INTI JOURNAL | eISSN:2600-7920
Vol.2019:008

Predictive Modeling for Student Performance Data Using Decision Tree and Support Vector Machine

Maryam Khanian Najafabdi[1], Sarasvathi Nagalingham[1], Sayed Mojtaba Tabibian[1]

[1]Faculty of Information Technology, INTI International University,
Persiaran Perdana BBN, Putra Nilai,
71800 Nilai, Malaysia.

Email: Maryam.najafabadi@newinti.edu.my, saras.nagalingam@newinti.edu.my

Abstract

This paper is aimed to present a conceptual understanding that summarizes higher education analytics lifecycle. This paper explores the establishment of new architecture of technologies, experts, standards, and practices in the complex data infrastructure projects among higher education institutes. The research on higher education analytics converges with the demands from industry to improve the learning education systems by considering the teaching and learning analytics capabilities enhancing the efficiency of higher education. The exploitation of massive volume campus and learning information could be a crucial challenge for the planning of campus resources, personalized curricula and learning experiences. In the field of higher education, institutions look to a future of the unknown and vast speed advancement of technology. Moreover, with more strategic data solutions used in decision making with the over increasing social needs and political changes at national and global, competition within and among universities increase. Higher education needs to expand local and global impact, increase financial and operational efficiency, create the new funding models in a changing economic climate and respond to the greater accountability demands to ensure the success of organizational at all levels and stay on top of the ranks. Research on higher education institutes is also important because it enables maximum benefit and perceptions on students' performance and learning trajectories to be determined as these two are important in adapting and personalizing curriculum and assessment. The findings of this paper provided a view about modeling students' performance classification by Machine Learning models and to identify which of the predictors in the dataset contribute towards good prediction on the students' performance.

Keywords

Predictive Modeling, Higher education, Students' performance, Massive data

International Conference on Innovation and Technopreneurship 2019
Submission: 17 July 2019
Acceptance: 2 August 2019

Introduction

Throughout the last decade, massive amount of large data has been produced. This can be seen in many papers and reports of the number of videos, images, text, new posts such as Facebook, Twitter, Instagram uploaded every minute (Chen and Zhang, 2014). The emergence of massive data has raised several questions among researchers and scholars. These questions are:

a) How to propose an environment that can cater continuous growth of data?
b) How to prepare functionality of systems when there are major crisis and risks?
c) How to propose economical and effective solutions?

These questions are mooted because most of the existing tools for data storage, processing and analysis are insufficient for the massive amount of data generated. Hence, there should be more applied solutions for Big Data as social media and other forms of media are increasingly becoming more prominent among users (Aljohani et al., 2018; Siddiqa et al., 2016).

Big data has been given a wide range of definitions by scholars and researchers. One common idea that these scholars and researchers share about big data is that it has four major characteristics which are known as the Four Vs. These Four Vs are volume, variety, veracity and velocity. Volume refers to large size of data that needs powerful storage capacity, fast accessibility and processing power. Variety refers to the different forms of data like video, textual, signals, audio and images. Veracity indicates imprecise or uncertainty in the datasets that maybe the cause from biasness, noise or/and abnormality of data. Meanwhile, velocity denotes the speed in which data flows from mobile devices, social networks, and internet of things (IoT) to its user or destination (Inoubli et al., 2018; Chen and Zhang, 2014; Najafabadi, Mohamed, & Mahrin, 2017). In addition, there are computer-based organizations in their wake of Big Data solutions have given rise to another Three Vs that are variability, value and visualization. Variability refers to the point of differences between data in datasets which in statistic can be measured using range, mean, variance and standard deviation. Value is associated to the meaningfulness of data towards the required solutions. Finally, the way meaningful data are represented and told is pertinent and give significant impact to the direction and decision making of an organization in which emphasizes on the last V known as Visualization (Mohamed et al., 2019; Prajapati et al., 2017; Najafabadi et al., 2017; Chen and Yeh , 2017).

Not only data can be generated from social media and many other sources, data storing has long been practiced by institutes, companies and organizations in managing their operation. For example, higher education institutes have huge data of their students, programs and facilities. With the advent of big data, higher education institutes like universities and colleges have used digital data technology applications to manage the information on their stakeholders and assets and have turned themselves into “smarter institutes of education” (Najafabadi, Mohamed & Onn, 2019).

International Conference on Innovation and Technopreneurship 2019
Submission: 17 July 2019
Acceptance: 2 August 2019

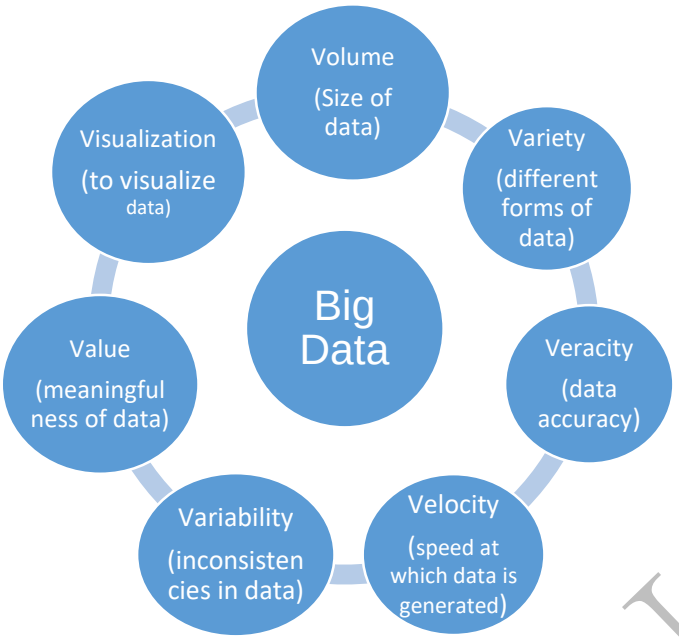

Figure 1. Big data features (Mohamed et al., 2019; Prajapati et al., 2017)

Big data has opened up more opportunities for higher education institutes to embark on outcome-based budgeting, evidence-based research work and explore new learning methodology that is apt. There have been many theories and methodologies that are developed in education and social science research works due to the surfacing of big data. Big Data analytics techniques involve a variety of disciplines, including data mining, social network analysis, statistics, machine learning, optimization methods, neural networks, pattern recognition, signal processing, and visualization techniques.Therefore, the drive behind this paper is to explore the establishment of new hidden architecture of technologies, standards, experts and practices in the complex data infrastructure projects among higher education institutes. This paper aims to demonstrate the capability of using machine learning models to predict student's final grade

Methodology

Extraordinary analytics techniques are required in Big Data to efficiently process huge number of data inside restricted run times. Data mining is the process of unfolding hidden knowledge and patterns and discovering relationships between the variables inside the data. It is a very powerful technique and is currently being used in numerous field including education – commonly known as Education Data Mining (EDM). EDM is mainly used for predicting student's academic performance which will then allow the academic institution to create a strategic program to boost the student's performance. It also allows for early intervention to take place should a student's performance is expected to drop. To analyse and compare the performance of two predictive models which are Decision Tree and Support Vector Machine (SVM) are used in predicting student's final grade. Decision tree resembles tree model analysis in making decision along with

International Conference on Innovation and Technopreneurship 2019
Submission: 17 July 2019
Acceptance: 2 August 2019

their possible outcomes in each branch. It as a popular model in decision analysis in predicting the outcomes of a scenario based on the sequential and hierarchical questions or variables of the model. Since the problem that this project tried to solve is a predictive modeling problem for target variable that is categorical, the type of support vector machine that is used is support vector classification (SVC).

The data used for this project was student performance data obtained from UCI machine learning repository (Paulo Cortez, University of Minho). The data consist of 2 datasets, which are Portuguese language (por) and Mathematics (mat). The dataset contains attributes such as student grades, demographic, social and school features for secondary level schools in Portugal. For 'mat' dataset there is a total of 396 cases (rows) and 33 variables (columns), meanwhile for 'por' dataset there is a total of 649 cases (rows) and 33 variables (columns).

Results and Discussion

Various researches have been carried out to identify the attributes and parameters that contributed in improvement in the academic qualities. The common evaluated factors are such as a student's personality, schools, families, and peers. In addition, involved demographic factors such as gender, age, family background and disabilities and external evaluation involving exam marks for specific subjects. There is also evidence that gender is one of the attributes that has influence on students' achievement where female students are commonly found to have a more positive attitude towards education as compared to male students.

The relationship between demographic and social variables indicated that these factors have tendency to affect differences in academic achievement or performance among students. In this research, our objective was to perform classification activities on students' performance dataset to understand how to model this case.

Based on what we analyzed, there are few factors that contribute the students' performance based on their grades. At first, we are doing Exploratory data analysis to find what is the highest correlation between the target variables. We just focus on two variables names Romantic and Address. Romantic refer to the status students that have kind of a relationship and students does not have a relationship. The results show significant correlation with 0.019. Besides that, for address its divided by urban and rural. The point is we want to know which one between urban and rural can give the highest effect on student performance. So, based on the analysis done in this research, it shows about 750 students get the better grades based on urban address meanwhile its only 280 students come from rural. This point of view we can conclude that urban students achieved a good grade. Based on the two table below, it shows model performance of Decision Tree and Support Vector Machine of 2 datasets which is Portuguese and Math. So, for Portuguese score dataset, the best selecting is Decision Tree with accuracy 0.7487179 and for datasets Math the best selecting is also Decision Tree with accuracy 0.6722689.

International Conference on Innovation and Technopreneurship 2019
Submission: 17 July 2019
Acceptance: 2 August 2019

Table 1. Performance for predictive model (Portugese dataset)

Datasets	**Performance**	**Decison Tree (por)**	**SVM (por)**
Training	Accuracy	0.6847826	0.7753303
Testing	Accuracy	0.6769230	0.6769230

Table 2. Performance for predictive model (Math dataset)

Datasets	**Performance**	**Decison Tree (por)**	**SVM (por)**
Training	Accuracy	0.7378854	0.8876811
Testing	Accuracy	0.7487179	0.6050420

Conclusions

This paper has examined the role of analytics in higher education. The significance of analytics and big data in higher education is to have a shift in the activities and improve teaching and learning processes. We have discussed the structure of higher education that has been aimed to clarify the benefits of analysis to learning and teaching within the educational setting. Learning analytics researchers have focused on employing its techniques toward enhancing the performance of students and university's behaviors. In conclusion, we can conclude that Data Mining can help in discovering relationships between the variables inside the data. We focus on 2 machine learning model names Decision Tree and Support Vector Machine in making classification and prediction. As discuss before for the classification task we divide by 6 Grade to classify the good grade students, fair and poor. Meanwhile for the prediction we use the performance of the 2 models above to determine which one is the best method. In last discussion it shows that Decision Tree models is the highest accuracy than Support Vector Machine.

Acknowledgements

The authors would like to thank faculty of information technology, INTI international university for their support and cooperation including researches and other individuals who are either directly or indirectly involved in this study.

References

Aljohani, N. R., Daud, A., Abbasi, R. A., Alowibdi, J. S., Basheri, M., & Aslam, M. A. (2018). An integrated framework for course adapted student learning analytics dashboard. Computers in Human Behavior.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314-347.

Submission: 17 July 2019
Acceptance: 2 August 2019

Chen, S. Y., & Yeh, C. (2017). The effects of cognitive styles on the use of hints in academic English: A learning analytics approach. Journal of Educational Technology & Society, 20(2), 251e264.

Inoubli, W., Aridhi, S., Mezni, H., Maddouri, M., & Nguifo, E. M. (2018). An experimental survey on big data frameworks. Future Generation Computer Systems.

Mohamed, A., Najafabadi, M. K., Wah, Y. B., Zaman, E. A. K., & Maskat, R. (2019). The state of the art and taxonomy of big data analytics: view from new big data framework. Artificial Intelligence Review, 1-49.

Najafabadi, M. K., Mohamed, A. H., & Mahrin, M. N. R. (2019). A survey on data mining techniques in recommender systems. Soft Computing, 23(2), 627-654.

Najafabadi, M. K., Mohamed, A., & Onn, C. W. (2019). An impact of time and item influencer in collaborative filtering recommendations using graph-based model. Information Processing & Management, 56(3), 526-540.

Prajapati, D. J., Garg, S., & Chauhan, N. C. (2017). MapReduce Based Multilevel Consistent and Inconsistent Association Rule Detection from Big Data Using Interestingness Measures. Big Data Research, 9, 18-27.

Siddiqa, A., Hashem, I. A. T., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., & Nasaruddin, F. (2016). A survey of big data management: Taxonomy and state-of-the-art. Journal of Network and Computer Applications, 71, 151-166.

International Conference on Innovation and Technopreneurship 2019
Submission: 17 July 2019
Acceptance: 2 August 2019