Recommendation system to improve students performance using machine learning

P Ramya, S G Balakrishnan and M Kannan

Department of Computer Science and Engineering, Mahendra Engineering College, Namakkal, Tamil Nadu, India. 637 503

E-mail: balakrishnansg@mahendra.info

Abstract. For any educational institutes, students are a crucial asset so as to supply graduates of great quality. To achieve this, it is becoming essential for every college, school or any other scholastic institute to examine the performance of students. Due to the unbelievable development in recent technology like social media, it’s going to prevent the students from their concrete track, and this is often one among the explanations for the students to accomplish poor in academic accomplishments and it even results in course drop outs. Predicting student’s performance will attentive the learner to know about their performance and it provides as a gamble to increase their performance in future. However academic performance may differ from student to student as every student has diverse level of performance. It becomes difficult to research student data by relating statistical techniques or other outdated management tools. Hence there’s a requirement to develop an automatic machine learning tool for student performance analysis that might analyze student performance and can guide them by displaying the suggestions where they need improvement.

1. Introduction

Predicting pupil data to efficiently improve both teacher and learner higher in coaching and studying. It additionally improves conversation between management, teachers and learner and allows to hint student’s action on more than one degrees like class assignments, seminars, internal assessments and very last examinations. This system shaped with the aid of three fields. They’re computing, Psychology and schooling. Within the latest years academic institutes are more and more the use of educational systems to analyze their performance to make techniques for similarly development and future movements. these actions specialize in identifying and developing techniques that may enhance scholar academic performance, which indirectly help the institutes to recognize more new students and hold antique students, the algorithms worried those systems are termed as Educational Data Processing or EDM. One vital aspect of EDM is predicting scholar overall performance, that is that the most place of this studies paintings. Predicting scholar overall performance can also be a challenge which specializes in inferring know-how from pupil performance facts to understand The results of prediction of pupil performance is to enhance route effectiveness, when the quantity of anticipated screw ups are excessive, it’d be way to ineffective syllabus, which can be reassessed and stepped forward. it can provide correct collection of mastering activity records like time-on-undertaking and assessment rankings, which permits beneficial development estimates for each the student and for that reason the trainer. Early warning of „at threat students’” are regularly obtained that may help trainers to boom specialize in them so on decorate their overall performance. This affords higher capacity to expect pupil performance at an early enough level to probably find methods to growth the achievement fee. Consequently, prediction of pupil overall performance helps easy model, personalization and intervention.
Students get drop out of faculties due to numerous reasons considered one of them is that the negative overall performance of the students in academics which finally ends up in terrible grades. A manner to save you one of these state of affairs is by means of early predicting the final phrase grades of students within the center of their studies from the preceding effects. If that is probably executed it'll be beneficial for the lecturers additionally due to the fact the students for improving the academic performance. by means of reading the predictions teachers will get an usual view approximately the performance of a particular pupil and in what all areas she or he want to pay attention greater so on achieving good grades and students will get an common concept about their research. Facts processing is widely utilized in educational field to seek out the issues rise up at some stage in this field. Student performance is of extremely good concern in the academic institutes wherein numerous factors may additionally affect the overall performance. For prediction the 3 required additives are: Parameters which have an effect on the scholar performance, processing strategies and 1/3 one is processing tool. Those Parameters could even be psychological, personal, and environmental. We behavior this study to require care of the schooling best of institute by minimizing the various effect of those elements on scholar’s overall performance. Throughout this Paper, Prediction of scholar performance is finished by means of making use of Apriori type strategies WEKA tool. by applying processing techniques on scholar statistics, we'll obtain expertise which describes the student performance. it's drawbacks of modifications in element affecting the performance, there's a danger for getting inaccurate consequences and it consumes longer for processing the challenge.

There are some of functions from the winning structures which might be employed at some stage in the design and implementation segment of the proposed device. These functions and functionalities include the interface, students’ performance prediction, illustration displays and report technology. an honest interface provides an consumer-friendly interface because it's trustworthy to be navigate and no longer complex. Meanwhile, the scholars’ performance prediction is protected into the proposed system to make sure the objectives are accomplished. Furthermore, the technology of stories in Portable Document Format (PDF) and example display like charts in PDF makes scholar performance analysis easier. From these capabilities observed in proposed device, all of the user requirements might be fulfilled. Prediction of pupil performance prediction is based totally on various factors of scholar’s sort of a private, network, mental and environmental variables. At some stage in last few years many researches are administered to expect college students’ educational performance. So, in the course of this section, we are taking few research articles into attention then examine them for numerous student's factors which have an effect on the student academics prediction. Almost thirty-forty research papers, article, e book chapters are considered for assessment. Farhana Sarkar and Hugh C Davis [1] in his studies confirmed that the institutional internal data sources (IDS) and external data source (EDS) gave the most effective result than the model supported most effective institutional internal scholar databases.

In another study, Angeline D M D used inner evaluation take a look at grade, project submission and Grade, accurate reaction, Self-confidence, interest in the actual route and diploma ambition for prediction of scholar's educational overall performance [2]. Ambeer Bard El Din Ahmet al. in his observe used the route of the scholar, HSD, mid-term marks, Lab check grade, seminar performance, mission, attendance, homework, pupil participation for prediction student performance prediction [3]. Fadhili Ahmad and Azwai Abdul Aziz collected information from the database of tutorial branch, Unisa that saved in Informix management gadget. They further used 9 distinct parameters like gender, race and hometown, GPA, family profits, college access mode, grades Malay Language, English, and mathematics [4]. Mashaei A Al-Barak and Mona S Al-Raegan amassed dataset of students from the knowledge technology department at kinfolk Saud university, Saudi Arabia for his or her analysis [5].
2. Related Work

Edin Osman Begovic and Mirza Salic (2012) amassed information from surveys within the midst of first-12 month’s students and for this reason the information taken in the course of the enrollment at the college of Tuzla. They further used the varied attribute for the prediction like Gender, family, Distance, high school, GPA, entrance exam, Scholarships, Time, materials, the web, Grade significance, earnings [6]. Waheela Asif and Mahmood okay. Pathan (2014) in his observe they used four academic batches of computing & Information Technology (CS&IT) department at NED university, Pakistan. They used HSC marks, marks in MPC, Math’s marks in HSC, marks in various challenge studied within the ordinary direction of a programming language, CSA, good judgment layout, OOP, DBMS, ALP, FAM, sad, arrangement etc for their evaluation [7]. Mohammed M. Abu Tari and Alana M. El-Hales (2012) in his observe tried to extract a few useful information from student's statistics of technology and era college –Khan Younis. They to begin with decided on unique attributes like Gender, date of beginning, location of birth, area of expertise, Enrollment yr, commencement year, city, area, address, smartphone wide variety, HSC Marks, SSC college type, HSC acquired the location, HSC 12 months, university CGPA for analysis. However after pre-processing of the records they determined that attribute like Gender, strong point, metropolis, HSC Marks, SSC college kind, college CGPA are maximum substantial [8].

Azwai Abdul Aziz and H.I.F Ahmad (2014) used first-semester scholar information of Bachelor of computing from college Sultan ZainalAbidin (UnisZA) for evaluation. They used the attributes like Gender, race, hometown place, college entry Mode, circle of relatives income for records collection. ok.DKolo and J.okAlhassan (2015) accrued computing pupil's records of Nigerian colleges of training. In his take a look at, they taken into consideration association path of computing is one of the principal essential subjects and subsequently collect records respective to this subject. They taken into consideration student attributes like scholar's grade, scholar's reputation, college students gender, economic electricity, mindset to gaining knowledge of as vital elements for the prediction of student overall performance prediction. Jyoti Bansode (2016) for predicting pupil lecturers overall performance accumulated statistics from Shah and Anchor Kutchhi Polytechnic, Chembur, Mumbai. They taken into consideration student attributes like discern's education, parent’ s career, class, SSC board, admission kind, SSC medium, SSC class, first-semester result, 2nd-semester, 0.33-semester, forth-semestre, the 5th-semester and sixth-semester end result as maximum massive attributes. R. Sumitha and E.S. Vinoth Kumar (2016) for his research accumulated information of round 350, BE (CSE) college students of KLN college of know-how era. To start with, they selected 33 attributes for evaluation, but finally attributes with the top ranking are considered for the type reason. The selected attribute is CGPA, arrears, attendance, SSC marks, Engineering reduce-off, medium-of-schooling and type of Board. Mrinal Pandey and S. Taruna (2016) for this have a look at used datasets from an engineering institution.

They included the statistics associated with the student’s academics attributes additionally as their demographics records. Maria Goga, colorKuyoro, NicolaeGoga (2015) used student statistics from Babcock college, Nigeria. at the idea of reviewed literature, they taken into consideration age, gender, figure's fame, discern's qualification, parent's occupations, SSC score, HSC score, CGPA first 12 months. Maria Koutina and Katia LidaKermanidis (2011) they attempted to seek out out the best strategies for predicting the ultimate word grade of the postgraduate college students of Ionian university Informatics, Greece. Midhun Mohan et. al. they in particular used sort of techniques for the overall prediction of the scholars’ overall performance over a huge extent of records. Those techniques are getting to know Analytics and Predictive Analytics. Learning Analytics specially offers with the data collection and records pre-processing phase where the required facts for constructing the prediction model is amassed from the CBSE faculties, they've used MySQL server for storing the big amount of statistics. Then within the info pre-processing step like records cleaning, information
transformation etc. are finished via using apache HIVE framework. From the pre-processed statistics needed statistics's are located out using MapReduce set of rules inside the Hadoop framework. Then comes the Predictive analytics component in which the real predictions are made the usage of the couple of linear regression version. Madhav S. Vyas et. al. makes use of a choice tree version for predicting the academic overall performance of students.

The information wanted for constructing the model is gathered and facts pre-processing is finished wherein the continual values are converted to discrete values and as a result the null values are removed. Then by means of the use of CART algorithm to the records the selection tree prediction version is beingbuild. And therefore the scholars with poor overall performance are expected out. Alana M. et. al. here they may be mainly concentrating on two methods Clustering and regression. Clustering is used for records pre-processing and regression is used for developing the predictions. In clustering similar data are grouped collectively, though which the facts type is being performed and the use of the regression technique a version is being for doing the predictions. Huda Al-Sheri et. al. This work offered two prediction models for the estimation of pupil’s performance in final exam. The paintings made use of the popular dataset furnished by the college of Minho in Portugal, which relate to the overall performance in math situation and it consists of 395 statistics samples. Forecasting the overall performance of college students are often beneficial in taking early precautions, on the spot movements, or choosing a student that's healthy a particular challenge. the necessity to discover higher models to understand higher overall performance cannot be overemphasized. maximum of in advance work at the identical dataset used K-Nearest Neighbor set of rules and accomplished low effects, whilst assist Vector gadget set of rules changed into rarely used, which takes place to be a totally popular and powerful prediction method to shape certain higher assessment, we applied each aid Vector machine algorithm and K-Nearest Neighbour set of rules at the dataset to are expecting the student’s grade then as compared their accuracy. Empirical studies outcome indicated that aid Vector device finished slightly higher results with coefficient of correlation of 0.96, while the k-Nearest Neighbor performed coefficient of correlation of 0.95. David de la Peña et. al. focuses on a logistic regression version.

The collected information are preprocessed and data cleansing is accomplished after that a reference model is built for each of the route the usage of logistic regression and are saved in the database which are then used for category. Here the whole pupil details are being used to categorise into two instructions specifically dropout or non-dropout. Febrianti Widyahastuti et. al. In here they have completed a analysis of linear regression and multilayer perceptron models the usage of WEKA tool package. they've used undergraduate scholar data in information gadget management. The dataset includes 50 statistics information. The information are processed in WEKA device itself and the algorithms needed for the model production also are available inside the tool named WEKA.

### 2.1 Random Forest
Random forests or random choice forests are an ensemble mastering method for classification, regression and different duties that works through constructing a choice tree multitude at schooling time and outputting the elegance it's the mode of the classes (category) or imply prediction (regression) of the character trees.

### 2.2 Decision Tree
Decision tree are graph structures, where every capacity choice creates a brand new node, ensuing in a tree-like graph (Quinlan, 1987). Figure 1.indicates an example of a decision tree. This tree is used to are expecting if a scholar has handed the examination by using looking at GPA and age values. “Yes” and “No” in the edges suggest whether the “GPA > three point zero” and “Age > 15” conditions are met. In machine learning, decision trees partition the info set in appropriate values till a tree structure has emerged. This technique is referred to as recursive partitioning (Strobl, 2009). Decision tree algorithm attempts to searching for out the only way to partition the data in order that parts are as
homogeneous as possible. If a completely homogeneous element is impossible, extra commonplace fee is selected. This procedure is proven by way of developing a decision tree from scholar information. The decision tree algorithm determines the partition places and the quantity of partitions.

![Decision Tree](image)

### 2.3 Naive Bayes Classifier

On this specific phase, we will talk the most finding of our meta-evaluation. on this meta-evaluation, we find that basically used statistics mining algorithm for prediction is decision Tree (DT), Naive Bayes (NB), artificial Neural Networks (ANN), Rule-based (RB) and k-Nearest Neighbor (KNN). In decision tree set of rules the maximum and minimal accuracy for predicting student's academic overall performance are ninety nine point nine and sixty six point eight percentage respectively. To locate the maximum prediction accuracy MariaGoga, shade Kuyoro and NicolaeGoga used the mixture of student’s characteristic like own family, PEP, EES, quit of first session end result.

In Naive Bayes set of rules, the maximum and minimum accuracy for predicting student's instructional overall performance are a hundred and sixty three point three percentage respectively. Maria Koutinaet al. used the diverse combination of scholar's attribute like Gender, Age, felony status, quantity of youngsters, career, job associated with the laptop, Bachelor, any other grasp, forty six Literature Survey on student’s overall performance Prediction in training the usage of statistics processing techniques pc literacy, Bachelor in informatics for getting maximum accuracy. In rule-based totally set of rules, the maximum and minimal accuracy for predicting scholar's academic performance are 96.7% and 55.0% respectively. To find the maximum prediction accuracy Maria Goga et al. used a aggregate of scholar's characteristic like own family, PEP, EES, cease of first consultation result. In k-Nearest Neighbor algorithm the utmost and minimum accuracy for predicting scholar’s instructional performance are a hundred% and seventy four% respectively. In artificial Neural Networks (ANN) the maximum and minimum accuracy for predicting student’s educational performance are 89.8% and sixty seven point six percentage respectively. To find the maximum prediction accuracy Mashael A. Al-Barrak and Mona S. Al-Razgan used a combination of scholar's characteristic like first mid-term examination in their first-12 month's route.

### 3. System Implementation

#### 3.1 Dataset and Attribute Selection

Future selection isn't the same as dimensionality reduction. Each methods searching for to reduce the amount of attributes in the dataset, but a dimensionality reduction technique accomplish that with the aid of developing new combinations of attributes, while future selection strategies consist of and exclude attributes present inside the statistics without converting them. Future selection strategies useful resource you to your undertaking to make an correct predictive model. They assist you by using deciding on capabilities in order to provide you almost as exact or higher accuracy at the same time as requiring less facts. function selection techniques are frequently wont to pick out and take away unneeded, irrelevant and redundant attributes from records that don't make contributions to the accuracy of a predictive version or might also really lower the accuracy of the version.
3.2 Preprocessing
Records preprocessing is an crucial step in gadget learning due to the fact the exceptional of know-how and therefore the useful facts that allows you to be derived from it immediately impacts the energy of our version to study; consequently, it is extraordinarily crucial that we preprocess our facts before feeding it into our model. In any real-international dataset, there are always few null values. It doesn’t clearly depend whether it's far a regression type or another kind of problem, no version can deal with those NULL values on its personal so we need to intrude. first of all, we’d want to see whether we have null values in our dataset or not. We will do that the use of the isnull() method. There are numerous approaches for us to deal with this problem. The easiest way to remedy this trouble is by way of dropping the rows or columns that include null values.

3.3 Normalization
Database normalization is that the process of structuring a relational database according with a series of so-known as regular paperwork so as to cut back statistics redundancy and enhance data integrity. It was first proposed with the aid of Edgar F. Codd as an critical part of his relational version. Normalization involves organizing the columns (attributes) and tables (relations) of a database to ensure that their dependencies are nicely enforced by database integrity constraints.

4. Model Building
The machine learning knowledge of roadmap is filled with trial and mistakes. Engineers and scientists, who're novices at the concept, will continuously tweak and alter their algorithms and models. All through this system, challenges will arise, specifically with coping with records and determining the right model. when building a system learning model, it’s essential to take into account that real-world facts is imperfect, differing forms of knowledge require different tactics and tools, and there'll always be exchange-offs whilst figuring out the proper version.

4.1 Algorithm Comparison
By way of selecting the top of the line set of rules in your trouble depends on its capabilities like pace, forecast accuracy, schooling time, quantity of statistics required to train, how clean is it to put in force, how difficult is to clarify it to others because an outsized venture of facts scientist is to discuss and give an explanation for patterns and ML algorithms and most significantly if the algorithm solves your trouble proven in figure 2.

```
1. LR: 0.798214 (0.113516)
2. LDA: 0.748214 (0.135397)
3. KNN: 0.773214 (0.107455)
4. CART: 0.683929 (0.081851)
5. NB: 0.748214 (0.192568)
6. SVM:0.760714(0.102020)
```

Figure 2. Algorithm Comparison

4.2 Logistic Regression
Logistic regression is that the maximum well-known gadget mastering set of rules after linear regression. In a number of the way, linear regression and logistic regression are similar. However the
maximum crucial distinction lies in what they are used for. Linear regression algorithms are wont to are expecting/forecast values but logistic regression is hired for type duties. in case you're shaky on the standards of linear regression. There are many category obligations accomplished routinely through people. as an example, classifying whether or not an e mail may be a junk mail or not, classifying whether a tumour is malignant or benign, classifying whether an internet website online is fraudulent or no longer, etc. those are regular examples in which machine gaining knowledge of algorithms could make our lives heaps simpler. A really simple, rudimental and useful set of rules for type is that the logistic regression algorithm. Now, let’s take a deeper test out logistic regression.

![Figure 3: True Positive Rate](image)

Logistic regression algorithm additionally makes use of an equation with unbiased predictors to predict a worth. The expected price are often everywhere between bad infinity to high-quality infinity. We need the output of the algorithm to be elegance variable, i.e. 0-no, 1-yes. consequently, we're squashing the output of the equation into a spread of [0,1]. To squash the anticipated price among 0 and 1, we use the sigmoid feature that takes the output(z) of the linear equation and provide to the characteristic g(x) which returns a squashed fee h, the well worth h will dwell the range of zero to at least one. To recognize how sigmoid characteristic squashes the values inside the variety, allow’s visualize the graph of the plotting shown in determine 3. Logistic regression makes use of an equation due to the fact the illustration, significantly like linear regression. input values (x) are combined linearly the usage of weights or coefficient values (referred to because the Athens letter Beta) to predict an output value (y). A key difference from linear regression is that the output fee being modelled is a binary cost (0 or 1) instead of a numeric value.

Following is an example logistic regression equation:

\[ y = \frac{e^{b_0 + b_1*x}}{1 + e^{b_0 + b_1*x}} \]

Where y is that the anticipated output, b0 is that the prejudice or intercept time period and b1 is that the coefficient for the only input value (x). each column to your enter report has an associated b coefficient (a consistent actual price) that has got to be discovered out of your training statistics. The real representation of the model that you honestly might save in reminiscence or during a file are the coefficients inside the equation (the beta value or b’s). Univariate evaluation is that the only form of reading statistics. “Uni” way “one”, so in other phrases your statistics has just one variable. It would not have an effect on reasons or relationships (not like regression) and its predominant reason is to describe; it takes statistics, summarizes that facts and finds styles within the information. Multivariate evaluation can be a hard and fast of techniques used for analysis of understanding units that include pretty one variable, and consequently the techniques are particularly precious while working with correlated variables.
4.3 Model Training
Training sets are wont to suit and song your models. Take a look at sets is forgot as "unseen" records to evaluate your models.

1. You must constantly cut up your data before doing whatever.
2. This is the best thanks to get reliable estimates of your models’ overall performance.
3. After splitting your statistics, don’t contact your take a look at set till you’re capable of pick out your very last version.

Comparing vs. Training overall performance permits us to avoid overfitting. If the model plays alright on the training information but poorly at the take a look at facts, then it’s overfit.

4.4 Prediction
Traditionally, device mastering models have not covered insight into why or how they obtained an final results. This makes it tough to objectively give an explanation for the selections made and actions taken supported those models. Prediction causes avoid the “black box” syndrome via describing which characteristics, or characteristic variables, have the nice effect on a model’s outcomes shown in Figure 4 and Figure 5.
5. Conclusion

Education is a crucial element in our society. Business Intelligence/Data Mining techniques, which permit a excessive stage extraction of information from facts, provide interesting opportunities for the education area. Specifically, several studies have used Business Intelligence/Data Mining techniques to decorate the standard of education and enhance school resource management. In this paper, we’ve addressed the prediction of secondary student grades of two core classes (Mathematics) by means of the use of beyond college grades (first and 2nd durations), demographic, social and other faculty associated information. Three exceptional data mining goals (i.e. binary/five-stage classification and regression) and 4 datamining methods, i.e. decision trees, Random Forests, Neural Networks and help Vector Machines, had been examined. Also, distinct input choices (e.g. without or with past grades) were explored. The obtained results reveal that it is possible to achieve a high predictive accuracy, provided that the first and/or second school period grades are known. This confirms the pupil success may be very stricken by previous performances. nonetheless, an evaluation to information furnished by way of the best predictive models has shown that, in some cases, there are other applicable capabilities, along with: school related (e.g. number of absences, reason to choose college, extra instructional college aid), demographic (e.g. scholar’s age, discern’s job and education) and social variables.

Future Enhancement

Digitalizing the data by providing management dashboard to the professors and students to track the data more efficiently in order to achieve more accuracy. This helps the students to know their level of performance easily. And it will be more comfortable for the professors to maintain the students data in a centralized storage.

6. References

[1] Farhana Sarker, Thanassis Tiropanis and Hugh C Davis 2013 Students’ performance prediction byeprints.soton.ac.uk/353532/1/Students’ mark prediction model.pdf.
[2] Angeline D M D 2013 Association rule generation for student performance analysis using an apriori algorithm, The SIJ Transactions on Computer Science Engineering & its Applications (CSEA) 112 – 16.
[3] AberBadr El Din Ahmed and Ibrahim SayedElarabyy 2014 Data Mining: A prediction for student’s performance using classification method, World Journal of Computer Application and Technology 112 – 16.
[4] Fadhilah Ahmad, NurHafieza Ismail and Azwa Abdul Aziz 2015 The prediction of students’ academic performance using classification data mining techniques, Applied Mathematical Sciences 96415 - 6426.
[5] Mashael A Al-Barraky and Mona S Al-Razgan 2015 predicting students’ performance through classification: a case study, I.J. Intelligent Systems and Applied Information Technology. 2 167-175.
[6] Edin Osmanbegović and Mirza Suljic2012Data mining approach for predicting student performance, Economic Review – Journal of Economics and Business, X 1.
[7] Raheela Asif, Agathe MerceronMahmood and K Pathan 2015 Predicting student academic performance at degree level: A case study, I.J. Intelligent Systems and Applications0149-61.
[8] Mohammed M Abu Tair and Alaa M El-Halees, 2012 Mining educational data to improve students’ performance: a case study, International Journal of Information and Communication Technology Research, ISSN 2223-49852 2.