Analysing the Process of Categorization to Anticipate Improvements in Performance through Data Mining

Minakshi Kandari[1], Pawar Prem[2], Jaiswar Gaurav[3]
Asst. Professor[1] and FYBMS[2,3]
Uttar Bhartiya Sangh's Mahendra Pratap Sharda Prasad Singh College of Commerce & Science, Mumbai, Maharashtra

Abstract: *The volume of data contained in instructional datasets is growing fast in today's environment. These data sets contain information related to the performance and progress of students. The presentation of higher education in India is a pivotal moment in academics for all students. This academic exhibition is influenced by various circumstances. Therefore, it is crucial to develop a proactive information gathering strategy for students' presentations in order to differentiate between high-achieving students and underperforming students.*

Keywords: Data Mining, Educational Data Mining, Predictive Model, Classification

I. INTRODUCTION

In educational contexts, the capacity to foresee a student's performance is crucial. Various factors, including individual, social, mental, and environmental components, have an impact on students' academic performance. Data Mining is a highly promising approach for reaching this goal. Data mining techniques are utilized to analyze vast quantities of data in order to uncover distinctive patterns and relationships that are valuable for autonomous guidance. Arrangement is a proactive method of extracting valuable information by predicting the advantages of data based on real outcomes obtained from various sources. Classification is the process of categorizing data into predefined categories or classifications. Controlled learning is commonly used to describe a situation where the data is not fully resolved before analyzing it. The instructor should provide additional support to the exceptional students in order to enhance their performance in the future. The present study was designed with the objective of assisting underperforming students in higher education by establishing the following goals:

- Creation of a vast source of knowledge containing predictive attributes.
- Verification of the constructed model for prospective higher education students intending to enrol in Indian universities or institutions.
- Identification of several factors that impact a student's learning behavior and performance throughout their academic journey.

II. CONTEXT AND PRIOR RESEARCH

Alaa's observations suggest that Information Mining can be applied in the field of education to enhance our comprehension of learning interactions. This is achieved by specifically identifying, eliminating, and analyzing features associated with a student's learning system. This field of study is commonly referred to as Educational Data Mining. Han and Kamber discuss the functionality of information mining software, which enables users to analyze data from various angles, categorize it, and summarize the links identified during the mining process. Pandey and Pal conducted an assessment of student performance by selecting 600 students from multiple universities affiliated with Rd. R. M. L. Awadh University, located in Faizabad, India. It was determined whether or not new students would perform using Bayes Classification based on their class, language, and background proficiency. The hypothesis said that the performance of the understudy is connected with several factors, including their attitude towards class involvement, the amount of time they spend on review consistently after school, their family income, their mother's age, and their mother's education. The basic direct relapse evaluation revealed a significant association between characteristics such as the mother's level of education and the student's household income with the academic performance of the student.

The data mining process

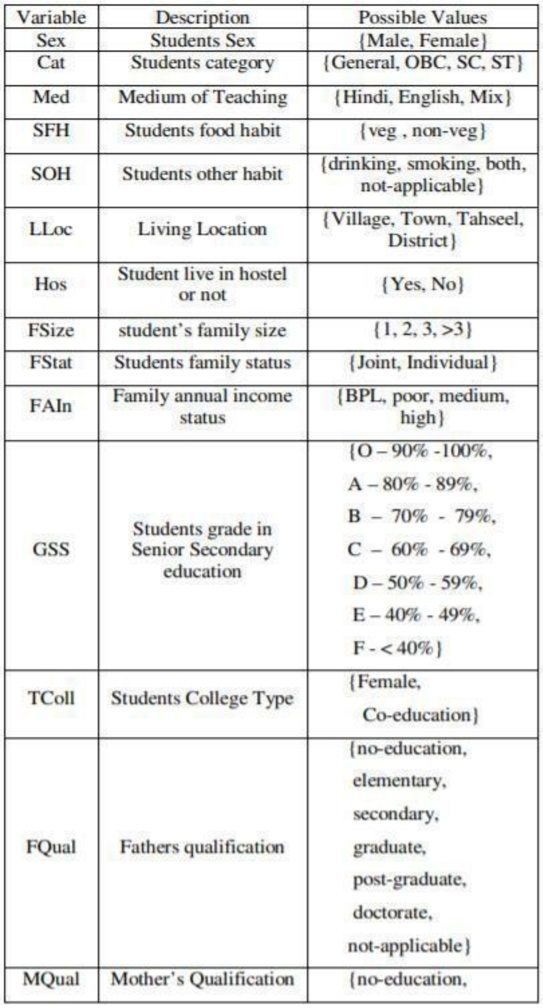
This research collected data from multiple prestigious universities and organizations that partnered with Rd. R. M. L. Awadh University in Faizabad, India. The data is analyzed using an orderly technique to predict the performance of the student. The following advancements are implemented in order to apply this procedure:

Preparation of Data

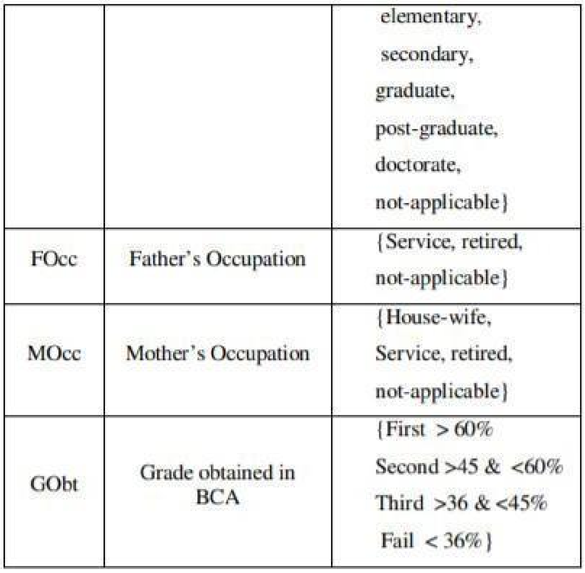
The data used in this evaluation were gathered from several schools on the examination procedure for PC Applications division obviously BCA (Bachelor of Computer Applications) of meeting 2009-10. The initial information size is 290. In this step, information from several tables was combined into a single table, and errors in the joining process were removed.

Transformations and data selection

Only the fields required for information mining were picked in this process. A few specific factors were considered. While some of the data on the factors was deleted from the database. Table 1 contains a list of all the indicator and response components obtained from the data set.



Variable	Description	Possible Values
Sex	Students Sex	{Male, Female}
Cat	Students category	{General, OBC, SC, ST}
Med	Medium of Teaching	{Hindi, English, Mix}
SFH	Students food habit	{veg , non-veg}
SOH	Students other habit	{drinking, smoking, both, not-applicable}
LLoc	Living Location	{Village, Town, Tahseel, District}
Hos	Student live in hostel or not	{Yes, No}
FSize	student's family size	{1, 2, 3, >3}
FStat	Students family status	{Joint, Individual}
FAIn	Family annual income status	{BPL, poor, medium, high}
GSS	Students grade in Senior Secondary education	{O – 90% -100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%}
TColl	Students College Type	{Female, Co-education}
FQual	Fathers qualification	{no-education, elementary, secondary, graduate, post-graduate, doctorate, not-applicable}
MQual	Mother's Qualification	{no-education,



		elementary, secondary, graduate, post-graduate, doctorate, not-applicable}
FOcc	Father's Occupation	{Service, retired, not-applicable}
MOcc	Mother's Occupation	{House-wife, Service, retired, not-applicable}
GObt	Grade obtained in BCA	{First > 60% Second >45 & <60% Third >36 & <45% Fail < 36% }

Fig 1 Student related variables.

The following are the domain values for some of the variables used in this study:

Drug - This report focuses solely on the degree universities and businesses in India's Uttar Pradesh area. The method of guidelines is either Hindi or English or a mix of both (Both Hindi and English).

Got - Marks/Grade obtained in a BCA course and announced as a response variable. It is also divided into five class esteems: First - >60%, Second - >45%, Third - 36% and 45%, and Fail - 40%. SOH - In today's culture, undesirable idiosyncrasies are rapidly spreading among college students. Understudies' additional propensities include drinking, smoking, both, or being inappropriate.

SOH - In today's culture, undesirable idiosyncrasies are rapidly spreading among college students. Understudies' additional propensities include drinking, smoking, both, or being inappropriate.

GSS - A student's grade in Senior Secondary School. Students in state board show up for five topics, each with 100 impressions. Grades are assigned to all students based on the following criteria: O - 90% to 100%, A - 80% - 89%, B - 70% - 79%, C - 60% - 69%, D - 50% - 59%, E - 40% - 49%, and F - 40%.

Size-. According to India's population statistics, the average number of children in a family is 3.1. As a result, the maximum family size is set at ten, and the possible range of attributes is one to ten.

Application of Mining Models

For information disclosure from data sets, various computations and processes such as Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour technique, and so on are used.

Order is one of the most commonly focused on challenges by data mining and AI (ML) professionals. It entails predicting the worth of a (global) attribute (the class) based on the benefits of several qualities (the foreseeing credits). There are several grouping techniques. The Bayesian Classification computation is used in this review.

Bayes order has been proposed, which is based on the Bayes rule of contingent likelihood. The Bayes rule is a method for determining the likelihood of a property given the arrangement of information as proof or information. The Bayes rule, often known as the Bayes hypothesis, is

$$P(h_i \mid x_i) = \frac{P(x_i \mid h_i)P(h_i)}{P(x_i \mid h_i) + P(x_i \mid h_2)P(h_2)}$$

The approach is labelled "innocent" since it anticipates independence between different property estimations. The credulous Bayes arrangement is both a separate and predictive type of computation. The probabilities are computed, and they are then used to forecast class enrolment for an objective tuple. The gullible Bayes technique has a few advantages: It is simple to use; unlike other order moves, just one sweep of the preparation information is necessary; efficiently manage mining esteem by simply dismissing that possibility

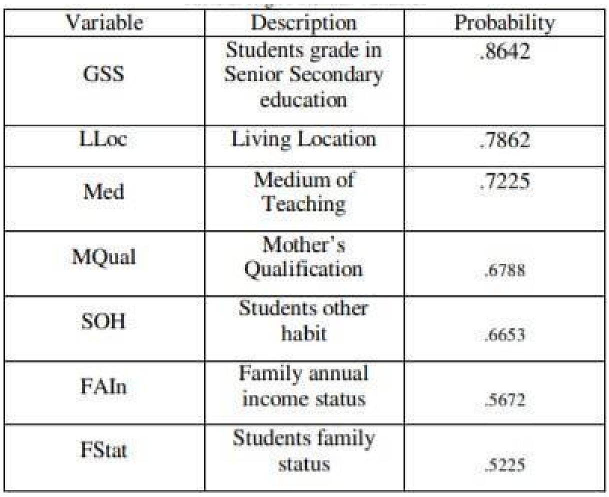
The guileless Bayes classifier has the advantage of requiring a small amount of preparation information to evaluate the boundaries (means and changes of the components) required for arrangement. Since autonomous factors are recognised, only the fluctuations of the factors for each class remain uncertain, rather than the entire covariance grid. Regardless of their guileless design and obviously erroneous suspicions, gullible Bayes classifiers have performed excellently in a variety of mind-boggling verifiable conditions. We picked five-degree universities affiliated with Rd. R. M. L. Awadh University, Faizabad, UP, India, for the present review. Two of the five-degree institutions were metropolitan-based, independent, and co-instructive, one was rural-based, assisted, and female, and the other two were provincial-based, supported, and co-instructive. The instances for our study were 300 BCA course understudies (226 men, 74 women) from these five colleges who appeared in the 2010 assessment. All data linked with understudy section, academic and budgetary elements was obtained directly from the 300 understudies via survey and University information base. These understudies' imprints were obtained from the University Examination cell. The credulous Bayes computation, given a preparation set, first estimates the earlier likelihood P (Ch) for each class by counting how frequently each class occurs in the preparation material. To determine P, each quality worth xi may be built up (xi). The probability P (xi | Ch) can also be calculated by counting how frequently each value occurs in the class in the preparation information. The restricted and earlier probabilities generated from the preparation set are used to create the expectation when describing an objective tuple. At that moment, multiply P (it | Ch) by to calculate P (it), we can assess the likelihood that it belongs to each class. The contingent probabilities for each characteristic esteem result in the possibility that it belongs to a class. The class with the highest probability is chosen for the tuple.

$$P(t_i \mid c_j) = \prod_{k=1}^{p} (x_{ij} \mid c_j)$$

To design the understudy execution forecast model, the present study used information mining as an apparatus and guileless Bayes order computation as a process. The separated element choosing technique was used to select the optimal subset of factors based on the probabilistic upsides.

III. CONCLUSION

In the current evaluation, those factors with likely esteems more than 0.50 were given careful consideration, and the most influential elements with high likelihood esteems were displayed. These highlights were used to build forecast models. MATLAB was used for variable determination as well as forecast model construction.



Variable	Description	Probability
GSS	Students grade in Senior Secondary education	.8642
LLoc	Living Location	.7862
Med	Medium of Teaching	.7225
MQual	Mother's Qualification	.6788
SOH	Students other habit	.6653
FAIn	Family annual income status	.5672
FStat	Students family status	.5225

Fig 3 high potential variables

It has been shown that pupils' performance is significantly reliant on their grade in the Senior Secondary Examination.

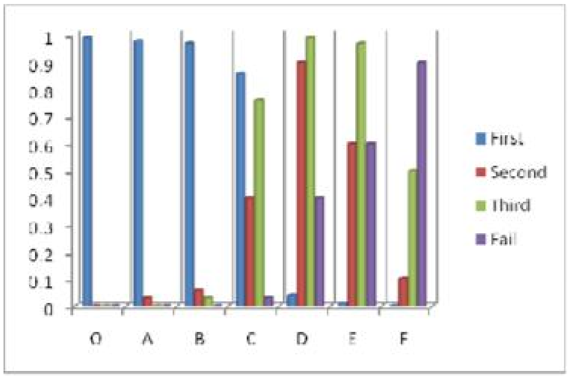


Fig 2: Relationship between GSS and Got

The medium of instruction is discovered to be the third high potential variable for student achievement. The mother tongue of students in Uttar Pradesh is Hindi. Students are more at ease in Mixed and Hindi languages than in English. The association between students' medium of instruction and their BCA test grade.

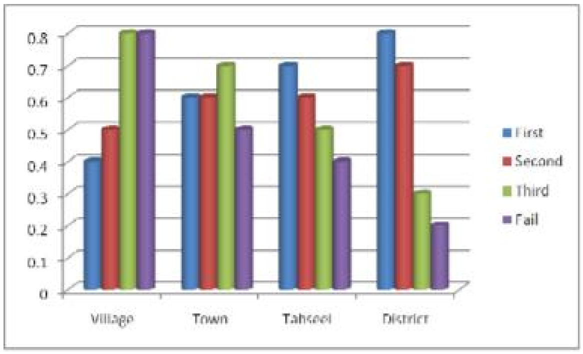


Fig 2: Relationship between LLC and Got

In this research, a Bayesian arrangement approach is used on an understudy data set to forecast the understudy division based on previous year data. This review will help the understudies and instructors work on the understudy division. This evaluation will also seek to separate those understudies who required special treatment in terms of decreasing bombing allocation and making the appropriate move at the right moment. The current research demonstrates that understudies' academic exhibits do not always rely on their own labour. Our investigation reveals that several aspects have a significant influence on understudy' performance. This offer will build on existing techniques by using pieces of expertise.

REFERENCES

[1]. AI-Radande. A., AI-Sawka, E.M., and AI- Najjar, M. I., “Mining Student Data using Decision Trees”, International Arab Conference on Information Technology (ACIT'2006), Alaa tell-tales, “Mining Students Data to Analyse e- Learning Behaviour: A Case Study”, 2009.

[2]. Bray, M. The Shadow Education System: Private Tutoring and Its Implications for Planners, (2nd ed.), UNESCO, PARIS, France, 2007.

[3]. David Hand, Heikki, Manni Padraic smith, “Principles of Data Mining” PHI

[4]. Galit.et.al, “Examining online learning processes based on log files analysis: a case study”. Research, Reflection and Innovations in Integrating ICT in Education 2007.

[5]. Hamm. and Kamber, M., "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, 2006.