ESTIMATING THE PROBABILITY OF DEPRESSION IN ADOLESCENT BASED ON INTERNAL FAMILY FACTORS USING BINARY LOGISTIC REGRESSION AND NAÏVE BAYES

Muhammad Hasan Sidiq Kurniawan[1)]
Achmad Fauzan[2)]
[1,2)] Statistics Department, Faculty of Mathematics and Natural Science, Universitas Islam Indonesia, Yogyakarta, Indonesia
e-mail: achmadfauzan@uii.ac.id

ABSTRACT

Depression is a common mental health issue affecting approximately 300 million people globally, with severe cases potentially leading to death. Adolescents with depression are reported to have a 30-fold increased risk of suicide compared to other age groups, making early identification and intervention in this age group essential. Internal family factors play a crucial role in influencing adolescent mental health, including variables such as family type, parenting style, residence status, birth order, and parental occupation. This study aims to identify which internal family factors significantly impact the likelihood of depression and to estimate the probability of depressive episodes in adolescents. By understanding these factors, preventive measures can be better tailored to reduce adolescent depression in the future. The data utilized in this study is primary data obtained through a sampling process employing the simple random sampling method applied to first-year university students. A binary logistic regression model was employed to analyze the significance of each family-related factor. Findings indicate that parenting style and parental occupation are among the most significant factors associated with adolescent depression. The novelty of this study lies in the exploration of internal family factors that are rarely examined comprehensively in previous research, such as the effects of family type, birth order, and parental occupation. Additionally, the study adopts a dual-method approach, combining logistic regression and Naive Bayes, to provide a robust and comparative analysis of predictive accuracy. Probability estimates were conducted using both binary logistic regression and Naive Bayes methods. Results from these analyses suggest that a democratic parenting style tends to foster more stable mental health in adolescents, while adolescents with parents employed in the private sector or similar occupations face a higher likelihood of depression. Both methods demonstrated high predictive accuracy, with 96.31% for logistic regression and 96.93% for Naive Bayes.

Keywords: Binary Logistic Regression, Naïve Bayes, Depression, Adolescent, Family.

INTRODUCTION

Depression is a mental health condition characterized by a persistent disturbance in mood, often resulting in profound feelings of sadness and a loss of interest in previously enjoyable activities. Over 300 million people globally are estimated to live with depression. According to a 2014 report by the World Health Organization (WHO), the prevalence of depression has risen significantly. In severe cases, depression can lead to life-threatening outcomes, including suicide. Adolescents with depression are particularly vulnerable, facing a suicide risk approximately 30 times higher than that of other age groups (Stringaris, 2017). Numerous factors contribute to depression in adolescents, with internal family dynamics being among the most significant. The family plays a critical role in preventing

depression in adolescents; a supportive and well-functioning family environment helps teenagers feel secure and reduces their stress levels, thereby lowering the risk of depression (Wang et al., 2021).

Parental attentiveness is essential in helping children feel secure, which can lower the risk of depression. Thus, family structure—whether nuclear or extended—is a crucial factor to consider. In large families or situations where a child does not live with their parents, some children may receive less attention, which may increase the likelihood of depression. Additionally, birth order within large families may influence mental health outcomes. Parenting style also plays a significant role, as poor parenting practices can lead to elevated stress in adolescents (Van Assche et al., 2017). A family's social status, closely linked to income and the primary breadwinner's occupation, also affects adolescent mental health (Gao et al., 2023). The occupation of the household head indirectly influences adolescents' well-being, potentially impacting their mental health. Given the significant effects of these factors, it is essential to analyze them to estimate the likelihood of depression in adolescents. Identifying the most influential factors can support targeted prevention efforts aimed at reducing the risk of adolescent depression.

This study will employ binary logistic regression to identify factors that significantly influence depression in adolescents. Logistic regression is selected because it enables the prediction of the probability of an event occurring in a binary-dependent variable (Riffenburgh & Gillen, 2020) and allows for examining the combined influence of multiple predictors. Probability estimates will then be derived based on these significant factors. Additionally, probability estimation will be conducted using the Naïve Bayes method, which was chosen for its ability to estimate probabilities based on the posterior distribution derived from the sample data (Rezaei & Jabbari, 2022). This method also provides the conditional probability of depression occurring based on the observed predictors. Another reason for using the Naïve Bayes method is to deal with the imbalance in the dataset, as the Naïve Bayes can be adjusted to handle such cases (Badar et al., 2022). Meanwhile, the logistic regression will give biased results if there are imbalances in the data (Hendriyana et al., 2022). Although in this paper we were using the standard Naïve Bayes method, hopefully, the standard Naïve Bayes will also produce fewer bias results compared to the logistic regression method when the imbalance occurs.

Previous studies on adolescent depression have primarily focused on comparing depression levels based on abuse-related factors, such as emotional, physical, and sexual abuse (Freier et al., 2022; Najafi et al., 2022; Onyeaka et al., 2022). Additionally, other demographic factors, including gender, age, income, race, and education level, have been analyzed using logistic regression to understand their influence on depression.

Based on this background, the present study aims to analyze factors influencing depression in adolescents, focusing on internal family variables, including family type, parenting style, living status, birth order, and head of household occupation. Since all data are categorical, dummy variables will be incorporated into the logistic regression model. The best-fitting regression model will be selected based on the Akaike Information Criterion (AIC). A secondary objective of this study is to estimate the probability of depression based on significant factors identified through the regression analysis. To provide a comparison, probability estimation will also be conducted using the Naïve Bayes method. The accuracy of probability estimates for both methods will be evaluated using training and test datasets. By estimating the probability of adolescent depression, this study aims to assess the influence of internal family factors, thus informing targeted interventions such as family counseling programs facilitated by relevant authorities. Unlike previous studies, this research highlights several key differences, particularly in terms of the variables analyzed and the focus of the study's object.

Furthermore, this study employs a more comprehensive statistical approach by utilizing logistic regression modeling to examine relationships between variables, which is then critically compared to the performance of the Naïve Bayes model. This methodological framework is designed to provide deeper insights and support decision-making through more accurate and reliable analysis.

METHOD

This study utilizes primary data collected through direct measurement. A total of 2,278 individuals aged 17 to 22 years (adolescents and young adults) from the Special Region of Yogyakarta were observed to determine the presence of depression. The data utilized in this study is primary data obtained through a sampling process employing the simple random sampling method applied to first-year university students. Depression was assessed using the Eysenck Personality Inventory (EPI) questionnaire, a tool designed to evaluate an individual's psychological stability and propensity for anxiety. The EPI consists of 57 items, divided into three subscales: 24 items on Extroversion (E), 24 on Neuroticism (N), and 9 on Lie (L). Individuals scoring above 12 on all three components (E, N, and L) are classified as experiencing anxiety, which may indicate a predisposition toward depression. In this study, depression status serves as the dependent variable, while the internal family factors under investigation are treated as independent variables. The operational definitions of these variables are presented in Table 1.

Table 1. Operational definition of variables

Variables	Operational Definition of Variables
Depression Status (Y)	The psychological condition of teenagers. • Depression • Not depressed
Family type (X_1)	Types of families where teenagers come from. Consisting of nuclear families and extended families. • Nuclear family: a family consisting of parents and their children only. • Extended family: a family consisting of the nuclear family and other relatives, such as grandparents, uncles, and so on.
Parenting Style (X_2)	Parenting patterns are applied by parents in educating their children. • Authoritarian (strict). • Permissive (indulgent). • Democratic.
Living Status (X_3)	Family members living with the teenager. • Parents. • Other relatives/family. • Siblings (without parents). • Living alone (boarding house/dormitory).
Birth Number (X_4)	A person's position in the family and their birth order.
Head of household's occupation (X_5)	Livelihood is carried out by the head of the family. • Educator.

Variables	Operational Definition of Variables
	• State Civil Apparatus (ASN). • Private. • Self-employed. • Others

This study will begin by visualizing the data using pie and bar charts for categorical factors, followed by a descriptive statistical analysis to summarize key data characteristics. Subsequently, a series of binary logistic regression models will be developed, starting with individual models for each factor to assess their specific effects on depression risk. A combined binary logistic regression model will then be constructed to examine the collective influence of all factors. To enhance model simplicity and interpretability, the principle of parsimony will be applied, eliminating non-significant factors and starting with those of least significance. The refined model will be compared with the partial logistic regression models developed for each factor to assess differences in predictive power and significance. The AIC will be calculated for each model to determine the optimal balance between goodness-of-fit and model complexity, identifying the best regression model.

Based on the selected model, conditional probabilities of depression will be estimated, and these estimates will be further compared with those derived using the Naïve Bayes method. Prediction accuracy for both the logistic regression and Naïve Bayes models will be assessed using training and test datasets, allowing for a comparative analysis of model performance. The findings will then be interpreted based on the best-performing method, offering insights into the influence of internal family factors on adolescent depression and informing potential interventions for prevention. In summary, the steps for data analysis using binary logistic regression are as follows. **Step 1.** Creating a logistic regression model for each independent variable being analyzed. **Step 2.** Conducting a significance test for each independent variable based on the model created in Step 1. **Step 3.** Creating a logistic regression model by including all independent variables. **Step 4.** Testing the significance of each independent variable in the model. **Step 5.** Eliminating the most insignificant independent variable and then creating the logistic regression model without the eliminated independent variable. **Step 6.** Calculating the Akaike Information Criterion for the model. **Step 7.** Repeat Step 4 until Step 6 until there is no insignificant independent variable in the model. **Step 8.** Calculating the probability of depression using the latest model.

The binary logistic regression analysis will provide information about the independent variables that significantly affect depression. Based on these results, these variables will be analyzed using the Naïve Bayes method to predict the likelihood of depression. The steps are as follows. **Step 1.** Calculating the prior probability based on the frequency of each class of depression in the dataset. **Step 2.** Calculating the conditional likelihood using every independent variable for each class of depression. **Step 3.** Based on the results of Step 1 and 2, calculate the posterior probability for each class using the Bayes Theorem for each class of depression. **Step 4.** Using the result in Step 3, choose the maximum posterior probability as the prediction.

RESULTS AND DISCUSSION

Descriptive Statistics Analysis

Among the 2,278 adolescents surveyed, 94 were identified as experiencing depression. This indicates a relatively low prevalence of depression among adolescents in Yogyakarta. As shown in Figure 1, only 4% of surveyed adolescents reported depressive symptoms.

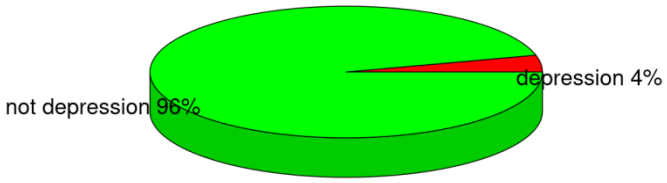

Figure 1. Pie Chart of the Percentage of Adolescents with Depression

Although depression cases constitute only 4% of the total observed sample, an in-depth investigation into internal family factors that may contribute to adolescent depression is essential. This analysis aims to further reduce the occurrence of depression and mitigate its potential consequences linked to family dynamics. Moreover, the presence of 94 adolescents experiencing depression is a noteworthy finding. By examining the incidence of depression based on the factors under investigation, preliminary conclusions can be drawn about which specific family-related factors influence the occurrence of depression among adolescents.

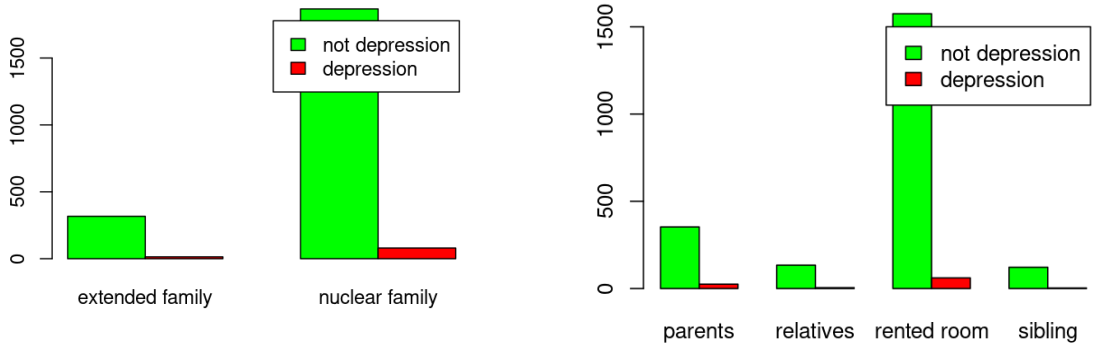

Figure 2. The number of depressions vs. non-depression in relation to the Family Type and Living Status variables

Teenagers living in nuclear families tend to experience higher rates of depression compared to those residing with extended families, likely due to the positive influence of broader family support in preventing adolescent depression (Fitria & Maulidia, 2018). Adolescents in extended families benefit from the support of relatives beyond their immediate family, such as uncles, grandparents, and aunts, which helps stabilize their psychological well-being and reduces the risk of depression. The crucial role of family in mitigating depression is further underscored by findings that adolescents living alone in dormitories or boarding houses are more susceptible to depression, as illustrated in Figure 2. Another living arrangement associated with adolescent depression is residing solely with parents, which may

increase risk when influenced by poor parenting practices (Safitri & Hidayati, 2013). As shown in Figure 3, the majority of adolescents who do not experience depression are those raised in a democratic parenting environment, where they are given the freedom to make independent decisions under parental guidance. Although some individuals within this parenting style experience depression, it is likely that additional factors contribute to these cases. Overall, the prevalence of depression is comparatively lower among adolescents raised with a democratic parenting style than with other styles. Beyond parenting style, parental occupation also plays a role in influencing depression, as socioeconomic factors significantly impact mental health.

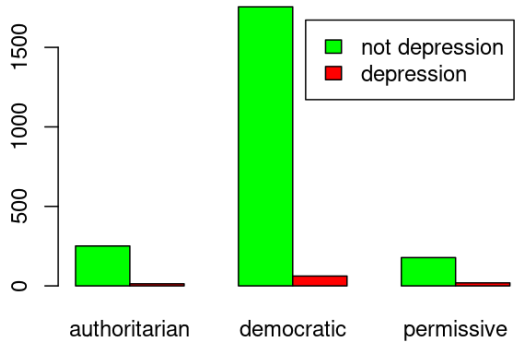

Figure 3. The number of depressions vs. non-depression in the parenting style variable

The family's socioeconomic status, largely determined by the type of employment serving as the primary income source, can contribute to the mental health outcomes of adolescents (Dewi et al., 2021).

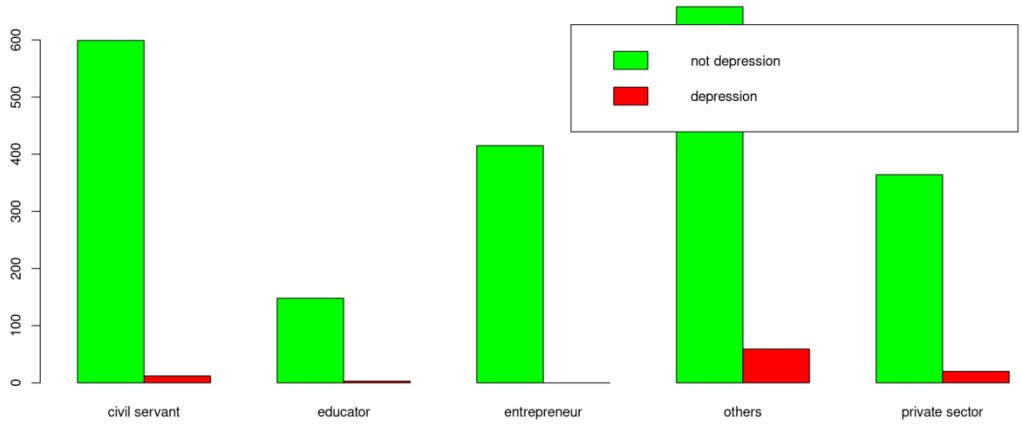

Figure 4. The number of depressions vs. non-depression in the variable of Head of Household's Occupation

Figure 4 illustrates that adolescents whose heads of households are entrepreneurs generally exhibit lower rates of depression. In contrast, the highest incidence of depression is observed among adolescents whose parents are categorized under "other occupations," which includes those who are

deceased, retired, or unemployed. These circumstances significantly impact the family's economic stability, which, in turn, indirectly affects the mental health of their children.

Binary Logistic Regression

The binary logistic regression model is used to examine the effect of an independent variable X on a categorical, dichotomous dependent variable Y (Cornish et al., 2023). Logistic regression was chosen because it can analyze the relationship between independent variables, whether categorical or continuous, and a binary dependent variable, such as depression or no depression. Logistic regression provides straightforward interpretations through log-odds coefficients, which can be translated into probabilities, facilitating the understanding of how each internal family factor influences the risk of adolescent depression. Additionally, this model enables statistical significance testing to identify the most influential factors. For example, a study by Chiu et al. (Chiu et al., 2021) highlighted the benefits of logistic regression in detecting the risk of major depressive episodes with significant severity, further supporting its applicability in studies of mental health. This model predicts the probability of a "success" event. If the probability of a successful event is defined as $P(Y = 1)$, the binary logistic regression model for the population is presented in Equation 1.

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon \tag{1}$$

with $\varepsilon \sim N(0, \sigma^2)$. The estimation for $\hat{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ an be obtained by minimizing $\sum_{i=1}^{n} \varepsilon_i^2$ so that the estimate of the regression model will be obtained in Equation (2).

$$\widehat{\log}\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i \tag{2}$$

Furthermore, the estimated probability of a successful event occurring with the known independent variable X is presented in Equation (3).

$$P(Y=1|X) = \frac{\exp\left(\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i\right)}{1 + \exp\left(\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i\right)} \tag{3}$$

So, the probability of a failure event is presented in Equation (4).

$$P(Y=0|X) = \frac{1}{1 + \exp\left(\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i\right)} \tag{4}$$

to find out whether the independent variable has a significant influence on the dependent variable, a Wald test can be carried out with the following test statistics (Kotimah & Wulandari, 2014).

$$W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \quad (5)$$

with $H_0: \beta_i = 0$ and a significance level of α, H_0 will be rejected if $|W| > Z_{\frac{\alpha}{2}}$. Based on the W value obtained, the p-value can be calculated, which is the error value obtained from the researcher which can be calculated by calculating the area of the normal curve based on the value of the test statistic. If the p-value is less than α, then H_0 will be rejected or the independent variable has a significant effect. The results of the univariate significance test in the logistic regression model for each of the factors studied at a significance level of α = 5% are. All independent variables analyzed in this study are categorical, except for the Birth Order variable. Consequently, dummy variables are created for the categorical variables to estimate the logistic regression model effectively. Table 2 displays results for a single dummy variable for each independent variable, based on the assumption that if one dummy variable shows significance, then the entire independent variable exerts a significant influence.

Table 2. The results of the significance test for each independent variable

Variable	W	$Z_{\frac{\alpha}{2}}$ Value	Decision
Family type (nuclear family)	-0.10		Not rejected H_0
Parenting Style (permissive)	4.04		Rejected H_0
Living status (parents)	1.28	1.96	Not rejected H_0
Birth number	-0.46		Not rejected H_0
Head of household's occupation (educator)	-2.48		Rejected H_0

According to the results in Table 2, only Parenting Style and Head of Household's Occupation significantly impact the likelihood of depression.

Table 3. The results of the significance test for the independent variables in the regression model with all independent variables

Variable	W	$Z_{\frac{\alpha}{2}}$ Value	Decision
Familiy Type (nuclear family)	-0.28		Not rejected H_0
Parenting style (authoritarian)	1.80		Not rejected H_0
Parenting style (permissive)	4.13		Rejected H_0
Living status (rented roomi)	-0.17		Not rejected H_0
Living status (parents)	0.91		Not rejected H_0
Living status (sibling)	-0.66	1.96	Not rejected H_0
Birth order	-0.83		Not rejected H_0
Head of household's occupation (Educator)	-2.68		Rejected H_0
Head of household's occupation (Civil servant)	-4.67		Rejected H_0
Head of household's occupation (Private sector)	-2.04		Rejected H_0
Head of household's occupation (Entrepreneur)	-0.03		Not rejected H_0

When a regression model is constructed including all independent variables listed in Table 3, only the Parenting Style and Parental Occupation variables demonstrate a significant influence on the likelihood of depression. Even after applying a parsimonious elimination of non-significant variables and comparing the AIC values, Parenting Style and Parental Occupation remain the only variables with a significant impact. AIC, which quantifies prediction error, can be calculated as follows (Chang et al., 2018).

$$AIC = -2\log(likelihood.model) + 2k \quad (6)$$

k represents the number of parameters. A lower AIC value suggests a better model fit, with Table 4 comparing AIC values across different models.

Table 4. Comparison of AIC values between regression models

Regression Model	AIC
Full Model	713.42
Model without Family Type variables	711.5
Model without Family Type, Residence Status variables	710.8
Model without Family Type, Residence Status, and Birth Order variables	709.75

Thus, the best logistic regression model is to use the Parenting Pattern and Parental Occupation variables.

Table 5. Estimation of the logistic regression model with the smallest AIC

Variable	Coefficient Parameter	W Value	$Z_{\frac{\alpha}{2}}$ Value	Decision
Intercept	-2.62	-17.16	1.96	Reject H_0
Parenting Style (Authoritarian)	0.59	1.87		Failed to reject H_0
Parenting Style (Permissive)	1.19	4.23		Reject H_0
Occupation (Educator)	-1.55	-2.58		Reject H_0
Occupation (Government Employees)	-1.55	-4.79		Reject H_0
Occupation (Private)	-0.54	-2.01		Reject H_0
Occupation (Self-Employed)	-17.19	-0.03		Failed to reject H_0

Based on Table 5, the logistic regression equation is as follows.

$$\log\left(\frac{P(Y = depression)}{1 - P(Y = depression)}\right) = -2.62 + 0.59X_2(\text{Authoritarian}) + 1.19\,X_2(\text{Permissive}) - 1.55\,X_5(\text{Educator}) - 1.55X_5(\text{Government Employees}) - 0.54X_5(\text{Private}) - 17.19X_5\,(\text{Self} - \text{Employed}) \quad (7)$$

Equation (7) is used to calculate $\log\left(\frac{P(Y=depression)}{1-P(Y=depression)}\right)$. Let

$$A = -2.62 + 0.59X_2(\text{Authoritarian}) + 1.19\,X_2(\text{Permissive}) - 1.55\,X_5(\text{Educator}) \\ -1.55X_5(\text{Government Employees}) - 0.54X_5(\text{Private}) \\ -17.19X_5\,(\text{Self} - \text{Employed})$$

To calculate the probability of depression, Equation (7) can be written as follows:

$$P(Y = depression) = \frac{\exp(A)}{1 + \exp(A)} \tag{8}$$

Independent variables with positive coefficients will increase the probability of depression, while the negative coefficients will decrease the probability. For example, Permissive parenting style will increase the probability of depression more than an Authoritarian parenting style because the coefficient for Permissive is higher. Meanwhile, parent's occupation as educator or government employees tend to lower the probability of depression because their coefficients are negative.

To find out whether the model is suitable for predicting cases of depression, the Hosmer-Lemeshow test will be carried out. With H_0 being the regression model fits the model, the test statistic for the Hosmer-Lemeshow test is distributed χ^2 with degrees of freedom $D - 2$ where D is the number of groups observed (Jawa, 2022).

$$HL = \sum_{d=1}^{D} \frac{(O_{1d} - E_{1d})^2}{N_d P(Y = depression)_d (1 - P(Y = depression)_d)} \tag{9}$$

with O_{1d} is the number of depression events in the d-th group, E_{1d} is the expected value of depression events in the d-th group, and N_d is the number of observations in the d-th group. The test results obtained are $HL = 0.33$ and the value is less than $\chi^2_{db:8;\alpha=5\%} = 15.51$ so that it fails to reject H_0. So, it can be said that the regression model in equation (7) is suitable for the data. Furthermore, to determine the accuracy of the prediction using the model. The data owned will be divided into training data and test data. It is determined that the training data is 80% of the total data selected randomly. The results of the model on the training data also show that only the Parenting Pattern and Parental Occupation variables have a significant effect as written in Table 6.

Table 6. Parameter significance test using training data

Variable	W Value	$Z_{\frac{\alpha}{2}}$ Value	Decision
Intercept	-15.39	1.96	Reject H_0
Parenting Style (Authoritarian)	1.71		Failed to reject H_0
Parenting Style (Permissive)	3.65		Reject H_0
Occupation (Educator)	-2.12		Reject H_0

Variable	W Value	$Z_{\frac{\alpha}{2}}$ Value	Decision
Occupation (Government Employees)	-4.17		Reject H_0
Occupation (Private)	-1.25		Failed to reject H_0
Occupation (Self-Employed)	-0.03		Failed to reject H_0

Subsequent predictions were made for the test dataset based on the model derived from the training data. Since no cases of depression were identified in the test data, the accuracy was calculated as the proportion of adolescents correctly predicted not to be depressed relative to the total observations in the test set.

Table 7. Confusion matrix table for depression prediction using Logistic Regression

Actual	Prediction	
	Depression	No-Depression
Depression	1	0
No-Depression	16	418

Results indicated that, among 434 adolescents without depression, only 16 were incorrectly predicted as depressed, yielding an accuracy rate of 96.31%. The high accuracy suggests that the logistic regression model is reliable for predicting depression, with an error rate below 5%. However, the proportion of adolescents experiencing is only about 4%, which may result in a biased conclusion from logistic regression. Therefore, the obtained results will also be compared with Naïve Bayes method.

As shown in Table 6, the reference category for the Parenting Style variable is Democratic parenting. The positive regression coefficient for the Parenting Style variable indicates that authoritarian and permissive parenting styles are associated with a higher likelihood of adolescents experiencing depression, whereas democratic parenting is associated with a lower likelihood. If, for example, the parents hold occupations categorized as "Other," the probability of depression for each parenting style can be calculated accordingly.

Table 8. Probability of depression based on parenting patterns

Parenting Patterns	Probability of Depression
Authoritarian	0.12
Permissive	0.19
Democratic	0.07

Calculate the probability of depression based on parenting patterns, can be done using Equation (8), by adjusting A according to the parenting patterns being studied, and setting the values of variables other than parenting patterns to zero. Therefore, we will get:

$$A_{Authoritarian} = -2.62 + 0.59(\text{Authoritarian} = 1) + 1.19\ (\text{Permissive} = 0)$$
$$A_{Permissive} = -2.62 + 0.59(\text{Authoritarian} = 0) + 1.19\ (\text{Permissive} = 1)$$

$$A_{Democratic} = -2.62 + 0.59(\text{Authoritarian} = 0) + 1.19\ (\text{Permissive} = 0)$$

Then the probability of depression can be calculated using Equation (8) and replace A with $A_{Authoritarian}$, $A_{Permissive}$, and $A_{Democratic}$. As shown in Table 8, permissive parenting is associated with a higher likelihood of depression among adolescents. This may be due to the lack of parental engagement, which leaves adolescents without a supportive outlet for their concerns. Conversely, an overly authoritarian approach, where parents interfere excessively in their children's lives, can also contribute to adolescent stress. Therefore, the most effective parenting style to reduce the incidence of depression appears to be democratic parenting, which allows children freedom while maintaining parental guidance. Following this, the probability of depression will be calculated based on parental occupation for adolescents raised under a democratic parenting style.

Table 9. Probability of depression based on parents' occupation

Occupation	Probability of Depression
Educator	0.02
Government employees	0.02
Private	0.04
Self-Employed	2.47×10^{-9}
Others	0.07

Similar to the parenting style variable, the probability of depression in **Table 9** can be calculated similarly, by adjusting the value of A and substituting it into Equation (8).

$$A_{educator} = -2.62 - 1.55\ (\text{Educator} = 1) - 1.55(\text{Government Employees} = 0) - 0.54(\text{Private} = 0) - 17.19\ (\text{Self} - \text{Employed} = 0)$$

$$A_{gov.emplo} = -2.62 - 1.55\ (\text{Educator} = 0) - 1.55(\text{Government Employees} = 1) - 0.54(\text{Private} = 0) - 17.19\ (\text{Self} - \text{Employed} = 0)$$

$$A_{private} = -2.62 - 1.55\ (\text{Educator} = 0) - 1.55(\text{Government Employees} = 0) - 0.54(\text{Private} = 1) - 17.19\ (\text{Self} - \text{Employed} = 0)$$

$$A_{self.emp} = -2.62 - 1.55\ (\text{Educator} = 0) - 1.55(\text{Government Employees} = 0) - 0.54(\text{Private} = 0) - 17.19\ (\text{Self} - \text{Employed} = 1)$$

$$A_{other} = -2.62 - 1.55\ (\text{Educator} = 0) - 1.55(\text{Government Employees} = 0) - 0.54(\text{Private} = 0) - 17.19\ (\text{Self} - \text{Employed} = 0)$$

Table 9 indicates that adolescents with self-employed parents are more likely to not experience depression, because the probability is the lowest (2.47×10^{-9}). The highest probability of depression is observed among adolescents whose parents fall under the "other" category because the probability is the highest among other categories for parents' occupation, that is 0.07. That means that among 100 household with parents' occupation as "other", seven of them will have adolescents with depression. The "other" category includes those who are unemployed, retired, or deceased. This is understandable, as parents who struggle to meet the family's daily needs may inadvertently increase stress levels within the household, contributing to adolescent depression. Adolescents with

unemployed parents tend to be at higher risk of experiencing depression due to being neglected by their parents. In some cases, these adolescents also often face abuse from their parents (Judd et al., 2023). Adolescents with parents working in the private sector also have a higher risk of depression compared to those with parents employed as government workers or educators. This suggests that the demanding schedules of private-sector jobs may affect parental availability, potentially leading to neglect of adolescents' emotional needs. Consequently, adolescents in such households may be more vulnerable to depression, especially when parenting practices are impacted by parents' time constraints. Also, parents working in the private sector often have shift-based jobs, which can affect e quality time available for the family. Parents with irregular shifts may have less quality time with their families, which has been shown to increase the likelihood of depression in adolescents (Han & Miller, 2009).

Naïve Bayes Methods

To evaluate the effectiveness of the logistic regression model, depression probability will also be estimated using the Naïve Bayes method. Naïve Bayes was chosen for its fast and efficient probabilistic approach, based on Bayes' Theorem with the assumption of independence among independent variables. Although this assumption is not always entirely realistic, Naive Bayes often delivers highly accurate classification results, particularly for datasets with relatively simple dimensions like this study. For instance, research by (Haque et al., 2021) demonstrated that Naive Bayes can effectively predict depression in children and adolescents with high accuracy, while recent findings Samanvitha et al. (2021) showcased the method's effectiveness in analyzing textual data to detect depressive symptoms.

Let $S = \{s_1, s_2, \dots, s_n\}$represent an n-observation sample consisting of both dependent and independent variables, where the categorical dependent variable serves as the label, indicating whether an individual is experiencing depression. In the Naïve Bayes approach, accurately determining this label is essential. Let H denote the hypothesis that an observation S belongs to the depression group. The probability of this hypothesis, $P(depression|S)$ will be estimated (Muflikhah et al., 2018). Using Bayes' theorem, the following probability calculations are obtained.

$$P(depression|S) = \frac{P(S|depression).P(depression)}{P(S)} \tag{10}$$

Observation S is classified into the depression group if P(depression | S) is greater than (not depressed | S). In this approach, the dataset is divided into training and test sets, with 80% allocated to training. The analysis focuses on Parenting Style and Parental Occupation variables, enabling a comparison of classification accuracy between the Naïve Bayes method and logistic regression analysis. To analyze using the Naïve Bayes method, first, we calculate the prior probability as follows.

$$P(Depression = Yes) = \frac{94}{2278} = 0.0413$$
$$P(Depression = No) = 1 - P(Depression = Yes) = 0.959$$

Next, we are calculating the conditional likelihood for Parenting Pattern (PP) and Parental Occupation (PO)variables for each class of depression and each category of PS and PO. For the Parenting Pattern variable, we will have the results as shown in **Table 10.**

$$P(PP = Democratic|Depression = Yes) = \frac{62}{94} = 0.640$$

$$P(PP = Democratic|Depression = No) = \frac{1755}{2184} = 0.804$$

$$P(PP = Authoritarian|Depression = Yes) = \frac{13}{94} = 0.138$$

$$P(PP = Authoritarian|Depression = No) = \frac{251}{2184} = 0.115$$

$$P(PP = Permissive|Depression = Yes) = \frac{19}{94} = 0.232$$

$$P(PP = Permissive|Depression = No) = \frac{178}{2184} = 0.082$$

Table 10. Estimation of the probability of depression based on parenting patterns

	Parenting Patterns		
	Democratic	Authoritarian	Permissive
Depression	0.64	0.14	0.23
Not Depressed	0.81	0.11	0.08

Table 10 shows that the chance of depression is higher in permissive and authoritarian parenting patterns. While democratic parenting patterns have a higher chance of non-depression. This conclusion is in line with what has been obtained using logistic regression analysis. Using the same formulation for the Parental Occupation (PO) variable, we will get the results shown in **Table 11.**

Table 11. Estimated probability of depression based on occupation

	Occupation				
	Educator	Government employees	Private	Self-Employed	Others
Depression	0.03	0.11	0.2	0	0.66
Not Depressed	0.07	0.27	0.17	0.19	0.30

As shown in Table 11, parental occupations categorized as "other" and "private" are associated with a higher likelihood of adolescent depression, while other job categories are more frequently linked to a lower risk of depression. Using the test dataset of 456 observations, all of whom were not depressed, there were 14 incorrect predictions, yielding an accuracy rate of 96.93% for the Naïve Bayes method. This accuracy is slightly higher than that obtained from logistic regression analysis; however, the difference between the two methods is not statistically significant. Consequently, both methods demonstrate similar accuracy, leading to consistent conclusions based on the interpretation of results.

CONCLUSION

The binary logistic regression and Naïve Bayes methods demonstrate comparable accuracy in predicting the probability of depression, with accuracy rates of 96.31% and 96.93%, respectively. Both methods identify parenting style and parental occupation as the most influential internal family factors associated with adolescent depression. A democratic parenting style is recommended, as it allows adolescents to feel supported and valued without feeling constrained. Regarding occupation, parents employed in the private sector should ensure they provide adequate attention and care to prevent feelings of neglect in their children. For adolescents whose parents are retired or deceased, addressing potential depression requires additional support and attention from other family members.

REFERENCE

Badar, M., Fisichella, M., Iosifidis, V., & Nejdl, W. (2022). Discrimination and Class Imbalance Aware Online Naïve Bayes. *Information Sciences*. https://doi.org/10.48550/arXiv.2211.04812

Chang, L., Zhou, Z., Chen, Y., Xu, X., Sun, J., Liao, T., & Tan, X. (2018). Akaike Information Criterion-based conjunctive belief rule base learning for complex system modeling. *Knowledge-Based Systems*, *161*(July), 47–64. https://doi.org/10.1016/j.knosys.2018.07.029

Chiu, I.-M., Lu, W., Tian, F., & Hart, D. (2021). Early Detection of Severe Functional Impairment Among Adolescents With Major Depression Using Logistic Classifier. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.622007

Cornish, R. P., Bartlett, J. W., Macleod, J., & Tilling, K. (2023). Complete case logistic regression with a dichotomised continuous outcome led to biased estimates. *Journal of Clinical Epidemiology*, *154*, 33–41. https://doi.org/10.1016/j.jclinepi.2022.11.022

Dewi, Y., Relaksana, R., & Siregar, A. Y. M. (2021). Analisis Faktor Socioeconomic Status (Ses) Terhadap Kesehatan Mental: Gejala Depresi Di Indonesia. *Jurnal Ekonomi Kesehatan Indonesia*, *5*(2), 29–40. https://doi.org/10.7454/eki.v5i2.4125

Fitria, Y., & Maulidia, R. (2018). Hubungan antara Dukungan Sosial Keluarga dengan Depresi pada Remaja di SMPN Kota Malang. *Prosiding Seminar Nasional Hasil Penelitian Dan Pengabdian Epada Masyarakat III*, *September*, 270–276.

Freier, A., Kruse, J., Schmalbach, B., Zara, S., Werner, S., Brahler, E., Fegert, J. M., & Kampling, H. (2022). Supplementary data for the mediation effect of personality functioning – Gender differences, separate analyses of depression and anxiety symptoms and inferential statistics of the relationship between personality functioning and different types of child. *Data in Brief*, *42*, 108–272. https://doi.org/10.1016/j.dib.2022.108272

Gao, R., Liang, L., Yue, J., Song, Q., Zhao, X., Fei, J., Hu, Y., Lv, J., Yuan, T., Guo, X., Meng, C., & Mei, S. (2023). The relationship between Chinese adults' self-assessments of family social status in childhood and depression: A moderated mediation model. *Journal of Affective Disorders*, *320*(June 2022), 284–290. https://doi.org/10.1016/j.jad.2022.09.115

Han, W., & Miller, D. P. (2009). Parental work schedules and adolescent depression. *Health Sociology Review*, *18(1),* 36-49. http://dx.doi.org/10.5172/hesr.18.1.36

Haque, U. M., Kabir, E., & Khanam, R. (2021). Detection of child depression using machine learning methods. *PLOS ONE*, *16*(12), e0261131. https://doi.org/10.1371/journal.pone.0261131

Hendriyana, Karo, I. M. K., & Dewi, S. (2022). Analisis Perbandingan Algoritma Support Vector Machine, Naïve Bayes, dan Regresi Logistik untuk Memprediksi Donor Darah. *Jurnal Teknologi Terpadu*, 8 (2), 121-126.

Jawa, T. M. (2022). Logistic regression analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arabia. *Alexandria Engineering Journal*, *61*(10), 7995–8005. https://doi.org/10.1016/j.aej.2022.01.047

Judd, N., Hughes, K., Bellis, M. A., Hardcastle, K., & Amos, R. (2023). Is parental unemployment associated with increased risk of adverse childhood experiences? A systematic review and meta-analysis. *Journal of Public Health, 54 (4),* 829-839. https://doi.org/10.1093/pubmed/fdad069

Kotimah, K. M., & Wulandari, P. S. (2014). Model Regresi Logistik Biner Stratifikasi Pada Partisipasi Ekonomi Perempuan Di Provinsi Jawa Timur. *Jurnal Sains Dan Seni Pomits*, *3*(1), 2337–3520.

Muflikhah, L., Ratnawati, D. E., & Putri, R. R. M. (2018). *Data Mining* (Cetakan Pe). Universitas Brawijaya Press.

Najafi, K., Khoshab, H., Rahimi, N., & Jahanara, A. (2022). Relationship between spiritual health with stress, anxiety and depression in patients with chronic diseases. *International Journal of Africa Nursing Sciences*, *17*(August), 100463. https://doi.org/10.1016/j.ijans.2022.100463

Onyeaka, H., Ajayi, K. V, Muoghalu, C., Eseaton, P. O., Azuike, C. O., Anugwom, G., Oladunjoye, F., Aneni, K., Firth, J., & Torous, J. (2022). Access to online patient portals among individuals with depression and anxiety. *Psychiatry Research Communications*, *2*(4), 100073. https://doi.org/10.1016/j.psycom.2022.100073

Rezaei, N., & Jabbari, P. (2022). Chapter 6 - Naïve Bayes' classifiers in R. In N. Rezaei & P. Jabbari (Eds.), *Immunoinformatics of Cancers* (pp. 71–85). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-822400-7.00010-5

Riffenburgh, R. H., & Gillen, D. L. (2020). 17 - Logistic regression for binary outcomes. In R. H. Riffenburgh & D. L. Gillen (Eds.), *Statistics in Medicine (Fourth Edition)* (Fourth Edition, pp. 437–457). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-815328-4.00017-6

Safitri, Y., & Hidayati, N. E. (2013). Hubungan Antara Pola Asuh Orang Tua Dengan Tingkat Depresi Remaja Di Smk 10 November Semarang. *Jurnal Keperawatan Jiwa*, *1*(1), 11–17.

Samanvitha, S., Bindiya, A. R., Sudhanva, S., & Mahanand, B. S. (2021). Naïve Bayes Classifier for depression detection using text data. *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, 418–421. https://doi.org/10.1109/ICEECCOT52851.2021.9708014

Stringaris, A. (2017). Editorial: What is depression? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *58*(12), 1287–1289. https://doi.org/10.1111/jcpp.12844

Van Assche, E., Moons, T., Cinar, O., Viechtbauer, W., Oldehinkel, A. J., Van Leeuwen, K., Verschueren, K., Colpin, H., Lambrechts, D., den Noortgate, W., Goossens, L., Claes, S., & van Winkel, R. (2017). Gene-based interaction analysis shows GABAergic genes interacting with parenting in adolescent depressive symptoms. *Journal of Child Psychology and Psychiatry*, *58*(12), 1301–1309. https://doi.org/https://doi.org/10.1111/jcpp.12766

Wang, D., Jiang, Q., Yang, Z., & Choi, J. K. (2021). The longitudinal influences of adverse childhood experiences and positive childhood experiences at family, school, and neighborhood on adolescent depression and anxiety. *Journal of Affective Disorders*, *292*(June), 542–551. https://doi.org/10.1016/j.jad.2021.05.108