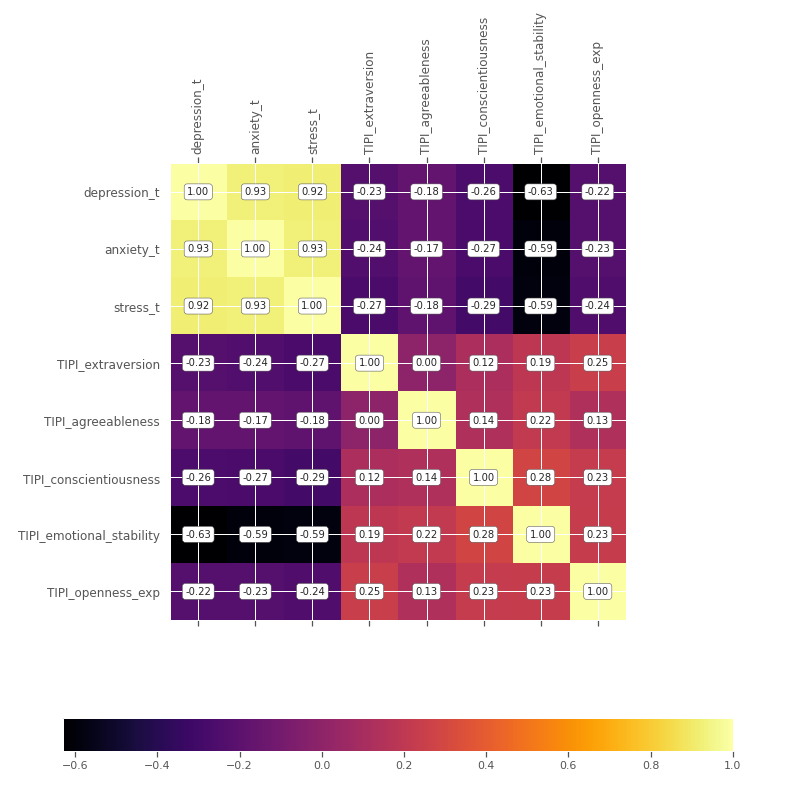


The TIPI is a ten-item measure of personality traits and characteristics conceptualized by the five-factor model (FFM). It assesses the big-five traits—Emotional Stability, Extraversion, Openness, Agreeableness, and Conscientiousness—using two items for each scale. Here are some of the observations as per our data.



Below we can see the correlation coefficients between variables. Each cell in the table shows the correlation between two variables.

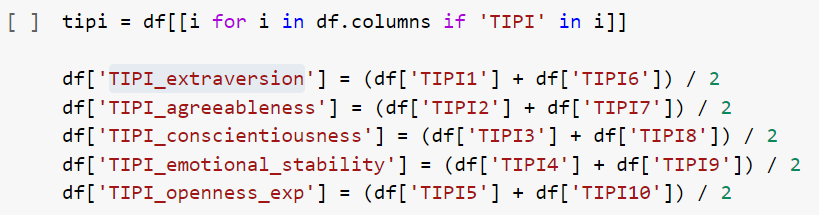
In this matrix, we can see that Emotional stability is negatively correlated with DASS items. Emotionally unstable (or less stable) individuals suffer more from stress, anxiety, and depression.



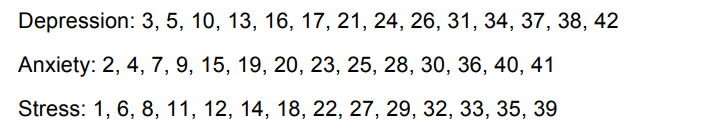
In EDA, we have analyzed different factors that can be the reasons for Stress, Anxiety & Depression.

**Model building** :

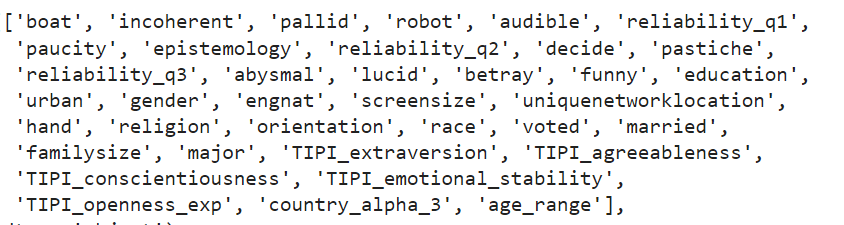
After visualizing the data to find insights, we have done feature engineering to remove the unwanted columns and also add new columns to the dataset. The initial data had multiple components to it: the answers to the questions, the time taken to answer the questions and the orientation of the questions, extra information about the user like (race, orientation, screen size), vocabulary assessment, and personality information. The important part of it is the answers to the question, which was analyzing stress, anxiety, and prediction and the additional information to know better about them. The personality information is grouped from the ten personality inventory to the big five personality dimensions like Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience, which can be seen in the below figure.



The Dass question, which consists of 42 questions, is divided into 3 different datasets each for Stress, Anxiety, and Depression. The below figure shows the breakdown of the questions .



The severity of the illness is scored differently based on the research done (References 8 ) and removed any null values present in the data to enable the model to learn better insights into the data.

All the models are constructed with answers from all 42 questions in the DASS-42 questionnaire, then the data is combined with the other parameters and fed to five different machine-learning models.The other we combined with the answers are shown in the figure below.  


**Decision Tree:**

Decision is taken as a base model for the model building and we have explored advanced models which are described below.

The parameters are hypertuned to get the highest accuracy.



**Random Forest:**

A Random forest algorithm is then implemented on top of the decision trees. This algorithm uses ensemble learning techniques to construct multiple decision trees and checks for overfitting and resolves it.



**Naive Bayes:**



**K-nearest neighbors:**

The parameter n\_neighbors is used for hyper parameter tuning.



**Support vector machines:**

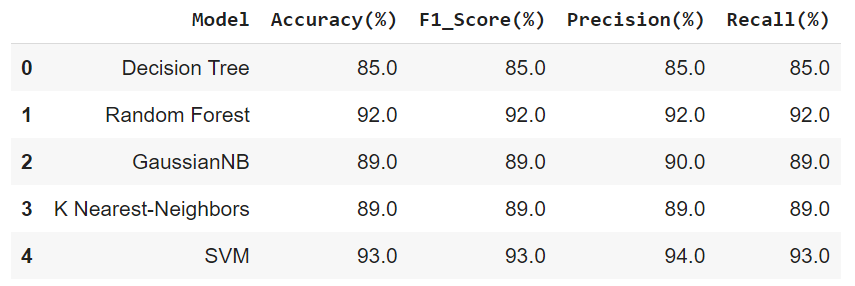
The parameters C, gamma, and kernel are hyper tuned to find the best parameters.



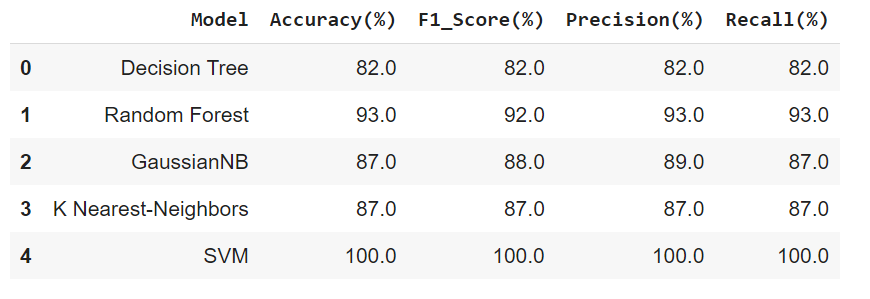
All three different types of data are fed to all five models creating a customized adaptivity to the data. The hyperparameters of the models are tuned using RandomSearchCv and GridSearchCV

The performance of the 5 models can be seen below.

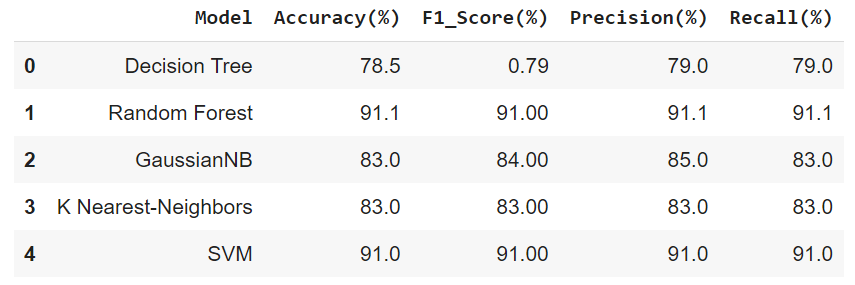
For Depression:



For Stress:



For Anxiety :



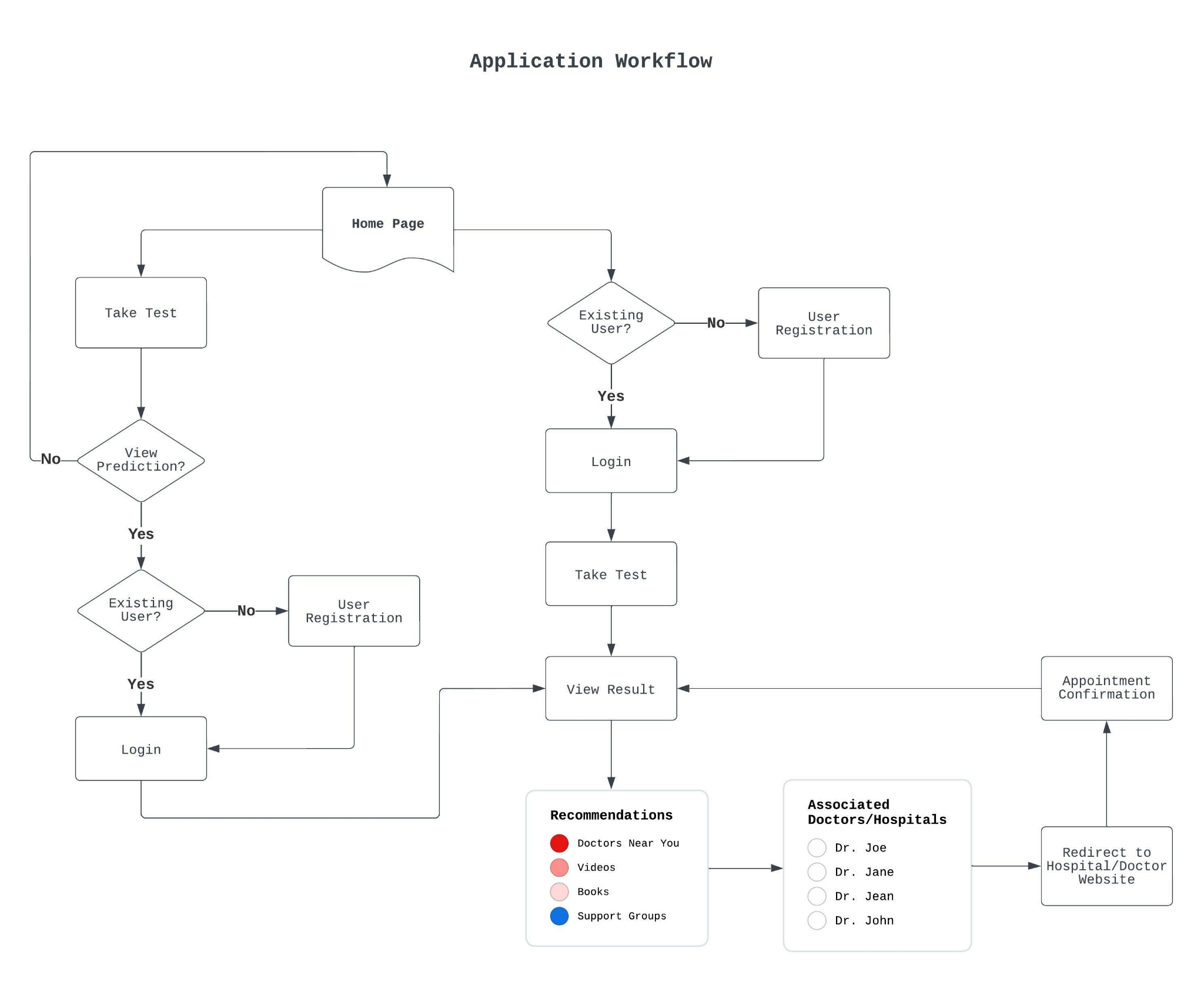
The output classes are labeled Low, Medium, High, Extreme for all three mental health disorders.

The user will get the result with the graphs depicted below:



From the model that was implemented, we chose the best model as the SVM( Support Vector Machine) for all three different types of data that we are trying to predict.

**USER WORKFLOW**

****

* The User/Customer starts on the **Home page** of our website.
* The user clicks the **Take Test** button, which redirects the user to the instructions page for the test.
* Clicking Next, the user answers pre-test questions like gender, age group, and occupation.
* On clicking Next, the user sees the questionnaire, where they answer 42 questions related to Stress, Anxiety, and Depression.
* On the last page of the questionnaire, the user clicks the **Results** button to view the test results**.**
* The user must be signed in or registered to see the results. If not, the application shows a popup asking the user to log in or sign up.
* Once the user logs in or becomes a member, they can view the prediction result.
* If the user is already an existing member, they can view the prediction result without restrictions.
* The member then sees a list of recommended specialists to **Book** appointments.
* The member can also view recommendations for motivational/meditation videos, mindfulness coaching options, support groups, and books related to mental wellness.