To successfully complete this assignment for the \*\*Hotel Booking Prediction\*\* task, here'

### \*\*1. Understand the Assignment Brief and Dataset\*\*

- \*\*Objective\*\*: Build a machine learning classifier to predict whether a hotel booking will be canceled.

- \*\*Dataset\*\*: You will be working with real-world hotel booking data from two hotels in Portugal. It includes over 100,000 records and more than 30 features, covering bookings from July 2015 to August 2017.

- \*\*Challenges\*\*: The data is noisy and inconsistent (common with real-world datasets), and your goal is to clean and process the data to build a classifier that achieves more than 70% accuracy.

- \*\*Assessment Type\*\*: Individual work (no collaboration). The final submission should be in Jupyter Notebook format, both as PDF and `.ipynb`.

### \*\*2. Data Pre-processing (30%)\*\*

Start by cleaning the dataset. This is crucial for improving the quality of your predictions.

- \*\*Understand the Dataset\*\*:

- Load the data and inspect it.

- Use basic data checks like `describe()` and `info()` to get a sense of the distribution of data types and potential issues such as missing values.

- Summarize the dataset’s features and relationships.

- \*\*Handle Missing and Inconsistent Data\*\*:

- Identify missing values using `.isnull().sum()`.

- Choose how to handle them (imputation with mean/median/mode, removal of rows/columns, etc.), and justify your decisions.

- \*\*Extract Insights\*\*:

You need to calculate several statistics, including:

- \*\*Cancellation percentages\*\* for City and Resort hotels.

- \*\*Most ordered meal types\*\*.

- \*\*Returning guests\*\* (how many guests made multiple bookings).

- \*\*Room types and cancellations\*\*: See if certain room types have higher cancellation rates.

- \*\*Customer types\*\* (families, couples, etc.) and their behavior patterns.

### \*\*3. Exploratory Data Analysis (EDA) (15%)\*\*

Visualizing data helps uncover trends and patterns.

- \*\*Create at least three visualizations\*\*:

- \*\*Line graph\*\*: Plot the number of guests over time (monthly data).

- \*\*Bar chart\*\*: Depict guest stay durations.

- \*\*Pie chart\*\*: Show where most guests come from (guest origins).

- \*\*Tools\*\*: Use libraries like `matplotlib`, `seaborn`, or `plotly` for your plots.

### \*\*4. Feature Engineering (25%)\*\*

Improve the dataset by transforming or selecting important features.

- \*\*Feature Selection and Conversion\*\*:

- Drop irrelevant columns and convert data types if needed (e.g., date to datetime, categorical data to numeric).

- \*\*Outlier Analysis\*\*:

- Use boxplots to visualize outliers and decide whether to handle them or leave them in the dataset.

- \*\*Binning\*\*:

- You may group continuous data into bins to simplify analysis, e.g., stay duration categories (short, medium, long).

- \*\*Scaling\*\*:

- Check if feature scaling is needed (especially for models like SVM or KNN that are sensitive to the scale of data). Use normalization or standardization if required.

### \*\*5. Classifier Training (20%)\*\*

Now, train the machine learning model.

- \*\*Model Selection\*\*:

Use basic classifiers like:

- \*\*Decision Tree\*\*

- \*\*Random Forest\*\*

- You can explore other models if needed, such as \*\*Logistic Regression\*\* or \*\*Support Vector Machines (SVM)\*\*.

- \*\*Splitting the Dataset\*\*:

- Use `train\_test\_split()` from `scikit-learn` to split the dataset into training and testing sets (maintaining class balance using \*\*stratified sampling\*\*).

- Aim for a model with an accuracy of at least 70%.

- \*\*Hyperparameter Tuning\*\* (optional but beneficial):

- Use cross-validation and grid search (`GridSearchCV`) to fine-tune your model parameters and improve performance.

### \*\*6. Feature Importance (10%)\*\*

Once the model is trained, identify the most important features influencing the prediction.

- \*\*Importance Visualization\*\*:

- Use feature importance methods (like `.feature\_importances\_` in Random Forest) to determine which features are contributing most to the model.

- Visualize this importance using bar plots or heatmaps.

### \*\*7. Finalize and Submit\*\*

- Ensure your \*\*Jupyter Notebook\*\* is clean, well-commented, and easy to follow. The code should include:

- Data loading and preprocessing

- EDA and visualizations

- Feature engineering

- Model training and evaluation

- Feature importance analysis

- \*\*Submission Format\*\*: Submit the notebook as both \*\*PDF\*\* and \*\*`.ipynb`\*\* files. Do not compress the files or submit them in any other format, as this will lead to marks being deducted.

### \*\*Marking Scheme (Summary)\*\*:

- \*\*Data Pre-processing\*\*: 30% (missing values, inconsistent values, outlier analysis)

- \*\*Exploratory Data Analysis\*\*: 15% (charts and insights)

- \*\*Feature Engineering\*\*: 25% (selection, binning, scaling)

- \*\*Classifier Training\*\*: 20% (splitting, training, accuracy)

- \*\*Feature Importance\*\*: 10% (discussion and visualization of important features)

### \*\*References and Guidance\*\*:

- Refer to the \*\*academic integrity rules\*\* and ensure that your submission is original.

- You can use AI tools to help during research but ensure all sources and ideas derived from AI are properly referenced.

By following these steps and adhering to the assignment brief, you should be able to meet the requirements for the coursework.

Sure! Here’s a brief and concise outline of how to approach the \*\*Hotel Booking Prediction\*\* assignment:

### 1. \*\*Understand the Task\*\*

- \*\*Goal\*\*: Predict hotel booking cancellations using machine learning.

- \*\*Dataset\*\*: Over 100,000 records with 30+ features about bookings from two hotels (Resort and City) in Portugal.

- \*\*Deliverables\*\*: Submit a Jupyter notebook as both PDF and `.ipynb` formats.

### 2. \*\*Data Pre-processing (30%)\*\*

- \*\*Understand Data\*\*: Check for missing values, data types, and inconsistencies.

- \*\*Clean Data\*\*: Handle missing values and justify your choices.

- \*\*Insights\*\*: Calculate key metrics (e.g., cancellation percentages, popular room types, frequent guests).

### 3. \*\*Exploratory Data Analysis (EDA) (15%)\*\*

- Create visualizations:

- \*\*Line graph\*\*: Number of guests per month.

- \*\*Bar chart\*\*: Guest stay durations.

- \*\*Pie chart\*\*: Guests' geographic origins.

### 4. \*\*Feature Engineering (25%)\*\*

- \*\*Select/Drop Features\*\*: Keep relevant features, remove/convert others.

- \*\*Handle Outliers\*\*: Use boxplots to visualize and address outliers.

- \*\*Binning/Scaling\*\*: Simplify data and normalize if necessary.

### 5. \*\*Classifier Training (20%)\*\*

- \*\*Choose Model\*\*: Use a basic classifier (e.g., Decision Tree or Random Forest).

- \*\*Split Data\*\*: Use stratified train-test split to maintain class balance.

- \*\*Accuracy Target\*\*: Aim for more than 70% accuracy.

### 6. \*\*Feature Importance (10%)\*\*

- \*\*Identify Key Features\*\*: Use feature importance to find which features influence cancellations the most.

- \*\*Visualize\*\*: Show feature importance with a plot.

### 7. \*\*Submission\*\*

- Submit your \*\*Jupyter Notebook\*\* in both PDF and `.ipynb` formats (no other formats allowed).

By following this outline, you can stay organized and focus on key aspects of the assignment.