A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Rodney Wardle* |  |
| *SBS23057* |  |
| *Strategic Thinking* |  |
|  |  |
| *Assessment Due Date : 17/05/2024* |  |
| *Submitted: 02/12/2024* |  |

**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.

Table of Contents

[List of Tables 3](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714465)

[Introduction 4](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714466)

[Business Objectives 4](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714467)

[Hypothesis 5](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714468)

[Scope and Methodology 5](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714470)

[Success Criteria 7](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714471)

[Exploratory Data Analysis (EDA) 7](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714472)

[Descriptive Statistics 7](file:///C:\\Users\\User\\AppData\\Local\\Microsoft\\Windows\\INetCache\\Content.Outlook\\3DY5WAT2\\KatieRogers_SThinkingHDip_CA3.docx" \l "_Toc181714473)

[Data Visualisation 9](file:///C:\\Users\\User\\AppData\\Local\\Microsoft\\Windows\\INetCache\\Content.Outlook\\3DY5WAT2\\KatieRogers_SThinkingHDip_CA3.docx" \l "_Toc181714474)

[Data Preparation 10](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714475)

[Data Cleaning 10](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714476)

[Feature Engineering 11](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714477)

[Model Selection 11](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714478)

[Hyperparameter Tuning and Cross-Validation 13](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714479)

[Training and Validation 13](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714480)

[Evaluation Metrics 14](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714481)

[Data Source and Ethical Considerations 15](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714482)

[Summary of Findings 15](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714483)

[Challenges Encountered 23](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714484)

[Results and Analysis 24](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714485)

[Conclusion 25](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714486)

[References 26](file:///C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\3DY5WAT2\KatieRogers_SThinkingHDip_CA3.docx#_Toc181714487)

Table of Figures

# Introduction

An examination of the “The movement of people” using the UN Data sources

The International Organisation for Migration has been gathering and collating relevant data on the movement of people since 2017 and this data is available through the Demographic Yearbook data collection (unstats.un.org, n.d.).

The Data collected is regarded as accurate and reliable and is the work of the United Nations Statistics Division (UNSD). There are four key impacts of the statistics produced by UNSD:

* Collects and disseminates official national data on international migrant flows and stocks through Demographic Yearbook data collection
* Produces international standards and methods related to international migration statistics.
* Assists countries in enhancing their capacity on migration statistics.
* Coordinates statistical programmes and activities through the United Nations Expert Group on Migration Statistics

Business Objectives

The 5 main objectives for this capstone project included the following:

1. How can data science be used to analyse the growing number of asylum seekers around the world.
2. Examination of available data may help to predict the future applications for asylum seekers across the world – not just the number of applications but also the routes and preferred destinations of people on the move.
3. To develop a machine learning model to estimate the number of asylum applications.
4. Compare the estimates of applications both supervised and unsupervised and a description of exactly what this entails.
5. To take a deeper look into the global figures to allow for some examination of the movement of peoples with in Europe, America, Asia and Oceania (Australia).

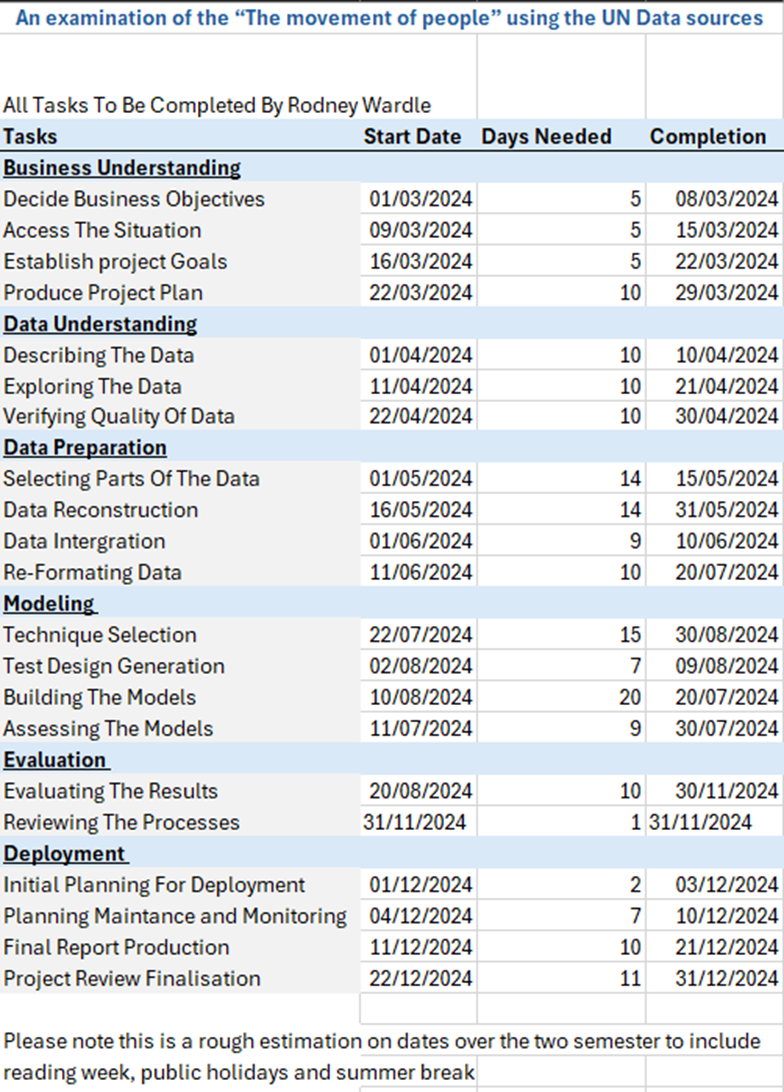
Hypothesis

Machine learning models can be used successfully to help predict the current and future applications for asylum seekers around the global through the UN. It will allow countries prepare for the influx of application especially down to conflict and political issues globally which force people to flea their home country.

Scope and Methodology

Over the two semesters work has been carried out on this dataset regarding aslymn seekers data globally through the UN following the CRISP-DM methodology.

I have followed this time line as close as I could



Of specific interest in this project are the statistics relevant to The United Nations High Commissioner for Refugees (UNHCR) (www.unhcr.org, n.d.) who also collects and compiles data on asylum seekers and refugees more specifically on asylum applications, refugee status determination, recognition rates, refugee populations and movements, demographic characteristics (age and sex) as well as major refugee locations (camps, centres, urban areas, etc.).

We need to import all the necessary libraries to allow us go through the whole project management process following industry standard CRISP-DM.

“CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology.” (Smart Vision Europe, 2017)

Success Criteria

In this capstone project the criteria for seeing how successful the project runs is based off key areas.

* Model performance: The machine learning algorithms need to achieve high accuracy, recall and f1-score.
* Interpretability: Will key stack holders such as government and UN be able to understand the results following deployment on the models an be able to move forward and plan successfully with the whole handling of asylum seeker applications globally.

Exploratory Data Analysis (EDA)

IBM define Exploratory Data Analysis as “is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods.” It allows us to work with a data source in order to get the answers we need for example in this case with the project we want to establish and predict future asylum seekers applications globally.

Descriptive Statistics

In this particular stage of CRISP DM we need to try and understand the data in front of us from the very start as the understanding of this data is imperative to processing the data as needed, creating a machine learning algorithm for the said data

This data has come from the official UN website data finder. The data is regarding asylum seeker application around the EU but specifically between the years 2018-2023 which is 5 years all together.

The UN data finder website provides a data dictionary. This is a comprehensive dictionary.

[unhcr.org/refugee-statistics/methodology/data-content/](https://www.unhcr.org/refugee-statistics/methodology/data-content/)

By using the .head() we can see the first few rows of the data so we can get an introduction to the dataset and try and understand it more. We can see that there are 10 columns which equates to 10 features.

A screenshot of a computer

Description automatically generated

We can get more basic information on the dataset by using .info() function. We can see that we have 1 numerical value as an integer and 9 objects which are categorical data. Due to this dataset having so many categorical value features I am going to use Label Encoder.

A screenshot of a computer program

Description automatically generated

By using the .describe function we can establish the basic statistics for the dataset on asylum seekers applications. It tells us the mean I, standard deviation applied of 3391, minimum of 5 and maximum values in the 25863.

A screenshot of a computer screen

Description automatically generated

Data Visualisations

Data Preparation

Data Cleaning

Data Preparation

Modelling

Evaluation

Deployment and Conclusions

Github repository: https://github.com/RodneyWardle2023/CapstoneProjectCA2RodneyWardle-SBS23057/

# 

# References

unstats.un.org. (n.d.). UNSD — Demographic and Social Statistics. [online] Available at: <https://unstats.un.org/unsd/demographic-social/sconcerns/migration/>.

Smart Vision Europe (2017). Building and Applying Predictive Models in IBM SPSS Modeler training webinar. [online] Smart Vision - Europe. Available at: <https://www.sv-europe.com/crisp-dm-methodology/>.

IBM (2023). What is a Decision Tree | IBM. [online] www.ibm.com. Available at: <https://www.ibm.com/topics/decision-trees>.

Team, G.L. (2023). Label Encoding in Python - 2023. [online] Great Learning Blog: Free Resources what Matters to shape your Career! Available at: <https://www.mygreatlearning.com/blog/label-encoding-in-python/#:~:text=Label%20encoding%20is%20a%20technique>.

Kaushik, S. (2019). An Introduction to Clustering & different methods of clustering. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>.