STUDENT PERFORMANCE PREDICTION SYSTEM

Vedang Gharat[1], Saurabh Rathod[2], Chandan Saw[3], Aditi Malkar[4]

[1]Student, Department of Information Technology, Universal College of Engineering, Vasai
[2]Student, Department of Information Technology, Universal College of Engineering, Vasai
[3]Student, Department of Information Technology, Universal College of Engineering, Vasai
[4]Assistant Professor, Department of Information Technology, Universal College of Engineering, Vasai

---***---

ABSTRACT: *The research tend to all grasp that once ten grade that have been got to pick out a selected stream to figure for our future however few students area unit unable are to determine the steam during which the factors that determine their academic fields in line with their share of ultimate examination. Range of student drop out is one of the parameters that will be wont to live student performance,*

Find an appropriate data processing rule. Data processing is to show data into helpful data. Data processing has been wide wanting to predict student's performance, and it that applied during this field typically referred to as academic data processing. Numerous data processing strategies like the Associative tree and call tree area unit is used. Classification and prediction play a vitally significant role and important role within the data processing. one in all the info mining algorithms used is that the C4.5 rule that is that the high data processing rule. This project is especially involved concerning the prediction of student's performance in examination in line with that a selected field is appointed to them. They need to find out in. Predicting the long run of a student could be a nice concern towards the upper education managements.

Keywords: **Data Processing, C4.5 Algorithm, Information Pre-Processing ,Data Mining , Information Gathering**

I. INTRODUCTION

Data Mining: Data Mining is the process of mechanically discovering helpful data in giant information repositories. data processing techniques area unit utilized to find important patterns and rules from giant quantities of knowledge that unknown. There are a unit several techniques, strategies, and rules in data processing to extract specific info. Recently, data processing has been widely applied in several applications, like client relationships, engineering, medical, marketing, education, etc. Classification and prediction area unit 2 sorts of information analysis those are wont to extract models describing necessary information categories or to predict future information trends

Process of knowledge Mining:

1. The information collected is held on within the data warehouses
2. The info that is loaded inside the info warehouses has then dedicated a server or a cloud
3. Analysts, management, and IT skilled decide or confirm that however will they organize the info.
4. It kinds and organizes the info as per the user's result.
5. the tip user reports the info within the kind of visual or written format

The four stages of knowledge mining:

1. Information Gathering: information Gathering, additionally referred to as information assortment. Information assortment worries with the correct acquisition of data; though strategies could disagree reckoning on the sphere, the stress on making certain accuracy remains similar. The first goal of any information assortment endeavour is to capture quality information or proof that simply interprets to wealthy information analysis which will result in credible and conclusive answers to queries that are displayed. correct information assortment is critical to confirm the integrity of the analysis, no matter the sphere of study or information preference

2. Information pre-processing: the information pre-processing is that the vital method within the data processing. The information Pre-Processing could be a method wherever the info is encoded to bring it in such a state wherever the machine will simply analyse or are aware of it. This consists of knowledge Cleansing, information Integration, information Transformation, information reduction, information Discretization. information Pre-Processing is that the methodology that's evidenced to resolve problems.

3. Classification: Classification is that the thanks to organizing and kind the info in numerous or distinct categories. it's a method of finding a model that describes and distinguishes information categories and ideas. Initially, a model is formed that's supported the info distribution. The model created is classified as new information. A category will predict for brand new information. Classification is employed for nominal and separate values.

For Example, before functioning on a project, we'd like to ascertain whether or not it's 'Risky' or 'Safe' to more approve it.

Classifiers are categorized into 2 completely different parts:

a. Discriminative: Discriminative is that the basic categorified that is employed to see only one class in one row of knowledge. It doesn't think about distributions rather it focuses on quality of knowledge.
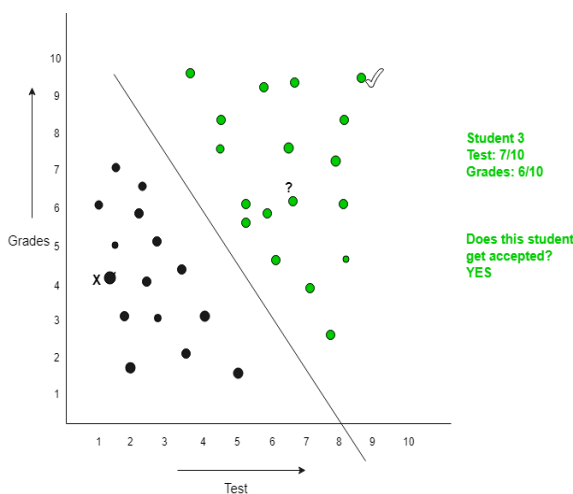
Example: supply Regression

Acceptance of a student at a University (Test and Grades got to be considered)
Suppose their area unit few students and therefore the results of them area unit as follows:

Student 1: check Score: 9/10, Grades: 8/10 Result: Accepted
Student 2: check Score: 3/10, Grades: 4/10, Result: Rejected
Student 3: check Score: 7/10, Grades: 6/10, Result: to be tested

b. Generative: This model supports the distribution of individual categories and tries to find out the model that helps to get the info behind the scenes by estimating assumptions and distributions of the model. want to predict the unseen information.

Example: Naive mathematician Classifier

4. Prediction: Same as classification, a model is formed supported information distribution. This model is employed to predict future or unknown values. Prediction in data processing is to spot information points strictly on the outline of another connected information worth. it's not essentially associated with future events however the used variables area unit unknown. Prediction derives the connection between a

factor you recognize and a factor you would like to predict for future reference.

Data Analysis task is an example of numeric prediction wherever the model constructs continuous worth operate or ordered value. The model is a predictor

For example, in MasterCard fraud detection, a history of knowledge for a selected person's MasterCard usage has got to be analysed. If any abnormal pattern was detected, then it ought to be reported as 'fraudulent action'.

II. SYSTEM ARCHITECTURE
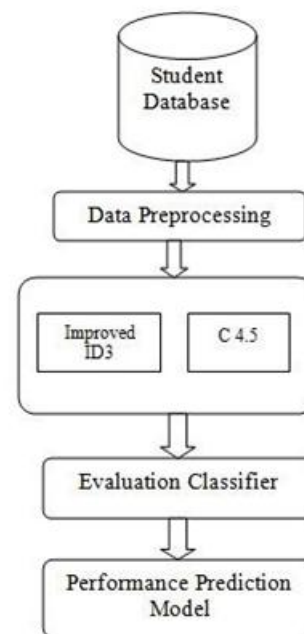
The database of the scholars of the college, created for examining their score and predicting the academic fields they'll be admitted into. Experimentation is going to be done employing coaching info. The coaching info consists of names and marks of scholars.

Data pre-processing: information pre-processing could be a data processing technique that's accustomed to rework the information during a helpful and economical format.

a. **Data Cleansing**: the information cleansing is that the technique wherever the raw data to be born-again into the knowledge must be analysed and smart data is unbroken and unhealthy data is eliminated. This offers the user the power to search out incomplete and inaccurate information. Inaccurate records from database or a dataset is recognised as unfinished, less reliable, inaccurate or non-relevant components

of the information and so restoring, remodelling, and remove the dirty or crude information.

b. **Data Integration**: information integration could be an information pre-processing technique that mixes information from multiple heterogeneous information sources into a coherent information store and provides a unified read of the info. Information Integration could be a method wherever information sets, files, or information cubes square measure integrated.

There square measure chiefly two major approaches for information integration – one is "tight coupling approach" and another is that the "loose coupling approach".

c. **Data Transformation**: This step is taken to rework the info into acceptable forms appropriate for the mining method. This involves the subsequent ways:
1. Normalization: it's done to scale the info values during such vary (-1.0 to 1.0 or 0.0 to 1.0)
2. Attribute Selection: From the given set of attributes, new attributes square measure generated or created

3. Concept Hierarchy Generation: Here attributes square measure born-again from level to higher level within the hierarchy.

d. **Data Reduction**: data processing could be a technique that's accustomed to handling an oversized quantity of knowledge. Whereas handling an oversized quantity of knowledge, analysing becomes troublesome. To induce eliminate this research tend to use information Reduction. It aims to extend storage potency and cut back information storage and analysis prices.

The various steps to information reduction are:

1. **Data Cube Aggregation**: Aggregation operation is employed for information to construct the info cube.

2. **Attribute set Selection**: The necessary Attributes ought to be used and therefore the remaining attributes ought to be discarded. For activity attribute choice, one will use the tier of significance and p-value of the attribute. The attribute having a p-value bigger than the importance level is discarded.

3. **Numerosity Reduction**: Numerosity Reduction permits to store the model of information

rather than whole data, as an example, Regression Models.

4. **Dimensionality Reduction**: victimization cryptography mechanisms and cut back the dimension's information. It is lossy or lossless. If once reconstruction from compressed information, the original information is retrieved, such reduction referred to as termed lossless reduction else it's called lossy reduction. The 2 effective ways of spatial property reduction are:

   a. Wavelet transforms
   b. PCA (Principal Element Analysis).

e. **Data Discretization**: information discretization could be a method of changing continuous information attributes values into a finite set of intervals with tokenism loss of knowledge. Discretization is that the method of swing values into buckets so their square measure a restricted range of potential states. The buckets themselves square measure treated as ordered and separate values. You'll discretize each numeric and string column. There square measure many ways that you just will use to discretize information.
Discretization is performed quickly on associate degree attributes to produce a graded partitioning of the attribute values, called a planning hierarchy.

   1. **top-down discretization**
   If the method starts by 1st finding one or many points (called split points or cut points) to separate the whole attribute varies, and so repeats this recursively on the ensuing intervals, then it's known as top-down discretization or ripping.

   2. **Bottom-up discretization**
   If the method starts by considering all of the continual values as potential split-points, removes some by merging neighbourhood values to create intervals, then it's known as bottom-up discretization or merging.

III. ALGORITHMS

C4.5: the C4.5 formula employed in data processing could be a call Tree Classifier which might be used to get a call, supported an exact sample of knowledge. In short, the C4.5 is employed to get a call tree classifier. If you have the knowledge that may accurately predict the longer term, then the knowledge concerning the event that you have been expected isn't new. But, if matters go South associate degreed a surprising outcome happens, it counts as helpful and necessary info.

C4.5 formula general operating steps square measure as follows.

Assume all samples within the list work into constant class. If the condition is true, it will produce a leaf node of the choice tree to pick out a selected category.

None of the options offers any info gain C4.5 creates a call node to a higher place the tree victimization the first moment of the category.

An instance of a previously-unseen category is encountered and so C4.5 creates a call node to a higher place the tree victimization the first moment.

It can say that the quantity of knowledge gained is reciprocally proportional to the chance of the event happening. Therefore, it is like because the Entropy will increase, the knowledge decreases. This is often as a result of Entropy refers to the chance of an occasion.

Say the research tend to square measure staring at a coin toss. The chance of expecting any aspect of a good coin is five hundredths. If the coin is unfair such either the chance of feat a HEAD or TAIL is one.00 then research will predict that the entropy is minimum because of none type of trials, algorithm will predict the result of the coin toss.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) = -\sum_{i=1}^{2} (1/2) \log_2 (1/2) = -\sum_{i=1}^{2} (1/2) \times (-1) = 1$$

In the aforethought graph below, the research, tend to notice that the utmost quantity of knowledge gained because of maximized uncertainty of a selected event is once the chance of each of the events is equal. Here, p=q=0.5

E = entropy of the system event

p = chance of HEAD as associate degree outcome

q = chance of TAIL as associate degree outcome

Entropy for a 2-states variable with state probability p and q with p+q=1
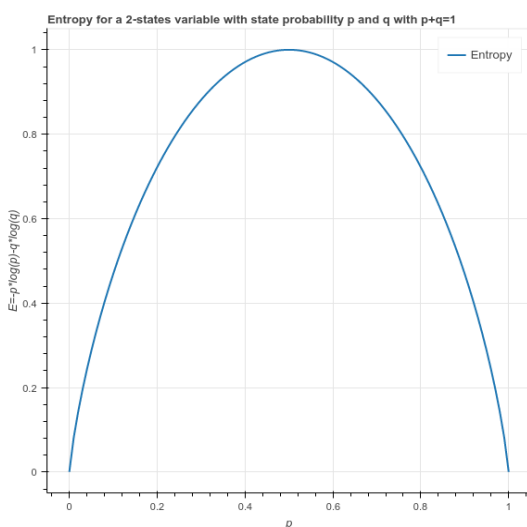
Advantages of C4.5 algorithm:

1. It employs Single Pass Pruning Process to Mitigate overfitting.
2. It works with Continuous and Discrete data.
3. C4.5 can also handle the issues of incomplete data very well.
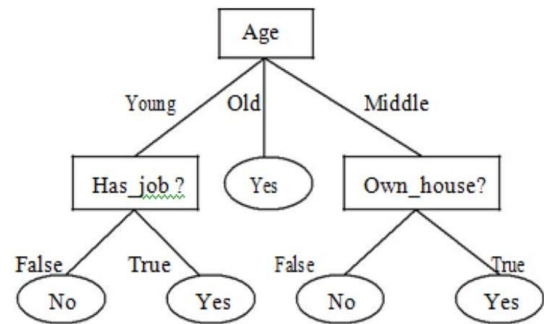
IV. CONCLUSION

The proposed system predicts the result of the students based on their current and previous performance. Overall, the student performance analysis system is proposed using data mining technique of classification to predict the performance of current students using C4.5 Algorithm. The result showed that multilayer perceptron gives good prediction results than linear regression. The results of the data mining algorithms for the classification of the students based on the attributes selected reveals that the prediction rates are not uniform.

V. REFERENCES

1] H. Aldowah, H. Al-Samarraie and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis", *Telematics Informat.*, vol. 37, pp. 13-49, Apr. 2019

[2] "Student's Performance Prediction Using Weighted Modified ID3 Algorithm" - International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882   Volume 4, Issue 5, May 2015.

[3] N. Ruijs and H. Oosterbeek, "School choice in Amsterdam: Which schools are chosen when school choice is free?", *Educ. Finance Policy*, vol. 14, no. 1, pp. 1-30, Jan. 2019

4] D. Gaevi, S. Dawson, T. Rogers, and D. Gasevic, learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success, The Internet and Higher Education, 2016, vol. 28, pp. 68-84.

[5] S-H. Lin, Data mining for student retention management, Journal of Computing Sciences in Colleges, 2012, vol. 27, no. 4, pp. 92-99.

[6] M. Saqr, U. Fors and M. Tedre, "How learning analytics can early predictunder-achieving students in a blended medical education course", *Med.Teacher*, vol. 39, no. 7, pp. 757-767, 2017

[7] R. Sumitha, E.S. Vinothkumar "Prediction of Students Outcome Using Data Mining Techniques" International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue6, June 2016

[8] Zhibing Liu, Huixia Wang, Hui Zan "Design and implementation of student information management system." 2010 International symposium on intelligence information processing and trusted computing. 978-0-7695- 4196-9/10 IEEE

[9] S.R. Bharamagoudar, Geeta R.B., S.G. Totad, "Web Based Student Information Management System", International Journal of Advanced Research in Computer and Communication EngineeringVol.2, Issue6, June2013