1. Introduction

- This project aims to analyze the impact of gender, race, parental level of education and test preparation to determine student's academic performance by making use of a dataset downloaded from Kaggle that has all the necessary features.
- The objective is to analyze how all these factors except test preparation influence the academic performance and test preparation status, providing insights for educational enhancements and to finally develop a ML algorithm to predict students' preparedness for new entries..

2. Literature Review

- The research analyzes the relationship between socio-econimic factors and students' academic performance. As a result of the study, after thorough investigation, the researchers identified that students facing social and economic and social challenges were more likely to experience academic difficulties. The research also emphasizes the need to target support mechanisms to address the address academic inequalities arising from socioeconomic disadvantages, [1]
- The study finds the impact of demographic and academic factors on predicting students' performance by taking into consideration factors such as gender, race, parental education, and performance in subjects like math, reading, and writing. The results encourage educational strategies and interventions to be tailored to address specific demographic and academic factors that impact students' academic performance. [2]
- The analysis revolves around finding the connection between how good the students are with technology and computers and their family's socioeconomic status. Steps like data collection by asking the students about their skills with technology and then using this data to find patterns or differences related to their skills and their socioeconomic status, took place. The goal of this analysis was to make sure that every student has the same opportunity to be good at using technology regardless of their background. [3]
- The research focuses on predicting students' academic achievements by considering school related and socio-demographic factors. The researchers collected data on aspects like student's socio-demographic background and type of schools they attended. By analyzing this data, a predictive algorithm was formed that predicted how well students would perform academically. [4]
- The study reveals a correlation between socio-economic status and academic performance by showcasing students with higher socioeconomic status exhibiting better academic performance. The study sheds light on the inequalities in academia and gives information that can be used to create rules in the future that can ensure fair and equal chance at education to everyone. [5]
- The analysis examines connection between socioeconomic status and academic performance to provide valuable insights into the educational landscape of the developing countries. The findings of this analysis may have implications for educational practices and policies aimed to promote students' academic success. [6]

3. Problem Formulation

- The goal of this project is to investigate the influences on students' academic performance by examining factors like gender, race, parental level of education and lunch providing a nuanced perspective on social and demographic conditions influencing students' performance, then to decide from these factors if the student is prepared for the test or not hence uncovering patterns that offer valuable insights for students, their parents and professors which further lead to fabrication of a ML algorithm that can predict if the students' are prepared or not for the test depending on the input collected from the user.
- This binary classification of test preparation allows a thorough examination of the factors contributing to students' preparedness for tests. It also showcases how each one of these factors plays a vital role in students' test preparation status.

4. **Data Description**
   - The dataset, downloaded from Kaggle, consists of 1,000 entries comprising information on student performance.
   - It includes attributes like: gender, race, parental level of education, lunch type and test scores of math, reading and writing. The data underwent preprocessing steps as follows, handling missing values (result: there were none), renaming columns (result: renamed for simplified understanding for example renamed race/ethnicity to race), addressing duplicate entries (result: deleted the duplicate entries), data standardization (result: renamed some college to college) and mapping categorical variables(result: mapped prepared to 1 and unprepared to 0). A 2 by 2 grid was used for illustrating confusion matrix outcomes, to make it more easily understandable. Finally, a user-friendly interface that collects required data from the user and gives predicted class as an output helps use this algorithm for future use and new entries.

5. **Methodology**
   - The project uses machine learning techniques like Naive Bayes, Logistic Regression, Random Forest and Decision Tree classifier.
   - The goal of this project was to classify if the students are prepared for the tests or not.
   - Thus, application of above mentioned classification methodologies is justified. It allowed a comprehensive evaluation of factors influencing students' academic performance.
   - The models are trained and evaluated using accuracy metrics to find out their effectiveness in predicting the binary classification of test preparation and according to the results, went forward with the one that showed the most accuracy to finally develop an ML predictive algorithm that dynamically takes input from the user on run time.

6. **Solution Implementation**
   - **Tools**
     - ❖ Jupyter Notebook: My project was executed on this platform.
     - ❖ Python: The code to carry out the data mining process for my project was written in python.
     - ❖ Pandas: I employed pandas to load my .csv file.
     - ❖ Scikit-learn: I made use of this library to test and train the dataset and implement classification models.
     - ❖ Seaborn/Matplotlibs: I included this library to show the result of the confusion matrix via heatmaps.

   - **Breakdown with code snippets**
     - ❖ Data Preprocessing
       - ➢ Missing values were checked for each attribute -> df.isnull().sum().
       - ➢ Columns were renamed for clarity -> df = df.rename(columns={'race/ethnicity': 'race', ...}).
       - ➢ Duplicate rows were identified and removed -> df.drop_duplicates().
     - ❖ Feature Engineering
       - ➢ Categorical variables were mapped to binary labels -> df['test_performance'] = df['test_performance'].map({'prepared': 1, 'unprepared': 0}).
       - ➢ Columns like 'some college' and 'none' were replaced for uniformity -> df['parent_education'] = df['parent_education'].replace('some college', 'college').
     - ❖ Data Splitting
       - ➢ The dataset was split into training and testing sets for model evaluation -> F_train, F_test, C_train, C_test = train_test_split(F_encoded, C, test_size=0.2, random_state=40).
     - ❖ Model Training and Evaluation

- ➢ Various classifiers (Naive Bayes, Logistic Regression, Random Forest, Decision Tree) were trained and evaluated for accuracy -> model_accuracies = train_and_evaluate_models(models, F_train, F_test, C_train, C_test).

❖ Confusion Matrix Visualization

- ➢ Confusion matrices were generated for each classifier to visualize model performance using heatmaps -> cm = confusion_matrix(C_test, C_pred), sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', cbar=False).

❖ Predictive Algorithm

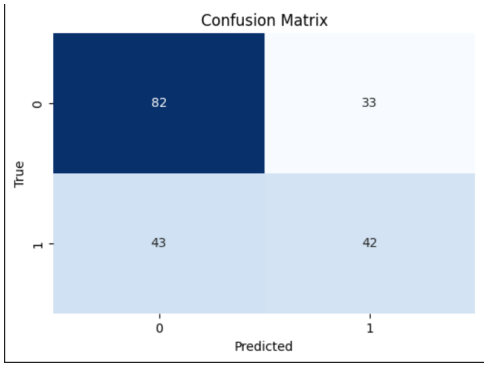
❖ User-friendly interface was created to take new entries from the user at runtime and classify that entry based on the ML predictive algorithm formed during training and testing of data -> predicted_class = logistic_regression_model.predict(new_entry_encoded), print("Predicted class:", predicted_class_label).

7. Experimental Setup

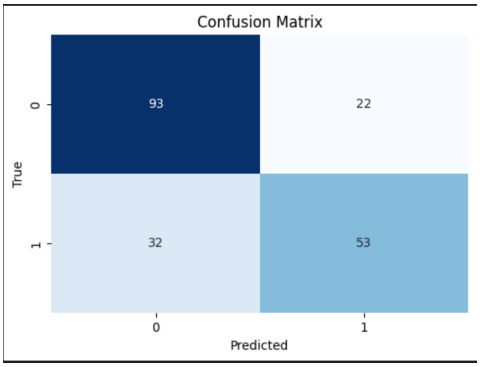
- The dataset was split using the train_test_split function from scikit-learn library.
- 80% of the data was used for training the models and the remaining 20% for testing.
- I reached this bifurcation after trying out various combinations of numbers of train and test set data and chose the one that gave the best results, that is showed the most accuracy amongst all models.
- Talking about accuracy, that is what I chose as the primary metric using accuracy_score function from scikit-learn library.
- I selected this metric because it showcased the model's ability to correctly classify input data into prepared and unprepared class labels which was exactly what I wanted to move forward with my project.
- Four classifiers (Naive Bayes, Logistic Regression, Random Forest and Decision Tree) were trained using the training set obtained in the previous step and tested using the test set.
- At last, the one with the best predictive performance was selected, which in this case was Logistic Regression with accuracy of 0.73.
- Using the train and test dataset formed by performing Logistic Regression Model a predictive algorithm was formed whose function is to predict the students' preparedness based on the data provided by the user on runtime.

8. Results

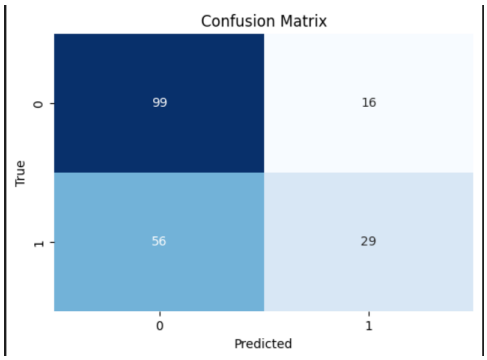
- Result consists of predictive accuracies of all the models, graphical representation of confusion matrix outcome for all and a predictive algorithm with a user-friendly interface. The results are as follows:
- Naive Bayes showed an accuracy of 62%, moderate predictive performance.



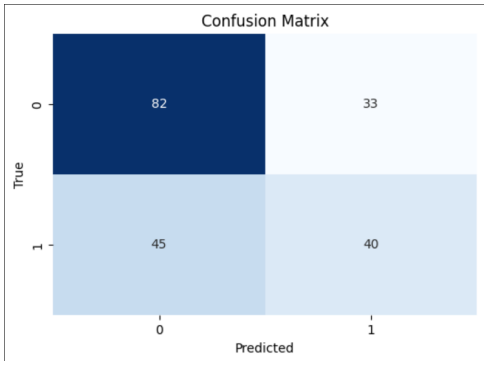
- Logistic Regression showed an accuracy of 73%, stronger predictive performance.



- Random Forest showed an accuracy of 65%, moderate predictive performance.



- Decision Tree showed an accuracy of 60%, lowest predictive performance.



- To summarize, Logistic Regression demonstrated the highest accuracy in predicting students' test preparation status, followed by Random Forest, Naive Bayes, and Decision Tree as we can see in the output of respective confusion matrices.
- Therefore, Logistic Regression Model was employed to develop a predictive algorithm that classified students to be prepared or unprepared depending on the input given by the user in the form of: gender, race, parental level of education, lunch type, math score, reading score and writing score.

```
Enter gender (male/female):  female
Enter race (group A/B/C/D/E):  group E
Enter parental level of education (college/master's degree/bachelor's degree/associ
ate's degree/high school):  associate's degree
Enter lunch type (standard/free):  free
Enter math score (0-100):  50
Enter reading score (0-100):  56
Enter writing score (0-100):  54
Predicted class: unprepared
```

9. Discussion

- Data Cleaning and Preprocessing
  - ❖ Checked for missing values in each column. It seems there are no missing values.
  - ❖ Renamed columns for better readability.
  - ❖ Handled some categorical data, such as replacing 'some college' with 'college' in the 'parental level of education' column and converting 'none' to 'unprepared' and 'completed' to 'prepared' in the 'test preparation course' column.

- Handling Duplicates
  - ❖ Checked for and removed duplicate rows.

- Feature Engineering
  - ❖ Went through all the combinations of features and selected the set which gave out the most accuracy of classification models, that were, gender, race, parental level of education, lunch, math score, reading score and writing score.
  - ❖ Mapped the 'test preparation' column to binary labels, where 'Prepared' is mapped to 1 and 'Unprepared' to 0.
  - ❖ Encoded categorical variables using one-hot encoding.

- Data Splitting
  - ❖ Split the data into training and testing sets. This is crucial to assess the model's performance on unseen data.

- Model Training and Evaluation
  - ❖ Used four different classifiers: Naive Bayes, Logistic Regression, Random Forest, and Decision Tree.
  - ❖ Trained each model on the training set and evaluated their accuracy on the test set.
  - ❖ Displayed the accuracy for each model. The Logistic Regression model performed the best among the models you tried.

- Confusion Matrix Visualization
  - ❖ Plotted confusion matrices for each classifier, providing insights into how well the models are performing in terms of true positives, true negatives, false positives, and false negatives.

- Insights
  - ❖ The accuracy results indicate how well each model is performing overall.
  - ❖ Analyze the confusion matrices to understand the types of errors the models are making.

- Predictive Algorithm
  - ❖ Predictive algorithm with a user-friendly interface that predicts if the students' are prepared or unprepared for the test based on the input from the user.
  - ❖ The ML algorithm was trained using the Logistic Regression Model since it showed the most accuracy.

- Alignment and Differences with Existing Literature
  - ❖ My project's findings are a mix of what other studies found and some new discoveries. Like previous research [1][2][3][4][5][6], my results show that a student's family background and certain personal details can affect how well they do in school. However, I also found some new things that others might have missed like how the type of lunch the student has had before the exam also affects students' academic performance. Differences in findings could be because of where and when the studies were done or the data they used. It's essential to know the project's limits, like the data I used, and future studies could look more into how things change over time or how technology impacts students. This helps us see where my project adds to what we know and suggests areas where we could learn even more.

10. Limitations

- Limitation: The analysis is based on a limited number of features that may oversimplify the complex factors influencing student performance and might result in underfitting.
- Reflection: While chosen features provide valuable insights, a more detailed dataset containing numerous more socio-economic and demographic factors could offer a deeper understanding.

- Limitation: The dataset misses the influence of changes over time, hence the insights derived from it now is valuable but maybe it won't be as useful in the future.
- Reflection: Incorporating temporal dimension to keep the dataset up to date to reveal evolving trends and patterns will help in more informed decision making.

11. Future Work

- Suggestion: Include additional socio-economic and demographic factors in the dataset.
- Relevance: Offers more detailed understanding.
- Feasibility: Numerous databases available on the internet that can be merged with the current one.

- Suggestion: Incorporate a temporal dimension.
- Relevance: Captures change over time and keep the analysis up to date and relevant to current time.
- Feasibility: Finding a database that has features like year semester analysis, monthly progress and seasonal influence or add these features in the dataset yourself after observing these features in students but that might take time.

12. Citations

[1] Considine, G., & Zappalà, G. (2002). The influence of social and economic disadvantage in the academic performance of school students in Australia. Journal of Sociology, 38(2), 129-148. https://doi.org/10.1177/144078302128756543.

[2] Bilal, M., Omar, M., Anwar, W. et al. The role of demographic and academic features in a student performance prediction. Sci Rep 12, 12508 (2022), DOI: https://doi.org/10.1038/s41598-022-15880-6.

[3] Albert D. Ritzhaupt, Feng Liu, Kara Dawson & Ann E. Barron (2013) Differences in Student Information and Communication Technology Literacy Based on Socio-Economic Status, Ethnicity, and

Gender, Journal of Research on Technology in Education, 45:4, 291-307, DOI: 10.1080/15391523.2013.10782607.
[4] Tamara Thiele, Alexander Singleton, Daniel Pope & Debbi Stanistreet (2016) Predicting students' academic performance based on school and socio-demographic characteristics, Studies in Higher Education, 41:8, 1424-1446, DOI: 10.1080/03075079.2014.974528.
[5] Carlos Felipe Rodríguez-Hernández, Eduardo Cascallar, Eva Kyndt, Socio-economic status and academic performance in higher education: A systematic review, Educational Research Review, Volume 29, 2020, 100305, ISSN 1747-938X, https://doi.org/10.1016/j.edurev.2019.100305.
[6] Erdem, C., & Kaya, M. (2023). Socioeconomic status and wellbeing as predictors of students' academic achievement: Evidence from a developing country. Journal of Psychologists and Counselors in Schools, 33(2), 202-220. doi:10.1017/jgc.2021.10.