**Tripadvisor European Restaurants Analysis and Rating Prediction**

**Group 18:**

Sushmitha Jogula (NUID: 001546751)

Gautham Rajsimha Pulipati (NUID: 001572432)

**Introduction:**

**Background:**

From planning and booking to taking a trip, Tripadvisor is the world's largest travel website and guidance platform, assisting hundreds of millions of people in becoming better travelers by making trip planning easy, thanks to more than 988 million reviews and opinions from almost 8 million companies. The Tripadvisor European Restaurants dataset consists of information related to all the restaurants in the main European countries. This dataset includes 1083397 restaurants along with their attributes and features such as average rating, cuisine types, awards, popularity ranking, location information, number of reviews, open working hours etc. The data is retrieved from the publicly available [Tripadvisor](https://tripadvisor.com/) website by scraping the data in early May 2021.

**Motivation and Goal:**

Being avid travelers and frequent users of the Tripadvisor website, we thought this would be the apt dataset to explore for our project and to understand various factors of restaurants that contribute in making a restaurant successful and appreciated by the users, while performing a comparative analysis of common features of restaurants across several European countries such as average rating, open hours, awards, locations etc.

Our goal is to analyze and predict the performance and success of a restaurant based on various factors such as cuisine types, location, price range, rating for food, service, and value, average rating, popularity index, awards etc. by applying several machine learning models.

* Using machine learning models to predict the rating value of the restaurant
* Compare the predicted rating value with the true rating value and identify the most efficient algorithm with the highest accuracy rate for our dataset
* Visualize and analyze the performance and accuracy of all the applied models

**Methodology:**

* Perform data wrangling and data preprocessing to clean the raw data after it is read and before training the data
* Identify any missing or inconsistent values in the target columns or feature matrix and handle it accordingly
  + Either discard the column from the prediction process
  + Or fill with mean values
* Perform Exploratory Data Analysis (EDA) on the raw dataset to analyze and visualize the fundamental attributes in the dataset
* Apply machine learning models to predict the restaurant’s performance and success
* Analyze the performance of applied learning models and review the best among them in terms of accuracy

The Machine Learning models that will be used in this project are:

1. **LINEAR REGRESSION**

Linear Regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of variables do a good job in predicting the outcome variable, and (2) which particular variables are significant predictors of outcome variable and in what way do they impact the outcome variable?

1. **DECISION TREE**

Decision Tree builds regression or classification models in the form of a tree structure. It breaks a dataset into smaller and smaller subsets while an associated decision tree is incrementally developed at the same time. The final result is a tree with decision nodes and lead nodes. Decision trees can handle both numerical and categorical data.

1. **SUPPORT VECTOR MACHINE**

Support Vector Machines (SVM) can be used as a regression method, maintaining all the main features that characterize the algorithm (maximum margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification. The error is minimized by individualizing the hyperplane which maximises the margin.

**DataSets**:

##### The dataset has 1083397 unique restaurant links with 42 columns. The size of data is 679.68 MB. Below are the details of this dataset:

|  |  |
| --- | --- |
| Column Name | Column Description |
| restaurant\_link | Unique TripAdvisor Link |
| restaurant\_name | Restaurant Name on TripAdvisor |
| original\_location | Original Location displayed on TripAdvisor |
| country | Country name retrieved from original\_location |
| region | Region name retrieved from original\_location |
| province | Province name retrieved from original\_location |
| city | City name retrieved from original\_location |
| address | Address displayed on TripAdvisor |
| latitude | Latitude coordinate |
| longitude | Longitude Coordinate |
| claimed | Restauarnt business claimed on TripAdvisor |
| awards | Award Names |
| popularity\_detailed | Popularity detailed ranking |
| popularity\_generic | Popularity generic ranking (among all places to eat in the area) |
| top\_tags | Top tag names |
| price\_level | Level of price in current currency |
| price\_range | Range of price in current currency |
| meals | Types of meals |
| cuisines | Types of Cusines |
| special\_diets | Types of special diets |
| features | Restaurant features |
| vegetarian\_friendly | Is the restaurant vegetarian friendly? (Yes or no) |
| vegan\_options | Does the restaurant have vegan options? (Yes or no) |
| gluten\_free | Does the restaurant have gluten free options? (Yes or no) |
| original\_open\_hours | Original open hours on trip advisor |
| open\_days\_per\_week | Number of open days per week retrieved from original\_open\_hours |
| open\_hours\_per\_week | Number of hours per week retrieved from original\_open\_hours |
| working\_shifts\_per\_week | Number of working shifts per week retrieved from original\_open\_hours |
| avg\_rating | Average restaurant rating |
| total\_reviews\_count | Total count of the reviews |
| default\_language | Default language displayed while scraping |
| reviews\_count\_in\_default\_language | Total reviews count in default language |
| excellent | Excellent reviews count in default language |
| very\_good | Very\_good reviews count in default language |
| average | Average reviews count in default language |
| poor | Poor reviews count in default language |
| terrible | Terrible reviews count in default language |
| food | Food rating |
| service | Service rating |
| value | Value rating |
| atmosphere | Atmosphere rating |
| keywords | Popular keywords |

**Data Source:**

<https://www.kaggle.com/stefanoleone992/tripadvisor-european-restaurants>

**References:**

<https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/>

<https://www.saedsayad.com/decision_tree_reg.htm>

<https://www.saedsayad.com/support_vector_machine_reg.htm>