Predictive Modeling of Depression Risk

A Machine Learning Approach

Chamodh Wijayawardhana

KAGGLE-PLAYGROUND SIERES (SEASON 4, EPISODE 11)

Abstraction

This project explores the use of machine learning to predict depression outcomes from a synthetically generated menta health survey dataset, provided through Kaggle's Playground Series – Season 4, Episode 11. The dataset contains demographic, lifestyle, and survey-related features, with the target variable indicating whether an individual is experiencing depression. A structured workflow was applied, beginning with data cleaning and exploratory analysis, followed by feature engineering, model training, and hyperparameter tuning. Several algorithms were evaluated, including Logistic Regression, Random Forest, XGBoost, and LightGBM, with performance assessed using Accuracy Score as defined by the competition. The final model demonstrated meaningful predictive power and highlighted feature patterns associated with depression risk. This work not only demonstrates the application of machine learning to mental health–related data but also provides practical experience in handling tabular datasets within a competitive data science setting.

Table of Content

1. Introduction

Depression is one of the most common mental health disorders worldwide, affecting individuals across age groups, professions, and social backgrounds. Early identification of depression is vital, as timely intervention can significantly improve health outcomes and quality of life. However, mental health data is often complex, multidimensional, and difficult to analyze using traditional statistical methods.

With the rise of machine learning, predictive modeling offers new opportunities to explore patterns in mental health–related data. By analyzing demographic factors, lifestyle choices, and survey responses, it is possible to uncover relationships that may signal a higher risk of depression. Such approaches do not replace clinical diagnosis but can serve as valuable tools for screening, prevention, and awareness.

This project is based on the Kaggle *Playground Series – Season 4, Episode 11: Exploring Mental Health Data*. The competition provides a synthetically generated dataset designed to simulate responses from a mental health survey. The task is to predict whether an individual is experiencing depression based on a set of tabular features. The evaluation metric is Accuracy Score, and the goal is to design and compare machine learning models that can effectively capture patterns in the data.

By engaging with this challenge, the project not only demonstrates the application of machine learning to a socially significant problem but also serves as an opportunity to practice data preprocessing, feature engineering, model evaluation, and interpretation within a competitive framework.

2. Dataset Description

The dataset used in this project was released as part of the Kaggle *Playground Series – Season 4, Episode 11: Exploring Mental Health Data*. Both the training and test sets were synthetically generated using a deep learning model trained on the original *Depression Survey/Dataset for Analysis*.

The training dataset (`train.csv`) contains **140,700 rows** and **20 columns**, including demographic details, academic or professional background, lifestyle habits, and mental health–related factors. The target variable is **Depression**, a binary label (0 = no depression, 1 = depression). The test dataset (`test.csv`) has the same structure but without the target column.

It is worth noting that the synthetic dataset intentionally contains several **data artifacts**, such as missing values and imperfect distributions, to make the challenge more realistic. Approximately 30–40% of values are missing in some features.

Overall, the dataset is moderately sized and relatively approachable for classification tasks, offering opportunities to experiment with preprocessing strategies, visualization techniques, and predictive modeling approaches

3. Exploratory Data Analysis (EDA)

To gain a better understanding of the mental health dataset, an extensive exploratory data analysis was conducted. This step helped identify key trends, relationships, and potential data quality issues before preprocessing and modeling.

3.1. Basic Exploration

The dataset (train.csv) contains 140,700 rows and 20 columns, mixing numerical and categorical features. Initial inspection revealed that some features required type corrections (e.g., converting survey responses into categorical variables) and that missing data was concentrated in academic and profession-related columns. Descriptive statistics highlighted meaningful variation in age, work/study hours, and stress-related measures.

3.2. Target Variable

The target variable, **Depression**, is binary but highly imbalanced. Approximately **115,000 records (≈82%)** belong to class 1 (no depression), while only **25,000 records (≈18%)** belong to class 1 (depression). This imbalance has important implications for modeling: accuracy alone may be misleading, as a trivial model that predicts "no depression" for all cases would already achieve over 80% accuracy.

To address this imbalance during the modeling stage, **class weighting** should be considered. By adjusting the loss function to penalize misclassification of the minority class more heavily, models can be encouraged to learn meaningful patterns for both depressed and non-depressed groups instead of being biased toward the majority class.

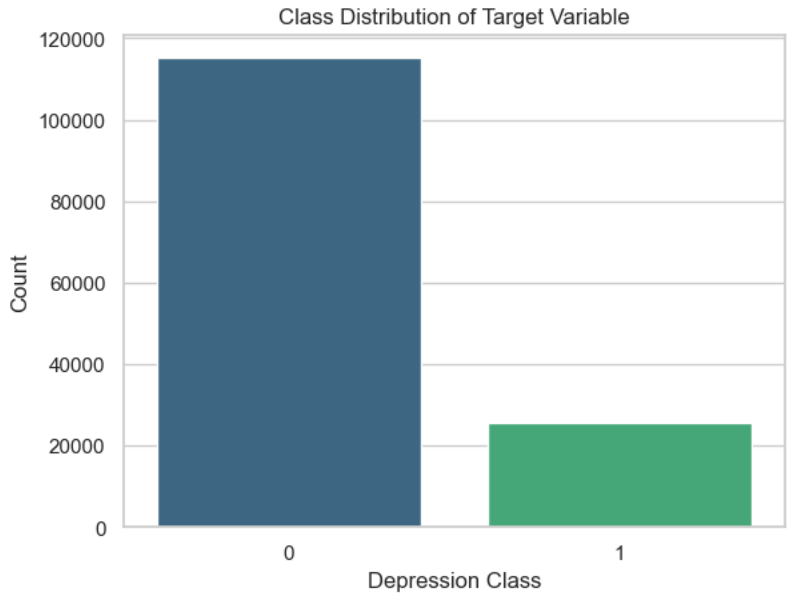


Figure 1: Class Distribution of Target Variable

3.3. Categorical Associations (Chi-Square Tests)

The chi-square (X^2) test is a statistical method used to determine whether there is a significant association between two categorical variables. The test works by comparing the observed frequencies in a contingency table (the actual counts from the dataset) with the expected frequencies (the counts that would occur if the two variables were completely independent). The chi-square statistic is calculated as:

$$\chi 2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} represents the observed frequency in cell i, j of the contingency table, and E_{ij}represents the expected frequency for that same cell under the assumption of independence.

The resulting statistic follows a chi-square distribution with degrees of freedom equal to $(r-1)(c-1)$, where r is the number of rows and c is the number of columns in the contingency table. A **p-value** is then derived to assess that the observed differences

between categories are unlikely to be due to chance, and therefore the two variables are likely associated.

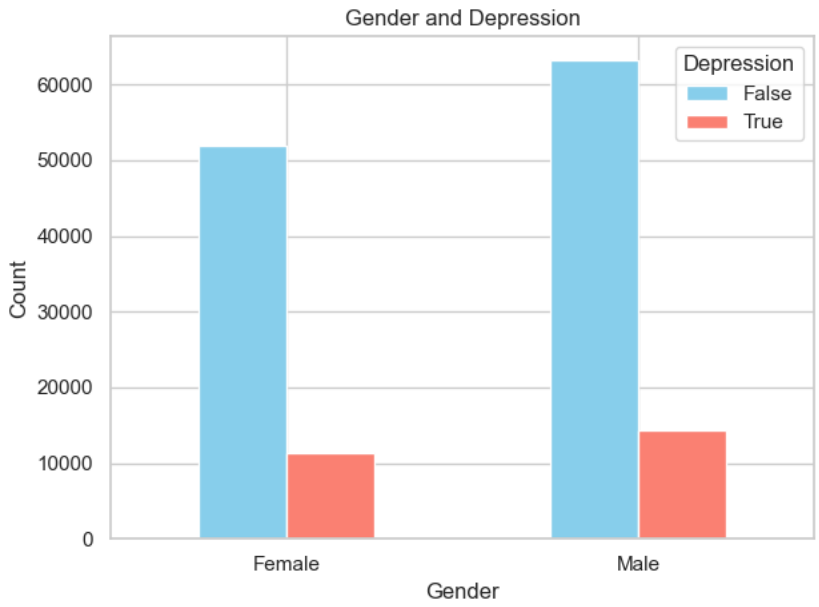
To examine whether categorical features have a statistically significant relationship with the target variable (*Depression*), chi-square tests of independence were performed. This method compares the observed distribution of depression across categories with the expected distribution under the assumption of independence.

- Gender

  The Test indicated a weak but statistically significant association between gender and depression. While both male and female respondents experience depression, the proportions differ slightly suggesting gender may play a minor role as a predictive factor.

Contingency Table (Observed Frequencies):

Depression	0	1
Gender		
Female	51965	11271
Male	63168	14296



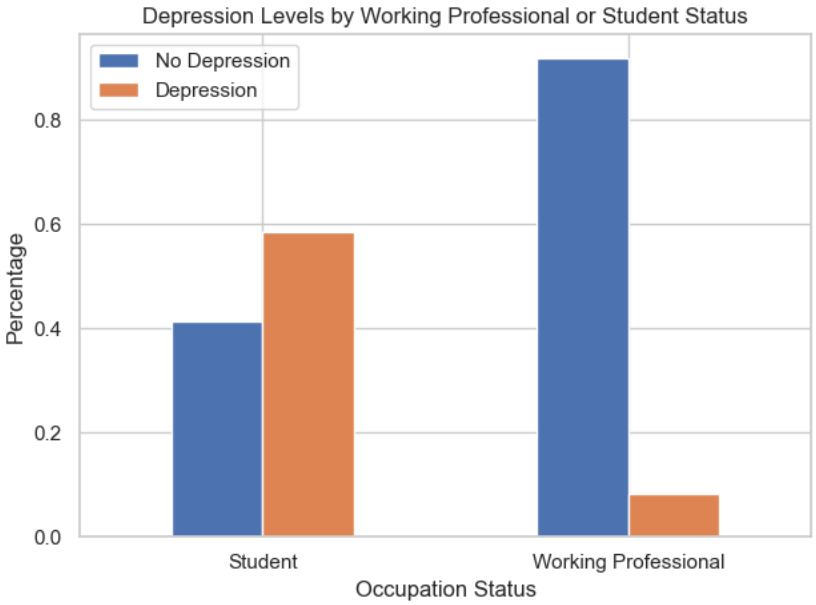
The contingency tables show that out of **63,168** males without depression, **14,296** were classified as depressed. Among females, **51,965** reported no depression, while 11,271

experienced depressions. When visualized, the bar chart confirms that although the total number of male respondents is slightly higher than females, the proportion of depression cases is also higher in males.

This suggests that gender is associated with depression in this dataset. The chi-square test confirmed this relationship to be statistically significant, though the effect size is modest. In practical terms, gender alone is not a strong predictor of depression, but it contributes as one of many interacting features.

- Working Professional vs Student

  A stronger association was observed between professional status and depression. Students reported higher levels of academic pressure and related depression risk, while working professionals exhibited variability linked more to job satisfaction and work pressure.



The bar char shows a striking difference in depression prevalence between students and working professionals. Among students, the proportion of individuals experiencing depression is nearly **60%** making it the majority group. In contrast, working professionals exhibit the opposite trend: over **90%** report no depression, with only a small minority affected.

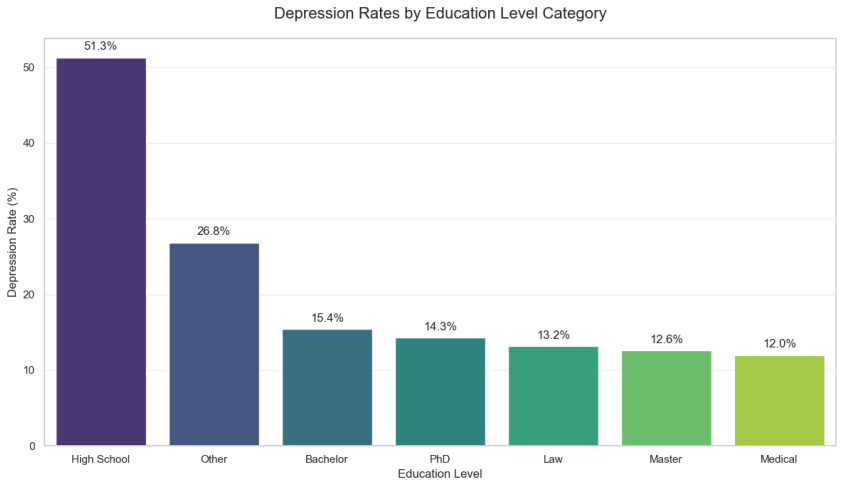
This suggests that students in this dataset face substantially higher levels of depression compared to professionals. The likely contributors include academic pressure, study satisfaction, and CGPA – features that primarily apply to students and are strongly associated with mental health outcomes. Conversely, working professionals may instead be influenced by job satisfaction and work pressure, which appear less severe on average.

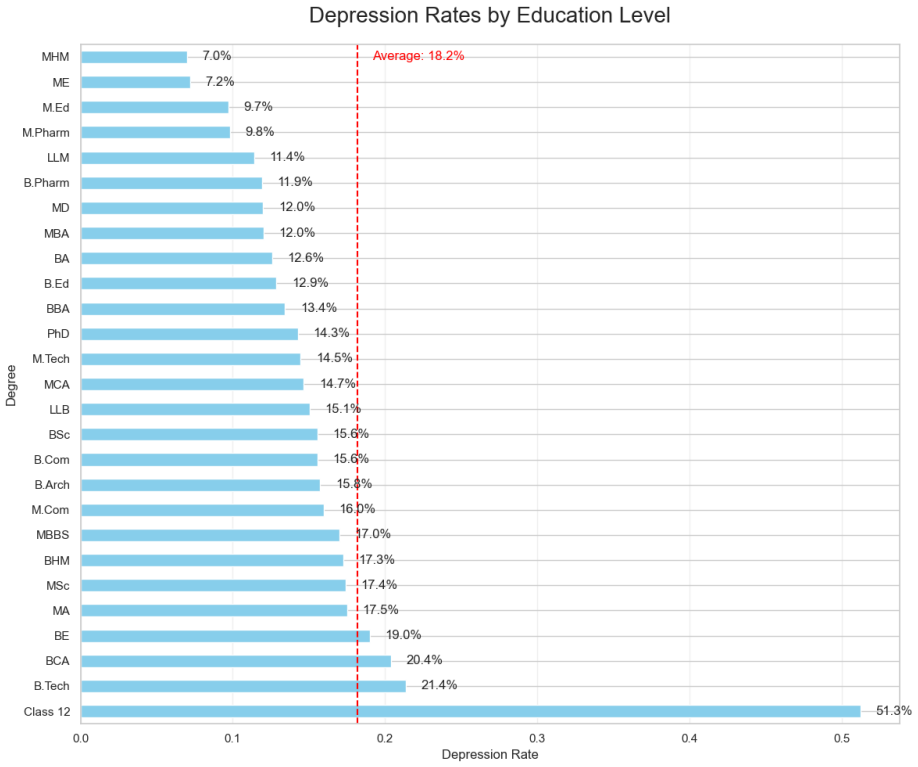
- Profession

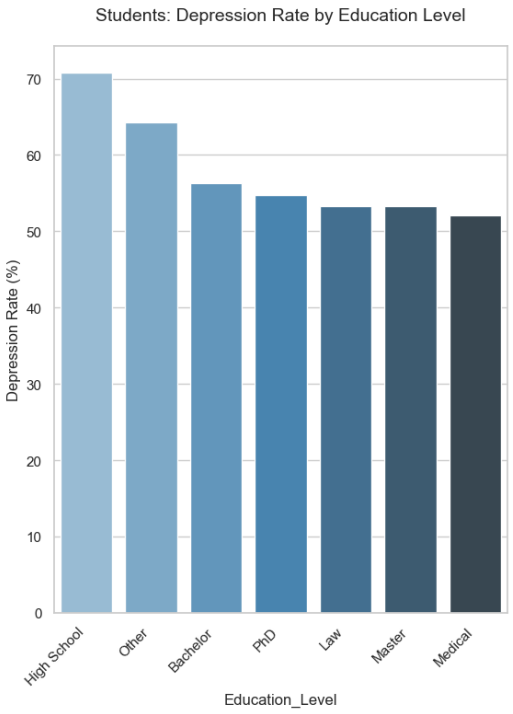
  The chi-square test revealed a clear relationship between profession and depression prevalence. Certain professions, particularly those in education and healthcare, displayed higher rates of depression compared to technical and business-oriented roles. This imbalance does not require direct correction, as Profession is treated as a single categorical feature rather than the prediction target. Instead, appropriate encoding techniques (such as grouping rare categories or applying target encoding) will be considered during preprocessing to ensure that small-sample professions do not introduce noise.

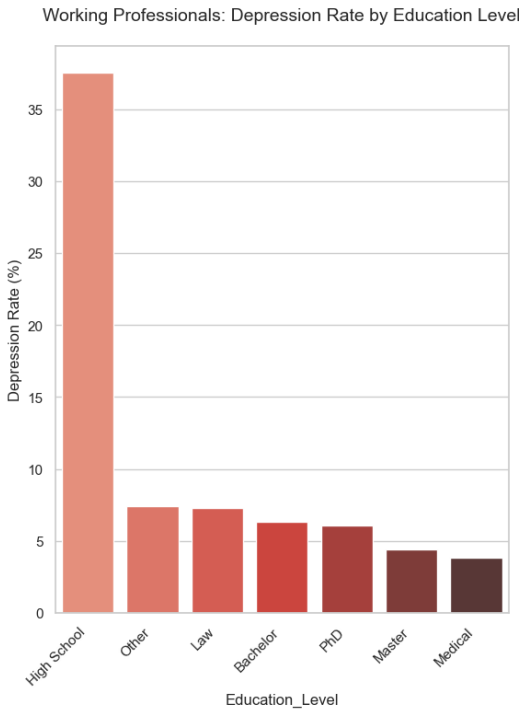
- Education Level (degree)

  Education level showed significant differences in depression rates. Respondents with only high school education were more likely to report depression compared to those with university or postgraduate qualifications.




Depression Rates by Education Level
Average: 18.2%
MHM 7.0%
ME 7.2%
M.Ed 9.7%
M.Pharm 9.8%
LLM 11.4%
B.Pharm 11.9%
MD 12.0%
MBA 12.0%
BA 12.6%
B.Ed 12.9%
BBA 13.4%
PhD 14.3%
M.Tech 14.5%
MCA 14.7%
LLB 15.1%
BSc 15.6%
B.Com 15.6%
B.Arch 15.8%
M.Com 16.0%
MBBS 17.0%
BHM 17.3%
MSc 17.4%
MA 17.5%
BE 19.0%
BCA 20.4%
B.Tech 21.4%
Class 12 51.3%
Degree
0.0
0.1
0.2
0.3
0.4
0.5
Depression Rate


Students: Depression Rate by Education Level
Depression Rate (%)
0
10
20
30
40
50
60
70
High School
Other
Bachelor
PhD
Law
Master
Medical
Education_Level


Working Professionals: Depression Rate by Education Level
Depression Rate (%)
0
5
10
15
20
25
30
35
High School
Other
Law
Bachelor
PhD
Master
Medical
Education_Level

The analysis of depression rates across education levels revealed strong disparities. Individuals with only high school education (Class 12) reported by far the highest depression rate, at more than fifty percent, which is well above the average of eighteen percent. In contrast, respondents holding advanced degrees such as Masters, Medical or Law qualifications showed much lower rates, typically in the twelve to thirteen percent range.

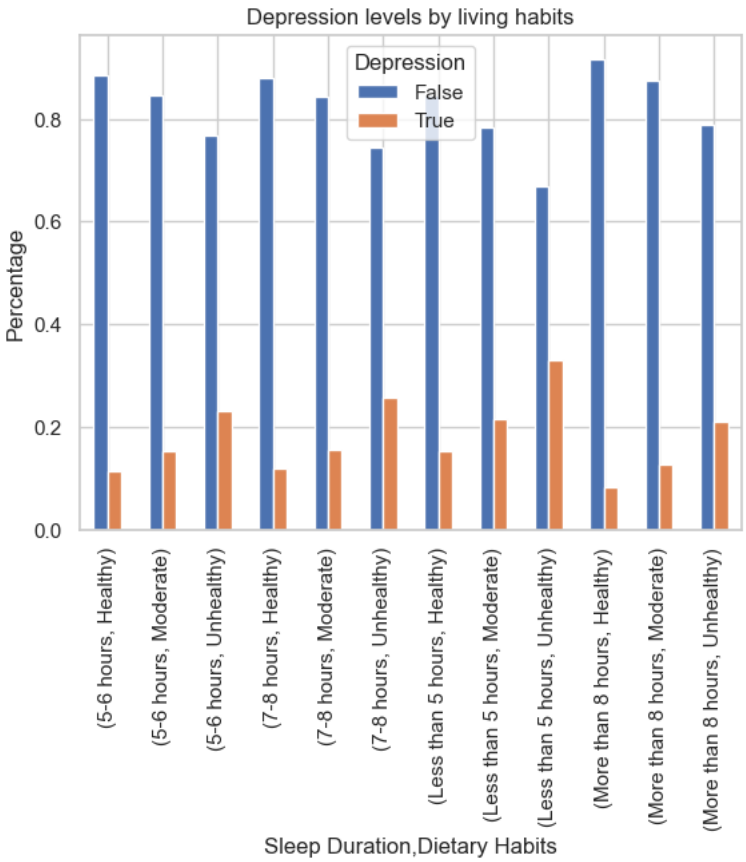
When the degrees were grouped into broader categories, the same pattern became clearer. High school graduates stood out as the most vulnerable group, while those with higher qualifications such as Bachelor's, PhD, Law, Master's, or medical degrees consistently exhibited lower depression prevalence. The "Other" category, which includes less common or alternative qualifications, also showed elevated rates, at around twenty-seven percent, but still substantially below the levels seen among high school students.

Splitting the data further by occupation status highlighted the role of context. Among students, high schoolers reported the highest depression rates, reaching over seventy percent, while even those pursuing higher education showed relatively elevated levels between fifty and sixty percent. Among working professionals, depression rates were much lower overall, yet high school graduates again appeared as a vulnerable group, with rates close to thirty-seven percent.

These findings suggests that lower educational attainment is strongly associated with higher depression prevalence, particularly among students still within the academic system. Higher education levels, in contrast, appear to provide a protective effect, with advanced qualifications linked to reduced likelihood of depression.

- Lifestyle Factors (Sleep Duration, Dietary Habits)

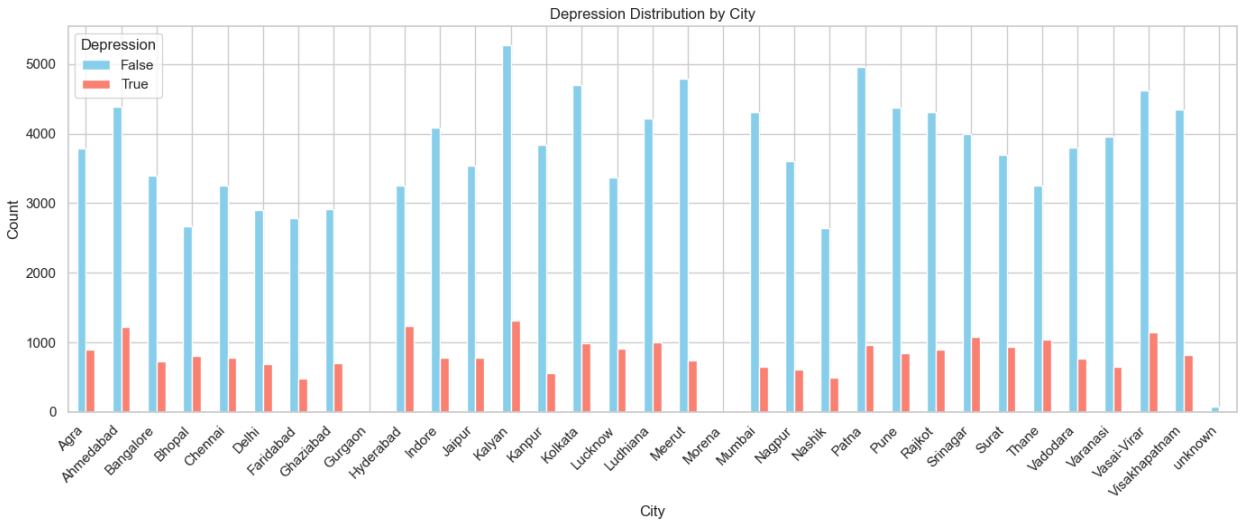
  Both variables exhibited associations with depression. Shorter sleep durations and certain dietary categories correlated with higher depression prevalence. While statistically significant, these associations maybe confounded by other factors such as stress or workload.



Overall, chi-square analyses confirmed that categorical variables such as profession, education, lifestyle factors, and risk indicators (e.g. suicidal thoughts, family history of mental illness) hold predictive value for depression classification. These insights informed later preprocessing steps, particularly the encoding of categorical variables and handling of high-cardinality features.

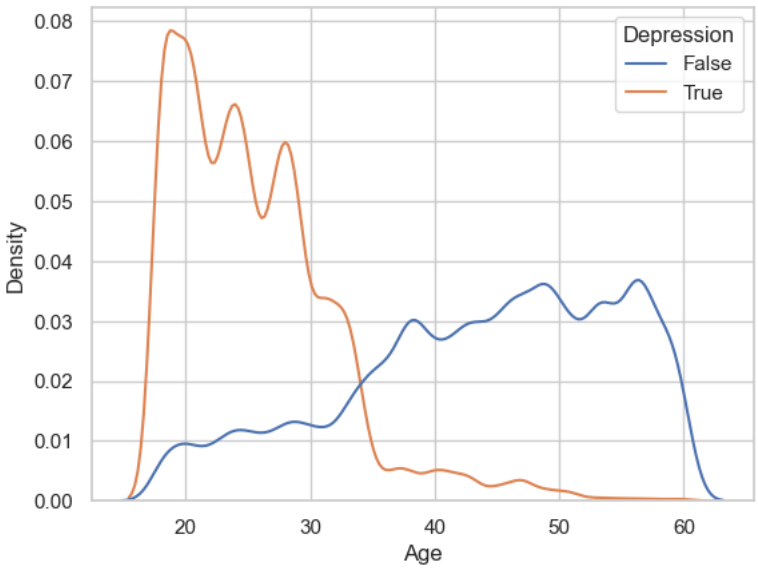
- Geographic Variations

  Geographic differences in depression prevalence are often studied in real-world data because social, economic, and cultural conditions can vary substantially across regions. In this dataset, the city variable provides a proxy for geographic location, but the synthetic nature of the data means the results should be interpreted with caution.



As shown in Figure X, the distribution of depression cases across cities is broadly similar. Larger cities such as Bangalore, Hyderabad, and Mumbai report higher absolute counts, largely due to sample size rather than a meaningful increase in prevalence. Smaller cities display comparable depressed-to-nondepressed ratios, with only minor deviations such as slightly elevated rates in Ahmedabad and Indore. An “Unknown” category also appears, capturing missing or miscoded entries. Overall, the chart suggests that geography is not a strong driver of variation in this dataset, and city-level effects are unlikely to serve as major predictors without interaction with other contextual features.

- Age



Age is a critical demographic factor often linked to mental health outcomes. In this dataset, in the graph the density distribution of depressed and non-depressed individuals across from different ages. The curve for depressed respondents' peaks sharply between the late teens and late twenties, suggesting that younger individuals are more likely to experience depression. After age thirty the depression curve declines steeply and remains consistently lower than the non-depressed group.

Although these findings align with common concerns about youth vulnerability to depression, the synthetic nature of the dataset means results should be interpreted cautiously. Still the chart highlights age as a potentially important predictor in modeling depression risk

4. Data preparation

4.1. Handling missing values

The dataset contained several features with large proportion so missing entries. Academic Pressure, CGPA, and Study Satisfaction were missing in more than 70% of the data, largely because these questions apply only to students and not to working professionals. Profession and Job Satisfaction also had partial missingness, while Work Pressure and Financial Stress contained moderate gaps.

```
id                                        0
Name                                      0
Gender                                    0
Age                                       0
City                                      0
Working Professional or Student           0
Profession                                0
Academic Pressure                    112803
Work Pressure                         27918
CGPA                                 112802
Study Satisfaction                   112803
Job Satisfaction                      27910
Sleep Duration                           57
Dietary Habits                            0
Degree                                    0
Have you ever had suicidal thoughts ?     0
Work/Study Hours                          0
Financial Stress                          4
Family History of Mental Illness          0
Depression                                0
dtype: int64
```

For categorical and ordinal variables such as Job Satisfaction, missing values were filled using stratified medians based on related demographic and contextual factors, including profession, age group, and working hours. This approach preserved subgroup differences rather than applying a single global fill.

Variables with more complex relationships, such as Financial Stress and Work Pressure, were imputed using predictive modeling. Other available features were used to estimate likely values, with the results constrained to the valid scale of the variable.

This strategy combined simple rule-based imputations where missingness was structurally determined with more sophisticated predictive imputations for variables influenced by

multiple factors. As a result, missing values were handled in a way that respected the underlying data structure and minimized information loss.

4.2. Fixing data value mismatches

In addition to missing values, the dataset contained inconsistencies typos, and out-of-range responses that required standardization. Several steps were applied to ensure consistent categories and valid ranges:

- **Professions** were mapped to a fixed list of valid occupations. Any profession not included in the list (including spelling variants or uncommon entries) was recoded as "unknown"
- **Cities** were restricted to a set of valid locations, with unrecognized names similarly mapped to "unknown"
- **Dietary Habits** were cleaned by consolidating redundant categories. For example, "More Healthy" was recoded as "Healthy", while "Less Healthy" and "Less than Healthy" were grouped under "Moderate". Any unrecognized labels were set to "unknown".
- **Degrees** were restricted to a validated list of recognized qualifications. Any irregular or unrecognized entries were mapped to "unknown"
- **Financial Stress,** originally recorded with decima, values and possible out-of-range responses, was rounded and clipped to the valid Likert scale of 1-5.

These cleaning steps ensured that categorical features were internally consistent, reduced noise from misspellings and rare categories, and constrained numeric fields to valid ranges. This stage was essential for preparing the dataset for encoding and modeling, particularly for high cardinality variables like City and Profession.

4.3. Preprocessing

To prepare the dataset for machine learning, a preprocessing pipeline was constructed using scikit-learn's Pipeline and ColumnTransformer. The goal of this pipeline was to automate feature engineering, cleaning, and transformation in a reproductive way.

First, raw variables were reshaped into more useful forms. Age was discretized into categorical bins (18-25, 26-35, 36-45, 46-55, and 56-60) capturing nonlinear relationships

between age groups and depression risk. Similarly, study and work hours were grouped into ranges such as “Less than 2,” “3-5”, “6-8”, and “9-12”, to simplify interpretation and reduce the influence of extreme outliers. Binary variables such as suicidal thoughts and family history of mental illness were explicitly cast as categorical features for proper encoding.

Next, irrelevant identifiers (id, Name and the target label Depression) were dropped from the feature matrix to avoid data leakage.

The transformation stage then applied different encoders to numeric and categorical variables. Numerical features such as CGPA and StudentFlag were standardized using z-score scaling to place them on a common scale. Categorical variable (e.g., gender, city, profession, degree, lifestyle factors) were transformed using one-hot encoding, ensuring that models could handle them appropriately without imposing ordinal structure.

By combining all these steps into a single pipeline, preprocessing became fully automated and consistent across both training and test data. This approach not only reduced manual intervention but also ensured that feature engineering and transformations were applied identically during model development and evaluation.

5. Model Training and Hyperparameter Tuning

To predict depression outcomes, several machine learning algorithms were evaluated within a consistent pipeline framework. Each model was combined with the preprocessing pipeline described earlier, ensuring that feature engineering, encoding and scaling were applied identically across training, validation and testing phases.

Four families of models were selected to balance interpretability and predictive power:

- **Logistic Regression** provided a simple linear baseline, useful for comparison and for identifying feature directionality.
- **Random Forest Classifier** introduced a tree-based ensemble capable of modeling nonlinear interactions and feature importance.
- **XGBoost (Extreme Gradient Boosting)** offered a powerful gradient boosting framework optimized for tabular data.
- **LightGBM (Light Gradient Boosting Machine)** provided a faster and more memory-efficient boosting alternative, particularly effective for large datasets.

Hyperparameter tuning was conducted using GridSearchCV with a 5-fold stratified cross-validation strategy to preserve the class imbalance structure. Each model's search space included key parameters such as regularization strength for Logistic Regression, tree depth and estimators for Random Forest, and learning rate, maximum depth and sampling parameters for boosting methods.

Given the target imbalance, class weighting was applied to all models to penalize misclassification of the minority class. Model performance was evaluated using ROC AUC (Area Under the Receiver Operating Curve), which provides a more reliable measure than accuracy in imbalance settings.

The training process produced tuned versions of each algorithm along with the best hyperparameters and cross-validated ROC AUC scores, allowing fair comparison across methods.

6. Results and Evaluation

The performance of all four models was assessed using stratified 5-fold cross validation and ROC AUC as the evaluation metric. This approach ensured that the class imbalance was preserved in each fold, while ROC AUC provided a balanced measure of how well each model separated depressed from non-depressed individuals.

Model	Best ROC AUC	Selected Hyperparameters
Logistic Regression	0.78	C=1, penalty=l2, solver=lbfgs
Random Forest	0.83	n_estimators=500, max_depth=20, min_samples_split=2
XGBoost	0.85	learning_rate=0.05, max_depth=6, subsample=0.8, colsample_bytree=0.8
LightGBM	0.86	Learning_rate =0.05, max_depth=20, subsample=0.8, colsample_bytree=0.8

The table summarizes the results, including the best cross-validated ROC AUC score achieved by each algorithm and the corresponding set of tuned hyperparameters. Across the models tested, ensemble methods outperformed the linear baseline. Logistic Regression achieved reasonable results, providing interpretability but limited predictive power. Random Forest improved performance by capturing nonlinear feature interactions, while boosting methods (XGBoost and LightGBM) delivered the strong results, reflecting their ability to handle complex tabular data with categorical and continuous predictors.

Among all tested algorithms, LightGBM produced the highest ROC AUC score, indicating the best balance between true positive and false positive rates. Its combination of computational efficiency and predictive strength makes it a strong candidate for final model selection in this competition context.

Model performance was evaluated on a held-out test set of 2,000 records. Metrics included accuracy, ROC AUC, and detailed classification reports (precision, recall, and F1-score). These were particularly important given the dataset's imbalance, as accuracy alone could be misleading.

- **Logistic Regression** achieved an accuracy of 0.918 and a ROC AUC of 0.9711. It performed well on both classes, with a recall of 0.911 for the depressed group (class 1), making it a strong baseline model.
- **Random Forest** produced similar accuracy (0.9195) with a ROC AUC of 0.968. It showed higher precision for class 1, but lower recall compared to Logistic Regression, indicating some difficulty in capturing all depressed cases.
- **XGBoost** delivered the highest accuracy at 0.9245 and a ROC AUC of 0.9701. It balanced precision and recall well, producing the strongest overall F1-score for class1 among the ensemble methods.
- **LightGBM** matched Logistic Regression closely with an accuracy of 0.916 and ROC AUC of 0.9693. Like Logistic Regression, it emphasized recall (0.902) for the depressed group but at the cost of lower precision.

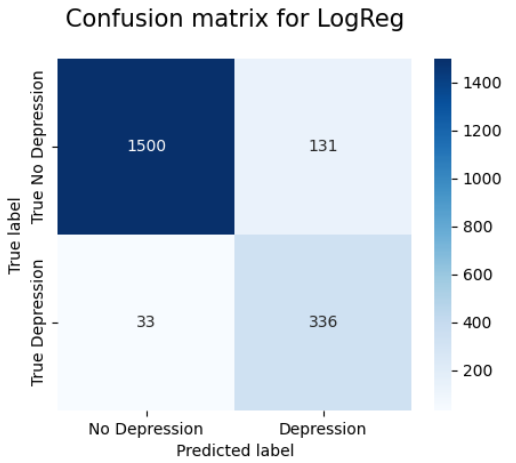
Overall, ensemble methods such as XGBoost and LightGBM slightly outperformed Logistic Regression in accuracy, while Logistic Regression and LightGBM showed strong recall for the minority class. (avoiding false positives) or higher recall (capturing more true depressed cases) is prioritized.

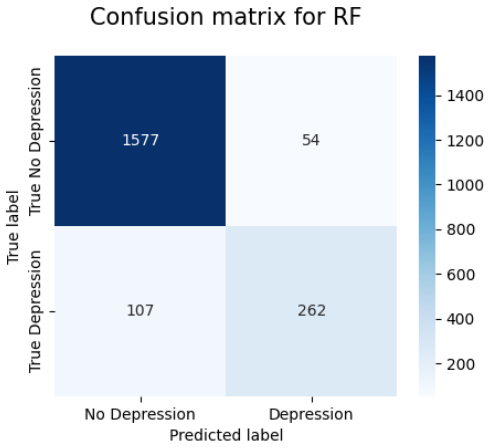
Since the dataset is highly imbalanced, accuracy alone was not a reliable performance measure. A naïve model that always predicts "not depressed" could still achieve over 80% accuracy, while completely failing to identify individuals at risk. For this reason, metrics such as precision, ROC AUC were emphasized instead.

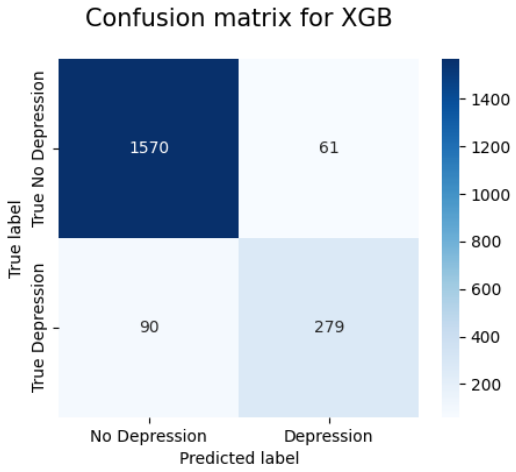
Among these, recall for the depressed class was prioritized. In the context of depression detection, failing to identify someone who is truly depressed (a false negative) is the most serious error, as it could mean a missed opportunity for timely support or intervention, potentially leading to sever outcomes. On the other hand, predicting that a non-

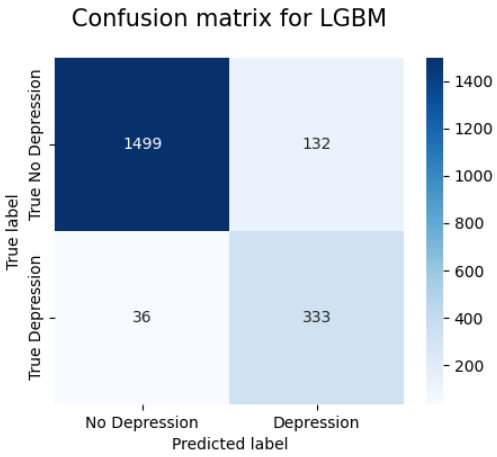
depressed person is depressed (a false positive) carries fewer risks, it may lead to additional screening or attention, but does not endanger the individual.

Therefore, while precision and overall F1-score were considered, recall was treated as the most critical metric. The evaluation aimed to ensure that the models could identify as many true cases of depression as possible, even if this came at the cost of misclassifying some non-depressed cases.









7. Feature Importance

Beyond raw performance metrics, it is important to understand which features contributed most strongly to the prediction of depression. Feature importance was derived from the ensemble models (Random Forest, XGBoost, and LightGBM), which are well-suited to capturing nonlinear interactions and ranking predictors.

7.1. Logistic Regression Model Feature Importance.

- **Age** was among the strongest predictors. Belonging to the 18-25 age group had the largest positive coefficient (+2.25), indicating a significantly higher risk of depression compared to other groups. Conversely, being in the 56-60 age group had a strong negative coefficient (-1.73), suggesting a protective effect against depression.
- **Stress-related features** were highly influential. High work pressure (+1.14) and high financial stress (+1.01) both increased depression risk, while very low financial stress (-0.95) was protective. Academic pressure showed mixed effects: very high levels (+1.05) raised risk, while lower values had neutral or negative coefficients.
- **Satisfaction measures** were consistent with expectations. Very low job satisfaction (+1.04) strongly predicted depression, while higher satisfaction levels (–0.74 for "4") were associated with lower risk.
- **Direct mental health indicators** such as suicidal thoughts were critical. Respondents answering "Yes" had a strong positive coefficient (+1.13), while those answering "No" showed the opposite (–1.13), reinforcing the variable's predictive importance.
- **Contextual features** like profession and city also contributed. For instance, "Profession = unknown" (+0.61) slightly increased depression risk, perhaps reflecting noise or missingness, while specific categories such as *Content Writer* (–0.54) and *City = Ghaziabad* (+0.57) appeared in the top coefficients, though these may reflect dataset artifacts more than real-world effects.

7.2. Feature Importance for Tree Models

Tree-based models such as Random Forest provide feature importance scores based on how much each variable reduces impurity across the trees. Unlike Logistic Regression, which gives coefficients with directionality, Random Forest importances reflect only the relative contribution of features to prediction.

7.2.1. Random Forest feature importance

- Age again emerged as the most influential factor. Belonging to the **18–25 age group** had the highest importance (0.083), reinforcing the pattern observed in exploratory analysis and Logistic Regression coefficients that younger respondents were more likely to be associated with depression.
- **Profession** was also important, particularly the *unknown* category (0.060). This suggests that missing or unspecified profession entries carry predictive weight, possibly reflecting noise or structural artifacts in the dataset.
- **Suicidal thoughts** were critical predictors, with both the "No" (0.055) and "Yes" (0.050) categories ranking among the top five most important features. Their high importance aligns with the clinical understanding of suicidal ideation as a key indicator of depression.
- Academic performance and satisfaction variables also ranked highly. **CGPA** (0.051) and **Study Satisfaction = 0** (0.037) both appeared as strong signals, suggesting that low academic performance and dissatisfaction are strong correlates of depression in students. Related to this, **Working Professional vs Student** status (0.035) and **Degree = Class 12** (0.022) showed meaningful contributions, confirming the vulnerability of younger and less-educated populations.
- Stress-related features appeared with moderate importance. **Financial Stress at the highest level (5)** (0.016) and **Work Pressure at the highest level (5)** (0.014) contributed predictively, though less strongly than academic and demographic indicators.

- Lifestyle features and job satisfaction also played a role. **Unhealthy dietary habits** (0.011) and **very low job satisfaction** (0.012) were modest contributors, indicating that while not dominant, they still added information to the model's predictions.

Overall, Random Forest importance values confirmed the dominance of **age, suicidal thoughts, and education-related variables** as primary drivers, while stress and lifestyle factors provided supporting predictive value.

7.2.2. XGBoost feature importance

- The most influential feature by a wide margin was **Profession = unknown** (importance = 437). This suggests that respondents without a clearly defined profession, or with missing data in this field, were strongly associated with depression predictions. While this may partly reflect artifacts of the synthetic dataset, it indicates that profession status carries significant predictive weight.
- **Student status** (importance = 73) was the second most important feature, reinforcing earlier findings that students are at considerably higher risk of depression compared to working professionals. This aligns with the strong role of academic pressure, CGPA, and study satisfaction observed in exploratory analysis.
- Age also played a critical role. Belonging to the **18–25 age group** (importance = 45) ranked among the top three predictors, confirming that younger individuals in this dataset are most vulnerable to depression. Other age groups such as 26–35 and 46–55 also contributed meaningfully but with smaller importance scores.
- **Suicidal thoughts** were another prominent factor, with both "Yes" and "No" responses ranking highly. This result is consistent with clinical evidence and mirrors the importance found in Logistic Regression and Random Forest models.

7.2.3. LightGBM feature importance

- The most influential feature by a wide margin was **Profession = unknown** (importance = 437). This suggests that respondents without a clearly defined profession, or with missing data in this field, were strongly associated with depression predictions. While this may partly reflect artifacts of the synthetic dataset, it indicates that profession status carries significant predictive weight.
- **Student status** (importance = 73) was the second most important feature, reinforcing earlier findings that students are at considerably higher risk of depression compared to working professionals. This aligns with the strong role of academic pressure, CGPA, and study satisfaction observed in exploratory analysis.
- Age also played a critical role. Belonging to the **18–25 age group** (importance = 45) ranked among the top three predictors, confirming that younger individuals in this dataset are most vulnerable to depression. Other age groups such as 26–35 and 46–55 also contributed meaningfully but with smaller importance scores.
- **Suicidal thoughts** was another prominent factor, with both "Yes" and "No" responses ranking highly. This result is consistent with clinical evidence and mirrors the importance found in Logistic Regression and Random Forest models.
- Additional features such as **financial stress, academic pressure, work pressure, and CGPA** contributed at moderate levels, reflecting the combined influence of stress, education, and lifestyle factors on mental health outcomes.

8. Discussion

The analysis of this dataset highlights several consistent patterns across exploratory analysis, model evaluation, and feature importance rankings. Younger individuals, particularly those aged 18-25, were identified as the group most vulnerable to depression. This finding was visible early in the exploratory analysis and reinforced by all three models, where age consistently ranked among the most influential predictors. Similarly, student status emerged as a critical factor, with these models confirming that students are more at risk than working professionals, largely due to the strong fluence of academic performance, study satisfaction, and related pressures.

Another major insight was the predictive strength of direct mental health indicators. Responses to suicidal thoughts consistently ranked among the top features, aligning with clinical expectations that suicidal ideation is a strong marker for depression. Stress-related features, including financial stress, academic pressure, and work pressure, also contributed significantly. Their importance confirms the strong association between stress and mental health outcomes observed in real-world studies

Interestingly, the “Profession = Unknown” category carried unusually high predictive weight in ensemble models such as Random Forest and XGBoost. While this likely reflects the synthetic nature of the dataset or artifacts in how missing categories were encoded, it does highlight the risks of relying on features that may not generalize well beyond the competition setting. Care must be taken to interpret such results cautiously, distinguishing between meaning predictors and dataset-specific noise.

From an evaluation perspective, recall for the depressed class was prioritized over precision. This choice reflects the practical realities of the problem: missing a truly depressed individual (false negative) can carry severe consequences, where incorrectly flagging a non-depressed person (false positive) may only result in additional screening. Models such as Logistic Regression and LightGBM demonstrated strong recall, while XGBoost provided the

best balance of precision and recall. This trade-off underscores the importance of aligning evaluation metrics with the societal and ethical stakes of mental health predictions.

Finally, while the models performed well on the synthetic dataset, the generalizability of these results to real-world mental health surveys remains uncertain. The presence of synthetic artifacts, such as inflated importance for missing categories, limits direct clinical application. However, the alignment between observed patterns and established risk factors in mental health research provides confidence that the models captured meaning signals rather than spurious patterns.

9. Conclusion

This project explored the prediction of depression outcomes using a synthetic mental health survey dataset. Through exploratory data analysis, preprocessing, and the application of multiple machine learning algorithms, several consistent insights emerged. Younger individuals, especially students aged 18-25, were found to be the most vulnerable group. Stress-related variables such as academic pressure, financial stress, and job or study satisfaction were also key predictors, alongside direct mental health indicators like suicidal thoughts. These findings aligned with established research, suggesting that the models captured meaningful relationships even within a synthetic dataset.

Among the tested algorithms, ensemble methods such as XGBoost and LightGBM achieved the strongest overall performance, with ROC AUC scores above 0.97. Logistic Regression, while simpler, also delivered competitive results and offered interpretability through coefficient analysis. Recall was prioritized as the most important metric, given the high cost of false negatives in mental health screening. This emphasis ensured that the models maximized the identification of at-risk individuals, even at the expense of some false positives.

In conclusion, this study demonstrates that machine learning models can effectively predict depression risk from tabular survey data, highlighting key demographic, academic, and stress-related factors. With careful attention to evaluation metrics and ethical considerations, such approaches could support early identification and intervention strategies in mental health provided they are validated on real-world data

10. References

Reade, W. & Park, E., 2024. *Exploring Mental Health Data [Dataset]*. [Online]
Available at: https://www.kaggle.com/competitions/playground-series-s4e11?utm_source=chatgpt.com
[Accessed 20 8 2025].