EasyChair Preprint

№ 2016

Student Performance Analysis at Secondary Level Using Machine Learning

Prashant Johri, Pallavi Goel, Sandesh Kumar Srivastava, Sashant Suhag and Sanskriti Jadaun

November 21, 2019

Student Performance Analysis at Secondary Level Using Machine Learning

Prashant Johri[1] , Pallavi Goel[1] , Sandesh Kumar Srivastava[1], Sashant Suhag[1] , Sanskriti Jadaun[1]

johri.prashant@gmail.com, pallavi30nov@gmail.com, sandesh.srivastava7@gmail.com, sashantsuhag23@gmail.com, sanskritijadaun14@gmail.com

Galgotias University, India-201310.

Abstract:

In today's era of information and technology, Big Data & Machine Learning are emphasising a very great impact on every sector i.e. education, healthcare, entertainment etc. Researches have shown that education is the most important juncture in an individual's life, with the help of Big Data Analytics it has become very easy to identify the consequences and challenges evolving in the education sector with every passing day. This paper mainly emphasises on two factors, which are, reasons for the variations in marks of the student from secondary school level to higher secondary school level and secondly, solutions to solve the above-mentioned problem using Machine Learning and develop a model which may be capable of predicting the marks easily at secondary level with the least amount of errors involved.

Keywords: Big Data, Big Data Analytics, Machine Learning, Hadoop, R programming, Regression Coefficient(R-Squared), Education sector, Secondary school, Higher Secondary school

1. Introduction

The emergence of Big Data and its analysis using Machine Learning in the education sector has shown remarkable results. Through the use of Big Data Analytics, it has become very easy to address the challenges evolving in the education sector. After studying the data of the result of past three years of students studying in Galgotias University, India, we noticed that there is a very big variation in most of the student's secondary school and higher secondary school marks. There may be many reasons for the deterioration in the student's marks which would be discussed in the paper and the ways to improve the performance of students. We have taken a sample the data of marks of students studying at Galgotias University and implemented Machine Learning on the given dataset for studying the scenario and get desired consequences where we get a model using which we can predict the marks of the students at secondary level.

Big Data has become a "buzz word" in recent years and has become prominent as an emerging technology. This apparent technology is something that can "lead to the rapid development of new capabilities" and inspire about new opportune for and problems in addressing the global issue" [1]. In simple language, Big Data is very large structured or unstructured data which is growing very rapidly every passing day.

In a 2001 analysis report [2], META Group (now Gartner) introduced data growth challenges and opportunities as a three-dimensional thought, i.e. increasing volume (increase inside the standard of data), velocity (growing of data at degree fearsome speed), and variety (types of data). After the formation of the "3Vs" by Gartner, all the sectors where big data was developing, began to use the concept of the "3Vs" model for outlining "what is big data" [3]. In 2012, Gartner updated its definition: "Big Data is high-volume, high-velocity and/or high-variety info assets that demand of cost-efficient, innovative varieties of information science that perpetually modify increased insight, higher cognitive process, and method automation" [4]. Gartner's definition of the 3Vs became standard worldwide and continues to be universal. Consistent with associate degree agreement with a compatible definition that states that "Big Data represents the data assets characterised by such a High Volume, Velocity and Variety to want specific Technology and Analytical ways for its transformation into Value" [5]. After the emergence of 3 Vs, two more Vs came into the picture i.e. "Value" and "Variability" were accessorial by some alliances to explain it [6]. modification summons by some companies [7].

Big data can be delineated by the subsequent characteristics [8],[9]:

Volume: It describes in what amount the data is being received or being transferred. Normally, it deals with data in petabytes.

Variety: It defines the sort of data (structured and unstructured data) which is being employed or processed.

Velocity: During this context, we tend to get to grasp concerning the speed in which the data is being generated to fulfil the approaching or assigned challenges.

Variability: It essentially defines the variation and inconsistency of data which might damage the information.

Value: This term describes the worth of the data and is taken into account to vital from corporates perspective as a result of each data has some value.

Fig 1: Five V's of Big Data

If we glance at today's scenario Big data is growing with terribly tremendous speed and is evolving in each field wherever data keeps progressing. A number of sectors where Big Data is in trend square measure Education, healthcare, recreation and communication etc.

Big Data isn't solely being highlighted within the IT sectors and corporate world, however additionally in education. A late correct vital similarity of the many approaches taken by 35 charter colleges in NYC has discovered that one in every of the highest five policies connected with significant academic effects was the utilization of data to guide instruction [10]. Several technologies square measure evolving with the collaboration of Big Data in it. One in each of the technologies from that emergence is cloud computing. These new technologies are very

helpful and can improve the structure of educational services and sectors, can provide students access to study material at allowing cost, online teachers, and a faction of co-learners [11]. Additionally, as West says, “Big Data can support the classic educational system helping teachers to analyse what students know and what techniques are most effective for each pupil.” If we have a tendency to add this manner then it might be very simple to take care of the students and the lecturers, to get acknowledged about the new techniques and methods evolving in the education sector [12]. Therefore, Data analytics being a brand-new technology will offer students and teachers a quick assessment of their academic performance. With the study of these methods, we can also provide students and teachers with a great platform for the analysis of various educational patterns and extricate valuable knowledge and information. Similarly, with the assistance of an outsized variety of data set it can be predicted, which student desires a lot of attention for the improvement in their performance from the education system so that the rate of failure may be evaded and rate of performance of every student may be enhanced [13][14].

2. Literature Review:

In 2012, Gartner updated its definition: "Big Data is high-volume, high-velocity and/or high-variety info assets that demand of cost-efficient, innovative varieties of information science that perpetually modify increased insight, higher cognitive process, and method automation" [4]. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful information for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. Big data and its analysis are at the center of modern science and business. These data are generated from online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, social networking interactions, science data, sensors and mobile phones and their applications. They are stored in databases grow massively and become difficult to capture, form, store, manage, share, analyze and visualize via typical database software tools.[15]

Najafabadi et al. [16] focused on deep learning, but noted the following general obstacles for machine learning with Big Data: unstructured data formats, fastmoving(streaming) data, multi-source data input, noisy and poor-quality data, high dimensionality, scalability of algorithms, imbalanced distribution of input data, unlabelled data, and limited labelled data. Similarly, Sukumar [17] identified three main requirements: designing flexible and highly scalable architectures, understanding statistical data characteristics before applying algorithms; and finally, developing ability to work with larger datasets. Both Najafabadi et al. [16] and Sukumar [17] reviewed aspects of machine learning with Big Data; however, they did not attempt to associate each identified challenge with its cause. Moreover, their discussions are on a very high level without presenting related solutions. In contrast, our work includes a thorough discussion of challenges, establishes their relations with Big Data dimensions, and presents an overview of solutions that mitigate them.

Junfei Qiu referred Machine leaning as a highly interdisciplinary field building upon ideas from many different kinds of fields such as artificial intelligence, optimization theory, information theory, statistics, cognitive science, optimal control, and many other disciplines of science, engineering, and mathematics. Because of its implementation in a wide range of applications, machine learning has covered almost every scientific domain, which has brought great impact on the science and society. It has been used on a variety of problems, including recommendation engines, recognition systems, informatics and data mining, and autonomous control systems.[18]

3.2. Big data analysis using machine learning:

Big Data analysis using Machine Learning is one of the most commonly used data analyses. In this form of data analysis, the machine is trained on the basis of the data that may be classified or unclassified. In the entire process of analysis, the dataset is trained on a machine learning algorithm and model is derived on the basis of the results obtained.

This model is then tested with a new dataset set which may or may not be classified. Certain inferences are obtained from the results produced on this data and on that basis a new and error less mode is defined which is capable of using machine learning to give the desired results.

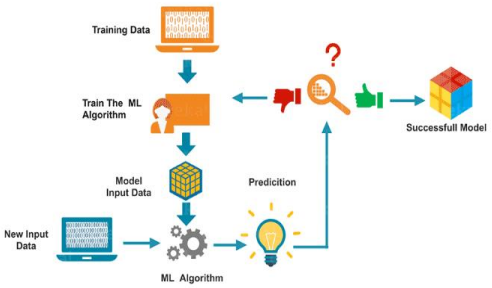

Fig 3: Big Data analysis using Machine Learning

Importance of Big Data Analytics in Education Sector

Big data analytics helps in recognising the opportunities and challenges occurring in every firm. In education, Big Data Analytics provides various opportunities for learning and teaching experience for both students and instructors. Through the implementation of data analytics in education, it becomes very easy to identify the flaws arising in educational institutes and also helps students to enhance their academics and perform better [19].

Benefits of Big Data Analytics in Education:

1. With the help of analytics, we can predict future marks according to their past records so that they can enhance their results and perform better.

2. Students must take any course according to their calibre, with the help of big data analytics we can find courses according to students’ interest so that they can perform better.

3. Matching students to various Companies on the basis of their skills and abilities.

4. Prioritise Transparent Education Financing by which students must be able to indulge themselves in various academic activities and can find better courses for their higher education according to their capability.

3.3. Tools and Methodology Used

In this paper, we have used R and Hadoop the two important platforms of Big Data Analytics for analysing the marks of the students from their secondary school level to the higher secondary school level.

4. HADOOP & R Programming

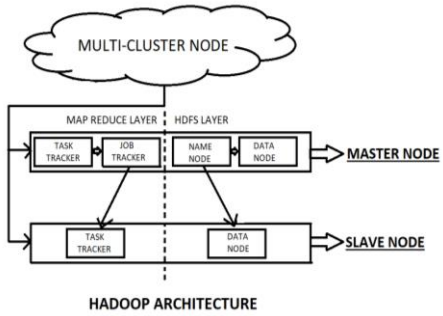

Fig 4: Hadoop Architecture

[20] Hadoop is a platform which was introduced to manage the big amount of data. It is a framework which is used to deal with the processing of large data sets. Hadoop was developed by Google's MapReduce which is a framework where an application breaks down into various parts. The Apache Hadoop ecosystem comprises of the Hadoop kernel, MapReduce, HDFS and numbers of various components like Apache Hive and Base.

R is an [21] open-source statistical programming language and environment which freely available for access to all mainstream operating systems. R basically with the help of data sets and programs compute statistics and graphs. It is very easy to access as it is free and is also very easy to understand.

5. Proposed Work

Every year when class 10th and 12th results get declared the various media channels focus on the passing rate of students as compared to previous year results, but no one focuses on the deviation in students' marks, which is growing every year. This paper shows the analysis of variation in students marks from their middle school to high school. For this, we have taken data sets of class 10th and 12th results of students of India. After studying the results of a student's class 10th and 12th of the past 3 years, we came across noticing the deviation in the marks of most of the students. With the help of R and Hadoop, we have done the analysis on student's marks. After analysing the graph of each year, we came across noticing that there is a deviation in most of the student's marks. We use dataset shown in Fig. 5, Fig.6 and Fig. 7.

I. No.	Name	10th CGPA/%	12th%
1	Lokesh Vaishnav	88.50	77.20
2	Meghna Kumar	90.04	87.00
3	Mohit Kumar Dhiman	78.00	84.00
4	Neha Varshney	98.00	86.00
5	Nishant Singh	88.00	65.50
6	Parikshit Anand	90.05	82.40
7	piyush kumar	88.20	79.00
8	PRAFULLA MISRA	91.66	83.60
9	Rishal Sinha	96.00	75.80
10	Saurav Chandra	77.00	65.20
11	SHAANI ARYA SRIVASTA	82.02	68.00
12	shammy raza	77.68	82.70
13	Shanu chauhan	94.25	77.60
14	SHIVAM KUMAR SINGH	96.00	82.00
15	Shivani	86.20	70.00
16	Shubham Lal	76.00	71.30
17	Shubham Modi	95.00	92.20
18	Siddhesh Ratna Singh	90.02	72.20
19	Sumedh Kumar	88.60	73.00
20	Suyash Ratnam	74.00	80.00

Fig 5: Data set of marks of 2015 students

.No.	Name	10th CGPA/%	12th/Diploma%
1	Swapnil shekhar	72.00	64.60
2	Tanisha katiyar	91.64	77.40
3	Tanisha Srivastava	90.06	85.75
4	UTTAM KUMAR	82.00	71.00
5	VIJAY PRATAP PANDEY	81.50	68.02
6	Vikas singh	77.00	68.00
7	Vishal Keshri	84.50	60.00
8	VISHAL YADAV	63.66	61.80
9	YASHI GROVER	94.00	81.00
10	Abhishek kumar	71.00	64.00
11	Abhishek Kumar Sahu	86.00	73.60
12	Aman Chandra	82.00	78.00
13	Amrit lal gupta	72.00	60.05
14	Hemant Parihar	86.00	82.60
15	ishita kathpalia	85.20	75.00
16	Kunwar Harshit Shah	90.00	82.00
17	MAYANK PRIYADARSHI	77.20	78.60
18	Meenakshi meenu	91.20	85.00
19	MOHAMMAD SHAMIM SIDDIQ	92.00	88.60
20	Narendra Vikram Pal	76.50	79.00

Fig 6: Data set of marks of 2016 students

S. No.	Name	10th %/CGPA	12th/Diploma %
1	Aanchal mishra	95.00	83.00
2	Abhay Kumar	80.00	77.00
3	Abhay Saini	88.00	74.60
4	Abhijeet Panwar	78.00	62.80
5	Aishwarya Bhatia	95.60	89.50
6	amit kumar	68.60	64.60
7	ananya tiwari	90.00	83.40
8	Ankit Raja	91.00	91.00
9	Apala Agarwal	77.60	79.75
10	Chetan Malhotra	86.00	76.00
11	Chinmay Mourya	88.00	73.50
12	Deepak Tomar	90.25	81.00
13	Deepanshi Kumar	92.00	79.00
14	Devesh Gupta	76.00	66.00
15	Devesh Ratna Singh	94.60	83.90
16	Dhruw Kumar	77.78	84.40
17	Divyanshu Pathak	92.40	83.20
18	KANIKA GOSAIN	94.00	88.00
19	KISHOR KUMAR	61.50	74.60
20	Kushal Mishra	72.60	65.00

Fig 7: Data set of 2017 students

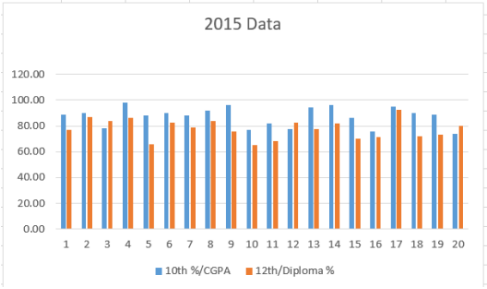

Fig 8: Graph showing 2015 results in variation

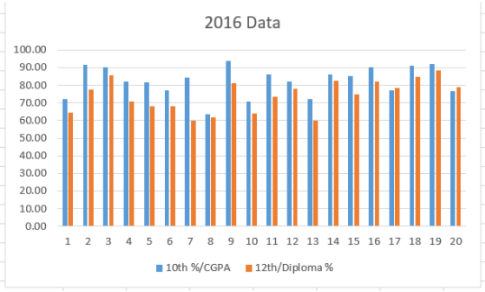

Fig 9: Graph showing 2016 results in variation

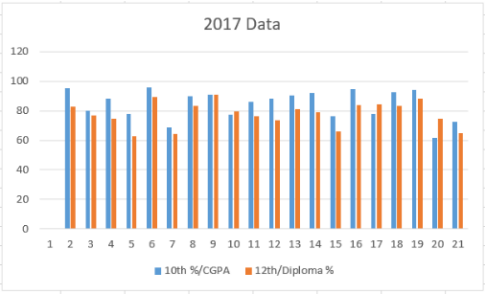

Fig 10: Graph showing 2017 results in variation

Percentage decrease in 2015:
Average marks of students in 10th standard= 84.44
Average marks of students in 12th standard= 78.01
Decrease in marks= 84.44 - 78.01 = 6.43
Percent decrease= 6.43 / 84.44 * 100 = 7%
Percentage decrease in 2016:
Average marks of students in 10th standard= 82.27
Average marks of students in 12th standard= 74.20
Decrease in marks= 82.27 - 74.20 = 8.07
Percent decrease= 8.07 / 82.27 * 100 = 9%
Percentage decrease in 2017:
Average marks of students in 10th standard= 87.26
Average marks of students in 12th standard= 77.73
Decrease in marks= 87.26 - 77.73 = 9.53
Percent decrease= 9.73 / 87.26 * 100 = 10%
NOTE: -
The value of percentage decrease has increased in every year i.e. student marks in class 12 are degrading more every year than student marks in class 10.

6.1. Analysing the changes using Machine learning
Regression Analysis is one of the most common methods used in machine learning. It is a form of predictive modelling technique the helps us find the relation between a dependent variable and an independent variable. In other words, it involves graphing a line over a set of data points that most closely fits the overall shape of data.
Three major uses of regression analysis are:

- Determining the strength of predictors.
- Forecasting an effect, and
- Trend forecasting

Where is the Linear regression used?

- Evaluating Trends and Sales Estimates
- Analysing the impact of price changes
- Assessment of risks in financial services & insurance domain.

Computing or predicting marks of students at secondary level:
Let us consider the equation of the straight line, that is:

$$Y = m X + C$$

Here we assume Y as the prediction we need to make according to the given values of X.
What we do here is we plot a graph of Y vs X and hence compute:

- The slope or the **m** value of the equation from the graph.
- Y-intercept or the **C** value from the graph.

Pseudocode to generate the slope and y-intercept:

```
Compute_slope_intercept (){
Read [reads the dataset] //reads the dataset
X ←[marks of tenth],Y ← [marks of twelfth]
Mean_x ← mean marks of class tenth
Mean_y ← mean arks of class twelfth
Loop ← length (X){
numerator = numerator + (X[loop] – Mean_x) * (Y[loop] – Mean_y)
denominator = denominator + Square_Of [(X[loop] – Mean_x)]} //end
loop
Slope = numerator/denominator
Intercept = Mean_y – (Slope * Mean_x)}//end function
```

6.2. Regression Coefficient:

- R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model.

- In investing, R-squared is generally interpreted as the percentage of a fund or security's movements that can be explained by movements in a benchmark index.
- An R-squared of 100% means that all movements of a security (or dependent variable) are completely explained by movements in the index (or the independent variable(s) you are interested in).
- An R-squared of 70% or less, indicates that the movement of the index is not followed and the model does not work for all the indexes.
- An R-Squared value ranging between 70% to 90% indicates that the model works for but the predictions done by the model are not very accurate and may have chances of error.

Pseudocode to generate regression coefficient:

```
Regression (){
Read [reads the dataset] //reads the dataset
X ←[marks of tenth], Y ← [marks of twelfth]
Mean_x ← mean marks of class tenth
Mean_y ← mean arks of class twelfth
Loop ← length (X){
y_pred = Intercept + Slope * X[loop]
rsq_t = rsq_t + Square_Of [(Y[loop] – Mean_y)]
rsq_r = rsq_r + Square_Of [(Y[loop] – y_pred)]}//end loop
R_Square_Value = 1 – (rsq_r / rsq_t)}//end Regression
```

Graphs to compare the R-Squared values:

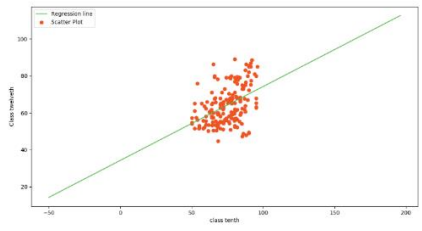

Fig. 11: Graph showing scattered points and a very less R Squared value

The above Fig. 11 shows the graph of the tenth vs twelfth marks of students. Since the plotted points are very scattered and the computed value of R-square here is **0.18852674369757294**. Therefore, the R-squared value remains as 18.85% which is comparatively a very low R-Squared value.

Therefore, the above model is not capable of being a perfect model and hence the predictions cannot be possible.

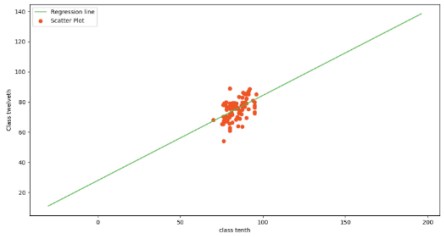

Fig.12: Graph showing points close to the regression line and an average R-Squared value

In Fig. 12 the graph of tenth vs twelfth marks of students gives the point very near to the best fit line. Since the points are very close the R-Squared value here comes out to be **0.71992631797897244**. Therefore, the R-Squared value remains as 71.99% which is comparatively an average R-Squared value.

As the R-Squared value is more than 70%, we can conclude that the model here is capable of predicting the value but also as the R-Squared value is not very high., it is not capable of predicting correctly on all the indexes.

Limitations of R-Squared:

R-squared will give you an estimate of the relationship between movements of a dependent variable based on an independent variable's movements. It doesn't tell you whether your chosen model is good or bad, nor will it tell you whether the data and predictions are biased. A high or low R-square isn't necessarily good or bad, as it doesn't convey the reliability of the model, nor whether you've chosen the right regression. You can get a low R-squared for a good model, or a high R-square for a poorly fitted model, and vice versa.

6.3. Reasons for Deviations in Marks

As we can see, in Fig. 8, Fig. 9 and Fig.10 that we have three graphs with students' class 10^{th} and 12^{th} marks. In each of, we examined that most of the students scored less than their 10^{th} marks. After doing a small survey to know the reasons for the variation in marks with some students of each year, we discovered many factors which prime reasons for the deteriorations in are students' marks that are as follows.

- **Distraction:** Mostly students get access to any or all the technologies i.e. mobile phones, web etc after scoring good results in results in 10^{th}. And as we know excess of anything is bad, similarly when students at once get the access of technologies they indulge most the time over their or internet which on the other hand hamper their result.
- **Curriculum Design:** If you undergo the curriculum of CBSE, you will observe that the syllabus till 10^{th} the syllabus is very easy and when it comes to $12^{th}$the syllabus increases and most of the students are not able to cope up with it.
- **The choice for subject selection:** This is a very important factor for the failure of students in scoring good marks. Most of the students without knowing the future consequences select subjects. Some students due to their parents' pressure choose a subject in which they don't have an interest and, in the end, hamper their result. Some students take subjects in which they are not interested because their friends are choosing another subject.
- **Preparation for various exams:** In class 12th everyone sets a goal of scoring good marks in the entrance exam. Which is an MCQ type exam and the board exams are descriptive type. On the other hand, in 10th student has no other aim than scoring in boards. So as per the goal at that time student prepares. Once the student gains the habit of solving MCQs, writing the long answers may bore him/her. And also, he/she might begin framing their own sentences round the vital terms that result in lesser marks.

6.4. Solutions for Improvement in Result

For solving the problem of marks deterioration of students we must work on the above-mentioned factors which are hampering students' class 12^{th} results. Students must be awarded about the consequences of excessive use of technologies and must be advised to focus on their studies to enhance at performance for their future endeavours. Listed below are some points which must be taken under consideration to improve the results of students.

- Students must be given subjects on their interest. They should not be pressurized to take any subject which they don't understand.
- Should indulge themselves more in studies than getting distracted.
- Should focus on one thing in spite of thinking about doing many things at one time.
- Should be stressed free and focus on studies.
- Students should set a goal that they have to score better than their past results and then prepare for exams accordingly.

7. Result

After analysing the data of students, we come across several factors which resulted in the deviation of marks of students. Students need to overcome all these factors to score good marks at secondary level. If we do not improve now, this trend will ruin students' academics marks in the years to follow.

With the help of big data analysis, we can now analyse the marks of several years and using machine learning with this analysis, create a model which can predict the marks of students at secondary level with greater accuracy and hence this may help the student focus more on their studies and how they can improve their marks if needed. As we can see in fig.13 there is a very slight difference in the predicted value vs the obtained value, hence the students can evaluate themselves on the basis of this prediction and can work accordingly to score better marks.

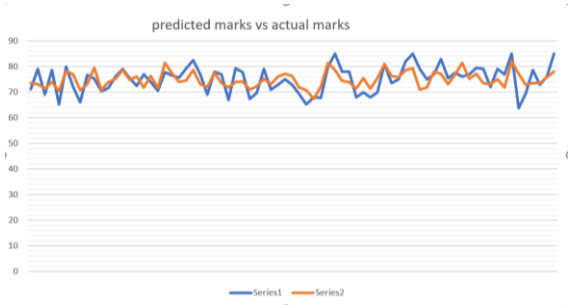

Fig 13: Variation in predicted values Vs Obtained Values

	A	B	C	D
1	10th%/CGPA	12th %/CGPA	predictded 12th%/CGPA	difference
2	81	71.29	73.59497624	-2.304976
3	80	79	73.03113919	5.968861
4	77	69	71.33962803	-2.339628
5	81.8	78.6	74.04604589	4.553954
6	75.17	65.2	70.30780622	-5.107806
7	89.17	80	78.20152498	1.798475
8	87	72	76.97799857	-4.977999
9	76	66	70.77579097	-4.775791
10	80	76.8	73.03113919	3.768861
11	91.33	75.2	79.41941301	-4.219413
12	76	70.4	70.77579097	-0.375791
13	81.6	71.6	73.93327848	-2.333278
14	84	76.2	75.28648741	0.913513
15	90	79	78.66950973	0.33049
16	83.33	75.4	74.90871658	0.491283
17	85.4	72.4	76.07585928	-3.675859
18	78	77	71.90346508	5.096535
19	86	74	76.41416151	-2.414162
20	77.67	70.5	71.71739885	-1.217399

This way the Big Data Analysis using the concept of Machine Learning can stop the ongoing trend of the decreasing marks and hence is capable to bring a change henceforth.

8. Conclusion and Future Work

After studying we came to a result that, Big Data along with Machine Learning are important platforms which are very much helpful for people in every sector. When it comes to the education field, Machine Learning with its Analytics techniques plays a very vital in shaping an individual present and future by predicting their future performance with the help their past records. In this growing age of information and technology, Machine Learning is overcoming the traditional challenges to enhance the efficiency of education.

As in this paper, author worked on big data analytics using machine learning to analyse students results and predicted reasons which are hampering students present result when we compared it with their past results. Furthermore, with this growing trend of analysis technique, we can predict the future of education on both students and teacher's perspective to give better results and enhance student's performance.

References

[1] S. A. Hossain, "Big data analytics in education: Prospects and Challenges," in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)*, 2015, p. 1.

[2] D. Laney, "3d data management: Controlling data volume," *Veloc. Var.*, vol. 6, no. 01, p. 2001, 2001.

[3] J. G. Enr\'\iquez, F. J. Dom\'\inguez-Mayo, M. J. Escalona, J. A. Garc\'\ia Garc\'\ia, V. Lee, and M. Goto, "Entity Identity Reconciliation based Big Data Federation-A MDE approach," 2015.

[4] "Big Data Analytics - Gartner Tech Definitions.".

[5] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of Big Data based on its essential features," *Libr. Rev.*, vol. 65, no. 3, pp. 122–135, 2016.

[6] F. Martin-Sanchez and K. Verspoor, "Big data in medicine is driving big changes," *Yearb. Med. Inform.*, vol. 23, no. 01, pp. 14–20, 2014.

[7] S. Grimes, "Big data: Avoid 'wanna V'confusion," *Inf. Week*, vol. 7, no. August 2013.

[8] M. Hilbert, "Big data for development: A review of promises and challenges," *Dev. Policy Rev.*, vol. 34, no. 1, pp. 135–174, 2016.

[9] P. Ongsulee, V. Chotchaung, E. Bamrungsi, and T. Rodcheewit, "Big Data, Predictive Analytics and Machine Learning," in *2018 16th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 2018, pp. 1–6.

[10] A. Bhatia and G. Vaswani, "Big data-a review," *IEEE Int. J. Eng. Sci. Res. Technol. IJESRT*, 2013.

[11] D. M. West, "Big data for education: Data mining, data analytics, and web dashboards," *Gov. Stud. Brookings*, vol. 4, no. 1, 2012.

[12] G. Siemens and D. Gasevic, "Guest editorial-Learning and knowledge analytics.," *Educ. Technol. Soc.*, vol. 15, no. 3, pp. 1–2, 2012.

[13] N. Elgendy and A. Elragal, "Big data analytics: a literature review paper," in *Industrial Conference on Data Mining*, 2014, pp. 214–227.

[14] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make a big impact: Findings from a systematic review and a longitudinal case study," *Int. J. Prod. Econ.*, vol. 165, pp. 234–246, 2015.

[15] [2013 International Conference on Collaboration Technologies and Systems (CTS)----by Seref Sagiroglu ; Duygu Sinanc]

[16]M.M.Najafabadi,F.Villanustre,T.M.Khoshgoftaar,N.Seliya,R.Wald, and E. Muharemagic, ''Deep Learning Applications and Challenges in Big Data Analytics,'' J. Big Data, vol. 2, no. 1, p. 1, Feb. 2015.

[17] S. R. Sukumar, ''Machine learning in the big data era: Are we there yet?'' in Proc. 20th ACMSIGKDD Conf. Knowl. Discovery Data Mining, Workshop Data Sci. Social Good (KDD), 2014, pp. 1–5.

[18] Qiu et al. EURASIP Journal on Advances in Signal Processing (2016) 2016:67

[19] X. Sun *et al.*, "Towards delivering analytical solutions in the cloud: Business models and technical challenges," in *2011 IEEE 8th International Conference on e-Business Engineering*, 2011, pp. 347–351.

[20] D. Fallik, "For big data, big questions remain." 2014.

[21] H. S. Bhosale and D. P. Gadekar, "A review paper on big data and Hadoop," *Int. J. Sci. Res. Publ.*, vol. 4, no. 10, pp. 1–7, 2014.

[22] K. Kelley, K. Lai, and P. Wu, "34 USING R FOR DATA ANALYSIS A Best Practice for Research," 2008.

[23] J. Bughin, M. Chui, and J. Manyika, "Clouds, big data, and smart assets: Ten tech-enabled business trends to watch," *McKinsey Q.*, pp. 75–86, 2010.