**Final Report: Predicting Student Performance Using Data Mining**

**Team Members**

1. Jasmine Liamsuwan
2. Shivam Kanungo
3. Kweku Darko

**Project Summary**

Understanding what drives students to success has always been a challenge for educators, and this project takes a data-driven approach to shed light on this question. By analyzing a dataset of 6,607 student records, we built and evaluated four machine learning models which include Decision Tree, Random Forest, and Neural Network (Multilayer Perceptron), to predict whether a student would pass or fail based on behavioral and environmental factors.

Among the models, the Random Forest emerged as the most accurate, achieving a validation accuracy of 91.09%. Its ability to capture intricate relationships between variables makes it ideal for high-stakes decisions. However, the Decision Tree model offered something unique like simplicity and interpretability, allowing educators to easily pinpoint the key factors influencing outcomes, such as attendance and hours studied. These findings provide actionable insights to help schools and teachers identify and support struggling students before it is too late.

**Background**

Student performance is shaped by various factors, including how much time they dedicate to studying, their attendance, and the support they receive from their parents. Schools often struggle to identify students who need help before they fail. This project seeks to address that challenge using data-driven techniques.

The dataset includes variables such as study habits, attendance, and sleep hours, offering a broad perspective on what impacts academic success. By analyzing these factors and predicting performance we aim to provide solutions and ways to help students from failing classes or school.

**Business Objectives**

The project had several goals. First, we aimed to identify the key factors that influence student performance. Knowing what matters most, whether it is attendance, study habits, or family support that can help teachers focus their effort effectively. Second, we wanted to create predictive models that classify students into pass or fail categories, enabling early interventions. Third, by interpreting these models, we aimed to provide actionable insights for improving teaching strategies and supporting students. Finally, the models and methodologies were designed to be scalable so they could be applied to other schools and education systems in the future.

**Process Followed for Selecting and Gathering Data**

**Data Overview**

The dataset consisted of 6,607 records and captured a variety of factors that influence academic performance. Key variables included:

* Exam\_Score: The target variable, categorized as “pass” (score > 60) or “fail” (score < 60).
* Hours\_Studied: Weekly study hours.
* Attendance: Class attendance percentage.
* Parental\_Involvement: Level of family support (Low, Medium, High).
* Sleep\_Hours: Average nightly sleep hours.
* Access\_to\_Resources: Availability of educational resources (Low, Medium, High).

To ensure the data was clean and reliable for analysis, the following steps were taken:

* Missing values in variables like Sleep\_Hours (7%) and Attendance (5%) were filled using mean imputation.
* Outliers were capped to reasonable ranges, such as limiting Hours\_Studied to a maximum of 40 hours per week.
* Categorical variables like Parental\_Involvement were encoded, while numerical variables were standardized.
* The dataset was split into a 70% training set and a 30% validation set to evaluate model performance.

**Initial Exploration**

Before diving into model building, it was essential to understand the dataset and its intricacies. This phase, often overlooked, is critical because it sets the foundation for robust analysis and reliable predictions. During the initial exploration, we examined each variable to identify potential issues, patterns, and trends that could affect model performance. Here’s what we uncovered:

**Missing Data**

One of the first steps in the exploration phase was to assess whether the dataset had missing values. Two key variables, Sleep\_Hours and Attendance, had small gaps in their data:

* Sleep\_Hours were missing for about 7% of the students.
* Attendance had 5% missing values.

To address this, we used mean imputation, replacing the missing values with the average values for these variables. This approach ensured that we preserved as much data as possible without introducing bias or distorting the overall distribution.

**Outliers**

Outliers can distort model performance, so we examined each numerical variable for extreme values. Two variables stood out:

* **Hours\_Studied:** While most students reported studying between 5 and 30 hours per week, a few outliers claimed to study over 70 hours weekly. These values were unrealistic and likely errors, so we capped this variable at 40 hours per week.
* **Attendance:** A handful of students had attendance percentages above 100%, which clearly indicated data entry errors. We corrected these values by capping attendance at 100%.

By handling these outliers, we ensured the dataset represented realistic student behaviors without the noise introduced by extreme values.

**Patterns in Study Hours**

One of the most striking findings was the relationship between study hours and student performance. When we analyzed the distribution of study hours:

* Students studying fewer than 5 hours per week had a failure rate of over 70%.
* On the other hand, those studying more than 10 hours per week were significantly more likely to pass their exams.

This trend highlighted the critical role that consistent study habits play in academic success. It also suggested that study hours would likely be one of the most influential predictors in our models.

**Attendance Trends**

Attendance proved to be another critical variable. During exploration, we noticed a strong correlation between attendance rates and exam outcomes:

* Students with attendance below 60% were much more likely to fail their exams. In fact, this group had one of the highest failure rates across the dataset.
* Conversely, students with attendance rates above 80% consistently performed well, with a high likelihood of passing their exams.

This insight underscored the importance of classroom engagement and suggested that attendance would be a key focus in any intervention programs designed to improve academic outcomes.

**Parental Involvement and Resources**

We also explored the role of parental involvement and access to educational resources:

* Students with high levels of parental involvement performed better overall. However, while this factor was important, it was secondary to variables like attendance and study hours.
* Similarly, access to resources such as study materials and internet connectivity positively influenced performance but didn’t outweigh the impact of attendance or study habits.

**Data Visualization**

To better understand these patterns, we created visualizations like histograms and scatterplots during this phase. For example:

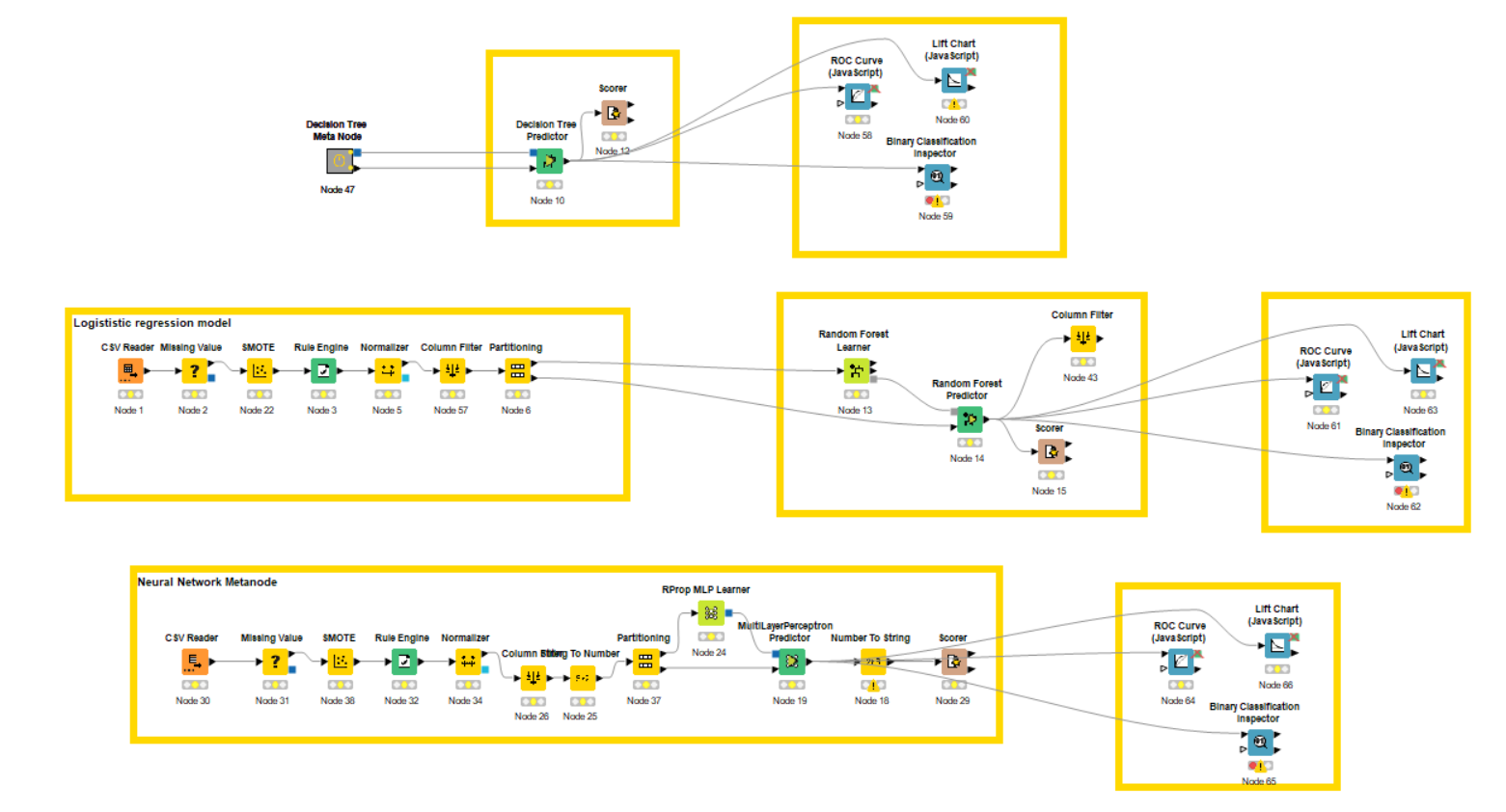
* A scatterplot of Hours\_Studied vs. Exam\_Score revealed a clear upward trend, reinforcing the importance of regular studying.
* Boxplots of Attendance grouped by passing or failing outcomes illustrated the stark contrast between students with high and low attendance rates.

**Key Takeaways from Initial Exploration**

The exploration phase revealed several critical insights:

1. Attendance and Study Hours are the most influential variables and likely to play a significant role in model predictions.
2. Addressing missing data and outliers ensured the dataset was clean and reliable for analysis.
3. Parental involvement and access to resources, while beneficial, are secondary predictors compared to attendance and study hours.

This thorough understanding of the data gave us the confidence to move forward with preprocessing and model building, knowing that we had accounted for potential issues and identified the variables most likely to impact performance. Let me know if you'd like additional details on specific aspects.

**Results and Analysis**

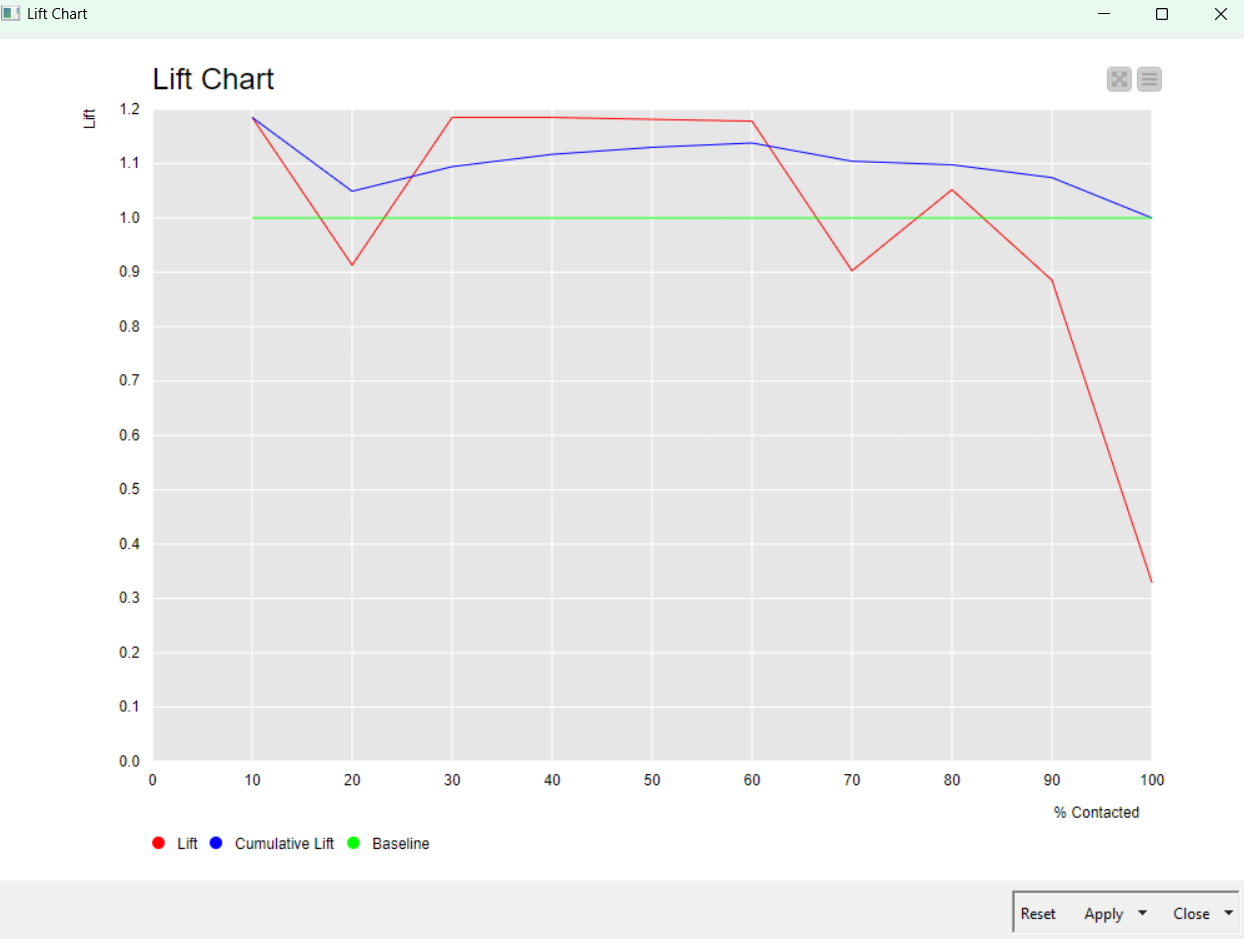
In this project, we explored how three machine learning models which include Decision Tree, Random Forest, and Neural Network to perform in predicting student success. Each model offered unique insights, and while they varied in complexity and interpretability, all provided valuable tools for identifying key factors influencing academic performance. Here’s a breakdown of their results and what they mean for real-world applications.

**Decision Tree Model**

The Decision Tree model demonstrated a strong validation accuracy of 90.82%, making it both reliable and easy to interpret. One of its most significant advantages is its simplicity, which allows educators to understand how decisions are made and to act on the insights it provides.

* **Key Insights**:  
  The Decision Tree highlighted Attendance and Study Hours as the most important predictors of success. Students with attendance rates above 80% were very likely to pass, while those below 60% were at a much higher risk of failing. Study habits also stood out because students studying fewer than 5 hours per week were overwhelmingly in the failing group, while those studying more than 10 hours per week performed significantly better.
* A graph of a bar chart

  Description automatically generated with medium confidence**Performance**:  
  Out of 10,247 students, the Decision Tree correctly classified 9,306, as shown in the confusion matrix (Appendix 1). It did, however, misclassify 941 cases, which highlights some limitations in edge cases. The lift chart (Appendix 2) shows how well the model predicted at-risk students, consistently outperforming random guesses.

**(Appendix 1)**

**(Appendix 2)**

* **Visualization**:  
  The Decision Tree model’s simplicity makes it perfect for educators who need clear explanations of why a student might be struggling. It’s easy to follow and ideal for designing targeted interventions.

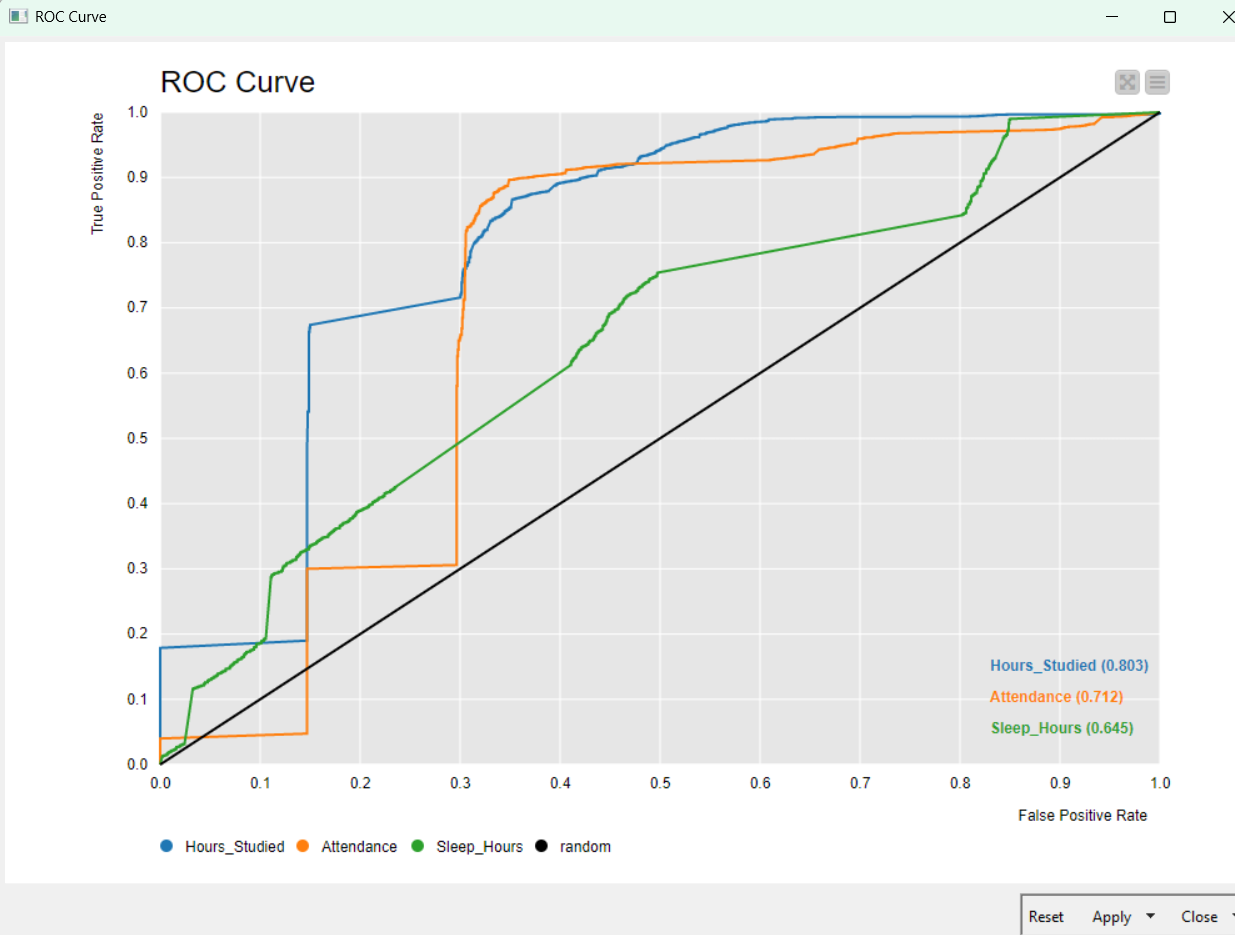
**Random Forest Model**

The Random Forest model emerged as the most accurate, achieving a validation accuracy of 91.09%. By combining multiple decision trees, this model enhanced predictive power and reduced the risk of overfitting, making it ideal for deployment in real-world scenarios.

* **Key Insights**:  
  Like the Decision Tree, the Random Forest model highlighted Attendance and Study Hours as the most critical predictors. The ensemble approach allowed it to capture more complex relationships between variables, providing a more robust and comprehensive view of student performance.
* A graph with red and blue lines

  Description automatically generated**Performance**:  
  The lift chart (Appendix 3) shows that the Random Forest consistently outperformed baseline predictions, particularly in identifying students who were likely to fail. The ROC curve (Appendix 4) highlights its strong classification ability, with an AUC of 0.804 for Hours Studied and 0.713 for Attendance. These scores indicate the model’s high reliability in distinguishing between passing and failing students.

**(Appendix 3)**



**(Appendix 4)**

**Neural Network Model**

The Neural Network model achieved the same validation accuracy as the Random Forest model, 91.09**,** but its complexity and computational intensity make it less practical for everyday use in educational settings.

* **Key Insights**:  
  The Neural Network captured intricate patterns in the data that other models might have missed, contributing to its high accuracy. However, its decision-making process is opaque, making it difficult to understand or explain the reasoning behind its predictions.
* A graph and diagram of a graph

  Description automatically generated with medium confidence**Performance**:  
  The confusion matrix (Appendix 5) shows that the Neural Network correctly classified 9,334 students, slightly outperforming the Decision Tree. Its lift chart (Appendix 6) and ROC curve (Appendix 7) also demonstrate its effectiveness in predicting success and failure, with AUC values like the Random Forest.

**(Appendix 5)**

**A graph with lines and numbers

Description automatically generated**

**(Appendix 6)**

**A graph showing a line graph

Description automatically generated with medium confidence**

**(Appendix 7)**

**Comparative Accuracy Analysis**

The table below summarizes the accuracy of each model, highlighting their strengths and trade-offs:

| **Model** | **Validation Accuracy** | **Key Strength** |
| --- | --- | --- |
| Decision Tree | 90.82% | Simple and interpretable, ideal for educators. |
| Random Forest | 91.09% | High accuracy and robustness, suitable for deployment. |
| Neural Network | 91.09% | Accurate but less interpretable; better for research. |

**Key Takeaways**

Each model offers something valuable:

* The **Random Forest** is the most accurate and reliable, making it the best choice for large-scale applications.
* The **Decision Tree** provides unmatched simplicity and interpretability, making it ideal for educators who want clear insights to guide student support.
* The **Neural Network**, while equally accurate, is less practical due to its complexity and lack of transparency.

Ultimately, these models highlight the importance of Attendance and Study Hours as the key drivers of academic success. By focusing on these factors and leveraging these tools, schools can take proactive, data-driven steps to support at-risk students and help them achieve their potential.

**Conclusions and Recommendations**

This project has shown the incredible potential of machine learning in predicting student success, with Attendance and Study Hours standing out as the most critical factors influencing performance. These findings underscore what many educators have observed anecdotally, students who show up consistently and dedicate time to studying are more likely to succeed. The three models we each used brought something unique to the table. The Random Forest model, with its 91.09% accuracy, proved to be the most robust and reliable for large-scale applications, offering a dependable tool for identifying students at risk. The Decision Tree model, while slightly less accurate at 90.82%, provided clear, interpretable insights that educators can easily act upon, making it a perfect fit for intervention planning. Meanwhile, the Neural Network model, though equally accurate to Random Forest, lacked the transparency required for practical use in educational settings where explainability is crucial.

To turn these findings into action, schools should focus on improving attendance and supporting effective study habits. Attendance tracking systems can help identify students who are falling behind, enabling schools to intervene early with personalized support, whether that’s counseling, tutoring, or mentorship programs. Encouraging better study routines through workshops or peer mentorship can also make a big difference. The Random Forest model is ideal for larger systems where accuracy and reliability are paramount, while the Decision Tree model is better suited for educators looking for transparent, actionable insights to guide their strategies. However, predictive models alone aren’t enough which means schools must pair these tools with thoughtful, timely interventions to ensure students get the help they need.

Ultimately, this project highlights the importance of combining data-driven approaches with a human touch. By focusing on attendance and study habits, and leveraging the right predictive tools, schools can proactively support their students and help them build a foundation for long-term success.

Ultimately, this project isn’t just about algorithms or accuracy scores because it’s about creating opportunities for students to thrive. By focusing on these recommendations, educators can turn data into meaningful action, helping students not only pass their exams but build a foundation for long-term success.