

**SUICIDE IDEATION PREDICTION FOR MILITARY PERSONNEL
USING SVM ALGORITHM**

at

**Sathyabama Institute of Science And Technology
(Deemed to be university)**

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology degree in Information Technology

By

**SHALINI.T
(Reg.No.37120067)
SELCIYA.M
(Reg.No.37120066)**



**DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119**

MARCH – 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai –600119

www.sathyabama.ac.in



DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of SHALINI.T(37120067) SELCIYA.M(37120066) who carried out the project entitled **"SUICIDE IDEATION PREDICTION FOR MILITARY PERSONNEL USING SVM ALGORITHM"** under my supervision from November 2019 to April 2020.

Internal Guide

Ms.VIMALI.J.S.,M.Tech.,(Ph.D).,

Head of the Department

Dr.R.SUBHASHINI., M.E., Ph.D. ,

Submitted for Viva voce Examination held on_____

Internal Examiner

External Examiner

DECLARATION

I am SHALINI.T(37120067) and M.SELCIYA(37120066),hereby declare that the Project Report entitled "**SUICIDE IDEATION PREDICTION FOR MILITARY PERSONNEL USING SVM ALGORITHM**" done by me under the guidance of Ms.Vimali.J.S.,M.Tech.,(Ph.D)., is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in Information Technology.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D, Dean**, School of Computing , **Dr.R.Subhashini M.E., Ph.D. , Head** of the Department of Information Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Ms.Vimali.J.S.,M.Tech.,(Ph.D).**,for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

Abstract

Suicides in the Military have increased over the last ten years. High mental stress may cause suicide ideations which are crucially driving suicide attempts. However, traditional statistical methods could only find a moderate degree of correlation between psychological stress and suicide ideation in non-psychiatric individuals. This paper utilizes machine learning techniques including gradient boosting regression tree, support vector machine and multilayer perception to predict the presence of suicide ideation by six important psychological stress domains of the military males and females. We are going to use 4 steps which are 1. Data collection, 2.Data Preprocessing, 3.Feature Extraction, 4.Classfication.The proposed algorithms can improve the performances of accuracy, sensitivity, specificity, precision. This SVM algorithm provides the best predictions of suicide ideation. This can be compare with the other algorithms which can help to improve the predictions. The Support Vector Machine algorithm is a most popular supervised learning algorithm which is used for classification and regression. This can improve the accuracy of suicide ideation predictions which can help to decrease the suicides in military.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO
	ABSTRACT	v
	LIST OF FIGURES	vii
	LIST OF ABBREVIATIONS	viii
1	INTRODUCTION	1
	1.1 Project Outline	
	1.2 Introduction	
2	LITERATURE SURVEY	3
3	AIM AND SCOPE OF THE PROJECT	4
	3.1 Existing System	5
	3.2 Proposed System	6
	3.3 System Architecture	7
4	SOFTWARE DEVELOPMENT AND METHODOLOGY	8
	4.2Software requirements	
	4.3Methodology	11
	4.4Algorithms	16
	4.5 Dataflow Diagrams	18
5	RESULTS AND DISCUSSION	24
	5.2 Performance Analysis	25
6	CONCLUSION AND FUTURE WORK	26
	6.1 conclusion	
	6.2 future work	27
	REFERENCE	
	APPENDIX	
	A.Plagiarism report	
	B.Base Paper	
	C.Source Code	
	D. Output Screenshots	

LIST OF FIGURES

FIGURE NO	TITLE	PAGENO
3.3	System Architechture	7
4.3.1	Data Collection and Preprocessing	10
4.5	Dataflow diagram for suicide ideation prediction	17
4.5.1	Dataflow diagram for level 1 Preprocessing	18
4.5.2	Dataflow diagram for level 2 Feature Extraction	19
4.5.3	Dataflow diagram for level 2 Splitting Data	20
4.5.4	Dataflow diagram for level 2 Classifying Data	21
4.5.5	Data Flow Diagram Level 2 for Accuracy Prediction	22
4.5.6	Usecase Diagram	23
4.5.7	Sequence Diagram of module	24

LIST OF ABBREVIATION

SVM-Support Vector Machine

DNN-Deep Neural Network

KNN-K-Nearest Neighbour

CHAPTER -1

INTRODUCTION

1.1 PROJECT OUTLINE:

TITLE OF THE PROJECT :- SUICIDE IDEATION PREDICTION FOR MILITARY PERSONNEL USING SVM ALGORITHM.

1.2 INTRODUCTION

Military personnel have greater psychological stress and are at higher suicide attempt risk compared with the general population. High mental stress may cause suicide ideations which are crucially driving suicide attempts. However, traditional statistical methods could only find a moderate degree of correlation between psychological stress and suicide ideation in non-psychiatric individuals. This paper utilizes machine learning techniques including logistic regression, decision tree, random forest, gradient boosting regression tree, support vector machine and multilayer perceptron to predict the presence of suicide ideation by six important psychological stress domains of the military males and females. The accuracies of all the six machine learning methods are over 98%. Among them, the multilayer perceptron and support vector machine provide the best predictions of suicide ideation approximately to 100%. As compared with the BSRS-5 score ≥ 7 , a conventional criterion, for the presence of suicide ideation ≥ 1 , the proposed algorithms can improve the performances of accuracy, sensitivity, specificity, precision, the AUC of ROC curve and the AUC of PR curve up to 5.7%, 35.9%, 4.6%, 65.2%, 4.3% and 53.2%, respectively; and for the presence of more severely intense suicide ideation ≥ 2 , the improvements are 6.1%, 26.2%, 5.8%, 83.5%, 2.8% and 64.7%, respectively.

CHAPTER-2

LITERATURE SURVEY

2.1Body composition in the military services: Standards and methods (2015)

The development of standards for body composition in the U.S. Navy and the methods of body composition assessment in use by the military services today. Navy's policy that judgments about appearance are subjective.

2.2Application of three-dimensional body scanner: Observation of prevalence of metabolic syndrome (2015)

This retrospective cross-sectional study correlates blood pressure, blood glucose, lipid and uric acid levels with anthropometric measurements. It does not give accurate value.

2.3Do measures of body fat distribution provide information on the risk of type 2 diabetes in addition to measures of general obesity (2016)

To investigate which anthropometric measurements of obesity best predict type 2 diabetes in a population of Pima Indians and whether additional information on diabetes risk. manually corrected in case of failure of automatic detection.

2.4Waist circumference and abdominal sagittal diameter as surrogates of body fat distribution in the elderly: Their relation with cardiovascular risk factors(2017)

Recent works on automatic land-marking try to make the automatic analysis more robust, for example against pose variation If the legs are not in contact under the knee and arms are not adherent to the trunk under the elbow.

2.6SCAPE: shape completion and animation of people (2015)

An approach to derive geometric descriptors of a body consists of creating a statistical model of shape and pose from annotated examples. observation and palpation by expert anthropometrists which are difficult to locate only through marker-less digital acquisitions.

CHAPTER-3

3.1 AIM AND SCOPE OF THE PROJECT

Our Aim Military Stress-ML should meet the basic requirements, i.e., predict the churning of the military, find the accuracy of the above algorithms and find the efficient one. Efficient model predicts the Military Personnel and helps in relieving the Military Personnel from the stress and help them in retaining in the respective in the respective Military service.

3.2 EXISTING SYSTEM

Stressed military, tend to cease doing work in a given time period which has become a significant problem for many firms. There are numerous predictive modeling techniques for predicting military personals behavior and their satisfactory level. These vary in terms of statistical technique (e.g., neural nets versus logistic regression), variable selection method (e.g., theory versus stepwise selection), and number of variables included in the model.

We have highlighted below algorithm and its predicting techniques

- KNN algorithm

3.2.1K Nearest Neighbour

KNN algorithm is one of the simplest classification algorithm. Even with such simplicity, it can give highly competitive results. KNN algorithm can also be used for regression problems. KNN algorithm which when implemented helps to predict wither the Military will churn or not. But the accuracy level is not so high.

Limitations

The major disadvantage in KNN is determining the K value . Distance based classifier makes a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class.

Another problem happens due to data scarcity. For any possible value of a feature, you need to estimate a likelihood value by a frequents approach. This can result in probabilities going towards 0 or 1, which in turn leads to numerical instabilities and worse results. In this case, you need to smooth in some way your probabilities or to impose some prior on your data.

3.2PROPOSED SYSTEM

- SVM algorithm

3.2.1SVM ALGORITHM

SVM algorithm developed by Vapnik, is based on statistical learning theory. In some classification cases, we try to find an optimal hyper-plane that separates two classes. When the two classes of points in the training set can be separated by a linear hyper-plane, it is natural to use the hyper-plane that separates the two groups of points in the training set by the largest margin. In order to find an optimal hyper-plane, we need to minimize the norm of the vector w , which defines the separating hyper-plane. This is equivalent to maximizing the margin between two classes.

Limitations

Military stress prediction is a problem of classification between “happy” and “stressed” Military. But when the number of the negative examples is too small, the generalization performance of SVM classifier must be weak, and the error rates is proved unsatisfactory.

In this paper, we review the existing works on Military stress prediction in three different perspectives: datasets, methods, and metrics. Firstly, we present the details about the availability of public datasets and what kinds of military details are available for predicting Military stress, which leads to churn. Secondly, we compare and contrast the various predictive modeling methods that have been used in the literature for predicting the churners using different categories of Military records, and then quantitatively compare their performances. Finally, we summarize what kinds of performance metrics have been used to evaluate the existing churn prediction methods. Analyzing all these three perspectives is very crucial for developing a more efficient churn prediction system for telecom industries.

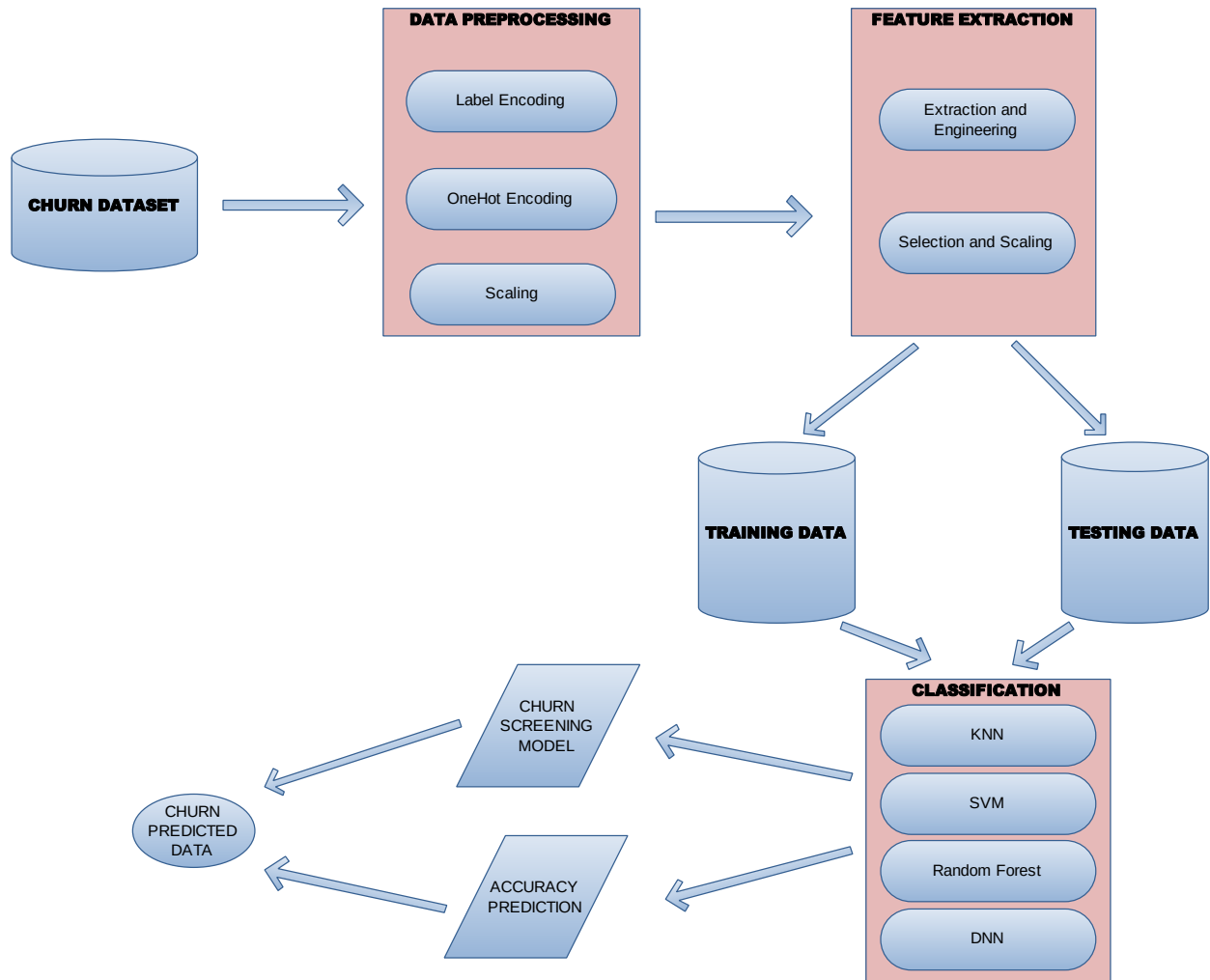
This is because of Military stress and un-satisfaction with the job. From a machine learning perspective, military stress (probable churn) prediction is a supervised (i.e. labeled) problem defined as follows: Given a predefined forecast horizon, the goal is to predict the future churners over that horizon, given the data associated with each subscriber in the network.

Churn Prediction is a phenomenon which is used to identify the possible churners in advance before they leave the network. This helps the CRM department to prevent subscribers who are likely to churn in future by taking the required retention policies to attract the likely churners and to retain them. Thereby, the potential loss of the company could be avoided.

The input for this problem includes the data on past calls for each mobile subscriber, together with all personal and business information that is maintained by the service provider. In addition, for the training phase, labels are provided in the form of a list of churners. After the model is trained with highest accuracy, the model must be able to predict the list of churners from the real dataset which does not include any churn label. In the perspective of knowledge discovery process, this problem is categorized as predictive mining or predictive modeling.

We are analyzing the Churn data with DNN – Deep Neural Network and Random forest . As both the algorithms are a deep investigator of data , the accuracy of the prediction can be increased .

3.3 SYSTEM ARCHITECHTURE:



CHAPTER-4

4.1 SOFTWARE DEVELOPMENT AND METHODOLOGY

4.2 SYSTEM REQUIREMENTS:

4.2.1 HARDWARE REQUIREMENTS

- SYSTEM : PENTIUM DUAL CORE
- HARD DISK : 120 GB
- MONITOR : 15" LCD
- INPUT DEVICE : KEYBOARD, MOUSE
- RAM : 8 GB

4.2.2 SOFTWARE REQUIREMENTS

- OPERATING SYSTEM : WINDOWS 10
- PROGRAMMING LANGUAGE : PYTHON
- TOOLS : ANACONDA, SPYDER, PYTHON
- DATABASE : KERAS LIBRARY

4.3 METHODOLOGY

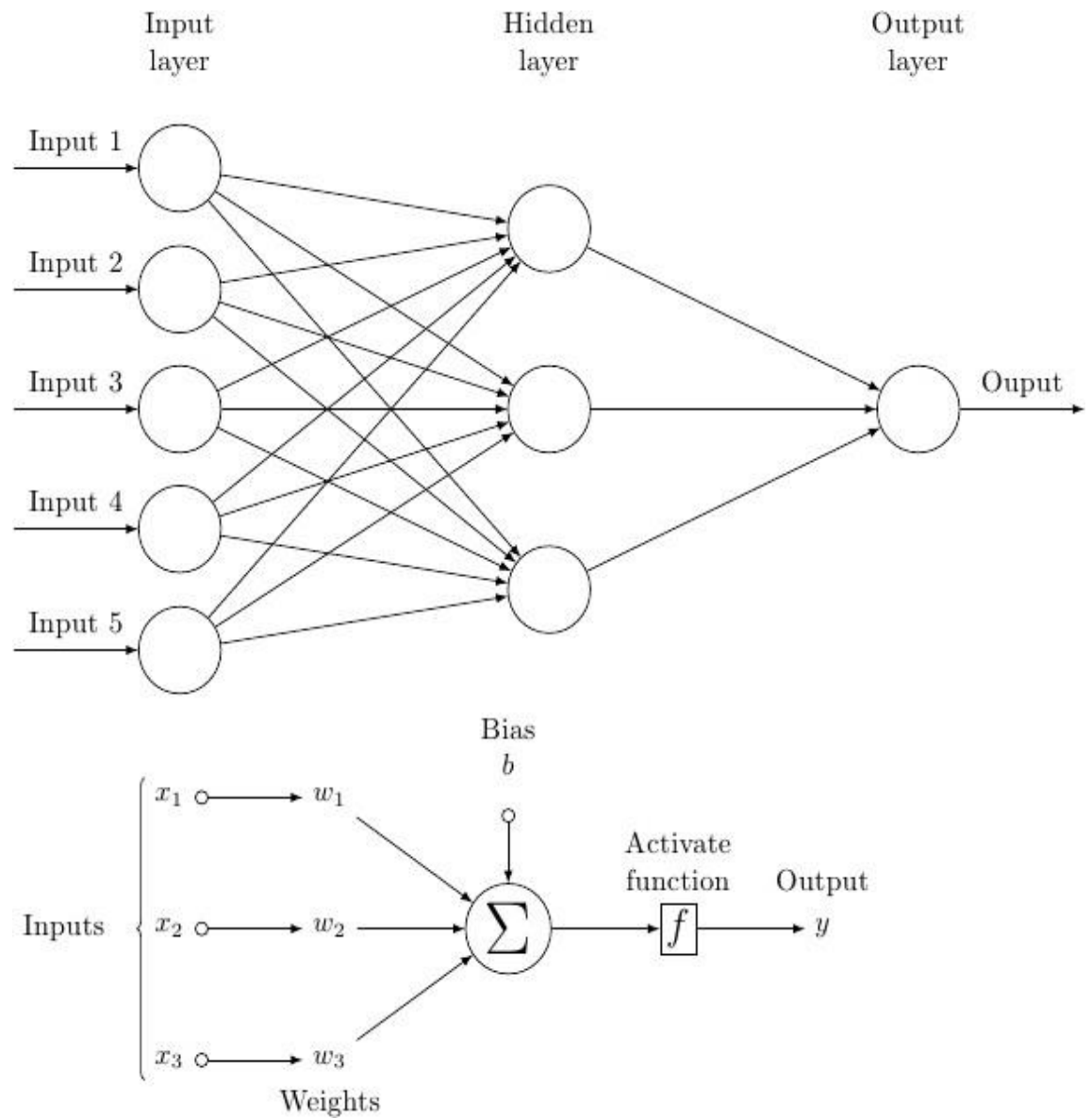
4.3.1 DATA DISCOVERY MODEL

Data extraction :

We have collected the Telecom Industry Churn dataset . We are using this data info to predict the churning of military from that particular telecom industry for the current year and make appropriate preparations. Importing the Pandas gives massive functionality to work on data. Along with Pandas importing all the required libraries are even imported.

Preprocessing :

The data has to be Cleaned before loading into the Neural Classifier .For each customer, I collected a total of 972 variables, divided into major categories shown in [Table 1](#). The full list of variables is provided in [S1 Table](#). All data elements were obtained from the enterprise data warehouse, using SQL queries to extract relevant raw-data in comma-separated value format. All subsequent processing was done in R. The link to a repository containing the de-identified dataset and R scripts are available in [S1 Text](#). Below, I summarize the processing steps for each category.



4.3.1 DATA COLLECTION AND PREPROCESSING

Adding the output layer

```
classifier.add(Dense(output_dim = 1, init = 'uniform', activation = 'sigmoid'))
```

First argument is Optimizer, this is the algorithm I use to find optimal set of weights. This algorithm is Stochastic Gradient descent(SGD). Among several types of SGD algorithm the one which I will use is 'Adam'. If I go in deeper detail of SGD, I will find that SGD depends on loss thus our second parameter is loss. Since our dependent variable is binary, I will have to use logarithmic loss function called 'binary_crossentropy', if our dependent variable has more than 2 categories in output then use 'categorical_crossentropy'. I want to improve performance of our neural network based on accuracy so add metrics as accuracy

I can train our model on training data. I use fit method to fit our model. Batch size is used to specify the number of observation after which I want to update weight. Epoch is nothing but the total number of iterations. Choosing the value of batch size and epoch is trial and error there is no specific rule for that.

LabelEncoding :

Before running a model, we need to make this data ready for the model. To convert any kind of categorical text data into model-understandable numerical data, we use the Label Encoder class, all we have to do is to label encode the first column. We import the Label Encoder class from the sklearn library, fit and transform the first column of the data, and then replace the existing text data with the new encoded data.

Onehotencoding :

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. This is why we use one hot encoder to perform "binarization" of the category and include it as a feature to train the model.

StandardScaler:

StandardScaler is available in ScikitLearn we are fitting and transforming StandardScaler method on train data. We have to standardize our scaling so we will use the same fitted method to transform/scale test data.

4.3.2 CLASSIFICATION MODEL

Data visualisation and descriptive statistics:

The extracted features are usually plotted against the output to check its relation to the output .

Data splitting into training (80%) and test sets (20%):

In statistics and machine learning we usually split our data into two subsets: **training data** and **testing data** (and sometimes to three: train, validate and test), and fit our model on the train data, in order to make predictions on the test data. When we do that, one of two things might happen: we overfit our model or we underfit our model. We don't want any of these things to happen, because they affect the predictability of our model—we might be using a model that has lower accuracy and/or is ungeneralized (meaning you can't generalize your predictions on other data).

4.3.3 ANALYTICS USING MODEL

Overfitting

Overfitting means that model we trained has trained “too well” and is now, well, fit too closely to the training dataset. This usually happens when the model is too complex (i.e. too many features/variables compared to the number of observations).

This model will be very accurate on the training data but will probably be very not accurate on untrained or new data. It is because this model is not generalized (or not AS generalized), meaning we can generalize the results and can't make any inferences on other data, which is, ultimately, what we are trying to do. Basically, when this happens, the model learns or describes the “noise” in the training data instead of the actual relationships between variables in the data. This noise, obviously, isn't part in of any new dataset, and cannot be applied to it.

Underfitting

In contrast to overfitting, when a model is underfitted, it means that the model does not fit the training data and therefore misses the trends in the data. It also means the model cannot be generalized to new data. As you probably guessed this is usually the result of a very simple model (not enough predictors/independent variables). It could also happen when, for example, we fit a linear model to data that is not linear. It almost goes without saying that this model will have poor predictive ability (on training data and can't be generalized to other data).

Train / Test Split

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset.

4.3.4 PREDICTION USING MODEL

Predicting admissions based on the test data set

Predicting the test set result. The prediction result will give you probability of the military churning out of the Telecom industry for the given period . We will convert that probability into binary 0 and 1.

The evaluation of model performance based on predictions made on the test data

This is the final step where we are evaluating our model performance. We build Confusion matrix to check the accuracy of model.

4.4 ALGORITHMS

The machine learning algorithms that has to be applied to the training data to build the models are

4.4.1 SVM ALGORITHM

SVM algorithm developed by Vapnik, is based on statistical learning theory. In some classification cases, we try to find an optimal hyper-plane that separates two classes. When the two classes of points in the training set can be separated by a linear hyper-plane, it is natural to use the hyper-plane that separates the two groups of points in the training set by the largest margin. In order to find an optimal hyper-plane, we need to minimize the norm of the vector w , which defines the separating hyper-plane. This is equivalent to maximizing the margin between two classes.

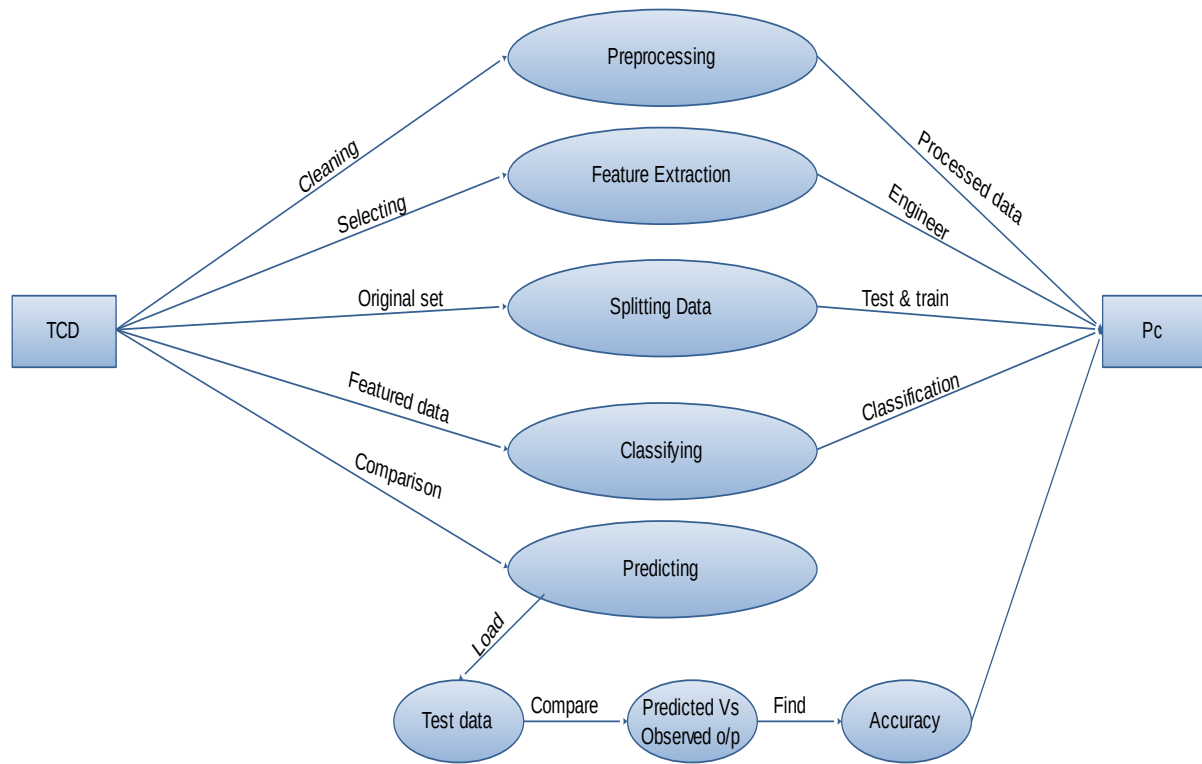
4.4.2Deep Neural Network

Logistic regression is suitable for predicting a binary dependent variable, such as positive/negative; deceased/alive; or in this study, admit/not admit. The technique uses a logistic link function to enable the calculation of the odds of an outcome occurring. The second algorithm that was used was a random forest, specifically recursive partitioning from the RPART package . The RPART package is an implementation based on the model presented by Breiman and colleagues . This algorithm splits the data at each node based on the model presented by Breiman and colleagues .

This algorithm splits the data at each node based on the variable that best separates the data until either an optimal model is identified or a minimum number of observations exists in the final (terminal) nodes .

The resulting tree can then be pruned to prevent overfitting and to obtain the most accurate model for prediction . The third algorithm was a SVM which creates two vectors and tries to classify the data based on it . The last algorithm is KNN where the N defines the number of nearest neighbours and the value changes as and when the n changes .Thus all the algorithms can increase the prediction and accuracy of the Churn prediction in Telecom Industry .

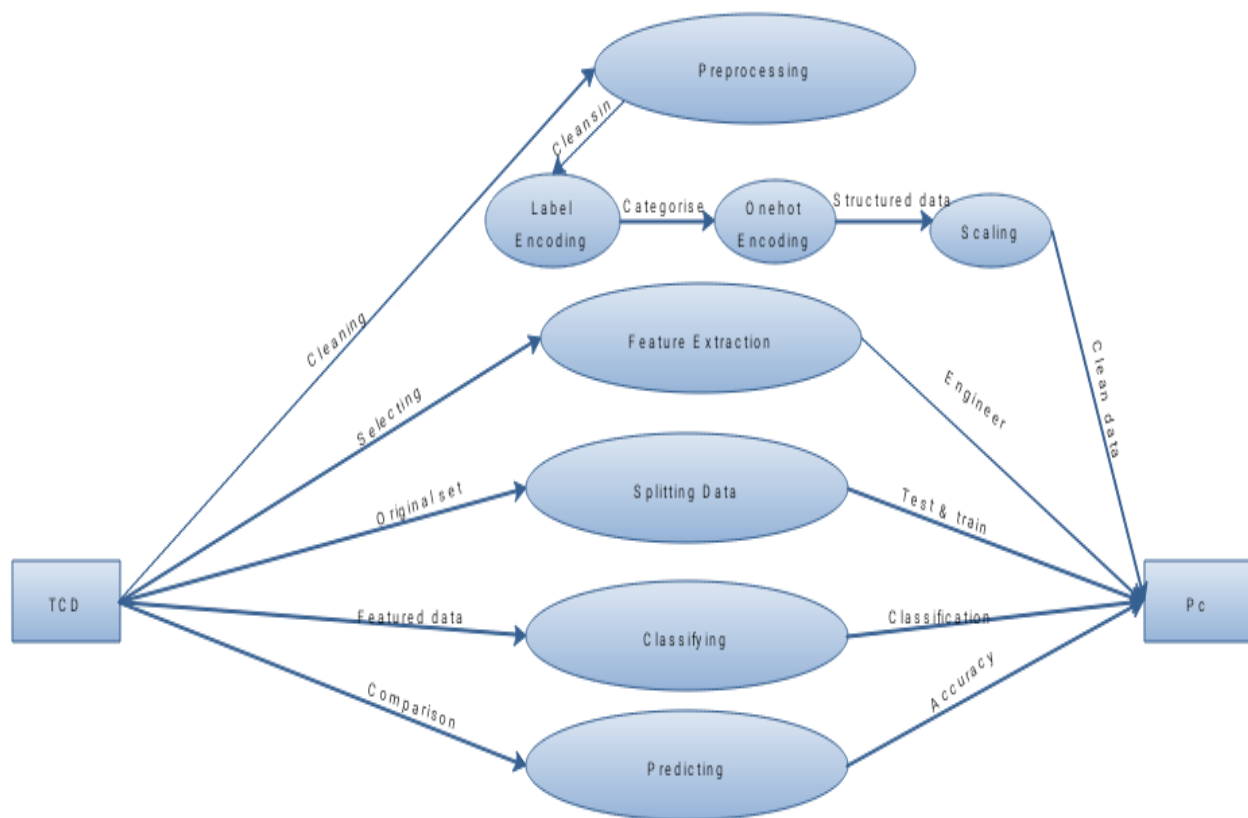
4.5 DATA FLOW DIAGRAM



4.5 Dataflow diagram for suicide ideation prediction

4.5.1 DATA FLOW DIAGRAM LEVEL 1 FOR DATA PREPROCESSING

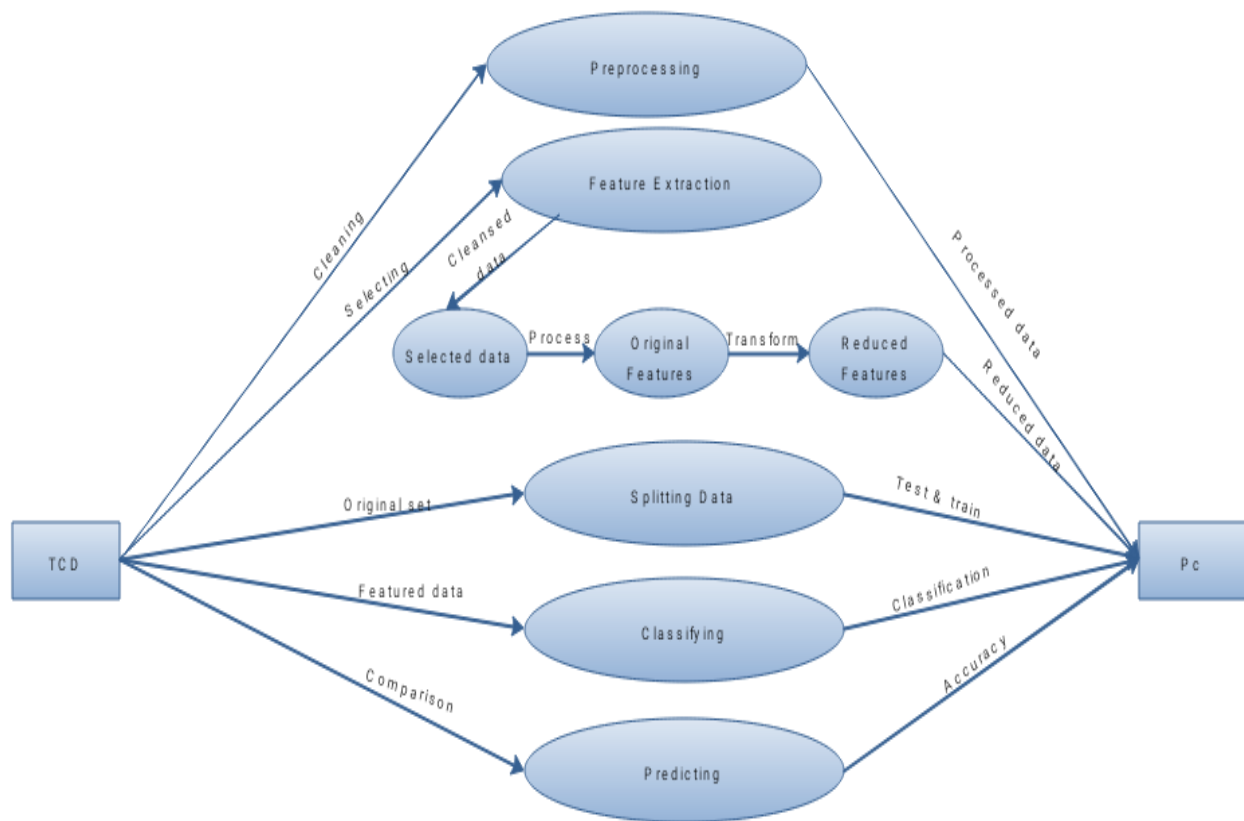
From the below diagram clearly explains the data is cleaned before loading to the Neural Network Classifier. Import the Label Encoder class from the sclera library, fit and transform the first column of the data, and then replace the existing text data with the new encoded data. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. This work fitting and transforming StandardScaler method on train data.



4.5.1 DATA FLOW DIAGRAM LEVEL 1 FOR DATA PREPROCESSING

4.5.2 DATA FLOW DIAGRAM LEVEL 2 FOR FEATURE EXTRACTION

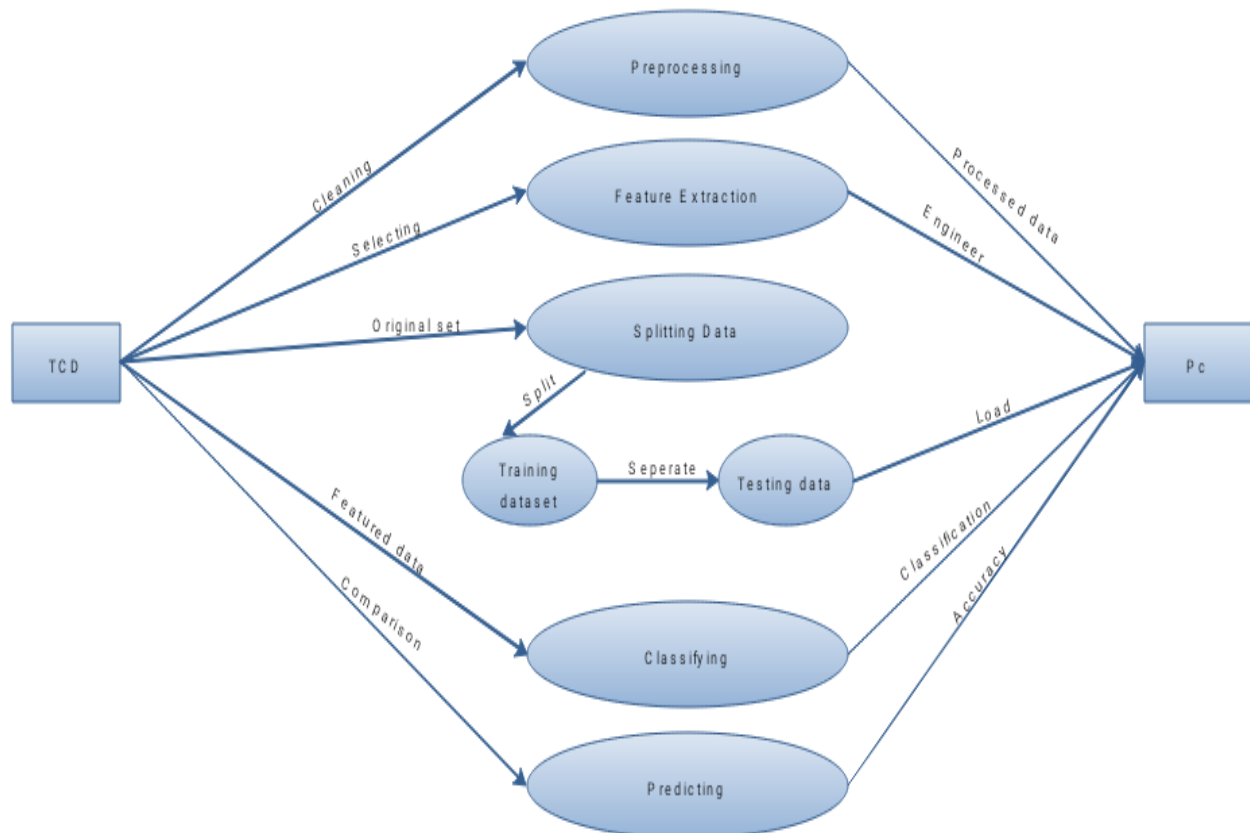
From the below diagram explains the Feature extraction by Layering of data. For feature extraction, first select the data and process data for original feature. Then reduce the feature by transform refining data. This can be done by selecting the data by CSV file and process the data for original feature and it was transformed and refining data to reduced feature, finally it predicts the accuracy of churning.



4.5.2 DATA FLOW DIAGRAM LEVEL 2 FOR FEATURE EXTRACTION

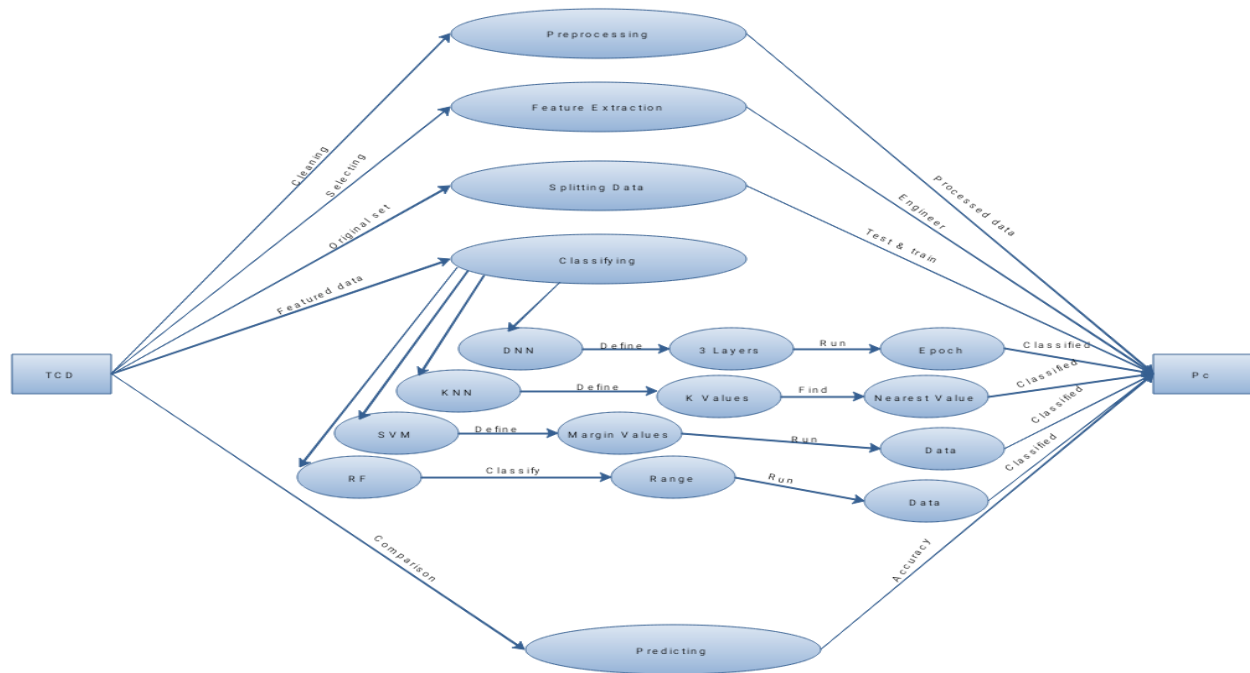
4.5.3 DATA FLOW DIAGRAM LEVEL 2 FOR SPLITTING DATA

From the below diagram explains the Splitting data by two ways are Training dataset and Testing dataset. In statistics and machine learning we usually split our data into two subsets: **training data** and **testing data** (and sometimes to three: train, validate and test), and fit our model on the train data, in order to make predictions on the test data. When this work do that, one of two thing might happen: we overfit our model or we underfit our model. It don't want any of these things to happen, because they affect the predictability of our model— Work might be using a model that has lower accuracy and/or is ungeneralized .



4.5.3 DATA FLOW DIAGRAM LEVEL 2 FOR SPLITTING DATA

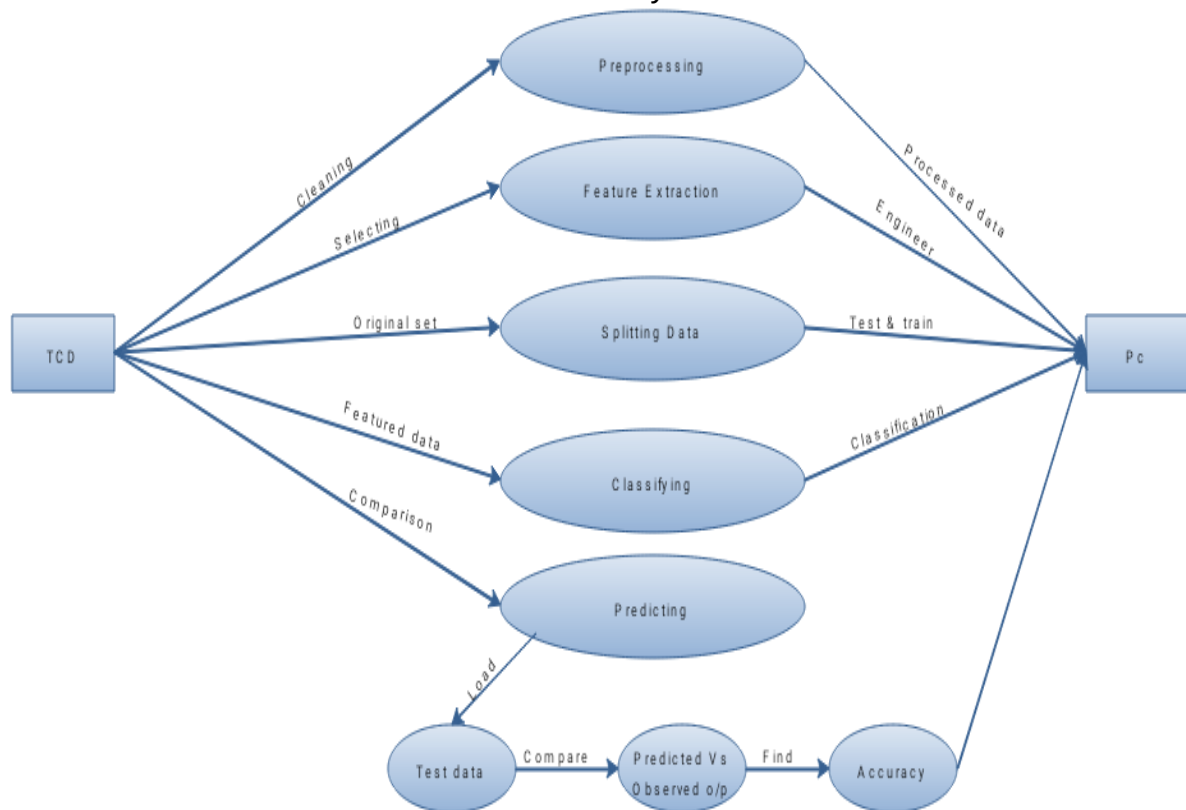
4.5.4 DATA FLOW DIAGRAM LEVEL 2 FOR CLASSIFYING DATA



From above the diagram explains that Classification of data through DNN (Deep Neural Network) Algorithm Neural networks use randomness by design to ensure they effectively learn the function being approximated for the problem. Randomness is used because this class of machine learning algorithm performs better with it than without. The most common form of randomness used in neural networks is the random initialization of the network weights. A deep neural network is a neural network with a certain level of complexity, a neural network with more than two layers. Deep neural networks use sophisticated mathematical modeling to process data in complex ways. A neural network, in general, is a technology built to simulate the activity of the human brain — specifically,

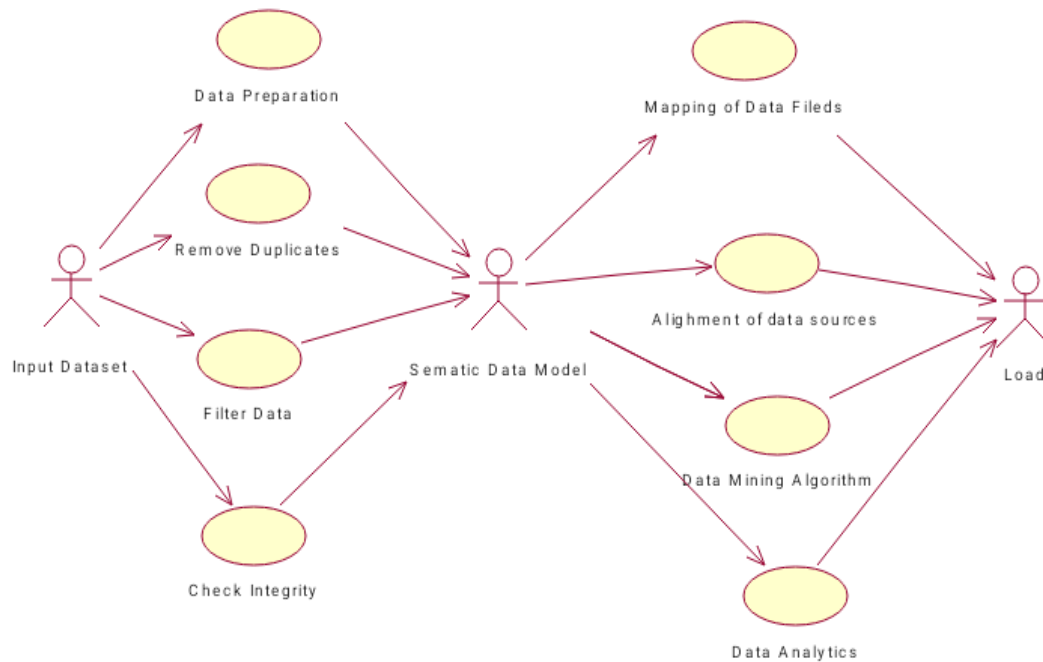
4.5.5 Data Flow Diagram Level 2 for Accuracy Prediction

From the above diagram explains how the accuracy emergency department prediction occurs using the classifier, list with predicted values and plot the rainfall graph. **Predicting** the test set result. The prediction result will give you probability of customers churning the telecom industry. This will convert that probability into binary 0 and 1. 1 for Churning the Telecom Industry and 0 for not churning. This is the final step where this work evaluating our model performance. It predicts the Confusion matrix to check the accuracy of model.



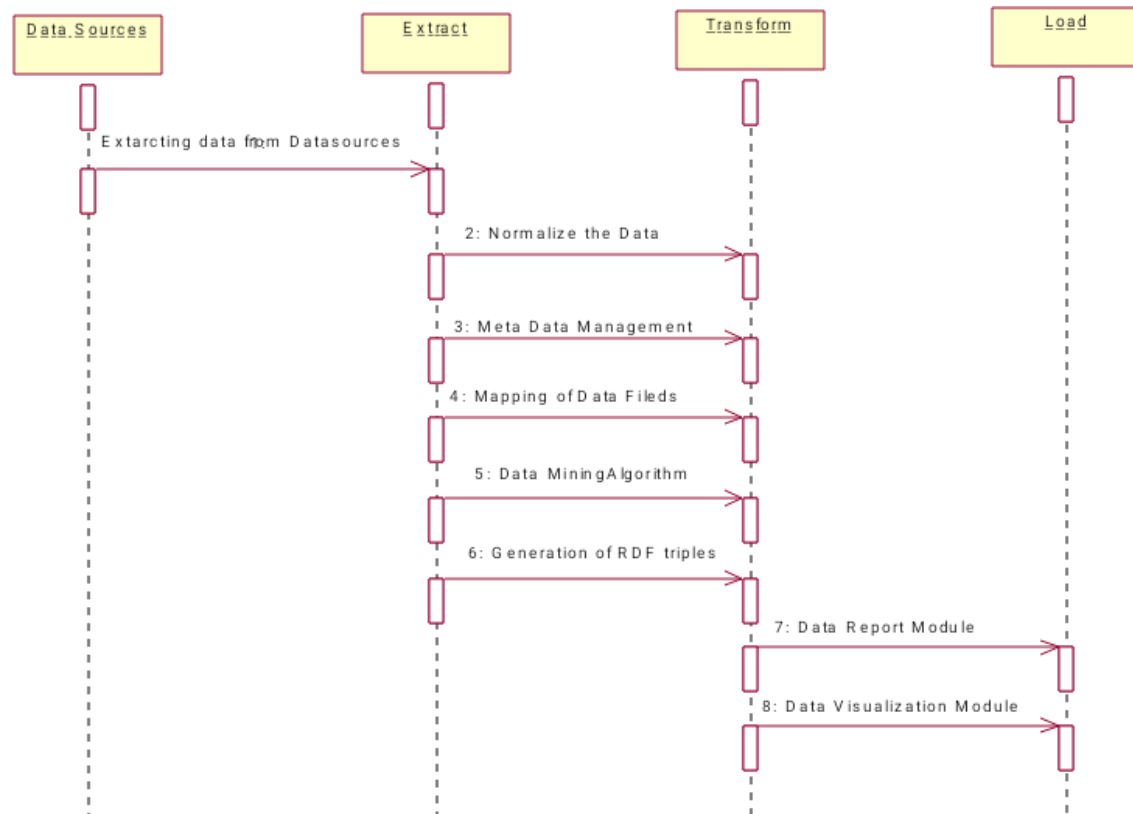
4.5.5 Data Flow Diagram Level 2 for Accuracy Prediction

4.5.6 Usecase Diagram



From the above diagram shows that it provides a global view of the actors involved in a system and the actions that the system performs, which in turn provide an observable result that is of value to the actors. Actors are an abstraction of the entity that interacts with the software. Here the system is partitioned simpler functions. The use case diagram of this software is as follows.

4.5.7 Sequence Diagram of module1



In the above sequence diagram, the entities used are data sources, extract, transform, load. The data source extract data from database which normalize the data and metadata management that performs mapping of data files. Data Mining Algorithm were used to generate RDF triple subnet dataset which transforms the data rate.

CHAPTER- 5

5.1 RESULTS AND DISCUSSION

A total of 14 202 tweets were collected and based on the features only 1897 tweets were selected for training and testing. We made use of 34 different features. It is seen from the result that different algorithms give different accuracy on the same dataset. Table 4 compares the performance of five different machine learning classifiers Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Decision Tree, Logistic Regression and SVM.

The confusion matrix to indicate the positive and negative results of the dataset from the above five algorithms are also shown below. In the confusion matrix TN (True Negative), TP (True Positive) shows the correct predicted values and FN (False Negative), FP (False Positive) shows the error values. The Twitter data is divided into training and testing in the ratio of 7:3, and after dividing it the training data contains 1327 tweets and the testing data contains 570 tweets. In case of Multinomial Naïve Bayes algorithm out of 570 tweets 133 tweets are the error values and 437 tweets are the correct predicted values, in Bernoulli Naïve Bayes algorithm out of 570 tweets 123 tweets are the error values and 447 tweets are the correct predicted values, in Decision Tree algorithm out of 570 tweets 118 tweets are the error values and 452 tweets are the correct predicted values, in Logistic Regression algorithm out of 570 tweets 116 tweets are the error values and 454 tweets are the correct predicted values and in Support Vector Machine algorithm out of 570 tweets 118 tweets are the error values and 452 tweets are the correct predicted values

The comparative performance evaluation shows that Logistic Regression gives best accuracy with 79.65%, precision with 48.28% and a very low recall with 12.17% on the other hand Bernoulli Naïve Bayes gives accuracy of 78.42%, the highest precision with 50% and comparatively impactful recall with 30.89%. Each Classifier was evaluated against a manually annotated gold standard set producing precision, accuracy and recall statistics as shown in Table 5. As precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned.

In case of our dataset it is clearly visible that Logistic Regression have the best precision and accuracy but the worst recall, on the other hand Multinomial Naïve Bayes have the best recall but not precision and accuracy, on the other hand Bernoulli Naïve Bayes gives accuracy of 78.42%, the highest precision with 50% and comparatively impactful recall with 30.89%. However, the accuracy of Logistic Regression is similar in comparison to some ensemble learning approaches like Voting Ensemble that is 79.65% but the recall increases in case of Random Forest with 32.17% having accuracy of 78.77% and precision of 46.25%. Further, the proposed work is evaluated with some ensemble approaches.

So we applied: • Random Forest which gives an Accuracy of 78.77%, Precision of 46.25% and Recall of 32.17%. • AdaBoost which gives an accuracy of 79.47%, Precision of 47.06% and Recall of 13.91%. • Voting ensemble which gives an accuracy of 79.65%, Precision of 48.28% and Recall of 12.17%. In case of Voting Ensemble algorithm out of 570 tweets 116 tweets are the error values and 454 tweets are the correct predicted values, in AdaBoost algorithm out of 570 tweets 117 tweets are the error values and 453 tweets are the correct predicted values and in Random Forest algorithm out of 570 tweets 121 tweets are the error values and 449 tweets are the correct predicted values .Next section contains the performance evaluation and the comparative analysis of different machine learning algorithms.

5.2 PERFORMANCE ANALYSIS

PERFORMANCE ANALYSIS FOR DEEP NEURAL NETWORK AND SUPPORT VECTOR MACHINE

Here various predicting models described for ED . This work has chosen DNN – Deep Neural Network. Neural networks use randomness by design to ensure they effectively learn the function being approximated for the problem. Randomness is used because this class of machine learning algorithm performs better with it than without.

The most common form of randomness used in neural networks is the random initialization of the network weights. Although randomness can be used in other areas, here is just a short list:

- Randomness in Initialization, such as weights.
- Randomness in Regularization, such as dropout.
- Randomness in Layers, such as word embedding.
- Randomness in Optimization, such as stochastic optimization.

A deep neural network is a neural network with a certain level of complexity, a neural network with more than two layers. Deep neural networks use sophisticated mathematical modeling to process data in complex ways. A neural network, in general, is a technology built to simulate the activity of the human brain – specifically, pattern recognition and the passage of input through various layers of simulated neural connections.

Many experts define deep neural networks as networks that have an input layer, an output layer and at least one hidden layer in between. Each layer performs specific types of sorting and ordering in a process that some refer to as “feature hierarchy.” One of the key uses of these sophisticated neural networks is dealing with unlabeled or unstructured data. The phrase “deep learning” is also used to describe these deep neural networks, as deep learning represents a specific form of machine learning where technologies using aspects of artificial intelligence seek to classify and order information in ways that go beyond simple input/output protocols.

CHAPTER-6

6.1 CONCLUSION

Uses machine learning techniques for several psychological stress dimensions training on the prediction of suicide ideation. This paper utilizes six machine learning techniques to predict the presence of any or more severely intense suicide ideation of military personnel. Normalization of input data and imbalanced classification strategy facilitate the prediction of machine learning methods. The experimental results show that the techniques of MLP and SVM provide the best performance for the two predictions.

6.2 FUTURE WORK:

The future approach is to investigate tweets to identify suicidal ideation using machine learning classifiers. As most of the individuals express suicidal ideation on social media platform frequently and motivated and instigated by others. Data is collected from a popular social networking site known as Twitter.

- Clinical data, which includes patient's description about insomnia or sleep deprivation, chronic pain, high anxiety, impulsivity, bipolar disorder, hopelessness etc.
- Semantics of written Blog posts,
- Tweets and interaction within the group and others.
- Shared photos and videos from Snapchat, Instagram, Facebook and so on.

REFERENCES

- A. M. Gadermann, C. C. Engel, J. A. Naifeh, M. K. Nock, M. Petukhova, P. N. Santiago, W. Benjamin, and A. M. Zaslavsky, and R. C. Kessler, "Prevalence of DSM-IV major depression among U.S. military personnel: meta-analysis and simulation," *Military Medicine*, vol. 177, no. 8, pp.47-59, August 2012.
- L. K. Richardson, B. C. Frueh, and R. Acerno, "Prevalence estimates of combat-related PTSD: critical review," *Australian and New Zealand Journal of Psychiatry*, vol. 44, no. 1, pp.4-19, January 2010.
- W. Berger, E. S. F. Coutinho, I. Figueira, C. Marques-Portella, M. P. Luz, T. C. Neylan, C. R. Marmar, and M. V. Mendlowiz, "Rescuers at risk: a systematic review and meta-regression analysis of the worldwide current prevalence and correlates of PTSD in rescue workers," *Social Psychiatry and Psychiatric Epidemiology*, vol. 47, no. 6, pp.1001-1011, June 2012.
- N. T. Fear, M. Jones, D. Murphy, L. Hull, A. Cliversen, B. Coker, L. Machell, J. Sundin, C. Woodhead, N. Jones, N. Greenberg, S. Landau, C. Dandeker, R. J. Rona, M. Hotopf, S. Wessely
- N.S. Tzeng, C. K. Chen, T. S. Wang, H. A. Chang, Y. C. Kao, H. W. Yeh, W. S. Chiang, and S. Y. Huang, "Forensic psychiatric evaluation for military absenteeism in Taiwan," *Journal of the American Academy of Psychiatry and the Law*, vol.44, pp.352-358, September 2016.
- K. E. Bachynski, M. Canham-Chervak, S. A. Black, E. O. Dada, A. M. Millikan, and B. H. Jones, "Mental health risk factors for suicides in US Army, 2007–2008," *Injury Prevention*, vol. 18, pp.405-412

APPENDIX

A.PLAGIARISM REPORT

shaliniabishek18@gmail.com.pdf			
ORIGINALITY REPORT			
13%	11%	12%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
PRIMARY SOURCES			
1	Submitted to Indian Institute of Technology, Madras Student Paper		2%
2	Submitted to Polk State College Student Paper		1%
3	Submitted to Georgia Southern University Student Paper		1%
4	Submitted to Rutgers University, New Brunswick Student Paper		1%
5	www.scribd.com Internet Source		1%
6	Submitted to Peninsula College Student Paper		1%
7	Submitted to Southern New Hampshire University - Continuing Education Student Paper		1%
8	www.frontiersin.org Internet Source		1%
9	www.psc.isr.umich.edu Internet Source		1%

Suicide Ideation Prediction For Military Personnel Using SVM Algorithm.

Shalini.T
Department Of Information
Technology,
Sathyabama Institute Of Science And
Technology,
Chennai,India.
thangaduarishalini@gmail.com

Selciya Nayagam.M
Department Of Information
Technology,
Sathyabama Institute Of Science And
Technology,
Chennai,India.
selciyanayagam99@gmail.com

Vimali.J.S.
Department Of Information
Technology,
Sathyabama Institute Of Science And
Technology,
Chennai,India.
vimali.j@sathyabama.ac.in

Abstract— In military one of the leading cause result in death is suicide. Suicide in military have increased over the last ten years .The stress generated by the war is one of the main reasons these soldiers choose suicide. This article concludes with a set of recommendations and addressing suicide in military. "Nobody really knows why suicide rates remain about on par with that one of the nonmilitary population.Suicide among uniformed service members and veterans is on the rise, despite increased attempts to address possible causes of the issue.Over time psychologists have developed a number of models to predict which people are most likely to commit suicide.The current work reviews a number of models and techniques proposed recently and offers a Bayesian machine learning model for prediction of suicide.The purpose of this project is to stop these every preventable tragedies. This article ends with a series of guidelines and a summary of military suicide.

Keywords—Veterans, Bayesian, Suicide,Psychologists.

I. INTRODUCTION

Suicide poses a significant health burden worldwide.Suicide is one of the major cause of death in the military.Military service is compulsory in many countries,mainly due to conscription system.Because of physical training,also such responsibilities military personnels are vulnerable to psychological stress.We know that In the last few years, there has been a substantial rise in the number of suicide attempts.

However, suicide rates among service members and veterans have risen steadily over the last decade. Many of the studies merely gather data on the effectiveness of previously tried intervention approaches.. It is possible for a prediction model to achieve the defined diagnostic accuracy by using a lower threshold for discrimination tests, but this will most likely come at the expense of a lower positive predictive value.

Because most previous studies on predicted suicide or suicide attempts had small sample sizes, they were limited..Our focused contribution is in building a dataset of 2,871 posts written by military personnel corresponds to 13,997 expert annotations of suicidal risks level.Here, challenge step is to create the datasets of military personnels and we have to predict that ,what is the reason behind their suicide , we should find out the average.In this, system utilizes the machine learning techniques including deep

neural network,support vector algorithm etc.Multilayer perception Suicidal ideation was projected using six major psychological stress domains in military personnel.

II. LITERATURE SURVEY

[1]Selfreport measures that could be used in populationbased researchof adults. Measures that require an interviewer or a clinician to administer are not included,which result in 19 eligible selfreport Measures were identified; however, Batterham did not mention diagnostic accuracy. [2]At the Early Psychosis Unit, patients received inpatient specialised treatment.For 24 month periods,there is a centre for intervention and prevention. In the other clinics,they are comparing the program with mental healthcare and nonspecialist.Which is resulted in Across9.7 years, 4.1 percent intervention versus 4.5 percent normal treatment (P=.84),proposed by Harris. [3]There is many types of services like community services ,delivery system.They have implemented suicide prevention,commander guidelines for mental health centers, which is started in 1997. Some suicide prevention programs ,events based on suicide all are proposed by Knox. [4]Many cases in STBhaving onsets. Many well planned plans,within the year which will occur in the ideation process. Many post enlistment attempts were associating with perenlistment,There target is for intervention and early screening.From predeployment of PPDS survey we found that suicide ideation from prevalences

Men had 12.7 percent and women had 20.1 percent lifetime suicide attempts. [5] Genetic risk are in so much interested because of targeting prevention. Here we have 4 hundred something cases to 8 thousand something cases which come to know from the research.For attempting some events ,researches used to having some samples of attempting suicide. [6] According to previous studies, a substantial number of suicide victims visited a treatment centre at least once a month. We looked into whether or not this is the case among soldiers.

Method: 608 Regular soldiers also getting attacked and died on duty times in the Indian Army between 2007 and 2010, compared to 5,690.Frequency of care centre will contact the report based on suicide.[7]They have understood the complexity of risk for suicide through the machine learning algorithms.Using the single constellation of factors to predict the suicide in last 47 years of search. [8] There have been a

13
XXX-N-XXXX-XXXX-N-XX/SSXX-00 ©20XX IEEE

slew of papers written about this population's mental and physical health. The effects of body research must be measured. They've come to assess and synthesise previously published research after a three-year hiatus. A Research Proof Map is also available is also associated to women veterans' mental and physical health comorbidities. [19]Scale for Suicide IdeationWorst (SSIW) Joint Commission approved SSIW for inpatient and outpatient areas use: This 19item rating scale developed by Beck ,they will measure the patient's intensity attitude. They will plans to attempt suicide in depression period. The instrument is validated and reliable and takes 10 Page 1 minutes to complete.[10]Liu noted that although experiencing nightmares was significantly there is a relationship between some night based diseases Richardson and colleagues analysed a treatment-seeking group of mainly veterans.

III. SYSTEM ARCHITECTURE



IV. METHODOLOGY

The STARRS consolidated ALLArmy Survey was used to pick participants. We have collected information from a few other websites. Three portion surveys make up [6] unified AAS. The surveys are listed in depth, as well as the methods used for data collection and weighing the combined data to, and there is a correction of selection probabilities and differences that is between the sample. Here we gonna use five to six types of methodologies namely Data collection,Data preprocessing,Data visualization and discriptive Statics,Data splitting into and test sets,Deep neural network,Predicting admissions based on test.

A.Data Collection

We have to collect the datas from Telecom churn dataset and use this data to predict.The collected data will include information relevant to predicting suicidal behaviour,such as the presence of depression and feelings of hopelessness.A search was conducted using machine learning and algorithms related to suicide prediction, suicide ideation and suicide risk factors.

B.Data Pre-Processing

Data preprocessing is a format that transforms raw data into the format that is necessary. Data preprocessing can be performed using a number of techniques, including data washing, data reduction, and data integration. The data is gathered from a variety of sources, each with its own collection of attributes and formats. Since some data is duplicated and includes attributes that aren't useful, we transform the raw data into required formats.

C.Data Visualization and Descriptive Statistics

In this process,we can explain this as a graphical representation of data and information. charts,graphs, and maps,data visualization tool are all elements,which is providing an accessible way to understands trends, outliers and patterns in data.Here, extrated To verify the data's relationship to the output, it was plotted against it. Here, used for describing the basic features of data is known as discriptive statistics.

D.Data Splitting

Here, we usually going to splitup the datas int two types:Training data and Testing data.Some time we also include validating the data.We get 80% of training datas and 20% of testing datas.During this process of splitting two things will happen which is Is it possible that we have overfitted or underfitted our model? These types of problems have an effect on our predictability. Overfitting: If a model is underfitted, it does not match the training data and thus misses patterns in the data. In order for models to be applied to other data, they must be based on training data.

Name	Type	Size	Value
X	Float64	(120289, 18)	[1.0000e+00 0.0000e+00 0.0000e+00 ... 0.0000e+00 0.0000e+00 0.0000e+00 ...]
X_test	Float64	(30058, 18)	[1.0000e+00 0.0000e+00 0.0000e+00 ... 0.0000e+00 0.0000e+00 0.0000e+00 ...]
X_train	Float64	(120289, 18)	[1.0000e+00 0.0000e+00 0.0000e+00 ... 0.0000e+00 0.0000e+00 0.0000e+00 ...]
dataset	DataFrame	(120289, 12)	Columns names: Size ID, Condition, Flat number, Year, Month, Day, Time ...
test	DataFrame	(30058, 12)	Columns names: Size ID, Condition, Flat number, Year, Month, Day, Time ...
Y	Float64	(120289, 1)	[0. 0. 0. ... nan nan nan]
Y_test	Float64	(30058, 1)	[0. 0. 0. ... 0. 0. nan]
Y_train	Float64	(120289, 1)	[0. 0. 0. ... 0. 0. 0.]

E) Predicting Admission

The prediction result give you the probability of military churning out of Telecom industry for given period.We will convert that probability into binary 0 and 1. F) The Model Efficiency Assessment This is the final step in the process of assessing the efficiency of the ouc model. To check the model's accuracy, we construct a confusion matrix.

1]Deep Neural Network

Logistic regression is suitable for predicting a binary dependent variable, such as positive, negative or in this study. A Deep neural network is a neural network where we have some level of complexity, a network is with above two layers. First argument is Optimizer, this is the algorithm we use to find optimal set of weights. Here we are using Adam which belongs to SGD algorithm. Based on accuracy, we have to improve the performances of our neural network. Epoch is total number of iteration. There is no rules for error. The algorithm splits the data at each node based on variables.

```
...: print("Accuracy" + str(accuracy_score(y_test, y_pred)))
...: print(classification_report(y_test, y_pred))
...:
[[ 588 276]
 [ 285 2895]]
Accuracy: 0.9057625712804562
precision recall f1-score support
0 0.74 0.77 0.76 768
1 0.94 0.95 0.94 3088
avg / total 0.90 0.90 0.90 3856
```

2]Random Forest

Random forest is a well-known machine learning algorithm that comes under the supervised learning category. An ensemble classifier is one that employs several models of classification. Random forest is specifically recursive partitioning from the RPART package. The bootstrap aggregating technique is the training algorithm used in our random forest process. When constructing each individual tree, it employs bagging and randomness. Random forest helps to boost decision tree accuracy by reducing overfitting. It is capable of handling both classification and regression issues. The decision tree number is the hyper parameter that needs to be optimised.

3]Support Vector Machine

This SVM can be used to solve problems involving classification and regression. In an N-dimensional space, this will locate a hyperplane that specifically classifies the data points. Hyperplanes are decision lines that assist in the classification of data points. Which creates two vector and tries to classify the data based on it.

```
[[ 3 757]
 [ 1 3887]]
precision recall f1-score support
0 0.75 0.80 0.81 768
1 0.90 0.90 0.90 3088
avg / total 0.79 0.80 0.72 3856

In [46]: print("Accuracy" + str(accuracy_score(y_test, y_pred)))
Accuracy: 0.885251425688123
```

4]K-Nearest Neighbours

KNN is one of the simplest algorithm used for regression and classification problem. The distance between test data and each row of training data will be determined. The array will be sorted, in ascending order and get the top K rows. At last, it will return the predicted class. Where N defines the number of nearest neighbours and the value changes as and when the n changes. Thus all the algorithms can increase the prediction and accuracy of the churn prediction in Telecom Industry.

V. EXPERIMENTAL RESULT

In this paper, we tackle the problem of suicidal risk of military personnel. Our research can be first step toward proper intervention programs and institutional support for soldiers with mental health issues. If we enter header codes in console we get the details related to military personnels like gender, licensing, senior citizen and it's value. Now that depression is there or not. In the next step, we get the value of x and y. The proposed model is the value of the accurate in estimating number if suicide. We have collected 14,202 data based on that only 1887 data were selected for testing and training purpose. In this we used 34 different features. Different algorithm give different accuracy on same data. Each classifier was evaluated. As precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned.

```
Python console
...: print("Accuracy" + str(accuracy_score(y_test, y_pred)))
...:
[[1028 188]
 [128 340]]
Accuracy: 0.875
In [36]:
```

VI. CONCLUSION

Our research utilises machine learning techniques to predict suicidal ideation through many psychological stress dimensions. The finding also suggest that the suicidal ideation of military personnel may lessen the number of suicidal ideation of this neglected personnel. In this paper, we investigated the problem of suicidality detection in online user-generated content. This work will greatly assist in the detection of military personnel they all are at risk of committing suicide.

VII. REFERENCE

- [1] Botherhorn, Philip J., et al. "A systematic review and evaluation of measures for suicidal ideation and behaviors in population-based research." *Psychological assessment* 27.2 (2015): 503.
- [2] Harris, Russell P., et al. "Current methods of the US Preventive Services Task Force: a review of the process." *American journal of preventive medicine* 20.3 (2001): 21-35.
- [3] Koss, Kerry L., et al. "Risk of suicide and related adverse outcomes after exposure to a suicide prevention programme in the US Air Force cohort study." *BMJ* 327 7428 (2003): 1376.
- [4] Nock, Matthew K., et al. "Risk factors for the transition from suicide ideation to suicide attempt: Results from the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS)." *Journal of abnormal psychology* 127.2 (2018): 139.
- [5] Stein, Murray B., et al. "Genomewide association studies of suicide attempts in US soldiers." *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 174.8 (2017): 786-797.
- [6] Ursano, Robert J., et al. "The Army study to assess risk and resilience in servicemembers (Army STARRS)." *Psychiatry: Interpersonal and Biological Processes* 77.2 (2014): 107-119.
- [7] Franklin, Joseph C., et al. "Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research." *Psychological Bulletin* 143.2 (2017): 187.
- [8] Coombs, Suzanne K., et al. "Clinical complexity in women veterans: A systematic review of the recent evidence on mental health and health comorbidities." *Behavioral Medicine* (2019): 1-19.
- [9] Beck, AT, Brown, GK, Steer, RA. Psychometric characteristics of the Scale for Suicide Ideation with psychiatric outpatients. *Behaviour* 1997.
- [10] Lee, X. 2004 Sleep and adolescent suicidal behavior. *Sleep*, 27(7), pp.1331-1338.

Machine Learning Based Suicide Ideation Prediction for Military Personnel

Gen-Min Lin, *Member, IEEE*, Masanori Nagamine, Szu-Nian Yang, Yueh-Ming Tai, Chin Lin, Hiroshi Sato

Abstract—Military personnel have greater psychological stress and are at higher suicide attempt risk compared with the general population. High mental stress may cause suicide ideations which are crucially driving suicide attempts. However, traditional statistical methods could only find a moderate degree of correlation between psychological stress and suicide ideation in non-psychiatric individuals. This paper utilizes machine learning techniques including logistic regression, decision tree, random forest, gradient boosting regression tree, support vector machine and multilayer perceptron to predict the presence of suicide ideation by six important psychological stress domains of the military males and females. The accuracies of all the six machine learning methods are over 98%. Among them, the multilayer perceptron and support vector machine provide the best predictions of suicide ideation approximately to 100%. As compared with the BSRS-5 score ≥ 7 , a conventional criterion, for the presence of suicide ideation ≥ 1 , the proposed algorithms can improve the performances of accuracy, sensitivity, specificity, precision, the AUC of ROC curve and the AUC of PR curve up to 5.7%, 35.9%, 4.6%, 65.2%, 4.3% and 53.2%, respectively; and for the presence of more severely intense suicide ideation ≥ 2 , the improvements are 6.1%, 26.2%, 5.8%, 83.5%, 2.8% and 64.7%, respectively.

Index Terms—Machine Learning Techniques, Psychological Stress, Suicide Ideation

This work was supported by Ministry of Science and Technology, Taiwan, under grant MOST 107-2221-E-809-002-MY3.

Gen-Min Lin is with the Department of Internal Medicine, Hualien Armed Forces General Hospital, Hualien 97144, Taiwan; Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei 11490, Taiwan; Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA. (e-mail: jamner507@yahoo.com.tw or gen-min.lin@northwestern.edu).

Masanori Nagamine is with the Division of Behavioral Science, National Defense Medical College Research Institute, Tokorozawa City, Saitama 359-8513, Japan. (e-mail: maganori.nagamine@gmail.com).

Szu-Nian Yang is with the Department of Psychiatry, Tri-Service General Hospital-Beitou Branch, National Defense Medical Center, Taipei 11490, Taiwan. (e-mail: ysn56725@ms4.hinet.net).

Yueh-Ming Tai is with the Department of Psychiatry, Tri-Service General Hospital-Beitou Branch, National Defense Medical Center, Taipei 11490, Taiwan. (e-mail: Samytai@hotmail.com).

Chin Lin is with the Graduate Institute of Life Sciences, National Defense Medical Center, Taipei 11490, Taiwan. (e-mail: xup6lap0629@gmail.com).

Hiroshi Sato is with the Artificial Intelligence Laboratory, Department of Computer Science, National Defense Academy, Yokosuka, Kanagawa 239-8686, Japan. (e-mail: h.sato@nda.ac.jp).

I. INTRODUCTION

MILITARY personnel are vulnerable to psychological stress because of tense physical training, multiple deployments and responsibilities. The prevalence of major depression was reported, ranging from 2.0% to 37.4% in the US military [1], and that of combat-related posttraumatic stress disorder (PTSD) was reported 2.0%-17.0% among US military war veterans [2]. A meta-analysis showed consistent results that the worldwide pooled prevalence of PTSD in rescue workers was up to 10.0% [3]. The symptoms of mental disorders developed frequently in those of continued combat exposure and those of repeated deployments [4]. The association between military absenteeism and mental health problems has been discussed in [5]. The rate of suicide attempt among active duty US Army personnel has been increasingly higher than that in the civilians [6]. According to an analysis for 27,501 military participants in [7], 14.3% of survey respondents reported suicide ideation and 3.0% committed suicide. In other words, 21% of those with suicide ideation had suicide attempt. As is known, previous studies have revealed a relationship between suicide ideation and psychological stress [8],[9]. To early predict the presence of suicide ideation and further prevent the behavior of suicide are essential and important in the military.

With the technology improvement and the availability of various kinds of real world big data, artificial intelligence (AI) grows fast accordingly. The academics have made great efforts on the computerized algorithms to deal with big data. Machine learning, a combination of AI and computations, could provide accurate diagnosis of diseases and predict the outcomes [10]-[17]. For instance, the circuits for seizure classification and detection by machine learning are implemented in [18]. Recognition of heart murmurs could be achieved by deep neural networks [19]. In addition, Ambale-Venkatesh *et al.* identified the top-20 risk factors of incident cardiovascular events by the random survival forest which performance was better than the conventional risk calculators [20]. Therefore, using machine learning and deep learning techniques has become an efficient and reliable tool for clinical practice by physicians globally.

High mental stress may cause suicide ideations which are crucially driving suicide attempts. However, traditional statistical methods find merely a moderate correlation between psychological stress and suicide ideation. Machine learning



Fig. 1. Schematic diagram of proposed method

could provide better performance of the prediction of suicide ideations. In this paper, we utilize a large sample of the military members for several machine learning techniques by taking the psychological stress dimensions into consideration to predict the presence of suicide ideation.

The schematic diagram of the proposed method in this paper is illustrated in Fig. 1. A binary probabilistic classifier of machine learning algorithm can determine whether the military persons, through their questionnaires, have suicide ideations. Machine learning provides an effective manner for early warning and prevention of suicide by automatic suicide ideation detection.

The rest of this paper is organized as follows. The materials are shown in Section II. In Section III, the proposed algorithms by machine learning techniques are demonstrated. Section IV shows the experimental results. Section V concludes this paper.

II. MATERIALS

This study used a historical cohort of 3,546 military men and women aged 18-50 years, with an average of 29 years of age from the cardiorespiratory fitness and hospitalization events in armed forces (CHIEF) study performed in the Hualien Armed Forces General Hospital, the only one military referral hospital of Eastern Taiwan, in 2014. All participants carried out a formal health examination including a blood routine, biochemical tests and chest X-ray. The participants self-reported a questionnaire for their mental status and experience of substance use as well. The study design of CHIEF study has been described in detail previously [21]-[26].

TABLE I
FIVE-POINT LIKERT-TYPE SCALE OF FIVE PSYCHOPATHOLOGICAL DOMAINS OF THE BSRS-5 SCORE AND SUICIDE IDEATION

	Not at all	A little bit	Moderately	Quite a bit	Extremely
Five-point Likert-type Scale	0	1	2	3	4

TABLE II
STATISTICS OF THE BSRS-5 SCORE AND THE FIVE PSYCHOPATHOLOGICAL DOMAINS

	BSRS-5	Anxiety > 1	Depression > 1	Hostility > 1	Interpersonal Sensitivity > 1	Insomnia > 1
Total (N=3546)	1.91±2.80	5.30%	5.22%	7.16%	4.23%	8.21%
Suicide Class 0 (N=3418)	1.66±2.42	3.89%	3.77%	5.71%	3.25%	6.82%
Ideation ≥ 1 Class 1 (N=128)	8.71±3.59	42.97%	43.75%	46.09%	30.47%	45.31%
Suicide Class 0 (N=3504)	1.81±2.62	4.57%	4.65%	6.65%	3.80%	7.65%
Ideation ≥ 2 Class 1 (N=42)	10.45±3.83	66.67%	52.38%	50.00%	40.48%	54.76%

The matrices of the suicide ideation severity against the scales of the BSRS-5 score and each psychopathological domain are exhibited in Fig. 2. For the extreme of suicide

ideation 4, there is no obviously higher scales in the BSRS-5 score and each psychopathological domain.



balance the majority data of those without suicide ideations. The six proposed machine learning algorithms are individually trained by optimizing the corresponding hyperparameters.

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information. Authorized licensed use limited to: University of Canberra. Downloaded on May 01, 2020 at 11:40:31 UTC from IEEE Xplore. Restrictions apply.

It has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/BHI.2020.2988295, IEEE Biomedical and Health Informatics

III. PROPOSED METHODS

The six input factors of psychological stress for machine learning include BSRS-5 score, anxiety, depression, hostility, interpersonal sensitivity and insomnia. This paper uses six machine learning techniques including logistic regression (LR), decision tree (DT), random forest (RF), gradient boosting decision tree (GBDT), support vector machine (SVM) and multilayer perceptron (MLP) for the prediction of the presence of suicide ideation of the military members. The system diagram of proposed method is illustrated in Fig. 3.

A. Data Pre-Processing

To solve the phenomenon of different dynamic ranges for the six input variables, we apply the normalization of Min-Max scaling [27], [28] to normalize input data into the interval 0–1. Min-Max normalization executes a linear transformation on the original data. Each of the actual data d of feature x is mapped to a normalized value adjusted in the range of 0 to 1 as Eq.(1).

$$\text{Normalized}(d) = d' = \frac{d - \min(x)}{\max(x) - \min(x)}, \quad (1)$$

where $\min(x)$ and $\max(x)$ denote the minimum and maximum values of the input feature x , respectively. d' represents the normalized data.

10-fold cross validation is utilized in this paper as shown in Fig. 4. The data numbers illustrated by one fold are detailed in Table III. For the prediction of the presence of any suicide ideation (suicide ideation ≥ 1), the numbers for training and test sets are 3191 (class 0: 3080, class 1: 111) and 355 (class 0: 338, class 1: 17), respectively. For the prediction of more severely intense suicide ideation (suicide ideation ≥ 2), the numbers for training and test sets are 3191 (class 0: 3157, class 1: 34) and 355 (class 0: 347, class 1: 8), respectively. This imbalance in the dataset between class 0 and class 1 is obvious. This problem is addressed by applying the synthetic minority over-sampling technique (SMOTE) [29]. The training data for class 1 are pre-processed by SMOTE to 3080 and 3157 for the two predictions, respectively, as shown in Table III.



Fig. 4. Data partition of 10-fold cross validation for training and test sets

TABLE III
NUMBERS OF TRAINING AND TEST DATA

		Class 0	Class 1	Total
Suicide Ideation ≥ 1	Training Set	3080	111	3191
	Pre-processed by SMOTE	3080	3080	6160
Suicide Ideation ≥ 2	Training Set	3157	34	3191
	Pre-processed by SMOTE	3157	3157	6314
	Test Set	347	8	355

B. Machine Learning Models

1. Logistic regression

Logistic regression (LR) [30], a classification algorithm used to assign observations to a discrete set of classes, is applied in our method and illustrated in Fig. 5. Logistic regression (LR) is a linear model defined as Eq. (2)-(3).

$$z = w_0x_0 + w_1x_1 + \dots + w_nx_n + w_{n+1}, \quad (2)$$

$$y = \text{sigmoid}(x) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}, \quad (3)$$

where $x_1 - x_n$ denote the input variables and $w_0 - w_n$ are the weights being learned. n is 6 for BSRS-5 and its five psychopathological domains. Logistic regression transforms its output y using the logistic sigmoid function to return a probability value.

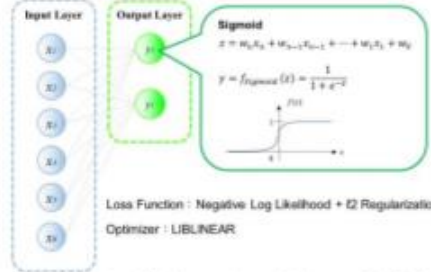


Fig. 5. Illustration of logistic regression used in the proposed method

The loss function consists of loss term and regularization term. The loss term for learning the weight vector is negative log-likelihood as Eq. (4)-(5). m denotes the sample size. X_i indicates the vector of features and Y_i means the class. The regularization term l_2 -norm is used in our method as Eq. (6) to penalize large weights in order to prevent overfitting. The optimized hyperparameter α is chosen in our algorithm by grid search. We minimize $J_{l_2}(w)$ to find the optimal weight vector w by LIBLINEAR (A Library for Large Linear Classification) [31].

$$L(w) = \sum_{i=1}^m P(Y_i = 1|X_i, w)^{Y_i} (1 - P(Y_i = 1|X_i, w))^{1-Y_i}, \quad (4)$$

$$m = 6160, 6314 \text{ for Suicide Ideation } \geq 1, 2,$$

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information. Authorized licensed use limited to: University of Canberra. Downloaded on May 01, 2020 at 11:40:31 UTC from IEEE Xplore. Restrictions apply.

It has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/BHI.2020.2988295, IEEE Biomedical and Health Informatics

$$J_0(w) = -\frac{1}{m} \log(L(w)). \quad (5)$$

(bagging) technique. RF builds multiple CART models with different samples and different initial variables. In each

The matrices of the suicide ideation severity against the scales of the BSR5-5 score and each psychopathological domain are exhibited in Fig. 2. For the extreme of suicide

ideation 4, there is no obviously higher scales in the BSR5-5 score and each psychopathological domain.

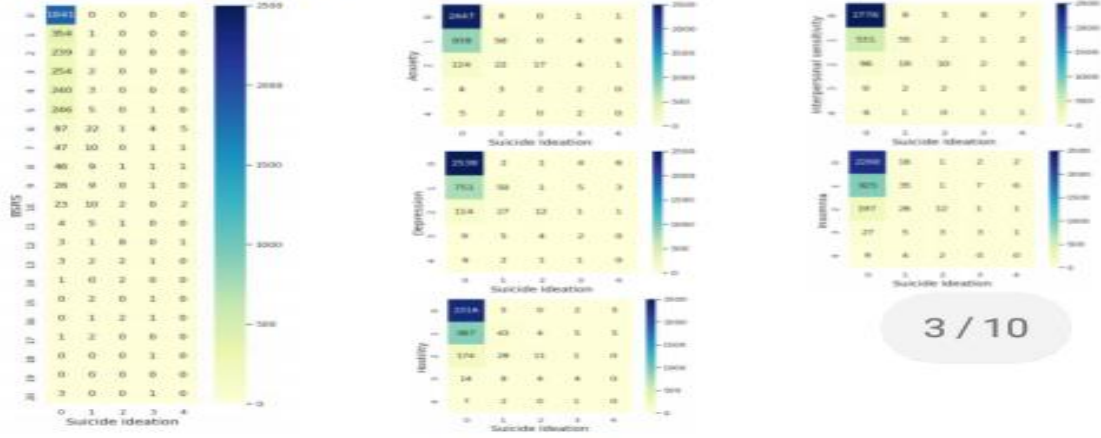


Fig. 2. Data distribution of suicide ideation vs. psychological stress

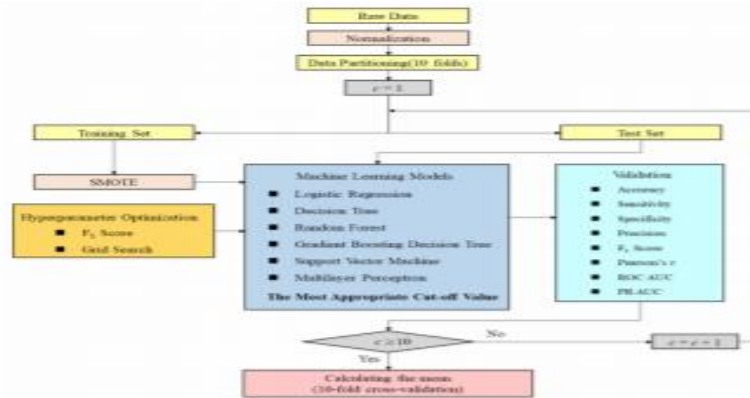


Fig. 3. The flowchart of the proposed method. The input data are pre-processed by mini-max normalization and then partitioned to training and test sets for 10-fold cross validation. The synthetic minority over-sampling technique (SMOTE) for the minority data (those with suicide ideations) in training set is performed to balance the majority data of those without suicide ideations. The six proposed machine learning algorithms are individually trained by optimizing the corresponding hyperparameters.

III. PROPOSED METHODS

The six input factors of psychological stress for machine learning include BSR5-5 score, anxiety, depression, hostility, interpersonal sensitivity and insomnia. This paper uses six machine learning techniques including logistic regression (LR).

TABLE III
NUMBERS OF TRAINING AND TEST DATA

		Class 0	Class 1	Total
Suicide Ideation ≥ 1	Training Set	3080	111	3191
	Pre-processed by SMOTE	3080	3080	6160
	Test Set	338	17	355

$$J_0(w) = -\frac{1}{m} \log(L(w)), \quad (5)$$

$$J_{reg}(w) = J_0(w) + \alpha \sum_{i=0}^n w_i^2. \quad (6)$$

2. Decision tree

Decision tree (DT) [32], which establishes classification models in the form of a tree structure, with CART (Classification and Regression Tree) [33] algorithm is used in our method and shown in Fig. 6.



Fig. 6. Illustration of decision tree used in the proposed method

Each root node represents one of the six features and a split point on that feature. DT partitions the data (the parent node) into two subsets (the child nodes) by minimizing $Gini(T)$ impurity shown as Eq. (7)-(8) to create decision points for classification task.

$$Gini(T) = \sum_{i=1}^2 \frac{N_i}{N} gini(D_i), \quad (7)$$

$$gini(D_i) = 1 - \sum_{j=1}^2 p_j^2, \quad (8)$$

where T indicates one of the six features. N denotes the total sample number of the parent node. N_i represents the sample number falling on the i -th subset D_i (child node) and p_j indicates the percentage of each category j in D_i after T classification. The two resulting subsets change into the new parent nodes, which are subsequently separated further into two child nodes. This procedure continues until all leafs are classified. The leaf nodes of the tree contain an output prediction (class 0 or class 1). A simpler tree is built by pre-pruning processing to shorten the branches of the tree and avoid over-fitting. The maximum tree depth is the hyperparameter optimized in our algorithm.

3. Random forest

Random forest (RF) [34], an ensemble machine learning technique, constructs multiple decision trees and collects them together for classification. The training algorithm adopted in our method for random forest is the bootstrap aggregating

(bagging) technique. RF builds multiple CART models with different samples and different initial variables. In each decision tree, a random subset of the features is taken into consideration for splitting a node. The individual trees are not correlated with each other and thus the trees in random forest of our method are not pruned. The final prediction result is according to the majority-votes model from the multiple DTs. RF combines the merits of feature selection and bagging. The decision tree number is the hyperparameter to be optimized.

4. Gradient boosting decision tree

Gradient boosting decision tree (GBDT) [35] is also an ensemble machine learning method and constructs multiple additive decision tree models. The DTs fitting the gradient on pseudo residuals of previous cumulative models are repeatedly trained to minimize mean squared error. This sequential stepwise manner combines the performance of weak learners (i.e., DT here) in an iterative fashion into a single strong learner to increase the accuracy of prediction. Our algorithm uses the maximum tree depth as the hyperparameter to be optimized to avoid over-fitting.

5. Support vector machine

Support vector machine (SVM) [36] with linear kernel (Linear SVM) is used for our proposed method. A data point is viewed as a 6-dimensional vector and we separate such points with a hyperplane. This linear SVM constructs the maximum-margin hyperplane so that the distance from it to the nearest training data point of any class (class 0 or class 1) is maximized.

If the training set is not linearly separable, soft-margin SVM allows the fat decision margin and some outliers inside or on the wrong side of soft-margin SVM, which minimize against margin. Regularization, regularization term aims to fit training set data and over-fitting. ℓ_2 -norm is utilized in SVM for our method. The regularization hyperparameter is optimized in our algorithm to control overfitting.

6. Multilayer perceptron

Multilayer perceptron (MLP) [37] consists of an input layer, hidden layers and an output layer for our algorithm as illustrated in Fig. 7. In fully connected MLP, each node in one layer connects with a certain weight to every node in the following layer as shown in Eq. (9)-(10).

$$v_i = u_{i0}x_n + \dots + u_{i1}x_1 + u_{i0}, \quad (9)$$

$$h_i^k = f_{relu}(v_i), \quad (10)$$

where n represents the number of input or the number of the neuron in previous hidden layer; v represents the weighted sum of the connections. h_i^k denotes the output of the i -th node (neuron) of the k -th hidden layer. The activation function rectifier linear unit (Relu) is used for each node in hidden layers. The output layer y is determined by the logistic regression function.



TABLE V
PERFORMANCE COMPARISONS OF SIX MACHINE LEARNING METHODS AND PREVIOUS WORK FOR THE BSRS-5 SCORE ≥ 7

		Accuracy	Sensitivity	Specificity	Precision	F ₁ score	ϕ	ROC AUC	PR AUC
Suicide Ideation ≥ 1	Logistic Regression	99.8%	100%	99.8%	96.3%	98.0%	81.9%	99.9%	97.9%
	Decision Tree	98.4%	77.7%	99.2%	80.7%	78.2%	71.7%	88.4%	79.6%
	Random Forest	98.7%	87.4%	99.1%	80.6%	82.8%	74.1%	97.7%	83.9%
	Gradient Boosting Decision Tree	99.0%	87.7%	99.4%	86.5%	86.6%	77.3%	98.2%	81.5%
	Support Vector Machine	100%	100%	100%	100%	100%	88.7%	100%	100%
	Multilayer Perceptron	100%	100%	100%	100%	100%	87.4%	100%	100%
	BSRS-5 Score ≥ 7 [9]	94.1%	64.1%	95.4%	74.8%	44.4%	42.4%	95.7%	46.8%
Suicide Ideation ≥ 2	Logistic Regression	99.9%	100%	99.9%	94.1%	96.7%	88.0%	99.9%	93.1%
	Decision Tree	98.9%	78.8%	99.2%	56.4%	61.2%	53.9%	90.4%	43.4%
	Random Forest	99.5%	90.8%	99.6%	71.3%	77.9%	69.2%	99.5%	71.6%
	Gradient Boosting Decision Tree	99.6%	90.8%	99.2%	80.1%	84.7%	69.2%	98.7%	81.3%
	Support Vector Machine	99.9%	100%	99.9%	96.6%	98.1%	87.1%	99.9%	93.3%
	Multilayer Perceptron	99.9%	100%	99.9%	96.7%	98.0%	90.8%	99.9%	97.9%
	BSRS-5 Score ≥ 7 [9]	93.8%	73.8%	94.1%	13.2%	22.1%	42.4%	97.1%	33.2%

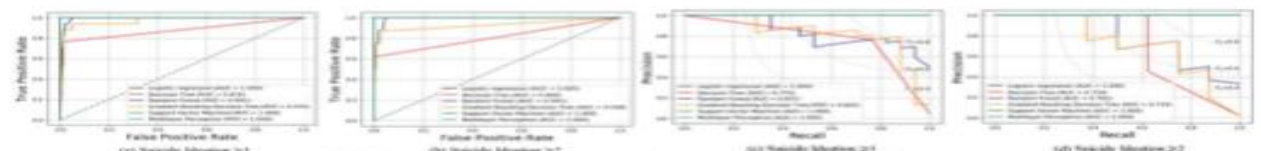


Fig. 8. Receiver Operating Characteristic Curves and Precision Recall Curves

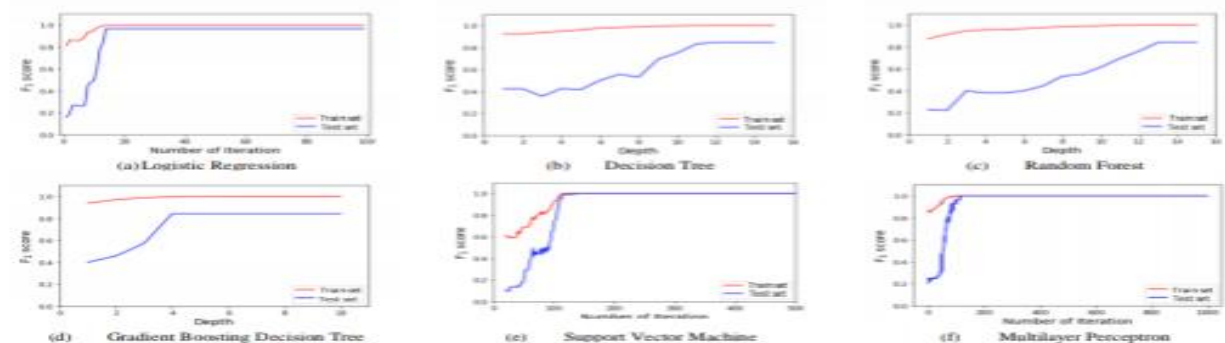


Fig. 9. A set of learning curves for each of the six machine learning methods

as the average of the 10-fold cross validation in Table V. The p values for all techniques are <0.0001 . The accuracies for all of the six machine learning methods are over 98%. The results show that the MLP and SVM provide the best predictions. The second is the LR, and then the three tree-based methods. The RF and GBDT yield overall better results than that of the DT. The prediction results for suicide ideation ≥ 1 are better than those of suicide ideation ≥ 2 .

The ROC curves and the PR curves and the corresponding AUCs are shown in Fig. 8, for one of the 10-fold cross validation. The areas under ROC and PR curves reach to 1 for

MLP and SVM. The tendency is similar with the results shown in Table V. The learning curves for the six proposed methods for suicide ideation ≥ 1 are shown in Fig. 9. The learning curves depict an improvement in F₁ score when there are changes in the number of iteration (logistic regression, support vector machine and multilayer perceptron) or tree depth (decision tree, random forest, gradient boosting regression tree). All curves for the six proposed methods exhibit convergence well.

The distributions of variable importance for six input parameters except MLP are shown in Fig. 10. The average values of feature importance of the 10-fold cross validation are



validation. The areas under ROC and PR curves reach to 1 for values of feature importance of the 10-fold cross validation are

© (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information. Authorized licensed use limited to: University of Canberra. Downloaded on May 01, 2020 at 11:40:31 UTC from IEEE Xplore. Restrictions apply.

been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2000000, IEEE Jour
Biomedical and Health Informatics

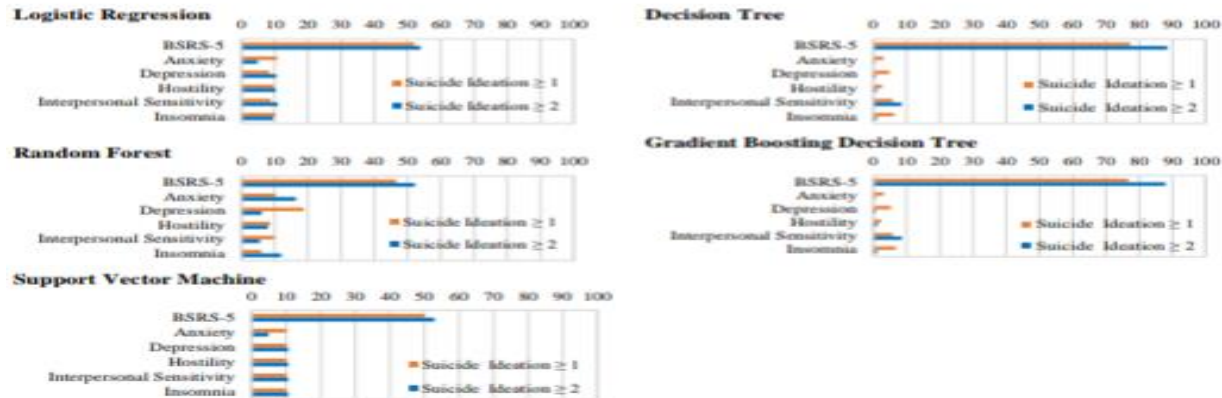


Fig. 10. Feature Importance for the machine learning methods in the proposed algorithm

TABLE VI
FEATURE IMPORTANCE FOR SIX INPUT PARAMETERS OF SIX MACHINE LEARNING METHODS

		BSRS-5	Anxiety	Depression	Hostility	Interpersonal Sensitivity	Insomnia
Suicide Ideation ≥ 1	Logistic Regression	52.15%	10.71%	8.41%	9.66%	8.83%	10.24%
	Decision Tree	77.33%	3.08%	4.88%	2.59%	5.83%	6.29%
	Random Forest	46.60%	10.19%	18.71%	8.54%	9.98%	5.98%
	Gradient Boosting Decision Tree	76.76%	3.11%	5.29%	2.25%	5.83%	6.76%
	Support Vector Machine	50.00%	10.00%	10.00%	10.00%	10.00%	10.00%
Suicide Ideation ≥ 2	Logistic Regression	53.89%	4.93%	10.44%	10.26%	10.89%	9.59%
	Decision Tree	88.54%	0.41%	0.67%	0.93%	8.46%	0.99%
	Random Forest	52.19%	16.36%	6.05%	7.95%	5.52%	11.93%
	Gradient Boosting Decision Tree	87.98%	0.47%	0.81%	1.11%	8.55%	1.08%
	Support Vector Machine	52.86%	4.86%	10.57%	10.57%	10.57%	10.57%

isted in Table VI. All of the five machine learning techniques take the BSR5-5 score as the most important variable and the proportion is even larger for the case of suicide ideation ≥ 2 . For both of DT and GBDT, the anxiety dimension has less importance compared to other three methods.

For the screening instrument to predict suicide ideation proposed in [8], the sensitivities for psychiatric group, community group and general medical group are 83.76%, 21.57%, and 10.57%, respectively, and the specificities for the three groups are 72.17%, 99.49%, and 99.88%, respectively. As compared with the BSR5-5 score ≥ 7 in [9], a conventional criterion, for the presence of suicide ideation ≥ 1 , the proposed algorithms can improve the performances of accuracy, sensitivity, specificity, precision, the AUC of ROC curve and the AUC of PR curve up to 5.7%, 35.9%, 4.6%, 65.2%, 4.3% and 53.2%, respectively; and for the presence of more severely intense suicide ideation ≥ 2 , the improvements are 6.1%, 26.2%, 5.8%, 83.5%, 2.8% and 64.7%, respectively. Instead of only considering the BSR5-5 score in screen of suicide ideation like [8] and [9], our algorithm additionally takes the five psychopathological domains which are related to the BSR5-5 score, i.e., anxiety, depression, hostility, interpersonal

sensitivity and insomnia as the input variables, and our schemes incorporating machine learning techniques provide better results than those of [8] and [9].

In addition, we add several critical physiological data including age, sex, body height, body weight, waist circumference, heart rate, systolic blood pressure, diastolic blood pressure and physical activity on the initial inputs of the BSR5-5 score and related five psychopathological domains in our proposed models. We find that the performances are only improved by incorporating these nine physiological data into the model of logistic regression. All of the performances of logistic regression regarding accuracy, sensitivity, specificity and precision reach 100% for suicide ideation ≥ 1 and 99.9% for suicide ideation ≥ 2 . However, for the other five machine learning methods, the performances are not better with additional inputs of these physiological data.

As the incidence of suicide is very low, a meta-analysis reveals that the utility of machine learning for predicting later suicide is limited by low predictive value and modest sensitivity [41]. Machine learning techniques for the BSR5-score and related five psychopathological domains can be aimed for later suicide attempts in future work.

© (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information. Authorized licensed use limited to: University of Canberra. Downloaded on May 01, 2020 at 11:40:31 UTC from IEEE Xplore. Restrictions apply.

been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2000000, IEEE Jour
Biomedical and Health Informatics

V. CONCLUSION

Our study uses machine learning techniques for several

- [12] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease," *IEEE*

intense suicide ideation (≥ 4), the improvements are 0.17%, 20.27%, 5.8%, 83.5%, 2.8% and 64.7%, respectively. Instead of only considering the BSRS-5 score in screen of suicide ideation like [8] and [9], our algorithm additionally takes the five psychopathological domains which are related to the BSRS-5 score, i.e., anxiety, depression, hostility, interpersonal

as the incidence of suicide attempts is relatively low, a meta-analysis reveals that the utility of suicide ideation for predicting later suicide is limited by low positive predictive value and modest sensitivity [41]. Machine learning techniques for the BSRS-score and related five psychopathological domains can be aimed for later suicide attempts in future work.

V. CONCLUSION

Our study uses machine learning techniques for several psychological stress dimensions training on the prediction of suicide ideation. This paper utilizes six machine learning techniques to predict the presence of any or more severely intense suicide ideation of military personnel. Normalization of input data and imbalanced classification strategy facilitate the prediction of machine learning methods. The experimental results show that the techniques of MLP and SVM provide the best performance for the two predictions. Most of the proposed machine learning techniques take the BSRS-5 score to be the most pivotal variable and the five psychopathological dimensions are also adopted as the features to improve the screening. As compared with the prior study using the BSRS-5 score only, the machine learning techniques can improve the performances of predicting suicide ideation. This work can substantially help to screen out the military personnel at high suicide risk for suicide prevention.

REFERENCES

- [1] A. M. Gadermann, C. C. Engel, J. A. Naifeh, M. K. Nock, M. Petukhova, P. N. Santiago, W. Benjamin, and A. M. Zaslavsky, and R. C. Kessler, "Prevalence of DSM-IV major depression among U.S. military personnel: meta-analysis and simulation," *Military Medicine*, vol. 177, no. 8, pp. 47-59, August 2012.
- [2] L. K. Richardson, B. C. Frueh, and R. Acerno, "Prevalence estimates of combat-related PTSD: critical review," *Australian and New Zealand Journal of Psychiatry*, vol. 44, no. 1, pp. 4-19, January 2010.
- [3] W. Berger, E. S. F. Coutinho, I. Figueira, C. Mangens-Ponella, M. P. Luz, T. C. Neylan, C. R. Marmar, and M. V. Mendlowitz, "Rescuers at risk: a systematic review and meta-regression analysis of the worldwide current prevalence and correlates of PTSD in rescue workers," *Social Psychiatry and Psychiatric Epidemiology*, vol. 47, no. 6, pp. 1001-1011, June 2012.
- [4] N. T. Fear, M. Jones, D. Murphy, L. Hull, A. Clervsen, B. Coker, L. Maclell, J. Sandin, C. Woodhead, N. Jones, N. Greenberg, S. Lundan, C. Dandekar, R. J. Rona, M. Hotopf, S. Wessely, "What are the consequences of deployment to Iraq and Afghanistan on the mental health of the UK armed forces? A cohort study," *The Lancet*, vol. 375, no. 9728, pp. 1783-1797, May 2010.
- [5] N. S. Tzeng, C. K. Chen, T. S. Wang, H. A. Chang, Y. C. Kao, H. W. Yeh, W. S. Chiang, and S. Y. Huang, "Forensic psychiatric evaluation for military absenteeism in Taiwan," *Journal of the American Academy of Psychiatry and the Law*, vol. 44, pp. 352-358, September 2016.
- [6] K. E. Bachynski, M. Canham-Chervak, S. A. Black, E. O. Dada, A. M. Millikan, and B. H. Jones, "Mental health risk factors for suicides in US Army, 2007-2008," *Injury Prevention*, vol. 18, pp. 405-412, December 2012.
- [7] K. L. Zuremaki, S. L. Bernecker, P. M. Gutierrez, T. E. Joiner, A. J. King, H. Liu, J. A. Naifeh, M. K. Nock, N. A. Sampson, A. M. Zaslavsky, M. R. Stein, R. J. Ursano, and R. C. Kessler, "Assessment of a risk index for suicide attempts among US Army soldiers with suicidal ideation," *JAMA Network Open*, vol. 2, no. 3, e190766, March 2019.
- [8] F. W. Lung and M. B. Lee, "The five-item brief symptom rating scale as a suicidal ideation screening instrument for psychiatric inpatients and community residents," *BMC Psychiatry*, vol. 8, no. 53, July 2008.
- [9] C. C. Ma, and Y. M. Tai, "Cut-off values of five-item brief symptom rating scale in evaluating suicidality among military recruits," *Taiwanese Journal of Psychiatry*, vol. 28, no. 2, pp. 109-120, June 2014.
- [10] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule extraction from support vector machines using ensemble learning approach: An application for diagnosis of diabetes," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 2, pp. 728-734, March 2015.
- [11] M. Shahin, B. Ahmed, S. T. B. Hamida, F. L. Mulafier, M. Glos, and T. Penzel, "Deep learning and ensemble: assisting clinicians with their diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 6, pp. 1546-1553, November 2017.
- [12] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 173-183, January 2018.
- [13] B. Lei, Peng Yang, Y. Zhao, F. Zhou, D. Ni, S. Chen, X. Xian, and T. Wang, "Deep learning for brain disease diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1661-1673, July 2019.
- [14] G. M. Lin, M. J. Chen, C. H. Yeh, Y. Y. Lin, H. Y. Kuo, M. H. Lin, and M. C. Chen, S. D. Lin, Y. Guo, A. Ran, and C. Y. Cheung, "Transforming retinal photographs to entropy images in deep learning to improve automated detection for diabetic retinopathy," *Journal of Ophthalmology*, vol. 2018, Article ID 2159702, September 2018.
- [15] X. Du, R. Tang, S. Yan, Y. Zhang, and S. Li, "Direct segmentation-based full quantification for left ventricle via deep multi-task regression learning network," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 942-948, May 2019.
- [16] A. Akay and H. Hess, "Deep learning: current and emerging applications in medicine and technology," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 906-920, May 2019.
- [17] M. J. Chen, P. H. Yan, M. T. Hsieh, C. H. Yeh, C. H. Huang, C. M. Yang, and G. M. Lin, "Machine learning of PM2.5 and PM10 concentrations to relate with outpatient visits for upper respiratory tract infections in Taiwan: a nationwide analysis," *World Journal of Clinical Cases*, vol. 6, no. 8, pp. 200-206, August 2018.
- [18] L. Feng, Z. Li and Y. Wang, "VLSI design of SVM-based seizure detection system with on-chip learning capability," *IEEE Transactions on Circuits and Systems*, vol. 12, no. 1, pp. 171-181, February 2018.
- [19] J. P. Dominguez-Morales, A. F. Jimenez-Fernandez, M. J. Dominguez-Morales, and G. Jimenez-Moreno, "Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 1, pp. 24-34, February 2018.
- [20] B. Ambale-Venkatesh, X. Yang, C. O. Wu, K. Liu, W. G. Hundley, R. McClelland, A. S. Gomes, A. R. Folsom, S. Shea, E. Guallar, D. A. Bluemke, and J. A. C. Lima, "Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis," *Circulation Research*, vol. 121, no. 9, pp. 1092-1101, October 2017.
- [21] K. W. Chen, F. C. Meng, Y. L. Shih, F. Y. Su, Y. P. Lin, F. Lin, J. W. Lin, W. K. Chang, C. J. Lee, Y. H. Li, C. B. Hsieh, and G. M. Lin, "Sex-specific association between metabolic abnormalities and elevated alanine aminotransferase levels in a military cohort: The CHIEF study," *International Journal of Environmental Research and Public Health*, vol. 15, no. 3, pp. 545, March 2018.
- [22] Y. J. Chen, K. W. Chen, Y. L. Shih, F. Y. Su, Y. P. Lin, F. C. Meng, F. Lin, Y. S. Yu, C. L. Han, C. H. Wang, J. W. Lin, T. Y. Hsieh, Y. H. Li, and G. M. Lin, "Chronic hepatitis B, nonalcoholic steatohepatitis and physical fitness of military males: CHIEF study," *World Journal of Gastroenterology*, vol. 23, no. 25, pp. 4587-4594, July 2017.
- [23] G. M. Lin, Y. H. Li, C. J. Lee, J. C. Shiang, K. H. Lin, and K. W. Chen, Y. J. Chen, C. F. Wu, B. S. Lin, Y. S. Yu, F. Lin, F. Y. Su, and C. H. Wang, "Rationale and design of the cardiorespiratory fitness and hospitalization events in armed forces study in eastern Taiwan," *World Journal of Cardiology*, vol. 8, no. 8, pp. 464-471, August 2016.
- [24] K. Z. Tsai, J. W. Lin, F. Lin, F. Y. Su, Y. H. Li, Y. P. Lin, Y. K. Lin, C. L. Han, C. B. Hsieh, and G. M. Lin, "Association of betel nut chewing with exercise performance in a military male cohort: the CHIEF study," *Journal of the Royal Army Medical Corps*, vol. 164, pp. 399-404, November 2018.
- [25] J. W. Lin, K. Z. Tsai, K. W. Chen, F. Y. Su, Y. H. Li, Y. P. Lin, C. L. Han, F. Lin, Y. K. Lin, C. B. Hsieh, and G. M. Lin, "Sex-specific association between serum uric acid and elevated alanine aminotransferase in a military cohort: the CHIEF study," *Endocrine, Metabolic and Immune Disorders - Drug Targets*, vol. 19, no. 3, pp. 333-340, 2019.
- [26] W. H. Chao, F. Y. Su, F. Lin, Y. S. Yu, and G. M. Lin, "Association of electrocardiographic left and right ventricular hypertrophy with physical fitness of military males: the CHIEF study," *European Journal of Sport Science*, vol. 19, no. 9, pp. 1214-1220, October 2019.
- [27] Y. K. Jain and S. K. Bhattacharya, "Min max normalization-based data perturbation method for privacy protection," *International Journal of Computer and Communication Technology*, vol. 3, no. 4, pp. 45-50, 2014.

- [28] Y. Suresh, L. Kumar and S. K. Rath, "Statistical and machine learning methods for software fault prediction using CK metric suite: A comparative analysis", *ISRN Software Engineering*, pp. 1-15, March 2014.
- [29] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "SMOTE: synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol.16, pp. 321-357, June 2002.
- [30] D. W. Hosmer, and S. Lemeshow, "Applied logistic regression," *John Wiley & Sons*, 2004.
- [31] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871-1874, 2008.
- [32] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp.81-106, 1986.
- [33] L. Breiman, "Classification and regression trees," *Routledge*, 2017.
- [34] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, October 2001.
- [35] J. H. Friedman and J. J. Meulman, "Multiple additive regression trees with application in epidemiology," in *Proceedings of the 8th Biennial CDC and ADSTR Symposium on Statistical Methods Issues Associated with Complicated Designs and Data Structures*, vol. 22, no. 9, pp.1365-1381, April 2003.
- [36] B. Scholkopf, "Learning with kernels: support vector machines, regularization, optimization, and beyond," *MIT Press*, 2001.
- [37] J. Patterson and A. Gibson, "Deep learning: a practitioner's approach," *O'Reilly Media*, 2017.
- [38] F. Habibzadeh, P. Habibzadeh, and M. Yadollahie, "On determining the most appropriate test cut-off value: The case of tests with continuous results," *Biochemia Medica*, vol. 26, no. 3, pp. 297-307, October 2016.
- [39] J. Davis and M. Goadrich, "The relationship between precision recall and roc curves," in *Proceedings of International Conference of Machine Learning*, pp. 233-240, June 2006.
- [40] K. Hajian-Tilaki, "Receiver operating characteristic curve analysis for medical diagnostic test evaluation," *Caspian Journal of Internal Medicine*, vol. 4, no. 2, pp. 627-635, Spring 2013.
- [41] C. M. McHugh, A. Corderoy, C. J. Ryan, I. B. Hickie, and M. M. Large, "Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value," *BJPsych Open*, vol. 5, no. 2, pp. e18, March 2019.

C.Source Code

```
import numpy as np
import pandas as pd
import matplotlib

import theano
import tensorflow
import keras

#dataset = pd.read_csv('C:\\Users \\ Churn.csv', encoding = "ISO-8859-1",
header=None)
dataset = pd.read_csv('D:\\ Churn\\ Churn.csv')
X = dataset.iloc[:, 0:20].values
y = dataset.iloc[:, 20].values

from sklearn.preprocessing import LabelEncoder, OneHotEncoder
labelencoder_X_0 = LabelEncoder()
X[:, 0] = labelencoder_X_0.fit_transform(X[:, 0])

labelencoder_X_2 = LabelEncoder()
X[:, 1] = labelencoder_X_2.fit_transform(X[:, 1])
```

```
labelencoder_X_2 = LabelEncoder()  
X[:, 2] = labelencoder_X_2.fit_transform(X[:, 2])
```

```
labelencoder_X_3 = LabelEncoder()  
X[:, 3] = labelencoder_X_3.fit_transform(X[:, 3])
```

```
labelencoder_X_4 = LabelEncoder()  
X[:, 4] = labelencoder_X_4.fit_transform(X[:, 4])
```

```
labelencoder_X_5 = LabelEncoder()  
X[:, 5] = labelencoder_X_5.fit_transform(X[:, 5])
```

```
labelencoder_X_6 = LabelEncoder()  
X[:, 6] = labelencoder_X_6.fit_transform(X[:, 6])
```

```
labelencoder_X_7 = LabelEncoder()  
X[:, 7] = labelencoder_X_7.fit_transform(X[:, 7])
```

```
labelencoder_X_8 = LabelEncoder()  
X[:, 8] = labelencoder_X_8.fit_transform(X[:, 8])
```

```
labelencoder_X_9 = LabelEncoder()  
X[:, 9] = labelencoder_X_9.fit_transform(X[:, 9])
```

```
labelencoder_X_10 = LabelEncoder()  
X[:, 10] = labelencoder_X_10.fit_transform(X[:, 10])
```

```
labelencoder_X_11 = LabelEncoder()  
X[:, 11] = labelencoder_X_11.fit_transform(X[:, 11])
```

```
labelencoder_X_12 = LabelEncoder()  
X[:, 12] = labelencoder_X_12.fit_transform(X[:, 12])
```

```
labelencoder_X_13 = LabelEncoder()  
X[:, 13] = labelencoder_X_13.fit_transform(X[:, 13])
```

```
labelencoder_X_14 = LabelEncoder()  
X[:, 14] = labelencoder_X_4.fit_transform(X[:, 14])
```

```
labelencoder_X_15 = LabelEncoder()
X[:, 15] = labelencoder_X_15.fit_transform(X[:, 15])
```

```
onehotencoder = OneHotEncoder(categorical_features = [1])
X = onehotencoder.fit_transform(X).toarray()
X = X[:, 1:]
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2)
```

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

```
import keras
from keras.models import Sequential
from keras.layers import Dense
```

```
classifier = Sequential()
```

```

classifier.add(Dense(output_dim = 6, init = 'uniform', activation = 'relu', input_dim
= 20))
classifier.add(Dense(output_dim = 6, init = 'uniform', activation = 'relu'))
classifier.add(Dense(output_dim = 1, init = 'uniform', activation = 'sigmoid'))
classifier.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics =
['accuracy'])
classifier.fit(X_train, y_train, batch_size = 10, nb_epoch = 100)
y_pred = classifier.predict(X_test)
y_pred = (y_pred > 0.5)

```

```

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)

```

```

from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
lda = LDA(n_components=1)
X_train = lda.fit_transform(X_train, y_train)
X_test = lda.transform(X_test)
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(max_depth=2, random_state=0)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)

```

```

from sklearn.metrics import confusion_matrix

```

```
from sklearn.metrics import accuracy_score
```

```
cm = confusion_matrix(y_test, y_pred)
```

```
print(cm)
```

```
print('Accuracy' + str(accuracy_score(y_test, y_pred)))
```

```
""" Knn """
```

```
from sklearn import metrics
```

```
from sklearn.neighbors import KNeighborsClassifier
```

```
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.model_selection import train_test_split
```

```
""" Chi Feature Extraction Start"""
```

```
from sklearn.feature_selection import SelectKBest
```

```
from sklearn.feature_selection import chi2
```

```
X = X.astype(int)
```

```
X_train= X_train.astype(int)
```

```
X_test= X_test.astype(int)
```

```
chi2_features = SelectKBest(chi2)
X_kbest_features_train = chi2_features.fit_transform(X_train, y_train)
print('Original feature number:', X.shape[1])
print('Reduced feature number:', X_kbest_features_train.shape[1])
```

```
X_kbest_features_test = chi2_features.fit_transform(X_test, y_test)
print('Original feature number:', X.shape[1])
print('Reduced feature number:', X_kbest_features_test.shape[1])
```

""SVM ""

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
from sklearn.svm import SVC
svclassifier = SVC(kernel='linear')
svclassifier.fit(X_train, y_train)
```

```
from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

"""Random Forest """

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
lda = LDA(n_components=1)
X_train = lda.fit_transform(X_train, y_train)
X_test = lda.transform(X_test)
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(max_depth=2, random_state=0)

from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score

cm = confusion_matrix(y_test, y_pred)
print(cm)
print('Accuracy' + str(accuracy_score(y_test, y_pred)))
```

D.Implemented Screenshots

Name	Type	Size	Value
X	float64	(19289, 18)	[[1.0000e+00 0.0000e+00 0.0000e+00 ... 0.0000e+00 0.0000e+00 0.0000e+00 ...
X_test	float64	(3858, 18)	[[0.89053431 -0.25121133 -0.60962648 ... -0.53825969 -0... -0 ...
X_train	float64	(15431, 18)	[[-1.12292136 3.98071223 -0.60962648 ... -0.1270457 -0... -0 ...
dataset	DataFrame	(19289, 15)	Column names: Site ID, Condition, Plot number, Year, Month , Day , Tim ...
tmp	DataFrame	(19289, 15)	Column names: Site ID, Condition, Plot number, Year, Month , Day , Tim ...
y	float64	(19289,)	[0. 0. 0. ... nan nan nan]
y_test	float64	(3858,)	[1. 1. 1. ... 1. 1. nan]
y_train	float64	(15431,)	[0. 0. 1. ... 1. 1. 0.]

Variable explorer
File explorer
Help

B1. SPLITTING OF DATA

```

+00
Epoch 36/100
15431/15431 [=====] - 2s 120us/step - loss: nan - acc: 0.0000e
+00
Epoch 37/100
15431/15431 [=====] - 2s 120us/step - loss: nan - acc: 0.0000e
+00
Epoch 38/100
15431/15431 [=====] - 2s 120us/step - loss: nan - acc: 0.0000e
+00
Epoch 39/100
15431/15431 [=====] - 2s 119us/step - loss: nan - acc: 0.0000e
+00
Epoch 40/100
15431/15431 [=====] - 2s 110us/step - loss: nan - acc: 0.0000e

```

B2. EPOCH COUNT FOR PREDICTION

```
15431/15431 [=====] - 2s 119us/step - loss: nan - acc: 0.0000e
+00
Epoch 96/100
15431/15431 [=====] - 2s 120us/step - loss: nan - acc: 0.0000e
+00
Epoch 97/100
15431/15431 [=====] - 2s 119us/step - loss: nan - acc: 0.0000e
+00
Epoch 98/100
15431/15431 [=====] - 2s 119us/step - loss: nan - acc: 0.0000e
+00
Epoch 99/100
15431/15431 [=====] - 2s 120us/step - loss: nan - acc: 0.0000e
+00
- . . . . .
```

B3.EPOCH COUNT FOR PREDICTION

```
Console 1/A X
...:
[[ 3 757]
 [ 1 3097]]
      precision    recall  f1-score   support

 0       0.75       0.00       0.01       760
 1       0.80       1.00       0.89      3098

avg / total       0.79       0.80       0.72      3858

In [69]: print('Accuracy' + str(accuracy_score(y_test, y_pred)))
Accuracy0.8035251425609123
```

B4. ACCURACY FOR SVM ALGORITHM