Hotel Booking Cancellation Classification Project

1. Project Overview

This project aims to build a full machine learning pipeline that predicts whether a hotel booking will be canceled or not. Students are required to perform data preprocessing, exploratory data analysis (EDA), feature engineering, model building, feature selection, and performance evaluation.

The target variable for prediction is **is_canceled**.

2. Dataset Description

Target Column

- **is_canceled**: Indicates whether the booking was canceled (1) or not (0).

Input Feature Columns

- hotel: Type or name of hotel where the booking was made.
- lead_time: Number of days between booking date and arrival date.
- arrival_date_year: Year of arrival.
- arrival_date_month: Month of arrival.
- arrival_date_week_number: Week number of arrival date.
- arrival_date_day_of_month: Day of the month of arrival.
- stays_in_weekend_nights: Number of weekend nights included.
- stays_in_week_nights: Number of week nights included.
- adults: Number of adults.
- children: Number of children.
- babies: Number of babies.
- meal: Type of meal plan selected.
- country: Country of origin of the customer.
- market_segment: Booking segment.
- distribution_channel: Distribution channel of booking.
- is_repeated_guest: Whether the customer is a repeated guest.
- previous_cancellations: Number of previous cancellations.
- previous_bookings_not_canceled: Number of successful previous bookings.
- reserved_room_type: Reserved room type.
- assigned_room_type: Assigned room type.
- booking_changes: Number of updates to the booking.
- deposit_type: Type of deposit.
- agent: Booking agent ID.
- company: Company ID.
- days_in_waiting_list: Days the booking was on a waiting list.
- customer_type: Type of customer.
- adr: Average daily rate.
- required_car_parking_spaces: Number of required parking spaces.
- total_of_special_requests: Number of special requests.

- reservation_status: Final reservation status.
- reservation_status_date: Date of final status.
- city: City of the hotel.

3. Tasks Required in This Project

A. Data Preprocessing & Cleaning

Students must:

- Handle missing values
- Fix incorrect data types
- Remove duplicates if found
- Handle outliers
- Convert categorical values to numerical (One-Hot, Label Encoding)
- Convert dates to usable features

-
- **B. Exploratory Data Analysis (EDA)**
- Students should generate visual charts such as:
- Distribution of cancellations
- Booking trends by month, week, city
- Relationship between ADR and cancellations
- Correlation heatmap
- Lead time analysis
- Categorical plots (hotel type, market segment, customer type)
- Students must write insights for every important graph.

C. Checking Data Balance

- Plot distribution of the target variable
- If imbalanced, apply SMOTE or undersampling/oversampling

D. Feature Engineering

Possible tasks:

- Extract useful features from dates
- Create new features (e.g., total_stay = weekend + week nights)
- Drop irrelevant or highly correlated features

- **E. Feature Selection using Genetic Algorithm**

- Students must:Apply a Genetic Algorithm to select the best subset of feature and Test its impact on model performance

- **F. Model Building**

- Students must build and compare:

- K-Nearest Neighbors (KNN)

- Decision Tree Classifier

- Neural Network (MLPClassifier)

- Each model must:

- Use train/validation/test split

- Be tuned using validation data (not test data)

- **G. Performance Evaluation**

- Students must compute:

- Accuracy

- Precision

- Recall

- F1 Score

- Confusion Matrix

- Models should be compared and best one selected.

-

- **4.Expected Deliverables:Deadline to submit these requirements is Thursday 12/20/2025 11:59pm**

- Students must submit:

- Python notebook (.ipynb)

- Full project report (PDF/Word)

- Charts and insights

- Final model comparison table

- Explanation of chosen features (after GA)

-