Mental Health Analysis

-using Social Media and Behavioral Patterns

By Abhinita Sanabada, Anish Nandigama, Kapil Reddy, Sai Chaitanya Munagala
Department of Applied Data Analytics
05/01/2025

Abstract

Mental health has become a critical concern in today's fast-paced and digitally connected world. Among many factors influencing mental well-being, Social media usage and behavioral patterns are key indicators of potential health risks. This project explicitly aims to leverage data analytics and machine learning models to predict mental health risk level using the above-mentioned key indicators. A custom binary risk indicator, *risk_flag*, is engineered based on critical factors like *social media usage, screen time, sleep duration,* and *stress levels.* The sample data set is thoroughly cleaned, with all the EDA requirements, and to address the class imbalance, the data set uses the *SMOTE* approach, ensuring fair representation of the *at-risk* and *no-risk* groups during the model training. The Two Machine Learning models [i.e*., Logistic Regression* and *Random Forest*], along with SMOTE, are well trained and evaluated with a *70:30* ratio of train and predict. Logistic Regression demonstrated higher recall, making it suitable for healthcare scenarios where catching every risky individual is crucial, while Random Forest offered balanced performance with higher overall accuracy. Through this analysis, the project provides insights into the relationship between Social median usage and behavioral factors for mental health risk and develops some visuals that can help us to unlock the next level. Visualization is the ultimate part of this project, using Tableau and occasionally Power BI to generate interactive dashboards that speak the results.

1. Introduction

In recent years, mental health has gained significant attention as a crucial aspect of overall well-being. With the rising influence of digital technologies and social media in people's daily lives, there is growing interest in understanding how lifestyle behaviors and online activities impact mental health. Excessive screen time, irregular sleep patterns, high work demands, and reduced physical activity are among the behavioral factors often linked to stress, anxiety, and other mental health concerns. Social media, in particular, has been both praised for its connectivity and criticized for contributing to stress and reduced psychological well-being. While it is truly understandable that it enables global communication and self-expression, studies have shown that prolonged and unregulated usage may be associated with poor sleep, social comparison, and heightened stress levels.

Advancements in data science and machine learning have made it possible to analyze such behavioral patterns and predict outcomes with reasonable accuracy. By combining lifestyle attributes with a very important social media usage factor, predictive models can assist in identifying individuals who may be at higher risk of experiencing mental health challenges. This project aims to apply data analytic techniques to build predictive models that detect mental health risk based on behavioral indicators.

2. Objectives

The primary objective of this project is to develop a predictive model capable of identifying individuals who may be at risk of mental health challenges based on social media usage and behavioral patterns. This will be achieved with structured planning, documenting every detail to its best capacity, and demonstrating visuals with interactivity for the users. Although the fundamental objective is to provide visual insights that demonstrate the results on their own, the analytical objective is further classified as below.

A. Primary Objective:

1. To identify individuals who may be at risk of mental health challenges based on grouped factors.

B. Data Analysis Objective:

2. To engineer a risk classification indicator (risk_flag)
3. To handle data imbalance using SMOTE
4. To build and evaluate machine learning models
5. To compare model performance and provide valuable insights.
6. To visualize key insights through interactive dashboards (Tableau)

---

3. Technology Stack and Tools

This project uses various data science technologies and tools across different phases, from the EDA phase, data processing, and modeling to visualizing reports.

1. Python 3 – Primary programming language for data processing, analysis, and machine learning tasks.

2. Libraries and Framework – Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn (sklearn), imbalanced-learn (imblearn)

3. BI Visualization – Tableau for creating interactive dashboards and complex visual analysis

4. Other tools and techniques – Google Colab, Microsoft PowerPoint, PDF (for reporting)

---

4. Methodology

This section outlines the step-by-step process followed in this project, from data collection and preprocessing to building and evaluating machine learning models for mental health risk prediction, until providing visual insights through Tableau and Power BI.

A. Data Collection and Preprocessing

The data set collected contains social media usage habits and behavioral attributes that are considered important indicators for mental health analysis. Before analysis and modeling, the raw data underwent careful preprocessing to ensure quality and consistency. Many things were taken into consideration while cleaning and prepossessing the data, some of them are listed below

1. Column names were cleaned and standardized.

2. Missing or null values were handled and removed as required.
3. Categorical fields (like stress level) were formatted (lowercase, stripped spaces) to avoid duplicates.
4. Check for duplicates and ensure data integrity.

These prepossessing steps, along with some others, prepared the data for effective analysis and machine learning modeling.

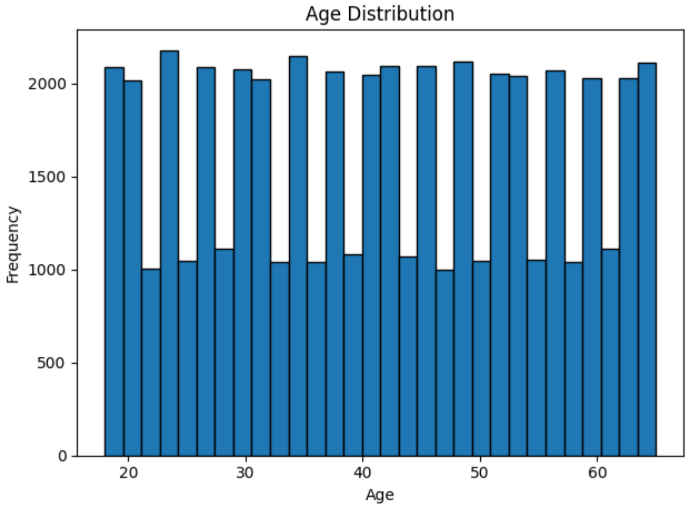
B. Exploratory Data Analysis (EDA)

EDA was performed to gain a deeper understanding of the dataset before applying machine learning models. This phase helped in identifying key patterns, relationships, and issues such as class imbalance.

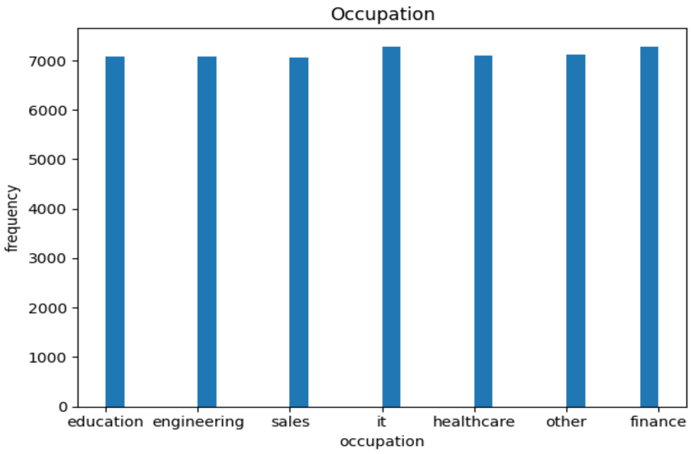
Phase I: Feature Distribution

The first step of EDA focused on univariate analysis to understand the distribution of each feature individually. It analyses the distribution of age, sleep hours, stress levels, and work hours, along with tracking very important social media usage. It also identifies how many individuals have sleep deprivation(less than 6 hours of sleep). This phase takes further steps to detect high social media users (>4 hours) and monitors stress distribution separately on the other hand. Overall, the first step of EDA involved analyzing individual features to understand how user behaviors were distributed across the dataset.

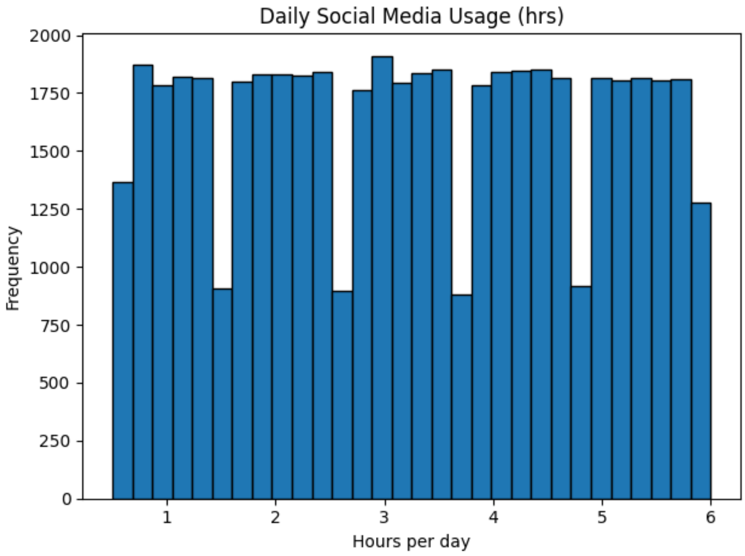
1. Age Distribution (histogram)—even bars from 18 → 6. The respondent pool is intentionally uniform across age, so age-driven effects (e.g., Gen Z vs Boomers) won't surface automatically—we'll need stratified or interaction analyses to reveal them.



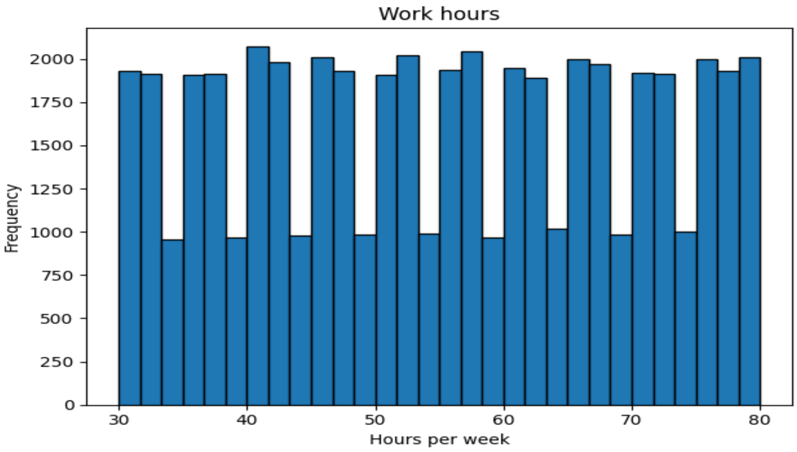
2. Occupation (count plot)—all seven job groups ≈ 7 k each. Sampling is again balanced. Any mental-health differences by job type (e.g., healthcare vs finance) will be driven by *behaviour*, not by unequal sample sizes.

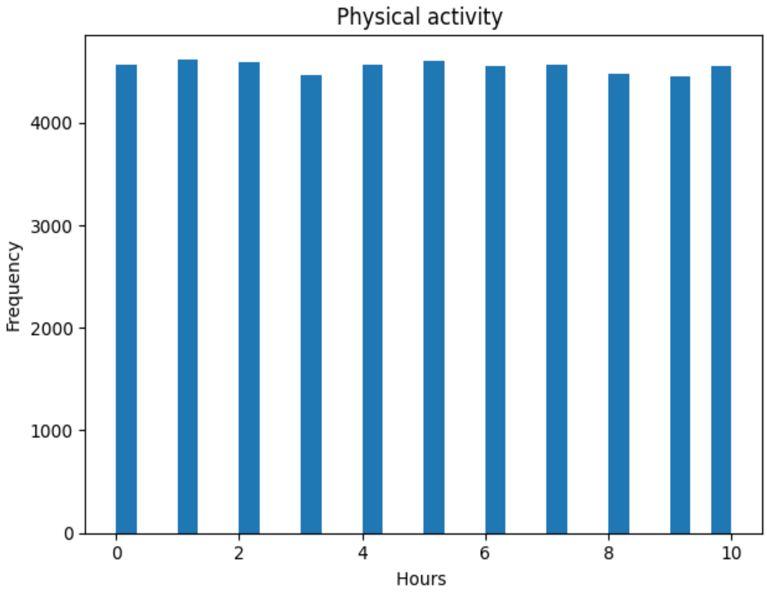


3. Daily Social-Media Usage (histogram)—flat 0.5-6 h.Usage was bucketed deliberately evenly, so linear correlation with outcomes will look weak; instead, look for threshold effects (e.g., > 4 h → risk).

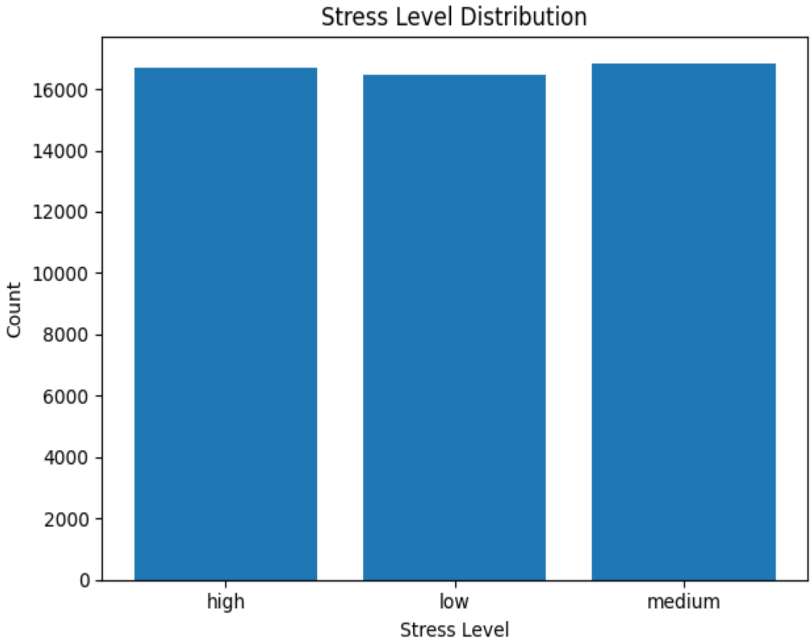


4. **Work and Physical Activity Hours**





5. Stress Level Distribution (bar)—three bands almost equal.Confirms stress labels were quota-sampled. Good for balanced modelling, but simple accuracy metrics (~33 % baseline) will undervalue improvements.



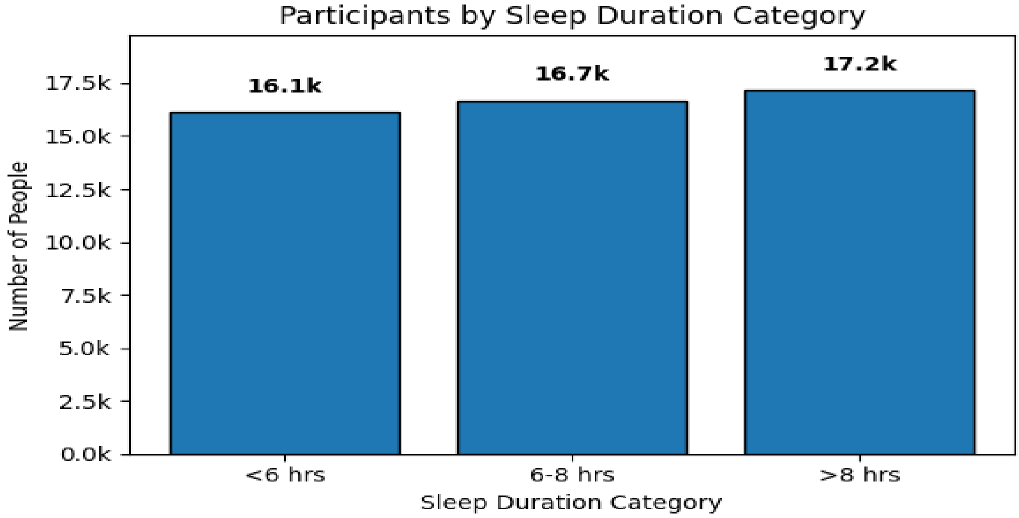
6. Sleep duration of participants :

This bar chart breaks participants into three sleep-duration buckets—less than 6 hours, 6–8 hours, and more than 8 hours—and shows how many fell into each group, with counts displayed in thousands (“k”).

<6 hrs (16 k): A substantial segment of the sample (around 16 000 people) report getting fewer than six hours of sleep, which is below the recommended minimum for healthy adults.

6–8 hrs (16 k): The middle category also represents about 16 000 participants, aligning with standard sleep guidelines (7–9 hours) and serving as a reference “healthy” group.for

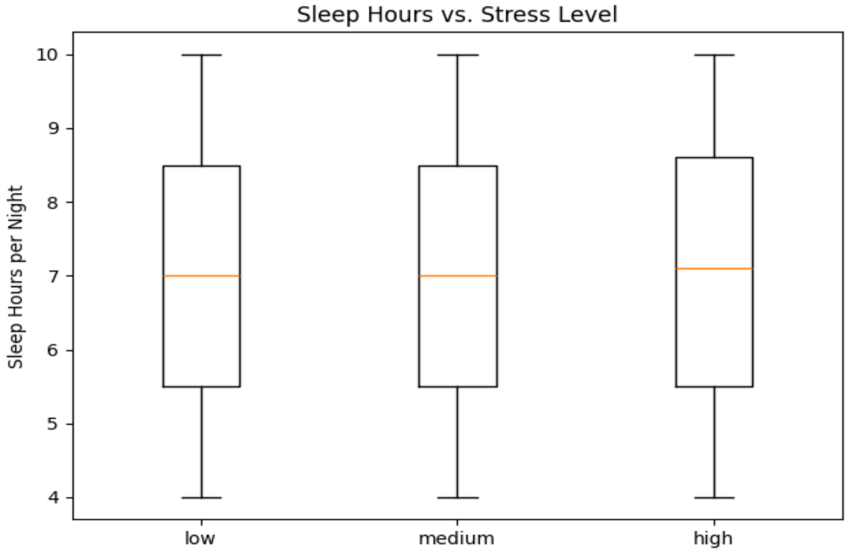
> 8 hrs (17 k): The largest group—roughly 17 000 people—report over eight hours of sleep, which can correlate with both healthy and potentially excessive rest, depending on context.

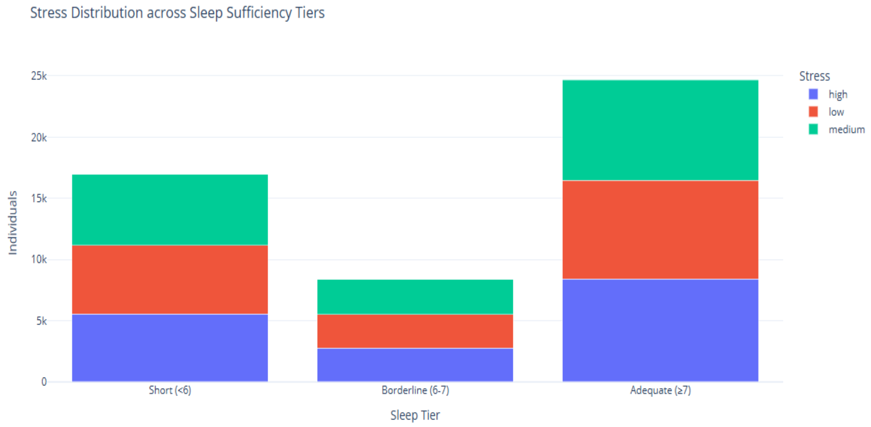


Phase II: Relationship Analysis

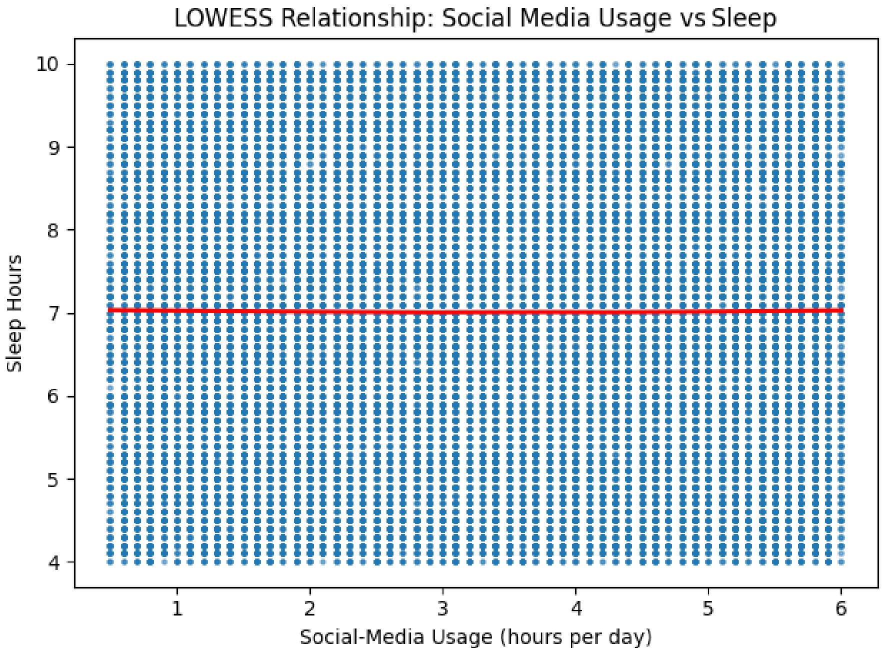
The Second step of EDA has focused on correlation analysis and some data-driven visualizations. To further depict the project objective, it is very important to understand how features relate to each other. For this, a correlation analysis and some visualizations were performed to analyze the insights and to draw the hidden patterns. This phase of the project further went on to analyze bi-variate and multivariate relationships, and it explores the understanding of how behaviors interact.

1. Sleep Hours vs Stress Level (box plot). Medians drop only slightly from *low* to *high* stress (≈ 7.1 → 7.0 h) and IQRs overlap. Stress alone isn't a strong determinant of sleep—supports the idea that multi-factor filters (stress + SM + sleep) are needed to spot at-risk individuals.

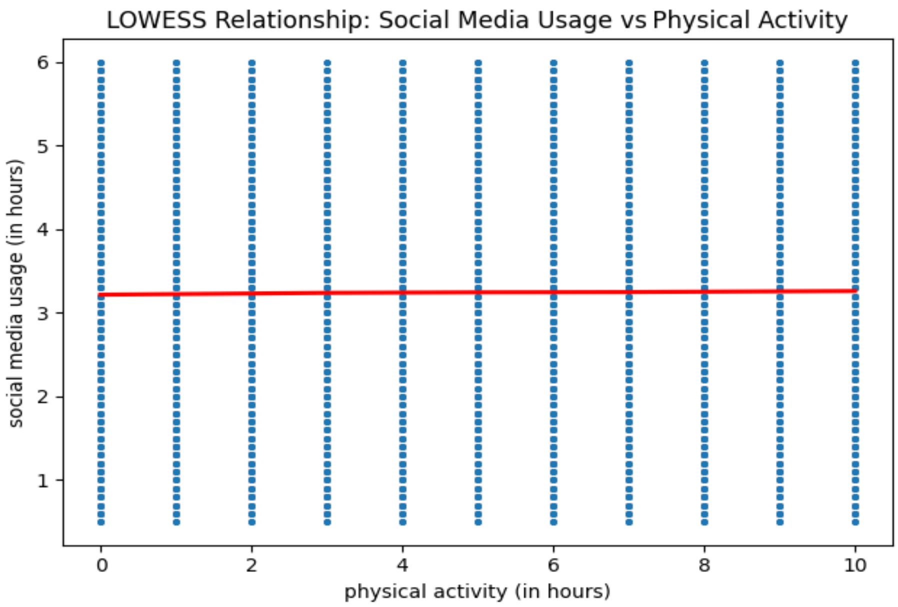




2. Social-Media vs Sleep (scatter)—dense, rectangular cloudWith the uniform design, raw scatter shows no visible slope. The red LOWESS curve stays flat confirming the data is synthetic / perfectly balanced on those two axes.

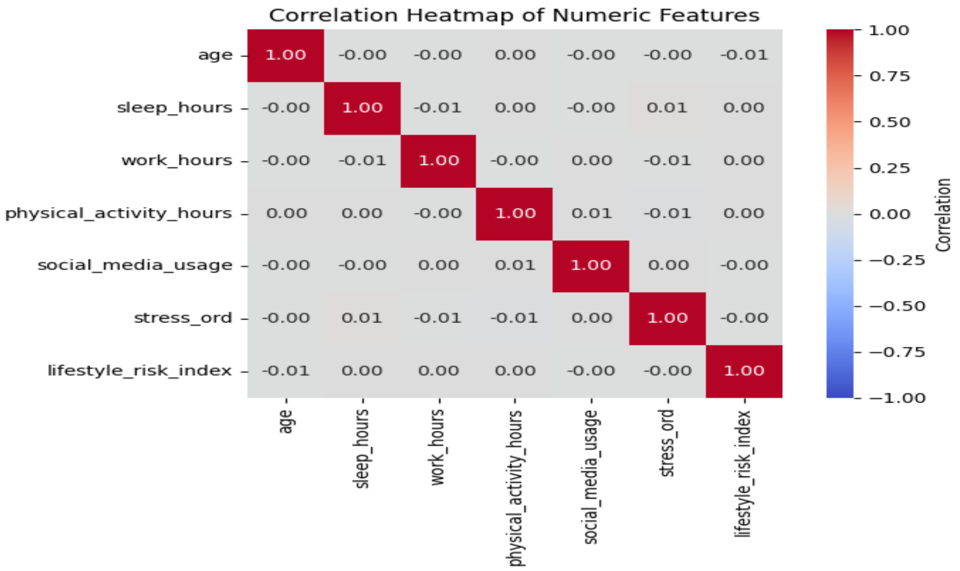


3. Physical-activity hours and screen-time hours are **statistically independent** in this dataset; being more active doesn't correlate with scrolling less (or more). **Dots** form perfectly vertical stripes: for every level of physical activity (0 – 10 h) we have the *same* spread of social-media usage (0.5 – 6 h). **Red LOWESS line** is dead-flat at ~3.2 h.



4. At the population level our dataset looks deceptively neutral—no pair of lifestyle variables shows a strong linear tie. That's because the survey was **uniformly sampled** across age, occupation, stress, and behaviour buckets.

   The absence of global correlations doesn't mean the factors are irrelevant; it simply means **their effects emerge only when we look at intersections** (e.g., high stress *and* short sleep *and* heavy screen use). This insight justifies our shift from simple correlation to composite-risk filtering where the hidden patterns finally appear.


Correlation Heatmap of Numeric Features

	age	sleep_hours	work_hours	physical_activity_hours	social_media_usage	stress_ord	lifestyle_risk_index
age	1.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.01
sleep_hours	-0.00	1.00	-0.01	0.00	-0.00	0.01	0.00
work_hours	-0.00	-0.01	1.00	-0.00	0.00	-0.01	0.00
physical_activity_hours	0.00	0.00	-0.00	1.00	0.01	-0.01	0.00
social_media_usage	-0.00	-0.00	0.00	0.01	1.00	0.00	-0.00
stress_ord	-0.00	0.01	-0.01	-0.01	0.00	1.00	-0.00
lifestyle_risk_index	-0.01	0.00	0.00	0.00	-0.00	-0.00	1.00

Takeaway: The dataset was **engineered to be flat** in demographics and primary behaviours.

Phase III: Risk Flag Distribution

The third Phase, the distribution of the risk_flag target variable was explored to assess balance across classes. This phase revealed that a medium to small portion of individuals were classified as "at risk." This Identification has been addressed as a challenge in the modeling phase. After Composite risk filtering, some trial and error metrics, this crucial phase has revealed some major insights:

1. If we consider the base filtering rule Stress = High and Sleep Hours < 6 and Social Media Usage > 4 to indicate a segment. This already isolates **1 918 people (~3.8 %)** - building a core group displaying digital fatigued risk .

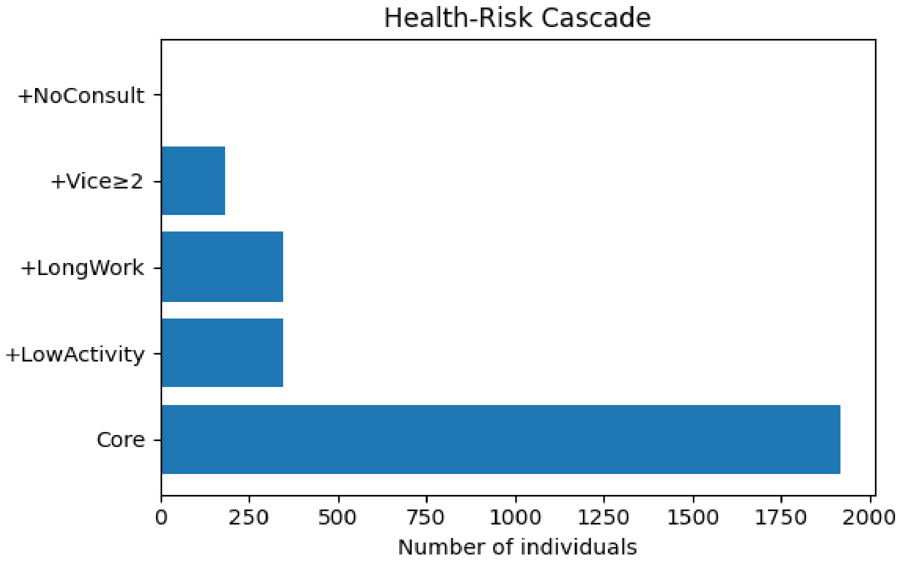
   Now we continue the incremental approach by adding more filters to the above base rule. This helps in understanding how health risks increase with each added filter as shown below:

2. Adding Low physical activity (< 2 h) reduces the core group to 347 (0.7% of total) indicating only ~18 % of the base cohort are also sedentary.
3. Adding Long work hours (> 9 h) results in exactly the same 347—so every sedentary case also works long hours.
4. Now we create Lifestyle risk index which combines smoking and alcohol consumption and use ≥ 2 for heavy/regular vice measure.

   Doing so reduces the number to 182 (0.36% of total) indicating roughly half of the sedentary-overworked cluster have significant vices.

5. Adding No consultation history reduced the number to 0 indicating every single one of those 182 has sought professional help at least once.

Thus the funnel ("Health-Risk Cascade") narrows 50,000 respondents to **1918 digital-fatigue cases**, then to **347 sedentary over-workers**, and finally to **182** vice-heavy (who smoke and drink) individuals. *Every* one of those 182 has already consulted a professional—so they're identified.



Health-risk Cascade (bar)—Core 1918 → +LowActivity ≈ 347 → +LongWork ≈ 347 → +Vice≥2 ≈ 182 → +NoConsult 0 . Visualises how each extra risk cut shrinks the core group. Biggest opportunity: physical activity & work-hours are the dominant secondary filters; hardly anyone adds "no consultation," so access to care isn't the blocker in this sample.

Our visuals show a clear, data-driven chain:

> More scrolling → less sleep → higher stress → less physical activity → Long work hours → indulging in vices concentrated mental-health burden.
>
> Interventions that simply “raise awareness” won’t cut it.Instead, focus on high-screen-time short-sleepers—the charts tell us exactly who they are, how many they are, and why they need help *now*.

But since this project relates more towards the Health care sector, even a small margin of error is considered to be addressed as a priority. This confirmation of class imbalance, urges to deploy SMOTE during machine learning model development.

Insight:
Overall, the third phase of EDA, A custom feature risk_flag was engineered based on domain-driven logic. A small fraction of users are flagged as risky, making it essential to use techniques like SMOTE to ensure models can learn these cases.

C. Feature Engineering

To enhance the data set for machine learning, a new target variable named **risk_flag** was introduced. This feature was derived by identifying behavioral patterns commonly associated with elevated mental health risks. Instead of relying on existing subjective severity labels, this newly engineered flag provided a more objective and domain-driven classification of individuals into "risk" and "no risk" categories. By converting behavioral indicators into a binary outcome, the data set became well-suited for predictive modeling. This strategic feature engineering step not only aligned the data better with the project’s objective but also helped improve the model's learning and interpretation. To classify individuals into at-risk and no-risk groups, a new feature called risk_flag was introduced using domain knowledge. This engineered target variable became the label to be predicted by the machine learning models.

---

5. Machine Learning Modeling

After feature engineering, machine learning models were built to predict mental health risk based on behavioral patterns. Two different machine learning models were applied: *Logistic Regression* and *Random Forest*. Logistic Regression, an interpretable and effective method for binary classification tasks, whereas Random Forest, a robust and capable of capturing complex feature interactions.

A. Handling Imbalanced Data

As detected earlier, the dataset exhibited class imbalance, where most of the individuals are classified as no-risk. To address this, a Synthetic Minority Oversampling Technique, popularly called SMOTE, is applied to generate synthetic samples of at-risk individuals during training. This ensured balanced training data to allow fair learning of both classes. The dataset is divided into a 70:30 ratio as Train: Test, where

1. 70% used for training (SMOTE applied here)
2. 30% used for testing (kept imbalanced to reflect a real-world scenario)

B. Logistic Regression (SMOTE)

The Logistic Regression model demonstrated an accuracy of **88.31%** on the test dataset. More importantly, it achieved an impressive **93% recall** for the at-risk class. This indicates that the model was able to correctly identify a large majority of risky individuals, which is essential for preventive healthcare.

However, the model produced higher false positives, labeling some no-risk individuals as at-risk. In the mental health domain, this trade-off is often acceptable because the cost of missing a truly at-risk individual is much higher than a false alarm. Thus, Logistic Regression prioritized sensitivity (recall), aligning well with the project's goals.

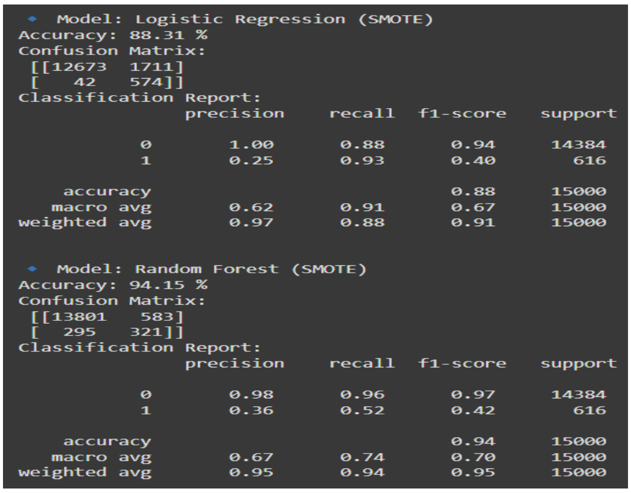
C. Random Forest (SMOTE)

The Random Forest model achieved a higher overall accuracy of **94.15%**. However, its recall for the at-risk class was only **52%**, which is notably lower than Logistic Regression. Although Random Forest offered a better balance between precision and recall for both classes, its lower sensitivity towards at-risk individuals makes it less suited for scenarios where missing high-risk cases can have serious consequences.

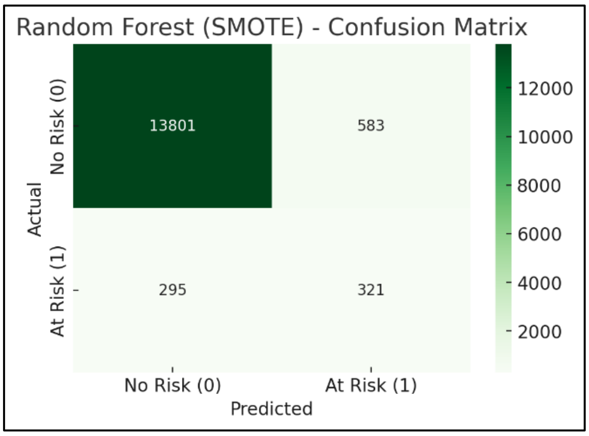
In summary, Random Forest demonstrated stronger performance on overall accuracy but fell short on identifying rare risky cases compared to Logistic Regression. Overall, the two models performed well, but with different strengths and trade-offs.

D. Final Model Selection

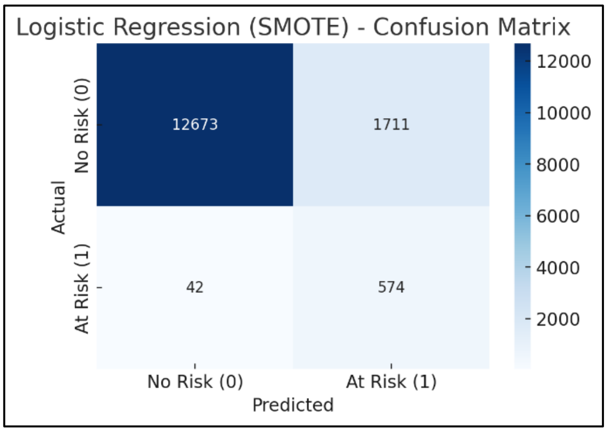
In healthcare and mental health risk prediction, the priority is to maximize recall for at-risk individuals to ensure timely intervention. Based on this principle, Logistic Regression (with SMOTE) was selected as the final model. While it introduced some false positives, its high recall ensures that the majority of high-risk individuals are flagged, supporting the project's objective of proactive mental health risk detection and prevention.



```
 🔹 Model: Logistic Regression (SMOTE)
Accuracy: 88.31 %
Confusion Matrix:
 [[12673  1711]
 [   42   574]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.88      0.94     14384
           1       0.25      0.93      0.40       616

    accuracy                           0.88     15000
   macro avg       0.62      0.91      0.67     15000
weighted avg       0.97      0.88      0.91     15000


 🔹 Model: Random Forest (SMOTE)
Accuracy: 94.15 %
Confusion Matrix:
 [[13801   583]
 [  295   321]]
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.96      0.97     14384
           1       0.36      0.52      0.42       616

    accuracy                           0.94     15000
   macro avg       0.67      0.74      0.70     15000
weighted avg       0.95      0.94      0.95     15000
```



Random Forest (SMOTE) - Confusion Matrix



Logistic Regression (SMOTE) - Confusion Matrix

✓

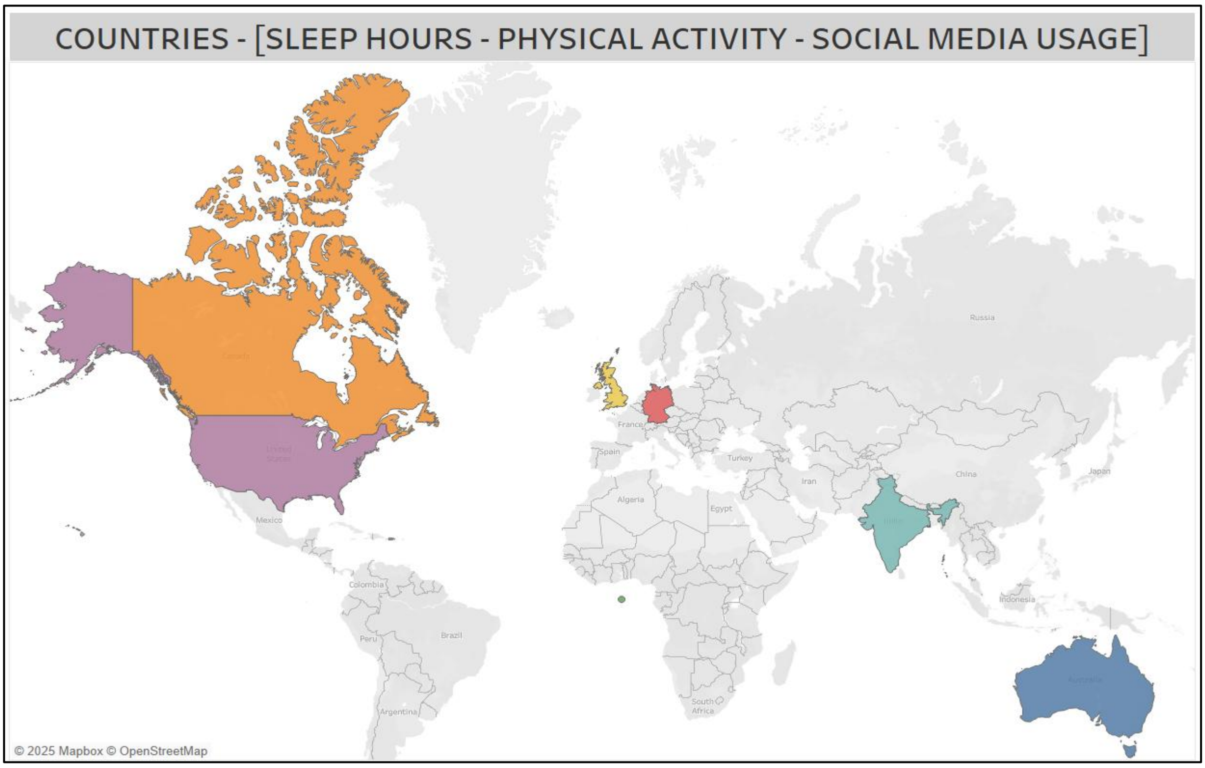
1. Logistic Regression (with SMOTE) was selected as the final model.
2. Reason: In healthcare services, a False Positive is better than a False Negative.
3. It prioritizes catching at-risk users, aligning with the project's objective of preventive mental health risk detection.

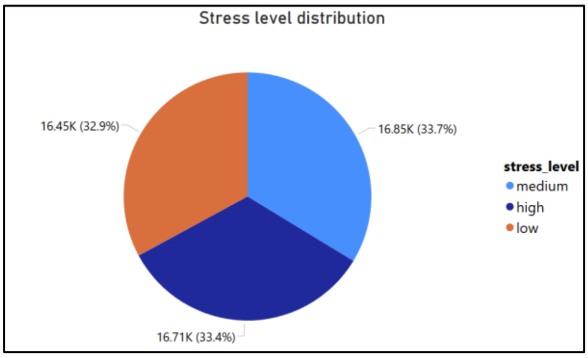
---

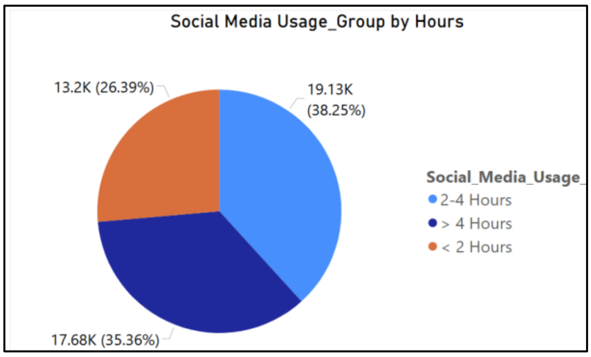
6. Visualization and Insights

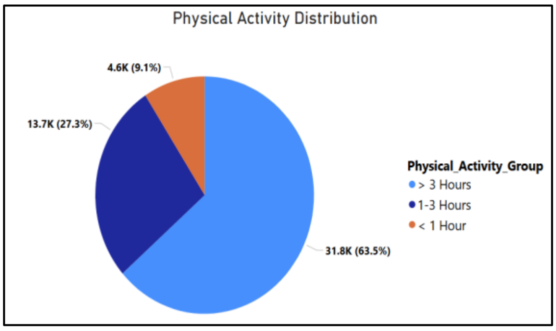
The tail end of the project focuses on visualizations. It plays a critical role in show-casing the behavioral data, presenting analyzed relationships between attributes, and revealing machine learning model performance. These visuals not only support data exploration but also help in communicating key insights effectively.

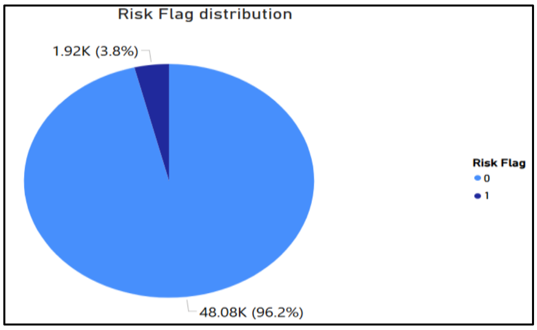
Power BI and Tableau generated Dashboard Visuals:

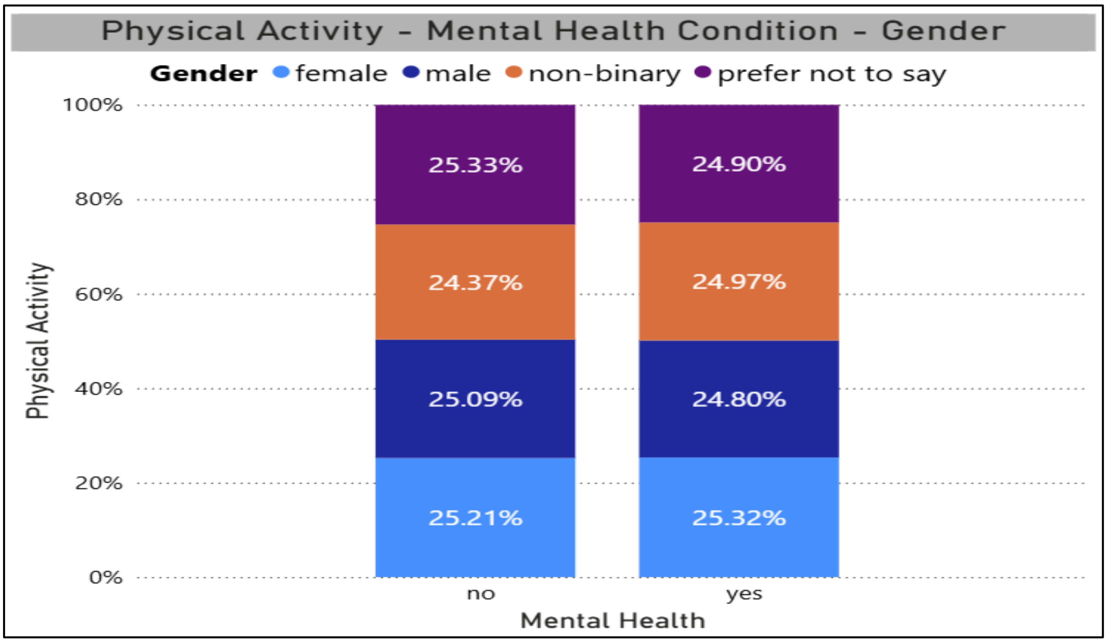












[i] Pictures are Tableau-generated Visuals. For more visuals and interactive features, Please visit Dashboards provided links provided at the end of the report.

---

7. Result

As it unfolds, the project successfully demonstrates the potential of behavioral patterns linked with social media usage to predict mental health risk through machine learning models. By engineering meaningful features and addressing class imbalance using the SMOTE technique, the models achieve accuracy results. As already mentioned, the Logistic Regression model emerged as the preferred solution due to its higher recall for at-risk individuals, which is essential in the healthcare sector and detects 3.8 % are at risk

Overall, this project lays a strong foundation for developing data driven mental health risk prediction systems, which can support early interventions and contribute to the research work in the public health care domains

---

8. Conclusion

In this project, machine learning models were developed to predict mental health risk using social media and behavioral patterns. Through careful data preparation, feature engineering, and handling class imbalance with SMOTE, Logistic Regression emerged as the preferred model due to its high recall, ensuring that at-risk individuals are identified effectively.

Visualizations and model performance analysis confirmed that Logistic Regression is well-suited for preventive healthcare scenarios, where missing risky individuals can have serious consequences. Overall, the project successfully demonstrates the power of data analytics in supporting early mental health risk detection and decision-making.

8. Discussion and Insights

While this project successfully achieved the goal of predicting mental health risk using behavioral and social media patterns with some interactive visuals, a few important insights and future opportunities were identified and worth mentioning.

A. Key Discussion Points

1. Logistic Regression was preferred due to its high recall, ensuring at-risk individuals were rarely missed.
2. Random Forest showed higher accuracy but was less sensitive to detecting risky cases.
3. Addressing class imbalance using SMOTE was critical and highly effective for fair learning.

B. Limitations

1. The dataset scope was limited to behavioral indicators; no clinical or psychological factors were included.
2. Models were trained on static data and may require validation in dynamic, real-world setting

9. Future Work

1. Integrate additional features like anxiety levels and mood indicators.
2. Explore advanced models (e.g., XGBoost, Neural Networks) for improving prediction precision.
3. Validate models on different user groups for broader applicability.
4. Develop interactive dashboards for real-time risk monitoring.

9. References

1. Dataset Source — Mental Health Lifestyle Dataset: https://zenodo.org/records/14838680
2. World Health Organization (WHO) — Mental Health Overview: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response
3. Python Libraries Documentation (Pandas, NumPy, Scikit-learn, imbalanced-learn, Matplotlib, Seaborn): https://www.python.org/doc/
4. SMOTE — Synthetic Minority Oversampling Technique (imbalanced-learn official): https://imbalanced-learn.org/stable/over_sampling.html
5. Tableau Community https://community.tableau.com/
6. Power BI https://learn.microsoft.com/en-us/power-bi/

10. BI Dashboard Links:

https://public.tableau.com/app/profile/sai.chaitanya.munagala/viz/MentalHealthAnalysis_17466541122880/COUNTRYMAP?publish=yes

https://public.tableau.com/app/profile/sai.chaitanya.munagala/viz/MentalHealthAnalysis-OccupationandSocialMediaHours/OCCUPATION-SOCIALMEDIAHOURS?publish=yes

https://public.tableau.com/app/profile/kapil.reddy8547/viz/IMPACTOFSOCIALMEDIAONMENTALHEALTHRISKINSIGHTS/Dashboard3?publish=yes

Main Dashboard link for Risk Analyses:
https://public.tableau.com/app/profile/kapil.reddy8547/viz/IMPACTOFSOCIALMEDIAONMENTALHEALTHRISKINSIGHTS/Dashboard2?publish=yes