A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* |  |
| *Student Number* | 2024247 |
| *Module Title* |  |
| *Assessment Title* |  |
| *Assessment Due Date* |  |
| *Date of Submission* |  |

**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.

Chapter 1: Introduction

* 1. Motivation (background domain) 1 page
  2. Research problem
  3. Research questions
  4. Research hypothesis (Ho, Ha)
  5. Research objectives

paragraph

Chapter 2: Literature review/ related works

2.1 Technologies you are using (machine learning)

Paragraph (summarise your related work)

Chapter 3: Methodology

Write one or two sentences that describe the aim of your project.

Add architectural diagram of your project.

3.1 Dataset information

3.2 Data analysis and preprocessing

3.3 Model training (3 models)

3.3.1 Neural networks

3.3.2 Random Forest

3.4 Model evaluation

Chapter 4: Results and Discussion

4.1 Results from ED

4.2 Classification result from machine learning algorithms

4.3 Regression result from machine learning algorithms

Paragraph (discuss your result and compare with research hypothesis)

Chapter 5: Conclusion (brief about research problem, aim, solution, research, contribution to the body of knowledge, and future work).

Contents

[Introduction 1](#_Toc158384946)

[Chapter 1 1](#_Toc158384947)

[Chapter 1.1 1](#_Toc158384948)

[Chapter 1.1.1. 1](#_Toc158384949)

[References 2](#_Toc158384950)

# Introduction

## Motivation

“Over the last decade, Iceland has experienced an unprecedented tourism boom, establishing itself as a highly attractive destination due to its unique natural beauty, culture, and safety”, as noted by Johannesson et al. (2010, p.278), “tourism in Iceland has evolved from a marginal activity to a central component of the national economy”. This growth in tourism has brought significant economic opportunities, but also important challenges related to strategic planning. Understanding the profiles of visitors travelling to Iceland non only allows for the improvement of tourism services but also supports the implementation of more effective public policies. In this context, data analysis has become an important tool for transforming information into applicable knowledge. Techniques such as clustering enable the segmentation without the need for pre-labelled data, helping to uncover hidden patterns. Applying these methods to the study of tourists visiting Iceland can offer new insights into their behaviours, thus contributing to more informed decision-making by industry stakeholders.

## Research Problem

Currently, there is limited profiling of tourist visiting Iceland. Available information often focuses on general statistics by country of origin, without deeper segmentation based on variables such as age, length of stay, and income. This lack of segmentation limits the ability of tourism institutions to develop targeted marketing campaigns and promote a more personalised experience.

## Research Question

Given the above problem, the research question which will guide this research is: How can clustering techniques be applied to identify relevant tourist segments visiting Iceland, based on age, length of stay and average income?

## Research Hypothesis

Null Hypothesis (H0): There are no clear segmentation patterns among tourists based on the analysed variables.

Alternative Hypothesis: (Ha): There are significant segmentation patterns among tourist, and these can be identified using unsupervised clustering algorithms.

## Research Objectives

The main objective of this project is to explore the effectiveness of unsupervised algorithms in segmenting tourists visiting Iceland according to 3 variables: age, length of stay, and income. To achieve this general aim, there are some other objectives such as, analyse and prepare a tourism dataset, apply and compare 3 clustering models, evaluate their performance, visualise and interpret the segments generated by each method to identify relevant tourist profiles, and discuss the implications of the findings.

# Literature Review

The use of machine learning techniques in tourism research has grown significantly over the past years, offering new ways to analyse tourist behaviour, segment markets, and support strategic decision-making. Unsupervised learning techniques have gained attention due to their ability to uncover patterns without the need for labelled data. Tourism is a complex industry influenced by diverse demographic, behavioural and economic factors. Traditional statistical methods have often fallen short in capturing this complexity. According to the UNWTO (2023), “big data and advanced analytics offer the opportunity to shift from reactive to proactive destination management”. This has encouraged researchers to adopt machine learning models to improve decision-making in areas such as tourist flow prediction, experience personalisation, and visitor segmentation.

## Clustering Techniques

Clustering is a type of unsupervised machine learning that groups data points into clusters based on similarity. It is particularly well-suited to tourism studies, where traveller behaviour can be segmented based on attributes such as demographics, spending habits, and travel motivations. Several studies have demonstrated the value of clustering for segmenting tourist markets. For instance, Dolnicar (2022) reviewed numerous data-driven segmentation studies and concluded that clustering techniques offer more meaningful tourist profiles than traditional demographic approaches. Her work highlights the importance of including behavioural and psychographic data, such as travel motivations and preferences, in addition to basic variables like age or income.

The present study compares three clustering methods (K-Means, Hierarchical, and DBSCAN) to a tourism dataset from Iceland in the year 2023. While prior work has largely focused on larger-scaled datasets, this research seeks to explore how even small, structured dataset can yield valuable insights into tourist behaviour based on age, length of stay, and income. By testing different clustering methods and looking at the groups they create, this study adds to the discussion on how data can help improve tourism planning. The results can help with making better decisions and creating more suitable services for visitors in Iceland.

To sum up, many studies have shown that clustering methods can help understand and organise tourist data better than traditional ways. Researchers now use more detailed data, like behaviour and preferences, instead of just age or income. This project uses three clustering methods to study a real dataset. Even though the dataset is small, it can still give useful information about tourist groups. The literature shows that machine learning, especially unsupervised methods, can support better planning and decision-making in the tourism industry.

# Methodology

The methodology guiding this project is CRISP-DM, a widely used and standard approach in data mining, consisting of six main phases:

## Business Understanding

In this phase, the objective of the project is defined. The main goal it to analyse tourist behaviour patterns in Iceland, segmenting them based on age, length of stay, and income. The resulting insights will be used to improve decision-making within Iceland’s tourism sector, helping to personalise experiences and enhance destination management.

## Data Understanding

Chapter 3: Methodology

Write one or two sentences that describe the aim of your project.

Add architectural diagram of your project.

3.1 Dataset information

3.2 Data analysis and preprocessing

3.3 Model training (3 models)

3.3.1 Neural networks

3.3.2 Random Forest

3.4 Model evaluation

# References