**BASIC INFORMATION**

**Title of Project:** Travel Rating Reviews

**Student Name:** Mohit Gupta

**Branch :** Artificial Intelligence & Data Science B1B

**Enrolment Number:** 05119011921

**Signature:**

**Email ID:** mohit.051190111921@ipu.ac.in

**Contact Number:** 8527716857

**Google Drive Link:** https://drive.google.com/drive/folders/1b\_0bxVLPg6K4AE2C\_3OYC1Jf1EHX7wuG?usp=sharing

**Google Website Link:**

**YouTube Video Link:**

**GitHub Link:**

**Title**: **Travel Rating Reviews**

**Abstract:**

Presently, the web has upset the manner in which individuals plan their movements, with online travel rating stages giving significant experiences and client created audits. However, it can be difficult and time-consuming to manually analyze and classify a large number of reviews. A machine learning-based method for automatically categorizing travel rating reviews is the goal of this project. To determine the sentiment and rating associated with each review, we make use of a variety of classification models, such as Logistic Regression, Support Vector Machines (SVM), Gaussian Naive Bayes, K-Nearest Neighbors (KNN), and Decision Trees. Moreover, we investigate the use of bunching models, for example, K-Means and Agglomerative Hierarchical clustering to bunch comparable audits in view of their substance. Preprocessing the review data, finding relevant features, and training the models on a labeled dataset are all part of the project. Using appropriate metrics like accuracy, precision, recall, and F1-score, we assess each model's performance. The findings demonstrate that a variety of classification models can accurately predict travel review sentiment and rating. Additionally, the clustering models facilitate the identification of patterns and the grouping of reviews that are alike, facilitating a deeper comprehension and analysis of the review dataset. The discoveries of this venture can be used by movement stages, the travel industry associations, and voyagers to effectively investigate and decipher huge volumes of movement rating surveys, empowering further developed direction and customized suggestions.

**Keywords:**

Classification,

K-Means Clustering,

Agglomerative Hierarchical clustering,

Logistic Regression,

Support Vector Machine(SVM) Classifier

**1. Introduction:**

Client audits are often regarded as critical hotspots for promoting knowledge and emotion examination (Dickinger and Mazanec Citation2015). The purpose of sentiment analysis is to create a system for analysing and evaluating consumer feedback on websites, blogs, Twitter, or Instagram. Client surveys have recently had a positive impact on business advancement and attracting new customers, thanks to the expansion of online frameworks. As a result, review classification emerges as the most essential technology for textual data arrangement. Survey order is defined as distributing new records to a set of predefined classes based on grouping designs . In reality, consumer sentiment is critical to the hospitality business because it helps build customer interactions, manage customer preferences, and provide higher-quality services that are more personalized to each customer's demands. Sentiment research also assists businesses in developing marketing strategy, executing successful marketing campaigns, and forecasting client happiness.

As a result, hotels now have access to a substantial amount of information through both internal, organised systems such as CRM systems and external, unstructured systems such as social networks and websites. To take use of the increased availability of unstructured data via social networks and websites, previously unknown patterns in a huge dataset must be discovered and managed.

An examination of this volume of information necessitates the association's involvement of tremendous assets in Large Information developments that have been involved well in various fields. As a result, audit mining and sentiment analysis can be developed as important tools in this association.

Machine learning technologies are widely used to analyse and predict emotions. There are three layers of sentiment analysis: record level, sentence level, and element level. According to Tripathy, Agrawal, and Rath (2016), sentiment analysis at the document level determines whether a document is positive, negative, or neutral. This study took into account sentiment analysis at the document and feature levels. Reviews are sorted using machine learning techniques. In this paper, six techniques for AI were utilized to characterize opinions. To determine how well these methods work, their accuracy has been investigated. Additionally, feature sentiment analysis has been used to examine the customer's perspective.

We have used various Classification Model ie, Logistic Regression, SVM, Gausian Naive Bayes, KNN, Decision Tree

**Logistic Regression :**

Logistic Regression is a well-liked statistical model for problems with binary classification. It demonstrates the relationship between the input variables and the probability of belonging to a particular class. Strategic Relapse makes use of the calculated ability to predict the likelihood that an example will belong to a particular class. It can manage straight and nonlinear connections between the objective variable and the information highlights.

**Support vector machines, or SVMs, are:**

SVM is a good machine learning algorithm for classification and regression. It finds a hyperplane in the feature space that best divides the data points into distinct classes. The edge between the hyperplane and the closest data of interest from each class is expected to increase with SVM. It can deal with data that can be separated linearly or nonlinearly by means of kernel functions that translate the input features into higher-dimensional spaces.

**Gaussian Naive Bayes:**

Gaussian Naive Bayes is a probabilistic classifier considering Bayes' speculation with an assumption of opportunity between the components. It anticipates a Gaussian (typical) distribution of the elements. By combining the preceding likelihood and the contingent likelihood of the elements, it calculates the likelihood of an event occupying a particular class. Despite its naive assumption of feature independence, Gaussian Naive Bayes frequently performs well, particularly when working with small to medium-sized datasets.

**K-Nearest Neighbors (KNN):**

Non-parametric classification and regression algorithm known as KNN. New instances are classified based on the majority vote of the K closest neighbors in the training set. The distance metric (e.g., Euclidean distance) is used to choose the area between events. KNN can manage nonlinear choice limits and makes no suppositions about the fundamental information conveyance. However, large datasets may result in substantial computational expenses.

**Decision Trees:**

Decision Trees are adaptable ML models that can be used for both gathering and backslide tasks. In their tree-like flowchart structure, each internal node represents a decision based on a feature and each leaf node represents a class label or a predicted value. Decision Trees recursively divide the data and divide the feature space based on the most informative features in order to maximize information gain or Gini impurity. They can handle both mathematical and straight-up information, are easy to imagine, and can be interpreted. However, they may be susceptible to overfitting if they are not properly regularized.

All of these models have its benefits and losses, and their show could move depending upon the dataset.

**K-Means CLustering:**

K-Means CLustering is an iterative calculation that expects to parcel a dataset into K unmistakable groups. Each data point is assigned to the closest centroid after K points are chosen at random as the initial centroids. The centroids are then refreshed by ascertaining the mean of the places in each bunch. This cycle rehashes until combination, bringing about K groups.

**Agglomerative Hierarchical clustering:**

Agglomerative Hierarchical clustering is a bottom-up method in which each data point is initially placed in its own cluster before being iteratively merged into larger clusters. It begins by merging the closest pairs of data points and calculating the distances between each pair. Until all of the data points are in a single cluster, this process continues, resulting in a dendrogram that depicts the sequence of merges.

K-means clustering assumes equal-sized, spherical clusters and requires a predefined number of clusters. Agglomerative Hierarchical clustering can handle clusters of various sizes and shapes without requiring a specific number of clusters to be specified. However, large datasets may incur significant computational costs.

**2. Proposed Methodology [ Pictorial Diagram and Explanation of each sub modules]**

**Diagram / Flow Chart**

**a. Datasets**

Dataset is related to Travel Rating Reviews. It comprises the various features on which rating of the place depends. It has mainly float values in all features.

**b. Pre processing**

Our dataset has 2 columns that are not useful. Our data set has 2 NULL values and one string value that makes it an object. All other features are float values. And there are some outliers but those are also useful. Shape of dataset is(5456, 26).

EDA gives the distribution of the features.

**c. Feature Scaling**

Feature Scaling is done using Standard Scaler. It will standardize the data and make the mean of the data features nearly the same.

**e. Train Test Split**

We divide the data in Training and Testing data so that we can check the accuracy and test our data.

**e. Model Training and Testing with the help of Clustering and Classification**

We make clusters using Clustering and make labels from 1 to 5 as rating of the tourist places. Then We classify and train using Classifiers like Logistic Regression.

**f. Performance Measure**

There are various Performance metrics that can help to measure performance like Confusion Matrix, Classification Report, F1 Score and Accuracy Score, etc.

**3. Result & Discussion**

After clustering(Kmeans clustering and Agglomerative Hierarchical clustering) the data set has the labels to classify it into ratings from 1 to 5.

Then we apply the various Classification models after splitting the dataset into training and testing data -

Logistic Regression

SVM

Gausian Naive Bayes

KNN

Decision Tree

After applying the model we calculate the accuracy scores.

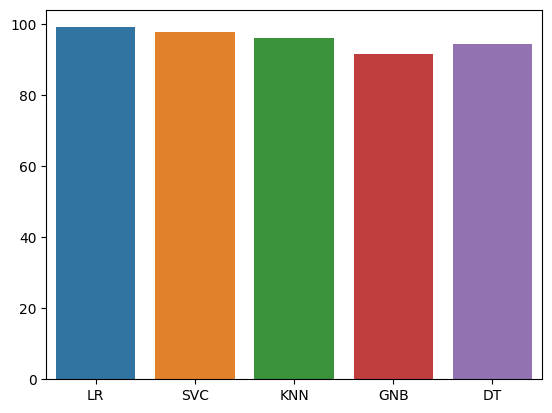
**LR :- 99.04692082111437 %**

**SVC :- 97.65395894428153 %**

**KNN :- 96.04105571847508 %**

**GNB :- 91.42228739002933 %**

**DT :- 94.28152492668622 %**



We can see the best accuracy is shown by Logistic Regression on our data.

**Accuracy - 99.12023460410558 %**

(At the point of creating model)

To further increase the accuracy we can apply hyperparameter tuning using Grid Search Cross Validation.

**After Hyperparameter tuning**, we have best parameters as -

**best\_model = LogisticRegression()**

**best\_parameters = {'penalty': 'l2', 'C': 100, 'tol': 0.001}**

**best\_score = 99.09565305653325 %**

**After applying the Hyperparameter** tuning parameters the **Accuracy** becomes **99.26686217008798 %**.

Our RMS error is also very very less - 0.21661214442955293

**Performance Matrix**

1. **Classification Matrix**

**precision recall f1-score support**

**1 0.99 0.99 0.99 135**

**2 0.98 1.00 0.99 243**

**3 1.00 0.99 0.99 455**

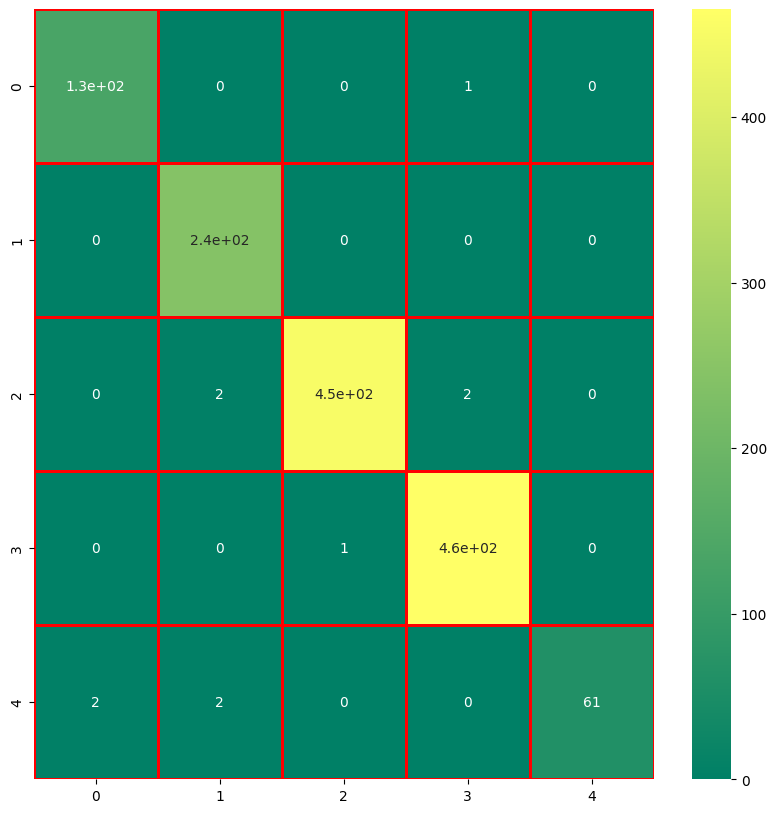
**4 0.99 1.00 1.00 466**

**5 1.00 0.94 0.97 65**

**accuracy 0.99 1364**

**macro avg 0.99 0.98 0.99 1364**

**weighted avg 0.99 0.99 0.99 1364**

1. **Confusion Matrix**

4. **Conclusion**

From the Travel Rating Reviews Project we could conclude that the Best Model that fits the data is **LOGISTIC REGRESSION CLASSIFIER** with accuracy **99.26789 %**.

Clusters and no. of data points in them

1 253

2 940

3 1833

4 584

5 1846

**Future Work**

Certainly! Here are a few expected future headings for an AI project on Travel rating reviews:

1. Modifying Models: Investigate methods for further improving the classification models' performance. Hyperparameter fine-tuning, experimenting with various methods of feature extraction, or incorporating advanced methods like ensemble learning or neural networks are all examples of this.
2. An Aspect-Based Analysis of Emotions: Learn more about travel review aspect-based sentiment analysis. Distinguish and dissect opinions connected with explicit parts of movement encounters like facilities, administration, area, or conveniences. This gives more granular bits of knowledge to voyagers and organizations for designated upgrades.
3. Personalization and User Profiling: Make user profiles based on reviews from travel rating sites. Create individualized travel experiences or recommendations for each user by analyzing patterns and preferences. This makes travel platforms more enjoyable and engaging for users.
4. Analyses in real time: Create a method for analyzing travel rating reviews in real time. Consistently interact and update surveys to acquire quick bits of knowledge into client opinion. Travel companies are able to respond or intervene more quickly as a result of this.
5. Analyses of Multilingualism: Add the ability to handle reviews in multiple languages to the project. To accurately analyze travel rating reviews in various languages and regions, employ multilingual natural language processing methods.
6. Including Data from Other Sources: To provide additional context and enrich the analysis of travel rating reviews, integrate external data sources like news articles, social media feeds, or weather data. The overall travel experience and its impact on customer sentiment are better understood as a result of this.
7. Model Expendableness: Increase transparency and confidence in the analysis by introducing AI techniques that can be explained. Provide explanations for the predictions made by the models so that users and businesses can comprehend the factors that influence a travel review's sentiment or rating.

The ML project on travel rating reviews can improve accuracy, granularity, personalization, real-time analysis, multilingual capabilities, and model explainability by pursuing these future directions. By providing actionable insights and enhancing decision-making processes, these advancements benefit travelers, travel businesses, and the travel industry as a whole.

**References**

1. <https://www.sciencedirect.com/science/article/abs/pii/S0261517716301698>
2. <https://dl.acm.org/doi/abs/10.1145/3443279.3443282?casa_token=kkHukxFTpVEAAAAA:Xh69qWzHtzWdcLarv7YP4VhOCIaaeeRkpVNgupGEJ_Dsk30HFdKp5_a0tXM6Kc_wVHo2HaWrNNZ2>
3. <https://pubsonline.informs.org/doi/abs/10.1287/serv.2017.0189>
4. <https://dl.acm.org/doi/abs/10.1145/2914586.2914629?casa_token=QPlvByg3Zw0AAAAA:e8k9NJUhusWkonMYpBhgk2xse2W5EXjMTilHMJUJNtZGcW9DhMbfQnZ0uyjo5Qa93A-a3vI-8lYx>
5. <https://www.emerald.com/insight/content/doi/10.1108/JHTI-02-2022-0078/full/html>
6. <https://dl.acm.org/doi/abs/10.1145/3038912.3052632?casa_token=Wnk-_ZNGjKAAAAAA:1QgkiM46VGWzAahoaO62Qz5I8J3kyusHbKcAjWgznISYjQuZpMcESGontq5KOWo6KcGHzejCphNV>
7. <https://thesis.eur.nl/pub/65503/Thesis-final-draft-Marjon-van-de-Lindeloof-504427.pdf>
8. <https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE07261363>
9. <https://www.sciencedirect.com/science/article/pii/S0736585317305543>
10. <https://www.sciencedirect.com/science/article/pii/S0378720621001610>
11. <https://link.springer.com/chapter/10.1007/978-3-319-56538-5_50>
12. <https://ieeexplore.ieee.org/abstract/document/7373661/>
13. <https://arxiv.org/abs/1605.05362>
14. <https://link.springer.com/chapter/10.1007/978-3-030-50641-4_4>
15. <https://ieeexplore.ieee.org/abstract/document/7484319/>