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

- Predicting the patient’s disease based on the medical results
- Predicting the success rate of a movie based on social media comments
- Predicting the polling results based on the sample surveys
- Predicting the attrition rate based on student feedback

Several researchers have proven that classification algorithms are comparatively 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].

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...
Predicting Academic Performance of Tertiary Students using Classification Algorithm

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.
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**

<table>
<thead>
<tr>
<th>Grades</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;=</td>
<td>Distinction</td>
</tr>
<tr>
<td>&gt;= 65 and &lt;75</td>
<td>First Class</td>
</tr>
<tr>
<td>&gt;=55 and &lt; 65</td>
<td>Second Class</td>
</tr>
<tr>
<td>&gt;=40 and &lt; 55</td>
<td>Third Class</td>
</tr>
<tr>
<td>&lt;40</td>
<td>Fail</td>
</tr>
</tbody>
</table>

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.

\[
P(C_j | x_1, x_2, \ldots, x_d) = \frac{P(x_1, x_2, \ldots, x_d | C_j)P(C_j)}{
\sum_{j=1}^{n} P(x_1, x_2, \ldots, x_d | C_j)P(C_j)}
\]

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**

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Attribute Name</th>
<th>Attribute Type</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal Data</td>
<td>Family Income</td>
<td>Nom</td>
<td>Three values (Low Class, Middle Class, High Class)</td>
</tr>
<tr>
<td>Pre-Universit y Data</td>
<td>School Type</td>
<td>Nom</td>
<td>Private, Public</td>
</tr>
<tr>
<td>Senior High School Grade</td>
<td>Nom</td>
<td>Distinction, First Class, Second Class and Third Class</td>
<td></td>
</tr>
<tr>
<td>University Data</td>
<td>Distance Travelled</td>
<td>Nom</td>
<td>Far, Moderate and Nearby</td>
</tr>
<tr>
<td>Semester 1 Marks</td>
<td>Nom</td>
<td>Distinction, First Class, Second Class and Third Class</td>
<td></td>
</tr>
</tbody>
</table>

**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 correctly 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.

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


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.