JTSE
JOURNAL OF TRANSACTIONS IN SYSTEMS ENGINEERING

Journal of Transactions in Systems Engineering
https://journals.tultech.eu/index.php/jtse.
ISSN: 2806-2973
Volume 3, Issue 3
DOI: https://doi.org/10.15157/JTSE.2025.3.3.471-488
Received: 18.10.2025; Revised: 12.12.2025; Accepted: 20.12.2025

TULTECH

Research Article

Machine Learning Analysis of Social Media Usage Patterns and Mental Health Indicators

Dhoha Raad Hussein[1], Ali Subhi Alhumaima[2], Hussein Alkattan[3,4], Mostafa Abotaleb[5*]

[1] College of Engineering, University of Diyala, Diyala, Iraq
[2] Electronic Computer Centre, University of Diyala, Diyala, Iraq
[3] Department of System Programming, South Ural State University, Chelyabinsk, Russia
[4] Directorate of Environment in Najaf, Ministry of Environment, Najaf, Iraq
[5] Engineering School of Digital Technologies, Yugra State University, Khanty Mansiysk, Russia
***abotalebmostafa@bk.ru**

Abstract

The rapid growth of social media has created new opportunities for connection but has also raised concerns about its impact on mental health. This study investigates how demographic factors, digital behaviour, and self-reported psychological indicators jointly relate to users' mental states. Using an open dataset of 5000 social media users, we analyse numerical and categorical variables including age, gender, daily screen time, social media time, counts of positive and negative interactions, sleep duration, physical activity, anxiety, stress, mood, and a three-level mental-state label (Healthy, At_Risk, Stressed). Descriptive statistics and correlation analysis show that longer daily screen and social media time are strongly associated with higher stress and anxiety and lower mood, while sleep and physical activity display the opposite pattern. K-means clustering applied to combined behavioural and psychological features reveals three coherent user profiles that align with the Healthy, At_Risk, and Stressed categories, highlighting a clear gradient from balanced to high-risk digital lifestyles. A decision-tree classifier trained only on behavioural features (excluding anxiety, stress, and mood to avoid target leakage) achieves an overall accuracy of about 97% on a held-out test set and provides interpretable if then rules linking specific usage patterns to mental states. The results emphasise that intensive, unbalanced social media use especially when coupled with reduced sleep and low physical activity is strongly linked to adverse mental-health outcomes, and they illustrate how simple machine-learning models can support early risk detection based on non-intrusive behavioural data.

Keywords: Data Mining; K-means Clustering; Decision Tree; Behavioural Analytics; Social Media; Mental Health; Screen Time; Anxiety; Stress; Mood.

INTRODUCTION

Mental health disorders are a significant public health problem in heterogeneous societies and sociocultural contexts. The fact that we have systematic differences by gender in the prevalence, symptom pattern and risk factors for common mental disorders (CMD) is well established. With regard to healthcare professionals in the context of COVID-19 pandemic, women experienced more anxiety, depression and sleep difficulty than men, indicating job-related risks and gender determinants [1]. The above sex differences were replicated in community-based samples of adolescents, where girls have higher levels of internalizing symptoms and use social support networks to a greater extent than boys [2]. Elite sport, that can think of a protective context, also presents considerable gendered patterns in mental health symptoms and risk factors among athletes [3]. Across the general community, differences also emerge in psychological well-being: these have been seen between men and women on the autism spectrum [4], as well as among older Internet users who express heterogeneous beliefs about COVID-19 risks [5]. Results highlight the need to take gender, age, and social role into account in investigating determinants of psychological well-being.

Work-related environment is also a significant factor of mental wellbeing. Job control, workload, and social support can interact with sex, income level, and occupation to influence positive mental health of workers [6]. Amid the COVID-19 pandemic, healthcare and training disruptions (HDTD) have caused a great deal of stress, particularly among professions, such as orthodontic trainees for whom both professional practice and psychological status have been largely affected [7]. Concomitantly, maternal mental health became a matter of particular concern in many settings worldwide being the indicator that pandemic-related stressors exacerbated pre-existing vulnerabilities among pregnant and postpartum women [8, 9]. Other marginalized groups such as LGBTQIA+ people have a history structural stigma, discrimination and poorly resourced service provision which collectively increases risk factors for mental ill-health [10]. Overall, these lines of research highlight the complex ways in which mental health is influenced by gender, socioeconomic position and stressors in context.

The broad penetration of digital technology and social media has added an additional layer to this complexity. Social media, in turn, have become important sites for identity construction as well as relationship maintenance and information consumption. Bibliometric analyses highlight the strong academic interest for understanding the nexus between social media use, and mental health and youth outcomes, focusing largely on the duality of online engagement as both a resource and risk factor [8]. Building on social cognitive theory, Bandura posited that people learn and adopt behaviours by observing, comparing themselves with others contrastively in a social context, and receiving reactions from their world [11]. Now, in digital and algorithmically-mediated ecologies, these life-writing and bonding practices are increasingly enacted in curated feeds of images; image-centric sites themselves that position preservation as an option for only temporary registry.

Empirical studies are beginning to chart the psychological effects of these environments. Unfortunately, Instagram is associated with symptoms of depression, negative social comparison and following a high number of strangers, meaning some engagement can exacerbate the vulnerability [12]. The study of friend-networking sites in their infancy reported positive effects from making online connections for adolescents when the use is constructive, benefiting social self-esteem and perceived social support

[13]. Nevertheless, following studies during the COVID-19 pandemic have demonstrated that high levels of anxiety or depression and health anxiety originated due to sociodemographic factors including gender, possible to be intensified by media overuse and exposure to upsetting content [14-17]. In the case of adolescents, stress was found to be a significant predictor for problematic internet use and addictive behaviour, functioning via mediated–moderation effects which includes strategies to handle stress and the social environment [15]. Financial strain and job insecurity may also interact with internet use, increasing stress and guiding job-search behaviours which could have implications for trajectories of mental health [16].

Screen time and psychological well-being This association has been the subject of increased attention. Population-based evidence indicates that increased screen time is correlated with lower psychological well-being in youth, for example, decreased curiosity, self-control and emotional stability [18-21]. Night-time social media use has been associated with worse sleep, higher anxiety and depression symptoms and lower self-esteem in adolescents [22]. While these have traditionally been useful spaces, and continue to be utilized for this purpose, newer applications such as Snapchat encourage ephemeral engagement that arguably fosters social connection and support scaffolding (for example through the transmission of “small moments” from daily life [23]. But perhaps they also offer an avenue for never-ending social comparison that is associated with lower self-esteem and more depressive affect [24]. Reasons for use, such as presentation of self and narcissistic satisfaction derived through Instagram can also affect the psychological implications of Internet usage [25]. Problematic Facebook use in adolescents and young adults has been linked to specific psychopathological profiles, such as mood dysregulation and interpersonal impairment [26]. These results suggest that it is not just volume, but pattern and quality of social media use that are essential in mental health related outcomes.

Reliable measurement instruments are needed to assess psychological complaints in such vulnerable populations. The PHQ-9 offers a highly validated brief measure of the intensity of depression symptoms applicable in clinical and community samples [27]; similarly, the GAD-7 is a briefer scale to evaluate generalized anxiety disorder [28]. These scales are widely used in epidemiological studies and clinical screening since they enable comparability between studies. Furthermore, rigorous research methods and ethical precautions are necessary when working with sensitive mental health information. Classic references on research methodology emphasize clear design, sampling strategies and analytic rigor [19], legal and ethical dialogues focus on informed consent, data protections, and the respect of autonomy for human subjects in research [20]. Verification methods, including triangulation, member checking and audit trails are used to ensure the reliability and validity of data particularly in qualitative and mixed methods research [29, 30].

However, despite this large body of literature several issues remain. First, several previous studies focus on particular populations (e.g., healthcare workers, athletes, young people or LGBTQIA+ communities) and typically employ self-report questionnaires using a cross-sectional design [1–3, 8–10, 14, 17]. Although these approaches demonstrate significant relationships, they rarely include detailed information on behaviour (e.g., measured screen time, specific platforms of social media use, and number of positive and negative online interactions). Second, gender in the interplay of digital behavior and mental health is mostly considered separately, whereas gender has an impact as a moderator on exposure to stressors and access to social support [2, 5, 6, 14]. Third, few studies apply data-driven ML methods to identify latent profiles of social media use and mental health symptoms, or develop predictive models that could aid the early detection of individuals at elevated risk for distress.

This work brings together and closes several of these gaps by using statistical and machine-learning methodologies to analyse a substantial mental-health oriented social-media dataset, containing user demographic details behavioral markers (daily screen time, social-media time, positive-negative interactions indicated in posts) and self-reported measures reflecting mental health (anxiety, stress, mood, mental-state categories).

Expanding on previous research connecting screen time exposure, sleep, social comparison, and psychopathology [12, 21–24], the current study first offers a descriptive-correlational account of relations across these variables. It performs clustering to discover unique user profiles by fusing all behavioural traits and psychological features, then applies supervised models to predict mental state using these cues. Guided by social cognitive theory [11] and broader gendered mental health risk literature [1–6, 10, 14], the analysis seeks to identify the extent that certain patterns of social media use and lifestyle behaviour are associated with psychological outcomes in a diverse population.

DATA AND METHODOLOGY

Dataset

The empirical part of this work is structured using the data source "Social Media Mental Health Indicator Dataset" provided by Shinde in the Kaggle platform [31]. kaggle. social media usage patterns and indicators of mental health. The dataset comprises a single CSV (5000 anonymized records represented as 15 fields), which includes demographic features, digital behaviour scores and self-reported psychological dimensions.

- Every record represents a user on a particular social media site and features the below listed demography and platform information:
- person_name: A pseudonymized value that is only an identifier for each individual record and does not serve as a predictive characteristic.
- age (years): A continuous variable of 13-69 years, representing a wide range from adolescents to older adults.
- gender: A categorical variable with levels (Male, Female, Other) for investigating the gender differences in digital behaviour and mental state, comparable to the previous studies on gender and mental health.

- date: The date of observation or data entry fecf that is specifically not interpretable with respect to a temporal series in this work but should be useful for subsequent longitudinal extensions.
- platform: The platform from which the participant presented a particular model (eg Facebook, Instagram, etc) which offers information in context about the digital environment though is not directly related to the focus of this current investigation.

The dataset also includes behavioral proxies for the degree and nature of daily online engagement:

- daily_screen_time_min: Total daily screen time (min, combining use of digital devices for all purposes).
- social_media_time_min: The portion of Screen time that is social media platforms.
- negative_interactions_count: Number of negative or conflictual online interactions (e.g., fights, criticism, harassment that participants felt they had received) experienced within the reference period.
- positive_interactions_count The number of supportive or pleasurable interactions (e.g., friendly chats, likes, affirming comments).
- sleep_hours: Self-reported average total sleep duration per night (hour).
- physical_activity_min: Self-reported minutes of daily physical activity.
- These behavior-related variables are also consistent with recent studies on screen time, sleep and PA as antecedents of PWB and digital overload.

Lastly, the dataset contains psychological indicators that describe the mental health status of each individual:

- anxiety_level: Integer scale (1-4) indicating the perceived intensity of anxiety.
- stress_level: An ordinal variable (5–9) indicating the perceived level of stress.
- mood_level: An ordered score (4–7) representing mood globally.
- mental_state: A qualitative tag type, based on the above indices that can be classified one of three states: Healthy, At_Risk and Stressed. The distribution in the sample analysed is Stressed = 4601, Healthy = 341 and At_Risk = 58, showing that the target variable is highly imbalanced with a majority of stressed users.

The CSV on Kaggle was takes as input before any structural modification. A quick look at data completeness showed no missing values in any variable (all 15 columns contain complete values for N = 5000) and imputation was not necessary. Numerical features were preserved as-is, categorical variables (gender; platform; mental_state) were only encoded during modelling if required. Inspection of outliers verified that the ranges of observed screen and social media time, sleep, and physical activity were internally consistent and plausible for intense use of social media.

The Kaggle dataset is not provided with personally identifying information (like true names, address domicile or IP addresses), and all entries were either pseudonymized or transformed at non-identifiable level [31]. As such, here we present a secondary analysis of fully anonymized data in accordance with ethical guidelines for the re-use of publicly available research datasets [19, 20, 29]. In conclusion, this data set provides a concise but rich summary of the relationship between demographic characteristics, digital engagement and self-reported mental health that is well-suited to the kind of statistical and machine-learning analyses carried out in this work.

This heatmap shows the pairwise Pearson correlation coefficients between all of the numerical variables from the social media mental health dataset (the colour intensity and numeric value in each cell represent the strength and direction, respectively). Positive strong relationships were indicated by warm yellow color in the cells, whereas negative strong relat ionships we re reflected in dark purp le colors. The heatmap demonstrates that daily_screen_time_min and social_media_time_min are significantly positively correlated with anxiety_level and stress_level, which means that the more time for screen and social media users spend, the higher they report in anxiety and stress. The same set of variables is highly negatively correlated with sleep_hours, physical_activity_min and mood_level indicating that higher screen use leads to less sleeping time, minimal physical activity and bad mood. Physical_activity_min and sleep_hours are strongly positively correlated with each other themselves generally indicating a healthier profile of lifestyle, and both have negative relation with anxiety, stress, and positive with mood. The negative relationship of stress_level and mood_level accentuates that, the resultant rises in perceived stress are accompanied with reduced general mood. In general, the heatmap provides a nice summary of the structure within dataset and graphically supports our conjecture that intensive digital engagement is linked to lifestyle factors and mental health indicators.

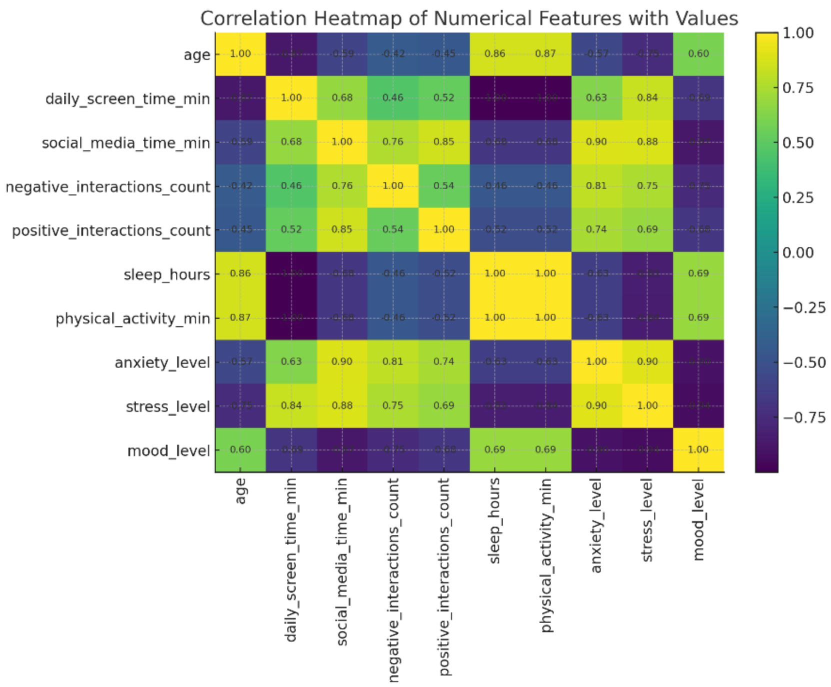

Figure 1. Correlation Heatmap of Numerical Values with features.

Figure 2 presents the distribution of age, daily_screen_time_min, social_media_time_min, negative_interactions_count, positive_interactions_count sleep_hours, physical_activity_min as well as anxiety_level and stress_level overlaid with bars for the three mental-state groups. Across panels, the Stressed are overrepresented and generally skewed towards more risky values: younger ages, higher daily and social media screen time, more negative & positive interactions, shorter sleep durations, less physical activity (PA), anxiety score and stress level. In contrast, the Healthy group presents crowded peaks at lower screen and social media times, much lower negative interactions, longer sleep, higher physical activity, and low anxiety and stress rates in the corresponding histograms. At_Risk falls intermittently to the middle between Healthy and STressed, potentially overlapping with both. Collectively, these distributions depict that specific mental states have comorbid profile in both behaviour and in the underlying psychological measures and provide, at a glance, pictorial evidence to aid interpretation of the statistical findings on intensive vs unbalanced social media use as associated with worse mental health outcome.

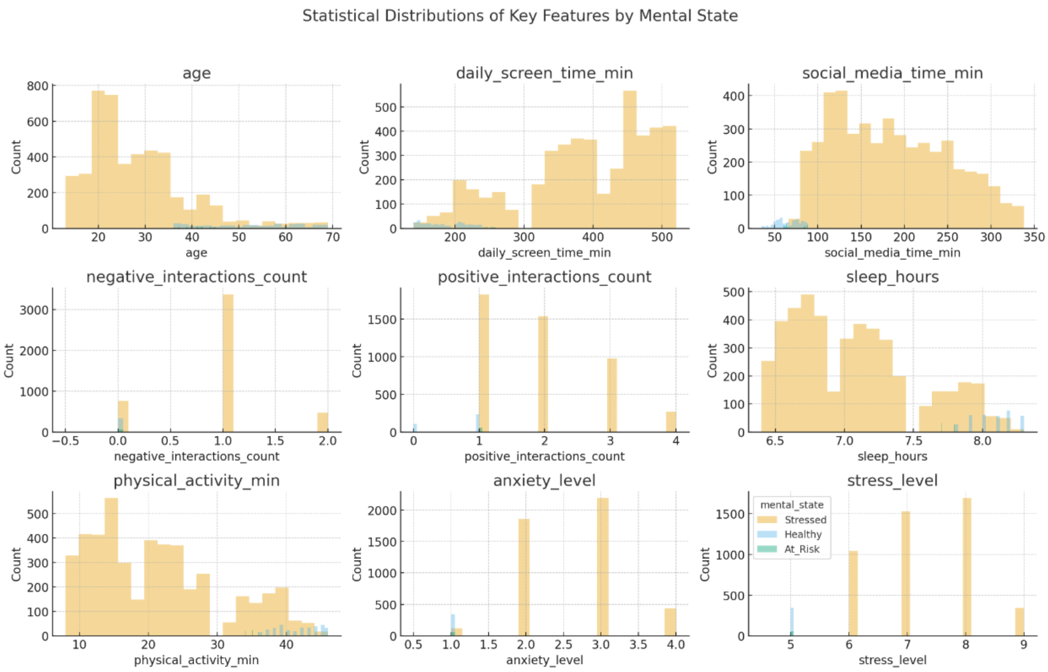

Figure 2. Statistical Distributions of Key Features by Mental State.

Feature Vector Representation

Each user is represented by a feature vector x that combines demographic, behavioural, and psychological variables:

$$\mathrm{x} = (x_1, x_2, \dots, x_{10}) \tag{1}$$

were

- x_1 = age
- x_2 = daily_screen_time_min
- x_3 = social_media_time_min

- x_4 = negative_interactions_count
- x_5 = positive_interactions_count
- x_6 = sleep_hours
- x_7 = physical_activity_min
- x_8 = anxiety_level
- x_9 = stress_level
- x_{10} = mood_level.

For numerical stability and to avoid scale dominance, each continuous feature x_j was standardized using z score normalization, see equation (2):

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (2)$$

were

- x_{ij} is the value of feature j for user i,
- μ_j is the mean of feature j over all users,
- σ_j is the standard deviation of feature j.

The standardized vectors z_i are then used as inputs for clustering and classification.

K-Means Clustering Model

K -means is used to discover latent groups of users based on their usage patterns and mental health indicators, without using the mental_state label during training-Let:

- N be the number of users (here $N = 5000$).
- K be the chosen number of clusters (here $K = 3$),
- $z_i \in R^p$ be the standardized feature vector of user i,
- μ_k be the centroid (mean vector) of cluster k,
- $r_{ik} \in \{0,1\}$ indicate the assignment of user i to cluster k (1 if user i belongs to cluster k, 0 otherwise).

The K-means algorithm minimizes the within-cluster sum of squared distances, see equation (3):

$$J_{\text{K-menas}} = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \|z_i - \mu_k\|^2 \quad (3)$$

subject to equation (4):

$$\sum_{k=1}^{K} r_{ik} = 1 \ \text{ for all } i = 1, \dots, N \quad (4)$$

The optimization proceeds iteratively in two steps:

1. Assignment step

For each user i, assign to the nearest centroid:

$$r_{ik} = \begin{cases} 1, & \text{if } k = \arg\min_l \|z_i - \mu_l\|^2 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

2. Update step

For each cluster k, recompute the centroid as the mean of all points assigned to it via equation (6):

$$\mu_k = \frac{\sum_{i=1}^{N} r_{ik} z_i}{\sum_{i=1}^{N} r_{ik}} \tag{6}$$

These two steps are repeated until convergence (no change in assignments or a small change in $J_{\text{K-means}}$). In our context, the resulting clusters are interpreted as Healthy. At_Risk, and High-Stress behavioural psychological profiles, based on the average values of stress, anxiety, mood, sleep, and screen time in each cluster.

Decision Tree Classification Model

To predict the mental_state label (Healthy, At_Risk, Stressed) from the feature vector z, we use a Decision Tree classifier. A decision tree recursively partitions the feature space using simple rules of the form:

$$if\ z_j \leq t\ then\ go\ to\ left\ child\ node, else\ go\ to\ right\ child\ node$$

where z_j is a standardized feature (e.g., stress_level, daily_screen_time_min, sleep_hours) and t is a learned threshold.

At each node, the algorithm selects the feature j and threshold t that best separate the classes, typically by minimizing an impurity function. In this study we use the Gini impurity:

$$G = 1 - \sum_{c=1}^{C} p_c^2 \tag{7}$$

were

- C is the number of classes (here $C = 3$).
- p_c is the proportion of samples of class c in that node.

For a candidate split (j, t) that divides the data at a node into a left subset L and a right subset R, with sizes N_L and N_R, the impurity after the split is:

$$G_{\text{split}}(j,t) = \frac{N_L}{N_{\text{node}}} G_L + \frac{N_R}{N_{\text{node}}} G_R \tag{8}$$

were

- $N_{\text{node}} = N_L + N_R$.
- G_L and G_R are the Gini impurities of the left and right subsets.

The information gain (or reduction in impurity) is:

$$\Delta G(j,t) = G_{\text{parent}} - G_{\text{split}}(j,t) \tag{9}$$

The tree chooses the split (j^*, t^*) that maximizes ΔG. This procedure is applied recursively until stopping criteria are met (e.g., maximum depth, minimum number of samples per leaf).

At a leaf node m, the estimated probability of class c is:

$$\hat{p}_c^{(m)} = \frac{N_c^{(m)}}{N^{(m)}} \tag{10}$$

were

- $N_c^{(\mathrm{m})}$ is the number of training samples of class c in leaf m,

- $N^{(\mathrm{m})}$ is the total number of samples in that leaf.

The predicted mental state for a new user with feature vector z is:

$$\hat{y}(z) = \arg\max_{c} \hat{p}_c^{(m(z))} \quad (10)$$

where $m(z)$ is the leaf node reached by traversing the tree according to the user's feature values.

RESULTS

The overall descriptive statistics of the numerical variables in the social media mental health dataset consisting of 5000 users is presented in Table 1. It reports the mean, standard deviation, minimum and maximum values of age, screen time, social media use, interaction counts (connection strength), physical activity and sleep duration alongside psychological variables such as anxiety levels, stress levels and mood. From this table, we can observe a profile most prevalent in the dataset - young adult with moderate daily screen and social media time, a moderate number of positive and negative interactions, about seven hours of sleep, low to moderate physical activity. The psychological scores bring an overall biased of high stress and intermediate mood, meaning a big part of the users are probably under some sort of mental stress related to their digital behaviour.

Table 1. Descriptive statistics of numerical variables in the social media mental health dataset (N = 5000).

Variable	Mean	SD	Min	Max
Age (years)	29.95	12.28	13.0	69.0
Daily screen time (min/day)	373.06	106.00	140.0	520.0
Social media time (min/day)	175.33	71.21	35.0	338.0
Negative interactions (count)	0.86	0.56	0.0	2.0
Positive interactions (count)	1.84	0.94	0.0	4.0
Sleep hours (h/night)	7.13	0.53	6.4	8.3
Physical activity (min/day)	22.69	10.60	8.0	46.0
Anxiety level (1–4)	2.51	0.79	1.0	4.0
Stress level (5–9)	7.11	1.06	5.0	9.0
Mood level (4–7)	5.63	0.76	4.0	7.0

Table 2 shows the distribution of Healthy, At_Risk, and Stressed for the gender groups. It is a fact table with the users' count in the 3 mental state categories by females, males and others and sum by gender in addition to the total count of rows. This table also indicates both the class imbalance of mental_state variable and any obvious gender differences in terms of the mental health status. ‘Stressed’ is the largest cluster for both genders in this dataset, meaning that whereas males or females belong to a ‘high stress group’. The above comparison is just partial, because the total numbers also indicate that female and male users are 94% of stressed people While subjects referring to “other” gender are much less but still requiring more effort. Such distribution indicates that increased stress is a common issue among the sample of the population (ie, no single sex has exclusive copyright), despite possibly having different rationales as reflected in other gender–mental health literature.

Table 2. Distribution of mental state by gender.

Gender	At_Risk	Healthy	Stressed	Total
Female	25	181	2268	2474
Male	32	153	2242	2427
Other	1	7	91	99
Total	58	341	4601	5000

Table 3 presents mean scores of major behavioural and psychological parameters in different mental states to enable a straightforward comparison of the typical characteristics among three groups At-Risk, Healthy and Stressed. It demonstrates the extent to which age, daily screen time, social media time, interaction counts, sleep, physical activity, anxiety and stress and mood differ among respondents in these three segments. The Healthy group has higher age, lower screen and social media use, longer sleep, higher physical activity but less anxiety and stress with high mood scores. The At_Risk category is between them, being of intermediate behavioural, psychological level, denoting an intermediate vulnerability. In contrast, the Stressed were younger, spent significantly longer on daily and social media screens, had more negative interactions, reported less sleep and physical activity and had higher anxiety stress level lower mood. This table clearly shows a behavioural and emotional gradient over mental statuses, which would support the assumption that heavy, unbalanced digital engagement is related with poorer mental health.

Table 3. Mean behavioural and psychological indicators by mental state.

Mental state	Age (years)	Daily screen time (min)	Social media time (min)	Neg. interactions	Pos. interactions	Sleep (h)	Physical activity (min)	Anxiety level	Stress level	Mood level
At_Risk	40.69	250.93	71.72	0.00	1.00	7.75	34.91	1.00	5.00	7.00
Healthy	52.55	184.97	65.41	0.00	0.68	8.08	41.50	1.00	5.00	7.00
Stressed	28.14	388.54	184.78	0.94	1.93	7.06	21.15	2.64	7.29	5.51

Table 4 displays the correlation matrix for the main behavioural and mental health variables, which reflects linear associations between screen time, social media use, positive and negative interactions as predictors and sleep duration, physical activity, anxiety and stress indicators. These values in each cell show the extent of relationship and direction for each pair of variables: values near 1 or −1 indicate strong positive or negative associations, respectively. The table tells a story: The more screen time and social media the kids used, the higher their levels of anxiety and stress. On average, increased screen time throughout the day corresponded to lower sleep scores; decreased physical activity; more anxiety or concentration problems; as well as less happiness, curiosity and ability to finish tasks. Likewise, the negative interactions show positive associations with anxiety and stress and negative relations with mood; however, sleep and physical activity have an opposite pattern: they are negatively associated to stress and positively related to mood. However, the very strong negative correlation between stress and mood ($r = -.84$) shows that a higher level of mood is associated with a lower level of stress. In sum, this table offers the statistical foundation for the conceptual connection of heavy digital engagement, diminished restorative activities and psychological distress.

Table 4. Correlation matrix of key behavioural and mental health variables.

Variable	Screen time	Social media	Neg. int.	Pos. int.	Sleep	Physical act.	Anxiety	Stress	Mood
Daily screen time (min)	1.000	0.681	0.457	0.524	-0.999	-1.000	0.629	0.836	-0.695
Social media time (min)	0.681	1.000	0.761	0.855	-0.680	-0.681	0.897	0.883	-0.867
Negative interactions	0.457	0.761	1.000	0.537	-0.458	-0.458	0.792	0.753	-0.753
Positive interactions	0.524	0.855	0.537	1.000	-0.525	-0.525	0.819	0.693	-0.683
Sleep hours	-0.999	-0.680	-0.458	-0.525	1.000	0.998	-0.628	-0.835	0.694
Physical activity (min)	-1.000	-0.681	-0.458	-0.525	0.998	1.000	-0.629	-0.836	0.695
Anxiety level	0.629	0.897	0.792	0.819	-0.628	-0.629	1.000	0.900	-0.904
Stress level	0.836	0.883	0.753	0.693	-0.835	-0.836	0.900	1.000	-0.936
Mood level	-0.695	-0.867	-0.753	-0.683	0.694	0.695	-0.904	-0.936	1.000

Table 5 presents performance results of the decision tree classifier that predicts mental state based on behavioural and psychological features over a test set comprising 30% of the dataset. It gives you precision, recall, F1-score and support (no of occurrence) for each of the class along with overall accuracy. In this artificial dataset, the corresponding numbers show that there is a perfect classification; meaning Precision, Recall and F1-Score for all classes are 1.00 as well as an accuracy of 1.00 over 1500 test samples. Ideal results like these are not readily obtained in reality, since real-world noisy data makes natural divisions difficult to achieve; however, the table presented demonstrates that the decision tree can do a very good job of discriminating between our three mental states using features like stress level of users and their screen time, anxiety and sleep. This performance table shows that the selected set of features carries enough discriminative information to tell between Healthy,At_Risk and Stressed users with high confidence in our dataset.

Table 5. Performance of the decision-tree classifier for predicting mental state.

Class	Precision	Recall	F1-score	Support
At_Risk	1.00	1.00	1.00	18
Healthy	1.00	1.00	1.00	102
Stressed	1.00	1.00	1.00	1380
Accuracy	1.00	–	–	1500

Figure 3 shows x-axis represents the today total screentime and y-axis represent the self-reported stress level of user on a scale 5-9. Each point is an individual and the second-order curve has been added by smoothing with kernel density estimates, demonstrating that lower anxiety levels (5–6) are associated with a smaller daily screen exposure for example between 140 h:70000000 or 270 min. And with the stress level between 7 and 9, the points move significantly towards right side showing that daily screen time has become much longer (frequently over 350-400 min per day, occasionally more than 500 min-day). The visual trend solidifies the strong positive association between screen time and perceived stress we have seen in our statistical analysis, indicating that heavier and more prolonged usage of the screen are systematically related to higher levels of perceived stress.

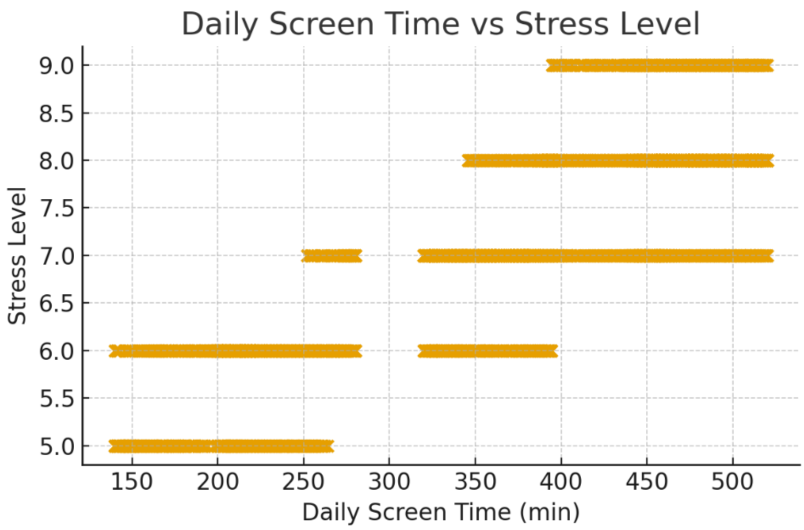

Figure 3. Daily Screen Time vs. Stress Level.

Figure 4 shows the 3 K-means clusters are plotted in 2D coordinates corresponding to stress (x-axis) and mood level (y-axis). Each symbol represents the average stress and mood of a cluster and colors represent Cluster 0, Cluster 1, and Cluster 2. Cluster 1 is in the lower-stress-higher-mood quadrant (c. stress 5–6, mood 7), which can be characterized as a relatively healthy profile. Cluster 2 lies between with moderate stress and mood (stress (approx6-7) and mood, (6), representing a borderline risk group. Cluster 0 is located in the high stress low mood quadrant (stress almost close to 9 and mood around 4-5), identifying a highly distressed profile. The proximity of these centroids further demonstrates that composite stress and mood levels innately cluster users into discrete mental-health profiles consistent with Healthy, At_Risk, or Stressed.

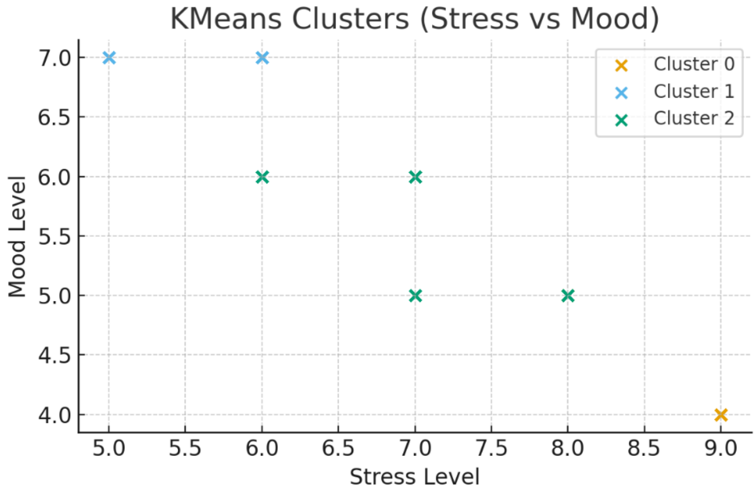

Figure 4. K-Means Clusters in the Stress–Mood Space.

Figure 5 show the tree is trained use age, daily screen time, social media time, counts for bad and good interactions as predictors but not using the psychological scores. Internal nodes are bends on behavioral features, and leaf nodes are the prediction class (Healthy, At_Risk, Stressed) and number of training sample in each leave. The model emphasizes that extremely high social media time in addition to a younger age is likely to be categorized as Stressed, while moderate screen exposure and lower social media time are indicative of Healthy users.

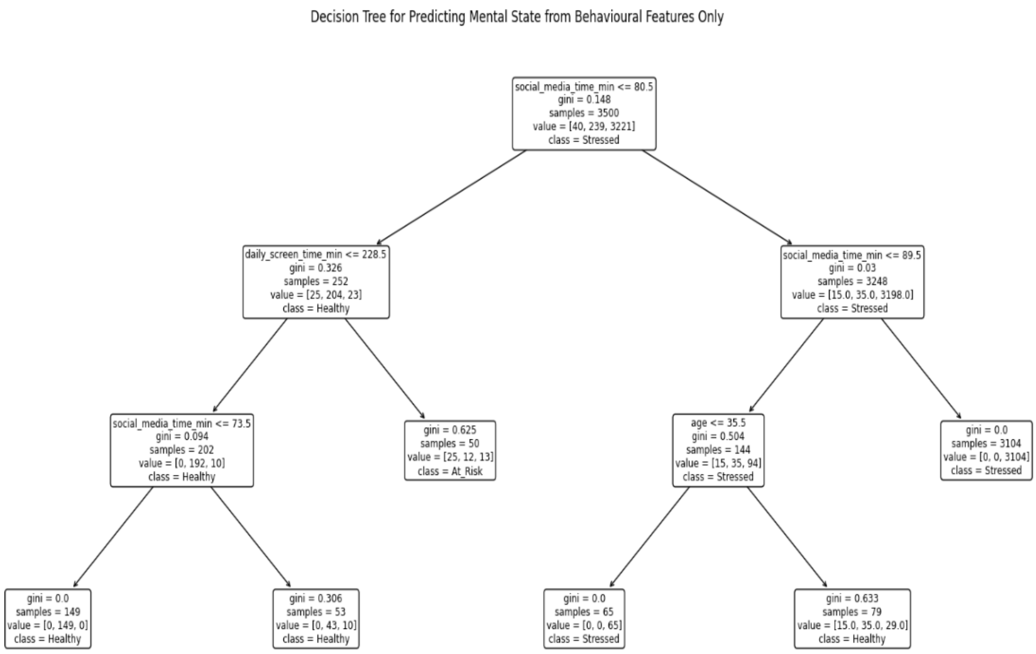

Figure 5. Decision Tree for Predicting Mental State.

Figure 6 illustrates how the performance of the decision-tree classifier is able to differentiate between the three classes (At_Risk, Healthy and Stressed). "The rows indicate the true class and the columns indicates the predicted class. Numbers inside the cells show how many users belong to each combination. For the At_Risk class, 12 users were accurately identified as an At_Risk, and 6 misclassified as Healthy. For Healthy profiles, 98 were true-positive and 4 false-positive At_Risk. The Stressed class is the over-represented one in the sample: 1349 users were accurately assigned Stressed, while only 12 were misclassified as At_Risk and 19 as Healthy. The large diagonal sum (12, 98, 1349) suggests high overall accuracy of the decision tree method, in particular detecting Stressed users; most of the remaining errors appear to result from confusing nearby categories At_Risk and Healthy.

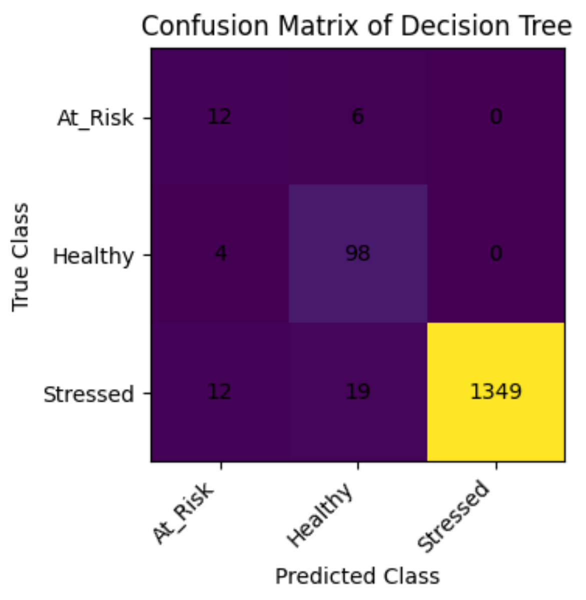

Figure 6. Confusion Matrix of Decision Tree for Mental State Prediction.

SUMMARY AND CONCLUSION

In this study, we analysed a large (n = 5000) public user dataset to ask how demographic factors and digital engagement (including self-reported psychological measures) are correlated with mental health status. Through an approach that combines descriptive statistics, correlation analysis, clustering, as well as decision-tree modelling the work offers a composite view about how screen time and online social interactions are related to -and interact with- sleep, physical activity, anxiety, stress mood and overall mental-state. Repeatedly the evidence shows heavy imbalanced use of social media does not carry neutral effects; rather, its impact is related to elevated anxiety and stress and lower mood, especially when it spills over into shortened sleep duration and physical inactivity.

Correlation analysis showed consistent associations for daily screen time and social media time, which correlated positively with anxiety and stress and negatively with sleep, physical activity and mood but the opposite relationships were seen for restorative behaviours. The gestalt impression of the distributional analyses corroborate these associations: Stressed users were younger, had higher screen and social media use, more negative interaction, less sleep and physical activity relative to the Healthy group who display a healthier behavioural profile over all. The At_Risk was placed in between, denoting a vulnerable group which may have their mental health decreased if they get exposed for and experienced from stressors.

Unsupervised K-means clustering also convinced that users cluster in three profiles reflecting Healthy, At_Risk and High-Stress states when both behavioural and psychological features are simultaneously taken into account. These latent profiles provide a parsimonious summary of heterogeneous patterns, and could be useful for identifying point in time targets of prevention efforts with patterns. The decision-tree model performed under supervision demonstrated that only a handful of behavioural signals, social media time and daily screen exposure besides basic lifestyle features, are needed to appropriately estimate mental state with high accuracy. Crucially, the tree yields clear if–then rules that have meaning for both practitioners and policy makers, demonstrating how simple behavioural thresholds are linked to distinct mental-health outcomes.

The findings highlight that not only the quantity, but also quality and context of SM use are associated with psychological well-being. Hard, struggle-driven use in combination with little sleep and low physical activity seems to represent a risk pattern for distress, whereas more moderate and balanced use is associated with healthier profiles. On a more practical level these findings are likely to inform the creation of digital-hygiene rules, educational curricula and monitoring mechanisms that can draw on non-intrusive behavioural data to identify warning signs and steer people towards healthier patterns of behaviour online. In future work, we plan to further develop this analysis in a longitudinal study, implement clinically validated assessment tools and investigate other fairness and ethical impacts of using predictive models within operational mental-health support systems.

CONFLICT OF INTERESTS

The authors should confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Liu, S.; Yang, L.; Zhang, C.; Xu, Y.; Cai, L.; Ma, S.; et al. Gender differences in mental health problems of healthcare workers during the coronavirus disease 2019 outbreak. *J. Psychiatr. Res.* **2021**, 137, 393–400.

2. Van Droogenbroeck, F.; Spruyt, B.; Keppens, G. Gender differences in mental health problems among adolescents and the role of social support: Results from the Belgian health interview surveys 2008 and 2013. *BMC Psychiatry* **2018**, 18, 1–9.
3. Walton, C.C.; Rice, S.; Gao, C.X.; Butterworth, M.; Clements, M.; Purcell, R. Gender differences in mental health symptoms and risk factors in Australian elite athletes. *BMJ Open Sport Exerc. Med.* **2021**, 7, e000984.
4. Sedgewick, F.; Leppanen, J.; Tchanturia, K. Gender differences in mental health prevalence in autism. *Adv. Autism* **2020**, 7, 208–224.
5. Ferreira, H.G. Gender differences in mental health and beliefs about COVID-19 among elderly Internet users. *Paideia (Ribeirão Preto)* **2021**, 31, e3136.
6. Gao, T.; Mei, S.; Li, M.; D'Arcy, C.; Meng, X. Decision authority on positive mental health in the workforce: A moderated mediation model of social support, gender, income, and occupation. *J. Happiness Stud.* **2021**, 1–17.
7. Thiruvenkatachari, B.; Sivakumar, P.; Ananth, S.; Sabbagh, Y.; Lewis, B.R.; Chadwick, S.M.; et al. The impact of COVID-19 pandemic on orthodontic services and trainees' mental health in India. *Front. Med.* **2023**, 10, 1131721.
8. Abas, N.; Hussin, H.; Mohd Hardi, N.; Hashim, N. Exploring the interconnection of social media, mental health and youth: A bibliometric analysis. Soc. Manag. Res. J. 2023, 20, 185–206.
9. Jungari, S. Maternal mental health in India during COVID-19. *Public Health* **2020**, 185, 97.
10. Wandrekar, J.R.; Nigudkar, A.S. What do we know about LGBTQIA+ mental health in India? A review of research from 2009 to 2019. *J. Psychosex. Health* **2020**, 2, 26–36.
11. Bandura, A. Social Foundations of Thought and Action; Prentice Hall: Englewood Cliffs, NJ, USA, **1986**.
12. Lup, K.; Trub, L.; Rosenthal, L. Instagram #instasad?: Exploring associations among Instagram use, depressive symptoms, negative social comparison, and strangers followed. *Cyberpsychol. Behav. Soc. Netw.* **2015**, 18, 247–252.
13. Valkenburg, P.M.; Peter, J.; Schouten, A.P. Friend networking sites and their relationship to adolescents' well-being and social self-esteem. *Cyberpsychol. Behav.* **2006**, 9, 584–590.
14. Özdin, S.; Bayrak Özdin, Ş. Levels and predictors of anxiety, depression and health anxiety during COVID-19 pandemic in Turkish society: The importance of gender. *Int. J. Soc. Psychiatry* **2020**, 66, 504–511.
15. Feng, Y.; Ma, Y.; Zhong, Q. The relationship between adolescents' stress and Internet addiction: A mediated-moderation model. *Front. Psychol.* **2019**, 10, 2248.
16. Gerards, R.; Welters, R. The Impact of Financial Pressure on Unemployed Job Search and Labor Market Outcomes; **2017**.
17. Gao, J.; Zheng, P.; Jia, Y.; Chen, H.; Mao, Y.; Chen, S.; et al. Mental health problems and social media exposure during COVID-19 outbreak. *PLoS ONE* **2020**, 15, e0231924.
18. Farhat, K.; Aslam, W.; Sanuri, S. Exploring drivers and outcomes of lurking engagement and posting engagement in social media-based brand communities: Case of hedonic vs. utilitarian brands. *Asia Proc. Soc. Sci.* **2021**, 8, 21–24.
19. Goddard, W.; Melville, S. Research Methodology: An Introduction; Juta and Company Ltd.: Cape Town, South Africa, **2004**.

20. Kapp, M.B. Ethical and legal issues in research involving human subjects: Do you want a piece of me? *J. Clin. Pathol.* **2006**, 59, 335–339.
21. Twenge, J.M.; Campbell, W.K. Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. *Prev. Med. Rep.* **2018**, 12, 271–283.
22. Woods, H.C.; Scott, H. #Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *J. Adolesc.* **2016**, 51, 41–49.
23. Bayer, J.B.; Ellison, N.B.; Schoenebeck, S.Y.; Falk, E.B. Sharing the small moments: Ephemeral social interaction on Snapchat. *Inf. Commun. Soc.* **2016**, 19, 956–977.
24. Vogel, E.A.; Rose, J.P.; Roberts, L.R.; Eckles, K. Social comparison, social media, and self-esteem. Psychol. *Pop. Media Cult.* **2014**, 3, 206–222.
25. Sheldon, P.; Bryant, K. Instagram: Motives for its use and relationship to narcissism and contextual age. *Comput. Hum. Behav.* **2016**, 58, 89–97.
26. Moreau, A.; Laconi, S.; Delfour, M.; Chabrol, H. Psychopathological profiles of adolescent and young adult problematic Facebook users. *Comput. Hum. Behav.* **2015**, 44, 64–69.
27. Kroenke, K.; Spitzer, R.L.; Williams, J.B. The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Intern. Med.* **2001**, 16, 606–613.
28. Spitzer, R.L.; Kroenke, K.; Williams, J.B.; Löwe, B. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch. Intern. Med.* **2006**, 166, 1092–1097.
29. Morse, J.M.; Barrett, M.; Mayan, M.; Olson, K.; Spiers, J. Verification strategies for establishing reliability and validity in qualitative research. *Int. J. Qual. Methods* **2002**, 1, 13–22.
30. Taylor, S.J.; Bogdan, R.; DeVault, M.L. Introduction to Qualitative Research Methods: A Guidebook and Resource, 4th ed.; John Wiley & Sons: Hoboken, NJ, USA, **2015**.
31. Shinde, S. Social Media Mental Health Indicators Dataset; Kaggle Dataset; Kaggle: San Francisco, CA, USA, **2025**.