# 

**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Programming for DA  Statistics for Data Analytics  Machine Learning for Data Analysis  Data Preparation and Visualization |
| **Assessment Title:** | MSC\_DA\_CA1 - Population in Ireland |
| **Lecturer Name:** | David Gonzalez  Marina Iantorno  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Gerard Heraghty |
| **Student Number:** | sba23347 |
| **Assessment Due Date:** | 12/11/2023 |
| **Date of Submission:** | 10/11/2023 |

**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. |

**Group ID - MSc in Data Analytics**

Author: Gerard Heraghty

e-mail: [sba23347@student.cct.ie](mailto:sba23347@student.cct.ie)

Student ID: sba23347

Table of Contents

[Abstract 4](#_Toc150477074)

[Introduction 4](#_Toc150477075)

[Methodology 5](#_Toc150477076)

[Business Understanding 6](#_Toc150477077)

[Data Understanding 6](#_Toc150477078)

[Data Preparation 7](#_Toc150477079)

[**Slimming the datasets** 7](#_Toc150477080)

[**Dealing with null values** 7](#_Toc150477081)

[**Restructuring the migration dataset** 7](#_Toc150477082)

[**Merging the dataset** 7](#_Toc150477083)

[**Feature engineering** 7](#_Toc150477084)

[**Visualizing the data** 8](#_Toc150477085)

[Statistical Analysis 9](#_Toc150477086)

[**Descriptive Statistics** 9](#_Toc150477087)

[**Binomial Distribution** 10](#_Toc150477088)

[**Poisson Distribution** 11](#_Toc150477089)

[**Normal Distribution** 12](#_Toc150477090)

[Modelling 14](#_Toc150477091)

[**Regression Models** 14](#_Toc150477092)

[**Classification Models** 19](#_Toc150477093)

[Conclusion 22](#_Toc150477094)

[References 23](#_Toc150477095)

## Abstract

*The overall aim of this project is to investigate and analyse the leading contributors to population growth in Ireland and identify appropriate machine learning algorithms to forecast future populations. Univariable and multivariable linear regression models were applied and compared based on their prediction performance. It was concluded that the univariable linear regression model had the best performance metrics and was used for forecasting for the next 10 years. Analysis and forecasting using this model was also conducted on the age dependency ratio in Ireland for the same period. A comparison between 2 classification models was also performed as part of this project. The task was to classify the largest contributor to population growth for a given year. A KNN model with tuned hyperparameters outperformed logistic regression in this binary classification.*

## Introduction

Population forecasts are used to summarize knowledge of a population that already exists to aid decision makers to develop strategic policies for the future development of a region. This includes planning the future provision of services, infrastructure, housing, education and healthcare expenditure.

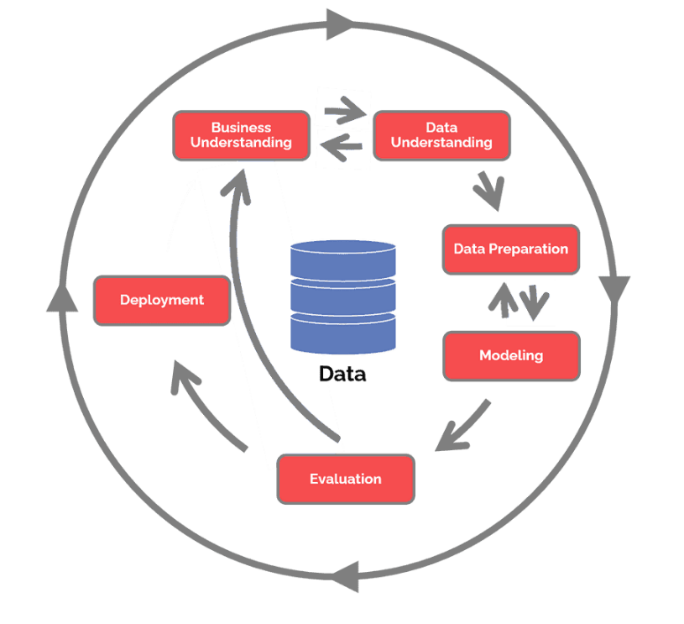
The National Planning Framework published under Project Ireland 2040, sets out the high-level strategic planning and development for the country to 2040 based on official population projections. The initiative involves investment of €116 billion to be prioritised for the future needs of the population (Government of Ireland, 2018). This highlights the need for accurate population forecasting models.

Irelands current age structure compares favourably to other EU countries, with the joint lowest median age (37.7 years) and the lowest share of population aged 65+ (Department of Finance, 2021). According to this report the composition of Irelands age demographic is set to significantly change. Since an aging population will have sizable implications for the public finances given the expected demand on pension and healthcare related spending, forecasting this demographic is vital.

Major components affecting population growth in Ireland include net migration, births and deaths (Central Statistics Office, 2022). These three components are traditionally used to forecast the population using the cohort-component method (Morgenroth, 2008). Although this method has certain advantages, it tends to present a biased result in the forecasting. This report sets out to analyse the effectiveness of various machine learning models in accurately forecasting future population demographics and the dominant components responsible for population change.

## Methodology

For this project a structured 6 phase approach in line with the Cross Industry Standard Process for Data Mining (CRISP-DM) framework was followed. The phases are seen in Figure1 (Hotz, 2023).



**Figure 1:CRISP-DM life cycle**

1. **Business Understanding:** Understand what you want to achieve and assess the requirements and resources available to successfully achieve the project objectives.
2. **Data Understanding:** Collect the required data, exploring and describing the loaded data.
3. **Data Preparation:** Select only relevant data which analysis is to be carried out on. Necessary data cleaning is performed on the datasets. New features or combined datasets are created.
4. **Modelling:** Machine learning models are applied to the data, assessed and hyperparameters tuned appropriately to optimise results.
5. **Evaluation:** Models results are assessed, determine if there is a reason why any model is deficient. Review work accomplished and determine the next steps.
6. **Deployment:** Evaluation results put into action. (i,e final report document).

## Business Understanding

The aims of this project were to analyse the different components of population growth, compare the performance of machine learning models as a tool for population forecasting, and to investigate the changing age demographic in Ireland.

Population demographic datasets are made available by the Central Statistics Office (CSO) (CSO Datasets, 2023). To achieve the aim of this report 4 datasets were chosen containing data in relation to population, migration, births and deaths.

## Data Understanding

Multiple pandas functions were run on each dataset to ascertain some level of understanding about the data. These included head(), to give a snapshot of the first 5 rows of data, shape(), to display the number of rows/columns, info(), to indicate the type of each column and nunique(), to show the number of unique values for each column. Descriptive statistics were displayed for each raw dataset using the describe() function. This gives an indication of some basic statistics such as mean, standard deviation, min, max and percentiles. Any null values were indicated through a combination of isnull() and sum().

During the understanding phase, simple visualizations were created using the raw data to further provide understanding. Correlation heatmaps and pairplots were utilised to highlight any significant relationships between the variables in the dataset. histograms and box plots were used to highlight the spread of the numerical values.

It was observed from the data understanding phase that the datasets that were selected were already incredibly clean with few null values, no duplicates, no mismatched types.

Each dataset loaded contained one numerical colum. This numerical column, “VALUE”, is defined by the categorical values that make up the rest of each dataset, therefore it can be classed as the dependent variable.

## Data Preparation

### **Slimming the datasets:**

There were some columns that were irrelevant to the analysis such as “STATISTIC Label” and “UNIT”, which were common across all datasets. Only data related to “both sexes” was kept for each dataset as this is not a demographic to be analysed, therefore the “sex” column can be dropped after this query is executed. In addition to this, only specific “Age Groups” categories are of interest to this project. This ensured each dataset had the same age group values so they can be merged. The deaths dataset needed further manipulation as it contained quarterly data (section 1.2.3 notebook).

### **Dealing with null values:**

Null values were only observed within the *migration* dataset. It is observed that only “Net migration” values are null. Since net migration is calculated as immigration minus emigration, all these null values can be calculated from the existing data. A “for loop” using iterrows was utilised to iterate through each row, identify when “value” is null and fill that null value with the calculated net migration (section 1.2.2 notebook)

### **Restructuring the migration dataset:**

The migration dataset was restructured to pivot the emigration, immigration and net migration values into their own columns instead of being combined in the “*value”* column (section 1.2.2 notebook). This was done so enable the merging of the all the datasets later in the preparation phase.

### **Merging the datasets:**

The *pop\_df* dataset was created by merging all four datasets using “*year*” and “*ages*” as the key. The pandas .merge() method was used for this, with the *how* parameter set to “left”. This type of merge was selected as it keeps the integrity of the table on the “left” side of the join and only joins the data from the “right” table that matches the key set using the *on* parameter. This results in the desired table (section 1.5 notebook).

### **Feature engineering:**

The age dependency ratios (old, young and total) were determined using the *dep\_ratio* python function defined and detailed in section 1.5.1 of the notebook. Once each dependency ratio is calculated, only data related to “All ages” is selected from the dataset, as the age groups are no longer needed. The dependency ratio values are then merged to the resultant dataset.

“NaturalGrowth” feature was created by calculating the difference between births and deaths and will be used in the modelling phase of this project.

### **Visualizing the data:**

Visualizations were generated in close accordance with Tufte’s principles of data visualisation (Tufte, 2001). This ensures that the information is accurately conveyed without distortion or misrepresentation.

Figure2 provides insight into the key drivers of population change in Ireland. These three variables were plotted on the same chart is to provide insight into relationships between the variables. A scatter plot was determined to portray the data coherently. Grid lines were removed to reduce non-data ink.

It is observed that the net migration is closely correlated with the population change. Natural growth appears steady throughout the duration but is tending downwards in the last decade.

A graph of a number of people

Description automatically generated

**Figure 2**

In Figure3 a scatter plot was also determined to be the most effective way at displaying the data. Y axis gridlines were included in this plot as it makes it easier to visualise the trends. The graph shows that there is a sharp increase in the old age dependency ratio in the last 15 years. The young age dependency also appears to be trending downwards in recent years.

A graph showing the age of a person

Description automatically generated

**Figure 3**

## Statistical Analysis

### **Descriptive Statistics:**

Descriptive statistics consist of three basic categories of measurement: central tendency (mean, median, mode), variability (variance, standard deviation) and frequency distribution (count). Since all the data being dealt with as part of this project is numerical, frequency distribution was ignored.

Using the .describe() pandas method on the dataset, the descriptive statistics was produced for each dataset. The descriptive statistics table for the *pop\_df* dataset is visible in the Appendix or section 3.1 notebook. The mean provides an insight into the central tendency of each variable while the standard deviation, min, max and percentiles give an insight into the distribution of the data.

The mean is not always a good measure of central tendency. It is more appropriate to use the median as this measure when the data is skewed or the data contains outliers. From the distribution plot in Figure4, it is observed that the youth\_dependency\_ratio is skewed to the right. The median would therefore be a better measure of central tendency for this variable. Using .median() this is determined to be 32.3623.

A graph of a number of individuals

Description automatically generated with medium confidence

**Figure 4**

Box plots can be utilized for visually comparing descriptive statistics of variables. Figure5 displays two box plots. The first comparing *immigration* and *emigration*. The second comparing *births* and *deaths*. Box plots are useful for identifying the mean, illustrating the spread of the data and the presence of any outliers. It can be observed from the *immigration* box plot that there is one outlier value from the dataset with a value of 151100. It is also seen from the plots that the smallest value of *births* is still much greater than the largest value of *deaths*. Therefore, the natural change in the data is always positive (i.e there are always more people being born than dying in Ireland).

**Figure 5: Box plots comparisons of related variables**

A diagram of a box with a line and a line

Description automatically generated with medium confidenceA diagram of a box plot

Description automatically generated

### **Binomial Distribution:**

The Binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial, and the probability of success is the same for each outcome. This discrete distribution is used to obtain the probability of observing *x* successes in *n* trials, with a probability of success *p* (Weiss. 2016).

By looking at the *tot\_dependency\_ratio* it can be deduced whether the majority age demographic of the population is that of a dependent (i.e 0 –14 years, 65+ years). A **successful** outcome is determined if *tot\_dependency\_ratio* is >50%, and a **failure** if it is <=50%. The probability of success, *p*, for each outcome is determined as the count of all values that have a tot\_dependency\_ratio > 50% divided by the total number of values in the dataset. This was calculated to be 0.6666667. The number of trials, *n,* is selected as 10 random values. A binomial probability distribution can now be determined.

To do this using python (section 3.2 notebook) the binomial probability mass function from scipy.stats is used. This function gives the probability that a discrete random variable is exactly equal to some value (SciPy Documentation, 2023). The resultant distribution is shown in Figure6.

A graph of a bar graph

Description automatically generated with medium confidence

**Figure 6: Binomial Probability Distribution**

It is observed that the distribution is skewed to the left. This is due to *p* > 0.5, i.e there is a greater probability of a dependency ratio > 50% being observed.

The shape of the distribution becomes more symmetrical as the number of trials, *n,* is increased. Figure7 below illustrates the distributions when *n* is increased to 30. This is in line with the Central Limit Theorem which states that as *n* increases, the binomial distribution with *n* trials and probability *p* approximates a normal distribution.

A graph of a number of individuals

Description automatically generated

**Figure 7: Binomial Probability Distribution with larger number of trials**

### **Poisson Distribution:**

The Poisson distribution is a discrete distribution that measures the probability of number of events occurring within a specific time period. The main characteristic of this distribution is there is no limit to the domain (i.e number of events). Poisson distribution only takes 1 parameter, λ, which indicates the average number of events occurring in the time period.

By considering the average number of births occurring per year a Poisson distribution can be applied since there is no defined limit to the number of births that can occur during the time period (1 year), the occurrences of each event is random and independent.

The Poisson probability mass function (pmf) may be used to obtain the probabilities of each random variable. The random variables were defined as a range, *k,* around the calculated mean, λ = 60.078 (mean number of births per year in thousands). For large values of λ (i.e λ > 20) a normal distribution approximates the Poisson distribution (Frost, 2021), therefore from the empirical rule we can assume that 99.7% of the data lies within 3 standard deviations of the mean. The range of values used for plotting this distribution was chosen within 3 standard deviation of the mean.

A graph of a number of persons

Description automatically generated

**Figure 8: Poisson Probability Distribution**

The probability distribution of the occurrence of different numbers of births per year is illustrated in Figure8. Due to the large mean value, it is evident from the plot that the distribution is very symmetrical.

The cumulative distribution function (*cdf*) may be used to determine the probability that of a number of events occurring is less than or equal to a specified value. Using the *cdf* function, P(X<= 50k) was determined to be 0.1059, which indicates quite a low probability that births will be below 50,000 each year.

### **Normal Distribution**

The normal distribution is a continuous distribution that is symmetrical about its mean. This distribution calculates a cumulation of probabilities, therefore the probability of an exact value cannot be calculated, it will always be greater or less than.

Various checks for normality were conducted to ensure the variable to be analysed can be approximated using a normal distribution (S. Gandhi et al. , 2019). Histograms were plotted for each variable in the dataset with the density curve (kde) set to true to give an approximation for the distributions. From this it was observed that both *net\_migration* and *deaths* exhibited characteristics of a Gaussian. QQ (Quantile-Quantile) plots were constructed using the *net\_migration* and *deaths* data. The distribution on the y-axis is the normalized sample data and the x-axis reflects the theoretical quantiles. Any normally distributed data will roughly follow a straight line. It is observed that *net\_migration* roughly follows a straight line. The Shapiro-Wilk normality test is applied to the *net\_migration*. This hypothesis test was chosen as it is more appropriate for small sample size (<50) than other tests. The results of this test show a p-value >0.05 which indicates a normal distribution.

To analyse the probability density of the *net\_migration* variable, the mean, standard deviation and migration values are passed through the normal *pdf*. The results of this are illustrated in Figure9.

A graph of a normal distribution

Description automatically generated

**Figure 9: Normal Distribution of Net Migration**

The probability densities for each net migration value are symmetric about the mean (red line). Using the normal survival function (sf), the probability that the net migration is greater than 0 (i.e immigration > emigration) is calculated to be 0.675.

## Modelling

Supervised learning is where labelled datasets are used to train algorithms to classify or predict outcomes accurately (Delua, 2021). This type of learning learns relationships and patterns between the input variables and target variable. There are two types: regression and classification. Regression models are useful for predicting numerical or continuous values based on different data points, while classification algorithms learn from the data to predict discrete outcomes. Since all the data that is dealt with in this project is labelled, supervised learning models are the best techniques to aid in the analysis.

### **Regression Models:**

One of the aims of this project is to compare the performance of different machine learning models as population forecasting tools. Given the population (*pop\_estimate*) is a continuous variable and not some selected discrete value it is appropriate to apply regression models to this problem. 4 different regression models were utilized (section 4.1 notebook) for population forecasting: Linear (univariable and multivariable), Ridge, Lasso.

For splitting the data into training and test data the *train\_test\_split* function from sklearn was used. 80% was used for training while 20% was used for testing. The data was split this way to ensure there was enough training data due to the overall small number of values in the dataset. Since the dataset is time series the data should not be shuffled to ensure the model is tested against the most recent values (Radecic, 2021).

Linear regression is used to study the linear relationship between a target variable and one or more independent variables. The prediction formula is defined as:

y = ω0 + ω1x1 + ω2x2 + … + ωnxn

where x[1…n] are the independent variables, ω[0…n] are the parameters that are learned by the model and y is the predicted target variable. (Schneider et al. 2010)

#### Univariable Linear Regression:

Univariable linear regression studies the relationship between the target (*pop*\_estimate) and one single independent variable. *Year* was taken as the independent variable for this case as it has the strongest correlation to the target variable. The straight line learned from the training data is fit to the test data and is observed in Figure10.

A graph with blue dots and green lines

Description automatically generated

**Figure 10: Univariable Linear Regression**

#### Multivariable Linear Regression:

Given the limitation of univariable linear regression that one single independent variable (i.e *year*) does not suffice to explain the target *pop\_estimate*, multivariable linear regression was also analysed. In multivariable linear regression, the target is described as linear function of multiple independent variables (i.e xn from the equation above)

Feature selection was initially performed before running this model. Adding redundant features can negatively impact the overall accuracy of the model. Only variables that possess a moderately high correlation with the target are considered. A threshold of 0.5 is selected (Gupta, 2023). This value was selected as it removes some poorly correlated variables, while keeping enough to perform the multivariable model. Figure 11 illustrates the correlation values between the independent variables and the target.

A bar graph with blue bars

Description automatically generated

**Figure 11**

From the independent variables, if two variables are correlated with each other, then one of these pair of variables should be removed from the model (Gupta, 2023). After this process is completed the features to be passed to the model are *emigration, immigration,* and *year.* Figure12 illustrates the estimated population values when plotted against the data.

A graph with blue dots and green lines

Description automatically generated

**Figure 12**

It is obvious from Figure12 that the model is overfitting given how close each training data point is to the respective prediction. The score for the training and test data confirms this. for the training was 0.9772 and for the test was 0.05575. The score shows how well the data fit the regression model, with 1 being the best score. The high score for the training data and the very low score for the test data is indicative of the overfitting. A potential cause of this overfitting is that the model is too complex (i.e there are too many variables compared to the number of observations in the data). Such a model tends not to generalize well to new data.

#### Regularization:

To try reducing this overfitting regularization was applied. Regularization works by constraining the coefficients (ω) in the model which aims to discourage more complex models from overfitting. There are two types of regularization: **Ridge** (L2) and **Lasso** (L1). They are both governed by a tuning parameter α which determines how much the model is penalized. For Ridge regression the penalty is calculated by the sum of the squared coefficients multiplied by α, while the Lasso penalty is calculated by the sum of the absolute coefficients multiplied by α. The larger α is, the less complex the model becomes and the more each coefficient is penalized (Gupta, 2017). Hyperparameter tuning to determine the optimal value of α is carried out for both regularizations using GridSearchCV.

After applying both types of regularization to the dataset used in the multivariable linear regression, the overfitting issue was not addressed. The scores are displayed in Table1.

A blue and white rectangular box with numbers

Description automatically generated

Table 1: Model accuracy comparisons for multivariable regression

Table1 shows regularization resulted in worse test accuracy scores. An explanation for regularization not being able to address this issue is that the model is too complex for the given data and task. The data may have too much variation or noise resulting in poor predictions.

#### Model Comparison:

Figure13 illustrates the comparisons between the regression models based on the score and the root mean squared error. The RMSE is defined as the average difference between values predicted and the actual values. It gives an idea of how well the model can predict the target. The lower the RMSE the better. The full table of regression metrics is shown in section 3 of appendix. Univariable linear regression is seen to outperform the rest of the models in all metrics.

A screenshot of a graph

Description automatically generated

**Figure 13: Regression model comparisons**

#### Forecasting population and dependency ratios:

Given that univariable linear regression was the most accurate model, this will be used for the purpose of running forecasts. The resultant forecasts are illustrated in Figures 14, 15 and 16. With this model, the population is forecasted to reach 5,646,660 by 2034. The old age dependency ratio is shown to continue an upward trajectory while the youth dependency ratio shows a downward trend in the next decade.

A graph with numbers and a line

Description automatically generated

**Figure 14**

A graph with numbers and lines

Description automatically generated

**Figure 15**

A graph of a number of people

Description automatically generated with medium confidence

**Figure 16**

### **Classification Models:**

The aim of this part of the project is to analyse the accuracies of different classification models in determining the dominant component of population growth. To do this a new variable *pop\_change\_comp* is derived (section 4.2 notebook) and merged to the dataset. Since the two principal components of population growth are net migration and natural growth (Canudas-Romo et al. 2022), this categorical variable is set to *Natural* where natural growth is greater than net migration, and *Migration* where net migration is larger. It is appropriate to treat this as a binary classification problem as the target variable (*pop\_change\_comp*) is a discrete variable that only has two possible outcomes. Since machine learning models work better with numerical values, label encoding is used to transform the *pop\_change\_comp* categorical values into numerical values.

Feature selection was also performed for these models. In this case it is not appropriate to measure the correlation between the independent variables and the target variable due to the discrete nature of the target. The correlation was evaluated between the independent features, with one feature from any pair with strong correlation (>75%) removed.

#### **Evaluation Metrics:**

A confusion matrix may be used to evaluation performance. This NxN matrix (N is the number of classes) compares the actual target values (columns) with those predicted by the models (rows). True Positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) can be determined from this matrix. This will give an indication at how well the model is performing at predicting each class, and how often the model is predicting the wrong class.

Comparison metrics may be evaluated such as accuracy (how often the classifier correctly predicts class), precision (measures the accuracy of positive predictions),recall (measures the completeness of positive predictions) and f1-score (the mean of the precision and recall) (Agrawal, 2023).

2 classification models are utilized: Logistic regression and K-nearest neighbours (KNN). The reason these two models were chosen is because they are generally considered amongst the most popular models for binary classification (Brownlee, 2020) and are quite simplistic.

#### **logistic regression:**

Logistic regression makes use of the sigmoid function to transform the input variables into a probability between 1 and 0 (indicating both classes in the target). To implement this in python the LogisticRegression function from sklearn is utilized. This produced the resulting confusion matrix:

A diagram of a logistic regression

Description automatically generated

**Figure 17: Logistic Regression Confusion Matrix**

The confusion matrix shows that in total 7 values were correctly predicted. The overall accuracy of the model was determined to be 0.636364.

#### **K-Nearest Neighbors (KNN):**

The KNN model uses proximity to make classifications about the grouping of an individual data point. The model determines the *k* (i.e number of neighbours) training data points that are closest to the test data point using a distance function (i.e Euclidean) and assigns the test data with the most common class label (Shahraki et al., 2017).

To implement this model using python the KNeighborsClassifier from sklearn is utilized. The model is initially run with the default number of neighbors (k=5), so see how well it performs before this hyperparameter is tuned. The initial confusion matrix obtained (section 4.2.2 notebook), indicates a high number of incorrect predictions. The initial accuracy is 0.45455, which is lower than logistic regression.

In order to tune the *k* (number of neighbors) hyperparameter, a range of values for *k* is input into the model. By plotting the range in training and testing accuracies for each value of *k (*Figure 18), the optimal number of neighbors may be deduced.

A graph with blue and orange lines

Description automatically generated

Figure 18

*k*=2 and *k*=4 may be appropriate values to select for this hyperparameter to optimise accuracy. However, the KNN classifier will perform differently depending on how the data is split. Using k-fold cross-validation will resample the data a selected number of times resulting in a better approximation for the optimal *k.* GridSearchCV is used to determine the best value of *k* by passing in a range of *k* values to be tested against each sample of data (section 4.2.3 notebook). The optimal *k* is determined to be 2.

The model is rerun with this updated hyperparameter, which results in the following confusion matrix.

A diagram of a number of blue squares

Description automatically generated

Figure 19

The tuned KNN model is much better than the model using the default number of neighbors. In this case a total of 8 classes are predicted correctly. The overall accuracy of the model was determined to be 0.727273.

#### **Model Comparisons:**

The below table and chart indicate the results of each of the classification models utilized on the dataset as part of this project. Based off all evaluation metrics KNN (2 neighbors) is the best model at classifying the target variable (*pop\_change\_comp*).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **model** | **accuracy** | **precision** | **recall** | **f1\_score** |
| Logistic Regression | 0.636364 | 0.666667 | 0.666667 | 0.666667 |
| KNN (5 neighbors) | 0.454545 | 0.5 | 0.666667 | 0.571429 |
| KNN (2 neighbors) | 0.727273 | 0.8 | 0.666667 | 0.727273 |

A graph of different colored bars

Description automatically generated

Figure 20

## Conclusion

Through the analysis of multiple datasets provided by the CSO, key insights into the components driving population growth, the changing age demographic and the outlook for Irelands future population were obtained. KNN classification model outperformed logistic regression in a simple model to classify the dominant component of population growth for a given year. By comparing multiple regression models to predict the population, it was concluded that univariable linear regression outperformed multivariable linear regression. Overfitting encountered by the latter could not be resolved with ridge or lasso regularization. The population of Ireland was forecast from 2023 – 2033 using the univariable model. This predicts a population of 5,646,660 in 2033. The old age dependency was also forecast using the same model, with the ratio increasing from 23.7% to 29.03% in the next 10 years.

This result highlights that although Irelands population is on an upward trajectory the age demographic is getting older. This should be taken into consideration by government departments and other decision makers for the future provision of services and investment.

## References

Government of Ireland (2018). Project Ireland 2040 - National Planning Framework. Available at:

<https://www.gov.ie/en/campaigns/09022006-project-ireland-2040/?referrer=http://www.gov.ie/2040/> (accessed 1 November 2023)

Department of Finance, Government of Ireland (2021). 2021 Ageing Report

Available at:

<https://economy-finance.ec.europa.eu/system/files/2021-05/ie_-_ar_2021_final_pension_fiche.pdf> (accessed 1 November 2023)

Central Statistics Office (2022). Census of Population 2022 – summary results

Available at:

<https://www.cso.ie/en/releasesandpublications/ep/p-cpsr/censusofpopulation2022-summaryresults/populationchanges/#:~:text=Population%20change%20is%20driven%20by,Ireland's%20population%20increased%20by%20387%2C274>.

(accessed 1 November 2023)

Morgenroth E. (2008). ESRI Report – Demographic Projections for the period until 2021

Available at: <https://www.esri.ie/system/files/publications/BKMNEXT166.pdf>

(accessed 1 November 2023)

CSO Datasets (2023). Population Estimates

Available at: <https://data.cso.ie/> (accessed 3 November 2023)

Hotz N. (2023). What is CRISP-DM?

Available at: <https://www.datascience-pm.com/crisp-dm-2/> (accessed 3 November 2023)

Weiss N. A. (2016). Introductory Statistics 10th Edition

SciPy Documentation (2023). , Available at:

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.binom.html>

Jim Frost (2021). Statistics By Jim – Poisson Distribution

Available at: <https://statisticsbyjim.com/probability/poisson-distribution/>

(accessed 6 November 2023)

Gandhi S., Mishra P., Pandey, C M. et al. (2019). Descriptive Statistics and Normality Tests for Statistical Data, Available at:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350423/#:~:text=The%20two%20well%2Dknown%20tests,the%20normality%20of%20the%20data>.

(accessed 6 November 2023)

Delua J. (2021). IBM - Supervised vs. Unsupervised Learning: What’s the Difference.

Available at:

<https://www.ibm.com/blog/supervised-vs-unsupervised-learning/>

(accessed 6 November 2023)

Radecic D. (2021). Towards Data Science – Time series from scratch, train/test splits and Evaluation Metrics , Available at:

<https://towardsdatascience.com/time-series-from-scratch-train-test-splits-and-evaluation-metrics-4fd654de1b37#:~:text=Train%2Ftest%20splits%20in%20time%20series,-In%20machine%20learning&text=For%20example%2C%20if%20you%20had,(2%20years)%20for%20testing.&text=And%20that's%20all%20there%20is%20to%20train%2Ftest%20splits>. (accessed 7 November 2023)

Schneider A, Hommel G. et al. (2010). Linear Regression Analysis, Available at:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992018/#:~:text=Univariable%20linear%20regression%20studies%20the,and%20b%20is%20its%20slope>.

(accessed 7 November 2023)

Gupta A. (2023). Feature Selection Techniques in Machine Learning, Available at:

<https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/> (accessed 8 November 2023)

Gupta P. (2017), Regularization in Machine Learning, Available at:

<https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a>

(accessed 8 November 2023)

Canudos-Romo V. (2022). The Components of Change in Population Growth Rates

Available at:

<https://pubmed.ncbi.nlm.nih.gov/35156686/>

(accessed 9 November 2023)

Brownlee J (2020). 4 Type of Classification Tasks in Machine Learning., Available at:

<https://machinelearningmastery.com/types-of-classification-in-machine-learning/>

(accessed 9 November 2023)

Agrawal S K. (2023). Metrics to Evaluate you Classification Model to take the right decisions., Available at:

<https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/#:~:text=Accuracy%20simply%20measures%20how%20often,the%20total%20number%20of%20predictions>.

(accessed 9 November 2023)

H R Shahraki et al. (2017). *K* Important Neighbors: A Novel Approach to Binary Classification in High Dimensional Data, Available at:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5742505/#:~:text=K%20nearest%20neighbors%20(KNN)%20are,classification%20in%20high%20dimensional%20problems>.

(accessed 9 November 2023)

Tufte, E (2001). The Visual Display of Quantitative Information. Cheshire: Graphic Press

APPENDIX:

1. Descriptive Statistics:

A table with numbers and text

Description automatically generatedA table with numbers and symbols

Description automatically generated

1. A group of graphs showing different distribution

   Description automatically generated with medium confidenceA group of blue and black graphs

   Description automatically generated with medium confidenceDistributions of pop\_df variables:
2. Comparison metrics for regression models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **models** | **r\_squared\_test** | **r\_squared\_train** | **rmse\_test** | **mae\_test** |
| univariate\_LR | 0.72775525 | 0.92572691 | 83884.066 | 1.359899 |
| multivariate\_LR | 0.05703462 | 0.97712671 | 156116.065 | 2.596025 |
| multivariate\_LR\_ridge | 0.024784259 | 0.97711622 | 158763.284 | 2.642507 |
| multivariate\_LR\_lasso | 0.05670609 | 0.97712671 | 156143.266 | 2.596496 |