EXPLORING LEARNING ANALYTICS OF SKILL BASED COURSES IN ONLINE USING MACHINE LEARNING ALGORITHMS

Ms. Syed Nazia Banu[1], U. Kavitha[2], K. Kalyani[3], S. Vyshnavi[4], E. Sankeerthana[5]

[1]Assistant Professor in Department of Computer Science and Engineering, Santhiram Engineering College, Nandyal, Kurnool, Andhra Pradesh, India.

[2,3,4,5]Student, Department of Computer Science and Engineering, Santhiram Engineering College, Nandyal, Kurnool, Andhra Pradesh, India.

DOI: https://www.doi.org/10.58257/IJPREMS33543

ABSTRACT

Online learning has attracted a large number of participants because it has no limit to enrollment and regardless of personal background and location. Predicting academic performance is an important task for the students in university, college, and school, etc. Machine Learning is a field of computer science that makes the computer to learn itself without any help of external programs. The dataset used in this project is stored in a SQL database and accessed using queries as and when required. There are two approaches for machine learning techniques one is supervised learning and the other one is unsupervised learning.

In unsupervised learning, K-means clustering is being used and in supervised, ensemble techniques like Random Forest and XgBoost algorithm are implemented. Nowadays evaluating the student performance of any organization is going to play a vital role to train the students. All of the above algorithms were combined and used for student evaluation and a possible suggestion to the student is provided to improve their career.

Keywords: Machine Learning, K-Means, XG Boost, Random Forest, Ensemble method.

1. INTRODUCTION

The academic performance of students holds significant importance within educational institutions, often serving as a primary metric for assessing excellence. While some scholars argue that academic performance can be gauged through learning assessments and participation in extracurricular activities, many agree that past academic achievements and grades are strong indicators of future success. Online learning platforms offer a plethora of resources, including lecture videos, online assessments, discussion forums, and live video sessions, making learning more accessible and flexible for participants worldwide

2. LITERATURE SURVEY

A) Karimi, Hamid et al. “A Deep Model for Predicting Online Course Performance.” (2020)

Online learning has attracted a large number of participants because it has no limit to enrollment and regardless of personal background and location. One of main goals of education is improving students’ learning gain. However, the completion rates for online learning are notoriously low.

We focus on predicting students’ learning performance early and help instructors to provide intervention in-time. We propose a deep online learning performance prediction model incorporate clickstream and demographic data of students. The experiments on the Open University Learning Analytics Dataset (OULAD) show that fusion of learner demographic information can make up for inadequate online learning behavior data early and improve prediction performance. And our model can achieve reliable performance both in intra-course and inter-course outcome prediction.

Summary: This journal discusses about scoring and performance predictions in online courses.

B) Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR).

As the computer technology and computer network technology are developing, the amount of data in information industry is getting higher and higher.

It is necessary to analyze this large amount of data and extract useful knowledge from it. Process of extracting the useful knowledge from huge set of incomplete, noisy, fuzzy and random data is called data mining. Decision tree classification technique is one of the most popular data mining techniques. In decision tree divide and conquer technique is used as basic learning strategy. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label.

The topmost node in the tree is the root node. This paper focus on the various algorithms of Decision tree (ID3, C4.5, CART), their characteristic, challenges, advantage and disadvantage.

Summary:

In this paper, we learn about Decision Tree, types of Decision tree (ID3, C4.5, CART etc..). It also discusses about the advantages and disadvantages of Decision Tree.

C) Manju & Mathur, Bhawana. (2014). Comparative Study of K-Means and Hierarchical Clustering Techniques. International journal of Software and Hardware Research in Engineering.

Clustering is a process of keeping similar data into groups. Clustering is an unsupervised learning technique as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. Many types of clustering methods are hierarchical, partitioning, density –based, model-based, grid –based, and soft-computing methods. In this paper compare with k-Means Clustering and Hierarchical Clustering Techniques. Strength and weakness of both Clustering Techniques and their methodology and process.

Summary:

In this paper, we learn clustering algorithms like K means and Agglomerative clustering and their comparisons.

D) Kabakchieva D (2012) Student performance prediction by using data mining classification algorithms. IJCSMR

This paper presents the results from data mining research, performed at one of the famous and prestigious Bulgarian universities, with the main goal to reveal the high potential of data mining applications for university management and to contribute to more efficient university enrolment campaigns and to attracting the most desirable students. The research is focused on the development of data mining models for predicting student performance, based on their personal, pre-university and university-performance characteristics.

The dataset used for the research purposes includes data about students admitted to the university in three consecutive years. Several well-known data mining classification algorithms, including a rule learner, a decision tree classifier, a neural network and a Nearest Neighbor classifier, are applied on the dataset. The performance of these algorithms is analyzed and compared.

3. METHODOLOGY

In this section, the methodology was adopted in order to predict Student Performance. More specifically in section A, the dataset information is described. And Section B consists of evaluation metrics.

Dataset Information:

The xAPI Edu-Data dataset contains a variety of student related data, including nationality, place of birth, across different types of feed backs. here are the main features of the xAPI dataset with brief description:

- Gender: Student’s gender.
- Nationality: Student’s Nationality.
- Place of Birth: Student’s place of birth.
- Educational Stage: Educational level student belongs to.
- Grade ID: Grade student belongs to.
- Section ID: Students section ID.
- Topic: Course topic.
- Semester: School year semester.
- Relation: Parent responsible for student.
- Raised hands: How many times the student raises his/her hand in the classroom.
- Visited Resources: How many times the student visits course content.
- Announcements View: How many times the student checks the new announcement.
- Discussion: How many times the student participates on discussion groups.
- Parent Answering Survey: Parent answer the survey which are provided from school or not.
- Parent school Satisfaction: The degree of parent satisfaction from School.
- Duration Student Absence Days ion: The number of absent days for each student.

4. IMPLEMENTATION AND ANALYSIS

In this section, the implementation details are mentioned to predict the student's performance. It contains the model selection, and the prediction that has done, and its accuracy is shown.

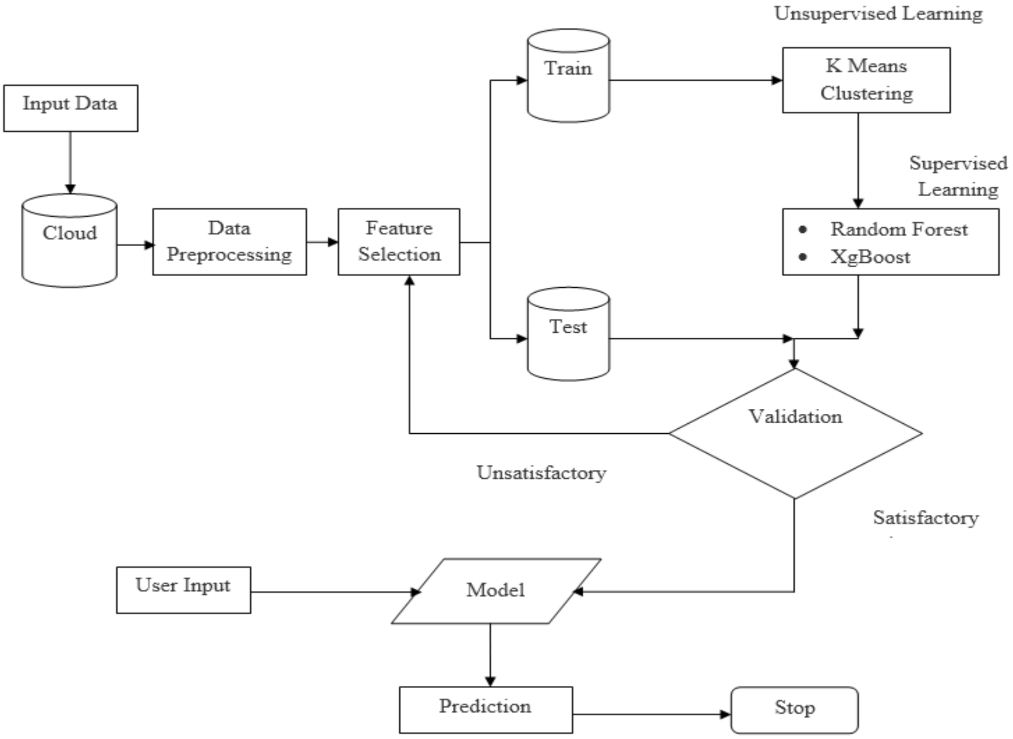


Figure:1 System Architecture

Model - 1: K-Means Clustering:

The algorithm will categorize the items into k groups of similarity. To calculate that similarity, we will use the Euclidean distance as measurement. First we initialize k points, called means, randomly. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.

We repeat the process for a given number of iterations and at the end, we have our clusters.

Model - 2: Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

Model - 3: XgBOOST:

XgBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

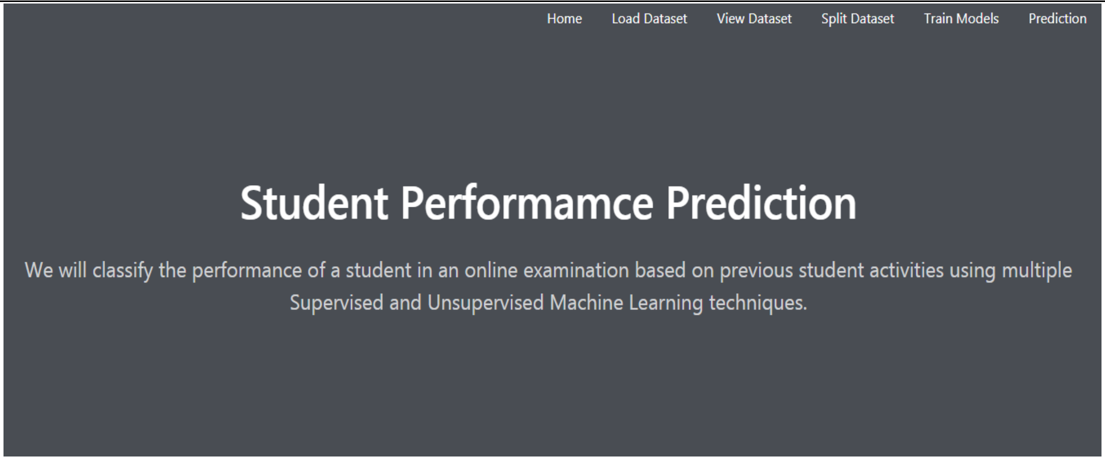
The implementation of the model supports the features of the scikit-learn and R implementations, with new additions like regularization. Three main forms of gradient boosting are supported:

- **Gradient Boosting** algorithm also called gradient boosting machine including the learning rate.
- **Stochastic Gradient Boosting** with sub-sampling at the row, column and column per split levels.
- **Regularized Gradient Boosting** with both L1 and L2 regularization.

5. RESULTS

We will classify the performance of a student in an online examination based on previous student activities using multiple Supervised and Unsupervised Machine Learning techniques. This research explored the possibility of predicting student's exact grade, success and failure on the basis of different input variables. The final results demonstrate the effectiveness of the combined approach in predicting student performance in online course. High prediction accuracy and reliability are achieved, contributing to improved educational outcomes and student support.

After executing the program, in the terminal click "python main.py" and click on Enter. In the terminal we will get a link which directs into a new page. Click on the link to display the below page.

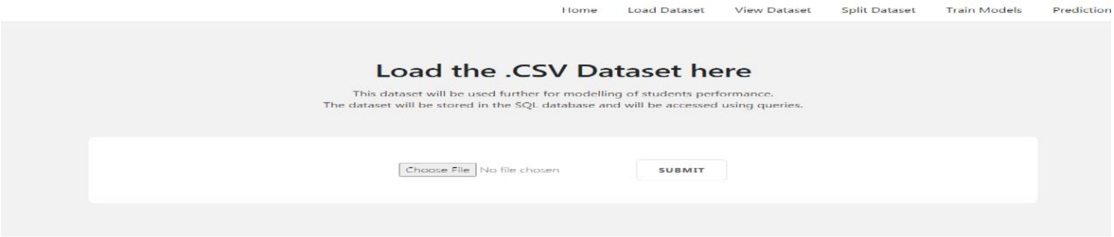


Student Performamce Prediction

We will classify the performance of a student in an online examination based on previous student activities using multiple Supervised and Unsupervised Machine Learning techniques.

Fig:2 Output Screen.

At the navigation bar, click on "Load dataset", such that the below screen will be displayed. Here, we can upload the student's dataset by clicking on the "Choose File" and click on "Submit".



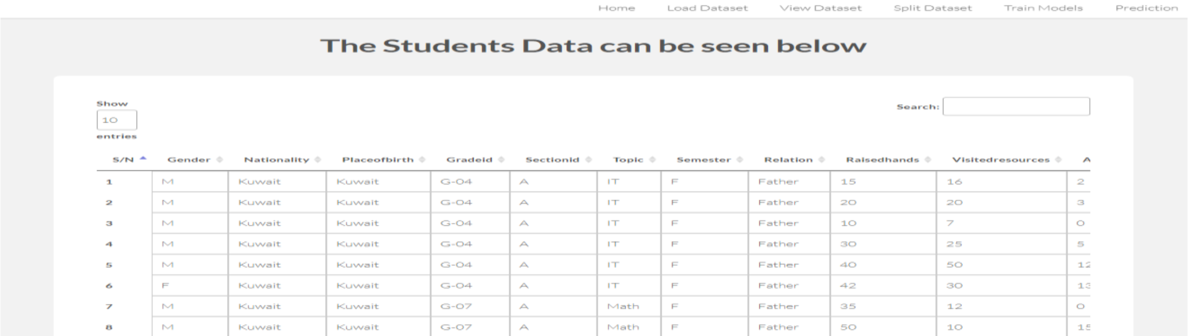
Load the .CSV Dataset here

This dataset will be used further for modelling of students performance.
The dataset will be stored in the SQL database and will be accessed using queries.

No file chosen

Fig:3 Load Dataset

At the navigation bar, click on "Load dataset", such that the below screen will be displayed, such that the dataset we uploaded will be visible.



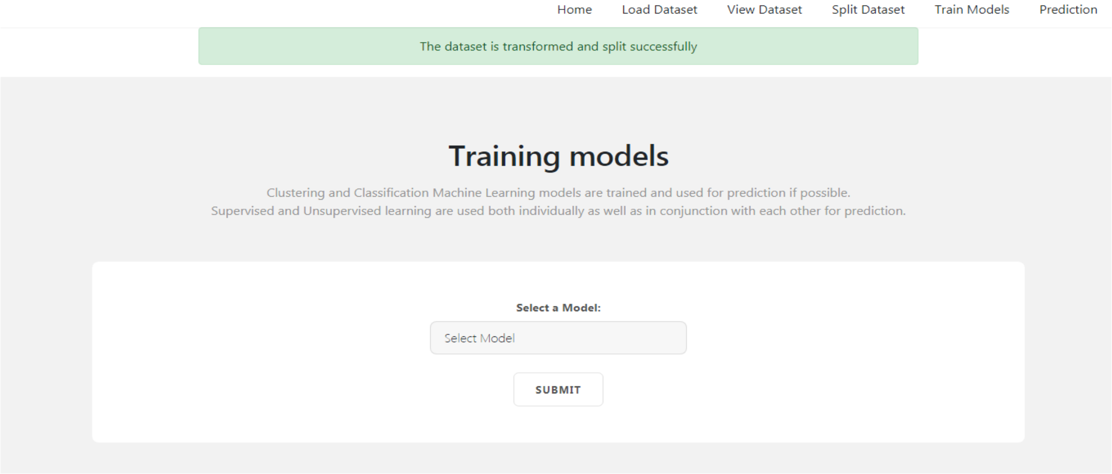
The Students Data can be seen below

Show 10 entries

S/N	Gender	Nationality	Placeofbirth	Gradeid	Sectionid	Topic	Semester	Relation	Raisedhands	Visitedresources	A
1	M	Kuwait	Kuwait	G-04	A	IT	F	Father	15	16	2
2	M	Kuwait	Kuwait	G-04	A	IT	F	Father	20	20	3
3	M	Kuwait	Kuwait	G-04	A	IT	F	Father	10	7	0
4	M	Kuwait	Kuwait	G-04	A	IT	F	Father	30	25	5
5	M	Kuwait	Kuwait	G-04	A	IT	F	Father	40	50	12
6	F	Kuwait	Kuwait	G-04	A	IT	F	Father	42	30	13
7	M	Kuwait	Kuwait	G-07	A	Math	F	Father	35	12	0
8	M	Kuwait	Kuwait	G-07	A	Math	F	Father	50	10	15

Fig: 4 View Dataset

After viewing the Dataset click on "Split Dataset", such that the dataset we upload will be split based on the similarities and it will automatically redirect the page to "Train Models" page shown below.



The dataset is transformed and split successfully

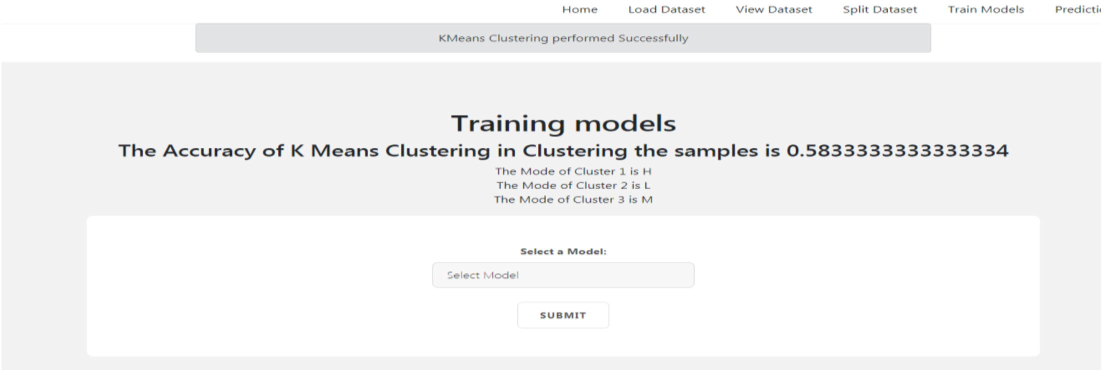
Training models

Clustering and Classification Machine Learning models are trained and used for prediction if possible.
Supervised and Unsupervised learning are used both individually as well as in conjunction with each other for prediction.

Select a Model: Select Model

Fig:5 Split Dataset and selecting the Training Model

In the “Training Models” page select your required training model that you want to use to know the accuracy and click on “Submit”, it gives the accuracy for each of the model



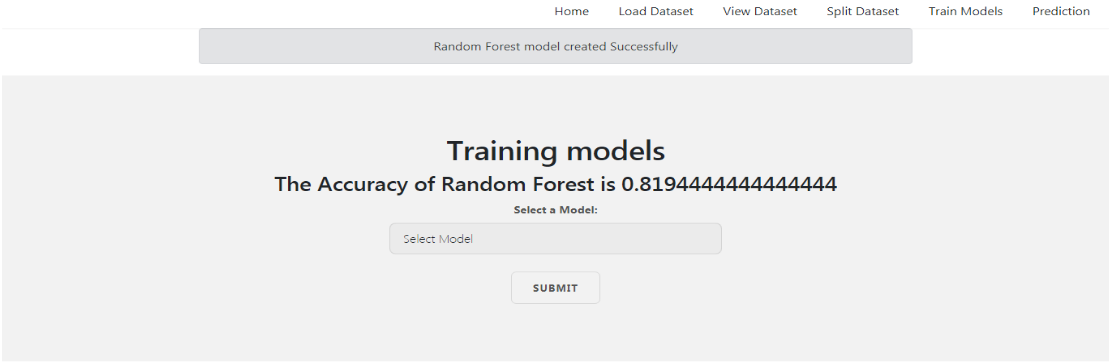
KMeans Clustering performed Successfully

Training models

The Accuracy of K Means Clustering in Clustering the samples is 0.5833333333333334

The Mode of Cluster 1 is H
The Mode of Cluster 2 is L
The Mode of Cluster 3 is M

Fig: 6 Accuracy using K-Means Clustering

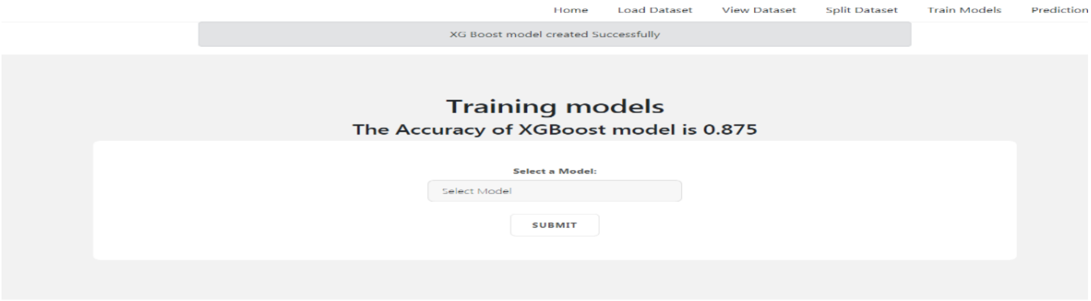


Random Forest model created Successfully

Training models

The Accuracy of Random Forest is 0.8194444444444444

Fig:7 Accuracy using Random Forest



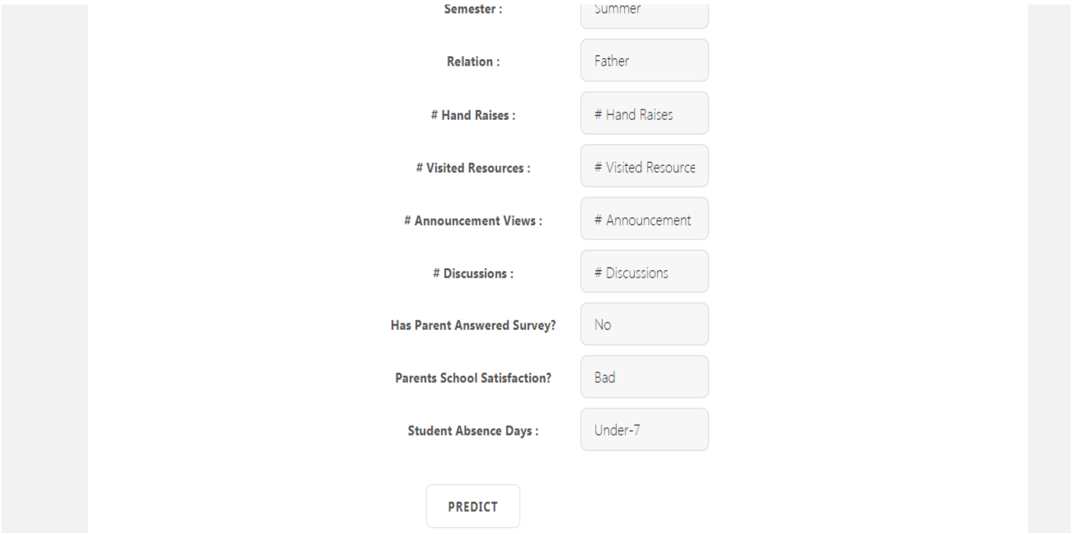
XG Boost model created Successfully

Training models

The Accuracy of XGBoost model is 0.875

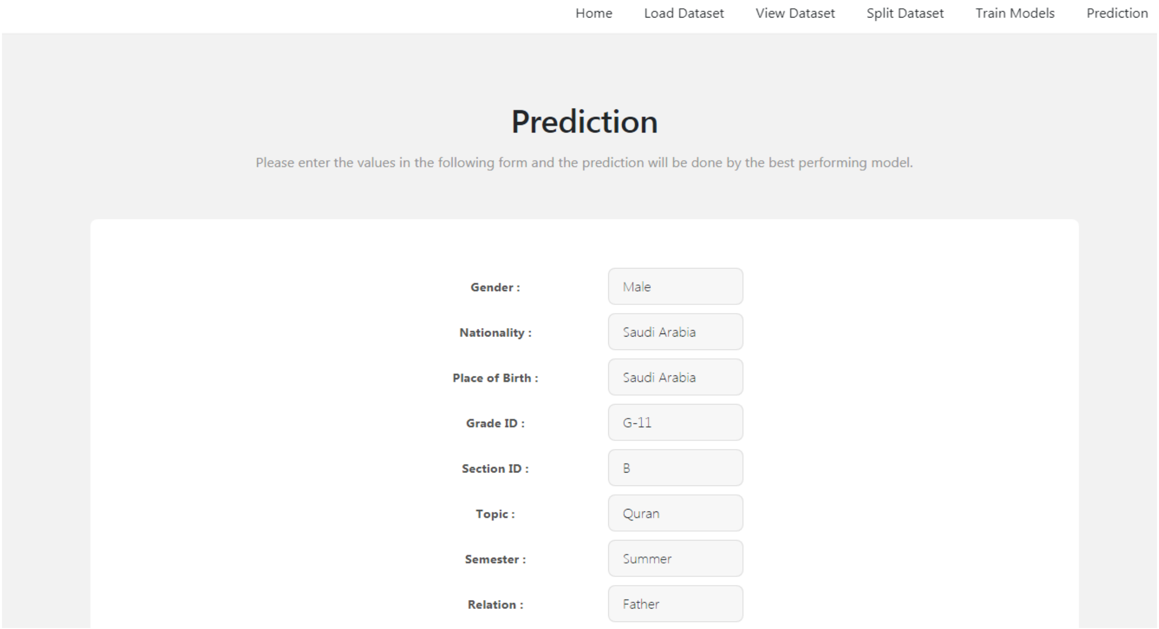
Fig: 8 Accuracy using XgBoost

Now, Click on “Prediction” on the top right below screen will be displayed. Now, enter all the values of specific student and click on “Predict”.



Semester :	Summer
Relation :	Father
# Hand Raises :	# Hand Raises
# Visited Resources :	# Visited Resource
# Announcement Views :	# Announcement
# Discussions :	# Discussions
Has Parent Answered Survey?	No
Parents School Satisfaction?	Bad
Student Absence Days :	Under-7

Fig: 9 Giving the required inputs



Prediction

Please enter the values in the following form and the prediction will be done by the best performing model.

Gender : Male

Nationality : Saudi Arabia

Place of Birth : Saudi Arabia

Grade ID : G-11

Section ID : B

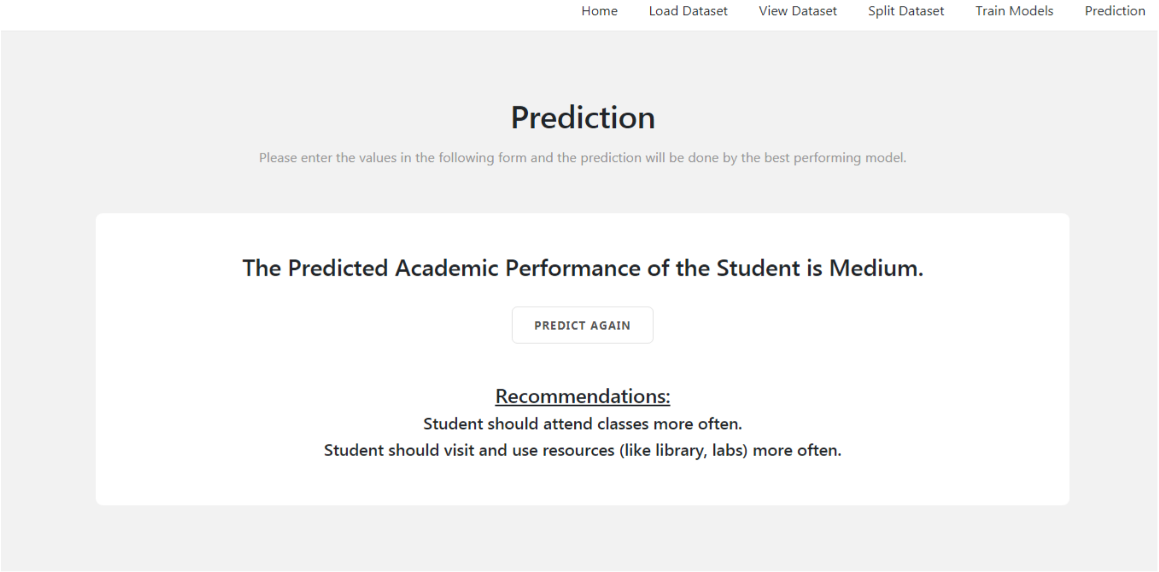
Topic : Quran

Semester : Summer

Relation : Father

Fig:10

Based on the information you have given, we will get the below prediction of the specified student.



Prediction

Please enter the values in the following form and the prediction will be done by the best performing model.

The Predicted Academic Performance of the Student is Medium.

Recommendations:

Student should attend classes more often.

Student should visit and use resources (like library, labs) more often.

Fig:11 Prediction of the Model

6. CONCLUSION

In this application, we have preprocessed the data by removing the null values and encoding all the variables. We used an unsupervised and 2 supervised learning methods. K-Means clustering is the unsupervised algorithms which we have used here. Random Forest and XgBoost are the 2 supervised algorithms used for actual classification of the students' performance. The best model was the XgBoost model with hyper parameters tuning. Clustering algorithms cannot be explicitly used for classification. But we can use them in conjunction with supervised techniques to be used for prediction. The dataset used was the dataset with student's online course performance against their activities previously.

7. REFERENCES

[1] Mahammad, F. S., & Viswanatham, V. M. (2020). Performance Analysis Of Data Compression Algorithms For Heterogeneous Architecture Through Parallel Approach. The Journal Of Supercomputing, 76(4), 2275-2288.

[2] Karukula, N. R., & Farooq, S. M. (2013). A Route Map For Detecting Sybil Attacks In Urban Vehicular Networks. Journal Of Information, Knowledge, And Research In Computer Engineering, 2(2), 540-544.

[3] Farook, S. M., & Nageswarareddy, K. (2015). Implementation Of Intrusion Detection Systems For High Performance Computing Environment Applications. Inter National Journal Of Scientific Engineering And Technology Research, 4(0), 41.

[4] Sunar, M. F., & Viswanatham, V. M. (2018). A Fast Approach To Encrypt And Decrypt Of Video Streams For Secure Channel Transmission. *World Review Of Science, Technology And Sustainable Development*, *14*(1), 11-28.

[5] Mahammad, F. S., & Viswanatham, V. M. (2017). A Study On H. 26x Family Of Video Streaming Compression Techniques. *International Journal Of Pure And Applied Mathematics*, *117*(10), 63-66.

[6] Devi,S M. S., Mahammad, F. S., Bhavana, D., Sukanya, D., Thanusha, T. S., Chandrakala, M., & Swathi, P. V. (2022).” Machine Learning Based Classification And Clustering Analysis Of Efficiency Of Exercise Against Covid-19 Infection.” Journal Of Algebraic Statistics, 13(3), 112-117.

[7] Devi, M. M. S., & Gangadhar, M. Y. (2012).” A Comparative Study Of Classification Algorithm ForPrinted Telugu Character Recognition.” *International Journal Of Electronics Communication And Computer Engineering*, *3*(3), 633-641.

[8] Devi, M. S., Meghana, A. I., Susmitha, M., Mounika, G., Vineela, G., & Padmavathi, M. Missing Child Identification System Using Deep Learning.

[9] V. Lakshmi Chaitanya. "Machine Learning Based Predictive Model For Data Fusion Based Intruder Alert System." Journal Of Algebraic Statistics 13, No. 2 (2022): 2477-2483.

[10] Chaitanya, V. L., & Bhaskar, G. V. (2014). Apriori Vs Genetic Algorithms For Identifying Frequent Item Sets. International Journal Of Innovative Research &Development, 3(6), 249-254.

[11] Chaitanya, V. L., Sutraye, N., Praveeena, A. S., Niharika, U. N., Ulfath, P., & Rani, D. P. (2023). Experimental Investigation Of Machine Learning Techniques For Predicting Software Quality.

[12] Lakshmi, B. S., Pranavi, S., Jayalakshmi, C., Gayatri, K., Sireesha, M., & Akhila, A. Detecting Android Malware With An Enhanced Genetic Algorithm For Feature Selection And Machine Learning.

[13] Lakshmi, B. S., & Kumar, A. S. (2018). Identity-Based Proxy-Oriented Data Uploading And Remote Data Integrity Checking In Public Cloud. International Journal Of Research, 5(22), 744-757.

[14] Lakshmi, B. S. (2021). Fire Detection Using Image Processing. Asian Journal Of Computer Science And Technology, 10(2), 14-19.

[15] Devi, M. S., Poojitha, M., Sucharitha, R., Keerthi, K., Manideepika, P., & Vasudha, C. Extracting And Analyzing Features In Natural Language Processing For Deep Learning With English Language.

[16] Kumar Jds, Subramanyam Mv, Kumar Aps. Hybrid Chameleon Search And Remora Optimization Algorithm-Based Dynamic Heterogeneous Load Balancing Clustering Protocol For Extending The Lifetime Of Wireless Sensor Networks. Int J Commun Syst. 2023; 36(17):E5609. Doi:10.1002/Dac.5609

[17] David Sukeerthi Kumar, J., Subramanyam, M.V., Siva Kumar, A.P. (2023). A Hybrid Spotted Hyena And Whale Optimization Algorithm-Based Load-Balanced Clustering Technique In Wsns. In: Mahapatra, R.P., Peddoju, S.K., Roy, S., Parwekar, P. (Eds) Proceedings Of International Conference On Recent Trends In Computing. Lecture Notes In Networks And Systems, Vol 600. Springer, Singapore. Https://Doi.Org/10.1007/978-981-19-8825-7_68

[18] Murali Kanthi, J. David Sukeerthi Kumar, K. Venkateshwara Rao, Mohmad Ahmed Ali, Sudha Pavani K, Nuthanakanti Bhaskar, T. Hitendra Sarma, “A Fused 3d-2d Convolution Neural Network For Spatial-Spectral Feature Learning And Hyperspectral Image Classification,” J Theor Appl Inf Technol, Vol. 15, No. 5, 2024, Accessed: Apr. 03, 2024. [Online]. Available: Www.Jatit.Org

[19] Prediction Of Covid-19 Infection Based On Lifestyle Habits Employing Random Forest Algorithm Fs Mahammad, P Bhaskar, A Prudvi, Ny Reddy, Pj Reddy Journal Of Algebraic Statistics 13 (3), 40-45

[20] Machine Learning Based Predictive Model For Closed Loop Air Filtering System P Bhaskar, Fs Mahammad, Ah Kumar, Dr Kumar, Sma Khadar, ...Journal Of Algebraic Statistics 13 (3), 609-616

[21] Kumar, M. A., Mahammad, F. S., Dhanush, M. N., Rahul, D. P., Sreedhara, K. L., Rabi, B. A., & Reddy, A. K. (2022). Traffic Length Data Based Signal Timing Calculation For Road Traffic Signals Employing Proportionality Machine Learning. Journal Of Algebraic Statistics, 13(3), 25-32.

[22] Kumar, M. A., Pullama, K. B., & Reddy, B. S. V. M. (2013). Energy Efficient Routing In Wireless Sensor Networks. International Journal Of Emerging Technology And Advanced Engineering, 9(9), 172-176.

[23] Kumar, M. M. A., Sivaraman, G., Charan Sai, P., Dinesh, T., Vivekananda, S. S., Rakesh, G., & Peer, S. D. Building Search Engine Using Machine Learning Techniques.

[24] “ Providing Security In Iot Using Watermarking And Partial Encryption. Issn No:

[25] 2250-1797 Issue 1, Volume 2 (December 2011)

[26] The Dissemination Architecture Of Streaming Media Information On Integrated Cdn And P2p, Issn 2249-6149 Issue 2, Vol.2 (March-2012)

[27] Provably Secure And Blind Sort Of Biometric Authentication Protocol Using Kerberos, Issn: 2249-9954, Issue 2, Vol 2 (April 2012)

[28] D.Lakshmaiah, Dr.M.Subramanyam, Dr.K.Satya Prasad," Design Of Low Power 4- Bit Cmos Braun Multiplier Based On Threshold Voltage Techniques", Global Journal Of Research In Engineering, Vol.14(9),Pp.1125-1131,2014.

[29] R Sumalatha, Dr.M.Subramanyam, "Image Denoising Using Spatial Adaptive Mask Filter", Ieee International Conference On Electrical, Electronics, Signals, Communication &Amp; Optimization (Eesco-2015), Organized Byvignans Institute Of Information Technology, Vishakapatnam, 24 Th To 26th January 2015. (Scopus Indexed)

[30] P.Balamurali Krishna, Dr.M.V.Subramanyam, Dr.K.Satya Prasad, "Hybrid Genetic Optimization To Mitigate Starvation In Wireless Mesh Networks", Indian Journal Of Science And Technology,Vol.8,No.23,2015. (Scopus Indexed)

[31] Y.Murali Mohan Babu, Dr.M.V.Subramanyam,M.N. Giri Prasad," Fusion And Texure Based Classification Of Indian Microwave Data – A Comparative Study", International Journal Of Applied Engineering Research, Vol.10 No.1, Pp. 1003-1009, 2015. (Scopus Indexed)