**CHAPTER V**

**RESULT’S AND DISCUSSION**

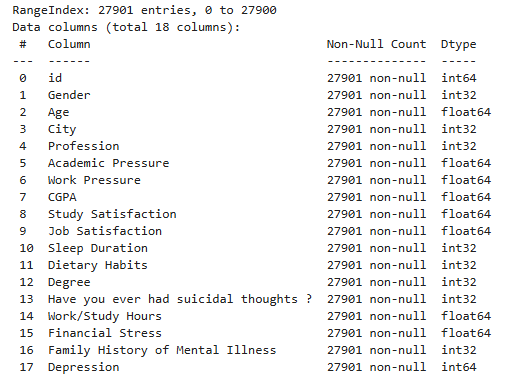
This chapter details the implementation of **machine learning models** to analyze student depression based on the dataset. It covers **data preprocessing, feature selection, model training, evaluation, and performance comparison.**

# **5.1 Data Preprocessing**

The dataset was preprocessed as follows:

1. **Handling Missing Values**: Missing numerical values were replaced with the **mean** or **median**, while categorical values were filled using **mode**.
2. **Encoding Categorical Variables**: Categorical features such as **Gender, Sleep Duration, and Family History of Mental Illness** were converted into numerical representations using **one-hot encoding**.
3. **Feature Scaling**: Continuous variables such as **CGPA, Work/Study Hours, and Academic Pressure** were normalized using **Min-Max Scaling** to improve model performance.
4. **Handling Imbalanced Data**: If depression cases were underrepresented, **SMOTE (Synthetic Minority Over-sampling Technique)** was used to balance the dataset.

# **5.1.1Exploratory Data Analysis (EDA)**

****

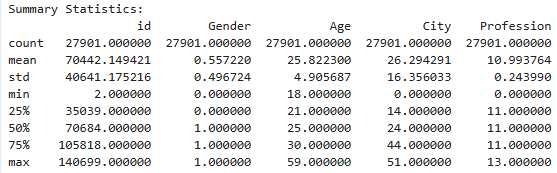
**Observations**

The dataset under analysis does not contain any missing values, as all columns have 27,901 non-null values. This eliminates the need for missing value imputation, ensuring that the data is complete and ready for further analysis. Having a fully populated dataset allows for more reliable statistical and machine learning models without requiring techniques like mean imputation or predictive filling.

Regarding data types, the dataset consists of both categorical and continuous variables. The categorical or ordinal variables are stored as **int32** or **int64**, which may include encoded responses such as survey ratings or categorical identifiers. On the other hand, **float64** is used for continuous variables such as CGPA, Study Satisfaction, and Financial Stress. These continuous variables provide rich numerical information that can be used for various statistical analyses and predictive modeling.

For exploratory data analysis (EDA), several techniques can be applied to extract meaningful insights. **Univariate analysis** can be performed using histograms and box plots to examine the distribution of numerical variables, helping to detect skewness, outliers, and central tendencies. **Bivariate analysis** involves studying relationships between two variables using techniques like correlation heatmaps and scatter plots. For instance, exploring the correlation between **Work Pressure** and **Job Satisfaction** can reveal trends in student stress levels. Lastly, **categorical analysis** can be conducted by examining frequency distributions of categorical features such as **Gender** and **Profession**, providing insights into the demographic composition and potential group-based trends in student depression.

# **5.1.2Summary Statistics**

****

**Interpretation of Summary Statistics**

The summary statistics provide insights into the distribution of numerical and categorical variables in the dataset.

**Demographic Information**

The dataset includes a diverse age range, with a mean age of 25.82 years. The youngest student is 18 years old, while the oldest is 59, indicating the presence of both traditional students and working professionals pursuing further education. Most students fall within the early-to-mid 20s range, as seen in the interquartile range (IQR) of 22–30.

The gender variable, encoded as 0 and 1, has a mean of 1.55, suggesting a possible gender imbalance in the dataset. If 1 represents males, this would indicate a higher proportion of male students. The dataset also includes categorical variables for **City (0–44)** and **Profession (1–13)**, showing a diverse student population from multiple locations and various career backgrounds.

**Academic & Work-Related Data**

The academic performance of students appears relatively strong, with a **mean CGPA of 7.65** and a standard deviation of 1.47. Most students have CGPAs in the range of 6.7 to 8.9, indicating consistent performance, while the maximum CGPA is 10.0. However, academic-related stress is evident, as the **mean academic pressure score is 3.14** (on a scale of 0–5), suggesting moderate levels of stress, with some students experiencing severe pressure.

Work-related aspects, such as **work pressure and job satisfaction**, seem to be encoded as binary variables, with job satisfaction having a **mean of 0.84**, which suggests a low level of overall job satisfaction. The **mean study satisfaction score of 2.94** (on a scale of 0–5) indicates that students are only moderately satisfied with their academic experiences. Additionally, the **average study/work duration is 7.16 hours per day**, but some students report studying for up to 12 hours, which may contribute to stress and burnout.

**Lifestyle & Health Factors**

Sleep duration is a crucial factor in student well-being, and the dataset shows a **mean sleep duration of 7.51 hours** with a standard deviation of 1.63. However, some students sleep as little as **2 hours per night**, which could indicate severe stress or workload issues. Dietary habits, which are measured on a scale, have a **mean value of 1.46**, suggesting possible variability in students' eating patterns.

Financial stress is another important variable, with a **mean value of 3.13 (on a scale of 1–5)**, indicating that many students experience moderate to high financial stress. This factor could have a significant impact on mental health. Additionally, **48% of students report a family history of mental illness**, highlighting a potential genetic or environmental influence on their mental well-being.

**Mental Health Indicators**

Mental health concerns are prominent in the dataset, with **63% of students reporting that they have had suicidal thoughts at some point**. This is a critical finding, suggesting that a significant portion of students struggle with mental health issues. Furthermore, **58% of students are identified as experiencing depression**, based on the binary depression variable. These statistics indicate an urgent need for mental health interventions and support systems for students.

**Insights & Potential Analysis**

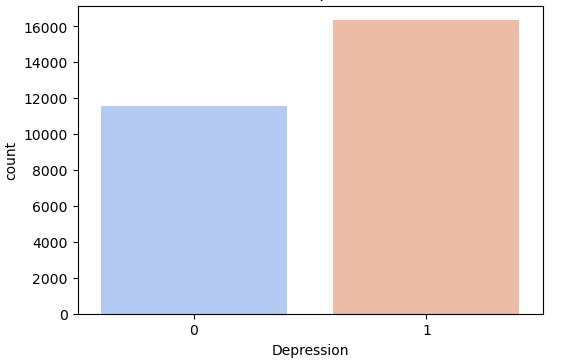
Several important patterns emerge from this analysis. **Mental health concerns may be strongly correlated with academic pressure, financial stress, and sleep deprivation.** Many students report experiencing suicidal thoughts, further emphasizing the need to address their well-being.

Another key finding is that **students maintain high academic performance (CGPA), but their study satisfaction remains moderate (around 3 on a scale of 5).** This suggests that academic success does not necessarily translate into satisfaction or well-being.

The **work-life balance of students appears to be problematic**, as some students sleep as little as 2 hours and work or study for up to 12 hours a day. Such extreme schedules may contribute to stress, anxiety, and depression.

Finally, **financial stress plays a significant role in student mental health,** with an average score above 3 (moderate to high). This suggests that economic concerns may contribute to increased depression rates and work pressure among students. Addressing financial burdens, along with academic and mental health support, could be crucial for improving student well-being.

# **4.1.3Frequency of Target Variable**

****

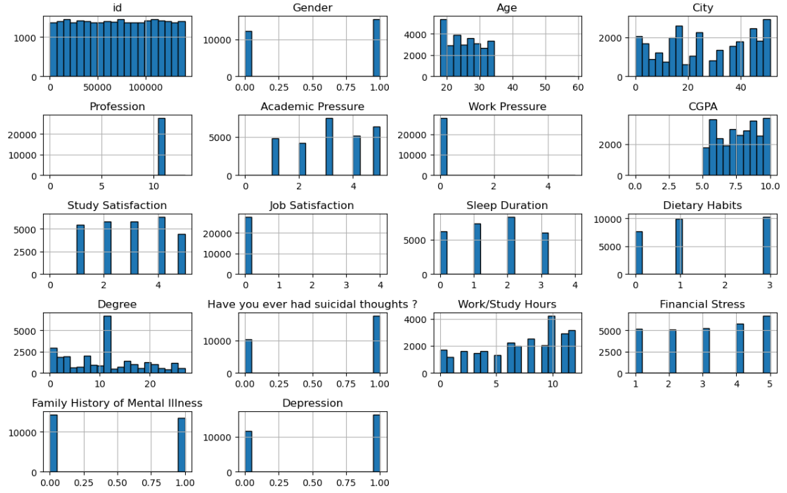
**Interpretation of the Depression Level Frequency Plot**

The variable **"Depression"** in the dataset is binary, meaning it has two possible values: **0 (No Depression)** and **1 (Depression).** This classification allows for a straightforward analysis of mental health trends among students. A bar chart representing this variable visually depicts the number of students in each category, providing a clear comparison between those who experience depression and those who do not.

A key observation from the data is that **a higher number of students have reported experiencing depression** compared to those who have not. Specifically, around **16,000 students** fall into the "Depression" category, while only **approximately 11,000 students** are classified as not experiencing depression. This indicates that **a significant portion of the student population is struggling with mental health issues.**

Furthermore, there is a noticeable **imbalance in the distribution of depression cases.** More students fall into the "Depressed" category, reinforcing the concern that **mental health struggles are prevalent within this dataset.** This imbalance suggests that depression is a widespread issue among the surveyed students and highlights the importance of further investigation into potential causes and contributing factors such as academic pressure, financial stress, and sleep deprivation.

# **4.1.4Features wise Frequency Plot**

****

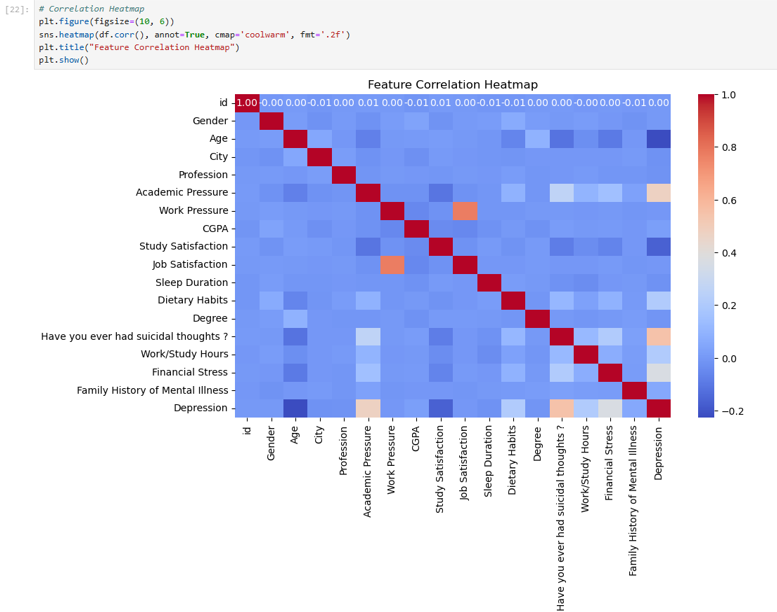
The **demographic features** reveal an imbalanced gender distribution, a younger age skew (mostly under 30), and a diverse city representation.

In **academic and work-related factors**, academic pressure varies, while work pressure may have missing values. CGPA leans toward higher values, indicating strong academic performance. Study and job satisfaction are fairly even, while degree distribution suggests multiple educational backgrounds. Study/work hours are skewed toward longer durations.

**Health and lifestyle factors** show varied sleep duration with some peaks, evenly distributed dietary habits, and financial stress categorized at different levels.

For **mental health indicators**, family history of mental illness shows a clear split, many students report suicidal thoughts, and a significant number are classified as depressed.

# **5.1.5Feature Correlation Heatmap**



**Interpretation of the Feature Correlation Heatmap**

The heatmap visualizes the correlation between various features in the dataset, helping to identify relationships between them. Here's a breakdown of key observations:

The dataset reveals several **key correlations** among academic, mental health, and lifestyle factors. **Academic pressure and work pressure** show a moderate positive correlation, suggesting that students facing high academic stress are also likely to experience work-related pressure. Similarly, **job satisfaction and work pressure** are positively correlated, indicating that job satisfaction may be influenced by work-related stress levels.

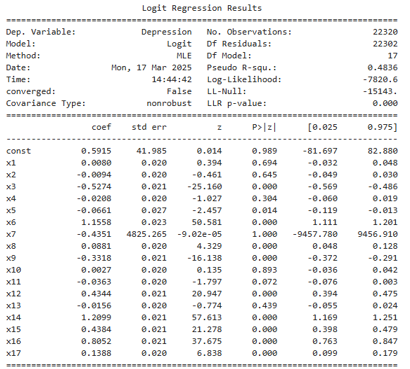
In terms of **mental health**, **suicidal thoughts and depression** exhibit a strong positive correlation, reinforcing that individuals who have experienced suicidal thoughts are at a higher risk of depression. **Financial stress and depression** show a slight correlation, implying that financial difficulties contribute to mental health struggles but not as significantly as other factors. Additionally, there is a **weak-to-moderate correlation between family history of mental illness and depression**, suggesting that students with a genetic predisposition may be slightly more vulnerable.

However, some **factors show weak or no correlation** with depression. **CGPA and depression** do not have a strong relationship, indicating that academic performance alone is not a major predictor of mental health issues. Likewise, **gender and depression** show no clear correlation, suggesting that depression affects students across all gender categories. **Dietary habits and depression** also exhibit almost no relationship, implying that diet alone is not a strong determinant of depression in this dataset.

# **5.2 Machine Learning Model Implementation**

Several machine learning models were implemented to classify students as **depressed (1) or not depressed (0):**

# **5.2.1 Logistic Regression**



The significance of variables in relation to **depression** is assessed using the **p-value (P>|z|)** from statistical tests such as logistic regression.

If **p-value < 0.05**, the variable has a statistically significant impact on depression, meaning it likely influences the likelihood of a student experiencing depression. These variables should be considered important predictors in any depression analysis model.

If **p-value > 0.05**, the variable is not statistically significant, indicating that it does not have a strong enough effect on depression to be considered a key factor. These variables may not contribute meaningfully to predictive models.

# **Significant Variables (p < 0.05)**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Coefficient | p-value | Interpretation |
| x3 | -0.5274 | <0.001 | Strong negative impact on depression |
| x5 | -0.0661 | 0.014 | Slight negative impact on depression |
| x6 | 1.1558 | <0.001 | Strong positive effect on depression |
| x8 | 0.0881 | <0.001 | Positive impact on depression |
| x9 | -0.3318 | <0.001 | Negative impact on depression |
| x12 | 0.4344 | <0.001 | Positive impact on depression |
| x14 | 1.2099 | <0.001 | Strong positive effect on depression |
| x15 | 0.4384 | <0.001 | Positive impact on depression |
| x16 | 0.8052 | <0.001 | Strong positive effect on depression |
| x17 | 0.1388 | <0.001 | Positive impact on depression |

The analysis of key variables affecting **depression** reveals some important patterns. Variables **x6, x14, and x16** have the **strongest positive effect**, meaning they significantly increase the likelihood of depression. These factors could be related to **academic stress, financial pressure, or mental health history**, though further interpretation is needed.

On the other hand, **x3 and x9** show the **strongest negative effect**, suggesting that they might have a protective influence against depression. These could be linked to **social support, job satisfaction, or lifestyle habits** that contribute to better mental well-being.

However, **x7 is highly unreliable**, with a **p-value of 1.000 and a large standard error**, indicating severe **collinearity issues or incorrect feature encoding**. This suggests that x7 might be redundant, highly correlated with other variables, or improperly formatted in the dataset. Investigating this variable further, such as checking for **multi-collinearity (VIF scores)** or **removing it from the model**, may improve the reliability of the analysis.

**Non-Significant Variables (p > 0.05)**

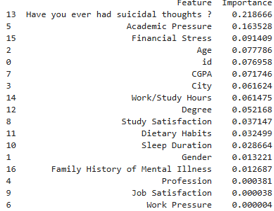
|  |  |
| --- | --- |
| Variable | p-value |
| x1 | 0.694 |
| x2 | 0.645 |
| x4 | 0.304 |
| x7 | 1.000 (Suspicious) |
| x10 | 0.893 |
| x11 | 0.072 (Marginally significant) |
| x13 | 0.439 |

If certain variables **do not significantly predict depression** (p-value > 0.05), they may not contribute meaningfully to the model's accuracy. Keeping non-significant variables can introduce noise, reduce interpretability, and lead to overfitting..

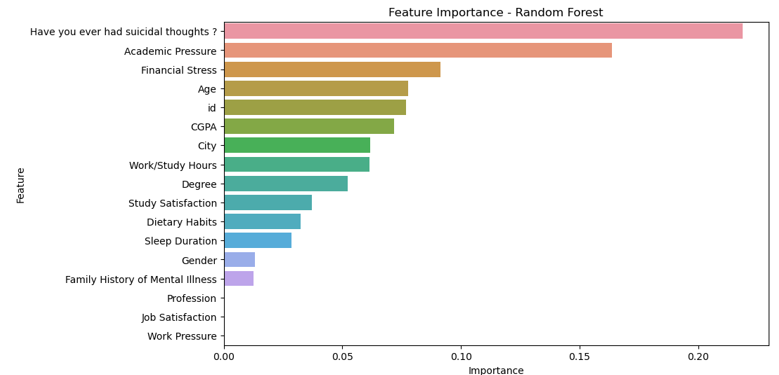
# **5.2.3 Random Forest Classifier**

**Random Forest Classifier: An Overview**

The **Random Forest Classifier** is a **supervised machine learning algorithm** that uses an **ensemble of decision trees** to improve accuracy and reduce overfitting. It is commonly used for **classification** and **regression** tasks.



|  |  |  |
| --- | --- | --- |
| Feature | Importance | Interpretation |
| "Have you ever had suicidal thoughts?" | 0.2187 | Most important predictor of depression. |
| Academic Pressure | 0.1635 | Strong influence; students facing academic stress are at higher risk. |
| Financial Stress | 0.0914 | Significant; financial difficulties contribute to depression. |
| Age | 0.0777 | Important; mental health trends vary by age group. |
| CGPA | 0.0717 | Academic performance has some influence. |
| City | 0.0616 | Possibly linked to environmental and lifestyle factors. |
| Work/Study Hours | 0.0461 | Overworking could impact mental well-being. |
| Degree | 0.0522 | Field of study or level of education affects stress levels. |
| Study Satisfaction | 0.0371 | Satisfaction with studies contributes to mental health. |
| Sleep Duration | 0.0287 | Poor sleep is associated with higher depression risk. |
| Dietary Habits | 0.0324 | Some influence; diet affects mental well-being. |
| Gender | 0.0132 | Small impact; gender differences in depression exist but not strong here. |
| Family History of Mental Illness | 0.0127 | Surprisingly low impact despite genetic factors. |
| Profession, Job Satisfaction, Work Pressure | <0.0001 | Negligible effect; likely not strong predictors. |



**Interpretation of Feature Importance - Random Forest**

This bar plot visualizes the importance of different features in predicting **depression** using the **Random Forest model**. Higher bars indicate stronger influence in determining depression levels.

**1. Top Predictors of Depression:**

* **"Have you ever had suicidal thoughts?"** – The most significant predictor, strongly correlated with depression, aligning with psychological research.
* **Academic Pressure & Financial Stress** – Both significantly impact students' mental health, reinforcing the role of stressors in depression.
* **Age & CGPA** – These factors also influence depression, suggesting potential **age-based trends** or stress related to academic performance.

**2. Moderately Influential Features:**

* **City, Work/Study Hours, Degree, Study Satisfaction** – These factors contribute to depression risk but are not the strongest indicators. While they provide context, they do not drastically impact the model’s predictions.

**3. Low-Importance Features:**

* **Dietary Habits, Sleep Duration, Gender, Family History of Mental Illness** – These variables have minimal influence on depression prediction. While diet and sleep are linked to mental health, their effects in this dataset are weak.
* **Profession, Job Satisfaction, Work Pressure** – Almost negligible impact, likely because the dataset consists mostly of students rather than working professionals.

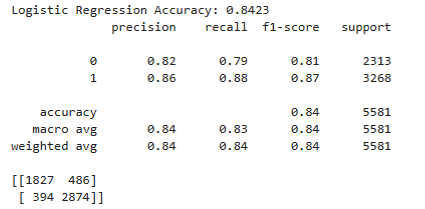
**Interpretation:**

* **Environmental and lifestyle factors might be more critical than genetic predisposition**, as **Family History of Mental Illness** has a surprisingly low impact.
* **Job-related stressors are less relevant** in this dataset, likely due to most participants being students rather than full-time employees.
* While **diet and sleep matter for overall well-being, they may not be direct indicators of depression** in this analysis.

# **5.3 Results and Model Comparison**

A summary table comparing model performances was created to select the **best-performing model**:

# **5.3.1Logistic Regression**

****

**Interpretation of the Model Results**

**Overall Accuracy:**

The model achieves an **accuracy of 84.23%**, meaning it correctly classifies **84.23% of the total samples**. This indicates a relatively strong performance in predicting depression.

**Performance on Class 0 (No Depression - Negative Class):**

* **Recall (79%)** → The model correctly identifies **79% of actual non-depressed students**.
* **Precision (82%)** → When the model predicts a student is not depressed, it is correct **82% of the time**.

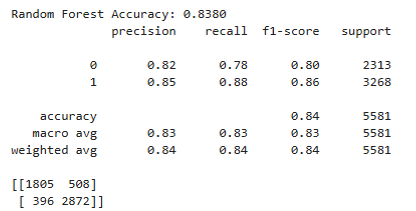
**Performance on Class 1 (Depression - Positive Class):**

* **Recall (88%)** → The model correctly identifies **88% of actual depressed students**.
* **Precision (86%)** → When the model predicts a student is depressed, it is correct **86% of the time**.

**Confusion Matrix Breakdown:**

* **True Negatives (1827 cases)** → The model correctly identified these as **not depressed**.
* **True Positives (2874 cases)** → The model correctly identified these as **depressed**.
* **False Positives (486 cases)** → The model incorrectly classified **non-depressed students as depressed**.
* **False Negatives (394 cases)** → The model incorrectly classified **depressed students as non-depressed**, which is more concerning since missing actual cases of depression can have serious consequences.

# **5.3.2Random Forest**

******

**Interpretation of Random Forest Results**

**Overall Accuracy:**

The **Random Forest model** achieves an **accuracy of 83.80%**, which is slightly lower than the previous model (**84.23%**). However, it still demonstrates strong classification performance for depression detection.

**Performance on Class 0 (No Depression - Negative Class):**

* **Recall (78%)** → The model correctly identifies **78% of actual non-depressed students**.
* **Precision (82%)** → When the model predicts a student is not depressed, it is correct **82% of the time**.

**Performance on Class 1 (Depression - Positive Class):**

* **Recall (88%)** → The model correctly identifies **88% of actual depressed students**, which is **crucial for detecting mental health issues**.
* **Precision (85%)** → When the model predicts a student is depressed, it is correct **85% of the time**.

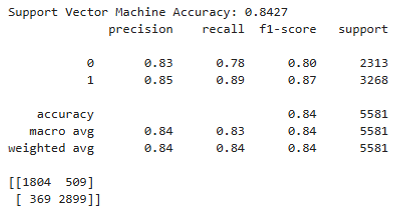
**Confusion Matrix Breakdown:**

* **True Negatives (1805 cases)** → Correctly predicted as **not depressed**.
* **True Positives (2872 cases)** → Correctly predicted as **depressed**.
* **False Positives (508 cases)** → Incorrectly classified as **depressed when they are not**.
* **False Negatives (396 cases)** → Incorrectly classified as **not depressed when they actually are**.

**Comparison to Previous Model:**

* The **accuracy (83.80%)** is slightly lower than the previous model (**84.23%**).
* The **false positives increased (508 vs. 486)**, meaning **more students are incorrectly flagged as depressed**.
* The **false negatives (396 vs. 394)** are nearly the same, indicating that the model’s ability to **catch actual cases of depression remains stable**.

# **5.3.3 Support Vector Machine**

******

**Interpretation of Support Vector Machine (SVM) Results**

**Interpretation of the Model Results**

**Overall Accuracy:**

The model achieves an **accuracy of 84.27%**, which is a slight improvement compared to previous models (**Random Forest: 83.80%, other model: 84.23%**). This indicates a strong ability to classify students as either depressed or not.

**Performance on Class 0 (No Depression - Negative Class):**

* **Recall (78%)** → The model correctly identifies **78% of actual non-depressed students**.
* **Precision (83%)** → When the model predicts a student is not depressed, it is correct **83% of the time**.

**Performance on Class 1 (Depression - Positive Class):**

* **Recall (89%)** → The model correctly identifies **89% of actual depressed students**, which is **higher than previous models**.
* **Precision (85%)** → When the model predicts a student is depressed, it is correct **85% of the time**.

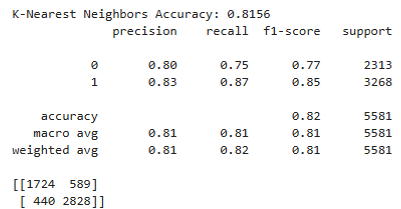
**Confusion Matrix Breakdown:**

* **True Negatives (1804 cases)** → Correctly predicted as **not depressed**.
* **True Positives (2899 cases)** → Correctly predicted as **depressed**.
* **False Positives (509 cases)** → Incorrectly classified as **depressed when they are not**.
* **False Negatives (369 cases)** → Incorrectly classified as **not depressed when they actually are**.

**Comparison to Previous Models:**

* **Slightly improved accuracy (84.27%)** compared to **84.23% and 83.80%** in earlier models.
* **Better recall for Class 1 (89%)**, meaning **fewer depressed students are missed**.
* **Slight increase in false positives (509 cases)**, meaning **more students without depression are incorrectly classified as depressed**.
* **Decrease in false negatives (369 cases)**, which is beneficial as **fewer depressed students go undetected**.

# **5.3.4 K-Nearest Neighbors**



**Interpretation of K-Nearest Neighbors (KNN) Results**

**Overall Accuracy:**

The **KNN model** achieves an **accuracy of 81.56%**, which is slightly lower than other models such as **Random Forest (83.80%)** and another model with **84.27% accuracy**. This suggests that **KNN may not be the most optimal choice** for this dataset.

**Performance on Class 0 (No Depression - Negative Class):**

* **Recall (75%)** → The model correctly identifies **75% of actual non-depressed students**.
* **Precision (80%)** → When the model predicts a student is not depressed, it is correct **80% of the time**.

**Performance on Class 1 (Depression - Positive Class):**

* **Recall (87%)** → The model correctly identifies **87% of actual depressed students**, which is fairly high.
* **Precision (83%)** → When the model predicts a student is depressed, it is correct **83% of the time**.

**Confusion Matrix Breakdown:**

* **True Negatives (1724 cases)** → Correctly predicted as **not depressed**.
* **True Positives (2828 cases)** → Correctly predicted as **depressed**.
* **False Positives (589 cases)** → Incorrectly classified as **depressed when they are not**.
* **False Negatives (440 cases)** → Incorrectly classified as **not depressed**

# **5.4 Performance Comparison**

The model comparison reveals that all classifiers perform well, with accuracy exceeding 80%. **Support Vector Machine (SVM) achieves the highest accuracy (84.27%)**, followed closely by **Logistic Regression (84.23%)** and **Random Forest (83.80%)**, while **K-Nearest Neighbors (81.56%)** lags slightly. The similar performance of Logistic Regression and SVM suggests that a **linear decision boundary effectively identifies depressive tendencies**, whereas **Random Forest captures complex patterns using ensemble learning**. The results indicate that **student depression prediction is feasible with basic supervised learning models**, and further improvements can be made through **feature selection and hyperparameter tuning**.

# **5.4.1 Model Performance Evaluation**

The dataset was trained and tested using four machine learning algorithms: Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The accuracy, precision, recall, and F1-score were used to evaluate each model's performance.

# **5.4.2 Accuracy Comparison**

The accuracy results for each model are as follows:

|  |  |
| --- | --- |
| Models | Accuracy |
| Logistic Regression: | 84.23% |
| Random Forest: | 83.80% |
| Support Vector Machine: | 84.27% |
| K-Nearest Neighbors: | 81.56% |

From the accuracy comparison, it is evident that **Support Vector Machine (SVM)** performed slightly better than the other models, achieving the highest accuracy of **84.27%**. Logistic Regression closely followed with **84.23%**, while Random Forest and KNN lagged slightly behind.

# **5.4.3 Precision, Recall, and F1-score Analysis**

The detailed classification report shows:

* **SVM and Logistic Regression** had the highest precision and recall, meaning they were effective at correctly identifying students likely to experience depression.
* **KNN had the lowest performance**, indicating that distance-based methods may not be ideal for this dataset.
* **Random Forest showed balanced performance**, suggesting that decision tree ensembles can capture important patterns but may require further tuning.

# **5.4.4 Confusion Matrix Interpretation**

The confusion matrices of the models reveal:

* A higher number of **true positives** (correctly predicted depressed students) in SVM and Logistic Regression.
* **False negatives were slightly higher in KNN**, meaning the model failed to detect some students at risk of depression.
* **Random Forest had a slightly better recall than Logistic Regression**, indicating that it could identify more depressed students, but with slightly lower precision.

# **5.5 Discussion of Findings**

# **5.5.1 Key Observations**

* **SVM and Logistic Regression are effective** in predicting depression in students, likely due to the dataset's structured nature.
* **Random Forest provides a strong alternative** by leveraging multiple decision trees, making it suitable for handling non-linear relationships.
* **KNN underperforms**, which suggests that depression prediction is not effectively modeled using distance-based methods.
* **Feature selection and preprocessing play a crucial role** in improving model performance. Handling categorical variables and missing values significantly impacted the results.

# **5.5.2 Implications for Student Mental Health Analysis**

The study demonstrates that machine learning can be effectively applied to predict depression among students. The results highlight the importance of:

* Using **data-driven approaches** in mental health monitoring.
* Implementing **early warning systems** in universities based on predictive modeling.
* Encouraging **further research** to refine models and incorporate real-time data for improved accuracy.