A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Rodney Wardle* |  |
| *Student Number* |  |
| *Strategic Thinking* |  |
|  |  |
| *Assessment Due Date : 17/05/2024* |  |
| *Submitted:* |  |

**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

[Business Understanding 1](#_Toc166668370)

[Data Understanding 2](#_Toc166668371)

[Data Preparation 3](#_Toc166668372)

[Modelling 3](#_Toc166668373)

[Evaluation 3](#_Toc166668374)

[Deployment and Conclusions 3](#_Toc166668375)

[References 4](#_Toc166668376)

[NOTE: The table of contents above has been included for your convenience. To refresh the table, simply click on it, then select 'Update Table' using the mouse. You can choose to update either the page numbers exclusively or the entire table as needed.]

# Business Understanding

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.).

We need to import all the necessary libraries to go through the whole project management process.

**“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)

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

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.).

This research project aims to address these following objectives:

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)Business Understanding

This project is following the CRISP DM project management methodologies.

This capstone project is going to span two semesters and by the end all the objectives listed above will have been examined.

To start the project off I plan on examining one of these main objectives from the original proposal.

1. To develop a machine learning models to estimate the number of asylum seeker applications.

For now looking first at the total number of asylum application around Europe over the past five years and make predictions for future years. In the next semester I plan on comparing this to the number of applications globally and form educated comparisons. This was established when I began reviewing the data from the UN data finder website.

I have tried to stick to the original time frame or as close as can be. It was easy to follow the required steps to complete the project.

A screenshot of a data table

Description automatically generated

Time line for capstone project

Data Understanding

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 a introduction to the dataset and try and understand it more. We can see that there are 10 columns which equates to 10 features.

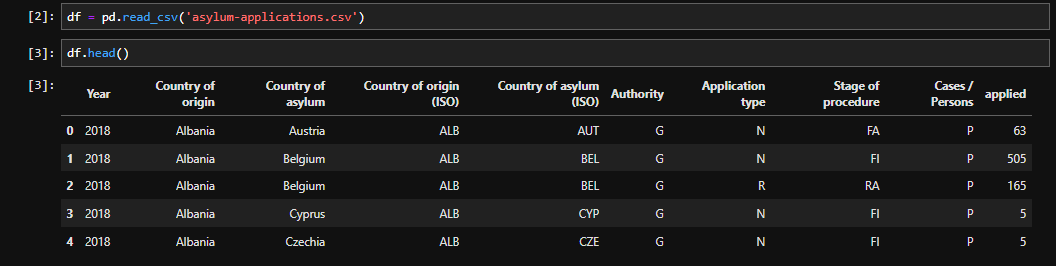


Figure 1. .head() function on the dataset

We want to find out how many observations and features we have so we use the function .shape and 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

Description automatically generated

Figure 2. .shape and .info function on the dataset

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, minimum and maximum values in the dataset.

A screenshot of a computer

Description automatically generated

Figure 3. Describe function to get the descriptive statistics on the given dataset.

To allow me understand the data further I am going to populate some visuals of the dataset to get more insight to the given dataset for this particular business objective for the capstone project.

A blue bar graph with numbers

Description automatically generated

Figure 4. Bar Chart for feature Stage of Procedure

This feature on stage of procedure is heavily skewed to the left especially on the FI step of the application process. FI stands for First instance decisions

A graph with a blue rectangular bar

Description automatically generated with medium confidence

Figure 5. Boxplot of the applied for asylum around the eu.

IA screen shot of a graph

Description automatically generated

Figure 6. country of origin vs country of asylum.

Data Preparation

Data preparation is a crucial step in the CRISP DM protocol. Before we put the dataset through machine learning models the data needs to be cleaned, no matter what we do we will never have 100% cleaned data.

A screenshot of a computer program

Description automatically generated

Figure 7. Dealing with missing data in the dataset.

We are firstly checking for is the any null values in the dataset and we can see that Stage of procedure has 65 null values. To help this we can make a list of missing value formats and remove from dataset if any of the missing formats are found.

A screen shot of a computer

Description automatically generated

Figure 8. Label encoder for dealing with the categorical data in the data set.

Computers only under stand numerical value and not categorical values, in order to preform any machine learning models on this data we need to deal with the categorical and encode it. Due to the amount of categorical data present it was decided to use the LabelEncoder.

Label encoding is a technique used in machine learning and data analysis to convert categorical variables into numerical format. It is particularly useful when working with algorithms that require numerical input, as most machine learning models can only operate on numerical data. (Team, 2023)

Modelling

There are several machine learning models that could be used on this dataset. For now I have used a decision tree on the target variable for the number of asylum applications in Europe for the past 5 years.

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes. (IBM, 2023)

In this particular case we could use decision trees and clustering algorithms due to the values on country of origin, country of asylum being sought in.

A screenshot of a computer program

Description automatically generated

Figure 9. Modelling the feature applied by using a decision tree.

For the train test split we drop the column first and the column applied for X and then add the response to y. I use 30% with a random state of 99.

The decision tree classifier is then created and fitted. We can see the precision and recall is 1 and the support is 19.

A black screen with white text

Description automatically generated

Figure 10 . Confusion matrix and accuracy score on the specified feature applied

Evaluation

A screenshot of a computer

Description automatically generated

Figure 11 . Evaluating the decision tree.

A screenshot of a computer program

Description automatically generated

Figure 12 . Plotting the tree on the gini

A screen shot of a computer program

Description automatically generated

Figure 13. Setting the decision tree

The decision tree classifier is set to gini. It has a random state of 100 and a maximum depth of 3. The score is then printed of the x test and y test,

To be perfectly honest this is quite difficult to interpret so I need to revisit what would be the best machine learning models for this particular context of predicting the number of asylum seeker applications around the EU. Only by experimenting did I discover this .

A screenshot of a computer

Description automatically generated

Figure 14. report on the decision tree classifier

A screenshot of a computer

Description automatically generated

Figure 15. Confusion matrix on the decision tree classifier

For semester two I intend to work further on the following:

* Look at different models such as clustering as this would work better with the given data and context as there is a set application process for applying to become an asylum seeker. When encountering an unsupervised learning problem initially, confusion may arise as you aren’t seeking specific insights but rather identifying data structures. This process, known as clustering or cluster analysis, identifies similar groups within a dataset. (Kaushik, 2019)
* Compare the EU figures on applications for asylum with the figures globally
* Did covid have any impact on asylum applications, maybe it took longer than usual to go through the process.
* Compare between supervised and unsupervised models on the asylum application figures.
* Try to merge other datasets that have similarities such as the decisions on the applications.

Deployment and Conclusions

This capstone project main objective was to examine the asylum seeker applications around the EU for the past five years. I followed the CRISP DM frame work for project management. Business understanding played an important role as it helped get more precise on the objectives. Following this I had to prepare and understand the data. There were missing values and a big proportion of the dataset was categorical. This complicated matters as for example the application process is abbreviated as is and then it had to be encoded so this made it much harder to make understand and evaluate the findings.

Following on from this initial set up for the machine learning models it is apparent that more precise models are going to be researched , create the other models, compare and contrast them and see what one is better for predicting the asylum application figures from the UN.

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/>.