Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230681

Application of machine learning on depression prediction and analysis

Hongbin Lyu

New Channel-Shenzhen Bona School, Shenzhen, Guangdong, China

hl13g20@soton.ac.uk

Abstract. In this paper, machine learning methods based on Python are applied to predict and analysis the possibility of depression, which provide the possibility whether the individuals suffer depression. Four kinds of machine learning methods, including, supported vector machine, naïve Bayes, random forest and neural network, that make predictions are all based on 23 types of features from the public depression dataset and after processing the data, some relevant figures have also been drawn. The results show that the random forest obtains the best performance among all the methods. Random forest achieves 86.8% accuracy, which validates the effectiveness of machine learning based depression prediction. Furthermore, the important factors which lead to depression are also analysed. Among the predictions of the depression dataset, these figures show that the most influential features are the age, level of education, living expenses and the gender does not have any effect.

Keywords: Machine Learning, Data Analysis, Depression.

1. Introduction

With the emergence of deep learning-based methods [1,2,3,4], several methods with innovative underpinnings have been developed, such as AlexNet [5], VGG [6], ResNet [7], DenseNet [8] and inception [9] network. Depression is one of the most common psychological disorders nowadays, characterised by continuous and prolonged depressive moods as the main clinical feature. It is the most significant type of psychological disorder in nowadays, and it is sometimes difficult to be diagnosed. This project based on machine learning and features of depression dataset to predict the possibility of depression. The depression dataset is consisting as a study about the life conditions of people who live in rurales zone, it includes 23 kinds of features and a total of 1432 rows of objectives. After exploring the data, some features values of the data are standardized and four machine learning models are constructed and trained to predict the results. The comparison results of machine learning models indicate the most influential features of the depression dataset. There have been a variety of studies on the prediction and classification of depression based on machine learning. Shuang Gao et al, used popular machine learning methods for brain imaging classification and prediction in their study species and outlined studies specifically on MDD (Major Depression Disorder) [10]. Md. Sabab Zulfiker et al. investigated six different machine learning classifiers in their study, utilizing socio-demographic and socio-psychological data to determine if a person is depressed. In addition, three distinct feature selection techniques were employed to extract the most pertinent characteristics from the dataset, as well as several model assessment criteria, such as accuracy. F1 scores etc. have also been calculated. Of these,

DOI: 10.54254/2755-2721/5/20230681

the AdaBoost classifier outperformed all other methods with an accuracy of 92.56%[11] .This paper will able to analyse the dataset to identify groups that are vulnerable to depression and to increase the rate of diagnosis of depression to a meaningful extent.

2. Method

2.1. Support Vector Machine (SVM)

Emergence of deep learning, SVM had recognized as an effective and successful deep learning algorithm of the preceding decade or so. Since SVM is a linear classifier, the items to be classified has to be linearly distinct. The objective of classification learning in machine learning is a sequence of sample feature data. It is not feasible to determine a linear segmentation line or segmentation plane and execute SVM classification unless the sample data is linearly separable. SVM classification is possible if sampling up with the digital are linearly indistinguishable and there is no linear segment line or segment plane. SVM seeks to identify the optimal hyperplane for feature space modification, as well as the purpose of the SVM method is to optimize classification margins. In addition, SVM has a solid theoretical base, is mostly independent of probability, and produces efficient classification, hence considerably simplifying the typical classification and regression tasks. However, it is challenging to develop SVM classifiers for huge training samples. In most data mining applications, classification issues involving several classes must be handled, however SVM classifiers cannot directly categorize multiple features.

2.2. Naive Bayes

A Naive Bayes algorithm is a supervised learning technique that applies Bayes' theorem to classification problems. It is mostly utilized for text categorization, which requires massive training set. The Nave Bayes Classifier is among the simplest and most effective classification algorithms, aiding in the creation of machine learning with the ability to make quick predictions. It is a classifier, so means it makes predictions due to the possibility of categorizing an item. However, since Naive Bayes assumes that almost all features are independently or uncorrelated, it is incapable of learning feature connections.This is the formula for Bayes' theorem:

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Posterior probability, often known as P(A|B), is the possibility that the hypothesis A is supported by the event that was seen as B.

The likelihood probability, denoted by the symbol P(B|A), is the chance that the evidence suggests that the likelihood of a hypothesis is accurate.

Prior Probability, abbreviated as P(A), refers to the likelihood of the hypothesis being true before being exposed to any supporting evidence.

Marginal Probability, often known as the probability of evidence, is represented by the symbol P(B).

2.3. Random Forest

As shown in Figure 1, random forest is a supervised learning strategy for classification or regression algorithms in machine learning. It is a classifier that combines the outcomes of numerous decision trees performed to varied subsets of a dataset in terms of the forecasting accuracy of the dataset.

Inside a random forest, which contains of several decision trees, there is no correlation between unique decision trees. When conducting a classification task, every decision tree in the forest classifies newly presented input samples independently. So, every decision tree will acquire its own classifier, with the random forest selecting the classifier with the largest number of classifications. Random forest is able to analyze high-dimensional or feature-rich data without the need for dimension reduction and feature selection. Furthermore example, random forest is very simple to construct and can compensate for errors in unequal data sets. Accuracy can be maintained despite the absence of a considerable number of features. As a result, the property weights provided by the random forest for such as data are unreliable.

DOI: 10.54254/2755-2721/5/20230681

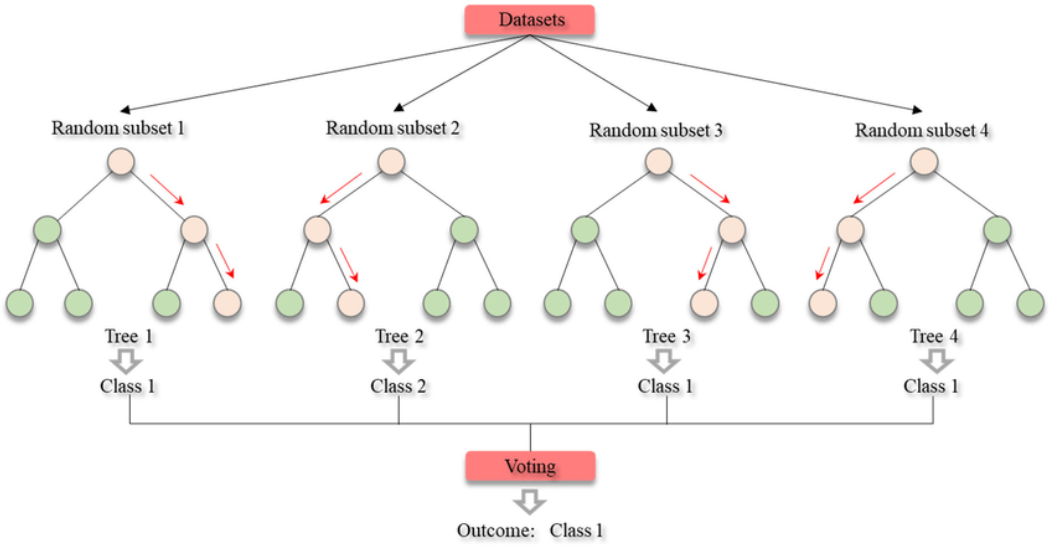

Figure 1. Pipeline of random forest.

2.4. Neural Network

As seen in Figure 2, a neural network is composed consisted of a hidden layer, an input layer, and an output layer that seeks to simulate the activity of the human brain. The hidden layers may be thought of as a representation of the learned input data. These layers help the neural network understand the incoming data's characteristics. Neural networks are interconnected networks of simple, flexible units that imitate the organic nervous system's reactions to inputs. The neuron is the most fundamental element of a neural network, and its design is fully inspired by the information propagation process of neurons in nature. In 1943, the psychologist McCulloch and the mathematician Pitts, inspired by biological neurons, created the notion of artificial neurons, sometimes known as perception machines.

A general neural network contains the following parts, the input layer, the hidden layer and the output layer. The more layers there are and the greater the number of hidden layer nodes, the deeper the features can be learned by the neural network under a non-linear activation function. All model variables are inputs. The goal of the neural network determines the number of interconnected hidden layers. Hidden layers are functions of input variables' weights. Each hidden unit in a network with numerous hidden layers is a function of the preceding layer's weighted sum. Hidden layers' output layer contains target (dependent) variables. The output layer divides input data into numerous nodes for most image classification tasks.

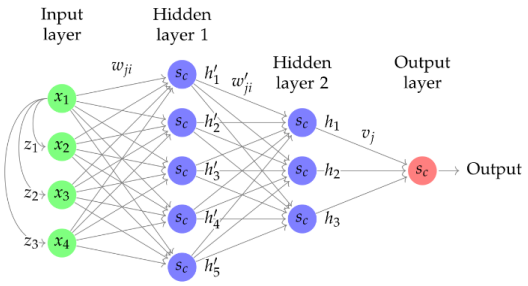

Figure 2. Pipeline of convolution neural network.

With artificial neural networks, the activation function is what maps inputs to outputs by making the neurons in the network nonlinear. In a deep neural network with many hidden layers, the activation function connects the weighted sum of the units in one layer to the values of the units in subsequent

DOI: 10.54254/2755-2721/5/20230681

levels. Each of the obfuscated levels uses the same activation method. The activation function is often a nonlinear one. Each feature's (input variable's) relative importance in making an accurate prediction is reflected in the weights. The deviations also make the networks more selective and the model more flexible, and they help to explain the connection between the feature and the intended output.

2.5. Feature Selection

A standard machine learning problem is predicting the value corresponding to a data sample based on the sample's properties. If the sample contains too few characteristics, new features are evaluated; however, in fact, there are frequently too many features, necessitating a suitable reduction. Typically, the final dataset contains noise and other unnecessary characteristics that not only do not assist us in achieving our goals, but may even impede the selection of our model. Reducing features is of great practical importance, not only to reduce overfitting, to reduce the number of features (dimensionality reduction), and to increase the model's generalization, as well as to enhance the model's interpretability, to strengthen the relationship between features and eigenvalues, to accelerate the training of the model, and, overall, to achieve better performance. Before model selection and training, it is necessary that the data be properly organized and sorted by task.

In this dataset, there are many large values and 20 missing values. After cleaning and normalizing the data and filtering out irrelevant features, the Figure 3 is listed:

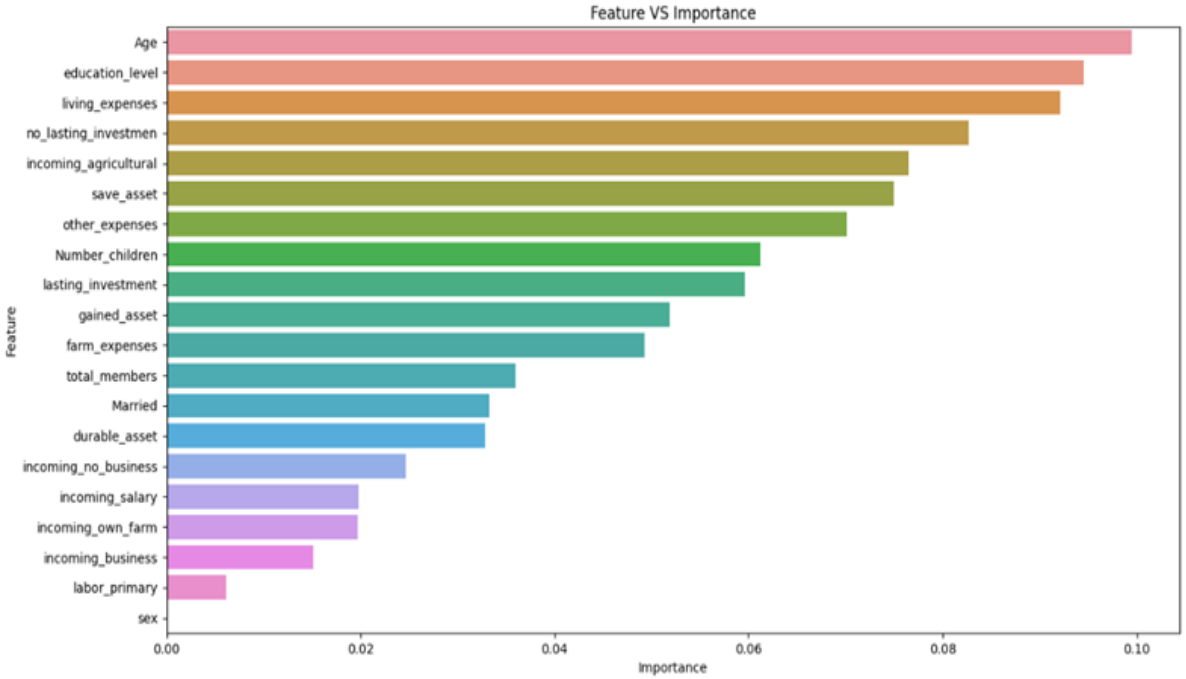

Figure 3. Illustration of feature importance.

Table 1. Top-5 feature importance in the dataset.

Feature	Importance
Age	0.099430
education level	0.094603
Living expense	0.090053
No lasting investment	0.082716
Incoming agricultural	0.076744

DOI: 10.54254/2755-2721/5/20230681

Table 1 shows the 20 relevant features after processing and their influence on the prediction results (Depressed).

3. Results

3.1. Dataset Description

The dataset concerns the analysis of depression and it is composed as a study on the living conditions of people living in rural areas. The dataset contains 23 features, 21 of which are integers and 2 of which are Survey_id and Ville_id. A total of 1432 rows of objectives. Table 2 indicates all the features in the dataset.

Table 2. Feature description.

Feature item	Mean	Std. Dev.
Survey_id	**715**	**413**
Ville_id	**76.3**	**66.4**
sex	**0.92**	**0.27**
Age	**34.8**	**14**
Married	**0.77**	**0.42**
Number_children	**2.88**	**1.87**
education_level	**8.69**	**2.92**
total_memberd	**4.97**	**1.79**
gained_asset	**33.6m**	**20m**
durable_asset	**27.2m**	**18.2m**
save_asset	**27.4m**	**17.7m**
living_expenses	**32.5m**	**21m**
other_expenses	**33.7m**	**21.7m**
incoming_salary	**0.18**	**0.38**
incoming_own_farm	**0.25**	**0.43**
incoming_business	**0.11**	**0.31**
incoming_no_business	**0.26**	**0.44**
incoming_agricultural	**34.5m**	**20.8m**
farm_expenses	**35.5m**	**21.1m**
labor_primary	**0.21**	**0.41**
lasting_investment	**33m**	**21.2m**
no_lasting_investmen	**33.6m**	**21.6m**
depressed	**0.17**	**0.37**

DOI: 10.54254/2755-2721/5/20230681

3.2. Comparison Results.

3.2.1. Confusion Matrix

Through machine learning, the Confusion Matrix might be called the error matrix or the probability matrix. When it comes to supervised learning, confusion matrices are often called "matching matrices," while when it comes to unsupervised learning, they are usually called "matching matrices." Most of the time, it is often used to make comparisons classifier performance with actual measured values to figure out how accurate an image is. The accuracy of the classification results can be shown in a confusion matrix.

Figure 4 are the confusion matrix of SVM, Naïve Bayes, and Random Forest respectively:

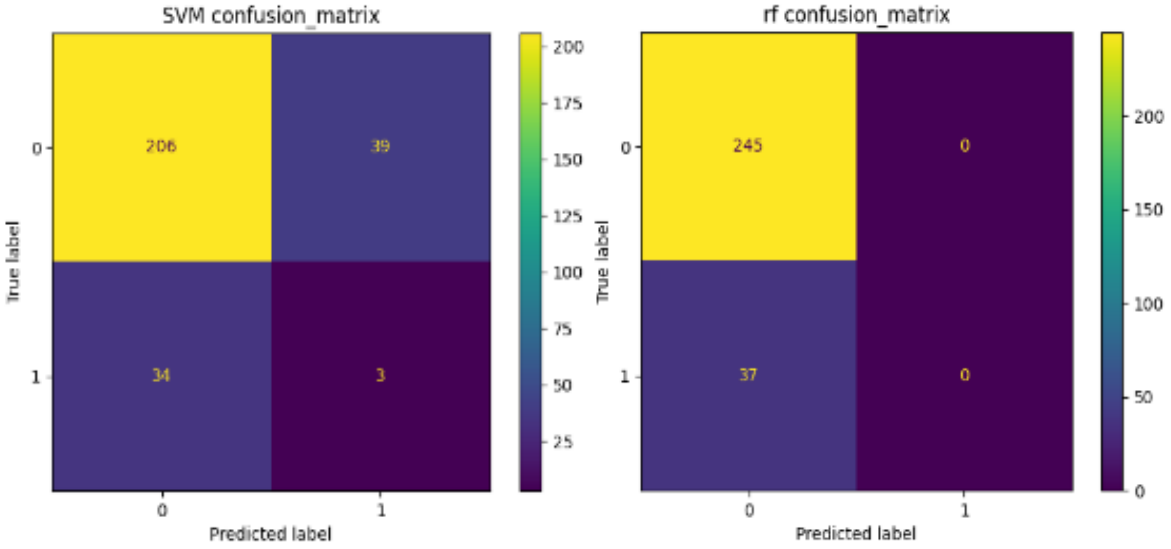

Figure 4. Confusion matrix.

The confusion matrix is organized so that each column represents a different anticipated category, and the sum of the values in each column represents the expected proportion of the data that falls into that group. Each row represents a different real data attribution category, with the overall quantity of data corresponding to the predicted number of occurrences in that category, and each column corresponding to the actual number of occurrences in that category.

Four Categories:

TP, or True Positive, means genuine excellence. The model's predicted class matches the true class of the sample.

Class for false negatives; abbreviated FN. The true sample class is positive, but the model assigns it to the opposite category.

In other words, it was a TP, or a true positive. Even when the true category of the sample is negative, the model interprets it as positive.

Class for true negatives; abbreviated 'TN'. The model has correctly determined that the true class of the sample is negative.

3.2.2. Receiving Operating Characteristic (ROC) Curve

Curve of the receiver's operating characteristics, commonly known as the sensitivity curve. The ROC curve is used to analyse the ability of the binary classifier, with TPR as the x-axis and FPR as the y-axis, coordinate points are obtained at different thresholds and the individual coordinate points are connected to obtain the ROC curve.

DOI: 10.54254/2755-2721/5/20230681

Formula of TPR & FPR:

$$\mathrm{TPR} = \frac{\mathrm{TruePositives}}{(\mathrm{TruePositives}+\mathrm{FalseNegatives})} \tag{2}$$

TPR is the proportion of expected positive samples that are really positive compared to the total set of positive samples. (Examples with positive actual results).

$$\mathrm{FPR} = \frac{\mathrm{FalsePositives}}{(\mathrm{FalsePositives}+\mathrm{TrueNegatives})} \tag{3}$$

FPR is the proportion of samples that were predicted to be positive but turned out to be negative relative to the total number of negative samples. (the actual result is negative) Figure 5 depicts the ROC curves for SVM, Nave Bayes, and Random Forest.

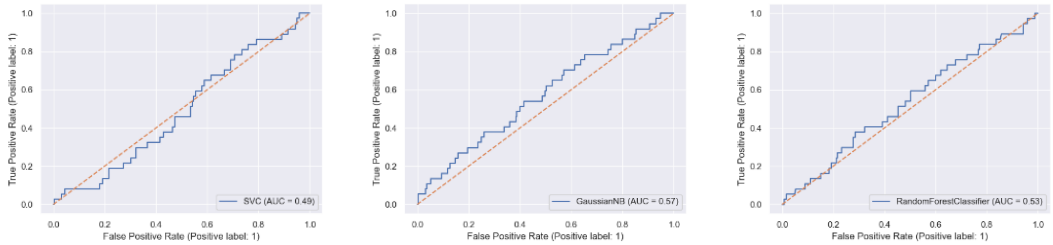

Figure 5. ROC curves of those three methods.

3.2.3. F1 Score

The F1 Score is a statistical measure of a binary classification model's precision. The F1-score average of the model's accuracy and recall, ranging from 1 to 0. The F1 score measures the classification model's precision and recall. Before and after feature selection, the results are shown in Tables 3 and 4.

$$\mathrm{F1-Score} = 2 * \frac{\mathrm{Precision}*\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{4}$$

Table 3. Performance before feature selection.

Method	Accuracy	Recall	F1 Score
SVM	0.7411347517730497	0.08108108108108109	0.0759493670886076
Naïve Bayes	0.7624113475177305	0.13513513513513514	0.12987012987012989
Random Forest	0.8687943262411347	0.00	0.00
Neural Network	0.7872340425531915	0.16216216216216217	0.16666666666666669

Table 4. Performance after feature selection.

Method	Accuracy	Recall	F1 Score
SVM	0.6879432624113475	0.40540540540540543	0.2542372881355932
Naïve Bayes	0.851063829787234	0.02702702702702703	0.045454545454545456
Random Forest	0.8687943262411347	0.00	0.00
Neural Network	0.7907801418439716	0.13513513513513514	0.14492753623188406

4. Conclusion

In this paper, the task of depression prediction is handled by machine learning based methods and exploit the effectiveness of feature selection on it. First, four different methods, including support vector machine, random forest, naïve Bayes and CNN are utilized to predict the individuals whether they suffer the depression or not. A public dataset is introduced to evaluate those methods. Then, the author selects the most important features based on feature correlation. Feature selection methods can compress the

DOI: 10.54254/2755-2721/5/20230681

feature dimensions and remove the data noise, thus achieving a satisfying performance. In the experiment, random forest is the best method according to the accuracy, which achieves 86.8% accuracy. After applying feature selection, most of the methods has an obvious promotion, which can validate the effectiveness of feature selection. The Random Forest classifier still had the highest accuracy, although it had the lowest Recall and F1 Score of the four models. The SVM classifier's accuracy decreased after feature selection, but its Recall and F1 Score improved significantly. Finally, the conclusion can be described that the analysis of the dataset revealed that the 10 most influential features were age, education level, living expense, no lasting investment, incoming agriculture, saved asset, other expenses, number of children, lasting investment and gained asset, while gender had a negligible effect on the results.

References

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
[2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.
[3] Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., ... & Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR), 51(5), 1-36.
[4] Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. IEEE Access, 7, 53040-53065.
[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.
[6] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778)
[8] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).
[9] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).
[10] Gao, S., Calhoun, V.D. and Sui, J., 2018. Machine learning in major depression: From classification to treatment outcome prediction. CNS neuroscience & therapeutics, 24(11), pp.1037-1052.
[11] Zulfiker, M. S., Kabir, N., Biswas, A. A., Nazneen, T., & Uddin, M. S. (2021). An in-depth analysis of machine learning approaches to predict depression. Current research in behavioral sciences, 2, 100044.