**REPUBLIC OF TUNISIA Ministry of Higher Education and Scientific Research**



**PRIVATE HIGHER SCHOOL OF TECHNOLOGY & ENGINEERING**

**Computer Engineering Training Program for Engineers**

**BI-ML Project Report**

**Prediction of Fake Reservations for British Airways**

**Made By**

TORKHANI Oumayma

KHALED Ghassen

BEN HMIDA Mohammad Ali

TOKOGMON Djavph David

**Academic Year :** 2022/2023

**Table of Contents**

Contents

[Introduction 6](#_Toc132624318)

[Chapter 1: Introduction 7](#_Toc132624319)

[1.1 Background 7](#_Toc132624320)

[1.2 Objectives 7](#_Toc132624321)

[1.3 Problem Statement 7](#_Toc132624322)

[1.4 Web Scraping and Dataset 7](#_Toc132624323)

[1.5 Definition of Business Intelligence 7](#_Toc132624324)

[1.6 Definition of Machine Learning 8](#_Toc132624325)

[1.7 Steps for Developing a Machine Learning Model 8](#_Toc132624326)

[1.8 Scrum Methodology 8](#_Toc132624327)

[1.9 CRISP-DM for Machine Learning 9](#_Toc132624328)

[1.10 Conclusion 9](#_Toc132624329)

[Chapter 2: Business Intelligence (BI) 10](#_Toc132624330)

[2.1 Definition and Quick Differentiation of BI 10](#_Toc132624331)

[2.2 Data Source 10](#_Toc132624332)

[2.3 Data Preparation 10](#_Toc132624333)

[2.4 Creating Database Objects 11](#_Toc132624334)

[2.4.1 ETL Source to Staging 11](#_Toc132624335)

[2.4.2 ETL Staging to DW 11](#_Toc132624336)

[2.5 Semantic Layer 11](#_Toc132624337)

[2.5.1 SSAS Tabular and Multidimensional 11](#_Toc132624338)

[2.5.2 Creating Power BI Report 11](#_Toc132624339)

[2.6 Authoring 12](#_Toc132624340)

[2.7 Presentation 12](#_Toc132624341)

[2.8 Conclusion 12](#_Toc132624342)

[Chapter 3: Machine learning (ML) 13](#_Toc132624343)

[Conclusion 13](#_Toc132624344)

**List of Figures**

**No table of figures entries found.**

**List of Tables**

**No table of figures entries found.**

# Introduction

In the airline industry, fake reservations can cause a lot of problems for both airlines and passengers. To prevent this, we worked on a project to use Business Intelligence techniques to analyze the data of British Airways' reservation system and predict the likelihood of a reservation being fraudulent. By using machine learning and data analysis tools, we were able to collect and analyze data from multiple sources to identify patterns and predict fraudulent behavior. In this project, we focused on the steps of data collection, cleaning, modeling, analysis, and visualization. The insights we gained from this project can help British Airways and other airlines to detect and eliminate fraudulent activities, which will improve the customer experience and increase the trust of their customers in the airline industry.

This report is divided into four parts: Introduction, Business Intelligence, Machine Learning, and Conclusion.

- In the Introduction section, we provide background information on the project and discuss the definitions of machine learning and Business Intelligence.

- The Business Intelligence section focuses on data collection, cleaning, and analysis.

- The Machine Learning section covers the development and evaluation of the machine learning model.

- In the end we will get a Conclusion that summarizes the key findings and provides recommendations for detecting and eliminating fraudulent activities in the airline industry.

# Chapter 1: Introduction

## 1.1 Background

In the airline industry, fake reservations can cause a lot of problems for both airlines and passengers. Fraudulent activities such as fake bookings, invalid credit cards, and identity theft can lead to significant losses for the airlines and inconvenience for the passengers. Therefore, it is crucial for airlines to detect and eliminate fraudulent activities to improve the customer experience and increase customer trust.

## 1.2 Objectives

The main objective of this project is to develop a machine learning model to predict the likelihood of a reservation being fraudulent for British Airways. Specifically, the project aims to:

•Collect and preprocess data from multiple sources

•Develop a machine learning model to classify reservations as fraudulent or legitimate

•Evaluate the performance of the model using appropriate metrics

•Provide insights and recommendations based on the results

## 1.3 Problem Statement

The problem addressed in this project is the need for British Airways to detect and prevent fraudulent activities in their reservation system. The current methods of manual checking and random audits are time-consuming and not efficient enough to detect all fraudulent activities. Therefore, the airline needs an automated system that can analyze data and identify potential fraudulent activities with high accuracy and efficiency.

## 1.4 Web Scraping and Dataset

To collect data for this project, we will use web scraping to extract information from online booking platforms. One example of a website that we can use to obtain our dataset is Expedia.com. We will scrape data such as flight details, passenger information, and booking history. We will also collect additional data from internal sources within British Airways. The collected data will be preprocessed and used to train and evaluate the machine learning model.

## 1.5 Definition of Business Intelligence

Business Intelligence (BI) refers to the technologies, applications, and practices used to collect, integrate, analyze, and present business information. BI tools are designed to help organizations make informed decisions by providing insights into their operations, customers, and competitors. BI involves the use of data mining, data analysis, and data visualization tools to identify patterns, trends, and anomalies in data.

## 1.6 Definition of Machine Learning

Machine learning is a subfield of artificial intelligence that focuses on the development of algorithms and statistical models that enable computer systems to improve their performance on a specific task by learning from data, without being explicitly programmed. The process of machine learning involves feeding a large amount of data to a model, which learns from the data and improves its performance over time.

## 1.7 Steps for Developing a Machine Learning Model

Developing a machine learning model involves several steps, including:

1.Data collection: This involves gathering relevant data from various sources, such as databases, APIs, and web scraping.

2.Data preprocessing: This involves cleaning, transforming, and organizing the data to prepare it for analysis.

3.Feature selection: This involves identifying the most relevant features that can help the model make accurate predictions.

4.Model selection: This involves selecting the appropriate machine learning algorithm that can best solve the problem.

5.Model training: This involves feeding the data to the model and adjusting its parameters to optimize its performance.

6.Model evaluation: This involves testing the model on a new dataset and measuring its performance using appropriate metrics.

7.Model deployment: This involves integrating the model into a production environment and making it available for use.

## 1.8 Scrum Methodology

Scrum is an agile project management framework that emphasizes collaboration, flexibility, and continuous improvement. In Scrum, the project is divided into short iterations called sprints, which typically last between two to four weeks. During each sprint, the team works on a set of tasks and delivers a potentially shippable product increment. Scrum also emphasizes regular team meetings, including daily stand-up meetings, sprint planning meetings, sprint review meetings, and sprint retrospective meetings. The Scrum framework provides a flexible and iterative approach to project management, which can help teams respond quickly to changing requirements and deliver high-quality products.

## 1.9 CRISP-DM for Machine Learning

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a popular methodology for developing machine learning models. It consists of six phases:

Business understanding: In this phase, the problem to be solved is defined and the project objectives are established.

Data understanding: In this phase, the data is collected and analyzed to identify patterns, trends, and potential issues.

Data preparation: In this phase, the data is cleaned, transformed, and organized to prepare it for analysis.

Modeling: In this phase, the machine learning model is selected and trained on the data.

Evaluation: In this phase, the model is evaluated on a new dataset to assess its performance and identify areas for improvement.

Deployment: In this phase, the model is integrated into a production environment and made available for use.

CRISP-DM provides a structured and systematic approach to machine learning that can help ensure the accuracy, reliability, and usability of the model. By following the CRISP-DM methodology, machine learning teams can increase their chances of success and reduce the risk of project failure.

## 1.10 Conclusion

In this chapter, we introduced the background and objectives of the project, as well as the problem statement and definitions of machine learning and Business Intelligence. We also discussed the importance of web scraping and data sources in collecting data for the project. In the next chapter, we will discuss the methodology used in this project, including the steps taken to collect, clean, and analyze the data.

# Chapter 2: Business Intelligence (BI)

## 2.1 Definition and Quick Differentiation of BI

Business Intelligence (BI) refers to the tools, techniques, and applications used to collect, integrate, analyze, and present business data. BI solutions are designed to help organizations make data-driven decisions and gain actionable insights into their operations. The primary objective of BI is to provide business users with timely and accurate information that can help them make informed decisions.

The BI process consists of several steps, including data collection, data integration, data analysis, and data presentation. Each step in the BI process requires specialized skills and tools to ensure that the data is accurate, consistent, and useful to decision-makers.

## 2.2 Data Source

The first step in the BI process is data collection, which involves gathering data from various sources and storing it in a centralized location. In this project, we used two different data sources: Kaggle and British flight reviews.

The Kaggle dataset contained information on airline passenger satisfaction ratings, including attributes such as in-flight service, check-in, onboard experience, and arrival and departure procedures. This dataset was collected from various airlines worldwide, including major and low-cost carriers.

The second data source was British flight reviews, which we obtained by web scraping online review websites. This dataset contained detailed information on flight routes, flight numbers, flight durations, flight frequencies, and flight ratings.

Both datasets were collected in Excel format, and we used Excel to store and manipulate the data before importing it into our data warehouse.

## 2.3 Data Preparation

Data preparation involves transforming raw data into a format suitable for analysis. This process typically includes data cleaning, data integration, and data transformation.

In our project, we performed several data preparation tasks, including :

Data cleaning : We removed any duplicates, missing values, or irrelevant data from our datasets.

Data integration : We merged the two datasets using a common key, which was the flight route.

Data transformation: We converted the data into a format suitable for our data warehouse and BI tools.

## 2.4 Creating Database Objects

After the data preparation phase, we created the necessary database objects for our data warehouse. This involved designing the schema, creating tables, and defining relationships between them.

### 2.4.1 ETL Source to Staging

We then used ETL (extract, transform, load) processes to transfer the data from the source systems to the staging area of our data warehouse. This involved extracting the data from the Excel files, transforming it to fit the data warehouse schema, and loading it into the staging area.

### 2.4.2 ETL Staging to DW

Next, we used ETL processes to move the data from the staging area to the data warehouse proper. This involved extracting the data from the staging area, transforming it further if necessary, and loading it into the appropriate tables in the data warehouse.

## 2.5 Semantic Layer

The semantic layer is a layer of abstraction between the physical data model and the business user. It provides an intuitive and consistent view of the data to the business user, shielding them from the complexities of the underlying data structure. The semantic layer is implemented using a variety of tools, including SQL Server Analysis Services (SSAS) Tabular and Multidimensional.

### 2.5.1 SSAS Tabular and Multidimensional

SSAS is a Business Intelligence tool from Microsoft that provides OLAP (Online Analytical Processing) and Data Mining functionality for business users. It has two main modes: Tabular and Multidimensional. In the Tabular mode, the data model is created using a relational model and is optimized for in-memory processing. In the Multidimensional mode, the data model is created using a dimensional model and is optimized for disk-based processing.

### 2.5.2 Creating Power BI Report

Power BI is a business analytics tool from Microsoft that provides interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards. Creating a Power BI report involves connecting to the data source, selecting the appropriate data elements, and designing the report layout. Power BI provides a range of visualizations, including charts, tables, and maps, that can be customized to meet the needs of the user.

## 2.6 Authoring

Authoring involves designing and developing the Business Intelligence solution, including the data model, reports, and dashboards. It requires a deep understanding of the business requirements, the underlying data sources, and the tools being used to develop the solution. Effective authoring involves a collaborative effort between the Business Intelligence team and the business stakeholders to ensure that the solution meets the needs of the business.

## 2.7 Presentation

Presentation is the final stage of the Business Intelligence process and involves delivering the reports and dashboards to the end users. This can be done through a variety of channels, including email, web portals, and mobile devices. Effective presentation requires a deep understanding of the end users and their needs, as well as an understanding of the technology being used to deliver the reports.

## 2.8 Conclusion

In conclusion, Business Intelligence is a critical process that enables organizations to transform data into actionable insights. It involves a range of activities, including data source identification, data preparation, semantic layer development, report authoring, and presentation. Effective Business Intelligence requires a collaborative effort between the Business Intelligence team and the business stakeholders to ensure that the solution meets the needs of the business. In the next chapter, we will discuss the Data Mining process, which is a key component of Business Intelligence.

# Chapter 3: Machine learning (ML)

# Conclusion