Volume 4 Nomor 2, November 2025
https://journal.upy.ac.id/index.php/ASTRO/index

Analysis Comparison of Depression Levels Based on Gender and Academic Factors of Students

Miranti Verdiana*[1], Eko Dwi Nugroho[1] , Leslie Anggraini[1], Radhinka Bagaskara[1], Winda Yulita[1], Aidil Afriansyah[1], Muhammad Habib Algifari[1]
[1]Program Studi Teknik Informatika, Institut Teknologi Sumatera, Indonesia
E-mail: ***[1]miranti.verdiana@if.itera.ac.id** , [1] eko.nugroho@if.itera.ac.id , [1]leslie.anggraini@if.itera.ac.id , [1] radhinka.bagaskara@if.itera.ac.id, [1] winda.yulita@if.itera.ac.id , [1]aidil.afriansyah@if.itera.ac.id , [1]muhammad.algifari@if.itera.ac.id

Abstract

This study aims to analyze the level of depression among university students by examining gender and several academic indicators. The dataset includes responses from 27,901 students across various regions, with variables covering age, gender, academic pressure, study satisfaction, work/study hours, CGPA, and depression status. The analytical methods applied in this study include the chi-square test to evaluate the association between gender and depression status, point-biserial correlation to examine the relationship between numeric variables and depression, and logistic regression to develop a prediction model. The chi-square test results revealed no significant relationship between gender and depression (p = 0.774), indicating that depression affects both genders. In contrast, academic pressure exhibited the strongest correlation with depression status (r = 0.47), followed by work/study hours (r = 0.209) and study satisfaction (r = -0.168). The Logistic Regression model constructed using the four most relevant variables demonstrated satisfactory performance, achieving 75.5% accuracy and 82.1% recall in identifying students experiencing depression. These findings highlight the critical role of academic-related factors—particularly academic pressure—in influencing students' mental health. Therefore, targeted academic support strategies are essential to mitigate depression risks in higher education environments.

Keywords: *Student Depression, Academic Pressure, Gender, Logistic Regression, Mental Health Prediction*

1. INTRODUCTION

College students' mental health has been a growing issue in recent decades, particularly with the increasing academic pressures and social demands faced by college students. The World Health Organization (WHO) states that depression is one of the leading causes of mental health disorders in productive age groups, including among college students [1]. College students are often faced with academic challenges, career pressures, future uncertainty, and rapid life transitions, all of which contribute to the risk of psychological disorders such as depression and anxiety [2].

In Indonesia, the issue of depression among college students is also a concern because it can impact academic performance, social relationships, and overall quality of life. Surveys from several universities indicate that academic pressure, workload, and lack of emotional support are the main triggers for depressive symptoms [3]. Furthermore, many cases of mental disorders go undetected due to limited monitoring systems and the stigma surrounding mental health.

The specific problem addressed in this research is identifying the relationship between academic factors and demographic characteristics (such as gender and age) and levels of depression in college students. This research also aims to develop a predictive model for student depression status based on significant variables using logistic regression. With this predictive model, it is hoped that educational institutions can conduct early detection and provide appropriate interventions for students at high risk of depression.

This research is quantitative with a descriptive and predictive approach. Data were obtained from *Kaggle* in the form of a student survey dataset covering 27,901 rows of data. Analysis was conducted using the *chi-square test to analyze the relationship between gender and depression, point-biserial* correlation analysis for numeric variables, and the construction of a classification model using *logistic regression*. The findings of this study are expected to make a significant contribution to the development of a data-based early detection system for depression in higher education.

2. RESEARCH METHODS

Study This uses approach quantitative descriptive and inferential approach, with the objective to analyze the connection between academic factors and depression status in students, as well as building a predictive model using the regression method. The stages study includes :

2.1. Data collection

Dataset used in the study. This was obtained from Kaggle and developed by Adil Shamim. This dataset loading survey data on students who aim to explore the connection between academic pressure, demographic characteristics, and psychological conditions with depression status. Consisting of 27,901 student records, this dataset covers information like type, gender, age, city of origin, academic pressure, number of study hours or work, level of satisfaction with studying, GPA, and depression status experienced by students [4]. This dataset is very useful as a material study in an effort to understand factors that influence the mental health of students, in particular, to identify the risk of depression. The use of this data is expected can support development system prediction and early detection of depression data-based, which will later be beneficial for creating environment more academically healthy and supportive environment for students.

2.2. Analysis Statistics Descriptive

Analysis of descriptive statistics in research. This was done to describe the distribution and characteristics variables studied from data totaling 27,901 students. Stage this as an important base for understanding the data profile before doing analysis, inferential, and predictive model building [4].

The size centralization of data used includes the average (*mean*), median, and mode. The average is calculated for numeric variables like pressure, academics, age, work/study hours, and satisfaction with study, to mark the middle representing data distribution. The median is used for map position marking the middle in distribution, while the mode is applied especially on categorical variables, such as gender and city of origin of students, to see the category most [5].

The size of the distribution used includes range, variance, and standard deviation. Variable numeric, like pressure, academic, and work/study hours analyzed to evaluate how much big variations that occur in respondent data [6]. To detect possible outliers or significant data outliers, the Interquartile Range (IQR) is used, which is calculated as the difference between the third quartile (Q_3) and first quartile (Q_1) [7].

Visualization in the form of a *boxplot* is used to describe data distribution, detect *outliers*, and compare distribution between groups (eg, between student depression and depression).

The analysis results are descriptive. This gives information that the majority of respondents are students with pressure academic varies, the majority of respondents are in range mature young, and the proportion of students who experience depression needs to be addressed with special attention. Stage. This becomes base important for supporting testing hypotheses and predictive model development based on logistic regression at this stage.

2.3. Relationship and Correlation Test

After mapping data characteristics through analysis statistics, the stage furthermore is tests the relationship and correlation between the variables under research, especially between demographic, academic, and depression status of students. This test aims to identify variables that have a significant connection with depression so that they can be used in predictive model development [4][5].

2.3.1. Test of Relationship between Categorical Variables

To test the relationship between categorical variables such as gender and depression status, the Chi-Square (χ^2) test is used. This test examines the null hypothesis that there is no relationship between two categorical variables. The general formula for Chi-Square is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where $O_i$is the observed frequency and $E_i$is the expected frequency. If the p-value of the test result is < 0.05, then the null hypothesis is rejected, indicating a significant relationship [7].

2.3.2. Correlation Test between Numerical Variables and Depression

To measure the strength and direction of the relationship between numerical variables (academic pressure, study satisfaction, work/study hours, age) and depression status (binary variable), point-biserial correlation was used. (r_{pb}). The formula for point-biserial correlation is:

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s_X} \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

where $\bar{X}_1$and $\bar{X}_0$are the averages of groups 1 (depression) and 0 (no depression). Depression), $s_X$is the standard deviation numeric variables, $n_1$and $n_0$is the number of data for each group, and $n$is the total number of data [4].

2.4. Integration of Test Results into Prediction Models

After obtaining the results of descriptive analysis and testing of relationships and correlations between variables, the next stage is to integrate significant variables into the development of a predictive model. The method used in this study is *Logistic Regression*, which is suitable for modeling the probability of occurrence of categorical target variables, in this case, student depression status (*Depression* = 0 or 1). The selected predictor variables include academic pressure, age, work/study hours, and study satisfaction, which have shown significant relationships with depression status in previous analysis stages [5], [8].

Before modeling is carried out, the data that has gone through the pre-processing stage is divided into training data (80%) and test data (20%) using the *holdout method. Validation* to avoid *overfitting*. Categorical variables were converted to a numeric format using label *encoding*, while numeric variables were standardized to maintain a uniform scale [9]. The *Logistic Regression model* was constructed using the following equation:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where $p$is the probability of a student experiencing depression, $X_i$is the variable's predictor, and $\beta_i$is the regression coefficient that will be estimated [7]. This model was calibrated using training and test data; its performance was evaluated with test data. Evaluation done with count metric accuracy, *precision*, *recall*, *F1-score*, and uses a *confusion matrix* for map classification, right and wrong. Interpretation coefficient regression was done with the count *odds ratio* to evaluate how much big influence of variables on opportunity occurrence of depression [8].

3. RESULTS AND DISCUSSION

This chapter presents the results of the data analysis and discussion based on the methodology established in this study. The analysis was conducted comprehensively to understand data distribution patterns, identify relationships between variables, and develop a predictive model for student depression status. The primary objective of presenting these results is to answer the research questions: whether there is a significant relationship between academic and demographic variables and student depression status, and how effective the model is in predicting depression.

Each step of the analysis focuses not only on quantitative results but is also supported by in-depth interpretation and reinforcement with the latest literature. The presentation of data in tabular form, graphic visualizations, and model evaluation metrics provides a comprehensive and in-depth understanding of variable relationship patterns, with practical implications for developing mental health policy in higher education settings.

3.1. Analysis Results Statistics Descriptive

Analysis results, descriptive results are that of the 27,901 students who became respondents, distribution variables such as academic pressure, age, work/study hours, and satisfaction with studies vary. Average pressure is at the current level until height, distribution can be seen in Figure 1.

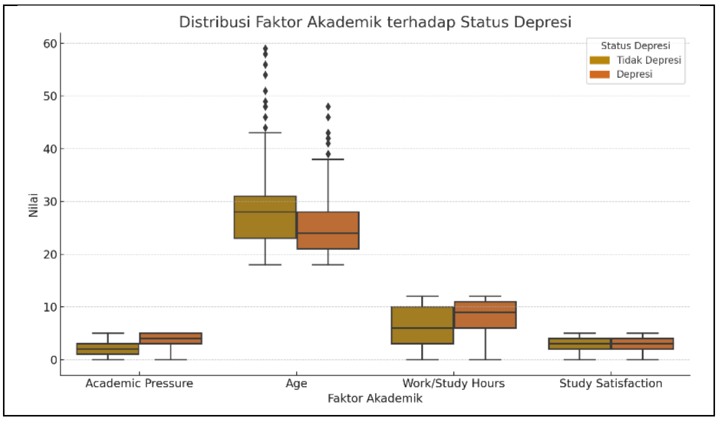

Figure 1. Boxplot of the Distribution of Academic Factors against Depression Status

Figure 1 provides a visualization distribution mark factor academic to depression status of students, with Non- Depressed and Depressed categories. The boxplot shows the distribution of four factors related to the main academic depression status in students, namely academic pressure, academics, age, work/study hours, and satisfaction with study. Based on visualization, visible that students who experience depression tend to have median pressure more academic tall compared to students who do not experience depression. This is also accompanied by wider data distribution in group depression, which indicates the existence of students with level pressure very high academic pressure. On the variable age, group students who experience depression generally have a lower median age, while group No depression shows a wider distribution , even though there are student outliers aged more old. This pattern indicates that young students Possible own vulnerability more vulnerable to depression.

Next, the distribution of the working/study hours variable shows that group depression has a lower median working/study hours more tall compared to group No depression. This strengthens the suspicion that long working/study hours potentially increase the risk of depression. Meanwhile, in the variable satisfaction study, the median value group depression tends more lower compared to the group with no depression, with a relatively small distribution. This shows that students who have low satisfaction studies low more risk of experiencing depression.

In a way, Overall, this boxplot gives an indication that high academic pressure, older age , long working/studying hours, and low study satisfaction are possible factors related to height prevalence depression in college students. However, the findings This nature descriptive and require testing statistics advanced For confirm the strength connection between the variables.

3.2. Results of the Test of the Relationship between Gender and Depression

The following is a table of contingency for the Chi-Square test between Gender and Depression Status, along with statistical test results :

Table 1. *Cross* Gender vs Depression Status *Tabulation*

Gender/Depression	**Not Depressed**	**Depression**
Female	5.133	7,221
Male	6,432	9.115

Table 1 presents the distribution amount students based on type, gender, and depression status. From the table, it can be seen that female students who don't experience depression totaled 5,133, while those who experienced depression as many as 7,221, while male students a man who do not experience depression totaled 6,432, while those who experienced depression as many as 9,115. In total, amount students who experience depression good in groups, women, and men compared to those who don't experience depression. This shows that the prevalence of depression is high enough among students, regardless of gender.

Analysis beginning based on table contingency, Table 1 provides a description distribution frequency amount students based on the combination of gender category and depression status. Although seen existence difference amount of student depression and not depression between group male and female groups, the distribution frequency is absolute, solely not yet enough to ensure existence significant connection in terms of statistics between the second variables.

In the context study, Table 1 only displays the proportion of students for every combination category, without giving information about whether the difference in distribution happens in a way coincidental (random) or, of course, shows a real pattern relationship. Therefore , it is necessary to test more formal statistics, namely the chi-square test, to evaluate the significant connection between variables [5].

The chi-square test allows researchers to test a hypothesis (H_0), which states that there is no connection between gender and depression status, with compare frequencies observations obtained from data with frequency expected frequencies happen If the second independent variable [4]. If the p-value of the test result is greater than small from the specified significance limit (for example, 0.05), then can concluded existence significant relationship. On the other hand, if the p-value is greater big, then there is enough proof to reject the null hypothesis.

Following is a results table from the chi-square test on gender data comparison of students who experience depression.

Table 2. Chi-Square Test Results

Chi-Square Statistics	0.083
Degrees of Freedom (df)	1
p-value	0.774

Chi-square test results between gender and depression status of students show a chi-square value of 0.083 with a p-value of 0.774, which is far more big from the significance limit of 0.05. This shows that No There is a significant relationship between gender and depression status in students in this dataset. The distribution of students by gender and depression status is almost balanced, so that there is quite a big difference in the influence connection both of them in a way statistical.

On testing, used degrees of freedom used were $df = 1$ because the total degrees of freedom were influenced by a combination of both variables, namely multiplication from the amount gender category minus 1 and the number of depression status minus 1, $df = (2 - 1)(2 - 1) = 1$.

3.3. Visualization Connection Variables

To understand the pattern connection between variables in this study, done analysis visualize the correlation through Figure 2, which presents a closeness relationship and direction correlation between major variables, including depression status of students, academic pressure of academics, age, work/study hours, satisfaction with studies, and GPA. Visualization. This aim For give a comprehensive description of the relatedness between influencing factors condition student mental health.

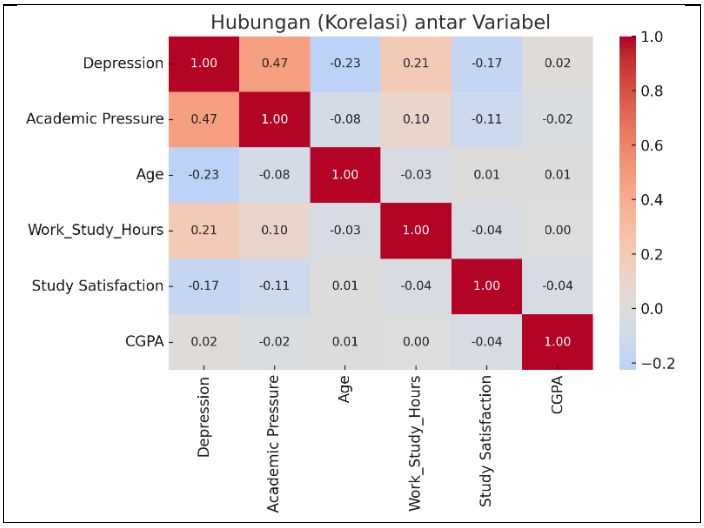

Figure 2. Relationship between Variables

Figure 2 shows the connection between variables, the main thing that was studied , namely *Depression* , *Academic Pressure* (Academic Pressure), *Age* (Age), *Work* / *Study Hours* (Work / Study Hours), *Study Satisfaction* (Study Satisfaction), and CGPA. Colors in the image represent closeness and direction connection between variables, with scale color from red (positive, strong) to blue (negative, strong).

Analysis results show that academic pressure own correlation strong positive correlation with depression ($r = 0{,}47$). This means that the higher the academic pressure students feel, the greater their likelihood of experiencing depression. This correlation confirms that academic pressure is a major risk factor for mental health disorders in students.

Next, work/study hours show a positive correlation with depression ($r = 0{,}21$), which indicates that excessive study or workloads can also potentially increase the risk of depression. This supports the understanding that a balance between academic and personal life is crucial for students' mental health.

On the other hand, the variable satisfaction studies own negative correlation ($r = -0{,}17$)against depression. This suggests that the higher a student's level of satisfaction with their studies, the lower their risk of developing depression. Study satisfaction serves as a protective factor, mitigating the negative impact of academic stress on mental health.

In Addition, age also shows a negative correlation ($r = -0{,}23$)with depression. This means that younger students tend to be more susceptible to depression than older students. This may be related to limited experience managing stress, uncertainty about the future, or the life transitions new students experience.

Interestingly, the CGPA variable has almost no significant correlation with depression ($r = 0{,}02$). This suggests that academic achievement (GPA) is not the primary factor directly related to students' mental health status. Academic stress and study satisfaction may have a greater impact than academic performance alone.

In a way, overall, this heatmap visualization gives a comprehensive understanding of the connection between variables that influence depression in college students. Findings. This supports existing literature that pressure and academic satisfaction studies are determinantly important for student mental health [8], and provide a base for developing more interventions directed at them.

3.4. Prediction Model Results: Logistic Regression

3.4.1. Classification Model Performance

As part of the evaluation performance of the depression status prediction model, the ROC *Curve* served to visualize model capabilities in differentiating between students who experience depression and those who do not. ROC *Curve* describes the *trade-off* between *True Positive Rate* (TPR) and *False Positive Rate (FPR)* at various classification thresholds, which gives a comprehensive understanding of model performance. Visualization. This not only shows how good a model is in identifying students with risk of depression, but also strengthens the reliability analysis conducted.

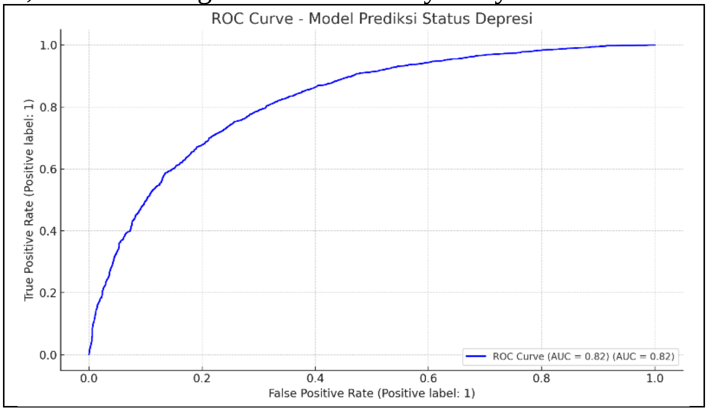

Figure 3. ROC *Curve* Prediction Depression Student

Figure 3 for the ROC *Curve* obtained shows predictive model capability in differentiating students who experience depression and those who do not. The blue line on the ROC curve shows that, along with the improvement mark *threshold*, occurs change in balance occurs between *True Positive Rate* (TPR) and *False Positive Rate* (FPR). The area under the curve (AUC) reached 0.82, which indicates that this model's own performance is sufficient for classification .

This AUC value provides practical meaning that If We in a way randomly choose one students who experience depression and one student who does not experience depression, then the model has a probability of around 82% to differentiate them with true. This shows the potential of the model as a tool to help in detecting students at risk of depression, although, of course still needs further development and testing more carry on.

In a way, overall, the interpretation from the ROC *Curve*. This strengthens the belief that the prediction model built can become a promising start for supporting early detection of depression in the educational environment.

Following is the *Threshold* Table with intervals of 0.1 to 0.9, with *True Positive Rate* (TPR) and *False Positive Rate* (FPR) of the prediction model :

Threshold, TPR, and FPR values of Prediction Depression Student

Threshold	**TPR**	**FPR**
0.1	0.9978	0.9240
0.2	0.9731	0.7426
0.3	0.9370	0.5698
0.4	0.8839	0.4379
0.5	0.8215	0.3363
0.6	0.7270	0.2386
0.7	0.6161	0.1605
0.8	0.4565	0.0862
0.9	0.2156	0.0260

Values : *The threshold* in Table 3 shows the probability limits used. To classify students into in category of depression or no depression. At every *threshold*, the calculated *True Positive Rate* (TPR) and *False Positive Rate* (FPR) show sensitivity, and the level of error classification is negative.

Table 3 has a connection with Figure 3; the ROC *Curve* depicts all combinations of TPR and FPR from various *thresholds*, forming a curve showing model performance. *The threshold table* gives concrete numbers that become the dots, dot, dot compiler the ROC curve. For example, *a threshold of* 0.5 produces TPR and FPR, which are one of the points on the ROC graph. The more good model, the

higher the TPR and increasingly lower FPR at all *thresholds*, which is reflected in an AUC approaching 1.
Connection This shows that the table *threshold* not only provides detailed figures, but also explains the mechanism formation ROC curve as a visual representation of predictive model performance.

3.4.2. Model Performance Evaluation

The next stage is the evaluation of the model obtained from the calculation of *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), and *False Negative* (FN) from the following *confusion matrix*.

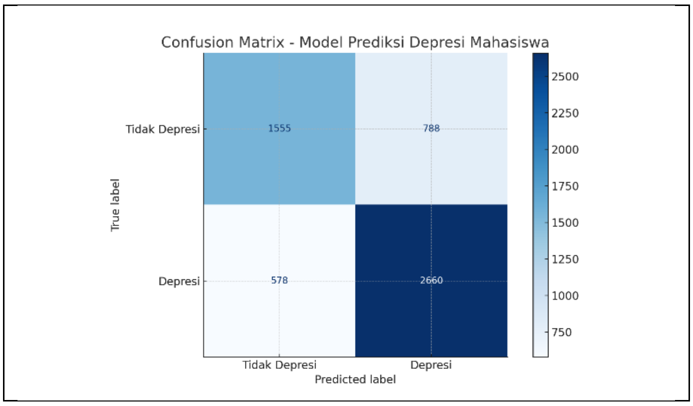

Figure 4. *Confusion Matrix* Prediction Depression Student

The confusion matrix displayed gives a description prediction model performance in classifying students into the categories of depression and not depression. Based on results analysis, the successful model classifies 2,660 students with depression as true (TP) and 1,555 students with no depression as true (TN). However, the model also produced 788 *False Positives,* namely classified students as depression, whereas they did not, and 578 *False Negatives*, namely students who should have been detected depression but were not.

Combination results classification. This was then counted to become the metric evaluation main, namely, accuracy, *precision*, *recall*, and *F1-score*.

Table 4. Model Evaluation

Model Evaluation	**Mark**
Accuracy	0.7552
Precision	0.7715
Recall	0.8215
F1 - Score	0.7957

Evaluation performance of the depression status prediction model shows sufficient results , with an accuracy of 75.5%, a *precision of* 77.1%, a *recall of* 82.1%, and an *F1-score of* 79.6%. Achievements This shows that the model has a good ability to differentiate between students who experience depression and what not depression. The proportion of students who experience depression in the dataset is said to be significant, which provides an opportunity to model to learn to distinguish patterns in groups from students who do not experience depression. This is seen from *recall* by 82.1%, which shows that a big part a real student experience depression is successfully identified with correct by the model.

However, the *precision* of 77.1 % shows that there are still several classified students as depressed by the model, even though they actually have no experience with depression (false positive). This possibility is caused by the presence of similar characteristics of some students who, even though in a way factual No depression, but own a close profile group depression, such as high academic pressure or long working / study hours. On the other hand, *false negatives* are recorded, showing that there are students who should have been detected to experience depression but were not classified as true. This is Possible because the variables predictors used (eg, academic pressure, satisfaction with study, and working / study hours) do not yet fully cover other factors that influence depression, such as psychosocial support , social factors, or health conditions.

F1-score of 79.6 % of the results obtained in the study. This shows that the model is not only accurate in predicting depression, but also quite sensitive in detecting when a real student experiences depression, without too many misclassifications. Findings . This is in line with the study, the latest showing that metric evaluation, like *F1-score* own important role in evaluating prediction model performance for depression, because it combines *precision* and *recall* For giving size size-sensitive balance to error classification [11]. Other studies also confirm that the model with an *F1-score* tall shows more capabilities , Good at detecting case depression, at the same time minimizing the amount of error prediction [12].

In a way, overall results evaluation shows that the model not only gives good performance classification, but also provides a clearer picture of the distribution and characteristics of students who experienced and did not experience depression in the dataset used. Evaluation. This emphasizes the importance of developing predictive models that are not only accurate but also sensitive and capable of minimizing error classification for supporting the detection of early depression among students.

3.4.3. Odds Ratio Integration

Table 2 shows the results of the *chi-square test*, which provides information on whether or not there is a significant relationship between gender and depression status. The *chi-square test* only detects the existence of a relationship between variables. Therefore, to measure the magnitude and direction of the relationship and provide a deeper understanding of the chi-square test results, further analysis using *the odds ratio is required.* The following are the results of the *odds ratio calculation using the Logistic Regression* model.

Table 5. Odds Ratio Values of Student Depression Prediction

Variables	Odds Ratio
Academic Pressure	2,279
Age	0.897
Work/Study Hours	1,126
Study Satisfaction	0.788

Obtaining *odds ratios* from the prediction model for student depression status provides a deeper understanding of the influence of key variables. Academic stress, for example, has *an odds ratio* of 2.28, meaning that every one-unit increase in academic stress increases a student's likelihood of experiencing depression by approximately 2.28 times. This reinforces the understanding that high academic stress is a major risk factor affecting student mental health.

Meanwhile, the age variable showed *an odds ratio* of 0.90, indicating that every additional year of age decreases the risk of depression by 10%. This means that younger students tend to be more susceptible to depression, possibly because they are still adapting to the academic demands and transitions of college life.

Furthermore, the variable of working or studying hours had *an odds ratio* of 1.13, meaning that every additional hour of studying or working per day increased the likelihood of depression by 13%. This suggests that excessive study load can contribute to an increased risk of depression, making it important for students to maintain a balance between academic and leisure time.

On the other hand, study satisfaction was shown to act as a protective factor, with *an odds ratio* of 0.79. This means that every one-unit increase in study satisfaction reduces the likelihood of depression by 21%. This underscores the importance of academic support, a conducive learning environment, and personal satisfaction in maintaining student mental health.

Overall, this interpretation of *odds ratios* not only provides statistical figures but also provides a more humanistic understanding of how academic pressure, age, study load, and study satisfaction interact to influence students' psychological well-being. These results are expected to form the basis for developing more targeted policies and interventions to support mental health in higher education settings.

4. CONCLUSION

Based on the research results conducted, can concluded as following.

1. Prediction models that use *logistic regression* have sufficient performance . Good at differentiating students who experience depression and those who do not, with an accuracy of 75.5%, a *precision* of 77.1%, a *recall of* 82.1%, an *F1-score of* 79.6%, and an AUC approaching 0.82. This indicates that the model is capable of identifying at-risk students' depression with enough effectiveness.
2. The model has high sensitivity (TPR) in detecting real student experiences depression, but little *precision* , more lower amount of classified students as depressed, whereas no *false positives*.
3. The results of the chi-square test strengthen findings. This shows existence significant connection between variables like academic pressure, study/work hours, satisfaction with study, and academic stress with depression status in students. Relationship. This is furthermore supported by the results' *odds* ratio, which gives a quantitative description of the size of the influence of each variable. In general detail, the pressure academic *odds ratio is* 2.279, which means increase risk of a student experiencing depression is more than doubled. Study hours or Working ($OR = 1{,}126$) also increases the risk, although not as much as academic stress. Conversely, study satisfaction ($OR = 0{,}788$) and academic stress ($OR = 0{,}90$) actually reduce the risk of depression, although the effect is not as great as academic stress. Factor risk main.
4. Analysis results. This shows that the combination of the chi-square test and the odds ratio can give a comprehensive description: chi-square validates the existence significant connection between variables with depression, while the odds ratio provides information on how much big influence each variable influences to risk of depression. This makes the model into more strong Good in a way statistics and interpretation.

5. SUGGESTION

Suggestions for further study should expand the variables used, such as financial support , social, and psychological factors, as well as involve data from various institutions, so that the results are more comprehensive and can be generalized. In addition, exploration methods will enable more predictions of advanced technologies, such as machine learning, which will enrich the analysis and improve the accuracy of the prediction model.

REFERENCE

[1] World Health Organization, “Depression,” 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

[2] A. K. Ibrahim, S. J. Kelly, C. E. Adams, and C. Glazebrook, “A systematic review of studies of depression prevalence in university students,” Journal of Psychiatric Research, vol. 47, no. 3, pp. 391–400, 2013. [Online]. Available: https://doi.org/10.1016/j.jpsychires.2012.11.015

[3] F. Putri and M. Amalia, “Faktor-faktor yang mempengaruhi depresi pada mahasiswa,” Jurnal Psikologi, vol. 15, no. 1, pp. 33–45, 2020.

[4] M. F. Triola, Elementary Statistics, 13th ed. Pearson, 2018.

[5] A. Field, Discovering Statistics Using IBM SPSS Statistics, 5th ed. SAGE Publications, 2018.

[6] D. M. Levine, D. F. Stephan, K. A. Szabat, Statistics for Managers Using Microsoft Excel, 8th ed. Pearson, 2020.
[7] D. S. Moore, G. P. McCabe, and B. A. Craig, Introduction to the Practice of Statistics, 9th ed. W.H. Freeman, 2017.
[8] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd ed. Wiley, 2000.
[9] S. García, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining. Springer, 2016.
[10] Q. Mou, J. Zhuang, et al., “The relationship between social anxiety and academic engagement among Chinese college students: A serial mediation model,” Journal of Affective Disorders, vol. 311, pp. 247-253, 2022. [Online]. Available: https://doi.org/10.1016/j.jad.2022.04.158
[11] Vu, Thien, et al., “Prediction of depressive disorder using machine learning approaches: findings from the NHANES,” BMC Medical Informatics and Decision Making, vol. 25, no. 83, 2025. [Online]. Available: https://doi.org/10.1186/s12911-025-02903-1
[12] X. Wang, et al., “Application of machine learning in depression risk prediction for connective tissue diseases,” *Scientific Reports*, vol. 15, no. 1706, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-85890-7