ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Pre-Project Report Web scraping:

- Analysis and Prediction of Air Travel Q1 2025-

Supervised by:

Dr. MELLOULI Nédra

Realized by:

Anis HENTIT – DIA 2

Hamza HALINE – DIA2

Project Context:

The aviation sector is one of the most dynamic and competitive industries, where travelers are constantly seeking solutions to optimize their flight choices in terms of price, duration, and convenience. Given the diversity of options available in the market, it is essential to provide analytical tools that help users make informed decisions. This project aims to address this need by automating the collection of flight data, analyzing it to extract valuable trends, and making it accessible through an interactive interface.

Project Objective:

The project aims to create a comprehensive solution for collecting flight data through scraping, structuring this data, and applying advanced Machine Learning or NLP techniques. The goal is to extract valuable insights, such as price variations based on days or airlines, and provide predictions to help travelers plan their trips. An interactive web application, developed using Streamlit, will enable users to explore the data and receive personalized recommendations.

Data Collection:

In this project, two distinct methods were used to collect the necessary data for flight analysis. These complementary techniques ensure maximum coverage and quality of the data.

1) **Web Scraping with Selenium:**
   The first method involves the use of Selenium, a Python library that enables interaction with dynamic web pages. By accessing the Kayak website, we simulated human navigation to extract information available from flight search results. This method allowed us to collect data such as: Departure date, Origin and destination airports, Departure and arrival times, Ticket price, Seat type offered…

2) **Using the Amadeus API:**
   To complement the data collected via scraping and ensure a more reliable and structured source, we integrated the Amadeus API. This method provides official data directly from reservation systems. Based on parameters such as departure date, origin airport, and destination airport, the API provided:
   - Detailed information about available flights,
   - Ticket prices and fare options,
   - Flight durations, number of stops, and airline information.

3) **Combining Both Methods:**
   These two techniques will be used jointly to ensure data collection that is both broad and precise. Data from scraping provides flexibility and extensive coverage, while the Amadeus API ensures reliability and standardization. The results from both methods will be

consolidated into a uniform format, enabling in-depth analysis and efficient exploitation of the collected information.

Data Preprocessing:

The raw data collected through scraping requires cleaning and structuring to make it suitable for machine learning algorithms. This process involves removing missing or erroneous values and normalizing formats, such as converting flight durations into minutes and ticket prices into numerical values. Additionally, categorical variables, such as airline names or seat types, are transformed into numerical representations using techniques like One-Hot Encoding. These steps ensure the data is consistent, well-structured, and ready for analysis and predictive modeling.

Machine Learning Application:

To enhance the user experience, we will develop a flight cost prediction system using machine learning techniques. This system will focus on accurately estimating the price of a flight based on various features such as the flight date, origin and destination airports, airline, duration, and number of stops. For this purpose, we will utilize models like Random Forest Regressor and Gradient Boosting Machines (e.g., XGBoost), which are highly effective for handling tabular data with non-linear relationships. Additionally, we will experiment with Neural Networks for more complex patterns in the data. These models will be trained and evaluated using processed flight data, enabling us to predict costs for new flight queries with high accuracy. The insights derived from these predictions will help users make informed decisions and optimize their travel budgets.

User-Friendly Application with Streamlit:

The final application will be developed using Streamlit, offering a seamless and interactive interface for users. The platform will allow users to input their preferences, such as the flight date, origin, and destination airports, directly into the app. Based on these inputs, the system will predict the cost of the flight using the implemented machine learning models. Users will also be able to view detailed explanations of the factors influencing the price, helping them understand how different features affect costs. This intuitive interface will make it easy for users to explore data, access predictions, and receive real-time cost estimates, ensuring an engaging and user-centric experience.