Project Title: Hotel Booking Cancellation Prediction

Goal / Description:
The goal of this project is to predict whether a hotel booking will be canceled based on customer and reservation details.
Hotel cancellations cause financial loss, reduce room utilization, and make planning difficult.
Using Data Analytics and Machine Learning, we will analyze hotel booking patterns and build models that help hotels identify high-risk cancellations in advance.

We will compare different classification models (Decision Tree, KNN, Logistic Regression, Random Forest, and Neural Network) and select the best-performing model.

Dataset:
Kaggle Hotel Booking Demand Dataset
This dataset contains 119,390 booking records from two hotels (City Hotel & Resort Hotel).
Each record indicates whether the booking was canceled or honored.

What the dataset includes (not limited to):
is_canceled (Target variable: 0 = Not canceled, 1 = Canceled)
lead_time
arrival_date_month, arrival_date_week_number
adults, children, babies
meal
country
market_segment
distribution_channel
is_repeated_guest
previous_cancellations
deposit_type
assigned_room_type
booking_changes
days_in_waiting_list
customer_type
reservation_status_date
total_of_special_requests

The dataset is clean, numeric + categorical, and ready for ML.

Models to Compare (Not Limited To):
We will run multiple classification models to compare performance:

Main Models:
Decision Tree (CART / C4.5 style)
K-Nearest Neighbors (KNN)
Logistic Regression (easy + interpretable)
Random Forest
Neural Network (simple MLP)
These models suit classification tasks and will help us understand which features influence cancellations most.
Method We Are Going to Use (CRISP-DM)
Same format/style as the uploaded file.

1. Business Understanding
Problem:
Hotels face financial loss due to unexpected cancellations. We want to predict cancellation probability and understand key factors that cause guests to cancel.

Stakeholders:
Hotel management, revenue managers, front-desk planners, marketing, finance.

Business Value:
Reduce last-minute cancellations
Better manage room availability
Plan overbooking strategies
Improve revenue forecasting

2. Data Understanding
We will explore dataset structure and patterns:
Key Explorations:
Cancellation rate by month
Cancellation by lead_time
Cancellation by deposit_type
Cancellation by customer type
Cancellation by room type
Correlation heatmap
Country-wise cancellations

Questions We Will Answer:
Who cancels more: repeated guests or new guests?

Does high lead time cause more cancellations?
Does deposit type reduce the cancellation rate?
Which months have the highest cancellations?

3. Data Preparation
Steps:
Remove or impute missing values
Convert categorical variables to numeric (one-hot encoding or label encoding)
Feature scaling (for KNN, NN)
Handle highly imbalanced data if needed
Create derived features (e.g., total_people = adults + children + babies)

4. Modeling
We will build and compare multiple models:

A. Classification Models (Main Task)
Decision Tree (CART / C4.5)
KNN Classifier
Logistic Regression
Random Forest Classifier
Neural Network (Simple MLP)

Why these models?
Easy to run
Good accuracy
Useful for large and small datasets
Professor-approved + taught in class

5. Evaluation
We will evaluate the models using:
Metrics:
Accuracy
Precision
Recall
F1-score
Confusion Matrix
AUC-ROC Curve (optional)

Business Interpretation:
What is the cost of incorrectly predicting a cancellation?
Which model identifies high-risk customers most accurately?

6. Deployment / Output (for PPT):
Although full deployment is not required:

We will produce:
Cancellation risk classification output
Feature importance chart (e.g., deposit_type, lead_time, special_requests)
Simple decision dashboard concept (optional)
This shows practical value to hotels.

Conclusion (for proposal):
This project uses real-world hotel booking data to predict customer cancellation behavior.
It fits perfectly with Data Analytics + Machine Learning requirements, supports multiple algorithms taught in class, and provides valuable insights for business planning.

All four team members can contribute easily by dividing tasks into:
Data cleaning
EDA & visualizations
Model building
Report & PPT documentation