Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 1, 2025

Identifying Depression Using Machine Learning

Jiang Lu *

Department of Tianjin University of Commerce, Tianjin, China
*** Corresponding author Email:** 2524493299@qq.com

Abstract: Depression is the leading cause of disability worldwide. However, accurately estimating the epidemiological factors that contribute to depression remains challenging. Deep learning algorithms can be used to assess the factors that contribute to the prevalence and clinical manifestations of depression. In this paper, five machine learning models, logistic regression, decision tree, random forest, SVM, and CatBoost, were used to assess depression in 5533 adult participants from the 2018 NHANES database. The results show that random forest is the best model for identifying depression, with the highest area under the working characteristic curve (AUC), followed by CatBoost and decision tree models. This suggests that using machine learning can accurately predict depression risk.

Keywords: Depression; Machine Learning; CatBoost; Random Forest; Decision Tree.

1. Introduction

Depression is a common mental disorder. According to the World Health Organization, approximately 35,020.14 million people suffer from this disease each year, and this situation has persisted for many years [1].It is estimated that 5% of adults suffer from depression, and the global prevalence of depression has been increasing in recent decades due to urbanization, overall population growth and changes in its age structure [2].Depression is the leading cause of disability worldwide and a major contributor to the global burden of disease [3].Therefore, large-scale national surveys have been conducted to determine the prevalence and risk factors of depression. In the United States, several national surveys have measured the incidence of depression in adolescents [4] and the general population [5]. The results show that depression has seriously affected people's daily life and physical and mental health, and even further affected family harmony and social relationships. Depression has placed a heavy burden on the patient's family and society, and therefore depression has become an important public health issue. Adults show complex pathogenesis of depression, which is affected by multiple risk factors such as gender, age, education level, chronic diseases, sleep and physical health [6]. Some previous studies have also identified the relationship between demographics [7], lifestyle behavior conditions [8] and physical exercise [9] and depression. Therefore, establishing a depression risk assessment model based on risk factors is conducive to the early detection and early treatment of people at high risk of depression.

Traditional machine learning methods, such as multivariate logistic regression, can help locate clinical manifestations of depression in these survey data. Dipnall used regression analysis in machine learning to find many biomarkers associated with depression in the National Health and Nutrition Examination Survey dataset [10]. Sohrab used the National Health and Nutrition Examination Survey data from 2005 to 2016 and used logistic regression to assess the temporal trend of depression prevalence [11].

However, traditional machine learning models have some limitations. For example, before building a regression model, we should fully understand our data attributes and model functions in order to successfully apply the model. Therefore, some other machine learning models can be combined for prediction. For example, Mudasir used a dataset of various depression signals from online social network (OSN) platforms (including Facebook, Twitter, and YouTube) to propose an efficient model based on artificial intelligence (AI) and deep learning (DL) to identify patients with depression on social media platforms. Experiments showed that the deep learning models LSTM and CNN as well as the hybrid (CNN+LSTM) model achieved high accuracy on all single and combined datasets [12]. Zhang Chenyang used five machine learning models to identify depression in middle-aged and elderly people using the National Health and Nutrition Examination Survey data from 2011 to 2018. The results showed that CatBoost was the best model for identifying depression, with the highest area under the operating characteristic curve (AUC) [13].

In this paper, we focus on investigating how machine learning algorithms can assess epidemiological, demographic, lifestyle, and other factors that contribute to depression in large survey datasets. We further compare the performance of the CatBoost algorithm with several traditional machine learning algorithms, such as decision trees and logistic regression. Our goal is to evaluate the utility of machine learning in identifying risk factors associated with depression in adults in large survey datasets, and to provide a scientific basis for early detection and early treatment of people at high risk of depression, making machine learning a valuable tool for clinicians to decide to care for their patients.

2. Method

2.1. Dataset and Study Population

This paper uses data from the 2017-2018 National Health and Nutrition Examination Survey (NHANES) to train machine learning algorithms and other machine learning classifiers. NHANES is a cross-sectional, population-based study of the non-institutionalized civilian population in the United States that assesses the health and nutritional status of adults and children [14]. Its data collection includes an in-home interview component and a physical examination component conducted in a mobile examination center. The interview includes demographic, socioeconomic, dietary, and health-related questions, and the physical examination

component includes medical, dental, physiological, medical examinations, and laboratory measurements [15]. All NHANES data are in the public domain and can be accessed on the website of the National Center for Health Statistics (https://www.cdc.gov/nchs/nhanes).

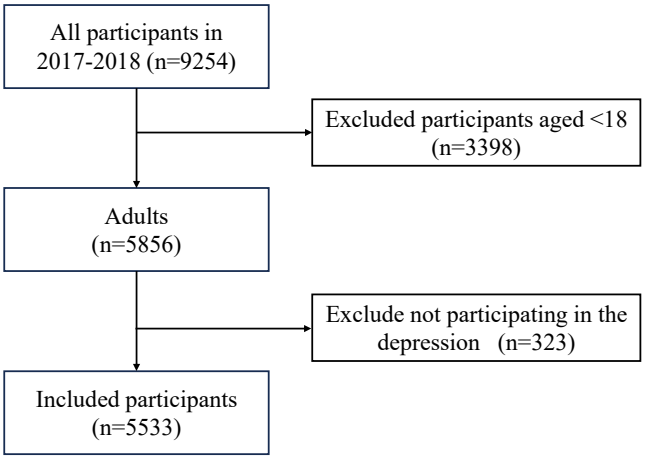

Figure 1. Flowchart of inclusion and exclusion criteria for study samples

The dataset used in this article initially included 9254 participants, and the inclusion criteria for the analysis sample were adults aged 18+ who answered questions related to depression. Therefore, we ultimately used 5533 participants as research subjects. The values "9", "99", "777", and "999" represent "don't know" answers to the variables and are therefore considered missing values. The specific inclusion and exclusion criteria are shown in Figure 1.

2.2. Disease Definition

Depression severity was determined by nine questions from the Patient Health Questionnaire (PHQ), a version of the PrimeMD diagnostic tool that is a brief, reliable, and valid measure. The PHQ-9 assesses symptoms of a major depressive episode in the previous 2 weeks and has a score range of 0-27. We defined a score greater than or equal to 10 as clinically relevant depression because it has reliable sensitivity and specificity for detecting major depression at this threshold. We eliminated people with missing answers to six questions. During 2017-2018, 5533 participants were assessed for depression and 472 were diagnosed with depression (8.53%).

2.3. Variable Selection

Table 1. Depression Baseline Table

Variable	Healthy(n=5055)	Depression(n=472)	p test
Age	49.78 (18.70)	49.93 (18.03)	0.87
Gender (%)			<0.001
Male	2478 (49.0)	188 (39.8)	
Female	2577 (51.0)	284 (60.2)	
Race (%)			<0.001
Mexican American	688 (13.6)	63 (13.3)	
Other Hispanic	718 (14.2)	89 (18.9)	
Non-Hispanic White	1701 (33.6)	193 (40.9)	
Non-Hispanic Black	1183 (23.4)	99 (21.0)	
Non-Hispanic Asian	765 (15.1)	28 (5.9)	
Education (%)			<0.001
<High School	994 (19.7)	122 (26.0)	
=High School	1252 (24.8)	131 (27.9)	
>High School	2799 (55.5)	217 (46.2)	
Marital (%)			<0.001
Single	2898 (60.3)	193 (43.0)	
Married	1054 (21.9)	162 (36.1)	
Divorced	853 (17.8)	94 (20.9)	
Poverty Index Ratio	2.57 (1.61)	1.96 (1.47)	<0.001
BMI (kg/m2)	28.69 (6.85)	30.41 (7.72)	<0.001
Smoke (%)			0.029
No	69 (3.5)	2 (0.8)	
Yes	1897 (96.5)	260 (99.2)	
Never	548 (11.8)	34 (7.2)	
Drinke (%)			0.002
Former Drinker	921 (19.8)	124 (26.3)	
Light Drinker	2057 (44.2)	200 (42.5)	
Moderate Drinker	856 (18.4)	84 (17.8)	
Heavy Drinker	268 (5.8)	29 (6.2)	
Albumin-to-Creatinine Ratio (mg/g)	47.18 (353.07)	71.42 (304.89)	0.154
Glycohemoglobin (%)	5.82 (1.05)	6.02 (1.40)	<0.001
Triglycerides (mmol/l)	1.25 (1.13)	1.44 (1.43)	0.019
LDL-C (mmol/l)	2.88 (0.94)	2.93 (1.05)	0.48

This paper selects indicators from four aspects: demographics, socioeconomic status, lifestyle, and some disease indicators. Among them, demographic variables are self-reported, and the body mass index (BMI) is calculated by measuring weight (kg) divided by the square of height (m); some disease indicators are selected from three aspects: liver and kidney, diabetes, and heart disease. The specific variables and variable levels are shown in Table 1.

3. Data Processing

In order to solve the problem of poor quality of the data set, it needs to be preprocessed accordingly. In this paper, the data set is processed step by step as follows: missing value processing, outlier processing, data imbalance processing, feature selection, and finally the data that can be used by the machine learning algorithm is formed. The processing of missing values in this paper uses R studio, and the model construction is analyzed and processed using scikit-learn (a third-party python package) in jupyter-notebook.

3.1. Feature Selection

Recursive Feature Elimination (RFE) is a feature selection method that aims to find the most influential features by recursively reducing the size of the feature set. It is done by recursively building the model and selecting the most important features (based on weights), removing the least important features, and then repeating the process on the remaining features until a preset number of features is reached or the model performance becomes stable.

3.2. Missing Value Handling

Missing data are present in almost every study, and the choice of method to handle missing data is usually not important when the proportion of missing data is less than 5%. However, in large epidemiological studies, it is not uncommon for the proportion of missing data to exceed this percentage, which may reduce statistical power, produce biased parameters, and increase the risk of type I errors. Therefore, it is very important to handle missing data correctly [16]. Multiple imputation is a useful and flexible strategy to address the problem of missing values. Multiple imputation is considered when missingness is not completely random, depending on the observed or unobserved values. However, this method is applicable even if the pattern of missing data is not random.

First, we perform missing value processing on the dataset to detect whether there are missing values in the data. The missing type analysis showed that there were 11,288 missing values, of which 12 cases were missing for education (DMDEDUC), with a missing rate of 0.2%; 273 cases were missing for marital status (DMDMARTL), with a missing rate of 4.9%; 734 cases were missing for poverty index ratio (INDFMPIR), with a missing rate of 13.3%; 52 cases were missing for body mass index (BMI), with a missing rate of 4.1%; 3,300 cases were missing for smoking status (ISSMOKE), with a missing rate of 59.6%; 406 cases were missing for drinking status (ALCOHOL), with a missing rate of 7.3%; 146 cases were missing for albumin-creatinine ratio (URDACT), with a missing rate of 2.6%; 272 cases were missing for glycohemoglobin (LBXGH), with a missing rate of 4.9%; 3040 cases were missing for triglycerides (LBDTRSI), with a missing rate of 54.9%; 3047 cases were missing for low-density lipoprotein cholesterol (LBDLDNSI), with a missing rate of 55.1%.

In order to observe the missing data of each variable more intuitively, we used the VIM package in R studio to draw a visualization of missing data. The left side is a histogram of the missing data ratio. It can be seen that the most missing values are in the attraction condition, followed by triglycerides and low-density lipoprotein cholesterol; the right side is a missing data pattern diagram, blue represents missing values, and red represents non-missing values. The specific visualization is shown in Figure 2:

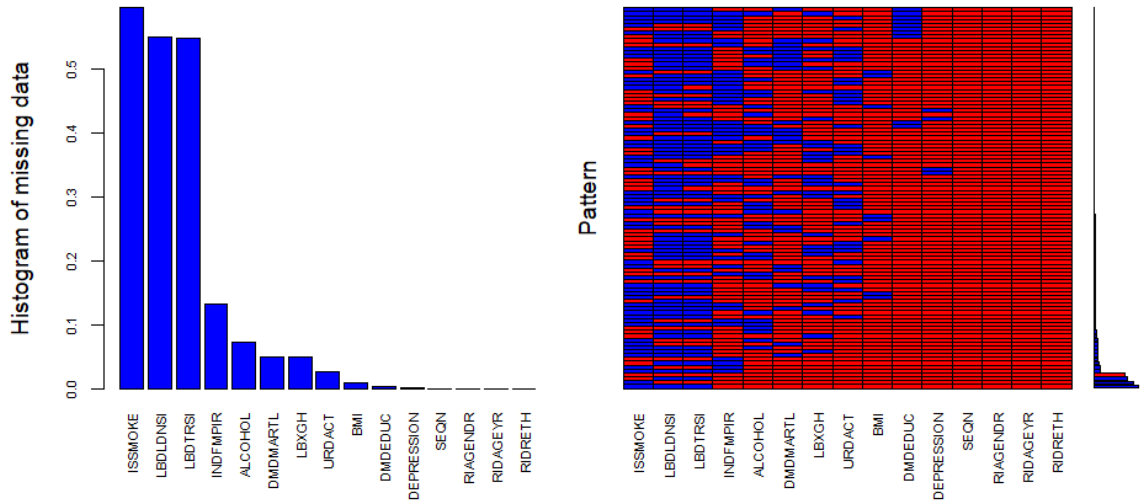

Figure 2. Missing data visualization

For missing data, this paper adopts the random forest method in the multiple imputation method to interpolate the missing data 5 times, and finally selects the set of data with the smallest AIC. Next, we check the results after interpolation. We use R studio to draw a density map to view the distribution of the interpolated data set and the observed data. As shown in Figure 3, the interpolation result is similar to the distribution type of the original data, so the interpolated data is of good quality and can be used for subsequent analysis.

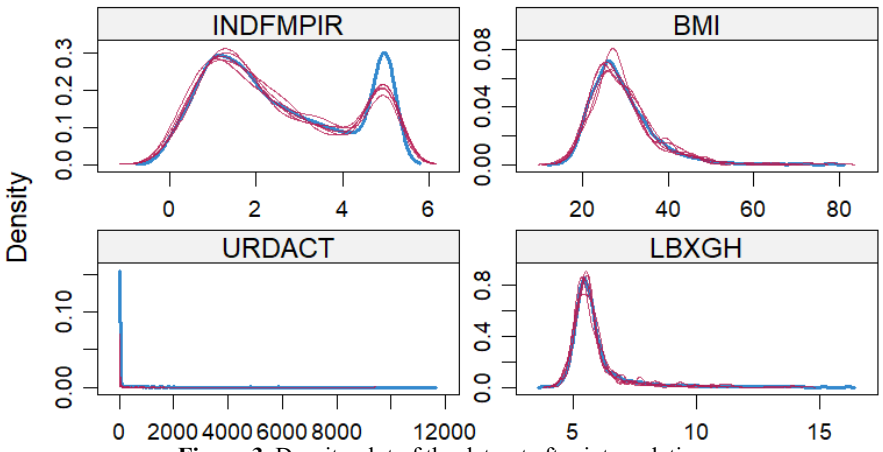

Figure 3. Density plot of the dataset after interpolation

3.3. Data Imbalance Handling

Next, we will deal with the problem of data imbalance. The problem of data imbalance is very common in data mining, and is often seen in anti-fraud detection, medical diagnosis, oil spill detection, facial recognition, outlier detection, etc. This is caused by the skewed nature of the data, and these problems will affect the process of machine learning in the classification process. In such a problem, the classes have different proportions of samples, where a large number of samples belong to one class and a smaller number of samples in the other class, which is usually the basic class, but unfortunately is misclassified by many classifiers. To date, a large number of studies have been conducted to implement different technologies and methods to solve the problem of unbalanced data [17]. The most commonly used methods are undersampling and oversampling. As shown in Figure 6, the number of non-depressed patients is much less than that of depressed patients, and the proportion of the minority class is 8.34%, which is a moderately unbalanced sample. Therefore, this paper uses the SMOTE oversampling method for processing.

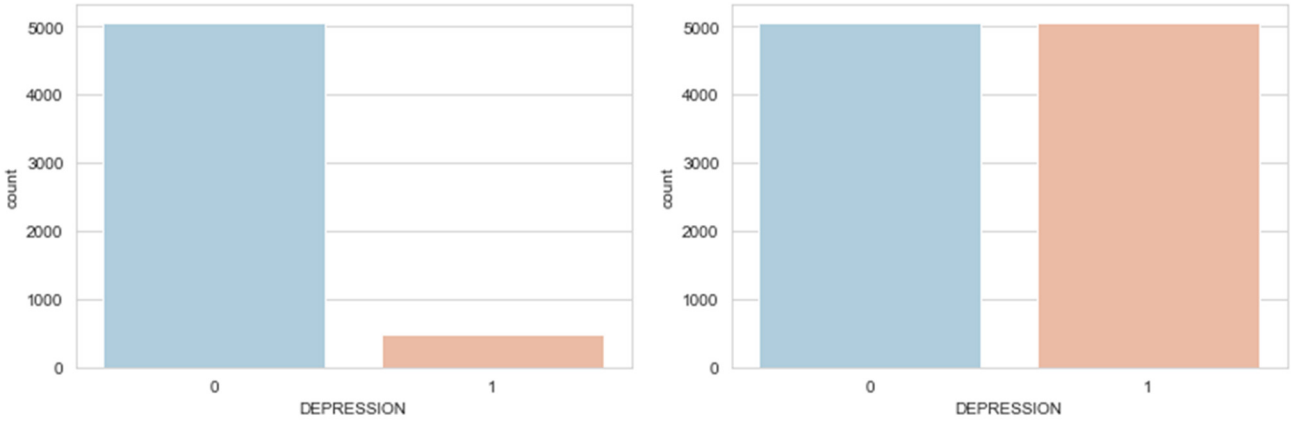

Figure 4. Bar graph of people suffering from depression

3.4. Data Standardization

After exploring the dataset, this article found that it is necessary to convert some categorical variables into dummy variables and scale all values before training the machine learning model. First, we used the get_dummies method in python to create dummy columns for categorical variables. Then, in order to eliminate the differences in dimensions and value ranges between features, the data needs to be standardized. Data normalization is to rescale the sample data according to a certain calculation method so that the value changes in a certain range. Data normalization can speed up the algorithm to find the optimal value. We use the StandardScaler method in sklearn.procession to process the data and standardize the data to fall in the [-1,1] interval.

4. Construction of Depression Prediction Model

According to the existing literature, machine learning algorithms are used to predict the risk of depression, which mainly include five models: logistic regression, decision tree, random forest, support vector machine and CatBoost. For depression data set, a machine learning model is used to predict whether people suffer from depression. The specific experimental process is shown in Figure 5.Before building the model, the data set needs to be divided. After preprocessing, 70% of the data is randomly selected as the training set, and 30% of the modeling data set is used as the test set.

4.1. Data Modeling and Forecasting

After the previous data preprocessing, our data already conforms to the input of the model algorithm. Next, we build models and test efficiencies on the basis of these data. In this paper, five machine learning models, logistic regression, decision tree, support vector machine, random forest and Catboost classification model, are used for prediction.

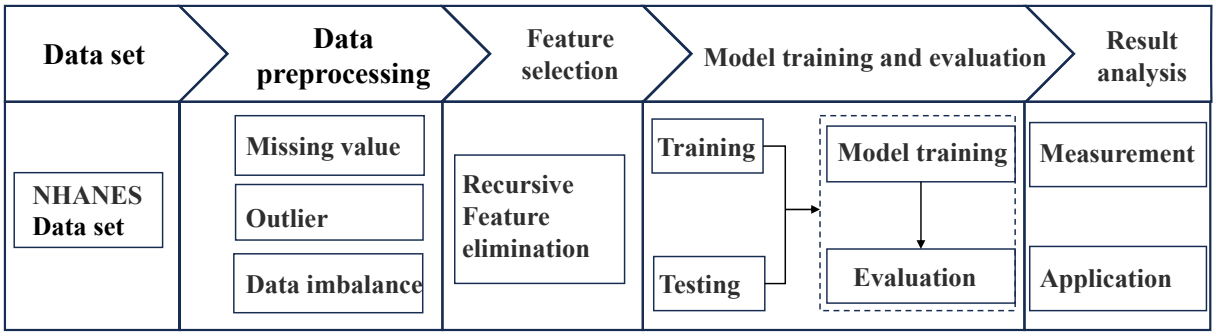

Figure 5. Flow chart of depression prediction model based on machine learning

After the construction of the model, Accuracy rate, accuracy rate, recall rate, F1-Score and AUC indicators are usually used to evaluate the prediction results of various machine learning algorithms for depression recognition. Accuracy is used to measure the prediction effect of a model and explain whether the model achieves the best prediction effect. By comparing the calibration accuracy of the model output results and combining the output results of the model algorithm fit, we found that among the five models, random forest and decision tree had the best performance, and the accuracy rate of detecting depression was 100%. This was followed by the CatBoost model, which detected depression with 96.32% accuracy. The accuracy of CatBoost model is the highest, 91.67%, followed by random forest, 88.84%. The predictive ROC curves of these five algorithms are shown in Figure 6.

Table 2. Evaluation of depression recognition and prediction results by various algorithms

Data	Index	Model				
		Logistic regression	Decision tree	Support vector machine	Random forest	Catboost
Training	ACC (%)	72.05	100	80.48	100	96.32
	AUC	0.788	1.000	0.885	1.000	0.993
	Precision	0.722	1.000	0.806	1.000	0.964
	Recall	0.721	1.000	0.805	1.000	0.963
	F1-score	0.720	1.000	0.805	1.000	0.963
Testing	ACC (%)	72.21	81.3	78.53	88.84	91.67
	AUC	0.782	0.813	0.856	0.953	0.967
	Precision	0.722	0.814	0.786	0.889	0.919
	Recall	0.722	0.813	0.785	0.888	0.917
	F1-score	0.722	0.813	0.785	0.888	0.917

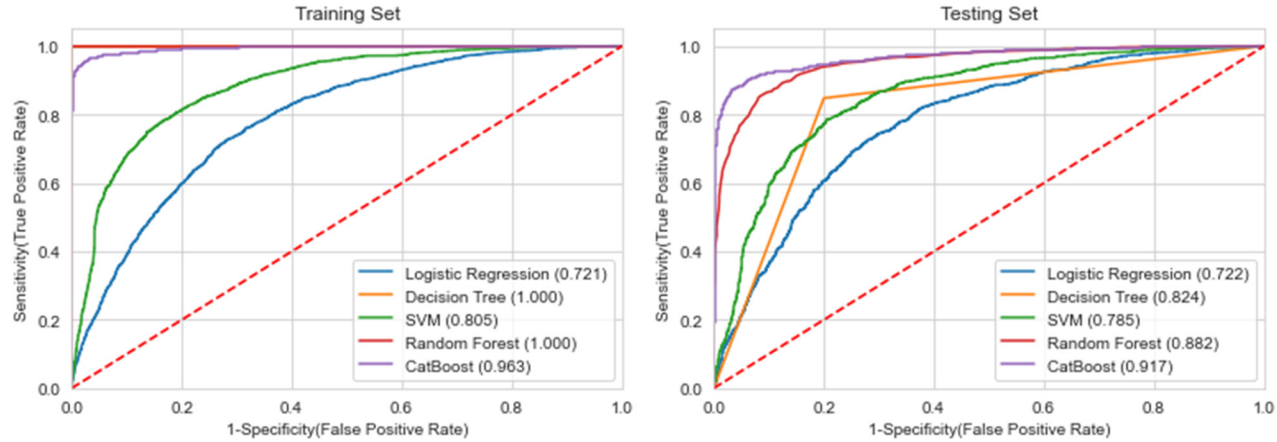

Figure 6. ROC curves of subjects of five models

5. Conclusion

5.1. Summary

In this study, the test set AUCs of the four machine learning models were all above 0.7, among which the CatBoost model had the highest accuracy. This result shows that the CatBoost algorithm can well identify the occurrence of depression from other health-related and demographic factors in large survey data sets. Unlike traditional statistical methods that rely on researcher input to specify variables relevant to a particular analysis, machine learning methods can identify which variables in a given data set are associated or unrelated to the outcome of interest. This is also the advantage of machine learning in building clinical prediction models. The possible applicability of machine learning models in clinical practice can be a web-based tool for assessing participants' risk of depression. Although it cannot replace some traditional screening tools (such as PHQ-9), we can use it to estimate the prevalence of depression in areas where there are no personal mental health surveys.

5.2. Strengths and Limitations

The data mining method of splitting the data file into training and validation minimizes the problem of overfitting, which is often problematic in traditional statistical techniques with a large number of predictors. Enhanced machine learning techniques can adapt to different types of variables and have been found to have high predictive accuracy. Shrinkage is also used to avoid overfitting.

At the same time, this study also has several limitations. First, this study treated depression as a binary variable. Therefore, we cannot evaluate the correlation between various factors and the severity of depression. Second, the prevalence of depression in 2018 may be underestimated due to missing values. Third, because NHANES is a cross-sectional survey (rather than a longitudinal survey), we cannot measure the prognosis of the disease or the future occurrence of depression in the population.

5.3. Directions for Further Work

Although the research of the paper has been completed and achieved the expected goals and initial success, there is still room for improvement and further improvement. Here are a few key points for brief discussion:

1. There are many algorithms in machine learning. This paper only uses a few of them. In the future research process, we will continue to study other model algorithms, summarize their respective characteristics and applicability, and integrate them and apply them to the prediction system.

2. The structure of the prediction system is still relatively simple. The data set used by the system is too structured and has certain limitations. We will continue to improve and enrich the system structure in the future, and try to use real-time data for more in-depth prediction research.

3. The data processing in this paper uses the jupyter scientific computing tool class. The processing process is completed by writing program algorithms. Some are too complicated and take up a lot of time. We will continue to strengthen the research and study of dedicated data mining tools in the future.

References

[1] K. Smith. Mental health: a world of depression[J]. Nature. 2014,515(7526): 181.

[2] James S L, Abate D, Abate K H, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017[J]. The Lancet, 2018, 392 (10159): 1789-1858.

[3] World Health Organization (WHO). Depression.https:// www.who.int/news-room/fact-sheets/detail/depression, 2022-09-01.

[4] Avenevoli S, Swendsen J, He J P, et al. Major depression in the national comorbidity survey–adolescent supplement: Prevalence, correlates, and treatment[J]. Journal of the American Academy of Child & Adolescent Psychiatry, 2015, 54 (1): 37-44.

[5] Center for Behavioral Health Statistics and Quality,.2016 National Survey on Drug use and Health: Detailed Tables. Substance Abuse and Mental Health Services Administration, Rockville, MD. 2017.

[6] A. Singh-Manoux, A. Dugravot, A. Fournier, J. Abell, et al. Trajectories of depressive symptoms before diagnosis of dementia: a 28-year follow-up study[J]. JAMA Psychiatry, 2017, 74 (7):712-718.

[7] Weissman M M, Bland R C, Canino G J, et al. Cross-national epidemiology of major depression and bipolar disorder[J]. Jama, 1996, 276(4): 293-299.

[8] Robbins R, Weaver M D, Czeisler M É, et al. Associations between changes in daily behaviors and self-reported feelings of depression and anxiety about the COVID-19 pandemic among older adults[J]. The Journals of Gerontology: Series B, 2022, 77(7): e150-e159.

[9] Creese B, Khan Z, Henley W, Corbett A, Vasconcelos Da Silva M, Mills K, et al. Loneliness, physical activity and mental health during Covid-19: a longitudinal analysis of depression and anxiety between 2015 and 2020. Int Psychogeriatr. 2021; 33:505–14.

[10] Dipnall J F, Pasco J A, Berk M, et al. Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression[J]. PloS one, 2016,11(2): e0148195.

[11] Iranpour S, Sabour S, Koohi F, et al. The trend and pattern of depression prevalence in the US: Data from National Health and Nutrition Examination Survey (NHANES) 2005 to 2016[J]. Journal of affective disorders, 2022,298: 508-515.

[12] Wani M A, ELAffendi M A, Shakil K A, et al. Depression screening in Humans with AIand deep learning techniques[J]. IEEE Transactions on Computational Social Systems, 2022.

[13] Zhang C, Chen X, Wang S, et al. Using CatBoost algorithm to identify middle-aged and elderly depression, national health and nutrition examination survey 2011–2018[J]. Psychiatry Research, 2021, 306: 114261.

[14] NHANES--About the National Health and Nutrition Examination Survey. Available online:https:// www.cdc. gov/nchs/nhanes/about_nhanes.htm (accessed on 28 December 2021).

[15] Reeder N, Tolar-Peterson T, Bailey R H, et al. Food insecurity and depression among USadults: NHANES 2005–2016[J]. Nutrients, 2022, 14(15): 3081.

[16] Heymans M W, Twisk J W R. Handling missing data in clinical research[J]. Journal of Clinical Epidemiology, 2022, 151: 185-188.

[17] Ali H, Salleh M N M, Saedudin R, et al. Imbalance class problems in data mining: A review[J]. Indonesian Journal of Electrical Engineering and Computer Science, 2019, 14(3): 1560-1571.