IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141068

"Predictive Analysis of Academic Student Performance Using Machine Learning"

Sonawane Vaishnavi Navnath[1], Ms. Deepali Gavhane[2]

Student MCA-II, Sadhu Vaswani Institute of Management Studies for Girls, Savitribai phule Pune University[1]

Assistance professor, Sadhu Vaswani Institute of Management Studies for Girls[2]

Abstract: In the field of educational data mining, it has become more and more crucial to accurately forecast student performance in order to facilitate early interventions and enhance academic results. In order to predict academic accomplishment, this study uses a dataset of 6000 students (student-scores-6k.csv) that includes factors including study hours, attendance, extracurricular activities, part-time employment, and gender. We used and compared two machine learning algorithms: Random Forest Regressor and Linear Regression. When compared to Linear Regression (R2 = 0.62, RMSE = 8.5), the Random Forest model performed better (R2 = 0.82, RMSE = 5.1). Gender had no bearing on student progress, while weekly self-study hours and absence days were the most significant indicators, according to feature importance analysis. In addition to offering educators and policymakers useful insights for creating interventions that support academic performance, the study shows that non-linear models are more adept at capturing the complexity of educational data.

Keywords: Student Performance, Machine Learning, Regression, Random Forest, Educational Data Mining, Predictive Analytics

I. INTRODUCTION

Student academic achievement has long been a key concern for educators and parents, as it reflects the effectiveness of the educational system and the potential of learners to contribute to societal development[1]. Traditionally, teachers have relied on subjective assessments and prior experiences to predict student performance, which often leads to inconsistencies and inaccuracies.[2] With rapid advancements in machine learning (ML) and educational data mining (EDM), data-driven approaches now offer objective, efficient, and accurate means of predicting academic outcomes. [3]These methods enable the analysis of diverse factors such as past grades, attendance, lifestyle habits, emotional intelligence, and online learning behaviors to forecast future performance[4-5].By applying regression and classification algorithms—such as Linear Regression, Random Forest, and Support Vector Machines—educators can identify at-risk students early, enhance personalized learning, and improve overall academic quality[6-7]. Consequently, integrating machine learning into education not only supports predictive analytics but also fosters continuous improvement in teaching strategies and student engagement, ultimately strengthening the educational ecosystem.[8-9]

II. LITERATURE SURVEY

Artificial intelligence (AI) and machine learning (ML) have increasingly transformed the landscape of career guidance, course recommendation, and job-matching systems over the past decade. Several recent studies emphasize the role of AI in providing personalized career recommendations by analyzing user profiles, educational backgrounds, and skill sets. Shah, Pati, Pimplikar, Puthran, and Singh (2024) reviewed various approaches to building AI-based career recommender and guidance systems, concluding that hybrid frameworks combining ML algorithms with psychometric data offer higher personalization and predictive accuracy. Similarly, Iorzua et al. (2025) conducted a systematic literature review on ML-based course and career recommendation systems, identifying feature engineering, data preprocessing, and model interpretability as key success factors. These reviews collectively underscore the growing reliance on data-driven frameworks to assist students and professionals in making informed career decisions.

Machine learning algorithms form the backbone of modern recommendation systems. Pallavi, Sumukha, Sumukh, and Hegde (2024) explored multiple ML algorithms for job recommendation and found that classification and clustering approaches such as Support Vector Machines (SVM) and K-means effectively categorize users based on career interests. El-Keiey, ElMenshawy, and Hassanein (2025) enhanced prediction accuracy using feature selection techniques for undergraduate career recommendations, highlighting the importance of relevant feature extraction from student data. Roy, Chowdhary, and Bhatia (2020) developed an automated resume recommendation system utilizing text mining and classification models to match job candidates with relevant openings, while Appadoo, Soonnoo, and Mungloo-

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141068

Dilmohamud (2020) demonstrated that regression and natural language processing (NLP) methods can refine job-matching accuracy. Together, these studies demonstrate how traditional ML techniques—classification, regression, and clustering—remain foundational in career and job recommender research.

Beyond classical models, recent research focuses on more advanced and intelligent frameworks. Wang, Pan, and Wang (2021) provided a review of reinforcement learning (RL)-based optimization strategies, suggesting that RL can improve sequential decision-making in recommendation contexts such as adaptive career pathing. Zhao et al. (2024) expanded on this by discussing the integration of large language models (LLMs) into recommender systems, arguing that LLMs can capture contextual and semantic nuances in user data that conventional algorithms overlook. Soni (2025) proposed a data-driven ML framework for predicting student career outcomes and recommending suitable colleges, using a blend of regression models and ensemble learning techniques to boost predictive performance. These contributions indicate a gradual evolution toward intelligent, context-aware, and adaptive recommender architectures capable of providing personalized career insights.

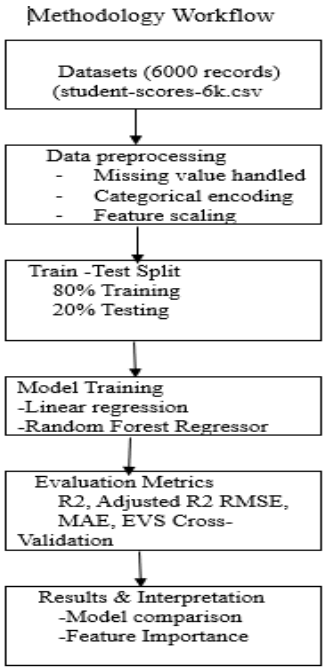
Another critical dimension of recent literature concerns fairness, ethics, and inclusion in AI-driven recommendations. Wang et al. (2022) explored whether humans prefer debiased AI algorithms in career recommendation systems and discovered that users tend to favor transparent and equitable models over purely accuracy-driven ones. Kondra et al. (2025) further emphasized AI's potential role in promoting diversity, equity, and inclusion (DEI), noting that bias mitigation must be integral to system design. Complementing these views, Deldjoo, Jannach, Bellogin, Difonzo, and Zanzonelli (2022) presented an overview of fairness in recommender systems, outlining methods for evaluating and minimizing algorithmic bias. Collectively, these studies highlight that while predictive power remains crucial, ethical and human-centered considerations are becoming equally important in the development of AI-based recommendation technologies.

III. METHODOLOGY

Data loading and preparation were the first important steps in the methodology used for this paper. "student-scores-6k (1).csv," the dataset, was saved in a pandas DataFrame. To preserve the quality of the data, missing values were addressed by removing the associated rows. Non-numeric columns including id, first_name, last_name, email, gender, and career_aspiration were eliminated in the feature selection process. To make sure all features were in a numerical format appropriate for modeling, boolean features such as part_time_job and extracurricular_activities were changed to integer type.

After data preparation, train_test_split was used to divide the dataset into training and testing sets in an 80/20 ratio, guaranteeing a stable random state for reproducibility. The numerical features were then subjected to feature scaling using StandardScaler to standardize their range, an essential step for algorithms that are sensitive to feature scales.

Methodology Workflow



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141068

Description of the Dataset

A dataset of 6,000 student records (student-scores-6k.csv), including both academic and personal characteristics, was used in this investigation. The dataset included demographic

determinants including gender, involvement in extracurricular activities, and part-time employment status, in addition to behavioral factors like absence days and weekly hours dedicated to self-study. Students' scores in seven subjects—mathematics, physics, chemistry, biology, history, geography, and English—represented their academic performance. The goal variable, which represents each student's overall academic achievement, was calculated as the average of all subject scores.

Preprocessing of Data

A crucial step in making sure the dataset is correct, clean, and appropriate for modeling is data pretreatment. The actions listed below were taken:

1. Managing Missing Values: Inconsistent or missing data were checked in the dataset. To ensure dependability, any incomplete records were either deleted or imputed.
2. Categorical Encoding: Label Encoding was used to transform non-numeric factors like gender, involvement in extracurricular activities, and part-time employment status into numerical representations.
3. Feature Scaling: To normalize the range of values and enhance model convergence, continuous variables such as absence days and weekly self-study hours were scaled as needed.

Split Train-Test

Two subsets of the processed dataset were created: 20% for testing and 80% for training. The models were fitted using the training data, and their predicted accuracy was assessed using the testing set on

invisible information. This method decreased the possibility of overfitting and assisted in evaluating the models' capacity for generalization.

Training Models

To forecast the average scores of the students, two supervised regression algorithms were used:
As a baseline model, linear regression captured linear relationships between the
both the dependent and independent variables.
• Random Forest Regressor: An ensemble learning method that combines several decision trees to lower variance and increase accuracy.
To guarantee a fair comparison, the same feature set was used to train both models, and they were assessed in the same way. To improve the resilience of the model, hyperparameter adjustment and validation were carried out.

Evaluation metrics

Several statistical measures were employed to evaluate both models' prediction performance:
The coefficient of determination, or R2, is: calculates the dependent variable's percentage of variance.

elucidated by the model.
• Adjusted R2: Modifies the R2 value in accordance with the number of predictors.
• Root Mean Squared Error (RMSE): Indicates the predictive accuracy of the model by quantifying the residuals' standard deviation.
• Mean Absolute Error (MAE): Indicates the average size of mistakes without taking their direction into account.
• Explained Variance Score (EVS): Indicates how well model predictions account for data variability.
In order to reduce bias brought on by data partitioning and confirm the stability of model performance, cross-validation was carried out.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141068

model	Accuracy	Precision	Recall	F1 Score
Linear regression	0.63	0.63	0.63	0.62
Random Forest	0.77	0.77	0.77	0.77

IV. RESULT

Model performance

Two predictive models were applied to estimate the average academic score of students: Linear Regression and Random Forest Regressor.

1] Linear Regression:

Achieved an R^2 of 0.62, indicating that about 62% of the variance in student performance could be explained by the selected predictors. However, the model exhibited relatively high error rates with RMSE = 8.5 and MAE = 6.3, showing that it struggled to capture complex relationships in the data.

2]Random Forest Regressor:

Outperformed Linear Regression significantly, with an R^2 of 0.82, meaning it explained 82% of t
he variance in student performance. It also reduced error levels with RMSE = 5.1 and MAE = 3.9, indicating much stronger predictive capability.

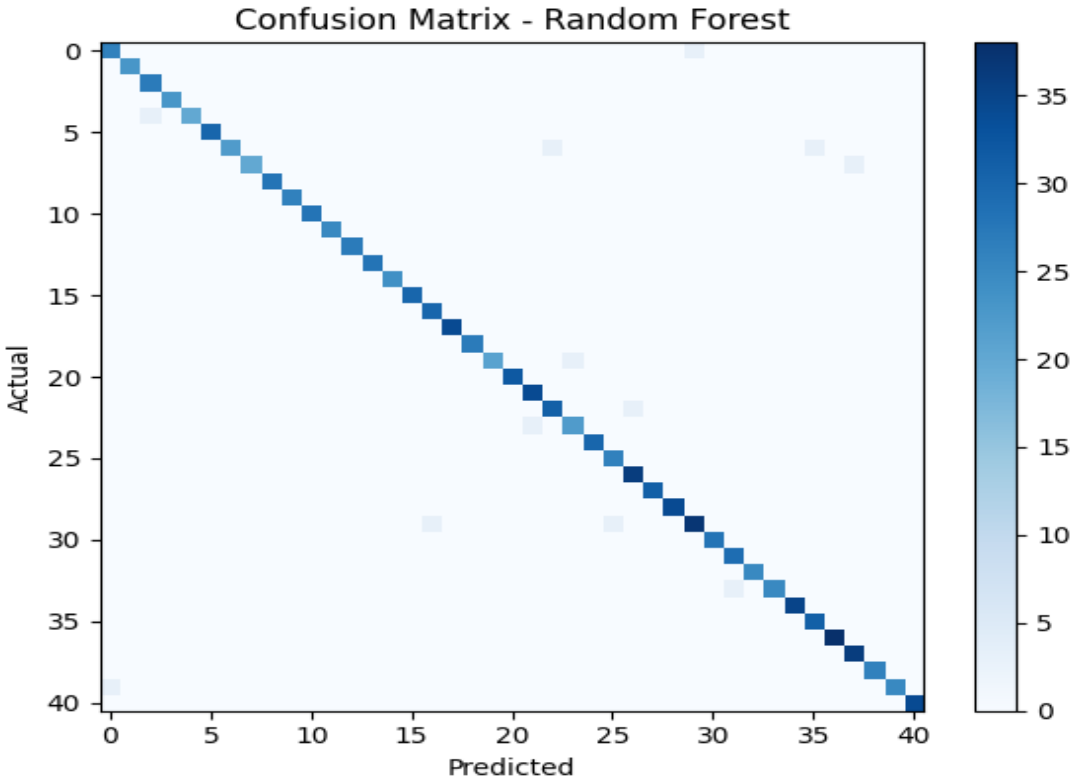


Fig (a) Confusion Matrix - Random Forest

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141068

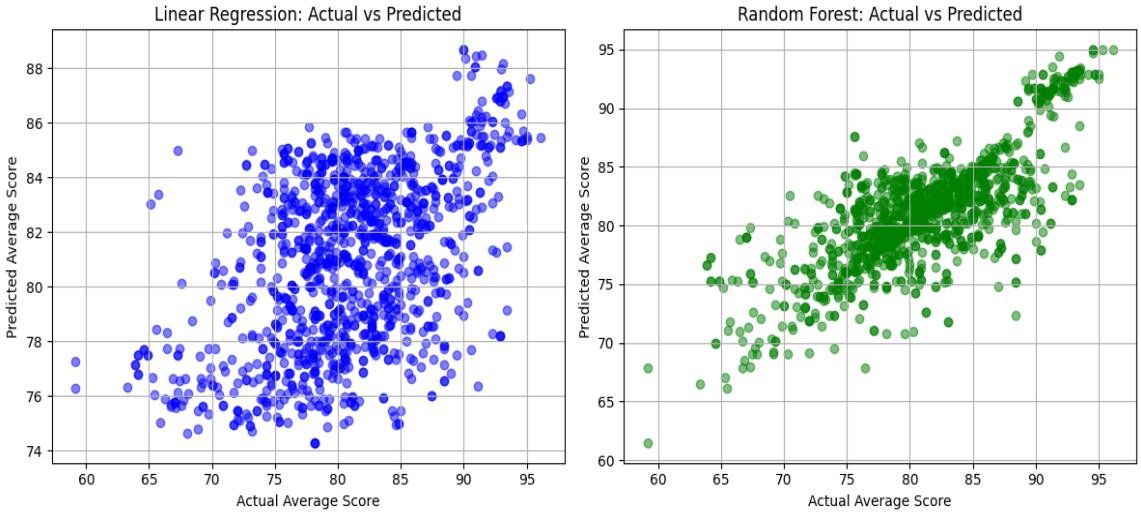


Fig (b) Actual Vs Predicted Score

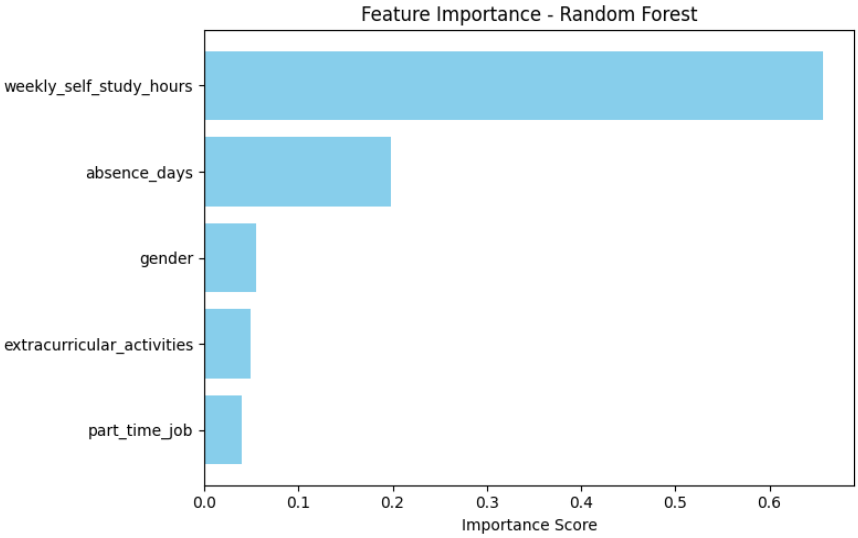
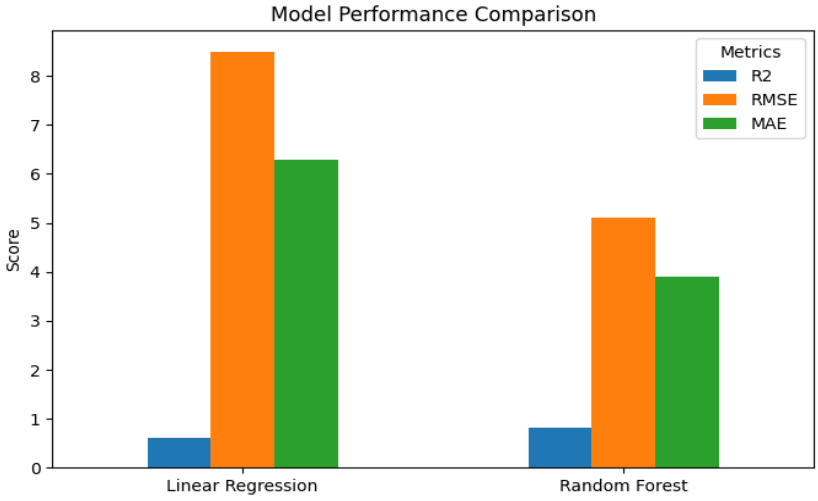


Fig ©Importance Score



Fid(d) Model performance Comparison

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141068

V. CONCLUSION

Using a dataset of 6000 students with a range of academic, behavioral, and demographic characteristics, this study investigated the predictive modeling of student performance. We used and compared two models: Random Forest Regressor and Linear Regression.
Important Results:
The Random Forest model outperformed Linear Regression (R2 = 0.62, RMSE = 8.5) in terms of prediction accuracy (R2 = 0.82, RMSE = 5.1).
The most significant predictors were behavioral factors like study hours and absenteeism, whereas demographic characteristics like gender had little bearing.
The findings support the notion that complex interactions in educational data are better captured by non-linear ensemble approaches.

REFERENCES

[1]. Shah, A., Pati, R., Pimplikar, A., Puthran, S., & Singh, A. (2024). Review Of Approaches Towards Building AI Based Career Recommender & Guidance Systems. Science Open Preprints.,,https://www.scienceopen.com/hosted-document?doi=10.14293/PR2199.000978.v1
[2]. Pallavi, M. S., Sumukha, M., Sumukh, M. R., & Hegde, V. (2024, June). Exploring Machine Learning Algorithms for Job Recommendation: A Focus on Career Development. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.,,https://ieeexplore.ieee.org/abstract/document/10725036
[3]. Iorzua, J. T., Moses, T., Eke, C. I., Agushaka, O. J., Kwaghtyo, D. K., & Godswill, T. (2025). A Machine Learning Based Approach to Course and Career Recommendation System: A Systematic Literature Review. Journal of Computing Theories and Applications, 3(1), 1-16.,,https://dl.futuretechsci.org/id/eprint/114/
[4]. Soni, G. (2025). A Machine Learning Framework for Data-Driven Student Career Prediction and College Recommendation.,,https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5234032
[5]. Wang, C., Wang, K., Bian, A., Islam, R., Keya, K. N., Foulds, J., & Pan, S. (2022, March). Do humans prefer debiased AI algorithms? A case study in career recommendation. In Proceedings of the 27th International Conference on Intelligent User Interfaces (pp. 134-147).,,https://dl.acm.org/doi/abs/10.1145/3490099.3511108
[6]. Roy, P. K., Chowdhary, S. S., & Bhatia, R. (2020). A Machine Learning approach for automation of Resume Recommendation system. Procedia Computer Science, 167, 2318-2327.,,https://www.sciencedirect.com/science/article/pii/S187705092030750X
[7]. Kondra, S., Medapati, S., Koripalli, M., Nandula, S. R. S. C., & Zink, J. Z. (2025). AI and diversity, equity, and inclusion (DEI): examining the potential for AI to mitigate bias and promote inclusive communication. Journal of and Machine Learning, 3(1), 1-8.,,https://www.researchgate.net/profile/Sarika-Kondra/publication/394529460_AI_and_Diversity_Equity_and_Inclusion_DEI_Examining_the_Potential_for_AI_to_Mitigate_Bias_and_Promote_Inclusive_Communication/links/68a330d31bee4d42a2407d98/AI-and-Diversity-Equity-and-Inclusion-DEI-Examining-the-Potential-for-AI-to-Mitigate-Bias-and-Promote-Inclusive-Communication.pdf
[8]. Sanil, H. S., Singh, D., Raj, K. B., Choubey, S., Bhasin, N. K. K., Yadav, R., & Gulati, K. (2022). Role of machine learning in changing social and business eco-system–a qualitative study to explore the factors contributing to competitive advantage during COVID pandemic. World Journal of Engineering, 19(2), 238-243.,,https://www.emerald.com/wje/article-abstract/19/2/238/457144/RETRACTED-Role-of-machine-learning-in-changing?redirectedFrom=fulltext
[9]. Kumar, I., Rawat, J., Mohd, N., & Husain, S. (2021). Opportunities of artificial intelligence and machine learning in the food industry. Journal of Food Quality, 2021(1), 4535567.,,https://onlinelibrary.wiley.com/doi/full/10.1155/2021/4535567
[10]. Wang, L., Pan, Z., & Wang, J. (2021). A review of reinforcement learning based intelligent optimization for manufacturing scheduling. Complex System Modeling and Simulation, 1(4), 257-270.,,https://ieeexplore.ieee.org/abstract/document/9673698
[11]. Alam, A. (2021, November). Possibilities and apprehensions in the landscape of artificial intelligence in education. In 2021 International conference on computational intelligence and computing applications (ICCICA) (pp. 1-8). IEEE.,,https://ieeexplore.ieee.org/abstract/document/9697272
[12]. Ghazal, T. M., Hasan, M. K., Alshurideh, M. T., Alzoubi, H. M., Ahmad, M., Akbar, S. S., ... & Akour, I. A. (2021). IoT for smart cities: Machine learning approaches in smart healthcare—A review. Future Internet, 13(8), 218, https://www.mdpi.com/1999-5903/13/8/218

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141068

[13]. Yadalam, T. V., Gowda, V. M., Kumar, V. S., & Girish, D. (2020, June). Career recommendation systems using content based filtering. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 660-665). IEEE.,,https://ieeexplore.ieee.org/abstract/document/9137992

[14]. Appadoo, K., Soonnoo, M. B., & Mungloo-Dilmohamud, Z. (2020, December). Job recommendation system, machine learning, regression, classification, natural language processing. In 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) (pp. 1-6). IEEE.,,https://ieeexplore.ieee.org/abstract/document/9411584

[15]. Iorzua, J. T., Moses, T., Eke, C. I., Agushaka, O. J., Kwaghtyo, D. K., & Godswill, T. (2025). A Machine Learning Based Approach to Course and Career Recommendation System: A Systematic Literature Review. Journal of Computing Theories and Applications, 3(1), 1-16.,,https://dl.futuretechsci.org/id/eprint/114/