
Predicting Academic Success in Nigerian University Admission Exams: A Data-Driven Approach to Student Performance
Veda Garg, Janelle Correa, Sofia Villafuerte

Table of contents

01

Introduction

What is JAMB ?

The Joint Admission and Matriculation Board (JAMB) exam is an entrance examination given to Nigerian students, in order to assess their eligibility & admission into tertiary-level education systems.

Students must have earned their Senior School Certificate in order to take the exam and be eligible for admission into a University or College.

The scoring for the exam ranges from 0 to 400 and the minimum scores for different institutions is as follows:

- Universities: 140
- Colleges of Education: 100
- Polytechnics, Innovation Enterprise Institutions: 100

Our Goal

Using the dataset of various different scores and other specific academic indicators, we want to analyze the relationship between each indicator and its effect on a students JAMB score. From there, we will train various learning models to predict the score of a student and analyze which features & model produce the most accurate results.

Key Stakeholders:

- Educational Institutions
- Policy Makers
- Students & Families seeking admission

02

Dataset

Raw Dataset

We got our dataset, “Student Performance on 2024 JAMB”, from Kaggle. The layout is as follows:

	JAMB_Score	Study_Hours_Per_Week	Attendance_Rate	Teacher_Quality	Distance_To_School	School_Type	School_Location	Extra_Tutorials
0	192	22	78	4	12.4	Public	Urban	Yes
1	207	14	88	4	2.7	Public	Rural	No
2	182	29	87	2	9.6	Public	Rural	Yes
3	210	29	99	2	2.6	Public	Urban	No
4	199	12	98	3	8.8	Public	Urban	No

Access_To_Learning_Materials	Parent_Involvement	IT_Knowledge	Student_ID	Age	Gender	Socioeconomic_Status	Parent_Education_Level	Assignments_Completed
Yes	High	Medium	1	17	Male	Low	Tertiary	2
Yes	High	High	2	15	Male	High	NaN	1
Yes	High	Medium	3	20	Female	High	Tertiary	2
Yes	Medium	High	4	22	Female	Medium	Tertiary	1
Yes	Medium	Medium	5	22	Female	Medium	Tertiary	1

Cleaning

Unnecessary Columns

The Dataset provided some unnecessary columns for our analysis, such as “Student_ID”, so we decided to drop those columns

Nulls

There was one specific columns with a lot of null values, “Parent Education Level”. Filled in null values with a new label: “Missing”

```
null check:
 JAMB_Score                         0
Study_Hours_Per_Week                0
Attendance_Rate                     0
Teacher_Quality                     0
Distance_To_School                  0
School_Type                         0
School_Location                     0
Extra_Tutorials                     0
Access_To_Learning_Materials        0
Parent_Involvement                  0
IT_Knowledge                        0
Age                                 0
Gender                              0
Socioeconomic_Status                0
Parent_Education_Level            891
Assignments_Completed               0
dtype: int64
```

```
Parent_Education_Level Check:
 Parent_Education_Level
Secondary     1556
Primary       1335
Tertiary      1218
Name: count, dtype: int64
```

EDA

Step 1:

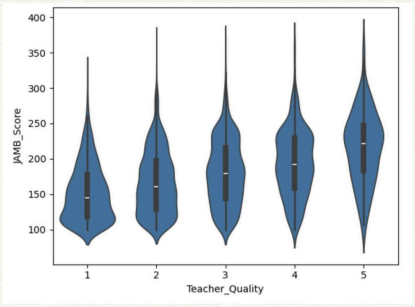
Plot each feature against JAMB_Score to check for any relationships

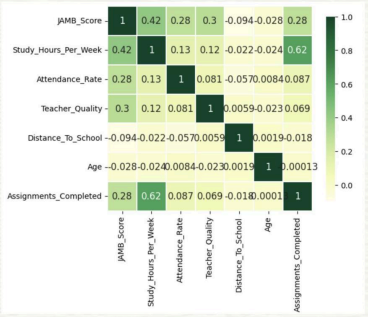
Step 2:

Determine whether feature would be useful for inclusion in model

Step 3:

Make correlation heatmap to ensure features are not related to one another

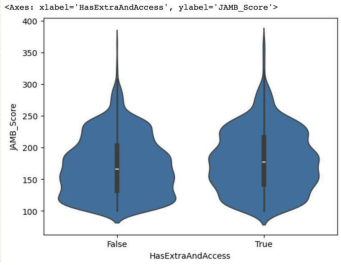




Feature Engineering

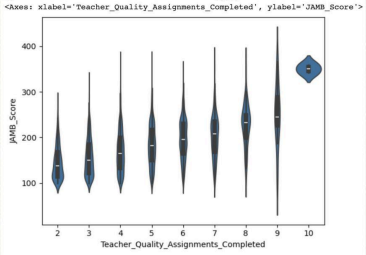
Categorization based upon students having BOTH Extra_Tutorials and Access_To_Learning_ Materials

```
medians                 JAMB_Score
HasExtraAndAccess
False                        166.0
True                         177.0
<Axes: xlabel='HasExtraAndAccess', ylabel='JAMB_Score'>
```



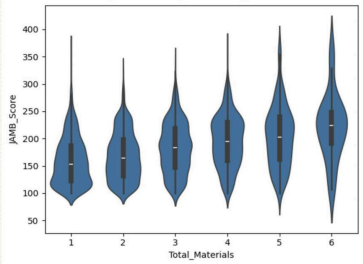
Sum of Teacher_Quality and Assignments_Completed

```
medians                                 JAMB_Score
Teacher_Quality_Assignments_Completed
2                                            138.0
3                                            150.0
4                                            165.0
5                                            182.0
6                                            196.0
7                                            208.0
8                                            233.0
9                                            245.0
10                                           350.5
<Axes: xlabel='Teacher_Quality_Assignments_Completed', ylabel='JAMB_Score'>
```



Assignments_Completed + 1 if Extra_Tutorials was “Yes”

```
medians                 JAMB_Score
Total_Materials
1                            153.0
2                            164.0
3                            183.5
4                            195.0
5                            202.0
6                            224.0
<Axes: xlabel='Total_Materials', ylabel='JAMB_Score'>
```



03

Methods

Preparing Data for Modeling

Objective: Further prepare data to accurately predict JAMB score categories.

1.) Data Transformation:

- Transformed the JAMB scores (ranging from 0 to 400) into 20 discrete categories, each representing a score range of 20 points.
- This transformation redefined the task as a multiclass classification problem.

2.) Feature Selection and Engineering

- Identified key features impacting JAMB scores, like Study Hours Per Week, Attendance Rate, Teacher Quality, and Parental Involvement
- Add new engineered features like Teacher Quality Assignments Completed, HasExtraAndAccess, Total Materials

3.) Categorical Data Encoding

- Used one-hot encoding for categorical variables like School Type (Public or Private) and Parent Education Level (Primary, Secondary, Tertiary).
- Applied label encoding to ordinal features like Parental Involvement and IT Knowledge

	Total_Materials	Parent_Involvement	IT_Knowledge	Attendance_Rate	Study_Hours_Per_Week	Teacher_Quality_Assignments_Completed	HasExtraAndAccess	School_Type_Public
0	3	0	2	78	22	6	True	True
1	1	0	0	88	14	5	False	True
2	3	0	2	87	29	4	True	True
3	1	2	0	99	29	3	False	True
4	1	2	2	98	12	4	False	True
...	...	...	...	...	...	...	...	...
4995	3	1	1	74	20	4	False	True
4996	1	2	2	80	0	3	False	True
4997	3	1	0	89	17	6	False	True
4998	1	1	2	96	15	3	False	True
4999	3	2	2	100	34	3	True	True

Models Used

We explored various machine learning models to handle multiclass classification

Random Forest Classifier:

- Effective in handling high-dimensional data and providing feature importance insights. Robust against overfitting for moderately complex problems.

Logistic Regression:

- A simple baseline model for multiclass classification, leveraging the One-vs-Rest approach for interpretability.

Gradient Boosting Classifier:

- Known for capturing complex, non-linear relationships and handling noisy data well.

Decision Tree Classifier:

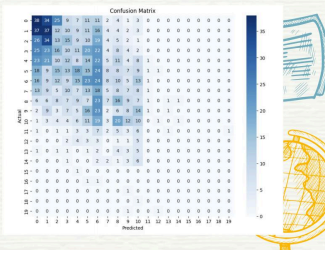
- Easy to interpret and implement, serving as a quick baseline model to understand feature splits and their contribution.

Neural Network:

- Capable of capturing complex interactions between features and making predictions for highly non-linear problems.

Classification Report:

	precision	recall	f1-score	support
0	0.18	0.26	0.22	145
1	0.19	0.26	0.22	145
2	0.10	0.09	0.10	138
3	0.09	0.07	0.08	145
4	0.07	0.06	0.06	139
5	0.09	0.10	0.10	146
6	0.10	0.17	0.13	145
7	0.09	0.05	0.06	104
8	0.14	0.15	0.15	108
9	0.11	0.08	0.09	97
10	0.10	0.11	0.10	95
11	0.00	0.00	0.00	34
12	0.00	0.00	0.00	19
13	0.00	0.00	0.00	18
14	0.00	0.00	0.00	15
15	0.00	0.00	0.00	1
16	0.00	0.00	0.00	2
17	0.00	0.00	0.00	1
18	0.00	0.00	0.00	1
19	0.00	0.00	0.00	2
accuracy			0.12	1500
macro avg	0.06	0.07	0.07	1500
weighted avg	0.11	0.12	0.11	1500



Evaluation & Insights

Model Performance:

The Linear Regressor achieved the lowest MSE and the highest R^2 score, making it the most effective model for predicting continuous JAMB scores in this setup. The Decision Tree Regressor performed poorly, likely due to overfitting on the training data.

Feature Contributions:

Feature importance from tree-based models (e.g., Random Forest, Gradient Boosting) highlighted:

- Attendance Rate, Parent Involvement, and Assignments Completed as the most predictive features.
- Features like Distance to School and School Location contributed minimally and may be excluded in future iterations.

Model Limitations:

- Moderate R^2 scores across models indicate that additional features or more refined preprocessing might be necessary.
- High variance in some features (e.g., Study Hours Per Week) might have contributed to inconsistencies.

04

Results & Conclusions

Results

Linear Regression and Neural Network Regression demonstrated better predictions in predicting JAMB scores compared to other models like Random Forest, Gradient Boosting, and Decision Tree Regression.

While these models were good...how did we make them better?
Our models were not terrible in terms of error but it is important to bring to light that there was limitations with the data. There were fewer scores that were in the top percentile than the rest.

→ **First optimize our two best models: Neural Network Regressor and Linear Regresssor**

+ We took out JAMB scores with more than 250 because the weights that affected the higher scores was greater and increased MSE significantly
+ Cross Validation was used to compare different models, in this case we made the cv 5

Overall, this lead to to reduce our Neural Network Regressor and Linear Regressor MSE scores significantly from 1592.72, 1588.40 to 1279.3 and 1293.6 respectively

Results

Attendance Rate,' 'Parent Involvement,' and 'Assignments Completed' best key predictors of JAMB scores!

Data Preprocessing: Removing outliers (JAMB Scores above 250) improved model performance, especially for the Neural Network model.

Model Selection: The optimized Linear Regression model is preferred due to its lower MSE and better generalization, offering a balance of performance and interpretability.

Improvements: Enhancing model accuracy by collecting more comprehensive data, including lifestyle, family dynamics, and past academic performance.

Conclusions

Future Implications:

→ **Save schools money:** Knowing what resources schools need in order to allocate funding in those sectors that most contribute to the success of the student

→ **School policy creation:** Attendance and parent involvement is one of the predictive factors that contribute to a student's score, so data can advocate for the creation of school policy that can encourage attendance and parental engagement

→ **Creating quality teachers**: Investing in high quality teachers is key for the success of student's JAMB scores. If schools were to give teachers fair pay and accessibility to materials, teachers can be incentivised more fairly to engage more with students

Real World Ready?:

- **Further testing: Not yet,** possible other factors to consider is students have different lifestyles, family dynamics, technological limitations→ we need more data points


Thank You!