E-*EXTENSION* EMPLOYABILITY OF SCHOLARS PURSUING POST GRADUATION IN AGRICULTURAL EXTENSION IN SAUS : USING DATA MINING TECHNIQUES

N. M. Vegad[1], R.S. Parmar[2] and N. B. Chauhan[3]

1 Assistant Professor, College of Agricultural Information Technology, AAU, Anand - 388110
2 Professor, College of Agricultural Information Technology, AAU, Anand - 388110
3 Professor & Head, Dept. of Agricultural Extension & Communication, BACA, AAU, Anand - 388110
Email : rsparmar@aau.in

ABSTRACT

e-Extension can be defined as the use of digital technologies to enhance one-to-one interactions. Since the invention of computer, the information available in every field has been digitized to be accessible by people using digital resources. Hence, data are growing rapidly in huge amount in every domain. One such domain of interest for researchers is agriculture field. To find interesting patterns from the database, we applied and analyzed concepts of data mining on the database of scholars pursuing post graduation in agricultural extension in SAUs of Gujarat in this paper. Data mining is the process of getting useful information by analyzing different kind of data. In this research, we have concentrated on classification models namely Rules based Decision Table, PART and JRip, Tree based J48, RandomForest, RandomTree, LMT and REPTree, Bayesian based NaiveBayes and Lazy based IBK and KStar with respect to their accuracy of correctly classified instances, incorrectly classified instances and very important Receiver Operating Characteristic (ROC) Area which helps in understanding the classification model and their results, which can also help other researchers in making decision for the selection in classification model based on their data and number of attributes. Experimental results show that KStar classifies better than other classification algorithms with 93 % predictability, followed by Random Forest with 92 % predictability while Decision Table classification is the lowest with 80 % predictability. The fitted models, using data mining approaches suggested that predictor variables namely job preference, interpersonal communication, information collection behaviour, self confidence and mother's education had higher influence on e-extension employability. Based on all the benchmarks used to measure the algorithms employed in this study, it was discovered that KStar performance is the most appropriate in terms of predictability (accuracy) based on this data. Focus was therefore laid on designing a predictive system on the most suitable algorithm which is Forest and KStar in this particular domain.

Keywords : *e-extension, employability of scholars, agricultural extension*

INTRODUCTION

E-Extension is using the power of online networks, computer communications and digital interactive multi-media to facilitate dissemination of agriculture technology. It includes effective use of management information system, decision support system, expert system, artificial intelligence based system and data mining (knowledge discovery in databases) based system. Data Mining Techniques are likely to continue as a key driver of agriculture development in India. Data Mining is the process of analyzing, extracting and predicting the meaningful information from enormous data to mine various pattern. Data mining is a rising technology with the development of artificial intelligence and database procedures which is used in different business organizations to improve the effectiveness and value of a business process (Manish 2009). Data Mining is a multidisciplinary field that merges artificial intelligence, computer science, machine learning, database management, mathematics algorithms and statistics (Liao SH 2003). Data Mining is a mixture of procedures such as neural network, decision trees or standard statistical procedures to identify pieces of information of data, and extracting these in such a way that they can be put to use in areas such as decision support, prediction and estimation (Wah and Bakar 2003). Data mining approaches uses classification algorithms and other statistics methods to find out, from massive data set, useful information for industry operation (Witten and Eibe 2011). Secondary data of most enterprises are millions, which are extremely difficult to analyze, it has becomes important to extract useful information from huge amount of data (Wei and Albert 2013). Ahamed *et al*. (2015) applied clustering techniques to predict the rice yield in the areas of Bangladesh. Mucherino *et al*. (2009) surveyed about different data mining techniques and how they can be useful in agriculture sector. Kumari and Chitra (2013) studied support vector machine algorithm as a classifier for predicting diabetes. Their experimental results showed that SVM can be successfully used for predicting diabetes diseases. Rajeswari and Arunesh (2016) made a comparative analysis of three algorithms namely

NaveBayes, JRip and J48. JRip classification algorithm gives better result for experimental dataset. Chiranjeevi and Ranjana (2018) conducted a comparative analysis of two algorithms like NaiveBayes and J48. Naive Bayes gives better result for experimental dataset. Raunak (2018) studied three algorithms namely Naive Bayes, zeroR and stacking. Naive bayes classification algorithm gives superior effect on experimental dataset. Kalekar et al. (2018) studied J48 decision tree algorithm showed an accuracy of 87.5% in classifying the soil according to its fertility, and can be used to recommend appropriate fertilizers. Bhuyar (2014) studied different classifier algorithms to predict fertility rate. Study indicates that J48 classifier perform better to predict fertility index. Elhamayed (2016) studied algorithms like J48, Decision Table, PART, NaiveBayes, and IKB. By analyzing the overall experiment results of the production dataset, it is concluded that J48 algorithm has produced the best classification performance than IKB and NaiveBayes and has produced slightly difference performance with Decision Table, PART classifiers. The present investigation has been taken up to study different data mining algorithms likes DecisionTable, PART, JRip, J48, RandomForest, RandomTree, LMT, REPTree, NaiveBayes, IBK and KStar and also compare predictability of fitted algorithms.

OBJECTIVE

To know **e**-*extension* employability of scholars pursuing post graduation in agricultural extension in SAUs : using data mining techniques

METHODOLOGY

The present study was carried out in all the four State Agricultural Universities of Gujarat. The ex-post facto research design was applied for the study. Data were collected from a random sample of 120 scholars pursuing post graduation in agricultural extension in four SAUs of Gujarat. The Dataset having 18 attributes namely academic performance, medium of education, native place, father's education, mother's education, annual family income, family occupation, involvement in extracurricular activity, library exposure, information collection behavior, attitude towards extension work, job preference, achievement motivation, self confidence, interpersonal communication, innovativeness, professional zeal , willingness to work in rural area and E-extension employability. E-extension employability is class label which categorized as VL - Very Low (0-20 Percent), L - Low (21-40 Percent), M- Medium (41-60 Percent), H - High (61-80 Percent) and VH - Very high (81-100 Percent). Open source Data mining tool WEKA version 3.8.1 was used for this research. This dataset prepared in Excel sheet with .CSV extension.

Feature selection is a preprocessing step to data mining that choose a subset of original features according to a certain evaluation criterion and is effective in removing effect of irrelevant data, removing redundant data, reducing dimensionality, increasing learning accuracy and improving result comprehensibility. The present study focuses on two feature selection techniques namely cfsSubsetEval and GainRatioAttributeEval, which is one of the important and frequently used in data preprocessing in data mining. Using these attribute selection algorithms we can select the best attributes out of 17 numbers of attributes of dataset that affect the e-extension employability.

The model building supported in this investigation is a classification in the search for the best model. The population for which a model is built is further divided into two parts namely training and testing. The ratio of the sample population is set at approximately 75 %, 25% with the objective to avoid occurrence of over-fitting and thus increase mode predictability in the dataset.

DecisionTable Classifier

Decision Table is an accurate method for numeric prediction from decision trees and it is an ordered set of If-Then rules that have the potential to be more compact and therefore more understandable than the decision trees. The reason to explore decision table is simplicity, less compute intensive algorithm than the decision-tree-based approach. The algorithm is found in the Weka classifiers under Rules.

PART Classifier

Part technique comes under the rules classification. It obtains the rules from partial tree built using J4.8 classifier technique. Therefore most of the time part and J4.8 gives same result for classification of given data. The algorithm, PART is found in the Weka classifiers under Rules.

JRip classifier

It is one of the basic and most popular algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error JRip (RIPPER) proceeds by treating all the examples of a particular judgment in the training data as a class, and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered.

J48 Classifier

J48 is an extension of C4.5 which is used to generate decision tree using C4.5 algorithm. Decision tree generated can be used for classification and hence also called statistical

classifier. The main thing that must be kept in mind while using algorithm is that the database must be properly organized and information must be correct for proper analysis.

RandomForest Classifier

Random Forest is an improvement over bagged decision a tree that disrupts the greedy splitting algorithm during tree creation so that split points can only be selected from a random subset of the input attributes. This simple change can have a big effect decreasing the similarity between the bagged trees and in turn the resulting predictions. Click the "Choose" button and select "RandomForest" under the "trees" group.

Random Tree Classifier

It is a supervised Classifier. it is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a **random** set of data for constructing a decision **tree**. In standard **tree** each node is split using the best split among all variables.

LMT Classifier

Classifier for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values.

REPTree Classifier

REPTree builds a decision or regression tree using information gain/variance reduction and prunes it using reduced-error pruning. Optimized for speed, it only sorts values for numeric attributes once. It deals with missing values by splitting instances into pieces, as C4.5 does. You can set the minimum number of instances per leaf, maximum tree depth, minimum proportion of training set variance for a split (numeric classes only), and number of folds for pruning.

IBK Classifier

Instance Based K-nearest neighbor uses K-NN for classification. It is also a lazy learner i.e. it delays construction of classifier until classification time.

KStar Classifier

K* is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function.

Naive Bayes Classifier

A Naive Bayes classifier is one of the classifiers in a family of simple probabilistic classification techniques in machine learning. It is based on the Bayes theorem with independence features. Each class labels are estimated through probability of given instance. It needs only small amount of training data to predict class label necessary for classification.

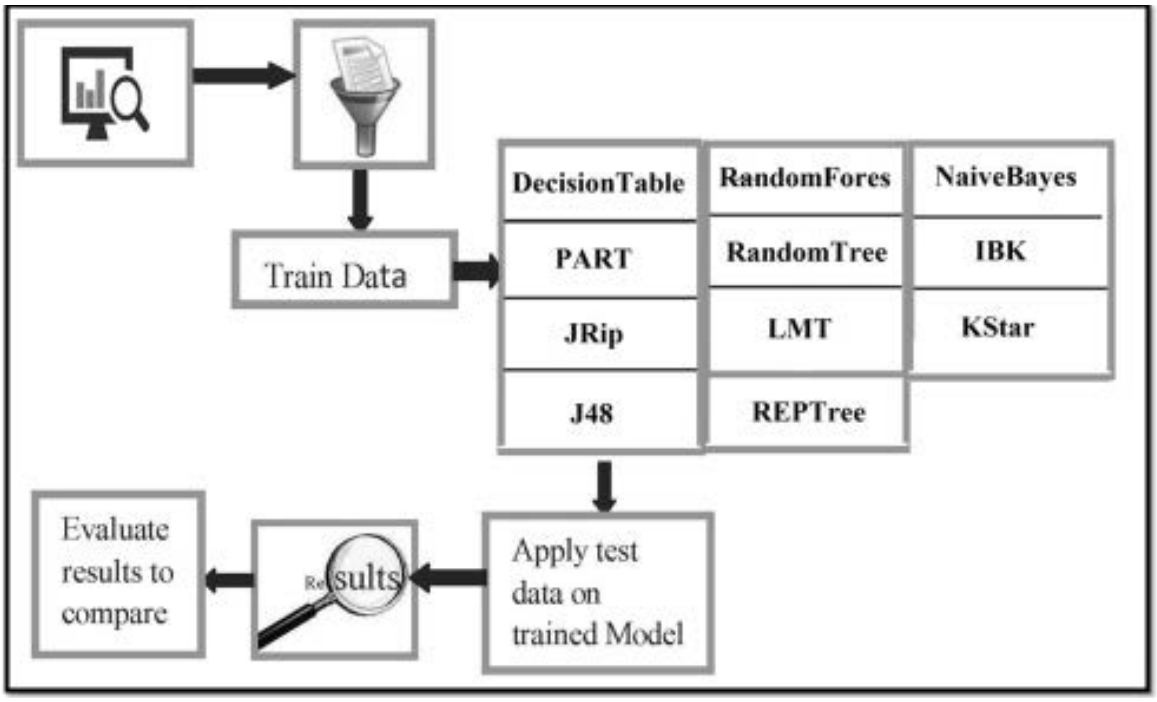

Fig. 1 Architecture of E-extension employability prediction system

The Fig 1 shows the entire architecture of e-extension employability prediction system. The raw data are used, which are then cleaned, attributes selected and sorted. The classification techniques like DecisionTable, PART, JRip, J48, RandomForest, RandomTree, LMT, REPTree, NaiveBayes, IBK and KStar are then implemented over the trained data. The results of each algorithm is noted from WEKA and compared with each other. Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Relative Absolute Error (RAE) values are taken into consideration for each case. Thereafter performance is measured using three factors namely Sensitivity, Specificity, and Accuracy.

RESULTS AND DISCUSSION

The software tool used for this study was WEKA an open source and free software used for knowledge analysis. From Weka toolkit, attribute selection algorithm cfsSubsetEval, which is used in data preprocessing. Using this algorithm we have selected 5 best attributes namely job preference, interpersonal communication, information collection behaviour, self confidence and mother's education out of 17 numbers of independent attributes of dataset that affect the e-extension employability. For this experiment, the performance of the models namely Rules based DecisionTable, PART and JRip, Tree based J48, RandomForest, RandomTree, LMT and REPTree, Bayesian based NaiveBayes and Lazy based IBK and KStar are examined.

The table 1 contains the results of efficiency analysis of each data classification model, showing kappa statistic, correctly classified and incorrectly classified instances. In addition, the table presents the values of Precision, Recall, True Positive Rate and False Positive Rate. The KStar model classifies instance correctly with an accurate rate of 93 %, this indicates that results obtained from training data are optimistic and can be relied on for predictions. This result informed the choice for the selection of the best classification model which is KStar in this case.

Table 1: Efficiency analysis of each data classification model

Performance Error	Different Model Algorithms										
	Rules Based			Tree Based					Lazy Based		Bayesian Based
	Decision Table	PART	JRip	J48	Random Forest	Random Tree	LMT	REP Tree	IBK	KStar	Naïve Bayes
Number of Selected Attributes	5	5	5	5	5	5	5	5	5	5	5
Kappa statistic	0.32	0.64	0.51	0.70	0.77	0.69	0.55	0.54	0.71	0.80	0.61
Correctly Classified Instances	80 %	88 %	84 %	90 %	92 %	89 %	84 %	85 %	90 %	93 %	86 %
Incorrectly Classified Instances	20 %	12 %	16 %	10%	08 %	11 %	16 %	15 %	10 %	07 %	14 %
TP Rate	0.80	0.88	0.84	0.90	0.93	0.89	0.84	0.86	0.90	0.93	0.86
FP Rate	0.53	0.25	0.36	0.24	0.21	0.19	0.28	0.38	0.19	0.18	0.23
Precision	0.78	0.87	0.83	0.90	0.92	0.89	0.84	0.85	0.90	0.93	0.87
Recall	0.80	0.88	0.84	0.90	0.93	0.89	0.84	0.86	0.90	0.93	0.86

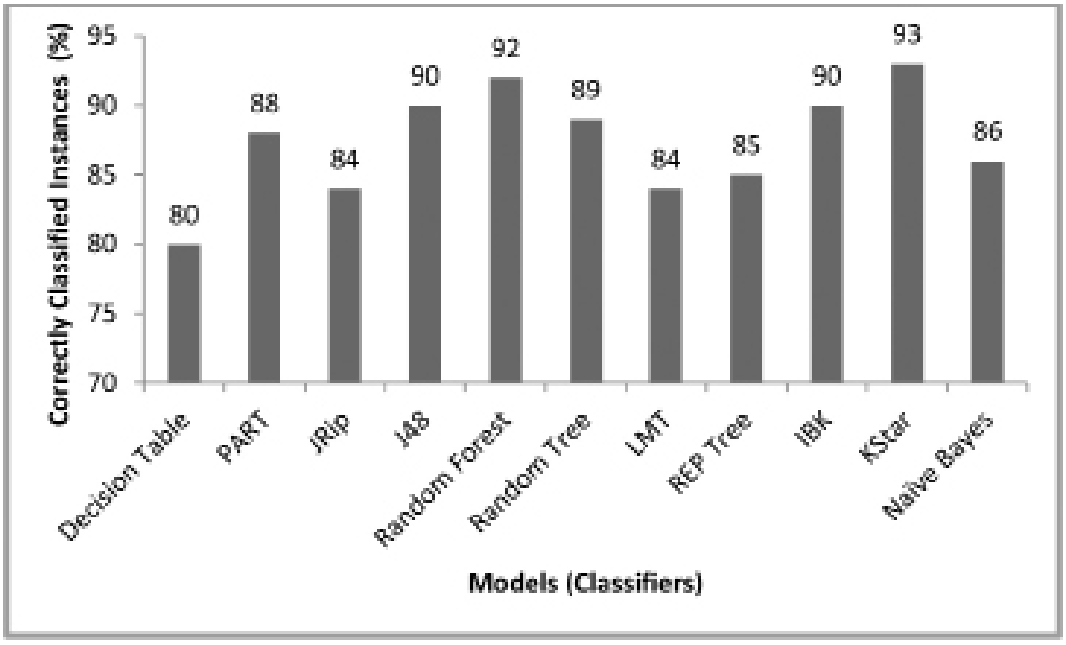

Fig. 2: Prediction accuracy for classification models

Figure 2 shows the prediction accuracy for different classification models. Out of eleven models used in this research work, KStar predicts better than other classification models with 93 %, followed by Random Forest with 92 % predictability. While Decision Table classification is the lowest with 80 % predictability.

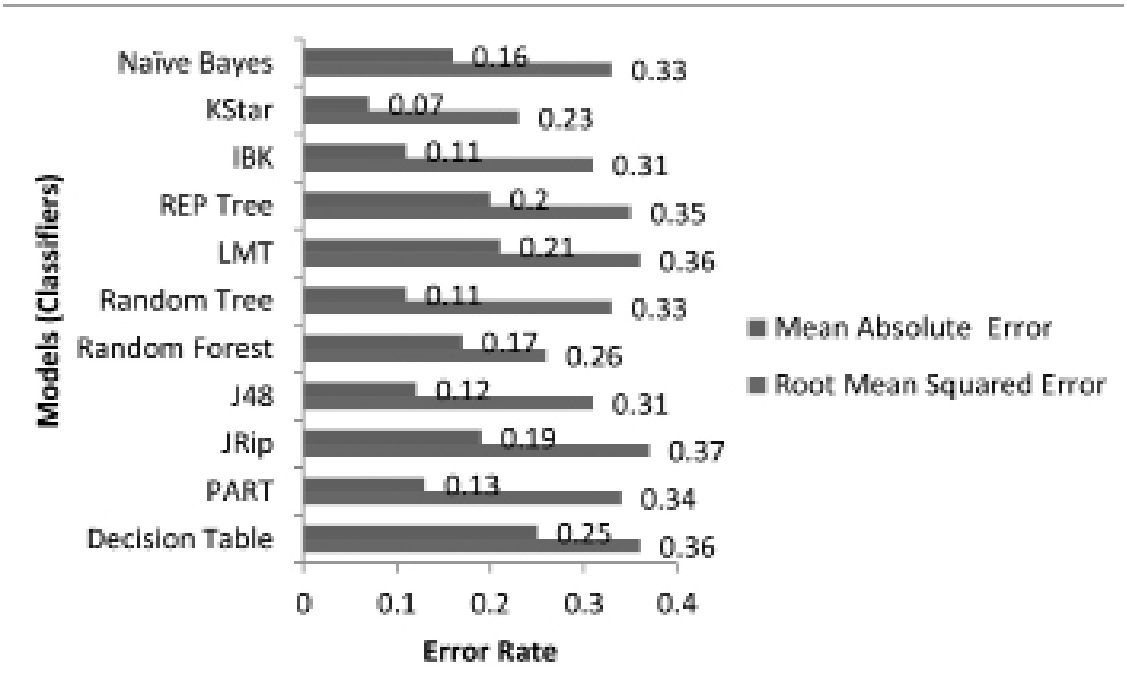

Fig. 3: Error rate for classification models

Figure 3 shows the error rates reported for different classification models, KStar has Mean Absolute Error (MAE) of 0.07 and Root Mean Squared Error (RMSE) of 0.23 respectively. This shows minimal error reported during the prediction processes while Decision Table has the high error rate of 0.25 and 0.36 MAE and RMSE respectively.

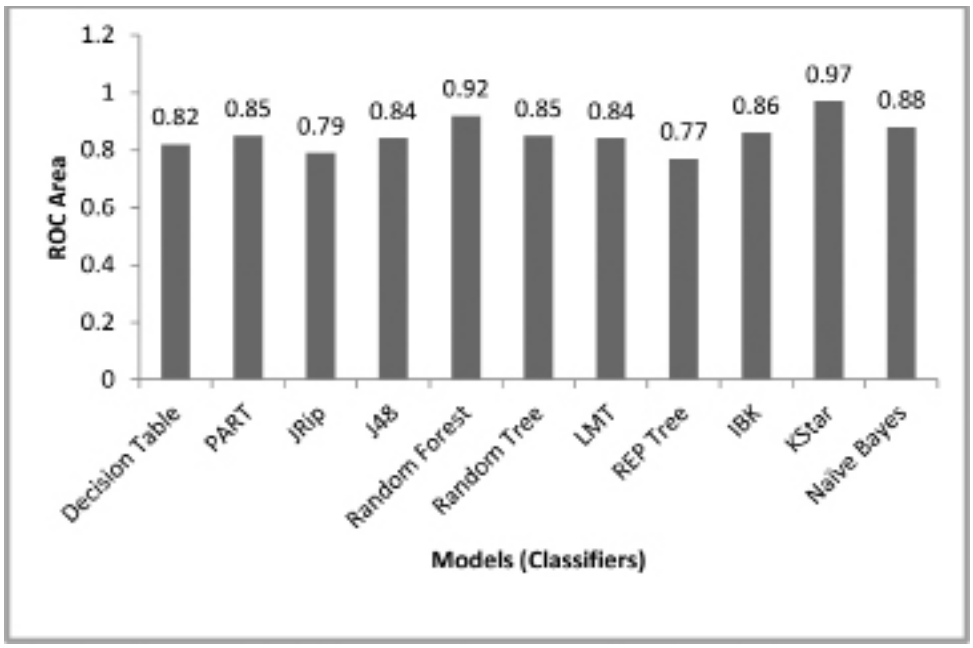

Fig. 4: ROC area for classification models

Figure 4 shows the Receiver Operating Characteristic (ROC) Area curve for different classification models. In figure 4, KStar recorded a high ROC area with 0.97 followed by Random Forest with 0.92. Therefore, the KStar performs better than other models when compared with classification accuracy and error rates to build the model metrics.

CONCLUSION

The aim of this paper is to study and compare different data mining approaches namely DecisionTable, PART, JRip, J48, RandomForest, RandomTree, LMT, REPTree, NaiveBayes, IBK and KStar for prediction of e-extension employability of the postgraduate scholars. The fitted models, using data mining approaches suggested that predictor variables namely job preference, interpersonal communication, information collection behaviour, self confidence and mother's education had higher influence on e-extension employability of the postgraduate scholars. Based on all the benchmarks used to measure the models employed in this study, it was discovered that KStar model with 93% accuracy is the most appropriate in terms of predictability based on this data. Focus was therefore laid on designing a predictive system on the most suitable algorithm which is KStar in this particular domain.

REFERENCES

Ahamed, A.; Mahmood, N.; Hossain, N; Kabir , M.; Das, K.; Rahman, F.; and Rahman, R. (2015). Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh. 16th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing (SNPD). : 1- 6

Bhuyar, V. (2014). Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District. *International Journal of Emerging Trends & Technology in Computer Science*., 3(2) : 200-203.

Boppana, J., H. M., Vinaya Kumar and Patel, J. B. (2019) Attitude of postgraduate students towards research. *Guj. J. Ext. Edu.* 30(1):87-89.

Elhamayed , S.A. (2016). Enhancement of Agriculture Classification by Using Different Classification Systems. *International Journal of Computer Applications.*, 3(1):08-12

Kalekar, A. ; Vispute, S.; Kokane, P.; Kamble, M. and Bokefode, K. (2018). Automated Generation and Analysis of Soil Health Card and Calculation of the Village Soil Fertility Index. *International Journal of New Technologies in Science and Engineering.,* 5(3) : 118-125

Kumari, AV. and Chitra, R. (2013). Classification of Diabetes Disease Using Support Vector Machine. *International Journal of Engineering Research and Applications* (IJERA)., 7(1) : 67:70.

Liao, SH. (2003). Knowledge Management Technologies

and applications Literature review from 1995 to 2002. Expert System with Application., 25:155-164

Manish, Jain. (2009). Data Mining: Typical Data Mining Process for Predictive Modeling. BPB Publications. , 235-241

Mucherino, A.; Papajorgji, P. and Pardalos, PM. (2009). A survey of data mining techniques applied to agriculture. Oper Res., 9 (2) : 121-140

Rajeswari, V. and Arunesh, K. (2016). Analysing Soil Data using Data Mining Classification Techniques. *Indian Journal of Science and Technology.*, 9(19) : 1- 4

Raunak, J. (2018). Applying Naive Bayes Classification Technique for Classification of Improved Agricultural Land soils. *International Journal for Research in Applied Science & Engineering Technology.*, 6(5): 189-193

Wah, TY. ; Abu, Bakar, Z. (2003). Investigating the Status of Data Mining in Practice. CiteSeerx College of Information Science and Technology Pennsylvania State University. ,105:110

Wei, Fan.; Albert, Bifet. (2013). Mining Big Data: Current Status, and Forecast to the Future. *SIGKDD Explorations.,* 14(2): 123:135

Witten, IH.; Eibe, Frank. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Second Edition (Morgan Kaufmann Series in Data Management Systems). , 189-303.

Received : May 2020 : Accepted : August 2020