**Machine Learning – Final Project**

***Dataset Description:***

Hotel Reservations Dataset.

Dataset Taken from: <https://www.kaggle.com/datasets/ahsan81/hotel-reservationsclassification-dataset>.

The file contains the different attributes of customers' reservation details.

Data Dictionary:

Booking\_ID: unique identifier of each booking

no\_of\_adults: Number of adults

no\_of\_children: Number of Children

no\_of\_weekend\_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

no\_of\_week\_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

type\_of\_meal\_plan: Type of meal plan booked by the customer:

required\_car\_parking\_space: Does the customer require a car parking space? (0 - No, 1- Yes)

room\_type\_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.

lead\_time: Number of days between the date of booking and the arrival date

arrival\_year: Year of arrival date

arrival\_month: Month of arrival date

arrival\_date: Date of the month

market\_segment\_type: Market segment designation.

repeated\_guest: Is the customer a repeated guest? (0 - No, 1- Yes)

no\_of\_previous\_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking.

no\_of\_previous\_bookings\_not\_canceled: Number of previous bookings not canceled by the customer prior to the current booking

avg\_price\_per\_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)

no\_of\_special\_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)

booking\_status: Flag indicating if the booking was canceled or not.

***Research Questions:***

**Q1.** What factors most influence the booking status (cancelled or not cancelled)?

**Q2.** How does the number of adults, children, and special requests affect the average price per

room?

**Q3.** What is the impact of the number of previous cancellations on the booking status?

**Q4.** Is it possible to predict an order cancellation?

***Techniques used:***

**Linear Regression:** is a supervised machine learning algorithm used for predicting a continuous target variable based on one or more predictor variables. The basic idea behind linear regression is to find the best linear relationship between the target variable and the predictor variables.

**Random Forest:** is an ensemble machine learning algorithm that can be used for both classification and regression problems. The basic idea behind Random Forest is to combine the predictions of multiple decision trees to improve the overall accuracy of the model.

**Logistic Regression:** is a type of supervised machine learning algorithm used for binary classification problems. The goal of logistic regression is to predict a binary outcome (e.g., yes/no, 0/1) based on one or more predictor variables.

**Decision Trees:** is a type of supervised machine learning algorithm used for both classification and regression problems. It works by creating a tree-like model that splits the data into smaller and smaller subgroups based on the values of the input features. Each split is represented by a node in the tree, and each leaf node represents a prediction for the target variable.

**KNN:** (K-Nearest Neighbors) is a machine learning algorithm that is commonly used for classification and regression problems. KNN is a non-parametric algorithm, meaning it doesn't make any assumptions about the underlying distribution of the data. KNN is a simple algorithm that uses the distance metric to identify the k-nearest neighbors to a new data point and classify it based on the majority class of the k-nearest neighbors.

**SVM:** (Support Vector Machines) is a machine learning algorithm that is commonly used for classification and regression problems. SVM is a linear model that tries to find a hyperplane that separates the data into two or more classes. The hyperplane is chosen in such a way that it maximizes the margin between the classes. SVM is a powerful algorithm that can handle both linear and non-linear problems, making it useful for various applications.

***Algorithms Description Summary:***

Q1. This algorithm performs a binary classification task on the dataset using random forest and decision tree classifiers.

It preprocesses the data by dropping a column, one-hot encoding categorical columns, and scaling the data.

The data is split into training and test sets, and the classifiers are trained on the training data.

Finally, the feature importance of the trained classifiers is displayed using the showInfo() function.

Q2. This algorithm prepares the dataset for training three regression models: Linear Regression, Random Forest, and Decision Tree.

It then splits the data into training and test sets, trains the models, makes predictions on the test set, and evaluates each model using mean absolute error.

Finally, it creates a scatter plot of the actual values and the predictions for each model, and it generates bar charts of the average price per room by the number of adults, children, and special requests.

Q3. This algorithm demonstrates the use of four different classification models (Random Forest, Logistic Regression and Decision Tree and KNN) to predict the booking status of hotel reservations based on the number of previous cancellations.

The dataset is loaded, prepared, and split into training and test sets using the pandas and scikit-learn libraries.

The accuracy of each model is then calculated and printed to the console.

Q4. In order to answer that question we want to use a few different ML techniques which used for classification and regression problems: Linear Regression, Logistic Regression, RandomForest, KNN, SVC.

The algorithm gets the dataset after preprocessing, and check the relations between the relevant features and the order cancellation label.

Each model will answer the question with the accuracy rate he predicted.

Finally, the algorithm reports the prediction accuracy of each model and checks if any of the models overfits the data.

***Results:***

What factors most influence the booking status (cancelled or not cancelled)?

Based on the feature importance values that the Random Forest Classifier and Decision Tree Classifier classes returned,

the feature that has the highest importance for predicting whether a booking will be canceled or not is the 'lead\_time' feature.

It has an importance score of 0.31132 (Random Forest) or 0.34635 (Decision Tree), which is higher than any other feature in the dataset (on both models).

Chart

Description automatically generated with medium confidenceChart, calendar

Description automatically generated with medium confidenceActually, It is very reasonable - The higher the lead\_time is, there is more chances of unpredictable events to cause for a cancellation.

Figure . Feature importances by Random Forest.

Figure . Feature importances by Decision Tree.

How does the number of adults, children, and special requests affect the average price per

room?

Based on the figures, Number of adults has the largest impact on avg price per room.

Chart, bar chart

Description automatically generatedChart, bar chart, histogram

Description automatically generatedAdditionally we can see by the figures that the 2 other features (Number of children, number of special requests) are not increasing monotonicly, that means they don’t impact directly on the average price per room.

Chart, bar chart

Description automatically generated

Chart, scatter chart

Description automatically generatedHere we can see the models prediction accuracy rate relative to the actual values:

What is the impact of the number of previous cancellations on the booking status?

Based on the results of the code and the fact that only the **no\_of\_previous\_cancellations** feature is used for training the models, we can infer that the number of previous cancellations has some impact on the booking status.

Looking at the accuracy scores of the different models, we can see that they are all relatively low (and even), with around 55% accuracy rate.

This suggests that the **no\_of\_previous\_cancellations** feature alone may not be sufficient to accurately predict the booking status, and other features may need to be considered.

The reason why we are getting the same accuracy for all four models (Random Forest, Logistic Regression, Decision Tree and KNN) is because we are using only one feature 'no\_of\_previous\_cancellations' to train the models, which may not have a strong correlation with the target variable 'booking\_status'. As a result, the models may not be able to capture the underlying patterns and relationships in the data, leading to similar accuracy scores.

Accuracy of Random Forest: 0.6669882839421089  
Accuracy of Logistic Regression: 0.6669882839421089  
Accuracy of Decision Tree: 0.6669882839421089  
Accuracy of KNN: 0.6669882839421089

Is it possible to predict an order cancellation?

Yes, its possible.

Based on the accuracy of all unoverfitted models, we can get a prediction of an order cancellation.

We can see the knn model is overfitting!

The model which has the best accuracy is Random Forest Classifier – with 0.90723638869745 accuracy rate.

Accuracy of Linear Regression: 0.8011026878015162  
Accuracy of Logistic Regression: 0.8071674707098553  
Accuracy of Random Forest: 0.90723638869745  
Accuracy of KNN: 0.8406616126809097  
KNN Classifier is overfitting !  
Accuracy of SVC: 0.8518263266712612