GRADE PREDICTION USING MACHINE LEARNING

[1] Sushmita Gaonkar

[1]Student
[1]CSE,
[1]Nhce, Bangalore, India.

Abstract : Large amount of data is generated across multiple fields such as education, medical, defenses, social media and so on. Machine Learning (ML) and data Mining (DM) are two techniques which can be used to identify the hidden patterns. One of the key areas of this application of Education Data Mining is developing the prediction model to predict college grades. We have built a model that predicts the future grade of the college based on the current activities they perform. We have used linear regression model which will help colleges to priorly know the grades which they will obtain so if grades are less than they expected they can improve their grades by improving the activities.

***IndexTerms* – Machine Learning, Data Mining, Education Data Mining, Prediction, Regression.**

I. INTRODUCTION

Large amount of data is produced across a variety of fields. The Big Data is used to collect and organize data and valuable data need to be extracted. We need a large amount of data to analyze the useful information. In order to get that we use Data Mining and Machine Learning Techniques to build a model which can analyze the patterns and give the required prediction or the results. Educational Data Mining (EDM) collects the data about the organization and learns their environment to provide a useful model to identify the patterns from the available data. Educational organization is always a concern about their grades and ranking of colleges, so by using ML and DM techniques we can used to develop a model which can help them to improve their college grades or ranking.

II. PREREQUISITES

Machine learning: It is an application of AI that provides ability to the system to learn automatically and improve from the experiences without being explicitly programmed. Numerous Machine learning algorithms are used to predict college performance as it is a very complex issue. Machine Learning is defined as: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”.

The prediction models will not only give the prediction about the grades of the college, but will also provide real time advice to resolve their difficulties. The key difference between human and computer is the ability to think and learn from experience. Humans learn from the experiences at other hand Computers execute human-made algorithms to get the knowledge, get trained and make the predictions. Machine learning algorithms are found to be very effective to particular type of learning. They are mostly useful in circumstances where the human doesn’t have sufficient amount of knowledge.ML algorithm predict the future based on the input given to it, it explores given input data and get trained on it and produce hypotheses.

In the further section, we discussed some of the machine learning algorithms.

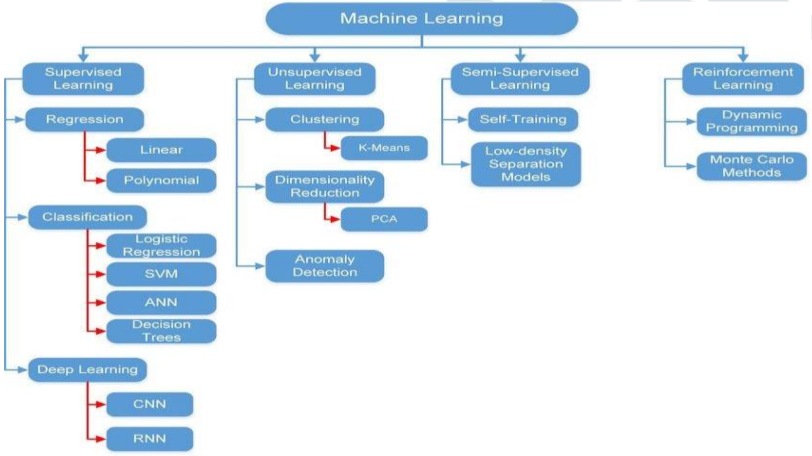

Fig.1 Machine Learning Algorithm

- Supervised learning algorithms: We train the machine using labeled data which is already tagged with correct answer , then machine is given the new set of data so that machine analyses the set of training samples and produces the correct output.
- Unsupervised learning algorithm: Here data is provided without any guidance to machine, here we don’t provide any labeled or classified data. Machine by itself has to sort unsorted data based on the similarities and patterns

Supervised machine learning consist of two processes: Classification and regression.

Classification: here incoming data is labeled based on former data sets and algorithm is manually trained to recognize type of object and based on the object data is categorized. Machine should learn how to distinguish data, binary recognition or image.

Type of classification algorithm:

1. Linear Classifiers: Logistic Regression, Naive Bayes Classifier
2. Nearest Neighbor
3. Support Vector Machines
4. Decision Trees
5. Boosted Trees
6. Random Forest
7. Neural Networks

1. Naive Bayes Classifier: Is based on Bayesian theorem. It classifies assuming that hidden attribute doesn't affect predictions and the value of the given attribute is independent.
2. Nearest Neighbor: it takes a group of labelled points and with the labelled points learns to label other points. To label a new point, it looks at the labelled points closest to that new point and has those neighbors point.
3. Support Vector Machines: Support Vector Machine (SVM) , to achieve the prediction decision it derive features from the variables, and then manage them in linear combination.
4. Decision Trees: It builds the model in the form of tree structure by breaking down the data-set into subsets. Finally tree has decision nodes and leaf nodes. Top most node is called as root node and bottom nodes are called leaf nodes which represents decision or classification.

Regression: In regression we predict the output by identifying the patterns. Here system has to understand values numbers groups etc.

Type of regression algorithm:

1. Linear
2. Polynomial

Linear Algorithm:

Linear regression is one of the most widely used algorithms in machine learning

Here dependent variables are continuous, independent variables are either continuous or discrete, and nature of the regression line will always be linear.

Polynomial regression:

A regression equation is said to be polynomial regression if power of independent variable is greater than one.

Y=A+B*X^2

In polynomial regression, best fit line is a curve that fits into data points, rather than the straight line.

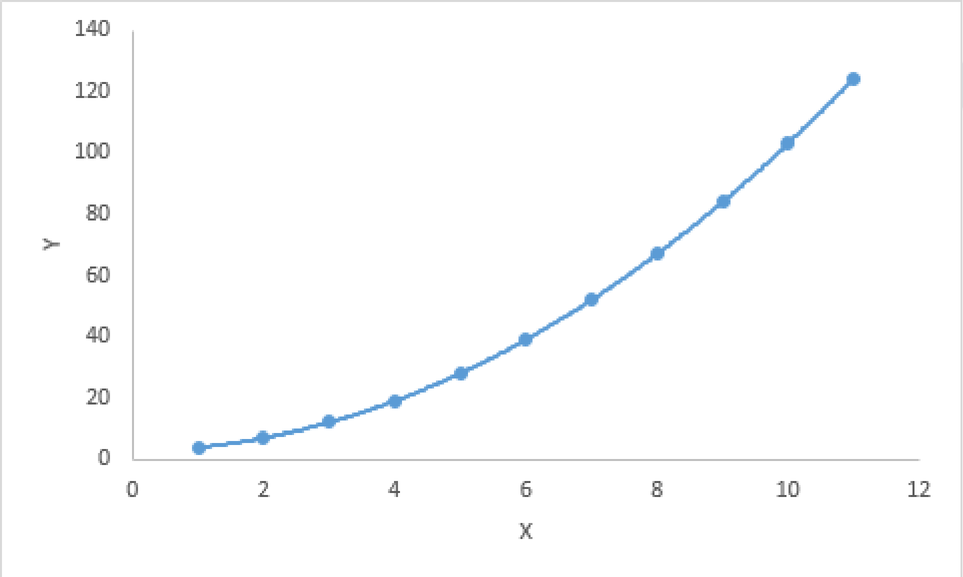

Fig.2 Polynomial Regression

There can be temptation to fit higher degree polynomial to get less errors which can result in overfitting so we have to plot a graph is such a way that it fits the nature of problem.

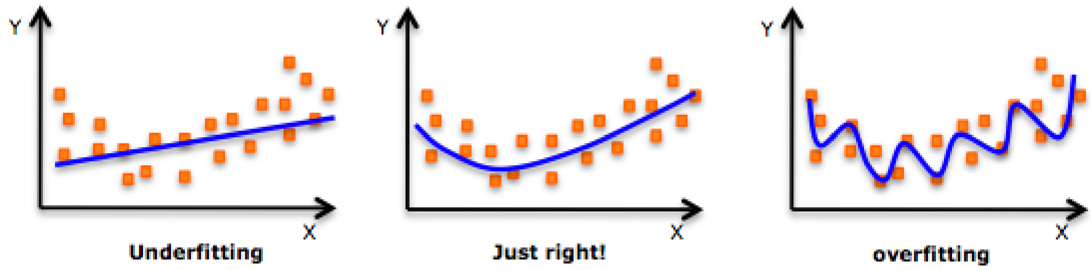

Fig.3 Possible Polynomial Graphs

III. METHODOLOGY

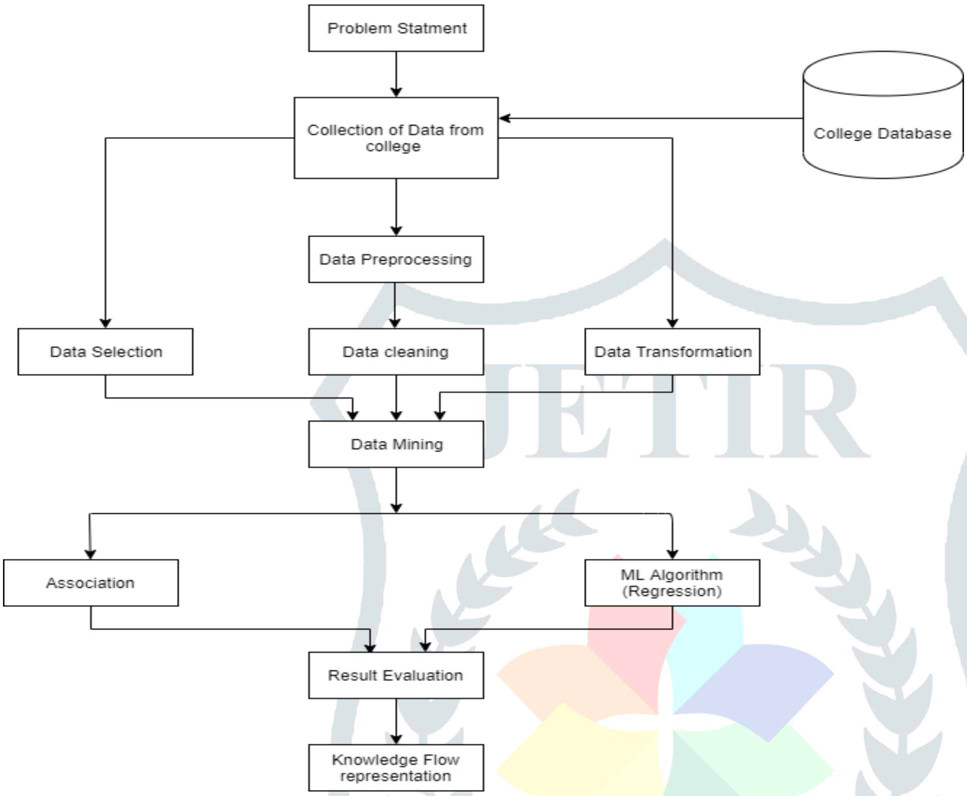

Fig.4 Data Flow

Problem Statement: Here we are looking forward to predict the college future grades based on their current activities they perform. Which will help the organization to improve their grade by improving the current activities.

Data Set: Once we get the problem statement, the next process is to collect the relevant data. We have collected the data from the internet sources, where colleges upload their self-study report (SSR).Based on the SSR which college has uploaded we have accessed the needed data.

Data Preprocessing: Once we collected the raw data we need to process it, before we do any analysis on that data. Usually data from such sources will be very huge, messy, duplicate data can be present, there can be some missing values. In order to get the accurate analysis we need to cross check the data.

The common errors which we came across :

- Missing values
- Unreverent data for our analysis such as address and the names of the students and staff etc.
- invalid entries

Data Mining: once we get the data cleaned and transform it to required format. We had to closely examine the factors which will directly affect the result and which indirectly affect so keeping both in mind we begin the further process to analyze the patterns more deeply. We have divided the whole data into 60:40 where 60% of data is taken as training set and the remaining 40% is used for testing.

Machine Learning Algorithm: At this step we have applied all our mathematical, statistical and technological knowledge to apply best fit Machine Learning algorithm to get the predictive model that predicts the grade of the college. From qualitative analysis we got the quantitative data and applied the machine learning algorithm on the quantitative data. Here we used Linear Regression Algorithm to predict the analysis.

Y=B0+B1*X + e,

Where Y is response variable

X is predictor variable

B0 is intercept parameter,

B1 slope parameter

e Is error term representing deviations of Y about A0+B1*X

This equation uses given predictor variables to predict value of targeted variables

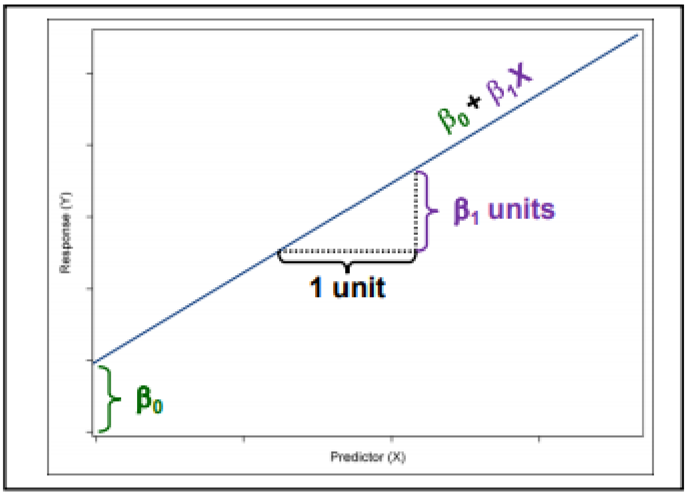

Fig.5 Linear Regression Graph

Once the machine is trained using 80% of data, then other 20% of data is used for testing the model. Result are evaluated by compare the association data and model output.

Below image shows the attributes used in training the data and their key values.

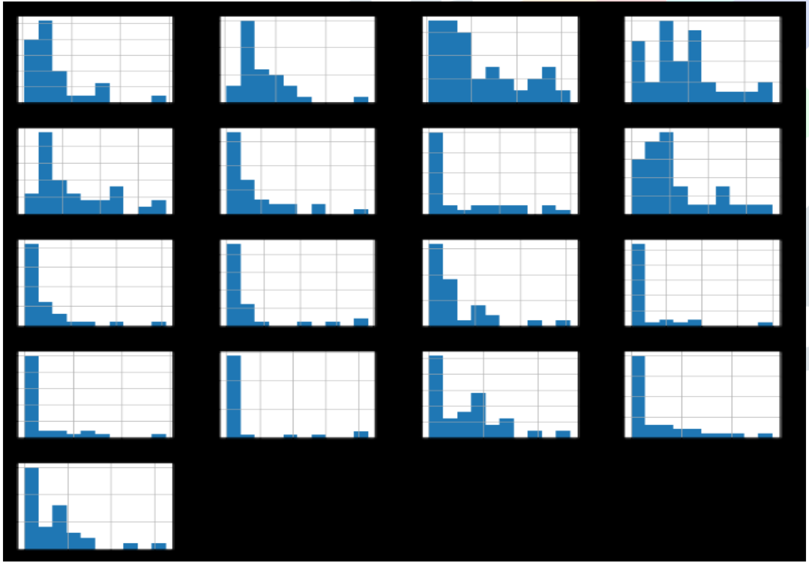

Fig.6 Attributes output

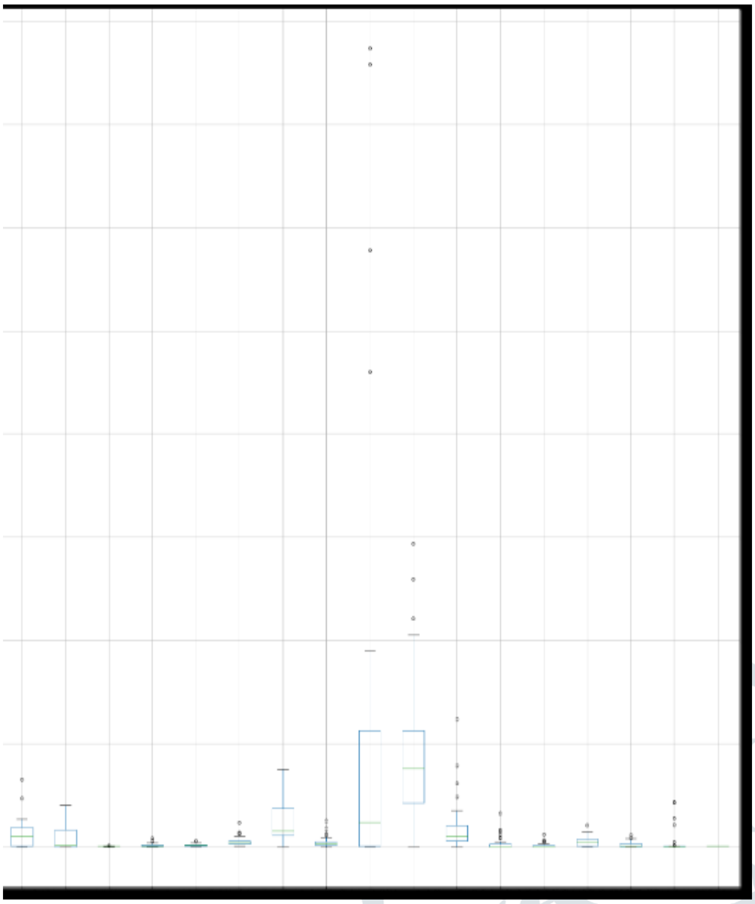

Fig.7 over fitting Display

IV. Conclusion

This paper presents machine learning algorithms that predict the college grades based on their current activities. The patterns are recognized and analyzed to offer recommendation to the organization. Regression Algorithms are used to predict the grades of the college.

The college performance is evaluated based on academic activities and infrastructure of the college. Data is been collected from various sources college database, SSR report, internet sources and so on. Once data was cleaned and processed it was given to regression model to evaluate. So based on output we conclude that for this kind of prediction linear regression is a best suet. As now we are using limited amount of data. In future we can get maximum data and can use different algorithm to get better prediction model.

References

[1] Samrat Singh, Dr. Vikesh Kumar , "Performance Analysis of Engineering Students for Recruitment Using Classification Data Mining Techniques ",IJCSET February 2013.

[2] M. Goyal and R. Vohra, “Applications of Data Mining in Higher Education”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue2, No 1, March 2012.

[3] Jason Brownlee ,"How to Save Your Machine Learning Model and Make Predictions in Weka", August 3, 2016.

[4] Neelam Naik & Seema Purohit, “Prediction of Final Result and Placement of Students using Classification Algorithm”International Journal of Computer Applications (0975 – 8887) Volume 56– No.12, October 2012

[5] Alaa M.El-Halees,Mohammed M. Abu Tair, “Mining Educational Data to Improve Students’Performance: A Case Study”,International Journal of Information and Communication Technology Research, 2012.

[6] B.K. Bharadwaj and S. Pal,“Data Mining: A prediction for performance improvement using classification”, International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[7] Suchita Borkar, K. Rajeswari, "Predicting Students Academic Performance Using Education Data Mining ", IJCSMC,Vol. 2, Issue. 7, July 2013, pg.273– 279.

[8] Randhir Singh, M.Tiwari, Neeraj Vimal,"An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education", 2013.

[9] D.Magdalene Delighta Angeline,"Association Rule Generation for Student Performance Analysis using Apriori Algorithm",The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 1, March-April 2013

[10] Mrs. M.S. Mythili, Dr. A.R.Mohamed Shanavas,"An Analysis of students’ performance using classification algorithms ",ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver .III (Jan. 2014), PP 63-69

[11] S. Anupama Kumar and Dr. Vijayalakshmi M.N “Implication of classification Techniques in Predicting Student’s Recital” International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011. [13] Nageswara Rao Thota, Srinivasa Kumar Devireddy, “Image Compression Using Discrete Cosine