Machine Learning Pre-Project
Classification of Users Holiday Preferences

ESILV A4 – DIA 3

Jérémie FORMEY DE SAINT LOUVENT, Alexis DUCROUX,
Terence FERNANDES, Arthur DA COSTA VIEIRA

1 Business Challenge and State-of-the-Art

The business challenge we chose to address is the classification of user preferences between **mountains vs. beach** vacations. The objective is to utilize machine learning models to detect a user's preference based on various personal data.
With the rise of personalized marketing, companies in travel, tourism, and related industries are increasingly looking to enhance customer satisfaction by tailoring experiences to individual preferences.

State of the Art

Predicting "mountains vs. beaches preferences" involves combining machine learning and personalization techniques widely used in tourism. Recent approaches focus on models like *logistic regression*, *decision trees*, and *support vector machines* for simpler data, while *deep learning models* (e.g., CNNs, RNNs) are applied for complex inputs like images and trip sequences.
Clustering techniques, such as *k-means*, are used to segment users into groups like "adventure seekers" or "relaxation seekers." Hybrid recommendation engines, combining collaborative and content-based filtering, improve personalization and address the "cold start" issue for new users.
Data enrichment techniques help generate context-aware recommendations, while evaluation metrics include both traditional ML metrics (e.g., accuracy, precision) and business metrics (e.g., CTR, ROI), ensuring robust insights to maximize user satisfaction and conversion rates.

2 Data Description and Sources

To implement this project, we will use a dataset called *Mountains vs. Beaches Preferences*, available on Kaggle.

Dataset Overview

The dataset aims to analyze public preferences between two popular vacation types: mountains and beaches. It provides insights into various demographic and lifestyle factors that may influence these preferences.
Data Characteristics:

- Type: Supervised learning problem (target feature present).
- Data Types: Mixed (categorical and numerical).
- Instances: 52,444
- Features: 13

The dataset is provided on Kaggle. It has been synthetically generated data to practice machine learning concepts.

3 Business Objectives and Project Scope

Business Objectives

The objective is to develop a predictive model that identifies whether users prefer mountain or beach destinations, which would support:

- **Personalized Recommendations**: Providing custom suggestions for trips, activities, or accommodations, improving booking conversion rates and customer satisfaction.
- **Targeted Marketing**: Data-driven advertising campaigns directed at specific user segments (e.g., families, adventure-seekers) to optimize marketing expenditure.

Scope of the Project

The project's success metric will be the **accuracy of predictions**. We will evaluate our model using various techniques, with the primary audience being:

- **Primary**: Travel agencies, hospitality providers, tourism marketers.
- **Secondary**: E-commerce sites, retail, and lifestyle brands.

4 Model Spectrum and Evaluation Techniques

This project involves a **classification problem**. We plan to test several models:

- **Decision Tree Classifier**: A commonly used classification model.
- **Support Vector Machine (SVM)**: Effective for small datasets with multiple variables.
- **Naive Bayes Classifier**: Useful for categorical data.
- **K-Nearest Neighbors (KNN)**: A clustering algorithm efficient on small datasets.
- **Ensemble Learning**: Using *Random Forest Classifier*.
- **Artificial Neural Network (ANN)**: Exploring deep learning for comparison.

Note: we will probably try some other algorithms.
We will evaluate models using the *Confusion Matrix*, *ROC Curve*, and *Stratified K-Fold Cross-Validation*.

5 Work Plan

The project is divided into **three parts**, with a deliverable at the end of each:

1. **First Solution**: Includes exploratory data analysis (EDA), data cleaning, visualization, and preprocessing. Each team member will explore different models to identify the best-performing ones.
2. **Second Solution**: Improving the solution by refining the models selected before, implementing additional algorithms, optimizing hyperparameters, and refining data preprocessing as needed.
3. **Third Solution**: Enhancing the solution by implementing innovative models (e.g., Neural Networks), followed by a final comparison and selection.
4. **Final part**: we will prepare the final report and a video presentation of the project.

Each team member will implement, test, and evaluate at least one algorithm. We will collaborate using **GitHub** or Google Colab. The final submission will include a notebook, a comprehensive report, and a video presentation.

6 Conclusion

In conclusion, we have decided to address the business case of classifying user preferences between beaches and mountains. The objective is to provide personalized recommendations and targeted marketing strategies. The dataset used is available on Kaggle, providing ample opportunity to apply machine learning techniques.

References

- Kaggle: Mountains vs. Beaches Preferences Dataset