Educational Data Mining

Submitted against receiving

IBM Professional Data Science Certificate

As

Final Capstone Project Report

Author Note

Vedant Bahel

Abstract

This report highlights the work that can be carried out in Education Domain. The objective is to advance and improve the current Educational System. Data Science along with other intelligent analysis and mining tools can be used to solve the purpose. This report highlights some of such domains where the concept of Data Science can be applied to improve the educational system.

Keywords: Educational Data Mining, Data Science and Machine Learning.

Introduction

Education is a foremost requirement for everyone around the world. There are many problems that Educational system in India faces. The concept of EDM is not new but it has not been popular in Asian countries, specially India. This paper focuses on EDM tools and techniques for institutions specially in the low economic zone and underdeveloped countries which cannot spent more fund on sophisticated software, so this approach shall give them general inferences to improve the quality of education.

The grade/performance prediction modules already have been proposed and implemented by some of the researchers [15]. But these existing concepts fails when we consider Asian Scenarios (specially India). The learning behaviors of students in India differs a lot from rest of the world. In India, the educational system is very diverse.

Today, there are numerous software in the market which can be used for the study of Educational data. But more work has been in the fields of web-based learning and distant education [13]. Due to systematic conduct of online education platform, it becomes easy to implement EDM on web-based learning platforms, basically online platforms [14]. But when on-campus education system

is considered, the situation becomes different. Even for on-campus scenarios, EDM is currently more focused on Data analysis and other intelligent mining tools are being skipped off or ignored.

In western countries, they have uniformity in the educational system throughout the country. Whereas in India, each university has numerous colleges under them where the functioning of each college differs a lot. Example: some institutes follow grade-based system, some follow credit-based system, etc. The educational system in India is in developing stages where they are undergoing many transition's due to frequent changes in policies. Moreover, due to relatively poor economic conditions, one's socio-economic status is likely to affect the educational practices in India [16].

There shall be some method that can tackle such problems. One of the major problems is unemployability [18]. During the past decade, there has been an exponential rise in the use of web-based learning platform for both students and teachers. With recent trend having more focus on intelligent web-based learner platforms that not only provides learning resources but also guides students on what to learn as per their interest and calibre, educational institutions should adopt such tools and methodologies with an aim to advance and improve their educational system [1]

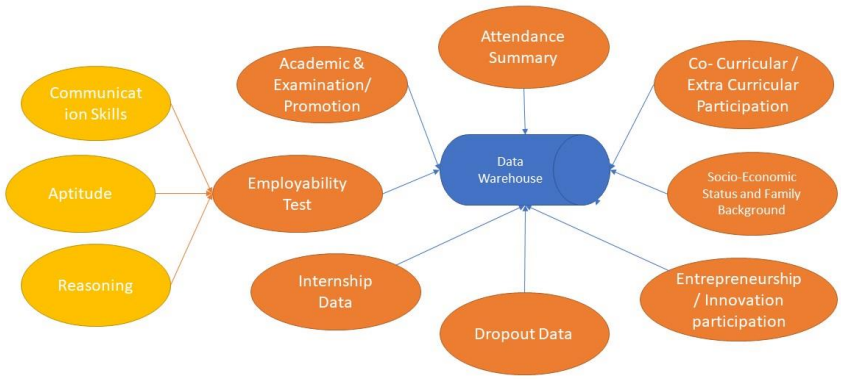
With jobs becoming lesser, the employers have become more particular about skills requirements, while job seeker have become more specific about their domains, as well. With this approach, it is of paramount importance that students be conscious about their interest, hidden skills if any and shall be provided with assistance in choosing relevant subjects in Choice based system otherwise it becomes challenging at both the ends – mentor who become clueless to guide to the student without understanding his interest and the student who is totally bewildered about choice of the subject and his own performance.

One of the ways to solve such problems is to study the vast database that institutes have in this

context. These databases include student's attendance, academic and cocurricular/extracurricular performance, course information, curriculum goals, personal interests, performance in assessment examinations conducted by various agencies, his/her performance in the examination prior to program undertaken – specially in state or common assessment/ board exams. Data analysis and data mining techniques can greatly assist institutes in drawing trends and inferences from educational data. The use of data mining in the educational system is also referred to as Educational Data Mining (EDM). This paper discusses and proposes concept related to EDM especially in Indian context so as to increase the efficiency and productivity of students in higher education. Also, it can be used to bring forward various insights that can conclude about the working and improvement methodologies for management, course developers and academicians of the institutes.

Methodology: Sub-Domain of Application

The combination of the concepts of statistical thinking, visualizations, mathematical modelling, clustering, pattern mining, classification and intelligent algorithms constitutes as the elements of EDM.



The concepts of EDM can be applied on various sub-domains of educational systems. This paper presents the application of EDM on following areas:

- Early drop out prediction
- Future Performance Prediction
- Early measures towards Performance Improvement based on prediction
- Domain Modelling
- Student Feedback based Inferences & Changes in Policies
- Graduate Admission Prediction & Course Suggestion Module

A. Early Drop Out Prediction

With Increasing number of students opting for higher education, there has also been increase in dropout rate due to various reasons. The reason may be due to poor performance, socio-economic status, related to family background, etc. Lack of interest, lower score and such related features can be studied through data warehouse. A university wouldn't prefer students who are likely to drop-out. A classification model can be developed to predict which student is likely to drop out depending upon the student's pre-admission data. The preadmission data like SSC Marks, HSC Marks, Working Status of Parents, Socio-economic status, etc from the data warehouse can be used to predict the required result. If the model predicts the student likely to dropout, it can assist Universities to find ways to avoid such situation to happen. This will ensure potential intake of candidates for that specific educational program. [11]. In [22], author talks about on track drop-out prediction system. Sometime a student's performance may deteriorate after admissions or during the academic program. If at that time the model predicts about a student likely to dropout than preventive measures can be taken to avoid the predicted situation to take place.

B. Future Performance Prediction

It is observed that the learning habits possessed by student mostly remains same throughout his studies. Depending upon the previous academic record of that student his/her future performance can be predicted [19]. A Machine Learning model can be trained on the data of past students from an institute to map relation between the input features and the output performance. There may be numerous input features including the background of the student, father's occupation, mother's occupation, etc. which are of equal significance for determining the efficiency of the model. It is very important to discover which input parameters are important or relevant to the study. Choosing wrong input parameters may result in underfitting/overfitting of the model. Hypothesis testing was used to evaluate the dependency of input parameters for calculating the grade of the student [10]. A Linear Regression model trained on past data of similar students can predict the student's current semester's Cumulative Grade Point Average (CGPA). Also, a classification model can be used to group students depending on the type of students may be into classes namely: Performer and Underperformers [20].

C. Early Measures towards Performance Improvement based on Predictions

On the basis of quarterly performance of a specific student his/her scope of improvement can be studied. If the performance of specific student in a subject is less than the class average that can infer the need of improvement of the student in that specific subject [21]. Also, visual data analytics can be used to determine the reason why a group of students are performing better than the other

students in the class. The learning habits of such students can be considered as ideal learning habits that can be implemented by others in order to improve their own performance. A suggestion module can help students improve in fields where they are showing poor performance. The model will itself judge which subjects are to be kept on focus line and as per that it will automatically suggest the student to necessarily attend those specific classes, take more assignments or video lecture in specific field. GHRCE conducts Remedial classes for weak students. The model will also suggest weaker students to attend remedial classes and warn them about detention probability based on their attendance data.

D. *Domain Modelling*

In higher education the involvement of students in extracurricular and co-curricular activities also matters a lot to determine their overall success rate. There are various extra-curricular/ co-curricular domains in a particular institute example: Drama, Arts, Various Clubs, Technical societies, cultural societies, etc. And it becomes difficult for a student to choose which one to join or which one will be beneficial for him/her. An intelligent clustering model can help students predict the club or domain in which his/her performance may prove to be remarkable. Each of the target classes requires different set of skill sets. The skill set of a particular student can be judges from the data warehouse. If a particular student has a good technical knowledge and has research orientation than he/she can be suggested to take research as his/her part time work, students with sports history can be suggested to go for that sport as his/her extra-curricular activity. This concept can also help institute to mine talent from all the students without manual evaluation of the

databases.

E. Feedback based Inferences & Changes in Policies

In educational institutes and other teaching environments, educators and the management easily obtain student's feedback on the learning system. This enables management to keep the teacher's evaluation process on track [2]. There are different types of students depending upon their learning practices. They collectively may or may not be able to understand what a specific teacher is teaching but may be able to understand from some other teacher. This can be easily bought into the picture from the annual feedback data given by students on the course and learning experience throughout the year. Also, the overall performance of the class when compared from some other class with similar quality of student may also reflect the same result. This study can be used by management to allot a specific teacher for specific course to a specific group of students to ensure optimal performance from the system.

There are some requirements of students that institute is unable to recognise. These requirements may be of utmost importance. Such a requirement of students can be understood from the analytics of data. For Example: Let's assume an institute that arranges a class for an aptitude for all students. And conducts weekly tests on same. The test data can be used to analyse the trend in performance. Even after attending the classes there isn't much improvement in the general performance then that concludes that there are some requirements to be fulfilled in order to solve such problems. The demand for books in the library or the resources can also be understood using the same concept. The lack of lab resources can also be understood.

All educational institutes have some or the other policies to maintain smooth functioning of

educational system. For example, institute uses some or the other educational software to enhance the productivity of students and improve their performance in certain areas. Sometimes such software is not found useful by the students but are still functional in the institute unnecessarily increasing the complexity of the system. With the use of certain tools, algorithms and data visualization techniques, such software can be bought in the focus of management so as to rule out them or take decision either removing the software or thinking out any other solution for the same.

F. Graduate Admission Prediction & Course Suggestion Module

After Under Graduation, more than 50% of students plans to go for higher education. A course suggestion module can be created that can suggest student which degree course to opt after UG. Example: A person having good management skills will be suggested to take MBA, a person having good technical skills can be suggested to take M. Tech with specific domain, a person interested in research can go for PhD, etc. The parameters for this test will be the performance of student in UG including the employability tests and other tests that student took. GHRCE conducts Aspiring minds, Co-Cubes, Cambridge BEC tests, etc for their students. The score of these tests if taken as input parameters can significantly improve the model. A machine learning model can be designed which can study various parameters like entrance exam score, previous transcripts, Letter of Recommendations (LOR), Statement of Purpose (SOP), etc to calculate the chance of admit or can give a prediction about the university. The SOP and LOR originally be in textual format. They need to be scored so as to feed as parameter for the intelligent model. This scoring can be done using text analysis. Authors also propose this text analysis model to score the SOP & LOR. This model can significantly help students judge their own profile and improve at mock level.

```
[5] #Data Understanding
    df.describe()
```

	IDNumber	10th	10B	12th	12B	Sgpa1	Sgpa2	Sgpa3	Sgpa4
count	238.000000	238.000000	238.000000	238.000000	238.000000	238.000000	238.000000	238.000000	238.000000
mean	12516.533613	77.725134	1.369748	70.679750	1.344538	7.357227	7.855126	7.670966	7.485000
std	1359.381023	9.695173	0.621222	9.004876	0.767889	1.094575	1.177280	1.266680	1.615555
min	1184.000000	44.000000	1.000000	7.900000	1.000000	4.500000	4.450000	4.000000	0.890000
25%	12283.500000	71.017500	1.000000	64.800000	1.000000	6.570000	7.050000	6.720000	6.680000
50%	12637.500000	79.335000	1.000000	70.076923	1.000000	7.330000	7.910000	7.830000	7.835000
75%	13076.000000	84.180000	2.000000	75.200000	1.000000	8.100000	8.820000	8.692500	8.595000
max	13393.000000	96.000000	4.000000	94.000000	4.000000	9.750000	9.820000	10.000000	10.000000

```
[6] df.info()

    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 238 entries, 0 to 237
    Data columns (total 13 columns):
     #   Column    Non-Null Count  Dtype
```


```
[6] df.info()
```


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 238 entries, 0 to 237
Data columns (total 13 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   IDNumber  238 non-null    int64
 1   10th      238 non-null    float64
 2   10B       238 non-null    int64
 3   12th      238 non-null    float64
 4   12B       238 non-null    int64
 5   Sgpa1     238 non-null    float64
 6   Sgpa2     238 non-null    float64
 7   Sgpa3     238 non-null    float64
 8   Sgpa4     238 non-null    float64
 9   Sgpa5     238 non-null    float64
 10  Sgpa6     238 non-null    float64
 11  Sgpa7     238 non-null    float64
 12  Sgpa8     238 non-null    float64
dtypes: float64(10), int64(3)
memory usage: 24.3 KB
```

```
[8] #data visualization
    ax = sns.boxplot(x=df["Sgpa1"])
```

Result

Data mining and KDD techniques help to extract a lot of valuable information from raw data, offering numerous opportunities in Indian educational domain. These techniques can be applied to improve the quality of educational system. Educational Data Mining needs to consider the all

aspects of learner in order to put forward logical and efficient inferences. Data mining tools can easily bring insights from educational data. Standardisation and pre-processing are very important techniques to be carried out when EDM is considered. Through this paper, the authors discussed the ways in which EDM can be applied over various areas in education.