ISBN 978-93-84422-80-6
17th IIE International Conference on Computer, Electrical, Electronics and Communication Engineering (CEECE-2017)
Pattaya (Thailand) Dec. 28-29, 2017

Building a Data Mining Model & Comparative Analysis to Target Student Admission

Arif Uddin Ahmed Palash[1] and MasudKarim[2]

[1] Daffodil International College

[2]Daffodil Institute of Science & Technology, Dhaka, Bangladesh, palash.a@yahoo.com, masud.mkarim@gmail.com

***Abstract*:** *Data mining tools and techniques are being developed and applied in different sector to predict information, market survey, decision making, knowledge discovery etc. Previously unknown hidden pattern and theory are found out from large database and data warehouse using data mining function. Different classifiers are frequently being used regarding such issues. In this research we have proposed a data mining model to identify prospective students for admission. We have used five popular classifier algorithms: Naïve Bayes, K nearest neighbor (KNN), Decision Tree (J48), Support Vector Machine (SVM) and ZeroR. The results of these algorithms are analyzed comparatively. The model is tested with aprospective student admission database. Performancesof these classifiers are also illustrated and analyzed.From this study we have found that support vector machine shows comparatively better accuracy.*

Keywords: *Data Minging, SVM, k Nearest Neighbor, Naive Bayes, Student Data, Classification, Machine Learning.*

1. Introduction

Data mining techniques play a strong role in marketing and target group selection. In private educational institutes like training centers, language centers, libraries, consulting firm of student admission etc and even private universities depend on student admission for smooth operation and revenue. They have a target number of student admission. Generally there is a query or information desk where visitors and interested person come to receive information and counseling. From these interested people and visitors some take admission and others do not. But the staff and management don't know who are going to get admitted there. So they waste more time to communicate all visitors and follow up in regular basis. Their marketing and publicity process also expands and increase operational cost.

But if they can identify the target group then they can save both time and money. Revenue will also increase. We have applied data mining functions to build such model and from the model we can target the peoplewho are to be admitted.

2. Related Work

In [3] V.Thavavel and S.Sivakumar presents text mining in distributed environment. They proposed a framework to analyze privacy preservation for distributed data mining. Unstructured data is converted into structured form using XML. Then they have applied their proposed method and data mining tools over the structured data.

In [8] Dr. S.Vijayaraniand others uses a medical dataset for classification. The data set is collected from UCI repository to predict heart disease. They have analyzed three through classification algorithms: logistics, multilayer perception and sequential minimal optimization. They have analyzed performance of these

classification functions. WEKA data mining tool is used for comparative analysis. They have shown that logistics classification algorithm provides the best result in this case. True positive (TP) rate, F Measure, ROC area and Kappa statistics are used to measure the accuracy.

In [9] AshokkumarVijaysinhSolankiand others applied open source data mining tool WEKA to predict sickle cell disease. They have used decision tree classifications. They also presented a comparison of two algorithms, J48 and Random Tree. After the experiments, it was shown that using Random tree is better than J48. Random tree produces details decisions by comparing to J48 which is very much useful for further classification of each node. They have emphasized genetic Sickle Cell Disease (SCD). This research is helpful to the society of medical sector and government department for the improvement of medical sectors.

In [10] Lambodar Jena and others used different classification algorithms in their research. They applied these algorithms on chronic kidney disease related fields. The data set is downloaded from UCI machine learning repository. They have used WEKA data mining tools for classification and describe a comparison analysis among the performances of six algorithms. Based on the analysis the researchers have illustrated that the multilayer perceptron has the best accuracy comparing Naïve Bayes, SVM, J48, conjective rules and decision tables. The objective of this research is to predict the target class accuracy for each case in the data. The researchers found out suitable algorithm for diagnosis and prediction of chronic kidney diseases.

In [11] ArunaGovada and others have proposed an algorithm for classification. Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm is used to handle noisy data sets. Their proposed algorithm implements RIPPER at local level and then combines the output of local level into global level. In global level they have implemented distributed environment. They have used five distinct datasets distributed in different nodes. The accuracy of the algorithm is calculated by parameters, time taken for rule generation, accuracy in each iteration and testing accuracy.

3. Research Background

3.1. A. Student AdmissionTarget

Most of the educational business have target of a certain number of student admission. Admission depends on student's interest about the course. To take admission they have queries about the courses and institution. Students take admission based on answers of these queries.

3.2. B. The Dataset

To test our model we have used a dataset of student's queries. After the query they have decided to choose course and admission. The data set is collected from [14] which is a collection of data about visitors or students those have visited the institute physically or called to take information. We have taken two years of data such as: 2015 and 2016. Different staffs have handled the queries. For making the data usable for WEKA tools the data set is preprocessed.A sample of dataset is in table I.

The status column shows the admission status. If the visitor takes admission then status shows AD otherwise Nad. Theseare the classes where we need to classify.

TABLE I: A Sample of Dataset

month	contact	course	session	Qtype	staff	status
1	016	GD	march	P	anawar	Ad
1	019	Exel	march	NA	anawar	Nad
1	016	Apps	march	M	anawar	Nad
1	017	GD	march	T	anawar	Nad
1	16	DWEC	march	M	anawar	Nad
1	019	DWEC	march	P	anawar	Nad
3	019	COP	june	M	roisul	Nad
3	017	IFY	june	M	farhana	Nad
3	017	IFY	june	P	al-amin	Ad
3	016	DWEC	june	P	anawar	Ad
3	017	DWEC	june	P	roisul	Ad

3.3. Classification and Classification Algorithms

The power of data mining is physically shown by data mining algorithms. These algorithms create data mining model from data. The model is created by analyzing the data, looking for specific types of patterns or trends. From this patterns or trends results appear. The algorithms also use the results of this analysis to define the optimal parameters for creating the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics. Then we get the actionable knowledge [3] [10]. There are many algorithms developed for classification. Different algorithms have different uses. Choosing the right algorithm is also a challenging task. Different algorithms have different objectives to perform different business goal [3] [11]. Sometimes algorithms are in same style but different in result presentation. It is also a difficult to classify the data mining algorithms in specific fashion.

4. Proposed Models

Due to increasethe flow of data and information, data mining algorithms are becoming popular for classification and knowledge discovery [1][4][5]. Knowledge grid framework is using to discover target group [3]. To identify a target group, our proposed model works in some steps.

First previous data are collected and stored in a database. Generally educational institutes have procedure and rules to preserve information of interested people and visitors. It may require preprocess the data stored in database. Whenever collecting the data there are many information and all the information is not necessary for applying data mining tools. Sometimes we need to convert data into other format or change the data type. After the preprocessing data, we will apply data mining tools & techniques to build different models. From these models, we select the best one. The selection of model depends on accuracy, time taken to build the model, classification correctness and others performance of data mining. The figure 1 shows a typical framework to select the target group.

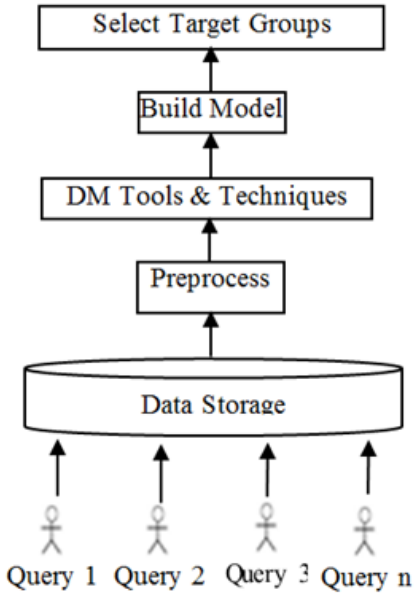

Fig 1A: Typical Diagram of Proposed model.

5. Experiment Setup

To build the modelwe have applied classifier algorithm. Different powerful classifier algorithms [4] [8] [12] are being used by data mining researchers. Some researchers use a combination technique with classification algorithms [3] [5] [6] [8] [11]. We have used fivealgorithms to predict prospective students for admission. WEKA data mining tool is used in this research for experiment [3] [8] [10]. WEKA is freely available JAVA based tool. There is a collection of classification algorithms in this tool. In this research performances of these

algorithms are also compared with each other for finding the best classifier.As we have say a brief in section III, the datasetcontains 1805 instances of different queries, which is used to test the model [14]. ARFF file is created from the dataset. ARFF is Attribute Relation File Format is an ASCII text file. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for usesof the WEKA machine learning software. This file describes a list of instances sharing a set of attributes.We have used 66% instances for training and 34% instances for testing. A collection of data of two years is used to test the model.

TABLE II: Dataset Description

Attributes	Descriptions
Month	Numeric value, identify month like February for 2, September for 9 etc.
Contact	Numeric value, Shows the first three digit of contact number to identify the phone operator that's he is using.
Course	Nominal value, identifies interested course or query subjects.
Session	Nominal value, for session in year.
Qtype	Nominal value,shows query type e.g P for physically visit, M for mobile call etc.
Staff	Nominal values. It is the staff who handles the query or provides information to visitor.
Status	Nominal value for target class.

6. Experimental Result & Discussion

We have find out results using WEKA. As for experiment we have used five popular classifier algorithms, Naïve Bayes, K nearest neighbor (KNN), Decision Tree(J48), Support Vector Machine (SVM) and ZeroR. For using WEKA the datasets is converted into ARFF file format. After the training and testing, performanceare found as the below.

TABLE III: Classifier Performance

Classifier	Kappa Static	Correctly Classified (%)	Incorrectly Classified (%)	Time to Build Model(in Second)
Naïve Bayes	0.2553	81.759	18.241	0.03
KNN	0.2073	77.5244	22.4756	0.02
J48	0	82.899	17.101	0.19
SVM	7.1	84.202	15.798	7.1
ZeroR	0	82.899	17.101	0.02

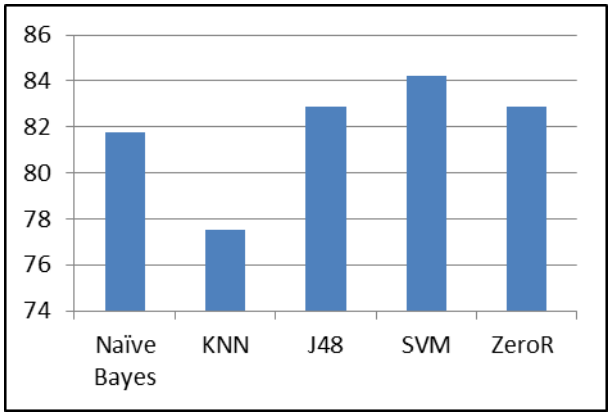

Fig 2: Accuracy of Classifiers

After the training and testing we have seen that Support Vector Machine comparatively shows better result.Its' accuracy is also topmost in this scenario. On the other hand for ROC area Naïve Bayes creates more area to predict accurate classes.

TABLE IV: Detail Accuracy of Classifier

Classifier	Precision	Recall	F Measure	ROC Area
Naïve Bayes	0.793	0.818	0.802	0.772
KNN	0.775	0.775	0.775	0.651
J48	0.687	0.829	0.751	0.5
SVM	0.828	0.842	0.79	0.553
ZeroR	0.687	0.829	0.751	0.5

From the above table and classification results, it is understood that KNN and J48 bring good result. So for shortage of page limit we have illustrated Classification curve and ROC curve of these two classifiers. ROC curve is shown for predicting class of admitted student.

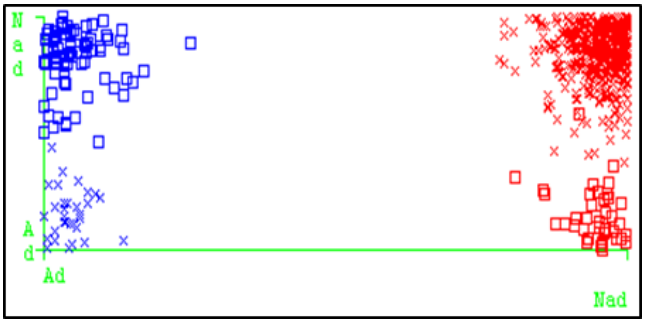

Fig 3: Classification by Naïve Bayes

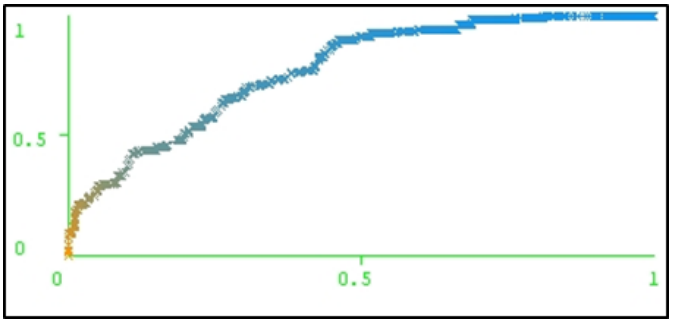

Fig 4: ROC curve by Naïve Bayes

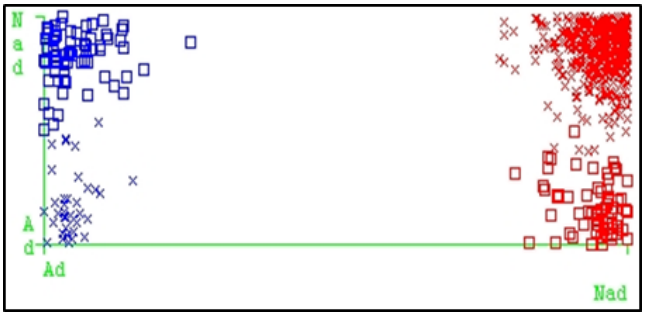

Fig 5: Classification by KNN

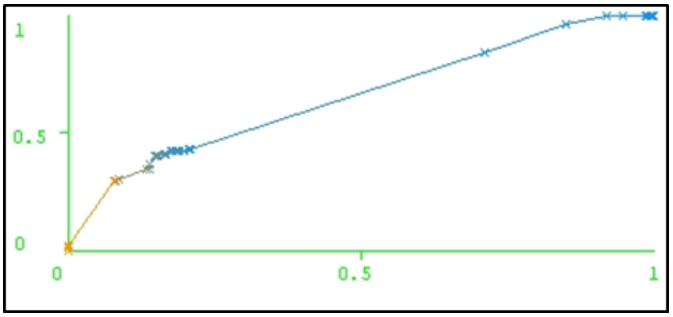

Fig 6: ROC curve by KNN

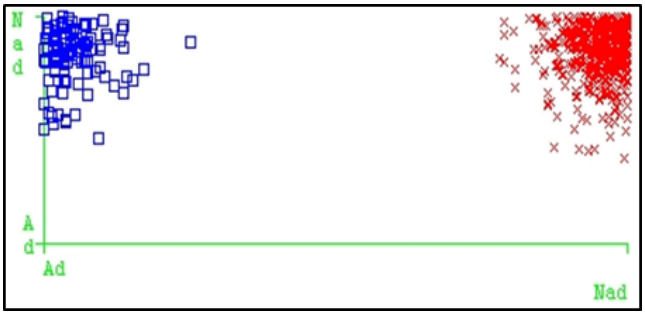

Fig 7: Classification by J48

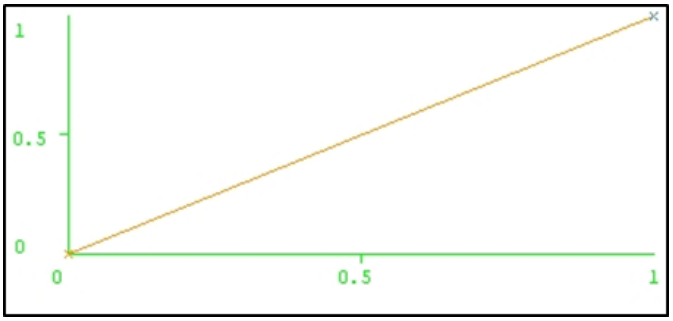

Fig 8: ROC curve by J48

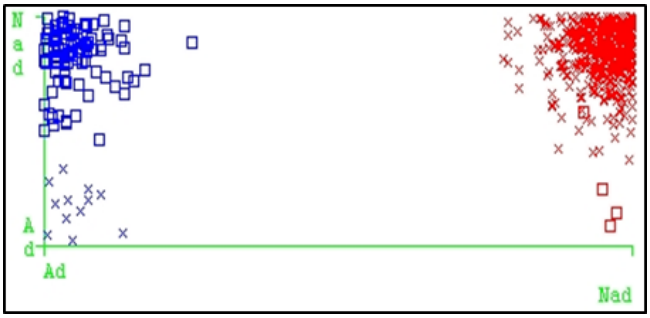

Fig 9: Classification by SVM

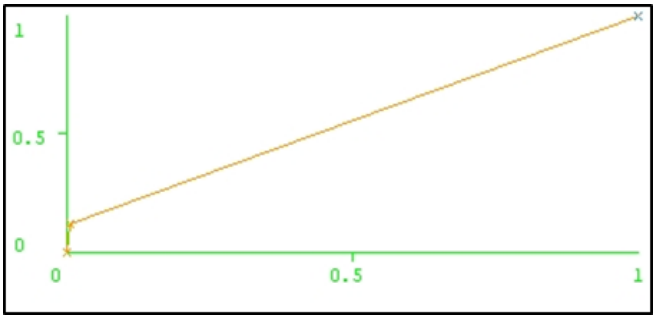

Fig 10: ROC curve by SVM

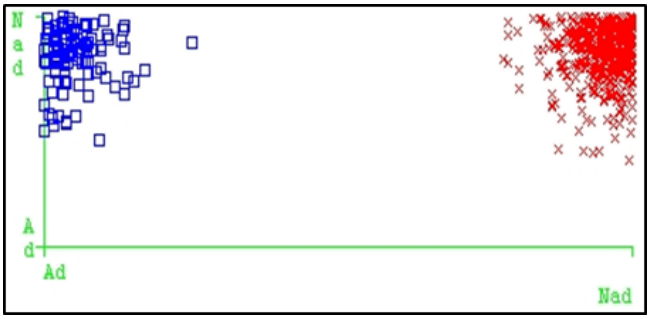

Fig 11: Classification by ZeroR

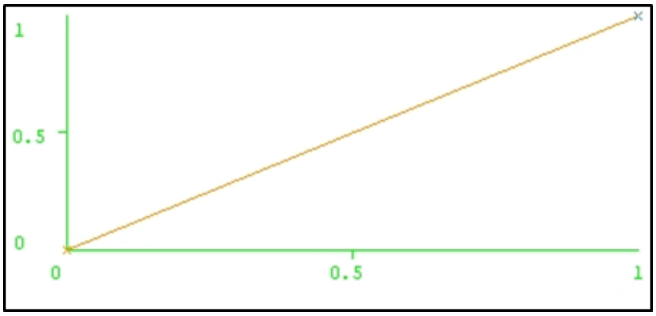

Fig 12: ROC curve by ZeroR

7. Conclusion

In this study we have analyzed the performance of five classifier algorithms. Using the algorithm we have proposed a model to identify a target group. The model needs to be built based on more collected data. The model will be more effective if it can be implemented as distributed fashion. For further development it may be redesigned for distributed environment.

8. Acknowledgement

We thank to Daffodil Institute of Information Technology and its students for experimental task. The institute has supported us by providing experimental data.

9. References

[1] MinghaoPiao, HeonGyu Lee, CoucholPok, Keun Ho Ryu, “A Data Mining Approach for Dyslipidemia Disease Prediction Using Carotid Arterial Feature Vectors”,2010 2nd International Conference on Computer Engineering and Technology, Volume 2.

[2] DaraneeThitiprayoonwongse, PrapatSuriyaphol, NuanwanSoonthornphisaj, “A Data Mining Framework for Building Dengue Infection Disease Model”, The 26th Annual Conference of the Japanese Society for Artificial Intelligence, 2012.

[3] V.Thavavel and S.Sivakumar, “A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment”, International Journal of Computer Science Issues, Vol 9, Issue 1, No 2, January 2012, ISSN (Online): 1694-0814.

[4] Mrs. R. Vidhu, Mrs. S. Kiruthika, “A New Feature Selection Method for Oral Cancer Using Data Mining Techniques”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 1, January 2016

[5] A.ShameemFathima ,D.Manimegalai and NisarHundewale, “A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue”, International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011, ISSN (Online): 1694-0814.

[6] Hem JyotsanaParashar, Singh Vijendra, and NishaVasudeva, “An Efficient Classification Approach for Data Mining”, International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.

[7] R. Thanigaivel and K. Ramesh Kumar, “Boosted Apriori: an Effective Data Mining Association Rules for Heart Disease Prediction System”, Middle-East Journal of Scientific Research 24 (1): 192-200, 2016, DOI: 10.5829/idosi.mejsr.2016.24.01.22944.

[8] Dr. S.Vijayarani, S.Sudha, “Comparative Analysis of Classification Function Techniques for Heart Disease Prediction”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 3, May 2013, ISSN (Print) : 2320 – 9798, ISSN (Online): 2320 – 9801.

[9] AshokkumarVijaysinhSolanki, “Data Mining Techniques UsingWEKA classification for Sickle Cell Disease”, International Journal of Computer Science and Information Technologies Vol.5 (4), 2014, 5857-5860.

[10] Lambodar Jena, Narendra Ku. Kamila, “Distributed Data Mining Classification Algorithms for Prediction of Chronic-Kidney-Disease”, International Journal of Emerging Research in Management &Technology, ISSN: 2278-9359 (Volume-4, Issue-11), November 2015.

[11] ArunaGovada, Varsha S. Thomas, IpsitaSamal, Sanjay K. Sahay, ”Distributed multi-class rule based classification using RIPPER”, 2016 IEEE International Conference on Computer and Information Technology, DOI 10.1109/CIT.2016.111.

[12] Nikhil N. Salvithal, Dr. R. B. Kulkarni, “Evaluating Performance of Data Mining Classification Algorithm in Weka”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 10, October 2013 ISSN 2319 - 4847

[13] Jun Wang, Jin-Mao Wei, Zhenglu Yang and Shu-Qin Wang, “Feature Selection by Maximizing Independent Classification Information”, IEEE Transactions On Knowledge And Data Engineering,2016. DOI 10.1109/TKDE.2017.2650906.

[14] Daffodil Institute of Information Technology, House-02, Road-01, Sector-06, Uttara, Dhaka, Bangladesh.