Diagram

Description automatically generated

|  |
| --- |
| Hotel rating  prediction |
| **TEAM ID: SC\_30**  **TEAM EMAIL:** [**ahmedmohsen2492@gmail.com**](mailto:ahmedesmail4716@gmail.com) |

# Team members

|  |  |  |
| --- | --- | --- |
| **Ahmed Esmail Mohamed**  **Ebrahim Tarek Mohamed**  **Ahmed Mohsen Fathelbab**  **Ebrahim Ahmed Abdelmaboud**  **Ahmed Mohamed Kamal**  **Yasmine Khaled Atta** | **20201700024**  **20201700009**  **20201700067**  **20201700007**  **20201700080**  **20201701154** |  |

# preprocessing

|  |
| --- |
| split the dataset20% for testing (58063 rows) 80% for training (232252 rows)Hotel\_address **It was divided into two columns (hotel\_country, hotel\_city) using regular expressions, pycountry and GeoText.** Review\_DateIt was converted to DateTime first then extracted the year only to column (Review\_year) then Review\_Date column converted to ordinal.Negative\_Review, Positive\_ReviewApplied sentiment analysis to classify if the review positive or negative using vader.Review\_Total\_Negative\_Word\_CountsIf the review already negative so it filled by the word count but if the review is positive so it filled by zero.Review\_Total\_Positive\_Word\_CountsIf the review already positive so it filled by the word count but if the review is negative so it filled by zero.TagsIt was divided into multiple columns (type\_of\_trip, with\_a\_pet, people, Room\_Type, nights, submitted\_from\_mobile).days\_since\_reviewit contains days or day word after the number and both have been removed.Lat, lngBoth contains null data and filled with average.Label encoding (LabelEncoder model)DEFINITIOn: is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.Applied on (Hotel\_Name, Reviewer\_Nationality, type\_of\_trip, people, hotel\_country, hotel\_city).Feature selectionApplied based on the correlation with Reviewer\_Score column and all the columns with correlation more than or equal 0.2 have been selected (Average\_Score, Review\_Total\_Negative\_Word\_Counts, Review\_Total\_Positive\_Word\_Counts). -Then we create a heat map of the correlation matrix using the heatmap function from Seaborn that allows us to visualize the strength and direction of the correlations between different variables in the data:Chart, treemap chart  Description automatically generated11. Handling Outliers  DEFINITIOn: An outlier is an observation in a dataset that lies an abnormal distance from the other observations in the dataset. Outliers can occur due to measurement error, data entry errors or natural variation in the data. Outliers can affect the results of statistical analysis and machine learning models, so it is important to identify and handle them appropriately. 1- using statistical method for detecting outliers : Text  Description automatically generated 2- Using box plot Visualization for detecting outliers : Chart, box and whisker chart  Description automatically generatedChart, box and whisker chart  Description automatically generatedGraphical user interface, chart, box and whisker chart  Description automatically generatedChart, box and whisker chart  Description automatically generated3- removing outliers using the Interquartile Range (IQR) method. 12. Feature scaling (MinMaxScaler)DEFINITIOn: It involves transforming the feature values of a dataset so that they fall within a specified range, typically between 0 and 1.Applied on selected features. |
|  |

# models

|  |
| --- |
| Linear Regressiondefinition: is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variableMSE: 0.032908836186091354score: 0.29922637317226874best parameters: {'fit\_intercept': True, 'normalize': True}Polynomial Regressiondefinition: Polynomial regression is a kind of linear regression in which the relationship shared between the dependent and independent variables Y and X is modeled as the nth degree of the polynomial. This is done to look for the best way of drawing a line using data points.- Used pipeline to merge linear regression and polynomial features.MSE: 0.027699723107936558score: 0.41997005121604314best parameters: {'linear\_\_fit\_intercept': False, 'linear\_\_normalize': True, 'poly\_\_degree': 7, 'poly\_\_include\_bias': True}Visualize the expected vs. predicted VALUES in Polynomial Regression model :Chart  Description automatically generatedRandom Forest Regressiondefinition: a supervised learning algorithm that uses ensemble learning method for regression.MSE: 0.027724169074332266score: 0.4372377176111383best parameters: {'max\_depth': 10} Visualize the expected vs. predicted VALUES in Random Forest Regression model : Chart  Description automatically generatedRidge Regressiondefinition: Is A Technique Which Is Used For Analyzing Multiple Regression Where The Data Suffers From Multicollinearity.MSE: 0.03290908982472024score: 0.29923119202735904best parameters: {'alpha': 0.0001, 'fit\_intercept': True, 'normalize': True, 'solver': 'sag'}Gridsearchcvdefinition: a technique for finding the optimal parameter values from a given set of parameters in a gridApplied to find best parameters for each model. |

# conclusion

|  |
| --- |
| In the Hotel Rating Prediction project , we aim to build a machine learning model that can predict the rating of a hotel based on various features such as address, Average Score, and customer reviews.To accomplish this task, we first need to gather and preprocess the data. This involves collecting data from various sources such as hotel booking websites and customer review platforms, cleaning the data to remove any missing or incorrect values, and transforming the data into a format suitable for machine learning.Next, we can use various machine learning algorithms such as linear regression, Ridge Regression, and random forest regression to train and test our models. We can also use techniques such as feature selection, Feature scaling and gridsearchcv to optimize the performance of our models.Once we have trained and tested our models, we can use them to predict the ratings of new hotels based on their features. We can also evaluate the performance of our models using metrics such as mean squared error (MSE) and score.Overall, the Hotel Rating Prediction project using Python requires a combination of data preprocessing, machine learning, and data analysis skills to successfully build an accurate and reliable model that can predict hotel ratings. |