VOL. 10, NO. 1, JANUARY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

UTILIZATION OF DATA MINING TECHNIQUES FOR DIAGNOSIS OF DIABETES MELLITUS - A CASE STUDY

Thirumal P. C. and Nagarajan N.
Department of IT, Coimbatore Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
E-Mail: thirumalciet@gmail.com

ABSTRACT

Data mining looks through a large amount of data to extract useful information. The most important and popular data mining techniques are classification, association, clustering, prediction and sequential patterns. In health concern businesses, data mining plays an important role in early prediction of diseases. In general to detect a disease numerous tests must be conducted in a patient. The usage of data mining techniques in disease prediction is to reduce the test and increase the accuracy of rate of detection. One of the most common diseases among young adult is Diabetes mellitus. This develops at a middle age and more common in obese children and adolescents. In order to reduce the population with diabetes mellitus it should be detected at an earlier stage, hence a quick and efficient detection mechanism has to be discovered. The principle of this study is to apply various data mining techniques which are noteworthy to prediction of diabetes mellitus and extract hidden patterns from the PIMA Indian diabetes dataset available at UCI Machine Learning Repository.

Keywords: data mining, diabetes mellitus, association, classification, decision trees.

1. INTRODUCTION

Knowledge discovery in databases (KDD) converts the low level data in to high level knowledge. It is an iterative process that consists of sequence steps of processes. The phases of KDD includes selection, pre-processing, Transformation, Data Mining, Interpretation/Evaluation. The target data's are selected and integrated from many heterogeneous sources. The aim of selection is to focus on correct subset of variable and data samples. Pre-processing phase selects data with respect to data mining in hand. Real world data's may be incomplete, noisy, incomplete and inconsistent. Tasks involved in pre-processing are Data Cleaning, Data integration, Data transformation. Data cleaning fills the missing values, smoothens the noisy data, identify outliers and resolve if any inconsistencies. Data integration integrates multiples data's or files. Normalization and Aggregation are done in data transformation [1].

Data Mining is the process of extraction of useful knowledge from huge amounts of data to predict using techniques such as classification, clustering and association. A data mining system may generate numerous patterns. The discovered patterns can be similar to prior knowledge or expectations. Data Mining has two primary goals: Prediction and Description. Prediction involves variables in the dataset to predict unknown values or future values whereas Description focuses on finding patterns that describes the data interpreted by humans [2]. Data mining techniques can be implemented in hardware and software to add values for existing information and can be integrated with new products and systems. Disease prediction plays an important application in data mining.

Health care industry generates large amounts of complex data's such as patient history, hospital resources, electronic records, information about medical devices etc. These data's serves as a key resource to process and analyse for knowledge extraction that enables the decision making and to save cost. Research using data mining techniques have been applied in diagnosis of various diseases such as cardiovascular diseases, AIDS, diabetes and asthma. In this paper we have focused on Type-2 diabetes [3]. Several life style factors may affect the incidence of diabetes mellitus. Obesity and weight gain increases the risk and physical inactivity in turn again elevates the risk. Cigarette smoking is associated with a small level of increase and moderate alcohol consumption decreases the risk of diabetes. Various data mining techniques such as Naïve Bayes classifier, Decision Tree, Support Vector Machines, and k Nearest Neighbor have been investigated and developed to measure of the efficiency of those in type-2 diabetes prediction.

2. RELATED WORK

Insulin is most important hormone in the body. It converts the sugar, starch and other food items in to energy needed for daily life. If the body does not produce insulin the redundant amount of sugar will be driven out by urination. This disease is called as Diabetes. Causes of diabetes are always a mystery even though obesity and lack of exercise plays a vital role. In November 2007, 20.8 million children's and adults (approximately 7% of the population) in USA were affected by diabetes. In early the ability to diagnose diabetes plays a major role in treatment process [4]. Diabetes is common in both developed and developing countries. There was estimated 175 million people in 2000 and was expected to increase to 354 million by 2030. It is also given that by 2030, 85 percent of world's diabetic patients will be in developing countries. In India the diabetes count is expected to increase from 31.7 million in 2000 to 79.4 million in 2030 [5].

Many researchers have been made in the past decade to diagnose the diabetes mellitus. Applying data mining techniques in medical field is a typical task. The data mining begins with a hypothesis and the results are adjusted to fit the hypothesis in medical research. This

differs from the standard data mining task in which datasets without apparent hypothesis is commonly used [2]. Artificial neural networks, decision trees and logistic regression have been used for prediction of diabetes among two sets of people from china. The classification accuracy achieved by logistic regression was 73.23%, decision tree with 77.8% and ANN with 77.87%. Decision tree proved as best followed by regression and ANN [6]. Drug treatment for diabetic patients have been investigated with the logistic regression model and it is shown that the drug treatments in young age group causes side effects and patients of old age group need drug treatment immediately since no other alternatives are available [7]. Fuzzy ontology to model the diabetes knowledge and establish relations based on the structure of fuzzy diabetes ontology. It is also shown that it works effectively in decision support applications [8].

Diabetes mellitus decreases the resting of blood flows through the skin by disturbing regulation of skin microcirculation [9]. Obesity and lack of exercise plays a significant role in cause for diabetes [10]. The availability of health records and monitoring those leads to accumulation of data which are used by practitioners, health care decision makers, physicians. Since diabetes is a lifelong disease an individual patient record may be massive and difficult to handle [11]. Poor generalization ability is a major issue of data mining in healthcare industry because of the lack of input data and re processing [12]. Information gain method and feature selection can be used in collaboration with adaptive neuro fuzzy interference system in diagnosis of new patients. This proposed approach has gained 98.24% accuracy when compared to other data mining algorithms [13]. Data mining techniques can be used to predict the fast glucose level (FGL) thereby predicting the fluctuations in glucose level [14].

3. MATERIALS AND METHODS

3.1. Pima Indians diabetes data set

This dataset is obtained from the UCI Repository of Machine Learning Databases. All the patients here are females and atleast 21 years old who are living in phoenix, Arizona, USA. This larger database was held by the National Institutes of Diabetes and Digestive and Kidney Diseases. This consists of two classes which are represented by binary variable '0' or '1'. Here '1' represents the positive test diabetes and '0' represent the negative test for diabetes. The database has 768 patients with 9 numeric variables. There are 268 (34.9%) positive cases which belong to class '1' and 500 (65.1%) cases in class'0'. There were no missing values. Five patients had glucose of 0, 11 patients have BMI of 0, 28 patients have blood pressure of 0, 192 persons have skin fold thickness of 0, 140 have serum insulin level of 0. These zero values are a part of missing values. The attributes in this dataset is given in Table-1. The database is designed such that 8 attributes contribute to the result of 9th attribute.

Table-1. Attribute in Pima Indian diabetes dataset.

At No.	Abbreviation	Description	Type	Unit
A1	PREG NANT	Number of times pregnant	Numeric	-
A2	GTT	2-hour OGTT plasma glucose,	Numeric	mg/dl
A3	BP	Diastolic blood pressure	Numeric	mmHg
A4	SKIN	Triceps skin fold thickness	Numeric	mm
A5	INSULIN	2-hour serum insulin,	Numeric	mm U/ml
A6	BMI	Body mass index(kg/m)	Numeric	Kg/m^2
A7	DPF	Diabetes pedigree function	Numeric	-
A8	AGE	Age of patient(years)	Numeric	-
Class	DIAB ETES	Diabetes onset within 5 years (0, 1).	Numeric	-

3.2. Pre-processing and sampling

The statistical analysis Pima Indian Diabetes dataset is shown in Table-2 and Table-3. Range of values differs widely as seen in Table-2. Hence a normalization method has to be implemented. Here we have used 'weka.filters. Discretize' method to normalize the data. Result of normalization is shown in Table-2.

Table-2. Before normalization.

Attributes no.	Mean	Standard deviation
Atr_1	3.84	3.37
Atr_2	120.89	31.97
Atr_3	69.1	19.35
Atr_4	20.53	16.0
Atr_5	79.79	115.24
Atr_6	31.99	7.88
Atr_7	0.47	0.33
Atr_8	33.24	11.76

Table-3. After normalization.

Attributes no.	Mean	Standard deviation
Atr_1	0.226	0.19
Atr_2	0.608	0.16
Atr_3	0.566	0.15
Atr_4	0.207	0.16
Atr_5	0.094	0.13
Atr_6	0.477	0.11
Atr_7	0.168	0.14
Atr_8	0.204	0.19

3.3. Data analysis

The distribution of attribute values with respect to class attribute '0 or 1' is shown in Figure-1.

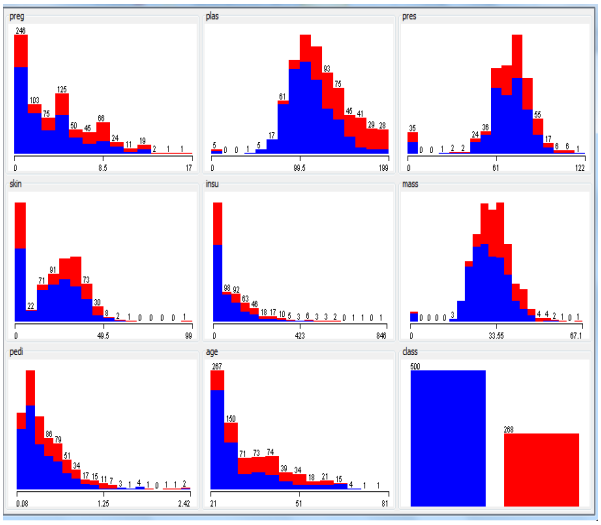

Figure-1. Attribute value distribution with respect to class variables.

The blue color denotes the prevalence of diabetes. It is clear from the above figure that most of the diabetic patients who are pregnant are in the values of 0 to 1.5, have plasma in the range 99.5 to 103.5, having pressure in the range 65 to 71, have skin fold thickness in the range 0 to 7, insulin levels in the range 0 to 50, BMI in the range 27 to 30, pedigree function 0.25 to 0.50 and they belong to the age of 21 to 25.

3.4. Mining the dataset

Weka is popular machine learning software developed in Java at University of Waikato, New Zealand. It is a open source software available at GNU (General Public License). It consists of visualization tools and algorithms which are used in data analysis and predictive modelling with graphical user interface for easy functionality access [15]. Weka supports several data mining tasks such as data pre-processing, clustering, regression, visualization and feature selection. The attributes available in weka are of one of these types Nominal: One of predefined list of values, Numeric: A real or integer number, String, Date, Relational. Key Features of this tool are open source and platform independent. This consists of various algorithms for data mining and machine learning [16].

4. METHODOLOGY

Classification is the process of identifying a new observation category set on the basis of training set of data that contains observations whose category is known. Cluster analysis technique is used to group the objects according to its similarity. Studies have been made to compare the different techniques of classification which have been developed so far. In this study we have compared different classifiers Naïve Bayes, Decision tree (c4.5), k-Means, SVM and k Nearest Neighbor classifier.

4.1. Naive baayes classifier

Naive Bayes classifier is a well known type of classifiers. A set of programs that assign a class of predefined set to an object under construction based on the descriptive attributes. This is done by using a probabilistic approach which computes class probabilities and predicts most probable classes. The principle of basic Naïve bayes is described as follows: A training set of patient data's are taken. Marginal probabilities of symptoms are given by $P(S_i)$ and diseases are given by $P(d_j)$ and conditional probabilities of symptoms on all the diseases $P(s_i|d_j)$ are computed by counting their frequencies in the data. For a given set of symptoms ($s=\{S_i\}$) for a patient, the posterior probability for diagnosis done for the patients are calculated by,

$$P(d_j|S) \propto P(d_j)\pi_{s_i \in S} \frac{P(s_i|d_j)}{P(s_i)} \quad (1)$$

Diagnostic score for each disease diagnosis is given as,

$$P(d_j|S) \propto P(d_j)\pi_{s_i \in S} P(s_i|d_j) \quad (2)$$

Conditional probability for a symptom S_i for a disease d_j is given by,

$$P(s_i|d_j)=\frac{p(s_i \cap d_j)}{p(d_j)} \quad (3)$$

Here $p(d_j)$ is the number of patients in the dataset with disease $p(d_j)$ and $p(s_i \cap d_j)$ is the count of patients with both s_i and d_i. Naïve bayes classifier is formed when Bayesian network is applied to classification problem. Bayes network is used for classification of diabetes dataset obtained an accuracy of 72.3% [17]. Naïve bayes have acquired a classification accuracy of 76.3% in the same dataset [18].

4.2. C4.5 Algorithm

C4.5 algorithm (known as "J48"in Weka) is used to generate decision tree during classification. System that builds classifier is one of the commonly used tools in data mining. This separates the data's taken in to inspection in to branches for building a tree to improve classification accuracy. The inputs belong to one of small number of classes with fixed set of attributes and output a classifier that predicts the class to which case belongs accurately [19]. The decision trees used for classification are referred to as statistical classifier. C4.5 accounts for unavailable values, continuous attribute value ranges, rule derivation, pruning of decision trees, etc. Features of C4.5 algorithm

- Handles training data with missing values of attributes
- Handles different cost attributes
- Pruning the decision tree after its creation
- Handles attributes with discrete and continuous values

Let $S=s_1$, s_2 …. set be the training data which consists of already classified samples. Each sample $Si=x_1$, x_2,, be the vector where x_1, x_2.. represents the features or attribute of the sample. Training data is a vector $C=c_1$, c_2… where c1, c2 represent the class to which each of the sample belongs [20].

4.3. SVM

Support vector machined are supervised learning algorithm. SVM naturally points to hyper plane that separates the classes of data. SVM not only separate entities in to correct classes but also identifies instances which establish classification not supported by data. SVM are insensitive in the distribution of training samples of each class. It can also be extended to perform numerical calculations. SVM can be extended in two ways. First is extending svm to execute regression analysis and the goal of this type of extension is to produce a linear function that can fairly accurate the target function. Another type of extension is learning to rank elements other than producing a classification for individual elements. Ranking can also be reduced to combining pairs of instance and produce a+1 estimate if the pair is in correct ranking order in addition to -1 otherwise.

4.4. k-NN Algorithm

k-NN approach has been used in different data analysis applications such as pattern recognition, data mining, databases and machine learning due to its simplicity and high accuracy. It has been recognized as one of the top 10 algorithms in data mining [22]. kNN classification classifies instances based on similarity. It is a type of lazy learning algorithm where the function is approximated locally and computation is deferred until classification. kNN is mainly used for classification and clustering. Many researches found that the kNN algorithm accomplishes good performance in their experiments on various datasets. Pima Indian diabetes dataset is complex due to its missing values. k-NN method replaces the missing values with the corresponding values from the neighboring column in Euclidean Distance. If the corresponding value from the nearest neighbor is also missing it takes from other immediate next column value. This model is simple and highly competitive when compared to other techniques. Shortcoming of kNN is the lack of probabilistic semantics which allow posterior predictive probabilities to be employed. For example, assigning variable losses in a consistent manner.

kNN has been modified by many authors to improve its efficiency. A classwise k nearest algorithm have been designed and tested against Pima Indian diabetes dataset. Here testing data is classified in to class label corresponding to the lowest distance. Accuracy achieved for C-kNN is 78.16% [23]. K means and kNN algorithm have been combined as amalgam kNN model to classify PIDD. Here the quality of data is improved by removing the noise and thereby increasing the efficiency. K-means removes the incorrectly classified instances and classification is done using k nearest neighbor. The choice of k in kNN depends upon the data. Larger values of k reduce noise on classification. A good k value is selected by cross validation technique. By increasing the k value with tenfold cross validation the classification accuracy of 97.4% have been obtained.

5. EVALUATION AND TESTING

5.1. Cross validation

Cross validation is a data mining technique used to evaluate the performance of classification algorithms. It is used to evaluate error rate for learning techniques. The dataset is portioned in to n folds; each fold is used for testing and training. The procedure repeats for n times in testing and training. In a 10 fold cross validation the data is divided in to 10 parts where each parts are approximately same to form the full dataset. Each term is held out and during the learning scheme which trained on remaining nine-tenths, the error rate is calculated in the holdout set. Learning procedure executes 10 times on training sets and finally the error rates for 10 sets are averages to yield an overall error rate.

5.2. Confusions matrix

Confusion matrix is used to present the accuracy of classifiers obtained through classification. It is used to show the relationship between outcomes and predicted classes (Refer Table-4).

Table-4. Confusion matrix.

Confusion matrix		**Targeted values**	
		Positive	**Negative**
Model	Positive	a	b
	Negative	c	d

Here 'a' represents number of correct instance that is of negative instance; 'b' represents the number of

incorrect prediction that is of positive instance. c is the number of incorrect predictions that is of negative instance. d is the correct predictions that is of positive instance.

5.3. Performance evaluation

Two metrics are used to evaluate the performance of data mining algorithms namely recall and false positive rate (FPR). In case of pima Indian diabetes dataset recall will reflects the number of diabetic patients which are correctly classified and the formula to calculate recall is given below:

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (4)$$

False Positive rate will reflect number of non-diabetic patient records and is calculated by:

$$\text{FPR} = \frac{FP}{FP+TN} \quad (5)$$

Accuracy is the percentage of predictions that are seems to be correct. Precision is the measure of accuracy provided that a class has predicted.

$$\text{Accuracy} = \frac{TP+FN}{TP+FP+TN+FN} \quad (6)$$

$$\text{Positive precision} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Negative Precision} = \frac{FN}{FN+FN} \quad (8)$$

$$\text{Error Rate} = \frac{FP+FN}{TP+FP+TN+FN} \quad (9)$$

Best learning algorithms are selected based on the performance of classifiers in reference with recall and low false positive rate (FPR). False negative implies the non-diabetic patient records and false positive gives the non-diabetic patient record which is classified as diabetic patient record.

6. RESULTS AND DISCUSSIONS

The accuracy of learning system needs to be evaluated before it is being used. Limited data availability makes estimating accuracy a difficult task. Choosing a good evaluation technique is very important in machine learning system environment. There are several methods for evaluation that divides data in to training set and testing set. In this paper we have used classification and clustering techniques on Pima Indian diabetes dataset and measured performance of those algorithms. The resultant value for the above dataset using data mining classification algorithms is shown in the Table-5.

Table-5. Accuracy comparison of algorithms.

Algori-thm	Accuracy (%)	TP	FP	Precision	Recall
Naïve bayes	77.8646	0.83	0.317	0.83	0.83
C4.5	78.2552	0.864	0.369	0.814	0.864
SVM	77.474	0.775	0.309	0.77	0.775
kNN	77.7344	0.892	0.437	0.792	0.892

C4.5 algorithm outperforms the other algorithms with the accuracy of 78.25%. Naïve Bayes ranks second with accuracy of 77.8 % and k-NN scores 77.73%, finally SVM acquired 77.4% as accuracy. Figure-2 shows the comparison of accuracy of algorithm.

Figure-2. Accuracy comparison of algorithms.

It is clear from the graph knn was the least among classification algorithms. kNN is lazy learning algorithm since it stores training examples and delays the processing until a new instance is classified.

7. CONCLUSIONS

The prevalence of diabetes is increasing among young adults and old age people. The present study concludes that the elderly diabetic patients can be given assessment and treatment plans that suits their needs and lifestyle. Simple awareness measures such as low sugar diet, proper diet can avoid obesity. The main goal of this study is to get best algorithms that describes given data in multiple aspects. These algorithms are necessary for automatic classification tools. The automatic design tools will help the experts to reduce wait in line. In this paper, several data mining algorithms such as Naïve Bayes, Decision trees, k Nearest neighbor and SVM have been discussed and tested with pima Indian diabetes dataset. From the experiments it is concluded that kNN provides lower accuracy when compared to other algorithms. This can be further enhanced and expanded by incorporating with other classification algorithms and can be designed in such a way that it removes the property of laziness.

REFERENCES

[1] Prakash Mahindrakar *et al*. 2013. Data Mining in Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges. Int. Journal of

Engineering Research and Applications. 3(6): 937-941. (ISSN: 2248-9622)

[2] S.Vijiyarani and S.Sudha. 2013. Disease Prediction in Data Mining Technique - A Survey. www.ijcait.com International Journal of Computer Applications and Information Technology. 2(1). (ISSN: 2278-7720).

[3] Mythili T., Dev Mukherji, Nikita Padalia and Abhiram Naidu. 2013. Article: A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). International Journal of Computer Applications. 68(16): 11-15.

[4] Huy Nguyen Anh Pham and Evangelos Triantaphyllou. 2008. Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization. R. Lee and H.-K. Kim (Eds.). Computer and Information Science, SCI 131. pp. 11-26.

[5] S. Wild, G. Roglic, A. Green, R. Sicree and H. King. 2004. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care. 27(5): 1047-1053.

[6] X. H. Meng, Y. X. Huang, D. P. Rao, Q. Zhang and Q. Liu. 2013. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung Journal of Medical Sciences. 29(2): 93-99.

[7] Abdullah A. Aljumah, Mohammed Gulam Ahamad and Mohammad Khubeb Siddiqui. 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences. 25: 127-136.

[8] Chang-Shing Lee. 2011. A Fuzzy Expert System for Diabetes Decision Support Application. IEEE transactions on systems, man, and cybernetics - part b: cybernetics. 41(1).

[9] Anburajan M. Changes of Skin Temperature of Parts of the Body and Serum Asymmetric Dimethylarginine (ADMA) in Type-2 Diabetes Mellitus Indian Patients. 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, August 30 - September 3.

[10] J.S. Petrofsky, M. Prowse and E. Lohman. 2008. The influence of ageing and diabetes on skin and subcutaneous fat thickness in different regions of the body. The Journal of Applied Research. 8(1): 55-61.

[11] Riccardo Bellazzi and Ameen Abu-Hanna. 2009. Data mining technologies for blood glucose and diabetes management. Journal of diabetes science and technology. 3(3): 603-612. PMID: 20144300.

[12] AbuKhousa. E. 2012. Predictive data mining to support clinical decisions: An overview of heart disease prediction system. IEEE Transaction on Innovations in Information Technology (IIT). pp. 267-272.

[13] Deepali Chandna. 2014. Diagnosis of Heart Disease Using Data Mining Algorithm. International Journal of Computer Science and Information Technologies. 5(2): 1678-1680.

[14] Yamaguchi Masaki, *et al*. 2006. Prediction of blood glucose level of type 1 diabetics using response surface methodology and data mining. Medical and Biological Engineering and Computing. 44(6): 451-457.

[15] Jagtap Sudhir B. 2013. Census Data Mining and Data Analysis using WEKA. arXiv preprint arXiv: 1310. 4647.

[16] An Introduction to the WEKA Data mining system - http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf.

[17] Guo Yang, Guohua Bai and Yan Hu. 2012. Using Bayes Network for Prediction of Type-2 Diabetes. Internet Technology and Secured Transactions, 2012 International Conference for. IEEE.

[18] Koklu Murat and Yavuz Unal. Analysis of a Population of Diabetic Patients Databases with Classifiers. human resources. 1: 2.

[19] Lakshmi K. R. and S. Prem Kumar. 2013. Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability. International Journal of Scientific and Engineering Research. 4(6).

[20] Adhatrao Kalpesh, *et al*. 2013. Predicting Students' Performance Using ID3 and C4. 5 Classification Algorithms. arXiv preprint arXiv:1310.2071.

[21] Karegowda Asha Gowda, M. A. Jayaram and A. S. Manjunath. 2012. Cascading K-means clustering and K-nearest neighbor classifier for categorization of diabetic patients. International Journal of Engineering and Advanced Technology. 1: 147-151.

[22] Wu Xindong, *et al*. 2008. Top 10 algorithms in data mining. Knowledge and Information Systems. 14.1: 1-37.

[23] Y. Angeline Christobel and P.Sivaprakasam. 2013. A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, 2(3).