DAVAO DEL NORTE STATE COLLEGE
Institute of Information Technology

TERMINAL REPORT

Research Title: **Mining Academic Data to Predict Drop-outs using Machine Learning Algorithm**

Proponents: Ariel O. Gamao
Marie Cielo Funa
Ronald S. Decano
Jumar Buladaco

Funding Source: Davao del Norte State College
Duration: March 2019 – December 2019
Total Project: P 4,250.00

ABSTRACT

The academic database is considered as the heart and soul of every higher education institution. This database contains a vast amount of useful information that is useful for analysis. Algorithms for machine learning play a significant role in mining academic databases and have been proven to be effective when applied in the academic field. Prediction models are made using relevant classification algorithms for dropout analysis. The success of the prediction model depends on the performance of the feature selection algorithm used for dimensionality reduction. The study utilized relevant classifier algorithms, namely: Support Vector Machine (SVM), Decision Tree (DT), and K Nearest Neighbor (KNN), where both algorithms exhibit better performance in predicting dropouts. The results of the simulation revealed that both models underwent a stringent classification process to come up with the accuracy rates concerning each classifier algorithm. It is observed that KNN has a higher accuracy of 91.4% compared to SVM and DT with 87.8% and 84.9%, respectively, using 10-fold cross-validation. This study utilized students' cumulative dataset to come up with a predictive model for dropout analysis of Davao del Norte State College, Davao del Norte, Philippines.

Keywords: Dropout analysis, classifier algorithms, support vector machine (SVM), k-nearest neighbor (KNN), decision tree.

1. INTRODUCTION

Knowledge mining from large databases, especially from the field of academics, plays a significant role in all aspects of human life [1]. From a global perspective, it is an incontestable fact that the nation's progress is likely dependent on the education of its citizens [2]-[3]. Democratizing access to Higher Education (HE) has increased the diversity of students, and this new situation requires a deeper understanding of the students' paths leading to them dropping out or completing their courses [4]. The student's performance prediction is an important research topic because it can help prevent students from dropping out before final exams and identify students that need additional assistance[5]-[6].

Machine learning algorithms have been proven to be effective when applied in the academic field [7]. Classification is a data mining technique used to predict group membership for data instances in different types of problems [8]-[9]. Some classification algorithms may perform quite well in general, but these may be easily outperformed by other algorithms in terms of performance when dimensionality reduction technique or feature selection is not correctly performed [10].

Classification is a data mining (machine learning) technique used to predict group membership for data instances. There are several classification techniques that can be used for classification purposes. In this paper, we present the basic classification techniques. Later we discuss some major types of classification methods, including Bayesian networks, decision tree induction, k-nearest neighbor classifier, and Support Vector Machines (SVM) with their strengths, weaknesses, potential applications, and issues with their available solution [11].

The Support Vector Machine is a theoretically superior machine learning methodology with great results in classification of high- dimensional datasets and has been found competitive with the best machine learning algorithms. In the past, SVMs have been tested and evaluated only as pixel-based image classifiers. Moving from pixel-based techniques towards object-based representation, the dimensions of remote sensing imagery feature space increases significantly [12].

With increasing amounts of data being generated by businesses and researchers, there is a need for fast, accurate, and robust algorithms for data analysis. Improvements in database technology, computing performance, and artificial intelligence have contributed to the development of intelligent data analysis. Support vector machines are a specific type of machine learning algorithm that is among the most widely-used for many statistical learning problems, such as spam filtering, text classification, handwriting analysis, face and object recognition, and countless others. Support vector machines have also come into widespread use in practically every area of bioinformatics within the last ten years, and their area of influence continues to expand today [13].

Supervised Machine Learning (SML) is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. Supervised classification is one of the tasks most frequently carried out by intelligent systems. This paper describes various Supervised Machine Learning (ML) classification techniques, compares various supervised learning algorithms as well as determines the most efficient classification algorithm based on the data set, the number of instances and variables (features). Seven different machine learning algorithms were considered: Decision

Table, Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Neural Networks (Perceptron), JRip and Decision Tree (J48) using Waikato Environment for Knowledge Analysis (WEKA)machine learning tool. To implement the algorithms, a Diabetes data set was used for the classification with 786 instances with eight attributes as the independent variable and one as dependent variable for the analysis. The results show that SVM was found to be the algorithm with the most precision and accuracy. Naïve Bayes and Random Forest classification algorithms were found to be the next accurate after SVM accordingly [14].

The Support Vector Machine (SVM) has attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. Sims has been employed in a wide range of real-world problems such as text categorization, hand-written digit recognition, tone recognition, image classification, and object detection, microarray gene expression data analysis, data classification. It has been shown that Sims is consistently superior to other supervised learning methods [15].

The success of the classification process depends on the quality of datasets. Features may contain unreliable data, which may lead the classification process to produce undesirable results; thus, a feature selection approach is considered a solution for this kind of problem [16]. Also, the selection of appropriate highlights assumes a fundamental job in the selection process [17]. Feature selection (FS) is an essential machine learning technique for classification applications to achieve an optimal subset of input features [18]. The accuracy of prediction is a primary challenge in training a model [19]. Feature selection techniques do not alter the original features of the variables, but merely selects a subset of them [20]. Also, feature selection is a crucial task in applying machine learning in various fields. Also, the increase of data dimensionality poses a significant challenge to many existing feature selection methods concerning effectiveness and efficiency [7].

The kNN (k-nearest neighbors) classification algorithm is one of the most widely used non-parametric classification methods. However, it is limited due to memory consumption related to the size of the dataset, which makes them impractical to apply to large volumes of data. Variations of this method have been proposed. Such as condensed KNN, which divides the training dataset into clusters to be classified, other variations reduce the input dataset in order to apply the algorithm. This paper presents an adaptation of the kNN algorithm, of the type structure less NN, to work with categorical data. Categorical data, due to their nature, can be compressed to decrease the memory requirements at the time of executing the classification [22].

In this paper, the researchers performed experiments through data mining techniques using the existing students' cumulative record of the Guidance and Testing Office of the Davao del Norte State College, Davao del Norte. The dataset contains the students' cumulative record, covering SY2016-2017 to SY2018-19. The study utilized classification algorithms to explore the best model as means in the analysis of students' dropout to be used by academic administrators in crafting policies as an intervention in minimizing dropout rates.

Objectives of the Study

The study aimed to explore the use of the existing dataset from the Guidance and Testing Office of Davao del Norte State College, Davao del Norte, to be used to train a model through a classification algorithm. Furthermore. Furthermore, the study aimed to utilize relevant classifier algorithms, namely: SVM, DT, and KNN, to determine the best model that can be used to predict students' dropout.

II. Material and Methods

CONCEPTUAL FRAMEWORK

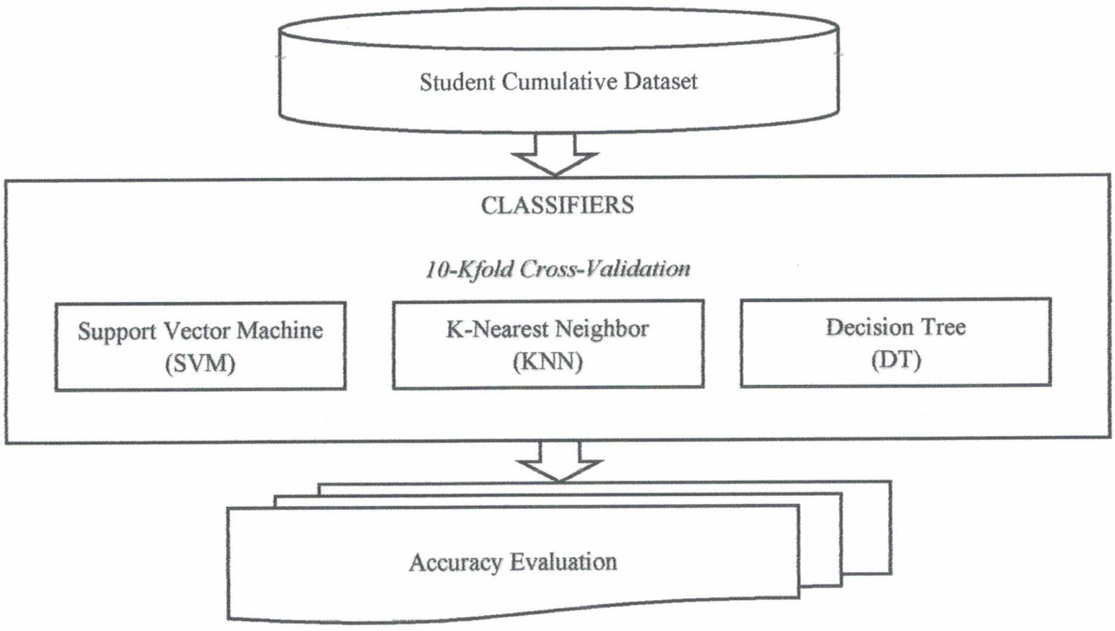

Figure 1. Dropout Prediction Framework

The study utilized the existing dataset available at the Guidance and Testing Office of the Davao del Norte State College, Davao del Norte as the basis for the creation of the appropriate prediction model in analyzing students' cumulative data using who most likely to drop from college (DNSC), as shown in Figure 1.

A. Data Source

The study utilized a cumulative record of freshmen college students of the Guidance and Testing Office of the Davao del Norte State College containing the personal and family data, socioeconomic status, educational data, record of outstanding performance, and tests taken, such as OLSAT and SATT. The said dataset is an official record of the Guidance Office of the college, where students were required to fill in the cumulative student form during the enrollment period, and this contains 45 (see Table 3).

Table3: Attributes of Students' Cumulative Dataset

Dataset	Attributes		
	No. of Features	No. of Classes	No. of Instances
Student Cumulative Dataset	45	2	1862

In the experiment, the study utilized two classification algorithms, namely: Naïve Bayesian and Decision Tree, using a 10-fold cross-validation strategy to arrive on the best model with the best prediction accuracy. The said model can be used in the analysis of student dropouts using the students' cumulative record for the academic administrations to come-up with necessary remediation activities to avoid if not minimize dropouts for students in the whole duration of students' journeys in the college [23], [24], [25], [26].

A. *Data Preparations*

The researchers will update the database stored in the Guidance Office to include the Institute Programs. Once the database is ready for data mining, the researchers will obtain those data as a dataset of this study. Good data preparation is a key prerequisite to successful data mining. Experience suggests that data preparation takes 60 to 80% of the time involved in a data mining study [27].

B. Data selection and Transformation

After the database preparation, fields from the students' cumulative records will be explored, which will be used for mining academic data. The extracted fields shall be considered as predictors for student's dropouts.

C. Implementation of the Mining Model

In this study, MATLAB was used in the simulation process. MATLAB, which stands for MATrix LABoratory, is a state-of-the-art mathematical software package, which is used extensively in both academia and industry. It is an interactive program for numerical computation and data visualization, which, along with its programming capabilities, provides a very useful tool for almost all areas of science and engineering. Unlike

other mathematical packages, such as MAPLE or MATHEMATICA, MATLAB cannot perform symbolic manipulations without the use of additional Toolboxes. It remains, however, one of the leading software packages for numerical computation [28].

III. Results and Discussion

In the experiment, the study utilized two classification algorithms, namely: Support Vector Machine (SVM), Decision Tree (DT), and K Nearest Neighbor (KNN) using a 10-fold cross-validation strategy to arrive on the best model with the best prediction accuracy. The said models can be used in the analysis of student dropouts using the students' cumulative record for the academic administrations to come-up with necessary remediation activities to avoid if not minimize dropouts for students in the whole duration of students' journeys in the college.

Both models underwent a stringent classification process to come up with the accuracy rates concerning each classifier algorithm. It is observed that KNN has a higher accuracy of 91.4% compared to the SVM and DT with 87.8% and 84.9%, respectively, as shown in Table 2.

Table 4: Prediction Accuracy Result of Students' Cumulative Dataset

Classifier Algorithm	Accuracy (%)
SVM	87.8
Decision Tree	84.9
KNN	91.4

4. CONCLUSION

With the experiments conducted in the study, the goal of developing a predictive model for the analysis of dropouts using the students' cumulative record was achieved through the use of the relevant classifier algorithms such as SVM, DT, and KNN. Both algorithms were able to explore the existing dataset in proving the relevance of the said algorithms with their respective accuracy rates.

Thus, the accuracy achieved through the study can be used by academic administrators of the Davao del Norte State College in crafting academic policies that enhance students' performance for both curricular and extra-curricular activities to minimize dropouts. Future researchers may consider improving the dataset in order to optimize the model to be used for decision-making.

REFERENCES

[1] S. Chokkadi, M. S. Sannidhan, K. B. Sudeepa, and A. Bhandary, "A study on various state of the art of the art face recognition system using deep learning techniques," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 4, pp. 1590–1600, 2019.

[2] L. A. Choudhary AI, "Economic Effects of Student Dropouts: A Comparative Study," *J. Glob. Econ.*, vol. 03, no. 02, pp. 2–5, 2015.

[3] I. S. Makki and F. Alqurashi, "An adaptive model for knowledge mining in databases 'EMO_MINE' for tweets emotions classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 7, no. 3, pp. 52–60, 2018.

[4] J. R. Casanova, A. Cervero, J. C. Núñez, L. S. Almeida, and A. Bernardo, "Factors that determine the persistence and drop out of university students," *Psicothema*, vol. 30, no. 4, pp. 408–414, 2018.

[5] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, "Using machine learning to predict student difficulties from learning session data," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 381–407, 2019.

[6] A. K. P. Moore, "DROPPED OUT: FACTORS THAT CAUSE STUDENTS TO LEAVE BEFORE GRADUATION," *ABA J.*, vol. 102, no. 4, pp. 24–25, 2017.

[7] M. Kumar and A. J. Singh, "Performance analysis of students using machine learning & data mining approach," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 3, pp. 75–79, 2019.

[8] E. R. S.Neelamegam, "An Overview of Classification Algorithm in Data mining," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 12, pp. 255–257, 2015.

[9] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019.

[10] J. Gama and P. Brazdil, "Characterization of Classi cation Algorithms 2 Normalizing Meta-Datasets," no. October 2013, pp. 1–12, 2000.

[11] A. Soofi and A. Awan, "Classification Techniques in Machine Learning: Applications and Issues," *J. Basic Appl. Sci.*, vol. 13, no. August, pp. 459–465, 2017.

[12] A. Tzotsos and D. P. Argialas, "Support Vector Machine Classification for Object-Based Image Analysis A SUPPORT VECTOR MACHINE APPROACH FOR OBJECT BASED IMAGE," no. January 2008, 2014.

[13] H. Bhavsar and M. H. Panchal, "A Review on Support Vector Machine for Data Classification," vol. 1, no. 10, pp. 185–189, 2012.

[14] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, 2017.

[15] D. Srivastava, "DATA CLASSIFICATION USING SUPPORT VECTOR," no. February 2010, 2019.

[16] E. M. Mashhour, E. M. F. El Houby, and K. T. Wassif, "Feature Selection Approach based on Firefly Algorithm and," vol. 8, no. 4, pp. 2338–2350, 2018.

[17] M. I. Mohmand, A. Bhaumik, M. Humayun, and Q. Shah, "Science Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse31852019.pdf The Performance and Classifications of Audio-Visual Speech Recognition by," vol. 8, no. 5, 2019.

[18] L. Zhang, L. Shan, and J. Wang, "Optimal feature selection using distance-based discrete firefly algorithm with mutual information criterion," *Neural Comput. Appl.*, vol. 28, no. 9, pp. 2795–2808, 2017.

[19] R. Moazenzadeh, B. Mohammadi, S. Shamshirband, and K. W. Chau, "Coupling a firefly algorithm with support vector regression to predict evaporation in northern iran," *Eng. Appl. Comput. Fluid Mech.*, vol. 12, no. 1, pp. 584–597, 2018.

[20] S. Dash and B. Patra, "Feature selection algorithms for classification and clustering in bioinformatics," *Artif. Intell. Concepts, Methodol. Tools, Appl.*, vol. 3, no. January, pp. 2071–2091, 2016.

[21] A. A. Yahya, "Feature Selection for High Dimensional Data : An Evolutionary Filter Approach Feature Selection for High Dimensional Data : An Evolutionary Filter Approach," no. December, 2014.

[22] D. Compression, "Compressed k NN: K-Nearest Neighbors with Data Compression," pp. 1–20, 2019.

[23] S. Geiser and M. V. Santelices, "Validity of high-school grades in predicting student success beyond the freshman year: High school record vs. standardized tests as indicators of four-year college outcomes," *CSHE Res. Occas. Pap. Ser.*, p. 35, 2007.

[24] B. A. Friedman and R. G. Mandel, "Motivation predictors of college student academic performance and retention," *J. Coll. Student Retent. Res. Theory Pract.*, vol. 13, no. 1, pp. 1–15, 2011.

[25] F. J. da Costa, M. de S. Bispo, and R. de C. de F. Pereira, "Dropout and retention of undergraduate students in management: a study at a Brazilian Federal University," *RAUSP Manag. J.*, vol. 53, no. 1, pp. 74–85, 2018.

[26] E. M. Sosu and P. Pheunpha, "Trajectory of University Dropout: Investigating the Cumulative Effect of Academic Vulnerability and Proximity to Family Support," *Front. Educ.*, vol. 4, no. February, pp. 1–10, 2019.

[27] P. Jermyn, M. Dixon, and B. J. Read, "Preparing Clean Views of Data for Data Mining," pp. 1–15, 2014.

[28] C. Xenophontos, "A Beginner's Guide to MATLAB," 2012.