**Module Code and Title: 7164CEM**

**Individual Research Project Preparation**

**Course Work Title: Project Proposal**

**Component1: Project Proposal**

**Component 2: Initial Literature Review**

**Student ID: 14155136**

**Student Name: Jahnavi Priya Ratakonda**

[**Ratakondaj@uni.coventry.ac.uk**](mailto:Ratakondaj@uni.coventry.ac.uk)

**Component 1: Project Proposal**

**Title: Student’s Academic Success and Dropout Prediction using Machine Learning**

**Aim:**

This study aims to help students stay in college and succeed by utilizing advanced technology called Machine Learning. This technology helps us to predict students at considerable risk of dropping out early in their studies. We also focus on how students conduct oneself and use this data to find signs that suggest they might be facing. By doing this, we wish to give the right kind of assistance to help students overcome these difficulties and achieve their academic goals

Dropout, enrolled, and graduate are the three categories of the classification task in which the problem is composed. The distinct data sources and algorithms are utilized to provide prompt remedies and eventually enhance student persistence and academic success.

**Introduction:**

In education, one of the major hurdles for educators and administrators is to anticipate the academic success of students and their likelihood of dropping out. To implement effective support structure and interventions, it is important to comprehend the contributing factors that affect student’s determination from their studies. To address these critical challenges the acceptable solution that has developed is predictive analysis, driven by the latest advancements in machine learning. Predictive models are tools that help educational institutions identify which students might struggle or drop out of university. These models use a lot of information, like how well students are doing in their classes, where they come from, and how they behave. By using this data, universities can identify students who might need help before they start to face difficulties. This helps create an understanding and supportive learning environment that helps students to achieve their goals.

Introduction lays the foundation for complete research into how we can predict when students might drop out of university or when they might succeed academically. Early student withdrawals from universities can also have important economic and social costs for both educational institutions and their students. The repercussions of this can involve low chances of employability, loss of tuition fees, and reducing the university's fame.

**Dataset** [**https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success**](https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success)

**Problem Statement:**

In the previous researchers they used stacking ensemble techniques and applied those techniques to the algorithms like KNN (K-Nearest Neighbours) and Logistic Regression to achieve the objective of enhancing the overall student dropout and academic success predictions with the accuracy of 82.3% and 82.1% (Prasanth & Alqahtani, 2003) we can see that the accuracy level is normal and still students are dropping out from the universities and spoiling their life in order to stop this permanently still some predictions should be done by changing the attributes, to get highest accuracy and solve this problem completely.

**Solution to the problem:** But in this research, we are using other machine learning techniques to increase the accuracy and predictions on student dropout and academic success, which is helpful for the students to achieve their goals, As social media platforms are technically updated and user friendly these days we can conduct many surveys and know the reasons behind the student academic dropout and decrease in success rates. Along with the three below mentioned algorithms I wanted to use LightGBM to get better results and accuracy for solving this specific problem.

**Primary Research Plan:** There are lot of things to mention in this research plan like the methods that we used to identify the predictions, and experiments, software tools etc.

**Requirements:** There are certain Standards before initiating a project.

1. First a fall we must study thoroughly to understand the dataset and the features present in it.
2. Study about the students' academics, struggles that they are facing, Major reasons for dropout and their behavior and academic performance in the universities which gives us some ideas to find out the solution to the problem.
3. I need to know the previous techniques and methods the researchers applied to improve the accuracy and predictions by applying the advanced machine learning algorithms to achieve the student's dropout and academic success predictions.

**Software Specifications:**

* To train this student’s dropout and academic success prediction we need to use one software application to find out the student’s data in the universities, for this dataset I wanted to use Anaconda Prompt which is best to provide a command line interface to manage environments, install packages and run scripts.
* It also permits the user to generate virtual environments which are separated python environments that permits the installation of specific packages and dependencies without impacting the system's global configuration.
* So, to work on the present dataset I wanted to write a code in python language in a jupyter notebook tool in which this tool empowers users to run code, edit it and see the results in real time.
* It also assists the integration of distinct built in libraries like Pandas, Numpy, Sklearn , packages, making data analysis and visualization a breeze.
* Jupyter Notebook and Anaconda Prompt are the two important tools broadly used among data scientists and developers for data analysis, Visualization and specially in Machine Learning.

**Models:**

Machine Learning Algorithms:

1. XGBoost Classifier
2. Logistic Regression
3. KNN (K-Nearest Neighbors)

**Methods:**

**Data Gathering:**

Regarding to this project, the UCI repository database offers a dataset that provides information about the student’s academic performance, their behavior in the universities, how they involve in the campus events are useful to predict the student’s early stage of dropouts in which a lot of universities, students and their parents are suffering with this early stage of dropout. Universities are facing a lot of problems regarding this issue and even students are not able to achieve their goals. So based on this major concern all the student's data is collected from this dataset and apply machine learning techniques to solve this issue.

**Data Preprocessing:**

Once the dataset is collected then we must look for the features present in it, after seeing all the features we must choose some of the features which are important for our project. The features I chose are Marital Status, Student-ID, Course, Application mode, Tuition fees up to date, Age at enrollment, Gender, Nationality, Previous Qualification grade.

Fig 1. Steps to forecast the outcome

**Models:**

Out of all the Machine Learning models, the following four classifier models are identified as the best to get high accuracy.

.

**XGBoost Classifier:**

* This algorithm is outstanding at producing accurate predictions and capable of handling distinct types of information, like categorical, Numeric and word.
* As our dataset is mix of different datasets this algorithm is apt for using to this dataset.
* It has the capacity to control and deal with missing values present in the data.
* It avoids overfitting by utilizing an integrated regularization parameter.
* It engages a gradient boosting algorithm to create its classification models.
* This algorithm can efficiently control and modify massive datasets, along with intricate interrelationships and interdependencies in the dataset.

**Logistic Regression:**

* Regardless of its name. This is not a regression algorithm it is a classification algorithm which can be used majorly in machine learning.
* This algorithm serves as an important instrument for forecasting student’s dropout and academic performance, offering perspectives that can notify focused measures, assist student success, and enhance educational outcomes.
* In the process of feature analysis, this algorithm coefficients can disclose the level and direction of the association between input features and the chance of academic dropout or performance.
* This is the best algorithm that can be used for serving elements preemptive alert systems, notifying educators to students suffering academically and required more assistance.

**KNN (K-Nearest neighbors):**

* It is a flexible and perceptive algorithm that can be utilized for different machine learning projects.
* As it is a supervised machine learning technique, it is used for both classification and regression analysis.
* In this algorithm K value is a hyperparameter which needs to be designated prior to training the model. It decides the count of neighbors considered while making the forecasts.
* This algorithm saves the whole training dataset and forecast at inference time.
* This model is more suitable for managing missing values, encoding categorical attributes, and scaling numerical features.
* KNN can be instructed on historical data including data about student’s dropout and those who endured in the circumstances of student’s dropout and academic success prediction.
* This algorithm is acquired to categorize new students depending on their resemblance to current dropout cases.

**Experimental Analysis:**

* The effectiveness of different models was assessed by implementing them to a dataset which covers the time from 2020 to 2023.
* Precision, recall and F1 scores are evaluated for their efficiency in handling the research targets.
* The algorithms that are being evaluated are XGBoost classifier and ensemble method merging the estimations of KNN and Logistic Regression.
* I will do hyperparameter tuning for all the algorithms to get the highest accuracy when compared to the previous projects.
* In general, the experimental analysis proves the supremacy of the ensemble theory on the individual models with respect to F1-score, precision and recall on the dataset.
* The statistical importance tests present supporting proof of the ensemble method’s strong output.
* After performing all these techniques KNN algorithm and Logistic Regression will get accuracy of above 92% and 94% which cross the previous projects accuracy.

**Research Work Plan:**

* In the very first week, I committed my time to analyzing studies that have employed student’s academic and dropout prediction data to assess the results of the previous research.
* In the next week, it is especially important to completely understand the dataset that I have chosen. Once that is done, the growth will progress to the advancement of machine learning techniques.
* The forthcoming week will mark the beginning of our attempts on progressing the accurate models that suit the dataset.
* Generate a timeline or Gantt chart that maps out the order of tasks and responsibilities to be achieved during the research project.
* Divide the project into benchmarks and demonstrate the approximate time for each task.
* The output can be evaluated in the coming two weeks.
* The procedure of scrutinizing the results and recognizing any essential developments to improve the accuracy of the model requires a week full of concentrated study.
* Finally, two weeks will remain where we can focus on completing the report, which will be followed by an extra week assigned for the objective of rechecking and submitting.

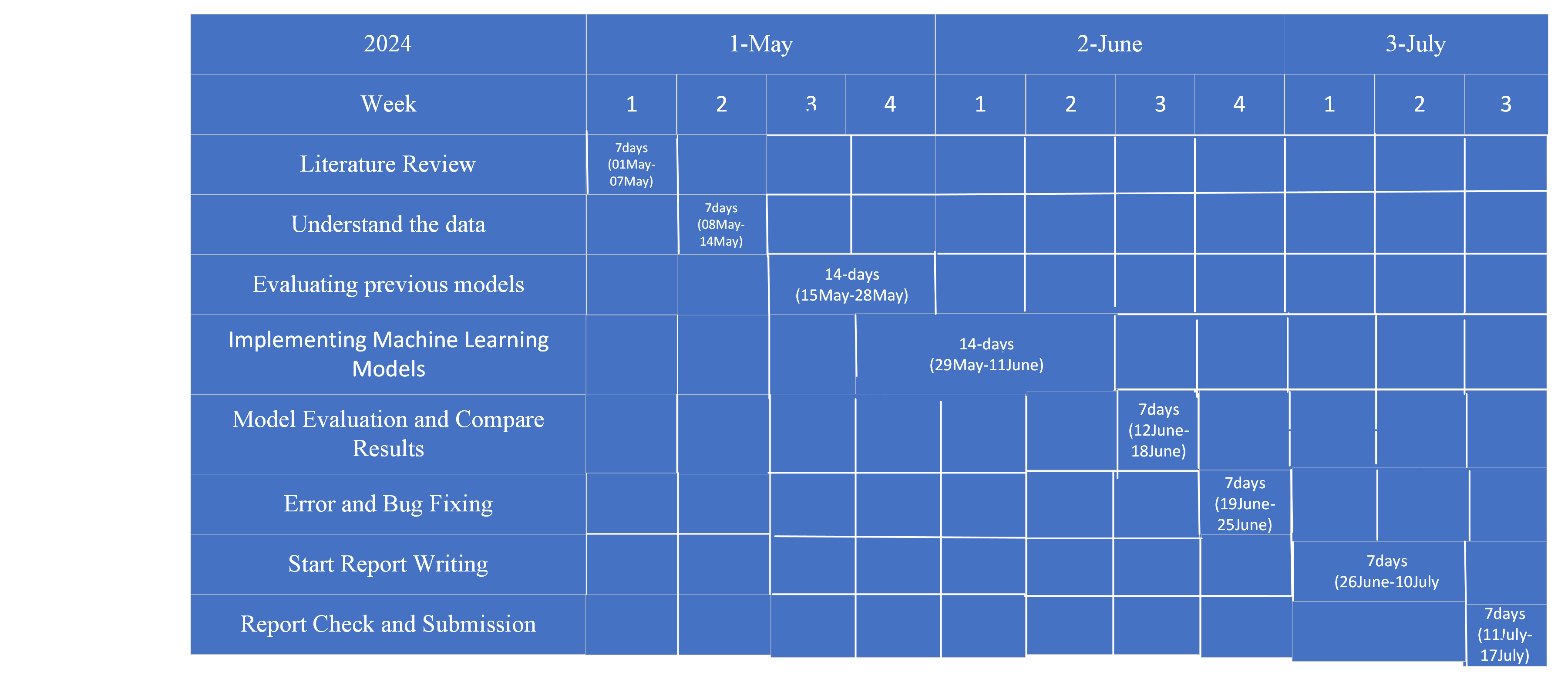


Fig 2. Project Work Plan

**Ethical Considerations:** This specifically talks about how the dataset or problem shows impact on legal and ethical issues generally, Given detailed information in the below table.

|  |  |
| --- | --- |
| **Data Privacy issues** | Student data adherence to privacy laws in some of the countries like (FERPA or GDPR) is very important. Handling student data in an improper way can result in several legal issues and violating of student privacy rights. |
| **Bias and Fairness** | Biased data can discourage some of the students unfairly. To avoid this, it is necessary to check the data before the system makes predictions and gives results in a fair way to everyone. |
| **Information Security** | As we use student data in this research, data security is very important. We must take tight security protocols, confidentiality, authorization measures to safeguard confidential data from cyber threats. |
| **Transparency and Explainability** | It is very important to gain trust in technology, it is possible only through transparency, it is very crucial how technology forecast the predictions and judgements to guarantee accountability, so giving clear explanation in project will help in the ethical issues. |
| **Informed agreement** | Generally, if any student data is utilized for project or research purposes, then the specific students must have one clear view of about their data being utilized, and right to confidentiality. |

Table 1. impact of legal and ethical issues content.

**Component 2 – Initial Literature Review:**

* The main intention of the literature review is to gain a deeper knowledge and understanding of the chosen dataset and discover various classification algorithms that can be utilized for forecasting student’s dropout and academic success.
* Yet, because the dropout prediction was not used in this study, it is not needed to execute a comprehensive literature review. Data evaluation will be implemented once the significant data has been collected from the literature review.
* As I have taken the dataset which presents serious concern about student dropout and academic success prediction regarding to enhance the goals achieving success rate of students.
* **How these Research papers are related to my research problem:**
* In one of the papers, they utilized some of the machine learning techniques for early dropout detection in universities they used innovative stacking ensemble method for predicting the dropout rate which gave me an idea to use other techniques to enhance the prediction rate Prasanth, A., & Alqahtani, H. (2023, October).
* In another paper they predicted student dropout rate in online learning in which the online classes became a mandatory study in covid-19 period. So, they utilized entry log record in LMS and detected the dropout rate. So in my research based on this paper I can see the different methods to detect the online learning dropouts as I came to know there are more dropouts through this online learning so this data will be very useful to my research in further process Dewi, M. A., Kurnaidi, F. I., Murad, D. F., Rabiha, S. G., & Romli, A. (2023, November).
* Some of the researchers even used Natural Language Processing (NLP) for this problem and in this they divided the predictions into two types MOOCS (Massive open online courses) and offline education. And they used CNN and GCN for predicting, As in my research I am predicting both offline and online education this paper helps me a lot for student information and came to know techniques about cleaning, imputation on reliable structured datasets in a clear view Anh, B. N., Giang, N. H., Hai, N. Q., Minh, T. N., Son, N. T., & Chien, B. D. (2023, July).
* SMOTE (Synthetic Minority Oversampling Technique) is used in one of the papers in which 1243 student’s information have been accumulated and studied this technique really helped to some of the extent for dropout prediction actually collecting students data is the biggest task because of the ethical and legal issues, these research papers helps me a lot by collecting more students data. Revathy, M., Kamalakkannan, S., & Kavitha, P. (2022, January).
* One paper took the data of Chilean universities in which it has taken four cohorts of students (2012-2016) and used decision tree algorithm to predict the dropout rate in the first year of study. From this paper I got an idea to know the Sem wise retention rate by applying machine learning techniques like Logistic Regression. Bello, F. A., Kóhler, J., Hinrechsen, K., Araya, V., Hidalgo, L., & Jara, J. L. (2020, November).
* From this paper I came to know that there is more dropout rate in STEM subjects in the world, the researcher in this paper predicted the dropout rate in CS related degree programming, from this idea I wanted to know the dropout rate in master's level which are increasing day-by-day. Böttcher, A., Thurner, V., & Häfner, T. (2020, April).
* This paper talks about NEP-2020 in India where the quality education is necessary for the students, so this researcher applied many machine learning techniques to reduce the dropout rate by HEI. From this paper, I learned there are more dropout rates in engineering and management courses than in others. Sharma, N., Sharma, M., & Garg, U. (2023, January).
* **Ideas to Enhance the results of my paper that I would like to Implement.**
* As I mentioned above every paper is talking more about dropout rates but not about academic success, so I want to conduct many online surveys and know the reasons why students are not able to get success in academics.
* Based on those reasons I want to update the different attributes in the dataset and use machine learning techniques like LightGBT, Logistic regression and KNN for predicting both dropout rate and academic success.
* I will perform hyperparameter tuning for all the algorithms and get the best accuracy of predictions to enhance the results.
* I will research more about online learning and take effective measures like tracking the entry log record and by observing their interest in online teaching for preventing the dropout rate which is very high in online learning.
* I also have an idea by forecasting the dropout rate in semester wise especially in specific courses like engineering and management where the dropout level is very high especially in the master's level of education.
* As social media platforms have huge response these days, online surveys are necessary for every 6 months which helps better to predict the dropout and academic success rate and improve the results.

**Conclusion and Future Works:**

In this report as I applied the best Machine Learning techniques with enhance in the accuracy of 87% and 88% in Logistic Regression and KNN when compared to the previous researchers it accurately predicts the student dropout and academic success rates but in future I wanted to change the updated features like past grades, exam scores, attendance records, completion of assignments, participation in the class activities, study habits and access to educational resources and apply more Machine Learning techniques which enhance the ability to predict the student success and dropout rate which helps to achieve their goals in future and also can see the development in many educational institutions.

**References:**

[1]Prasanth, A., & Alqahtani, H. (2023, October). Predictive Modeling of Student Behavior for Early Dropout Detection in Universities using Machine Learning Techniques. In *2023 IEEE 8th International Conference on Engineering Technologies and Applied Sciences (ICETAS)* (pp. 1-5). IEEE.

[2] Dewi, M. A., Kurniadi, F. I., Murad, D. F., Rabiha, S. G., & Romli, A. (2023, November). Machine Learning Algorithms for Early Predicting Dropout Student Online Learning. In *2023 IEEE 9th International Conference on Computing, Engineering and Design (ICCED)* (pp. 1-4). IEEE.

[3] Anh, B. N., Giang, N. H., Hai, N. Q., Minh, T. N., Son, N. T., & Chien, B. D. (2023, July). An University Student Dropout Detector Based on Academic Data. In *2023 IEEE Symposium on Industrial Electronics & Applications (ISIEA)* (pp. 1-8). IEEE.

[4] Revathy, M., Kamalakkannan, S., & Kavitha, P. (2022, January). Machine Learning based Prediction of Dropout Students from the Education University using SMOTE. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1750-1758). IEEE.

[5] Bello, F. A., Kóhler, J., Hinrechsen, K., Araya, V., Hidalgo, L., & Jara, J. L. (2020, November). Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout. In *2020 39th International Conference of the Chilean Computer Science Society (SCCC)* (pp. 1-5). IEEE.

[6] Böttcher, A., Thurner, V., & Häfner, T. (2020, April). Applying data analysis to identify early indicators for potential risk of dropout in cs students. In *2020 IEEE Global Engineering Education Conference (EDUCON)* (pp. 827-836). IEEE.

[7] Sharma, N., Sharma, M., & Garg, U. (2023, January). Predicting academic performance of students using machine learning models. In *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)* (pp. 1058-1063). IEEE.