Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 10, Issue. 4, April 2021, pg.50 – 62

ACADEMIC PERFORMANCE ANALYSIS

[1]Mr. M. Thirunavukkarasu; [2]B.J.S.S Sriram; [3]Javvaji Chandrasekhar Reddy
[1]Assistant Professor, Dept of CSE, SCSVMV (Deemed to be University), Kanchipuram, TamilNadu, India
(mthiru@kanchiuniv.ac.in)
[2]Student, Dept of CSE, SCSVMV (Deemed to be University), Kanchipuram, TamilNadu, India
(11179A025@kanchiuniv.ac.in)
[3]Student, Dept of CSE, SCSVMV (Deemed to be University), Kanchipuram, TamilNadu, India
(11179A098@kanchiuniv.ac.in)
DOI: 10.47760/ijcsmc.2021.v10i04.008

ABSTRACT: In this, we mainly focus on the analysis of the students' performance in academics not only external exams, but also the overall academic performance of each and every student. We segregate and calculate the performance of students using data. We then predict the performance of the students who are going to pass and fail based on previous result and also the predicted marks of a student using algorithm namely Linear Regression and SVM algorithm. In this project, we mainly focus on the analysis of the students' performance and then predict the results through them using training data and then test data of academics not only external exams, but also the overall academic performance of each and every student.

INTRODUCTION:

In many of the colleges, when we see the academic performance analysis is done, but there is no system that predicts the student's performance in advance. Of which if student fails in an Exam. Here we consider both internal and external marks for analyzing academic performance of a student in the college which is analyzed using SVM algorithm and then Linear Regression algorithm.

PURPOSE:

The main purpose of this is to provide a system which produces the required data to students and as well as the professors. The required data includes, the students were required to obtain their results of their assignment marks, internal and external examinations. And not only provide access to students to know their results, the lectures

are provided a report of how many students are going to pass and fail in the oncoming examinations. By knowing the students` performance the lecturers can assist the dull students to perform better in the coming exams. These predictions are done using the previous results of those students.

SCOPE:

There are quite a few things that can be polished or be added in the future work. We have opted to use two data mining classifies in this project namely the Naive Bayes classifier. There are more classifiers such as the Bayesian network classifier, Neural Network classifier and C4.5 classifier. Such classifiers were not included in this paper and could be counted in future to give a more data to be compared with. Though, we have taken into consideration the academic data of many students, there are still many students and sample amount of input data that could further be used. With more and more demand for not only students but also performance prediction as a whole, there is a lot of data that can be taken into consideration for more accurate results. There is a lot of scope for student performance prediction in the data mining world.

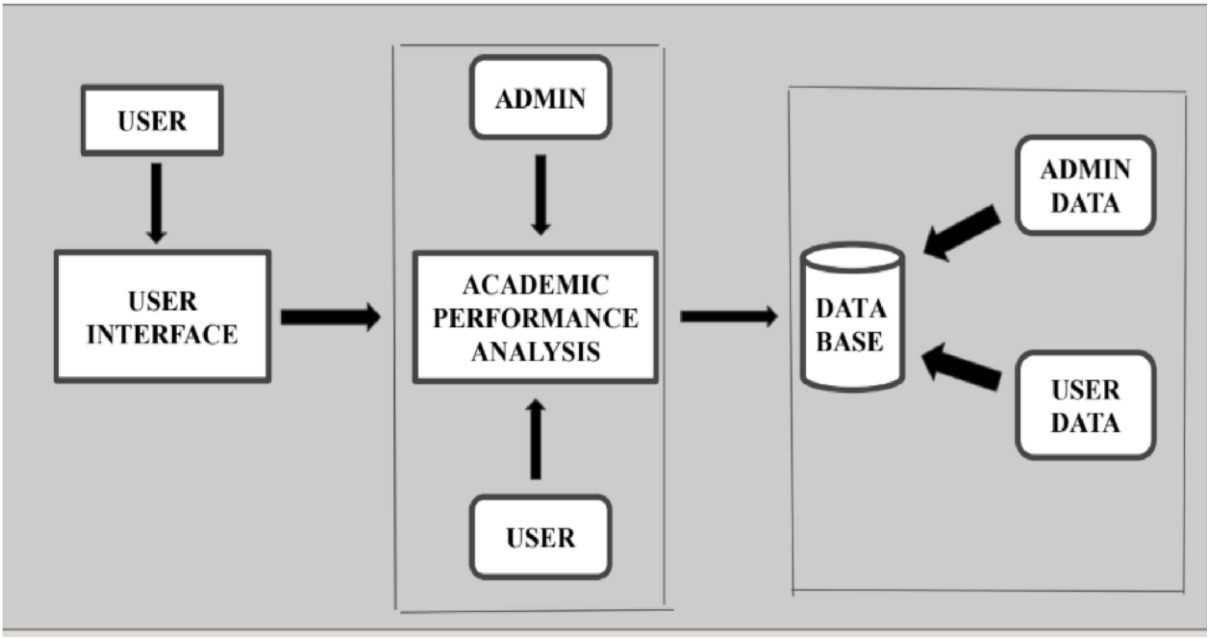
SOFTWARE REQUIREMENTS:

- Operating System : Windows XP/Windows 7/8/10
- Scripting languages : Python
- Python Packages : Numpy, Pandas, Matplotlib
- Tools : Visual studio code, Xampp

HARDWARE REQUIREMENTS:

- Processor : Intel Pentium Inside
- RAM : minimum 1 GB
- Secondary Storage : minimum 160 GB
- Key Board : Standard Windows Keyboard
- Mouse : Two or Three Button Mouse

System Architecture:



LITERATURE SURVEY

AUTHOR	YEAR	CONTRIBUTIONS OF AUTHOR	RESULT	PUBLISH AT
1.Mr Santosh B Akki	2018	The author deployed various data mining techniques to discover the education areas for Improvement.	It helps the organizations to manipulate and interact With data reports of static picture of the existing data.	3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2018.

2.Chew Li Sa, Dayang Hanani bt. AbangIbrahim , EmmyDahlian a Hossain Mohammad bin Hossin	2015	The author deployed various data mining techniques to discover the education areas for the improvement.	In this system is developed to predict student academic performance in course	3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2015.
3.Irina Sitova	2015	This research is to explore the applicability of data mining techniques in the area of simulation result analysis.	Data mining techniques may provide better interpretation of simulation output.	Data mining techniques may provide the better interpretation of simulation output.

Modules:

- **ADMIN**
- **FACULTY**
- **STUDENT**

Descriptions:

Admin:

In this module, admin has to login with valid username and password. After login successful he can do some operations such as add faculty, view faculty, add students,

view students. Adding faculty includes the details of the faculty such as name, email, phone number, gender, subject, address. Likewise adding student includes student roll number, name, email, password, course, gender and year.

Admin also creates the login credentials for the students and faculty.

Faculty:

In this module, faculty has to login with valid username and password. After login successful he can do some operations such as add marks, view marks, predict marks. Adding marks includes the details of the student and their subject marks respectively. In prediction we have two parts view pass prediction and view marks prediction. View pass prediction will show whether a student gets pass or fail and view marks prediction will predict the approximate marks of a student in the exams.

Student:

Here the student can only view his marks with his respective username and password.

Project Modules:

1. Data Preparation Module
2. Data Loading Module(combinational loads into the system)
3. Data Allocation/Publishing to threads
4. Centroid Assessment Of threads
5. Classification Results

1. Data Preparation Module

The process of preparing data for Machine Learning algorithm comprises the following:

I . Data Selection

ii. Data Preprocessing

iii. Data Transformation

Data Selection:

Steps involved in Data Selection involves:

- There is a vast volume, variety, and velocity of available data for a Machine Learning problem.

- This step involves selecting only a subset of the available data.
- The selected sample must be an accurate representation of the entire population.
- Some data can be derived or simulated from the available data if required.
- Data not relevant to the problem at hand can be excluded.

Data Preprocessing:

Let's understand Data Preprocessing in detail below.

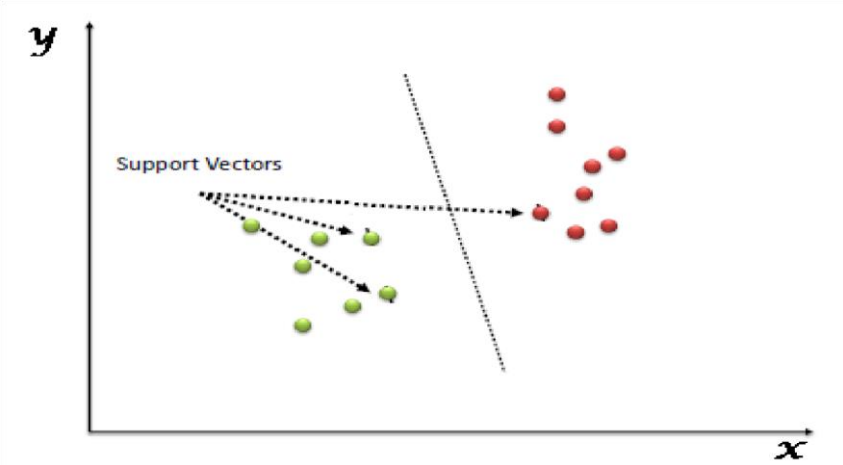
After the data has been selected, it needs to be preprocessed using the given steps:

- Formatting the data to make it suitable for ML (structured format)
- Cleaning the data to remove incomplete variables
- Sampling the data further to reduce running times for algorithms and memory requirements.

ALGORITHMS:

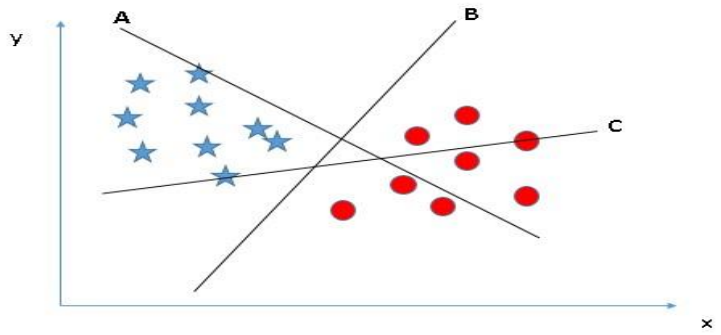
Support Vector Machine (SVM):

It is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).



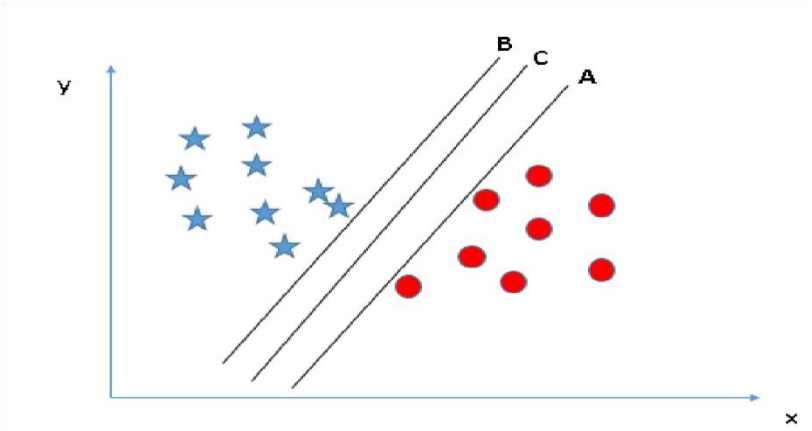
Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

Identify the right hyper-plane (Scenario-1): Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.

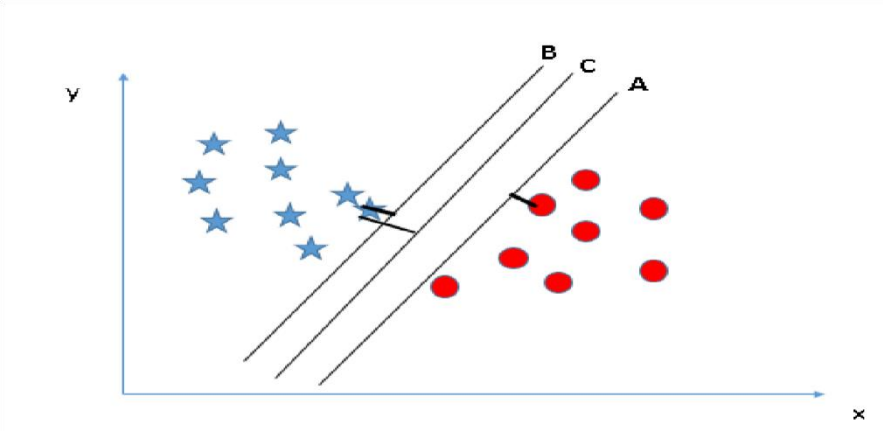


You need to remember a thumb rule to identify the right hyper-plane: "Select the hyper-plane which segregates the two classes better". In this scenario, hyper-plane "B" has excellently performed this job.

Identify the right hyper-plane (Scenario-2): Here, we have three hyper-planes (A, B and C) and all are segregating the classes well. Now, how can we identify the right hyper-plane?

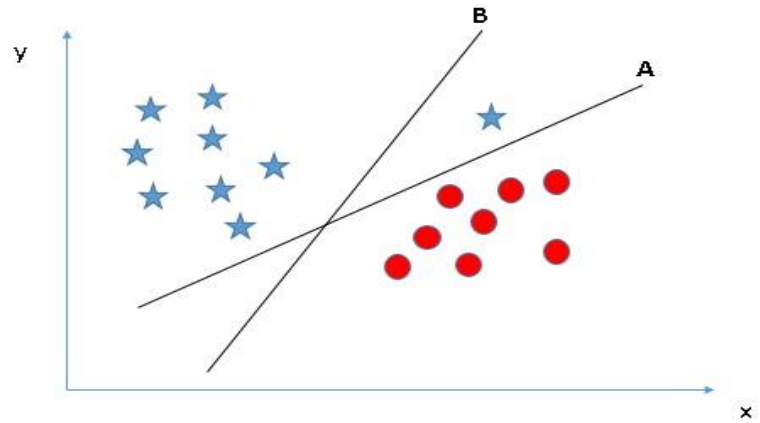


Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**.



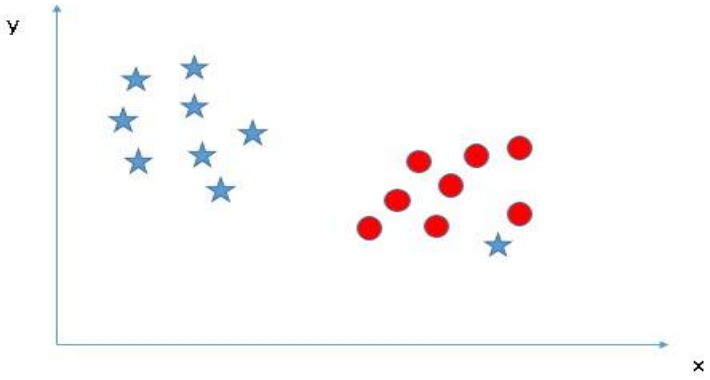
Above, you can see that the margin for hyper-plane C is high as compared to both A and B. Hence, we name the right hyper-plane as C. Another lightning reason for selecting the hyperplane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss-classification.

Identify the right hyper-plane (Scenario-3): Hint: Use the rules as discussed in previous section to identify the right hyper-plane

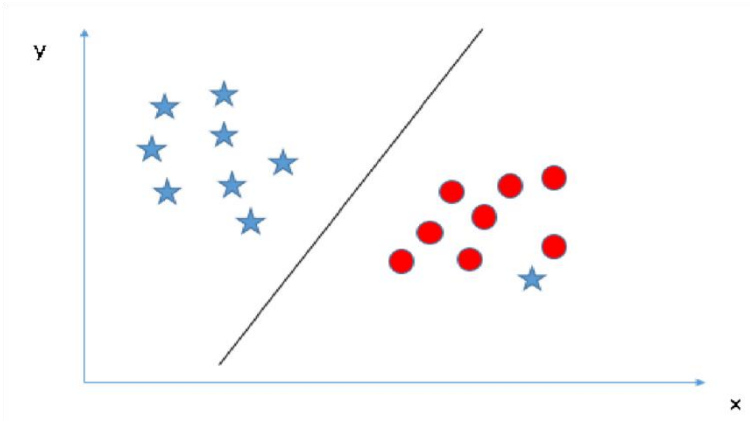


Some of you may have selected the hyper-plane **B** as it has higher margin compared to **A.** But, here is the catch, SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has classified all correctly. Therefore, the right hyper-plane is **A.**

Can we classify two classes (Scenario-4)?: Below, I am unable to segregate the two classes using a straight line, as one of the stars lies in the territory of other(circle) class as an outlier.



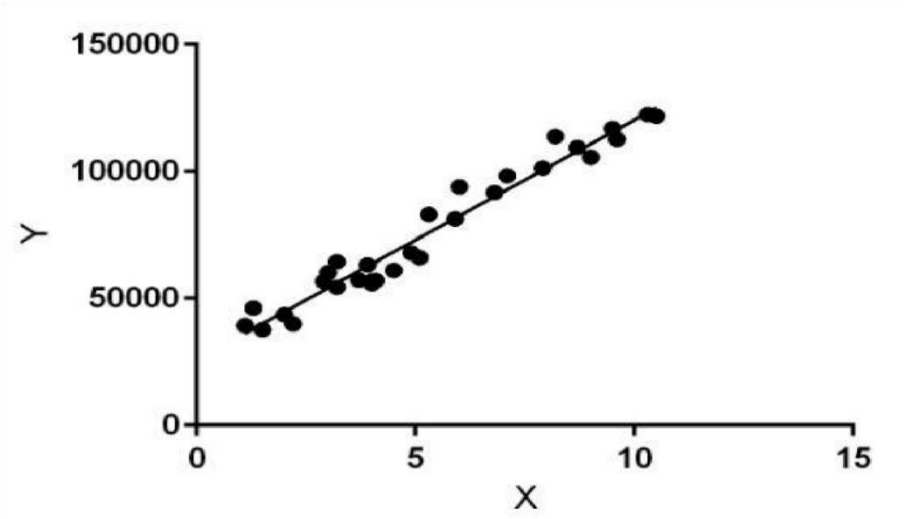
As I have already mentioned, one star at other end is like an outlier for star class. The SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin. Hence, we can say, SVM classification is robust to outliers.



Linear Regression:

It is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a *X + b.

Before knowing what linear regression is, let us get ourselves accustomed to regression. Regression is a method of modeling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.



Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that

models the points the best. The line can be modelled based on the linear equation shown below. y = a_0 + a_1 * x • Y – Dependent Variable

• a – Slope

• X – Independent variable

The motive of the linear regression algorithm is to find the best values for a_0 and a_1.

Before moving on to the algorithm, let's have a look at two important concepts you must know to better understand linear regression.

Cost Function:

The cost function helps us to figure out the best possible values for a_0 and a_1 which would provide the best fit line for the data points. Since we want the best values for a_0 and a_1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).
Minimization and Cost Function

We choose the above function to minimize. The difference between the predicted values and ground truth measures the error difference. We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error (MSE) functions. Now, using this MSE function we are going to change the values of a_0 and a_1 such that the MSE value settles at the minima.

Gradient Descent:

The next important concept needed to understand linear regression is gradient descent. Gradient descent is a method of updating a_0 and a_1 to reduce the cost function (MSE). The idea is that we start with some values for a_0 and a_1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the Values.

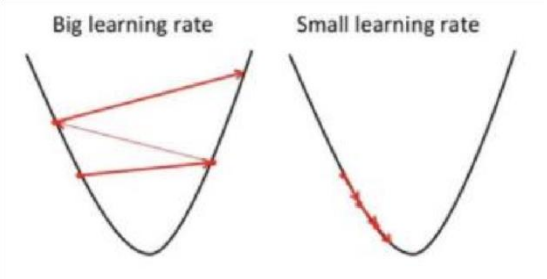


Fig: Gradient Descent

You may be wondering how to use gradient descent to update a_0 and a_1. To update a_0 and a_1, we take gradients from the cost function. To find these gradients, we take partial derivatives with respect to a_0 and a_1. Now, to understand how the partial derivatives are found below you would require some calculus but if you don't, it is alright.

You can take it as it is.

$$a_0 = a_0 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i)$$

$$a_1 = a_1 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i) \cdot x_i$$

The partial derivate are the gradients and they are used to update the values of a_0 and a_1. Alpha is the learning rate which is a hyper parameter that you must specify. A smaller learning rate could get you closer to the minima but takes more time to reach the minima, a larger learning rate converges sooner but there is a chance that you could overshoot the minima. The best way to understand linear regression is to relive this experience of childhood. Let us say, you ask a child in fifth grade to arrange people in his class by increasing order of weight, without asking them their weights! What do you think the child will do? She/he would likely look (visually analyze) at the height and build of people and arrange them using a combination of these visible parameters. This is linear regression in real life! The child has actually figured out that height and build would be correlated to the weight by a relationship, which looks like the equation above. In the equation, these coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line.

2) Unsupervised Learning:

In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Unsupervised learning algorithms are extremely powerful tools for analyzing data and for identifying patterns and trends. They are most commonly used for clustering similar input into

logical groups. Unsupervised learning algorithms include Kmeans, Random Forests, and Hierarchical clustering and so on.

Examples of Unsupervised Learning: Apriori algorithm, K-means

3. Reinforcement Learning:

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions.

Example of Reinforcement Learning: Markov Decision Process.

Similarly, there are four categories of machine learning algorithms as shown below

- Supervised learning algorithm
- Unsupervised learning algorithm
- Semi-supervised learning algorithm
- Reinforcement learning algorithm

However, the most commonly used ones are supervised and unsupervised learning. Purpose of Machine Learning: Machine learning can be seen as a branch of AI or Artificial Intelligence, since, the ability to change experience into expertise or to detect patterns in complex data is a mark of human or animal intelligence. As a field of science, machine learning shares common concepts with other disciplines such as statistics, information theory, game theory, and optimization. As a subfield of information technology, its objective is to program machines so that they will learn. However, it is to be seen that, the purpose of machine learning is not building an automated duplication of intelligent behavior, but using the power of computers to complement and supplement human intelligence.

CONCLUSION

This prediction algorithm serves as a good benchmark to monitor the progression of student's performance in higher institution. It also enhances the decision making by academic instructors to monitor the candidate's performance semester by semester by improving on the future academic results in the subsequence academic session. Thus with the help of this Prediction algorithm the instructors develop a good understanding of how well or how poorly the students in their classes will perform so instructors can take proactive measures to improve student learning.

REFERENCES

[1]. https://github.com/divyansh21april/Student-Performance-Analysis-/blob/master/student/student-por.csv
[2]. predicting student performance using data mining techniques
[3]. Student Performance Evaluation Using Data Mining ...
[4]. www.researchgate.net › publication › 329213058_Student_Performance...
[5]. **student academic performance using data mining techniques**
[6]. **A Student Performance Prediction Model Using Data Mining ...**
[7]. www.sciencepubco.com › index.php › ijet › article › view

