1. Introduction

Depression is a common mental health condition with meaningful effects on daily functioning and quality of life. As digital health tools become more widely used, it is increasingly important to identify behavioural factors that may be linked to changes in depressive symptoms. Wearable devices and smartphones provide continuous data on sleep, physical activity, and lifestyle patterns, creating new opportunities to examine how these behaviours relate to mental health. This project combines descriptive analyses, clustering, and supervised modelling to examine how behavioural and lifestyle features relate to PHQ-9 changes across three-month intervals. The central research question is: *To what extent do behavioural, lifestyle, demographic, and clinical features help differentiate participants whose PHQ-9 scores improve from those whose symptoms remain stable or worsen over a three-month interval?* Clustering is used to identify natural groupings in sleep, activity, and lifestyle variables, while Logistic Regression and Random Forest models try to identify the extent to which these features can distinguish individuals whose symptoms improve from those who remain stable or worsen. Although behavioural clusters reflect meaningful variation in daily routines, they do not correspond to distinct symptom-change trajectories. Moreover, predictive models achieve only modest accuracy, indicating that short-term PHQ-9 changes are very difficult to infer from the person-generated health data. These findings highlight both the potential and the limitations of using passive behavioural data to understand short-term mental-health dynamics.

1.1 Depression background

Depression is a very widespread and disabling mental health condition. According to the World Health Organization, about 5.7% of adults worldwide were affected in 2025 (World Health Organization, 2025). WHO defines depression as a common mental disorder involving a persistently low mood or a loss of interest and pleasure in activities for long periods. It goes beyond ordinary mood changes and can significantly disrupt daily functioning, relationships, academic performance, and work. While depression can develop in anyone, people exposed to trauma, prolonged stress, or major life events are at increased risk, and women tend to be affected more often than men.

1.1.1 Depression among students

Mental health issues are also highly prevalent among university students. Data from the Healthy Minds Network shows some improvement from 2022 to 2025, with moderate to severe depressive symptoms decreasing from 44% to 37% and severe depression dropping from 23% to 18% (University of Michigan School of Public Health, 2025). Anxiety, loneliness, and suicidal ideation also declined. However, even with these improvements, the percentages remain very high; more than one-third of students still struggle with depression, and large numbers continue to report significant anxiety and loneliness. At the same time, levels of “flourishing,” which measure positive psychological well-being, have slightly decreased. This suggests that reductions in distress do not necessarily mean students are feeling better overall.

1. 2 Literature Review

Research has explored how person-generated health data (PGHD) from wearables and smartphones can help track and understand changes in depressive symptoms. A key study, using the same dataset, is Makhmutova et al. (2022), who developed the PSYCHE-D model to predict three-month changes in depression severity using behavioural, lifestyle, medication, and demographic features derived from PGHD. Their findings showed that signals from sleep, activity patterns, and lifestyle adjustments contain meaningful information about whether symptoms are likely to improve or worsen, although predictive performance remained moderate due to the complexity of mental-health trajectories. The study, moreover, highlights the value of multi-domain behavioural data, providing an analytical framework closely aligned with the goals of this project.

Price et al. (2023) expanded this line of work by focusing on how depression symptoms fluctuate across a full year rather than simply improving or worsening. Using stacked ensemble models, they demonstrated that both demographic variables and wearable-derived features, specifically sleep regularity, sleep duration, and physical activity, independently contribute to understanding long-term symptom variability. They emphasized on fluctuation, rather than direction: depression is defined not only by severity but also by the stability of symptoms over time, a perspective that informs how behavioural patterns may reflect mental-health dynamics.

Other studies support the role of digital phenotyping in capturing meaningful signals related to depression. Pedrelli et al. (2020) showed that passive sensor data such as phone usage, sleep, activity, and physiological measures can estimate clinician-rated depression severity (HDRS-17) with moderate accuracy, providing early evidence that everyday behaviour reflects clinical symptom changes. Sun et al. (2023) similarly found that variations in mobility, sleep, and home-stay behaviour correlate with shifts in PHQ-8 scores, although the strength and direction of these associations differ substantially across individuals. They also highlighted common challenges in PGHD research, including missing data, irregular sensor measurements, and temporal misalignment.

These studies show how behavioural data can capture both short-term changes and longer-term variability in depressive symptoms and that machine-learning models can identify informative patterns, even if accuracy remains pretty low. Recurring behavioural markers such as sleep consistency, sleep duration, mobility, and physical activity appear especially relevant. At the same time, many practical challenges, i.e. data quality, individual differences, and complex longitudinal structure, highlight the importance of careful preprocessing and modelling choices in this project.

2. Problem formulation

Understanding which behavioural factors are associated with changes in depressive symptoms is an important question in mental-health research. Daily habits such as sleep duration and regularity, physical activity, general lifestyle routines are known to influence well-being, yet their month-to-month relationship with validated clinical measures is not well characterised. With the increasing availability of wearable devices that continuously capture these behaviours, it is now possible to examine these patterns with greater precision than traditional survey-based methods allow.

This question is also relevant within the context of student life. Academic schedules, irregular sleep routines, and fluctuating workloads create conditions where well-being can change rapidly. Observing these dynamics firsthand motivates an interest in whether objectively measured behaviours correspond to meaningful shifts in depressive symptoms.

This project analyzes the DiSCover dataset to investigate how behavioural signals collected through wearables relate to changes in PHQ-9 depression scores across three-month intervals. Rather than aiming to build a highly accurate predictive model, the focus is on identifying patterns in the data and examining whether certain behaviours are linked to improvement, stability, or a worsening in symptoms. By exploring these relationships, the project seeks to provide a clearer understanding of how everyday behavioural patterns may reflect or influence mental well-being, particularly in student populations.

3.Dataset Description

This project uses data from the DiSCover study, a year-long digital health research initiative that tracked the behaviour and mental health of over 10,000 U.S. adults between 2018 and 2020. Participants wore Fitbit devices and completed regular online surveys capturing lifestyle factors, stress, medication changes, and depressive symptoms. The dataset used in this analysis contains 35,694 rows and 156 columns, with each row representing one *participant-month* (e.g., "34_12" corresponds to participant 34 in month 12). All identifiers are anonymised. The dataset is also demographically unbalanced: it includes 26,881 women, 8,735 men, and 78 individuals classified as "other." This strong skew toward female participants is important to note, as it may introduce bias into analyses involving sex or comparisons across demographic groups. The age distribution in the dataset spans a broad range, from 19 to 79 years old, with an average age of about 39. The interquartile range (31–45) shows that most participants fall into early to mid-adulthood, while older adults are present but less common.

3.1 Behavioural Features from Wearables

A substantial portion of the dataset consists of Fitbit-derived daily activity and sleep metrics. The activity data include measures such as steps while awake (mean, sum, and IQR), light and moderate-to-vigorous physical activity (LPA and MVPA), rolling-window activity features, activity streaks, and counts of sedentary days. The sleep data cover weekday and weekend sleep duration, time in bed versus time asleep, sleep efficiency, sleep timing indicators such as the main sleep start hour and associated ranges or IQRs, and counts of hypersomnia or hyposomnia episodes. They also include a large set of engineered sleep features, that model behavioural consistency. Many of these engineered variables appear in triplets (score, intercept, and coefficient), reflecting the regression-based approach used to characterise behavioural patterns.

3.2 Survey, Lifestyle, and Demographic Variables

The dataset also includes monthly survey responses and participant characteristics. Lifestyle factors capture information on stress, meditation, alcohol reduction, and eating or activity behaviours. Medical information covers comorbidities such as diabetes or migraines, medication changes and dosing, and non-medication interventions. Demographic variables include sex, birthyear, height, weight, BMI, insurance status, and household composition.

3.3 Patient Health Questionnaire-9

A central component of the dataset is the PHQ-9 (Patient Health Questionnaire-9), a clinical instrument for assessing depression severity. The PHQ-9 consists of nine questions, each scored from 0 to 3, producing a total score ranging from 0 to 27. The scores correspond to standard clinical categories: 0–4 indicate minimal depression, 5–9 mild depression, 10–14 moderate depression, 15–19 moderately severe depression, and 20–27 severe depression. The PHQ-9 was collected approximately every three months in the DiSCover study, which allows each participant's data to be structured into three-month intervals containing a PHQ-9 score at the start, a PHQ-9 score at the end, and behavioural data from the intervening months. The difference between the start and end scores forms the main outcome used in this project: PHQ-9 change, representing improvement, worsening, or stability in depressive symptoms.

3.4 Missing Data Overview

The dataset exhibits substantial variability in missingness across variable, which is expected and typical of person-generated health data where device adherence fluctuate. The full table of the percentages of missing data can be found in the Appendix. Here we report only the most relevant results.

Highly missing features (>70%)

Some variables have extremely high missingness, including steps_mvpa_iqr at 86% missingness, the MVPA engineered features (scores, intercepts, and coefficients) at approximately 76%, and the PHQ-9 category labels (phq9_cat_start and phq9_cat_end) at roughly 70%. The PHQ-9 was collected less frequently,every three months, compared to behavioural data, so this level of missingness is expected.

Moderately Missing Sleep Features (40–60%)

These include several sleep-related metrics that show large gaps, such as sleep_main_start_hour_adj_iqr and sleep_ratio_asleep_in_bed_iqr at approximately 56% missingness, as well as numerous engineered sleep score, intercept, and coefficient features at around 43%. These gaps likely reflect inconsistent nighttime wear of the Fitbit device.

Partially Missing Features (20–40%)

These cover variables such as sleep_asleep_iqr and sleep_in_bed_iqr at roughly 38% missingness, step variability and rolling-window metrics at about 20–30%, and light activity modelling features at approximately 25%.

Low Missingness (<5%)

These consist largely of core behavioural and demographic features, including step means or sums, sleep weekday and weekend means, and BMI, height, and education, all of which are more than 95% complete.

Complete Data (0%)

Several important variables have no missing data, including demographic factors such as sex, birthyear, and race categories; comorbidities; lifestyle variables such as stress, meditation, and alcohol reduction; medication and non-medication event variables; and recent activity features such as steps_mvpa_sum_recent.

3.4 Implications for Analysis

The missingness patterns have several consequences for the analysis: all of the engineered variables have very high percentages of missing data and are not relevant for the analysis. Moreover, many sleep variability and engineered behavioral features are too incomplete to use without major imputation. PHQ-9 category fields cannot be used directly due to sparse availability. Clustering and modelling require filtering rows with sufficient coverage across key behavioural features. Selection of variables must balance data completeness with behavioural relevance. These considerations guide the preprocessing steps described in the Methods section.

3.6 Data Exploration

3.6.1 Categorisation of variables

As an initial step in exploring the dataset, the variables were organised into conceptually meaningful groups. To construct these groups, variable names were scanned for descriptive patterns that consistently appear in the dataset's naming conventions. The grouping procedure was as follows:

Sleep variables: All variables whose names contain the substring "sleep" were assigned to the sleep group. These include nightly duration metrics, timing measures, weekday and weekend averages, and sleep-related behavioural indicators.

Activity variables: Variables containing "step" in their names were grouped as activity features. This includes measures of total steps, MVPA and LPA summaries, rolling-window step metrics, and activity/sedentary day counts.

PHQ-9 variables**:** All variables containing "phq" were collected into a separate group. Because PHQ-9 assessments occurred only every three months, these features have higher missingness than behavioural predictors, but it makes sense.

Lifestyle variables: Variables containing the substring "life" were classified as lifestyle features. These include self-reported stress levels, meditation practice, eating/activity habits, and alcohol-related behaviour changes.

Medication and treatment variables**:** Variables containing the substring "med" were grouped under medication,. iy medication start/stop indicators, dosing, migraine treatments, and non-medication interventions.

Demographic variables: Demographic features do not follow a common naming pattern, so they were defined manually. This group includes: sex, birthyear, height, weight, BMI, pregnancy status, and insurance status.

3.6.2 Missingness analysis after preprocessing

This grouping strategy enabled the calculation of the average percentage of missing values, providing a high-level overview of data completeness across behavioural, clinical, and demographic categories.

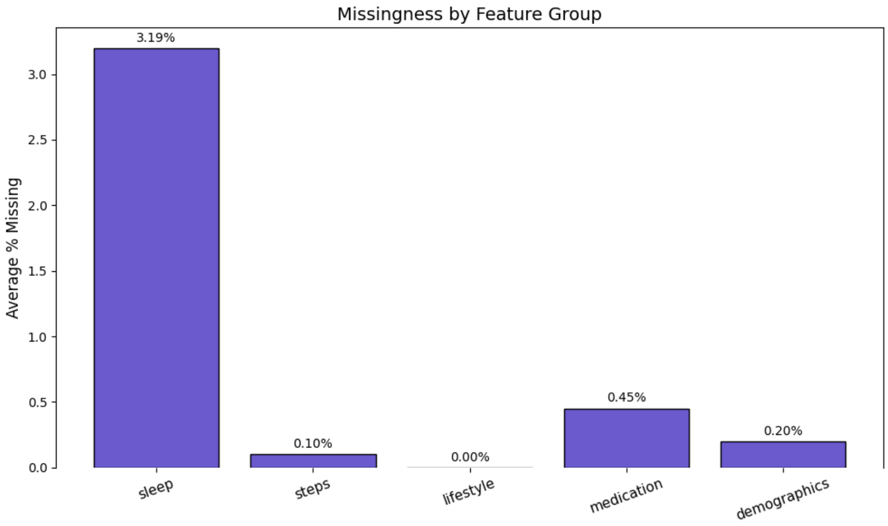


Figure 1: Missingness Comparison - Categories of data

The resulting plot shows that:

Sleep features exhibit low missingness (~3%); some imputation method needs to be analyzed. Step/activity features are nearly complete (<0.1%).
Lifestyle and demographic features are also nearly complete.
Medication variables have higher missingness (~0.45%). This is problematic, considering confounding factors.

3.6.3 Distribution of PHQ-9 Scores

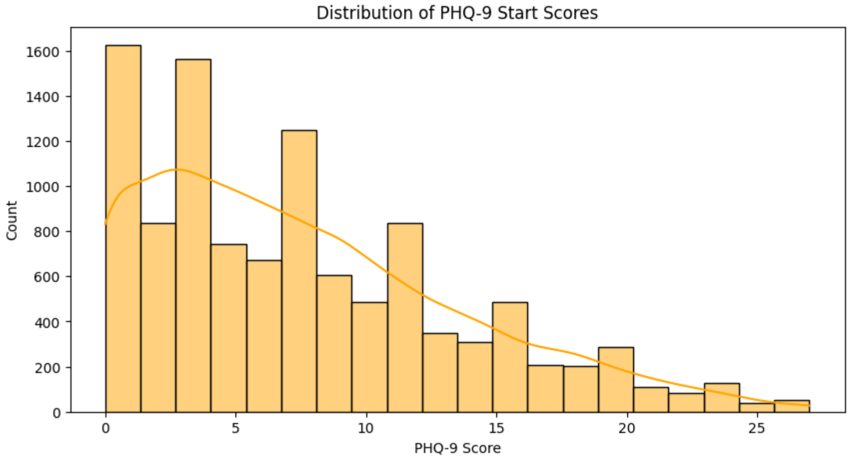


Figure 2: PHQ-9 Start Scores Distribution

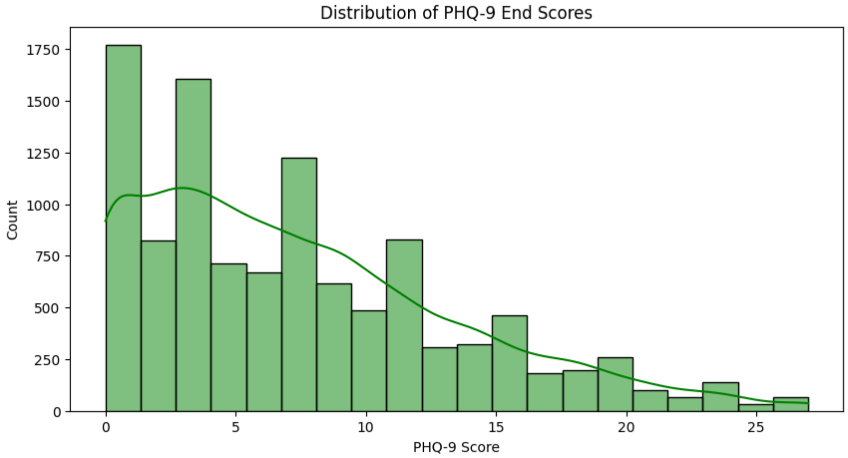


Figure 3: PHQ-9 End Scores Distribution

The distributions of PHQ-9 start scores and PHQ-9 end scores were examined, to understand baseline depressive symptom severity and how it evolves over the three-month intervals used in this study.

Both distributions show a clear right-skewed pattern, with the majority of participants reporting relatively low to moderate depressive symptoms. For the PHQ-9 start scores, the largest concentrations appear between 0 and 5, indicating that many participants began the interval with mild symptoms. The density curve shows a gradual decline as scores increase, with few participants exhibiting moderate to severe depression at baseline.

The distribution of PHQ-9 end scores exhibits a similar overall shape but with slightly higher counts in the lowest score range. This shift suggests that, across the sample, depressive symptoms tended to decrease modestly over the three-month intervals.

Nonetheless, the long right tail remains, indicating that a subset of participants continued to experience moderate, moderately severe, or severe symptoms at the end of the period.

These distributions provide important context for interpreting PHQ-9 change values. While many participants show improvement, the wide range of both starting and ending values highlights considerable heterogeneity in depression trajectories. This variability motivates the subsequent analysis of behavioural factors to understand which patterns may be associated with improvement, stability, or deterioration in symptoms.

3.6.4 Change in PHQ-9 Over Time

A change score was computed as the difference between the PHQ-9 end and start values, to investigate how depressive symptoms evolved over each three-month interval.

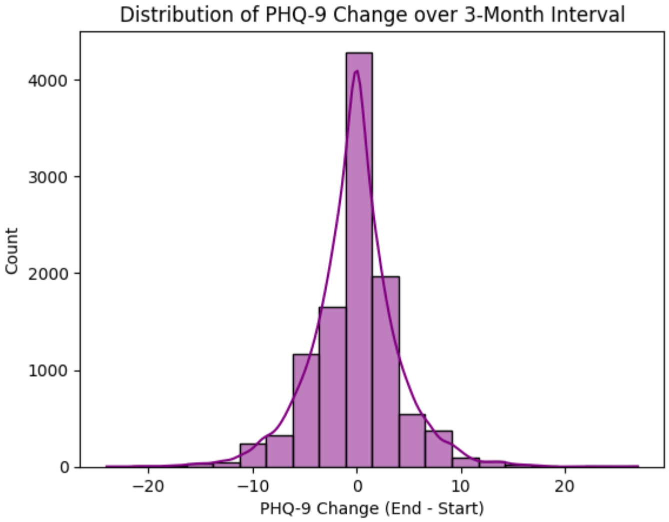


Figure 4: Distribution of PHQ-9 Change

The resulting distribution is centered around zero and displays an approximately symmetric shape with heavier tails, indicating substantial variability across participants. Most individuals experienced relatively small changes, either slight improvement (negative values) or slight worsening (positive values). The peak around zero suggests that for a large portion of the sample depressive symptoms have not changed over the three months. At the same time, the presence of both large negative and large positive values indicates that some participants experienced some notable improvements or deteriorations. This heterogeneity reinforces the value of examining behavioral correlates, as individual trajectories differ widely despite similar baseline scores.

3.6.7 Correlation analysis

A correlation analysis was conducted using PHQ-9 change as the outcome and a subset of behavioural, lifestyle, and demographic features as predictors as to explore potential associations between behavioural variables and changes in depressive symptoms. The heatmap visualises the strongest correlations among these variables.

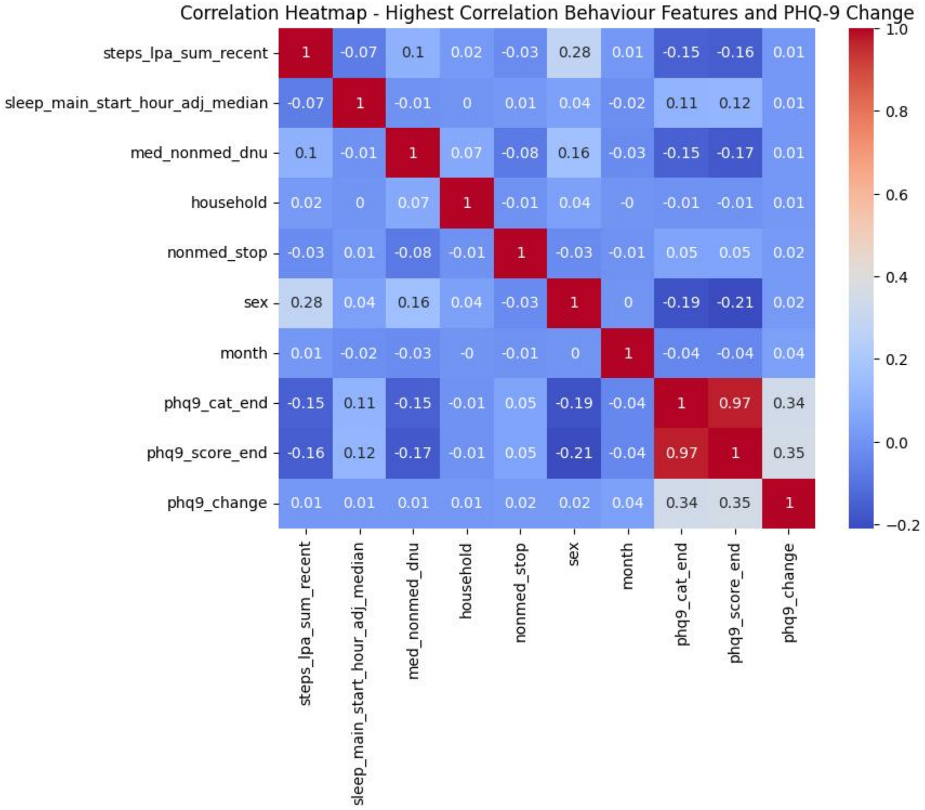


Figure 5: Correlations with PHQ-9 Change

The correlations between behavioural features and PHQ-9 change are very small, generally close to zero. This indicates that no single behavioural metric shows a strong linear association with short-term symptom change. This finding could be indicative that behavioural signals tend to be weakly predictive at the individual-feature level due to noise and the complexity of mental-health trajectories.

A few non-behavioural variables show slightly stronger associations. Sex has a modest positive correlation with activity measures, reflecting known population differences in our dataset, though its correlation with PHQ-9 change itself remains very small. Medication- and lifestyle-related variables (e.g., med_nonmed_dnu, nonmed_stop) also exhibit minimal relationships with symptom change.

The strongest correlations in the matrix appear between PHQ-9 start, end, and categorical depression severity labels. It is interesting to note that the correlation between PHQ-9 start and change scores is moderately negative: individuals with higher starting scores tend to show larger decreases (improvement). This will be studied further in the following sections.

Taken together, the correlation analysis indicates that short-term changes in depression are not strongly explained by any single behavioural metric in the dataset. This is expected: it highlights the need for multivariate or non-linear modelling approaches, reflecting the complexity of predicting mental-health trajectories.

3.6.8 Behavioural trends across PHQ-9 severity categories

In this section, an analysis of how different behavioural trends affect PHQ-9 severity categories, to understand what the major factors could be influencing the improvement or worsening of depression severity.

Sleep

To explore whether sleep behaviour differed across individuals who improved, remained stable, or worsened in depressive symptoms, some sleep-related metrics were compared across PHQ-9 change groups.

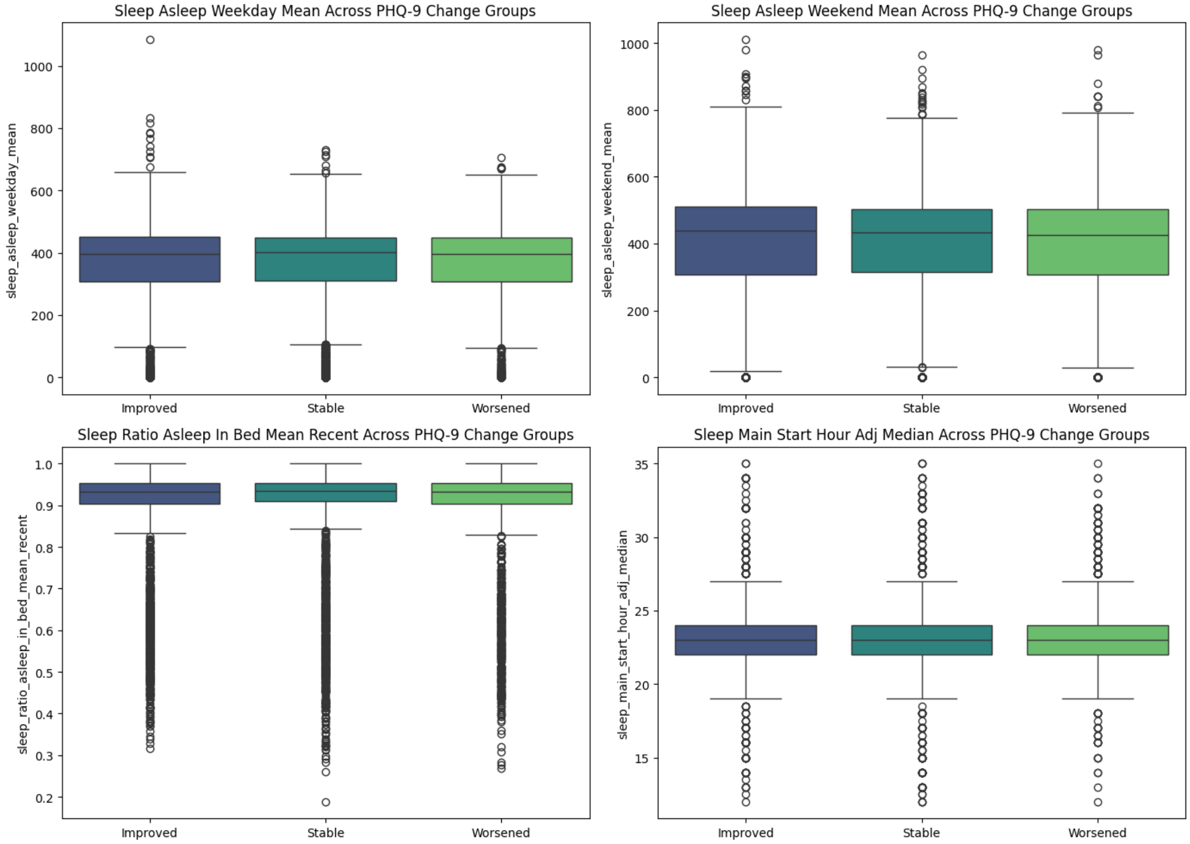


Figure 6: Sleep Boxplots across PHQ-9 Change Groups

The sleep patterns look very similar across the three PHQ-9 change groups. Weekday and weekend sleep duration differ only slightly, and sleep efficiency remains consistently high for all groups. Typical sleep start times cluster around the same hour, with no noticeable shift for those who improved or worsened. These small differences match the boxplots, which show substantial overlap. The sleep summaries in this dataset do not appear to differentiate individuals based on whether their depressive symptoms improved, were stable or worsened over the three-month interval.

Stress

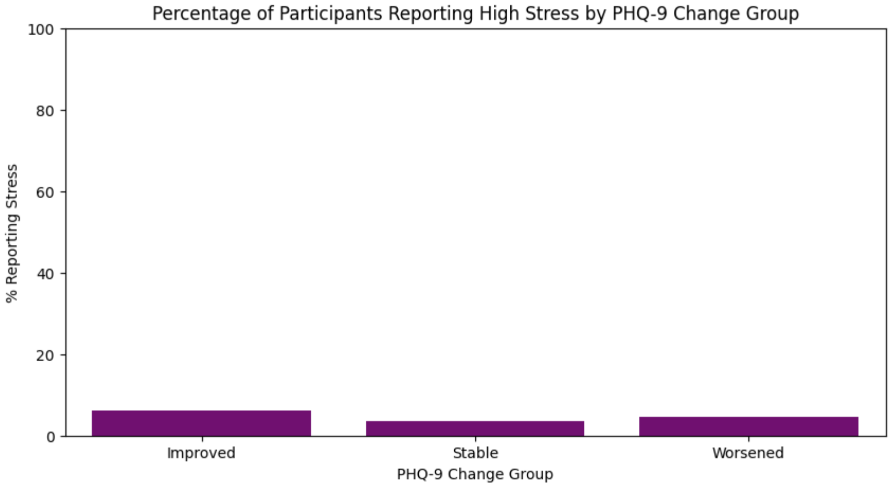


Figure 7: Stress differences across PHQ-9 change groups

The proportion of participants reporting high stress is low across all PHQ-9 change groups, and the differences between them are minimal. Individuals who improved, remained stable, or worsened show very similar stress-reporting rates, suggesting that this single stress indicator does not strongly distinguish between short-term changes in depressive symptoms in this dataset.

Sex

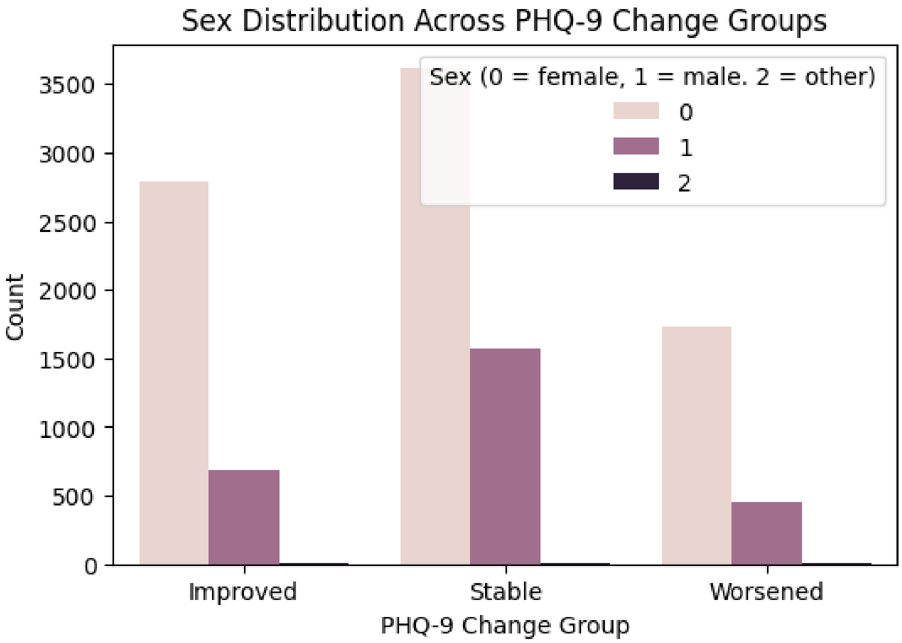
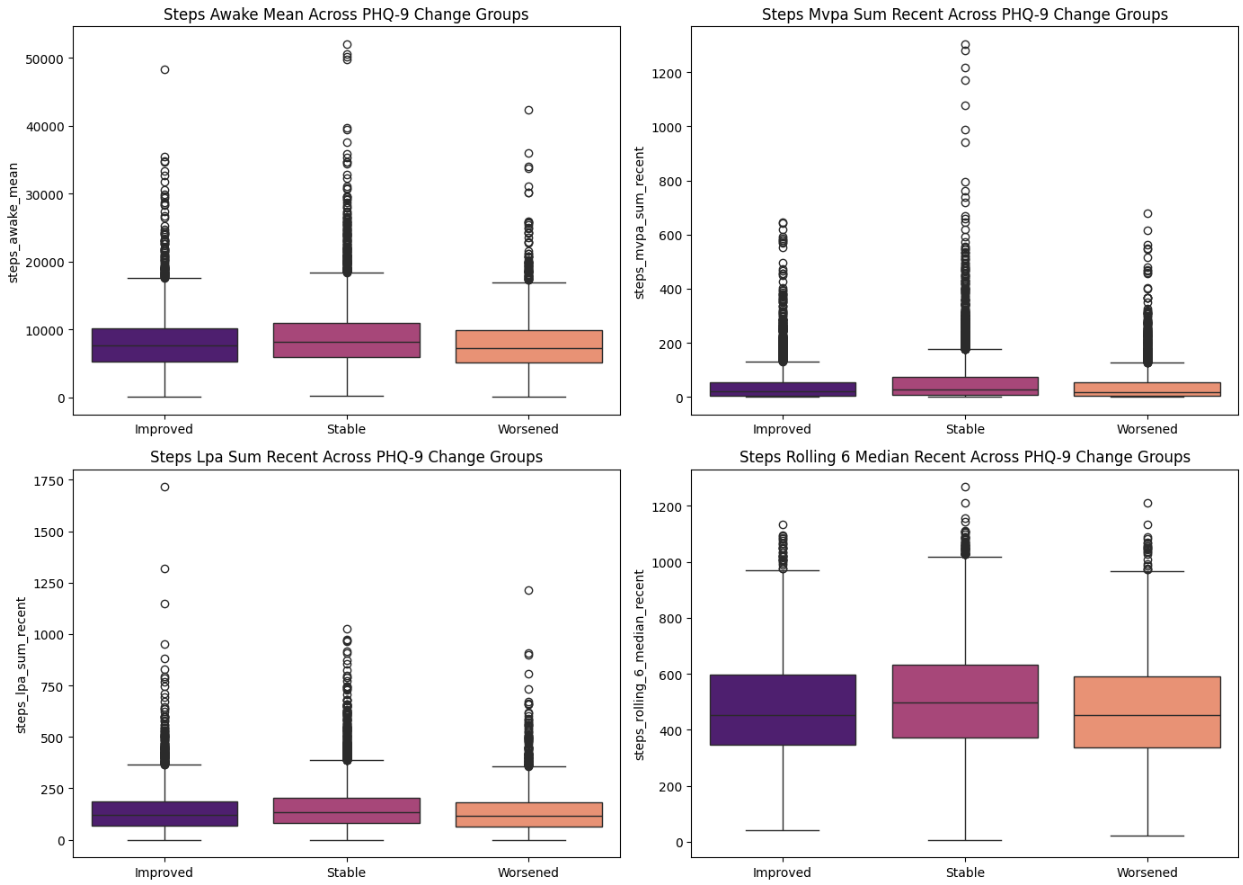


Figure 8: Sex distribution across PHQ-9 change groups

The plot shows that females (coded as 0) represent the clear majority across all PHQ-9 change groups). This reflects the composition of the dataset: three-quarters of participants identify as female. Male participants (coded as 1) make up a smaller proportion in every group, while individuals identifying as “other” (coded as 2) appear very infrequently. The relative proportions of each sex category remain similar across change groups, suggesting that the distribution of improvement, stability, or worsening in PHQ-9 scores does not differ substantially by sex. It is important to highlight the strong imbalance in sample size when interpreting group comparisons, as patterns may be more reflective of female participants.

Physical Activity – Steps Metrics

An analysis was performed also on four variables reflectings different dimensions of participants’ physical activity. steps_awake_mean captures overall daily movement, while steps_mvpa_sum_recent and steps_lpa_sum_recent distinguish between higher-intensity and light activity. steps_rolling_6_median_recent summarizes activity consistency over time, providing a more stable reflection of habitual behaviour than single-day measures.


Steps Awake Mean Across PHQ-9 Change Groups

Steps Mvpa Sum Recent Across PHQ-9 Change Groups

Steps Lpa Sum Recent Across PHQ-9 Change Groups

Steps Rolling 6 Median Recent Across PHQ-9 Change Groups

Figure 9: Boxplot of activity variables across PHQ-9 change

Across all four step-related features the three PHQ-9 change groups show approximately similar activity levels, with only small differences between them. Participants in the Stable group tend to record slightly higher average daily steps and more light and moderate-to-vigorous activity, while the Improved and Worsened groups have nearly overlapping distributions. Rolling 6-day median steps follow the same pattern, with the Stable group showing a modestly higher median. Overall, these results suggest that step-based activity measures do not clearly distinguish between improvement, stability, or worsening in depressive symptoms.

Trauma History

The trauma variable reflects whether participants reported a history of traumatic events.

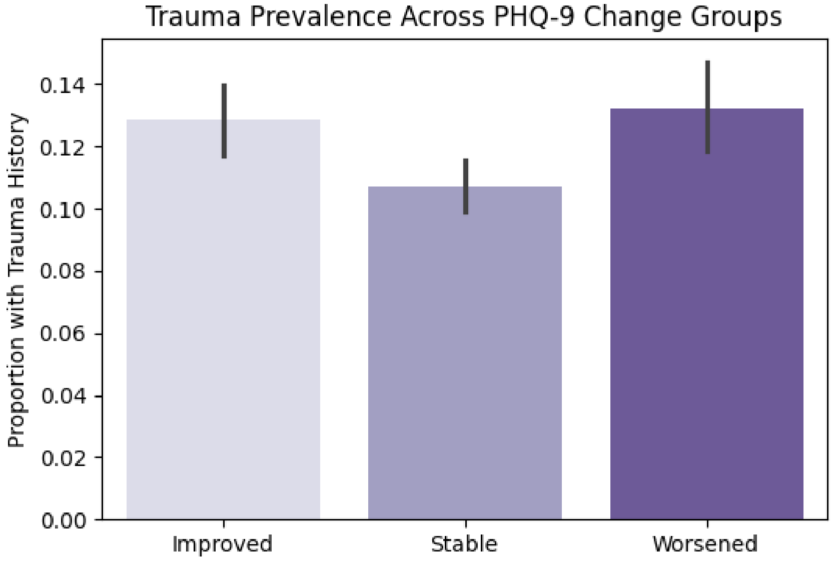


Figure 10: Trauma history prevalence across PHQ-9 change groups

Across groups trauma prevalence is relatively similar, but those whose symptoms worsened show a slightly higher rate compared with the stable group, while the improved group falls in between. These differences are small, however the pattern is consistent with the hypothesis that trauma history may make individuals more vulnerable to increases in depressive symptoms over time.

Age

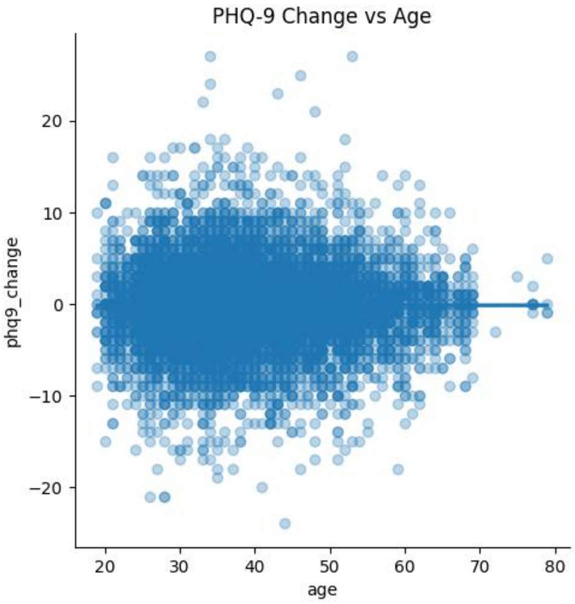


Figure 11: Scatterplot of age and PHQ-9 change

The scatterplot shows no clear relationship between age and PHQ-9 change. Participants across all age groups appear to experience similar ranges of improvement and worsening, with most values clustered around zero. Within this dataset short-term changes in depressive symptoms are not strongly associated with age.

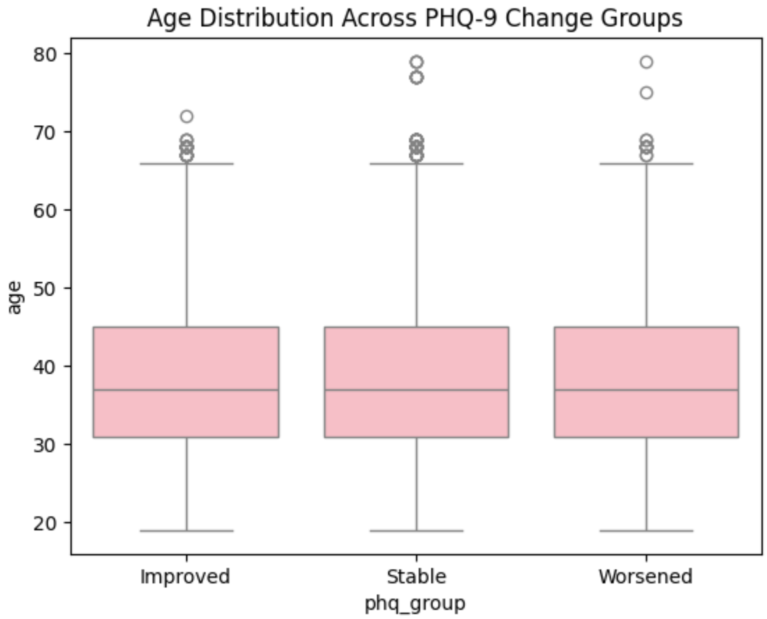


Figure 12: Age boxplot across PHQ-9 change groups

The age distributions across the three PHQ-9 change groups are highly similar. Median age is nearly identical in all groups, and the spread of ages, from early adulthood to older adulthood, does not differ

4.Methods

4.1 Preprocessing

4.1.1 Variables selection

After loading the dataset, engineered Fitbit features were removed. These included automatically generated modelling outputs containing terms such as “score”, “coeff”, “intercept”, “iqr”, and “range”. These variables do not represent direct behavioural measurements and were excluded using a pattern-based filter, with PHQ-9 start and end scores explicitly protected from removal.

The remaining variables were then prepared for analysis by constructing a set of clinically meaningful outcomes and person-level attributes. PHQ-9 change was computed as the difference between the end and start scores for each three-month interval, and rows lacking either value were removed to ensure a complete outcome. A categorical PHQ-9 group (Improved, Stable, Worsened) was generated from this change score.

A set of additional participant-level variables was created directly from the raw data. Age was calculated from the birth year using 2019 as a reference, reflecting the midpoint of the study period. A binary migraine indicator (has_migraine) was defined as 1 if a participant reported a migraine diagnosis or any migraine days; otherwise, it was set to 0.

Feature selection focused on retaining behavioural, demographic, lifestyle, and selected clinical variables that were easily interpretable, sufficiently complete, and relevant for examining behavioural patterns in relation to PHQ-9 change. Several subsets of variables were tested during the modelling process, and the final set was chosen because it offered the best balance between completeness, interpretability, and model performance. Variables that were redundant or carried overlapping information were removed. The full list of variables used in the modelling steps is provided in the Appendix.

The retained feature set included three main categories.
First, behavioural measures from the Fitbit device captured habitual sleep and activity patterns. These consisted of average sleep duration (weekday and weekend), time in bed, sleep efficiency ratios, sleep onset timing, and counts of hypersomnia or hyposomnia events. Activity-related features included mean daily step counts and counts of active versus sedentary days.

Second, demographic attributes such as age, sex, race, height, weight, and BMI were retained because they may reflect underlying differences in behaviour or mental-health trajectories. Socioeconomic indicators including education, insurance type, household size, and financial stress items were also included due to their known associations with mental-health outcomes.

Third, lifestyle and clinical history variables that are linked to psychological well-being were kept. These included stress levels, meditation practices, alcohol-reduction behaviour, eating-related activity, trauma history, and the migraine indicator (has_migraine) derived during preprocessing. Moreover, comorbidity indicators (e.g., diabetes, arthritis) were retained to capture broader health status.

4.1.3 Handling missing data

After limiting the dataset to rows with both a PHQ-9 start and end score, the remaining variables contained only small amounts of missing data. Sleep-related Fitbit variables contained the most missing observations, though still under ten percent of the dataset. Activity variables, demographic information, and lifestyle fields were largely complete.

Given this distribution, a simple imputation strategy was appropriate. Continuous behavioural and demographic variables, i.e. sleep metrics, activity measures, and BMI, were imputed using the median of each variable, which performs reliably for mildly skewed data. Binary and categorical fields, such as sex, pregnancy status, and insurance type, were imputed using the most frequent category, although these variables were rarely missing. This approach retains almost all observations and avoids introducing strong assumptions into the data.

4.2 Modelling techniques

4.2.1 Unsupervised: Clustering

Before moving to supervised prediction, clustering was used to explore whether participants formed distinct behavioural groups based on sleep, activity, lifestyle, demographic, and clinical variables. These variables were standardised using z-scores, and k-means clustering was applied to the scaled data. To select an appropriate number of clusters, silhouette scores were computed for values of *k* between 2 and 10. The silhouette score measures how well-separated the clusters are, with higher values indicating clearer group structure; a four-cluster solution showed the best overall separation.

After fitting the model, cluster labels were assigned to each participant and a principal component analysis (PCA) was performed to visualise the clusters. PCA reduces the many input features into two components that capture the largest share of variability in the data, making it possible to visualize the clusters in a two-dimensional space. Each cluster's behavioural and clinical profile was summarised by computing feature means and corresponding z-scores relative to the full sample.

The goal was to determine whether natural groupings in the behavioural data aligned with patterns of PHQ-9 change. The clusters reflect clear differences in daily behaviour, but they did not correspond to distinct symptom-change trajectories, indicating that unsupervised structure alone was insufficient for identifying improvement. This motivated the subsequent use of supervised models to understand whether any predictive signal could be extracted from the selected variables.

4.2.2 Supervised: Predictive modelling

Following preprocessing, the cleaned dataset was used to model the relationship between behavioural, demographic, lifestyle, and clinical variables and the likelihood of PHQ-9 improvement. Because the PHQ-9 change score is continuous but highly noisy, and because behavioural predictors were only weakly correlated with PHQ-9 outcomes, the modelling task was formulated as a binary classification problem. Participants were labelled as *Improved* if their PHQ-9 score decreased by at least two points over the three-month interval, and *Not Improved* otherwise. This threshold reflects a small but clinically relevant reduction in depressive symptoms. We focus on individuals whose PHQ-9 scores decreased over time. Identifying which participants improved allows us to test whether the selected features carry any signal about symptom change.

The dataset was imbalanced, with 7,374 participants classified as *Not Improved* and 3,492 as *Improved*, which makes correct identification of the minority class more difficult

Two modelling approaches were implemented: a Logistic Regression model representing a basic and interpretable statistical method, and a Random Forest Classifier representing a more flexible, non-linear machine-learning model. Both models were trained on the same set of selected predictor variables outlined in Section 4.1.2.

Before model fitting, the dataset was split into training and test sets using an 80/20 split, meaning 80% of the data was used to train the models and 20% was held out for evaluating how well they generalised to unseen cases. Stratification was applied to maintain the same proportion of *Improved* and *Not Improved* participants in both sets, preventing the minority class from being under-represented during training. Continuous predictors were standardised with *z-score scaling* for the logistic regression model so that variables measured on different numerical scales contributed evenly during optimisation. Tree-based models were trained on the unscaled predictors, as they are not sensitive to feature scaling.

For Logistic Regression, the model was fitted with maximum number of iteration equal to 1000 to ensure convergence with the number of predictors in the dataset, and balanced class weight to compensate for the imbalance between *Improved* and *Not Improved* participants. This weighting increases the influence of the minority class during training so the model does not default to predicting the majority category.

For the Random Forest, the number of estimators was set to 800, meaning the model was built from 800 individual decision trees. The maximum depth of the trees was left unlimited, allowing them to grow as needed. To reduce noisy splits, trees were required to have at least 5 samples before a split could be made and at least 2 samples in each final leaf. At each split, the model considered only a subset of predictors (the square root of the total), which increases diversity across trees and helps generalization (performance on unseen data). As with Logistic Regression, a balanced class weighting was used so that the minority class (*Improved*) received greater influence during training.

Model performance was evaluated on the test set using accuracy, precision, recall, and F1-score. Given the class imbalance and the clinical relevance of identifying individuals who improve, particular attention was paid to recall for the *Improved* class. Feature coefficients (for Logistic Regression) and feature importances (for the Random Forest) were extracted to assess the relative contribution of each variable.

4.4 Evaluation Metrics

Performance was evaluated using accuracy, precision, recall, and F1-score on the test set. Because the dataset contained more *Not Improved* than *Improved* cases, accuracy alone is not informative; it would remain high even if the model predicted the majority class for nearly everyone, completely ignoring the minority class. Recall for the *Improved* class was examined, as it measures how well the model identifies participants whose PHQ-9 scores decreased. Confusion matrices were used to inspect the types of errors the models made, and model outputs, Logistic Regression coefficients and Random Forest feature importances, were reviewed to see which variables influenced predictions and whether these patterns were clinically plausible. Moreover, ROC–AUC curves were plotted to study the models' ability to separate *Improved* from *Not Improved* cases across different classification thresholds. The ROC curve shows how the true-positive rate and false-positive rate change when the decision threshold is varied, and the AUC calculates the area under the curve, summarizing this into a single number. A higher AUC indicates that the model ranks *Improved* cases above *Not Improved* cases more reliably, regardless of the specific probability cutoff used.

5.Results

5.1 Clustering

Unsupervised clustering was applied to explore whether participants naturally grouped into distinct behavioural profiles. Silhouette scores were calculated for values of *k* from 2 to 10.

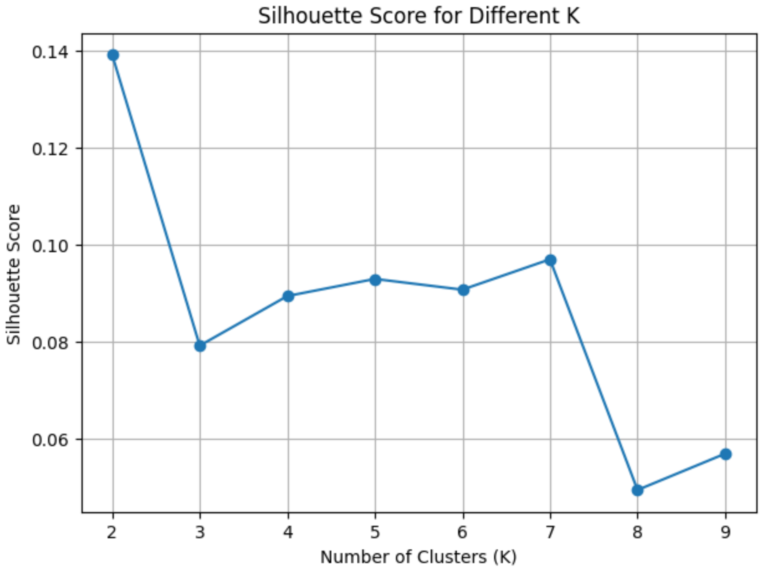


Figure 13: Silhouette score (k from 2 to 10)

The silhouette score peaked at k = 2 but remained stable for k = 4, which offered a good balance between separation quality and cluster interpretability. The silhouette scores were low however, indicating that the behavioural structure in the data is present but not strongly separated.

After assigning cluster labels, the clusters were visualised using a two-dimensional PCA projection.

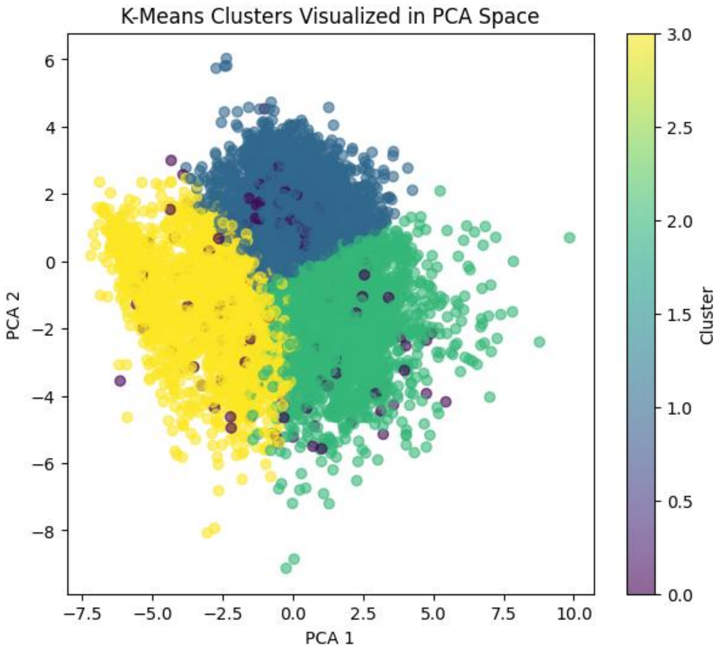


Figure 14: PCA visualization of clusters

The PCA plot shows that the four clusters occupy different regions of the 2-D projection, meaning that k-means identified consistent behavioural groupings, although these groups overlap, which is typical for high-dimensional health-behaviour data.

Cluster sizes were uneven: one large group (Cluster 1: 5,195 participants), two medium groups (Cluster 2: 3,255; Cluster 3: 2,139), and one small group (Cluster 0: 277).

Cluster profiles based on behavioural and clinical variables showed distinct patterns across the four groups. Cluster 1, the largest cluster in the sample, had the highest activity levels, with more steps and more active days, as well as longer sleep durations and higher sleep efficiency. Cluster 2 showed the opposite pattern, with lower daily activity and more sedentary days, combined with longer sleep periods and higher time in bed. Cluster 3 stood out for having the shortest and most irregular sleep, along with lower activity levels. Cluster 0, the smallest group, showed intermediate patterns but had noticeably higher values on several clinical indicators, including comorbidities and migraine history.

Z-scored profiles made these contrasts clearer. Cluster 1 consistently showed positive deviations from the population mean on activity metrics, while Cluster 3 showed strong negative deviations on both sleep and activity measures. Cluster 2 displayed above-average sleep duration and time in bed but low physical activity, and Cluster 0 showed positive deviations mainly on clinical variables rather than behavioural ones.

Average PHQ-9 change was similar across all clusters, ranging from –0.09 to –0.29, indicating that none of these behavioural or clinical patterns aligned with meaningful differences in symptom improvement. This reinforces that, although clustering captured real structure in behaviour and health status, these groupings did not correspond to distinct depression-change trajectories. It also reflects the bigger difficulty of determining which specific behaviours contribute to improvement or worsening, since depression is influenced by many factors, such as psychological, environmental, and contextual, that are not fully captured in wearable or survey data.

5.2 Logistic Regression

The Logistic Regression model achieved an accuracy of 0.57 on the test set. Despite the class imbalance, the model achieved a recall of 0.50 for the *Improved* class and 0.60 for the *Not Improved* class. Precision for the *Improved* class was lower (0.37), indicating that many predicted improvements were incorrect. These patterns reflect the limited signal available in the predictors for distinguishing between the two groups.

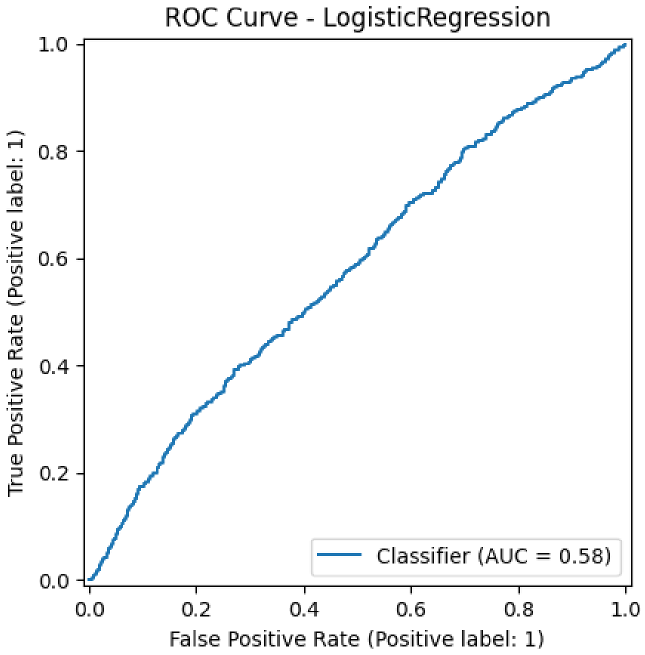


Figure 15: ROC Curve for Logistic Regression

The ROC curve for the Logistic Regression model showed an AUC of 0.58. An AUC close to 0.5 indicates performance similar to random guessing, and a value of 0.58 suggests that the model has only limited ability to rank *Improved* cases above *Not Improved* cases across different thresholds. This aligns with the modest recall and precision values and reinforces that the predictors provide only weak separation between the two outcome groups.

Logistic Regression coefficients represent changes in the log-odds of the outcome per unit increase in the predictor. When exponentiated into odds ratios, values above 1 indicate that higher values of the predictor are associated with greater odds of PHQ-9 improvement, while values below 1 indicate reduced odds of improvement. Odds ratios close to 1 show that the predictor has little practical influence on the outcome. The coefficients and odds-ratios for this model can be found in the Appendix.

Some predictors showed odds ratios slightly above 1, such as spending more time in bed on weekends, engaging more often in eating-related activities, or having a higher proportion of time asleep while in bed on weekends, suggesting small increases in the likelihood of improvement. Others, such as the amount of time actually spent asleep on weekends, older age, or being male, showed odds ratios below 1, indicating slightly lower odds of improvement. These effects were low, and most odds ratios were close to 1, suggesting that the behaviours and personal characteristics did not strongly differentiate participants who improved from those who did not.

5.3 Random Forest

The Random Forest model reached an accuracy of 0.65 on the test set. As with the Logistic Regression model, most predictions were assigned to the *Not Improved* class, reflected in the high recall for that class (0.93) and very low recall for the Improved class (0.06). Precision for the *Improved* class was also low (0.29), indicating that the model struggled to reliably identify individuals whose PHQ-9 scores decreased.

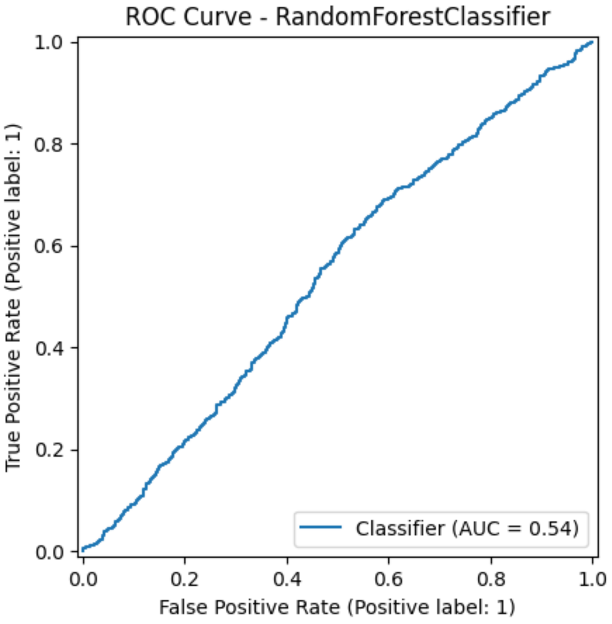


Figure 16: ROC Curve for Random Forest Classifier

The ROC curve for the Random Forest model showed an AUC of 0.54. This value is only slightly above 0.5, indicating performance close to random chance when ranking *Improved* versus *Not Improved* cases across different thresholds. These results show that allowing for non-linear relationships and interactions between predictors did not improve performance, indicating the presence of noise in data or confounding factors.

Feature importance in a Random Forest reflects how much each predictor reduces the classification error across all trees in the ensemble. Predictors that contribute more to splitting the data receive higher importance scores. The list of the most relevant (first 20) feature importance values can be found in the Appendix.

Feature importance values indicated that several sleep-related measures, such as how much time participants spent in bed or asleep on weekdays and weekends, contributed most to the model's decisions. However, feature importance does not indicate the direction of the relationship (whether it contributed to a positive or negative prediction); only how useful a feature was for making splits. Daily step counts and demographic factors such as BMI, weight, height, and age also had measurable influence. However, even the most influential features explained only small portions of the variability in the outcome; this is consistent with the difficulty of predicting symptom improvement from the available data.

6.Conclusion & Discussion

This project investigated whether behavioural, lifestyle, demographic, and clinical variables from the DiSCover study could help describe or predict short-term changes in depressive symptoms. The analysis combined exploratory clustering with supervised models to assess whether patterns in sleep, activity, and related factors aligned with changes in PHQ-9 scores over three-month intervals.

Clustering revealed that the dataset does contain meaningful variation in daily behaviour. Participants differed in activity levels, sleep duration and efficiency, and several clinical indicators. These differences formed consistent clusters that were visible in a PCA projection. However, the clusters showed very similar average PHQ-9 changes, indicating that the relevant behavioural patterns in the dataset did not correspond to distinct symptom-change trajectories. Clustering was therefore useful for understanding population structure, to improve data exploration, but did not provide evidence of clear behavioural pathways to improvement or worsening.

Supervised models showed very modest performance. Logistic Regression identified small associations between certain behaviours and PHQ-9 improvement, but the model struggled to reliably distinguish *Improved* from *Not Improved* cases. The Random Forest model detected non-linear patterns and assigned higher importance to several sleep- and activity-related features, but the gains in performance were minor and recall for the *Improved* class remained low. These results suggest that, using the available predictors and a three-month time window, it is challenging to isolate behavioural signatures that consistently correspond to symptom change.

Several factors may help us explain these outcomes. The dataset reflects complex real-world conditions, and depressive symptoms are influenced by many elements not captured here, i.e. psychosocial context, or moment-to-moment emotional states. Behavioural measures from wearables may not directly reflect the processes that drive changes in depression scores. The three-month interval may further dilute short-term patterns that occur at shorter, finer temporal scales.

There are also limitations specific to the present analysis. The modelling strategy prioritised interpretable methods and did not explore more complex or black-box approaches that might capture more structure. Feature selection was carried out manually to ensure interpretability and completeness, but this approach may not yield the most predictive subset of variables; automated feature-selection or penalised models could be explored in future work. Moreover, the population-level modelling used here does not account for individual differences, and personalised or longitudinal models might reveal stronger relationships.

Future analysis could add automated feature-selection methods, more flexible models such as gradient boosting or temporal neural networks, and designs that operate at finer temporal scales (daily or weekly). Personalised modelling approaches may also reveal some stronger relationships by accounting for individual behavioural baselines rather than pooling all participants together.

Overall, the results show that, within this dataset and modelling framework, behavioural and demographic variables offer only limited ability to distinguish participants who improve from those who do not over a three-month period. At the same time, the clustering results demonstrate that meaningful behavioural variation exists, and future work could explore whether these patterns become more informative with additional contextual data, alternative modelling approaches, or more fine-grained temporal information.

7. Appendix

This appendix contains supplementary material that supports the analyses presented in the main report.

Percentage of Missing Values for Each Variable in the DiSCover Dataset

Feature	Missing (%)
steps_mvpa_iqr	86.29
steps__mvpa__sum__intercept_	75.82
steps__mvpa__sum__score_	75.82
steps__mvpa__sum__coeff_	75.82
phq9_cat_end	69.56
phq9_cat_start	69.56
phq9_score_start	69.56
phq9_score_end	69.56
sleep_main_start_hour_adj_iqr	56.49
sleep_ratio_asleep_in_bed_iqr	56.49
sleep__awake__sum__score_	43.24
sleep__awake_regions__countDistinct__intercept_	43.24
sleep__awake_regions__countDistinct__score_	43.24
sleep__awake__sum__intercept_	43.24
sleep__awake__sum__coeff_	43.24
sleep__awake_regions__countDistinct__coeff_	43.24

sleep__nap_count__score_	43.22
sleep__main_efficiency__score_	43.22
sleep__main_efficiency__intercept_	43.22
sleep__main_efficiency__coeff_	43.22
sleep__main_start_hour_adj__coeff_	43.22
sleep__main_start_hour_adj__score_	43.22
sleep_ratio_asleep_in_bed__score_	43.22
sleep_ratio_asleep_in_bed__intercept_	43.22
sleep_ratio_asleep_in_bed__coeff_	43.22
sleep__nap_count__coeff_	43.22
sleep__nap_count__intercept_	43.22
sleep__main_start_hour_adj__intercept_	43.22
steps_lpa_iqr	38.21
sleep_asleep_iqr	37.73
sleep_in_bed_iqr	37.73
steps_awake_sum_iqr	31.55
sleep__total_asleep_minutes__score_	25.83
sleep__total_in_bed_minutes__coeff_	25.83
sleep__total_in_bed_minutes__intercept_	25.83
sleep__total_in_bed_minutes__score_	25.83
sleep__total_asleep_minutes__intercept_	25.83
sleep__total_asleep_minutes__coeff_	25.83
steps__light_activity__sum__coeff_	25.54
steps__light_activity__sum__score_	25.54
steps__light_activity__sum__intercept_	25.54
steps__mvpa__sum__intercept	22.70
steps__mvpa__sum__coeff	22.70
steps__mvpa__sum__score	22.70
steps__awake__sum__coeff_	20.55
steps__awake__sum__intercept_	20.55
steps__awake__sum__score_	20.55
steps__rolling_6_sum__max__coeff_	20.48
steps__rolling_6_sum__max__score_	20.48
steps__rolling_6_sum__max__intercept_	20.48
steps__not_moving__sum__score_	20.27
steps__not_moving__sum__intercept_	20.27
steps__not_moving__sum__coeff_	20.27
steps__streaks__countDistinct__intercept_	20.19
steps__streaks__countDistinct__score_	20.19
steps__streaks__countDistinct__coeff_	20.19
sleep__main_start_hour_adj__intercept	9.81

sleep__main_start_hour_adj__score	9.81
sleep__main_start_hour_adj__coeff	9.81
sleep__nap_count__score	9.81
sleep_ratio_asleep_in_bed__score	9.81
sleep__awake_regions__countDistinct__score	9.81
sleep__main_efficiency__score	9.81
sleep__awake__sum__score	9.81
sleep__awake__sum__intercept	9.81
sleep__awake_regions__countDistinct__intercept	9.81
sleep__awake_regions__countDistinct__coeff	9.81
sleep__awake__sum__coeff	9.81
sleep__nap_count__intercept	9.81
sleep__nap_count__coeff	9.81
sleep__main_efficiency__coeff	9.81
sleep__main_efficiency__intercept	9.81
sleep_ratio_asleep_in_bed__intercept	9.81
sleep_ratio_asleep_in_bed__coeff	9.81
sleep_ratio_asleep_in_bed_weekend_mean	9.62
sleep_ratio_asleep_in_bed_mean_recent	7.02
sleep_ratio_asleep_in_bed_weekday_mean	4.40
sleep_main_start_hour_adj_range	4.17
sleep_main_start_hour_adj_median	4.17
sleep_asleep_weekend_mean	3.90
sleep_in_bed_weekend_mean	3.90
sleep__total_in_bed_minutes__coeff	2.23
sleep__total_asleep_minutes__score	2.23
sleep__total_asleep_minutes__coeff	2.23
sleep__total_asleep_minutes__intercept	2.23
sleep__total_in_bed_minutes__intercept	2.23
sleep__total_in_bed_minutes__score	2.23
sleep_asleep_mean_recent	1.63
sleep_in_bed_mean_recent	1.63
sleep_asleep_weekday_mean	1.02
sleep_in_bed_weekday_mean	1.02
money_assistance	0.75
insurance	0.74
steps__light_activity__sum__score	0.73
steps__light_activity__sum__intercept	0.73
steps__light_activity__sum__coeff	0.73
pregnant	0.51
money	0.36

steps_rolling_6_median_recent	0.33
steps_rolling_6_max_recent	0.33
birth	0.28
steps__awake__sum__intercept	0.28
steps__awake__sum__score	0.28
steps__awake__sum__coeff	0.28
steps__rolling_6_sum__max__coeff	0.28
steps__rolling_6_sum__max__score	0.28
steps__rolling_6_sum__max__intercept	0.28
educ	0.10
bmi	0.06
steps_awake_mean	0.06
height	0.05
steps__not_moving__sum__score	0.03
steps__not_moving__sum__intercept	0.03
steps__not_moving__sum__coeff	0.03
weight	0.01
Variables with 0% missingness	
steps__sedentary_day_count_	0.00
sleep__hypersomnia_count_	0.00
sleep__hyposomnia_count_	0.00
steps__active_day_count_	0.00
steps__streaks__countDistinct__intercept	0.00
steps__streaks__countDistinct__score	0.00
steps__streaks__countDistinct__coeff	0.00
race_asian	0.00
steps_mvpa_sum_recent	0.00
birthyear	0.00
race_other	0.00
steps_lpa_sum_recent	0.00
sex	0.00
race_white	0.00
race_black	0.00
race_hispanic	0.00
trauma	0.00
comorbid_gout	0.00
comorbid_migraines	0.00
comorbid_ms	0.00
comorbid_osteoporosis	0.00
household	0.00
comorbid_cancer	0.00

comorbid_diabetes_typ1	0.00
comorbid_diabetes_typ2	0.00
med_stop	0.00
med_start	0.00
meds_migraine	0.00
num_migraine_days	0.00
med_dose	0.00
nonmed_start	0.00
nonmed_stop	0.00
life_meditation	0.00
life_red_stop_alcoh	0.00
life_activity_eating	0.00
med_nonmed_dnu	0.00
life_stress	0.00
comorbid_neuropathic	0.00
comorbid_arthritis	0.00
participant_id	0.00
month	0.00

Variables used for modelling

Sleep variables

- sleep_asleep_weekday_mean
- sleep_asleep_weekend_mean
- sleep_in_bed_weekday_mean
- sleep_in_bed_weekend_mean
- sleep_ratio_asleep_in_bed_weekday_mean
- sleep_ratio_asleep_in_bed_weekend_mean
- sleep_main_start_hour_adj_median
- sleep__hypersomnia_count_
- sleep__hyposomnia_count_

Activity variables

- steps_awake_mean
- steps__active_day_count_
- steps__sedentary_day_count_

Lifestyle variables

- life_meditation

- life_stress
- life_activity_eating
- life_red_stop_alcoh

Demographic and socioeconomic variables

- age
- sex
- race_white
- race_black
- race_hispanic
- race_asian
- race_other
- educ
- height
- weight
- bmi
- pregnant
- birth
- insurance
- household
- money
- money_assistance

Clinical history

- trauma
- has_migraine
- comorbid_cancer
- comorbid_diabetes_typ1
- comorbid_diabetes_typ2
- comorbid_gout
- comorbid_ms
- comorbid_osteoporosis
- comorbid_neuropathic
- comorbid_arthritis

Logistic Regression Coefficients

Feature	Coefficient	Odds Ratio
sleep_in_bed_weekend_mean	0.218404	1.244089
life_activity_eating	0.183159	1.201005
sleep_ratio_asleep_in_bed_weekend_mean	0.153019	1.165347
has_migraine	0.119087	1.126468
comorbid_neuropathic	0.104601	1.110268
comorbid_arthritis	0.088390	1.092414
money	0.086559	1.090416
bmi	0.057875	1.059583
race_white	0.039728	1.040527
sleep__hyposomnia_count_	0.036199	1.036862
comorbid_diabetes_typ1	0.033519	1.034087
race_asian	0.025951	1.026291
life_stress	0.022751	1.023012
sleep__hypersomnia_count_	0.019555	1.019748
comorbid_ms	0.018616	1.018790
race_hispanic	0.015525	1.015646
trauma	0.013596	1.013689
household	0.012705	1.012786
steps_awake_mean	0.012336	1.012412
steps__sedentary_day_count_	0.009741	1.009788
height	0.008475	1.008511
birth	0.008352	1.008387
race_black	0.007848	1.007879

comorbid_cancer	0.004756	1.004767
educ	-0.001535	0.998467
life_meditation	-0.003774	0.996233
life_red_stop_alcoh	-0.004222	0.995787
insurance	-0.004718	0.995293
sleep_in_bed_weekday_mean	-0.005098	0.994915
sleep_main_start_hour_adj_median	-0.012458	0.987620
comorbid_osteoporosis	-0.012882	0.987200
comorbid_diabetes_typ2	-0.015577	0.984543
sleep_asleep_weekday_mean	-0.016821	0.983320
race_other	-0.018100	0.982063
steps__active_day_count_	-0.018411	0.981758
comorbid_gout	-0.022066	0.978176
pregnant	-0.027039	0.973323
money_assistance	-0.048715	0.952453
weight	-0.051344	0.949951
sleep_ratio_asleep_in_bed_weekday_mean	-0.055785	0.945743
age	-0.068798	0.933515
sex	-0.090212	0.913737
sleep_asleep_weekend_mean	-0.179871	0.835378

Random Forest Feature Importance

Feature	**Importance**
steps_awake_mean	0.073175
sleep_ratio_asleep_in_bed_weekday_mean	0.070656

sleep_ratio_asleep_in_bed_weekend_mean	0.068864
sleep_in_bed_weekday_mean	0.067864
sleep_asleep_weekday_mean	0.066720
sleep_in_bed_weekend_mean	0.065263
sleep_asleep_weekend_mean	0.064905
bmi	0.063194
weight	0.057060
age	0.054103
height	0.047471
sleep_main_start_hour_adj_median	0.038705
household	0.028128
steps__active_day_count_	0.027865
steps__sedentary_day_count_	0.027564
sleep__hyposomnia_count_	0.023938
educ	0.023540
money	0.016999
has_migraine	0.012667
life_activity_eating	0.012469

8.References

Makhmutova, M., Kainkaryam, R., Ferreira, M., Min, J., Jaggi, M., & Clay, I. (2022). Predicting changes in depression severity using the PSYCHE-D (Prediction of Severity Change-Depression) model involving person-generated health data: Longitudinal case-control observational study. JMIR mHealth and uHealth, 10(3), e34148. https://doi.org/10.2196/34148

Price, G. D., Heinz, M. V., Song, S. H., et al. (2023). Using digital phenotyping to capture depression symptom variability: Detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. Translational Psychiatry, 13, 381. https://doi.org/10.1038/s41398-023-02669-y

Pedrelli, P., Fedor, S., Ghandeharioun, A., Howe, E., Ionescu, D. F., Bhathena, D., Fisher, L. B., et al. (2020). Monitoring changes in depression severity using wearable and mobile sensors. Frontiers in Psychiatry, 11, 584711. https://doi.org/10.3389/fpsyt.2020.584711

Sun, S., Folarin, A., Zhang, Y., Cummins, N., Garcia-Dias, R., Stewart, C., et al. (2023). Challenges in using mHealth data from smartphones and wearable devices to predict depression symptom severity: Retrospective analysis. Journal of Medical Internet Research, 25, e45233. https://doi.org/10.2196/45233

World Health Organization. (2025). Depression. World Health Organization (WHO). Retrieved from https://www.who.int/news-room/fact-sheets/detail/depression

University of Michigan School of Public Health. (2025). College student mental health shows third consecutive year of improvement. Healthy Minds Network / University of Michigan. Retrieved from https://sph.umich.edu/news/2025posts/college-student-mental-health-third-consecutive-year-improvement.html