🌟 Project Requirements Document: Data Science & Streamlit Application

🎯 1. Project Overview

The core objective of this project is to develop a comprehensive data analysis and machine learning solution. This solution will utilize a provided dataset and must integrate three key components: **Exploratory Data Analysis (EDA)**, **Machine Learning Model Training** (specifically Logistic Regression and Random Forest), and an **Interactive User Interface** built with Streamlit.

📊 2. Data Requirements & Preprocessing

To ensure a robust and reproducible analysis, the following data handling standards must be met:

- **Primary Source**: Utilize the provided dataset exclusively.
- **Data Quality**: Implement validation for data types and robust strategies for handling missing values.
- **Reproducibility**: All preprocessing steps must be clearly documented for full reproducibility.
- **Summary Generation**: Automatically generate and present key summary statistics and data distributions.

🔍 3. Exploratory Data Analysis (EDA)

The EDA section must provide a deep understanding of the dataset through both textual and visual analysis.

Required Outputs:

Category	Requirement	Details
Dataset Overview	Preview & Structure	Head, shape, and column data types.
Statistics	Summary Statistics	Standard descriptive statistics.
Quality Check	Missing-Value Analysis	Visualization and quantification of missing data.
Target Variable	Class Balance	Visualization of the target variable’ s distribution.

Required Visualizations:

- **Histograms**: At least **5** histograms to show the distribution of key features.
- **Boxplots**: At least **3** boxplots for outlier detection and comparison across categories.
- **Correlation**: **1** comprehensive correlation matrix heatmap.

Interpretation:

- **Insights**: All visualizations must be accompanied by **clear, written insights** that connect the visual evidence to meaningful conclusions.

🧠 4. Machine Learning Requirements

The solution requires the development and rigorous evaluation of two distinct classification models.

Required Models:

1. **Logistic Regression**
2. **Random Forest Classifier**

Training & Evaluation:

- **Methodology**: Utilize a robust method such as **Train/Validation Split** or **Cross-Validation**.
- **Performance Metrics**: The following metrics must be calculated and reported for both models:
    - Accuracy
    - Precision
    - Recall
    - F1 Score
    - ROC AUC (Area Under the Curve)
- **Visualizations**:
    - **ROC Curve** plot for each model.
    - **Confusion Matrix** for each model.

Model Comparison:

- A dedicated section must be included to compare the two models, detailing their **strengths and weaknesses** based on the evaluation metrics and business context.

💻 5. Streamlit UI Requirements

The Streamlit application serves as the interactive front-end for demonstrating the project’ s capabilities.

Core Features:

Feature	Description
Data View	Display `head()`, `describe()`, missing-value summary, and key EDA plots.
Model Training	Allow users to: • Select between Logistic Regression and Random Forest. • Input basic hyperparameters. • Display the full set of performance metrics upon training completion.
Prediction Interface	A form where users can manually input all feature values. The model must return the **Predicted Class** and the **Predicted Probability** (if applicable).

Usability Standards:

- **Instructions**: Provide clear and concise instructions for navigation and use.
- **Layout**: Implement a clean, professional layout, ideally utilizing a **sidebar for navigation**.

⚙️ 6. System & Technical Requirements

The project must adhere to the following technical specifications:

- **Environment**: Python **3.9+**
- **Key Libraries**:
    - `pandas`
    - `numpy`
    - `scikit-learn`
    - `matplotlib`
    - `streamlit`
    - `seaborn` (Optional, but recommended for advanced visualization)

- **Execution**: The application must be runnable via the standard command: `streamlit run app.py`

✅ 7. Deliverables & Acceptance Criteria

Deliverables:

1. **Complete EDA Report** (Code and documentation).
2. **ML Training Code and Results** (Model files, training scripts, and evaluation outputs).
3. **Streamlit UI Application** (Source code).
4. **Final Write-up** (Summarizing the approach, key findings, and recommendations).

Acceptance Criteria:

- **Completeness**: All required plots and performance metrics are present.
- **Functionality**: Both ML models are fully functional and can be trained/evaluated.
- **Stability**: The Streamlit application runs without errors.
- **Interactivity**: Predictions work seamlessly through the interactive interface.
- **Documentation**: All documentation is clear, professional, and complete.