Enriching the Students Learning activities by analyzing the skills using data mining techniques

Pranav kanth A[1]
CSE
Bannari Amman Institute of Technology, Tamil Nadu
India.
pranavkanth.cs17@bitsathy.ac.in

Prasanth M[2]
CSE
Bannari Amman Institute of Technology, Tamil Nadu
India
prasanthm.cs17@bitsathy.ac.in

Pravin A S[3]
CSE
Bannari Amman Institute of Technology, Tamil Nadu
India
pravin.cs17@bitsathy.ac.in

Abstract: Every educational institution will have their own strategy to manage the student's record. Analyzing the individual student's performance based on the various parameters is now a challenging task. This paper analyses and predicts student's performance using data mining techniques for the data sets of minimum number of students in the area of mathematics skill, reading and writing skills. The parameters such as parent's education, gender, race/ethnicity provide implications that affects the student marks either directly or indirectly This proposed system is an efficient approach which provides path for the instructors to take better decision about the individual students that enriches the students learning and makes the student to consistently improve himself to compete in today's world and to be a responsible citizen.

Keywords: students performance; data mining; data analysis; data visualization; factors affecting students performance; overall students performance

I. INTRODUCTION

Several decades back storing huge information was unimaginable, but advancements in science and technology has made it possible today. Hence the storage devices have become affordable, this led to the collection of huge datasets namely big data. These collected information can be used to make decisions, gain insights and to get an overview. In any educational institutions there are volumes of data like their marks, parental level of education, race/ethnicity and so on. These data can be qualitatively and quantitatively analysed using data mining techniques. The subjects which students study resides on three fundamental skills that is, mathematics, reading and writing skills. By using the test scores obtained in these subjects we can judge the quality of a student. This paper is focused on the collection of students data set and using data mining techniques to understand the factors affecting their scores. This research work primarily uses matplotlib (a library for visualisation) and seaborn (advanced visualisation) in python to perform data analysis and visualisation. It will also help the faculties to identify the result scores of the students in advance.

This paper is structured as follows: Literature Review, Methodology, Experimental Setup, Result & Analysis, Conclusion and Reference.

II. LITERATURE REVIEW

A. *Data mining and visualisation techniques*

Data mining is the process of deriving insights and to draw conclusions from the data Data mining is used to extract useful information from the given data. It deals with various kinds of patterns [9].

Data visualisation using matplotlib [10].

Data visualisation helps to represent the data in a pictorial way, the goal of visualisation is to communicate information in a easy and effective way to the users.

Many data mining research work has been conducted for the performance of students and classification technique are widely used in this analysis. Some of the research work is shown in the Table 1

Table 1

Name of the Author/Year	Tecnic/Attribute selected	Analysis/Usage
Oyelade, O. J, Oladipupo, O. O, Obagbuwa,	K-Means/Grade Point Average[7]	Cabhndidates' performance is impoved on future academic

I. C[7]/2010		session by monitoring his performance every semester.
Surjeet Kumar Yadav, Brijesh Bharadwaj, Saurabh Pal [1]/ 2012	Decision tree/ Previous Semester Marks, Class Test Grade, Seminar Performance, Assignment, Attendance, Lab work, and End Semester marks [1]	Decision tree accuracy along with the performance of students in various courses.
Mohammed M. Abu Tair, Alaa M. El-Halees [2]/2012	Classification technique/ student id to address and from matriculation GPA to college GPA and grade [2]	Low grades of graduate students need to be resolved on the basis of feedback provided to college management.
Saurabh Pal [4] / 2012	Predictive model to generate accurate prediction list [4]	This assessment yields most likely students who are most likely to drop out after first year of engineering.
T. Miranda Lakshmi , A. Martin, R. Mumtaj Begum, Dr.V.Prasanna Venkatesan [5] / 2013	Decision tree algorithm/ parent's Qualification, living location, economic status, family and relation support, resource accessibility [5].	The effect of various factors on student's performance along with the implementation of decision tree.
K. Shanmuga Priya, A.V.Senthil Kumar [6]/ 2013	Overall semester marks, practical lab, attendance , paper presentation, end semester marks [6].	Student performance improvement .

B. *Visualisation Technique(matplotlib)*

Matplotlib is a python package for plotting that generates quality graphs. Matplotlib is designed to create simple and complex plots with a few commands. Matplotlib works with pandas, numpy, and other extension code. With a few lines of code, it generates plots, histograms, power spectra, bar charts, error charts, scatter plots, line plots etc.

C. *Tools Used*

The analysis and visualisation is performed using jupyter notebook. The jupyter notebook is an open source web application that allows you to create and share documents that contains live core, equations, visualisations and narrative text.

III. METHODOLOGY

The proposed methodology involves collecting the data set, data pre processing, analysing the data and visualisation.

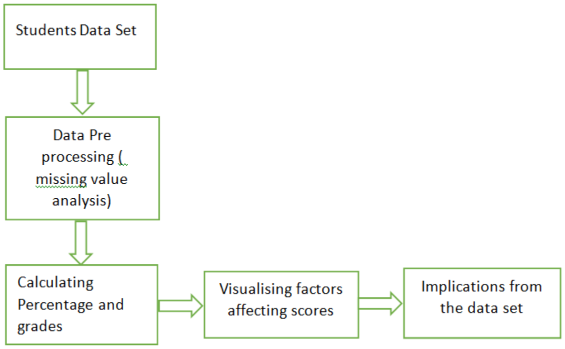

Students Data Set
Data Pre processing (missing value analysis)
Calculating Percentage and grades
Visualising factors affecting scores
Implications from the data set

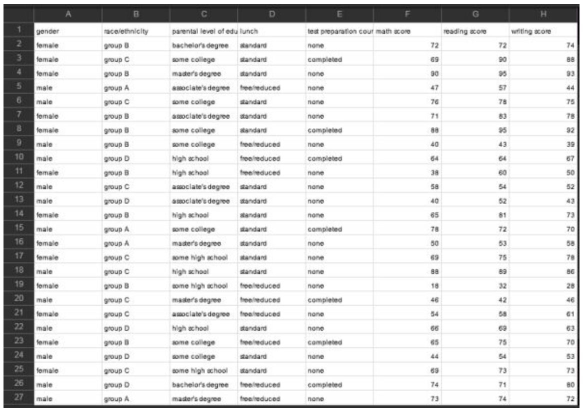
Figure 1 : Methodology Adopted

1. Firstly the most important thing for data mining is harvesting the data set. In this project the data set is obtained from an open source platform.
2. The data that has been collected is checked for null values.
3. The reading, writing and maths scores are added to calculate the percentage and also the grades.
4. Visualising various attributes like gender, parent's education, test preparation etc. with respect to the percentage.
5. Implications made from analysing and visualising the data set.

IV. EXPERIMENTAL SETUP

A. *Data Set*

The data set has been collected from the internet. The data set contains 1000 rows and 11 columns. The attributes listed in the data set are gender, race/ethnicity, parental level of education, lunch, test preparation course, maths score, reading score and writing score. The data set is pre processed for missing values. The grade and percentage are derived from the reading, writing and maths scores. The screen shot of the data set is given below



	A	B	C	D	E	F	G	H
1	gender	race/ethnicity	parental level of edu	lunch	test preparation cour	math score	reading score	writing score
2	female	group B	bachelor's degree	standard	none	72	72	74
3	female	group C	some college	standard	completed	69	90	88
4	female	group B	master's degree	standard	none	90	95	93
5	male	group A	associate's degree	free/reduced	none	47	57	44
6	male	group C	some college	standard	none	76	78	75
7	female	group B	associate's degree	standard	none	71	83	78
8	female	group B	some college	standard	completed	88	95	92
9	male	group B	some college	free/reduced	none	40	43	39
10	male	group D	high school	free/reduced	completed	64	64	67
11	female	group B	high school	free/reduced	none	38	60	50
12	male	group C	associate's degree	standard	none	58	54	52
13	male	group D	associate's degree	standard	none	40	52	43
14	female	group B	high school	standard	none	65	81	73
15	male	group A	some college	standard	completed	78	72	70
16	female	group A	master's degree	standard	none	50	53	58
17	female	group C	some high school	standard	none	69	75	78
18	male	group C	high school	standard	none	88	89	86
19	female	group B	some high school	free/reduced	none	18	32	28
20	male	group C	master's degree	free/reduced	completed	46	42	46
21	female	group C	associate's degree	free/reduced	none	54	58	61
22	male	group D	high school	standard	none	66	69	63
23	female	group B	some college	free/reduced	completed	65	75	70
24	male	group D	some college	standard	none	44	54	53
25	female	group C	some high school	standard	none	69	73	73
26	male	group D	bachelor's degree	free/reduced	completed	74	71	80
27	male	group A	master's degree	free/reduced	none	73	74	72

Figure 2 : Data Set Collected

A. Attributes

The attributes listed in the data set gender, race/ethnicity, parental level of education, lunch, test preparation course, maths score, reading score, writing score, percent, result and grade. The result and grade are calculated from the scores given in the data set.

B. Counts of the attributes

1. GENDER: The gender is classified into two types namely male and female. The total counts of male and female are 482 and 518 respectively.
2. Parental Level of education : This column contains information about the student's parent educational qualification.

   Some college: 226
   Associate's degree: 222
   High school: 196
   Some high school: 179
   Bachelor's degree: 118
   Master's degree:59
3. Race/ ethnicity : The students in the dataset comes under five race/ethnicity namely group a, group b, group c, group d, group e.

   Group A: 89
   Group B: 190
   Group C: 319
   Group D: 262
4. Lunch and test preparation : The lunch that students have is of two types namely standard, free/reduced. Some students have prepared for the test and some haven't.

C. *Jupyter Notebook tool*:

Jupyter notebook is used to do the analysis and visualisation in a web interface.

V. RESULT AND ANALYSIS

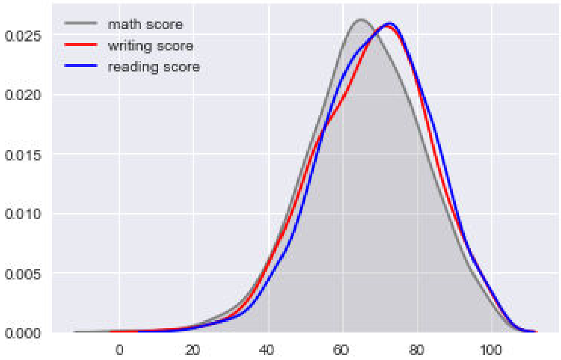


Figure 3 : Density plot representing marks

Figure 3 shows the kernel density plot for reading, maths and writing scores. From this plot it can be understood that the reading score has the highest mean followed by writing score and then the maths score. It also shows that the maximum score for the three attributes is 100.

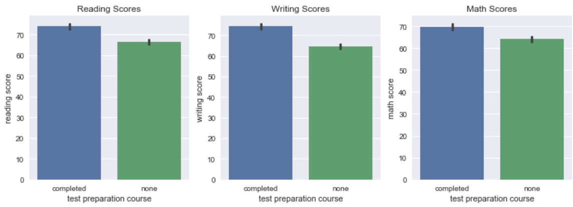


Figure 4 : Scores with respect to preparation

Figure 4 shows the marks obtained by the students in the three attributes with respect to their test preparation. It could be inferred from the above figure that the students who have prepared for the test has scored well.

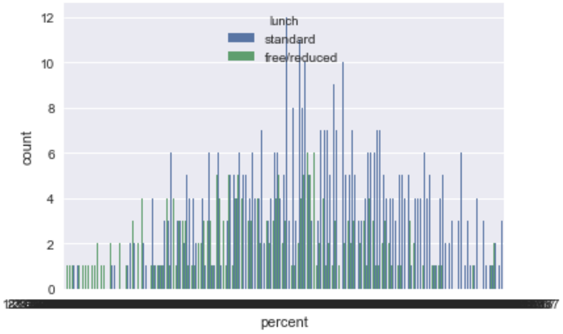


Figure 5 : Percentage with respect to lunch

This plot represents the percentage obtained by the students with respect to the lunch, surprisingly the lunch is related to the students percentage. The students who scored more percentage has standard lunch, when compared to the students having free/reduced lunch.

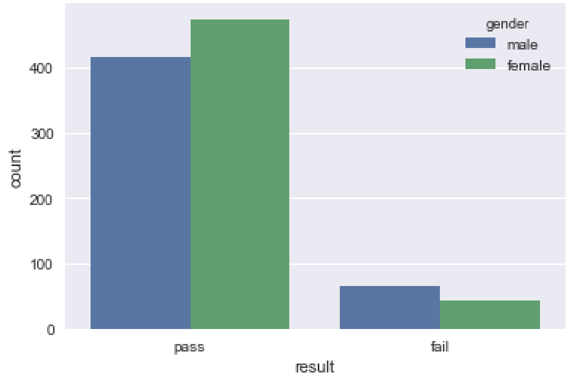


Figure 6 : Results based on gender

This plot shows the results obtained by the students, the students who have scored below 50% percentage is considered fail. This histogram tells us that the majority of the students who passed in the exam are female and the majority of the students who failed in the exam are male.

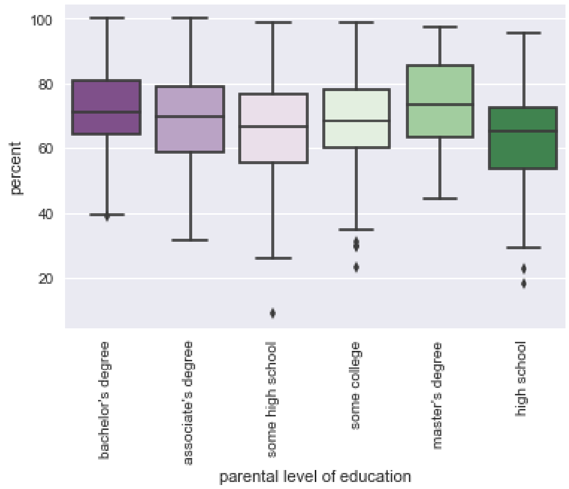


Figure 7 : Box plot on parental level of education

This box plot visualises the percentage with parental level of education. It can be seen from the plot that the students whose parents have completed the master's degree has the highest average and the students whose parents have completed the high school has the lowest average, hence parents level of education plays an important role here. The students whose parents have completed either bachelor's degree or associate's degree has the highest score.

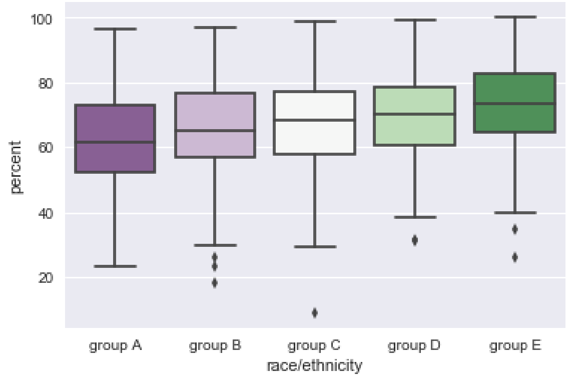


Figure 8 : Box plot on race/ethnicity

This box plot tells us the percentage obtained by the students with their race/ethnicity. The students in the group A has the lowest average and the students in the group E has the highest average. The students in the group E has scored the maximum percentage and the students in group C has the minimum percentage. This provides us with the insight that the students in the group E are likely to secure more marks.

A. Analysis of the percentage with grades

The percentage obtained by the students is classified into grades for a better understanding. The percentage is the sum of maths, reading and writing scores.

Table 2

Percentage	Grade
Percentage >= 80	A
Percentage >= 70	B
Percentage >= 60	C
Percentage >= 50	D
Percentage >= 40	E

With this classification of grades, plots are plotted for the race/ethnicity and parental level of education.

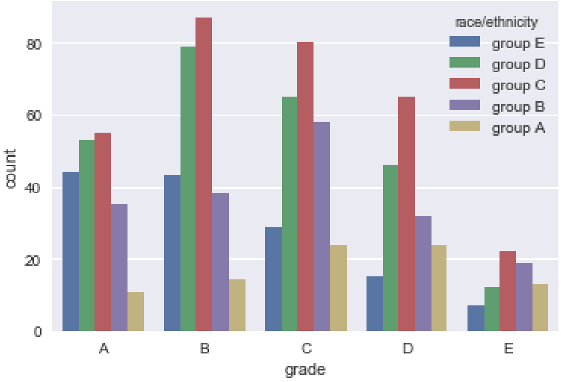


Figure 9 : Grades obtained with respect to race/ethnicity

This visualisation provides the relationship between the grades and the race/ethnicity. The 'A' grade has been mostly secured by the students belonging to group C followed by the students group D. It can be understood from the graph that majority of the students are from group C. The grade 'E' has been mostly scored by the students in the group C and minimum in group E. In this data set the students belonging to group E have performed well.

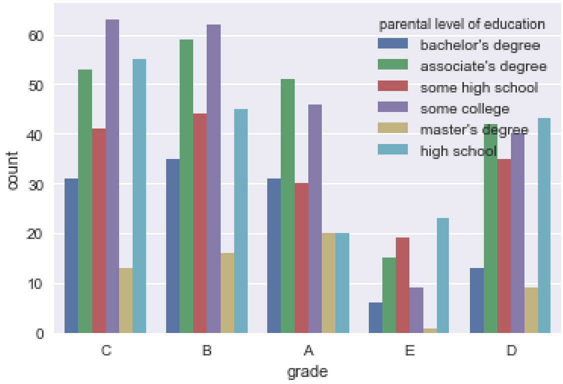


Figure 10 : Grades with respect to parent's education

From this graph the students whose parents with associates degree, master's degree and attended some college has scored well compared with the other parents.

VI. CONCLUSION

In this paper, we have done a basic analysis and visualisation to understand the factors influencing the students marks.

The analysis is helpful during the admission and placement process. The parameters taken for the analysis are: parental level of education, race/ethnicity, test preparation, lunch and gender. This work can be further extended by adding more attributes and use machine learning algorithms to build a model to predict whether the student will be eligible for a job/admission in an educational institution.

VII. REFERENCES

[1] Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. arXiv preprint arXiv:1202.4815.

[2] Tair, M. M. A., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. International Journal of Information, 2(2).

[3] Yadav, S. K., & Pal, S. (20 12). Data mining: A prediction for performance improvement of engineering students using classification. arXiv preprint arXiv:1203.3832.

[4] Pal, S. (2012). Mining educational data to reduce dropout rates of engineering students. Internatio nal Journal of Information Engineering and Electronic Business (IJIEEB), 4(2), 1.

[5] Lakshmi, T. M., Martin, A., Begum, R. M., & Venkatesan, V. P. (2013). An analysis on performance of decision tree algorithms using student's qualitative data. *International Journal of Modern Education and Computer Science (IJMECS)*, *5*(5), 18.

[6] Priya, K. S., & Kumar, A. S. (2013). Improving the Student's Performance Using Education al Data Mining. *Int. J. Advanced Networking and Applications*,*4*(04), 1680-1685.

[7] Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C. Application of k-Means Clustering Algorithm for prediction of Students' Academic Performance. (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 1, 2010

[8] Ishwank Singh ; A Sai Sabitha ; Abhay Bansal. Student performance analysis using clustering algorithm. 6th International Conference – Cloud Systems and Big Data Engineering (Confluence), 2016, IEEE Conference.

[9] Data Mining Comp 150 DW C hapter 7. 'Classification and Prediction' Dan Hebert.

[10] Niyazi Ari ; Makhamadsulton Ustazhanov. Matplotlib in python. 2014 11th International conference on Electronics, Computer and Computation (ICECCO).