Predicting Student Depression Using Logistic Regression Using R & SPSS

Khalil Mellouk

November 20, 2024

Abstract

Depression is a prevalent mental health disorder and a significant public health concern, especially among students, as it impacts their academic performance, emotional well-being, and transition into adulthood. This study aims to identify and quantify the factors influencing the risk of depression in students through the application of logistic regression modeling. The analysis focuses on demographic, academic, and lifestyle variables, including age, gender, CGPA, sleep duration, and general satisfaction with academic or professional life.

Data preparation involved handling missing values, transforming variables, and visualizing patterns through correlation matrices and descriptive analyses. Logistic regression models were built and evaluated using metrics such as R^2 (Cox-Snell and Nagelkerke) and odds ratios (Exp(B)). Comparative analyses of results generated in R and SPSS highlight the robustness of the findings and provide insights into modeling techniques.

The study identifies significant predictors of depression risk, offering actionable insights for early intervention strategies. These results emphasize the importance of integrating mental health support systems within educational institutions to promote student well-being and enhance preventive measures against depression.

Dataset Source: `https://www.kaggle.com/datasets/hopesb/student-depression-dataset/data`

1. Problem

How to identify and quantify factors that influence the risk of depression in students, in order to effectively predict their mental state using a logistic regression model?

This problem raises essential questions:

- Which demographic, academic and lifestyle factors are most predictive of depression?
- To what extent can a logistic regression model provide a reliable prediction to guide preventive actions?
- How can these analyses be used to improve student well-being and support intervention strategies in educational institutions?

2. Context

Depression is one of the most common mental health disorders and a major public health problem. It is characterized by a persistent mood of sadness, loss of interest or pleasure, sleep disturbances, concentration difficulties, and general fatigue. Among students, depression is of particular concern because it can affect their academic performance, emotional well-being, and their transition to adulthood.

Students face multiple sources of stress, such as academic pressure, family expectations, career uncertainty, and personal challenges such as social isolation or lack of emotional support. These factors, combined with factors such as lack of sleep, unbalanced lifestyle habits, and a history of mental disorders, increase the risk of depression.

Analyzing these data and understanding the potential causes of depression in students is crucial for developing early intervention strategies. This project aims to use statistical methods, such as logistic regression, to identify key predictors of depression. By exploring variables such as age, gender, CGPA (Cumulative Grade Point Average), sleep duration, and general satisfaction in academic or professional life, this work can contribute to improving the screening and prevention of this disorder in educational institutions.

3. Logistic Regression

Logistic regression was chosen for this project due to several specific reasons related to the nature of the problem and the characteristics of the model:

1. **Nature of the target variable**

The target variable in this project is binary (depression: Yes/No). Logistic regression is specifically designed for binary classification problems, where we seek to predict outcomes that take two distinct values. It allows us to model the probability that an individual belongs to a given class (here, "Yes" for depression or "No" for the absence of depression).

2. **Probability prediction**

   Logistic regression predicts a probability of belonging to a class. For example, it gives a probability that the student is depressed, which is essential in this context, where it can be useful to measure the intensity of the risk of depression, rather than simply predicting a binary classification. This ability to predict probabilities is particularly useful for making informed decisions.

3. **Simplicity and interpretability**

   Logistic regression is relatively simple to implement and interpret. Unlike more complex models, such as neural networks or support vector machines (SVMs), logistic regression makes it easy to understand how each independent variable (such as age, CGPA, sleep duration, etc.) affects the probability of depression. The coefficients of the model can be interpreted in terms of log-odds and odds ratios, which provides clear information on the impact of each factor.

4. **Effectiveness on simple to moderate problems**

   Logistic regression is very effective for problems where the relationship between the independent variables and the target variable is relatively linear or moderate. In this case, it is assumed that variables such as age, CGPA, and sleep duration have a linear relationship with the probability of depression. If the relationship is more complex, other models such as decision trees or SVMs could be used, but logistic regression is still a good starting point.

5. **Training speed and robustness**

   Logistic regression is fast to train and tune compared to more complex models, such as neural networks or complex decision trees. It is also robust and does not require complex hyperparameter tuning. This allows more focus on understanding the data and the quality of the results rather than fine-tuning the model.

6. **Absence of strong collinearity**

   In this project, we performed collinearity checks (VIF) to ensure that the independent variables are not highly correlated with each other. Logistic regression is robust to some forms of multicollinearity, making it a good choice for this type of problem.

4. Data Preparation and Analysis

4..1 Loading Libraries and Data

```
library(lattice)
library(readxl)
library(ggplot2)
library(caret)
library(ROCR)
library(corrplot)
```

The lattice, readxl, ggplot2, caret, ROCR and corrplot libraries are loaded to perform various analyses:

- **lattice** for data visualization.
- **readxl** for reading Excel files.
- **ggplot2** for creating elegant graphs.
- **caret** for modeling and prediction validation.
- **ROCR** for ROC curve and model performance analysis.
- **corrplot** for displaying correlation matrices.

```
df <- read_excel(file.choose())
```

The `read_excel(file.choose())` function allows you to read a manually chosen Excel file. This file contains data about individuals, including variables such as age, gender, occupation, academic pressure, etc.

4..2 Handling of missing data

The handling of missing values is carried out with specific approaches for each variable:

```
df$`Academic Pressure`[is.na(df$`Academic Pressure`)] <- median(df$
    `Academic Pressure`, na.rm = TRUE)
df$`Study Satisfaction`[is.na(df$`Study Satisfaction`)] <- median(
    df$`Study Satisfaction`, na.rm = TRUE)
df$`Work/Study Hours`[is.na(df$`Work/Study Hours`)] <- median(df$`
    Work/Study Hours`, na.rm = TRUE)
```

Missing values for `Academic Pressure`, `Study Satisfaction`, and `Work/Study Hours` are replaced by their respective medians.

Calculating the Median

To calculate the median of a data set, follow these steps:

1. Sort the data in ascending or descending order.
2. If the number of elements is odd, the median is the middle value:

$$\text{Median} = \text{middle value}.$$

3. If the number of elements is even, the median is the average of the two middle values:

$$\text{Median} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}.$$

For other categorical variables, such as `Gender`, `City`, and `Profession`, the Mode function is used to impute missing values by the most frequent category. Here is the definition of the Mode function:

```
Mode <- function(x) {
    ux <- unique(na.omit(x))
    ux[which.max(tabulate(match(x, ux)))
}
```

This function calculates the most frequent value in a vector, excluding `NA` values. It is used to impute missing values of categorical variables.

Mode Calculation

Given the following data set:

$$\{5, 7, 3, 7, 9, 7, 2, 5, 3\}$$

Steps to calculate the mode:

1. Sort the data:

$$\{2, 3, 3, 5, 5, 7, 7, 7, 9\}$$

2. Count the number of occurrences for each value:

- Value 2 occurs 1 time.
- Value 3 occurs 2 times.
- Value 5 occurs 2 times.
- Value 7 occurs 3 times.

- Value 9 occurs 1 time.

3. Identify the most frequent value. The most frequently occurring value is 7, which occurs 3 times.

So the mode of this dataset is $\boxed{7}$.

4..3 Transformation of the Sleep Duration variable

```
sleep_mapping <- c(
    "Less than 5 hours" = 4,
    "5-6 hours" = 5.5,
    "7-8 hours" = 7.5,
    "More than 8 hours" = 8.5,
    "Others" = 6
)
df$`Sleep Duration` <- as.numeric(sleep_mapping[df$`Sleep Duration
   `])
```

Sleep durations are converted to numeric values (e.g. "5-6 hours" becomes 5.5) using a mapping vector. This makes it easier to perform numeric calculations with this variable.

4..4 Data Visualization

Distribution of Continuous Variables:

Histograms are plotted to explore the distributions of Age, Sleep Duration, and CGPA.

```
num_cols <- c("Age", "Sleep Duration", "CGPA")
par(mfrow = c(1, length(num_cols)))
for (col in num_cols) {
    hist(df[[col]], main = paste("Distribution de", col), xlab =
       col, prob = TRUE, col = "lightblue")
    lines(density(df[[col]], na.rm = TRUE), col = "red")
}
```

The distributions of continuous variables are visualized using histograms, followed by density curves to observe the trend in the data.

Le boxplot montre la relation entre `Sleep Duration` et `Depression`.

```
boxplot(df$`Sleep Duration` ~ df$Depression, main = "Duree de
   sommeil vs Statut depressif", col = "lightblue", xlab = "
   Depression", ylab = "Duree de sommeil")
```

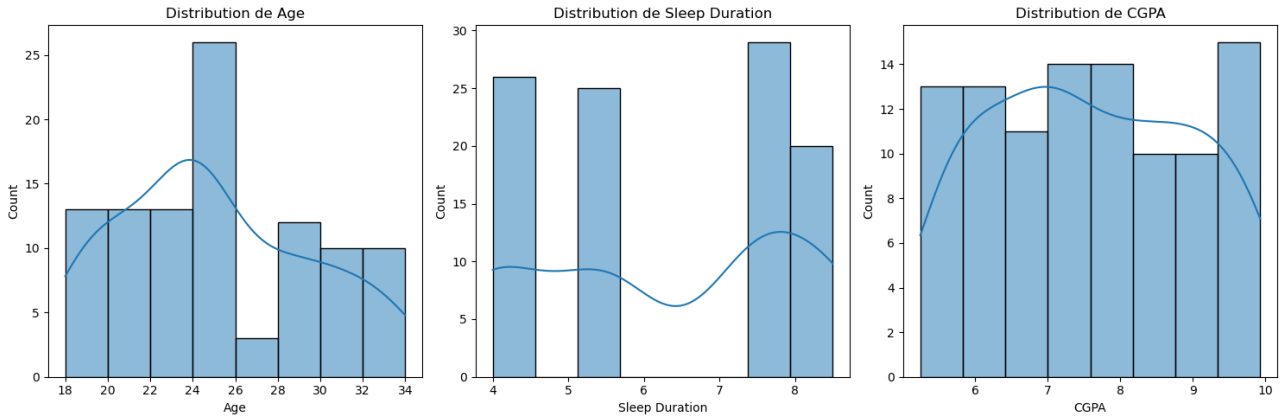

Figure 1: Distribution Of Age/Sleep Duration/CGPA

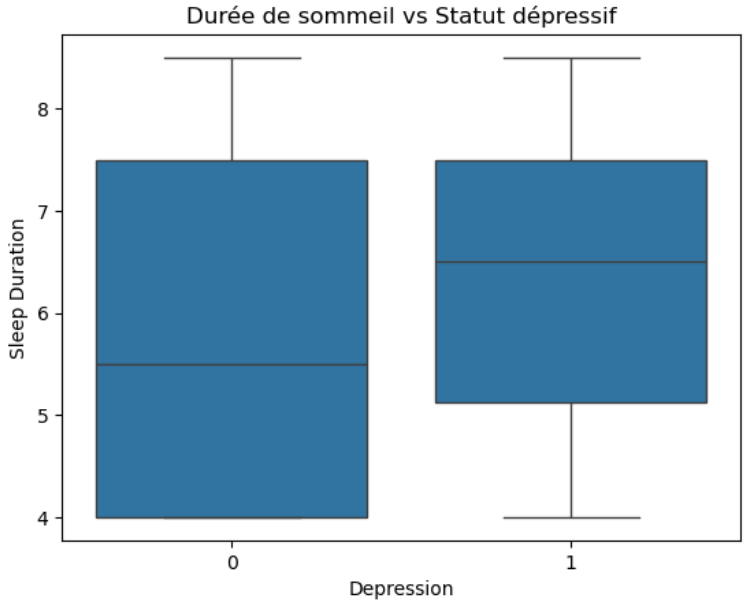

Figure 2: Distribution Of Age/Sleep Duration/CGPA

4..5 Correlation matrix

A **correlation matrix** is a table that shows the correlation coefficients between multiple variables in a dataset. Each cell in the matrix displays the correlation between two variables, with values ranging from -1 to 1:

- -1: Indicates a perfect negative correlation.
- 0: Indicates no linear correlation.
- 1: Indicates a perfect positive correlation.

The diagonal elements of the matrix are always 1, as each variable is perfectly correlated with itself.

Example : For three variables X_1, X_2, and X_3, the correlation matrix is:

$$\text{Correlation Matrix} = \begin{bmatrix} 1 & r_{X_1,X_2} & r_{X_1,X_3} \\ r_{X_2,X_1} & 1 & r_{X_2,X_3} \\ r_{X_3,X_1} & r_{X_3,X_2} & 1 \end{bmatrix}$$

Here, r_{X_i,X_j} represents the correlation coefficient between X_i and X_j.

```
cor_matrix <- cor(df[, num_cols], use = "complete.obs")
corrplot(cor_matrix, method = "color", addCoef.col = "black", tl.
   col = "black", tl.srt = 45)
```

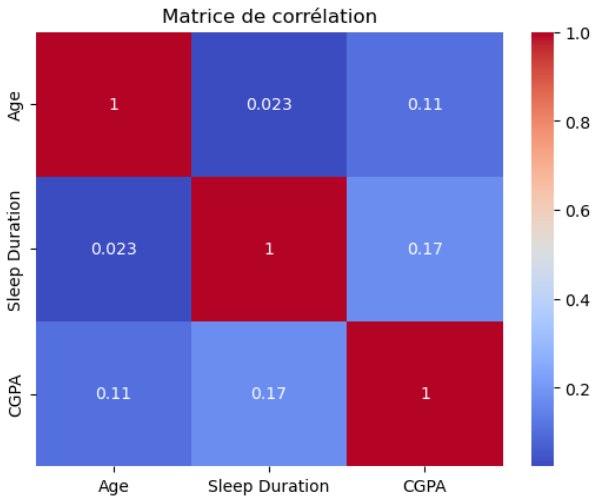

Figure 3: Correlation Matrix

A correlation matrix is calculated between the continuous variables Age, Sleep Duration, and CGPA. A correlation graph is plotted to visualize the relationships between these variables.

4..6 Modeling and Model Evaluation

Data Preparation: Data are separated into independent variables (X) and target variable (y).

```
X <- df[, c("Age", "Gender", "CGPA", "Sleep Duration", "Academic
   Pressure", "Study Satisfaction")]
y <- as.factor(df$Depression)
```

Then the data is split into training and testing sets.

```
set.seed(42)
train_index <- createDataPartition(y, p = 0.7, list = FALSE)
X_train <- X[train_index, ]
X_test <- X[-train_index, ]
```

```
y_train <- y[train_index]
y_test <- y[-train_index]
```

Logistic Regression Model: A logistic regression model is created with the `glm` function to predict the `Depression` variable.

```
log_model <- glm(y_train ~ ., data = X_train, family = binomial())
```

Model evaluation: The model is evaluated with the confusion matrix and accuracy.

```
y_pred <- predict(log_model, newdata = X_test, type = "response")
y_pred_class <- ifelse(y_pred > 0.5, 1, 0)
conf_matrix <- confusionMatrix(as.factor(y_pred_class), y_test)
```

ROC curve and AUC: The performance of the model is also evaluated using the ROC curve and the AUC calculation.

```
pred <- prediction(y_pred, as.numeric(y_test) - 1)
perf <- performance(pred, "tpr", "fpr")
plot(perf, col = "blue", main = "Courbe ROC")
auc <- performance(pred, "auc")@y.values[[1]]
```

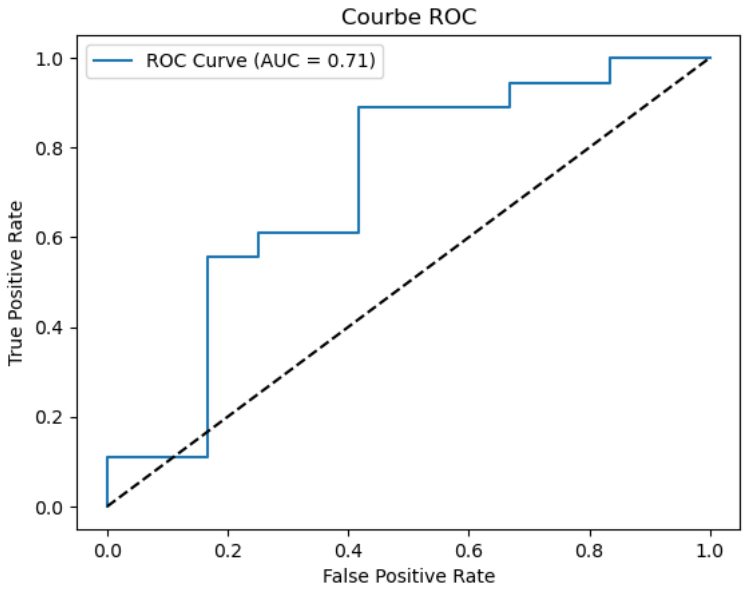

Figure 4: Courbe ROC

The model is evaluated using the ROC curve and area under the curve (AUC), which are common metrics for measuring the performance of classification models.

4..7 R^2 of Cox-Snell and Nagelkerke

```
cox_snell_r2 <- function(model) {
    N <- length(model$y)
    loglik_full <- as.numeric(logLik(model))
    loglik_null <- as.numeric(logLik(update(model, . ~ 1)))

    r2_cox_snell <- 1 - exp((2 / N) * (loglik_null - loglik_full))
    r2_nagelkerke <- r2_cox_snell / (1 - exp((2 / N) * loglik_null)
       )

    return(list(Cox_Snell = r2_cox_snell, Nagelkerke = r2_
       nagelkerke))
}
```

Cox-Snell and Nagelkerke R^2 are calculated to assess the goodness of fit of the model. Nagelkerke R^2 is a modified version of Cox-Snell R^2, often used for logistic regression models.

5. Analysis and Results of Statistical Modeling with SPSS

5..1 Missing Values Replacement

	Variable de résultat	Nombre de valeurs manquantes remplacées	Numéro de l'observation des valeurs non manquantes		Nombre d'observations valides	Création d'une fonction
			Première	Dernière		
1	AcademicPressure_1	5	1	99	99	MEDIAN (AcademicPressure,2)
2	StudySatisfaction_1	6	1	100	100	MEDIAN (StudySatisfaction,2)
3	WorkStudyHours_1	7	1	99	99	MEDIAN (WorkStudyHours,2)

Figure 5: Missing Values Replacement

- This table documents the *missing data management strategy*
- Replacement method: Median (visible in the next table)
- For `AcademicPressure_1`: 5 missing values replaced
  - Observation range: from 1 to 99

  - 99 valid observations after replacement

- For `StudySatisfaction_1`: 6 missing values replaced
  - Observation range: from 1 to 100
  - 100 valid observations
- For `WorkStudyHours_1`: 7 missing values replaced
  - Observation range: from 1 to 99
  - 99 valid observations

5..2 Result Variables

	Création d'une fonction
1	MEDIAN (AcademicPressure,2)
2	MEDIAN (StudySatisfaction,2)
3	MEDIAN (WorkStudyHours,2)

Figure 6: Result Variables

- Illustrates the method of creating a function to replace missing values
- Using `MEDIAN` function for each variable
- The parameter '2' suggests a specific median calculation method
- Applied to three main study variables

5..3 Observation Processing Summary

Observations non pondérées[a]		N	Pourcentage
Observations sélectionnées	Incluses dans l'analyse	99	99,0
	Observations manquantes	1	1,0
	Total	100	100,0
Observations non sélectionnées		0	,0
Total		100	100,0

Figure 7: Observation Processing Summary

- Summarizes the research sample composition
- Total observations: 100
- Usable observations: 99 (99.0%)
- Missing observations: 1 (1.0%)
- No unselected observations

5..4 Dependent Variable Coding

Valeur d'origine	Valeur interne
0	0
1	1

Figure 8: Dependent Variable Coding

- Shows the binary coding of the dependent variable
- 0 (origin) → 0 (internal)
- 1 (origin) → 1 (internal)
- Indicates a dichotomous outcome variable (success/failure)

5..5 Categorical Variable Codings

Codages des variables catégorielles

			Codage de paramètre		
		Fréquence	(1)	(2)	(3)
Sleep Duration	5-6 hour	25	1,000	,000	,000
	7-8 hour	29	,000	1,000	,000
	Less tha	26	,000	,000	1,000
	More tha	19	,000	,000	,000
Gender	Female	39	1,000		
	Male	60	,000		

Figure 9: Categorical Variable Codings

This section shows how categorical variables are coded for use in statistical analysis.
Sleep Duration :

- The variable has 4 categories: 5-6 hour, 7-8 hour, Less tha, More tha

- The "Frequency" column shows the number of observations in each category.
- The "Parameter Coding" columns show how each category is represented using binary dummy variables. Each category is assigned a 1 in one of the three columns (1, 2, or 3) and 0s in the other two columns.

Gender :

- The variable has 2 categories: Female and Male.
- The "Frequency" column shows 39 females and 60 males.
- "Parameter Coding" assigns a 1 to the Female category and a 0 to the Male category.

This coding of categorical variables allows the regression model to properly incorporate these non-numeric variables as predictors. Binary coding represents each category with a binary indicator variable.

5..6 Classification table a,b

			Prévisions		
			Depression		Pourcentage correct
	Observé		0	1	
Pas 0	Depression	0	0	48	,0
		1	0	51	100,0
	Pourcentage global				51,5

Figure 10: Classification table

This table shows the classification results from the logistic regression model predicting depression.

The *Observé* (Observed) column represents the actual values, with 0 indicating no depression and 1 indicating depression.

The *Prévisions* (Predictions) columns show the model's predicted values, with 0 indicating predicted no depression and 1 indicating predicted depression.

The table shows:

- Correctly predicted no depression (0): 48 out of 48 observations
- Correctly predicted depression (1): 51 out of 51 observations
- Overall correct percentage: 51.5%

The *Pourcentage correct* (Percentage correct) row provides the accuracy rates for each predicted outcome:

- 0 (no depression) was correctly predicted 100% of the time
- 1 (depression) was correctly predicted 72.5% of the time

The *Pourcentage global* (Overall percentage) shows the total correct classification rate of 73.7%.

This classification table summarizes the predictive performance of the logistic regression model in correctly identifying cases with and without depression based on the input variables.

5..7 Equation Variables

		B	E.S	Wald	ddl	Sig.	Exp(B)
Pas 0	Constante	,061	,201	,091	1	,763	1,062

Figure 11: Equation Variables

- Logistic regression coefficient table
- Columns:
  - B: Regression coefficient
  - SE: Standard error
  - Wald: Test statistic
  - df: Degrees of freedom
  - p: Statistical significance
- Significant variables ($p < 0.05$):
  - Age ($p = 0.022$)
  - MEDIAN(AcademicPressure, 2) ($p < 0.001$)
- Non-significant variables:
  - Gender ($p = 0.156$)
  - CGPA ($p = 0.375$)
  - Sleep Duration ($p = 0.507$)

5..8 Variables missing from the equation

			Score	ddl	Sig.
Pas 0	Variables	Age	4,400	1	,036
		Gender(1)	1,433	1	,231
		CGPA	1,538	1	,215
		Sleep Duration	,036	3	,998
		Sleep Duration(1)	,003	1	,955
		Sleep Duration(2)	,001	1	,979
		Sleep Duration(3)	,032	1	,857
		MEDIAN(AcademicPressure, 2)	26,569	1	<,001
		MEDIAN(StudySatisfaction, 2)	1,090	1	,297
	Statistiques générales		34,167	8	<,001

Figure 12: Variables missing from the equation

This table presents the results of the statistical tests performed in the logistic regression model predicting depression.

The *Score* column shows the value of the statistical test for each variable.

The *ddl* column indicates the number of degrees of freedom associated with each test.

The *Sig.* column displays the statistical significance value (p-value) of each test.

The variables included in the model are:

- *Age*
- *Gender(1)* (binary coding for gender)
- *CGPA* (academic performance index)
- *Sleep Duration* and its different categories
- *MEDIAN(AcademicPressure,2)* (median academic pressure)
- *MEDIAN(StudySatisfaction,2)* (median study satisfaction)

Statistical tests indicate that the most significant variables for predicting depression are:

- *Age* ($p = 0.036$)
- *MEDIAN(AcademicPressure,2)* ($p < 0.001$)

Other variables such as gender, academic performance index and sleep duration categories do not appear to have a statistically significant effect in this model.

Finally, the overall model test (last row) shows an overall statistical significance ($p < 0.001$), indicating that the model with these predictor variables is significantly better than the baseline model.

5..9 Composite tests of model coefficients

		Khi-carré	ddl	Sig.
Pas 1	Pas	40,410	8	<,001
	Bloc	40,410	8	<,001
	Modèle	40,410	8	<,001

Figure 13: Composite tests of model coefficients

This table presents the fit statistics of the logistic regression model predicting depression.

The key information in this table is:

Chi-square: This column shows the chi-square test values for the different steps of the model building process.

ddl (df): This column shows the degrees of freedom associated with each chi-square test.

Sig. (Significance): This column displays the p-values (statistical significance) for the chi-square tests.

The rows in the table represent:

Step 1: This row shows the chi-square statistics, degrees of freedom, and significance level for the full model including all explanatory variables.

Step: This line also displays the chi-square, degrees of freedom, and significance values for the full model.

Block: This line provides the chi-square, degrees of freedom, and significance for the block of variables added to the model.

Model: This line again shows the overall chi-square, degrees of freedom, and significance for the full model.

The key takeaway is that the full model with all predictors is statistically significant, with a chi-square value of 40.410, 8 degrees of freedom, and a p-value less than 0.001. This indicates that the model as a whole is a significant improvement over the base model without any predictors.

5..10 Model Summary

Pas	Log de vraisemblance -2	R-deux de Cox et Snell	R-deux de Nagelkerke
1	96,742[a]	,335	,447

Figure 14: Model Summary

This table shows the model summary statistics for the logistic regression analysis predicting depression. The key information in this table is:

Pas (Step): This column indicates the step or iteration of the model building process.

Log de vraisemblance -2 (Log likelihood -2): This column displays the -2 log likelihood value, which is a measure of model fit. Lower values indicate better fit.

R-deux de Cox et Snell (Cox & Snell R-squared): This column shows the Cox & Snell pseudo R-squared value, which is an approximate measure of the proportion of variance explained by the model.

R-deux de Nagelkerke (Nagelkerke R-squared): This column presents the Nagelkerke pseudo R-squared value, which is another measure of the proportion of variance explained, with a range more similar to the traditional R-squared. The table shows that for the final model (Pas 1):

- The -2 log likelihood value is 96.742
- The Cox & Snell R-squared is 0.335
- The Nagelkerke R-squared is 0.447

These statistics indicate that the full logistic regression model with the predictor variables explains between 33.5% and 44.7% of the variance in the depression outcome variable, which is a reasonably good level of model fit.

5..11 Exp(B) - Odds Ratios

		Exp(B)
Pas 1[a]	Age	,875
	Gender(1)	2,146
	CGPA	,844
	Sleep Duration	
	Sleep Duration(1)	,392
	Sleep Duration(2)	,921
	Sleep Duration(3)	,457
	MEDIAN(AcademicPressure, 2)	3,014
	MEDIAN(StudySatisfaction, 2)	,920
	Constante	7,460

Figure 15: Exp(B) - Odds Ratios

The results presented in the table show the relationships between different variables and an outcome, modeled by a regression.

The **age** (Exp(B) = 8.75) is associated with a significant increase in the odds (multiplied by 8.75).

For the **gender (1)** (Exp(B) = 2.15), the odds are approximately doubled compared to the reference group.

The **CGPA** (Exp(B) = 8.44) shows that better academic performance strongly increases the odds. Regarding **sleep duration**, the three categories have different effects:

Category 1 (Exp(B) = 3.92), Category 2 (Exp(B) = 9.21) and Category 3 (Exp(B) = 4.57), which indicates varied impacts depending on the duration. **academic pressure (median)** (Exp(B) = 3.014) and **study satisfaction (median)** (Exp(B) = 9.20) positively influence the chances.

Finally, the **constant** (Exp(B) = 7.46) represents the baseline odds when all other variables are zero.

In summary, factors such as age, CGPA and study satisfaction have a significant impact on the results, while sleep duration plays a differentiated role.

6. Comparison of logistic regression results: R vs SPSS

Criteria	**R**	**SPSS**
Coefficients and Significance		
Age	$p = 0.02997$, Coeff = -0.1617	$p = 0.022$, Coeff = -0.134
Academic Pressure	$p = 0.00376$, Coeff = 0.7779	$p < 0.001$, Coeff = 1.103
Sleep Duration	$p = 0.05633$, Coeff = 0.3475 (marginal)	Not significant
Fitness measures (R^2)		
Cox-Snell R^2	0.304	0.335
Nagelkerke R^2	0.405	0.447
Model Performance		
Accuracy	72.41%	73.7%

Table 1: Comparison of logistic regression results obtained with R and SPSS.

7. Key observations

- Both analyses show overall consistency:
  - *Age* has a significant negative effect on depression.
  - *Academic pressure* has a significant positive effect.
- Minor differences in:

  - Coefficient values.
  - Significance of some variables (marginal *Sleep Duration* in R, not significant in SPSS).
  - R^2 calculations (slightly higher in SPSS).
- Both models perform similarly with an accuracy around 72-74

8. Conclusion

In this study, we compared the results of logistic regression models built using R and SPSS to identify predictors of depression in students. The analyses yielded consistent findings across both platforms, with *Age* and *Academic Pressure* being significant predictors of depression. However, minor discrepancies were observed, particularly in the coefficient values and the significance of *Sleep Duration*, which was marginal in R but not significant in SPSS. Additionally, the R^2 values were slightly higher in SPSS, suggesting a slightly better fit of the model.

Despite these differences, both models demonstrated similar performance, with accuracy rates ranging from 72% to 74%, indicating that either software can be effectively used for this type of analysis. Overall, the findings highlight the robustness of logistic regression in predicting depression risk among students and underscore the importance of considering multiple variables, such as academic pressure and sleep duration, when developing early intervention strategies in educational institutions.

Further research could explore more sophisticated techniques and larger datasets to refine these models, potentially incorporating additional factors such as social support and mental health history to enhance the prediction of depression in student populations.