**CCT College Dublin**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for Data Analytics*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *Population Estimation With Phyton Algorithms: Irish population forecast for 2024* |
| **Lecturer Name:** | *Sam Weiss/* *David Gonzalez*  [*Taufique Ahmed*](https://moodle.cct.ie/user/view.php?id=13427&course=1)*/* *Marina Iantorno*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Deniz Ozsoy Ugur |
| **Student Number:** | 2023195 |
| **Assessment Due Date:** | 12/11/2023 |
| **Date of Submission:** | 12/11/2023 |

**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. |

**Abstract**

*The population of the countries should be known to be more consistent with the social, economic, political, and environmental decisions regarding the future of them. For this reason, censuses are made with great cost and the future population of the country is estimated with different techniques. In this study, different phyton algorithms are used to forecast population. Ireland's years 1926-2023 The population estimated data set was modelled. The data set used was taken from the Irish Central Statistics Office. Machine learning applications, statistics, EDA applications and basic codes Python The programming language and Jupiter program interface are listed one by one.*

**Introduction**

With the development of technology, information produced by people and companies it has created a large data pile. This mass of data is rapidly increasing day by day is growing. For example, there are currently approximately 1 trillion web pages. One hour of video is uploaded to the video provider YouTube every second, and approximately 10 years of content is uploaded per day. 1 million per hour by Walmart company Trading is taking place and the company has more than 2.5 petabytes of data in its database. In 2023 the volume of data generated on the Internet in just one week has exceeded that of the entire 20th century is greater than the volume of data produced. These examples can be multiplied, but the main point is to sort through all this data pile with efficient algorithms extracting and using meaningful information. Studies conducted so far have used a variety of estimation techniques forecasting has been done. These techniques include engineering methods, statistical, it can be grouped under three main headings, methods, and artificial intelligence methods. Making future predictions with so much data is easier and faster with artificial intelligence.

In this study, the population of Ireland in 2024 was tried to be estimated, and the applied models and results were compared with each other. To apply and analyse machine learning models, the data set was made suitable for use, it was tried to be optimized statistically, and the closest predictions to reality were tried to be made with machine learning algorithms. The models applied in the subsequent results were checked with model success tests. Make a prediction for population, the country's population statistics, birth, death and migration rates, data on life expectancy at birth and sex ratio at birth are used. Although the data set used has a small number of variables, it has been observed that estimates regarding the richness of the data set do not give inconsistent or unbalanced results compared to previous years.

There is a strong relationship between the population dynamics of countries and their level of development. Countries use population data to plan national needs such as education, health, investment, and housing. About the country with population statistics, age-gender information such as distribution, workforce, education, place of birth; regarding the structure of the population, birth, data such as death and migration rates are available. Countries are responsible for their citizens future they produce population projections to plan their needs. Thus, the future of the country social, economic, environmental, and political decisions that will affect it can be taken healthier. For this reason, it is very important that the population projections produced are consistent. It is important the use of machine learning algorithms allows the estimated population it can make a significant contribution to increasing its accuracy and consistency.

In this study, machine learning, methodology, data collection, data preparation and visualisation, statistical methods, data optimization and model establishment stages, application, performances There are sections on comparison, results, and recommendations.

**Programming For Data Analytics**

An attempt was made to divide the relevant data set into specific tasks and make the codes more readable and manageable. Functional programming was performed during the data processing and analysis stages. Using fewer variables shortens and optimizes data processing processes. Data structure is important when working on large data sets. Appropriate data structures and algorithms were selected to store and process data for population estimation. The complexity of the dataset, the requirements of the project, and the size of the dataset were considered. By using the specified programming functions, the project was made more organized, effective, and manageable.

**Data Preparation & Visualisation**

The data set that will be given to the model for learning during the data preparation phase prepared. Storage, cleaning, formatting for preparation of data setting It has been merged, transformed, or reviewed. Data preparation process It varies depending on the purpose of the study for each data set. In recording the data setting, quality, technical Parts have been reviewed. Results of machine learning application without using clean data is open to debate. Data inconsistent and erroneous data in the set was cleared. By cleaning these, only the data set it could be ready for machine learning algorithms to use. Identifying target data using a database of compiled data and variables evaluated knowledge discovery. In the next stage, data cleaning was performed to better process the data you are working on predictions on imprecise data are for future predictions of similar erroneous, missing, attribute mismatch information cleaning has been done.

The focus of the next phase is to convert the pre-processed data into a usable format. This is achieved by narrowing down the diversity and collecting data Features for future evaluation. Data is managed and organized here. They are often combined into a single type. In the next stage, deformed data in the data set is eliminated. Visuals were created with the relevant data. As a result of the Grouping, Convergence and Regression process, a method (or methods) was selected according to expectations and wishes. In the last stage, Data were determined and explained based on interpretation. In short, data is cleaned, transformed, divided into relevant attributes, and framed in visual language. This allows evaluation of selected outputs. Let's detail the transactions carried out within the framework of these explanations.

First of all, we added the libraries to the interface. Warnings that will occur in the stream are prevented. A search was conducted for both the addition of the data frame and the missing data situation. It was observed that there was no missing data in the relevant data set. However, the values ​​in the data set differed between object and integer. Additionally, when the relevant data set was examined with. head () and. tail() function, it was seen that there was unnecessary data in the dataset. These unnecessary columns are deleted with the. drop() function. The relevant dataset columns are 'Statistic Label', 'Year', 'Single Year of Age', 'Sex', 'Unit' and 'Value'.

'Statistic Label' and 'Unit' columns are considered unnecessary for data. It was deleted from the data. In the 'Single Year of Age' column, the ‘all ages’ row is written separately for men, women and both genders, for example, the population number appears to be double in 1926 because each is stated separately. For this reason, unnecessary information was deleted from the relevant data set and new data sets were created. The df1 data set only shows the total population by year for both genders.

While creating the second data set, the data set was created by considering the population values ​​specified separately for both genders. The titles were renamed from the created df1 and df3 data sets (created step by step from the df and df2 data sets). When checked with the (.describe) function, all data types must be the same in order to be used in machine learning and statistical algorithms. All data is converted to integer type with the Label Encoder() function. The columns and rows of the data sets related to the. shape function has been checked. Converting some observations in the data set to a single type it was necessary. The data speaks the same language and must be expressed in the same units; required for images and graphics. This prevents the machine from making mistakes during the learning phase. After the data collection, cleaning and transformation process, there is no problem in terms of the data set or model during the running of the models. If detected, retrospective steps can be re-evaluated, and corrections can be made. Possible solutions to the problem using the models most suitable for solving the problem is reported. If data transformation operations have been applied to the data set, the resulting the units of the results are reconverted to express the desired meaning. However, such a problem was not encountered in this data set.

After the data cleaning and transformation process, necessary applications were made to interpret and present the data with visualization, graphs and diagrams. These visual displays of information communicate complex data relationships and data-driven insights in an easy-to-understand way.

A graph with a number of blue squares

Description automatically generated with medium confidence

In this box plot we can see that the population density was between after 1980 and after 2010. On the x-axis, two outliers that require further investigation are observed. Graphs showing the development and change of the population by years are indicated, as well as the 10 years with the highest population are also indicated.

We visualized the population growth in Dublin over the years. To understand all years and the years with the highest population, the 10 years with the highest population were examined. There may be many reasons why the population decreases or increases at certain times, but these situations will be ignored for now.

A graph of a number of bars

Description automatically generated with medium confidence

The image above is a graphical bar chart showing the values ​​in the weak night column. Character size values ​​are stored in drawing characters. Store in a sns bar chart with x, y, palette, and data values. Displays a labelled axis subplot.

A graph of different colored bars

Description automatically generated with medium confidence

There is no significant difference between the number of highly populated units in Dublin City and the number of units in the other top 10 population values. A lower value was observed only in the first five years compared to other years. We can also express the relevant observations as follows; After the first five observations, the population tends to increase.

A graph of different colored rectangles

Description automatically generated

**Machine Learning for Data Analytics**

Follow-up when a problem wants to be solved with machine learning algorithms There are main steps required. To solve the problem successfully, follow these steps: must be applied sequentially. Although these steps are explained in different ways in the literature, expresses similar meanings. The KDD method refers to a broader process of data mining. It can be used in telecommunications, e-commerce, science, for example genetic data analysis. SEMMA, on the other hand, describes the application process of machine learning algorithms in five steps. He explained. After these steps are completed, the model for the desired purpose is created. available. It can be used in the field of retail, energy, human resources, for example, recruitment processes. Additionally, the approach called CRISP-DM is used in the data mining application process. This approach Daimler, Integral Solutions Ltd. (ISL) with more than 200 leading companies in the world. Developed by the consortium. data mining machine the CRISP-DM approach is like the machine learning application process. It can also be used in the implementation of learning algorithms. It can be used in areas such as marketing, health, and finance. for example, stock forecast. The methods used in the application process were mentioned, but the steps were not detailed, in short, these steps are collecting the data, preparing it, running this data in the model, and reporting the results obtained. In addition to these steps, the SEMMA model includes training and evaluation of the model. In general terms, these applications include the analysis, preparation and evaluation of data and the analysis and evaluation of the results obtained.

Supervised learning method was used in the relevant data set, and supervised learning was used using the training set. In this way, it is possible to compare what is taught with reality, and the error difference between the error differences for each input algorithm is tried to be minimized. It can be defined as the difference between the output produced by this device and its actual value. At the stage of teaching the training set to the system, the error difference for each data is determined by comparing it with the predetermined error threshold value during the training process. If the error value is less than this threshold value, the system continues to be trained. This means that the training of the system is completed when it reaches the intended threshold value. After the model was trained, it was asked to predict the next year with past data and the resulting values were found to be consistent. These models aimed to give an output as close as possible.

Machine learning applications, which store data too large to be processed manually, have been used to program using models and obtain meaningful inferences from them. Performance comparisons have been tried to be optimized by using this sample data and more accurate predictions have been tried to be made. In addition, to understand the results of these applications, predictions can be made for the future and meaningful conclusions can be drawn for the past. In this example, statistical applications were first made to continue the problem to learn the model. The size of the data used was chosen effectively and efficiently, considering issues such as allocating sufficient space for storing the solutions and shortening the solution downtime of the computers used to solve the problem. The inferences made after learning the model must be meaningful. In some applications, the results obtained as interruptions are important, the time required to be performed and the size of the data to be processed are important.

For this purpose, data optimization and cleaning were carried out. Since the population values in the data set naturally follow a certain trend over the years, training the model with this method with all past values reduces the model performance. The parameter value is selected as 5. The data set, which went through the preparation stages with the Lasso, Linear Regression, Ridge methods Regression, Random Forest and KNN algorithms determined previously for population estimation, was compiled using Python software with Jupiter interface. It was analysed through its language and performance results were stated.

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Version** | **Rmse-Mse** | **Model Success** |
| Linear Regression | Filtered data set | 6733.640653 | cv:10 7544.11033042413 cv:5 7086.50487136968 |
| Random Forest | Filtered data set | train: 0.999, test: 0.996 |  |
| KNeighbors Classifier | Filtered data set | 65896051106 | Training set score: 0.6819873135648631 Test set score: 0.6375685806925153 |
| Ridge | Filtered data set | 209577.1111 | Training set score: 0.78 Test set score: 0.92 |
| Lasso | Filtered data set | Mse: 43922565488.75813 | Training set score: 0.78 Test set score: 0.92 Number of features used: 1 |

To enrich the data set we have and increase the number of variables (features), improvements were made in Linear regression, Lasso and Ridge Regression models to enable the machine to learn better. The bias/variance dilemma was considered when making these improvements. As a result of the experiments, the parameters that gave the best results were selected. Accordingly, training the model, Lasso and Ridge regression models gave the best results. However, another observation is that the model overfitting, learning in some applications in the test and train models created, and as a solution to this, the best predicted results are reported.

A graph with blue dots and red line

Description automatically generated

When the performance results were evaluated, Lasso and Ridge Regression models trained on the entire data set and filtered data set were compared with other models. It seems to give better results; this is because Lasso and Ridge Regression models are more suitable for time series data set in terms of working principle. Random forest, which is mostly used in classification problems, detects trends and seasonality on data sets containing time series, as it makes decisions by dividing into branches when making each decision. Additionally, KNN is easier to use than other algorithms, but it shows overfitting. For these reasons, regression algorithms give better results. Especially when evaluating filtered regression algorithms in the analyses performed on the data set, the Ridge Regression model gave more successful results. Compared to Linear Regression, Ridge Regression seems to give overlearning, and in this dataset, it also gave slightly better results by minimizing the problem. After the Linear and Ridge Regression models were trained with the entire data set, the results obtained after training with only the total population data were obtained. When the results were evaluated, a significant increase in performance was observed.

The diagram above recreates the relationship diagram as shown. The help of correlation plots in Python is that scatter plots show the direction, strength, and relationship between two quantitative variables. From this heat map, there is a very weak, close to zero or even negative correlation between the number of units and the years. Therefore, there is almost no linear relationship between the four variables.

A screenshot of a graph

Description automatically generated

The diagram above reproduces the relationship diagram as shown. The help of correlation plots in Python helps scatter plots show the direction, strength, and relationship between two quantitative variables. The correlation coefficient helps measure the strength of the relationship. Correlograms are used to quantify the features of a given dataset. These statistics are most important in technology, Python, and science to have powerful tools for computational purposes. NumPy, Pandas, and SciPy are comprehensive, well-documented, and fast association methods.

**Statistics For Data Analytics**

If a Bernoulli trial occurs n times (trials are independent and the probability of each success number is (p), the resulting random variable X is called a Binomial random variable. That is, the number of n-independent Bernoulli trials leads us to the Binomial distribution. The image above is used to show the code used to perform the binomial discrete distribution in Python. Math library modules are imported into the Python notebook to provide the necessary math operations on the dataset data. The factorial of the data is created using mathematics. The factorial function in a Python notebook. The probability of each outcome is developed using the expression shown in the figure above.

In a probability distribution called the normal distribution, the values of a variable are distributed symmetrically around the mean. This shows that the values are equally likely to vary above or below the mean, and that the values that are like the mean are very concentrated.

A graph of a number of blue rectangular objects

Description automatically generated with medium confidence

Bernoulli experiment is the basis of discrete distributions. If a random experiment has one of two discrete results, this experiment is called Bernoulli experiment. The results of the Bernoulli trial were unsuccessful; It is expressed as healthy, broken, or desired and undesirable. In the relevant dataset, the gender column was evaluated to compare the two cases. In the application, it was considered that the experiments could be repeated under the same conditions. The gender column was chosen so that events had only two outcomes. The probability of success (p) should not change from trial to trial. Each trial must be independent from each other.

A graph of a number of people

Description automatically generated with medium confidence

There are two important factors that affect the normal distribution plot. These are the mean value and standard deviation. In practice, it is accepted that many random variables that occur independently of each other show a normal distribution.

A graph showing value and value

Description automatically generated

A graph showing value and distribution

Description automatically generated

In a probability distribution called the normal distribution, the values ​​of a variable are distributed symmetrically around the mean. This shows that the values ​​are equally likely to vary above or below the mean, and that the values ​​that are like the mean are very concentrated. The image above is used to demonstrate the code in the Python notebook for developing the normal distribution.

A graph with a line going up

Description automatically generated

A graph with a line graph

Description automatically generated

In the Poisson distribution, time is divided into such small pieces that more than one event does not want to occur in these small pieces of time. In other words, the event either occurs or does not occur within that narrow time frame. For this reason, while the binomial distribution deals with the number of successes in n experiments, the Poisson distribution deals with the number of outcomes of interest in a certain interval. For the Poisson distribution to be used, events occurring in two discrete time intervals are independent of each other. The probability of the event of interest occurring within the defined interval is fixed and does not change.

**Conclusion**

In this study, population estimation was made with different machine learning algorithms and statistical applications. Different models have been developed and the best performing models have been implemented. According to the results obtained, machine learning algorithms reduce the population with a margin of error of one in a thousand. Predicting why it is difficult to predict a country's population by minimizing and analysing uncertainties in demographic data with machine learning algorithms can better predict the future population of that country. For this reason, the use of machine learning algorithms in population estimation provides and will make a significant contribution to the country. Thus, planning the country's national needs will be easier and social, economic, and environmental decisions will be made more consistently.

Population Using machine learning algorithms to extrapolate generated population estimates will improve machine learning algorithms to deliver more meaningful results with large data sets. However, due to the nature of the problem, population statistics and demographic data for country indicators have a limited data set. Especially in underdeveloped countries. Although the data set is insufficient, the consistency of the existing data is also questionable. In addition, the limited values in the relevant data set also pose a problem. Increasing the consistency of the data collected and the number of data will improve the population estimate. For example, when making a prediction for the population of 2050, the birth rates for each year for the intervening 27 years are variables such as births, deaths, and migrations that need to be predicted for each year obtained after this input data. These rates are estimated by machine learning. This data can be used in the production components method. It is anticipated that this study can be expanded by increasing the number of years. Using machine learning algorithms can make a significant contribution to predicting what will happen next for many years. Of course, the basis, analysis and evaluation of machine learning can be achieved through statistical applications. We can only perform the visualization and optimization of the selected data set and the statistical applications to be applied by making predictable predictions.

**References**

Thomas, R. K. (1997). "Using demographic analysis in health services planning: A case study in obstetrical services." RAND-PUBLICATIONS-MR-ALL SERIES

Tekin, A. T., Ozkale, N. L., & Oztaysi, B. (2019). Big Data Concept in Small and Medium Enterprises: How Big Data Effects Productivity. In Industrial Engineering in the Big Data Era Springer, Cham

Siegel, J.S. ve Swanson, D.A. (2004) The Methods and Materials of Demography. 2nd Edition, Elsevier Science & Technology, San Francisco, USA.

McCarthy, J., ve E. Feigenbaum, (1990), Arthur Samuel: Pioneer in machine learning: AI Magazine,

Box, G. E. and N. R. Draper (1987). Empirical model-building and response surfaces, John Wiley & Sons.

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., & Pelletier, F. (2012). "Estimating trends in the total fertility rate with uncertainty using imperfect data: Examples from West Africa." Demographic research

Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., & Saidur, R. (2014). A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renewable and Sustainable Energy Reviews,

Shmueli, G. ve K. C. Lichtendahl (2016). Practical time series forecasting: A handson guide, Axelrod Schnall Publishers.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). "Cross-validation." Encyclopedia of database systems

Bergmeir, C. ve J. M. Benítez (2012). "On the use of cross-validation for time series predictor evaluation." Information Sciences

Taylor, S. J. and Letham, B.(2018). Prophet: Forecasting at scale,

<https://data.cso.ie/table/PEA11>

<https://miuul.com/> - Machine Learning