**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation and Visualization  Statistics for Data Analysis  Programming for Data Analysis  Machine Learning for Data Analysis |
| **Assessment Title:** |  |
| **Lecturer Name:** |  |
| **Student Full Name:** | DIANA FLORA NAMAEMBA |
| **Student Number:** | 2023385 |
| **Assessment Due Date:** | 12th November 2023 |
| **Date of Submission:** | 11th November 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. |

**INTRODUCTION**

Mortality rate is the number of deaths in a given population at a particular period with a particular disease as the underlying cause. The study of mortality rates is vital in public health research. Neoplasms, Diseases of the circulatory system, diseases of the respiratory system and external causes of injury and poisoning contribute to high mortality rates in Ireland. According to (*New research reveals where and how people die in Ireland - News & Events | Trinity College Dublin*, 2021) cancers which form part of neoplasms account for 31% of deaths, Diseases of the circulatory system account for 29% of deaths, Diseases of the respiratory system account for 13% of deaths and 4% of the deaths are as a result of external injuries and poisoning. These diseases are ranked as 2,9,10 and 19 respectively in the International Statistical Classification of Diseases and Related Health Problems (ICD)-10 disease classification.(*ICD-10 Version:2019*, no date) .

A neoplasm is an abnormal growth of tissues. Neoplasms can either be cancerous or benign(*ICD-10 - Wikipedia*, no date). Diseases of the circulatory system are diseases that affect the heart and blood vessels and make it hard for blood to flow throughout the body.(*Circulatory System Diseases: Risk Factors & Symptoms*, no date). Diseases of the respiratory system are diseases or disorders that affect the lungs and airways in turn affecting the human respiration.(*Respiratory disease | Definition, Causes, & Major Types | Britannica*, 2023). An injury is either a physical or physiological harm of the body that is caused by the body interacting with energy either in thermal, mechanical, electrical, chemical or radiant form or from extreme pressure in an amount that the body in either its physiological or physical form cannot tolerate. Lack of elements such as oxygen can also cause Injury. Poisoning on the other hand is damage to the body by toxins.(*22 - Injury, poisoning or certain other consequences of external causes - ICD-11 MMS*, 2023).

Modelling mortality is very important for the economy, life, demography and social insurance because mortality rates help determine various things such as insurance products prices, insurance liabilities’ etc.(Deprez, Shevchenko and Wüthrich, 2017) The use of machine learning models in modelling mortality has recently emerged. There has been application of various machine learning models, such as stochastic mortality models for estimation and forecasting the mortality rates.(Deprez, Shevchenko and Wüthrich, 2017). Machine learning models have also been used for mortality trend prediction. (‘11-23-22\_Mortality-trend-prediction-using-ML’, 2022). Machine learning techniques have also allowed for the study of the adequacy of the mortality rates that have been estimated.(Deprez, Shevchenko and Wüthrich, 2017) .Mortality rate modelling also helps assess the quality fit of the estimated mortality rates estimated using stochastic methods.

For this study I propose to predict future mortality rates of the 4 ICD-10 Diagnostic diseases (Neoplasm, diseases of the circulatory system, diseases of the respiratory system and the External causes of injury and poisoning) using various machine learning approached and using past data.

**METHOD**

**Data Selection**

The data used for this study was obtained from the Central Statistics Office (CSO) - Ireland. The Central Statistics Office (CSO) is Ireland's national statistical office and it impartially collect, analyze and make available statistics about Ireland’s people, society and economy. The office houses a database with open access datasets for use. Data used was the Mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease Dataset- (G0315). The data had 6 features and 1892 observations. The features included one continuous variable and 5 categorical variables. All the categories were nominal features.

**Dataset Construction**

The dataset was defined in several steps: -

1. **Area of Residence**

All the areas of residence were retained except the area of residence called state. Thus only 42 areas of residence were included in the dataset.

1. **Years**

All years were retained except the year 2017, because it had missing mortality rates for the whole year and for all area of residence.

1. **ICD Diagnostic group.**

All the 4 ICD diagnostic groups were retained.

1. **Mortality rate**

The mortality rate feature had 172 missing data that were dropped. A total of 1680 observations were retained after dropping the missing data points.

**Data Preparation and Visualization**

Data preparation involved data preprocessing, feature selections and data splitting. Data preprocessing steps included: handling missing data, recategorizing features and addressing data imbalance.

**Step 1: Handling missing data.**

Initially the data consisted of 1892 observations and 6 features. There was one feature mortality rate that had missing data for the whole year of 2017. The missing data points were therefor dropped because any other missing data handling technique would make the data incorrect. These process of handling missing data ensured the data was reliable and meaningful in analysis. As a result, the data was reduced to 1720 observations and 6 features.

**Step 2: Removing features that were not used**

From the 6 features a total of 2 featured “statistic label” and “unit” were dropped because they were labels for 2 features in the dataset. The “area of residence” had a category called “state” that was an average of the mortality rate for each year for all area of residence for each disease. The category was also dropped since it skewed the data and was not meaningful for the analysis.

**Step 3: Encoding categorical data**

The dataset had two variables (“Area of Residence” and “The IDC-10 Diagnostic group”) that were categorical variables. The variables were encoded using the one-hot encoding method. The one-hot encoding was preferred because

**Step 4: Feature selection**

Feature selection is ranking the features from the dataset according to their importance based on the calculation of the information gain values for each of the selected variables. In this study, we have used a random forest model to find the risk factors or important features that have a major contribution to child mortality. The higher information gain values tell us important variables that are highly correlated with the class of variable. We randomly selected the top eight ranked information values, which we used in the model building later.

**Step 5: Scaling/ normalization**

**Step 6: data splitting**

1. You must also rationalise justify and detail all the methods used to prepare the data for ML. **[0-30]**

**Data Visualization**

3. Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis. **[0-20]**

4. All design and implementation of your visualizations must be justified and detailed in full. **[0-30]**

The pop\_data was split into independent variables called X and the dependent variable called y. The X and y variables were then split into Training and test sets.

Training set: X\_train and y\_train which includes 70% of the X data and y data respectively.

Test set: X\_test and y\_test which include 30% of the X data and y data respectively.

**Programming: :** (Graded out of 100)

1. The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). **[0-50]**

**Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.**

1. Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. **[0-50]**

**Statistics**

*Descriptive Statistics*

Descriptive analysis was conducted. These made it possible to understand the data. The measures of central tendencies: mean, median and frequencies and the measures of dispersion were calculated for the mortality rate of each disease (Neoplasm, Diseases of the circulatory system, Diseases of the respiratory system and External causes of injury and poisoning).

**Mortality rate by ICD Diagnostic group**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ICD-10 Diagnostic group** | **Mortality rate (per 1000)** | | | |
| **n (Mean)** | **Median (SD)** | **IQR** | **Min-Max** |
| Diseases of the circulatory system | 420(2.12) | 2.11(0.35) | 1.93-2.30 | 1.00-3.92 |
| Diseases of the respiratory system | 420(0.80) | 0.79(0.14) | 0.71-0.87 | 0.35-1.48 |
| External causes of Injury and Poisoning | 420(0.36) | 0.35(0.10) | 0.29-0.41 | 0.10-0.75 |
| Neoplasms | 420(1.92) | 1.91(0.23) | 1.79-2.05 | 0.93-2.73 |

The diseases of the circulatory system had the highest mean mortality rate compared to the other diseases.

**Mortality rate by disease by year**

Summarise your dataset clearly, using relevant descriptive statistics and appropriate plots. These should be carefully motivated and justified, and clearly presented. You should critically analyse your findings, in addition to including the necessary Python code, output and plots in the report. You are required to plot at three graphs. [0-35]

Firstly, the normality of distribution was analysed by the Shapiro–Wilk test for continuous variables. Secondly,

intergroup comparisons between continuous variables with

normal distributions were performed by Student’s t-test and

presented as mean ± standard deviation, while continuous

variables with skewed distributions were compared using

the Mann–Whitney U-test and described as median with

interquartile range. Thirdly, categorical variables were

analyzed by the Chi-square test or Fisher’s exact test. Finally,

ordinal variables were compared by the Mann-Whitney

U-test. A probability level of P-value <0.05 was taken as

statistically significant.

Frontiers

**Normality test**

Normality test of distribution for continuous variables was analysed using the Shapiro-wilk test.

**Distributions**

Use two discrete distributions (Binomial and/or Poisson) in order to explain/identify some information about your dataset. You must explain your reasoning and the techniques you have used. Visualise your data and explain what happens with the large samples in these cases. You must work with Python and your mathematical reasoning must be documented in your report. [0-30]

**Normal Distribution**

Use Normal distribution to explain or identify some information about your dataset. [0-20]

Explain the importance of the distributions used in point 3 and 4 in your analysis. Justify the choice of the variables and explain if the variables used for the discrete distributions could be used as normal distribution in this case.

**Inferential Statistics** is a statistical approach that involves use of statistical techniques to draw inferences to the whole population from a sample. (Han, Kamber and Pei, 2011). This was conducted so that we can draw conclusions for the population of Ireland from the sample 1680. The inferential techniques employed were: -

**Programming: :** (Graded out of 100)

1. The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). **[0-50]**

**Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.**

1. Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. **[0-50]**

**Machine learning for Data Analytics**

1. Explain which project management framework (CRISP-DM, KDD or SEMMA) is required for a data science project. Discuss and justify with real-life scenarios. Provide an explanation of why you chose a supervised, unsupervised, or semi-supervised machine learning technique for the dataset you used for ML modeling. **[0 - 20]**

Supervised machine learning technique was used because

1. Machine learning models have a wide range of uses, including prediction, classification, and clustering. It is advised that you assess several approaches (at least two), choose appropriate hyperparameters for the optimal outcomes of Machine Learning models using an approach of hyperparameter tunning, such as GridSearchCV or RandomizedSearchCV. **[0 - 30]**

**K-Nearest Neighbors regression**

**Decision Tree Regression**

**Random Forest Regression**

**Linear Regression**

**Ridge Regression**

**Ridge Regression (gridSearchCV)**

**Lasso Regression**

**Lasso Regression (gridSearchCV)**

**Support Vector Machine**

**Support Vector Machine (gridsearchCV)**

1. Show the results of two or more ML modeling comparisons in a table or graph format. Review and critically examine the machine learning models' performance based on the selected metric for supervised, unsupervised, and semi-supervised approaches. **[0 - 30]**
2. Demonstrate the similarities and differences between your Machine Learning modelling results using the tables or visualizations. Provide a report along with an explanation and interpretation of the relevance and effectiveness of your findings. **[0 - 20]**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ML Algorithm** | **n (%)** | **Training Set (n)** | **Test Set (n)** |  |  |  |
| K-Nearest Neighbors regression |  |  |  |  |  |  |
| Decision Tree Regression |  |  |  |  |  |  |
| Random Forest Regression |  |  |  |  |  |  |
| Linear Regression |  |  |  |  |  |  |
| Ridge Regression |  |  |  |  |  |  |
| Ridge Regression (gridSearchCV) |  |  |  |  |  |  |
| Lasso Regression |  |  |  |  |  |  |
| Lasso Regression (gridSearchCV) |  |  |  |  |  |  |
| Support Vector Machine |  |  |  |  |  |  |
| Support Vector Machine (gridsearchCV) |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

**Programming: :** (Graded out of 100)

1. The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). **[0-50]**

**Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.**

1. Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. **[0-50]**