PRESIDENCY UNIVERSITY
Private University Estd. in Karnataka State by Act No. 41 of 2013
BANGALORE

A Project Report

On

"PSCS235-A One Stop Solution focusing on Tourism"

Batch Details

Sl. No.	Roll Number	Student Name
1	20211CSE0846	VAISHNAVI C
2	20211CSE0298	SHRUTHI V
3	20211CSE0308	RUTHIKA S SHETTY

School of Computer Science,

Presidency University, Bengaluru.

Under the guidance of,

Ms. Sreelatha P.K

Associate Professor

School of Computer Science,
Presidency University, Bengaluru

CONTENTS

INTRODUCTION

This project aims to create an innovative data mining solution capable of accurately forecasting tourists' travel preferences, specifically for optimizing the scheduling of domestic tour packages.

The project has three main objectives:

1. To build a clustering model that efficiently groups the collected data into distinct clusters for classification purposes.

2. To develop data mining models using predictive techniques that can forecast tourist travel clusters, enabling more effective planning of domestic tour packages.

3. To provide well-founded and actionable recommendations to the appropriate authorities.

LITERATURE REVIEW

The study by Jiantao Wu et al explores the impact of climate change on the tourism economy, a topic not yet fully realized despite increasing climate concerns. Using knowledge graph techniques, including weather data, the study aims to deepen understanding of the relationship between climate and tourism. Findings suggest that organizing climate and tourism data through knowledge graphs can provide valuable insights, potentially enhancing both quality of life and the resilience of the tourism industry. Method includes importing CSV datasets into a Neo4j knowledge graph (KG) using CYPHER's LOAD CSV command. Entities like "Airport" and relationships between "City" and "Weather Station" were mapped, with intermediate CSV files linking "Station" IDs to city names. Key properties, such as geodesic distances, were added to enhance data utility and calculation efficiency within the KG.

The data was collected from various resources like NOAA GHCND, AviationStack, Climateq, Simplemaps **[1]**.

This study by Olimpia Alcaraz et al investigates the intersection of physical and digital realms in tourism, introducing smart tourism destinations (STDs) that leverage technology and open data to enhance visitor experiences and inform decision-making. It demonstrates how integrating open data with local business campaign data can innovate tourism management and foster smart ecosystems through public-private collaboration. An AI-based search engine using word embeddings was developed to identify relevant open data, improving traditional data retrieval. The findings highlight the potential of this integration to enrich tourist experiences and support destination management strategies, contributing insights on combining retail and open data in a real case study.

The initial internal data used in this study are derived from local campaigns known as *bono consumo* (consumer voucher), a promotional campaign resulting from the health crisis caused by COVID-19. The initial private dataset was compiled by APYMECO, the local traders' association, which gathered data on the usage of consumer vouchers in the four editions of the campaign: October 2021, June 2022, September 2022, and November 2022. This dataset comprises more than 300,000 entries **[2]**.

This paper by Saman Forouzandeh et al introduces a novel approach to travel recommendation systems in the tourism industry, combining the Artificial Bee Colony (ABC) algorithm with Fuzzy TOPSIS. The Techniques for Order of Preference by Similarity to Ideal Solution (TOPSIS) is utilized as a multi-criteria decision-making method to optimize recommendations. Data were collected through an online questionnaire from 1,015 respondents on Facebook. In the first stage, the TOPSIS model identifies a positive ideal solution based on four key factors. In the second stage, the ABC algorithm searches for destinations to recommend the best tourist spot to users, enhancing the decision-making process for tourists.

The data was gathered through questionnaires provided to self-driven travelers. The authors distributed a survey to hotel visitors to gather data on the level of service. The data gathered by questionnaires, the exploration of popular topics, and the difficulty of materials were valued [3].

This paper by Tao Peng et al aims to enhance tourism demand forecasting accuracy by integrating social network data with traditional data sources. Using a web crawler, the authors collect social network data and apply sentiment analysis using the BERT model. The study builds a forecasting model based on Gradient Boosting Regression Trees, incorporating structured variables such as weather and holidays. Using Huang Shan as a case study, the authors conduct an empirical analysis comparing the model's performance against existing models, supported by an ablation study. Results indicate that incorporating social network data significantly improves forecasting accuracy for tourism demand.

Social network data acquisition is mainly achieved through web crawlers, which can collect and organize data on the Internet in accordance with established rules [4].

This study by İbrahim Topal and Muhammed Kürşad Uçar explores the growing importance of the tourism and travel sector in the global economy, emphasizing the influence of social media on consumer purchasing decisions. By analyzing historical user data from TripAdvisor, the research aims to employ artificial intelligence methods to identify profiles of consumers likely to prefer Turkey as a travel destination. This approach enables businesses to target the right audience and enhance the effectiveness of their promotional activities. Methods like F-Score Feature Selection Algorithm, classifiers such as Decision trees (DT), k Nearest Neighbors Classification Algorithm (KNN), Multilayer Feedforward Artificial Neural Networks (MLFFNN), Probabilistic Neural Networks (PNN), and Support Vector Machines (SVMs) were used.

The study used the travel data history of Chinese tourists taken from TripAdvisor. The data belong to a total of 624 users. The acquisition of historical data took place between 27 April and 11 May 2018 [5].

Nesreen K. Ahmed et al used models like MLP (Multilayer Perceptron) for classification/regression, RBF (Radial Basis Function) with Gaussian functions, GRNN (Generalized Regression Neural Network) using a Gaussian kernel, KNN (K-Nearest Neighbors) based on nearest neighbors, CART (Classification and Regression Trees) with decision trees, SVR (Support Vector Regression) using support vectors, and GP (Gaussian Processes) modeling data as a Gaussian process. This study explores machine learning methods for tourism demand forecasting, traditionally dominated by models like ARIMA and exponential smoothing. It evaluates the performance of seven machine learning models on Hong Kong’s inbound travel data and examines the impact of adding the time index as an input variable, comparing these models' effectiveness against conventional approaches.

In this study, data published in the study made by Law and Pine to forecast inbound travel demand for Hong Kong was used [6].

The study by Ram Krishn Mishra et al shows the use of SVR and Random Forest Regressor. SVR (Support Vector Regression), adapted from Support Vector Machines, is used for predicting real-number data, offering infinite possible solutions for continuous outputs. Random Forest Regressor is a tree-based model that splits data into nodes, with predictions made by averaging responses in terminal nodes for regression tasks. It improves prediction accuracy and reduces overfitting by constructing multiple decision trees on different sub-samples of the dataset, making it more robust than a single decision tree, which is prone to overfitting due to random noise. This study examines international tourist data from 2010 to 2020, analyzing multiple dimensions to identify valuable features for forecasting. Using Support Vector Regression (SVR) and Random Forest Regression (RFR), the research predicts global tourist arrivals, achieving forecasting accuracies of 99.4% and 84.7%, respectively. The study also addresses the impact of COVID-19 lockdowns on forecasting accuracy.

A substantial amount of data gathered by the government or other public entities is made available. These data sets are referred to as public data since they do not require specific authorization to use them [7].

The study by Noelyn M. De Jesus et al used time series data of tourist arrivals, particularly around the COVID-19 pandemic, splitting the dataset into three partitions for model training and testing. These partitions were based on key events like the first COVID-19 case (January 2020), travel suspensions (March 2020), and stricter entry restrictions (December 2020). The dataset was loaded into the Orange Data Mining tool, and a Multilayer Perceptron (MLP) neural network was used for time series prediction. The model’s performance was evaluated using metrics like MSE, RMSE, MAE, MAPE, and R^2. The best model was selected based on the highest R^2 and lowest MAPE, indicating how well the predictions matched the actual values. This research evaluates the predictive power of an artificial neural network (ANN) model for forecasting tourist arrivals, using tourism data from the Philippines spanning 2008-2022. The ANN was trained on three distinct data compositions and assessed with various time series evaluation metrics, achieving an R-squared value of 0.926 and a MAPE of 13.9%. The study found that including data from unexpected events, like the COVID-19 pandemic, improved model accuracy. The findings suggest that ANN can be a valuable tool for government and tourism stakeholders to support strategic and investment decisions.

The researchers collected the actual inbound tourist arrivals to Philippines between 2008-2022 from the Department of Tourism’s official website [8].

This article reviews machine learning techniques for predicting tourism, specifically analyzing prior studies in this domain. Bilal Sultan Abdualgalil et al discuss various machine learning techniques applied to tourism data analysis, focusing on two primary activities: association learning and classification learning. Key techniques include Logistic Regression and Linear Regression for predicting binary and continuous outcomes, respectively; Decision Trees and Random Forests for supervised classification and regression; Support Vector Machines for binary classification; and Naive Bayes for fast and effective classification. Additionally, KNN is highlighted for its simplicity in classifying data based on nearest neighbors, while K-Means Clustering is used for unsupervised grouping of data. Other methods like Dimensionality Reduction (e.g., PCA) simplify datasets, and Gradient Boosting and AdaBoost improve model accuracy through iterative refinement. The results showed higher prediction accuracy when using the first-quarter dataset, demonstrating its effectiveness for forecasting tourist numbers.

The dataset obtained from www.kaggel.com website was used [9].

This study by Dinda Thalia Andariesta et al presents machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic using multisource Internet data. In this study, data from the Indonesian Statistical Bureau, TripAdvisor, and Google Trends were used to develop prediction models for international tourist arrivals. The process involved data preprocessing, feature extraction, and forecasting model development using ANN, SVR, and Random Forest. These models were evaluated using RMSE, MAE, and MAPE to ensure accuracy. The ANN model used previous tourist data, online posts, and search volumes as predictors. The RF model, known for its reliability, averaged predictions from multiple decision trees to improve forecasting performance.

First, the researchers collected tourism data from the Indonesian Statistical Bureau Indonesia or BPS from January 2017 until June 2021. Next, we collect the data from a global online tourism platform, TripAdvisor [10].

ADVANTAGES AND DISADVANTAGES

1. Improving Tourism Analytics from Climate Data Using Knowledge Graphs

Advantages:

- Enhances tourism analytics by integrating climate and tourism data.
- Improves data integration with multi-source information to model relationships between climate and tourism.

Disadvantages:

- Scalability issues as more data sources are added.
- Limited real-time application focus due to reliance on historical data.

2. Augmenting Retail Data with Open Data for Smarter Tourism Destinations

Advantages:

- Enhances destination management with more strategic decision-making.
- Fosters smart tourism through public-private data integration.

Disadvantages:

- Data access challenges, as open data may not always be compatible or available.
- Privacy concerns with consumer data usage.

3. A Hybrid Method for Recommendation Systems based on Tourism with an Evolutionary Algorithm and TOPSIS Model

Advantages:

- Offers personalized recommendations tailored to user preferences.
- Optimizes travel choices efficiently by combining AI algorithms (TOPSIS and ABC).

Disadvantages:

- High computational demands due to the complexity of algorithms.
- Dependence on accurate, complete user data for high-quality recommendations.

4. Forecast Model of Tourism Demand by Social Network Data

Advantages:

- Captures public sentiment, adding real-time insights.
- Helps with planning during peak tourism times.

Disadvantages:

- Relies heavily on social media, which may not always reflect true demand.
- Data preprocessing is complex and time-consuming.

5. Hybrid Artificial Intelligence Based Automatic Determination of Travel Preferences of Chinese Tourists

Advantages:

- The AI model accurately predicts tourist preferences about 75% of the time.
- Reduces data size with feature selection, saving time and resources.

Disadvantages:

- Only focuses on Chinese tourists, so it doesn't apply to tourists from other countries.
- Relies heavily on TripAdvisor, which might not represent all tourist types.

6. Machine Learning for Tourism Demand Forecasting

Advantages:

- Accurate predictions with machine learning models.
- Handles complex data patterns well.

Disadvantages:

- Requires a lot of computational resources.
- Prone to overfitting with large datasets.

7. Machine Learning-Based Forecasting Systems for Worldwide International Tourist Arrivals

Advantages:

- The study provides high accuracy in predictions, with SVR achieving 99.4% and Random Forest achieving 84.7%.
- It addresses the impact of COVID-19 on tourism and adapts the forecasting models to consider pandemic-related disruptions.

Disadvantages:

- Limited generalizability for countries with missing or unavailable data due to challenges in data collection.
- COVID-19 created a disruption that the model could not entirely overcome, especially given the limitations of pre-pandemic data in adjusting to the sudden changes post-pandemic.

8. AI in Tourism: Leveraging Machine Learning in Predicting Tourist Arrivals in Philippines using Artificial Neural Network

Advantages:

- Utilizes multiple data compositions, improving the model's flexibility in various scenarios, including unexpected events like the COVID-19 pandemic.
- ANN's adaptability makes it valuable for tourism stakeholders for strategic and investment planning.

Disadvantages:

- ANN models require extensive computational resources, which could be challenging in real-time forecasting environments.
- The model's accuracy may vary significantly depending on the choice of data composition and preprocessing, which may not always be clear-cut for stakeholders.

9. Tourist Prediction Using Machine Learning Algorithms

Advantages:

- Offers a comparative study of multiple algorithms, including SVM, Naïve Bayes, and Decision Trees, which can provide insights into the best-performing algorithms under various conditions.
- Effective for identifying trends and patterns using ensemble approaches, which helps improve the robustness of tourism demand forecasts.

Disadvantages:

- The study is limited in geographical scope (focused on India) and may not generalize well to other regions without adjustments for local factors.
- There may be an over-reliance on machine learning techniques without a detailed consideration of socio-economic or event-driven factors, potentially limiting forecast accuracy in real-world scenarios.

10. Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach

Advantages:

- Combines multiple data sources (Google Trends, TripAdvisor) for better accuracy.
- Useful for understanding changes in tourist interest.

Disadvantages:

- Needs strong computational power, limiting accessibility.
- Internet data can be noisy, affecting accuracy.

OBJECTIVES

Objective 1:

Develop a robust clustering algorithm for tourist data segmentation

- Explore and implement various clustering algorithms (e.g., K-means, hierarchical clustering, DBSCAN) to effectively segment tourist data based on relevant attributes (e.g., demographics, preferences, behaviors).

Objective 2:

Construct predictive models for forecasting tourist travel clusters

- Develop and evaluate predictive models (e.g., time series analysis, regression models, classification models) to forecast the evolution of tourist clusters over time.

Objective 3:

Analyze and interpret the results of clustering and predictive models

- Conduct an in-depth analysis of the clusters identified and the predictions generated by the models, identifying key characteristics and trends.

Objective 4:

Develop actionable recommendations for tourism authorities

- Based on the clustering and predictive results, provide specific and actionable recommendations to tourism authorities, such as optimizing resource allocation, developing targeted marketing campaigns, and improving service offerings.

EXPERIMENTAL DETAILS/METHDOLOGY

Software used:

- Python 3.x
- Jupyter Notebook / Google Colab
- Pandas, NumPy (Data processing)
- Matplotlib, Seaborn, Plotly (Visualization)
- Scikit-learn, TensorFlow, PyTorch (Machine Learning)
- XGBoost / LightGBM (Advanced models)
- BeautifulSoup / Scrapy (Web scraping)
- APIs (Google Maps, OpenWeather, etc.)

METHODOLOGY

1. **Data Collection**: Gather data from social media, travel platforms, and user interactions.

2. **Data Pre-processing**: Clean, normalize, and extract relevant features.

3. **NLP**: Analyse text for sentiment, topics, and entities.

4. **Machine Learning**: Build models for recommendation, prediction, and clustering.

5. **Visualization**: Create interactive dashboards for insights.

OUTCOMES

1. Accurate Tourist Segmentation:

- The clustering model is expected to segment tourists based on demographics, preferences, and behaviors effectively. This will allow tourism authorities and companies to better understand different tourist groups and cater to their needs.

2. Improved Forecasting of Tourist Travel Patterns:

- Predictive models developed will help forecast future tourist behaviors and travel preferences, enabling tourism operators to plan more efficiently and anticipate the needs of tourists in real-time.

3. Optimized Tour Package Scheduling:

- By forecasting tourist travel clusters, the project aims to optimize the scheduling and customization of domestic tour packages, increasing efficiency in resource allocation and improving tourist satisfaction.

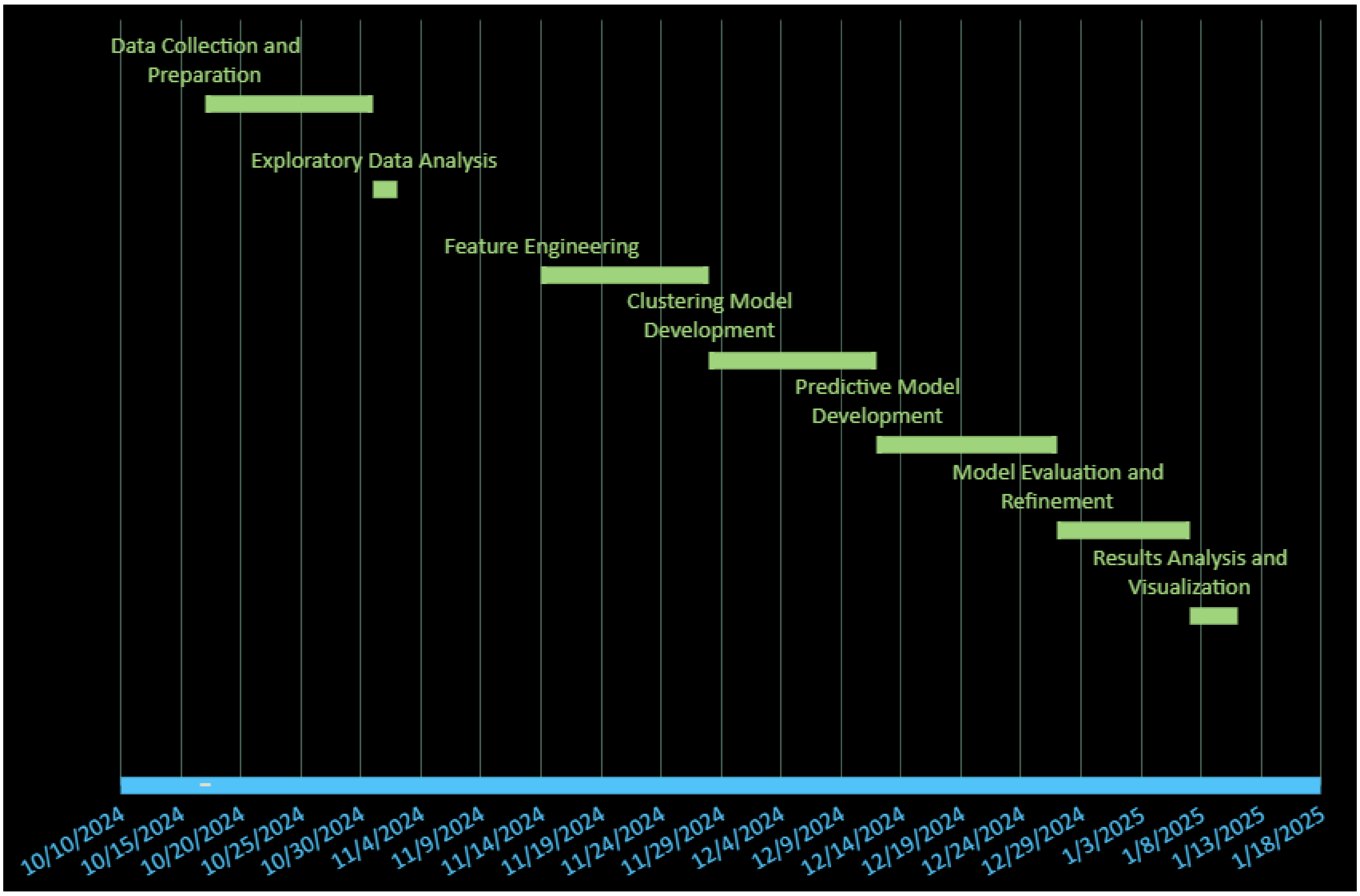
4. Actionable Insights for Tourism Authorities:

- The analysis of the clusters and predictions will provide tourism authorities with actionable insights, such as how to allocate resources, design targeted marketing campaigns, and improve service offerings.

5. Contribution to Sustainable Tourism:

- The project will contribute to sustainable tourism practices by optimizing travel routes and packages, reducing unnecessary travel, and promoting eco-friendly options aligned with sustainability goals.

TIMELINE OF THE PROJECT/ PROJECT EXECUTION PLAN



CONCLUSION

- This AIML project aims to enhance the tourism industry through advanced data analysis techniques. By building an efficient clustering model, we can categorize tourist data into distinct segments, improving understanding of traveler demographics and preferences.

- Additionally, our predictive models will forecast travel patterns, enabling better planning of domestic tour packages tailored to specific tourist needs. The actionable recommendations derived from our analyses will empower tourism authorities to make informed decisions, optimize resource allocation, and promote sustainable practices.

- Ultimately, this project seeks to create a more efficient and responsive tourism ecosystem, fostering growth and sustainability in the industry.

REFERENCES

[1]. J. Wu, J. Pierse, F. Orlandi, D. O'Sullivan and S. Dev, "Improving Tourism Analytics from Climate Data Using Knowledge Graphs," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 2402-2412, 2023, doi: 10.1109/JSTARS.2023.3239831.

[2]. O. Alcaraz, A. Berenguer, D. Tomás, M. A. Celdrán-Bernabeu and J. -N. Mazón, "Augmenting Retail Data with Open Data for Smarter Tourism Destinations," in *IEEE Access*, vol. 12, pp. 153154-153170, 2024, doi: 10.1109/ACCESS.2024.3480326.

[3]. S. Forouzandeh, M. Rostami and K. Berahmand, "A Hybrid Method for Recommendation Systems based on Tourism with an Evolutionary Algorithm and Topsis Model," in *Fuzzy Information and Engineering*, vol. 14, no. 1, pp. 26-50, March 2022, doi: 10.1080/16168658.2021.2019430.

[4]. T. Peng, J. Chen, C. Wang and Y. Cao, "A Forecast Model of Tourism Demand Driven by Social Network Data," in *IEEE Access*, vol. 9, pp. 109488-109496, 2021, doi: 10.1109/ACCESS.2021.3102616

[5]. İ. Topal and M. K. Uçar, "Hybrid Artificial Intelligence Based Automatic Determination of Travel Preferences of Chinese Tourists," in *IEEE Access*, vol. 7, pp. 162530-162548, 2019, doi: 10.1109/ACCESS.2019.2947712.

[6]. Ahmed, Nesreen & Gayar, Neamat & El-Shishiny, Hisham, “Tourism Demand Forecasting using Machine Learning Methods”, 2007.

[7]. Ram Krishn Mishra, Siddhaling Urolagin, J. Angel Arul Jothi, Nishad Nawaz and Haywantee Ramkissoon, “Machine Learning based Forecasting Systems for Worldwide International Tourists Arrival” *International Journal of Advanced Computer Science and Applications(IJACSA)*, 12(11), 2021, 10.14569/IJACSA.2021.0121107

[8]. Noelyn M. De Jesus and Benjie R. Samonte, “AI in Tourism: Leveraging Machine Learning in Predicting Tourist Arrivals in Philippines using Artificial Neural Network” *International Journal of Advanced Computer Science and Applications(IJACSA)*, 14(3), 2023, 10.14569/IJACSA.2023.0140393

[9]. Bilal sultan Abdualgalil and Sajimon Abraham, “Tourist Prediction Using Machine Learning Algorithms”, 2020.

[10]. Dinda Thalia Andariesta, Meditya Wasesa, “Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach”, *Journal of Tourism Futures*, 2022, doi: 10.1108/JTF-10-2021-0239.