International Journal of Innovation Scientific Research and Review
Vol. 05, Issue, 04, pp.4375-4380, April 2023
Available online at http://www.journalijisr.com
SJIF Impact Factor 2023: 6.599

Research Article

ISSN: 2582-6131

EFFICIENT MACHINE LEARNING ALGORITHMS FOR STUDENT'S ACADEMIC PERFORMANCE PREDICTION

[1,] * MOHANNAD E ALSOFYANI and [2]SOUHIR SGHAIER

[1]Department of Computer and Information Technology, Technical College of Ranyah, Technical and Vocational Training Corporation, Jeddah 24250, Saudi Arabia.
[2]Department of Science and Technology, University College of Ranyah, Taif University, Taif 21944, Saudi Arabia.

Received 21th February 2023; Accepted 22th March 2023; Published online 30th April 2023

ABSTRACT

Predicting students' academic performance is one of the crucial issues in the academic field. Several methods and practices have been applied for educational improvement since there is a lot of academic information related to students. In this paper, our model for predicting students' academic performance based on academic and demographic factors is developed to predict the final course grade at early stages. By applying several machine learning (ML) algorithms, Linear Regression (LR), Support Vector Machine (SVM), and Random Forest (RF). This model to the data of male students of Technical college. The dataset contains 890 instances for 199 students. The result showed that the prediction's Mean Absolute Percentage Error (MAPE) reached 0.04% and the academic factors had a higher impact on student's academic performance than the demographic factors.

Keywords: Machine learning, student's performance, academic performance, random forest, support vector machine, linear regression.

INTRODUCTION

Technical colleges are one of the most important places for students where they fulfill their goals and aspiration to be specialized in career-driven courses. After coming from High School, many students find it difficult to adapt themselves to the college's environment for study due to several factors such as the method of training and assessment in courses. In some cases, students avoid studies in college or start to change departments. Training and assessment strategies used in all courses at all levels play a significant role in students' academic performance during the period of study in college. The student's academic performance is a concern of college administration and faculty members, they should monitor their student's academic performance and identify the students with low performance to provide prompt, effective support that helps to enhance student's skills in each course. Hence, the demand for technical colleges has been increasing over recent years. The number of enrolled students has grown. Therefore, the prediction of students' academic performance method can assist technical colleges to provide the needed actions at the appropriate time to avoid the student failure rate. Many factors have been considered that affect the student's academic performance (e.g., academic, demographical, etc.) [1]. Machine learning techniques can be used to predict the academic performance of the students and identify the low performance of the students to enhance their performance at an early stage based on the factors. The objective of this study is to use machine learning techniques to develop a prediction model of college students' performance based on diverse features. The features consist of the academic and demographic data of the students in the Technical Computer Department at Technical College of Ranya. This paper achieves the goal by including feature engineering to create the dataset, data collecting, data preprocessing, evaluating machine learning models and finding the best prediction's MAPE, and identifying the most important factors that help to predict students' academic performance.

***Corresponding Author: MOHANNAD E ALSOFYANI,**
1Department of Computer and Information Technology, Technical College of Ranyah, Technical and Vocational Training Corporation, Jeddah 24250, Saudi Arabia.

RELATED WORK

Altabrawee in [2], focused on the performance of the students. They identified their low performance to take the appropriate actions to enhance their performance. The authors introduced two new attributes that focus not only on the use of the internet as a learning resource and the effect of the time spent by students on social networks on the student's performance, but also the academic, demographical, cultural, educational background, and psychological profile. Four machine learning algorithms have been used to build a model that can predict students' performance in computer science at MU university. The machine learning algorithms include ANN, Naive Bayes, Decision Tree, and Logistic Regression. The dataset has the information of 161 students collected from the students at the College of Humanities during 2015-2016 using a survey and the student's grade book. The result showed that the prediction accuracy using ANN yields 77.04% and five factors detected by Decision Tree as important factors which influence the performance of the students.

In [4], authors have provided a model to predict the student's performance in 3 courses of 1170 students. The total score of every course is 100 marks and the marks are distributed between class tests, antecedence, presentations, assignments, mid-term, and final exams. Their main goal is to predict the performance of the final exam based on students' past events reports. By applying machine learning algorithms K-Nearest Neighbours, SVC, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Linear Discriminate Analysis the future result can be predicted. The result showed that the prediction accuracy using Decision Tree Classifier is 94.44% and 89.74% accuracy using KNN.

In [6], authors have provided models to predict how students will perform in their exams, concerning an added feature which is the student's family background to Demographic, Academic, and Behavioral features. The dataset collected from the Kalboard which is an online Learning Management System (LMS) used across the world consists of 480 instances with 16 features. Five machine learning applied are as follows: Gaussian Naïve Bayes, Support Vector Machine, Random Forest, K-Nearest Neighbour, and Logistic

Regression and we achieved a prediction accuracy of 50.83%, 81.67%, 78.33%, 75.00%, and 74.17% respectively.

In [7], authors have determined the current state of the research on predicting student academic performance as a systematic literature review. They found the features that have been used to predict student performance which is split into five categories: demographic (e.g., age, gender), personality (e.g., self-efficacy, self-regulation), academic (e.g., high-school performance, course performance), behavioural (e.g., log data) and institutional (e.g., high-school quality, teaching approach). In [8], authors have predicted students' performance in the final exam in the Mathematic course, the dataset from the Portugal's University of Minho, which includes 395 data samples. The collected data consists of several features which are academic and marital. SVM is somewhat more accurate than k-NN. In [9], authors have developed a web-based system for predicting academic performance and the total score of a course based on academic and demographic factors. The dataset contains 842 observations for 168 students. Five machine algorithms are applied: Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbours (KNN), Artificial Neural Network (ANN), and Linear Regression (LR). The results showed that the RF algorithm yielded the lowest MAPE of 6.34%, using academic and demographic factors to predict students' academic performance.

Research Motivation

Our research aims to predict students' academic performance total score for each course by considering new factors such as average absence, health status, and distance from college, as well as academic features such as grades of midterm and final exams, assignments, and presentations. As mentioned in [4], these features lead to enhance the accuracy of the model. We are applying the model to a dataset of men from Technical College. Our data differ from other researchers due to several reasons, including the nature of the courses which consist of theoretical and practical parts. However, in our model, we consider these factors.

DESCRIPTION OF THE PROPOSED METHODOLOGY

This work uses several machine learning algorithms to solve regression problems. Fig. 1, illustrates the steps of system architecture data collection, data pre-processing, splitting dataset, and regression.

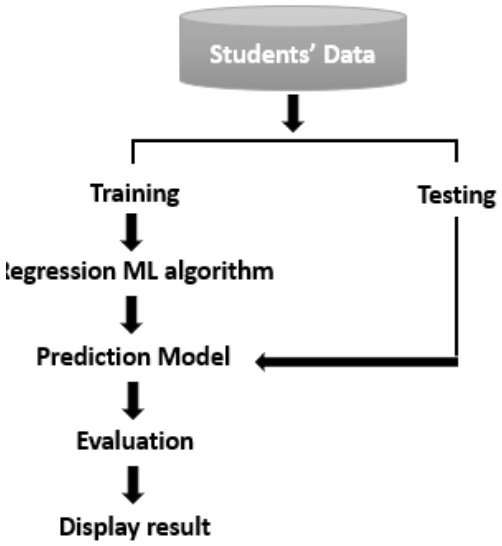

Fig.1. Architecture of the proposed system.

Data Source

In this study, we obtained data from the Computer Science department at Ranya Technical College for boys. We have 890 instances with 12 features for 199 students. The student's academic data were obtained from the Rayat Academic Affairs system. It was gathered during the first semester of the 2022-2021 academic year. All theoretical and practical courses taught were held on campus at the college. The features are partitioned into two categories: Academic, Demographic. The demographic data was collected via questionnaire, which are:

- Health status data: Whether or not the student has any diseases.
- Transportation data: Whether or not the student owns a car.
- Living place data: Includes how many minutes the student needs to travel from home to college.

Table 1 presents the diverse academic features and their corresponding Data domain and the description of the categories' dataset.

Table 1: Description of the categories' dataset.

NO.	Academic Features	Data Domain
1	Course Name	1-27
2	LAB1	0-10
3	LAB2	0-10
4	LAB3	0-10
5	LAB4	0-20
6	LEC1	0-10
7	LEC2	0-10
8	LEC3	0-10
9	LEC4	0-20
10	Total Score	60-100
11	ABSENCE	0-20
NO	Demographic Features	Data Domain
1	HEALTH STATUS	(1-0) Yes-No
2	DISTANCE	(10-45) minutes
3	OWN CAR	(1-0) Yes-No

It is worth noting that our academic features are numerical, whereas our demographic features are categorical. From these features, we will determine every student's performance by calculating our feature values (independent variables) to predict the target value which is the Total Score. These values are carried out by teachers based on students' performance in academic features.

Data preparations

To use the data, we must first pre-process it to remove the noise and outliers so that our model can work efficiently. To process our data in this study, we do the following steps:

Handle missing and outliers data

We must identify and handle missing values and outliers in our data to avoid inaccurate results and faulty conclusions. In our dataset, several students had grades below 60 out of 100 which led to missing values in other data columns and some students had over 20% average absence. These missing-values observations were removed.

Encoding the categorical data

Categorical data is a data type that can be divided into groups based on the names or labels. In the ML model, Independent and dependent

features are required to be numerical data. As a result, categorical data must be encoded before it can be used in the model. There are several techniques to encode categorical variables such as Label Encoding, One-Hot Encoding, and Hash Encoding.

In our dataset, we have 3 categorical features as shown in **Table1**. The health status feature describes the student's health status whether he has a chronic disease or not, and the own car describes whether the student has a car or not, these two features are considered a binary value, so we used Label-Encoder. The distance describes the expected time to get from home to college, it contains three parts the expected time is less than 10 minutes, more than 10 minutes and less than 25 minutes, and more than 45 minutes. Given our geographical and traffic knowledge in the area in which the college is located, if the student takes less than 25 minutes to reach the college that implies the student lives in the same city. On the other side, if the expected time is more than 45 minutes that implies that the student travel from outside the city. So, we combined the first two parts and we get a binary value, 1 assigned to less than 25 minutes and 0 assigned to more than 45 minutes.

Feature scaling

Features scaling is a method used to normalize a set of data features or independent variables, which leads to speeding up the training phase. It scaled the data points in a way to bound in the 0,1 range. You can perfume feature scaling in two ways:

Normalization as it is mentioned in equation (1):

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

X: Values in the dataset
X_{min}: The minimum value in the dataset
X_{max}: The maximum value in the dataset
X_{norm}: Normalized value in the dataset

Standardization as it is mentioned in equation (2):

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

x: The observation
μ: The mean
σ: The standard deviation

In this study, we will use Standardization for scaling our data features.

Splitting dataset

To build a machine learning model, we must split the original dataset into two phases. First, the training data set is used to train the model. Sconed, is a test data set that is used to predict the target value and evaluate the model. There are several validation techniques Hold-out and K-fold cross-validation used to calculate the true error rate. In our study, we found that Hold-Out 80:20 is more appropriate as shown in **Table 2**.

Table 2: Validation techniques.

Validation Technique	LR		SVM		RF	
Result of Predection	MAE	MAPE	MAE	MAPE	MAE	MAPE
Hold-out 80:20	6.25	0.07	8.80	0.11	3.87	0.04
5-Fold Cross Validation	6.74	0.09	9.25	0.12	3.89	0.05

Regression models and evaluation metrics

The main objective of the proposed system is the prediction of the student's total score based on several academic and demographic features considered as inputs and the prediction of the total score considered as outputs. Different regression algorithms (LR, SVM, RF) are compared to determine which algorithm is more appropriate. The evaluation metrics were used to compare the performance of all regression models. In this section, we will introduce an overview of ML models used in this study.

Linear regression

Regression methods are helpful to explore the relationship between a dependent variable and one or more independent variables. To make predictions, various regression techniques are available and these techniques driven by three factors as shown in Fig. 2.

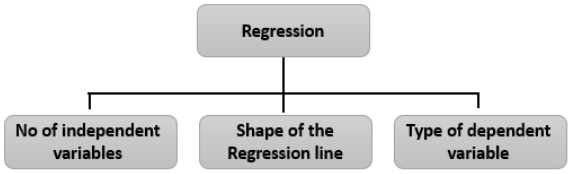

Fig 2. Factors lead regression techniques.

Polynomial regression is needed when there is no linear correlation between all the variables. Equation (3) presents a polynomial regression:

$$y = b_0 + b_1 x_1 + b_2 {x_1}^2 + b_2 {x_1}^3 + \cdots b_n {x_1}^n \quad (3)$$

In equation (3), y is the target value, and x is the feature variable. b0, b1, b2, bn are parameter factors. Polynomial regression is useful in many cases since the relationship between the target and feature variables is not required to be linear (Fig. 3).

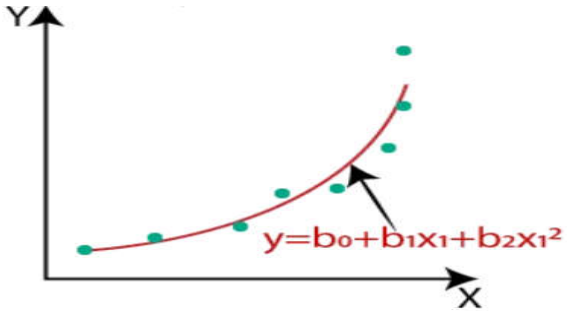

Fig. 3. Polynomial regression

Support vector machine

A support Vector Machine is a supervised machine learning algorithm capable of both classification and regression. It is mostly used to solve classification problems. SVM's basic idea is to find the best-fit line. The hyper plane with the greatest number of points is the best-fit line. The SVM tries to fit the best line within a threshold value. The threshold value is the distance between the hyper plane and the boundary line.

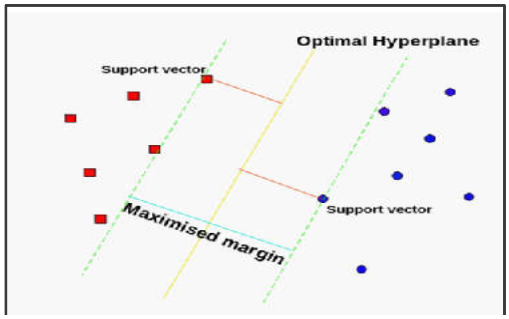

Fig. 4. Optimal hyperplane using SVM

Random Forest (RF)

RF is a supervised learning algorithm that was developed in 2001, RF is an ensemble learning method for classification, regression, and other tasks by constructing a group of randomly created decision trees and forecasting the class that is the mode of the classes (classification) or the mean (regression) of the individual trees.

EXPERIMENTAL SETTINGS

ML models were created using the LR, SVM, and RF algorithms. By using the hyper parameter value settings for algorithms as shown in **Table3**.

Table 3: Value of parameters of diverse machine learning algorithms.

Algorithms	Parameters	Value
LR	Intercept	True
	kernel	Rbf
SVM	C	25
	gamma	30
RF	n_estimators	200
	max_depth	13

Performance evaluation

All the supervised machine-learning algorithms mentioned in this study were deployed in Python. The accuracy of each model was determined using the performance metrics listed below.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is one of the most common performance metrics. It is used to calculate the model's prediction error. The MAE measures the average magnitude of the errors in a set of predictions. MAE is given by the Equation (4):

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (4)$$

Where y represents the prediction value, x represents the true value, and n represents the total number of data points.

Mean Absolute Percentage Error (MAPE)

MAPE is given by the Equation (5):

$$M = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \quad (5)$$

Where N represents the number of times the summation iteration happens, A represents the actual value, and F represents the prediction value.

RESULT AND DISCUSSION

Performance of the Regression Models

One of the objectives of this research is to know the best features that affect students' academic performance. Our dataset contains 7 features, the midterm exams score, and all demographic features (ALAB, ALEC, ALEC3, abs, diss, car, distance). In our model, authors predict the total score by using the features in our dataset. The following table shows the result of the regression models in terms of the prediction's MAPE and MAE. Random Forest had the lowest MAPE of.023 and the highest MAE of 1.869. The highest MAPE was obtained from the Linear Regression with 0.060 and an MAE of 4.866.

Table 4: Results of the regression model.

Algorithm	MAPE	MAE
Linear Regression	0.060	4.866
Random Forest	0.023	1.869
Support Vector Machine	0.025	2.06

Feature Importance

The idea of the feature importance is to identify the most important features that have a significant impact on student's academic performance of the dataset. The dataset contains many features, but some of the theses may not affect students' academic performance. After training the prediction model, the RF calculated the feature's importance.

Table 5: Ranking of different features.

Sequence	Feature Name	Ranking
1	ALEC	0.50105
2	ALAB	0.26973
3	Absence	0.06451
4	ALEC3	0.03354
5	Car	0.01650
6	Distance	0.01572
7	Disease	0.00859

It can be observed that the ALEC (average of midterm scores) is the highest rank which impacts the prediction process, followed by the ALAB. The demographic data Car, Distance, and Disease had little effect on students' academic performance. The reason is that most of the students in the dataset have their car, are from the same city, and are in a good health.

Based on the analysis of our data we found that the number of the students have disease less than the others, as it is shown in Fig. 5:

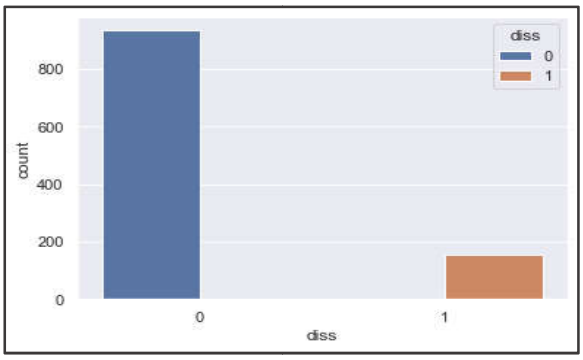

Fig. 5. Result of disease feature.

0 indicates students without disease, 1 with the disease.

Fig. 6 presents the number of students who have their cars equal to the other, 0 indicates students without cars and 1 with cars.

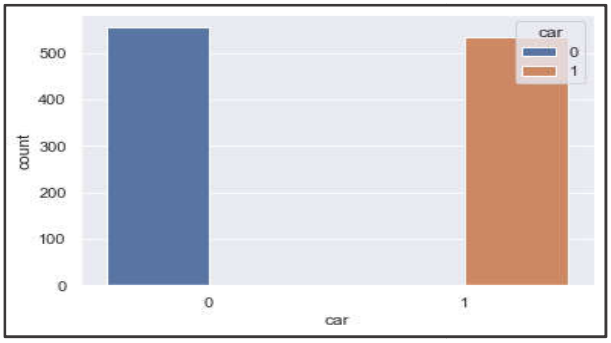

Fig.6. Result of car feature.

Fig. 6 presents the number of students traveling to the college, 1 indicates students from the same city cars and 2 outside the city.

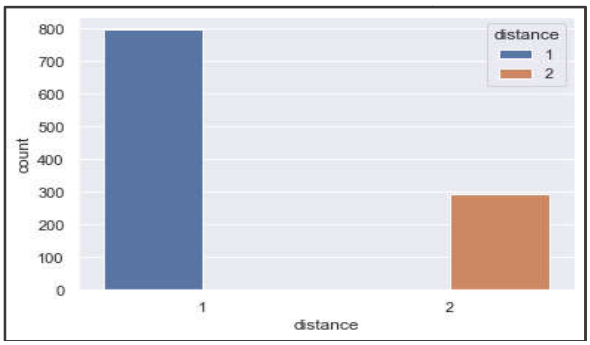

Fig.7. Result of distance feature.

Fig. 6 and Fig. 7 highlight the relationship between ALAB, and ALEC with final grade (ALL), there is a linear relationship between each other.

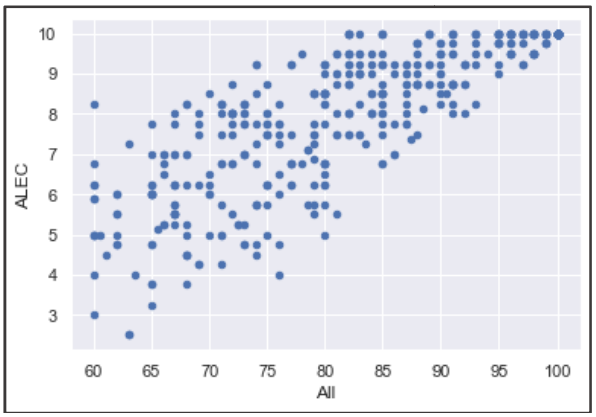

Fig. 6. Relationship between All and ALEC

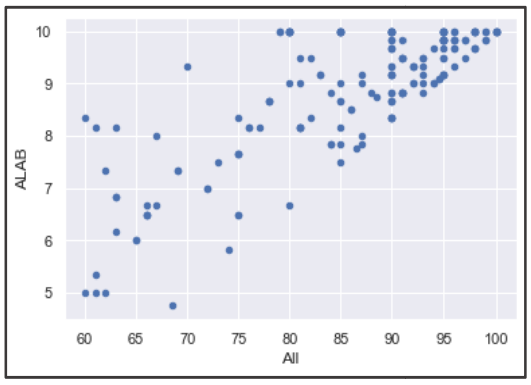

Fig. 7. Relationship between All and ALAB

Fig. 8, demonstrates the relationship between the practical and theoretical part of the course, there are negative trends between ALEC and ALAB, if the student gets a high score in ALEC that leads to getting a high score in ALAB.

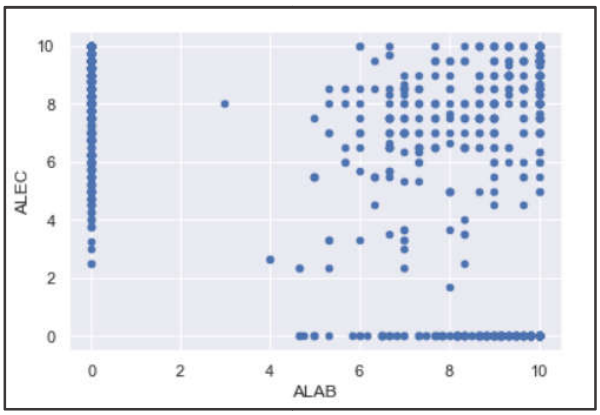

Fig. 8. Relationship between the practical and theoretical part of courses.

Discussion

By applying the importance feature technique, it was found that academic factors had a higher impact on student's academic performance, as in [2,4,9], while demographic factors had less impact. The midterm test score factor is the most significant in academic performance, this was consistent with the studies [6,8,9]. In addition, our study found that other factors such as health status (disses), own can (car), and the distance from college (distance) had the lowest impact on the students' performance. In our model, new features in the demographic factors were included, which are the absence of the course (absence) and if the student has his car or not (car). These factors have not been studied in previous research. Our result found that the impact of including these features on students' academic performance was a little similar to other demographic features.

CONCLUSION

In this study, we have developed an ML model to predict students' academic performance based on academic and demographic features. The data was collected from the Computer Science department at the Technical College of Ranya. We can get the most perfect result with LR, RF, and SVM and the lowest MAPE of 0.060,0.023, and 0.25. The model can be used to identify and improve the student's academic performance by notifying them in the early stages to enhance their drawbacks. Moreover, we used a dataset consisting of theoretical and practical courses as well as new factors that have not been studied in previous studies, which are the average absence and health status. In future work, the models can be developed to add more features such as the knowledge level of the English language. Our data contains the two-semester system, It would be a good idea to use the modern system of study in Saudi Arabia, which is the three-semesters system.

REFERENCES

1. Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education, 17, 1-21.

2. Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q. (2019). Predicting students' performance using machine learning techniques. JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences, 27(1), 194-205.
3. Hasan, H. R., Rabby, A. S. A., Islam, M. T., & Hossain, S. A. (2019, July). Machine learning algorithm for student's performance prediction. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.
4. Sixhaxa, K., Jadhav, A., & Ajoodha, R. (2022, January). Predicting Students Performance in Exams using Machine Learning Techniques. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 635-640). IEEE.
5. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., ... & Liao, S. N. (2018, July). Predicting academic performance: a systematic literature review. In Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education (pp. 175-199).
6. Alboaneen, D., Almelihi, M., Alsubaie, R., Alghamdi, R., Alshehri, L., & Alharthi, R. (2022). Development of a web-based prediction system for students' academic performance. Data, 7(2), 21.
7. Paramita, A. S., & Tjahjono, L. M. (2021). Implementing Machine Learning Techniques for Predicting Student Performance in an E-Learning Environment. International Journal of Informatics and Information Systems, 4(2), 149-156.
8. Figure3.https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18
9. Figure4.https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

****