Cellula Technologies

Data Preprocessing

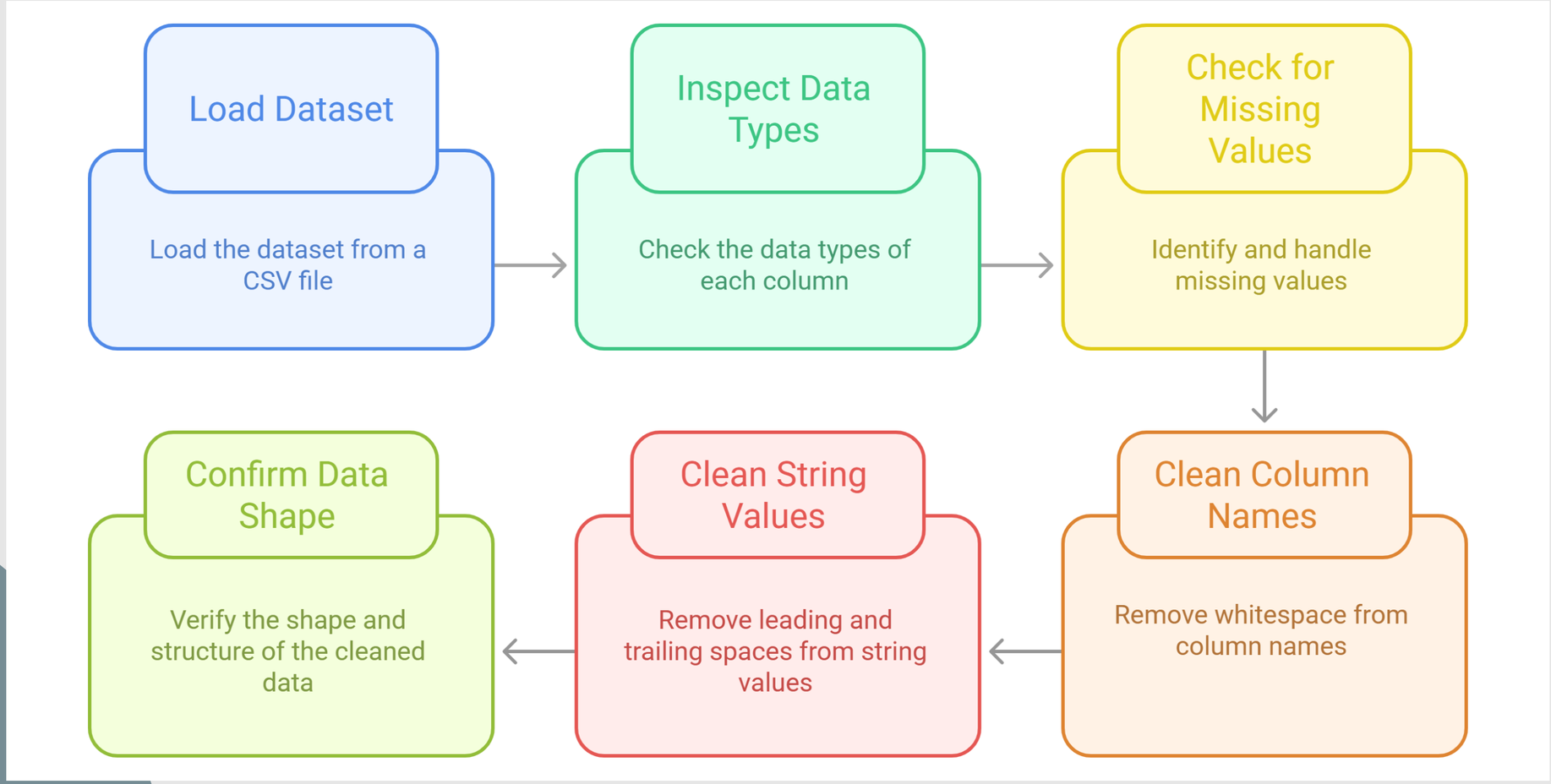
Presentation: Second Task

Group : ML_Group2

Submitted By: Dalia Nasser

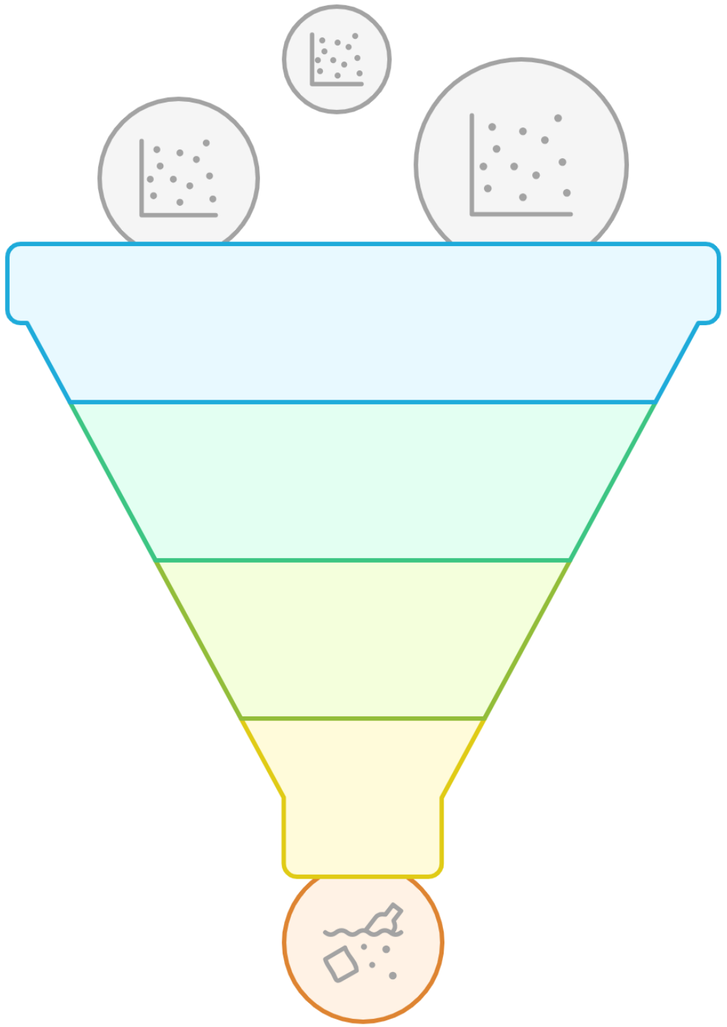
1. Data Preprocessing Sequence

The dataset includes **36,285 complete records** across **17 well-structured columns**, with a mix of numerical and categorical types and no missing values, ensuring it is clean and ready for analysis.



2. Outlier Removal Process

Outliers were removed using the **IQR method** on selected numeric columns, **reducing** the dataset to **20,525 clean records** while maintaining realistic values for **lead time**, **average price**, and **booking nights**, ensuring higher data quality for modeling.



Identify Numeric Columns

Select columns for outlier analysis

Apply IQR Method

Calculate IQR to detect outliers

Remove Outliers

Eliminate rows outside IQR bounds

Re-check Data

Verify data shape and statistics

IQR Method – Interquartile Range

- The IQR method **identifies** and **removes outliers** by keeping only values **within 1.5 times** the interquartile range, ensuring cleaner and more reliable data for analysis.
- The Calculation :

1. Calculate Q1 and Q3:

Q1 = 25th percentile of the data

Q3 = 75th percentile of the data

2. Compute the IQR:

IQR = Q3 − Q1

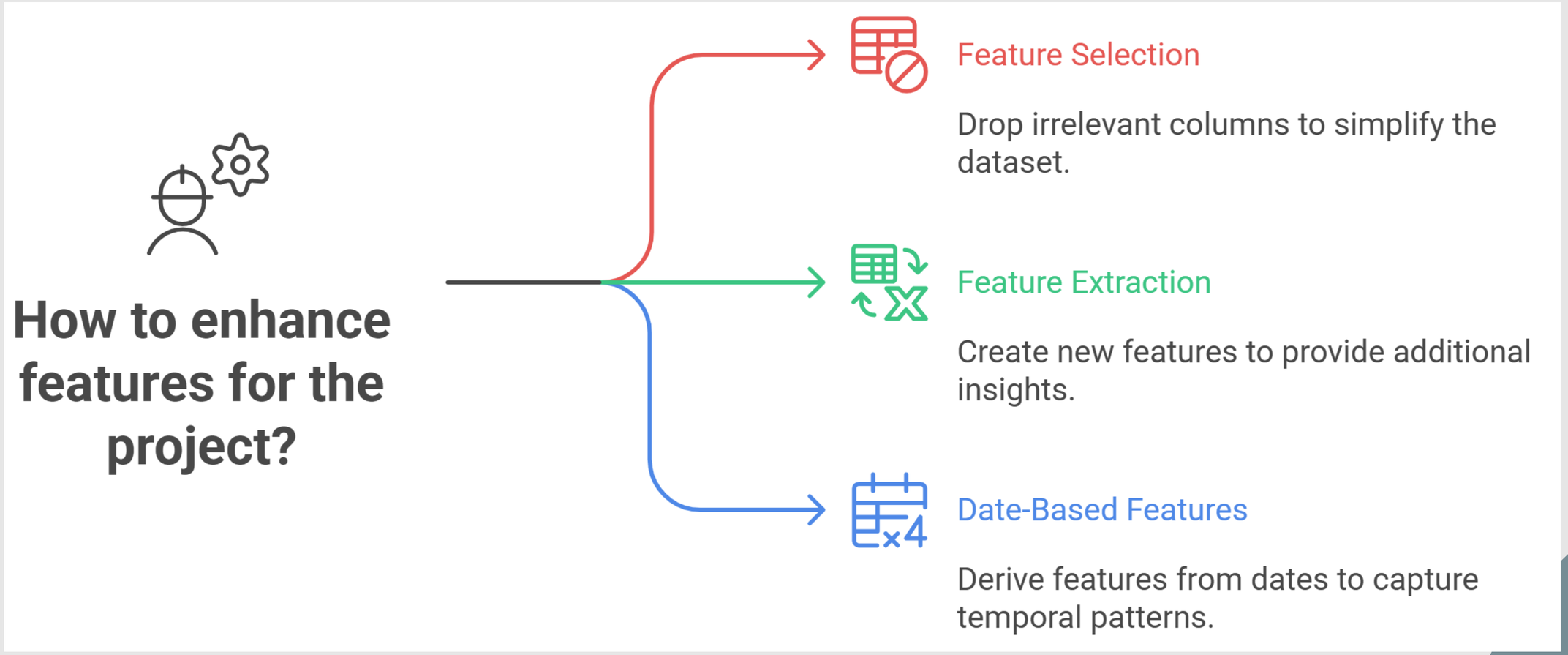
3. Define bounds:

Lower bound = Q1 − 1.5 × IQR

Upper bound = Q3 + 1.5 × IQR

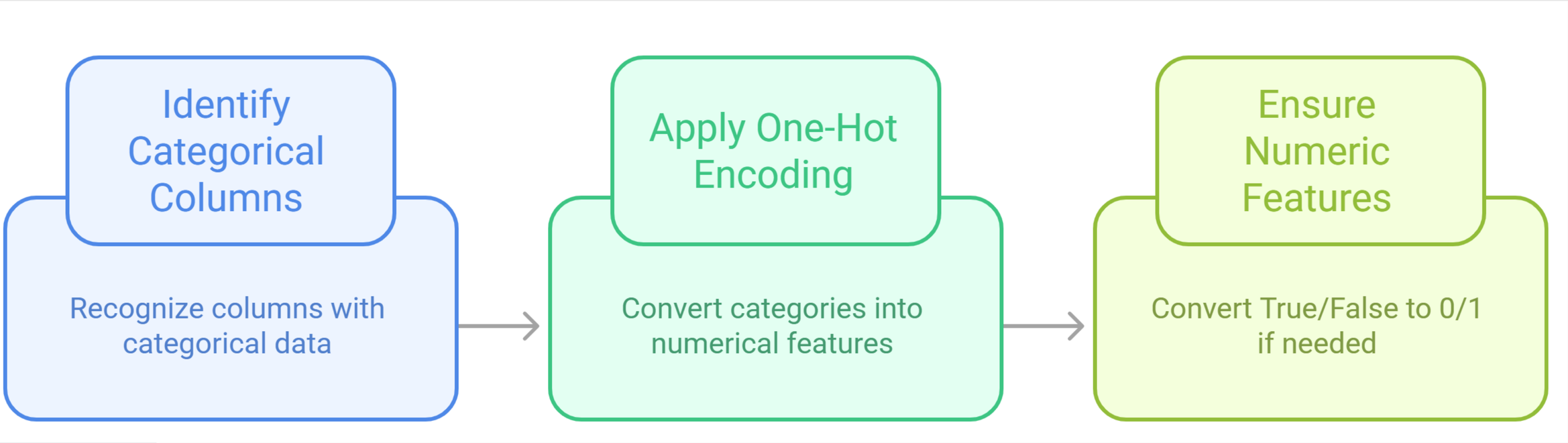
3. Feature Engineering

Feature engineering included dropping irrelevant columns like **Booking_ID and creating new features** such as total_guests by combining number of adults and children to enrich the dataset.



4. Categorical Data Transformation Process

Four categorical columns (**type of meal**, **room type**, **market segment type**, and **booking status)** were transformed using one-hot encoding, converting them into **multiple binary 0/1** columns to make the entire dataset fully numeric and suitable for machine learning models



5. Train-Test Split

The dataset was split into training and testing sets using an **80/20 ratio**, resulting in **16,420 training** samples and **4,105 testing samples**, ensuring balanced class distribution for model evaluation

Define Features and Target

Identify independent and dependent variables

Split Data

Divide data into training and testing sets

Encode Target Variable

Convert target to binary format

Apply Stratified Sampling

Ensure balanced target distribution

6. Modeling & Evaluating

Multiple classification models were evaluated, with **Random Forest achieving the highest test** set accuracy at **86.9%**, closely followed by **XGBoost at 86.6%**, while **Logistic Regression trailed at 80.2%**, demonstrating the advantage of ensemble methods for this prediction task.

Which classification model should be used for booking status prediction?

Random Forest

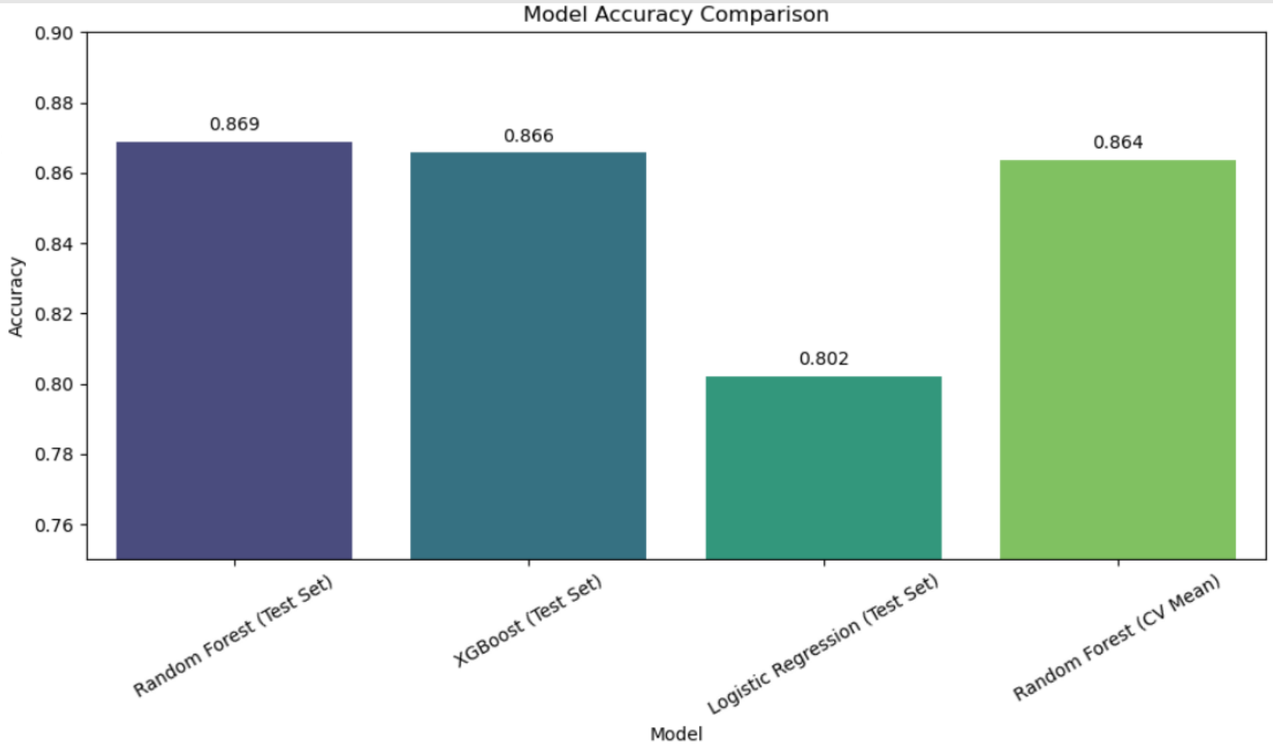
Offers high accuracy and robustness but can be computationally intensive.

Logistic Regression

Simple and interpretable but may not capture complex relationships.

XGBoost

High performance and scalability but requires careful tuning.



Thank You