Prediction of Mental Health Problems Based on Logistic Regression Model

Yining Wang *

Department of Bioinformatics, Dalian University of Technology, Dalian, 111000, China

* Corresponding Author Email: haiye@ldy.edu.rs

Abstract. This study aims to construct a logistic regression model using the "Mental Health Dataset" on Kaggle to predict mental health status and evaluate its accuracy based on the confusion matrix. The study selected 2,000 random samples and 16 variables, and after preprocessing the data, the model was built. Descriptive analysis found that "mental health history (binary)" was correlated to varying degrees with multiple variables, and ultimately 9 related factors were included in the model. The logistic regression results showed that social weaknesses, work interest, and mood swings were significantly related to the mental health status of college students. However, the model's predictive ability was limited, with an accuracy rate of 0.602, a Kappa value of -0.017, a sensitivity of only 0.053, and a specificity of 0.933. The study suggests that subsequent improvements in variable selection and model construction methods can enhance the predictive accuracy, providing a more reliable basis for early intervention and prevention of mental health problems.

Keywords: Logistic regression model; mental health prediction; confusion matrix.

1. Introduction

According to the World Health Organization (WHO), mental health refers to a condition in which individuals are aware of their own potential, can manage everyday challenges effectively, maintain productive and meaningful work, and contribute positively to their communities [1]. On the other hand, mental illness, or mental disorder, is characterized by alterations in a person’s cognition, emotions, or behavior, often resulting in significant distress and impairing their ability to function normally. Such disorders encompass a wide range of conditions, including anxiety disorders, depressive disorders, eating disorders, personality disorders, post-traumatic stress disorder (PTSD), psychotic disorders, and more [2]. The impact of mental disorders can be profound, affecting not only the individuals experiencing them but also those around them. In severe cases, mental illness may lead to tragic outcomes such as suicide, criminal behavior, or harm to others [3]. These consequences underscore the critical nature of mental health issues and the necessity of providing adequate mental health care and support.

Mental disorders account for a significant proportion of the global burden of disease (GBD), especially in low- and middle-income countries, where resources for diagnosis and treatment of mental disorders are severely inadequate. According to WHO, depression is one of the main causes of disability worldwide, affecting approximately 350 million people. The incidence of anxiety disorders is also very high, affecting about 264 million people [4]. Common mental health disorders include depressive disorders, anxiety disorders, schizophrenia, bipolar disorder, etc. Globally, depressive disorders affect a large number of people and are one of the main causes of disability; anxiety disorders are also widespread and affect a large number of people. Although the incidence of schizophrenia and bipolar disorder is relatively low, the burden they impose on individual patients and society is extremely heavy [5].

Under the COVID-19 pandemic, the mental health of children and adolescents is influenced by multiple factors. In terms of stress, social isolation and other disruptions to the HPA axis trigger neuroinflammation, and fear affects brain development. Regarding social isolation and diet, isolation stress leads to an increase in high-calorie food intake, which impacts the brain and gut microbiota. Among factors related to brain plasticity, restricted social interaction, exacerbated social inequality, and reduced contact with nature bring about a series of problems. Moreover, the pandemic's significant stress and the difficulty in fully utilizing coping resources due to social inequality jointly affect their mental health [6]. Research indicated that cross-diagnostic network intervention targeting potential personality risk factor may be an effective low-intensity intervention approach for preventing common mental disorders. This reflected the preventability of mental health issues, and early intervention through online means can effectively control the development of problems [7]. Therefore, it is of great importance to diagnose mental health problems as early as possible. Because early detection of problems enables prompt implementation of corresponding intervention measures to improve people's mental health issues.

In a previous study, Sahlan et al. processed data with Python and tools, used the Chi-square test to select factors related to mental health issues, and employed K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree to predict mental health problems among college students. The research determined that the Decision Tree was the optimal predictor, succeeded by KNN and then SVM [8]. Scholars Baba and Bunji used machine learning methods like logistic regression, elastic net, and LightGBM to analyze health survey data of Japanese undergraduate students in 2020 - 2021 and to predict students' mental health problems in the current and next year. The results showed that LightGBM performed well [9]. Additionally, Garriga et al. developed a machine learning model to continuously monitor patients for mental health crisis risk over a 28-day period with the help of electronic health records. The 6-month prospective study showed that the model had clinical value in 64% of cases, and it was the first time to achieve sustained prediction and clinical application exploration of multiple mental health crisis risks [10].

This paper aims to use one of the datasets from Kaggle called "Mental Health Dataset" to create a logistic regression model for predicting mental health conditions and examine its accuracy based on the confusion matrix.

2. Methods

2.1. Data Source

The data used in this study is sourced from the "Mental Health Dataset" dataset on Kaggle, which is provided by Jikadara. The dataset, which has an extensive array of features such as sentiment analysis scores and psychological metrics, provides possibilities for constructing predictive models. These models are designed to recognize or forecast mental health results relying on textual information. Such capabilities can prove beneficial for implementing early - stage interventions and providing support. The dataset contains 292365 samples and 17 variables in the .CVS format. The dataset's usability rating is 10.0, and it has been downloaded a total of 14940 times.

2.2. Variable Selection

The analysis in this paper uses only 2000 randomly selected observations and 16 variables from the dataset. Since the original variable named "Mental_Health_History" is not a numeric variable, it is difficult to use R for model analysis. Therefore, this paper creates a new variable in the dataset to replace this variable, turning the categorical variable into a numeric variable. By leveraging all available survey data (including gender, country, occupation, etc.), an algorithm is created to predict mental health issues. All 16 variables are listed in Table 1.

During the data preprocessing of this article, binary and multi-class variables are encoded as factors. It is planned to screen out variables with missing values exceeding 20% and handle the missing values using the "mice" package in R language. Since there are no variables with missing values exceeding 20% in the selected dataset, no variables are deleted.

Table 1. Different types of variables

Term	Type	Explanation
Gender	Categorical	0: Male, 1: Female
Country	Categorical	0: United States, 1: Poland, 2: Australia…
Occupation	Categorical	0: Housewife, 1: Corporate, 2: Student…
self_employed	Categorical	0: No, 1: Yes
family_history	Categorical	Do you have a family history of mental illness? 0: No, 1: Yes
treatment	Categorical	Have you sought treatment for a mental health condition? 0: No, 1: Yes
Days_Indoors	Categorical	0: 1-14 days, 1: 15-30 days, 2: 31-60 days…
Growing_Stress	Categorical	0: No, 1: Yes, 2: Maybe
Changes_Habits	Categorical	0: No, 1: Yes, 2: Maybe
Mood_Swings	Categorical	0: Low, 1: Medium, 2: High
Coping_Struggles	Categorical	0: No, 1: Yes
Work_Interest	Categorical	0: No, 1: Yes, 2: Maybe
Social_Weakness	Categorical	0: No, 1: Yes, 2: Maybe
mental_health_interview	Categorical	Have you sought treatment for a mental health condition? 0: No, 1: Yes, 2: Maybe
care_options	Categorical	0: No, 1: Yes, 2: Not Sure
Mental_Health_ History_Binary	Numeric	0: No/Maybe, 1: Yes

2.3. Method Introduction

In this study, a logistic regression model is applied to forecast mental health states, and a confusion matrix is utilized to evaluate the model's precision. The data collection is partitioned into a training subset and a testing subset. Specifically, the training subset encompasses 70% of the observations within the dataset, and the testing subset consists of the remaining 30%. The confusion matrix is then used to analyze the accuracy, sensitivity, and specificity of the logistic regression model.

The logistic regression model can be represented as:

$$P = \frac{exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\beta_m x_m)}{1+exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\beta_m x_m)} \tag{1}$$

Where β_0 is constant term， $\beta_1, \beta_2, \dots, \beta_m$ are partial regression coefficients. Perform a logit transformation on $f(x) = \frac{1}{1+e^{-x}}$, then $L(p) = ln\frac{p}{1-p}$. So, the logistic regression model can be expressed in the following linear form:

$$ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1, \beta_2 x_2 \dots, \beta_m x_m \text{cc} \tag{2}$$

The specific definition of the confusion matrix and the calculation formula are shown in Table 2 and Equation (3).

Table 2. Confusion Matrix

		Prediction Condition	
		Positive (P)	Negative (N)
Actual Condition	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TN+TP}{TP+FN+FP+TN} \tag{3}$$

Formula 1 represents the logistic regression model. Here, P is the predicted probability of the positive class. β_0 is the constant term, acting as an intercept, which is the predicted value when all independent variables $x_1, x_2, \dots, x_m$ are 0. $\beta_1, \beta_2, \dots, \beta_m$ are partial regression coefficients that quantify the influence of each independent variable on P. An increase in β_i (with other variables unchanged) affects P through the exponent part.

Formula 2 is the linear form after logit transformation. The left - hand side $ln\frac{p}{1-p}$, known as log - odds or logit, linearly combines β_0 and the products of β_i and x_i. This transformation simplifies the relationship between the predicted probability and independent variables, facilitating parameter estimation and variable - impact interpretation.

Formula 3 is a performance - evaluation metric for the logistic regression model. True Positive (TP) refers to the quantity of samples that are truly positive and have been accurately predicted as positive. True Negative (TN) represents the number of samples that are genuinely negative and have been correctly classified as negative. False Positive (FP) is the count of samples that are actually negative but have been mistakenly predicted as positive. False Negative (FN) indicates the number of samples that are truly positive yet have been wrongly predicted as negative. The given formula computes the ratio of the correctly predicted samples (the sum of TN and TP) to the total number of samples (the sum of TP, FN, FP, and TN), which shows the overall accuracy of the model's predictions.

3. Results and Disscusion

3.1. Descriptive Analysis

Figure 1 presents a variable correlation heatmap. In this heatmap, a redder shade implies a stronger positive correlation, while a bluer hue indicates a more pronounced negative correlation. From the heat map, "Mental_Health_History_Binary" shows varying degrees of correlations with multiple variables. Among them, the correlation with "Social_Weakness", "Work_Interest", “Mood_Swings”, and “Coping_Struggles” is relatively high; while the correlation with "Days_Indoors", "Growing_Stress", "Changes_Habits", "Occupation", and “Gender” is moderate.

Consequently, this paper incorporates only these nine factors into the logistic regression model. Other variables do not appear to have a significant enough correlation with the diagnostic outcome, so there is no need to include them in the model.

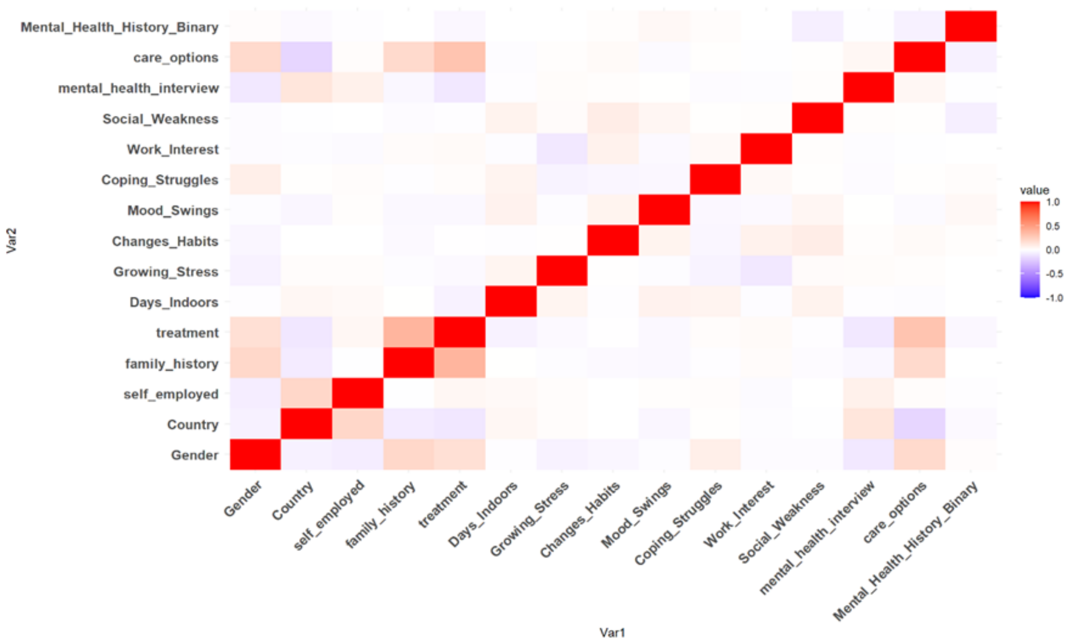

Fig. 1 Variables Correlation Heatmap

The bar chart in Figure 2 respectively presents the distribution of different levels of four variables: “Social_Weakness”, “Work_Interest”, “Mood_Swings”, and “Coping_Struggles”. The frequencies of the three levels of “Social_Weakness” are approximately the same; for “Work_Interest”, the frequencies of levels 1 and 3 are higher than that of level 2; the frequency of level 1 of "Mood_Swings" is the highest, while the frequencies of levels 0 and 2 are relatively low and close; the frequencies of the two levels of "Coping_Struggles" are basically the same. These distributions reflect the frequency characteristics of each variable in different states.

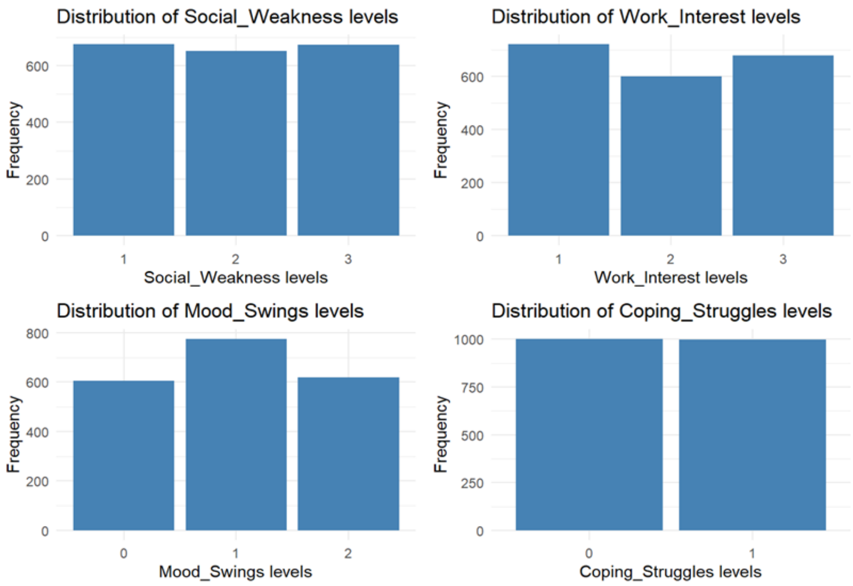

Fig. 2 Distribution levels bar plot

The histogram in Figure 3 shows the ditribution of the Mental Health History Binary variable, which is a binary variable. The graph shows a significant difference in the frequency distribution of its two values. The frequency of the value on the right (possibly corresponding to 1) is approximately 1,200, which is much higher than that of the value on the left (possibly corresponding to 0), which is about 800. This indicates that in the data, there are more observations where the Mental Health History Binary variable takes the value corresponding to the right side.

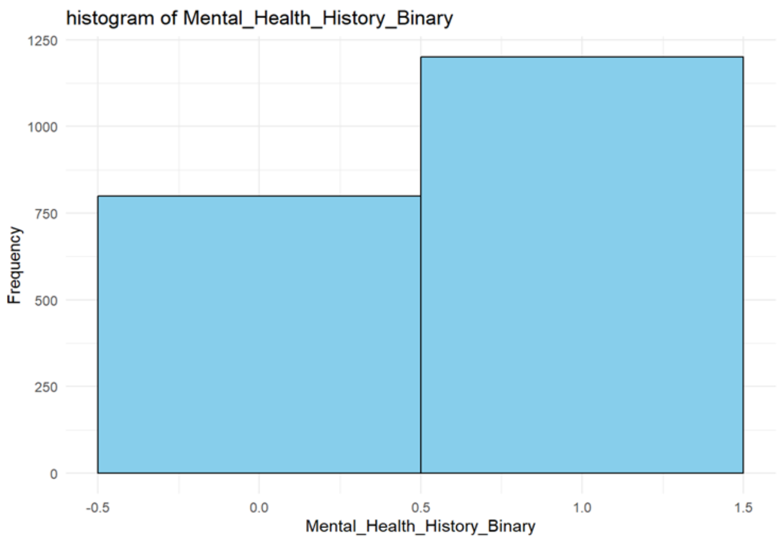

Fig. 3 The histogram of Mental_Health_History_Binary

The four mosaic plots in Figure 4 respectively present the relationships between "Social_Weakness", "Work_Interest", "Mood_Swings", "Coping_Struggles" and "Mental_Health_History_Binary". The graphs use color to indicate the magnitude and sign of the Pearson residuals, and all P-values are far less than 0.05 (e.g., $< 2.22 \times 10^{-16}$), suggesting that these four variables have extremely significant correlations with Mental Health History Binary, that is, their distributions under different values of Mental Health History Binary are significantly different.

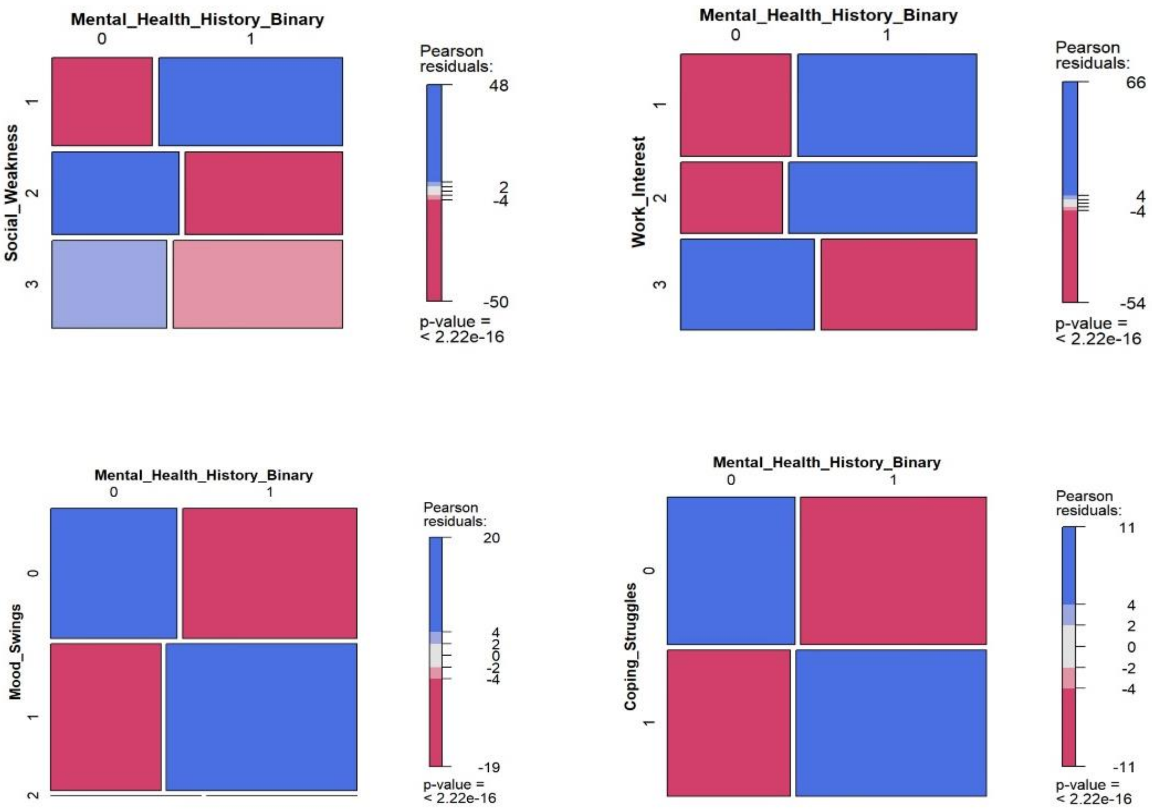

Fig. 4 The mosaic plot of four key variables

3.2. Logistic Regression Results

When conducting research on the mental health status, the results of logistic regression analysis provide crucial information for understanding the relationship between various factors and mental health problems. Judging from the regression coefficients and P-values (Table 3): The estimated coefficients for "Social WeaknessYes" and "Social WeaknessNo" are 0.217 and 0.153 respectively, with corresponding P-values of 4.052×10^{-102} and 7.862×10^{-54}. The P-values are far less than 0.05, indicating a significant association between social weakness and mental health status. The positive coefficients suggest that individuals with social weakness are more likely to experience mental health problems. The coefficients for "Work InterestYes" and "Work InterestNo" are -0.176 and -0.213 respectively, with P-values of 2.88×10^{-64} and 6.40×10^{-104}. This indicates that work interest is significantly related to mental health status, and the negative coefficients suggest that people with lower work interest are more likely to experience mental health problems. The coefficient of "Mood SwingsMedium" is 0.299, and that of "Mood SwingsLow" is -0.250, with their P-values being 4.59×10^{-177} and 2.29×10^{-136} respectively. This indicates that different levels of mood swings are significantly related to mental health status, and the impacts of medium and low levels of mood swings on mental health show different directions.

Table 3. Logistic Regression Information

term	estimate	std.error	statistic	p.value
Social_WeaknessYes	2.175×10^{-1}	0.010	2.146×10^{1}	4.052×10^{-102}
Social_WeaknessNo	1.533×10^{-1}	0.010	1.545×10^{1}	7.862×10^{-54}
Work_InterestYes	-1.756×10^{-1}	0.010	-1.693×10^{1}	2.883×10^{-64}
Work_InterestNo	-2.130×10^{-1}	0.010	-2.165×10^{1}	6.403×10^{-104}
Mood_SwingsMedium	2.985×10^{-1}	0.011	2.837×10^{1}	4.586×10^{-177}
Mood_SwingsLow	-2.502×10^{-1}	0.010	-2.485×10^{1}	2.286×10^{-136}

The prediction ability of the model is evaluated through the confusion matrix (Table 4, Figure 5): the accuracy rate of the model is 0.602, which means that when predicting the mental health problem, the proportion of correct prediction of the model is only 60.2%, and the overall prediction effect needed to be improved. The Kappa value is -0.017, which reflects the poor consistency between the model prediction results and the actual situation, and almost no substantial agreement has been reached. The sensitivity is only 0.053, indicating that the model's ability to correctly identify people with actual mental health problems is very low, and it may miss a large number of individuals with actual mental health problems. The specificity is 0.933, indicating that the model has a strong ability to correctly predict people without mental health problems, and can better identify people without mental health problems.

Table 4. Confusion Matrix Information

Accuracy	Kappa	Sensitivity	Specificity
0.602	-0.017	0.053	0.933

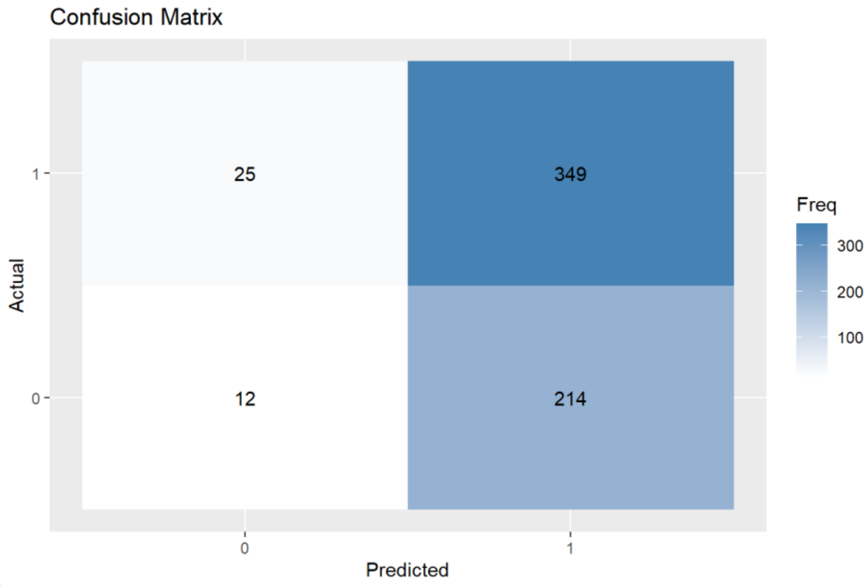

Fig. 5 Confusion Matrix

4. Conclusion

In conclusion, this research aims to construct a logistic regression model using the "Mental Health Dataset" on Kaggle for predicting mental health status. After carefully selecting 2,000 random samples and 16 variables, and performing data preprocessing, the model is established. Descriptive analysis reveals that "Mental_Health_History_Binary" has varying degrees of correlations with multiple variables, leading to the inclusion of 9 related factors in the model.

The logistic regression results demonstrate that social weaknesses, work interest, and mood swings are significantly related to the mental health status. However, the model's predictive performance leaves much to be desired. With an accuracy rate of only 0.602, it can correctly predict mental health problems in just 60.2% of cases. The Kappa value of -0.017 indicates a poor agreement between the model's predictions and the actual situation. The extremely low sensitivity of 0.053 means the model is highly likely to miss a large number of individuals with actual mental health issues. Although the specificity of 0.933 shows the model could effectively identify those without mental health problems, overall, its predictive ability is limited.

Future studies should focus on improving variable selection methods. This could involve exploring more advanced feature selection algorithms to ensure that the variables included in the model are more relevant and influential. Additionally, enhancing the model construction process, such as trying different optimization techniques or incorporating more complex model structures, might lead to better predictive accuracy. By doing so, this paper can provide a more solid basis for early intervention and prevention of mental health problems, ultimately contributing to better mental health outcomes for individuals.

References

[1] Michaud P A, Fombonne E. Common mental health problems. Bmj, 2005, 330(7495): 835-838.

[2] Chesney E, Goodwin G M, Fazel S. Risks of all-cause and suicide mortality in mental disorders: a meta-review. World psychiatry: official journal of the World Psychiatric Association (WPA), 2014, 13(2): 153-160.

[3] Banerjee D, Kosagisharaf J R, Sathyanarayana Rao T S. 'The dual pandemic' of suicide and COVID-19: A biopsychosocial narrative of risks and prevention. Psychiatry research, 2021, 295: 113577.

[4] Rehm J, Shield K D. Global Burden of Disease and the Impact of Mental and Addictive Disorders. Current psychiatry reports, 2019, 21(2): 10.

[5] Patel V, Chisholm D, Dua T, Laxminarayan R, Medina-Mora M E. Mental, Neurological, and substance use disorders: Disease Control Priorities. The International Bank for Reconstruction and Development, The World Bank, 2016.

[6] De Figueiredo C S, et al. COVID-19 pandemic impact on children and adolescents' mental health: Biological, environmental, and social factors. Progress in neuro-psychopharmacology & biological psychiatry, 2021, 106: 110171.

[7] Musiat P, Conrod P, Treasure J, Tylee A, Williams C, Schmidt U. Targeted prevention of common mental health disorders in university students: randomised controlled trial of a transdiagnostic trait-focused web-based intervention. PloS one, 2014, 9(4): 93621.

[8] Sahlan F, Hamidi F, Misrat M Z, Adli M H, Wani S, Gulzar Y. Prediction of mental health among university students. International Journal on Perceptive and Cognitive Computing, 2021, 7(1): 85-91.

[9] Baba A, Bunji K. Prediction of mental health problem using annual student health survey: machine learning approach. JMIR Mental Health, 2023, 10: 42420.

[10] Garriga R, Mas J, Abraha S, et al. Machine learning model to predict mental health crises from electronic health records. Nature Medicine, 2022, 28(6): 1240-1248.