ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Data Management of an Educational Institute

[1]Kunal Dulani, [2]Mohit Chanchlani, [3]Meeta Chanchlani, [4]Manoj Ahuja, [5]Richard Joseph.

[1,2,3,4]B.E, Dept. of Computer Engg, VES Institute of Technology, Mumbai, India

[5]Professor (Mentor), VES Institute of Technology, Mumbai, India

ABSTRACT: Data governance is a control that ensures that the data entry by an operations team member or by an automated process meets precise standards, much as a Business rule, a data definition and data integrity constraints in the data model. For this, the document level faculty score is calculated, along with student performance. All score calculations are done in accordance to the criteria specified by National Board of Accreditation (NBA). This project aims at providing a better insight to the result generated by feedback so that institution can strive to improve in those areas.

KEYWORDS: Educational Data Mining, Accreditation, Data Governance, Google Sheets Database.

I. INTRODUCTION

With rise of technological advancement everyday, we move more towards automating the manually managed previous processes. This has led to generation of large volumes of data or knowledge base. Data mining is successfully extracting the useful information from these knowledge bases. Data mining has become critical for any institution to get a better insight on the utility of its resources. The main function of data mining is to apply algorithms and extract patterns in data. There is increasing research interest in data mining and this has led to the emergence of a new field called educational data mining.

Using these techniques different types of knowledge can be found such as association rules, classifications and clustering.

In this paper we have proposed a system to replace the current manually managed data of Vivekananda Education Society’s Institute of Technology with online system that generates results of student performance in seconds that would have otherwise been required hours of manual computation. Data governance is a control that ensures that the data entry by an operations team member or by an automated process meets precise standards, much as a Business rule, a data definition and data integrity constraints in the data model.

For managing the data of this above mentioned institution we have first modeled the entire data sets to accept only clean data. This is achieved through the medium of google sheets and Google app scripts. Certain rules are defined to accept only data that is consistent with the rules that define the clean data sets. This overcomes the problem of inconsistency in manual data entry wherein each data item entered may be in different format then the previous one and data consistency is subject to circumstantial situations.

Taking the next step on the same guidelines, we have developed rules that check for satisfaction of rules stated by National Board of Accreditation in their Self Assessment Report (SAR). these rules will help develop patterns in the previous datasets and predict the ultimate quality of education imparted by the institute and also the student response and absorption level of the concepts. On having these predictions we will be able to judge the performance of students and the faculty and also predict their approximate performance in the near future. This prediction will help in estimating the college performance outcome after the accreditation review.

The main objective of the paper will be to help the college understand the students strengths and weakness and work accordingly. For this the achievement of Course Outcomes and Program Outcomes will be predicted and students will be helped to work into areas of their strength in situations like choosing their domain of interest.

II. Related work

[1] is designed to justify the capabilities of data mining techniques in context of higher education by offering a data mining model for higher education system in the university. In this research, the classification task is used to evaluate student’s performance and as there are many approaches that are used for data classification, the decision tree method is used here. Here, the classification task is used on student database to predict the students division on the basis of previous database. Study [2] presents an approach to classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system. Four classifiers were used to segregate the students by using a genetic algorithm The approach to classify students in our approach was inspired from here.

In [3] the applications of data mining in education for student profiling and grouping are discussed. They make use of Apriori algorithm for student profiling which is one of the popular approaches for mining associations i.e. discovering correlations among set of items. The other algorithm used, for grouping students is K-means clustering which assigns a set of observations into subsets. Here, data mining methods are used to find hidden patterns about student performance. The application of kmeans algorithm was understood from here.

III. Proposed algorithm

1. Google Apps Scripts

Google Apps Script is a scripting language that occupies little memory for application development on the Google Apps platform. Google Apps Scripts is based upon JavaScript 1.6 including some parts of 1.7 and 1.8 and provides subset of ECMAScript 5 API, however, the entire app is executed on Google which would rather have used client machine and memory in client site. According to Google, Apps Script facilitates simple methods to automate most of the tasks across Google products line up and third party apps and services. Apps Script is also the tool that powers the add-ons for Google Docs, Sheets & Forms.

Benefits of Google App Scripts are as follows:

- As it is based on JavaScript it is simple to learn.
- Debugging the App Scripts are done by debugger thats entirely on cloud and can be accessed in web browser.
- It can be used to make easy tools for an organization for itself.
- It can perform simple system administration tasks

2. Google Sheets

2.1. About Google Sheets

Google Sheets is a spreadsheet that is included in the web-based software office suite that is offered by Google within the Google Drive services. The suite allows users to create and edit documents online on the go and also collaborating with other users in real-time.

The app is available as web applications, as Google Chrome apps that work offline, and also as mobile apps for Android devices as well as iOS machines. The app offers compatibility for Microsoft Office file formats and other documents. The suite also consists of Google Forms (survey software), Google Drawings (diagramming software) and Google Fusion Tables (database manager).

2.2. Why Google Sheets over Database

Spreadsheets as well as databases offer to store and manage data. The basic elements in a spreadsheet or a database is a set of data values. Spreadsheets and databases differ in how to store and modify the data.

A spreadsheet stores data values in cells, multiple cells form up the rows and columns that are used to store multiple data values. Cells can refer to other cells, and the spreadsheet constitutes of cells that allow and can process on other cell values.

A database generally stores data values in tabular form. Each table is named and can have one or more than one

columns and rows. A row in a table is said to be a record. A record contains a value for each column in the table. Databases can be related with records in different tables. Using records we can update, filter and sort the data.
Databases offer up a large range of complexity in data manipulation, but this has to be coded in programming or SQL code. However, for basic data processing, spreadsheets have a large number of automated functions, which are easily accessible to people who do not have much technical experience. Thus spreadsheets can be easily filled up even by the students or the faculty as requirement arises.

2.3 R Programming

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

2.4 Shiny

- Shiny is a new package from RStudio that makes it incredibly easy to build interactive web applications with R.
- Build useful web applications with only a few lines of code - no JavaScript required.
- Shiny applications are automatically "live" in the same way that spreadsheets are live. Outputs change instantly as users modify inputs, without requiring a reload of the browser.
- Shiny user interfaces can be built entirely using R, or can be written directly in HTML, CSS, and JavaScript for more flexibility.
- Works in any R environment (Console R, Rgui for Windows or Mac, ESS, StatET, RStudio, etc.)
- Attractive default UI theme based on Twitter Bootstrap.
- A highly customizable slider widget with built-in support for animation. Pre-built output widgets for displaying plots, tables, and printed output of R objects.
- Uses a reactive programming model that eliminates messy event handling code, so you can focus on the code that really matters.

IV. Pseudo Code

```
1. kMeans:
   Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

   1) Randomly select 'c' cluster centers.
   2) Calculate the distance between each data point and cluster centers.
   3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the centers.
   4) Recalculate the new cluster center using:
      $v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$
```

where, ‘ci’ represents the number of data points in ith cluster.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

2. DBSCAN:

```
DBSCAN(D, eps, MinPts) {
 C = 0
for each point P in dataset D {
if P is visited
continue next point
    mark P as visited
    NeighborPts = regionQuery(P, eps)
    if sizeof(NeighborPts) < MinPts
      mark P as NOISE
    else {
      C = next cluster
      expandCluster(P, NeighborPts, C, eps, MinPts)
    }
  }
}

expandCluster(P, NeighborPts, C, eps, MinPts) {
 add P to cluster C
  for each point P' in NeighborPts {
    if P' is not visited {
      mark P' as visited
      NeighborPts' = regionQuery(P', eps)
      if sizeof(NeighborPts') >= MinPts
        NeighborPts = NeighborPts joined with NeighborPts'
    }
    if P' is not yet member of any cluster
   add P' to cluster C
  }
}

regionQuery(P, eps)
  return all points within P's eps-neighborhood (including P)
```

V. SIMULATION RESULTS

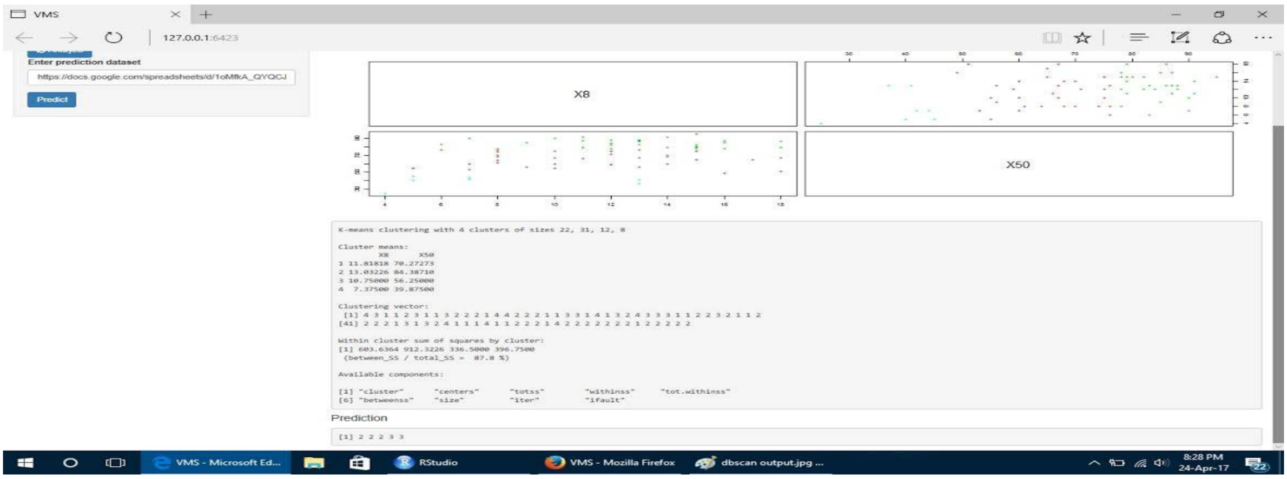
Data set was formed by collecting data through google forms. A sample of the data is given below

No.	ITM	ASM	ATT	TW	FG
1	13	B	50	17	P
2	14	B	75	20	P
3	16	A	75	22	P
4	16	A	75	23	P
5	14	B	80	20	P
6	15	B	90	25	P
16	16	A	75	19	P
17	17	A	40	16	P
18	11	B	60	13	F
19	15	C	100	15	P
20	12	A	70	16	P
21	18	B	75	17	P

7	16	A	75	19	P
8	17	A	75	20	P
9	20	A	75	18	P
10	14	B	75	16	P
11	15	A	75	17	P
12	16	A	75	20	P
13	14	A	80	21	P
14	12	B	50	24	F
15	15	A	80	18	P
22	17	B	60	18	P
23	16	B	75	20	P
24	18	B	75	24	P
25	14	C	80	23	P
26	15	B	75	20	P
27	16	A	80	16	P
28	17	C	80	18	P

Table No. 01
Example Scores of Students

The App deployed using R Studio and Shiny packages at https://vmsapp.shinyapps.io/vmsd/ gives the following output:


Enter prediction dataset

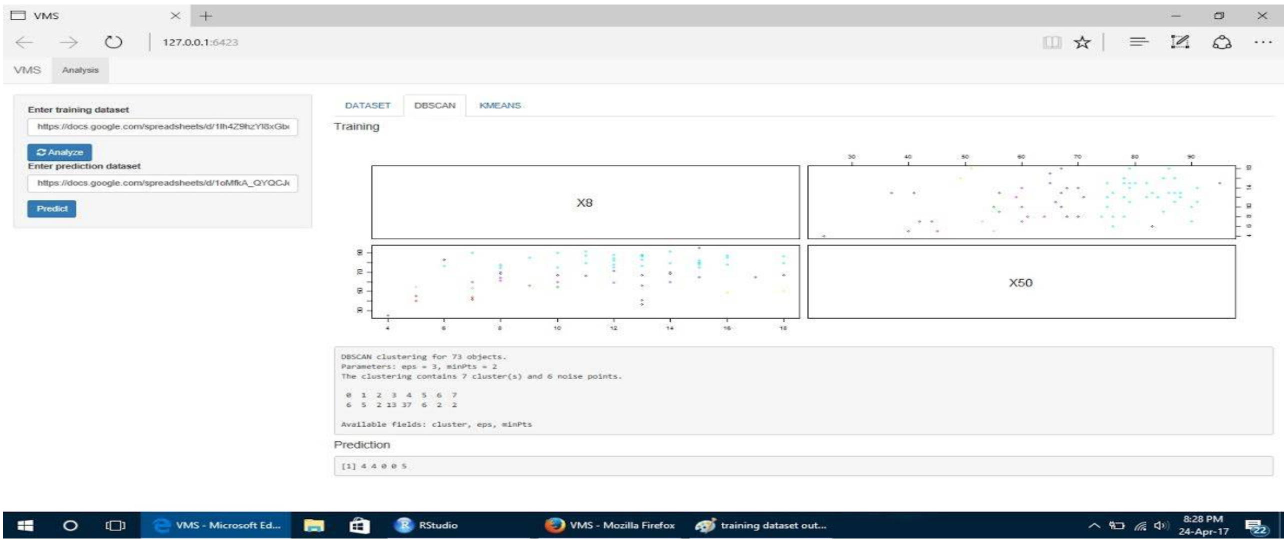
https://docs.google.com/spreadsheets/d/1oMfkA_QYQCJ

```
K-means clustering with 4 clusters of sizes 22, 31, 12, 8

Cluster means:
        X8       X50
1 11.81818 70.27273
2 13.03226 84.38710
3 10.75000 56.25000
4  7.37500 39.87500

Clustering vector:
 [1] 4 3 1 1 2 3 1 1 3 2 2 2 1 4 4 2 2 2 1 1 3 3 1 4 1 3 2 4 3 3 3 1 1 2 2 3 2 1 1 2
[41] 2 2 2 1 3 1 3 2 4 1 1 1 4 1 1 2 2 2 1 4 2 2 2 2 2 2 2 1 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 603.6364 912.3226 336.5000 396.7500
 (between_SS / total_SS =  87.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Prediction

```
[1] 2 2 2 3 3
```

k-Means Output


Enter training dataset

https://docs.google.com/spreadsheets/d/1Ih4Z9hzYl8xGbi

Enter prediction dataset

https://docs.google.com/spreadsheets/d/1oMfkA_QYQCJ

Training

```
DBSCAN clustering for 73 objects.
Parameters: eps = 3, minPts = 2
The clustering contains 7 cluster(s) and 6 noise points.

 0  1  2  3  4  5  6  7
 6  5  2 13 37  6  2  2

Available fields: cluster, eps, minPts
```

Prediction

```
[1] 4 4 0 0 5
```

DBSCAN Output

VI. CONCLUSION AND FUTURE WORK

This paper will help the college in supervising predicting all factors that are responsible for grading the college during accreditation based on the previous performance results and rules specified by National Board of Accreditation (NBA). For this, google sheets and google app scripts is utilized and Iterative Dichotomizer 3 (ID3) prediction algorithm is applied. After, the implementation of of this project the college is able to control the grade that will be obtained from the accreditation committee post the college analysis.

REFERENCES

1. Brijesh Kumar Baradwaj and Saurabh Pal “Mining Educational Data to Analyze Students Performance” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
2. Behrouz Minaei-Bidgoli, Deborah A. Kashy, Gerd Kortemeyer and William F. Punch “Predicting Student Performance: An Application of Data Mining Methods with an Educational Web Based System” 33'd ASEEIIEEE Frontiers in Education Conference.
3. Suhem Parack, Zain Zahid and Fatima Merchant “Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns” .
4. Yahya Eru Cakra and Bayu Distiawan Trisedya “Stock price prediction using linear regression based on sentiment analysis”.
5. N.D. Valakunde and M. S. Patwardhan “Multi-aspect and Multi-class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process”.
6. Kranti Ghag and Ketan Shah “Comparative analysis of the techniques for Sentiment Analysis”.

BIOGRAPHY

Kunal Dulani, Mohit Chanchlani, Meeta Chanchlani, Manoj Ahuja are all students final year students at Vivekanand Education Society’s Institute of Technology.

Richard Joseph is a Professor at Vivekanand Education Society’s Institute of Technology in the computer department.