Performance Analysis of Students in Educational Big Data using Deep Learning

Syyada Shumaila Khurshid[1],Ravi Kumar Singh Pippal[2]
[1]MTech Scholar, [2]Professor
[1]Department of Computer Science and Engineering, Vedica Institute of Technology, Bhopal, India
[2]Department of Computer Science and Engineering, Vedica Institute of Technology, Bhopal, India
shumailakhurshid74@gmail.com[1] ravesingh@gmail.com[2]

Abstract: The increased use of Technology in Education leads to the arrival of big data, the data from students, teachers as well as educational institutions. This paper discusses the implications of Machine Learning in the present system of education. Machine learning applications to be used in the current education system and also used in the educational database for prediction of the future trend in data. This study will help educational institutions to better allocate human and material as well as machine resources to improve the overall education system. In current advancement in technology such as big data, has proved promising advantages in every field. In Educational field it has showed its efficiency for deployment of new advancement in education. It can utilize for educational and personality development of students and promoted the better development of education. In this work DNN is used

Keywords: Data mining, Big Data, Educational Big Data, performance, Hadoop.

I. Introduction

Big Data is now widely used to describe and define the recent emergence and existence of large data sets. It can be found in many areas. The public, corporate and social sectors consistently receive and produce large amounts of data from a variety of sources and in different formats. Big data and analytics have added value to data in different contexts and have therefore proved to be an extremely useful approach to their potential impact both in industry in the form of business intelligence and analysis [4], and in science with data mining techniques in education to study and learn analysis [5]. Given the limited research on the use of big data and analytics in the context of the education system, we will introduce the reader to the new field of big data about education that puts big data into education and like data about education. education can be treated in different dimensions and from different dimensions. Perspectives to highlight various stakeholders such as policy makers, academic faculties, assessment specialists, researchers and students in computer science, engineering and computer science courses, and to promote data-driven activities to improve the quality of education.

One of the areas where volume, diversity and speed coexist in data is higher education. Large amounts of education data from various sources and in different formats are collected and generated daily in the higher education ecosystem. Education data varies from data obtained from student use and interaction with learning management systems and platforms (LMS) to learning activities and course information. which define a program such as learning objectives, programs, learning materials and activities, exam results and course assessment. include other types of data related to administrative, educational and quality improvement processes and practices.

The limited use of big data on education, as well as the size and nature of such data in the context of higher education, means that specific techniques must be applied to uncover useful new knowledge currently hidden in the data. [6]. More recently, big data and analytics have shown great promise in promoting various interventions in higher education. These measures concern "administrative decision making and the allocation of organizational resources", the prevention of failure of vulnerable pupils through early diagnosis, the development of effective teaching techniques and the revision of the traditional curriculum view. rethinking the various data collected by the LMS, social networks, learning activities and curriculum and regularly created as a network of relationships and connections. In particular, one of the areas identified where big data and analytics are suitably applicable to research and improvement in higher education is the curriculum and its content as an important part of big data on education [7]. Therefore, various work has been carried out towards developing an effective big data analytical approach on educational data. Figure 1 highlights the frequently addressed problems with its methodologies and limitations associated with analytical approach.

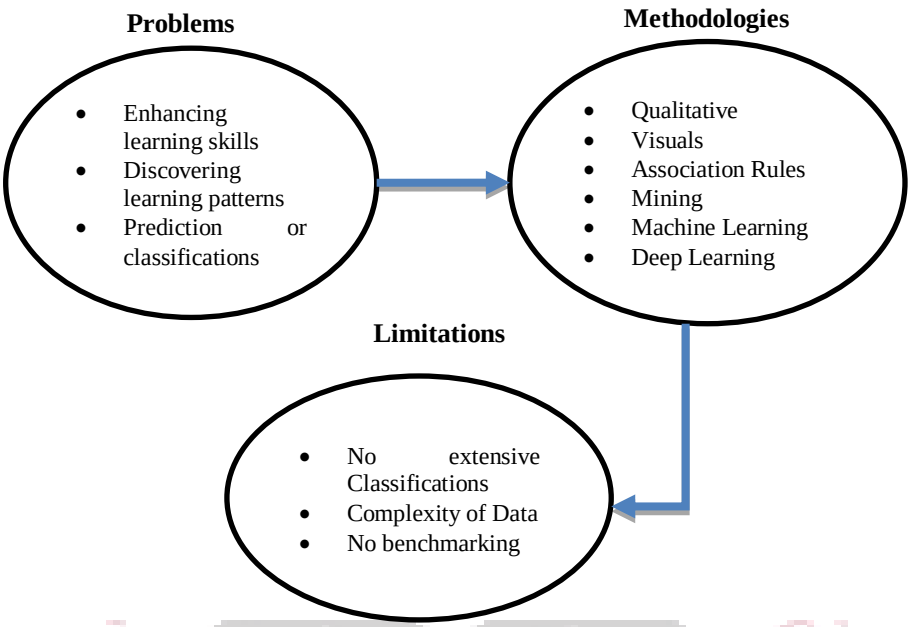

Figure 1: Observations from Existing Educational Big

The main focus of educational data mining and application of machine learning are described as below:

1. Predicting Future Behavior of Student: The Modeling can be used to create student Models based on learners' characteristics which may used detailed information such as their motivation and learning behavior.
2. Identify or Improving Domain Models: By applications of data mining or more precise Educational data mining used to find a new or improvements to existing models are possible.
3. Study of Educational Support: The learning system can be used to studying the effects of educational Support.
4. Scientific Knowledge of Learners: To build student models, in the field of Educational Data Mining research and the technology software and Machine Learning can be used.
5. Techniques of Machine Learning and Their Use in EBD: The two main techniques of Machine Learning supervised and unsupervised Learning and for what purpose they are used in educational database.

Big data analysis and mining is a process of extracting, refining, and analyzing cluttered data from aspects of visualization analysis, data mining algorithms, and predictive analysis [7].

A. Visual analysis

Visual analysis refers to analysis methods that clearly and effectively communicate information by means of graphical means. In the application of educational big data, it is mainly the correlation analysis of massive data, which is the process of correlation analysis of scattered and heterogeneous data to form a complete analysis chart.

B. Data mining algorithm

Data mining algorithms, which is the methods for calculating, processing, and analyzing data by creating a data mining model. It is the theoretical core of the entire education big data analysis. There are many kinds of data mining algorithms. Different algorithms applied to different types of data will show different results. In educational big data analysis and mining, usually the first step is to analyze the original data, and then find the mining algorithm for specific types of patterns and trends according to the application requirements.

C. Predictive analysis

Predictive analysis is one of the most important application areas of big data analysis. It achieves the purpose of predicting uncertain events by combining multiple advanced analysis functions. Predictive analysis of educational data often uses functions such as statistical analysis, predictive modeling, data mining, text analysis, optimization, machine learning, deep learning and other methods to help users discover data evolution trends, data patterns, and data relationships presented in the original data. Use these indicators to predict and provide decision-making basis for taking corresponding measures.

II. RELATED WORKS

Yu et al. [2] focuses on extracting payloads from rich online education data using transfer learning using Hadoop to build the Online Education Data Classification Framework (OEDCF) and design a Tr_MAdaBoost algorithm. This algorithm surpasses conventional classification algorithms, where the required data must be limited to independent and identically distributed data, since the correct classification can be achieved through instructions in line with this new algorithm, even if the distribution of the data is different. At the same time, OEDCF can use Hadoop's parallel processing architecture to significantly improve the efficiency of data processing, create favorable conditions for learning analytics, and promote personalized learning. and other activities in the era of big data.

Li et al. [3] Use of data such as information on student performance in a vocational school and consumption information from Campus One-Card. First, the original data is preprocessed so that the processed data can meet the basic requirements of data analysis. Then, after preprocessing, the data is discretized using data discretization technology so that the data can meet the a priori algorithm requirements for the assignment rules. Finally, the Apriori algorithm is used to find the correlation between students' academic performance and the consumption data of a card on campus.

Santos et al. [4] proposed a computational approach using educational data mining and various supervised learning techniques (decision trees, K-closest neighbors, neural networks, support vector machines, naive bayes and random forests) to evaluate and identify the behavior of different predictive models. profile of university students at risk in a Brazilian university context. The results of this article show that some algorithms can be used as tools to support decisions that reduce early school leaving.

Nadabi et al. [4] This project used J48, Sample Logistic and Naïve Bayes to predict student selection. The results of this project show that many students have to choose applied mathematics due to their performance level in mathematics.

Ketui et al. [6] focused on classification models for use in education data mining. Classification models are applied to identify the appropriate subject for science students. The experiment is designed to improve student performance by comparing the performance of five grading models and then predicting the appropriate academic performance in each major.

Shao et al. [7] proposed an optimized mining algorithm to analyze students' learning levels using dynamic data. The algorithm first uses optimized text classification technology to automatically adapt question texts to knowledge points to improve efficiency and quality. Then the subjective weighting method, combined with expert experience, is used to create the matrix of the student's learning level using knowledge points based on dynamic data from the student's records. Finally, the DBSCAN clustering algorithm is used to group students' personalized learning characteristics according to the learning level matrix.

III.PROPOSED METHODOLOGY

The proposed flow chart of this research work is given below in which we are going to analyze all the features by using single technique which is the part of machine learning. The given flowchart describes the complete process from initial stage to the final stage. As a result, the accuracy rate and precision rate is known. The procedure includes five steps as given in the flow chart, these steps are, the collection of educational big data which consist of huge information about the students. Second step is segregating the useful and useless information from the data, and this step also includes the removal of unwanted data from the given information. Third step is selecting the data which is to be predicted on the basis of given features, and enhancing the feature for different prediction values from the educational big data. Fourth step is the use of MapReduce, its complete working we have described above and the last step is DNN classifier which is deep neural network classifier. And finally, the output results as the accuracy rate and precision rate.

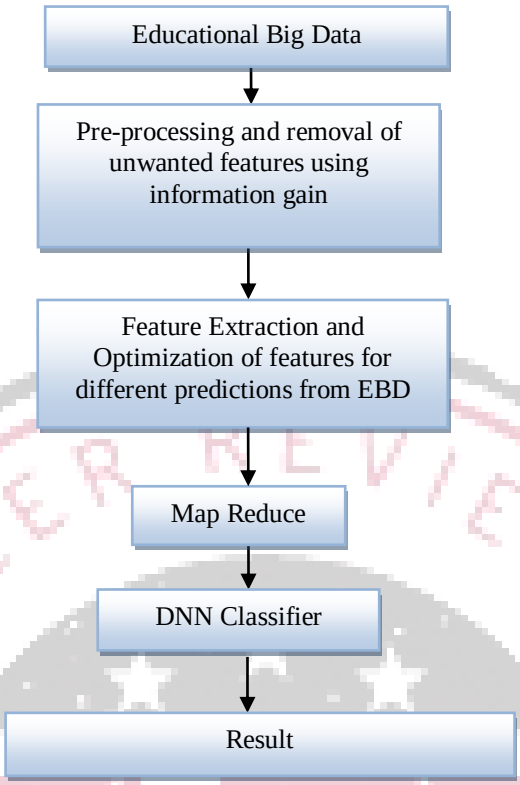

Figure 2: Proposed Flow Diagram

Many techniques are used in acquiring the features from the educational big data to analyze the student performance prediction, attendance shortfall and many more. The selection of proper initialization techniques significantly decreases the time required to converge when trained. Data feature metrics are initialized that are arbitrarily projected over the feature vector space without any training.

Algorithm:

1. Begin
2. Input: Institutional Data set, D
3. Output: SPP, SAS and AP ← D
4. Preprocess_Data (D) →{F_1, F_2,….F_n}
5. Info_gain{F_1, F_2,….F_n}←Score
6. Mapreduce(Score)
7. Output ←DNN(Score)
8. End

Preprocessing: This stage purpose is to preprocess the database file in which there is conversion of symbolic attributes in numerical is done.

Feature Extraction: After preprocessing feature extraction is performed. For best related feature extraction, information gain formula is used for feature selection. It evaluates the gain of each variable in the context of the target variable. In this slightly different usage, the calculation is referred to as mutual information between the two random variables.

$$Information_{gain} = E_{dataset} - \sum_{n=1}^{N} \frac{dataset_i}{dataset} (Entropy_{dataset_n})$$

Where, $E_{dataset}$= Entropy of dataset

$dataset_i$= i_{th} instance of dataset

$Entropy_{dataset_n}$= entropy of n_{th} subset of dataset

Regularization of network: Deep Neural Networks (DNN), when trained with large no. of parameters, confront the problem of overfitting. Over-fitting is nothing but the error which is caused due to addition of noise or when the function is unable to set into decided points this give rise to multiple error in the data and to predict and diminish such errors or fluctuations the networks are trained.

In dropout regularization, certain units from the neural network are selected randomly and ignored during training. To overcome the problem of over-fitting, a simple and powerful technique called dropout regularization is used. In dropout regularization, certain units from the neural network are selected randomly and ignored during training.

Initially, for each hidden layer, the probability of the dropout rate is fixed to 1.0; further, it is tuned between 0.5 and 1.0 over the validation set. Because dropped units are shared within the network, the same units of the hidden layer are dropped at each node. Basically the CNN has three layers, but in between some hidden layers also exists.

The hidden layers in CNN do not suffer from the problem of over-fitting as they are not directly involved in the process of prediction. However, due to the large size of the datasets, keeping dropout on the fully-connected layers are beneficial. To attain this, a standard softmax activation is used for the output layer that takes a node's vector z(v) as input and produces prediction y(v). The input given to this function is always a exponential form of every output element and then divided by the sum of all the output layer elements.

The output of the softmax function is equivalent to a categorical probability distribution, i.e., the probability that any of the classes are true.

$$f(x) = \frac{e^x}{\sum_{n=1}^{N} e^x N}$$

This method of calculation ensures that the sum of all the values in exponential form is equal to 1.

RESULT ANALYSIS

To evaluate the performance of methodology, the proposed algorithm is simulated in following configuration:

1. Pentium Core I5-2430M CPU @ 2.40 GHz
2. 4GB RAM
3. 64-bit Operating System
4. MATLAB Platform

For simulation result, the research is focused towards implementation of feature co-relation using information gain method and DNN. For executing this simulation, a dataset of students is created on excel and then imported for analysis.

A. Performance Parameters

1) Accuracy

The result analysis is performed to find accuracy of the proposed methodology and to decide the performance rate of proposed methodology.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

Where,

TP = True Positive, that means if student performance is good and predicted label also stands for good performance.

TN = True Negative, that means if student performance is poor and predicted label also stands for poor performance.

FP = False Positive, that means if student performance is poor and predicted label also stands for good performance.

FN = False Negative that means if student performance is good and predicted label also stands for poor performance.

2) Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (MAPE) is a measure of the predictive accuracy of a forecasting method in statistics, for example in estimating the trend. It usually expresses the precision in percentage and is defined by the formula:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{Target_{value} - Obtained_{value}}{Target_{value}}$$

3) R-squared (R^2)

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

A. Attendance Shortfall Analysis

Figure 3 represents the shortfall of attendance of students in each department of a university. This proposed model shows its efficiency in all respect either it is required for decision for promotion or to forecast the future shortfall of students in any department. So, that bunk nature of each student will be forecasted in prior and helps in deciding or preparing decisions to track such students and to enhance their performance.

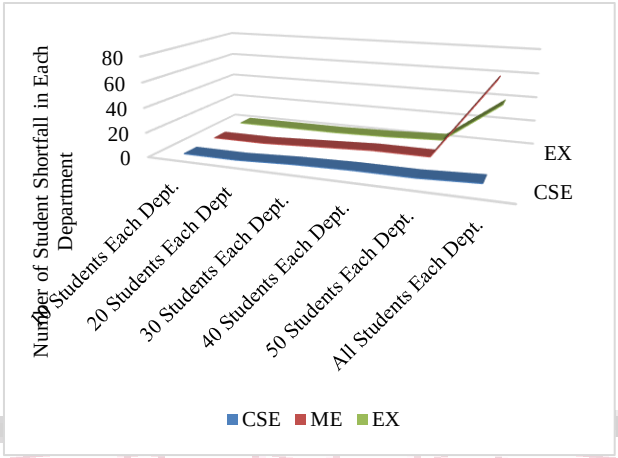

Figure 3: Student Shortfall Analysis

The Figure 4, figure 5 and figure 6 gives the result analysis of proposed work. The result is shown in terms of accuracy, MAPE and R2. For result analysis, test samples are taken variable such as 225, 300, 450 and 600.

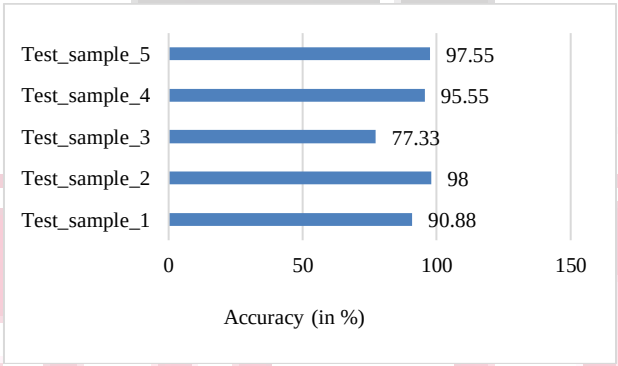

Figure 4: Accuracy Performance Evaluation on different test set

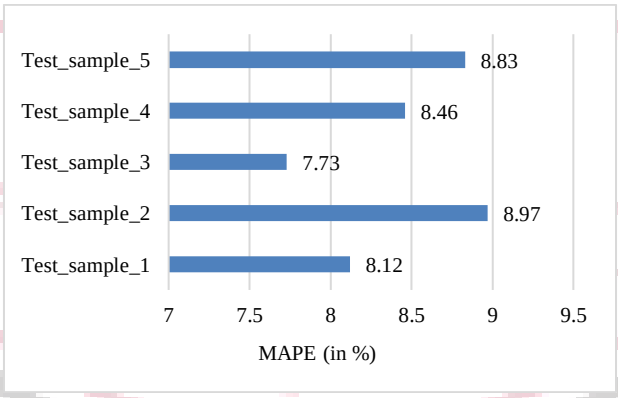

Figure 5: MAPE Performance Evaluation on different test set

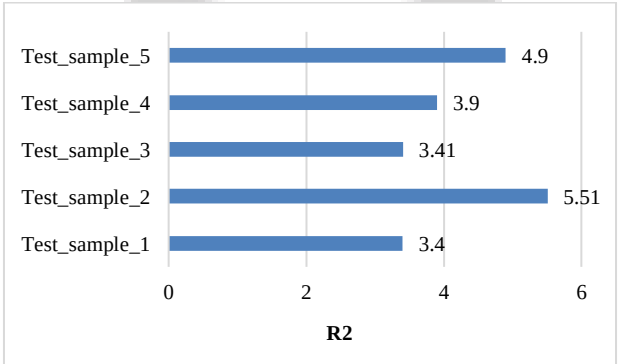

Figure 6: R^2 Performance Evaluation on different test set

The figure 7 gives a comparative result analysis of proposed work with existing work. The result shows the enhancement of proposed work with approx. 14% and the work is also extended towards the finding students which can be selected for appreciation which was not discussed in existing work.

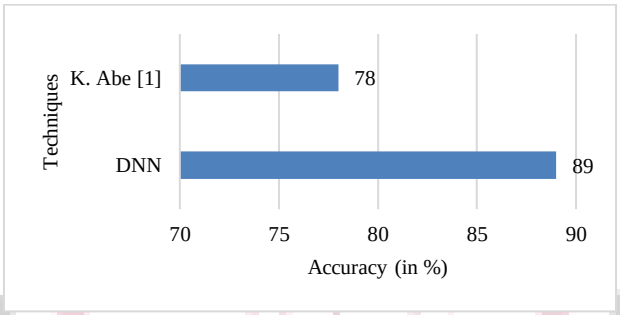

Figure 7: Comparative Accuracy Evaluation

V. CONCLUSION

Big data technology is now the forefront of scientific and technological development, and is an indispensable technology in education data mining. This paper first introduced Educational data mining (EDM) and big data technologies, and then introduced popular data mining algorithms and their applications. Then, according to the characteristics of educational data, combined with big data mining technology, the more practical educational big data mining algorithms and their core ideas are introduced. EDM research is developing rapidly, but the focus is different, and we make an outlook on the research of EDM. At present, there are less research on data preprocessing technology and data mining algorithms in the research system of EDM, this is the most significant problem. Data preprocessing methods is as important as data mining algorithms in EDM research. Some of the important facts analyzed and concluded in this work are stated as below:

i. In this research work, DNN is designed to predict performance/behavior of the student. The result shows that the accuracy of model is approx. 89% and shows improved performance with existing methods by approx. 14% in terms of accuracy.
ii. This model is quite efficient for finding eligible students for finding and sorting the best and deserving students.
iii. This model also gives motivational message to existing student to improve their performance.
iv. This model can also decide to give forecasting for shortage of attendance of students with respect to department.
v. This decision support system also efficient with respect to time.

In future this work would implement over internet of things (IoT) applications that can enrich the educational system with more technologies and applications. The application of IoT will enhance the technique for further analysis.

REFERENCES

[1] K. Abe, "Data Mining and Machine Learning Applications for Educational Big Data in the University," 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Fukuoka, Japan, 2019, pp. 350-355.
[2] L. Yu, X. Wu and Y. Yang, "An Online Education Data Classification Model Based on Tr_MAdaBoost Algorithm," in Chinese Journal of Electronics, vol. 28, no. 1, pp. 21-28, 1 2019.
[3] Z. Li, "New Employee Student Repast Big Data Analysis Research Application," 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Vientiane, Laos, 2020, pp. 583-586.
[4] K. J. de O. Santos, A. G. Menezes, A. B. de Carvalho and C. A. E. Montesco, "Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout," 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), Macei??, Brazil, 2019, pp. 207-208.
[5] S. S. Al-Nadabi and C. Jayakumari, "Predict the selection of mathematics subject for 11th grade students using Data Mining technique," 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), Muscat, Oman, 2019, pp. 1-4.
[6] N. Ketui, W. Wisomka and K. Homjun, "Using Classification Data Mining Techniques for Students Performance Prediction," 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), Nan, Thailand, 2019, pp. 359-363.
[7] Z. Shao, H. Sun, X. Wang and Z. Sun, "An Optimized Mining Algorithm for Analyzing Students' Learning Degree Based on Dynamic Data," in IEEE Access, vol. 8, pp. 113543-113556, 2020.