DATA-DRIVEN SOLUTIONS FOR ACADEMIC EXCELLENCE IN AN HYPOTHETICAL SECONDARY SCHOOL

Building a Predictive System to Enhance Student Performance in Final Exams

by

KENNETH ESSIEN, CHIDERA IGBOEJESI, AYOMIDE ADERONMU

9/10/2024

Abstract

Following the recent drop-in pass rate of secondary school students in their final exams, boards and ministry of education in Nigeria has undertaken some actions to curb this trend. The aim of this study therefore, was to identify the various factors that affect secondary school students. This research was carried out using synthetic data, due to the lack of readily available public data on students. However, various factors influenced our decision-making, such as interviewing students, teachers and parents. This gave us knowledge about the challenges the both students and school are faced with. Careful analysis, visualization and modelling was done and finally results and recommendations were given. This study makes use of tools like jupyter notebook, power bi, big query, streamlit. At the end, the school was provided with a resources for data collection, pipelining and warehousing solution.

Keywords: Synthetic Data, Data Analysis, Modelling, Jupyter Notebook, Power BI, BigQuery, Streamlit, Data Collection, Data Warehousing.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Education is considered imperative for not only the progress of the individuals, but also for the development of community and nation Kapur (2018). In recent years, Nigeria has seen troubling trends in the academic performance of senior secondary students. The Joint Admissions and Matriculation Board (JAMB) reported that, in the 2024 Unified Tertiary Matriculation Examination (UTME), 76% of students scored below 200 out of 400, which is less than 50%. Similarly, in the West African Senior School Certificate Examination (WASSCE), less than 50% of students achieved the necessary credits. These poor outcomes have sparked concern across the education sector. Meanwhile, some actions have been taken by CSOs, governments and other bodies to improve the quality of education, and by extension, students' performance. In order to ensure such does not arise in the upcoming standardize exams and to continuously improve the students' performance we have picked a school in Lagos to carry out research backed by data to develop our solution. The aim is to build a data driven solution to improve students'

exam performance. By leveraging data collected from students, teachers, and parents, we seek to understand the root causes of under performance and address them proactively.

1.2 Objective

The main objective of this study was to build a solution that leverages data to improve candidates performance in not just JAMB but all their upcoming final exams. It is about solving the existential problem of massive failure and using data to ensure students have great performance.

1.3 Sub objectives

i Identify possible problems students might be facing while writing these exams.

ii Generate data through any means based on the needs of your solution while ensuring that your data adequately reflects the state of the African education ecosystem.

iii Design an enterprise data solution for the school's data collection, pipelining, warehousing, automation, and reporting needs.

iv Create an optimized model that predicts the likelihood of a student passing or failing their upcoming exam based on their academic history.

v Make relevant recommendations to the stakeholders on how they can help improve the performance of the students based on your solution.

1.4 Research Questions

Asking questions is one of the first steps to analysis and problem solving in general. It's through attempting to answer these questions that we gain insights. The research questions for this study are:

i What are the common challenges faced by students in preparing for high-stakes exams like JAMB and WASSCE?

ii What types of data are most crucial for understanding and improving student performance in the African education ecosystem?

iii How can data be collected effectively from schools with limited technological infrastructure or no existing data management systems?

iv What data warehousing solutions are best suited for the school, considering scalability, cost-effectiveness, and ease of implementation?

v How can we model student academic performance prediction based on significant factors?

Chapter 2

Methodology

2.1 Research design

Table 2.1 shows an overview of the project.

2.2 Factors Selection

A systematic review of previous studies on predicting student academic performance prediction models, insights gotten from students, teachers has been used to identify the factors used in this study. The factors include student's demographic factors such as gender, academic performance, age Nambuya (2024), social factors and Parental factors such as education level, income levelSibomana et al. (2024) and school infrastructural factorsOmodan and Ekundayo (2024). Obadiah and Kelvin OmienoMusau et al. (2024) noted that academic performances is not always reliant on students' own effort but other factors that have significant influence over their academic performance.

Chapter	Description
Chapter One	This section provides the background, including statistics on JAMB and WAEC performance, and highlights the challenges students face in these exams. It explains the importance of using data-driven solutions to improve performance and how this study differs from past solutions. The main and sub-objectives outline the specific goals, including building a data infrastructure and predictive model.
Chapter Two	This section explains the research design and structure, detailing why synthetic data was used to model real-world scenarios. It covers variable selection, data generation techniques, and the structure of the database built for the school. It also describes feature engineering, sampling techniques, and how the data analysis and prediction model were developed and evaluated.
Chapter Three	This chapter describes the creation of a dashboard to visually track students' performance using various metrics. It emphasizes the automation of the dashboard, allowing the school to monitor real-time progress and intervene where necessary.
Chapter Four	The final solution is presented. This chapter outlines how the solution can be deployed for continuous student performance monitoring and prediction.
Chapter Five	The conclusion summarizes the findings and provides actionable recommendations for stakeholders to improve student performance. It includes steps on how to implement the solution effectively in the school's ecosystem.
Chapter Six	This section lists the academic and research sources cited throughout the study, supporting the methodologies, models, and techniques used.

Table 2.1: Chapter Descriptions for the Study

2.3 Data collection and Generation

During the initial 2-3 days of the hackathon, our team focused on researching and brainstorming approaches to acquire the necessary data for our project. We explored various sources, including academic papers, the West African Examinations Council (WAEC), the National Examinations Council (NECO), and the National Bureau of Statistics (NBS) websites, to identify relevant datasets. Although we were unable to find suitable publicly available data, insights gleaned from these resources, including scholarly papers and educational blogs, helped us determine the key data points required for the project.

2.3.1 Key Data Points to Collect

The data we planned to collect falls into four main categories:

- Student Data: Information about each student to identify patterns influencing their academic performance.
- Academic Data: Performance data, attendance records, and metrics on academic engagement.
- Socioeconomic Data: Information related to students' home environment and external factors impacting their studies.
- Resource Utilization Data: Insights into how students utilize learning resources at school and home.

Each category and the associated data points are outlined in more detail below.

2.3.2 Student Data

Student Bio-Data

- Name, Age, Gender, and Class Level.
- Contact Information (for follow-up and communication).
- Educational Background (e.g., primary school attended, history of class repetition).

Parental/Guardian Demographics:

- Parental Education Level and Occupation.
- Household Income Level (categorized as low, middle, high).
- Number of Siblings (to understand the division of parental attention and resources).
- Parental Involvement in Education (e.g., frequency of checking homework, attending PTA meetings).

Student Learning Behavior:

- Study Hours per Week (self-reported or observed by teachers).
- Access to Private Tutoring or Extra Lessons.
- Participation in Extracurricular Activities (sports, arts, clubs, etc.).

2.3.3 Academic Data

Historical Performance:

- Scores in past internal exams and external exams (e.g., WAEC, NECO results of past students).
- Subject-wise Performance Trends (to identify strong and weak subjects).
- Performance in Mock Exams and Weekly Tests.
- Class Participation Levels (categorized as active, passive, or indifferent).

Attendance Records:

- Attendance Rate (percentage of days attended).
- Patterns of Absence (unexcused absences, medical leaves).

Participation in School Programs:

- Involvement in remedial classes or academic support programs.
- Frequency and duration of participation in study groups or tutoring sessions.

2.3.4 Socioeconomic Data

Home Learning Environment:

- Access to Study Space (e.g., availability of a quiet study area, desk, and adequate lighting).

- Access to Learning Tools (e.g., textbooks, internet, computer/tablet).
- Frequency of Distractions at Home (e.g., household chores, sibling care).

Economic Background:

- Financial Ability to Purchase Study Materials.
- Access to Private Tutors or Educational Apps.
- Support for Educational Expenses (e.g., bursaries, government aid).

2.3.5 Resource Utilization Data

Library Usage and Resource Allocation:

- Frequency of Library Visits and Books Borrowed.
- Hours Spent in the Library per Week.
- Types of Resources Utilized (e.g., books, CBT software, learning videos).

Teacher Allocation and Hours Spent:

- Teacher Hours Spent per Student (for tutoring, mentoring).
- Teacher Subject Specialization (to identify gaps in subject-specific teaching).

Digital and Technological Exposure:

- Number of Hours Spent on Computer-Based Test Practice.

- Familiarity with CBT Platforms and Software.
- Internet Accessibility for Research and Online Learning.

2.3.6 Constraints

Given the limited availability of centralized, publicly accessible databases on student performance and demographic information, collecting data within the hackathon's time frame proved challenging. As a result, we opted to generate mock datasets while ensuring that the data reflected the reality of the African educational landscape as accurately as possible

2.3.7 Data Collection Instruments and Tools

In the absence of relevant publicly available data, we leveraged a few approaches to inform our mock dataset creation:

Focus Group Discussions (FGDs): We conducted brief discussions with secondary school students in our local communities and some of their parents (mostly neighbors). This provided valuable insights into the socio-economic conditions and parental involvement in students' education. During these discussions, we noted:

- **Resource Utilization:** Most Nigerian secondary schools lack structured tracking mechanisms for resource utilization, such as lab hours or library use. Students attend mandatory lab sessions based on a standardized timetable. Hence, resource utilization was excluded from our final dataset, as it would not add meaningful variance to our analysis.

- **Extracurricular Activities:** We found that schools typically allocate a uniform duration for extracurricular activities across the student body. Thus, tracking the exact hours spent on these activities was deemed redundant. Instead, we recorded students' participation in these activities as a binary variable (Yes/No) along with the type of activity engaged in.

In a real-life scenario, and given more time, we would have also used the following data collection methods:

Surveys and Questionnaires: In a real-life context, we would have developed separate surveys for students, teachers, and parents to collect both qualitative and quantitative data. The surveys would have been administered through:

- In-person meetings with school administrators, teachers, and students.
- Paper-based questionnaires for students and teachers.
- Digital forms (e.g., Google Forms) for convenience and ease of response tracking.

Academic Records and School Databases: We would have requested access to historical academic data, including grades, attendance, and disciplinary records, from the school's databases (if available) or implemented data entry mechanisms for schools with non-digitized records.

Observational Data: Observations during classroom sessions to understand teaching methods, student engagement, and participation levels. Monitoring of library usage and access to resources such as books and computers.

Student Performance Tracking: Tracking performance in periodic tests, mock exams, and assignments, along with data on hours spent in supervised study sessions or extracurricular learning activities.

2.3.8 Mock Data Generation Plan

Tools and Approach

Our team initially explored various open-source tools for generating mock datasets, including the popular Faker library in Python. While Faker is great for creating dummy data, we quickly realized it didn't quite fit our needs for this project. The data it produced, such as names and other attributes, felt too generic and foreign, lacking the local context of the Nigerian educational system.

After consulting with a few experienced data professionals, we concluded that the best approach would be to build a custom Python script tailored to our requirements. This allowed us to create datasets that reflect real-world Nigerian school scenarios more accurately, including localized names, school subjects, and socioeconomic data relevant to students in the region.

You can check out the mock data generation script on our GitHub page.

2.4 Data Structuring

The goal of this data structuring phase is to design an efficient and scalable data storage system that supports our data collection plan. This involves creating a database schema to organize and structure the data for easy access, analysis, and reporting.

Data Warehouse Tool

There are several tools available for data warehousing, such as PostgreSQL, MySQL, Snowflake, and Google BigQuery. To decide on the best option for our project, we evaluated each based on key criteria like deployment, cost, security, and availability.

On-Premise vs Cloud-Based Solutions We initially considered setting up an on-premise PostgreSQL database on one of the school's systems to keep costs low. However, this approach posed several risks:

- **Security Concerns:** An on-premise solution could be more susceptible to data breaches or compromises.
- **Data Loss:** There's also the potential risk of losing data if the local hardware fails or isn't properly backed up.

We also explored the option of using PostgreSQL on a remote server, given its flexibility and compatibility with various data models. However, this approach required running a virtual machine, which would involve additional costs for server maintenance and cloud infrastructure.

Given these factors, we realized that a cloud-based solution would be a better fit. Although it might incur some costs, it offers higher security, easier maintenance, and ensures data availability at all times.

Chosen Solution: Google BigQuery After weighing these considerations, we decided to use Google BigQuery for our data warehousing needs. Its robust cloud infrastructure,

ease of integration, and scalability made it the ideal choice for handling and analyzing the volume of data we were working with in this project.

Data Warehouse Design Approach

Designing an ERD

Given the variety of data sources and types (e.g., student bio-data, academic records, attendance, socioeconomic data), we'll use a star schema for the data warehouse. This schema will have Fact Tables to store quantitative data (e.g., scores, attendance) and Dimension Tables to store descriptive data (e.g., student demographics, subject details).

Data Warehouse Schema Design The data warehouse will consist of the following key components:

- **Fact Tables:**
  - **Fact_StudentPerformance:** Stores student performance metrics, including scores, exam results, and performance trends.
  - **Fact_Attendance:** Captures attendance data, including attendance rate.
- **Dimension Tables:**
  - **Dim_Student:** Contains student-specific information such as demographics, bio-data, parental information.
  - **Dim_Parent:** Stores parental demographic data, education level, and socioeconomic background.

- **Dim_Teacher:** Information about teachers, their specializations, years of experience, etc.
- **Dim_Subject:** Lists the subjects offered, subject codes, and categorization (e.g., core, elective).

2.4.1 Database Schema for the Data Warehouse

Let's break down each table and the fields it will contain:

1. dim_students (Student Dimension Table)

Stores information about students, their departments, and activities.

Field	Data Type	Description
ID (PK)	STRING	Unique identifier for each student.
First_Name	STRING	Student's first name.
Last_Name	STRING	Student's last name.
Age	INTEGER	Student's age.
Department	STRING	The department the student belongs to (e.g., 'Science & Mathematics').
Subjects	STRING	List of subjects the student is taking.
Parent_ID (FK)	STRING	Foreign key linking to the parent data.
Access_to_Technology	STRING	Indicates if the student has access to technology ('Yes' or 'No').
Extracurricular_Activities	STRING	Indicates participation in extracurricular activities ('Yes' or 'No').
Private_Home_Tutor	STRING	Indicates if the student has a private tutor ('Yes' or 'No').
Avg_Daily_Study_Hours	FLOAT	Average daily study hours of the student.

Table 2.2: dim_students (Student Dimension Table)

2. dim_subjects (Subject Dimension Table)

Stores information about all subjects taught at the school.

Field	Data Type	Description
ID (PK)	STRING	Unique identifier for each subject.
Subject_Name	STRING	Name of the subject (e.g., 'Mathematics').
Subject_Type	STRING	Type of subject ('Core', 'Trade', or 'Elective').
Class_Level	STRING	The class level the subject is offered in (e.g., 'SSS3').

Table 2.3: dim_subjects (Subject Dimension Table)

3. dim_teachers (Teacher Dimension Table)

Stores information about teachers and their specializations.

Field	Data Type	Description
ID (PK)	STRING	Unique identifier for each teacher.
First_Name	STRING	Teacher's first name.
Last_Name	STRING	Teacher's last name.
Qualification	STRING	Teacher's qualification (e.g., 'BSc', 'MSc', etc.).
Subject_Specialization	STRING	Subjects the teacher specializes in (comma-separated if multiple).
Years_Experience	INTEGER	Number of years of teaching experience.
Gender	STRING	Gender of the teacher ('Male' or 'Female').

Table 2.4: dim_teachers (Teacher Dimension Table)

4. dim_parents (Parent Dimension Table)

Stores information about the students' parents.

Field	Data Type	Description
ID (PK)	STRING	Unique identifier for each parent.
First_Name	STRING	Parent's first name.
Last_Name	STRING	Parent's last name.
Education_Level	STRING	Highest education level of the parent (e.g., 'BSc', 'Primary School').
Occupation	STRING	Parent's occupation (e.g., 'Teacher', 'Trader', etc.).
Socioeconomic_Status	STRING	Socioeconomic status ('High', 'Middle', 'Low').
Marriage_Status	STRING	Parent's marital status ('Married', 'Divorced', 'Single').

Table 2.5: dim_parents (Parent Dimension Table)

5. fact_student_performance (Student Performance Fact Table)

Stores information about student performance across subjects, terms, and years.

6. fact_attendance (Attendance Fact Table)

Stores student attendance rates across terms and years.

Field	Data Type	Description
ID (PK)	STRING	Unique identifier for each performance record.
Student_ID (FK)	STRING	Foreign key linking to the Dim_Students table.
Subject_Name	STRING	Name of the subject.
Year	STRING	Year of the performance (e.g., 'SS1', 'SS2', 'SS3').
Term	STRING	Term in which the performance was recorded ('First Term', 'Second Term', etc.).
Score	INTEGER	Score of the student in the subject.
Grade	STRING	Grade received for the subject (e.g., 'A1', 'B2', 'C6').

Table 2.6: fact_student_performance (Student Performance Fact Table)

Field	Data Type	Description
ID (PK)	STRING	Unique identifier for each attendance record.
Student_ID (FK)	STRING	Foreign key linking to the Dim_Students table.
Year	STRING	Year of the attendance record (e.g., 'SS1', 'SS2').
Term	STRING	Class term ('First', 'Second', 'Third').
Attendance_Rate	FLOAT	The attendance rate of students in percentage.

Table 2.7: fact_attendance (Attendance Fact Table)

Chapter 3

Model Development

3.1 Data Analysis

This study uses python programming language and jupyter notebook to perform various data analysis techniques on the data. The raw data gotten from the data warehouse was first aggregated to get the basic features needed. The aggregated data was analysed both qualitatively and quantitatively. The data was edited to eliminate inconsistencies, summarized and coded for easy classification in order to facilitate interpretation. Descriptive statistics was used in describing the sample data in such a way as to portray the typical respondent and to reveal the general response pattern. New features was also added, for example the grading scores for secondary schools in Nigeria according to WAEC. Feature distribution was done s well to gain some insight into the dataset and see how the features might be treated.

Various insights were gotten from this like; the relative distribution of attendance, study hours and yearly averages. This shows the range and gives us an idea on how the students are performing. No extreme outliers were observed.

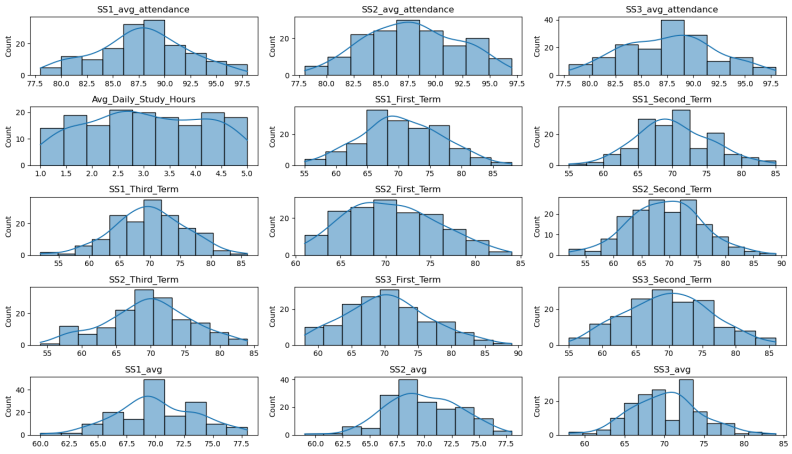


Figure 3.1: Numerical columns distributions.

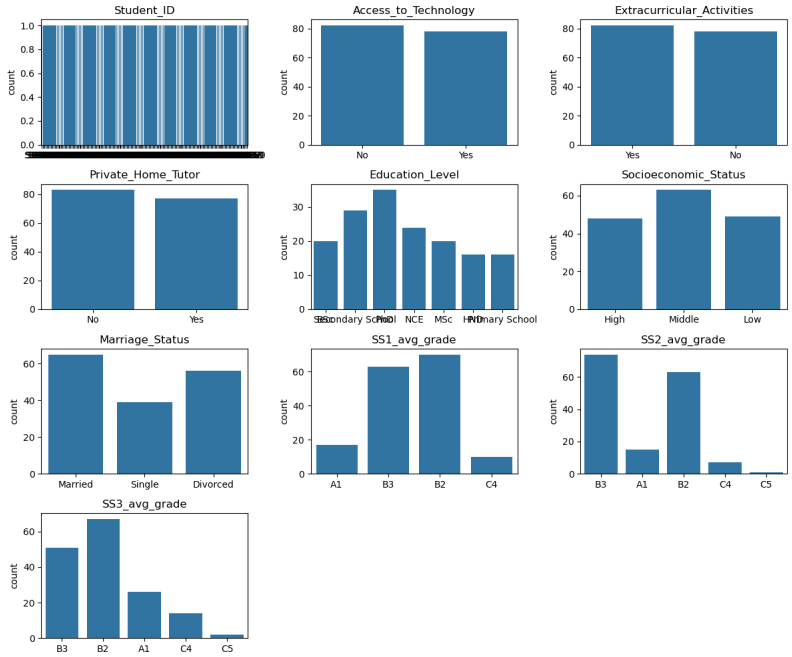


Figure 3.2: Categorical columns distributions.

The categorical data distributions depicted in the histograms provide insightful trends about various factors. For example, parents with secondary school education or NCE (Nigeria Certificate in Education) are the most represented, followed by those with MSc, HND, and Primary school education. This could indicate that a majority of the students come from households with moderate to advanced levels of parental education. To explore relationships between various factors (e.g., study habits, socio-economic status, and school infrastructure) and their impact on student performance, inferential statistics, including correlation and regression analyses, were employed.

3.2 Prediction Model Development

Machine learning is the process of equipping the computers with the ability to learn by using the data and experience like a human brain. The main aim of machine learning is to create models which can train themselves to improve, perceive the complex patterns, and find solutions to the new problems by using the previous dataÖzer Çelik (2024). Machine learning methodology follows the study the machine learning process to develop prediction models. The methodology provides a structured approach to developing prediction modelsMusau et al. (2024). The Machine learning process used in this study, is made up of six steps: creating the data set, data preparation(feature encoding and scaling), feature selection(variance and feature importance), model building(training models and model selection) and evaluation.

Stage One - Student Data Set This stage involves the creation of synthetic data using data generation models. This is discussed in the Data Generation Module.

Stage Two - Data Preparation This step includes all the necessary preprocessing

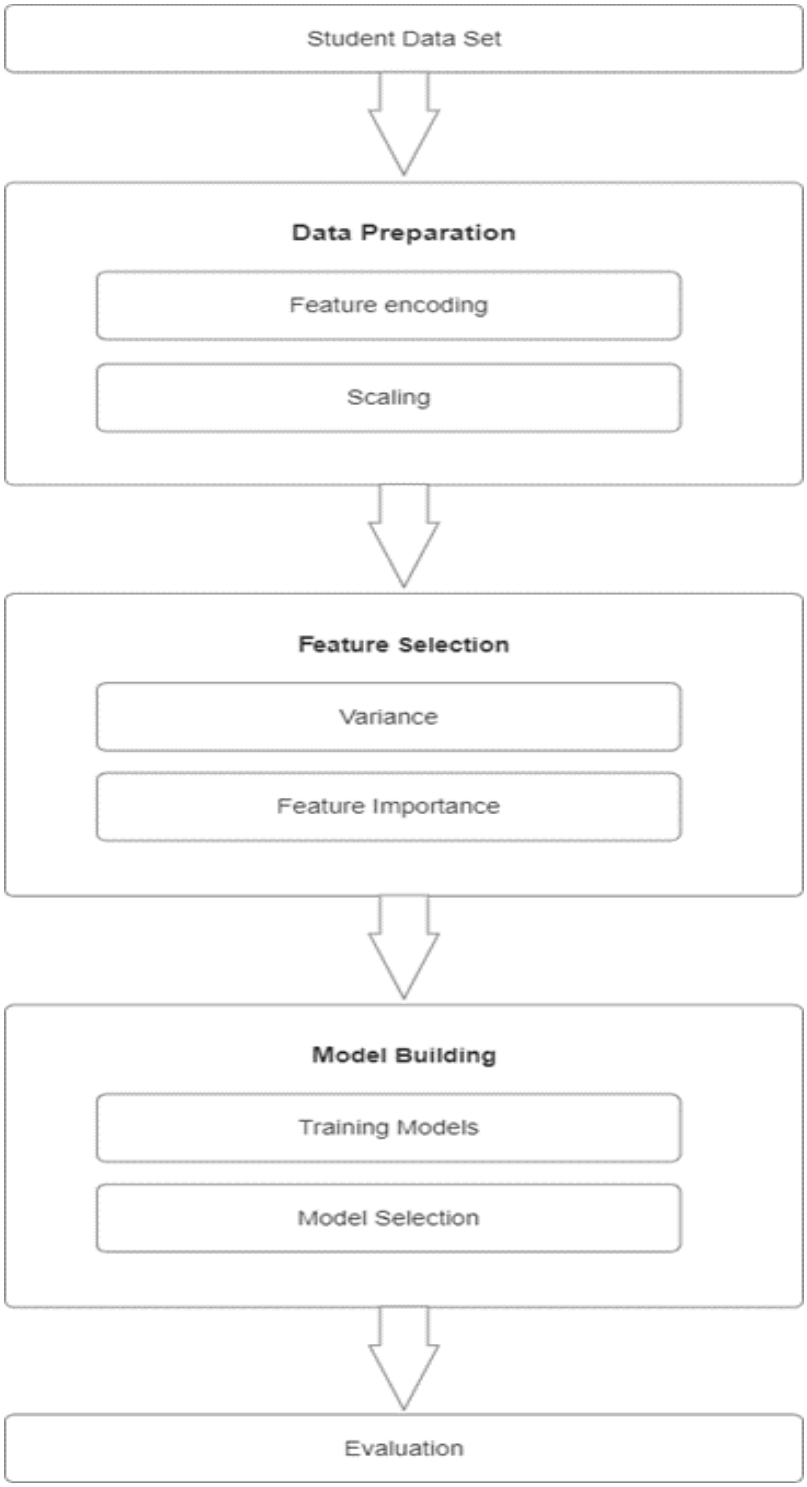


Figure 3.3: Model development Chart.

to make the data suitable for modeling, such as data cleaning, feature encoding, and scaling. These vital processes improve data quality, enhance model performance, and save time by making data modeling more efficient.

Stage Three - Feature Selection Feature selection is used to improve model performance by removing irrelevant features and selecting the most important ones. First, a variance check was carried out to assess the level of data dispersion or spread relative to the mean for each variable. Two columns were removed due to low variance. Feature importance analysis was then performed, which helps in deciding which features to include in the model, leading to better performance and efficiency.

Stage Four - Model Building This stage involves choosing an appropriate algorithm, defining the model structure, and fitting the model to the training data. The goal is to create a representation of underlying data patterns that can be used to make predictions or inform decisions. Multiple models are trained to determine which performs best on the validation data. Once a suitable model is selected, it can be used to predict student grades, indicating whether they pass or fail.

Stage Five - Evaluation Finally, the model is tested on unseen data to evaluate its performance. Key metrics used for evaluation include: `accuracy_score`, `confusion_matrix`, `mean_squared_error`, `mean_absolute_error`, and `r2_score`.

3.3 Model Evaluation

The evaluation of the prediction model involved assessing its performance on a validation dataset using several key metrics:

- **Accuracy:** The overall accuracy of the model is determined by comparing the

predicted values with the actual outcomes, providing a general sense of how well the model performed. This is usually measured as a percentage.

- **Mean Squared Error (MSE) and Root Mean Squared Error (RMSE):** These metrics measure the average of the squared differences between predicted and actual values. RMSE provides a more interpretable error in the same units as the predicted variable, where lower values indicate better model performance.

- **Mean Absolute Error (MAE):** This metric calculates the average absolute difference between predicted and actual values, offering insight into the average magnitude of errors.

- **R^2 Score:** The R^2 value measures how well the model explains the variance in the target variable, with values closer to 1 indicating a better fit.

- **Classification Report:** This report provides a detailed breakdown of model performance for classification, including precision, recall, and F1-score for each class.

- **Confusion Matrix:** A confusion matrix visualizes the performance of the classification model, showing the counts of true positives, false positives, true negatives, and false negatives. A heatmap of the confusion matrix was plotted to illustrate these results.

These metrics offer a comprehensive view of the model's performance, evaluating its accuracy, error rates, and classification capabilities.

3.4 Results

One of the main purposes of this study was to develop a machine learning model for prediction of students' academic performance. The chapter presents the research findings, results, and discussions of the study. The study starts with having interview sessions with students and teachers. Due to time constraints a proper data collection wasn't conducted but from the insights gotten, data generation was aided. Then data generation. Through synthetic data generation models. After carefully generating the data, the next step was data warehousing and structuring. This will form the basis of the schools' enterprise data solution such as data collection, pipe lining, warehousing, automation and reporting. Data aggregation was carried out as well. It Involves, collating all the needed features from all the data tables and joining them in a single dataset. This produced the students dataset. In order to make good predictions, understanding the data, analysis is a priority. Exploratory data analysis was carried out. Both descriptive, statistically and visually. This task gives a clear idea on things like the distribution, correlation, variance etc of the data. Model development which results in the prediction of the likelihood of a student passing or failing their upcoming exam considering the factors, comes next. After model training and selection, SVM model was chosen because of its high accuracy compared to the rest and its low variability.

SVM models which is a supervised machine learning algorithm, classifies data by finding an optimal line that maximizes the distance between each class. After training the model was tested on its ability to accurately predict the performance of students. On this it scored an accuracy point of 31According to Alexandra Barr (2023) models suffer from catastrophic forgetting and data poisoning when trained on synthetic data.

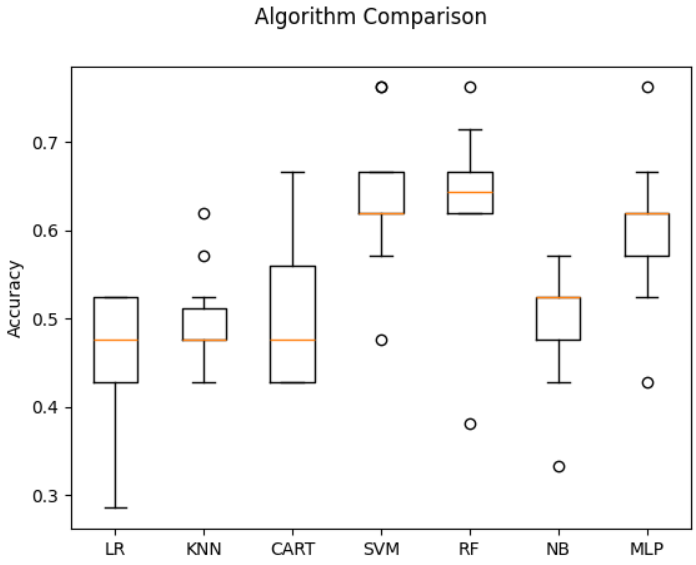


Figure 3.4: Algorithm Performances.

This and the fact that there seems to not be a relationship between the features and target variables. Among other things could be the reason for the performance of the model. Streamlit web app was used to display results of the model. Scores above 40 according to WAEC standards are deemed as pass marks while those below are termed fails. This was taken into consideration and executed. As a result, predictions on the students performances whether they will pass or fail can be gotten from the app HERE .

Chapter 4

Data Visualization

4.1 Overview

Data visualization is the graphical representation of information and data using visual elements like charts, graphs, maps, and dashboards. It provides an accessible way to see and understand patterns, trends, and insights within data by transforming complex datasets into visual formats. The goal of this visualization is to communicate information clearly and efficiently, helping school management make data-driven decisions and improvements.

Before this visualization was created, several important steps were taken to reach a conclusion. Such steps include:

i **Posing Relevant Business Questions:** These key questions include:

- How does student attendance correlate with overall academic performance across different terms and subjects?
- Which factors have the most significant impact on student performance?

- What are the performance trends by department and education level, and how do they vary across different academic years?
- How does participation in extracurricular activities or having a private tutor affect student outcomes?

ii **Data Preparation:** The business intelligence tool Power BI is used for preparation. After data is collected from the data warehouse, it is then loaded into Power BI to perform data cleaning. The steps taken in data cleaning include:

- Identifying and handling any missing or incomplete data points.
- Ensuring that no duplicated entries are present, especially in student-level data.
- Ensuring consistency by standardizing units, formats, and values.

After the data cleaning process, the next step is Data Modeling. In this phase, the data is organized into fact and dimension tables to make querying, analysis, and relationships between data entities more efficient. This structure ensures that data visualizations and reports run smoothly and perform well. Fact tables contain quantitative data, such as scores and attendance, which are often used to carry out calculations or aggregations. Dimension tables, on the other hand, hold descriptive information like student details, subjects, or departments, helping to filter and categorize the facts to generate meaningful insights.

After the data modelling process, the next process is Data Transformation. In this phase, additional metrics is calculated, such as total students, overall attendance

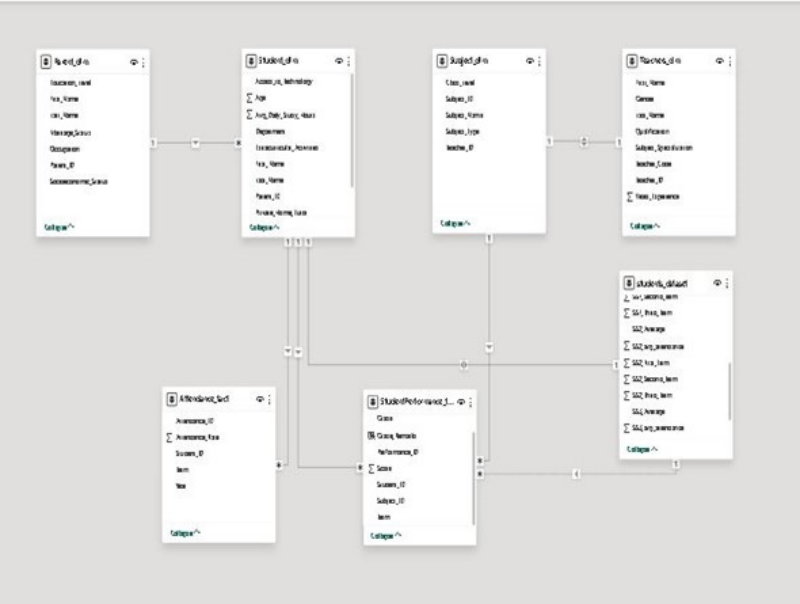


Figure 4.1: Diagrammatic representation of the Student Performance Data Model

rates, averages score, and pass rates etc. These measures /calculated columns were further validated to ensure accuracy and reliability for the next steps of analysis.

iii **Exploratory Data Analysis:** The goal is to explore and better understand the student performance, identify trends, and uncover relationships between variables such as scores, attendance rates, socioeconomic status, and access to technology Several important trends were discovered in the student performance data. Students received an average score of roughly 70, an attendance rate of roughly 88%, and a pass rate of 82%. In general, students who attended more classes did better; students from wealthier homes or with access to technology also did better. Students performed differently in each subject, with science and math showing the greatest improvement. As pupils advanced from SS1 to SS3, there was a noticeable decline in their results. When comparing departments, students studying Science and

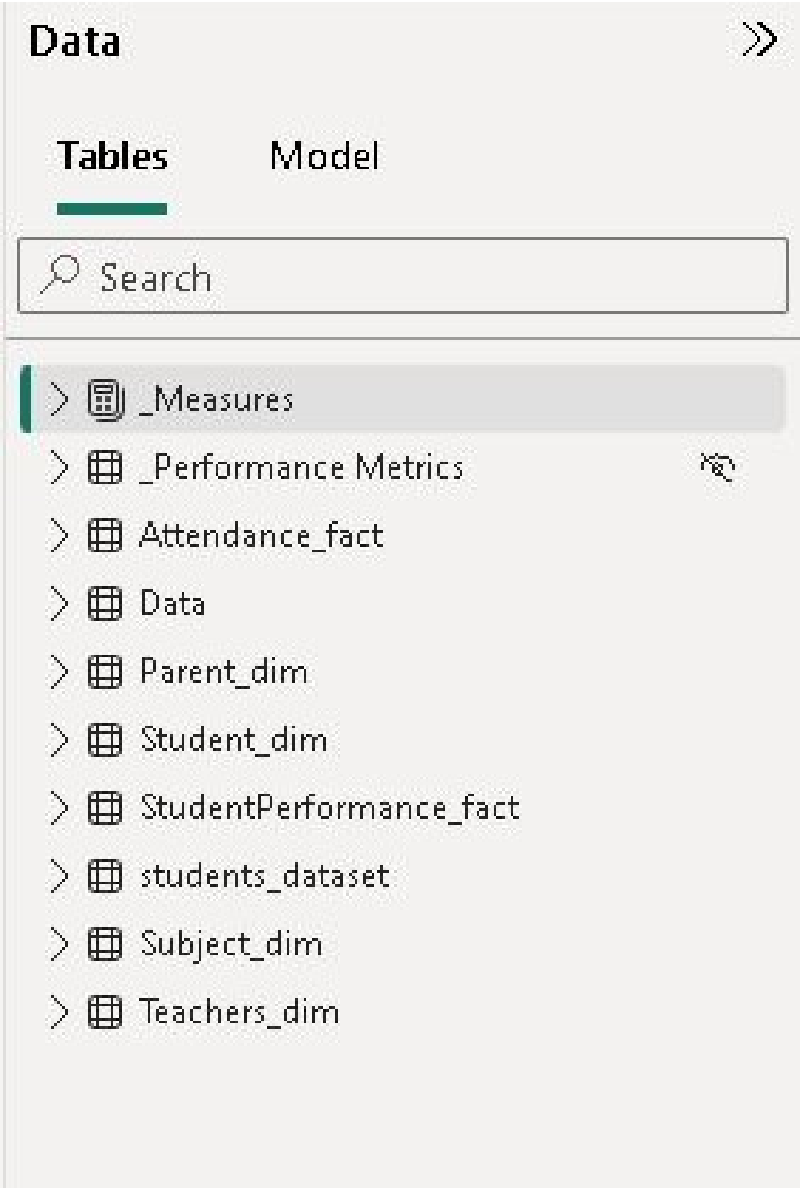


Figure 4.2: Calculated Dax Measures for Performance Analysis

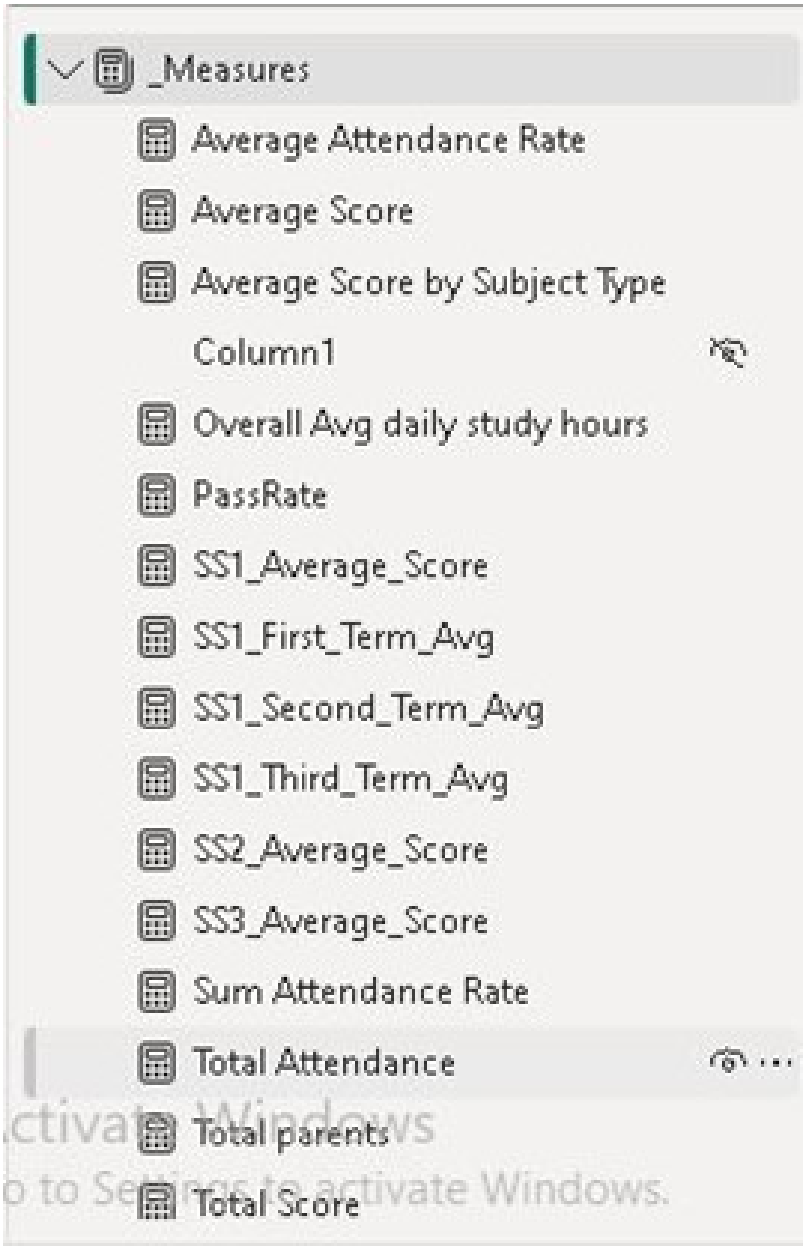


Figure 4.3: Data Tables and Measures for Performance Analysis

Mathematics did better than those studying Humanities and Commerce. Although taking part in extracurricular activities or hiring a private tutor helped students' results somewhat, attendance and access to technology had a greater influence. A few exceptions popped out: some students did well even though they had low attendance, and several departments or groups had persistent problems, indicating areas that may require assistance.

iv **Data Visualization:** To communicate these insights better or make the data easier to understand and actionable for stakeholders, visualization was carried out using simple bar charts and tables in PowerBI. Can be seen HERE.

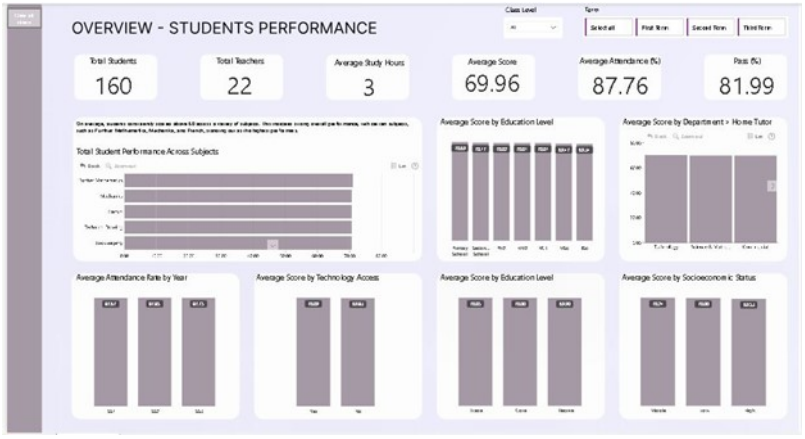
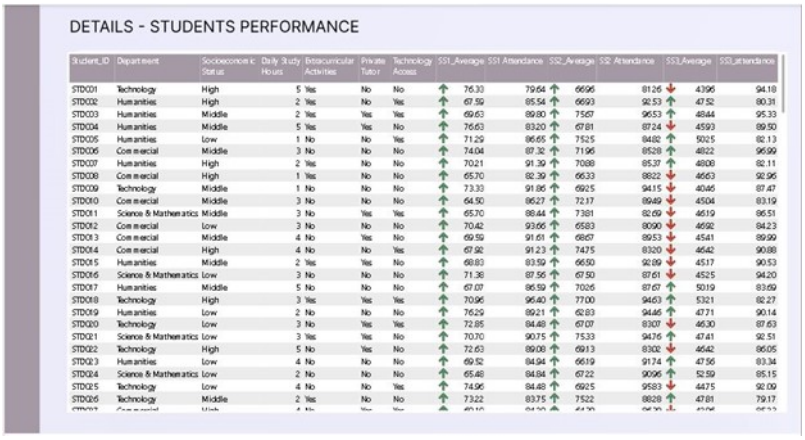


Figure 4.4: Overview Dashboard for Student Performance Metrics



DETAILS - STUDENTS PERFORMANCE

Student_ID	Department	Socioeconomic Status	Daily Study Hours	Extracurricular Activities	Private Tutor	Technology Access	SS1_Average	SS1 Attendance	SS2_Average	SS2 Attendance	SS3_Average	SS3_attendance
STD001	Technology	High	5	Yes	No	No	76.30	79.64	6696	81.26	4396	94.18
STD002	Humanities	High	2	Yes	No	Yes	67.59	85.54	6693	92.53	4752	80.31
STD003	Humanities	Middle	2	Yes	Yes	Yes	69.63	89.80	7567	96.53	4844	95.33
STD004	Humanities	Middle	5	Yes	Yes	No	76.63	83.20	6781	87.24	4593	89.50
STD005	Humanities	Low	1	No	No	Yes	71.29	86.65	7525	94.82	5025	82.13
STD006	Commercial	Middle	3	No	No	No	74.04	87.32	7196	85.28	4822	96.99
STD007	Humanities	High	2	Yes	No	No	70.21	91.39	7088	85.37	4808	82.11
STD008	Commercial	High	1	Yes	No	No	65.70	82.39	6633	88.22	4663	92.96
STD009	Technology	Middle	1	No	No	No	73.33	91.86	6925	94.15	4046	87.47
STD010	Commercial	Middle	3	No	No	No	64.50	86.27	7217	89.49	4504	83.19
STD011	Science & Mathematics	Middle	3	No	Yes	Yes	65.70	88.44	7381	92.69	4619	86.51
STD012	Commercial	Low	3	No	No	No	70.42	93.66	6583	80.90	4692	84.23
STD013	Commercial	Middle	4	No	Yes	No	69.59	91.61	6867	89.53	4541	89.99
STD014	Commercial	High	4	No	No	Yes	67.92	91.23	7475	83.20	4642	90.88
STD015	Humanities	Middle	2	Yes	Yes	No	68.83	83.59	6650	92.89	4517	90.53
STD016	Science & Mathematics	Low	3	No	No	No	71.38	87.56	6750	87.61	4525	94.20
STD017	Humanities	Middle	5	No	No	No	67.07	86.59	7026	87.67	5019	83.69
STD018	Technology	High	3	Yes	Yes	Yes	70.96	96.40	7700	94.63	5321	82.27
STD019	Humanities	Low	2	No	No	No	76.29	89.21	6283	94.46	4771	90.14
STD020	Technology	Low	3	No	Yes	Yes	72.85	84.48	6707	83.07	4630	87.63
STD021	Science & Mathematics	Low	3	Yes	Yes	No	70.70	90.75	7533	94.76	4741	92.51
STD022	Technology	High	5	No	Yes	No	72.63	89.06	6913	83.02	4642	96.05
STD023	Humanities	Low	4	No	No	No	69.52	84.94	6619	91.74	4756	83.34
STD024	Science & Mathematics	Low	2	No	No	No	65.48	84.84	6722	90.96	5259	85.15
STD025	Technology	Low	4	No	No	Yes	74.96	84.48	6925	95.83	4475	92.09
STD026	Technology	Middle	2	Yes	No	No	73.22	83.75	7522	88.28	4781	79.17

Figure 4.5: Tabular Representation of detailed student Performance Metrics

Chapter 5

Final Solution

5.1 Introduction

This chapter presents the final solution developed for the hypothetical secondary school aimed at enhancing student performance through a data-driven approach. The solution is a web application built using Streamlit, which facilitates the collection of data, automates data collation, updates existing records, and predicts student performance.

5.2 Web Application Overview

The web application serves as an integrated platform where data regarding students, their academic performance, and various socio-economic factors can be collected and analyzed. The application is designed to provide an intuitive user interface for various stakeholders including teachers, administrators, and parents.

5.2.1 Key Features

- **Data Collection:** The app allows users to input data through forms, ensuring that all necessary information such as student demographics, academic records, and resource utilization is captured efficiently.
- **Automated Data Collation:** Once data is collected, the app automatically organizes and collates the information into a structured database, reducing manual errors and saving time.
- **Updating Existing Records:** Users can easily update existing student records. This feature ensures that the data remains current and accurate, which is vital for effective analysis.
- **Performance Prediction:** Utilizing machine learning algorithms, the app predicts student performance based on historical data, identifying students at risk of underperforming and enabling timely interventions.
- **Power BI Dashboard:** A dynamic dashboard visualizes the key performance indicators (KPIs) of the school, offering insights into trends and facilitating evidence-based decision-making.

5.3 Technical Implementation

The implementation of the web application involved several key components:

5.3.1 Framework and Tools

The choice of Streamlit as the primary framework was driven by its simplicity and rapid development capabilities. Streamlit allows for the quick deployment of interactive web applications with minimal coding effort. The following tools and libraries were utilized:

- **Python:** For the backend logic and data processing.
- **Pandas:** For data manipulation and analysis.
- **Scikit-learn:** For implementing the machine learning models used for performance prediction.
- **big query:** For database interactions, facilitating the storage and retrieval of student data.
- **Power BI:** Visual data representation to track performance metrics. Can be seen HERE.

5.3.2 Architecture

The architecture of the application is designed to ensure scalability and efficiency. It consists of the following layers:

- **Frontend:** Built using Streamlit, providing a user-friendly interface for data entry and visualization.
- **Backend:** Handles data processing, modeling, and business logic. This layer integrates machine learning models to predict student performance based on the collected data.

- **Database:** A relational database (e.g., excel csv, big query) stores all the data, ensuring secure and efficient data management.

5.4 Deployment

The application was deployed on a cloud platform to ensure accessibility and reliability. Streamlit's deployment options were utilized, allowing for easy sharing of the application with stakeholders. The deployment process involved the following steps:

- **Environment Setup:** Configuring the cloud environment to run the Streamlit application.
- **Code Deployment:** Pushing the codebase to the cloud server.
- **Database Integration:** Connecting the application to the PostgreSQL database for data storage and retrieval.
- **Testing:** Conducting thorough testing to ensure the application functions as intended and is free of bugs.

5.5 Future Work

In future iterations of the project, additional features should be added such as:

- real-time data should be collected with the use of the app to update model and improve accuracy.

- real-time data visualization and advanced analytics could be incorporated to within the app to enhance the application's functionality and give real-time insights to school management.

Chapter 6

Conclusion and Recommendations

6.1 Overview

This chapter presents the objectives and achievements of the study. The chapter starts by systematically reviewing the research objectives and how the study has addressed each objective by answering the research questions associated with the objective. Finally, the chapter gives a conclusion, the contributions of the study, recommendations for future work and limitations of the study.

The objectives of the study were:

- Identify possible problems students might be facing while writing these exams.
- Generate data through any means based on the needs of the solution, ensuring that the data adequately reflects the state of the African education ecosystem.
- Design an enterprise data solution for the school's data collection, pipelining, warehousing, automation, and reporting needs.

- Create an optimized model that predicts the likelihood of a student passing or failing their upcoming exam based on their academic history.
- Make relevant recommendations to the stakeholders on how they can help improve the performance of the students based on the solution.

6.2 Objectives of the Study

6.2.1 Objective 1: Identify Possible Problems Students Might be Facing

This research objective led to the formulation of the research question: *What are the common challenges faced by students in preparing for high-stakes exams like JAMB and WASSCE?*

Following this, efforts were made to interview students, teachers and parents to determine factors that could contribute towards students' performance. The interviews and meetings had especially with the students and teacher was very helpful in giving understanding and insights into the current factors. For example, one student talked about not having private tutors and how that has affected her in mathematics because she needs more help in this area. Some other factors which include likeness for a particular teacher or not, ability of teachers etc were not considered so as not to make the model too complex. However, this can be taken into consideration on future reviews. The following list of factors were decided on as a result of the consultation and interviews with the various involved parties:

- Student_ID
- SS1_avg_attendance

- SS2_avg_attendance
- SS3_avg_attendance
- Access_to_Technology
- Extracurricular_Activities
- Private_Home_Tutor
- Avg_Daily_Study_Hours
- Education_Level
- Socioeconomic_Status
- Marriage_Status
- SS1_First_Term
- SS1_Second_Term
- SS1_Third_Term
- SS2_First_Term
- SS2_Second_Term
- SS2_Third_Term
- SS3_First_Term
- SS3_Second_Term

6.2.2 Objective 2: Generate Data Reflecting the African Education Ecosystem

This research objective led to the formulation of the research question: *What types of data are most crucial for understanding and improving student performance in the African education ecosystem?*

Due to the lack of data collection infrastructure in most schools, synthetic data was used. However, challenges include:

- Distribution issues: The generation of synthetic datasets often lacks sufficient consideration for demographic diversity Hao et al. (2024). This can lead to unbalanced data distributions in terms of performance grades, parental education level, attendance time etc.
- Lack of realism and accuracy: Synthetic data may not capture the complexity of real-world datasets and can potentially omit important details or relationships needed for accurate predictions Lamberti (2023).
- Lack of trends: Data generated artificially tend to miss out on capturing evolving trends or anomalies that arise over time.
- Overall Impact on research work: This could impact negatively on our work which is based on the premise of solving real life problems. On one hand the argument of inaccurate conclusions might arise and also there is the case of limited practical applications.

Despite these drawbacks of using synthetic data for this study, what we hope to show here is the process. Which is in line with the stated objectives, such that this template

can be replicated elsewhere. For example, a school starting its data warehousing can use the pipeline to generate insights on the performance of students based on its real-life data.

6.2.3 Objective 3: Design an Enterprise Data Solution

This research objective led to the formulation of the research question: *What data warehousing solutions are best suited for the school, considering scalability, cost-effectiveness, and ease of implementation?* The aim was to build a robust predictive framework that analyzes relevant data such as student demographics, attendance, study habits, socioeconomic background, and prior academic records to accurately forecast future performance. By employing machine learning algorithms, this model will enable the school to anticipate individual student success, identify at-risk students, and tailor interventions to improve academic results. Also data collection and updating records was automated.

6.2.4 Objective 4: Create an Optimized Prediction Model

This research objective led to the formulation of the research question: *How can we model student academic performance based on significant factors?*

For this reason, exploratory data analysis was done before building the model to view different aspects of the data. This way, the relative distributions, correlations, variance can be viewed. Machine learning models in general are capable of predicting results based on the learned patterns in the training data. Keyword is training data, which is why this project forms a template that should be built upon with real data. The prediction model built, predicts the grades of students from A1 – F9. To build the model, data preparation

was carried out with various techniques such as; scaling and feature encoding. Feature selection was done just before training multiple models. This was essential to determine the top 17 features to be used. After this the best model was selected from the results.

6.3 Recommendations

6.3.1 Enhance Academic Performance through Key Factor Identification

The study identified significant factors such as:

- SS3_avg_attendance
- SS2_avg_attendance
- SS1_avg_attendance
- SS1_First_Term
- SS1_Second_Term
- SS1_Third_Term
- Avg_Daily_Study_Hours
- Parental_Education_Level

Therefore, focus on the major contributing factors and work towards collection, monitoring and implementation of recommendations.

6.3.2 Student Attendance Monitoring

Implement stricter attendance policies to ensure student presence in class. Head of school should intensify on continues appraisal of school's system through regular supervision of school activities and prompt submission of their reports to appropriate authority Mohammed (2017).

6.3.3 Monitor Student Academic Performance Throughout the School Year

The first and foremost criterion for determining the success of a school in Nigeria is its academic performance Scholar (2024). Develop learning plans to support students throughout each term (first, second, third) across all academic years (SS1, SS2, SS3). Regular assessments and feedback mechanisms should be introduced to help students stay on track. Things like Mock examinations can help students prepare for the real exams.

6.3.4 Promote Effective Study Habits

Introduce peer study groups and access to additional learning materials. Learning is a change in behaviour. By planning effective study habit programs, students can learn how to study effectively to bring positive results.[Ogbodo (2024).

6.3.5 Engage Parents in the Academic Process

Educate parents on how they can support their children's education, especially in terms of study habits and providing a conducive learning environment at home. Teachers can

provide guidance on how parents can help with certain assignments and parents can provide feedback on areas where their child may need extra help.

6.3.6 Improve Access to Learning Resources and Technology

Ensure that students have access to technology that supports their learning, such as computers, internet access, and e-learning platforms. Conduct workshops and online courses to train teachers on using digital tools and resources uLesson Group (2024).

6.3.7 Incorporate Extracurricular Activities into the Academic Plan

Encourage students to participate in extracurricular activities, but ensure a balance between academics and other engagements. Monitor the impact of these activities on student performance and adjust participation levels if needed.

References

A. Barr. Synthetic vs real data: Why do models perform worse when trained on synthetic data? *Superfast AI Newsletter*, 2023. URL https://alexandrabarr.beehiiv.com/p/synthetic-data?__cf_chl_tk=.JazhLo520Eitpy5z9deXxtZhpLSQn47BekJBJh6b.s-1728385141-0.0.1.1-5268.

S. Hao, W. Han, T. Jiang, Y. Li, H. Wu, C. Zhong, Z. Zhou, and H. Tang. Synthetic data in ai: Challenges, applications, and ethical implications. *arXiv*, 2401(01629v1), 2024. URL https://arxiv.org/abs/2401.01629.

R. Kapur. Significance of education in human life. *ResearchGate*, 2018. URL https://www.researchgate.net/publication/324907980_Significance_of_Education_in_Human_Life.

A. Lamberti. The benefits and limitations of generating synthetic data. *Syntheticus*, 2023. URL https://syntheticus.ai/blog/the-benefits-and-limitations-of-generating-synthetic-data.

A. Mohammed. Influence of monitoring and evaluation strategies on teaching and learning in secondary schools in federal capital territory abuja, nigeria. 2017.

Obadiah Matolo Musau, Kelvin Omieno, and Raphael Angulu. Factors affecting secondary school students' academic performance. *Journal of Education and Practice*, 2024.

Onyara Beatrice Nambuya. School based factors influencing student's academic performance at kenya certificate of secondary education in teso south district. *East African Journal of Education and Social Sciences*, 2024.

R. O. Ogbodo. Effective study habits in the educational sector: Counselling implications. *Continuous Education, FCT College of Education*, 2024.

Bunmi Isaiah Omodan and Haastrup T. Ekundayo. Enhancing students' academic performance in secondary schools: The vicissitude of classroom management skills. *International Journal of Education and Research*, 2024.

Naija Scholar. Case study: Successful schools their strategies. *Disciplines In Nigeria*, August 2024. URL https://disciplines.ng/successful-schools-strategies.

Aimable Sibomana, Christian Bob Nicol, and John Sentongo. Factors affecting the achievement of twelve year basic students in mathematics and science in rwanda. *Journal of Educational Research*, 2024.

The uLesson Group. Education in nigeria: A practical technology integration guide for government agencies. *LinkedIn*, July 2024. URL https://www.linkedin.com/pulse/education-nigeria-practical-technology-integration-guide-government-kskcf#:~:text=Schools%20and%20government%20agencies%20can,integration%20of%20technology%20in%20classrooms.

Özer Çelik. A research on machine learning methods and its applications. *Journal of Machine Learning Research*, 2024.

Appendix A

Appendix

A.1 Student Academic Performance Prediction Questionnaire

The purpose of this questionnaire is to facilitate the collection of data that will be used to develop a machine learning model for predicting secondary school students' academic performance. The respondents are students in the secondary school.

A.1.1 Section I: General Information

- Date: ____ / ____ / 2024
- Class: _____

A.1.2 Section II: Student Demographic Information

i Gender: __ Male __ Female

ii What is your age? __ Below 14 Years __ 14–18 Years __ Above 18 Years

iii What is your religion? __ Christian __ Muslim __ Others

If others, please specify: _____

A.1.3 Section III: Family Information

[resume]Did you live with your parents? __ Yes __ No

If No, where are your parent(s)? __ Deceased __ Divorced __ Separated __ Am Adopted __ Others

If others, please specify: _____ Have you ever witnessed conflicts between your parents? __ Yes __ No

If Yes, how often? __ Very often __ Often __ Rarely What kind of family structure do you come from? __ Nuclear Family __ Single Parent __ Extended Family __ Step Family Which of the following best describes your family structure? Tick [✓] appropriately

iv
- I live with my parents, brothers, and sisters only __
- I live with my brothers, sisters, and one parent __
- I live with my parents, brothers, sisters, cousins, grandparents, uncles, and aunts __
- I live with a step-parent __
- I live with people who are not my real parents __

v Did you have any difficulties in fees payment? __ Yes __ No

vi Who used to pay your school fees? __ Parents __ Guardian __ Others

If others, specify _____

A.1.4 Section IV: Co-Curricular Activities

[resume]Did you participate in any co-curricular activities like games, drama, etc.

__ Yes __ No

If Yes, how often __ Daily __ Once a week __ Once monthly __ Not Often

__ Never Did you ever become a member of any school team __ Yes __ No

Did you ever represent your school in any co-curricular activity __ Yes __ No

A.1.5 Section V: Academic Information

[resume]Indicate the number of subjects taken in each of the following levels:

iii

Class	No. of Subjects Taken
SSS I	
SSS II	
SSS III	

ii Did you ever change schools while in junior secondary school __ Yes __ No

If Yes, how many times ____

iii Did you ever repeat a class while in school __ Yes __ No

If Yes, which class: _____