Machine learning techniques to improve the results of Student Performance

Sk Kaja Mohiddin, P Satish Kumar, S Amrutha Mani Sai, M V B T Santhi

Abstract: Calculating the student performance is most widely done in many institutions. It is very important for every institution to collect the performance of the student based on marks secured. Previously classification and clustering also used to get the results. The dataset used in this paper is student dataset with all student details such as name, marks, address etc. In this paper various machine learning techniques are implemented and analysed the performance of the student. Results show the comparison of the proposed system.

Index Terms: Student performance, educational data mining, performance prediction.

I. INTRODUCTION

The upcoming of information improvement in various fields has lead the wide volumes of data gathering in distinctive affiliations like records, reports, archives, pictures, sound, accounts, sensible data and distinctive new data social events.. The information amassed from various applications require fitting framework for expelling picking up from enormous reports for better crucial expert. Knowledge discovery in databases (KDD), routinely called data mining, goes for the introduction of strong data from tremendous accumulations of information [1]. The basic segments of information mining are applying unmistakable methods and algorithms so as to find and consider occurrences set away information [2]. Data mining and learning disclosure applications have a rich obsession because of its essentialness in crucial activity and it has changed into a fundamental part in different affiliations. Data mining strategies have been brought into new fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation limits, and so forth. There are developing investigation premiums in utilizing information mining in direction. This new rising field, called Educational Data Mining, worries with making methods that find picking up from data beginning from edifying conditions [3]. Educational Data Mining utilizes different systems, for example, Decision Trees, Neural Networks, Naive Bayes, K-Nearest neighbor, and different others. Using these systems unique sorts of learning can be found, for instance, affiliation measures, groupings furthermore, pressing. The discovered data can be used for need concerning enrollment of understudies in a particular course, offense of standard classroom appearing, zone of uncalled for methodology used in online examination, ID of unconventional properties in the result sheets of the understudies, measure about students? performance, and so on.

Manuscript published on 30 March 2019.
***Correspondence Author(s)**

Sk Kaja Mohiddin, computer science engineering, KLEF, Vaddeswaram , India.
P Satish Kumar, computer science engineering, KLEF, Vaddeswaram , India.
S Amrutha Mani Sai, computer science engineering, KLEF ,Vaddeswaram, India
M V B T Santhi, Associate Professor, computer science engineering, KLEF Vaddeswaram, India.

II. LITERATURE SURVEY

Dorina Kabakchieva [4] did same research in which he utilized depiction models made by utilizing four data mining tallies – One R Rule Learner, Decision Tree, Neural Network and K-Nearest Neighbour. The most ridiculous accuracy is developed by utilizing neural structures looked for after by Decision Tree model and K-NN show up. The Neural system show works better with the "Strong" class while other three worked amazingly with "Weak" class. The after effects of the entire model are separated and others for equivalent characteristics and data set.

Baradwaj and Pal [5] proposed an examination on around 50 understudies who attempted the course for a time of 4 years with attributes like "understudies last semester grades", "reliably centers", "lab assignments", "support", "last semester marks, and so on. They brought into utilization of the ID3 Decision tree for social occasion of understudy's information and a brief timeframe later build up a choice tree for precise figure of understudy's execution. Their standard point was to concentrate on the fields in which the understudy was horrible and decline the slip-up degree. They brought into utilization of the ID3 Decision tree for looking at the understudy execution since it is the most direct learning calculation. Abeer and Elaraby [6] in addition completed a fitting examination to organize and predict scholarly execution of a get-together of individuals over a time span of 6 years with different of highlights amassed from educational affiliations. As a yield, they could foresee the appraisal of understudies in the specific course and even can improve his execution by taking preparing in fragile fields to keep away from disappointment degree. Amjad Abu Saa [7] examine accepted that student’s performance relies upon scholastics similarly as rely upon other individual, social and additional curricular exercises. He close by Naïve Bayes estimation utilized three choice tree calculations for game-plan of information.

Specifically off the bat he completed an examination and amassed understudies information and a brief timeframe later pre-processed and investigated the information for data mining tries. In like manner, the data mining estimations were executed on the educational record to convey gathering models for foreseeing understudy's execution.

The examination work wrapped up by Ali Daud and Farhat Abbas [8] presents the understudy enlightening want process that utilizes four different kinds of characteristics explicitly: family use, family pay, understudy particular data and family resources. It adjusts the philosophy for trademark subset affirmation so as to see the most fundamental highlights for understudy informational execution figure. It is clear from the relative examination that their recommended characteristics are doable markers and accomplished F1-score on genuine understudy's information. They finished from the outcomes that family usage and individual data qualities basically impact the execution of the understudy in light of regular reasons.

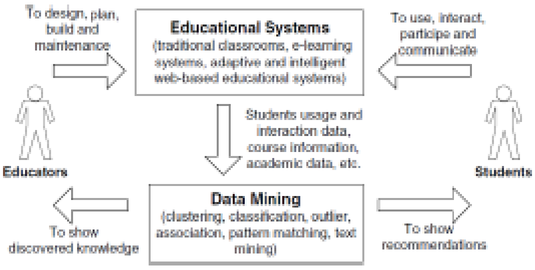

To design, plan, build and maintenance
Educational Systems (traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, particípe and communicate
Educators
Students usage and interaction data, course information, academic data, etc.
Students
Data Mining (clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Fig 1 the cycle of mining educational data

III. PROPOSED SYSTEM

A. Data Mining

Data mining is generally called the system of Discovery of Knowledge which suggests removing or mining information from tremendous gathering of data. It helps in choosing enchanting learning, for instance, anomalies, changes, affiliations, plans and basic structures from enormous volumes of information set away inside a couple of different kinds of databases as in data dispersion focuses or other information files open [9]. It's been noticeably used nowadays due to the availability of colossal volumes of data in electronic structure and there is a prerequisite for changing over such data into accommodating information and learning for immense applications. Decision Support, Artificial Intelligence, Machine Learning, Statistics and Database Systems and Business Management are a segment of the fields using its applications [10]. These procedures are used to chip away at enormous proportion of data for finding covered models and associations pleasing in essential initiative. While data mining and learning divulgence are commonly treated as same, data mining is extremely a vital bit of the information disclosure process. The all around requested strategy required for isolating information from data are showed up in Figure 2.

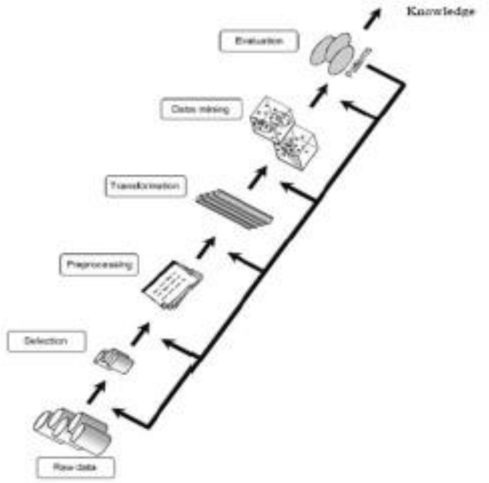


Fig 2. Various steps for mining useful information from data

B. Classification

Classification computation is a data mining methodology that makes us map into predefined grouping. It is a managed learning methodology which needs arranged planning data so it can making rules for characterizing test data into pre-sorted out order. [10] Its a 2 organize methodology. The fundamental stage as the learning stage, where the gathering rules are made and getting ready data is dismembered. The second stage as the game plan arrange, where test data is portrayed into predefined clusters as shown by the made rules. Since gathering counts requires predefined classes reliant on estimations of information section, we had made a section "execution" for all understudies, for which they may have an estimation of either "Extraordinary" or "Horrible"".

C. Clustering

Clustering algorithm generally means grouping certain course of components with the end goal that the components in a comparative cluster are more similar to one another than to those in various arrangements [9]. A couple of fields like pattern recognition, picture examination, machine learning and information recuperation insinuate this as a run of the mill strategy for authentic data examination. Clustering should be conceivable by a couple of systems that differentiate between the similar properties required between segments of a gathering and how to successfully find the segments of the clusters.

D. Classification over Clustering

In strategy of execution of clustering various classes can be found from the data and are examples of cloud apriori. As our major point is to separate understudies' execution into any of the predefined level - "Extraordinary" and "Bad", for which gathering was not reasonable, so we have used course of action method rather than clustering procedure.

E. Prediction of Results

Regression procedure can be used for looking at as regression examination can be used to show the association between one or more dependent and independent variables.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Unfortunately, various authentic issues are not only a prediction. A comparable model can be used while predicting results for both regression and classification. For instance, the CART (Classification and Regression Trees) decision tree figuring can be utilized to manufacture both course of action trees (to gather out and out response factors) and backslide trees (to assess endless response factors). Neural frameworks procedure can be used to make both classification and regression models.

F. Accuracy Measurement

Breaking down which technique for data mining is extraordinary altogether depend on how the customers has kept an eye on the issues. Generally every system execution is looked into by examining the precision of the results. Exactness estimation in Classification technique is done by choosing the dimension of set tuples morally justified class. Meanwhile, there might be some cost related with each off base undertaking to the wrong class which can be neglected.

IV. METHODOLOGY

We will describe algorithms that we have used to analyse student performance. Firstly, the dataset used for this research. Secondly, the attributes. Then about the algorithms linear regression, recursion and partion trees, support vector machine (linear and non linear).

V. EXPERIMENTAL RESULTS

A. Linear Regression

Linear regression is a key calculation and ordinarily used to imagine examination. This fall away from the faith measures are utilized to clear up the relationship between one ward variable and something like one independent segments.

B. Recursive Partitioning

Recursive partitioning wound up being especially worshiped and wide utilized tools for nonparametric regression and classification in several scientific fields. Particularly random forests that may alter gigantic measures of pointer factors even inside the closeness of front line affiliations are related with accomplishment in genetic science, clinical arrangement and bioinformatics at breaks the recent years. The motivation behind this work is to show the rules of the quality algorithmic dispensing ways still as late procedure upgrades, for instance their utilization for low and high dimensional data examination, regardless in like manner to indicate controls of the ways and potential catches in their application. Usage of the ways is depicted abuse clearly offered use inside the R structure for applied math computing.

C. Support Vector Machine (SVM)

The Support Vector Machine can be seen as a kernel machine. Thusly, you can change its direct by utilizing an other piece work. The most important kernel limits are

- the linear kernel
- the polynomial kernel
- the RBF (Gaussian) kernel
- the string kernel

D. Linear

The linear kernel is reliably prescribed for text classification. Content is regularly straightly indisputable. The majority of substance plan issues are straightly discernable. Linear kernel is indeed very well suited for substance strategy. Remember regardless that it isn't the standard strategy and for some condition utilizing another part may be better. The embraced framework for substance ask for is to attempt a straight piece first, in context on its extraordinary conditions. If however you search to get the best possible classification performance, it might be interesting to try the other kernels to see if they help.

E. Non-Linear Kernel

With what we have appeared of quite recently, data sets that are linearly separable(maybe with a few uncommon cases or some perplexity) are particularly managed. Regardless, what are we going to do if the educational record fundamentally doesn't permit depiction by a linear classifier? Enable us to take a gander at a one-dimensional case. The best instructive rundown is quick depicted by a straight classifier but the middle data set is not. We rather should probably choose a between time. One approach to manage this issue is to plot information on to a higher dimensional space and after that toutilize a linear classifier in the higher dimensional space. For instance, the base piece of the figure displays that a prompt separator can unquestionably depict the information on the off chance that we utilize a quadratic capacity to depict information into two estimations (a polar headings projection would be another acceptability). The general thought is to portray uncommon part space to some higher-dimensional segment space where the status set is specific. Obviously, we would need to do everything considered in propensities that protect material sections of relatedness between server farms, with the target that the resultant classifier should in any case total up well.

F. Decision Tree

It is a Supervised algorithm which initially reads and learn about the data and with the help of term a parameter e.g. Entropy we make decision about splitting of nodes and we finally construct a tree structure. And it will predict result for new sample. In the final tree structure each path represents a rule. We have different packages like Ctree, rpart etc. in R to implement decision tree.

G. Naïve Bayes

This technique is based on Bayesian theorem, in this we assume that there is no relation between the attributes that are present i.e. every attribute in the data set is independent. In this we use a mathematical concept called probability, with these probabilities we can predict the output. In this we use different probabilities like likelihood, prior probability etc.

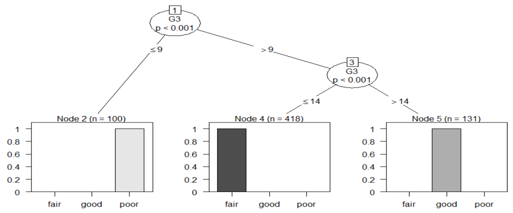
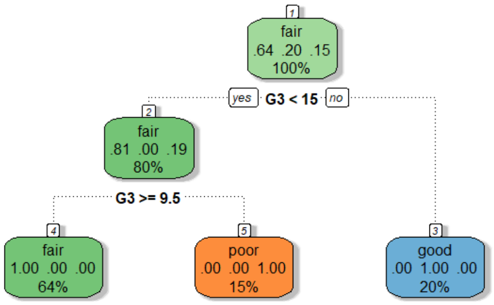
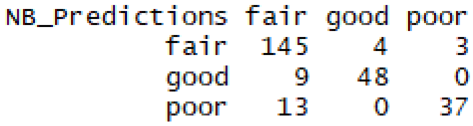


Fig 3. Decision Tree Final Tree Plot using ctree



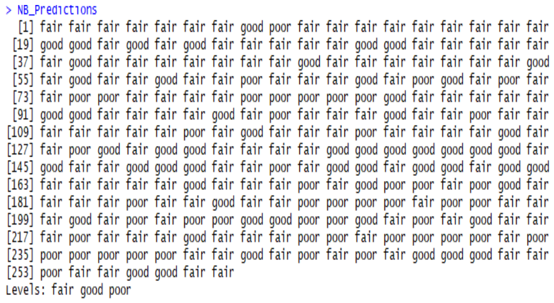
Rattle 2019-Jan-24 11:18:46 LENOVO

Fig 4. Decision Tree Final Tree using rpart



```
NB_Predictions fair good poor
          fair  145    4    3
          good    9   48    0
          poor   13    0   37
```

Fig 5. Prediction Table of Naïve Bayes



```
> NB_Predictions
  [1] fair fair fair fair fair fair fair good poor fair fair fair fair fair fair fair fair fair
 [19] good good fair good fair good fair fair fair fair fair good good fair fair fair fair fair
 [37] fair good fair fair fair fair fair fair fair good fair fair fair fair fair fair fair good
 [55] fair good fair fair good fair fair poor fair fair fair good fair poor good fair poor fair
 [73] fair poor poor fair fair fair fair poor poor poor poor poor good fair fair fair fair fair
 [91] good good fair fair fair fair good fair poor fair fair fair good fair fair poor fair fair
[109] fair fair fair fair fair poor fair good fair fair fair poor fair fair fair fair good fair
[127] fair poor good fair good good fair fair fair fair good good good good good good good fair
[145] good fair fair good good good fair fair poor fair good good fair good good fair good good
[163] fair fair fair fair fair good fair fair fair poor fair good poor poor fair poor good fair
[181] fair fair fair poor fair fair good fair fair poor poor poor poor fair poor poor fair fair
[199] fair good fair poor fair poor poor good good poor poor good fair poor fair good fair fair
[217] fair poor fair fair fair good fair fair fair poor poor fair poor poor poor poor fair poor
[235] poor poor poor poor poor fair fair good fair poor fair poor fair good good good fair fair
[253] poor fair fair good good fair fair
Levels: fair good poor
```

Fig 6. Naive Bayes Output

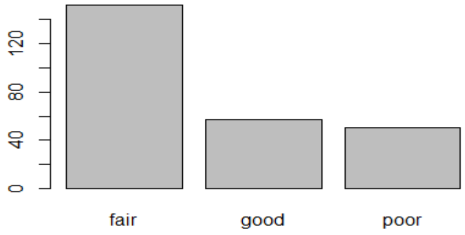


Fig 7. Plot for Naïve Bayes

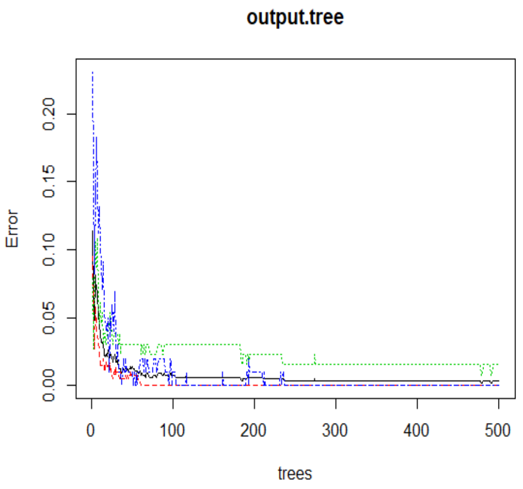
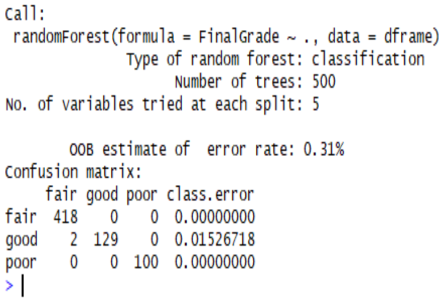


Fig 8. Random Forest Final performances



```
Call:
 randomForest(formula = FinalGrade ~ ., data = dframe)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 0.31%
Confusion matrix:
     fair good poor class.error
fair  418    0    0  0.00000000
good    2  129    0  0.01526718
poor    0    0  100  0.00000000
> |
```

Fig 9. Confusion Matrix for Random Forest

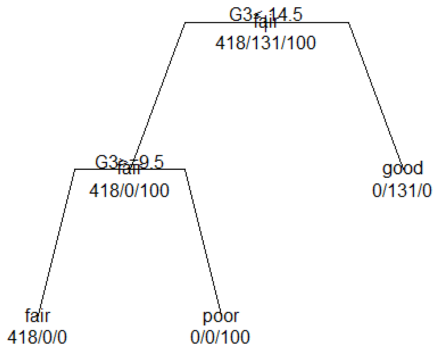


Fig 10. Recursion Partitioning Final Tree

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

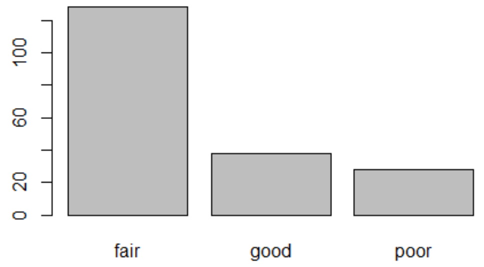


Fig 11. SVM Linear Final Plot

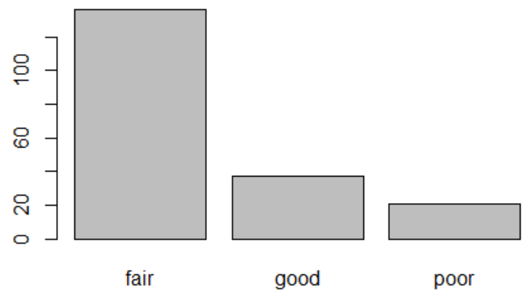


Fig 12. SVM Non-Linear Final Plot

VI. CONCLUSION

This paper explains about importance of utilizing machine learning systems for prediting students perfomace. Educational data mining is a rising field in research a region. It helps to analyze student information and give accuracy of various algorithms. In this paper, the performance based on the accuracy is shown. From the algorithms we used SVM linear has more accuracy. The SVM Linear shows the higher accuracy.

FUTURE SCOPE:

Not only the algorithms that used in this paper , there are good number of algorithms.so, this paper can further extended by using many other algorithms for suggesting algorithm which is more accurate than the algorithm proposed in this paper, this can also be extended for suggesting which algorithm is more effective in terms of space and time complexity**.**

REFERENCES

1. Heikki, Mannila, Data mining: machine learning, statistics, and databases, IEEE, 1996.
2. U. Fayadd, Piatesky, G. Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6, 1996.
3. J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann, 2000.
4. Kabakchieva, D., Stefanova, K., Kisimov, V. (2011). AnalyzingUniversity Data for Determining Student Profiles and Predicting Performance.Conference Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), 6-8 July 2011, Eindhoven, The Netherlands, pp.347-348.
5. Baradwaj, B.K. and Pal, S., 2011. Mining Educational Data to AnalyzeStudents’ Performance. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
6. Ahmed, A.B.E.D. and Elaraby, I.S., 2014. Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of
7. Computer Application and Technology, 2(2), pp.43-47.
8. Amjad Abu Saa, 2016. Educational Data Mining & Student’s Performance Prediction. International Journal of
9. Advanced Computer Science and Applications, Vol. 7, No. 5, 2016.
10. Ali Daud, Farhat Abbas, 2017. Predicting Student Performance using Advanced Learning Analytics. International World Wide Web Conference
11. Committee (IW3C2), published under Creative Commons, Pages 415-421.
12. Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques,Elsevier.
13. Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.

AUTHORS PROFILE:

Sk. Kaja Mohiddin, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. He is doing his research work in knowledge engineering.

P Satish Kumar, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram ,AndhraPradesh. He is doing his research work in knowledge engineering.

S Amrutha Mani Sai, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram ,AndhraPradesh. She is doing her research work in knowledge engineering.

M.V.B.T.Santhi received her B-Tech degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad, in 2003,M-Tech degree in Computer Science from Archarya Nagarjuna University, in 2010. She is an Associate Professor of Computer Science and Engineering Department of KL University. His research interest is Data Warehousing, Design and Analysis of Algorithms, DBMS, Big Data and Business Intelligence