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

[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)

[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" \l "_Toc181714482)

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

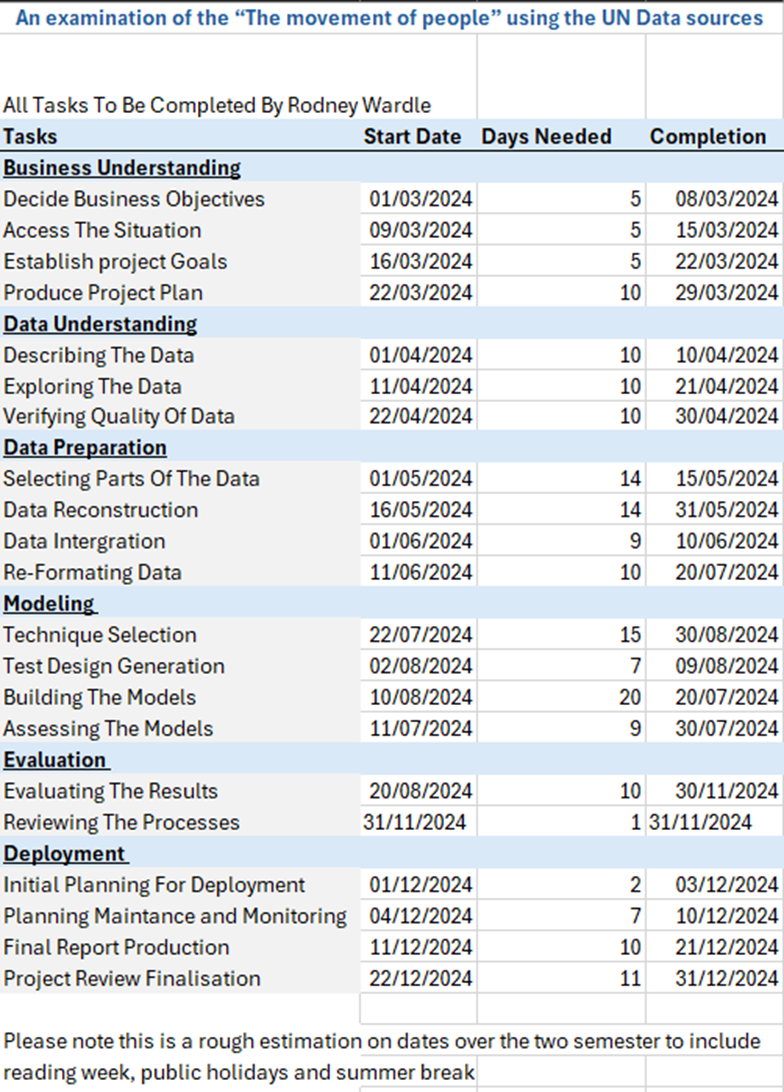
Hypothesis

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.
* One more

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

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

A graph of a stage of procedure

Description automatically generated

Data Preparation

Data Cleaning

According to the latest estimates by [Statista](https://bernardmarr.com/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/), 328.77 million terabytes of data are generated each day. It is estimated that 90% of the world's currently existing data was only generated in the last two years alone. In order for the machine learning algorithms to perform as they should do and have high accuracy etc its imperative that we process and clean this dataset.

We need to check for any missing values and it can be seen the feature Stage of procedure had 356 missing values. I found the best way to deal with these missing values was to create a list of missing value formats such as n.a. ,nan, unknown etc , that is assigned to the dataset and then you drop that list. I also double checked this by dropping any nan values or actual spaces in stage of procedure feature.

Machine learning models only work properly with data in numerical format. The features year and applied are numerical. The features Country of origin, Country of asylum, Authority, Application type , stage of procedure and cases/persons are all categorical data.

In order to deal with the categorical it was decided to use Label Encoder. It is a technique used in data analysis to convert the categorical variables in to numerical values. It uses the library scitkit-learn. It will turn the data such as stage of procedure from categorical in to numerical of 1,2,3,4,5,6,7,8. Label encoding works particularly well on tree based models such as Decision Trees. One big drawback is that it does not work with data that is non ordinal in nature so in cases like this another technique can be used which is called One Hot Encoding. You have to judge it based on the type of data and how it behaves.

Feature Engineering

Feature engineering, in data science refers to manipulation — addition, deletion, combination, mutation — of your data set to improve machine learning model training, leading to better performance and greater accuracy. Effective feature engineering is based on sound knowledge of the business problem and the available data sources.

The following feature engineering was performed on the dataset during the capstone project.

* Initially exploring the dataset to get an understanding of it.
* Dealing with missing data.
* Variable Encoding for categorical data.
* Feature selection on the dataset based off initial capstone project objectives.
* Feature Extraction – PCA was performed on this dataset.

Model Selection

One of the main objectives in this capstone project was to use appropriate machine learning models to help predict current and future asylum seeker applications globally both from a supervised and an unsupervised machine learning approach.

Before discussing the different machine learning models used on this dataset it is important to discuss in detail what is supervised and unsupervised machine learning models in order to understand the rationale behind the model selection during this capstone project.

According to IBM “Supervised learning is a machine learning approach that’s defined by its use of labeled data sets. These data sets are designed to train or “supervise” algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.” IBM also say unsupervised learning“ uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention”

In its simplest form supervised learning models are trained on labelled data which means features of the dataset are paired with the appropriate output, its main goal is for the model to learn a mapping function from the inputs and outputs and then generalize this mapping to unseen data.

There are several advantages and disadvantages to using supervised machine learning models over unsupervised machine learning models.

Advantages include the following:

* **High Accuracy and Predictability:** The models are trained with supervised learning which often provide more reliable and accurate predictions when the labelled data is sufficient and good quality.
* **Widely Applicable**: It works well on a wide range of tasks such as regression, classification and ranking problems.
* **Controlled Learning:** Supervised learning models operate in a structured manner since both the features and labels are well defined.
* **Automation of Repetitive Tasks:** It has the capability of automating the tasks involved in predictable patterns for example image recognition.
* **Flexibility in Algorithm Choice:** A wide variety of algorithms are available to cater for different types of data and problems that are complex in nature.
* **Availability of Evaluation Metrics:** Well established metrics such as accuracy, precision and recall enable clearer evaluation on model performance.
* **Ease of Interpretability (only in certain models**): Certain algorithms like decision trees, used on this capstone project , offer intuitive, interpretable outputs are useful in understanding the relationship in the data.

Disadvantages include the following:

* **Overfitting:** The models may perform really well on the training data but will normally fail to generalise the unseen data if overfitting occurs.
* **Dependency on Labelled Data:** It requires a large amount of labelled data, which can be time-consuming and expensive to collect and understand.
* **Limited Scope:** the models are only as good as their training data meaning If the data lacks diversity or fails to show real-world scenarios, the predictions can become biased and inaccurate.
* **Difficulty with Complex Patterns**: as supervised learning models can do brilliant with structured data it can struggle at times with extremely complex, high dimensional data unless advanced models such as deep learning are used.
* **Challenges with Scalability**: Supervised models that are trained on larger datasets can from a computationally expensive, specifically deep learning models.
* **Data Leakage Potentially:** If the datasets are not prepared properly during the training phase can lead to misleading high accuracy and poor real-world performance.

During semester one I worked on the supervised machine learning models. This dataset is classified in nature so I decided to use a Decision Tree Classifier algorithm. Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The main goal is to create a training model that can be used to predict the class or value of the target variable of the target variable.

One of the general objectives for this capstone project was to use both supervised and unsupervised machine learning models on the chosen features of the dataset and try and compare and contrast.

Unsupervised machine learning models in its simplest form involves analysing and learning the patterns from unlabelled data which is without predefined output labels, it focuses purely on uncovering the hidden structures, patterns or features in data

There are several advantages and disadvantages to using unsupervised machine learning models over supervised machine learning models.

Advantages include the following:

* **No Need For Labelled Data:** Unsupervised models work purely on the raw data which makes it cost effective and much better for exploratory analysis.
* **Handling Complex Data:** When we got high dimensional data it performs really well on it.
* **Versatility:** It can be deployed in many areas such as clustering and dimensionality reduction , in the case of this project PCA
* **Adaptive:** The models adapt very quickly to changes in the data and can be applied on dynamic situations.

Disadvantages include the following:

* **Overfitting**: When there is no target variable , it will fit to irrelevant structures and noise in the data.
* **Overgeneralization:** The model can make the data overly simplified which will miss important nuances or creating random groups.
* **Interpretability Difficulties:** The results can be extremely challenging to understand due to not having any predefined labels or outcomes in order to give context to the data.

In semester 2 I worked primarily on unsupervised machine learning models on the dataset.

The first model that was used was the Kmeans. K-means is a centroid-based clustering algorithm, where we calculate the distance between each data point and a centroid to assign it to a cluster. The goal is to identify the K number of groups in the dataset.  (reference needed). To create a K-means model is quite easy, firstly you need to choose the number of clusters you want,. You then initialize the centroid (centre of a cluster). You then assign the data points to the nearest cluster., this step is where we calculate the distance between data points X and the centroid using the Euclidean distance metric. You then re initialise the centroids by calculating average of all the data points within the cluster. You repeat the steps until we have the right amount of centroids and the data points in order to correct the clusters are no longer changing.

The second unsupervised learning model I picked was KNN (K-Nearest Neighbour). The KNN algorithm predicts responses for new data (testing data) based upon its similarity with other known data (training) samples. It assumes that data with similar traits sit together and uses distance measures at its core.

Hyperparameter Tuning and Cross-Validation

Hyperparameter tuning is the process of selecting the optimal set of hyperparameters for a machine learning model. It is an important step in the model development process, as the choice of hyperparameters can have a significant impact on the model's performance (REFERENCE NEEDED)

This tuning is not learned during the training stage of the model. These parameters are set normally before the training of the model takes place. The behaviour of the model can be set in many ways such as for example the learning rate, number of layers in a decision tree , the number of neurons present per layer, for support vector models you can select the kernel type, you can also set the tree depth or the number of estimators.

The main goal of the hyperparameter tuning is the try and find the right balance between values to maximise the models performance.

Evaluation Metrics

In order to evaluate the model’s performance I used several metrics which provided a detailed review of how well the different models where performing in order to predict current and future asylum seeker applications globally. Metrics are used to monitor and measure the performance of a model (during training and testing), and don’t need to be differentiable**.(reference needed)**

Accuracy is used to measure the overall appropriateness of the particular model by comparing the number of correctly instances to the total results predicted. It is generally used when the data is balanced meaning the classes are roughly equally represented. When the data is imbalanced it can actually make the results to be misleading.

Precision generally measures how frequent the model is able to correctly predict among all the positive predictions. I used it to try and minimise the false positives on the datasets.

Recall is a metric that measures the amount of actual positive instances that are correctly identified by the machine learning model. It works best with problems that have imbalanced data because it shows the models correctness when it comes to identifying the target group. You should not consider using recall when there are false negatives which means it will not account for the cases when the target is missed.

F1-score tries to be the average between the precision and recall metrics. It combines both metrics and balance the trade off between the two metrics. It looks at both false positives and false negatives , this comes in handy when the datasets are imbalanced and skewed.

The support metric refers to the actual number of occurrences of the occurrences within the dataset.

Challenges Encountered

During this capstone project I encountered several challenges which really made me think outside of the box for solutions.

* Originally in semester one I was only looking at the asylum seeker applications within EU but the problem was not enough data so in semester 2 I decided to go further and look at the data globally so I had to go and re prepare the data so this took further time obviously.
* I found it extremely difficult to interpret the results for the decision tree classification.
* Discovered with trail and error working on the dataset that this data is better suited to clustering models , but only discovered this following performing the decision tree classifier.

Data Source and Ethical Considerations

This data has come from the official UN website data finder. The data is regarding asylum seeker application around the Globally but specifically between the years 2018-2023 which is 5 years all together. During this time period there were less asylum seeker applications during the global Covid Pandemic.

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

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

Results and Analysis

Here are the results from the different machine learning models .

Conclusions

We have been able to show successfully that a machine learning algorithm can be used in the prediction of the amount of asylum seekers applications and to prepare for the future needs for the asylum seekers globally as a whole.

More work is required to get the accuracy up higher into 90% +.

More finetuning is needed to get the best results possible.

**Github repository:**

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

**Prerecorded presentation of poster:**

https://drive.google.com/file/d/1Go2PYogdF3EG7EmK9f7XoPbn7A6WtQkK/view?usp=drive\_link

# 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/>.