Digital Mental Health Assessment: Evaluating Depression and Anxiety Across Diverse Behavioral and Trauma-Induced Factors

Lakhera, Saksham
sakshaml@buffalo.edu
UB Person #50611360

Sivakumar, Sharan Raj
sivakum4@buffalo.edu
UB Person #505604183

Umredkar, Apurva
apurvara@buffalo.edu
UB Person #50592382

Vydadi, Rama Rao
ramaraov@buffalo.edu
UB Person #50604256

KEYWORDS

mental health, depression, anxiety, alcohol abuse, substance abuse, online gaming, big data, machine learning

ABSTRACT

Mental health disorders such as depression and anxiety are on the rise globally and are often influenced by factors such as alcohol and substance abuse, unresolved trauma from childhood, digital habits such as social media consumption and online gaming. In this work, we developed an interactive web platform to assist individuals to assess their mental health by evaluating these factors. The platform allows users to input data through an interactive assessment and leverages machine learning models trained on publicly available datasets and provides insights into their mental health, highlighting areas of concern and potential interventions. This tool is intended to assist both individuals and mental health professionals in identifying underlying triggers, promoting early diagnosis, and guiding therapeutic strategies. The scalability and accessibility of the platform make it a promising step toward digital mental health solutions.

1. INTRODUCTION

Taking care of mental health is essential for the overall well-being of an individual. It is often overlooked and leads to mental health disorders such as depression, anxiety, PTSD, OCD, ADHD etc. These disorders arise from various environmental and psychological factors. Around 9.2% of U.S. adults aged 18 or older had both a substance use disorder (SUD) and any mental illness in 2020. In 2023, around 16% of homeless individuals reported substance use disorders. This co-occurrence highlights the complex interplay between mental health and substance abuse.[1][2] . Children exposed to adverse childhood experiences (ACEs), such as abuse or neglect, have a higher likelihood of developing mental health disorders. These include post-traumatic stress disorder (PTSD), depression, and substance use disorders. Emotional disorders such as anxiety are common, with 4.4% of adolescents aged 10–14 and 5.5% of those aged 15–19 affected globally. Depression rates are 1.4% and 3.5% for the same age groups, respectively[3] . Excessive use of social media can increase the risks of depression and anxiety, particularly in adolescents and young adults. Cyberbullying, unrealistic comparisons, and sleep disruption are some contributing factors[4].

Despite the progress in many areas, seeking help for mental health issues is still often considered taboo in our society.

In this project, we explored publicly available datasets from SAMHSA and Kaggle, containing data gathered from surveys to assess the impact of these factors on mental health and its contribution to development of disorders. We developed several hypotheses to quantify the relationship between key mental health factors, with the goal of building predictive tools using machine learning models. We then deployed the classification and regression models using a multi-criteria decision making (MCDM) method in an interactive web platform built using the Streamlit framework for the UI and SQLite database engine for managing CRUD operations for persistent storage.

2. DATA

Data is the foundation for testing any hypotheses and developing machine learning models to validate them. In this section, we will discuss how the datasets for this mental health study were acquired and processed to make them suitable for exploratory data analysis and for fitting into machine learning algorithms.

2.1 DATASET ACQUISITION

For this project we acquired two datasets:

i. National Survey on Drug Use and Health (NSDUH), conducted annually by the Substance Abuse and Mental Health Services Administration (SAMHSA), provides nationally representative data on the use of tobacco, alcohol, and drugs; substance use disorders; mental health issues; and receipt of substance use and mental health treatment among the civilian, noninstitutionalized population aged 12 or older in the United States. NSDUH estimates allow researchers, clinicians, policymakers, and the public to better understand and improve the nation's behavioral health.[5]

ii. The Online Gaming & Anxiety dataset from Kaggle, which consists of data collected as a part of a survey among gamers worldwide. The questionnaire included questions commonly used by psychologists to assess individuals prone to anxiety, social phobia, and low or no life satisfaction. The questionnaire consists of several sets of questions asked as a part of psychological study. The original data was collated by Marian Sauter and Dejan Draschkow.[6]

The NSDUH dataset was loaded directly via a URL and processed in Parquet format, a columnar storage file format optimized for performance and efficiency. Parquet is widely used in big data processing due to its ability to store large datasets in a compact, compressed manner. The dataset consisted of data collected between 2015 to 2019 with over 2500 columns, each containing the answer to the questions in the survey.

The dataset from Kaggle was imported into our Python source code using the `opendatasets` library which uses the official Kaggle API. This dataset contained 55 columns consisting of demographics, information on the games played, and answers to the GAD-7, SWL and SPIN questionnaires.

2.2 CLEANING

After the datasets were loaded into our program, we analyzed them to uncover the underlying patterns in the dataset and identify anomalies in the dataset. With data processing and visualization libraries such as Pandas and Matplotlib, we checked the distribution of different data types.

For example, the following bar graph shows the distribution of data types in the NSDUH dataset:

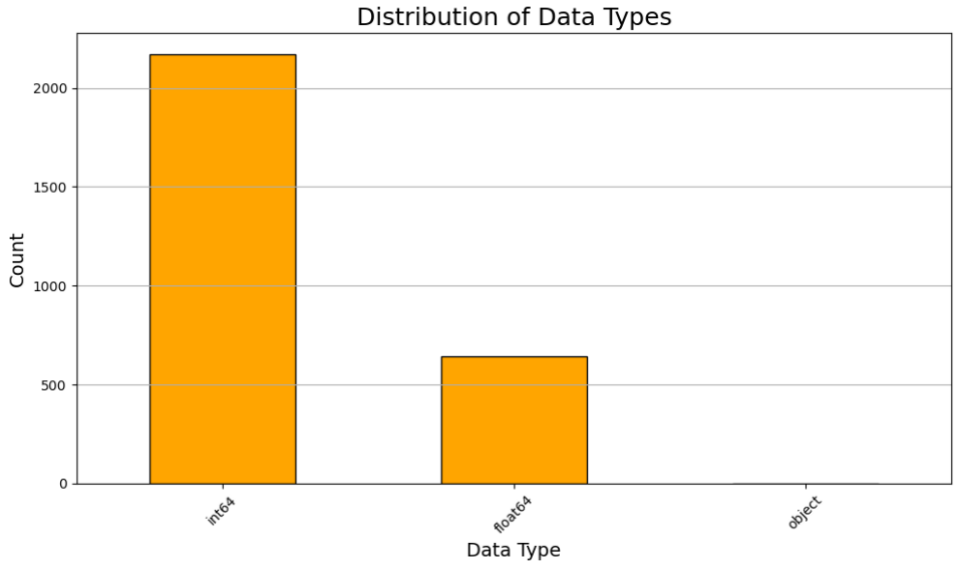


Fig 1

The following chart shows the columns with missing values in the online gaming & anxiety dataset:

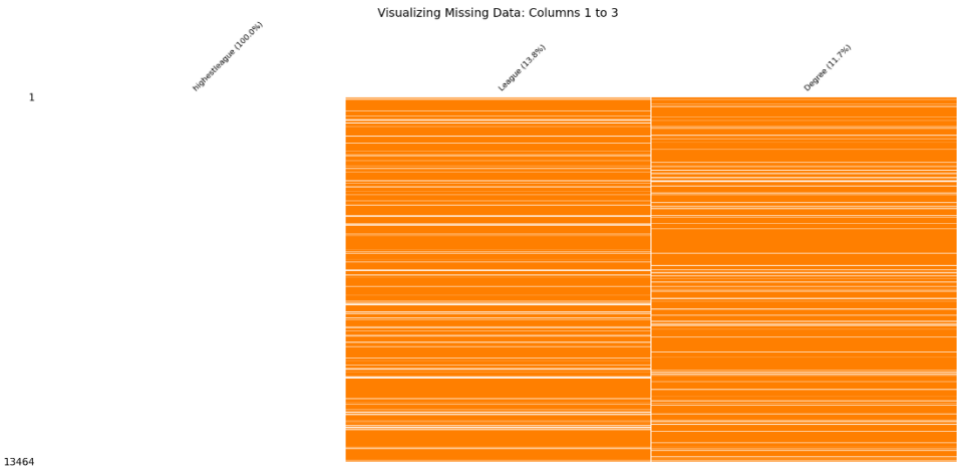


Fig 2

Once we identified the missing values, outliers, and inconsistencies we performed a thorough cleaning process to ensure that it was in a suitable format for training machine learning models. Following were the preprocessing steps:

i. Sparse columns, i.e. columns containing more than 10% of missing values, were identified and removed.
ii. Columns containing irrelevant or duplicate information were removed.
iii. Missing values were imputed to be replaced by the mean or mode of the column.
iv. Columns with object data types were encoded using label encoders.
v. Normalization using min-max scaler was performed to bring data to a uniform range.

3. EXPLORATORY DATA ANALYSIS

Once the data cleaning process was completed, we formulated several hypotheses aimed at understanding the relationships between key factors and mental health outcomes. These hypotheses

were essential in structuring our approach to analysis, allowing us to test specific assumptions and explore patterns in the data.

We validated the following hypotheses which were used as the foundation for developing our machine learning models and deployed into our data product.

3.1 Early drug use, particularly during adolescence, is a risk factor for the development of mental health issues.

In this hypothesis, we argue that either there is no relationship between early drug use and the development of mental health issues (null hypothesis) or that early drug use is associated with an increased risk of developing mental health issues (alternative hypothesis).

We discovered:

T-statistic = 43.341, which is very large. This indicates that the difference between early drug users and non-drug users (in terms of developing mental health issues) is much greater than what would be expected under the null hypothesis. In other words, this large T-value suggests a strong association between early drug use and the development of mental health issues.

P-value = 0. This indicates that the likelihood of observing this data if the null hypothesis were true (i.e., if there really is no association between early drug use and mental health issues) is virtually zero.

From this we inferred that the null hypothesis can be dismissed as:

- There is very strong evidence to suggest that early drug use, particularly during adolescence, is a significant risk factor for the development of mental health issues.
- The result implies that the observed relationship is not due to random chance but is likely a true effect in the population.

To support this inference, we also visualized the correlation between early drug usage with depression.

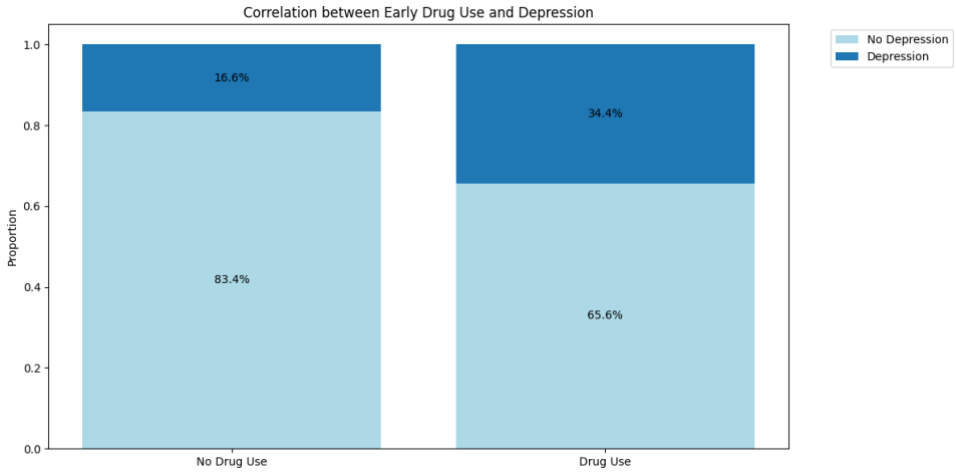


Fig 3

Key Observations:

- No Drug Use Group:
    - 83.4% of individuals who did not use drugs early are not depressed.
    - 16.6% of individuals in the "No Drug Use" group suffer from depression.
- Drug Use Group:
    - 65.6% of individuals who used drugs early are not depressed.
    - 34.4% of individuals in the "Drug Use" group suffer from depression.

We interpreted the above results as:

- Higher Depression in Drug Users: The proportion of individuals with depression is notably higher in the "Drug Use" group (34.4%) compared to the "No Drug Use" group (16.6%).
- Lower Depression in Non-Drug Users: Conversely, individuals who did not engage in early drug use show a much higher rate of being non-depressed (83.4%) compared to the "Drug Use" group (65.6%).

And we concluded that:

The chart visually supports the hypothesis that early drug use is associated with a higher likelihood of developing depression. There is a clear difference between the two groups, with a higher proportion of depression among those who used drugs early in life. This suggests that early drug use could be a risk factor for depression, as the relationship is clearly visible in the data.

Based on this hypothesis and findings we trained a binary classifier using the XGBoost model to check if an individual suffers from depression based on their substance consumption habits. The training process has been explained in the later sections.

3.2 Young online gamers suffer from higher anxiety levels.

From the online gamers & association with anxiety dataset acquired from Kaggle, we hypothesized that young gamers (age < 25) suffer from higher anxiety levels.

In Psychology, General Anxiety Disorder (GAD) has been divided into 7 categories:

1. GAD-1: Feeling nervous, anxious or on edge
2. GAD-2: Not being able to stop or control worrying
3. GAD-3: Worrying too much about different things
4. GAD-4: Trouble relaxing
5. GAD-5: Being so restless that it is hard to sit still
6. GAD-6: Becoming easily annoyed or irritable
7. GAD-7: Feeling afraid as if something awful might happen

The answer to GAD categories can be 0 - Not at all, 1 - Several days, 2 - More than half the days, 3 - Nearly every day.

To validate this hypothesis, we plotted radar charts across various age groups to quantify the anxiety levels using the answers to the GAD-7 questionnaire data from the dataset.

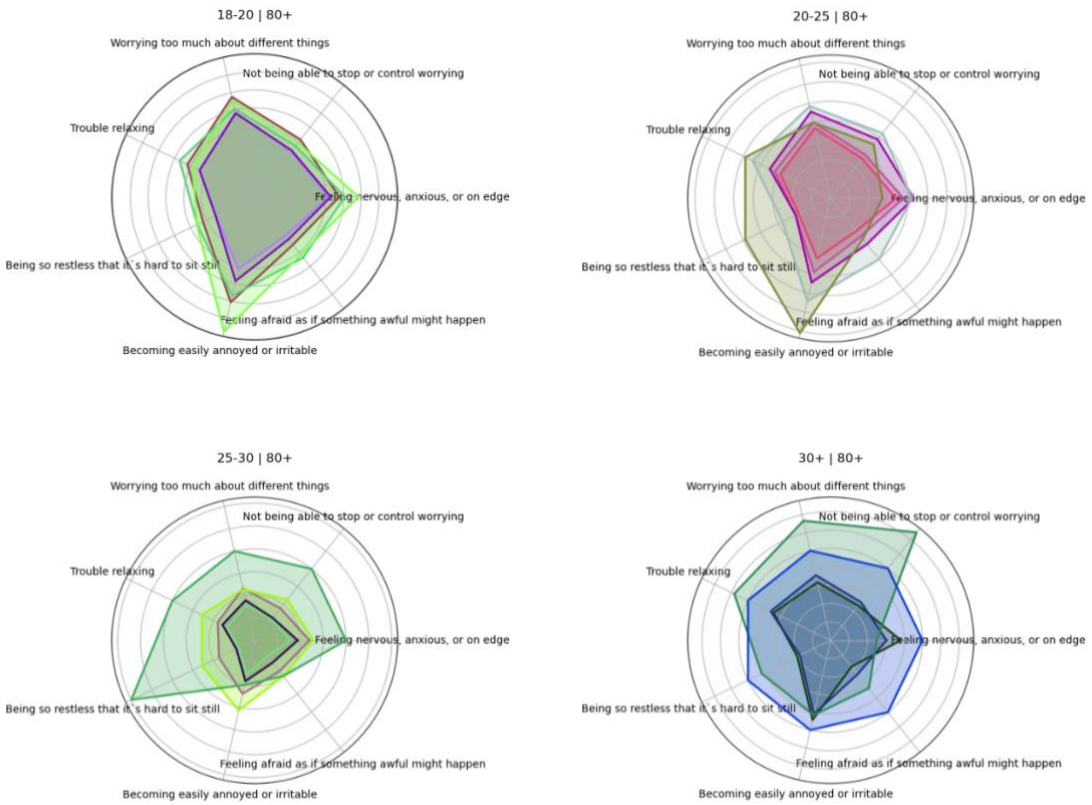


Fig 4

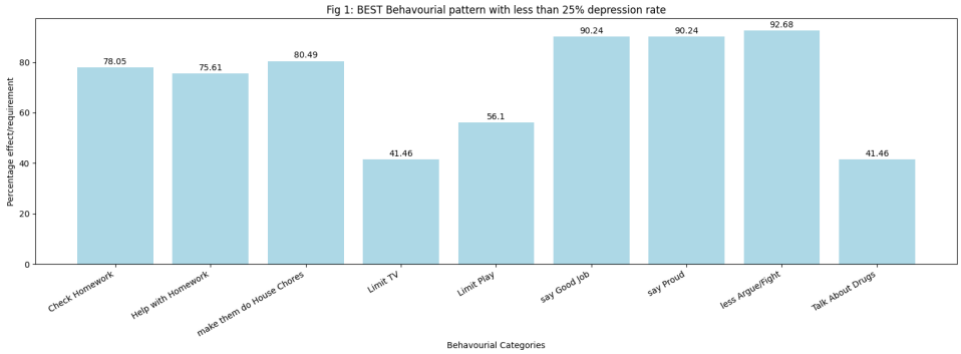
It can be concluded that young gamers from the age group 18-25 suffer from more anxiety in all areas.

- Age group 18-20 & 21-25 is feeling afraid as if something awful might happen (GAD-7).
- Age group 25-30 is mostly restless and find it hard to sit still (GAD-5).
- Ages 30+ are mostly not able to stop or control worrying (GAD-2).

Based on this hypothesis, we trained a multi-output MLP regression model which calculated the total anxiety, satisfaction with life and social phobia scores for online gamers.

3.3 Effect of bad parenting on a child's mental health.

In this hypothesis, we compared how behavioral patterns of parents affected the children and whether it caused them to be depressed. For this, we processed the depression rate columns from the dataset along with behavioral patterns such as making the child do homework, limiting screen time, showing appreciation, etc. We were able to plot the following visualizations:



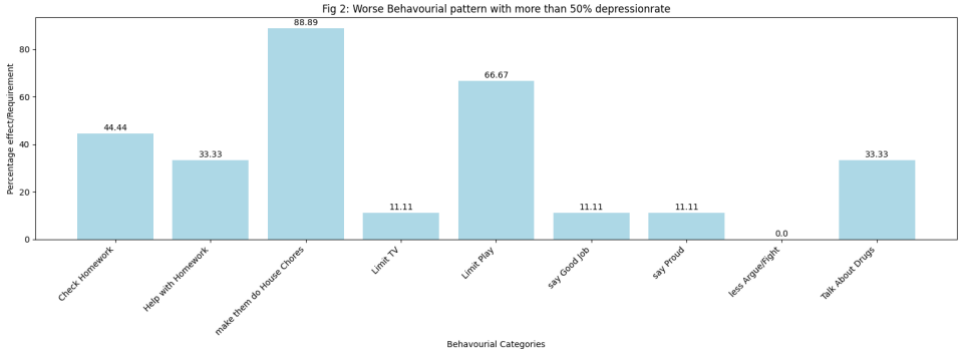


Fig 5

We were able to conclude that:

1. Praising your child whenever feasible significantly reduces depression, and the second graph shows that parents who do not do so have more depressed children on average.
2. Limiting your child's TV and playtime moderately has no negative impact on their mental health; nevertheless, if you tightly limit TV time while being flexible with their playtime, they may develop depression. As a result, somewhat strict restrictions on playing and watching television are beneficial to children's development.
3. Children are happy when their parents are involved in their work, such as assisting them with schoolwork and checking their homework, but parents who are not interested have more sad children on average.
4. For optimal mental health, parents should avoid arguing with their children and develop ways to prevent them from feeling the need to fight or dispute, as shown in Figure 2. Parents and children who frequently argue or fight with one other are more likely to suffer from depression. Fighting/arguing less with children has resulted in improved mental health outcomes.
5. Figures 1 and 2 show that on average, parents who talk to their children about the impacts of intoxicants on their health have better mental health. However, this component is not a substantial influence because there is only a minor rise.

Based on the results of this hypothesis, we trained a random forest classifier to identify if a child is depressed or not where we take the answers to the survey questions relating to the behavior of their parents towards them as the input.

3.4 Wealthy individuals (high socio-economic status) are less depressed.

In this hypothesis, we checked the impact of socio-economic status on mental health, i.e. we ask the question: are poor people depressed and rich people happy?

We were able to plot the following visualizations based on the data present in the NSDUH dataset.

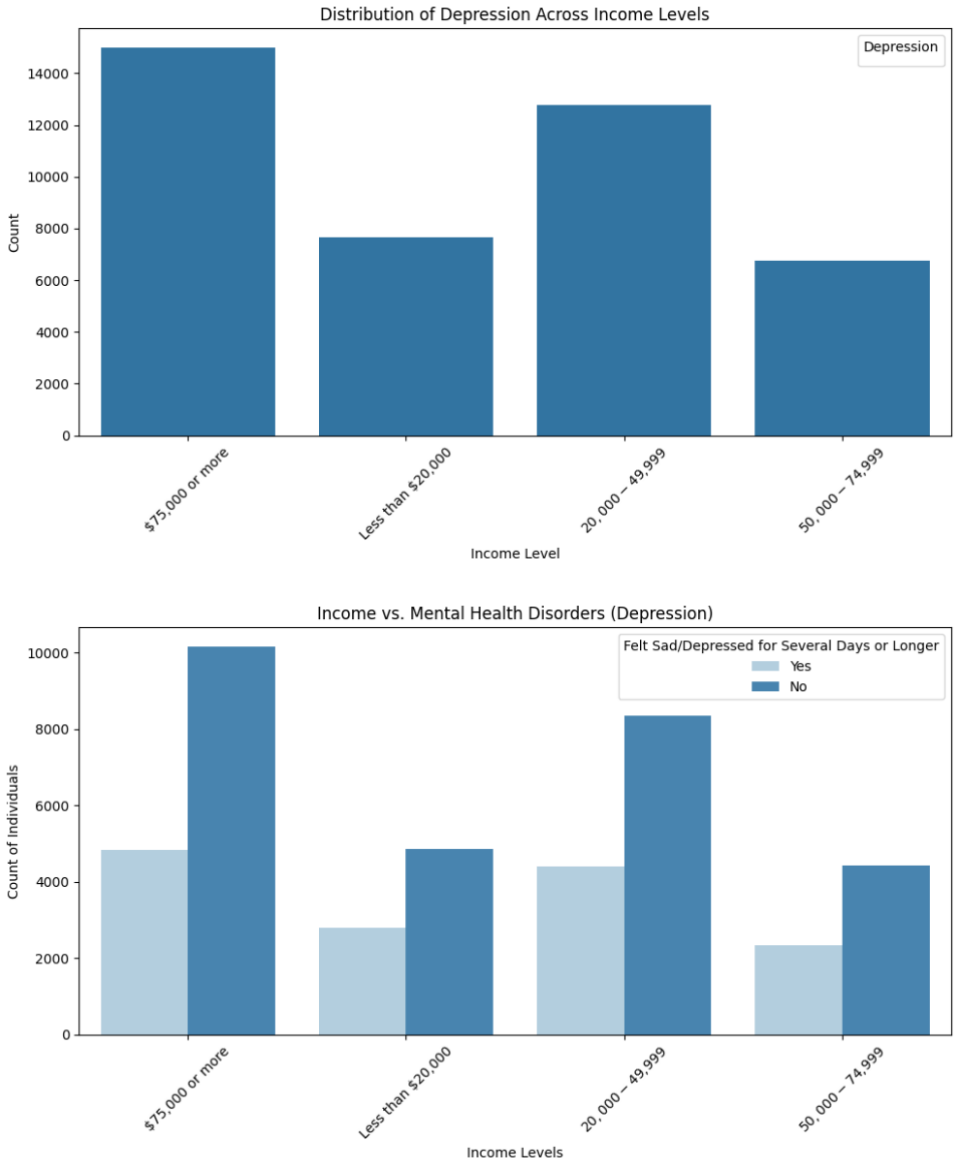


Fig 6

We can conclude that people whose income is less than $20,000 (which is the least among the others) are more likely to feel depressed.

Based on this hypothesis, we trained a binary classifier using the random forest algorithm to detect depression in a person based on their socio-economic status.

4. MODELING ML ALGORITHMS

We trained 4 machine learning models based on the hypotheses we formulated (section 3) for this study. These models are a crucial part of the backend of our data product.

4.1 XGBoost classifier to check for depression based on substance consumption habits.

Reason to choose the model: XGBoost uses an ensemble of decision trees, which allows it to model intricate patterns than logistic regression.

<write about which columns you chose and about any encoding/normalization you did>

We trained the model using a GridSearchCV to find the optimum set of hyperparameters.

Final set of model parameters:

- colsample_bytree: 0.8
- gamma: 0
- learning_rate: 0.01
- max_depth: 3
- min_child_weight: 1
- n_estimators: 300
- subsample: 1.0

Following is result of training the classifier:

- Classification report:

```
Best Cross-Validation Accuracy: 0.96
Test Accuracy: 0.97

Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.95      0.97      2568
           1       0.95      0.99      0.97      2674

    accuracy                           0.97      5242
   macro avg       0.97      0.97      0.97      5242
weighted avg       0.97      0.97      0.97      5242
```

Fig 7

- Confusion matrix:

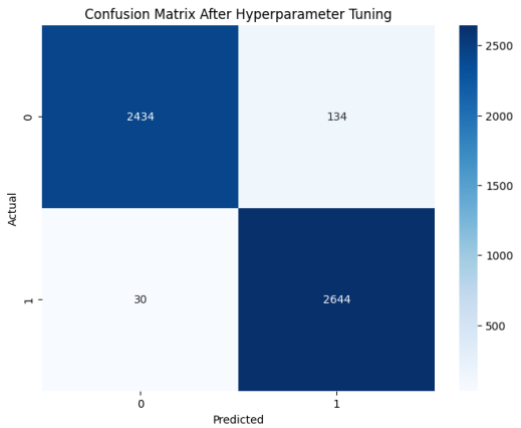


Fig 8

We were able to train the classifier with optimum set of hyperparameters to an accuracy of 97%.

4.2 MLP regressor to check the level of anxiety, satisfaction with life and social phobia.

Reason to choose the model: MLPs are more complex models with multiple hidden layers, allowing them to capture intricate interactions between features, which can be beneficial for complex datasets.

The columns of interest from the online gaming dataset for training this regression model were: Age, Game played, Hours played, and Country of residence for input and total GAD score, total SWL score and total SPIN score for output. The data in the game and country columns were converted to numerical data using a label encoder and the output columns were normalized using a MinMax scaler before fitting into the model.

Model architecture:

- hidden_layer_sizes = (50, 30)
- activation='relu'
- solver='adam'
- learning_rate_init=0.001
- max_iter = 500
- alpha = 0.01

Result after training:

```
MSE for GAD: 0.1623487330886398
MSE for SWL: 0.19444052729662445
MSE for SPIN: 0.04346593684242377
```

Fig 9

A low value of mean square error for the multiple outputs means that our regression model has trained well.

4.3 Random forest classifier to check for depression based on parental behavior.

Reason to choose model: Random Forest classifier have the upper edge in terms of high accuracy due to ensemble learning, robustness against overfitting, the ability to handle large datasets, and effective management of missing values. Additionally, we found that our dataset might have a good decision boundary.

We selected the following columns from the NSDUH dataset for training this model:

Column Name	Description
YEPCHKHW	Does your parent/guardian check your homework?
YEPHLPHW	Does your parent/guardian help you with homework?
YO_MDEA2	Have you lost interest in playing or any other favorite activities?
YEPCHORE	Does your parent/guardian make you do household chores?

YEPLMTTV	Have your parents limited your TV watching time in the past few months?
YEPLMTSN	Have your parents limited your playtime in the past few months?
YO_MDEA1	Have you been feeling discouraged while doing tasks you liked?
YEPGDJOB	Do your parents appreciate you when you have done a good/great job?
YEPPROUD	Do your parents let you know that they are proud of you?
YEYARGUP	How frequently do you have arguments with your parents in a month?
YEPRTDNG	Do your parents talk about the bad effects of drugs?
NEWRACE2	What is your race?

These features describe how parents behave with their kids and the environment they have created for them. The environment in which a kid grows up affects their mental health. The feature "race" is included as it provides insights into the cultural and societal context in which a child is likely to grow up. This serves as an indirect indicator of the environmental and community influences surrounding the child.

Normalization using MinMax scaler was performed, it followed integer encoding. Multiple hyperparameters were tested using trial and error to get the best results and finally the best model was selected with hyperparameter values given below.

Model architecture:

- n_estimators=30
- max_depth=17
- random_state=42

Result of training:

- Classification report:

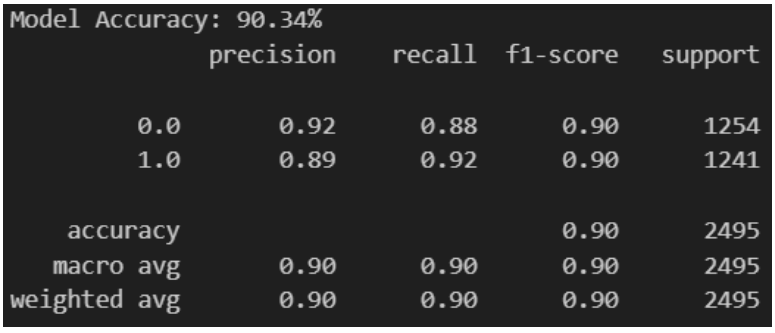


Fig 10

- Confusion matrix

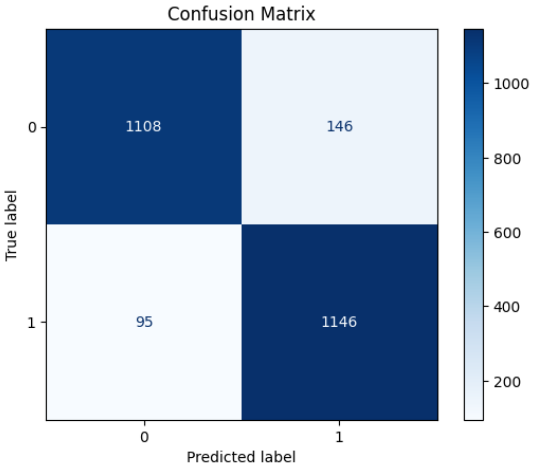


Fig 11

We were able to train the classifier with a classification accuracy of 90.34%.

4.4 Random forest classifier to check for depression based on socio-economic status.

For this classifier, we used the default hyperparameters that are set in the scikit-learn library.

The following features were selected for the Random Forest classifier:

- How often felt sad nothing could cheer you up: A categorical variable capturing the frequency of sadness.
- Employment status: The columns like *Employment_Employed* part time, *Employment_Other (incl. not in labor force)*, and *Employment_Unemployed* capture different types of workforce participation.
- Education: The two variables High_School_education and education_primary_education show the educational level of the respondents.
- Income: The Total_income_$50,000-$74,999, income_$75,000 or more, Total Less_than_$20,000 represents different category of household income.
- Geographic Context: We have also looked at further details about where one resides, whether he lives in the metropolitan or non-metropolitan area or a small metropolitan area.

These features were selected with consideration of the kind of data they offer which is psychological, socioeconomic as well as geographic.

These features were chosen because they provide a mix of psychological, socioeconomic, and geographic information.

Following are the results of the training:

- F1 scores, Recall and Classification report

```
Model Evaluation Metrics:
Accuracy: 0.912781954887218
Precision: 0.9140717339567406
Recall: 0.912781954887218
F1 Score: 0.9133013915713517
Confusion Matrix:
 [[4506  326]
 [ 254 1564]]
              precision    recall  f1-score   support

           0       0.95      0.93      0.94      4832
           1       0.83      0.86      0.84      1818

    accuracy                           0.91      6650
   macro avg       0.89      0.90      0.89      6650
weighted avg       0.91      0.91      0.91      6650
```

Fig 12

- Confusion matrix

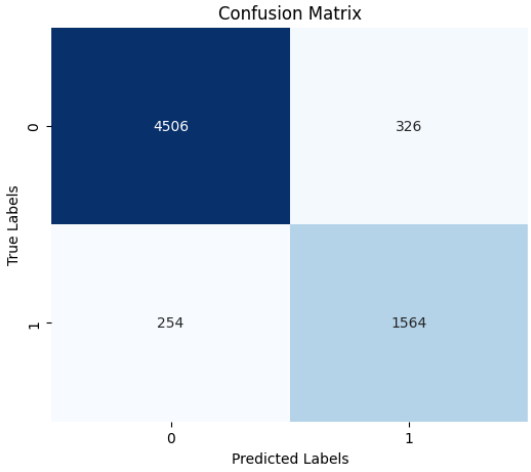


Fig 13

We were able to train the classifier with a classification accuracy of 91.27%.

5. DATA PRODUCT – WEB PLATFORM

We built an interactive web platform which integrated the machine learning models trained on the basis of the formulated hypotheses. Building this data product involved developing a frontend UI using Streamlit, integrating the models in the backend, and maintaining a persistent database using SQLite.

5.1 Walkthrough of the web-platform

Running the web-platform greets the user with an authentication page. We have offered the conventional option to either login as existing or sign up as a new user.

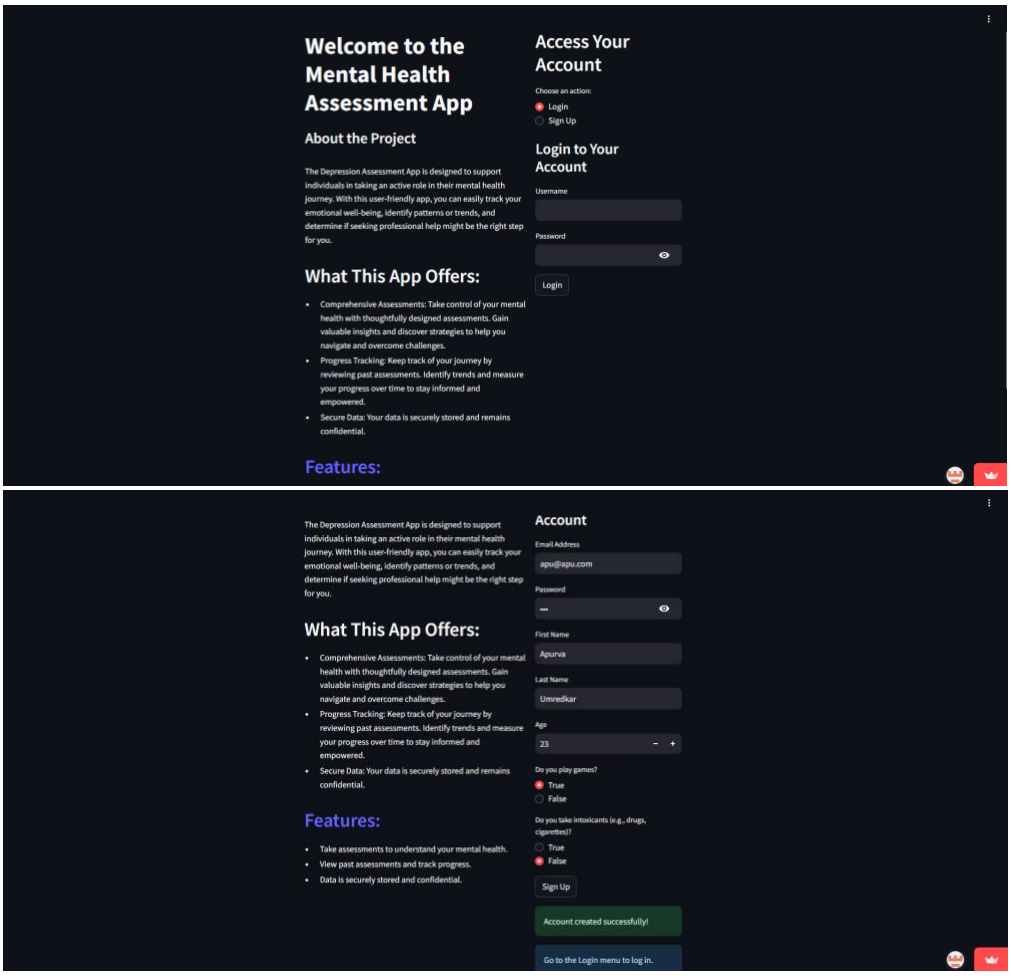


Fig 14

After a successful authentication, the user is redirected to a dashboard where they are choose to take an assessment to evaluate their mental health or view their previous assessments.

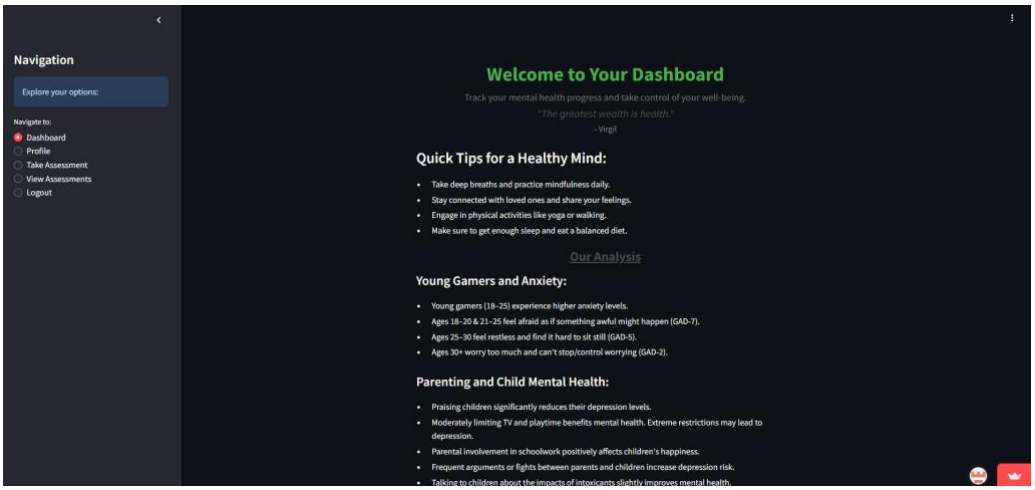


Fig 15

Page to modify profile details:

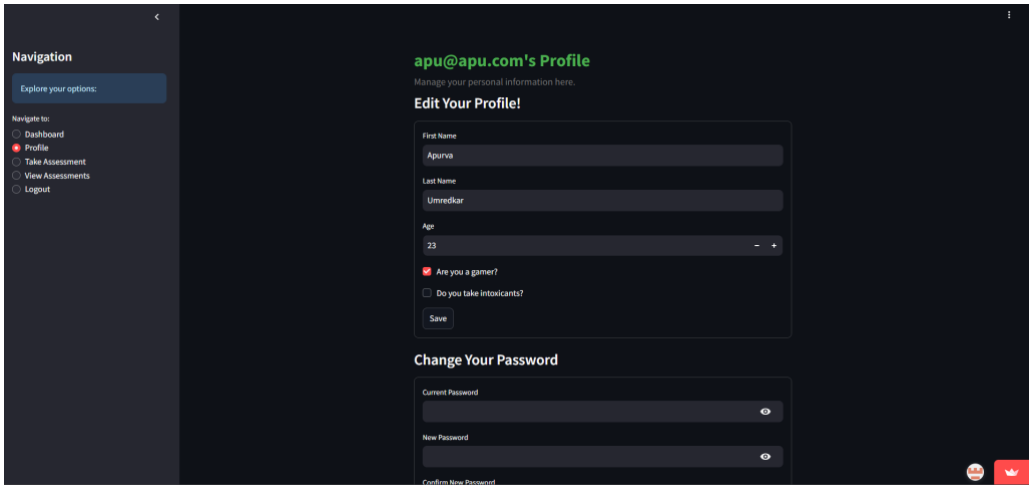


Fig 16

The user has to answer a set of questions presented to them.

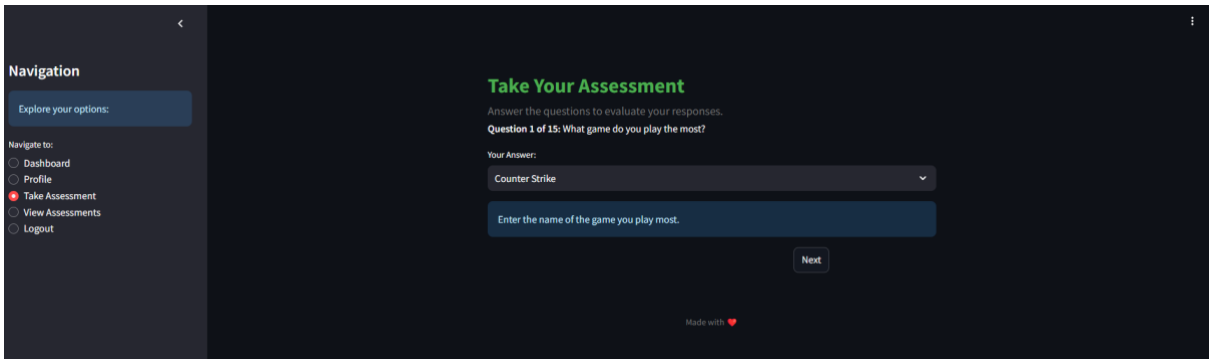


Fig 17

After completing the assessment, the results are presented to the user.

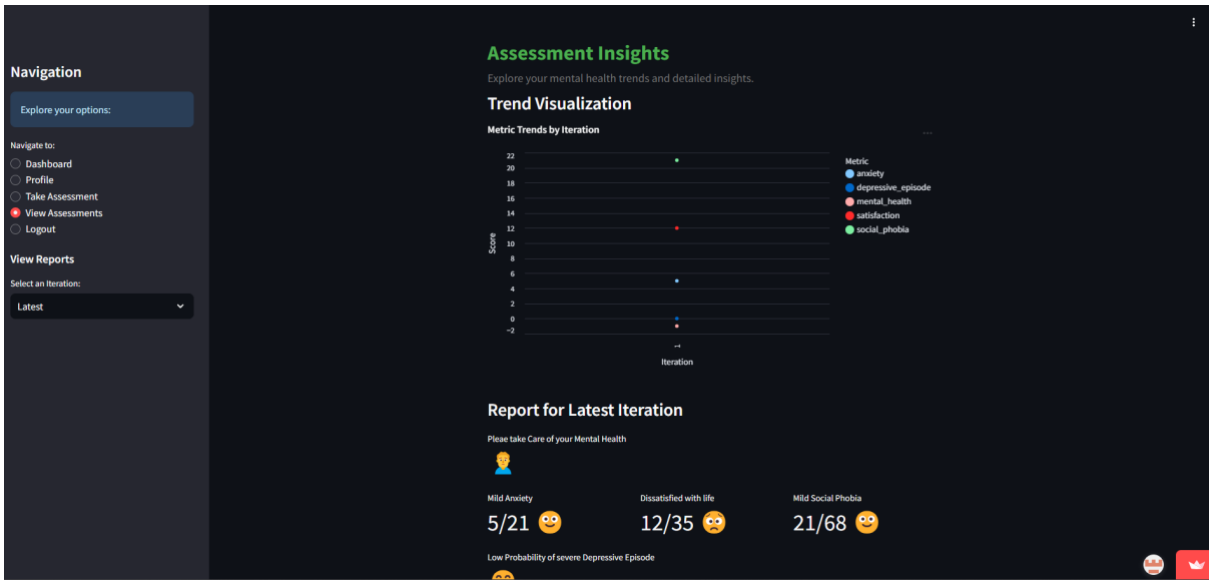


Fig 18

As the user keeps taking more assessments in the future, the platform helps them visualize the trend in their mental health. They have the option to view previous assessments and download the history.

Fig 19

5.2 Backend development

UI development:

The app uses Streamlit library to create its UI, The UI structure looks like

```
CSE587/
├── app.py
├── database.db
├── data.db
├── utils/
│   ├── auth.py
│   ├── database.py
│   ├── questions.py
│   └── utils.py
├── views/
│   ├── dashboard.py
│   ├── login.py
│   ├── profile.py
│   ├── signup.py
│   ├── take_assessment.py
│   └── view_assessment.py
├── code/
   ├── streamlit_app/
      ├── core/
         └── modelmanager.py
```

The project structure is organized to ensure modularity and maintainability, with distinct files and directories handling specific functionalities. Here's an overview of the file structure and their roles:

Root Directory (CSE587/)

- app.py: This is the main entry point of the application. It initializes the SQLite databases database.db and data.db. Additionally, it populates the "questions" table in data.db by executing the logic in questions.py located in the utils directory. The home page, served by app.py, provides options for user login and signup, which are managed by login.py and signup.py in the views directory.

utils/ Directory

- auth.py: Handles authentication-related functionalities, such as verifying user credentials.
- database.py: Manages database connections and operations, ensuring a seamless interaction with SQLite databases.
- questions.py: Responsible for populating the questions table in the database with predefined questions.
- utils.py: Contains utility functions to transform User input to data frame needed for ML models, it performs input mapping to features.

views/ Directory

- dashboard.py: Manages the dashboard UI after user login, displaying relevant information to the user.
- login.py: Handles the user login functionality, including input validation and authentication.
- profile.py: Provides the UI for user profile management, allowing users to update or delete their details. It supports CRUD operations on user data.
- signup.py: Handles user registration, including validation and adding user details to the database.
- take_assessment.py: Manages the UI for user assessments. It presents a set of questions based on the user's profile and records their answers in the assessment table in data.db. This module interacts with the **ModelManager** class to utilize machine learning models, feeding user inputs to the models and processing their results.
- view_assessment.py: Displays the user's assessment results, including their score and insights derived from the ML model.

Database and Persistence

All user data, including profile information and assessment results, is stored persistently in SQLite databases (data.db) and ML models are stored in database.db. Users have full control over their accounts, including the ability to update their details, reset their passwords, or delete their accounts, enabling seamless CRUD (Create, Read, Update, Delete) operations.

This modular and scalable architecture ensures efficient management of functionality and a smooth user experience.

We used data.db to store project-related tables, while database.db is dedicated to handling model-related data. This separation ensures scalability and maintainability; as the model grows in the future, managing it through a separate database is more efficient than combining all data into a single database.

Below is our table and database schema:

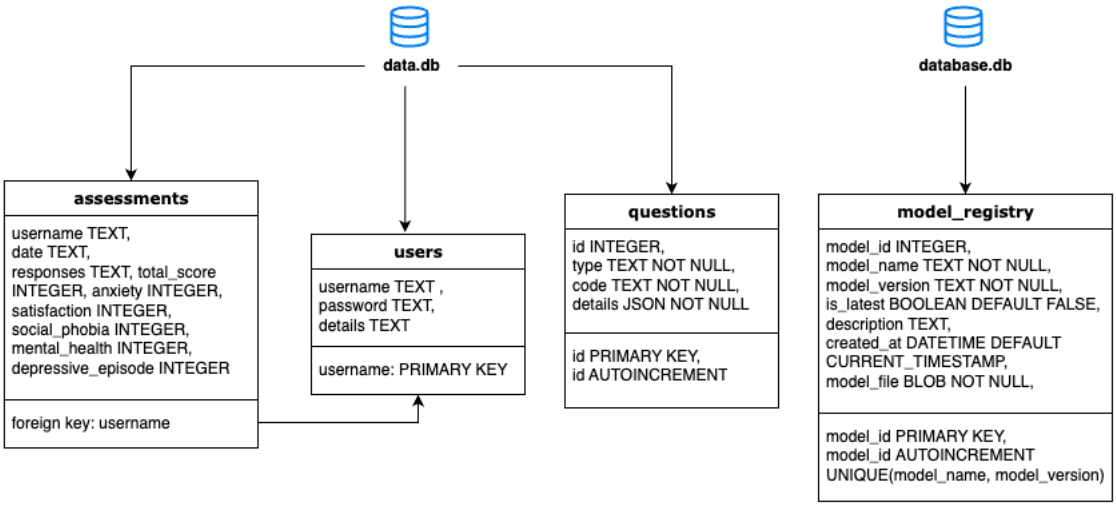


Fig 20

Model Persistence Management

The **ModelManager** class is a utility for efficiently managing machine learning models, including their storage, retrieval, versioning, and metadata. It uses SQLite as a backend to persistently store models as serialized pickle files in **database.db**, enabling easy management and reuse of trained models. The **model_registry** table tracks metadata such as model name, version, description, creation timestamp, and a flag indicating whether a model is the latest version. This versioning system ensures users can retrieve both the most recent and historical versions of models as needed.

The class provides methods for saving models (**save_model**), retrieving specific or latest models (**load_model**), and preloading frequently used models (**load_models**). When saving a model, the system automatically updates older versions of the same model to indicate they are no longer the latest. The **model_exists** method helps check for a model's presence in the registry, ensuring consistency during operations.

Beyond model management, **ModelManager** includes preprocessing utilities tailored to specific use cases like gaming mental health, youth drug abuse, child behavioral analysis, and general predictions. Each preprocessing function transforms raw user data into feature sets required by corresponding models. These functions ensure data consistency and simplify the process of preparing input for predictions.

The **run_models** method integrates the entire workflow, combining preprocessing and inference to generate predictions for a specified model. By automating model versioning, storage, and preprocessing, **ModelManager** serves as a powerful tool for managing machine learning pipelines and ensuring a streamlined workflow for data-driven applications.

In addition, it also contains final transformation functions which are required for each of the model to successfully execute.

REFERENCES

[1] Mental Health America – Youth Ranking 2024 https://www.mhanational.org/issues/2024/mental-health-america-youth-data

[2] Mental Health Statistics & Facts in 2024 https://womenonguard.com/statistics/mental-health/

[3] Mental health of adolescents https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

[4] Teens & Social media (APA) https://www.apa.org/monitor/2024/04/teen-social-use-mental-health

[5] NSDUH dataset by SAMHSA https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health

[6] Association of anxiety with online gaming dataset on Kaggle https://www.kaggle.com/datasets/divyansh22/online-gaming-anxiety-data