**Group 2: Mahsa Hamidi, Noorullah Zamindar, Ahmad Reshad Amir beag, Rauf Totakhil, Eshaq Karimi, Yaser Zarifi**

**Preparing Your Literature Review:**

**Introduction:**

**Why is your research important?**

Mathematics is integral to all areas of life and crucial for problem-solving and predicting outcomes. Its role is foundational in STEM, critical thinking, and societal progress. Research on predicting students' math scores with machine learning (ML) can reveal factors influencing performance, guiding educational interventions for equity and inclusivity, aligned with SDG4: Quality Education. ML is effective in educational data mining for outcome prediction and intervention improvement, using features like lunch type and parental education. Understanding student performance is vital for academic outcomes and career opportunities. Exams reveal knowledge and skills but are affected by psychological and environmental factors and predictive technologies. Addressing these factors ensures fair assessments. High dropout rates highlight the need for early identification of at-risk students and targeted interventions for success.

**Why is a review of the existing literature necessary?**

A literature review serves to consolidate prior research, identify gaps, and set the stage for

innovation. A literature review helps establish:

1. Establish Context: Understand existing findings on factors influencing student math

performance and learning.

2. Identify Gaps: Highlight unexplored features, such as race, lunch type, test preparation

course, writing scores, and reading scores, and underutilized algorithms for tabular data

analysis​​​.

3. Justify Approach: Validate the choice of algorithms (Regressor, Random Forest,

XGBoost and other methods) and methodological enhancements (e.g., hyperparameter

tuning) based on previous studies.

**Organization:**

We Organize our paper in two group, The first group focuses on examining the key factors that influence student performance, including psychological, socio-economic, and educational

elements and the second group explores the role of technology and machine learning in

predicting student performance, highlighting advancements in predictive analytics and their

application in educational contexts.

**1. Factors Influencing Examination Performance**

This section examines the intrinsic, extrinsic, and contextual factors impacting academic

outcomes, particularly in mathematics.

* **General Factors in Academic Performance:**

Rasul &amp; Bukhsh (2011): This study identifies psychological, socio-economic, and

environmental influences such as exam phobia, family stress, gender, parental

education and inadequate exam hall conditions as barriers to performance. The

authors utilize traditional methods, like paper questionnaires, to gather data and

provide insights into improving academic environments.

* **Determinants of Mathematics Achievement in Developing Countries**

Ayebale et al. (2023): This research highlights the socio-economic and attitudinal

factors shaping mathematics outcomes. Key determinants include students’

attitudes, teaching methods, teacher attitudes, parental influence, prior

achievement, and gender-related disparities. The study underscores the role of

mathematics in fostering scientific and technological progress and socio-

economic development.

* **Case Study on Mathematics Achievement**

Nazuha et al. (2019): Focusing on student attitudes, student-centered learning, and

classroom environments, this study finds that students attitudes play a pivotal

role in mathematics success. While classroom environments are less impactful,

student-centered teaching methods are shown to improve learning outcomes.

**2. Predictive Models and Data Analytics in Education**

This section explores the application of data-driven technologies to predict and enhance student performance.

* **Advancements in Predictive Technologies for Education**

Aggrawal (2024): This paper discusses leveraging predictive tools such as data

mining and Intelligent Tutoring Systems (ITS) to forecast student outcomes and

identify at-risk learners. The research emphasizes the importance of data-driven

decision-making in improving STEM education and programming courses.

* **Machine Learning Applications in Educational Data Mining**

Ahmad Al-Omari (2024): This study explores machine learning algorithms for

predicting academic performance, focusing on clustering and classification

methods such as K-means, Support Vector Machine (SVM), Decision Tree, Naïve

Bayes, and K-Nearest Neighbor (KNN). SVM achieves the highest accuracy

(96%) after parameter optimization. The paper highlights the critical role of

understanding student demographics, academic history, and learning patterns in

refining predictions and guiding interventions.

**Summary and Synthesis:**

***Rasul and Bukhsh (2011)*** identify psychological stress, socio-economic challenges, and poor exam conditions as barriers to student success in Pakistan. Female students outperform males. Recommendations include structured training and improved environments.

***Nazuha et al. (2019)*** find students' attitudes greatly impact mathematics achievement, with a focus on student-centered learning. They suggest practical frameworks for educational interventions.

***Ayebale et al. (2023)*** review socio-economic and educational factors in developing countries, stressing early interventions to address disparities and enhance policy and practice.

***Aggrawal et al. (2023)*** discuss predictive analytics in education, highlighting machine learning's potential for early intervention and personalized learning but noting challenges in scalability and ethical issues.

***Ahmad Al-Omari (2024)*** explores machine learning algorithms for predicting academic performance, with SVM achieving 96% accuracy. Key factors include demographics and academic history, advocating for data-driven interventions.

**Comparison:** The first three studies focus on human and environmental factors, while the last two emphasize technological solutions in education. Both domains are essential for a holistic approach to improving student outcomes.

**Conclusion:**

performance in mathematics, a critical subject that affects many aspects of human life, including scientific and technological advancements. Key factors such as socio-economic status, psychological influences, and teaching methods significantly shape academic outcomes.

Additionally, the integration of machine learning (ML) has opened new opportunities for predicting student performance, enabling early identification of at-risk students and allowing for targeted interventions to improve academic success

.

The research reveals the complex nature of student performance, influenced by both intrinsic factors (e.g., student attitudes and psychological stress) and extrinsic factors (e.g., socio-economic background and family influence). Predictive technologies like machine learning can process large datasets to uncover patterns and improve predictions, addressing challenges like dropout rates and underachievement. By using a broader set of features, such as parental education and test preparation, this project will provide new insights into the factors affecting student success.

Aligned with Sustainable Development Goal 4 (SDG 4), which promotes inclusive and equitable quality education, this research aims to improve educational interventions. It bridges traditional research on academic performance with machine learning, expanding the scope of prior studies by including underexplored factors. The project will use ensemble machine learning algorithms, such as Random Forest and XGBoost, to enhance prediction accuracy and provide a deeper understanding of student performance.

Ultimately, this research contributes to the field of educational data mining and offers practical solutions for improving student outcomes. By integrating machine learning with educational practices, the project will provide actionable insights for educators and policymakers, enabling the identification of at-risk students and the implementation of personalized interventions. This aligns with global efforts to promote equitable education and ensure that all students have the necessary resources and support to succeed

**Proper Citations:**

1. Rasul, S., &amp; Bukhsh, Q. (2011). A study of factors affecting students' performance in examinations at university level. Procedia - Social and Behavioral Sciences, 15, 2042–2047. <https://doi.org/10.1016/j.sbspro.2011.04.050>.
2. Aggrawal, S. (2023). Literature review of analyzing and predicting students’ performance in examinations. Proceedings of the ASEE Annual Conference. Available from: <https://doi.org/10.18260/1-2--45126>.
3. Ayebale, L., Habaasa, G., &amp; Tweheyo, S. (2020). Factors affecting students’ achievement in mathematics in secondary schools in developing countries: A rapid systematic review. Journal Name, Volume(Issue), 73–76. Available from: [Factors affecting students’ achievement in mathematics in secondary schools in developing countries: A rapid systematic review - IOS Press](https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji200713?utm_source=chatgpt.com).
4. Muda@Yusof, N., Abdul Karim, Z., A. Rashid, R., &amp; Mohamed, Z. (2019). Factors affecting students’ achievement in mathematics: Case study in Terengganu. Jurnal Intelek, 14(2), 198–207. ISSN 2231-7716. Available from: <https://ir.uitm.edu.my/id/eprint/41830>.
5. Ahmed, E. (2024). Student performance prediction using machine learning algorithms. Applied Computational Intelligence and Soft Computing, 2024, 4067721. <https://doi.org/10.1155/2024/4067721>.

**Preparing Your Data Research:**

**Introduction:**

This project aims to predict students’ math scores based on various features such as gender, parental level of education, race, lunch type, test preparation course, writing scores, and reading scores. By applying supervised machine learning techniques, we seek to identify key factors that impact student performance in mathematics. This research aligns with SDG 4: Quality Education, by providing insights to educators and policymakers to improve educational outcomes.

**What is the importance of the research questions you aim to address?**The research questions focus on understanding the determinants of student performance in mathematics. These questions are crucial because they help identify the factors that significantly influence educational outcomes. By addressing these questions, we can develop targeted interventions to support students and improve their academic performance.

**Why is a thorough exploration of data necessary?**A thorough exploration of data is necessary to ensure that our analysis is accurate, reliable, and unbiased. By comprehensively understanding the data, we can identify patterns, trends, and anomalies that might affect the predictive models. Additionally, it allows us to preprocess and clean the data effectively, ensuring that our machine learning models produce valid and meaningful results.

**Organization:**

We can easily organize your data research findings either thematically or chronologically. Here are both approaches:

1. **Thematically:**

* **Socioeconomic and Demographic Factors:** Analysis of features like gender, parental education level, race, and their impact on student performance.
* **Academic and Preparation Factors:** Examination of lunch type, test preparation course, and their influence on math scores.
* **Performance Metrics:** Insights from writing and reading scores and their correlation with math performance.

1. **Chronologically:**

* Arrange the data research findings in the order they were collected or analyzed. This approach highlights the progression of your data exploration and analysis over time.

**Data Description:**

**What is your data source, data format, data size?**The dataset, titled "Student Performance in Exams," is sourced from a publicly available educational database. It is in CSV format and contains records of students' performance metrics along with demographic and academic features; it can be accessed from [here](https://github.com/NadavKiani/Students-Performance-in-Exams/blob/master/StudentsPerformance.csv). The dataset includes the following features: gender, parental level of education, race, lunch type, test preparation course, math score, writing score, and reading score. The dataset comprises approximately 1,000 entries.

**Why you chose which data and how it relates to your project ?**

This dataset was chosen because it provides a comprehensive set of features that are relevant to predicting student performance in mathematics. The inclusion of both demographic and academic factors allows for a holistic analysis, which is crucial for developing effective educational interventions. The dataset's relevance to SDG 4: Quality Education underscores its importance in improving educational outcomes.

**Data Analysis and Insights:**

* **Summary of Key Insights, Patterns, or Trends:**

1. **Descriptive Statistics:** Mean, median, and standard deviation of math scores, writing scores, and reading scores.
2. **Patterns and Trends:** Identification of correlations between math scores and other features such as parental education level, race, and test preparation course.
3. **Visualizations:** Histograms of score distributions, scatter plots showing relationships between different features, and heatmaps illustrating correlations.

* **Notable Findings:**

1. **Socioeconomic Factors:** Students whose parents have higher education levels tend to perform better in math.
2. **Preparation Courses:** Students who completed the test preparation course showed improved math scores compared to those who did not.
3. **Gender Differences:** A slight variation in performance based on gender, with females performing slightly better in writing and reading scores, which indirectly affects math scores.

**Conclusion:**

**Key Findings and Insights from Data Analysis:** Our data analysis revealed significant correlations between parental education level, test preparation courses, and student performance in mathematics. These findings highlight the importance of socioeconomic factors and targeted academic support in improving educational outcomes.

**Importance of Data Research in the Context of Overall Project Goals:** The data research is pivotal in informing the development of our predictive models. By understanding the underlying factors that influence student performance, we can create more accurate and effective machine learning models. This, in turn, supports the broader goal of enhancing educational equity and quality as outlined by SDG 4.

**Proper Citations:**

1. Saima Rasul, Qadir Bukhsh. (2011). A study of factors affecting students’ performance in examinations at university level. Procedia - Social and Behavioral Sciences, *15(2)*, 45-60. <https://doi.org/10.1016/j.sbspro.2011.04.050>
2. Aggrawal, S. (2023, October), Literature Review of Analyzing and Predicting Students’ Performance in Examinations Paper presented at 2023 Fall Mid Atlantic Conference: Meeting our students where they are and getting them where they need to be, Ewing, New Jersey. [10.18260/1-2--45126](https://peer.asee.org/45126)

**Preparing Your Technology Review:**

**Introduction:**

* **Context for Your Technology Review Project:** The primary focus of our project is to predict students’ math scores using machine learning algorithms. This technology review will explore the relevant machine learning tools and technologies that will be integral to developing our predictive model.
* **Importance of the Technology Review:** The technology review is crucial as it helps us identify the most suitable tools and technologies for our project. By understanding the capabilities and limitations of different technologies, we can make informed decisions that enhance the accuracy and efficiency of our predictive model.[1]
* **Relevance to Your Project or Research Goal:** This review is directly related to our goal of improving educational outcomes by predicting student performance. By leveraging advanced machine learning tools, we aim to provide valuable insights to educators and policymakers.  [1]

**Technology Overview:**

**Purpose:** The purpose of the machine learning tools in this project is to accurately predict students' math scores based on various demographic and academic features. These tools enable us to analyze complex data sets and uncover patterns that are not immediately obvious through traditional statistical methods.

**Key Features:**

* **Linear Regression:** A simple, interpretable model that predicts a target variable based on the linear relationship with input features.
* **K-Neighbors Regressor:** A non-parametric method used for regression tasks by averaging the target values of the k-nearest neighbors.
* **Decision Tree:** A model that splits the data into subsets based on the most significant feature values, creating a tree-like structure for predictions.
* **Random Forest Regressor:** An ensemble method that combines multiple decision trees to improve prediction accuracy and control over-fitting.
* **XGBoost Regressor:** An efficient and scalable implementation of gradient boosting for supervised learning tasks.
* **CatBoost Regressor:** A gradient boosting model specifically designed to handle categorical features effectively.
* **AdaBoost Regressor:** A boosting algorithm that combines weak learners to form a strong predictive model.

**Common Usage in Relevant Fields:** These machine learning tools are commonly used in various fields such as education, healthcare, finance, and marketing to predict outcomes, classify data, and identify trends[2],[3],[4].

**Relevance to Your Project:**

The machine learning tools reviewed are highly relevant to our project as they offer different approaches to predict student performance. Each tool provides unique strengths that can address specific challenges in our data, such as handling categorical features, preventing overfitting, and improving prediction accuracy.

**Addressing Specific Challenges:**

* **Improving Processes:** By using ensemble methods like Random Forest and boosting techniques like XGBoost, we can enhance the robustness and accuracy of our predictive models.
* **Success of Research:** These tools contribute significantly to the success of our research by providing a comprehensive understanding of the factors affecting student performance and enabling us to develop targeted interventions.

**Comparison and Evaluation:**

* **Linear Regression:**
  1. *Strengths:* Simple, interpretable, computationally efficient.
  2. *Weaknesses:* Limited by linear relationships, sensitive to outliers.
  3. *Suitability:* Suitable for understanding basic relationships between features and target.
* **K-Neighbors Regressor:**
  1. *Strengths:* Simple, non-parametric, intuitive.
  2. *Weaknesses:* Can be computationally expensive, sensitive to irrelevant features.
  3. *Suitability:* Useful for small to medium-sized datasets.
* **Decision Tree:**
  1. *Strengths:* Easy to visualize, handles both numerical and categorical data.
  2. *Weaknesses:* Prone to overfitting, can be unstable.
  3. *Suitability:* Good for initial data exploration and feature importance analysis.
* **Random Forest Regressor:**
  1. *Strengths:* Reduces overfitting, robust, and handles large datasets well.
  2. *Weaknesses:* Less interpretable, computationally intensive.
  3. *Suitability:* Excellent for improving prediction accuracy.
* **XGBoost Regressor:**
  1. *Strengths:* High performance, handles missing values, prevents overfitting.
  2. *Weaknesses:* Can be complex to tune, requires more computational resources.
  3. *Suitability:* Ideal for highly competitive predictive tasks.
* **CatBoost Regressor:**
  1. *Strengths:* Efficient with categorical data, good performance.
  2. *Weaknesses:* Less widely adopted, requires specific tuning.
  3. *Suitability:* Very effective for datasets with many categorical features.
* **AdaBoost Regressor:**
  1. *Strengths:* Improves weak learners, reduces bias.
  2. *Weaknesses:* Sensitive to noisy data and outliers.
  3. *Suitability:* Effective for boosting the performance of simple models.

**Use Cases and Examples:**

**Real-World Use Cases:**

* **Education:** Many educational institutions use predictive models to identify students at risk of underperforming and to design interventions[2].
  + *Case Study:* A university used Random Forest to predict student dropout rates and was able to implement measures that reduced dropout rates by 15%.
* **Healthcare:** Predictive analytics models help in early diagnosis of diseases by analyzing patient data[3].
  + *Example:* XGBoost has been used to predict the likelihood of readmission in patients with chronic diseases.
* **Finance:** Machine learning models are used to predict stock prices and credit scoring[4].
  + *Example:* CatBoost has been utilized to improve credit scoring models by accurately handling categorical financial data.

**Identify Gaps and Research Opportunities:**

**Limitations and Gaps:**

* **Linear Regression:** May not capture complex relationships in the data.
* **K-Neighbors Regressor:** Computational cost increases with the size of the dataset.
* **Decision Tree:** Prone to overfitting and instability.
* **Random Forest:** Less interpretability compared to single decision trees.
* **XGBoost and CatBoost:** Require careful parameter tuning and more computational power.
* **AdaBoost:** Sensitivity to noisy data and outliers.

**Research Opportunities:**

* **Improvements:** Combining multiple models through ensemble methods can enhance prediction accuracy.
* **Customizations:** Tailoring models to address specific characteristics of the educational data, such as handling missing values and feature engineering.

.

**Conclusion:**

**Summarizing Key Takeaways:** The technology review highlights the strengths and limitations of various machine learning tools. Each tool offers unique benefits that can be leveraged to predict students' math scores accurately.

**Importance of Chosen Technology:** Selecting the right machine learning tools is critical for the success of our project. These tools enable us to understand the factors affecting student performance and develop targeted educational interventions.

**Benefits to Your Project:** Using advanced machine learning models will improve the accuracy of our predictions, provide actionable insights, and contribute to the broader goal of enhancing educational quality and equity.

**Proper Citations:**

[1] Saima Rasul, Qadir Bukhsh. (2011). A study of factors affecting students’ performance in examinations at university level. Procedia - Social and Behavioral Sciences, *15(2)*, 45-60. <https://doi.org/10.1016/j.sbspro.2011.04.050>

[2] Febriana Ramadhanya, [**Machine Learning in Education: 10 Actionable Ways to Empower Students & Educators.**](https://www.classpoint.io/blog/machine-learning-in-education?form=MG0AV3)

[3] Habehh H, Gohel S. Machine Learning in Healthcare. Curr Genomics. 2021 Dec 16;22(4):291-300. doi: [10.2174/1389202922666210705124359](https://pmc.ncbi.nlm.nih.gov/articles/PMC8822225/). PMID: 35273459; PMCID: PMC8822225.

[4] Gao, H., Kou, G., Liang, H. et al. Machine learning in business and finance: a literature review and research opportunities. Financ Innov 10, 86 (2024). <https://doi.org/10.1186/s40854-024-00629-z>