Personalized Learning Pathway for Students Using Different Machine Learning Approach

Arya Jalindar Kadam[1], Tejaswini Jaywant Durge[2], Ratnmala N. Bhimanpallewar[3]
Department of Information Technology
Vishwakarma Institute of Information Technology
Pune, India
[3]aryakadam348@gmail.com, [2]tejaswinidurge41@gmail.com, [1]ratnmalab@gmail.com

***Abstract*—This paper introduces a machine learning approach to create personalized learning paths for university students based on their academic and career performance. By analyzing data such as Cumulative Grade Point Average (CGPA), hackathon participation, internships, Aptitude test scores and student's domain of interest to design a customized road-map for their personal growth. It uses data from previously placed students, including their referred courses and resources, to suggest similar learning materials to students with matching academic and technical profiles. It then matches the students interests and suggests students with correct and reliable pathway.**

The introduced paper aims to assist students in enhancing areas where they require the most support. In addition to incorporating a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis, the primary objective is to pinpoint each student's individual progress and the areas where improvement is needed, the platform guides students toward taking real, actionable steps to close their skill gaps, boosting their confidence and making them more competitive in the job market. This approach not only helps students improve academically but also prepares them for better career opportunities.

***Index Terms*—Machine learning, SWOT Analysis, Personalized Road-map, Random Forest, K means, Decision tree**

Abbreviations: MSE (Mean Squared Error)

I. Introduction

In today's competitive world, students may struggle to identify the steps needed for academic progress and career preparation. Traditional lectures from teachers and peers can be helpful, but often lack the personalized approach needed to meet the unique and strong needs of each student. This is where machine learning provides solutions today, providing data-driven insights to create personalized learning experiences.

Traditionally, students have relied on a trial-and-error approach to their studies and careers. They seek advice from many people, including teachers, classmates, and seniors, for a roadmap to learn new technologies. However, the issue is that excessive data can degrade the performance of the model, just as too many suggestions can mislead students' paths. Additionally, this approach is overly time-consuming and uncertain because every student's level of understanding differs. Therefore, there is a risk of not receiving appropriate guidance for that domain. Furthermore, some seniors may not be from the same domain, and that will lead to inappropriate suggestions. Students will often consider suggestions that aren't reliable not even our domain-oriented.

Our study suggests a solution that is intended to solve all the above problems by giving them a road map with individual recommendations. The guidance includes a road map for various technical domains depending on the student's interest. Gathered data is particularly from seniors who are placed in reputed companies. Based on the academic performance and aptitude score Personalised guidance is based on the student's symmetry with placed seniors. Each student's profile is compared with the success stories of those who have a job and provide resources referred by placed students.

This strategy helps the students to concentrate on the right skills making them more confident when in search of a job. As they are clear with mindset and are now proficient in their domain of interest. Thus, such an approach is helpful as it gives the students more insights into what aspects to target, how one can manage the unknown, and the tactics to follow in approaching their future work. This is not only time-saving but also, offers the students a road map of their desired and achievable academic and/or career dreams. Thus, the students are well-prepared to compete for a job and gain their desired positions in leading organizations.

II. Literature Review

[1] This paper explores the K-means algorithm and its accuracy variation depending on the distribution of clusters. Making sure that centroids are well-spread from the beginning helps to improve clustering accuracy and avoid less optimal solutions.

[2] Research investigates how high-achieving students often follow fractal learning using network models and deep learning to examine these opportunities. According to these learning patterns, outcomes indicate that Deep Learning is capable of predicting student success.

[3] This research predicts student's performance using supervised machine learning algorithms based on different machine learning models. Their work provided early predictions, helping educators identify weaker students for targeted interventions.

[4] The classification results of random forest and J48 decision trees on 20 different datasets were compared. The study concluded that random forest performed better with larger datasets, while J48 is more efficient for smaller datasets,

illustrating how dataset size impacts model accuracy. Hence, Random Forest can work for this research.

[5]This paper focused on predicting engineering students' academic performance using C4.5, ID3, and CART algorithms. The model aimed to identify those students who are not performing well and offer them strategic and logical suggestions to enhance their outcomes. Particularly benefits academically challenged students.

[6] Paper proposed a predictive modeling approach to assess students' strengths and weaknesses through a SWOT analysis. Integrating machine learning and data visualization, the model helped students enhance academic proficiency.

[7] The study analyzed how student engagement influences academic performance in both traditional and nontraditional students. The research highlighted that higher engagement levels positively influenced relationships with faculty and peers, impacting academic success.

[8] explored personalized learning approaches for distance education using educational data mining. They emphasized self-paced learning and engagement through adaptive learning systems, improving knowledge retention and the learning experience.

[9] investigated complex systems in education, showing how deep learning networks can model learning pathways and accurately predict student performance. The study also demonstrated that learning pathways have emergent properties.

III. Methodology

A. Traditional Student Pathway Suggestion

The traditional way of finding personalised pathway to success are as follows:

Students use to collect data from various sources like their peers and teachers to find what suits their interest and what is the perfect path to follow in future. Then the student used to collect the data and analyse it and then follow respectively but even then there is no chance of getting success, as the student might fail in the middle of the pathway cause of various undiscovered difficulties.

This is not only time consuming but very confusing and does not provide a lot of assurance which leads to confusion and misleading. Sometimes its important to know the success of the person suggesting the pathway itself so as to assure it. Moreover, this approach depended heavily on external opinions and lacked a structured or data-driven process. Students had to rely on trial and error, often wasting valuable time and effort on paths that turned out to be unfit. The lack of personalized guidance meant that students could end up in pathways that didn't suit their individual learning styles, interests, or strengths. This uncertainty and lack of tailored direction made the traditional pathway gathering inefficient and frustrating, leading to missed opportunities or setbacks when the chosen route proved incompatible with their capabilities or aspirations.

B. Machine Learning Custom Pathway Suggestion

The new methodology for student pathway suggestion using machine learning:

This Application uses the earlier student data to train the model and analyse the students capacity and the reliability with the curriculum. By collecting student data from various students which include the column seen in the Table I.

As seen in the Table I there are three distinct columns hence there is a use of three different machine learning models in this algorithm.

TABLE I
Features for Training Models

Test Scores (9 cols)	**Coding and Platform Scores (5 cols)**	**Extra Activities (4 cols)**
These include the test scores from the syllabus and from other similar examinations	This includes other scores like on coding platforms	These include the extra activity scores that the student might have taken part in like hackathons and internships

1) Model 1 : Coding and Platform Scores: This part of the data is trained on Unsupervised models so as to make clustering of the data points. To find the effective unsupervised clustering model for this specific data set different tests were done. Below are the test outcomes:

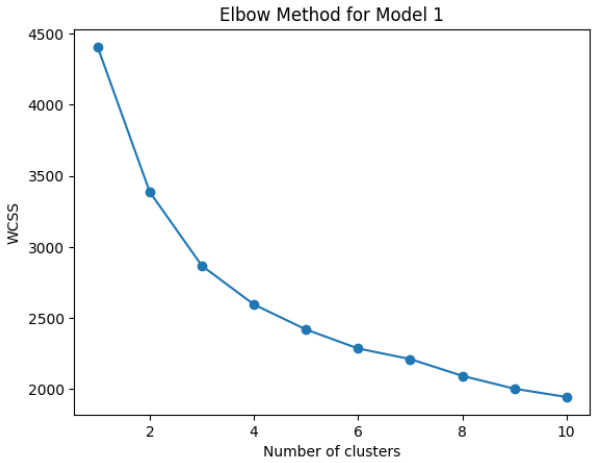

Fig. 1.

Firstly to decide the number of clusters elbow method for model 1 was taken and from the Fig 1 it was decided to consider the number of clusters to be 4.

From the Fig 2 it is clear that there are four distinct clusters formed from the Model 1 which can be further used for the prediction

From the Hierarchical clustering model which can be analysed from the Fig 3, 3 different clusters were identified and were plotted in the from of dendogram. Fig 2 it is clear that there are four distinct clusters formed from the Model 1 which can be further used for the prediction.

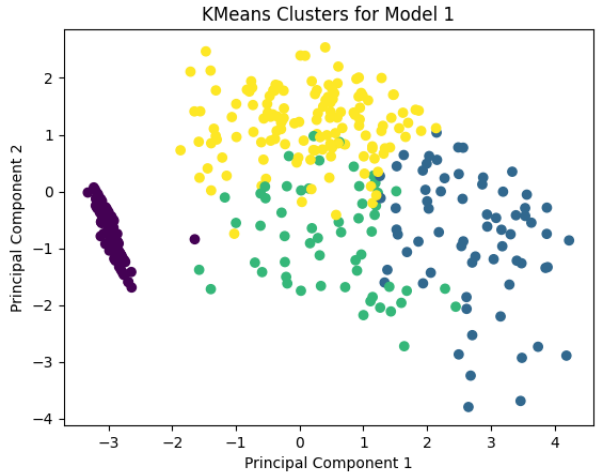

Fig. 2.

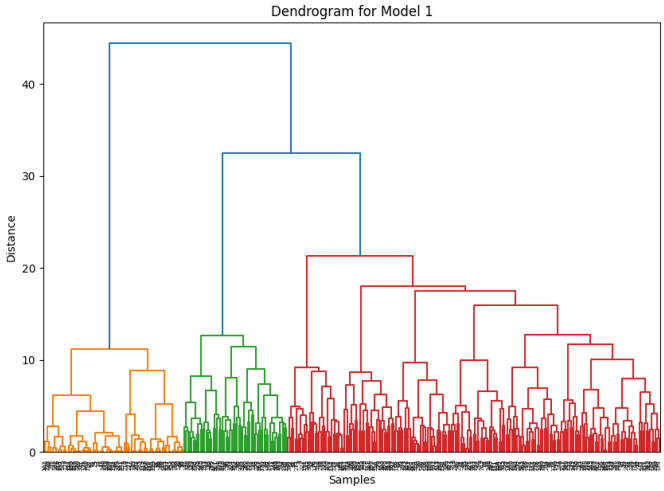

Fig. 3.

After Analysis the MSE of K Means is found to be less than that of Hierarchical clustering model. Hence the K Means model is selected as Model 1.

2) Model 2 : Extra Activities: This part of the data is trained on Unsupervised models so as to make clustering of the data points. To find the effective unsupervised clustering model for this specific data set different tests were done. Below are the test results:

Firstly to decide the number of clusters elbow method for model 2 was taken and from the Fig 4 it was decided to consider the number of clusters to be 3.

From the Fig 5 it is clear that there are four distinct clusters formed from the Model 1 which can be further used for the prediction. Similar observations are seen as per in Model 1. It is observed that between the 2 models the means square error for K means model was found to be less than the hierarchical clustering model and hence the K means was chosen.

3) Model 3: Test Scores: Several supervised learning models were tested, including Random Forest (RF), Decision Trees

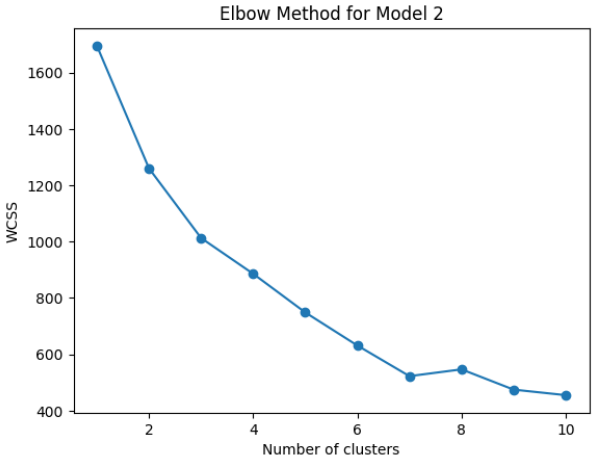

Fig. 4.

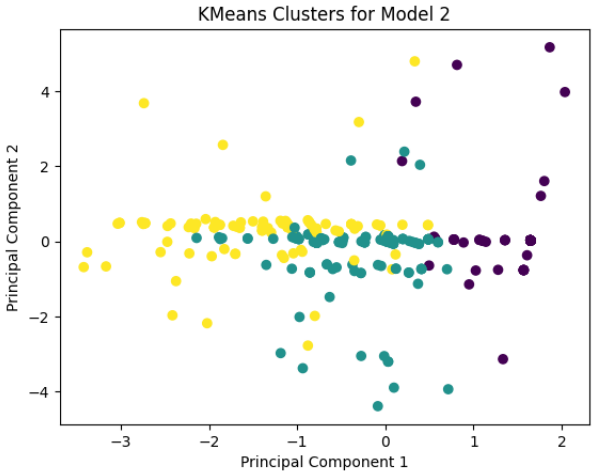

Fig. 5.

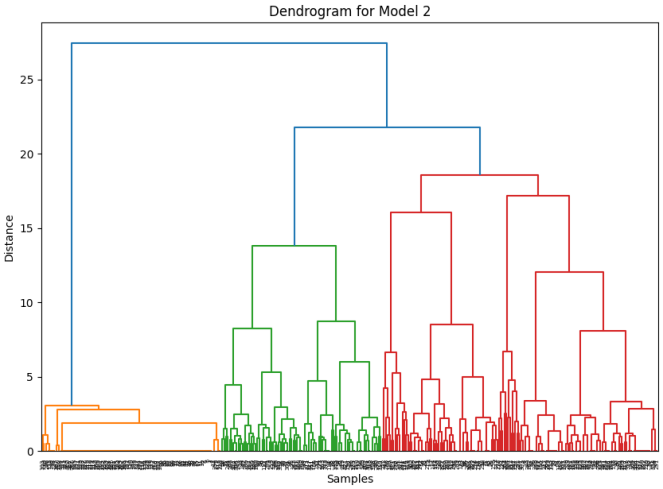

Fig. 6.

(DT), Decision Tree (DT), Deep Neural Networks (DNN) and Logistic Regression (LR). The comparison results are as follows:

MSE (Mean Squared Error)

- LR MSE: 86.2151
- RF MSE: 1.101
- DT MSE: 1.651
- DNN MSE: 17.866

From the from these results, it is clear that Random Forest (RF) significantly outperformed the other models, achieving the lowest MSE of 1.101. This indicates that RF provided the most accurate predictions, effectively capturing the underlying patterns in the data. Decision Trees (DT) also performed well, with an MSE of 1.651, though not as robust as RF. Deep Neural Networks (DNN) and Logistic Regression (LR) had notably higher MSE values, with LR being the least effective model for this dataset.

Given the superior performance of Random Forest, it was selected as the most suitable model for further predictions and analysis in this context.

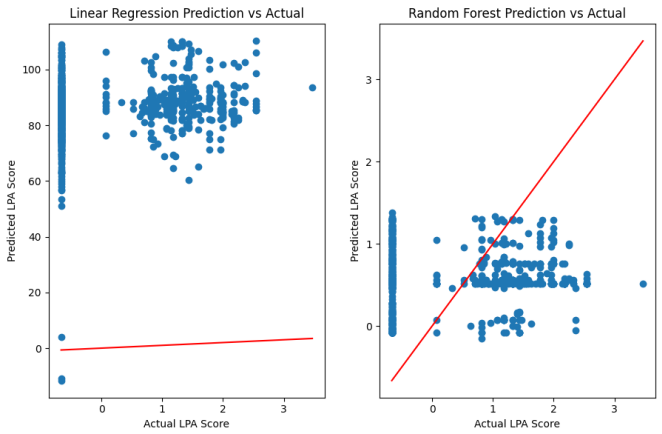

Fig. 7.

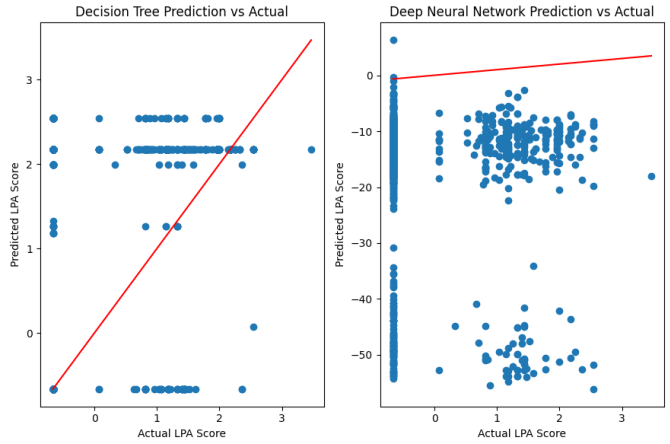

Fig. 8.

Fig 7 and Fig 8 shows that Random Forest can best predict the scores as it can fit the data more perfectly than the other models.

C. Final Pathway Suggestion

A student data is collected which includes all the described columns in Table I. This student data is collected from senior/alumni/graduates students from the institute. Along with the other data as described earlier it also includes the pathway which that specific student followed during his student life in that institute

Suppose we need to find the correct pathway for the student 'A'. The three different models can be used for three different tasks the

- Model 3: This model will be used to authenticate the student data and tell if the student's suggested pathway is reliable or not, whether it's up to the mark or not. This model will output a 1-100 score/rating that will decide whether this student has done better in the past or will do better in past. Which alternately tell us if its suggested pathway reliable.
  After the student rating for this dataset has been predicted Model 1 and Model 2 can be used to match the similarities of the student 'A' with this dataset students (senior/alumni/graduates).
- Model 1 and Model 2: By using the model predict method we could find the models clustering group for this 'A' student. We run the same predict method for all the students in the student dataset.

After this is done we can find the same student that matches the same interest as well as lies in the same cluster that the student 'A' lies in and suggests the dataset student's pathway if the dataset student rating is high enough.

IV. Algorithm: Understanding Regression

Regression analysis is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to predict the dependent variable based on the values of the independent variables. Here, we will explain the basic mathematics behind linear regression, which is one of the simplest forms of regression.

A. Linear Regression

Linear regression aims to fit a linear equation to observed data. The equation of a simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

where:

- y is the dependent variable (the outcome we are trying to predict).
- x is the independent variable (the predictor).
- β_0 is the y-intercept (the value of y when $x = 0$).
- β_1 is the slope of the line (the change in y for a one-unit change in x).
- ϵ is the error term (the difference between the observed and predicted values of y).

1) Estimating the Parameters: The parameters β_0 and β_1 are estimated using the method of least squares, which minimizes the sum of the squared differences between the observed and predicted values of y. The formulas for the estimates are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3}$$

where:

- $\hat{\beta}_1$ is the estimation of slope.
- $\hat{\beta}_0$ is the estimation of y-intercept.
- $\bar{x}$ is the mean of independent variable which is x.
- $\bar{y}$ is the mean of dependent variable which is y.
- n is number of observations.

2) Making Predictions: Once the parameters are estimated, we can use the regression equation to make predictions. For a given value of x, the predicted value of y is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{4}$$

where $\hat{y}$ is the predicted value of the dependent variable.

B. Evaluating the Model

The performance of the models can be measured by various ways but Mean Squared Error performs the best. This measures the average of the squared differences between the observed and predicted values. The formula for MSE is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

where:

- y_i is the observed value.
- $\hat{y}_i$ is the predicted value.
- n is the number of observations.

A lower MSE indicates a better fit of the model to the data.

C. Clustering with K-Means

K-Means clustering is used to partition data into k clusters. The algorithm aims to minimize the within-cluster sum of squares (WCSS), which is defined as:

$$\text{WCSS} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \|x_i^{(j)} - \mu_j\|^2 \tag{6}$$

where:

- k is the number of clusters.
- n_j is the number of points in cluster j.
- $x_i^{(j)}$ is the i-th point in cluster j.
- μ_j is the centroid of cluster j.

The centroids are updated iteratively to minimize WCSS until convergence.

D. Standardization of Data

Standardization is a preprocessing step that transforms the data to have a mean of 0 and a standard deviation of 1. The formula for standardization is:

$$z = \frac{x - \mu}{\sigma} \tag{7}$$

where:

- z is the standardized value.
- x is the original value.
- μ is the mean of the data.
- σ is the standard deviation of the data.

This ensures that the data is on a comparable scale, which is crucial for algorithms like K-Means and regression.

E. Random Forest for Classification

Random Forest is an ensemble learning method used for classification and regression. It operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. The prediction for a new instance x is given by:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{8}$$

where:

- T is the number of trees.
- $h_t(x)$ is the prediction of the t-th tree.

Random Forests reduce overfitting by averaging multiple trees, which improves generalization.

V. Results

An Interactive Application is made with the help of python and tkinter python-library. The UI of the application window can be seen in Fig 9

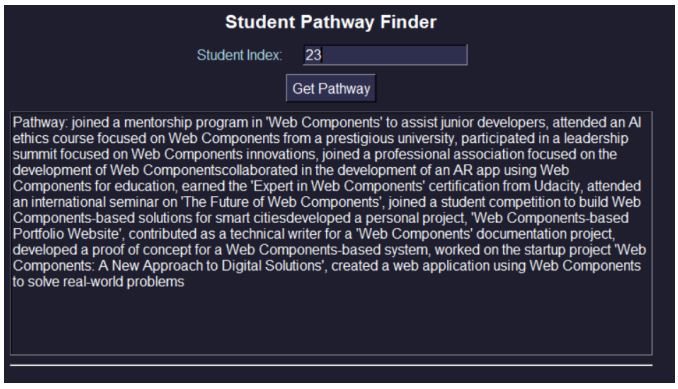

Fig. 9. UI Element of the Application

The application outputs in the form of strings in a simple text box. This includes the pathway suggested by the senior students who's interest matches with that of the tested student. By entering the students index number the student can get

its personalized pathway according to the machine learning done in the background. This tool provides a streamlined and data-driven alternative to the traditional trial-and-error method, significantly reducing the time and effort needed to find the best personalized pathway for students.

VI. Comparison with Other Methodologies

Analyzing the results of using several machine learning algorithms for predicting students' achievements and learning paths, it was revealed that certain differences both in terms of accuracy and time were observed between different methods. An article one availed itself of concept of K-means clustering in forming clusters of students that has similar characteristics where the study held that how the center was placed initially could be adjusted to increase the efficiency of clustering as it was believed that a number of solutions were suboptimal [1]. Likewise, deep learning models have shown how they can use the analysis of learning sequences to predict the students' performance revealing that successful learners follow complex, fractal-like learning patterns [2].

Some other related research works have used supervised learning methods such as decision tree, Bayesian network, and random forest for student performance prediction. These models helped identify early markers of the students who needed special attention to prevent them from falling behind [3]. Comparing between Random Forest and J48 in decision trees showed that Random Forest was perfect for large datasets than the J48 algorithm for the small datasets [4]. Another study was concerned with the forecast of student academic performance particularly on engineering students using C4. 5, ID3, and CART algorithms that assisted in the selection of students that were at risk of poor performance and where to get guidance [5].

Furthermore, several authors also introduced the models based on the SWOT analysis where ML solutions determined students' strengths and weaknesses and provided strategic action plans for academic achievements [6]. Student engagement was also examined in terms of impact on the academic performance of students and it was found that higher engagement level has a positive correlation towards the overall attitude students have towards faculty and vice versa affecting the students' output [7]. Specific tailored-learning strategies based on EDM have been illustrated to enhance learners' rates of retention and their level of interest, mainly in distance-learning situations [8]. Further, complex learning pathways have been modeled in deep learning networks and reliable predictions have been made about students' performance indicating that such pathways contain emergent properties [9].

VII. Conclusion

In our study, we described the application of machine learning when constructing a learning support schema for university learners. Using the data received from the tutorials, course results, number and kind of additional credit activities, and student's preferences, the platform offers the students recommendations for their future studies and occupations. To find the most precise pathways, the system processes the data with such advanced models as Random Forest, K-Means clustering, and Decision Trees.

In terms of the concept, it is more specific and well structured as compared to the traditional methods of seeking help from classmates, teachers, or seniors. It minimizes guesswork and increases the time they spend on activities they have to undertake in areas where they are weak, and their programs in anticipation of their careers. The inclusion of a SWOT, analysis introduces yet another strong feature that will enable one to see his weaknesses and how they can be dealt with, on the platform.

The created interactive application that simplifies pathway recommendations makes the integration of data and scientific approaches to learning with education effectively, which in turn enhances students' performance. That is way more advanced than conventional tutoring and mentoring which may only require a student to wait for their appointment with a tutor or an advisor who may then offer them an invaluable opinion and advice; it offers prompt opinions and proper advice which makes it a great progress in academic advising with students in our program having a complete and effective guide and preparation towards job market.

References

[1] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm," Phys Procedia, vol. 25, pp. 1104–1109, 2012, doi: 10.1016/j.phpro.2012.03.206.

[2] P. Ortiz-Vilchis and A. Ramirez-Arellano, "Learning Pathways and Students Performance: A Dynamic Complex System," Entropy, vol. 25, no. 2, Feb. 2023, doi: 10.3390/e25020291.

[3] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," in IOP Conference Series: Materials Science and Engineering, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1757-899X/928/3/032019.

[4] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," 2012. [Online]. Available: www.IJCSI.org

[5] S. Kumar and S. Kumar Yadav, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification," 2012. [Online]. Available: https://www.researchgate.net/publication/221710771

[6] Dr. L. A. Deshpande, "Revolutionizing The Education Industry: Swot Analysis And Predictive Modeling Approach To Enhance Students' Educational Proficiency," Educational Administration: Theory and Practice, pp. 3758–3765, May 2024, doi: 10.53555/kuey.v30i5.3530.

[7] A. S. Courtner, "Impact of Student Engagement on Academic Performance and Quality of Relationships of Traditional and Nontraditional Students," International Journal of Education, vol. 6, no. 2, p. 24, May 2014, doi: 10.5296/ije.v6i2.5316.

[8] V. Desai and K. Oza, "Personalized Learning Approach for Outcome Based Distance Education." [Online]. Available: https://www.researchgate.net/publication/378364655

[9] M. J. Jacobson, "Complex systems in education: Scientific and educational importance and implications for the learning sciences," 2006. [Online]. Available: https://www.researchgate.net/publication/220040401