Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 7
July 2023

Student Performance Prediction Using Regression Analysis & Feature-Based Opinion Mining on Student Feedback

Dr. M.V.Bramhe[*1] , Srushti Gohade[*2], Rupal Kapse[*3] , Sanchita Thamke[*4], Akash Mehar[*5] and Dr. Mohit Agarwal[*6]

[*1] Professor, Department of Information Technology, St. Vincent Pallotti College of Engineering and Technology, Nagpur (India)

[*2 3 4 5] Student, Department of Information Technology, St. Vincent Pallotti College of Engineering and Technology, Nagpur (India)

[*6] Assistant Professor, School of Computer Science Engineering and Technology, Bennett University, Noida (India)

Abstract- Predicting student performance in higher education is a crucial research area in today's technology-driven world. With advancements in technology, understanding, modeling, and predicting student performance has become increasingly beneficial and essential. However, accurately and robustly designing models for predicting student performance poses significant challenges. This study aims to explore the use of machine learning algorithms to predict student performance in higher education. The study collects and analyses data on various factors, such as student demographics, academic background, extracurricular activities, and engagement, using various machine learning algorithms. This approach can be beneficial for educators in identifying at-risk students and developing personalized interventions to improve their academic outcomes. This study emphasizes the importance of leveraging technology and machine learning techniques to predict and improve student performance in higher education. For educators and decision-makers, it offers insightful information on how to efficiently create interventions that enhance student success.
Feature-based opinion mining on student feedback is an important area of research that aims to extract valuable insights from large volumes of student feedback data. In this study, we performed sentiment analysis on student feedback comments using Natural Language Processing (NLP) approaches.

Keywords-Student performance, Machine Learning algorithms, Regression Analysis, Sentiment Analysis, Natural Language Processing (NLP), Data mining.

I. INTRODUCTION

Predicting students' academic performance is crucial for educators and administrators to support students' learning and academic success. Data mining techniques have become increasingly important in the field of education for analyzing and understanding students' performance. However, predicting students' academic imbalanced datasets commonly encountered in this field, and the lack of consensus on the most effective resampling methods for addressing this challenge.

Analyzing the data to derive meaningful insights can be a time-consuming and complex task for humans. However, data mining techniques can be used to efficiently and effectively analyze and discover

valuable insights from these data. By applying data mining techniques, researchers can uncover significant knowledge about students' academic performance and identify factors that may influence their success.

Overall, data mining techniques have the potential to provide educators and administrators with valuable insights into students' academic performance. However, careful consideration and evaluation of data mining techniques and methods are necessary to ensure that the insights derived from these techniques are accurate and meaningful for educational decision- making. Student feedback plays a crucial role in evaluating the effectiveness of teaching methods, curriculum, and infrastructure. Understanding the emotional responses of students toward their educational experience is vital in improving student engagement, motivation, and academic success. Traditionally, student feedback has been collected through surveys or interviews, which can be time- consuming and may not provide an accurate representation of the overall student sentiment. However, with the advancement of technology, sentiment analysis techniques have emerged as promising tools for analyzing student feedback.

Natural language processing and machine learning techniques are used in sentiment analysis to recognize and extract emotions and opinions from text data. Sentiment analysis can be used in the educational institutions to examine student's input from multiple sources, including course evaluations, social media, and discussion forums.

We analyzed student-provided data using machine learning methods to forecast student performance. We compared several algorithms, including linear regression, decision tree regression, Ridge regression, random forest regression, polynomial regression, to determine which produced the most accurate predictions. We evaluated the effectiveness of these algorithms using the root mean squared error (RMSE) values. After analyzing the results, we found that the polynomial regression algorithm provided the best results, with an RMSE value of 0.82, which was significantly lower than the other algorithms. This suggests that using polynomial regression is the most effective way to predict student performance based on the given data. These findings demonstrate the potential for machine learning algorithms to provide valuable insights into student performance, which could be used to improve educational outcomes and better support student success.

During another part of our research work we had conducted a feedback survey to understand how students felt about their courses and overall experience about their department. To gain insights from the data, we utilized opinion mining with the help of Natural Language Processing (NLP) tools such as NLTK. By analyzing the feedback, we were able to categorize it into positive, neutral, and negative sentiment categories. Opinion mining with NLTK is an effective way to gain insights from large amounts of feedback data. By categorizing feedback into positive, neutral, and negative sentiment categories, institutions can identify trends and patterns and make data- driven decisions to improve their services.

II. RELATED WORK

The [1] is primarily concerned with forecasting student academic achievement in Engineering subjects. Main objective of the paper is to use linear regression techniques to build a model which predicts the performance of the students in Engineering Discipline. The predictor or independent variables of the model contain how many hours spent on the internet in some activities based on the data collected. The output or dependent variable is the prediction of end semester examination grades i.e. CGPA (Cumulative Grade Points). Multiple measures are used to calculate and corroborate the models that were predicted along with the percentage of good predictions. The results show that the predicted model gives the better accuracy in prediction. a student's Cumulative Grade Point Average (CGPA), These models were used in the study to make precise predictions about how well students will succeed in Engineering courses and to pinpoint the key elements that influence their academic achievement in the course.

This study [2] suggests a two-layered LSTM (Long Short- Term Memory) model based supervised aspect-based opinion mining method. While the second layer specifies the direction of those expected features (positive, negative, or neutral), the first layer predicts the aspects that are mentioned in the feedback. The

model stands out from other LSTM models suggested for various domains due to its simplicity in terms of architecture. Both a conventional SemEval-2014 data set and a data set created from five years of student input from Sukkur IBA University were used to evaluate the model. Aspect extraction (91%) and sentiment polarity identification (93%), which are performed by the system in both tasks, are done with high levels of accuracy. The performance of the suggested methodology points to its potential use in automating the analysis of student feedback, which can aid educators and policymakers in identifying areas for improvement and enhancing the educational experience for students.

A crucial component of the training dataset for supervised machine learning algorithms is the traditional characteristics ofstudents, including their demographic data, academic history, and behavioral traits. Several supervised machine learning algorithms' performances, including Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine, K- Nearest Neighbor, Sequential Minimal Optimization, and Neural Network, were compared in the [3] study. The model was trained using datasets from courses in the bachelor's degree programs at the University of Basra's College of Computer Science and Information Technology
The study [4] suggests a fresh method for examining student comments given to professors. Based on specific quality indicators that have an impact on the feedback measure, this model utilizes a Nave Bayes Classifier to classify feedback as legitimate or invalid. The model's goal is to overcome the shortcomings of conventional models and offer a Faculty Effectiveness Index that is more exact and accurate. The validity of the feedback is used as the weight for calculating the index, which is generated as a weighted average of valid feedback measures. Using a Naive Bayes Classifier to divide feedback into valid and invalid categories, the model intends to increase the accuracy of faculty feedback analysis by only using the valid categories to calculate the Faculty Effectiveness Index.

In this study, [5] a sizable dataset of student evaluations of professor performance was analyzed using data mining and natural language processing. In order to extract common features and opinion words from the dataset, two well-known techniques, Apriori and Generalized Sequential Pattern (GSP) mining, were compared. After being crawled, pre-processed, and tagged, the student feedback data is then transformed into tri- model data files. Using the machine-learning application WEKA 3.7.10, both methods were applied to the prepared data. The algorithms' principles are then tested on different files to identify common traits and opinion words. The reviewed results showed that GSP is superior to Apriori for textual data mining. The [6] study created new multivariate linear regression models that can forecast student academic success in Engineering Dynamics using data gathered from 239 undergraduate students over the course of three semesters. Nearly all mechanical or civil engineering students are required to take this subject, making it a crucial topic of study. In order to forecast students' academic success in the course, the models employed their cumulative CGPA as the input or predictor variable. The study's findings indicate that the built predictive models may produce accurate predictions with a range of 44.4% to 65.6% and an average prediction accuracy of between 86.8% and 90.7%

The aim of this study [7] was to propose a methodology for monitoring and predicting success in higher education. The ultimate goal was to use this methodology to develop an individualized learning system. Based on the data collected in the study, the most commonly used technique for predicting student behavior was supervised learning, which was found to provide accurate and reliable results. The Support Vector Machine (SVM) algorithm was identified as the most frequently used algorithm by the authors and was found to offer the most accurate predictions. This is particularly important as there is only a limited amount of data available for further analysis

The Financial Opinion Mining and Question Answering Open Challenge, hosted at WWW 2018 in Lyon, France, provided the financial data used in this research [8], which proposed a transferred learning strategy for aspect categorization and a regression approach for sentiment prediction. Linear Support Vector Regressor produced the best results among the regression techniques used in the study, which also included BERT. In order to give readers a sense of recent developments in this field of study, the paper also contains a comparison of various available methodologies. Precision, recall, and F1- score for aspect classification, as well as MSE and R Squared (R2) metrics for sentiment prediction, are performance metrics that are used to assess how well the approach performs.

The [9] paper reviews earlier research that used a variety of analytical techniques to try to predict students' academic achievement. The cumulative grade point average (CGPA) and internal assessment were the datasets that were most frequently employed in these investigations. In educational data mining, the classification method was often employed, while Neural Networks and Decision Trees were the two most frequently employed methods for forecasting student achievement. The authors conducted additional study in their own setting as a result of the meta-analysis of these studies, which will assist the educational system

in more systematically tracking students' performance.

III. METHODOLOGY

Our research work was divided into five phases as data collection, data preprocessing, data visualization, data analysis and model building & testing using various machine learning algorithms for predicting the student performance.

Step 1: Data collection

To gather information on at-risk students and improve the learning environment proper data is needed. When collecting data for our project, we used Google Forms which is an effective and efficient way to gather information. The data is collected from the students of our college. The form was designed to gather information about certain parameters that may affect student performance in academics, such as attendance, study hours, extracurricular activities, and stress levels. By collecting this data, we aimed to createa model that can predict student performance based on these parameters, providing insights that can help improve academic performance and reduce stress levels for students.

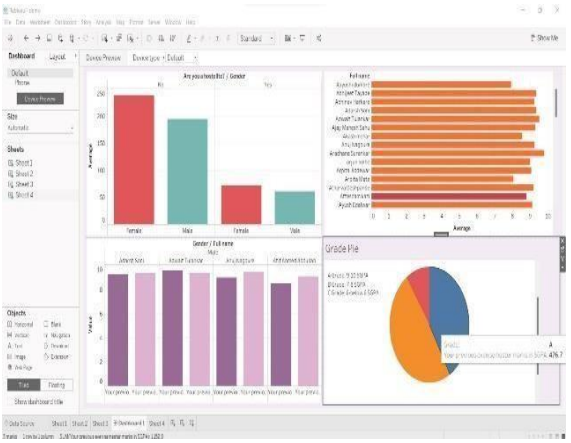
Step 2: Data Preprocessing

One of the most crucial activities that must be completed before a dataset can be used for machine learning is the pre-processing and cleaning of the data. The data from the real world is noisy, inconclusive, and inconsistent. It must therefore be cleaned.

Step 3: Data Visualization

Rephrasing information into a visual setting, such as a chart or graph, in order to make it simpler for the mortal brain to absorb and draw perceptivity from data is known as data visualization. Data visualization's major goal is to make it simpler to spot patterns, trends, and outliers in big data sets. After data has been collected, reused, and modeled, it must be imaged for conclusions to be made. For the data visualization, we've used the Tableau tool to more understand the relationship between colorful parameters. For conclusions to be drawn, data must first be collected, processed, and modelled before being visualized. For the data visualization, we have sed the Tableau tool to better understand the relationship between various parameters.

Fig 1. Data Visualization using Tableau



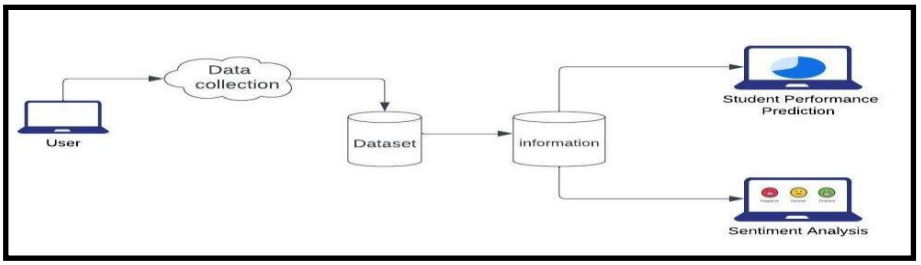
Step 4: Data Analysis

Data analysis involves using statistical and machine learning techniques to gain insights from the data and to identify patterns, relationships, and trends. This step is critical in turning the raw data into actionable information that can be used to make informed decisions. During data analysis, it is important to ensure that the data is clean, complete, and accurate. This may involve further preprocessing steps, such as removing missing values, scaling or normalizing the data, and performing feature engineering to create new features that can be used in the analysis.

Step 5: Model Building and Testing

After data analysis, the next step is to build models that can predict outcomes, detect patterns, or classify data. In this Step The models are tested to evaluate their accuracy and effectiveness. Training, validation, and testing sets were created using the provided data. The validation set is used to optimize the model once it has been trained, while the testing set is used to gauge the model's ultimate performance. Then, depending on how crucial they are to solving the issue at hand, the pertinent elements are chosen. Feature selection can enhance model performance and lessen overfitting. A suitable algorithm is used to train the model on the training set. Sets for training, validating, and testing have been created using the accessible data. The model is trained using the training set, validated using the validation set, then tested against the testing set to determine how well the model performed overall. Next, features that are pertinent to the issue at hand are chosen based on their significance. The model's performance and overfitting can both be enhanced by feature selection. Utilizing the proper ML based algorithm, the model is trained on the training set

System Architecture



IV. EXPERIMENT

We have conducted our experiment for student performance prediction using following hardware and software. We have used personal computer with Core i5 processor, 8 GB primary memory and NVidia GPU 940mc 2GB and Software Tools like Python 3.6, Anaconda Navigator, Jupiter Notebook and Tableau tool for the data visualization.

Various machine-learning techniques are used to train and test the obtained data in order to determine which model is most accurate at predicting student performance. In order to achieve this, we divided the data into training and testing sets and performed a variety of algorithms testing on the training set, including linear regression, decision trees, random forests, polynomial regression, Ridge regression, and lasso regression. The root means square error (RMSE), a commonly used statistic for gauging the accuracy of regression models, was then used to assess each algorithm's performance.

When comparing anticipated and actual values, the RMSE is calculated as thesquare root of the mean of the squared disparities. With declining RMSE values, the model's performance gets better. We were able to choose the best model for our project by comparing the RMSE values of various techniques. Following an analysis of the data, we discovered that the polynomial regression technique had the best outcomes, with an RMSE value that was 0.82, substantially lower than that of the other algorithms

	Model	RMSE	R2_Score(training)	R2_Score(test)	Cross-Validation
0	Linear Regression	0.479822	0.721212	0.421049	0.521390
1	Polynomial Regression (2nd)	0.821850	0.820091	-0.698501	0.521390
2	Ridge Regression	0.503268	0.703340	0.363088	0.330804
3	Lasso Regression	0.643864	0.025198	-0.042486	-0.139085
4	Decision Tree Regression	0.590519	1.000000	0.123104	0.309544
5	Random Forest Regression	0.483602	0.955445	0.411891	0.597527

Fig 2. Comparative analysis of Regression techniques

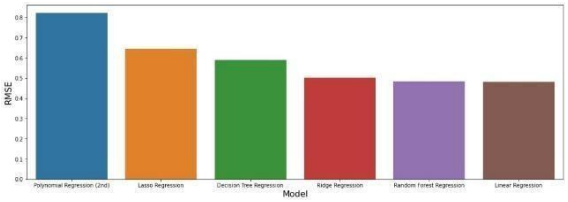


Fig 3. Comparison of RMSE value.

We trained our model using linear regression, which is a popular algorithm for predicting continuous variables. Linear regression works byfitting a linear equation to the data, which can then be used to make predictions based on new input variables.

After training the model, we used it tomake predictions on the testing data set. This involved inputting the features of a student, such as their GPA, study hours, and extracurricular activities, into the trained model, which then generated a predicted outcome i.e. their final grade or examscore.

```
[ ] from sklearn.linear_model import LinearRegression

    lireg=LinearRegression() # initialize the model
    lireg.fit(X_train,y_train) # fit he model
    y_pred=lireg.predict(X_test) # now predict
    y_pred

    array([9.03013429, 9.7105825 , 8.93118174, 8.50748337, 8.49141365,
           9.45106536, 7.75658635, 7.81703838, 9.07668749, 8.5303553 ,
           8.70247575, 8.49852955, 9.25883458, 9.14140341, 8.45711414,
           8.34720982, 9.2723246 , 9.28035713, 8.71033824, 9.28886115,
           9.22961514, 8.43899502, 9.16377803, 8.25105971, 8.68243001,
           8.28204787, 8.54190163, 8.58108513, 7.19803223, 9.15587668,
           8.43668234, 8.54808458, 8.45704203, 8.60327784, 8.6523481 ,
           8.56402223, 8.55489895, 9.28698121, 9.07067449, 9.04009721,
           9.04077159, 8.34851264, 9.08522015, 8.26764671])

[ ] print(y_pred)

    [9.03013429 9.7105825  8.93118174 8.50748337 8.49141365 9.45106536
     7.75658635 7.81703838 9.07668749 8.5303553  8.70247575 8.49852955
     9.25883458 9.14140341 8.45711414 8.34720982 9.2723246  9.28035713
     8.71033824 9.28886115 9.22961514 8.43899502 9.16377803 8.25105971
     8.68243001 8.28204787 8.54190163 8.58108513 7.19803223 9.15587668
     8.43668234 8.54808458 8.45704203 8.60327784 8.6523481  8.56402223
     8.55489895 9.28698121 9.07067449 9.04009721 9.04077159 8.34851264
     9.08522015 8.26764671]

[ ] X = data.iloc[:, :13].values
```

Fig 4. Predicted Values

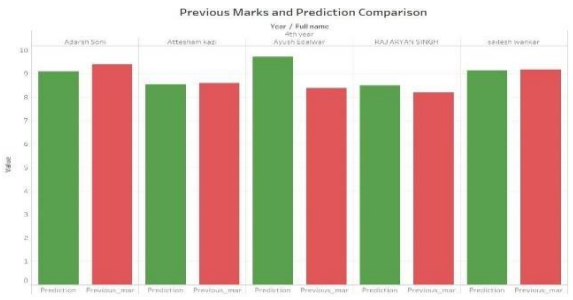


Fig 5. Accuracy comparison of the predictions

The second component of our research work was devoted to analyzing the survey responses of student feedback. We used the Natural Language Toolkit (NLTK) to do sentiment analysis on the textual data in

order to examine the emotions the students conveyed in their feedback. The first step in this process was to categorize the feedback into three categories: positive, neutral, and negative. To do this, each feedback response was subjected to a sentiment analysis algorithm using the NLTK toolkit. Each response received a score from this algorithm that represented the sentiment of the text, indicating whether it was favorable, negative, or neutral. After categorizing the feedback, we analyzed the results to gain insights into the students' overall sentiments about their academic experience. The use of sentiment analysis in our research work allowed us to gain a deeper understanding of the students' perspectives and experiences, and to use this information to make informed decisions about how to enhance the academic environment in the college.

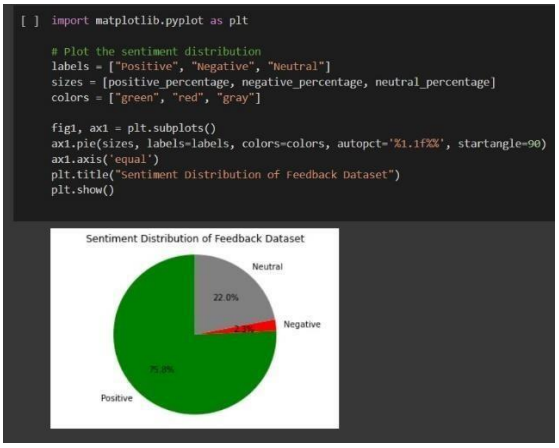
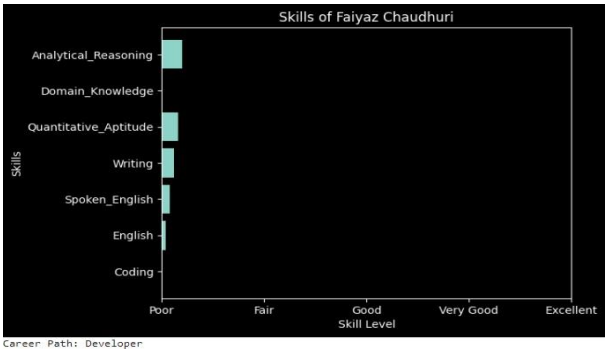


Fig 6. Sentiment Distribution of Feedback Dataset

We utilized the sample dataset for career path prediction, which included test scores in various topics such as domain knowledge, coding, quantitative writing, and english. We have analyzed the marks of each feature for every student in the dataset and used it to make predictions about their potential career paths. This involved identifying patterns and correlations between the students' test scores and different career paths, and drawing conclusions based on these findings. Each student's marks in these features can be used to identify their strengths and weaknesses, which can be further analyzed to determine their potential career paths. For instance, if a student scores high in domain knowledge and coding, they may be suitable for a career in computer science or software development. However, it is important to note that this dataset should not be the sole determinant of a student's career path. Other factors such as personal interests, experiences, and aspirations should also be taken into account. Therefore, a sample dataset for career path prediction based on test scores can be a useful starting point, but it should be combined with other



resources and guidance to provide a more comprehensive approach to career planning.

Fig 7. Career Path Prediction

Feedback analysis is a crucial process that helps institutions and organizations improve the quality of their services. By circulating a Google form that included various parameters such as punctuality, communication, interest in teaching, motivation, discipline, ethics, lab instructions, and lab evaluation, we were able to collect valuable data on the performance of the faculty members. By taking ratings from 1 to 5 for each particular faculty, we were able to identify the strengths and weaknesses of each individual in each area. The results of the feedback analysis allowed us to determine who was the best-performing faculty in each area, such as punctuality, discipline, or lab evaluation. This information can be used to recognize and reward the best-performing faculty members, as well as identify areas where improvements need to be made. By analyzing the feedback data, we can also identify common themes or issues that need to be addressed across the department, which can help to improve the overall quality of teaching and learning.

V. CONCLUSION

The method used for student performance analysis in our research is the regression model. Several Regression techniques were used to train and test the model, and a comparative analysis is conducted for the same. The feedback dataset is subjected to sentiment analysis using the Python Natural Language Toolkit (NLTK) package. Using sentiment analysis, we derstood

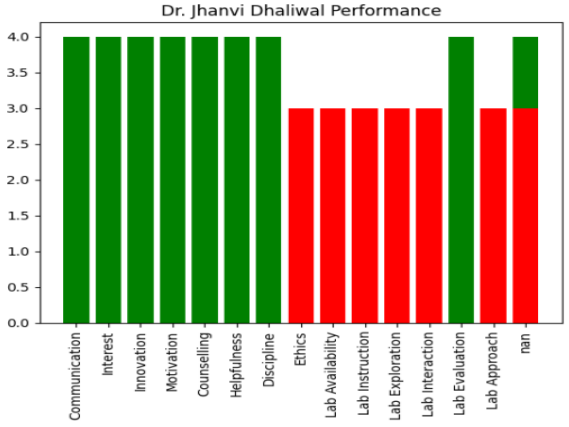


Figure 8. Feedback analysis Report

Overall feedback analysis is a valuable tool for improving the quality of teaching and learning. By collecting feedback from students and analyzing the data, institutions can identify areas for improvement and take action to address them, resulting in a better learning experience for students and a more effective teaching environment for faculty students' perception towards various aspects of education such as the curriculum, teaching methods, and infrastructure.

Predicting student performance can help teachers identify areas where their teaching methods are falling short. Teachers can use this information to adjust their teaching methods and improve student outcomes. Predicting student performance can help identify struggling students early on and provide intervention to prevent them from falling behind. Early intervention can help improve student outcomes and prevent the need for more intensive interventions later on. Student feedback analysis can be used to identify gaps in the curriculum and areas where more attention needs to be given.

Feedback mining can revolutionize how we collect, analyze, and act on student feedback. With the proper tools and techniques, educators and institutions can gain valuable insights into student experiences, identify areas for improvement, and provide personalized support to help students succeed. Feedback mining can identify common issues and challenges students face across different courses and programs. This can help institutions develop targeted interventions to address these issues.

REFERENCES

[1] Rajalaxmi R R, N. Krishnamoorthy, P Natesan, "Regression Model for Predicting Engineering Students Academic Performance", International Journal of Recent Technology and Engineering (IJRTE), Volume 7, Issue 653, April 2019

[2] Irum Sindhu, Sher Muhammad Daudpota , Kamal Badar, Maheen Bakhtyar, Junaid Baber, Mohammad Nurunnabi "Aspect- Based Opinion Mining on Student's Feedback for Faculty Teaching Performance Evaluation", IEEE Access, July 2019

[3] Ali Salah Hashim, Wid Akeel Awadh, Alaa Khalaf Hamoud "Student Performance Prediction Model based on Supervised Machine Learning Algorithms", 2nd International Scientific Conference (ISCAU), Volume 928, July 2020

[4] Sandhya Maitraa, Sushila Madanb, Rekha Kandwalc, Prerna Mahajan "Mining Authentic Student Feedback for Faculty using Naïve Bayes Classifier", International Conference on Computational Intelligence and Data Science(ICCIDS-2018)

[5] Ayesha Rashid "Feature Level Opinion Mining of Educational Student Feedback Data using Sequential Pattern Mining and Association Rule Mining", International Journal of Computer Applications, · November 2013

[6] Shaobo Huang, Ning Fang "Regression Models for Predicting Student Academic Performance in an Engineering Dynamics Course", American Society for Engineering Education, 2010

[7] M. S. Sassirekha, S. Vijayalakshmi "Predicting the academic progression in student's standpoint using machine learning", Journal of Control, Measurement, Electronics, Computing, Communications, Volume 63, Issue 04, 2022

[8] Ashish Salunkhe, Shubham Mhaske "Aspect Based Sentiment Analysis on Financial Data using Transferred Learning Approach using Pre-Trained BERT and Repressor Model", International Research Journal of Engineering and Technology, IRJET Volume 06, Issue 12, Dec 2019

[9] Wasyihun Sema Admass "Review on Predicting Student Academic Performance using Data Mining Classification Algorithm", Journal of Computer Engineering and Information Technology, Volume 10, issue 11, November 2021