Implementation of Machine Learning algorithms to classify university academic success

Efrén Jiménez Delgado
Department of Software Enginner
Universidad Técnica Nacional
San Carlos, Costa Rica
Email: ejimenez@utn.ac.cr

Authors Linnette Roldán Morales
Department of Software Enginner
Universidad Técnica Nacional
San Carlos, Costa Rica
Email: lroldanm@utn.ac.cr

Yesenia Calvo Araya
Department of Software Enginner
Universidad Técnica Nacional
San Carlos, Costa Rica
Email: ycalvo@utn.ac.cr

Abstract — **Machine learning is a software tool that allows information present in databases to be converted into intelligent decisions based on data and not on experience or feelings. This work presents a review of the algorithms used in the classification of student performance present in the database of the national technical university, specifically in the software engineering career. Information classification analysis using ma- chine learning algorithm has become a modern and powerful instrument, which could help academic institutions improve retention and performance rate of courses by students and with this obtain a more general view of student performance before starting the course based on information from other courses and general student variables in order to reduce the risk of failure. The main objective of this work is to describe the variables used by the National Technical University and the application of different automated learning algorithms in order to obtain the metrics that allow demonstrating the best algorithm among those studied. Finally, the article concludes with a classification algorithm that provides an accuracy of around 80% to 82% that can be replicated in other university institutions.**

Keywords - Machine Learning, Classification, ANN, Data Analytics, Random Forest, SVM, Decision Tree, KNN.

I. INTRODUCTION

Predictive analytics is the subset of statistical analysis which includes the process of information extraction from existing data sets by using various techniques to find the trends and patterns and give prediction of future outcomes. It has wide applications in the field of manufacturing, supply chain management, banks, retail and marketing industries. Recently, increasing number of educational institutes are applying predictive analytics for taking crucial decisions regarding improving student's enrollment, student's retention, lowering the dropout rate, maintaining long term relationship with alumni and for knowing the overall of placements in advance.

The National Technical University (UTN) was created by Law No.8638 signed on June 4, 2008 by the President of Costa Rica Dr. Oscar Arias Sanchez. The UTN was born with the objective of solving the need for technical and innovative careers that the country's productive sector demanded.

It currently has approximately 12,000 students enrolled in its different levels of higher education and its offer is mostly taught at its regional headquarters, outside the central plateau of the country [1].

The Software Engineering Career has approximately one thousand students distributed in three levels, diploma and bachelor's degree.

Taking advantage of the benefits of predictive analysis, it is desired to improve the academic performance of students and avoid desertion by attacking external agents that are identified that in the past have not allowed a good development of the learning process in the student body.

II. MACHINE LEARNING ALGORITHMS APPLY

Yi Tan says that machine learning demonstrated great success in building models for pattern recognition in domains ranging from computer vision to speech recognition and text comprehension and intelligence in video games.

In addition to these classical domains, machine learning, and in particular deep learning, are becoming increasingly important and successful in engineering and the sciences [2].

The field of machine learning in the classification branch is an important concept introduced by information technology, which ensures deduction and classification meanings and meaningful information from groups of information towards a predetermined goal according to Halil Ibrahim Bulbul [3]. Machine learning algorithms can be classified into two main streams: supervised and unsupervised learners. The goal of supervised learning is to predict the correct output vector for a given input vector.

2022 17th Iberian Conference on Information Systems and Technologies (CISTI)
22 – 25 June 2022, Madrid, Spain
ISBN: 978-989-33-3436-2

In cases with one or more continuous variables in the target tag. It is difficult to define the objective of learning without supervision. One of the main goals is to identify sensitive groups in the data input, called clustering, of similar samples. The result of the ML algorithm can be significantly improved through this processing step and is called feature extraction according to V. Porkodi [4].

The performance of the students will be classified if the student will pass or fail the course with the information of other students with the courses that they have previously taken before starting the course to be classified by the algorithm. Among the machine learning classification algorithms used we can find:

A. *K-Nearest Neighbour (KNN) Classifiers*

The KNN classifier is a non-parametric instance-based classifier. It is a lazy learning method which does not learn from training data, simply stores all the samples in the training data. These stored values are needed during the training phase. This algorithm is based on the nearest neighborhood estimation. The new cases are classified on the basis of similarity measure which is the distance metric. Most commonly used is Euclidean distance. Drawback of KNN classifier is large time required to find nearest neighborhood in a large training set. Hence dimensionality reduction step is done to overcome this. In KNN classifier, the class of x is found by following procedure [5]–[8].

1) Determine the k instances which are nearest to the class x based on the distance measure.

2) The next step is to allow this k instances to vote to find the class of x.

Here the number of closest neighborhood instances selection is crucial step in this algorithm which affects the overall performance of classifier

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}$$

B. *Support vector machines (SVM) Classifier*

The SVM can work on data set that belongs to both the linear and the nonlinear type, it working initiates with the conversion of the data fixed for training into a data of higher dimension data set by the nonlinear mapping, followed by the exploration for the linear optimum that splits the hyperplane to segregate the tuples from the data set. Utilizing the support vectors the hyper plane is found by the SVM. Multitudes of separating lines are drawn to detect the optimal one with the minimized errors the fig.1 below shows one such hyper plane that partition the data into two classes: class1 and class2 [8]–[11].

$$f(x) = \sigma\left(\sum_i \alpha_i \Phi\left(s_i\right) \cdot \Phi(x)\right) \tag{2}$$

C. *Artificial Neural Network (ANN)*

Neural Network models are one of the best models for numeric prediction [12]–[15]. They help to capture all the non-linear relationship between the dependent and independent covariates when used for regression. The neural network finds an output o(x) for given and current weights. It calculates the function

$$(x) = f\left(w_0 + \sum_{i=1}^{n} w_l x_i\right) = f\left(w_0 + w^T x\right) \tag{3}$$

here w0 denotes the intercept and w = (w1… wn) is the vector consisting of all synaptic weights without the intercept and x = (x1… xn) are the vector of all covariates. As the main focus is on supervised learning, the error E is calculated and weights are adjusted using the learning algorithm. The error is given as

(4)

$$\mathrm{E} = \frac{1}{2} \sum_{I=0}^{L} \sum_{h=1}^{H} \left(o_{lh} - y_{lh}\right)^2$$

D. *Random Forest*

Formally, a random forest is a predictor consisting of a collection of randomized base regression trees {rn (x, Θm,Dn), m ≥ 1}, where Θ1,Θ2 . . . are i.i.d. outputs of a randomizing variable Θ. These random trees are combined to form the aggregated regression estimate where EΘ denotes expectation with respect to the random parameter, conditionally on X and the data set n. In the following, to lighten notation a little, we will omit the dependency of the estimates in the sample, and write for example rr¯n(X) instead of r¯n (X,Dn). Note that, in practice, the above expectation is evaluated by Monte Carlo, that is, by generating M (usually large) random trees [16]–[19].

$$\bar{r}_n\left(\mathbf{X}, \mathcal{D}_n\right) = \mathbb{E}_\Theta\left[r_n\left(\mathbf{X}, \Theta, \mathcal{D}_n\right)\right] \tag{5}$$

Decision Tree is a classification algorithm that decides whether a specific value should be accepted or rejected, and it provides with the set of the IF-Then rules for transforming present state to future state. The tree structure is used to represent decision tree in which variant types of the nodes are connected by the branches where the topmost node is called as root node and the leaves are called decision node. An intelligent prediction and recommendation system was

ISBN: 978-989-33-3436-2

built which would predict the student's performance using current and past data and measure reaction of students to the given recommendations. Decision tree was used as classification technique to decide whether specific value should be accepted or rejected

$$\text{setDn} = \{(\text{Xi}, \text{Yi})\}\text{i} = \text{ln} \tag{6}$$

III. DATA

The data was collected from the data center of the National Technical University, specifically from the Software Engineering department from the relational database of its student enrollment system. The data consisted of around 2000 records and 51 columns. These columns are of a qualitative type and where you can find the sex, economic status, status of previous courses and marital status during the previous courses and how many times you have enrolled in the previous courses. The data is compiled into the CSV file to find the correlation between these variables and then the data is normalized using the minmax algorithm.

IV. EVALUATION

To analyze the performance of the system, several supervised learning algorithms were proposed and metrics such as precision (7), recall (8) and accuracy (9) were used. The performance of the proposed system results in a binary classification in which it classifies whether or not a student will successfully complete a course. The precision (7), recall (8) and accuracy (9) formula. The algorithm that gave the worst result was KNN with 75% accuracy, a recall of 73% and an accuracy of 74%. The second worst algorithm is the Artificial Neural Network with a precision of 82%, a recall of 74% and an accuracy of 78%, in addition the third worst algorithm is SVM with a precision of 88% and a recall of 86 and an accuracy of 87%. The best model is the Random Forest with a precision of 95% and a recall of 91% and an accuracy of 93%, as shown in table I.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{7}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{8}$$

$$\text{Accuracy} = \frac{\text{No. of rows correctly classified}}{\text{Total no. of rows}} \tag{9}$$

TABLE I. METRICS

Algorithm	Performance Metrics	Value
K-Nearest Neighbour (KNN)	Precision Recall Accuracy	75% 73% 74%
Support vector machines (SVM)	Precision Recall Accuracy	88% 86% 87%
Artificial Neural Network (ANN)	Precision Recall Accuracy	82% 74% 78%
Random Forest	Precision Recall Accuracy	95% 91% 93%

V. CONCLUSION

This article presents different examples of predictive analytics application in higher education institutes. It highlights the use of different machine learning algorithms used for the benefit of students, teachers and administrators. The experiment also shows that there is a strong relationship between student performance and its various variables found as civil and economic conditions. They can predict an accuracy of 75%-95%, which is why an option is presented to apply to other universities, due to their good numbers in the academic success of students and with this, being able to know in time which students have problems. to finish a course in the best way. Analytics coupled with machine learning has been shown to be a beneficial tool to explore opportunities and innovate in some gaps in the higher education system.

REFERENCES

[1] R. J., "Construyendo una esperanza. esbozo histórico de la creación y ´ desarrollo de la universidad técnica nacional," in ´ Sección de Noticias de ´ la página web de la Universidad Técnica Nacional. 2018.

[2] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, M. Walczak, J. Pfrommer, A. Pick, R. Ramamurthy, J. Garcke, C. Bauckhage, and J. Schuecker, "Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems," IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, 2021.

[3] H. I. Bulbul and Unsal, "Comparison of classification techniques used in machine learning as applied on vocational guidance data," in 2011 10th International Conference on Machine Learning and Applications and Workshops, vol. 2, pp. 298–301, 2011.

[4] V. Porkodi, D. Yuvaraj, J. Khan, S. A. Karuppusamy, P. M. Goel, and M. Sivaram, "A survey on various machine learning models in iot applications," in 2020 International Conference on Computing and Information Technology (ICCIT-1441), pp. 1–4, 2020.

[5] D. Bajpai and L. He, "Evaluating knn performance on wesad dataset," in 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 60–62, 2020.

[6] C. C, "Prediction of heart disease using different knn classifier," in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1186–1194, 2021.

[7] M. Manjusha and R. Harikumar, "Performance analysis of knn classifier and k-means clustering for robust classification of epilepsy from eeg signals," in 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 2412–2416, 2016.

[8] Maheshwar and G. Kumar, "Breast cancer detection using decision tree, na¨ıve bayes, knn and svm classifiers: A comparative study," in 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 683–686, 2019.

ISBN: 978-989-33-3436-2

[9] L. Demidova and M. Egin, “Improving the accuracy of the svm classification using the parzen classifier,” in 2018 7th Mediterranean Conference on Embedded Computing (MECO), pp. 1–4, 2018.

[10] Y. Yang, J. Wang, and Y. Yang, “Improving svm classifier with prior knowledge in microcalcification detection1,” in 2012 19th IEEE International Conference on Image Processing, pp. 2837–2840, 2012.

[11] S. K. Sahu, A. K. Pujari, V. R. Kagita, V. Kumar, and V. Padmanabhan, “Gp-svm: Tree structured multiclass svm with greedy partitioning,” in 2015 International Conference on Information Technology (ICIT), pp. 142–147, 2015.

[12] S. Khaparde, P. Kale, and S. Agarwal, “Application of artificial neural network in protective relaying of transmission lines,” in Proceedings of the First International Forum on Applications of Neural Networks to Power Systems, pp. 122–125, 1991.

[13] M. Malik and R. Kamra, “A novel pv based ann optimized converter for off grids locomotives,” in 2021 International Conference on Technological Advancements and Innovations (ICTAI), pp. 299–302, 2021.

[14] L. Zhang, L. Jia, and W. Zhu, “Overview of traffic flow hybrid ann forecasting algorithm study,” in 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), vol. 1, pp. V1–615–V1–619, 2010.

[15] T. Onoda, “Next day’s peak load forecasting using an artificial neural network,” in [1993] Proceedings of the Second International Forum on Applications of Neural Networks to Power Systems, pp. 284–289, 1993.

[16] Q. Zhou, W. Lan, Y. Zhou, and G. Mo, “Effectiveness evaluation of anti-bird devices based on random forest algorithm,” in 2020 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS), pp. 743–748, 2020.

[17] Y. Guo, Y. Zhou, X. Hu, and W. Cheng, “Research on recommendation of insurance products based on random forest,” in 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), pp. 308–311, 2019.

[18] P. Mekha and N. Teeyasuksaet, “Image classification of rice leaf diseases using random forest algorithm,” in 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, pp. 165–169, 2021.

[19] X. Chen, X. Li, P. Chen, Y. Liu, S. Sun, and J. Liu, “Research on android application detection based on static permission and random forest,” in 2020 5th International Conference on Control, Robotics and Cybernetics (CRC), pp. 181–184, 2020.

ISBN: 978-989-33-3436-2