**Abstract**

**Index:**

1. **Introduction**
2. **Data Preparation and Visualization**
   1. EDA
   2. Method justification
   3. Visualisations
      1. *Population growth*
      2. *Population Pyramids*
   4. Tuft Principles
3. **Statistical Analysis**
   1. Dataset summary
   2. Binomial and Poisson Distribution
   3. Normal Distribution
   4. Importance of Binomial and Poisson and variable justification. Discrete variables in a Normal Distribution
4. **ML**
   1. Project management framework CRISP-DM, KDD and SEMMA. Supervised, unsupervised, or semi-supervised machine learning technique.
   2. ML models
   3. Results of ML models selected
   4. Similarities and differences of ML models selected
5. **Programming**
   1. Programming paradigms used
6. **Conclusion**
7. **References**
8. **Appendix**

**1. Introduction**

The aim of this study is to highlight the population changes for the Republic of Ireland. Fifteen datasets are available to complete this study and here it goes some detail about how they are formed:

* Population changes we have seven datasets, four types of datasets, first one we have age group by sex, second age group by sex and region, third one year by component as in *“Annual births”, “Annual deaths” ….* Last type goes year by sex and nationality. This data will be used to understand how the population grew and is distributed in Ireland.
* Migratory flows, there are four datasets each one of them has a distinct attribute, first type age group by sex, second one country by sex, third education level and fourth one economic status. These datasets are key to identify which type of immigrants or emigrants get in and out of the country.
* Three datasets explain emigration, attributes across the datasets are almost the same, year by sex and nationality, the only difference between them is year length, 1996-2022, 2006-2022 and 2006-2023. I will be using one of the datasets to see if during *“The Celtic Tiger”* people living in Ireland decided to stay.
* One dataset is exclusively for immigration having year by sex and citizenship, I will be also using this data to get more clarity on *“The Celtic Tiger”* study.

We will not go beyond 1923, however we do have until 2023. In my opinion this is enough time to get a picture of the current demographics. We will get into detail in each of the sections.

**2. Data preparation and visualization**

2.1. EDA

2.2. Method justification

2.3. Visualizations

2.3.1 *Population Growth*

2.3.2 *Population Pyramids*

Population pyramid charts are an essential tool when analyzing demographics, we can visualize the age and gender distribution of a population. To plot this, we use horizontal bars by gender and age group (coderzcolumn.com, n.d.).

Here I wanted to see pyramid shape for 1926 being the first year with data, 1996 being at the beginning of the *“The Celtic Tiger”*, years 1995 to 2007 (Investopedia., n.d.) and 2023 being the current year.

Here are the results:

A chart of a pyramid

Description automatically generated

Figure: 1926 Pyramid Chart

A chart with a pyramid shape

Description automatically generated with medium confidence

Figure: 1996 Pyramid Chart

A chart with red and blue lines

Description automatically generated

Figure: 2023 Pyramid Chart

The first figure depicts a wide base, showing a big number of young people in the country however it shrinks quickly when climbing up into the older age groups, and it is very thin at the top, this pyramid shape relates to developing countries, making totally sense as the Republic of Ireland had experienced a civil war from 1922-1923 (John\_Dorney, 2012) and was still far from economic growth.

2.4. Tuft Principles

**4. ML**

4.1. Project management framework CRISP-DM, KDD and SEMMA.

As a project management framework, we are going to discuss each of them in detail. We have CRISP-DM (Cross Industry Standard Process project), KDD (Knowledge Discovery in Databases) and SEMMA (Sample, Explore, Modify, Model, Assess) models.

CRISP-DM has six phases, business/research understanding, data understanding, data preparation, modeling phase, evaluation, and deployment. A real-world example for this framework would be a hospital predicting the length of stay for its patients, based on a series of indicators. This method was conducted using Average Prediction, Multiple Regression, Decision Tree, Artificial Neural Network ensemble, Support Vector Machine and Random Forest. The implementation of this model lead to building successful predictions for the hospitals (Caetano et al., 2015).

A diagram of a data flow

Description automatically generated

Figure : CRISP-DM Flow

KDD in this framework we retrieve and analise the data stored in databases, the entire process consists of seven steps, data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation. A good example of this method is the use of machine learning in the prediction of sports results. A recent study has been conducted to predict football matches (Głowania et al., 2023).

**A diagram of data processing

Description automatically generated**

Figure : KDD Flow

SEMMA entails five steps, sample, explore, modify, model and assess. This method has been used for Crime Prediction and conducted using article neural networks machine learning concept (Forradellas et al., 2020).

A blue rectangle with white text

Description automatically generated

Figure: SEMMA flow

In PEA20.20231004T131025.csv we have labelled data that is why I have selected supervised machine learning techniques, we can train the model in regard to migration flows depending on people economic status, sex and inward or outward flow.

4.2. ML models

1. Programming
   1. Programming paradigms used