International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8076
***Article Received:** 24 June 2023 **Revised:** 18 August 2023 **Accepted:** 02 September 2023*

An Open-Source Platform for Real-Time Preliminary Diagnosis amongst Adults using Data Analytics

Kalpana Sharma[1], Sital Sharma[2], Sunil Dhimal[3], Biswaraj Sen[4], Ashis Pradhan[5]*, Vikash Kumar Singh[6]
[1]Computer Science and Engineering
Sikkim Manipal Institute of Technology,SMU
Majhirar East Sikkim, India
kalpana.s@smit.smu.edu.in
[2]Computer Science and Engineering
Sikkim Manipal Institute of Technology,SMU
Majhirar East Sikkim, India
sitalneo@gmail.com
sital.s@smit.smu.edu.in
[3]Computer Science and Engineering
Sikkim Manipal Institute of Technology,SMU
Majhirar East Sikkim, India
sunildhimal@gmail.com
[4]Computer Science and Engineering
Sikkim Manipal Institute of Technology,SMU
Majhirar East Sikkim, India
biswaraj.s@smit.smu.edu.in
[5]Computer Science and Engineering
Sikkim Manipal Institute of Technology,SMU
Majhirar East Sikkim, India
*corresponding author
ashis.p@smit.smu.edu.in
[6]Computer Science and Engineering
Sikkim Manipal Institute of Technology,SMU
Majhirar East Sikkim, India
Vikash.s@smit.smu.edu.in

Abstract— Depression can be defined as a mental health disorder characterized by persistently depressed mood, loss of interest in activities, causing significant impairment in daily life. Technical intervention to screen depression in non-clinical population which records, classify depression on the basis of severity and provide features or predictors that discriminate the classification of depression among non-clinical population comprising of college students is the main area of the study. Beck Depression Inventory – II (BDI-II), as per Diagnostic and Statistical manual of Mental disorder (DSM IV) is used to screen depression and its severity. Indicators are determined on the basis of how well the features or predictors can discriminate the classes of depression severity. Providing quality indicators which help in supporting the process can be considered as symptoms for screening depression. Descriptive analytics is used in order to find the underlying pattern of the responses captured, factor analysis groups variables on the basis of correlation between patterns of the responses to reduce dimension. The approach for supervised descriptive analysis method that takes BDI-II questions as features and refine the features using information gain and linear discriminant analysis as feature selection algorithm. The classification of severity of depression is done using Support vector machine (SVM)..

Keywords- Data analytics, Depression, Machine learning, Healthcare, Mental Health, Diagnosis of Depression.

I. INTRODUCTION

In the field of psychology and psychiatry, depression refers to sadness and other related emotion and behaviours (Khodayari-Rostamabad, A. et al 2010). Technical intervention to screen and analyse depression symptoms in a given set of population can give us the information about the population and their struggle with depression. Depression being a subjective mental disorder, whose effects and responses changes as per the environment,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8076
***Article Received:** 24 June 2023 **Revised:** 18 August 2023 **Accepted:** 02 September 2023*

location, ethnicity, age, gender etc. (Mair C. et al. 2010), is difficult to generalize as it's effect can be different and varies from being minimal to severe. Technical intervention to screen depression in a non-clinical population has not been well explored as depression may exist in non-clinical population who are not aware of the disorder or in denial. Technical intervention to screen depression among non-clinical population is important as it would improve diagnosis of depression as screening comprises of quality indicators (Firth, J. et al. 2017). There isn't any prevalent lab test to screen depression. Questionnaires can be evaluated to screen depression on the basis of the questionnaire's relevance to the severity of the depression. Clinical data for confirming the depression is not available and difficult to get. There are number of questionnaires that are used for screening depression. Beck Depression Inventory–II (BDI-II) is one of the questionnaires that are used for screening subjects with depression (Dozois, D. J. et al. 1998; Richter P. et al. 1998; Storch, E. A. et al. 2004).

BDI-II is used as a questionnaire to conduct the analysis as the questionnaire fits well with age group of 13 and above. BDI-II is a world-wide used self-rating scale for measuring depression. It consists of 21 questions each of which is scored from 0 to 3 in terms of intensity (Kendall, P. C., et al. 1987). Analysing non-clinical population on which screening of depression can give an idea about how the population under consideration respond to the survey and find underlying patterns that exists. Providing quality indicators for screening of depression classification of BDI-II responses can be used to provide an important feature that contributes to screen depression. BDI-II has been used in non-clinical patients with good classification (Osman A. et al. 2008). BDI-II is also used in Sikkim, India and is used for initial screening of depression among the patients.

Analysis using machine learning has been an area of research for diagnosis of mental disorders which diagnose a disorder with considerable level of accuracy. Studies have been made for improving the classification of depression by adding dimensionality to the classification than by just having pure categorical classification (Stengel, E. 1959; Beck A. T. et al 1961). There are researches conducted for classification of disorders like Major Depressive Disorder, Bipolar Disorder and Schizophrenia and Normal subjects using the data collected from electroencephalogram signals, classification of depression using EEG signals has also been proposed (Hosseinifard B. et al. 2013) using non-linear features from EEG signals. Classification of the disorder is based on maximum likelihood principle. Machine learning algorithms have been used for classification of depressive disorders using ensemble techniques [Ojeme B. et al. 2015) which comprises of algorithms like multi-layer perceptron, KNN and SVM and the accuracy was measured using confusion matrix (Deng X. et al. 2016); Ojeme B. et al. 2016).

Detection of depression from a given set of question involves those features that contribute the most for detection of depression, these features need to be identified which discriminates the class of depression. There is on-going research which may help in improving the classification of depression for screening of depression which are quality indicators that helps in improving the diagnosis process for depression, there are research that tries to find out the underlying pattern of the given responses by performing factor analysis that gives latent factors or variables that describes the response by grouping the features on the accounted variance.

II. MATERIALS AND METHODOLOGY

The following are the steps performed for the analysis of BDI-II responses to determine important FEATURES that are indicators to classify depression severity and underlying pattern existing among the non-clinical college students.

A. Strategy Used

1) Data Collection: Collect BDI -II responses using a platform designed using open-source tools available on the web (mobile compatible) at the link "http://dalabsmit.in".

2) Data Preprocessing: Check the data collected for any missing values, delete and process the response if it has missing data or inappropriate data. This step helps in cleaning missing/inappropriate data so that further analysis and processing is void of data validation error.

3) Outlier Detection: Screen the data for outliers for the response collected using the Mahalanobis distance, since the response comprises maximum of minimal depression data the outliers detected in this scenario are generally a part of the solution as it gives information, so it is not removed. The outliers are detected by calculating a cut-off score using one sample t-test to generate the p-value. This step helps in conforming data with the general population by dealing with outliers thereby helping in building a general trend in the depression related responses.

4) Data Classification: Classification models like Naive Bayes, Support Vector Machine, K-Nearest Neighbour (KNN) and Decision Tree are selected for classifying BDI responses. Out of the mentioned classification models, SVM outperforms other classification algorithm for BDI-II responses.

5) Feature Selection: Feature Selection is performed using the filter method approach. Information gain theory is used for selecting features as it gives subset of the features that are more relevant with the class using the concept on entropy [Peng, H. et al. 2005; Song, F. et al. 2010; Cai, J. et al. 2016). Evaluation is based on the concept of entropy. Linear discriminant analysis is a feature extraction methodology and the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8076
***Article Received:** 24 June 2023 **Revised:** 18 August 2023 **Accepted:** 02 September 2023*

same is used for determining discriminating features as features are selected on the basis of how well the features separates the classes. Evaluation is based on inter class distance.

6) Prioritize Features: The selected features are ranked according to the importance with the class and create a subset of features by classifying the subset of the feature set and the results are evaluated using performance matrices like accuracy, sensitivity and specificity.

7) Factor Analysis: Conduct factor analysis to determine the underlying pattern and latent factors that cause the responses using all the features observe the factor loading of the variables with the latent factor and calculate the reliability of the factors formed by using Cronbach's alpha (Bruce T. 2004; Muhktar, F., and Oei, T. P. 2008).

8) Factor Loading: Compare the factor loading and the importance of the feature and re-run factor analysis with the subset of relevant features. Accept the analysis result if the pattern formed is not distorted. The resultant features are indicators that describes the response of the non-clinical college student considering both feature class relevance and feature to feature interaction.

B. Design

The primary objective of the stated work is to provide a real-time process for early detection of depression in the college going demography. The high-level design diagram and process is explained in Figure 1.

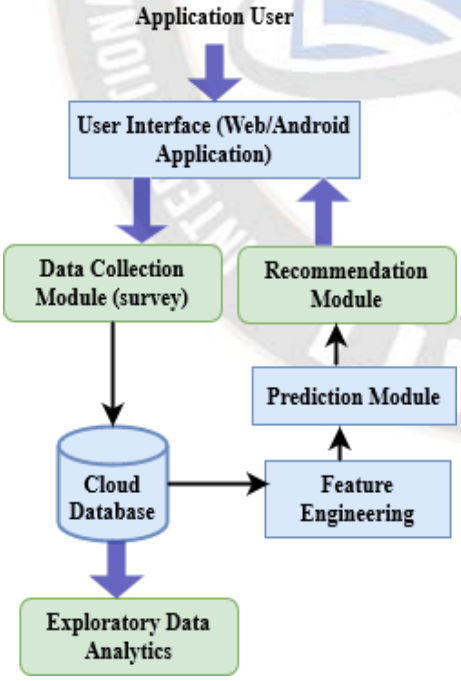

Figure 1. A high-level design diagram.

C. Data Collection

A web-based questionnaire is hosted on a portal/web application which can be easily accessed through a computer or a mobile device. It comprises of a customized BDI-II questionnaire expanded to include extra parameters pertinent to the demography (i.e. college going adults) it is targeted towards. These questionnaires are based on Diagnostic and Statistical Manual of Mental Disorder IV (DSM IV) (Brown, T. A., and Barlow, D. H. 2005) which is used worldwide to screen depression. The questionnaires assess the presence and severity in depressive symptoms. The responses carry a certain weightage, and the total sum of the responses suggests varying levels of depression. It can be self-administered; however, it is not a substitute for diagnosis by a trained professional. The respondents will also be assigned a unique ID to continually monitor their status. Like for instance with BDI-II, we can track improvements during treatment. This data is stored for each unique user on a cloud database that forms an integral component later during the processing of the data for improvement of the prediction system. The current study is done on data collected from 648 adults. The BDI – II questionnaire was augmented with additional questions carefully selected based on one-to-one communication between existing depression patients with the expert. Table 1 shows the different types of responses for different features.

TABLE I. RESPONSES OF DIFFERENT FEATURES

Sl. No.	New Features	Description	Responses (Weightage)
1	Doom	Thoughts of being better off without family and friends	Strongly Disagree (0), Disagree (1), Agree (2), Strongly Agree (3)
2	Expression	Expressing feelings with family and friends	Never (0), Sometimes (1), Most of the times (2), Frequently (3)
3	Usage	Mobile phone usage in 24 hours span	E-learning (0), Social Networking (1), Media (2), Gaming (3)
4	Usage Hours	Hours spent on the phone in 24 hours span	0-2 hours (0), 3-5 hours (1), 6-8 hours (2), more than 8 hours (3)
5	Belief	Degree of belief on the survey	Strongly Disagree (0), Disagree (1), Agree (2), Strongly Agree (3)

D. Exploratory Data Analytics

In this module the collected datasets are thoroughly analysed. The mean, median and mode have been plotted in the graph shown in Figure 2, to describe the data availability.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8076
***Article Received:** 24 June 2023 **Revised:** 18 August 2023 **Accepted:** 02 September 2023*

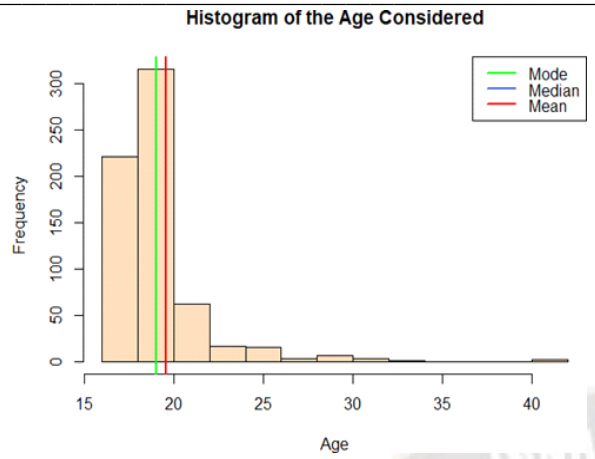

Figure 2. Summary of Age Group in BDI-II Survey.

The survey is taken by the students with an age range of 17 to 20 and few are taken by people whose age is above 20, so the distribution is not symmetric and it is skewed as shown in the graph, the survey is mainly concentrated on the age group of 17 to 20 as the data is coming from mostly students so the analysis would be carried out on the student data.

Gender and Depression Severity are two categorical data and are represented using contingency table. Numbers of female and male participants are 239 and 409 respectively, bar plot for representing severity of depression as 'A' represents minimal, 'B' represents mild, 'C' represents moderate and 'D' represents Severe, amongst the gender is shown in Figure 3.

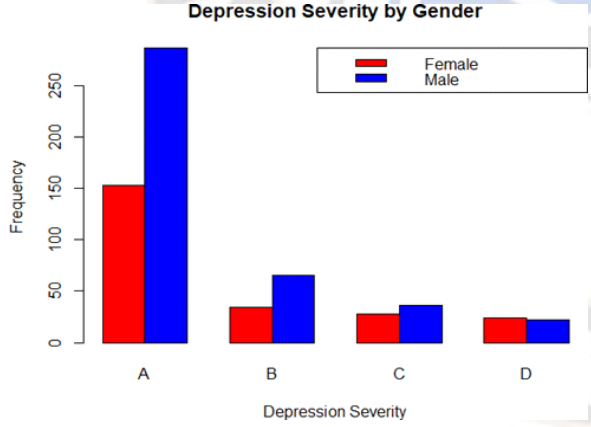

Figure 3. Depression Severity based on Gender.

E. Outlier Detection

For analysis purpose, R programming is used and tested with the given responses, for outlier detection Mahalanobis distance was calculated which checks for the pattern of responses and then calculates the distance from the centroid (Leys, C., et al. 2018). To compute Mahalanobis distance, following steps were undertaken:

Step 1: Calculate mean of the responses for each feature.
Step 2: Calculate the mean of means for all the feature set.
Step 3: Compute the Mahalanobis distance of each response from the centroid (means of means).
Step 4: Represent the Mahalanobis distance using boxplots which gives the presence of outliers.
Step 5: Calculate mean and median of the calculated Mahalanobis distance greater the difference effect of the presence of the outliers is more prominent.

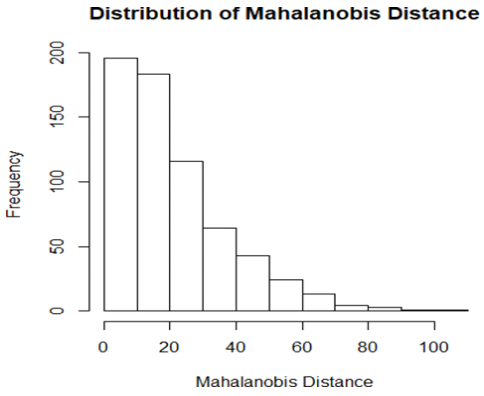

Figure 4. Distribution of Mahalanobis Distance

The mean and median of the Mahalanobis distance captured is 20.90 and 15.990 respectively, the difference between the two is 4.91 which means that the mean of the Mahalanobis distance is affected by the presence of outliers. The distribution of the Mahalanobis distance is chi-square distribution and is represented in the plot given in Figure 4. To compute the p-value-test results were obtained. One-sample t-test is conducted with obtained value as 31.7911, the p-value obtained for the t-test result is 0.04557. The cut-off score obtained is 33.025. Therefore, the score above 33.025 is considered as outliers and the value was 131. And 517 were outlier free data. However, the response collected had more observation of Minimal depression sample, so the Mahalanobis distance that was calculated was approximated and was affected by larger sample of minimal depression subjects so few outliers that was detected was a part of the analysis and provided an information about the population under study, so the outliers were not discarded and considered as a part of the solution. Total of 648 data was considered for further analysis.

F. Feature Engineering

After performing exploratory data analysis and examining the correlation amongst the predictors (features) which is used to predict classes we settled on including them in our prediction model. Feature selection is done using linear discriminant analysis (LDA) (Song, F., 2010) and information gain theory (Urbanowicz, R. J. et al. 2018). LDA concept gives the set of features that are discriminant in classifying depression severity the result is the given features, and the sum of the eigenvectors are calculated for the largest eigenvalues which is further

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8076
***Article Received:** 24 June 2023 **Revised:** 18 August 2023 **Accepted:** 02 September 2023*

computed by using the between class and total scatter matrices, the result is ranked according to its value, higher the value greater is the rank. For information gain the resulting feature with importance are given in Table 2.

TABLE II. RESULTING FEATURES WITH IMPORTANCE

Attributes	Importance	Rank
Concentration	0.22087563	1
Sadness	0.19475815	2
Pessimism	0.18575421	3
Decision	0.18544392	4
Guilty	0.18479543	5
Self-Dislike	0.18170023	6
Interest	0.17851609	7
Irritation	0.17483061	8
Failure	0.17287716	9
Tiredness	0.17096167	10
Blame	0.16171834	11
Punishment	0.15855645	12
Sleep-Pattern	0.1581059	13
Loss of Pleasure	0.15624814	14
Appetite	0.13188549	15
Cry	0.13079031	16
Self-Worth	0.12867226	17
Suicidal	0.12472239	18
Energy	0.10902671	19
Agitation	0.09678436	20
Sex	0.07432548	21

Feature selection using linear discriminant analysis concept gives the set of features that are discriminant in classifying depression severity the result is the given features and the sum of the eigenvectors that are calculated for the largest eigenvalues which is further computed by using the between class and total scatter matrices, the result is ranked according to its value, higher the value greater will be its rank. Table 3 gives the ranked feature set after computing feature selection using LDA.

TABLE III. ATTRIBUTES WITH EIGONVECTORS AND RANKS

Attributes	Sum of Eigen Vectors	Rank
Interest	0.74362737	1
Guilty	0.629092255	2
Suicidal	0.609664806	3
Pessimism	0.557566793	4
Concentration	0.550089506	5
Decision	0.525957157	6
Punishment	0.414465648	7
Blame	0.346724972	8
Tiredness	0.333698262	9
Agitation	0.321472182	10
Appetite	0.320554624	11
Sleep Pattern	0.256868538	12
Failure	0.238793587	13
Self-Worth	0.22854456	14
Loss of Pleasure	0.211633456	15
Irritation	0.208359236	16
Cry	0.187322355	17
Sex	0.124269588	18
Self-Dislike	0.081415915	19
Sadness	0.071862887	20
Energy	0.038644047	21

G. Prediction and Recommendation Module

Classification of responses are performed using supervised classification algorithms like Naive Bayes, K-Nearest Neighbour, Support Vector Machine (using svmLinear as the kernel trick), Decision Tree (Kotsiantis, S. B. et al. 2007). These basic algorithms and are chosen on the basis of its interpretability and accuracy. Within a specified amount of time the model is updated with new data to use batch prediction. That is, we schedule a service to run in some time interval that updates the model for output predictions. Finally, when a user is registered to our service, they can map their progress by repeatedly taking the survey and getting a score, over the treatment phase.

1) Classification of BDI-II Responses using Naïve Bayes Classifier: BDI-II Response classification using Naïve Bcc1ayes Algorithms as a classifier leads to classification of the responses considering the relationship between the features are independent to each other, it is based on the Bayes Theorem. Classification and performance results before feature selection using Naïve Bayes Classifier are given in Table 4 and Table 5.

2) Classification of BDI-II Responses using K-Nearest Neighbour (KNN) Classifier: BDI-II Response classification using KNN as a classifier leads to classification of the responses. KNN is a non-parametric algorithm. Due to its non-parametric property this algorithm has been used for solving real world classification problems. This classifier finds it's k nearest neighbour where k is number of neighbour data points. Classification and performance results before feature selection using the classifier are given in Table 6 and Table 7.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8076
***Article Received:** 24 June 2023 **Revised:** 18 August 2023 **Accepted:** 02 September 2023*

TABLE IV. CONFUSION MATRIX FOR NAÏVE BAYES CLASSIFIER

Predictions for Naïve				
	Minimal	**Mild**	**Moderate**	**Severe**
Minimal	87	1	0	0
Mild	12	23	6	0
Moderate	0	3	11	3
Severe	0	0	0	5

TABLE V. STATISTICS OF NAÏVE BAYES CLASSIFIER PERFORMANCE

Accuracy	0.8344
95% Confidence Interval	(0.7875, 0.9124)
Kappa	0.7183
Sensitivity for (Mild, Minimal, Moderate, Severe)	(0.9053, 0.8000, 0.63636, 0.75000)
Specificity for (Mild, Minimal, Moderate, Severe)	(1.0000, 0.9035, 0.95122, 0.98413)

TABLE VI. CONFUSION MATRIX FOR KNN CLASSIFIER

Predictions for KNN				
	Minimal	**Mild**	**Moderate**	**Severe**
Minimal	85	13	0	0
Mild	0	8	5	0
Moderate	0	1	6	0
Severe	0	0	0	6

TABLE VII. STATISTICS FOR KNN CLASSIFIER PERFORMANCE

Accuracy	0.84
95% Confidence Interval	(0.7711, 0.9052)
Kappa	0.6456
Sensitivity for (Mild, Minimal, Moderate, Severe)	(1.0000, 0.36364, 0.54545, 1.00000)
Specificity for (Mild, Minimal, Moderate, Severe)	(0.6667, 0.95098, 0.99115, 1.00000)

3) Classification of BDI-II Responses using Decision Tree (DT) Classifier: BDI-II Response classification using Decision tree as a classifier leads to classification of the responses. It is a non-linear classifier; it uses tree structure to model the relationship between among the features and its potential outcome. It is used to classify data into classes and represent the result in a flowchart like a tree structure. The decision tree classifies data in a dataset starting from the root it flows to the leaves which represents one class. Classification and performance results before feature selection using DT classifier are given in Table 8 and Table 9.

TABLE VIII. CONFUSION MATRIX FOR DT CLASSIFIER

Predictions for DT				
	Minimal	**Mild**	**Moderate**	**Severe**
Minimal	81	16	3	0
Mild	4	2	0	1
Moderate	1	5	11	7
Severe	0	0	0	4

TABLE IX. STATISTICS FOR DT CLASSIFIER PERFORMANCE

Accuracy	0.72
95% Confidence Interval	(0.6425, 0.7991)
Kappa	0.45
Sensitivity for (Mild, Minimal, Moderate, Severe)	(0.9419, 0.08696, 0.78571, 0.33333)
Specificity for (Mild, Minimal, Moderate, Severe)	(0.6122 , 0.95536, 0.89256, 1.00000)

4) Classification of BDI-II Responses using Support Vector Machine (SVM) Classifier: BDI-II Response classification using Support Vector Machine as a classifier leads to classification of the responses. SVM performs classification by creating a hyper plane the hyper plane needs to separate the classes well. A hyper plane should be such that it separates the classes and also the distance between the hyper plane and the support vectors should be more. SVM has a kernel trick which helps in separation of classes where it takes low dimensional input space and transforms into a high dimensional space. It converts non separable problem to separable problem by using function called kernels. It is more suitable in non-linearly separable problem. Classification and performance results before feature selection using SVM classifier are given in Table 10 and Table 11.

TABLE X. CONFUSION MATRIX FOR SVM CLASSIFIER

Prediction for SVM				
	Minimal	**Mild**	**Moderate**	**Severe**
Minimal	83	0	0	0
Mild	0	19	3	0
Moderate	0	0	10	2
Severe	0	0	0	12

TABLE XI. STATISTICS FOR SVM CLASSIFIER PERFORMANCE

Accuracy	0.94
95% Confidence Interval	(0.8433, 0.9881)
Kappa	0.909
Sensitivity for (Mild, Minimal, Moderate, Severe)	(1,0.80,0.80,1)
Specificity for (Mild, Minimal, Moderate, Severe)	(0.95,1,0.979,0.978)

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8076
***Article Received:** 24 June 2023 **Revised:** 18 August 2023 **Accepted:** 02 September 2023*

III. RESULTS AND DISCUSSION

On performing the factor analysis with all the features, the factors formed are given in Table 12. Factor loading of 0.35 and above is considered to have an appropriate level of correlation between the feature and the factor, with a minimum factor loading to be 0.30. Table 12 shows the highlighted factor loading above 0.30. The number of factors was chosen on the basis of Kaiser Criteria which says that number of eigenvalues above unity are considered to be as the number of factors is of Kaiser Criteria which says that number of eigenvalues above unity are considered to be as the number of factors.

TABLE XII. FACTOR LOADING FOR ALL FEATURES

Feature	Factor loading
Sadness	0.45
Pessimism	0.59
Failure	0.80
Loss of Pleasure	0.47
Guilty	0.36
Punishment	0.42
Self-Dislike	0.39
Blame	0.43
Suicidal	0.49
Cry	0.34
Agitation	0.51
Interest	0.51
Decision	0.45
Self-Worth	0.49
Energy	0.73
Sleep Pattern	0.54
Irritation	0.64
Appetite	0.57
Concentration	0.41
Tiredness	0.63
Sex	0.52

TABLE XIII. EVALUATING FACTOR ANALYSIS

Metrics	Value	Comment
Tucker Lewis Index (Before removing cry, after removing cry)	0.936, 0.938	Good
Factor 1 Reliability	0.88	Acceptable
Factor 1 Reliability	0.8	Acceptable

Where cry does not load to any factor which means that this feature is not important in terms of feature-to-feature relationship so it has been removed and factor analysis is performed to check the pattern is maintained as removing feature may result in losing of information and the original pattern may get distorted. The evaluation of factor analysis is given in a Table 13.

BDI-II Response classification using Naive Bayes Algorithms (Rish, I. 2001) as a classifier leads to classification of the responses considering the relationship between the features are independent to each other, it based on the Bayes Theorem. BDI-II Response classification using K-Nearest Neighbour (Weinberger, K. Q., and Saul, L. K. 2009) as a classifier leads to classification of the responses. KNN is a non-parametric algorithm. Due to its non-parametric property KNN has been used for solving real world classification problems. KNN classifier finds it's k nearest neighbour where k is number of neighbour data points. BDI-II Response classification using Decision tree (Safavian, S. R., and Landgrebe, D. 1991) as a classifier leads to classification of the responses. It is a non-linear classifier; it uses tree structure to model the relationship between among the features and its potential outcome. It is used to classify data into classes and represent the result in a flowchart like a tree structure. The decision tree classifies data in a dataset starting from the root it flows to the leaves which represents one class. BDI-II Response classification using Support Vector Machine (Kecman, V., 2005) as a classifier leads to classification of the responses. SVM performs classification by creating a hyper plane the hyper plane needs to separate the classes well. A hyper plane should be such that it separates the classes and also the distance between the hyper plane and the support vectors should be more. SVM has a kernel trick which helps in separation of classes where it takes low dimensional input space and transforms into a high dimensional space. It converts non separable problem to separable problem by using function called kernels. It is more suitable in non-linearly separable problem.

The accuracy comparison table for BDI-II response classification using Naive Bayes, KNN, Decision Tree and SVM is given in Table 14. It is being experimented that Support Vector Machine yields better result in comparison to other classification algorithm.

TABLE XIV. ACCURACY COMPARISON TABLE

Accuracy Matrix	Without Feature Selection	Feature Selection (Information Gain)	Feature selection (LDA)
Naive Bayes	0.85	0.83	0.82
KNN	0.70	0.84	0.84
DT	0.68	0.71	0.71
SVM	0.94	0.97	0.94

IV. SUMMARY AND CONCLUSION

Over the course of the work, features that are relevant to the class are obtained which gives increases the accuracy of the classification process. The feature selection methodology used was of filter method and the importance of the feature with respect to the class was generated based on information gain and linear discriminant analysis, the features were ranked according to the importance of the feature with respect to its class.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8076
***Article Received:** 24 June 2023 **Revised:** 18 August 2023 **Accepted:** 02 September 2023*

Classification of features selected using information gain as a feature selection method had better accuracy compared to that of linear discriminant analysis. Classification models like Naive Bayes, K-Nearest Neighbour, Support Vector Machine and Decision Tree are chosen based on its interpretability and accuracy, the best suited classification model was Support Vector Machine amongst the algorithms for classification of BDI-II responses. Using the feature subset obtained from feature selection method factor analysis was performed and compared to the results of the factor analysis performed based on all the features and concluded that loss of libido as a feature was least relevant to the class for classification and also had low factor loading on eliminating this feature considering both class-feature relevance and feature to feature interdependence, without distorting the underlying pattern or the pattern of the responses loaded in two factors. The factors obtained are the underlying pattern that causes the responses.

A. Difficulties encountered and tackeld

The first difficulty encountered was the collection of data; the population that was considered was of non-clinical college students, method to collect data was difficult. So in order to tackle with this problem a questionnaire was used namely Beck Depression Inventory-II (BDI-II). This questionnaire is used worldwide and has strong reliability in terms of internal consistency and validity when validated with other depression screening questionnaire. BDI-II is also used as screening measure to screen depression in Sikkim. The other difficulty was to determine the indicators that were relevant to the class on which it was classified but also maintained feature to feature relationship. So in order to achieve this feature selection was done on the basis of class relevance and then factor analysis was performed using the entire features as well as subset of feature and factor loadings was compared and feature subset was selected only if the eliminated features had low factor loading and it maintained the pattern created by the original datasets.

B. Limitation

The main limitation of the work is that this methodology can only be used to screen depression and not for diagnosis purpose. Indicators that are retrieved are based on the responses and the pattern of responses. There aren't any concrete processes to detect outliers of the responses collected. The result obtained by analysis of BDI-II gives analysis of only non-clinical students of a particular college, inference about all non-clinical college students as a whole cannot be made. This methodology also does not handle comorbidity issues i.e. a person being depressed due to other factors is not considered. The responses collected may be subjected to the "As is Scenario" i.e. How the person is feeling at that very moment when the survey was conducted.

C. Future work

The work can be enhanced by validating the use of Beck Depression Inventory-II with the use of other questionnaires that screen depression using the sample for conducting the survey. The factors that have been formed can be further scored and used for regression purpose. Beck Depression Inventory and Beck Anxiety Inventory can both be used to analyse the indicators, determine the relationship between these two questionnaires.

D. Special Observation

Experiment results indicate that on summarizing the data there are some information loss and also that the feature selection method using the filter method to rank the features depends upon the methodology used to determine the importance of feature with the class which gives different subset of features and the way to select the subset is using performance matrices like accuracy, specificity and sensitivity to select the feature selection methodology for the given scenario in hand.

REFERENCES

[1] Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961) 'An inventory for measuring depression', *Archives of general psychiatry*, 4(6), pp.561-571.

[2] Brown, T. A., & Barlow, D. H. (2005) 'Dimensional versus categorical classification of mental disorders in the fifth edition of the Diagnostic and statistical manual of mental disorders and beyond', Comment on the special section. Journal of abnormal psychology, 114(4), pp.551-556.

[3] Bruce T. (2004) 'Exploratory and confirmatory factor analysis: Understanding concepts and applications', Washington, DC, US: American Psychological Association. http://dx.doi.org/10.1037/10694-000.

[4] Cai, J., Wang, Z. J., Appel-Cresswell, S., & Mckeown, M. J. (2016) 'Feature selection to simplify BDI for efficient depression identification', In 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) (pp. 1-4). IEEE.

[5] Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016) 'An improved method to construct basic probability assignment based on the confusion matrix for classification problem', Information Sciences, 340, pp.250-261.

[6] Dozois, D. J., Dobson, K. S., & Ahnberg, J. L. (1998) 'A psychometric evaluation of the Beck Depression Inventory–II', Psychological assessment, 10(2), pp.83.

[7] Fayyad, J., Sampson, N. A., Hwang, I., Adamowski, T., Aguilar-Gaxiola, S., Al-Hamzawi, A., ... & Gureje, O. (2017) 'The descriptive epidemiology of DSM-IV Adult ADHD in the world health organization world mental health surveys', ADHD Attention Deficit and Hyperactivity Disorders, 9(1), pp.47-65.

[8] Firth, J., Torous, J., Nicholas, J., Carney, R., Pratap, A., Rosenbaum, S., & Sarris, J. (2017) 'The efficacy of smartphone-based mental health interventions for depressive

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8076
***Article Received:** 24 June 2023 **Revised:** 18 August 2023 **Accepted:** 02 September 2023*

symptoms: a meta-analysis of randomized controlled trials', World Psychiatry, 16(3), pp.287-298.

[9] Hosseinifard, B., Moradi, M. H., & Rostami, R. (2013), 'Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal', Computer methods and programs in biomedicine, 109(3), pp.339-345.

[10] Kecman, V. (2005) 'Support vector machines–an introduction', In Support vector machines: theory and applications, Springer, Berlin, Heidelberg, pp. 1-47.

[11] Kendall, P. C., Hollon, S. D., Beck, A. T., Hammen, C. L., & Ingram, R. E. (1987) 'Issues and recommendations regarding use of the Beck Depression Inventory', Cognitive therapy and research, 11(3), 289-299.

[12] Mohammed AL-Mafrji, A. A. ., Fakhrudeen, A. M. ., & Chaari, L. . (2023). Expert Systems in Banking: Artificial Intelligence Application in Supporting Banking Decision-Making. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 61–69. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2572

[13] Khodayari-Rostamabad, A., Reilly, J. P., Hasey, G., deBruin, H., & MacCrimmon, D. (2010) 'Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model', In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 4006-4009.

[14] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007) 'Supervised machine learning: A review of classification techniques', Emerging artificial intelligence applications in computer engineering, 160(1), pp.3-24.

[15] Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018) 'Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance', Journal of Experimental Social Psychology, 74, pp.150-156.

[16] Mair, C., Roux, A. V. D., Osypuk, T. L., Rapp, S. R., Seeman, T., & Watson, K. E. (2010) 'Is neighborhood racial/ethnic composition associated with depressive symptoms? The multi-ethnic study of atherosclerosis', Social science & medicine, 71(3), pp.541-550.

[17] Matheson F. I., Moineddin R., Dunn J. R., Creatore M. S., Gozdyra P, and Glazier R. H. (2006) 'Urban neighborhoods, Chronic stress, gender and depression', Social Science and Medicine, Vol.63, Issue 10, ISSN0277-9536, https://doi.org/10.1016/j.socscimed.2006.07.001, pp.2604-2616.

[18] Muhktar, F., & Oei, T. P. (2008) 'Exploratory and confirmatory factor validation and psychometric properties of the Beck Depression Inventory for Malays (BDI-Malay) in Malaysia', Malaysian Journal of Psychiatry, 17(1).

[19] Ojeme B., Akazue M. and Nwelih. E. (2015) 'Automatic Diagnosis of Depressive Disorders using Ensemble Techniques' Vol 8. No. 3(2), 2006-1781

[20] Ojeme, B., Mbogho, A., & Meyer, T. (2016)0 'Probabilistic expert systems for reasoning in clinical depressive disorders', In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 599-604.

[21] Osman, A., Barrios, F. X., Gutierrez, P. M., Williams, J. E., & Bailey, J. (2008) 'Psychometric properties of the Beck Depression Inventory-II in nonclinical adolescent samples', Journal of clinical psychology, 64(1), pp.83-102.

[22] Pathak, D. G. ., Angurala, D. M. ., & Bala, D. M. . (2020). Nervous System Based Gliomas Detection Based on Deep Learning Architecture in Segmentation. Research Journal of Computer Systems and Engineering, 1(2), 01:06. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/3

[23] Peng, H., Long, F., & Ding, C. (2005) 'Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy', IEEE Transactions on pattern analysis and machine intelligence, 27(8), pp. 1226-1238.

[24] Richter, P., Werner, J., Heerlein, A., Kraus, A., & Sauer, H. (1998) 'On the validity of the Beck Depression Inventory', Psychopathology, 31(3), pp.160-168.

[25] Rish, I. (2001, August) 'An empirical study of the naive Bayes classifier', In IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3, No. 22, pp. 41-46.

[26] Safavian, S. R., & Landgrebe, D. (1991) 'A survey of decision tree classifier methodology', IEEE transactions on systems, man, and cybernetics, 21(3), pp.660-674.

[27] Song, F., Mei, D., & Li, H. (2010, October) 'Feature selection based on linear discriminant analysis', In 2010 International Conference on Intelligent System Design and Engineering Application, Vol. 1, pp.746-749.

[28] Stengel, E. (1959) 'Classification of mental disorders', Bulletin of the World Health Organization, 21(4-5), 1959, 21, pp.601-663.

[29] Storch, E. A., Roberti, J. W., & Roth, D. A. (2004) 'Factor structure, concurrent validity, and internal consistency of the beck depression inventory', Second edition in a sample of college students. Depression and anxiety, 19(3), pp.187-189.

[30] Ms. Pooja Sahu. (2015). Automatic Speech Recognition in Mobile Customer Care Service. International Journal of New Practices in Management and Engineering, 4(01), 07 - 11. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/34

[31] Understanding Depression, National Institute of Mental Health, https://www.nimh.nih.gov/health/topics/depression/index.shtml. Available: National Institute of Mental Health Information Resource Centre [Accessed online: June 21 2019]

[32] Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018) 'Relief-based feature selection: Introduction and review', Journal of biomedical informatics, 85, pp.189-203.

[33] Weinberger, K. Q., & Saul, L. K. (2009) 'Distance metric learning for large margin nearest neighbor classification', Journal of Machine Learning Research, 10(2).