Journal of Harbin Engineering University Vol 44 No. 7
ISSN: 1006-7043 July 2023

The Impact of the Non-Intellectual Factors on Students' Educational Performance Using Data Mining Techniques.

Pandya Vishal Kishorchandra[1], Prof. (Dr.) Pandya Rajnikant A[2]
1 Assistant Professor, Shri V J Modha College of I.T, Porbandar, Gujarat, India, vkpandya26@gmail.com
2 Assistant Professor, National Forensic Sciences University, Gandhinagar, Gujarat, India, rapandya@sauuni.ac.in

Abstract – This study has been incorporated to investigate the data mining techniques that is regression analysis, here. In this study researcher has applied multiple regression. For this study, researcher had prepared a structured questionnaire for collecting data from students of class XI to make predict the educational performance of students of class X. Here, the researcher has basically seven hypotheses and three objectives that is to identified impact of factors, to know about importance of dependent and independent factors**,** to measure the impact of factors on the educational performance of the students. By applying multiple regression researcher comes to know that there is significant impact of independent variables on the dependent variable that is score or the educational performance of the students. A regression applied by researcher, Formula: $y = m_{x1} + m_{x2} + m_{x3} + m_{x4} + m_{x5} + m_{x6} + m_{x7} + b$ after applying formula researcher comes to know some positive impact of independent variables on the performance and some positive impact of independent variables on the educational performance of the students.

Key words – variables, education, educational performance, data mining, dependent, independent

Overview To Data Mining Techniques

Data miners apply cutting-edge data mining tools in order to find relevant patterns and connections in big data sets that were not previously discovered. These programs are able to mix The fields of statistical modeling and machine learning make use of a variety of mathematical approaches, including neural networks and decision trees. The fact that this is possible illustrates that data mining is not only capable of prediction but also analysis.

One of the many approaches that may be used to analyze data, data mining is only one of them. The mining of classification schemes and association rules are two of the most useful activities and methods associated with data mining. These methods, which include examining the actions of users, give insightful and perhaps valuable insights. In India's higher education institutions, faculty members and administrators have begun adopting data mining techniques in recent years to assist them in forecasting the academic achievement of their students. These approaches were developed by academics.

In this study, the educational performance of the students will be forecast during the school course itself. It will be helpful to identify slow learners to whom additional academic care is to be given. It will also be helpful to encourage moderate and effective learners for their better academic performance. (Shabnaz,2021) [VI] , (Noateng, 2020) [VIII]

Professionals in the area of data mining have committed their whole careers to mastering the skills necessary to analyze massive volumes of data and derive meaningful insights from them. They do this by utilizing a wide range of techniques and tools that lie at the intersection of database administration, statistical analysis, and automated learning technology.

Just a few of the many important data mining methods that have been created in recent years include association, classification, clustering, prediction, sequential patterns, and regression. In recent years, many more important data mining methods have also been developed. developed and put to use in modern data mining endeavors. (K.K.nivethithaa, 2021)[i], (raghunandan, 2011)[iv], (R, 2018)[ii], (R, 2018)[iii]

1. Data Minig Techniques

1. Classification:

It is possible to gain useful insights into the data and the metadata by utilizing this strategy. The frameworks for data mining are divided into several categories according to the data mining techniques that they use. This classification is determined by the method of data analysis that was utilized, which may include neural networks and machine learning, genetic algorithms and data visualization, statistics and data warehousing, and so on. The degree to which the user is involved in the data mining process may also be a factor that is taken into consideration throughout the categorization process. Some of the available possibilities include Interactive systems that are based on queries, may be self-directed, and encourage exploration.

2. Clustering:

The process of arranging data by establishing groups of objects that are similar is referred to as clustering. When characterizing the data using only a few clusters, it is normal to lose several confined characteristics; nonetheless, progress is still being made despite this loss. The data is clustered, and the resulting model is employed. Clustering, when seen in the context of data modeling, may be understood as having its roots in the domains of statistics, mathematics, and numerical analysis. This is because clustering is used to group data together. Clustering really serves an entirely unanticipated function when it comes to the applications of data mining. In the realm of science, some examples include computational biology, medical diagnostics, web analysis, customer relationship management, text mining, and information retrieval.

2. Outer detection:

Using this type of data mining, researchers are made aware of any data points within the dataset that do not adhere to the expected behavior or structure of the model. This strategy may be useful in a variety of disciplines, including those dealing with infiltration, detection, fraud detection, and so on and so forth. One other name for this type of analysis is "Outlier Analysis." A statistic that departs from the norm in an extraordinary manner is referred to as an outlier. The vast majority of collections of real-world data will have at least one out of the ordinary observation. The group of people that mine data extensively relies on various technologies that may detect outliers. Outlier identification is useful for a wide variety of tasks, including but not limited to the following: recognizing fraudulent usage of credit or debit cards; locating abnormalities in data collected by wireless sensor networks; etc.

3. Sequential Patterns:

The sequential pattern data mining technique is utilized so that patterns may be discovered in sequential data sets. Part of this procedure involves locating interesting subsequences that are included inside a larger set of sequences. The significance of a sequence may be established based on a variety of criteria, such as the duration of the sequence, the frequency with which it occurs, etc.

4. Prediction:

The process of prediction made use of a wide variety of data mining approaches, some of which included the detection of trends, clustering, classification, and other similar processes. It does a line-by-line study of the events or instances that came before in order to make accurate predictions about the future.

5. Regression in data mining

Data miners use a technique known as regression in order to make projections about the numbers that are included in a dataset. Forecasting a range of parameters, including price, may be accomplished through the use of regression. The analysis of consumer behavior, financial forecasting, and trend research are just a few examples of the various uses of this field of study. In this course, you will get an understanding of what regression is, as well as the many different types of regression and various uses of regression.

Regression

Use the supervised machine learning technique known as regression if you want to make accurate predictions about continuous-valued characteristics. Regression is a statistical method that may be used by any business to investigate

the relationship between a "target" variable and "predictor" variable. The process of sorting through information that might be beneficial in financial forecasting and time series modeling relies heavily on this resource. (raghunandan, 2011) [iv], (Noateng, 2020) [VIII]

The technique of regression involves attempting to match a series of data points to a straight line or curve using mathematical equations. The manner that this develops helps to close the gap between the data points and the answer as much as possible.

The most common types of statistical analysis are the linear and the logistic regression. On the other hand, there is a myriad of different types of regression that may be performed, each of which has the potential to be successful to varied degrees depending on the data that is available.

Using regression, it is possible to make predictions about any and all dependent data sets, provided that the independent variables are present. However, the trend can only be viewed for a certain amount of time. Regression is a valuable tool for generating predictions, but it does so under the assumption of certain requirements, such as the independence of the variables and their underlying normal distributions. In order for regression to be effective, these conditions must be met (Thakrar, Z.,2022)[XI] . In the event where A and B are two independent variables, the distribution of both of them together would be considered a bivariate distribution. There is a significant connection between these two aspects, despite the fact that one may argue that they are independent of one another. It is essential to acquire and make use of the marginal distributions of both A and B. Before really putting the Regression into action, it is essential to perform a comprehensive analysis on the data and run a number of preliminary tests to validate its suitability. You have the option of utilizing non-parametric testing in this kind of situation. (C.r.d.c,2010) [XII]

Variables used in the study Here, dependent variable is score obtained by the students in their educational performance and independent variables are age, gender, family type, family income, father's occupation, type of school, transportation to the school. (P. Vishal,2023)[X]

The following is a mathematical expression: $y = m_{x1} + m_{x2} + m_{x3} + m_{x4} + m_{x5} + m_{x6} + m_{x7} + b$

Where, Y= a component that is utilized in the process of regression analysis that is educational performance of the students or a score

X1= In a regression analysis, it is the first independent variable.

The x2= The second element in the examination of the regression.

The x3= the third variable taken into consideration in the regression analysis,

x4= the fourth element in the analysis of the regression,

x5= The fifth independent variable in the regression,

x6= the sixth element in the analysis of the regression,

x7= This is the seventth independent variable in the regression.

3. Objectives Of The Study

A. To identify the impact of variables

B. To know about importance of dependent and independent variables

C. To measure the impact of variables on the educational performance of the students

4. Hypothesis Of The Study

H1-According to the argument in Hypothesis, the academic achievement of students varies greatly with age.

H2 - The gender of a student has a significant impact on their level of academic achievement.

H3 - The socioeconomic status of the student's parents has a substantial impact on the student's level of academic achievement.

H4 - There is a considerable correlation between the amount of money brought in by families and the academic success of their children

H5 - The occupation of the father has a significant bearing on the degree to which his children achieve academically.

H6 - The schools that a student attends have a tremendous impact on the academic outcomes that student experiences.

H7 - The manner in which students get to school has a significant impact on the academic achievements of those students.

5. Phase Of Work

For the objective of this study, information was gathered from standard XI pupils at a number of different schools. The information gathered was only related to non-intellectual factors; additional information regarding the student's academic achievement needs to be gathered. The following steps generally classify the total activities:

1. Data collection and Data set preparation
2. Data Preprocessing and Handling Missing Value
3. Data Analysis and Interpretation

5.1 Data Collection and data set preparation

Data have been gathered by researchers from numerous associated institutions. The data set includes several Class X non-intellectual parameters. This data set contains roughly 512 records (work in progress). Additionally, information on the kids was gathered from the schools. This information is supplemented by information on the location and type of schools.

After combining all these data sets the resultant database record contain s approx. 20 and more parameters. Due to the fact that the data was gathered from many sources, it is important to properly clean the data by filling in missing numbers, reducing noise, locating and removing outliers, and resolving inconsistencies. The cleansed data are then converted into a table format that is appropriate for a data mining model.

5.2 Data Preprocessing and Handling Missing Value

The data collected and brought together is very huge and contains a lot of unwanted details. After gathering the data, researcher will work on data cleaning process, generally data cleaning reduces errors and improves data quality. Data in the real world is lots of potentially incorrect data like instrumental faulty, human or computer error etc. After collecting data researcher also face incomplete data like lacking attribute values;

- for example
  - Occupation=" "(Missing data)
  - Family Income="-10" (an error)
  - Age = "42" (Wrong input)

After data cleaning, then data integration process is the phase of combining data from several disparate sources. Researcher will work on data transformation process which is constructive, Destructive, Aesthetic and structural through this transformation research make usable format that can be useful in decision making process. Data reduction process is the process where researcher will collect data based on questionnaire from various schools and then compress data which will be prepare aggregate data.

In the process of data mining, the data is crucial. The effectiveness of the future mining procedure can be directly increased by the caliber of the data that has been acquired. However, data gathered in the real world frequently contains some missing values and is therefore incomplete. applying applicable missing value analysis principles and straightforward techniques like Mean/Mode. Missing values can be simply handled with imputation. (R, 2018) [XII]

5.3 Data analysis and interpretation

Researcher has used multiple regression test for data analysis and interpretation. Here, dependent variable is the score obtained by the students in their educational performance and independent variables are age, gender, family type, family income, father's occupation, type of school, transportation to school.

Formula: y = mx1 + mx2+ mx3+mx4+mx5+mx6+mx+ b

Where,

Y= regression's "dependent variable"

M= The slope of regressive behavior

X1= the very first variable that may be manipulated in a regression, x2= the second element to be included in the analysis of regression, x3= the third factor that determined the outcome of the regression, x4= the fourth important factor in determining the regression, x5= the sixth factor to be taken into consideration in the statistical model, x6= the sixth element to be included in the analysis of regression, x7= the eighth element taken into consideration in the statistical analysis, and B= constant.

Table no. 1.1 table showing Analysis of Regression Coefficients

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	65.992	4.313		15.303	.000
	A	3.915	1.491	.159	2.626	.009
	GD	4.186	1.696	.169	2.469	.014
	FT	-.010	1.620	.000	-.006	.995
	FI	-1.167	1.142	-.059	-1.022	.307
	FO	-.437	.795	-.027	-.549	.583
	TS	.237	.530	.021	.446	.655
	TTS	1.341	2.800	.028	.479	.632

a. Dependent Variable: score, A= AGE, GD=gender, FT= family type, FO= father's occupation, TS= type of school, TTS=transportation to school

Table no 1.2 table showing ANOVA f test of selected variables

Anova						
Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	2071.746	7	295.964	2.353	.023[b]
	Residual	63400.246	504	125.794		
	Total	65471.992	511			

a. Dependent Variable: score

b. Predictors: (Constant), TTS, GD, TS, FO, A,FI, FT

Residuals Statistics

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	72.4423	78.2690	74.9961	1.93556	512
Residual	-27.53433	25.54747	.00000	10.70737	512
Std. Predicted Value	-1.268	1.625	.000	.961	512
Std. Residual	-2.455	2.278	.000	.955	512

a. Dependent Variable: score

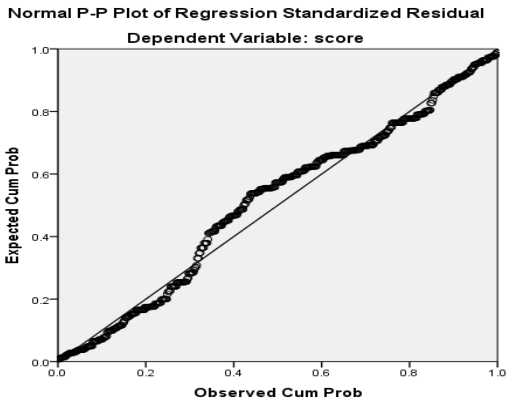
Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.178[a]	.032	.018	11.21580

a. Predictors: (Constant), TTS, GD, TS, FO, A,FI, FT

b. Dependent Variable: score

(here, A= AGE, GD=gender, FT= family type, FO= father's occupation, TS= type of school, TTS=transportation to school)

Graph no 1. Graph shows Normal probebility plot of regression analysis



Hypothesis Results

The study seeks to investigate the effect of age. Gender, family type, family income, father's occupation, type of school, transportation to the school on the performance of the students (score). Following hypothesis were proposed.

H1 - The age distribution of students has a significant impact on the academic progress of the students.

H2 - The gender of a student has a significant impact on their level of academic achievement.

H3 - The socioeconomic status of the student's parents has a substantial impact on the student's level of academic achievement.

H4 - The financial status of a student's parents has a major impact on the student's academic performance.

H5 - The occupation of the father has a significant bearing on the degree to which his children achieve academically.

H6 - The schools that a student attends have a tremendous impact on the academic outcomes that student experiences.

H7 - The manner in which students get to school has a significant impact on the academic achievements of those students.

The score was the independent variable, while the predictors were age, gender, and education level. gender, family type, family income, father's occupation, type of school, transportation to the school. The independent variables significantly predict student's performance (score), *F* (7.295) = 2.353. $p < 0.05$, which indicates that the seven factors under the study have a significant impact on student's performance (score). Moreover, R^2 = .032 depicts that the model explains 32% variance in the student's performance (score).

Calculating coefficients allowed us to identify the influence that each predictor had on (which was the effectiveness or rating of the overall) conclusion students).

1. H1 evaluates whether the age is significantly and positively effects on educational performance of students. The result revealed that age has significant and positive impact on student's performance. (B= 3.915, t= 2.626, p= .009) hence, H1 is supported.
2. H2 evaluates whether the gender is significantly and positively effects on educational performance of students The results show that gender has a significant positive impact on educational performance of the students (B= 4.186, t= 2.469, p= .014) hence, H2 is also supported.
3. H3 evaluates whether the family type is significantly and positively effects on educational performance of students. The results show that gender has a significant negative impact on educational performance of the students (B= -0.10, t= -,006 p= .995) hereafter, H3 is not supported.
4. H4 evaluates whether the family type is significantly and positively effects on educational performance of students. The results show that gender has a significant negative impact on educational performance of the students (B= -1.167, t= -1.122, p= .307) therefore, H4 also not supported
5. H5 evaluates whether the family type is significantly and positively effects on educational

performance of students. The results show that gender has a significant negative impact on educational performance of the students (B= -0,437 t= -,549 p= .583) hence, h5 is also not supported.

6. H6 evaluates whether the family type is significantly and positively effects on educational performance of students. The results show that gender has an impact that is extremely favorable to the academic achievement of children (B= 0.237, t= .446, p= .655) therefore, H6 is also not supported.

7. H7 evaluates whether the family type is significantly and positively effects on educational performance of students. The results show that gender has an impact that is extremely favorable to the academic achievement of children (B= 1.341, t= 0.479, p= .632) therefor, H7 is also not supported.

Conclusion And Future Work

Here researcher have collected the most affecting factors for the prediction of educational performance for students of class X. As a space for proposed work in our next researcher will test each and every factor against the model. For the development of the final model, each factor has to be analyzed and evaluated and finds its impact on final result.

References

1. S. V. K.K.nivethithaa, “Survey on data mining techinques, process and algorithms,” Journal of physics, 2021.
2. R. Pandya and A. Parikh, “A Predication Model and ITS Affecting Parameters For Prediction Of Rural Electricity Requirements,” International Journal of Research and Analytical Reviews (IJRAR), vol. 5, no. 3, pp. 151-154, 07 2018.
3. P. Rajnikant and P. Ajay, “Predicting doemstic electricity consumption using data mining for rural area of Gujarat.,” 2018.
4. Raghunandan and L. B. a. G.H., “A conceptual overview of data mining,” 2011.
5. Karunathilaka. and G. sajeewani, "Virtual team adaption: management perspective on individual differences," Businesses, 2022.
6. Shabnaz and Smina, “A Study on entrpreneurial intention of university students in bangladesh.”.
7. P. Rajnikant and P. Ajay, “Data Mining For Agriculture Electricity Consumption Forecasting For Rural Area Of Gujarat,” International journal of basic and applied research, vol. 8, no. 10, pp. 289-298, 10 2018.
8. Noateng, Simon, E. t. ampofo, E. a. sefah, A. b. moses and Azewara., “Pointificating the relationship between presenting styles and academic performance of senior high school students in sumasi metropolis, ghana,” Educational resaerches for police and practice, 2020.
9. Abdulmohesen, M. A. Alharbi and H. h. lahza, “Predicting malicious software in IOT environment based on machine learning and data mining techniques,” International journal of advanceed computer science and application, 2022.
10. P. Vishal and P. Rajnikant, “Exploring Correlates Of Student Well-Being And Academic Success: A Comprehensive Analysis,” Seybold Report, vol. 18, no. 5, pp. 1287-1298, 22 5 2023.
11. Thakrar, Z., & Gonsai, A.,”Comparing Fish Finding Techniques using Satellite and Indigenous Data based on Different Machine Learning Algorithms. In Advances in Information Communication Technology and Computing: Proceedings of AICTC 2022 (pp. 329-340).
12. c. R. v. d. c. a. t. c. Garca enrique, “Association rule mining in learning management systems,” chapman and hall/CRC data mining and knwoledge, 2010.