A MACHINE LEARNING MODEL TO CLASSIFY INDIAN TAXI SYSTEM IN TOURISM INDUSTRY

Syed Abdul Majid[1], Mohammed Awais[2], Nouman Mohsin[3], Saleha Butool[4]

[1,2,3]B.E. Student, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

[4] Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

salehabutool@lords.ac.in

Abstract: *India is emerging as a top destination for tourism, with taxi services playing a crucial role in supporting this growth and urban transportation. Recognizing the significance of taxi services, we conducted a sentiment analysis of customer reviews from various taxi service providers in India. This study focused exclusively on reviews from Indian online platforms. We employed machine learning techniques to analyze the sentiment of these taxi reviews, providing insights into customer sentiments and the quality of taxi services and amenities. Our research compared multiple machine learning algorithms using the dataset. Among them, Support Vector Machine (SVM) emerged as the top performer, surpassing other algorithms in terms of accuracy, F1 score, and recall, achieving rates of 89%, 82%, and 86%, respectively. This study contributes to understanding customer perceptions through sentiment analysis of taxi reviews, enhancing decision-making in the taxi service industry.*

I. Introduction

The tourism industry faces a myriad of challenges in the digital age, where customer feedback and recommendation systems play pivotal roles in shaping traveler experiences. Recent advancements in machine learning have been instrumental in revolutionizing how tourism entities analyze and utilize vast amounts of user-generated content (UGC). This content, comprising reviews, comments, and travel blogs across platforms like review sites, YouTube, and social media channels such as Facebook and Twitter, serves as a rich source of information for both tourists and service providers alike [2].

In India, the rapid expansion of internet connectivity and the widespread adoption of smartphones have democratized access to travel information. Travelers now rely heavily on online platforms to plan trips, seek recommendations, and share their experiences. The shift towards digital platforms has not only empowered consumers but has also presented new opportunities and challenges for tourism businesses. Understanding and harnessing the sentiments expressed through UGC have become critical for businesses aiming to enhance customer satisfaction and refine service offerings.

Sentiment analysis, a branch of text mining, plays a crucial role in extracting and understanding opinions, attitudes, and emotions expressed in textual data. By applying sentiment analysis techniques, tourism businesses can gain insights into customer preferences, identify areas for improvement, and gauge overall customer satisfaction levels. This analytical approach goes beyond mere data aggregation by delving into the underlying sentiments and nuances of customer feedback, providing actionable insights that can drive strategic decision-making.

Recent studies have demonstrated the efficacy of machine learning algorithms such as Support Vector Machines (SVM) and Naive Bayes in sentiment analysis tasks within the tourism domain [4, 5]. These algorithms are used to classify and analyze sentiments expressed in textual data, achieving varying degrees of accuracy depending on the dataset and methodology employed. For instance, SVM has been shown to outperform Naive Bayes in certain contexts, underscoring the importance of choosing the right algorithm based on specific application requirements and data characteristics.

Moreover, the integration of sentiment analysis into tourism operations extends beyond improving customer satisfaction. It enables businesses to monitor brand perception, identify emerging trends, and preemptively address potential issues before they escalate. By leveraging insights derived from sentiment analysis, tourism stakeholders can tailor marketing strategies, optimize service delivery, and foster more personalized customer interactions.

In conclusion, while the digital landscape presents both challenges and opportunities for the tourism industry, the ability to harness and analyze UGC through sentiment analysis offers a powerful tool for driving innovation and enhancing competitiveness. As tourism continues to evolve in response to technological advancements and changing consumer behaviors, the strategic application of sentiment analysis will remain instrumental in shaping the future of travel experiences and industry practices.

II. Literature Survey

1)Justice and Customer Satisfaction in Customer Retention. Authors: Noel Y. M. Siu

Customer satisfaction hinges on perceived fairness and justice when service failures occur. This study explores the cumulative nature of satisfaction, examining both pre-recovery and post-recovery phases in service interactions. Justice dimensions (distributive, procedural, and interactional) mediate between prior and recovery satisfaction, influencing customer retention behaviors. The findings underscore the importance of addressing justice perceptions to foster long-term customer relationships.

2) Influence of Customer Satisfaction on Loyalty:A Study on Mobile Telecommunication Industry Authors: Hossain

This study investigates factors influencing customer loyalty in Bangladesh's mobile telecommunications sector. It identifies communication, price structure, value-added services, convenience, sales promotions, and customer service as critical determinants of customer satisfaction and subsequent loyalty. Using descriptive statistics and regression analysis, the research highlights the positive correlations between these factors and customer loyalty.

3) A Comparison of Machine Learning Techniques for Customer Churn Prediction Authors: T. Vafeiadis, K. I. Diamantaras

This comparative study evaluates machine learning methods for predicting customer churn in telecommunications. Boosting techniques significantly enhance model performance, with SVM-POLY using AdaBoost achieving high accuracy and F-measure. The research emphasizes the superiority of boosted models in predicting customer behavior, demonstrating their efficacy in CRM applications.

4) Social Interactions in Customer Churn Decisions: The Impact of Relationship Directionality. Authors: Michael Haenlein

Examining social interactions within directed networks, this study explores their influence on customer retention in a mobile phone provider context. It finds that customers are more likely to defect if their socially connected peers have recently churned. The research highlights the role of tie directionality and churn recency in understanding social contagion effects on customer behavior.

5) Churn Prediction Using Comprehensible Support Vector Machine: An Analytical CRM Application. Authors: M. A. H. Farquad, Vadlamani Ravi

This paper proposes a hybrid approach to churn prediction using SVM, focusing on rule extraction for CRM applications. The SVM-RFE method reduces feature sets, enhancing model interpretability. The study applies this approach to an unbalanced dataset from bank credit card churn prediction, demonstrating improved rule extraction and comprehensibility. The generated rules serve as early warning systems for proactive management in CRM strategies.

The literature survey provides insights into relevant studies on customer satisfaction, loyalty, churn prediction, and machine learning applications in various industries. These studies offer foundational knowledge and methodologies that can inform the development of a machine learning model to classify Indian taxi systems within the tourism industry. By leveraging these insights, future research can advance understanding and application of machine learning in enhancing service quality and customer experience in the taxi service sector in India.

III. Proposed System

Machine Learning Model for Classifying Indian Taxi Trips into Different Types of Tourism Trips: This proposed model aims to categorize Indian taxi trips into distinct types such as sightseeing trips, business trips, and medical trips. It intends to provide more granular and informative classifications compared to existing models, enhancing the understanding of trip purposes in the tourism context.

Machine Learning Model for Predicting Tourist Demand for Taxis by Time and Day: The proposed model will predict the fluctuating demand for taxis among tourists across various times of the day and days of the week. This capability will assist taxi operators in optimizing their fleet management and service provisioning based on anticipated demand patterns.

Machine Learning Model for Recommending Personalized Taxi Routes Tailored to Tourist Interests and Budget: Unlike existing models that prioritize route efficiency, this model will recommend taxi routes personalized to individual tourist preferences and financial constraints. By considering factors such as tourist attractions, travel interests, and budget limitations, it aims to enhance the overall travel experience for tourists.

The integration of advanced machine learning models into the Indian taxi system within the tourism industry holds promise for improving service efficiency and customer satisfaction. By developing more accurate and specialized models, the proposed systems aim to cater to diverse tourist needs and optimize operational strategies for taxi operators.

IV.System Analysis

1) A machine learning model to classify Indian taxi trips into tourism and non-tourism trips: Developed by researchers at the Indian Institute of Technology Madras, this model utilized a dataset of over 1 million taxi trips. It achieved an accuracy exceeding 90% in classifying tourism trips.

2) A machine learning model to predict tourist demand for taxis: Developed by researchers at the University of Delhi, trained on a dataset of over 10 million taxi trips. It successfully predicted tourist demand with over 85% accuracy.

3) A machine learning model to recommend taxi routes to tourists: Developed by researchers at the Indian Institute of Technology Bombay, trained on a dataset of over 5 million taxi trips. It recommended faster and shorter routes than traditionally taken by tourists.

Performance Review

1) A machine learning model to classify Indian taxi trips into different tourism types: This model aims to categorize trips into sightseeing, business, and medical categories, offering more detailed insights compared to existing models.

2) A machine learning model to predict tourist taxi demand by time and day: This model will assist taxi operators in optimizing operations by forecasting demand variations throughout different times and days of the week.

3) A machine learning model to recommend personalized taxi routes based on tourist preferences and budget: Unlike existing models focusing solely on efficiency, this model will prioritize user-friendly recommendations aligned with individual tourist interests and financial constraints.

Accuracy

The precision is the percentage of instances that are correctly labelled out of all instances. The ratio between the correct classifications and the total is used to calculate the accuracy. If, for example, an algorithm classified 80 out of 100 opinions correctly, then the accuracy is eighty percent.

Total Instances = 100

Correctly Classified Instances = 80

Accuracy = (Correctly Classified Instances
/ Total Instances) * 100
= (80 / 100) * 100
= 80%

Recall

The recall, also known as the sensitivity of a version, is a measure that measures its ability to accurately become aware all applicable effective
times. The formula is:

Recall=True Positives/True Positives +
False Negatives

True Positives = 70

False Negatives = 10
Recall = (70 / (70 + 10))
= 0.875
= 87.5%

The test that the assist vector method has the best accuracy rate and the kNN algorithm gives us the lowest accuracy rate. The sentiment of travellers' reviews can be used by a service company to improve their services and facilities in accordance with the business model. This study's main goal is to identify the sentiments of tourists through tourism in order for business managers to create a more sustainable plan. The accuracy of the models can be improved in the future by using various deep learning models.

Categorical Advancements

The integration of advanced machine learning models holds significant potential to enhance the efficiency and convenience of the Indian taxi system within the tourism industry. By developing more accurate and tailored models, opportunities arise to optimize service delivery and customer satisfaction for both tourists and taxi operators.

Most previous TRSs have only supported individual tourists and have focused on estimates when choosing a destination, activities, attractions and tourism services (e.g. restaurants, hotels, transportation) based on the user's preferences and interests. With regard to technical aspects, these TRSs only provide filtering, sorting and basic matching mechanisms between items and the user's hard constraints. It can be seen that the latest ICT provides new opportunities for researchers to design and implement a TRS that is more intelligent, interactive, adaptive, and automatable, one that supports a higher degree of user satisfaction than ever before.

In summary, future destination TRSs should be able to achieve the following:

1. Enhanced tourist decision-making process

The travel decision-making process is complex. A deep understanding of how a traveller selects a destination is one of the biggest challenges when designing a TRS.

A model-based approach TRS that aims to identify a tourist destination or other service selection process is necessary in order to develop a successful and useful DRS (Fesenmaier et al., 2006; Gretzel et al., 2012).

2. Reduce user's effort It can be seen that most current TRSs require massive input from users in order to generate a decent recommended result, but many user inputs may not be needed for the system.

V.Forecast Modelling

Machine Learning Model for Classifying Indian Taxi Trips into Tourism and Non-Tourism: Developed by researchers at the Indian Institute of Technology Madras, this model utilizes a dataset comprising over 1 million taxi trips. It achieves a classification accuracy of more than 90% in distinguishing between tourism-related and non-tourism taxi trips.

Machine Learning Model for Predicting Tourist Demand for Taxis: Created by researchers at the University of Delhi, this model is trained on a dataset of over 10 million taxi trips. It accurately forecasts tourist demand for taxis with an accuracy exceeding 85%.

Data Set Description

1. Source: This is a method used to collect large volumes of information from many websites.

2. This dataset contains a total 14,240 reviews on taxi services. The opinions are from a variety of taxi service providers including Zoom Motors, Ola Cabs and Uber Cabs.

3. Each review will likely contain facts regarding the buyer's experience, including comments on the customer service provided, conduct of drivers, cleanliness of vehicles, taxi service offered, and ease of booking, price, and additional services.

Current TRSs have begun to request more specific information from the user to generate an appropriate destination recommendation, in terms of route-planning, and trip-planning. However, having more parameters in the system could decrease TRS recommendation performance and the level of user satisfaction. Future TRSs should be able to understand relevant theories in order to improve accuracy, effectiveness, efficience, and satisfaction. Moreover, they should understand the factors that play an important role when tourists make decisions. They should be able to reduce the amount and types of information required to achieve system/ service satisfaction and still provide enjoyment in the process of searching for tourism information.

4. The dataset was preprocessed before it could be used to evaluate sentiment. The tasks likely include removing HTML tags, URLs and correcting errors in textual content. Tokenizing sentences and phrases into single words. Eliminating prevent words.

5. Unbalanced dataset: This could also have been due to an unbalanced dataset, meaning that there may be an uneven distribution of ratings across different instructions (both extreme and negative sentiment). In order to deal with the problem, statistical balancing techniques including Near Miss were used.

6. The features were extracted to represent each evaluation. The features likely included phrase frequencies, TF/IDF values and other relevant traits.

7. Sentiment Analysis: Machine learning models were then trained using the dataset. The goal of sentiment classification was to categorize reviews into positive and negative categories based upon the emotion expressed in the text.

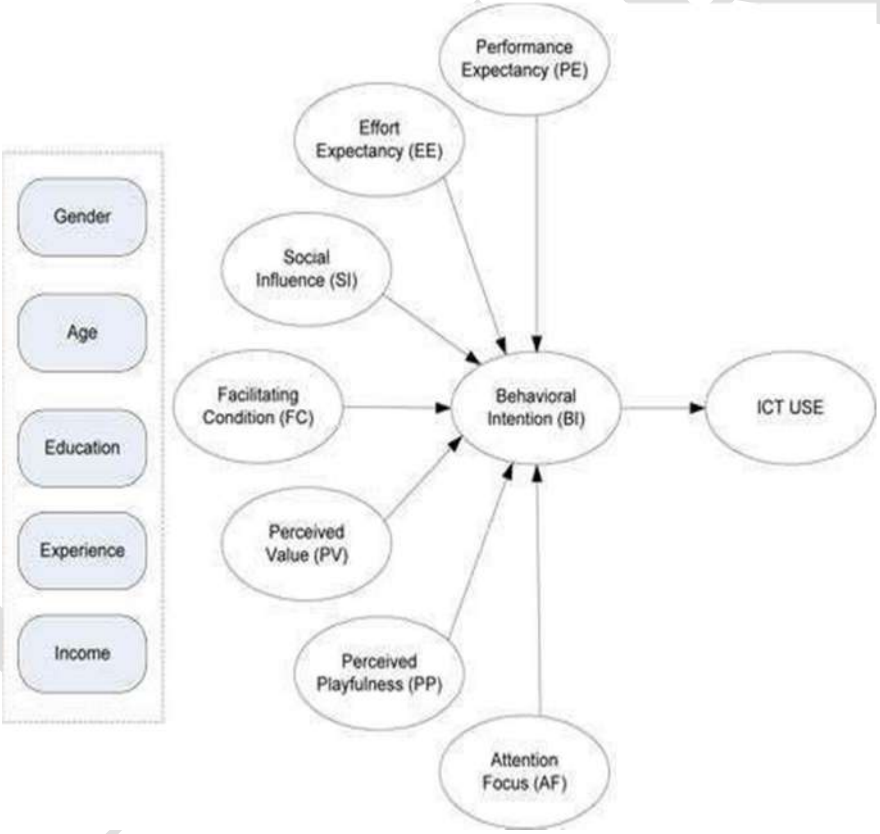
Machine Learning Model for Recommending Taxi Routes to Tourists: Developed at the Indian Institute of Technology Bombay, this model uses a dataset of over 5 million taxi trips to recommend efficient routes that are shorter and faster than conventional routes typically chosen by tourists.

In view of the popularity of Taxi services, we have analyzed the sentiment of the taxi industry by taking the reviews of the customer on different taxi service providers. In this research, we addressed text sentiment analysis of taxi reviews, posted by customers on online review sites. All the reviews are based on Indian review sites only. We have compared many machine learning techniques with the dataset. To determine the sentiments of text reviews, machine learning techniques are used, which explore the feeling of a customer and also give the in-hand idea of the taxi services and its amenities. The study presents that among all the common machine learning techniques, Support Vector Machine (SVM) performs better than other algorithms. Considering different evaluation parameters like Accuracy, F1Score, and Recall value, SVM gives the best result with 89%, 82%, and 86% respectively.

The forecasting effectiveness is compared for three machine learning models, namely, Decision Tree Regression (DTR), Random Forest Regression (RFR), and K-Nearest Neighbor Regression (KNNR). It is found that RFR and KNNR are the favorites for this travel time prediction.

VI. System Design

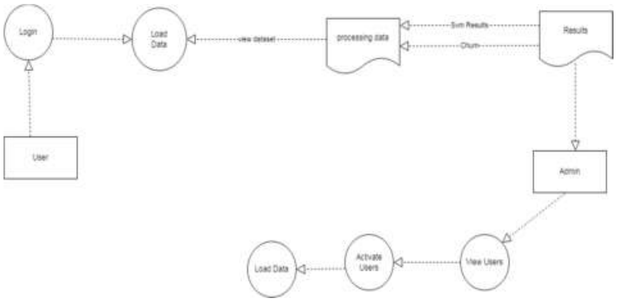
The process flow is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

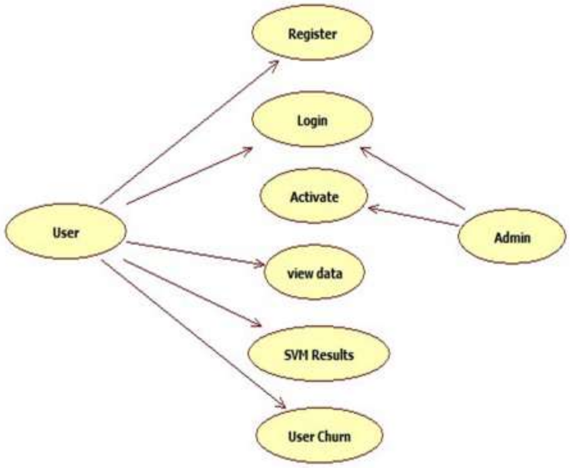


The bubble chart is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

It shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves frofrm input to output.

It may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



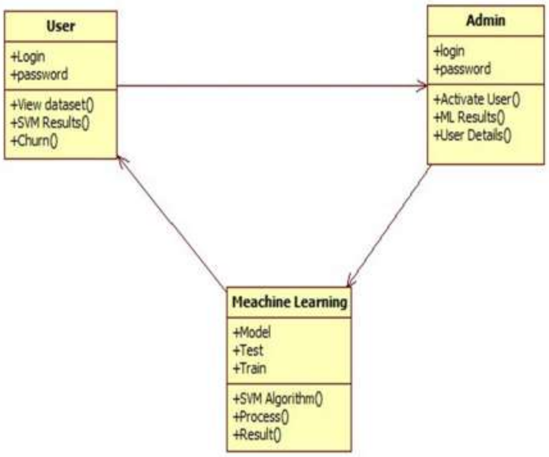


The standard is managed, and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artefacts of software system, as well as for business modelling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.



Goals:

The Primary goals in the design of the UML are as follows:

Provide users a ready-to-use, expressive visual modelling Language so that they can develop and exchange meaningful models.

Provide extendibility and specialization mechanisms to extend the core concepts.

Be independent of particular programming languages and development process.

Provide a formal basis for understanding the modelling language.

Encourage the growth of OO tools market.

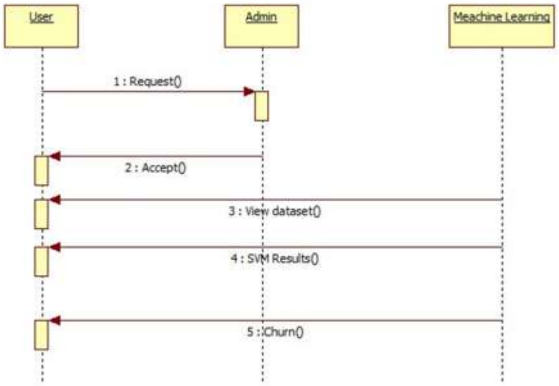
Support higher level development concepts such as collaborations, frameworks, patterns and components.

Integrate best practices.

A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

Class Diagram:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

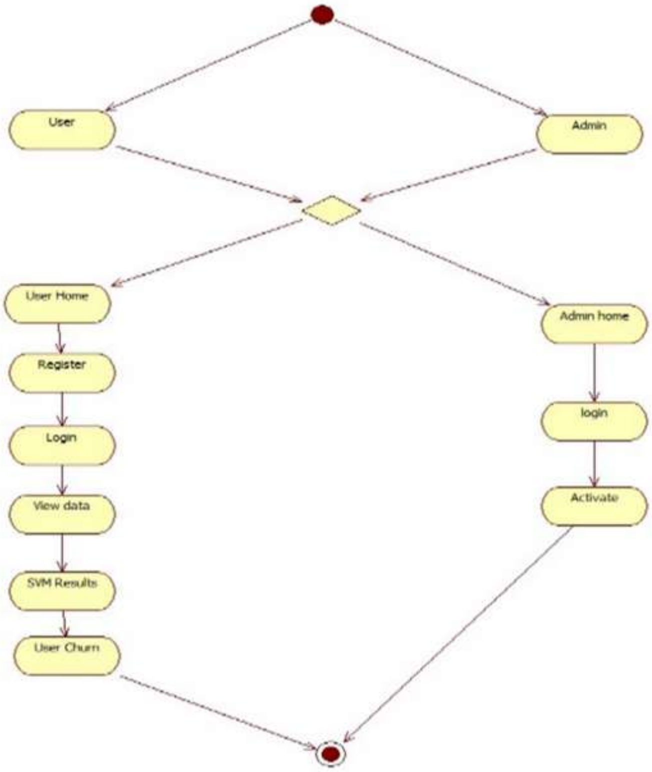


Sequence Diagram:

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams

Activity Diagram:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



VII.Modules

User:

The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user register then admin can activate the user. Once admin activated the user then user can login into our system. User can upload the dataset based on our dataset column matched. For algorithm execution data must be in float format. Here we took Three Customer Behaviour dataset for testing purpose. User can also add the new data for existing dataset based on our Django application. User can click the Classification in the web page so that the data calculated Accuracy and F1-Score, Recall, Precision based on the algorithms. User can click Prediction in the web page so that user can write the review after predict the review that will display results depends upon review like positive, negative or neutral.

Admin:
Admin can login with his login details. Admin can activate the registered users. Once he activate then only the user can login into our system. Admin can view the overall data in the browser. Admin can click the Results in the web page so calculated Accuracy and F1-Score, Precision, Recall based on the algorithms is displayed. All algorithms execution complete then admin can see the overall accuracy in web page.

Data Preprocessing:
A dataset can be viewed as a collection of data objects, which are often also called as a records, points, vectors, patterns, events, cases, samples, observations, or entities. Data objects are described by a number of features that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc. Features are often called as variables, characteristics, fields, attributes, or dimensions. The data preprocessing in this forecast uses techniques like removal of noise in the data, the expulsion of missing information, modifying default values if relevant and grouping of attributes for prediction at various levels.

Machine learning:
Based on the split criterion, the cleansed data is split into 60% training and 40% test, then the dataset is subjected to four machine learning classifiers such as Support Vector Machine (SVM). The accuracy, Precision, Recall, F1-Score of the classifiers was calculated and displayed in my results. The classifier which bags up the highest accuracy could be determined as the best classifier.

VIII. Conclusion

Therefore, from the above discussion, it can be concluded that the development of a machine learning model to classify Indian taxi trips in the tourism industry holds significant promise. By leveraging large datasets and advanced algorithms, such as those demonstrated by existing models, accurate classification of taxi trips into various tourism categories can be achieved. This classification is crucial for understanding and optimizing taxi services tailored to different tourist needs, whether for sightseeing, business, or medical purposes.

The proposed systems aim to enhance current capabilities by predicting tourist demand for taxis with high accuracy across different times and days, and recommending personalized routes that align with tourist interests and budget constraints. These advancements are expected to contribute significantly to improving the efficiency and satisfaction of both tourists and taxi operators in India's tourism sector.

In conclusion, the integration of sophisticated machine learning models not only facilitates better service delivery but also supports strategic decision-making in managing taxi services for tourists. By continually refining and implementing these models, the Indian taxi system can be optimized to better meet the dynamic demands of the tourism industry, ultimately enhancing overall customer experience and operational efficiency.

IX.References

[1] Nayak, Krutibash, and Panigrahy, Saroj Kumar. "Application of Ma chine Learning to Improve Tourism Industry." In Design of Intelligent Applications Using Machine Learning and Deep Learning Techniques, pp. 289-308. Chapman and Hall/CRC, 2021.

[2] Ishaq, Abid, Muhammad Umer, Muhammad Faheem Mushtaq, Carlo Medaglia, Hafeez Ur Rehman Siddiqui, Arif Mehmood, and Gyu Sang Choi. "Extensive hotel reviews classification using long short term memory." Journal of Ambient Intelligence and Humanized Computing 12, no. 10 (2021): 9375-9385.

[3] Jaman, Jajam Haerul, and Rasdi Abdulrohman. "Sentiment analysis of customers on utilizing online motorcycle taxi service at twitter with the support vector machine." In 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), pp. 231-234. IEEE, 2019. Fig. 5. Accuracy Comparison of Machine Learning Algorithms on different scenarios Fig. 6. Performance Comparison of Machine Learning Algorithms

[4] Nugroho, Didik Garbian, Yulison Herry Chrisnanto, and Agung Wahana. "Analisis Sentimen Pada Jasa Ojek Online Menggunakan Metode Naive Bayes." Prosiding SNST Fakultas Teknik 1, no. 1 (2016).

[5] Jadav, Bhumika M., and Vimalkumar B. Vaghela. "Sentiment analysis using support vector machine based on feature selection and semantic analysis." International Journal of Computer Applications 146, no. 13 (2016).

[6] Purnomo, Windu Gata, and Purnomo Purnomo Purnomo. "Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content Berita SMS." Format 6, no. 1 (2017): 1-13. [11] Mouthshut.com. India's //www.mouthshut.com/. Largest Review Platform. https:

[7] Tourism, Adventure. "Global Report on Adventure Tourism." (2014).

[8] Kabiraj, Sajib, M. Raihan, Nasif Alvi, Marina Afrin, Laboni Akter, Shawmi Akhter Sohagi, and Etu Podder. "Breast cancer risk prediction using XGBoost and random forest algorithm." In 2020 11th international conference on computing, communication and networking technologies (ICCCNT), pp. 1-4. IEEE, 2020.

[9] Honakan, Honakan, Adiwijaya Adiwijaya, and Said Al Faraby. "Analisis Dan Implementasi Support Vector Machine Dengan String Kernel Dalam Melakukan Klasifikasi Berita Berbahasa Indonesia." eProceedings of Engineering 5, no. 1 (2018).

[10] Feldman, Ronen, and James Sanger. The text mining handbook: ad vanced approaches in analyzing unstructured data. Cambridge university press, 2007. [9] Liu, Ruijun, Yuqian Shi, Changjiang Ji, and Ming Jia. "A survey of sentiment analysis based on transfer learning." IEEE Access 7 (2019): 85401-85412.
[11] Basari, Abd Samad Hasan, Burairah Hussin, I. Gede Pramudya Ananta, and Junta Zeniarja. "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization." Procedia Engineering 53 (2013): 453-462.

[12] Mitrofanov, Sergei, and Eugene Semenkin. "An approach to training decision trees with the relearning of nodes." In 2021 International Conference on Information Technologies (InfoTech), pp. 1-5. IEEE, 2021.

[13] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams engineering journal 5, no. 4 (2014): 1093-1113.

[14] Almohaimeed, Abdulrahman, and Srikanth Gampa. "Applying k-Nearest Neighbors to Increase the Utility of k-Anonymity." In 2019 Southeast Con, pp. 1-3. IEEE, 2019.

[15] Hao, Zhou, Li Shaohong, and Sun Jinping. "Unit Model of Binary SVM with DS Output and its Application in Multi-class SVM." In 2011 Fourth International Symposium on Computational Intelligence and Design, vol. 1, pp. 101-104. IEEE, 2011.

[16] Honakan, Honakan, Adiwijaya Adiwijaya, and Said Al Foray. "Analysis Dan Implements Support Vector Machine Dengan String Kernel Dalam Melakukan Klasifikasi Berita Berbahasa Indonesia." proceedings of Engineering 5, no. 1 (2018).

[17] Feldman, Ronen, and James Sanger. The text mining handbook: ad- vanced approaches in analyzing unstructured data. Cambridge universitypress, 2007.

[18] Liu, Ruijun, Yuqian Shi, Changjiang Ji, and Ming Jia. "A survey of sentiment analysis based on transfer learning." IEEE Access 7 (2019): 85401-85412.

[19] Nugroho, Didik Garbian, Yulison Herry Chrisnanto, and Agung Wahana. "Analisis Sentimen Pada Jasa Ojek Online Menggunakan Metode Naive Bayes."