A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Francisca Andrea Argandona Alvarado |
| *Student Number* | 2024247 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA 2 Capstone Report |
| *Assessment Due Date* | 15th December 2024 |
| *Date of Submission* | 15th December 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Introduction 1](#_Toc185153895)

[Objectives 2](#_Toc185153896)

[Problem Definition 2](#_Toc185153897)

[Methodology 3](#_Toc185153898)

[Business Understanding 3](#_Toc185153899)

[Data Preparation 4](#_Toc185153900)

[Data Collection 4](#_Toc185153901)

[Description and Exploration of the data 4](#_Toc185153902)

[Data Preparation 8](#_Toc185153903)

[Data construction and integration 8](#_Toc185153904)

[Data Selection 8](#_Toc185153905)

[Data Cleaning 8](#_Toc185153906)

[Modelling 9](#_Toc185153907)

[Evaluation 9](#_Toc185153908)

[Timeline 10](#_Toc185153909)

[Ethics and Bias 10](#_Toc185153910)

[References 11](#_Toc185153911)

[GitHub Link 11](#_Toc185153912)

# Introduction

The economy of a country can be influenced by different sectors such as manufacturing, agriculture, construction, and tourism, among others.

One of the countries that has been positively affected by tourism is Iceland. It is a Nordic country located in the Atlantic Ocean. The variety of its landscapes, like geysers, volcanoes, hot springs, and glaciers, catch the attention of people all over the world, regardless of which season it is.

In the last decades, Iceland's economy has been diversifying into manufacturing and service industries, where tourism belongs. Although it was hugely negatively affected by the COVID-19 pandemic, the number of tourists increased even more after it.

As we can see in the following graphic that shows the “Percentage change in the number of international visitor arrivals to Iceland from 2010 to 2022”, the international tourism volume has experienced every year growth of 146% in 2022.

Tourism in Iceland was not a big topic to talk about 10 years ago. Due to the development that this area has grown up too fast, it can be important for many travel agencies to know the preferences of the tourists that visit the island, depending on how long they stay, how many people they travel with, how long in advanced they book their flights and accommodations, among other things, a personalised package can be arranged for them so they can get a better experience and recommend this destination to some other people to support even more the economy’s country.

A graph with blue and white bars

Description automatically generated

# Objectives

In the beginning the main purpose of the following assignment was to predict how many tourists will visit Iceland per year in the next 5 years, nevertheless it will be changed to predict how likely the visitors from the most common countries that travel to the island, would recommend it to friends, family or colleagues, and what factors depends on their decision.

The reason of the objective changed is due to one of the variables to predict the future visitors is the price of the flights. That information depends on each country the tourists are from, the airlines and season. The timeline for this project is not enough to collect that information.

## Problem Definition

## 

The impression people get of a place where they went for holidays is important when part of the economy of the country depends on it. This is the case of Iceland.

One way to know if people liked the experience and services there is to know if they would recommend it to friends or family, and why are the reasons they would do it. Having the knowledge of this information, travel agencies can prepare a better trip package for the future visitors depending on their nationality, their incomes, how long they will stay, etc.

A diagram of data mining

Description automatically generated

# Methodology

The tools use for this assignment will be Jupyter Notebook to visualise the data using different plots and statistical summarize and to analyse and evaluate the model using python code.

For this project CRISP-DM will be implemented as methodology to solve the problem. The followings steps describe each stage of the project.

## Business Understanding

The landscapes, nature and culture of this country have caused an increase of the tourism industry getting attention from people all over the world. From the volcano hiking to northern light hunting, the variety of attractions this country has is diverse.

Due to the fast development of the tourism, the experience on this field may not be the best one. There are different aspects why people did not enjoy their stay, and normally when people do not like something it is very unlikely that they recommend it.

If they can get a good experience and they are likely to go back, that will be a positive result for the economy of the country in the future.

Finding the pattern why people would recommend Iceland as a country for holidays is an advantage for the tourism.

Due to the timeline for this project is short, I am going to analyse the behaviour of just people who come from 10 different countries.

## Data Preparation

### Data Collection

The source of the data collected for this assignment will be [Ferðamálastofa Icelandic Tourist Board](https://www.ferdamalastofa.is/en), which is an independent authority under the Ministry of Culture and Business Affairs. One of the ways they collect the data is through surveys among domestics and foreign tourists all year round to passengers who arrive at Keflavik International Airport.

### Description and Exploration of the data

The data collected belongs to surveys made in 2023, 2022, 2016, 2015, 2014 and 2013. The data are presented in excel and pdf and they are all in percentages. The data has not missing values, but it has some questions that were not answered from some visitors. All the data has just numerical values.

To analyse the data, I will select to 10 countries which are repeated in every year of the data collected.

Exploratory Analysis Data (E.D.A.) will be apply on this stage to get a better understanding of the data, I will visualise the data using different plots to see what kind of distribution each feature has, also boxplot will be used to see if there are any outliers. A statistical summarize will be also use to understand if the data is spread or not and to know the measures of location and dispersion.

As a first approach I can see the percentages of people who would recommend Iceland in 2016 (first picture) and 2022 (second picture). In 2016 the minimum of very likely was 65.1% from British and the maximum was 91.7% from Norwegian. In 2022 the minimum was 54.7% from China and the maximum was 78.5% from Sweden. It is possible to say that the percentage of people who recommend Iceland has decreased.

If I compared by the mentioned nationality on those years, I could see there is a difference in the results for people who would recommend Iceland very likely. Three of the countries increase the percentage meanwhile one (Norway) decrease it.

|  |  |  |
| --- | --- | --- |
| **Year/ Nationality** | **2016** | **2022** |
| **United Kingdom** | 65.1% | 75.4% |
| **USA** | 75.7% | 76.3% |
| **China** | No information | 54.7% |
| **Norway** | 91.7% | 68% |

A screenshot of a computer

Description automatically generated

A table with many small people

Description automatically generated with medium confidence

Another approach is the top nationalities who visit Iceland. In 2016 the top 3 countries were USA, Germany and Canada. In 2022 were USA, United Kingdom and Germany. There are also new countries which are in the list of 2022 which were not in 2016, like China.

Most of the people who visited Iceland are from European countries, but the country which has most of the tourist who went to Iceland is USA in both years.

A group of flags with numbers

Description automatically generated

A screenshot of a graph

Description automatically generated

Regarding the length of the stay in Iceland, in 2016 does not seem that there is a tendency as it is shown in 2022 table where the most repeated are 4-5 nights and 6-8 nights. USA, that is the country with more visitors in Iceland increase the number of nights they spent in the country.

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

## Data Preparation

### Data construction and integration

The first step on this stage will be to join the selected data in one csv file to get a better understanding of it in the Jupyter Notebook.

### Data Selection

The information that will be extract from the excel and the pdf will be the nationality, reason for trip, factors influencing the decision to go to Iceland, how long was it since they first considered go to Iceland, if they spend in a pre-paid package trip, which region of Iceland they visited and how long they stayed there, which kind of accommodation they stayed in, how many people they were traveling with, what mean of transport was the most used in their stay, how satisfied they were on different aspect regarding to tourism, which recreational activities they did in Iceland, household income and how likely they are to recommend Iceland as a travel destination to a friend, family member or colleague.

### Data Cleaning

Like it was mentioned before, there are no missing values to handle of any of the data collected. Due to the data has not many observations, if there are any outliers they will be kept. Because there are no categorical variables, there is no necessity to label it. Due to the whole data is in percentage, I will keep the data with the original numbers not using any kind of scale. If there are any duplicated data, it will be removed.

## Modelling

Because of the objective of this assignment if to predict how likely a tourist would recommend Iceland, which will be the target or dependant variable, I will use a supervised model. The model selected for this assignment will be k-neighbours, neural network and finally random forest to see the feature importance. All of them will be a classification problem, where the classes in the target can be low (0-6), medium (7-8) and high (9-10) where 0 is very unlikely to recommend and 10 very likely to recommend.

The data will be split into 3 testing sizes 10%, 20% and 25% for each model. To ensure that the model is not affected by overfitting, cross validation will be applied using different number of folders until get the one with the best amount of accuracy.

To improve the results of each model, GridsearchCV will be used to find the best hyperparameters and evaluate and train again the model with those hyperparameters.

After evaluating the models, I will plot the confusion matrix to get the amount of true positive, true negative, false positive and false negative for each class and which model has the best accuracy, precision and recall and finally I will make a comparison between each model and their results.

If the random forest gets good results, I will plot a histogram to see the feature importance and to know which are the most important that affect how likely the tourist are to recommend Iceland.

In case that none of the mentioned model have good results, different supervised classification algorithms will be implemented, using the same techniques, cross validation, test- training set and hyperparameters using GridsearchCV.

## Evaluation

After comparing the results of each model, I will choose the one that accomplish better the objective of the business criteria. A table comparing the results will be built to show a better understanding of them.

Because there is no boundary for the results of this assignment, I will set an accuracy of minimum 60% as a good result, with a precision and recall also with a minimum of 60%.

If it is needed, go back on previous steps to get better results may be a good idea in case the 60% is not reach. Understand if more data should be collected or use different techniques in the data preparation is required.

After analysing every step and being sure to probe that the results can get any better, the deployment can be implemented.

# Timeline

This assignment has a timeline of two semesters, from September 2024 to May 2025, where in the first semester the topic will be selected, followed by the methodology applied. A deep investigation of the business problem to solve and the collection of the data to solve and predict further results from it will be also applied. An overview of the data to understand how it will be use.

The second semester will be for the data understanding in depth using python code. Data preparation will be the next step followed by the evaluation of the models and finally the presentation of the full project and results.

A Gantt chart was made to get better results and show how long each step should take so there are no delays in the project. The Gantt chart is attached in the GitHub file.

# Ethics and Bias

It is important to analyse any ethical consideration that can affect the result of the project.

Personal information is not required for this project.

Being aware of past events such as, economic crises, pandemics, or natural disasters is important to understand that it can affect the results. New international treaties or visas can affect the prediction of new data for future predictions.

It is important to understand that the results will not have an accuracy of 100%, in other words the new predictions may not be always correct.

The data collected was applied to people who arrived in Iceland through the airport, and it does not considered people who arrived by sea.

# References

Bjarki Bents­son, J. (2024). *Will tourist numbers rise this year?* [online] www.islandsbanki.is. Available at: <https://www.islandsbanki.is/en/news/will-tourist-numbers-rise-this-year>.

Dirección General de Comunicación, Diplomacia Pública y Redes de España (2024). *Islandia Islandia OFICINA DE INFORMACIÓN DIPLOMÁTICA FICHA PAÍS*. [online] Available at: <https://www.exteriores.gob.es/Documents/FichasPais/ISLANDIA_FICHA%20PAIS.pdf>.

Ferðamálastofa (2020). *Icelandic Tourist Board*. [online] Icelandic Tourist Board. Available at:

<https://www.ferdamalastofa.is/en>.

Government of Iceland (n.d.). *Government of Iceland | Tourism in Iceland*. [online] www.government.is. Available at: <https://www.government.is/topics/business-and-industry/tourism-in-iceland/>.

López, A.M. (2024). *Change in international tourism in Iceland 2022*. [online] Statista. Available at: <https://www.statista.com/statistics/694121/tourism-industry-international-visitor-growth-iceland/>.

## GitHub Link

<https://github.com/CCT-Dublin/capstone-project-Pancha19.git>