International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Predicting Academic Performance of Tertiary Students using Classification Algorithm

Sujith Jayaprakash, Jaiganesh V

Abstract: ***Classification algorithms have paved the way for several recommender systems in the field of Medicine, Entertainment, Politics, Education, etc. Recently there is a growing interest among researchers to analyze or predict the academic progression of students from High Schools to Tertiary Education. Better performance of students will directly reciprocate in the growth of an institution. Hence, setting up a supervised learning system will act as a gauge to provide a benchmark education. This paper aims to recommend a system based on a predictive model which will aid the institution to measure the performance of students based on various parameters.***

Index Terms: ***Academic Progression of Students; Classification algorithm, Machine learning, Naïve Bayes Algorithm, Recommender System, Supervised Learning***

I. INTRODUCTION

Increasing interest in Data Science and Machine learning has made a revolutionary change in the business in terms of providing the flexibility in classifying the data and understanding the micro-level information of any real-world business. Information gathered are classified and processed to forecast or predict the futuristic growth of the organization. Many financial institutions and stock market brokers are embracing the support of various software packages to predict or forecast. Applications like Predict5!, DryRun, SalesForce and Veeva have provided the best possible solutions for businesses with the use of Machine Learning Algorithms. Amongst several machine learning algorithms, Classification algorithms have paved the way for several recommender systems in the field of Medicine, Entertainment, Politics, Education, etc. Nowadays, Machine learning algorithms are widely used in

a) Predicting the patient's disease based on the medical results
b) Predicting the success rate of a movie based on social media comments
c) Predicting the polling results based on the sample surveys
d) Predicting the attrition rate based on student feedback

Revised Manuscript Received on 30 July 2019.
* Correspondence Author
Sujith Jayaprakash, Research Scholar, Department of PG & Research Dr. N.G.P Arts & Science College, Coimbatore, India
Dr. Jaiganesh, V., Assistant Professor, Department of PG & Research Dr. N.G.P Arts & Science College, Coimbatore, India

Several researchers have proven that classification algorithms are comparably the best algorithms to develop a Recommender system. Recently there is a growing interest among researchers to analyze or predict the academic progression of students from High Schools level to Tertiary Education. Though there is a pressing need for such predicting systems, Institutions across the globe hasn't shown interest in implementing or maintaining it. From various researches, it is obvious that few universities from developed countries have taken their paramount step towards implementing these systems by foreseeing its dire need to maintain quality. IBM has come up with a cloud-based solution termed IBM Watson Analytics which will provide smart data analytics and visualization service that can be quickly used to discover patterns [1]. Similarly, Dublin City University developed a new system which predicts the performance of a student in a particular module using his academic data. It also compares the student's pattern in Moodle and an email has been sent on a weekly basis with suggestions on improvement in the modules they lack [2]. Oracle' People soft campus solutions 9.0 application provides various modules for tertiary education to maintain 360-degree information of a student. It also provides an analysis of student's performance through Academic Progress Tracker. This campus solution does not predict the performance of a student [3]. ProgressIQ is another student academic tracking system where institutions can get details about the student academic progress across the curriculum, learning outcomes, performance data etc., [4]. Although there are several third party software and solutions to monitor and keep track of the student's performance there is no actuality of a proper recommender system which is widely used in institutions. Moreover, these aforementioned applications largely keep the repository of student's current data and act as an analytical or statistical tool rather than a recommender. Hence, a system has to be built in to analyze the student performance based on current results and previous results along with the socio-economic data and psychological data to predict the results. Results from this robust system will provide timely feedback to improvise student's performance. In this research paper, we are trying to predict the grades of Semester 2 students by analyzing their previous grades and socioeconomic status. Naïve Bayes algorithm has been used to analyze and predict the data using Weka Tool.

II. LITERATURE REVIEW

The rationale behind this research is to identify the parameters that influence a student's progression and provide timely feedback to overcome the pitfalls.

The literature review reveals that quite a lot of researchers have shown interest on
this area and recommended solutions. Mueen et al proposed a system that can assist teachers in the early detect student who is expected to fail in a course. They used Data Mining techniques to predict and analyze a student's academic performance. Three techniques Decision tree (C4.5), Multilayer Perceptron, and Naïve Bayes were used. All these techniques were applied to student's data collected from undergraduate courses conducted in the duration of two semesters. In this study, three classification models were built to predict student academic performance and the result shows that Naïve Bayes classifier outperformed the other two classifiers by obtaining the overall prediction accuracy of 86% [5]. Kotsiantis et al proposed a prototype web-based support tool, which is developed using Naïve Bayes Algorithm that can automatically recognize students with a high probability of dropout [6]. Edin and Mirza used three supervised mining algorithm on the preoperative assessment data to predict success in a course (pass or fail) and the performance of the learning methods was evaluated based on their predictive accuracy, ease of learning and user-friendly characteristics. The results indicated that the Naïve Bayes classifier outperformed in the prediction decision tree and neural network methods [7]. Ashwin and Mariusz made a comparison between single model-based techniques and ensemble models, from their research it has been found that ensemble models not only gives better predictive accuracies on student performance but also provides better rules for understanding the factors that influence better student outcomes [8]. Amjad used multiple data mining tasks were used to create qualitative predictive models which were efficiently and effectively able to predict the students' grades from a collected training dataset. He used four decision tree algorithms that have been implemented, as well as, with the Naïve Bayes algorithm. Based on his research, he proposed that the student's performance is not totally dependent on their academic efforts, in spite, there are many other factors that have equal to greater influences as well [9]. Dorina proposed a data mining models for predicting student performance, based on their personal, pre-university and university-performance characteristics. Several algorithms are used to predict the performance of students [10].

III. RESEARCH APPROACH AND DATA SELECTION

This is an exploratory research wherein no model has been taken as a basis of the study. However, Cross-Industry Standard Process for Data Mining model has been used to tackle problems. CRISP-DM is one of the leading methodologies used by data miners to respond to the survey [11]. Using CRISP-DM, this research work has been divided into six different phases namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment

Fig. 1: CRISP-DM Model [12]

Implementation of recommender systems not only helps the students to improvise his grades but also improves the credibility of an institution and makes it a preferred choice of destination for young aspirants and this in-turn ratifies the initial phase of the model. The stated business problem is transformed into a data mining task which classifies the students final GPA based on their grades. The second phase is to understand the data involved and preprocess the data to remove unsolicited attributes. In this research work, we have collected the data of 150 students who are pursuing B.Com course. Data included several attributes like Registration Number, Student Name, Gender, Father Name, Mother Name, Occupation, Annual Income, School name, Locality, Junior and Senior High School grades and Semester grades. Data was collected from various sources like Student Admission Records and Academic Records. Students Junior/Senior High school grades, family details are collected from the admission records and the marks for various subjects of Semester 1 is taken from the Academic Records.

IV. DATA PREPROCESSING

Raw data collected from various sources are profoundly analyzed. Attributes like Registration No. Student Name, Father Name, Mother Name, Occupation and Gender are removed.

a) Junior/Senior High school names are classified into Private and Public Schools.
b) Students are classified into Low Class, Middle Class and Highly Class based on the Family income. Income less than 100k is considered as Low Class and income ranges from 100k to 200k is considered as Middle Class whilst the rest are tagged as High Class.
c) Based on the location/area of stay, we calculated the distance travelled by a student from his place of living to College. For calculating the distance Google Maps have been used. Location of the College and the student is entered in an excel sheet and passed as a source and destination parameter to Google Map link which has fetched the distance travelled in Kilometers. The formula used to calculate the distance is given below,

=CONCATENATE("http://maps.google.co.uk/maps?f=d&source=s_d&saddr=",SUBSTITUTE(A1,""","+"),"&daddr=",SUBSTITUTE(B1," ","+"))

- The values are then classified as nearby (<10km), Moderate (>10km and <20km) and Far (>20KM).
- Junior/Senior High Schools and Semester 1 grades are classified based on the below parameters,

Table 1: Classification of Grades

Grades	Classification
>=	Distinction
>= 65 and <75	First Class
>=55 and < 65	Second Class
>=40 and < 55	Third Class
<40	Fail

After the attributes are classified, the final dataset is prepared as shown below,

Table 2: Attribute Selection

Type of Data	Attribute Name	Attribute Type	Values
Personal Data	Family Income	Nom	Three distinct values (Low Class, Middle Class, High Class)
Pre-University Data	School Type	Nom	Private, Public
	Junior High School Grade	Nom	Distinction, First Class, Second Class and Third Class
	Senior High School Grade	Nom	Distinction, First Class, Second Class and Third Class
University Data	Distance Travelled	Nom	Far, Moderate and Nearby
	Semester 1 Marks	Nom	Distinction, First Class, Second Class and Third Class.

V. RESULTS AND DISCUSSION

In this research work, we have used the WEKA suite to classify the data and implemented the Naïve Bayes Algorithm. Waikato Environment for Knowledge Analysis (WEKA) software is used to implement machine learning algorithms in a dataset. WEKA has a collection of machine learning algorithms that can be used on a dataset. By applying this dataset in WEKA, it classifies them based on the predicted variable. Naïve Bayes Algorithm is used to develop the prediction mechanism in the proposed Recommender System. Naïve Bayes is a probabilistic classifier based on the Bayesian Algorithm. The reason behind using this algorithm is that it is scalable and maximum likelihood training can be done which takes linear time.

$$p(C_j \mid x_1, x_2, \ldots, x_d) \propto p(x_1, x_2, \ldots, x_d \mid C_j) p(C_j)$$

The first step in using Bayesian Algorithm is to develop a frequency table based on the proposed dataset and find the probability of each attribute based on the predictor variable. In total 150 instances with 7 attributes are chosen to build a training model. To estimate the accuracy of the chosen predictive model, cross-validation is used while building the training model. Test model used in the building is 10-fold cross-validation where a single subsample is retained as the validation data for the training model and remaining are used as training data. Frequency table generated during the training model building is given below,

Table 3.1: Frequency Table of the independent attributes

Frequency Table		Distinction(9)		First Class(21)		Second Class(19)		Third Class(7)	
Semester 1	Distinction	2	0.2	0	0	0	0	0	0
	First Class	6	0.7	13	1	2	0.1	1	0
	Second Class	0	0	8	0	9	0.5	0	0
	Third Class	1	0.1	0	0	8	0.4	6	1

Frequency Table		Distinction(9)		First Class(21)		Second Class(19)		Third Class(7)	
+2	Distinction	8	0.9	20	1	12	0.6	3	0
	First Class	1	0.1	1	0	4	0.2	1	0
	Second Class	0	0	0	0	1	0.1	2	0
	Third Class	0	0	0	0	2	0.1	1	0

Frequency Table		Distinction(9)		First Class(21)		Second Class(19)		Third Class(7)	
10	Distinction	8	0.9	21	1	14	0.7	6	1
	First Class	1	0.1	0	0	1	0.1	0	0
	Second Class	0	0	0	0	3	0.2	1	0
	Third Class	0	0	0	0	1	0.1	0	0

Frequency Table		Distinction(9)		First Class(21)		Second Class(19)		Third Class(7)	
Kilometer	Nearby	9	0.2	18	9	15	3.8	6	6
	Moderate	0	0	1	1	2	0.5	1	1
	Far	0	0	2	1	2	0.5	0	0

Frequency Table		Distinction(9)		First Class(21)		Second Class(19)		Third Class(7)	
Type of School	Private	9	0.2	21	11	17	4.3	5	5
	Public	0	0	0	0	2	0.5	2	2

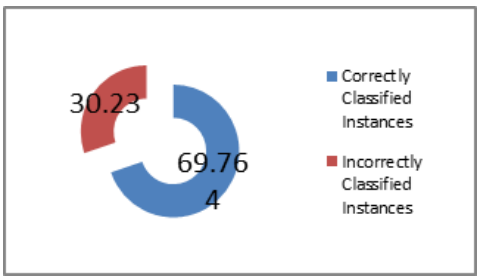
Frequency Table		Distinction(9)		First Class(21)		Second Class(19)		Third Class(7)	
Family Status	High Class	0	0	4	2	0	0	0	0
	Middle Class	4	0.1	4	2	2	0.5	2	2
	Low Class	5	0.1	13	7	17	4.3	5	5

Table 3.2: Observation from the Frequency Table

1. *Students who came from public schools are less likely to get Distinction or First Class compared to those coming from private schools.*
2. *Students who are travelling from the long-distance are less likely to get a distinction or first-class comparing to those travelling a moderate or short distance.*
3. *Also, from the 10th, +2 and Semester 1 frequency table it's obvious that students who secured Distinction or First class maintain the same designation in the higher semesters whilst students who have scored Second class lower or third class in their high school and semester 1 is less likely to get a distinction or first class.*

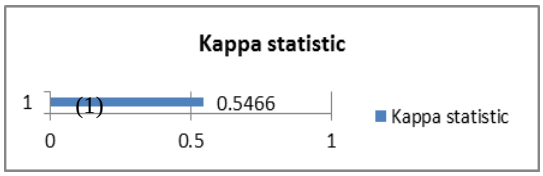
Evaluation on the training set is given below, We split the data into two sets wherein two-third of the data has been used to prepare a training model and the rest one third is used for the evaluation. After the training dataset is developed, the Bayesian algorithm is applied to classify the data and remove the redundancies. 69.7% of the data are correctly classified and 30.23 are incorrectly classified instances.

Fig 2: Correctly Vs. Incorrectly classified instances.



The test dataset is compared with the train dataset and prediction is made. Kappa Statistic result shows a prediction accuracy of 0.54 which provides a better accuracy rate when comparing both the datasets.

Figure 3: Kappa Statistic



VI. CONCLUSION

The objective of this research work is to develop a system that can act as a recommender for students as well as institution to predict the performance based on the grades. Historic data of a student helps us to develop a pattern which can be used for the prediction. The proposed system uses a Naïve Bayes algorithm which is robust and reliable. This algorithm is applied to the dataset which has a student's social status, travel distance, CGPA of junior/senior school grades and semester 1. Based on these attributes, a student's semester 2 CGPA is predicted. CGPA is classified as Distinction, First Class, Second Class and Third Class, the proposed system will collect the data based on the attributes and the collected data will be cross-checked from the trained dataset in the database to predict the CGPA. By using the Naïve Bayes Algorithm, the system produces good accuracy, however in the future research works we intend to use other classification algorithms that can yield better results compared to the prevailing system. Hence, from the research work carried out with the nominated attributes, it's evident that a recommender system can be built which will aid the students as well as institutions to monitor the performance. Few attributes which can contribute better results to this research are gender, social behaviour, classroom participation, frequency of accessing resources like Moodle, email and e-library of a student.

REFERENCES

1. IBM (2017, November 06). Transform learning experiences with Watson. Retrieved from https://www.ibm.com/watson/education
2. Erinc, M (2015, August 17). Software that helps to keep first-year students on track. Retrieved from https://www.irishtimes.com/news/education/software-that-helps-to-keep-first-year-students-on-track-1.2316190
3. Oracle (2017, November 06). PeopleSoft Campus Solutions. Retrieved from http://www.oracle.com/us/products/applications/peoplesoft-enterprise/campus-solutions/overview/index.html
4. ProgressIQ (2017, November 06).Why Choose ProgressIQ. Retrieved from https://www.progressiq.com/.
5. Mueen, A., Zafar, B and Manzoor, U. (2016). Modelling and Predicting Students' Academic Performance Using Data Mining Techniques. I.J. Modern Education and Computer Science, 36-42
6. Kotsiantis, S. B., Pierrakeas, C. J., Pintelas, P.E. (2003). Preventing Student Dropout in Distance Learning Using Machine Learning Techniques. Springer-Verlag Berlin Heidelberg, 267-274
7. Edin, O., Mirza, S. (2012). Data Mining Approach for Predicting Student Performance. Economic Review – Journal of Economics and Business, Vol. X, Issue 1, 3-12
8. Ashwin, S., Mariusz, N. (2016, April 8-9). Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance. Spring 2016 Mid Atlantic ASEE Conference.
9. Amjad, A. S., (2016). Educational Data Mining & Students' Performance Prediction. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 212-220
10. Dorina, K. (2012) Student Performance Prediction by Using Data Mining Classification Algorithms. International Journal of Computer Science and Management Research, Vol. 1, Issue 4, 686-690
11. Wirth, R & Hipp, Jochen. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.
12. The cross-industry standard process for data mining. (2019). Retrieved 18 February 2019, from https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

AUTHORS PROFILE

Sujith Jayaprakash is a research scholar at Dr N. G. P College of Arts and Science. His area of research is in machine learning algorithms, the academic progression of students, web mining, Use of education apps etc. He has over a decade of experience in Education Administration and academia.

Dr. Jaiganesh V. is currently working as a Professor at Dr N. G. P College of Arts and Science. His area of specialization includes Data mining and Machine learning. He has 19 years of teaching experience and guided several research scholars in the field of Data mining.