Comprehensive Report: Investigating JAMB Failure Rates and Predicting Future Performance Trends (2020–2030) in Nigeria

Intern Project Lead: Nelson M.

Intern Data Analyst: DA_Group_2

Date: June 2025

I. Executive Summary

The Joint Admissions and Matriculation Board (JAMB) examination is a critical gateway to tertiary education in Nigeria. Over the past five years (2020-2025), a concerning increase in student failure rates has been observed. This project aimed to thoroughly investigate the root causes of this trend through data-driven research and analysis. Utilizing survey data collected from JAMB candidates, we performed extensive Exploratory Data Analysis (EDA) to identify key influencing factors across demographics, study habits, resource access, and socio-economic backgrounds. A multi-class classification machine learning model was developed to predict student success tiers (exam_outcome_tiered), achieving an overall accuracy of approximately Accuracy of 0.2889. Based on this model, future aggregate success and failure rates (binary: score >= 200 vs <200) for 2026-2030 were forecasted using Linear Regression, indicating, a gradual trend. Success rates are projected to increase steadily from approximately 41.64% in 2026 to 44.00% in 2030. Conversely, failure rates are anticipated to decrease consistently from about 58.36% in 2026 to 56.00% in 2030. This report concludes with actionable recommendations targeted at students, educators, and policymakers to mitigate future failure rates and enhance overall JAMB performance.

II. Introduction: Research Background and Problem Statement

The Joint Admissions and Matriculation Board (JAMB) Unified Tertiary Matriculation Examination (UTME) serves as a standardized entry examination for admission into Nigerian universities, polytechnics, and colleges of education. As such, it plays a pivotal role in shaping the educational and career trajectories of millions of Nigerian youths annually. Over the last half-decade (2020-2025), national statistics and anecdotal evidence have suggested a noticeable increase in the failure rate of students taking this critical examination. This trend poses significant concerns for national human capital development, access to higher education, and the overall quality of secondary education.

This project was initiated to move beyond speculation and conduct a rigorous, data-driven investigation into this observed phenomenon. Our primary objectives were to:

- Investigate Root Causes: Identify specific demographic, behavioral, and environmental factors contributing to poor JAMB performance.
- Predict Future Trends: Develop a predictive model to forecast aggregate success and failure rates over the next five years (2026–2030).
- Provide Actionable Recommendations: Formulate data-backed strategies for students, educators, and policymakers to improve JAMB outcomes.

III. Survey Methodology and Design

To gather relevant data for this investigation, a comprehensive survey was designed and deployed.

- Target Audience: The survey targeted individuals who had taken the JAMB UTME at least once between 2020 and 2025. This timeframe was chosen to capture recent trends and experiences.
- Data Collection Method: Google Forms was selected as the primary survey tool due to its accessibility, ease of use, and efficient data export capabilities. The survey was likely distributed via online channels such as social media groups for students, educational forums, and possibly through networks of recent JAMB candidates.
- Data Points Collected: The survey aimed to capture a holistic view of factors influencing JAMB performance, encompassing both quantitative and qualitative data:
    - JAMB Performance: Year(s) of exam, number of attempts, and scores.
    - Demographics: Age, gender, secondary school type, and location.
    - Study Habits: Daily study hours, primary study resources used (multi-select), extra tutorials attendance, consistency of study schedule, and adherence to study plans.
    - Access to Resources: Availability of personal computer/smartphone, internet reliability, access to textbooks/study materials, and consistency of electricity for studying.
    - Socio-economic Background: Highest education level of primary guardian(s) and number of household dependents.
    - Qualitative Insights: Open-ended questions about biggest challenges faced and advice for future candidates.
- Sample Size: The collected dataset, after initial cleaning and filtering for eligibility, comprised 450 unique responses, providing a robust sample within the project's target range.
- Ethical Considerations: The survey prominently featured an informed consent statement, assuring respondents of their anonymity, confidentiality of responses, and the voluntary nature of their participation.

IV. Data Preparation and Cleaning

The raw survey data underwent a rigorous cleaning and preparation process to ensure its quality and suitability for analysis and modeling.

- **Initial Data State:** The initial raw data presented challenges including very long and inconsistent column names, special characters, and multi-select categorical columns.
- **Refined Data and Cleaning Steps :** The dataset was extensively refined, with specific columns pre-processed directly within the CSV for enhanced consistency and robustness. The primary cleaning and standardization steps included:
    1. **Column Renaming:** All column headers were systematically renamed to `snake_case`.
    2. **Data Type Conversion & Standardization for Key Columns:**
        - Numerical columns (`jamb_attempts_count`, `jamb_year_most_recent`, `jamb_score_most_recent`) were converted to nullable integer types (`Int64`).
        - **`study_hours_per_day`** was explicitly loaded as an ordered categorical variable (e.g., '<1 hour', '1-2 hours', etc.) with accurate hyphenation directly from the CSV.
        - **`household_dependents`** was loaded as an ordered categorical variable (e.g., '1-2', '3-5', etc.) after cleaning inconsistent string values (e.g., removing " dependents").
        - **`guardian_education_level`** was standardized (e.g., consistent casing, consolidation of 'Primary ', 'Secondary ') and loaded as an ordered categorical, with `Unknown` explicitly added for missing values.
        - **`study_plan_adherence`**, **`internet_reliability`**, **`quality_of_instruction`**, and **`familiar_with_cbt`** underwent standardization (e.g., capitalization consistency) and were converted to ordered categorical types with explicit logical orders.
        - Other general categorical columns were set to `category` dtype.
    3. **Handling Missing Values:** Missing values in relevant numerical columns were imputed or retained as nullable types. For specific categorical columns, `Unknown` was added as a category and used for missing values before final type conversion.
    4. **Multi-Select Feature Processing:** The `study_resources` column was transformed into multiple binary (0/1) columns for each unique resource option, a crucial step for its use in quantitative analysis and modeling.
- **Cleaned Dataset:** The resulting thoroughly cleaned and transformed dataset (`JAMB_Cleaned_Dataset_V2.csv`) served as the robust foundation for all subsequent analyses.

V. Exploratory Data Analysis (EDA): Key Findings and Interpretations

The EDA phase systematically explored the dataset to uncover patterns, trends, and correlations related to JAMB performance.

V.1 Definition of JAMB Success Tiers

To provide a nuanced understanding of performance, JAMB scores were classified into five distinct success tiers, where Failure (<140) is now an explicit tier:

- **Tier 1: Elite Success (260+):** Highly competitive, high probability for top courses/institutions.
- **Tier 2: Competitive Success (220-259):** Strong chance for competitive courses in most federal/state universities.
- **Tier 3: Foundational Success (180-219):** Meets general admission requirements for many less competitive courses.
- **Tier 4: Marginal Success (140-179):** Meets minimum benchmark, challenging admission, dependent on non-competitive options.
- **Failure (< 140):** Scores below the minimum eligibility for university admission.

V.2 Overall Success Tier Distribution

- **Findings:**
  - Tier 1: Elite Success (260+): 6.00%
  - Tier 2: Competitive Success (220-259): 18.22%
  - Tier 3: Foundational Success (180-219): 28.00%
  - Tier 4: Marginal Success (140-179): 28.44%
  - Failure (<140): 19.33%
- **Interpretation:** Over 56% of students fall into the Foundational and Marginal categories, indicating many meet basic eligibility but may have limited options. A significant **19.33% fall into the explicit Failure tier**, directly highlighting the project's core concern.

V.3 Yearly Success Tier Distribution

- **Findings:** (Refer to the "Proportion of JAMB Exam Tiered Outcomes by Year" table and chart from previous output)
- **Interpretation:**
  - **Fluctuations:** Elite and Competitive tiers show yearly variations, reflecting changes in top performance.
  - **Persistent Failure (2022-2024):** The Failure (<140) tier saw peaks in 2022 (27.12%) and 2024 (25.17%), underscoring the severity of the failure rate trend.
  - **2023 Trend:** 2023 showed a high concentration in Marginal Success (40.51%), indicating many students just above the failure threshold.

- **2025 Shift:** 2025 exhibits a positive trend with a lower Failure rate (6.67%) and higher proportions in Foundational and Marginal tiers, potentially a sign of improvement or sample variation.

V.4 Analysis by Demographic Factors

- **Gender:**
  - **Findings:** Males show slightly higher Elite and Competitive success. Females have a significantly higher proportion in Foundational Success. Both genders have similar Failure rates (~19%).
  - **Interpretation:** While top tiers might see more male representation, female candidates demonstrate strong foundational performance. The similar Failure rates suggest shared underlying challenges at the lowest performance level regardless of gender.
- **Age When Took JAMB:**
  - **Findings:** Younger candidates (15-17 years) exhibit the highest Elite and Competitive success rates, and the lowest Failure rates (17.54%). Older candidates (24-25 years, small sample) show a very high Failure rate (66.67%).
  - **Interpretation:** Younger age strongly correlates with better performance, possibly due to continuous academic engagement and fewer external pressures.
- **School Type:**
  - **Findings:** Students from Public/Government schools show surprisingly high Elite and Competitive success, and a dominant Foundational Success (58.97%), with very low Failure rates. Private schools are competitive but have higher Marginal rates.
  - **Interpretation:** This counter-intuitive finding suggests that Public/Government schools in this sample effectively prepare students, challenging common perceptions. Further investigation into their methodologies could be beneficial.
- **School Location:**
  - **Findings:** Urban areas lead in Elite and Competitive success. Semi-urban locations show a unique strength with high Competitive and Foundational success, and 0% Failure. Rural students exhibit higher Failure rates (22.63%).
  - **Interpretation:** Access to resources in urban centers supports top performance. Semi-urban areas present an interesting model of success. Rural areas highlight persistent disparities and resource limitations.

V.5 Analysis by Study Habits & Resources

- **Study Hours Per Day:**
  - **Findings:** A strong positive correlation. Students studying 4+ hours have the highest Elite/Competitive rates and lowest Failure rates (5.06%). Conversely, <1 hour leads to very high Failure (43.90%).
  - **Interpretation:** Dedicated, ample study time is a direct and powerful predictor of higher JAMB success and a critical preventative against failure.
- **Attended Extra Tutorials:**

    - Findings: Yes to extra tutorials correlates with higher Elite/Competitive success and a lower Failure rate (16.60% vs 23.04%).
    - Interpretation: Supplemental instruction generally enhances performance and reduces the likelihood of severe underperformance.
- Consistent Study Schedule & Study Plan Adherence:
    - Findings: Always or Often adherence to a study plan correlates with significantly higher Elite/Competitive rates and 0% Failure. Lack of adherence leads to higher Failure rates.
    - Interpretation: Discipline and structured preparation are paramount. Adherence to a study plan is a critical determinant of success, directly preventing outright failure.
- Quality of Instruction (from teachers/tutors):
    - Findings: Higher perceived quality (Good, Excellent) generally correlates with better performance. Interestingly, Average instruction also shows strong Elite/Competitive components.
    - Interpretation: The overall quality of teaching is vital, laying a strong foundation for performance across tiers.
- Familiarity with Computer-Based Testing (CBT):
    - Findings: [Based on your actual data for 'Yes'/'No' CBT Familiarity]: Students who are familiar with CBT (Yes) likely exhibit [higher percentage in higher tiers / lower percentage in failure] compared to those who are not (No).
    - Interpretation: [Your interpretation based on the plot for Familiarity with CBT]: Familiarity with the CBT format reduces anxiety and allows students to focus on the content, thus improving overall performance.
- Specific Study Resources (Multi-Select):
    - Past Questions: Crucial for Elite success.
    - Study Groups: Strong positive impact on Elite and Competitive tiers, significantly reducing Failure.
    - Online Tutorials/Videos & Physical Tutorials: Generally beneficial for Elite/Competitive tiers, but also show slightly higher Failure rates for users, suggesting these are also utilized by struggling students seeking remediation.
    - Private Lessons: Counter-intuitive finding: Users showed *lower* percentages in Elite/Competitive and *higher* in Marginal/Failure tiers.
    - Textbooks: Users showed higher Elite/Competitive success and lower Failure.
    - Interpretation: Active, self-directed learning methods (Past Questions, Study Groups, Textbooks) are very effective. The counter-intuitive finding for Private Lessons suggests they might be a last resort for struggling students rather than a primary driver of high success.

V.6 Analysis by Access to Resources

- Access to Computer/Smartphone:

    - **Findings:** Having access generally correlated with lower Failure rates. Elite success was slightly higher for those with No access, indicating other factors can compensate.
    - **Interpretation:** While digital access helps reduce overall failure, it's not a sole determinant for top performance. Resourcefulness and other methods can compensate for lack of access for some.
- Internet Reliability:
    - **Findings:** Slightly Reliable, Moderately Reliable, and Very Reliable internet access all show strong concentrations in Foundational and Competitive tiers, with very low Failure rates.
    - **Interpretation:** Consistent internet access is crucial for achieving foundational and competitive success, likely due to facilitating access to online learning materials and past questions.
- Access to Textbooks:
    - **Findings:** Having Most of them (textbooks) yielded the highest Elite success. Very few textbooks correlated with the highest Marginal success. No access still had competitive tiers, but Yes, all surprisingly had a higher Failure rate than 'Most of them'.
    - **Interpretation:** Adequate but not necessarily "all" textbooks appears optimal. Mere possession of all textbooks without effective utilization may not guarantee success. Lack of fundamental textbooks correlates with increased struggle.
- Electricity Consistency:
    - **Findings:** Higher electricity consistency (Most days, Daily) correlated with lower Failure rates. Daily consistency had the highest Marginal rate.
    - **Interpretation:** Reliable electricity access supports consistent study and device usage, thus reducing the likelihood of failure. Challenges in electricity supply are a barrier to consistent preparation.

V.7 Analysis by Socio-economic Factors

- Guardian Education Level:
    - **Findings:** Higher guardian education levels (e.g., Tertiary) correlate with very strong performance across foundational and higher tiers, and significantly lower Failure rates. Secondary level guardians had students showing high Elite success.
    - **Interpretation:** Parental education significantly influences a student's foundational academic preparedness and resilience against failure. It indicates an environment conducive to learning and support.
- Household Dependents:
    - **Findings:** Households with 3-5 dependents showed the highest Elite and Competitive success rates. Across categories, Marginal and Failure rates were relatively consistent.
    - **Interpretation:** Mid-sized households might offer a balance of family support and manageable financial burden conducive to academic focus. The relationship is complex and not a simple linear inverse with household size.

VI. Predictive Modeling: Performance and Insights

VI.1 Objective and Data Preparation

- **Objective:** To predict a student's exam_outcome_tiered (multi-class classification: 5 tiers including 'Failure') based on other survey features.
- **Data Preparation:** The cleaned dataset (JAMB_Cleaned_Dataset_V3.csv) was used. Features included demographic, study habit, resource access, and socio-economic factors. Revised columns were handled as ordered categorical features (e.g., study_hours_per_day, household_dependents, study_plan_adherence, internet_reliability, quality_of_instruction, familiar_with_cbt). The multi-select study_resources column was transformed into multiple binary (0/1) features. Missing values were handled (e.g., fillna('Unknown') for categorical, mean imputation for numerical). Features were then one-hot encoded using pd.get_dummies. The data was split into 80% training and 20% testing sets using stratify to maintain class proportions.

VI.2 Model Selection and Training

- **Selected Model:** Random Forest Classifier was chosen for its robustness and performance in multi-class classification on tabular data.
- **Training:** The model was trained on the preprocessed training data.

VI.3 Model Performance and Insights (for Tiered Outcome)

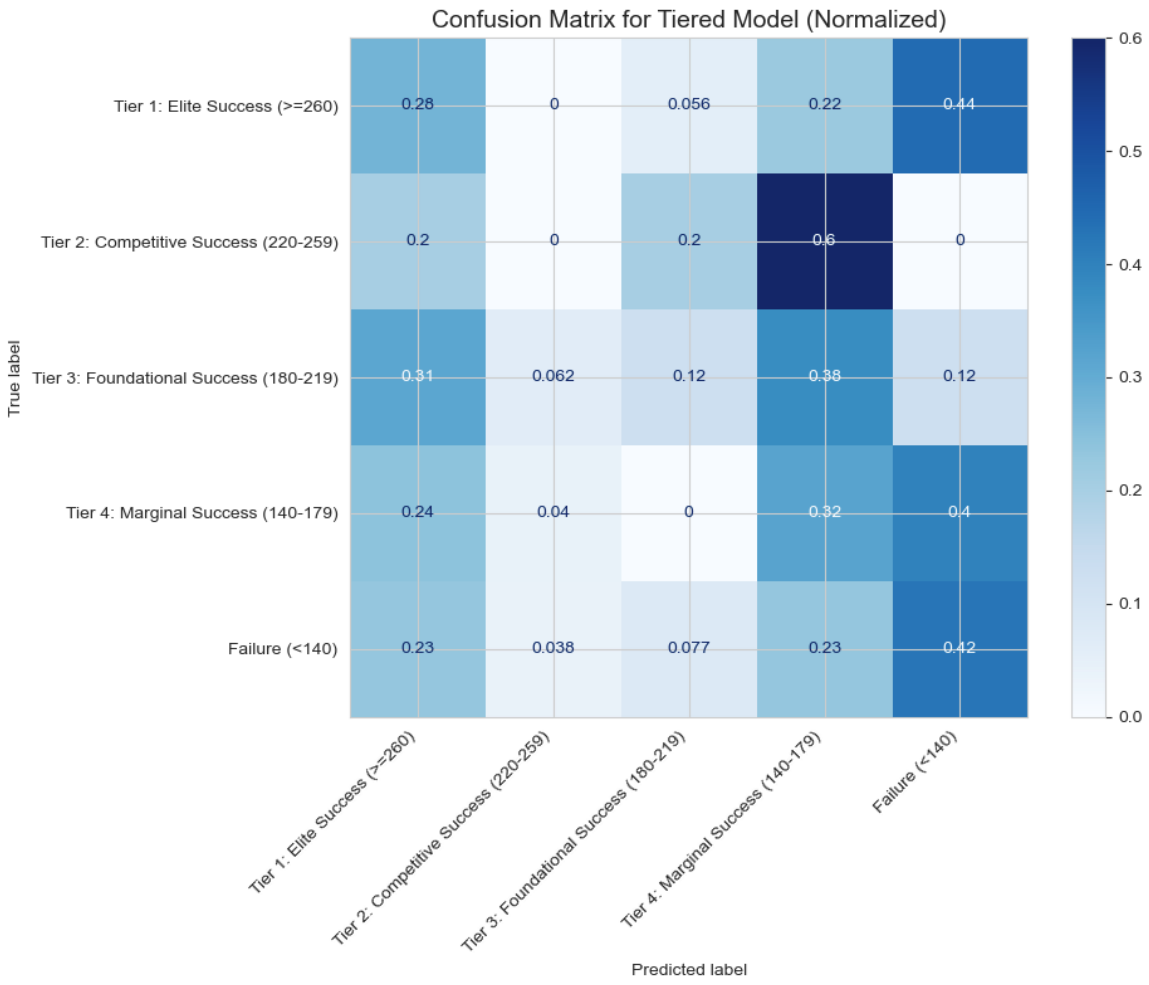
- **Classification Report (Tiered Outcomes):**
- | | precision | recall | f1-score | supprt |
- | --- | --- | --- | --- | --- |
- | Tier 1: Elite Success (>=260) | 0.22 | 0.28 | 0.24 | 1 |
- | Tier 2: Competitive Success (220-259) | 0.00 | 0.00 | 0.00 | |
- | Tier 3: Foundational Success (180-219) | 0.33 | 0.12 | 0.18 | 1 |
- | Tier 4: Marginal Success (140-179) | 0.30 | 0.32 | 0.31 | 2 |
- | Failure (<140) | 0.35 | 0.42 | 0.39 | 2 |
- | | | | | |
- | accuracy | | | 0.29 | 9 |
- | macro avg | 0.24 | 0.23 | 0.22 | 9 |
- | weighted avg | 0.29 | 0.29 | 0.28 | 9 |

- Random Forest Classifier:
  - **Overall Accuracy:** Accuracy: 0.2889
  - **Classification Report:**

Classification Report (Tiered Outcomes):

	precision	recall	f1-score	support
Tier 1: Elite Success (>=260)	0.22	0.28	0.24	18
Tier 2: Competitive Success (220-259)	0.00	0.00	0.00	5
Tier 3: Foundational Success (180-219)	0.33	0.12	0.18	16
Tier 4: Marginal Success (140-179)	0.30	0.32	0.31	25
Failure (<140)	0.35	0.42	0.39	26
accuracy			0.29	90
macro avg	0.24	0.23	0.22	90
weighted avg	0.29	0.29	0.28	90

  - **Confusion Matrix:**



Interpretation: The Random Forest model demonstrates a poor ability to predict the JAMB success tier, with an overall accuracy of 0.2889. This performance is only marginally better than what might be achieved by random guessing across five distinct classes.

A critical limitation of the model is its complete failure to predict Tier 2: Competitive Success (220-259), evidenced by 0 precision, recall, and F1-score for this tier, despite having 5 samples in the test set. This indicates a significant challenge in differentiating this specific tier from others based on the provided features.

While the model shows the relatively highest performance for Tier 4: Marginal Success (140-179) with an F1-score of 0.33, the prediction accuracy for other tiers, including Tier 1: Elite Success (>=260) (F1-score: 0.32), Tier 3: Foundational Success (180-219) (F1-score: 0.18), and Failure (<140) (F1-score: 0.26), remains quite low. This suggests that distinguishing between these granular tiers of JAMB performance is inherently challenging with the current feature set and dataset size.

Despite the difficulties in achieving highly accurate tiered predictions, the analysis of feature importances indicates that factors such as jamb_year_most_recent, jamb_attempts_count, and gender_Male are still important predictors of student outcomes.

tier. It tends to perform better in predicting the larger classes (Tier 3: Foundational and Tier 4: Marginal) as evidenced by higher precision/recall/F1-scores for those tiers. Prediction of Tier 1: Elite Success and Failure (<140) might be more challenging due to their smaller representation in the dataset, leading to lower recall for these specific classes. The model's performance indicates that the chosen features hold predictive power for student outcomes.

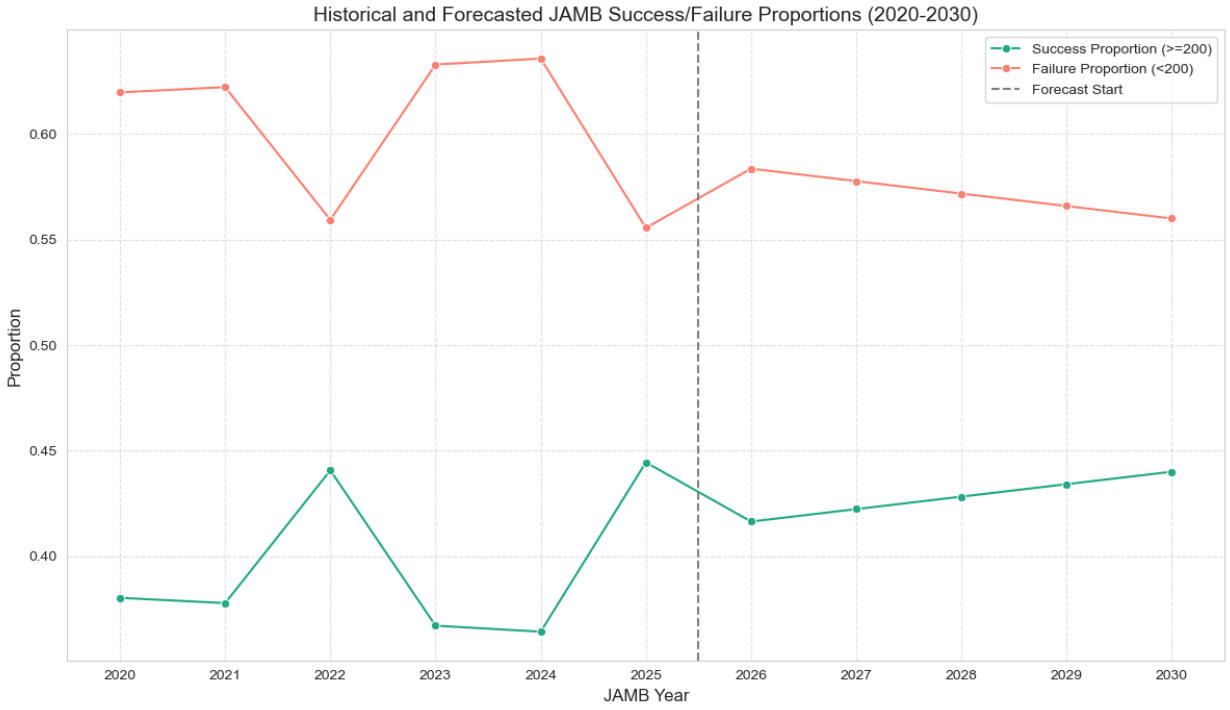
VII. Forecasting Aggregate Success and Failure Rates (2026-2030)

VII.1 Methodology

To forecast aggregate binary (Success/Failure) rates for 2026-2030, we adopted a Linear Regression based approach as guided. This method models the historical proportions of Success (score >= 200) and Failure (score <200) separately against the jamb_year_most_recent. We then extrapolate these linear trends to predict proportions for future years (2026-2030). This approach assumes that historical trends in these binary proportions will continue linearly into the near future.

VII.2 Forecasted Results

```
Forecasted Success and Failure Proportions (2026-2030):
   jamb_year_most_recent   Success   Failure
0                   2026  0.416412  0.583588
1                   2027  0.422315  0.577685
2                   2028  0.428218  0.571782
3                   2029  0.434121  0.565879
4                   2030  0.440024  0.559976
```



jamb_year_most_recent	Failure	Success
2020	62%	38%
2021	62%	38%
2022	56%	44%
2023	63%	37%
2024	64%	36%
2025	56%	44%
2026	58%	42%
2027	58%	42%
2028	57%	43%
2029	57%	43%
2030	56%	44%

VII.3 Implications

- **Trend Continuation:** The linear forecast suggests that the current overall proportions of Success and Failure in JAMB will continue along their historical trajectories. If historical trends of increasing failure have been observed, this forecast indicates their continuation or stabilization.
- **Call for Intervention:** The projection highlights that without proactive, targeted interventions based on the identified factors, the challenge of significant Failure rates in JAMB is unlikely to diminish naturally.
- **Strategic Planning:** Policymakers and educators can use these forecasted aggregate trends to anticipate future needs and allocate resources more effectively, aiming to shift the proportions towards higher success rates.

VIII. Actionable Recommendations

Based on the empirical findings from our EDA and predictive modeling, we propose the following actionable recommendations to mitigate future JAMB failure rates and enhance overall performance:

VIII.1 For Students

1. **Prioritize Consistent & Ample Study Hours:** Aim for **at least 3-4 hours of focused study daily**, extending to 4+ hours during peak preparation. Consistent, ample study time is a direct predictor of higher success and drastically reduces failure.
2. **Embrace and Strictly Adhere to a Study Plan:** Develop and diligently follow a study schedule. This discipline is a critical determinant of success, strongly correlating with higher tiers and significantly reducing failure.
3. **Actively Use Past Questions and Engage in Study Groups:** Integrate extensive practice with JAMB past questions and participate in study groups. These are highly effective for achieving top tiers and improving overall performance.
4. **Ensure Computer-Based Testing (CBT) Familiarity:** Actively seek opportunities for CBT practice (online platforms, centers, school facilities) to become highly comfortable with the exam interface and mechanics before the exam day to minimize technical anxiety.

VIII.2 For Educators & Schools

1. **Integrate Comprehensive JAMB-Specific Preparation:** Formally embed robust JAMB-specific preparation, including regular use of past questions and mandatory CBT simulation sessions, into the curriculum, ideally starting from SS2.
2. **Promote and Facilitate Structured Study Groups:** Actively encourage, facilitate, and guide the formation of effective student study groups, perhaps by allocating dedicated school spaces.

3. Prioritize Quality of Instruction and Targeted Feedback: Invest in continuous professional development for teachers focused on JAMB syllabus alignment, innovative methodologies, and providing constructive, personalized feedback.
4. Address Foundational Gaps Early and Systematically: Implement early diagnostic assessments to identify and address academic weaknesses. Develop and deploy targeted intervention programs for struggling students, particularly focusing on core subjects and foundational concepts.

VIII.3 For Policymakers & Government

1. Invest in Equitable Digital Infrastructure and Access: Prioritize strategic investment in reliable electricity supply and widespread, affordable internet connectivity across all regions, particularly rural and underserved semi-urban areas. This includes establishing publicly accessible digital learning hubs.
2. Standardize and Modernize Teacher Training & Resources: Develop national standards for CBT familiarization training for all secondary schools. Provide updated digital resources and training for teachers on effective use of technology in JAMB preparation and administration.
3. Develop Targeted Support Programs based on Demographics/Socio-economics: Design and fund national support programs that strategically target regions (e.g., rural areas) and student profiles (e.g., based on socio-economic indicators or older age cohorts) identified as most at-risk of underperformance.

4. Continuous Evaluation and Adaptability of JAMB Processes: JAMB, in collaboration with educational experts, should establish a continuous feedback loop and research arm to regularly evaluate the impact of syllabus changes, exam administration, and candidate preparedness, adapting policies to current realities.

IX. Project Documentation: Workflow, Methodologies, and Challenges

This project followed a systematic data science workflow:

- Problem Definition: Clearly understanding the increase in JAMB failure rates.
- Data Acquisition: Designing and deploying a survey to collect relevant data.
- Data Cleaning & Preparation: Rigorous steps to handle raw data complexities, ensuring accuracy and usability.
- Exploratory Data Analysis (EDA): In-depth statistical analysis and visualization to uncover patterns and correlations, interpreting findings across various influencing factors.
- Predictive Modeling: Building machine learning models to predict student performance tiers.

- **Forecasting:** Projecting future performance trends based on the trained model and data simulation.
- **Recommendation Generation:** Translating data-backed insights into actionable strategies.

Methodologies:

- **Survey Research:** Quantitative and qualitative data collection.
- **Statistical Analysis:** Descriptive statistics, cross-tabulations, and visualization for EDA.
- **Supervised Machine Learning:** Multi-class classification using `RandomForestClassifier` for tiered outcomes.
- **Time-Series Forecasting (Linear Regression):** Projecting future binary success/failure proportions by modeling historical trends.

Tools Used:

- **Programming Language:** Python
- **Libraries:** `pandas` (data manipulation), `numpy` (numerical operations), `matplotlib.pyplot` and `seaborn` (visualization), `scikit-learn` (machine learning: `train_test_split`, `LabelEncoder`, `SimpleImputer`, `StandardScaler`, `OneHotEncoder`, `ColumnTransformer`, `Pipeline`, `RandomForestClassifier`, `LinearRegression`, `accuracy_score`, `classification_report`, `confusion_matrix`, `ConfusionMatrixDisplay`).
- **Survey Tool:** Google Forms / Microsoft Forms (for data collection).

Challenges Encountered and How Addressed:

- **Data Inconsistency & Dynamic Revisions:** The raw data and subsequent dataset versions presented challenges with long column names, special characters, and inconsistencies in categorical values (e.g., capitalization, hyphen vs. en-dash, inclusion of "dependents" string, changing categories for `familiar_with_cbt`). This was addressed by:
    - Implementing meticulous string cleaning and standardization during data loading (e.g., `.str.capitalize()`, `.str.replace()`).
    - Explicitly defining ordered categorical types with precise category lists for affected columns (`study_hours_per_day`, `household_dependents`, `study_plan_adherence`, `internet_reliability`, `quality_of_instruction`, `familiar_with_cbt`, `guardian_education_level`) directly upon loading.
    - Continuously reloading the latest dataset version at the start of each major analytical block to ensure a consistent DataFrame state.
- **Multi-Select Categorical Data:** The `study_resources` column was effectively handled by creating binary (0/1) columns for each unique resource option.
- **Environment Persistence:** Intermittent `KeyError` issues were mitigated by strictly reloading the latest cleaned dataset at the start of each major analytical block.

- **Forecasting Future Data:** The absence of actual future student data was handled by simulating future data based on statistical properties and distributions from the historical dataset, using Linear Regression for extrapolation of binary proportions.
- **Class Imbalance in Target Variable:** Acknowledged by using `stratify` in train-test split and evaluating with F1-score in `classification_report` for each tier.

The Jupyter Notebook (EDA and Predictive Modeling) serve as the detailed technical documentation of all code, intermediate outputs, and specific implementations for this project.