UNCOVERING PATTERN IN EDUCATION DATA THROUGH MACHINE LEARNING*

BY

[1]MS. MUSKAN AGRAWAL♣, [2]MS. NEHA VEKHANDE

[1]B.Tech in Computer Engineering, Ajeenkya DY Patil University, Pune, India

muskan.agrawal@adypu.edu.in

[2]Assistant Professor, School of Engineering-ADYPU, Pune, India

neha.vekhande@adypu.edu.in

ABSTRACT

The main aim of machine learning is to extend accuracy, gain knowledge and maximize the performance of the machine on its task performed. Machine learning allows the system to be told new things from knowledge by making self-learning algorithms. In ML large problem is subdivided into several small tasks and at the end are combined to build a machine learning model. A sufficient amount of data can build a good model. Similarly, with the help of ML today every task has become very easy to perform with high accuracy. In this paper, we will see how machine learning has influenced education system in today's scenario. When every knowledge sector is going online machine learning, has proved to be boon to our education system by providing highest accuracy in students' performance, field of interest, as well as result prediction. Here, in this research the use of classification algorithms such as Decision Tree, Naïve Bayes Classifier, Random Forest and Logistic Regression will be learned. Whereas, Logistic Regression and Random Forest algorithms will be performed on the input data to transform specifically using Python Jupyter Notebook.

KEYWORDS

Machine Learning Algorithms, Artificial Intelligence, Educational Data Mining.

I.INTRODUCTION

Today, machine learning is altering the field of education in every aspect. It is fundamentally changing the way of teaching, learning and analyzing things. Education is no longer about teaching of texts or memorizing the manuscripts. Machine learning has turned out to be an innovative tool which is one of the strongest newer technology to find hidden insights without being explicitly programmed to do so. It not only works as a good predictive method but also

* Received 08 October 2021, Accepted 26 October 2021, Published 13 November 2021

♣ Corresponding Author

plays a significant role in broadening the learning system and enhancing the fundamentals of curriculum making learning more effective and resourceful. Machine learning methods are providing a new infrastructure to educational field by adapting new advanced technologies [1]. The overall assessment processes are upgraded. It is now more streamlined, accurate and unbiased with the help of machine learning technology.

The customized teaching and learning approach explore the student's background, reviews the individual aptitude, examines the learning and response time to provide the feedback to teacher. It acts like a black-box when the data is input, complex processing of data takes place and then the patterns and insights are generated as an output. This helps in contemplating student's attention as well as improves his overall participation. In simple terms, machine learning is proven to be of great importance for decision making process and analyzing the individual student's data. It has tremendously improved the learning experiences by providing the better results. With the help of machine learning algorithms, traditional learning systems is being suppressed where the goal was to complete the course without ensuring that everyone has got it [2]. Use of ML has improved the information perception and evaluated the learning progress.

II. LITERATURE SURVEY

Education is moving away from conventional chain of students. Textbooks have been exchanged with computers and e-books. Today classrooms are getting well-equipped with digital resources. This era is evolving technology was not at all possible without machine learning and artificial intelligence. Machine learning is playing a great role in development of computer programs without human intervention. One must believe that machine learning is future of the world. It is giving power to intelligence by predicting accurate outcomes. This paper research is completely based on educational data and how machine learning is helpful when implemented with educational dataset of different students. Some of the important outcomes of this research includes:

i.Support Teachers: Machine learning is very important for data mining. Earlier teachers were dependent on gradebooks. But now teachers can have access to each and every student's record at a single click in one repository. The application of machine learning can improve lessons by letting teachers know about in which topic the most of the students are facing problem.

ii.Predict Student Performance: Machine learning is one of the most efficient tools with the ability to predict accurate outcomes. By exploring the available student data and performing various algorithms on it, the model can identify the weaknesses and strength of a particular or

cluster of students. It learns from the education dataset and previous information to predict student performance.

iii.Customized Learning: The use of machine learning in Education Data Mining (EDM) can ease the learning process. Customized learning became trend in evolving world of technology. Digital resources started fulfilling the needs of current situation. Teachers are now able to assist which student requires how much time to learn a certain subject and what is the response time of each student. Visual learning and individualized instruction proved to connect more students towards learning. EDM can be used to group students and teachers according to their needs and availability [3].

iv.Test Students: The techniques of machine learning in educational field are providing AI based assessments to dispense constant feedback to both teachers as well as students. This not only calculates the progress of students but also puts check on their daily assessment and task. Thus, helps organizing the content effectively to grade all students fairly.

III. IMPLEMENTATION OF PYTHON JUPYTER NOTEBOOK

- Open Jupyter Notebook (in desired directory)
- Import the libraries and packages
- Load the dataset (education dataset)
- Find out the missing values and remove the blank records [5]
- Normalize data
- Split the dataset into training and test set
- Perform classification
- Apply the algorithm (Random Forest, Decision Tree, Naïve Bayes Classifier, Logistic Regression)
- Refine the values
- Observe the accuracy

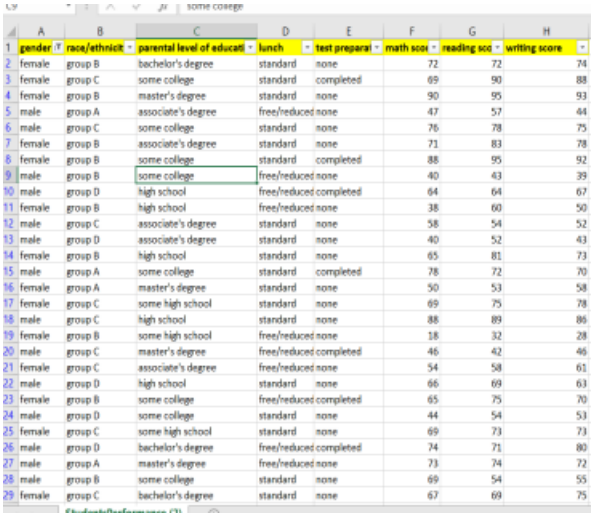
IV. DIFFERENT ML APPROACHES USED ON EDUCATION DATA

ML is a subset of AI that helps computers to learn from previous data available and predict intelligent decisions. It analyzes data from different perspectives and summarizes it into important information to identify hidden patterns and draw insights from large and complex datasets. To improve the learning process, use of knowledge and knowledge mining tools helps in extracting useful and unknown information about student result in repositories [4].

The main goal of this paper is to predict the student performances using different classification techniques and build models with the help of scikit-learn libraries in Python Jupyter Notebook.

This model uses training dataset to obtain better boundary condition and determines the target class. The research foresees to predict the accuracy of model classification algorithms such as Decision Tree, Random Forest, Logistic Regression and Naïve Bayes Classifier is performed.

Let's have a look on the dataset which we will be using in our research:



	A	B	C	D	E	F	G	H
1	gender	race/ethnicity	parental level of educati	lunch	test preparat	math scor	reading sco	writing score
2	female	group B	bachelor's degree	standard	none	72	72	74
3	female	group C	some college	standard	completed	69	90	88
4	female	group B	master's degree	standard	none	90	95	93
5	male	group A	associate's degree	free/reduced	none	47	57	44
6	male	group C	some college	standard	none	76	78	75
7	female	group B	associate's degree	standard	none	71	83	78
8	female	group B	some college	standard	completed	88	95	92
9	male	group B	some college	free/reduced	none	40	43	39
10	male	group D	high school	free/reduced	completed	64	64	67
11	female	group B	high school	free/reduced	none	38	60	50
12	male	group C	associate's degree	standard	none	58	54	52
13	male	group D	associate's degree	standard	none	40	52	43
14	female	group B	high school	standard	none	65	81	73
15	male	group A	some college	standard	completed	78	72	70
16	female	group A	master's degree	standard	none	50	53	58
17	female	group C	some high school	standard	none	69	75	78
18	male	group C	high school	standard	none	88	89	86
19	female	group B	some high school	free/reduced	none	18	32	28
20	male	group C	master's degree	free/reduced	completed	46	42	46
21	female	group C	associate's degree	free/reduced	none	54	58	61
22	male	group D	high school	standard	none	66	69	63
23	female	group B	some college	free/reduced	completed	65	75	70
24	male	group D	some college	standard	none	44	54	53
25	female	group C	some high school	standard	none	69	73	73
26	male	group D	bachelor's degree	free/reduced	completed	74	71	80
27	male	group A	master's degree	free/reduced	none	73	74	72
28	male	group B	some college	standard	none	69	54	55
29	female	group C	bachelor's degree	standard	none	67	69	75

Fig. 1 EDUCATION DATASET

```
gender                         0.0
race/ethnicity                 0.0
parental level of education    0.0
lunch                          0.0
test preparation course        0.0
math score                     0.0
reading score                  0.0
writing score                  0.0
dtype: float64
```

Fig. 2 FINDING OUT THE MISSING VALUES IN EDUCATION DATASET

A.DECISION TREE

Decision tree is a flowchart structure which is used to continuously split the data based certain parameters. The leaves of decision tree are termed as final outcomes and nodes are point at which the data is being split. This algorithm is useful for solving problems related to regression and classification.

WORKING: This algorithm creates training model which is used to predict the class and value of target variable. When the Decision Tree algorithm is applied on training dataset, the whole education dataset is considered as a root and feature values are taken to be categorical. Before the creation of model, it determines whether the values are in continuous form or not. In case of continuous values, it is discretized (transformed into discrete counterparts) to make it suitable

for numerical evaluation. The records are then distributed recursively on the basis of attribute values as root or internal nodes. This procedure works on the principle of statistical approach to achieve good results in student's performance.

B.NAÏVE BAYES CLASSIFIER

Naïve Bayes Classifier is one of the most simple and effective classification algorithms which is used to built models at fast pace as compared to other machine learning methods. This classifier is also known as probalistic classifier which is familiar to predict outcome on the basis of probability of an object. Naïve Bayes Classifier is named as Naïve because of the assumption that the occurrence of certain feature is independent of other features present.

WORKING: Naïve Bayes Classifier makes prediction with the help of probability to produce better results. When the implementation of Naïve Bayes algorithm is done on education dataset, it calculates the probability of data and separates the training data by class. Then summarization of dataset is made with the use of statistics. The mean and standard deviation is performed on the dataset to summarize it by class. Some of the Naïve Bayes algorithm used are Gaussian Naïve Bayes which includes normal distribution, Multinomial Naïve Bayes for multinomial distribution and Bernoulli Naïve Bayes for multivariate Bernoulli distribution.

*Note: In this paper, the implementation of below two algorithms is done on education dataset.

C.RANDOM FOREST

Random Forest is an ensemble technique where lot of decision trees are used. But each tree is a little different from one another. Here, when we get a new data, we take the majority vote of the ensemble to get a final result. It is considered to be one of the most powerful and highly accurate to build models without the use of normalization or scaling. It runs the trees in parallel so that the performance does not get affected.

WORKING: Random Forest algorithm is performed in two stages. In the first stage random forest is created and in the second one outcome is predicted from the random forest classifier created in first stage. Implementation of student data and then performing random forest creation by randomly selecting 'k' features from total 'm' features, where we usually say that k<<m. Using the best split point, the calculation of node and then splitting it into daughter nodes is done. Thus, by repeating the above steps the forest is built with 'n' number of trees [6]. Lastly, the classifier was operated to make predictions and to calculate each predicted target. The one with highest vote becomes the final prediction.

```
from sklearn.ensemble import RandomForestClassifier

# creating a model
model = RandomForestClassifier()

# feeding the training data to the model
model.fit(x_train, y_train)

# predicting the x-test results
y_pred = model.predict(x_test)

# calculating the accuracies
print("Training Accuracy :", model.score(x_train, y_train))
print("Testing Accuracy :", model.score(x_test, y_test))
```

Fig. 3 APPLYING RANDOM FOREST CLASSIFIER

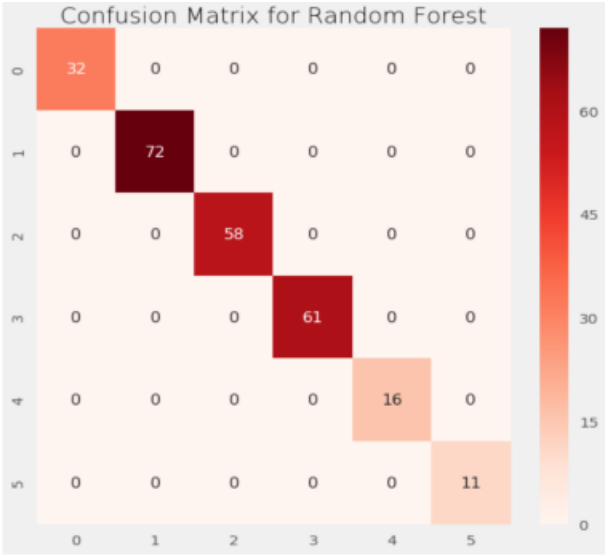


Fig. 4 CONFUSION MATRIX FOR RANDOM FOREST

D.LOGISTIC REGRESSION

Logistic Regression are adept to build statistical model using conditional probability and logistic functions. It comes into the picture whenever the target variable is given to be categorical. Logistic Regression deals in with binary classification problems with the help of statistics. Based on the previous information available it predicts the probability of event occurring with use of mathematical modelling.

WORKING: It is one of the most important supervised classification algorithms which is used to build neural networks on the basis of deep learning. To predict the target class, it calculates the logits. It works by measuring the relationship between the dependent and independent variables, by estimating probabilities using its underlying logistic function. It models the data by using sigmoid function and decision threshold. Once the threshold value is set the data is

classified to get the accurate prediction. These are one of the most complex and advanced algorithms which can easily predict the continuous value by finding out the correlation between categorical variables. If the data has more than two classes, SoftMax regression is used to predict the accuracy of model data [7]. The regression starts with mapping function in which it aligns the values to continuous output. Here the predicted data belongs to category of continuous values. Whereby, the nature if predicted data is ordered (in some sequence). Lastly, Root Mean Square Error is applied to identify the best fit side of dataset.

```
from sklearn.linear_model import  LogisticRegression

# creating a model
model = LogisticRegression()

# feeding the training data to the model
model.fit(x_train, y_train)

# predicting the test set results
y_pred = model.predict(x_test)

# calculating the classification accuracies
print("Training Accuracy :", model.score(x_train, y_train))
print("Testing Accuracy :", model.score(x_test, y_test))
```

Fig. 5 APPLYING LOGISTIC REGRESSION

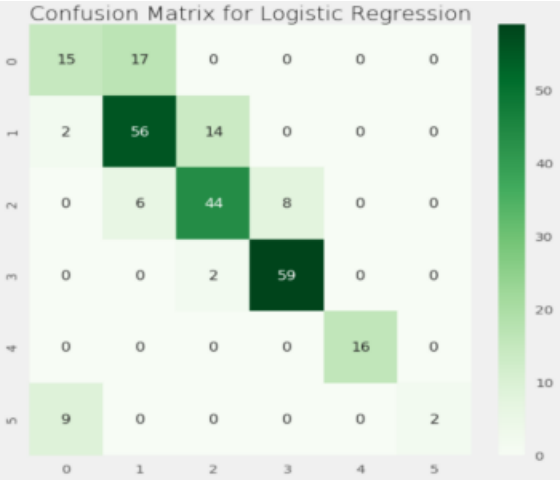


Fig. 6 CONFUSION MATRIX FOR LOGISTIC REGRESSION

OUTCOME: The above given dataset was used to study the student progress in different subjects (reading, writing, maths). The marks secured by students in these three subjects were used to build predictive models using machine learning algorithms. This study proved to be

essential parameter for detecting student's performance and their level of understanding towards particular subject. Finally, the comparison of total score of all students was done to predict the final graph between the students.

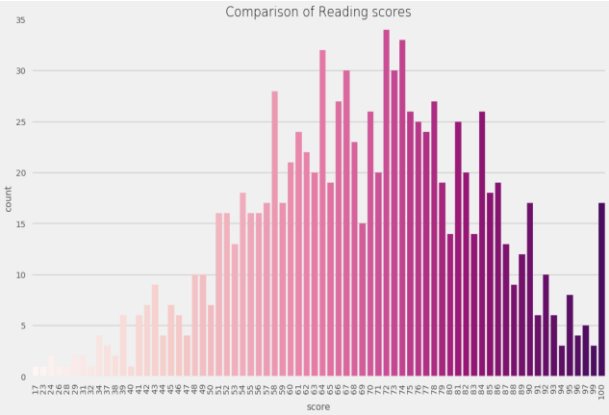


Fig. 7 READING SCORES MODEL

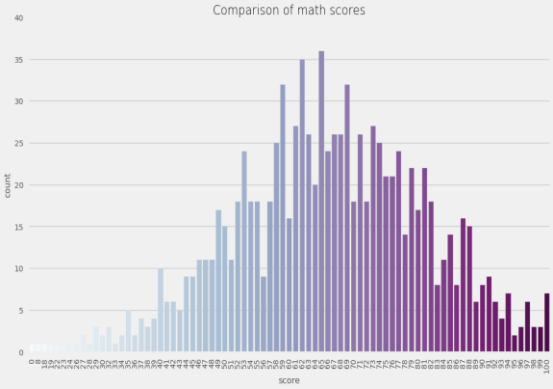


Fig. 8 MATH SCORES MODEL

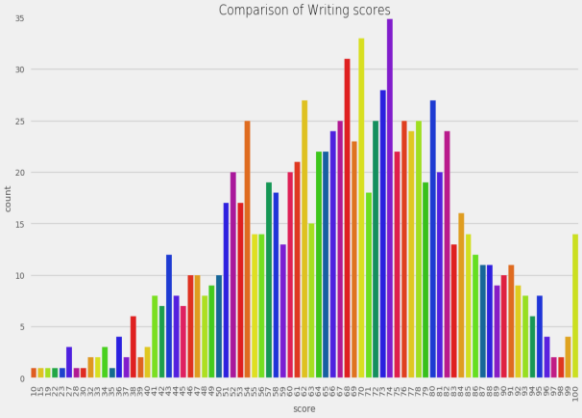


Fig. 9 WRITING SCORES MODEL

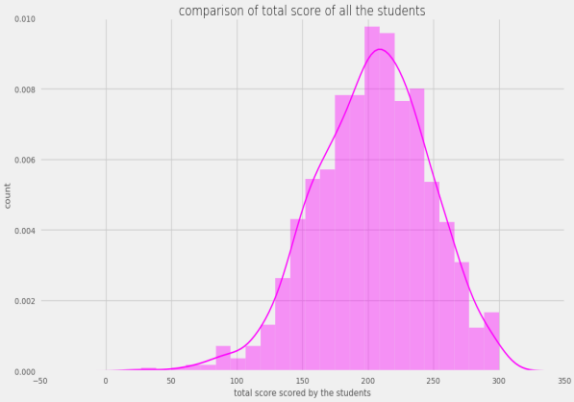


Fig. 10 COMPARISON OF TOTAL SCORE OF ALL STUDENTS

V. CONCLUSION

The main aim of this paper is to eradicate the traditional way of learning and to support machine learning techniques which can be used to generate much efficient learning system. In this research, an effort was made to predict a well-structured model which proposes all the features related to student's result. With the use of ML algorithms, the work of teachers as well as students becomes easier. Moreover, this technique will also amplify our learning outcomes. The application of three prediction strategies used in the paper was done to compare the classification methodology and to forecast the performance based on its properties.

In future this accuracy rate can be applied on actual data to modify some of the properties of the dataset. Today educational sector in the world is adapting machine learning methods to spot the struggling students and thereby, helps in increasing the success rate. Modification in teaching and learning experiences personalize engagement providing an individual to adapt for a better education. The information which cannot be gleaned by human brain can easily be done with data analytics using machine learning. These assists grappled students and at the same time challenges the gifted ones.

ACKNOWLEDGMENT

I am overwhelmed in all humbleness and would like to express all my special thanks of gratitude to all those who have helped me in putting these ideas concrete. I am very much thankful to my Prof. Neha Vekhande for her continuous support as well as our HOD Dr. Biswajeet Champaty for encouraging me and providing this golden opportunity to do the research on topic "Uncovering Patterns in Education Data through Machine Learning". The research has helped me in knowing about so many new things. It also inspired me to use machine learning algorithms to improve the current education scenario. I would also thank my parents who have helped me a lot in guiding me from time to time.

REFERENCES

[1]Vekhande, Neha & Hore, Debirupa. (2021). "UNDERSTANDING THE SCOPE OF RESEARCH IN EDU-TECH DOMAIN THROUGH ARTIFICIAL INTELLIGENCE AND DATA MINING.", Shodh Sarita, Volume 7, Issue 28, Pg. 175-181.

[2]Ryan S.J.D. Baker & Kalina Yacef. (2009). "The State of Educational Data Mining in 2009: A Review and Future Visions.", Journal of Educational Data Mining, Article 1, Volume 1, No 1, Pg. 3-16.

[3]Nor Azziaty Binti Abdul Rahman, Kian Lam Tan & Chen Kim Lim. (2017). "Supervised And Unsupervised Learning In Data Mining For Employment Prediction For Fresh Graduate Students", Journal of Telecommunication, Electronic, and Computer Engineering, Volume 9, No 2-12, Pg. 155-161.

[4]Bart Rienties, Henrik Kohler Simonsen & Christothea Herodotou. (2020). "Defining The Boundaries Between Artificial Intelligence in Education, Computer- Supported Collaborative Learning, Educational Data Mining, and Learning Analytics: A Need for Coherence.", Frontiers in Education, Article 128, Volume 5, doi: 10.3389/feduc.2020.00128

[5]Muskan Agrawal, Abhinav Tripathi, Vishal Dudgikar & Neha Vekhande. (2021). "Understanding Different Techniques Of Data Cleaning And Different Operations Involved.", Turkish Journal of Computer and Mathematics Education, Volume 12, No 11, Pg. 3820-3826.

[6]Wikipedia, "Educational Data Mining", https://en.wikipedia.org/wiki/Educational_data_mining visited on 26/05/2021.

[7]IntechOpen, "Machine Learning Advanced Techniques and Emerging Applications", https://www.intechopen.com/books/machine-learning-advanced-techniques-and-emerging-applications visited on 04/06/2021.