Predicting Vacation Preferences Based on Demographics Using Data Mining Techniques

Bedirhan Karaahmetli
Dept. of Computer Engineering
Dokuz Eylul University
Izmir, Turkey
bedirhankaraahmetli@outlook.com

Asst. Prof. Dr. Göksu Tüysüzoğlu
Dept. of Computer Engineering
Dokuz Eylul University
Izmir, Turkey
goksu@cs.deu.edu.tr

***Abstract*—This study focuses on predicting vacation preferences—specifically, whether individuals prefer mountain or beach destinations—based on demographic data using advanced data mining and machine learning techniques. A comprehensive dataset with 13 demographic features was rigorously prepared using feature engineering, scaling, and encoding to ensure optimal model performance. Six machine learning models—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, and K-Nearest Neighbors (KNN)—were implemented and evaluated using metrics such as accuracy, precision, recall, and F1-Score. Among the models, Logistic Regression emerged as the top performer, achieving perfect scores across all metrics, followed closely by SVM and Random Forest. Despite their simplicity, Decision Tree and KNN demonstrated strong performance, while Naive Bayes showed limitations due to its independence assumption. This study highlights the importance of robust preprocessing and model evaluation pipelines and offers valuable insights for leveraging machine learning in personalized tourism services.**

Keywords*—vacation preferences, demographic data, data mining, machine learning, feature engineering, predictive modeling, Logistic Regression, Random Forest, Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbors.

I. Introduction

In recent years, personalizing vacation experiences has gained significant attention, driven by advancements in artificial intelligence and data mining. Demographic data is often used to analyze and predict vacation preferences, enabling tailored recommendations for individuals. This study builds on such foundations to address a real-world problem: predicting whether an individual prefers mountain or beach vacations based on demographic information.

The study contributes to literature by focusing on a demographic dataset with 13 features, leveraging a balanced dataset for equitable model training, and adhering to rigorous preprocessing protocols. The work addresses potential biases and ensures that the data pipeline is well-prepared for subsequent machine learning tasks.

II. Literature review

The field of tourism research has been significant advancements in recent years, particularly in the application of data mining and machine learning techniques to predict and analyze vacation preferences. This literature review aims to synthesize current research on predicting individuals' vacation preferences between mountains and beaches based on demographic data, with a focus on data preprocessing, future engineering, and the application of data mining techniques in tourism preference analysis.

A. Demographic Influences on Vacation Preferences

Demographic factors have long been recognized as significant predictors of tourism behavior and preferences. Recent studies have highlighted the role of demographic characteristics such as age, gender, income, and education level in shaping tourists' destination choices and travel behaviors [1]. These factors can be leveraged to identify patterns and trends among different demographic groups, thereby enhancing our ability to predict vacation preferences [2].

B. Data Mining and Machine Learning in Tourism

The integration of data mining and machine learning techniques has revolutionized predictive analysis in tourism. These technologies allow for the extraction of meaningful patterns from complex datasets, enhancing the accuracy of prediction models [3]. Deep learning techniques have been successfully applied to predict tourist spot preferences, as demonstrated in studies focuses on regions like China [4].

Sentiment analysis, a subset of data mining, has proven particularly useful in gauging tourist sentiments and preferences to tourist destinations in Zhejiang Province by analyzing text data from reviews and social media [5]. This approach provides valuable insights into the emotional responses of tourists to different destinations and services, contributing to a more nuanced understanding of vacation preferences.

C. Data Preprocessing and Feature Engineering

Data preprocessing and feature engineering are critical steps in developing effective predictive models for vacation preferences. These processes involve transforming raw data into a clean and usable format and creating new features to enhance the model's predictive power [6]. The quality and relevance of the features used in a model can significantly impact its performance, making feature engineering a crucial aspect of predictive modeling [7].

Recent advancements in preprocessing methodologies, as discussed by Garcia et al. (2020), propose innovative approaches for handling large-scale demographic data [8]. These developments have the potential to improve the accuracy and reliability of vacation preference prediction models.

D. Gaps and Future Research Directions

While significant progress has been made in applying data mining and machine learning to tourism preference analysis, several gaps and opportunities for further research have been identified:

- Integration of diverse data sources: There is a need to combine data from various sources such as social media, online reviews, and transactional data to provide a more comprehensive understanding of tourist preferences [4].
- Real-time data processing: Developing capabilities for real-time data processing can enhance the responsiveness of tourism services to changing tourist preferences and behaviors [5].
- Personalization and recommendation systems: Further research is needed to improve personalization and recommendation systems in tourism, leveraging advanced machine learning techniques to offer more tailored experiences to tourists [9].
- Exploration of under-investigated demographic factors: Future studies could focus on more granular analyses, such as the impact of cultural background or lifestyle on travel preferences [2].

E. Conclusion and Study Contribution

This study contributes to existing literature by addressing several key aspects of vacation preference prediction:

- It focuses on a demographic dataset with 13 features, providing a comprehensive analysis of the factors influencing vacation choices between mountains and beaches.
- The study leverages a balanced dataset, ensuring equitable model training and addressing potential biases in prediction outcomes.
- By adhering to rigorous preprocessing protocols and employing advanced feature engineering techniques, this research aims to enhance the accuracy and reliability of vacation preference prediction models.
- The integration of data mining and machine learning techniques in this study contributes to the growing body of knowledge on applying these technologies in tourism research.

In conclusion, this literature review highlights the significant potential for predicting vacation preferences using demographic data and advanced analytical techniques. By addressing the identified gaps and leveraging the latest advancements in data preprocessing and feature engineering, this study aims to contribute valuable insights to the field of tourism research and enhance our understanding of the factors influencing vacation choices.

III. Description of the Dataset

The dataset used in this study contains 52,444 instances and 13 features. Each row represents an individual, and the features are designed to capture demographic, lifestyle, and environmental factors influencing vacation preferences. The dataset includes numerical, categorical (nominal and ordinal), and binary features. A summary of the features is as follows:

- **Age** (Numerical): Represents the age of the individual.
- **Gender** (Categorical, Nominal): Includes options like male, female, and non-binary.
- **Income** (Numerical): Reflects the annual income of the individual.
- **Education Level** (Categorical, Ordinal): Indicates the highest educational attainment, such as high school, bachelor’s, master’s or doctorate.
- **Preferred Activities** (Categorical, Nominal): Activities such as hiking, swimming, skiing, or sunbathing.
- **Vacation Budget** (Numerical): Budget allocated for vacations.
- **Location** (Categorical, Nominal): Type of residence (urban, suburban, rural).
- **Proximity to Mountains** (Numerical): Distance in miles from the nearest mountains.
- **Proximity to Beaches** (Numerical): Distance in miles from the nearest beaches.
- **Favorite Season** (Categorical, Nominal): Preferred vacation season (summer, winter, spring, fall).
- **Pets** (Binary): Indicates ownership of pets (0 = No, 1 = Yes).
- **Environmental Concerns** (Binary): Reflects environmental awareness (0 = No, 1 = Yes).

The target feature for the classification task is **Preference,** which indicates whether an individual prefers mountains or beaches as vacation destinations. This binary feature is encoded as follows:

- **0**: Prefers mountains
- **1**: Prefers beaches.

Initial distribution of the target variable can be seen from Fig. 1 below:

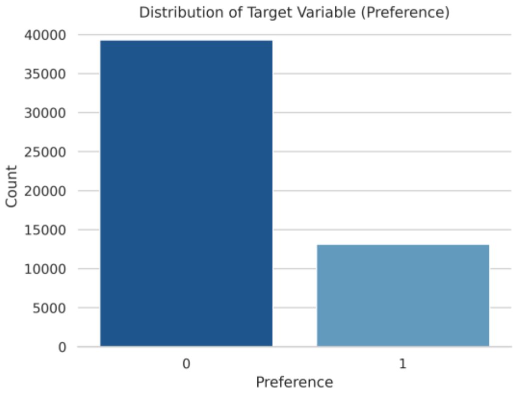

Figure 1 - Distribution of Target Variable

IV. APPLIED DATA PREPARATION TECHNIQUES

A. *Data Inspection*

- **Initial Inspection:** The dataset was examined for structure, missing values, and basic statistical summaries. No missing values were identified, eliminating the need for imputation.

- **Data Types:** The dataset was categorized into numerical, categorical, and binary features for targeted preprocessing.

B. *Statistical Analysis*

The following statistical properties were calculated for numerical features:

- **Mean**, **Median**, and **Mode** to understand central tendencies.
- **Variance** and **Standard Deviation** to evaluate the spread of data.
- **Min** and **Max** values to identify the range of each feature.

C. *Visualization*

The dataset was visualized to understand feature distribution and detect anomalies:

- **Boxplots**: Used for outlier detection in numerical features such as Age, Income, Travel Frequency, Vacation Budget. Boxplots can be seen below from Fig. 2.
- **Histograms**: Illustrated the distribution of numerical features post-normalization. Histograms can be seen below from Fig. 3.
- **Bar plots**: Showed the frequency distributions of categorical and binary features, such as Gender and Environmental Concerns.
- **Heatmap**: Used for selecting a highly correlated features with each other.

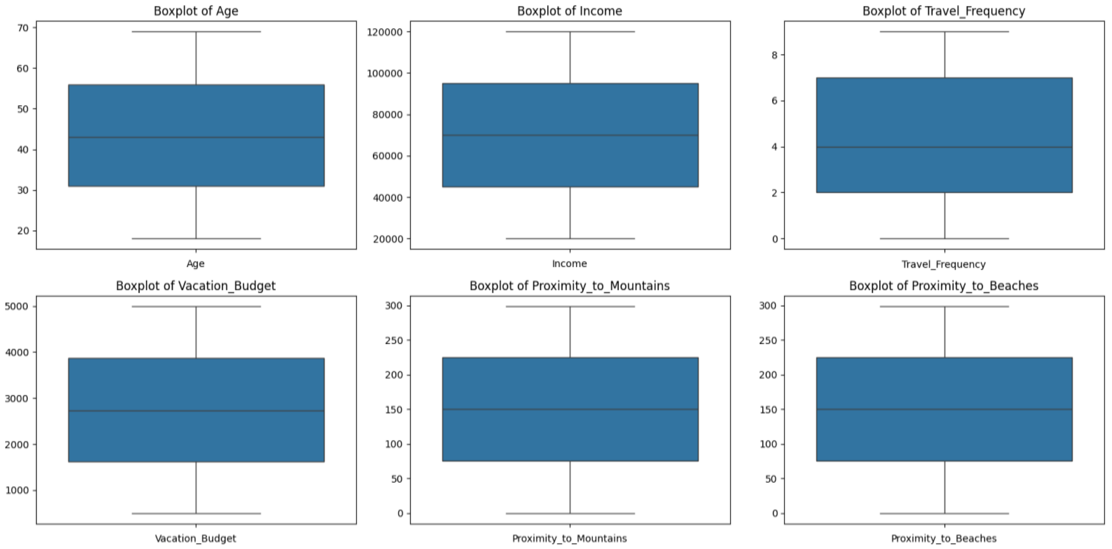

Figure 2 - Boxplots of numerical features

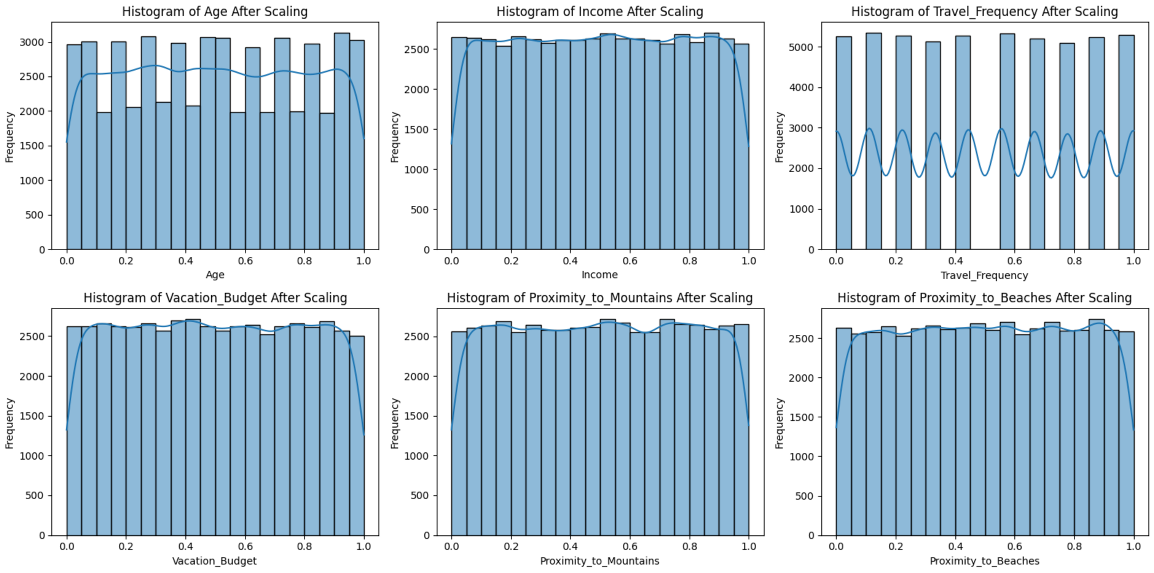

Figure 3 - Histograms for numerical features

D. Outlier Analysis

- **Boxplot Analysis**: No significant outliers were identified, suggesting that extreme values were not present in the numerical features.
- **Transformation**: Numerical features were normalized using **MinMaxScaler**, transforming values into the range of 0 to 1 to improve model performance.

E. Encoding Categorical Features

- **One-Hot Encoding**: Applied to categorical features, such as Gender and Location, to convert them into a numerical format while avoiding dummy variable traps. Following the encoding, the total number of features expanded to 22, enhancing the dataset's dimensionality for improved model performance.

F. Feature Selection

- **Correlation Matrix**: Assessed linear correlations between features to identify redundant attributes. No highly correlated features (correlation > 0.8) were detected. Correlation matrix can be seen below from Fig.4.
- **Recursive Feature Elimination (RFE)**: Used to identify the most influential features for prediction, selecting the top 10 features for modeling.

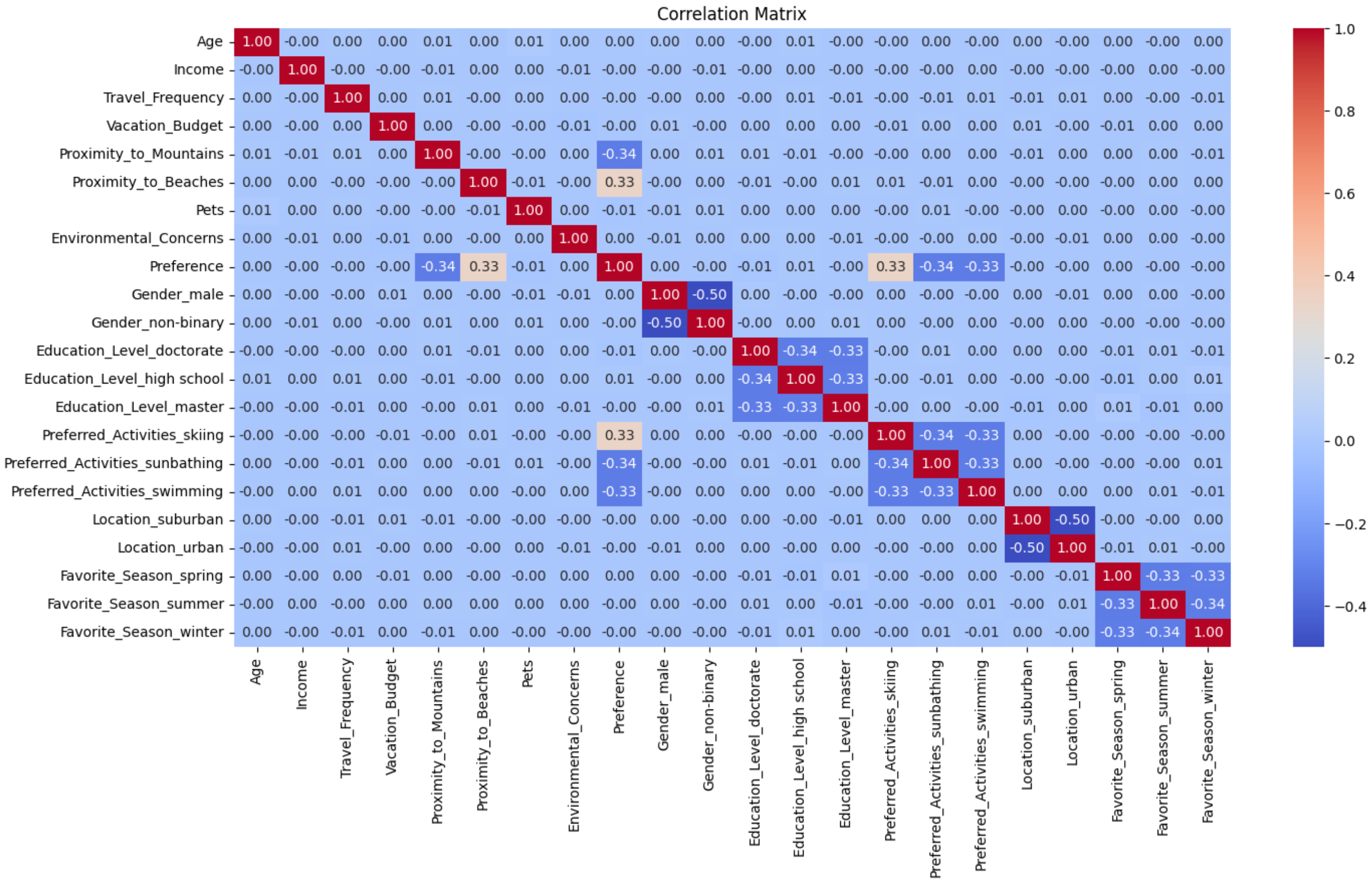

Figure 4 - Correlation matrix

G. Data Splitting

The dataset was partitioned into two primary subsets:

- 10% Testing Data: Reserved exclusively for final model evaluation to ensure unbiased performance assessment.
- 90% Training Data: Utilized for model development, further divided during cross-validation:
    - 1/9 Validation Set: Employed to tune hyperparameters and prevent overfitting.
    - 8/9 Training Set: Used for model training during each fold of the cross-validation process.

This approach ensures a robust evaluation framework while optimizing model performance.

V. Model training and evaluation

A. General Explanation of the Algorithm and Model Training Process

The goal of the algorithm is to build, evaluate, and compare multiple machine learning models for predicting vacation preferences based on demographic data. The process involves structured steps to ensure the models are well trained and their performance is thoroughly assessed:

1) Data Preparation:

The dataset is divided into training and testing subsets, with 90% of the data used for training and 10% reserved for final evaluation.

During training, cross-validation is performed to split the training data into smaller subsets. This approach ensures that the models are validated on different data splits, preventing overfitting and improving generalization.

2) Model Training and Hyperparameter Tuning:

Various machine learning models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and K-Nearest Neighbors (KNN), are implemented.

For each model except Naïve Bayes, hyperparameters are systematically adjusted using grid search to find the optimal configuration that maximizes performance.

3) Cross-Validation:

Each model undergoes cross-validation where the training data is split into multiple folds. Models are trained on subsets of the data and validated on the remaining fold. This process repeats all folds, and performance metrics are averaged to evaluate consistency.

4) Model Evaluation:

After training, each model is tested on reserved testing subsets. Metrics such as accuracy, precision, recall, F1 score, and ROC-AUC are calculated to measure performance.

Visual tools like confusion matrices, ROC curves, and learning curves are used to interpret results and diagnose potential issues like underfitting or overfitting.

5) Model Comparison and Selection

The models are compared based on their performance metrics. The one achieving the highest F1 Score and accuracy is selected as the best model.

6) Prediction on New Data

The best-performing model is retrained on the entire dataset and prepared for predicting preferences for new data instances.

B. How the Models Work and Their Roles

1) Logistic Regression:

Logistic Regression serves as a baseline model, leveraging the linear relationship between features and the log odds of the target variable to predict binary outcomes.

- **Hyperparameters**: L1 (lasso) and L2 (ridge) regularization were tuned through the penalty (alpha) parameter to prevent overfitting and improve generalization.
- **Best Parameters**: penalty= L1 (lasso), C= 10.
- **Performance**: Achieved the best results among the models, delivering perfect scores across metrics due to its ability to model linearly separable data effectively.

The confusion matrix for the Logistic Regression model is presented in Fig. 5:

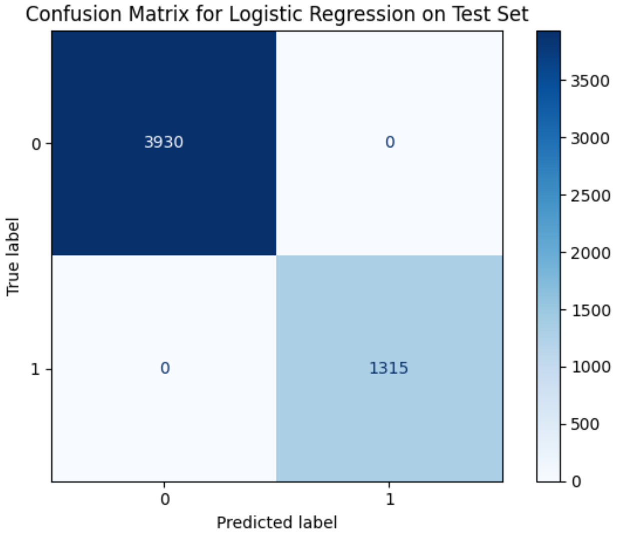

Figure 5 - Confusion matrix for logistic regression

2) Decision Tree:

Decision Trees classify data by creating a hierarchical structure of decisions based on feature splits, allowing for intuitive and interpretable predictions.

- **Hyperparameters**: Maximum depth of the tree (max_depth), and the minimum number of samples required to split an internal node (min_samples_split).
- **Best Parameters**: maximum depth = None, minimum number of samples = 5.
- **Performance**: While Decision Trees offer simplicity and interpretability, they are prone to overfitting, particularly on training data. However, they delivered competitive results in this analysis.

The confusion matrix for the Decision Tree model is shown in Fig. 6:

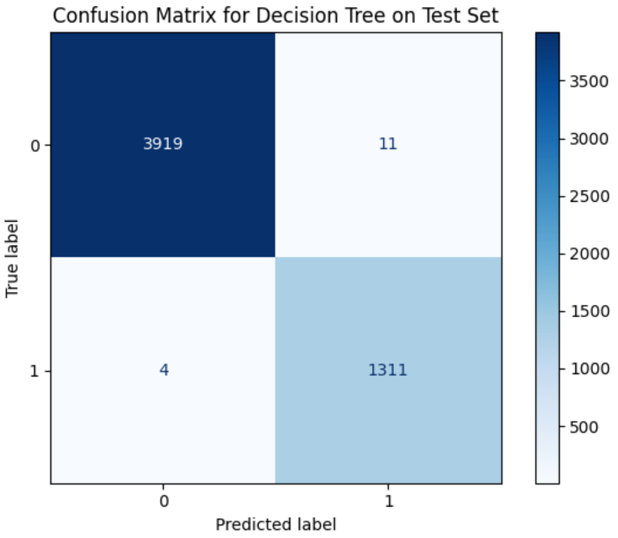

Figure 6 - Confusion matrix for decision tree

3) *Random Forest:*

Random Forest is an ensemble method that aggregates multiple Decision Trees to improve accuracy, reduce overfitting, and enhance generalization.

- **Hyperparameters**: Number of trees in the forest (n_estimators), maximum depth of each tree (max_depth), and the minimum number of samples required for splitting (min_samples_split).
- **Best Parameters**: Number of the decision trees in the forest = 100, maximum depth of each tree = None, minimum number of samples = 5.
- **Performance**: Delivered outstanding results, achieving metrics greater than 99%, demonstrating its robustness and reliability in handling complex data.

Confusion matrix of the random forest model can be seen below from Fig. 8:

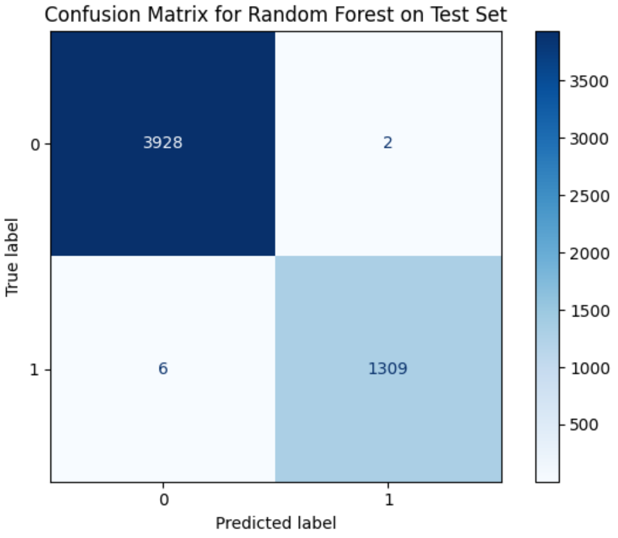

Figure 7 - Confusion matrix for neural network

4) *Support Vector Machine:*

Support Vector Machine (SVM) seeks to identify the optimal hyperplane that separates data points in feature space, making it highly effective for classification tasks.

- **Hyperparameters**: Kernel type (kernel), including linear and Gaussian, and penalty parameter (C) controlling the trade-off between margin size and classification error.
- **Best Parameters**: Kernel = linear, C = 1.
- **Performance**: Effective for linearly separable data, delivering excellent results, though computationally intensive for larger datasets.

The confusion matrix for the SVM model is shown in Fig. 9:

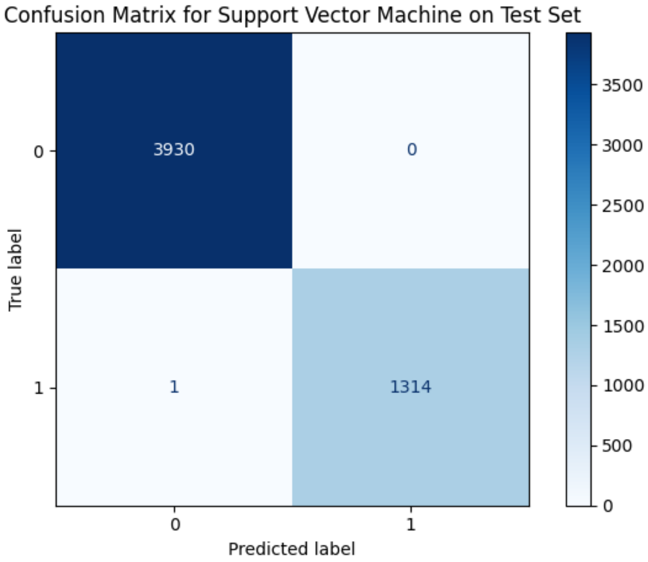

Figure 8 - Confusion matrix for SVM

5) *Naïve Bayes:*

Naïve Bayes is a probabilistic model based on Bayes' theorem that assumes independence between features, simplifying computation and making it efficient to implement [10].

- **Hyperparameters**: No specific hyperparameters are available for Gaussian Naïve Bayes, at it is a parameter-free model.
- **Performance**: Demonstrated weaker results compared to other models due to its strong independence assumption, limiting its ability to capture complex feature interactions in the dataset.

Confusion matrix of the random forest model can be seen below from Fig. 7:

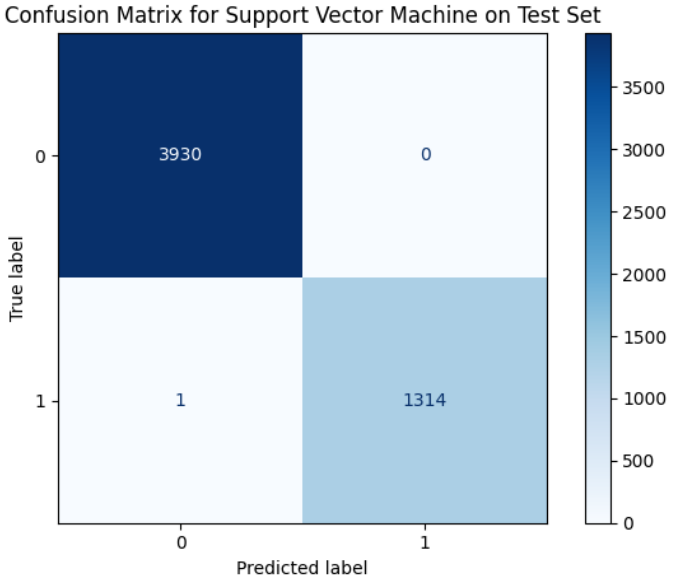

Figure 9 - Confusion matrix for random forest

6) *K-Nearest Neighbors:*

K-Nearest Neighbors (KNN) classifies data points based on the majority vote of their nearest neighbors, making it a non-parametric, instance-based learning algorithm

- **Hyperparameters**: Number of neighbors (k) tested with values from 3 to 9, and weighting (weights) options including uniform and distance-based weights.
- **Best Parameters**: k = 9, weights = distance.
- **Performance**: Performed effectively for small datasets, but computationally expensive for larger datasets due to the distance calculations involved.

Confusion matrix of k-nearest neighbors model can be seen below from Fig. 10:

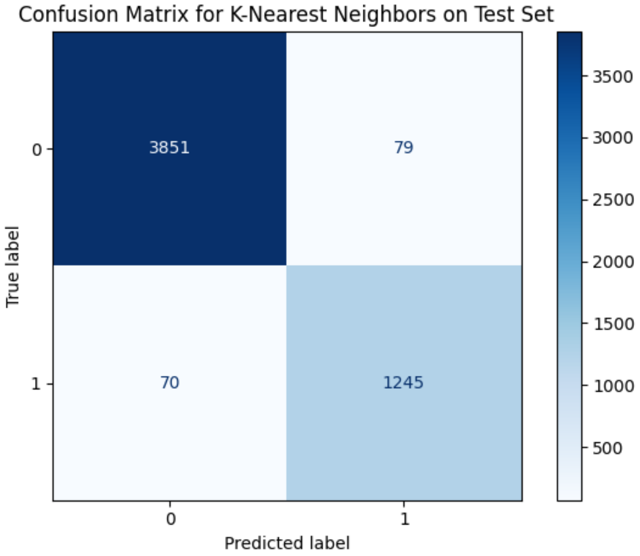

Figure 10 - Confusion matrix for KNN

C. *Comparative Analysis*

The models were evaluated on a holdout test set (10% of the data). The summarization of the models can be seen below from Fig. 11.

This comparison examines the performance of six machine learning models—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, and K-Nearest Neighbors (KNN)—across four evaluation metrics: Test Accuracy, Test Precision, Test Recall, and Test F1-Score.

Logistic Regression emerged as the top-performing model, achieving perfect scores across all metrics, demonstrating its capability to effectively capture the linear relationships in the dataset. Support Vector Machine (SVM) followed closely, with near-perfect scores, indicating its robustness in handling high-dimensional data and complex patterns. Random Forest also delivered excellent results, showcasing its ability to generalize well across the data due to its ensemble-based structure.

Decision Tree performed slightly lower than Random Forest but still achieved high accuracy and F1-Score, reflecting its effectiveness in splitting data hierarchically while managing overfitting to a certain extent. K-Nearest Neighbors (KNN) delivered satisfactory performance, particularly in smaller datasets, although its results were marginally lower compared to ensemble and kernel-based methods.

Naive Bayes, in contrast, demonstrated weaker performance compared to other models, particularly in accuracy and recall, which can be attributed to its strong assumption of feature independence. This limitation made it less effective for capturing complex dependencies within the dataset.

Overall, Logistic Regression stands out as the best-performing model, followed closely by Support Vector Machine and Random Forest. These results highlight the strengths of simpler linear models and robust ensemble or kernel-based methods, while also showcasing the limitations of simpler probabilistic models like Naive Bayes. The analysis provides a clear understanding of the balance between simplicity, interpretability, and predictive power.

VI. Conclusion

This study successfully demonstrated the application of data mining and machine learning techniques to predict individuals' vacation preferences based on demographic data. By leveraging a well-structured dataset and rigorous preprocessing protocols, six machine learning models were systematically trained, evaluated, and compared across key performance metrics, including accuracy, precision, recall, and F1-Score.

The results identified Logistic Regression as the top-performing model, achieving perfect accuracy, precision, recall, and F1-Score. Support Vector Machine and Random Forest closely followed, showcasing their robustness and ability to generalize effectively across complex datasets. Decision Tree and K-Nearest Neighbors provided strong yet slightly lower performance, with the latter showing limitations in large-scale datasets. Naive

Bayes, while computationally efficient, exhibited the weakest performance, likely due to its simplifying assumptions about feature independence.

The research underscores the value of a comprehensive data preparation pipeline, including feature engineering, normalization, and hyperparameter tuning, to enhance the performance of predictive models. The findings also demonstrate the effectiveness of linear, ensemble, and kernel-based approaches for addressing classification problems in structured demographic data.

Future work could extend this study by incorporating more diverse data sources, such as social media or transactional data, to capture additional factors influencing vacation preferences. Real-time prediction systems and personalized recommendation frameworks could further enhance the application of these models, particularly in the tourism industry.

In conclusion, this study provides a solid foundation for understanding how demographic data can inform vacation preference predictions. The insights gained emphasize the potential of machine learning models to drive data-driven decision-making in personalized tourism and marketing strategies, offering valuable contributions to both academia and industry.

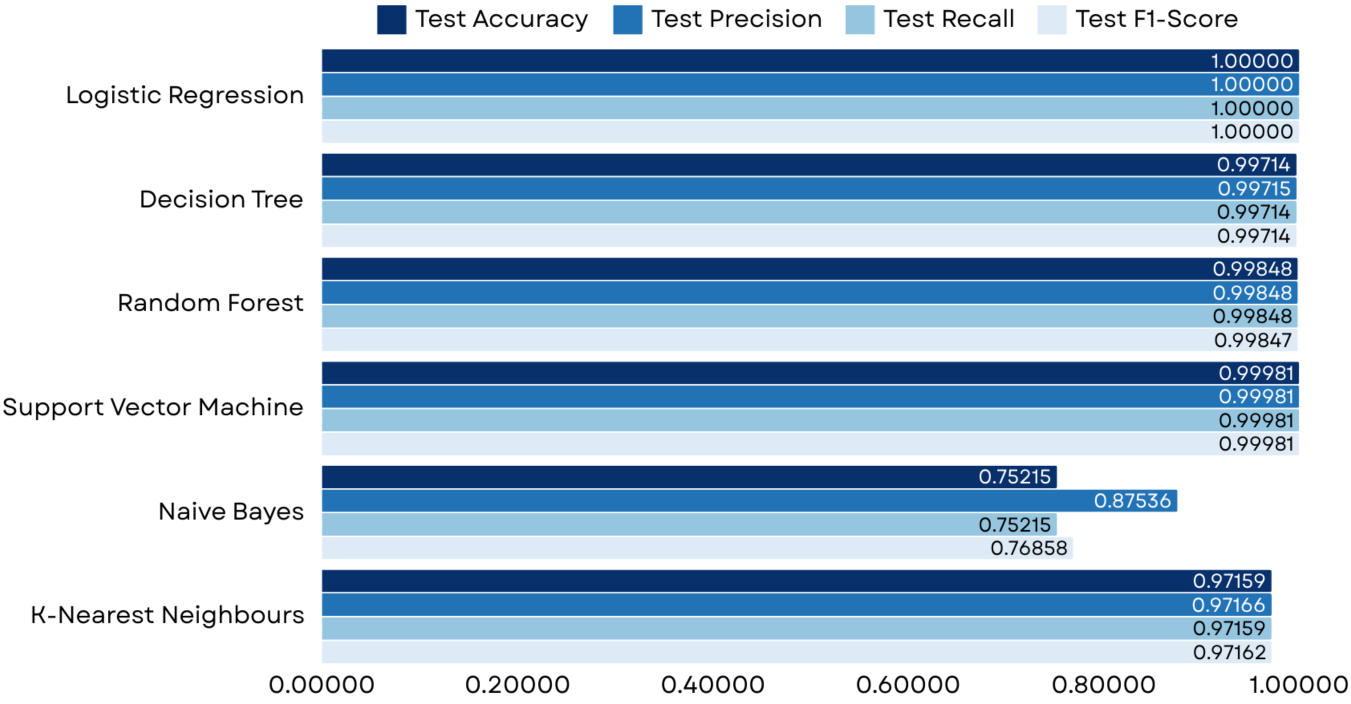

Figure 11 - Summarization of the metrics

References

[1] L. Cohausz, C.-M. De, Andrej. Tschalzev@uni-Mannheim. De, and H. Stuckenschmidt, “Investigating the importance of demographic features for EDM-predictions,” doi: https://doi.org/10.5281/zenodo.8115647.

[2] N. S. Kara and K. H. Mkwizu, “Demographic factors and travel motivation among leisure tourists in Tanzania,” International Hospitality Review, vol. ahead-of-print, no. ahead-of-print, Apr. 2020, doi: https://doi.org/10.1108/ihr-01-2020-0002.

[3] Nuno António, Ana de Almeida, and L. Nunes, “Data mining and predictive analytics for e-tourism,” Springer eBooks, pp. 1–25, Jan. 2022, doi: https://doi.org/10.1007/978-3-030-05324-6_29-1.

[4] L. Zuo, V. Savin, and Y. Wu, “Insight analysis of tourists’ behavior and preference by deep learning in tourism destination management,” pp. 99–103, Dec. 2023, doi: https://doi.org/10.1109/ici3c60830.2023.00029.

[5] F. Bi, “Analysis of tourism review information based on data mining technology,” Proceedings of the 2nd International Conference on Mathematical Statistics and Economic Analysis, MSEA 2023, May 26–28, 2023, Nanjing, China, 2023, doi: https://doi.org/10.4108/eai.26-5-2023.2334275.

[6] V. A. M, “Data preprocessing and feature engineering,” *Linkedin.com*, Jun. 16, 2024. https://www.linkedin.com/pulse/data-preprocessing-feature-engineering-vishnu-a-m-ppgjc/

[7] N. Babalık, “Importance of feature engineering in data preprocessing,” *Medium*, Sep. 26, 2024. https://medium.com/@nrmnbabalik/importance-of-feature-engineering-747065b4b0c7 (accessed Nov. 17, 2024).

[8] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, “Big data preprocessing: methods and prospects,” Big Data Analytics, vol. 1, no. 1, Nov. 2016, doi: https://doi.org/10.1186/s41044-016-0014-0.

[9] B. McKercher, “A dream vacation typology,” *International Journal of Tourism Research*, vol. 26, no. 3, May 2024, doi: https://doi.org/10.1002/jtr.26

[10] T. Bayes, “An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S,” *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418, Dec. 1763, doi: https://doi.org/10.1098/rstl.1763.0053.