AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Applying Selected Data Mining Techniques on a Higher Educational Data set Using Oracle Data Miner

[1]Sarah Abbas and [2]Md. Nasir Sulaiman

[1,2]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia,43300 UPM Serdang, Selangor, Malaysia.*

ARTICLE INFO

Article history:
Received 10 October 2014
Received in revised form
22 November 2014
Accepted 28 November 2014
Available online 1 December 2014

Keywords:
Classification techniques,Student's Characteristics, Oracle Data Miner

ABSTRACT

Nowadays data is superabundant in many areas and educational fields. Many researchers have given their attention to the presence of huge amounts of valuable data in the context of higher education. Extracting hidden information via the use of data mining techniques can play a major role in improving the level of education and enhancing organisational performance. The main target of this work is to compare the efficiency of three classification data mining techniques, namely, Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machine (SVM). This is based on Universiti Putra Malaysia students data characteristic and the effect of these characteristics on the choice of university considering the accuracy metric as a measurement. As a result, the NB technique prove to have the highest overall accuracy at 93.12% for correctly classifying the student preference choice compared to DT and SVM. The **c**onclusion of this study is that the best model generated can be applied to new data to examine individual cases and take immediate action to match with a marketing strategy.

To Cite This Article: Sarah Abbas, Md. Nasir Sulaiman., Applying Selected Data Mining Techniques on Higher Educational Data set Using Oracle Data Miner. ***Aust. J. Basic & Appl. Sci.**, 8(18): 321-327, 2014*

INTRODUCTION

Education is now a core factor for most developed countries. Higher education has a major contribution in term of the discovery of new technologies that can serve human needs (Bresfelean and Vasile Paul, 2009). Universities consider heuristic data as a main source to discover the knowledge required to help raise performance. This is possible with the availability of communications to support knowledge transformation and representation for the management of higher educational organisations. Most of the universities are concerned with the prediction of different activities for student behaviours (Luan Jing, 2004) which is achieved by focusing on the preferred courses and fields (Kalathur S., 2006). This is in addition to the type of interest and support required to allow the students to accomplish their academic journey. Scientific investigators (Vandamme J.P. *et al*., 2007) have considered studying the characteristics of the students by classify them into different groups with causes that apparently lead to low performance in the tests. This includes the circumstances surrounding the students and the amount of knowledge they have or they need to acquire. Accordingly, a good proposal can be suggested to minimize the gap and improve the educational levels (Pimentel E.P. *et al*., 2005). The availability of different kinds of learning system from online learning and intelligent educational systems and its ability to provide all the techniques to share the knowledge with students, as of many examples, registration system for courses selection per each student and learning management system in which will assist the student to provide their feedback about courses objective, material used, lecturers behaviour and techniques used for interactive learning (Krzysztof J. Cios *et al*., 2007). Data mining involves different techniques which vary in performance efficiency, while a positive effect relies on dataset quality and information availability. A data mining process is to be used to extract the knowledge and this will help to find the most effective model and can be consider as a basis for decision support systems (M. Prakash *et al*, 2014). This paper presents different classification methods based on the characteristics of the students from the heuristic dataset in order to achieve a model with the highest accuracy. The selected model will be used in future marketing matching strategies to increase the attraction for new students for the concerned university concerned.

The following sections are organised as follows. Section 2 presents related works concerning data mining applications in higher education. Section 3 provides an overview of different classification tasks using Naïve Bayes, Support Vector Machine and Decision Tree models. Section 4 depicts the experimental details and the

Corresponding Author: Sarah Abbas, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.
E-mail: sarah_kubba@yahoo.com

oracle data miner tool while Section 5 gives the results and discussion. Finally, the conclusion of the study will be provided in Section 6.

Related Works:

A previous researcher reviewed the means by which student retention could be maintained by studying all the aspects concerned that could affect the student's behaviour as seen from the academic point of view, and then proposed possible strategies that could be followed by academic management (C. H. Yu, 2010). Different factors could have an effect on the prediction of student retention such as student transfers with specific credit hours, the place of living and ethnicity. In another study (S. K. Yadav and S. Pal., 2012), student enrolment especially in Master of Computer Applications at an academic level formed the main subject in addition to the use of Iterative Dichotomiser 3 to achieve this prediction task. The resultant tree output shows that degree holders in mathematics will have more interest to enrol on such a course and perform better than science students without prior knowledge in mathematics. Other researchers (Aher and Lobo, 2011) used data set for degree students which include tests grades to classify the students into groups and predict student achievements in exams in future based on Decision Tree. The course managements have been examined from the data mining perspective for online study purposes according to the work done by C. Romero *et al*., 2008. In that study it was determined that a statistical graph can help the instructors to apply clustering techniques for similarity according to selected characteristics among the students based on heuristic data.

Classification Techniques:

To determine the a best predictive model by which different classification methods can be applied, the decision will be based on many criteria, such as data set size. Data preparation is also one of the important steps to build a classification and prediction model. For example, data cleaning, analysis, normalisation and data generalisation. Subsequently, a criterion is applied to evaluate the models based on speed, accuracy, robustness, scalability and interpretability (Micheline Kamber,2001). The objective of using different classification techniques is to find an efficient classification model with regard to accuracy of measurement from among the SVM, NB and DT models. This will help the management to apply the best model achieved in order to obtain a better road map from the marketing perspective for new applicants to the university.

A. Decision Tree:

The decision tree (DT) algorithm was developed at the beginning of the 1980s by Quinlan, who specialised in machine learning and named the technique as ID3. Later on, researchers attempted to extend this achievement and adding some modifications. Afterward, C4.5 was developed by Quinlan as inherited from ID3 and this was considered as a measurement for any comparison with supervised algorithms. Every node will have one directed input and can produce an outgoing path to a new node (Micheline Kamber, 2001). This case can be referred to as the testing node, while others represent the leaves which can be called decision nodes. The attribute values will be the key to split the instances and each test will have one attribute in most cases. Ranges will be used as the presentation for numeric data, although the leaf will contain the most likely possible target based on test result.

B. Support Vector Machine:

SVM performance is sensitive to noisy data, which may require the implementation of another algorithm such as symbolisation to remove it in order to considerably reduce the computational costs considerably (Gunn, S., 1998). Hyperplane (M.A.H *et al*., 2012) can be used by SVM to achieve an efficient data separation belonging to different classes to gain as far a distance as possible. In addition, the minimises the possibility of misclassification and unseen data points (V.N. Vapnik, 1995; V.N. Vapnik,1998). As mentioned by Gunn, S., 1998, there are two types of hyperplane for a regression case. The first dose not allow for training errors and no threshold will be decided, which is called hard margin, While the second is known as a soft margin as it allows for training errors in the case of non separable patterns.

C. Naïve Bayes:

The Naïve Bayes (NB) classifier is a probabilistic model as an application of the Bayes theorem. The early clarification of the details of the this classifier was undertaken by R. O. Duda and P. E. Hart in 1973 and later by P. Langley in the early of 90's (R. O. Duda and P. E. Hart, 1973) (P. Langley, 1992). The computation required to classify new objects is achieved by calculating the conditional probability based on the dependence of occurring of one event with other events. Prior probability is the pure probability for the event or class without dependency on the existence of other information. Posterior or inverse probability is the probability of the hypothesis or data that has been observed or with the application extra data (R. O. Duda and P. E. Hart,1973). he NB classifier is widely used in large dataset classifications.

Australian Journal of Basic and Applied Sciences, 8(18) December 2014, Pages: 321-327

Experiments:

In this experiment, the aim is to gain knowledge that will help in the prediction of future individual cases for those applicants who choose Universiti Putra Malaysia (UPM), as a low preference as compared to other universities. Subsequently, it will be used for marketing purposes to attract the new students to take up studies in this university.

A. Student Data set:

The selected data source has been retrieved from UPM student system which contains all information starting from early registration such as the student profiles, qualification backgrounds, skills and subject preference etc. In addition this includes a consequent follow up of the academic progress level till graduation. The UPM database schema contains many tables. However, this current work will be implemented based on one table which is the applicant course choice which includes many attributes. Some of the attributes are selected to fit the study requirements and the effect on the target model to be built (Muhammed Basheer *et al*., 2013), while some of the attributes have incomplete data and cannot be fixed. In Table 1, will show the nine selected attributes are shown with possible input data ranges to form the baseline for this work, and the extraction process will be done manually.

Table 1: Attributes of the data set

Attributes	Input Data	Attributes Data Summary
ID	starts from 1	Represents student Record ID
BUMISTATUS_ID	1 and 2	Local or International
CGPA_RANGE_ID	1-8	CGPA for students with eight distinct value
CHOICE_ID	1-8	Preference choice of student in the university
GENDER_ID	1and 2	Male, Female
LEVEL_OF_STUDY_ID	1-11	Level of study the applicant: Master, Degree, etc.
PROG_APPLIED_ID	1-1559	Faculties available in UPM
MARITALSTATUS_ID	1-5	Married, Single, Widow, Widower and others
AGE_APPLIED	17-55	Age of applicant

B. Data Preprocessing:

Pre-processing is an important step to be applied for the final selected attributes in order to ensure data validity. This ensures the data will be free of inconsistent or missing values to avoid any effect on the accuracy of the results for the output of the classification techniques. Figure1 shows the null values for a few attributes.

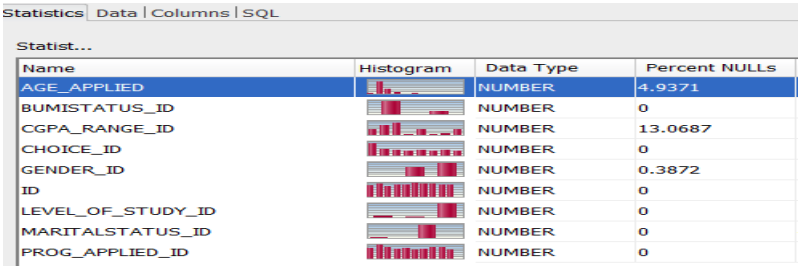

Fig. 1: Attributes with Null Values

- ***Data Cleaning:***

Null values will be treated by replacing them with the mode or median depending on the data type for the attributes as depicted in Table 2. Data discarding will be performed for the records which hold inconsistent or incorrect data.

Table 2: Attribute Types and Treatment.

Attribute	Type	Treatment
AGE	Ratio	Mean
GENDER	Nominal	Mode
MARITALSTATUS	Nominal	Mode
PROG APPLIED	Nominal	Mode
CGPA RANGE	Ordinal	Mode
BUMISTATUS	Nominal	Mode
LEVEL OF STUDY	Nominal	Mode

SQL statements were used to update the records and fix any data issues. The final data set size was 200k after discarding the inconsistent data.

Australian Journal of Basic and Applied Sciences, 8(18) December 2014, Pages: 321-327

- ***Data Transformations:***

In order to prepare the input data for the mining model, some attributes are in the form of continuous data. Thus, creating a range minimisation as small groups can reduce the effect on the accuracy of final the model. Age_applied will be binned with three groups "<22" , "22-30", "30-55". By performing this step the preparation is complete and the data set can be used to train and test the mining model.

C. ODM Tool:

The Oracle Data Miner (ODM) was used in this study to build a data mining classification model. The UPM student data set was stored in an Oracle base and ODM applied to support the data mining service based on SQL scripts inside the database of oracle. ODM is a graphical user interface. The main usage of ODM in analytics is to help organisations handle different analytical issues concerning their competitors. ODM can be used for prediction purposes as well as research as it allows the extraction of features and has the ability to obtain statistical information by exploring the data, building a mining model, testing and evaluating the results. The final generated model can be applied to new data and the rules used for example in decision support systems. Figure 2 depicts the ODM GUI with sample built classification model.

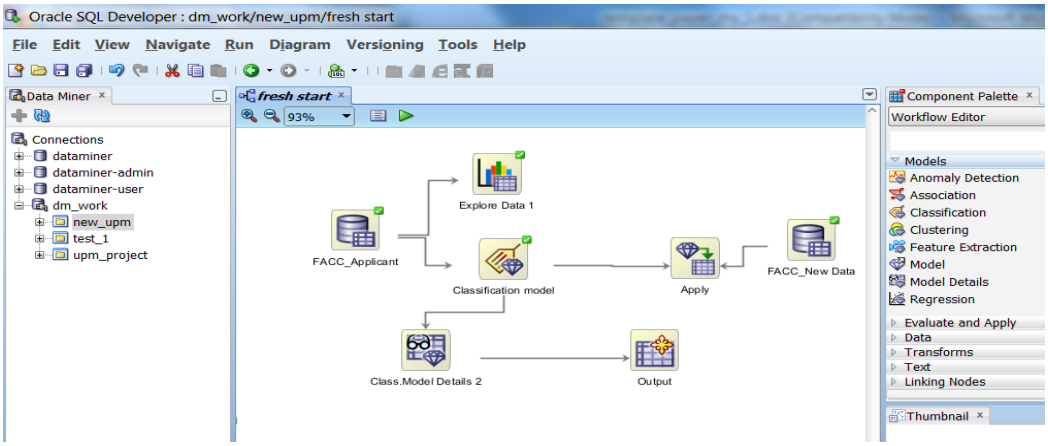

Fig. 2: ODM GUI Classification Model

RESULTS AND DISCUSSION

Three classification techniques (DT, SVM and NB) were applied in this study using the ODM tool. Algorithms were configured and general parameter settings shared for all three classifiers as follows: Target =Choice_ID, Case ID= ID. In DT, Entropy was used as a split node instead of Gini to avoid binary splitting and to allow the multiway splits (C. Romero *et al*., 2008). In Figure 3, the depth of the DT was set to 5 as can be seen in order to obtain a detailed information about the classes. Otherwise, the generated rules would not be helpful (Muhammed Basheer *et al*., 2013). The NB parameter settings, the default values were kept as shown in Figure 4. The linear kernel was chosen as part of the SVM algorithm settings, as it provides a good result for a data set with many attributes. In addition it was easy to interpret the result. It was appropriate to keep the default value for the tolerance value. However, for the active learning aspect it was felt that there was no need to maintain it in an active state because this would likely have a negative effect on the output results (http://www.oracle.com/technetwork/database/options/odm/odminer-release-notes-323189.html, 2011).

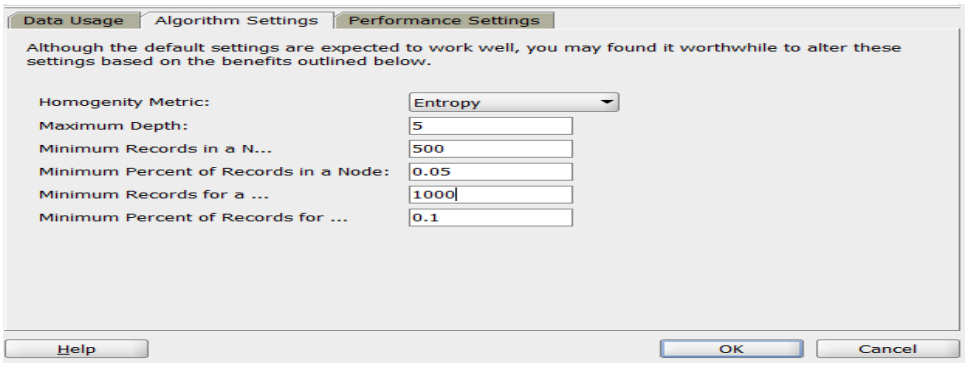

Fig. 3: DT Algorithm Settings

Australian Journal of Basic and Applied Sciences, 8(18) December 2014, Pages: 321-327

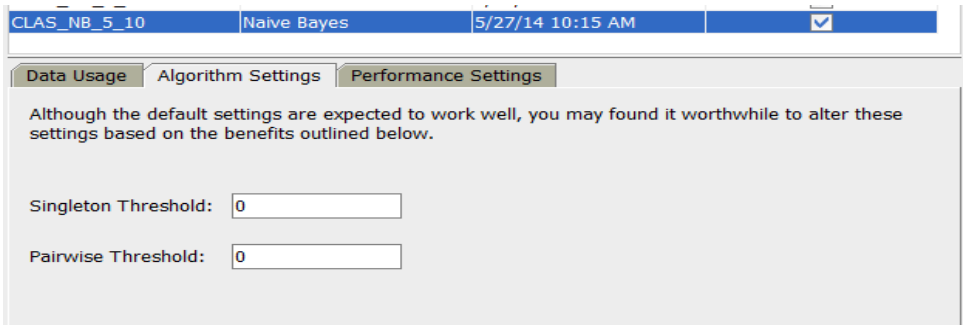

Fig. 4: NB Algorithm Settings

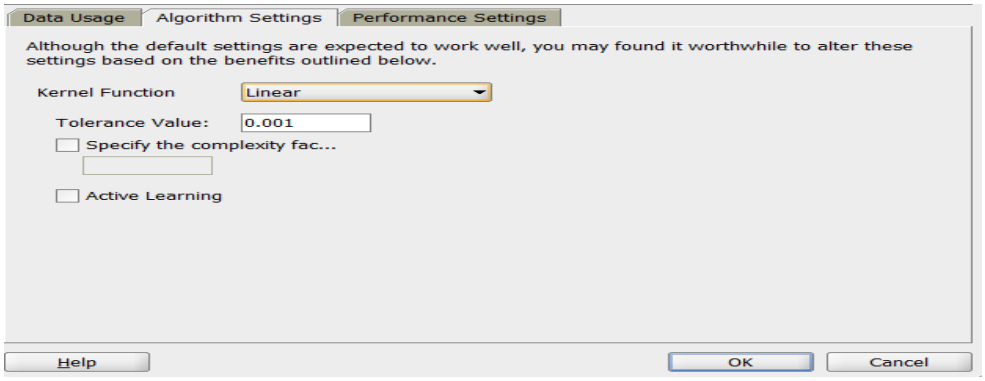

Fig. 5: SVM Algorithm Setting

the data set at 200k was split into three different percentages for training and testing and this was applied to all three algorithms in order to achieve the most accurate results for the targeted classes. In Figure 7, the final constructed models can be seen. The performance of the models has been evaluated using the overall accuracy measurement formula. This achieved the desired comparisons which can be easily shown from the ODM workflow performance tab in two different representations either by graph or by values. In Table 3, the data set percentage used can be seen with the output result for each model.

$$\text{Accuracy} = \frac{\text{Sum of correct classifications}}{\text{Total number of classifications}}$$

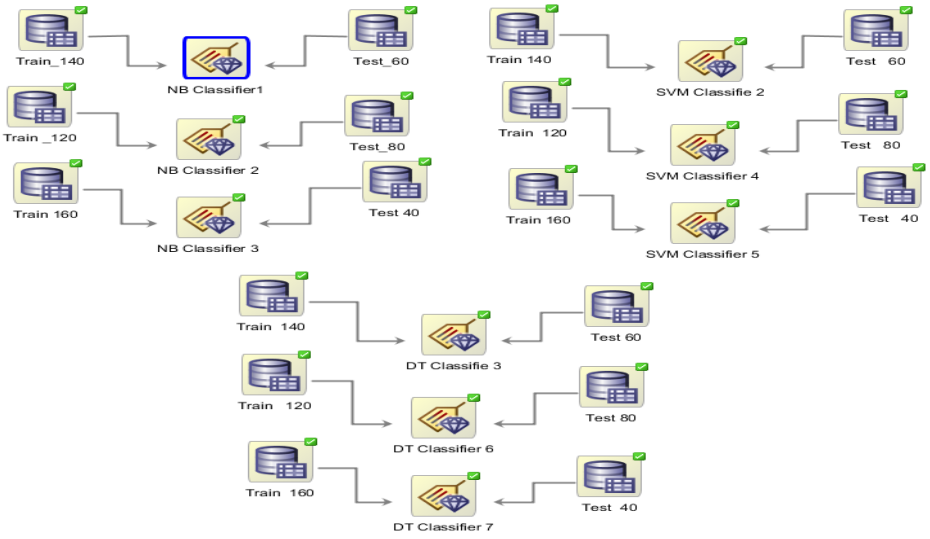

Fig. 6: Classification Models

Australian Journal of Basic and Applied Sciences, 8(18) December 2014, Pages: 321-327

Table 3: Models performance Comparisons

	Accuracy			Data Set	
Data Percentage	SVM	DT	NB	Test Data	Training Data
40%-60%	53.35	53.47	45.69	40000	160000
30%-70%	66.42	67.02	67.6	60000	140000
20%-80%	93.08	92.97	93.12	80000	120000

According to the accuracy results, the best accuracy arises from the percentage of 20% percentage for the testing data and 80% for the training data among the other percentages. The data set split had a great effect on the classification result as indicated in Table 3 and this was also confirmed by other researchers (Ratanamahatana *et al*., 2002). In addition, Figure 7 shows Naïve Bayes achieved the highest overall accuracy (93.12%) as compared to the Decision Tree (92.97%) and Support Vector Machine (93.08%).

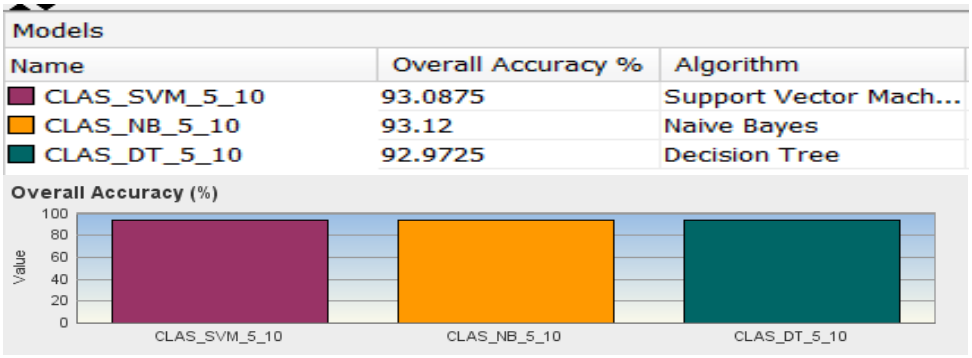

Fig. 7: Highest Overall Accuracy Graph For Three Models

The NB model could correctly classify 37,248 instances for the eight target class value of student preference choice out of 40,000 for the testing data while the result its 37,235 for the SVM and 37,189 for the DT model. In addition, the detailed correct prediction for each class from the performance matrix in ODM can be traced in Figure 8 . It can be seen that the highest percentage of 99.86 was achieved by the target value (choice_id) ='1', and this choice had the highest total in the applicant preference choice table because most of the students preferred UPM as their first preference choice. The other targeted class could be the main focus for immediately identifying individual cases and to take appropriate action. This especially applies to the values from 5 to 8 which represent the low preference choice from the applicants, while the other classes are considered to be in the acceptable range.

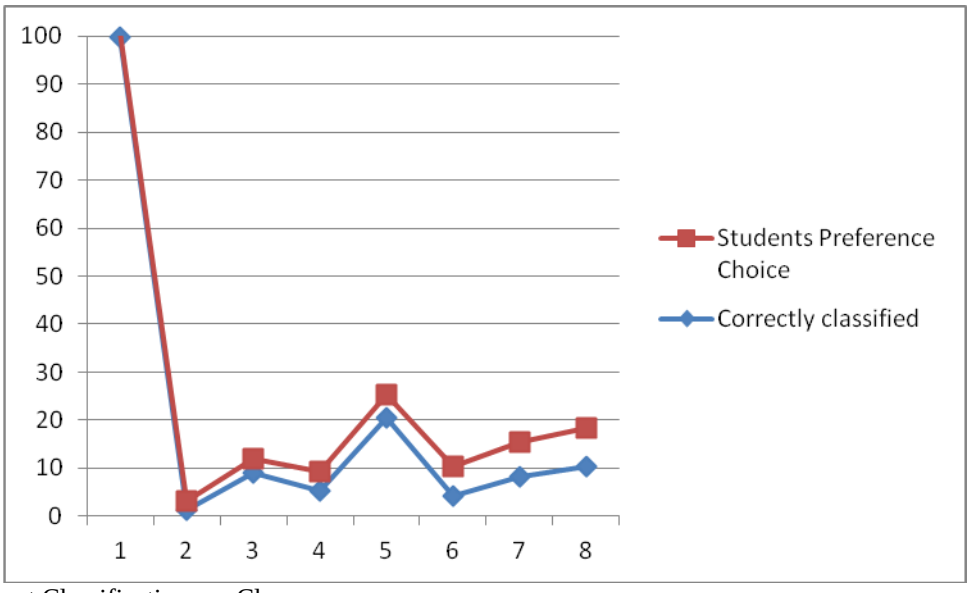

Fig. 8: Correct Classification per Class

Conclusion:

A performance evaluation has been undertaken for three classification techniques (SVM, DT and NB) using different percentages of training and testing data within same data set in order to meet the objective in using the ODM tool. The conclusion of this work is that the best classification technique based on the UPM applicant data set is Naïve Bayes (NB) among SVM and DT with an accuracy of 93.12 % in classifying the applicant preference when choosing a university. The resultant rules of the NB model can be extracted and used in identifying individual cases with a low preference choice based on their profile characteristics. The UPM database can provide a precious source of student data and can be capitalised upon for further knowledge discovery and many enhancements in higher educational level targets.

REFERENCES

Bresfelean and Vasile Paul, 2009. "Data Mining Applications in Higher Education and Academic Intelligence Management". *Theory and Novel Applications of Machine Learning.*

Yu, C.H., S. DiGangi, A. Jannasch-Pennell and C. Kaprolet, 2010."A data mining approach for identifying predictors of student retention from sophomore to junior year". Journal of Data Science, 8: 307-325.

Romero, C., S. Ventura and E. Garcia, 2008. ″Data mining in course management systems: Moodle case study and tutorial″.

Gunn, S.,1998."Support vector machines for classification and regression". *Technical report, ISIS group, University of Southampton.*

Kalathur, S., 2006."An Object-Oriented Framework for Predicting Student Competency Level in an Incoming Class", Proceedings of SERP'06 Las Vegas, pp: 179-183.

Krzysztof, J., Cios,Roman W. Swiniarski and Witold Pedrycz, A. Lukasz, 2007. "Kurgan Data Mining, A Knowledge Discovery Approach", *Springer US.*

Luan Jing, 2004. "Data Mining Applications in Higher Education", SPSS Exec. *Report. http://www.spss.com/home_page/wp2.htm*

Farquad, M.A.H. Indranil Bose, 2012. "Preprocessing unbalanced data using support vector machine". *Decision Support System, Elsevier.*

Prakash, M., G. Singaravel, 2014. "A Review on Approaches, Techniques and Research Challenges in Privacy Preserving Data Mining". Aust. J. Basic & Appl. Sci., 8(10): 251-259.

Micheline Kamber, Han and Jiawei, 2001."Classification and prediction. Data Mining Concepts and Techniques", 2nd ed. Jim Gray, Indian Reprint. Elsevier. Morgan Kaufmann pp: 285-375.

Muhammed Basheer Jasser, Fatimah Sidi, Aida Mustapha, Binhamid and Abdulelah Khaled, April 2013 "T. Article: Mining Students Characteristics and Effects on University Preference Choice: A Case Study of Applied Marketing in Higher Education", *.International Journal of Computer Applications.*

Pimentel, E.P., N. Omar, 2005."Towards a model for organizing and measuring knowledge upgrade in education with data mining", The 2005 IEEE International Conference on Information Reuse and Integration, Las Vegas, USA. pp: 56-60.

Langley, P., W. Iba and K. Thompson, 1992."An analysis of Bayesian Classifiers.", *in Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA.*

Ratanamahatana, Chotirat Ann, and Dimitrios Gunopulos, 2002."Scaling up the naive Bayesian classifier: Using decision trees for feature selection". *In: Proceedings of Workshop on Data Cleaning and Preprocessing (DCAP 2002), at IEEE International Conference on Data Mining (ICDM2002), Maebashi, Japan.*

Duda, R.O. and P.E. Hart, 1973. "Pattern classification and scene analysis", John Wiley Sons.

Aher, S.B. and L.M.R.J. Lobo, 2011."Data mining in educational system using weka". *In IJCA Proceedings on International Conference on Emerging Technology Trends, pages 20–25,New York,Foundation of Computer Science.*

Yadav, S.K. and S. Pal., 2012. "Data mining application in enrolment management: A case study", International Journal of Computer Applications, 41(5).

Vapnik, V.N., 1995."The Nature of Statistical Learning Theory", *Springer-Verlag, New York, USA.*

Vapnik, V.N., 1998."The Nature of Statistical Learning Theory", *2nd edition Springer-Verlag, New York, USA.*

Vandamme, J.P., N. Meskens, J.F. Superby, 2007. "Predicting Academic Performance by Data Mining Methods", Education Economics, 15: 405-419.