DETAILED PROJECT DOCUMENTATION

1. **PROJECT OBJECTIVE**

   To investigate the increasing failure rate of JAMB candidates from 2020 to 2025 and forecast future performance using supervised machine learning. The project aimed at:

   - Analyze student preparation behavior.
   - Identify root causes of poor performance.
   - Develop a predictive model for exam success.
   - Forecast pass rate trends for strategic planning.

2. **PROJECT PHASES**

   Survey Design and Administration

   - Developed a well-structured Google Form targeting post-JAMB students.
   - Questions designed to capture both quantifiable data and subjective insights.
   - Ensured anonymity and optional sections for open-ended responses.

   Data Collection and Cleaning

   - Downloaded data in Excel format.
   - Removed duplicates and inconsistent text inputs.
   - Converted categorical variables to numerical where needed (e.g. age ranges to midpoints).

   Exploratory Data Analysis (EDA)

   - Used histograms, bar chat, and line chart to discover trends.
   - Investigated how each variable correlates with scores and success.
   - Used scatter plots with color coding to highlight interactions.

   Model Development

   - Selected Random Forest due to its capability to handle mixed data types.
   - One-hot encoded categorical features.
   - Split dataset using an 80/20 train-test split.
   - Evaluated using accuracy, precision, recall, and confusion matrix.

   Forecasting

   - Aggregated yearly pass rates.
   - Built a simple linear regression model to project scores up to 2030.

3. TOOLS AND TECHNOLOGIES USED

- **Google Forms**: Data collection
- **Python**: Data cleaning, EDA, modeling
- **Libraries**: pandas, matplotlib, seaborn, scikit-learn
- **Jupyter Notebook**: Interactive workflow documentation
- **Excel**: Pre-cleaning inspection and formatting

4. KEY DECISIONS MADE

- Score threshold set at 200 for pass/fail classification.
- Used numeric midpoint approximations for ordinal text fields.
- Removed rows with critical missing values to preserve model integrity.

5. CHALLENGES ENCOUNTERED

- Non-numeric survey fields with inconsistent formatting.
- Imbalanced dataset (more passing students).
- Limited responses from certain regions.
- Need to balance simplicity vs. predictive power in feature engineering.

6. RESOLUTIONS

- Created mapping dictionaries for categorical fields.
- Applied stratified sampling and test splits to ensure generalizability.
- Noted geographical gaps and suggested expanded outreach for future surveys.
- Chose interpretable models (like Random Forest) to support stakeholder understanding.