# Project Documentation: JAMB Performance Analysis and Prediction

# 1. Workflow and Methodology

The project workflow began with the design and distribution of a survey using Google Forms. Survey questions were crafted to capture demographic, academic, and socio-economic factors that could affect JAMB performance. The form was then shared across various student platforms to gather responses.  
Following data collection, the steps below were implemented:

- Data Acquisition & Cleaning: Removal of duplicates, correction of categorical values, and handling of missing data.

- Exploratory Data Analysis (EDA): Utilized Seaborn and Matplotlib for visualizing distributions and relationships.

- Feature Engineering: Performed one-hot encoding and binary target encoding for categorical features.

- Predictive Modeling: Built a Logistic Regression classifier to predict exam success/failure.

- Forecasting: Estimated future pass/fail rates using model probabilities and simulated enrollment trends.

- Recommendations: Proposed actionable steps for students, educators, and policymakers based on findings.

# 2. Tools and Technologies Used

|  |  |
| --- | --- |
| Tool/Library | Purpose |
| Python | Core programming language |
| Pandas | Data manipulation and preprocessing |
| NumPy | Numerical operations |
| Matplotlib & Seaborn | Data visualization and EDA |
| Scikit-learn | Model training, evaluation, and encoding |
| Jupyter Notebook / VS Code | Coding environment and IDE |
| Excel (initially) | Dataset preview and survey structure |
| Google Forms | Survey creation and data collection |

# 3. Key Decisions Made

- Dropped problematic features (e.g., age group strings) when they caused encoding issues.

- Used one-hot encoding to avoid ordinal misrepresentation.

- Selected Logistic Regression for its simplicity and effectiveness in binary classification.

- Used a fixed random state for reproducibility.

- Refined feature selection iteratively to improve performance.

# 4. Challenges Encountered and Solutions

- Gathering survey responses was difficult and time-consuming, requiring persistent effort and outreach.

- Predictive modeling was challenging due to complex syntax, constant debugging, and understanding machine learning workflows.

- Faced conversion errors due to string-based categorical features like '19–21'. Resolved by one-hot encoding or dropping columns.

- Encountered difficulties installing necessary Python packages. Resolved by using pip and ensuring correct environments.

- Model tuning and interpretation required multiple iterations and a steep learning curve.