**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Programming for DA  Statistics for Data Analytics  Machine Learning for Data Analytics |
| **Assessment Title:** | MSC\_DA\_Integr\_Repeat\_Sem1 |
| **Lecturer Name:** | John, Sam, Muhammad |
| **Student Full Name:** | Joyce Simiao |
| **Student Number:** | 2019029 |
| **Assessment Due Date:** | 06/08/2023 |
| **Date of Submission:** | 06/08/2023 |

**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. |

**This project analyses Ireland and Portugal’s Sector Employment and Earnings.**

**Introduction**

The analysis of this project is about sector employment and earnings of Ireland and Portugal, the main idea is to compare both datasets to extract information and examine employment trends and economic activities within both countries, presenting a variety of employment figures, earnings, and exploring sub-sector in order to get insights.

The Ireland dataset does not cover a variety of public sector sub-sectors, in the file, it is possible to see only the Civil Service, Defence force, Garda Siochana, Education, Regional Bodies, health, Semi-State Companies, Comercial Semi-State Companies, Non – Comercial Semi-State Companies, and the total.

However, Portugal’s dataset contains more details and provides more information following these categories: agriculture, manufacturing, electricity, water supply, sewerage, waste management, construction, transportation, lodging, food services, information and communication, finance, real estate, professional services, administrative services, public administration, education, healthcare, arts and entertainment, Other services and Activities of extraterritorial organizations.

The main libraries were used for cleaning and data structuring, such as pandas, matplotlib, numpy, etc. The groupby function was effective aggregation for the result of the average employment across quarters and sub-sectors in the example of Ireland.

The exploration of Ireland and Portugal’s public sectors off employment patterns and perspectives. Inferential statistics and Machine learning were applied and used to improve the models.

**Raw Data**

The first data analyzed is about Ireland, having features information about Civil Service, Defence, Garda Siochana, Education, Regional bodies, health, Semi-state companies, Commercial semi-state companies, Non-Commercial semi-state companies, Total Public Sector including semi-state bodies, Total Public Sector excluding semi-State bodies, from 2008 until 2023, split in quarters.

The advantage of using this data is that it provides specific employment data for different sub-sectors of Ireland’s public sector, allowing a good analysis through the years.

However, there are some potential drawbacks associated with data. The dataset only includes employment figures for specific public sub-sectors. Additional data on other aspects of public sector employment (e.g., salaries, job roles, demographic information) would be useful for a more comprehensive and deep analysis.

However, Portugal’s dataset is just about September 2022, with features information about Agriculture, manufacturing, Electricity, water supply; sewerage; waste management and remediation activities, Construction, Wholesale and retail trade, Transportation and storage, Accommodation and food service activities, information and communication, Financial and insurance activities, Real estate activities, Professional; scientific and technical activities, Administrative and support service activities, Public administration and defence, Education, Human health and social work activities, Arts; entertainment and recreation, Other service activities, Activities of extraterritorial organizations and bodies.

Portugal’s dataset has more information and more details about the public sector money was applied.

To manage and gain insights from both datasets, data preparation, and cleaning were performed using pandas library, which is the best for cleaning rows and columns, handling missing values, and converting numeric columns appropriately, especially in Portugal because the dataset is an Excel file with many worksheets.

Portugal’s dataset is an interesting dataset, the features “Numbers of employers”, “total gross earnings”, “regular gross earnings”, and “base gross earnings” are numerical columns with the real number while the features “number of employees.1”, “Total gross earnings.1”, “Regular gross earnings.1”, “Base gross earnings.1”represent percentages, for that reason the percentage columns were drop.

**Cleaning Dataset**

In order to structure and enrich the information about the dataset for machine learning tasks, it was decided to impute missing data in employment values using appropriate imputation techniques such as mean or median, as appropriate, also necessary add a column with the average for Portugal dataset.

Dummy variables were created for categorical features (e.g., sub-sector) in order to convert them to a numeric format suitable for machine learning models.

Log transformation was applied to employment values to deal with skewness and bring the data closer to a normal distribution, which can help regression models.

To use in machine learning models, relevant features from the dataset, such as "STATISTIC," "TLIST(Q1)," "Quarter," and "C02741V03309," were extracted.

Data is loaded in CSV about Ireland, Excel about Portugal, and Jason about GDP in Europe.

The idea of presenting the GDP in Ireland is just to show the numbers have increased through the years over GDP, Even during covid-19 breakdown, Ireland is expecting to reach 5.5% GDP growth in 2023 and estimates 5.0 in 2024 which explains the facility to increase salaries every year. According to the website Cornmarket, public sector workers in Ireland are expected to receive a 7.5% pay increase by October 2023, obviously, this percentage varies according to the role performed.

A graph showing the growth of the economy

Description automatically generated

For data processing, the main library used was Pandas, which is a powerful library for handling tabular data. Other libraries such as numpy and matplotlib were also applied to organize and restructure the dataset. For data aggregation, the groupby function was used to calculate the average employment for each quarter and sub-sector efficiently in Ireland’s dataset.

For the dataset about Portugal, the action to clean the same is a bit tricky, especially because the dataset has different worksheets to clean and prepare the dataset for analysis, The first rows were skipped due to NAN information and the last rows also have the same information were dropped.

The quarter column was converted in datetime to ease the manipulation with the aggregation function.

It is possible to identify growth and decline in various sub-sectors over time based on an analysis of average employment trends. Further forecasting and sentiment analysis was performed to gain insights into future trends in Ireland’s public sector.

The columns TLIST and Quarter are quite similar, so I transformed the column quarter in datetime.

The graph below shows the evaluation through the years in each sector described in the legend, the calculation is based on the mean value of each sector per year.

A graph of different colored lines

Description automatically generated

Graph 1 – Ireland’s Average Employment

The following graph is based on the Sub-sectors and the newly created column “Average Euro”, which represents the mean value paid in September 2022 for each sector. The sector with the highest gross earnings was the supply of electricity; gas; steam; and air conditioning.

A graph of a bar chart

Description automatically generated with medium confidence

Graph 2 – Portugal’s Average Employment

**Statistics**

Inferential statistics are used to analyse the variables in the dataset and draw conclusions about the population values, with a T-statistic of 6.254 and a p-value of 5.972e-10, there is strong evidence to reject the null hypothesis, indicating a significant difference in average employment in Ireland between the "Civil service" and "Defence" sub-sectors. Below, is the difference between defence and garda, being the T-statistic of -2.313 and p-value of 0.020.

The T-statistic indicates that the difference in employment values is highly unlikely to be due to random chance, and the p-value confirms the findings' statistical significance. As a result, It concludes that there is a significant difference in employment levels between these four sub-sectors of Ireland's public sector.

Average mean Squared error: 321393645.41 The result was the evaluation of the performance of a machine learning model prediction based on a test set (y\_test, y\_pred), computing the mean squared error between the test set's true target values (y\_test) and the predicted values (y\_pred). The measures show how well the prediction matches the actual values on average. The process helps assess the model’s predictive accuracy and generalization ability over the unseen dataset.

The descriptive statistics were applied for both datasets but Ireland’s dataset has more information than Portugal’s, Ireland got the result for the mean value for the column ‘value’ 20153.78 for 5368.0 observations, while Portugal got the result for mean 426.60 for 21 observations.

**Sentiment Analysis**

The sentiment analysis came from a small extracted text from the website CSO.ie, that says about labour costs and percentages, being the text: “Average hourly other labour costs increased by 27.2% across all economic sectors to €4.58 in Q1 2023 from €3.60 in Q1 2022. A significant factor in this growth was the ending of the Employment Wage Subsidy Scheme (EWSS) on 31 May 2022. EWSS payments to enterprises were recorded as subsidies and refunds received and were deducted from other labour costs during the period Q1 2020 to Q2 2022.” (2023)

The overall sentiment score was 0.0498511.

Based on the provided text data and the overall sentiment score of 0.049, which is slightly positive, it concluded that the sentiment toward public sector employment in Ireland is positive, on the other hand, is a neutral and short text, indicating a balanced mix of positive and neutral sentiment.

Based on the second sentiment analysis with Ireland and Portugal, the scores between both countries were quite different, even explaining the same subject.

Ireland: The sentiment score for Ireland is 0.3182, indicating a positive attitude towards the country. People's attitudes and sentiments towards Ireland appear to be generally positive.

Portugal: has a sentiment score of 0.0, indicating a neutral sentiment. Positive and negative sentiments about Portugal are expressed in equal measure in the text data.

The sample data used for sentiment analysis was quite small just to observe and represent a sample, as described in the code.

According to the analysis, there is a lack of evidence-based recommendations for public-sector employment in Ireland. A lack of information directing specific addresses within Ireland’s public sector, so it would be possible to appropriate allocations of resources and efforts. Compared with Portugal’s dataset, it is possible to see they have more details about each sector, so Ireland could have divided better and explored more sectors and add these strategies in the dataset, focusing on the potential growth and investing, training, and developing the efficiency of the public sector.

One of the challenges for this project was the availability of data for another country. Finding relevant and reliable data for comparison with the same sectors, or similar sectors as Ireland. Also is critical to select the appropriate statistical test and machine learning for comparisons and find valid conclusions to ensure that every expectation was met.

**Machine Learning**

For the graph below, two regression models were applied, Linear Regression and Ridge Regression, in this code to predict employment values based on the provided dataset. Preparing the dataset by one-time encoding categorical features and tuning the Ridge Regression model's hyperparameters using GridSearchCV. Finally, using Mean Squared Error (MSE) and R-squared (R2) score metrics to evaluate both models and present the results in bar charts.

A comparison of a bar graph

Description automatically generated

To evaluate the impact of changes in the Portugal dataset, was implemented A/B testing using Python's scipy.stats.ttest\_ind() function. Dividing the dataset into control and experiment groups and simulating a change in the experiment group. After calculating the mean 'Average Euro' for both groups, performing a two-sample t-test to check for statistical significance. Based on the p-value, we determined whether the change had a significant impact or not.

Giving the result of 1151.42 for the control mean and for the experiment mean result was 1196.395.

Applying the inferential statistic, to compare “average euros” in Ireland and Portugal running a t-test to see if there were significant differences between the countries.

For machine learning tasks were used two approaches: Supervised Learning (SVM) for sentiment analysis and Unsupervised Learning (K-means) for clustering.

**Supervised Learning (SVM) for Sentiment Analysis:**  data divided into training and testing sets, created a TF-IDF vectorizer to convert textual data into numerical vectors, and defined the SVM model. Using GridSearchCV, It was performed hyperparameter tuning to find the best SVM model. After evaluating the model's accuracy and classification report, it gained insights into its performance for sentiment analysis.

**Unsupervised Learning (K-means) for Clustering:** The Principal Component Analysis (PCA) was applied for dimensionality reduction, transforming the textual data into a lower-dimensional space, then defining the number of clusters for K-means and creating the K-means model. Evaluating the clustering quality using the silhouette score which is the best metric to apply for this specific dataset and clustering, gaining insights into how well the data was clustered. The number of clusters selected was 2.

The silhouette score was 0.70, being a positive number because the range for the silhouette is from -1 to +1, where the higher number indicates better defined clusters.

**Conclusion**

The comparison of sector employment in Ireland and Portugal shows some interesting insights into the economy of the Countries.

The first dataset about Ireland includes details about Civil Service, Defence Force, Garda Siochana, Education, Regional Bodies, Health, Semi-State Companies, Commercial Semi-State Companies, Non-Commercial Semi-State Companies, and the Total Public Sector Employment. This enables a thorough examination of historical employment trends, even the dataset does not contain other sectors, job descriptions, or demographic data.

While Portugal represents a good sample of public payment, including more sectors but still do not show about demographic data.

Various data preparation and cleaning approaches were used to analyse and compare the sector employment in both nations. The data was organised and cleaned using the pandas package to make sure it was ready for analysis. The groupby function was used for Ireland.

Inferential statistics and machine learning approaches are used in the analysis to improve the models and get a better understanding of the data. By using these methods, it is feasible to point to employment patterns, investigate the effects of various industry sectors on the entire employment landscape, and forecast upcoming trends.

To handle categorical features, dummy variables were created by converting them into a numeric format suitable for machine learning models. This allows the models to effectively utilize these features in their computations and predictions.

To address skewness in the employment values, a log transformation was applied. This technique helps bring the data closer to a normal distribution, which is beneficial for regression models. By reducing the skewness, the models can make more accurate predictions using the transformed employment values.

To focus on relevant features for machine learning, a subset of columns like "STATISTIC," "TLIST(Q1)," "Quarter," and "C02741V03309" were chosen. These features help understand patterns and relationships in the data for more accurate predictions.

Inferential statistics were also applied to analyze the variables in the dataset and draw conclusions about the population values. For example, a T-statistic and a p-value were calculated to determine the significance of the difference in average employment between different sub-sectors in Ireland's public sector. The statistical analysis revealed a significant difference in employment levels and provided insights into the dynamics of the public sector in Ireland.

The performance of machine learning models was evaluated using metrics such as mean squared error (MSE) and R-squared score. These metrics helped assess the accuracy and generalization ability of the models in predicting the employment values.

A sentiment analysis was done to assess feelings about public sector employment in Ireland. Data from a website was used to calculate an overall sentiment score, showing a slightly positive sentiment.

The study also examined supervised and unsupervised learning techniques for machine learning problems. Support Vector Machine (SVM), a method of supervised learning, was used for sentiment analysis. The dataset was split into training and testing sets, and textual information was transformed into numerical vectors. The SVM model's precision and effectiveness in sentiment analysis were trained and tested.

Overall, the data cleaning techniques employed in this research ensured that the dataset was prepared and enriched for machine learning tasks. By handling missing data, transforming variables, selecting relevant features, and applying statistical analyses, valuable insights were gained, and accurate predictions could be made using machine learning models.

**References**

(No date) *Statistics portugal - web portal*. Available at: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine\_destaques&DESTAQUESdest\_boui=540817009&DESTAQUESmodo=2 (Accessed: 29 July 2023).

Bhardwaj, A. (2020) *Silhouette coefficient : Validating clustering techniques*, *Medium*. Available at: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c (Accessed: 06 August 2023).

*Como Ler Arquivo em Excel com Várias Abas no python [pandas]* (2021) *YouTube*. Available at: https://www.youtube.com/watch?v=y9ZUqbuZeI8&t=1430s (Accessed: 31 July 2023).

Cueto, M. del (2020) *Grid search in python from scratch- hyperparameter tuning*, *Medium*. Available at: https://towardsdatascience.com/grid-search-in-python-from-scratch-hyperparameter-tuning-3cca8443727b (Accessed: 04 August 2023).

*Data - CSO - central statistics office* (2023) *CSO*. Available at: https://www.cso.ie/en/releasesandpublications/ep/p-lfs/labourforcesurveyquarter32022/data/ (Accessed: 01 August 2023).

Datopian (no date) *Country, regional and world GDP (gross domestic product)*, *DataHub*. Available at: https://datahub.io/core/gdp (Accessed: 04 August 2023).

*Dissemination Standards Bulletin Board (DSBB)* (no date) *Dissemination Standards Bulletin Board*. Available at: https://dsbb.imf.org/sddsplus/dqaf-base/country/PRT/category/EMP00 (Accessed: 06 August 2023).

*Earnings and labour costs Q4 2022 (final) Q1 2023 (preliminary estimates) - CSO - central statistics office* (2023) *CSO*. Available at: https://www.cso.ie/en/releasesandpublications/ep/p-elcq/earningsandlabourcostsq42022finalq12023preliminaryestimates/ (Accessed: 01 August 2023).

*Economic Forecast for Ireland* (2023) *Economy and Finance*. Available at: https://economy-finance.ec.europa.eu/economic-surveillance-eu-economies/ireland/economic-forecast-ireland\_en (Accessed: 04 August 2023).

*EHQ10 - Public Sector Employment and earnings - Public Sector Employment and earnings* (no date) *Data.Gov.IE*. Available at: https://data.gov.ie/dataset/ehq10-public-sector-employment-and-earnings/resource/062d0798-b666-407f-b72d-077f989bc057 (Accessed: 27 July 2023).

*GRIDSEARCHCV | Hyperparameter tuning | Machine Learning with scikit-learn python* (2021) *YouTube*. Available at: https://www.youtube.com/watch?v=TvB\_3jVIHhg (Accessed: 01 August 2023).

Guest and Guest (2023) *What is the average salary in Portugal for 2023 ?*, *Time Doctor Blog*. Available at: https://www.timedoctor.com/blog/average-salary-in-portugal/ (Accessed: 02 August 2023).

*How to implement linear regression from scratch with python* (2022) *YouTube*. Available at: https://www.youtube.com/watch?v=ltXSoduiVwY&t=847s (Accessed: 31 July 2023).

Huddar, M. (2023) *How to compute silhouette coefficient – K means clustering in machine learning by Mahesh Huddar*, *YouTube*. Available at: https://www.youtube.com/watch?v=FGXkbawTHRQ (Accessed: 06 August 2023).

*Integração entre Python e excel USANDO pandas e o openpyxl* (2021) *YouTube*. Available at: https://www.youtube.com/watch?v=IT7zPluDADk&t=656s (Accessed: 03 August 2023).

Kumar, A. (2023) *Kmeans silhouette score python example*, *Data Analytics*. Available at: https://vitalflux.com/kmeans-silhouette-score-explained-with-python-example/ (Accessed: 06 August 2023).

Martini                  , J. and Augusto          , F. (2018) *How to use silhouette score in K-means clustering from sklearn library?*, *Stack Overflow*. Available at: https://stackoverflow.com/questions/51138686/how-to-use-silhouette-score-in-k-means-clustering-from-sklearn-library (Accessed: 05 August 2023).

Naik, K. (2018) *Principle component analysis (PCA) using sklearn and python*, *YouTube*. Available at: https://www.youtube.com/watch?v=QdBy02ExhGI&t=310s (Accessed: 31 July 2023).

*Public sector pay increases* (no date) *Cornmarket*. Available at: https://www.cornmarket.ie/public-sector-pay-increase/ (Accessed: 06 August 2023).

Shah, P. (2020) *My absolute go-to for sentiment analysis - textblob.*, *Medium*. Available at: https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524 (Accessed: 02 August 2023).

Sharma, A. (2020) *Principal Component Analysis (PCA) in Python tutorial*, *DataCamp*. Available at: https://www.datacamp.com/tutorial/principal-component-analysis-in-python (Accessed: 01 August 2023).

Stamer, J. (2018) *StatQuest: Principal Component Analysis (PCA), step-by-step*, *YouTube*. Available at: https://www.youtube.com/watch?v=FgakZw6K1QQ (Accessed: 04 August 2023).