Application For Comparison and Analysis of Rural to Urban IQ-Regions Based Study using the Data-Mining Technique

Sudharshan Duth P, Sudarshan Ganapati Bhat, Prajwal M L, Prasanna Hegde, John Joseph Jacgnet

***Abstract:** Predicting IQ in different Rural and Urban areas and comparing their respective education systems is very much challenging due to a set of factors affecting the results. In a country like India, there is a lack of existing system that analyzes and monitors the student profile. Here the IQ is generally suggested on the basis of one's own performance and social influence; to overcome this drawback of IQ prediction from student's profile we have come up with the solution which uses Data Mining Methodologies like Chi-square, Naïve Bayesin predicting and classifying IQ. Factors like answerability and the thinking capacity of the targeted student set is taken into consideration in the proposed model.In this paper we also suggest parameters and survey procedures by existing schools; the data collected will be used for comparing and identifying the regions with different IQ levels and in need of attention based on the results acquired.*

***Key Words:** Data Mining, Information, IQ test, Prediction Algorithm, mental age, Chi-square, Naïve Bayes.*

I. INTRODUCTION

Data mining in educational database is useful in understanding previous unknown patterns and current performance of student. Using Data mining Algorithms, the calculation as well as the comparison of student's IQ is possible by that collected data in certain predetermined Rural and Urban areas. Overall development of a student is based on one's own academic and social performance. Prediction system is the use of different data mining techniques in order to understand different patterns in surveyed student database and predict proposed output of IQ-Prediction.It allows a wider range of data mining techniques and algorithms to be implemented to find out various patterns in an acquired data set and classify the data according to system objectives.

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

Sudharshan Duth P, studied BCA at Amrita VishwaVidyapeetham, Mysuru campus.
Sudarshan Ganapati Bhat, studied BCA at Amrita VishwaVidyapeetham, Mysuru campus
Prajwal M L, Assistance Professor in the Department of Computer Science at Amrita VishwaVidyapeetham, Mysuru campus
Prasanna Hegde, studied BCA at Amrita VishwaVidyapeetham, Mysuru campus
John Joseph Jacgnet, studied BCA at Amrita VishwaVidyapeetham, Mysuru campus

Although many areas are being researched and challenges in the area are being taken up related to data mining, the prediction system is a much ignored field and the resources are limited even though the required data source is abundant in our country. The challenge is to conduct survey and Collect the required information, parameters and to maintain the acquired dataset's integrity. The classification of the huge data by implementing different data mining techniques poses as the biggest complexity. Accurate and efficient prediction of the collected data set without compromising the integrity of the data.

II. BACKGROUND STUDY

Tismy Devasia et al.,(2016)[1] The approach made the proposed system to generate the data set for predictive variables, and further identification of various features or factors which affected the performance of student's learning rate in their respective academic time.Validation of the model which is being proposed for educational institution with students' performance. Thomas J et al., (2016)[2] proposed a work related to the prediction of human heart diseases which has difficulty in implementing K-NN algorithm where it is a non-parametric in prediction system. Understanding the consistency of different data mining algorithms has been described.Mohamed SolimanHalawa et al.,(2017)[3] made a work on prediction system of students' personality which is useful in understanding the patterns and current performance of an individual in order to fetch the required result along with understanding various prediction parameters.KrisztianBuza et al., (2011)[4] worked on the IQ estimation to fetch the accurate time series classification that helped in understanding the implementation of different data mining algorithms. It has a brief methodology in comparing algorithms and its efficiency in the proposed system. DorinaKabakchieva et al.,(2013)[5] proposed their work on classification techniques in data mining on student performance prediction that helped understanding the parameters and implementation of Decision Tree technique. It helps in finding out any hidden patterns in the available data set that could be useful in predicting students' performance based on proposed survey characteristics. Avnish Kumar et al.,(2014)[6]Worked on the data mining techniques that helps in learning the relationship between the data, class and its labels. The paper gives the brief information about the data classification, which is normally done in two stages, the training stage and the testing stage. And these two stages are discussed in this paper with the advantage and disadvantages.

S Mohamed et al.,(2016)[7] worked in this paper that describes the use of data mining as well as the prediction techniques in the field of medical science. The disadvantage is that change in data could do an adverse effect. Proper adjustment in the treatment is to be made by considering the data. So that the conclusion is Data mining is successfully used in the medical field.N.V.

Krishna Rao et al.,(2016)[8]This paper gives the knowledge about the factors that we have to consider while collecting a huge set of data without any overlapping because large data collection could be time consuming and might be irrelevant. Algorithms like support vector machines, Ann helps in gathering the data and small dataset can also be considerable to test the following system.Huang Lan et al.,(2010)[9] worked in this paper deals with understanding the techniques like decision tree and data warehousing in behavior prediction. Regular data collection is required as this technique will be more efficient if handling of the dynamic data is possible. The customer needs could be predicted and matched if the collection and classification of the data is done in a fruitful way.Nguyen Thai-Nghe et al.,[10] have proposed a paper, Improving Academic Performance Prediction by Dealing with Class Imbalance, in which different features and parameters influencing in the proposed prediction system is dealt with, along with class imbalance at data and classifier levels and methods to overcome it like over-sampling, under-sampling and ensemble methods respectively are discussed.Cesar Hervas et al.,[11] have proposed the paper Data Mining Algorithms to Classify Students, where a proposed evaluation method for an imbalanced data set for finding Geometrical Mean of the acquired dataset depending on the variables is discussed. Neural networks and Decision tree are used for classifying the acquired dataset. The classifier model is found to be accurate and comprehensible to deal withincomplete data, missing data and noisy data. However, the accuracy degrades when the number of attributes increased and if the data set has more dimensions.

III. PROPOSED METHODOLOGY

A. *Collecting the required dataset.*

The initial stage will be the process of collecting the datasets that are required in implementing the specific functions of generating the IQ.

B. *Classification of data and storing.*

The classification procedure takes place by assigning values to the different parameters like answerability, duration along with the average scores obtained in each tests.

C. *IQ tests and data survey*

In order to compare IQ, a survey test has to be conducted in order to gain the required data set. The test should contain and be conducted with similar questions as well as time allotted to complete it. The proposed test would have questions regarding every aspect of the subjects being already dealt with in the respective institutions to give and fair opportunity to the students and also to maintain the integrity of the proposed goal of IQ comparison.

D. *Problems of IQ-comparison*

Poor data quality such as noisy data, missing values, incorrect values, inadequate data size and poor representation of data sampling. Efficiency and scalability of data mining algorithms to efficiently extract and implement information from abundant data set. Constant updating of models to handle data velocity or new incoming data. Processing of large, complex and unstructured data into a structured data.

E. *Analysis of IQ GENERATION*

IQ or the word ‘intelligence quotient’ is derived from the German term ‘Intelligenzquotient’ coined by psychologist William Stern. This concept which came into the light in 1912 is used to measure the mental ability of a person by considering his skills of answerability and thinking capacity. The present system provides the facility to generate the IQ of an individual considering their performance in various tests.The data is to be transformed with modeling towards achieving goal which has been discovered considering the score of the individual’s prediction of the IQ.

F. *Conducting the vary points in perspective views*

The task of collection includes conducting the survey with pre determined questionnaires and results of the each individual who had undergone with the test. The data collected here is from the Rural and urban schools of particular region. Data should consist of information about the students who are being tested and format has to be classified accordingly towards parameters.

G. *Formulation of techniques*

The classified data could be considered as the input for the GUI and it has to be stored in the SQL database. The formula used in the system generates the required arena of outcomes that are used for regional analysis of students’ mental performance.

The mathematical expression could be maintained under the laws of attentiveness.
IQ= (MA/CA)*100
In the above formula IQ represents Intelligence quotient of an individual.
MA represents the physical age of an individual i.e., mental age
CA represents the physical age of an individual i.e., Chronological age

H. *Pre-process testing*

The real time datasets are being tossed to the technique of chi-square testing. The testing based under comparison (▲) between expected and observed value. The difference value has to be migrated for the technique of ours.

$$x_c^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

IV. PROPOSED ALGORITHM

A. Naïve Bayesian classifieris also known as “Probabilistic Classifiers”, it works on Bayes’ theorem with objective assumptions between the features. With appropriate pre-processing, it is competitive in its domain. Requiring a number of parameters and variables in a learning problem it is very scalable.

Algorithm of Naïve Bayes in proposed system

Step1: Start
Step2: Scan the surveyed student data set i.e., conducted IQ-test results.
Step3: Apply the formula

$$P(b|A)=\frac{P(A|b)P(b)}{P(A)}$$

Where- P is probability
b is a class variable
A is dependent entity of size ‘n’
A=(a1, a2, a3, ….., an)

Step4: Note down the result and compare accordingly.
Step5: State the observed outcome.
Step6: Stop.

B. Decision Tree acts as a decision support methodologyfor decisions in various possible instences, including situational outcomes, resource costs etc. It has three nodes: Decision nodes, Chance nodes and End nodes. Every sub-branch of this tree structure shows a possible output of each testcase and its leaf node i.e., a child node.

Algorithm of Decision Tree in the proposed system

Step1: Start
Step2: Scan the surveyed student data set i.e., conducted IQ-test results.
Step3: Decide best attribute of the data set and place it as the root node of the tree.
Step4: Split the data set into subsets.
Subsets should be made in such a way that each subset contains data with the same value for an attribute.
Step5: Design the tree based on the respective parent and child nodes chosen.
Step6: Compare the data according to the designed tree and display the outcome.
Step7: Stop.

C.Comparison between Naïve Bayes and Decision Tree

Observation: These two algorithms are differentiated by three main things: speed, accuracy and interoperability.

Naïve Bayesian is an algorithm with supervised learning. It assumes Bayes theorem where it has more number of classes to predict text classification, recommendation system and others.

Pros:

- Implementation is easy for users.
- Requires only a small dataset to train.
- Accurate results are obtained for most of the cases.

Cons:

- As it is independent on dataset it is complex to predict its accuracy.
- The class occurrence and the probability of certain attributes together will be estimated to Zero.

Decision Trees- Easy to understand and has multiple parameters, those look at various issues like missing values, outliners, and dimensionality. They are non-parametric. Major disadvantage is it should be faster once trained (although both algorithms can train slowly depending on amount/dimensionality of the training data set). The tree uses inherently batch-learning algorithms, which also “throws away” the input features that it does not find useful.

V. RESULTS AND DISCUSSIONS

The system developed for the generation of the IQ for the particular student body by making a survey by considering the regional basis under the influence of some kind of bi-laws and ability of different people’s mindset. This helps in identifying and categorizing the change in mental ability. So the proposed system can help in various governmental and non-governmental subjective activities.

The setup which the system is using has been evolved around the configuration of a pre-requisite algorithm called Chi-Square testing technique. This classifies the data in order to maintain the ledger of finite data.

Fig:1. Architectural Flowchart

Algorithm for IQ generation

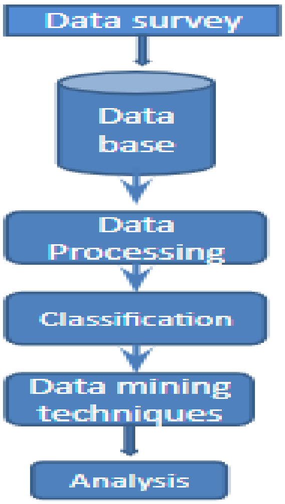


Figure: 1.1. Chi-Square Implementation with the dataset

Step1: Input the values into the GUI
Step 1.1: Check for the IQ using the parameters obtained by the tests
Step 1.2: Save the Input Data to the Database using GUI functions
Step 2: Access the database for recording the IQs and other details
Step 3: Gather the data to apply the chi-square test
Step 3.1: Find the observed and the expected values
Step 3.2: Using the formula generate the difference

Students	Urban	Rural	Total	Average
High IQ	15	6	21	0.21
Good IQ	20	14	34	0.34

Average IQ	10	20	30	0.3
Low	5	10	15	0.15
Total	50	50	100	
Expected	10.5	10.5		
	17	17		
	15	15		
	7.5	7.5		
Difference	0.019294			

The above table briefs on the distribution of rural and urban IQ over student body where their scores are systemized on the basis of prioritized level of ranks. With the help of average obtained using observed data, calculations of expected values are being generated. The differences are considered to be the outcomes for executing further enhancement of finding the chi-value.

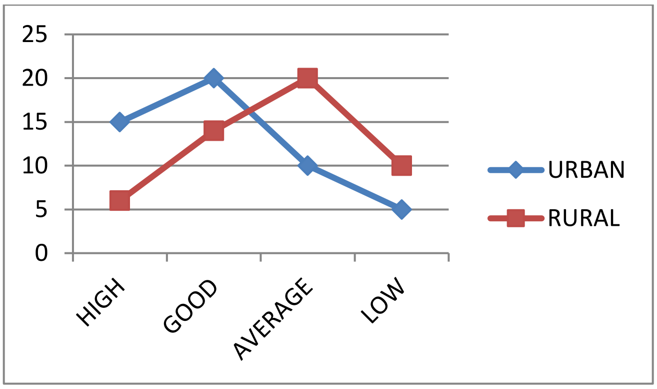


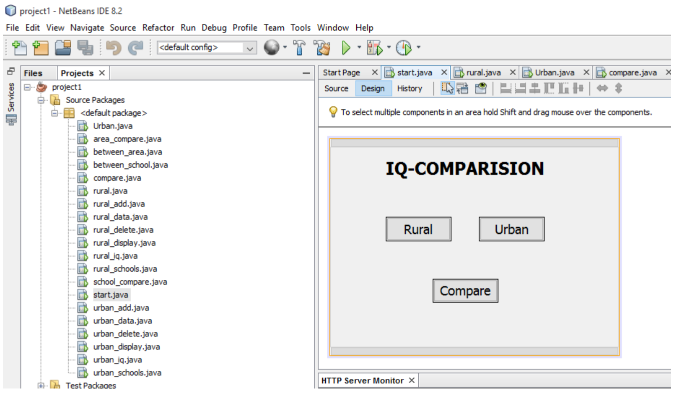
Figure: 1.2 Area distribution of IQ

The histogram chart follows up on area wise distribution of IQ based on their performances in respective tests. The dissimilarity clears the confusion between the knowledge based predictions about rural and urban areas.

As far the limitations are concerned, there it needs huge amount of collected data in order to face any numeric and logical problems. And even a static data set cannot be processed in the system because of change in academic progress of different kinds. So there needs to be so like dynamic collection of data sets to further classification.

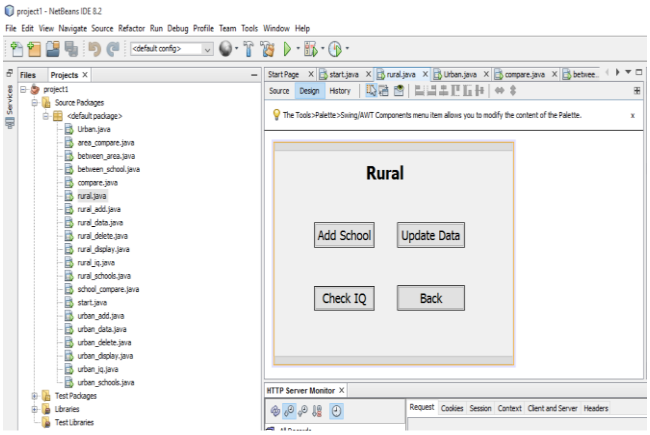
A regional basis has no technological infrastructure like computer systems and all. Even the students' involvement needs to be present to achieve this goal. So, these two could be the major drawbacks that can be faced during the presentation.

A. Implementation Screenshots



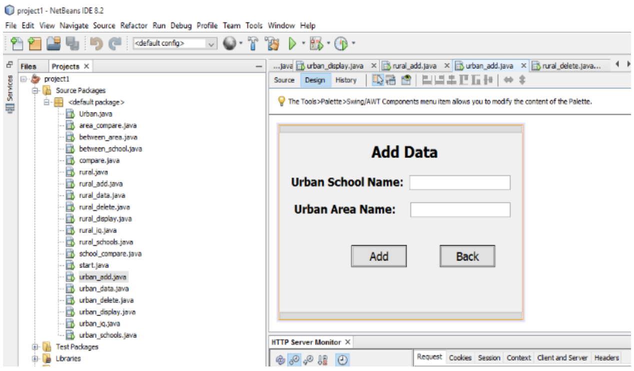
To select multiple components in an area hold Shift and drag mouse over the components.

Fig 2.1(a) GUI design of proposed system



The Tools>Palette>Swing/AWT Components menu item allows you to modify the content of the Palette.

Fig 2.1(b) rural database design

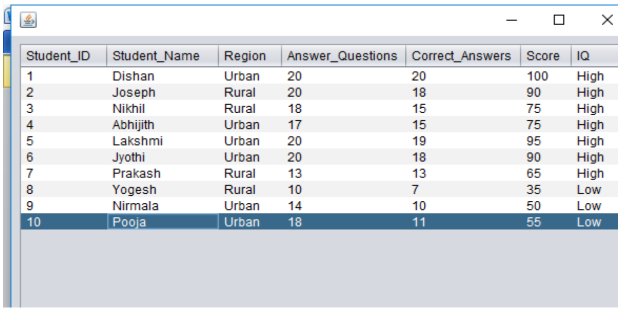


The Tools>Palette>Swing/AWT Components menu item allows you to modify the content of the Palette.

Urban School Name:

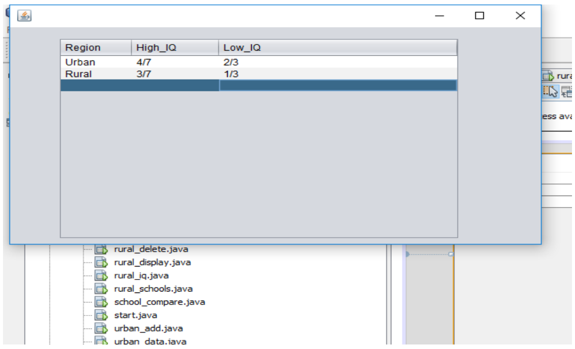
Urban Area Name:

Fig 2.1(c) urban database design



Student_ID	Student_Name	Region	Answer_Questions	Correct_Answers	Score	IQ
1	Dishan	Urban	20	20	100	High
2	Joseph	Rural	20	18	90	High
3	Nikhil	Rural	18	15	75	High
4	Abhijith	Urban	17	15	75	High
5	Lakshmi	Urban	20	19	95	High
6	Jyothi	Urban	20	18	90	High
7	Prakash	Rural	13	13	65	High
8	Yogesh	Rural	10	7	35	Low
9	Nirmala	Urban	14	10	50	Low
10	Pooja	Urban	18	11	55	Low

Fig 2.2(a) Naïve Bayes Dataset



Region	High_IQ	Low_IQ
Urban	4/7	2/3
Rural	3/7	1/3

Fig 2.2(b) Bayesian Calculated Table

REFERENCES

1. Ms.Tismy Desvasia,Ms.Vinushree, Mr.Vinayak Hegde,‘’Prediction of Students Performance using Educational Data Mining’’,Department of Computer Science, Amrita school of arts and sciences, Amrita Vishwa Vidyapeetham University, Mysuru, India.
2. J Thomas, Theresa Princy. R,’’Human Heart Disease Prediction System using Data Mining Techniques’’, Department of Information Technology,Christ University faculty of engineering, Bangalore, India.
3. Mohamed Soliman Halawa, Mohamed Elemam Shehab,’’Predicting Student Personality Based on a Data-Driven Model from Student Behavior on LMS and Social networks”, Arab academy for science,technology and marieime transport, Cairo, Egypt.
4. Krisztian Buza, Alexandros Nanopoulos, ‘’IQ Estimation for Accurate Time-Series Classification’’, University of Hildesheim, Hildesheim Germany .
5. Dorina Kabakchieva,‘’Predicting Student Performance by Using Data MiningMethods for Classification’’, Sofia University “St. Kl. Ohridski”, Sofia 1000,dorina@fmi.uni-sofia.bg
6. Avnish Kumar, Akshat Gawankar,“Student Profile and Personality Prediction Using Data mining Algorithm” RajivGandhi Institute of Technology-Mumbai,India ISSN No-2395-4396 Vol-3 Issue-2 Pageno-8.
7. S. Mohamed, B. Q. Huang, ‘’Prediction of NB-UVB Phototherapy Treatment Response of Psoriasis Patients using Data mining’’, M-T. Kechadi Insight Centre for Data Analytics, University College Dublin, Ireland, Email : sharifa.mohamed@ucdconnect.ie
8. N.V. Krishna Rao,Associate Professor, Dept. of CSE, Institute of Aeronautical Engineering, Hyderabad nvkrishnarao8778@gmail.com , Dr. N Mangathayaru, Professor, Dept of IT, VNR Vignana, Jyothi Institute Of Engineering & Technology, Hyderabad, Dr. M. Sreenivasa Rao, Professor of Computer Science, Director‘’Evolution and Prediction of Radical Multi- Dimensional E-Learning System using Data Mining Techniques’’,AAC, JNTUH, Hyderabad, India.
9. Huang Lan, Zhou Yu-Qin,‘’Research on Data mining Algorithms for Automotive Customers Behaviour Prediction Problem’’, College of Computer, Jilin University, Changchun, China.
10. Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme, “Improving Academic Performance Prediction by Dealing with Class Imbalance”, Information Systems and Machine Learning Lab University of Hildesheim, Hildesheim, Germany.
11. Cristobal Romero, Sebastian Ventura, Pedro G. Espejo and Cesar Hervas,”Data Mining Algorithms to Classify Students”, Computer Science department,Cordoba University, Spain.

AUTHORS PROFILE

Prajwal M L studied BCA at Amrita VishwaVidyapeetham, Mysuru campus. Interests include Data Mining, Soft Computing and Software Engineering.

Prasanna Hegde studied BCA at Amrita VishwaVidyapeetham, Mysuru campus. Interests include Data Mining, Distributed Computing.

P. Sudharshan Duthis serving as Assistance Professor in the Department of Computer Science at Amrita Vishwa Vidyapeetham, Mysuru campus. Interests include Bio-metrics, Digital Image Processing, Internet of things, Sentimental Analysis and Software Engineering; has published around 12 research papers in various domains such of Biomedical Image Processing, Sentimental Analysis and IOT in various scopus indexed international journals.

Sudarshan G Bhat studied BCA at Amrita Vishwa Vidyapeetham, Mysuru campus. Interests include Data Mining, Soft Computing and Distributed Computing.

John Joseph Jacgnet studied BCA at Amrita Vishwa Vidyapeetham, Mysuru campus. Interests include Data Mining, Distributed Computing.