**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation and Visualization  Statistics for Data Analytics  Programming for Data Analytics  Machine Learning for Data Analytics |
| **Assessment Title:** | Prediction of Mortality Rates for Cardiovascular Disease, Cancer, Diabetes, and Chronic Respiratory Diseases in the Population of Ireland |
| **Lecturer Name:** | 1. David McQuaid 2. Sam Weiss 3. Mohammed Iqbal 4. Taufique Ahmed |
| **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 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. |

**Contents**

[**List of Acronyms** 3](#_Toc150623317)

[**1.0 INTRODUCTION** 4](#_Toc150623318)

[1.1 Background 4](#_Toc150623319)

[1.2 Problem Statement 4](#_Toc150623320)

[**2.0 METHOD** 5](#_Toc150623321)

[2.1 Data Selection 5](#_Toc150623322)

[2.2 Level of Measurements of the Data Features 5](#_Toc150623323)

[2.3 Data Preparation and Visualization 5](#_Toc150623324)

[**2.3.1 Early Data Analysis (EDA)** 5](#_Toc150623325)

[**2.3.2 Data Visualization** 6](#_Toc150623326)

[**2.3.3 Data Preparation** 6](#_Toc150623327)

[**2.3.4 Data Visualization after Data Cleaning** 8](#_Toc150623328)

[**2.3.5 Programming in Data Preparation and Visualization** 9](#_Toc150623329)

[2.4 Statistics 10](#_Toc150623330)

[**2.4.1 Descriptive Statistics** 10](#_Toc150623331)

[**2.4.2 Coefficient of variation** 11](#_Toc150623332)

[**2.4.3 Density Probability Distribution** 11](#_Toc150623333)

[**2.4.4 Skewness and Kurtosis** 11](#_Toc150623334)

[**2.4.5 Cumulative Density Function** 12](#_Toc150623335)

[**2.4.6 Normality Tests (Normal Distribution)** 12](#_Toc150623336)

[**2.4.7 Programming for Statistics** 13](#_Toc150623337)

[2.5 Machine Learning 13](#_Toc150623338)

[**2.5.1 Data Splitting** 13](#_Toc150623339)

[**2.5.2 Supervised Machine Learning Algorithms Selection and model building** 14](#_Toc150623340)

[**2.5.3 Model Evaluation** 14](#_Toc150623341)

[**2.5.4 Programming in Machine Learning** 15](#_Toc150623342)

[**3.0** **REFERENCES** 16](#_Toc150623343)

# **List of Acronyms**

* MAE Mean absolute error
* MAPE Mean absolute percentage error
* MSE Mean square error
* R2 R-squared, coefficient of determination
* RMSE root mean square error
* SMAPE symmetric mean absolute percentage error
* ICD International Statistical Classification of Diseases and Related Health Problems
* CSO Central Statistics Office
* EDA Early Data Analysis
* MCAR missing completely at random
* CV coefficient of variation
* SD Standard Deviation
* IQR Interquartile ranges
* Min Minimum
* Max Maximum
* CDF cumulative density function
* KDD Knowledge discovery in databases
* SEMMA The sample, explore, modify, model, and assess
* CRISPM-DM The Cross-industry standard process for data mining
* ML Machine Learning

# **1.0 INTRODUCTION**

## 1.1 Background

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 that 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 system. (*Respiratory disease | Definition, Causes, & Major Types | Britannica*, 2023). An injury is either a physical or physiological harm to 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 oxygen or any other elements 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).

Modeling mortality is very important for the economy, life, demography, and social insurance because mortality rates help determine various things such as insurance product prices, insurance liabilities, etc.(Deprez, Shevchenko and Wüthrich, 2017) The use of machine learning models in modeling 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 modeling also helps assess the quality fit of the estimated mortality rates estimated using stochastic methods.

## 1.2 Problem Statement

The CSO of Ireland collected 1892 mortality rate values for cardiovascular disease, cancer, diabetes, or chronic respiratory diseases for the population of Ireland for 42 areas of residence. Certain Features/variables of the data have been defined. The aim is to build a predictive model for the mortality rates of the four diseases in a particular area of residence in a particular year using various machine-learning approaches and past data.

# **2.0 METHOD**

## 2.1 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 collects, analyzes, and makes available statistics about Ireland’s people, society, and economy. (*Population Changes - CSO - Central Statistics Office*, no date). 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.

## 2.2 Level of Measurements of the Data Features

There was a total of 6 features but only 4 were used in this study

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Variable Type** | **Level of Measurement** | **Study Use (Y/N)** |
| Statistic Label | Independent variable | Nominal | N |
| Area of Residence | Independent variable | Nominal | Y |
| ICD 10 Diagnostic Group | Independent variable | Nominal | Y |
| Year | Independent variable | Nominal-Integer | Y |
| Unit | Independent variable | Nominal | N |
| Mortality Rate (per 1000) | dependent variable | Numerical | Y |

## 2.3 Data Preparation and Visualization

### **2.3.1 Early Data Analysis (EDA)**

Early data analysis was performed to better explore the data. The primary aim of the EDA was to examine the data’s distribution, outliers, and any anomalies that would be used to generate specific hypotheses for testing. (*Secondary Analysis of Electronic Health Records*, 2016). EDA was important because it aimed to assist in pattern recognition. The steps involved in EDA included: -

1. Viewed the dataset head and tail, this displayed the top and bottom rows of the dataset giving an overview of the data contents and structure. (*head () and tail () Functions Explained with Examples and Codes*, 2023)
2. Checking the shape of the dataset: how many rows and columns the dataset had. The dataset had 1892 observations and 6 observations
3. Checking the dataset’s data types (continuous or discrete). The dataset had 5 categorical variables(year was a nominal integer) and only one continuous variable.
4. Calculate the summary statistics of both numeric and categorical variables.
5. Checking for any missing data and duplicates.
6. Checking for outliers

Ref: JupyterNotebook Line 1-12

### **2.3.2 Data Visualization**

Data visualization was then performed. This is because it’s the fastest way to learn about the data. (*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020). Various data visualization techniques were used. A pair plot was plotted because it helps understand the relationship between variables. There was no relationship between the continuous variables year and Mortality Rate. A heatmap was also preferred because it visualizes the relationship between an independent variable and a dependent variable. (*Secondary Analysis of Electronic Health Records*, 2016) . A heatmap for the correlation of the variables in the dataset was plotted. The results showed there was a weak negative correlation between year and mortality rate.

Ref. Python JupyterNotebook Line 13-14.

### **2.3.3 Data Preparation**

Data preparation involved data preprocessing steps: Data cleaning, feature selection, and data transformation.

(*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020)

#### **2.3.3.1 Data Cleaning**

Data Cleaning involved checking the following: -

1. ***Validity***- This involved checking the data types and the data ranges. This involved using the Python libraries pandas and NumPy. The data types for the four variables were integers and objects.
2. ***Completeness***- This involved checking the degree to which all the required data were known or available. Missing data was checked and handled appropriately
3. ***Uniformity***- This involved checking the degree to which the data was specified using the same units of measure.

(*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020)

The data cleaning steps included: -

**Step 1: Handling missing data.**

Initially, the data consisted of 1892 observations and 6 features. One feature ‘mortality rate’ had missing data for the whole year of 2017. The data was missing completely at random (MCAR) (Jakobsen *et al.*, 2017) therefore the missing data points were dropped because any other missing data handling technique would make the data incorrect because the mortality rate values were obtained for each year, each area of residence and for each disease. This process of handling missing data ensured the data was reliable, meaningful in analysis, and most importantly unbiased(Kang, 2013). 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 areas of residence for each disease. The ‘area of residence’ also had 8 provinces in the data, i.e., border, midland, west, Dublin, Mid-east, Mid-West, South-East, and South-West. (*cso ireland regions - Google Search*, no date) The provinces’ mortality rate values equaled the average of all the areas in that particular province.

The areas in each province are as shown below: -

* **Border** – Cavan, Donegal, Leitrim, Louth, Monaghan, Sligo.
* **Midland** – Laois, Longford, Offaly, Westmeath.
* **West** – Galway, Mayo, Roscommon.
* **Dublin** – Dublin City, Dún Laoghaire-Rathdown, Fingal, South Dublin.
* **Mid-East** – Kildare, Meath, Wicklow.
* **Mid-West** – Clare, Limerick, North Tipperary.
* **South-East** – Carlow, Kilkenny, South Tipperary, Waterford, Wexford.
* **South-West** – Cork, Kerry.

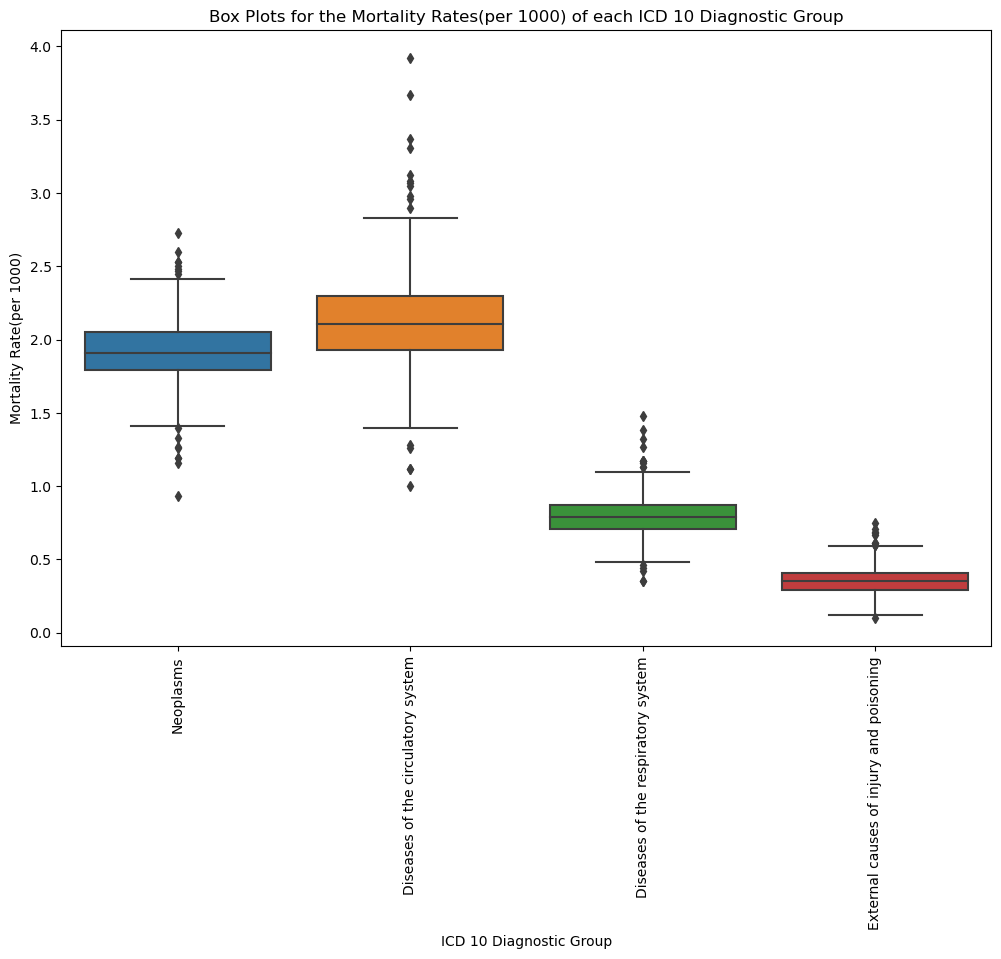
The provinces were therefore dropped because they skewed the data and were not meaningful for the analysis since the areas were already present in the data.

**Step 3: Encoding categorical data**

The dataset had two variables (“Area of Residence” and “The IDC-10 Diagnostic group”) that were categorical. The variables were encoded using the one-hot encoding method. The one-hot encoding was preferred because the categorical variables were nominal i.e., they did not have any order. (Dahouda and Joe, 2021). A total of 38 new columns were created, changing the shape of the data to 1360 observations and 40 features.

**Step 4: Handling Outliers**

A box plot was plotted for the Mortality rate variable. The results showed that each disease had outliers as shown below: -



Outliers affect significantly small sample size outcomes when one performs any robust statistical procedures such as coefficient of variation and variance, hence the need to handle them. (Cousineau and Chartier, 2010). There are over 20 techniques for handling outliers. The keep method (*which is a technique of handling outliers that acknowledges the presence of outliers but doing nothing about the outlier values considering any prior analysis*.) was chosen because the outlier was an influential outlier, i.e., accurate data points that are at a distance from the other points and are neither error nor interesting outliers.(Aguinis, Gottfredson and Joo, 2013)

#### **2.3.3.2 Feature selection**

Feature selection is the process of selecting features that are useful to the model. Multicollinearity was used to determine which features to keep. Features with a correlation coefficient of 0.95 and above would be removed.

(*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020)

A heatmap for the correlation of the features was plotted. The variable Mortality rate and year were checked for collinearity. There was a very weak negative relationship between the two variables hence both of them were included in the model.

#### **2.3.3.3 Data Transformation**

Data transformation is a data preparation technique that makes sure the data is in the best possible manner in the machine learning algorithms.

(*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020).

Data transformation can be achieved through the following steps: -

1. ***Rescaling***- is the process of rescaling the scale of all attributes if they are not of the same scale to the same scale.
2. ***Standardization***- is a technique that transforms attributes into a standard normal distribution.
3. ***Normalization***- is rescaling the observations to have a length of one.

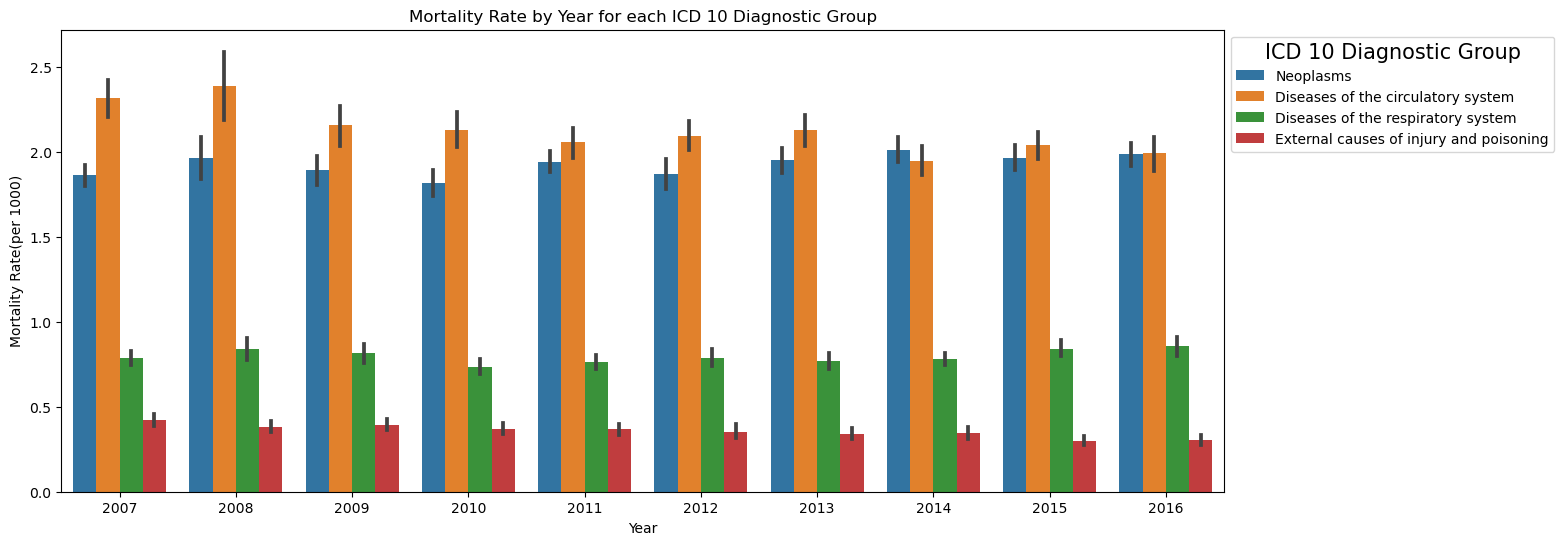
No data transformation technique was applied to the X (independent variables) because of the following reasons: -

1. The Year variable was a nominal integer, scaling it would make the feature lose meaning.
2. The area of residence and ICD-10 diagnosis group were encoded using one hot-encoding making them either 0 or 1. Thus there was no need to scale them.

### **2.3.4 Data Visualization after Data Cleaning**

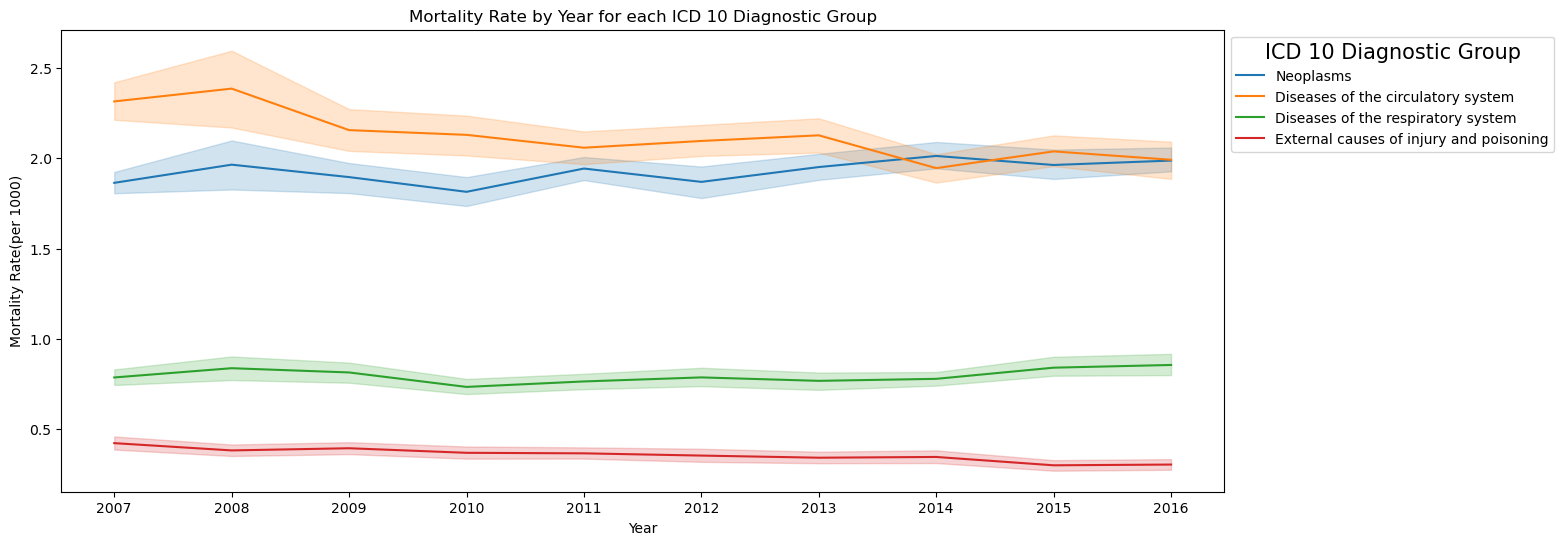
After cleaning the data several plots were made to answer the following questions: -

1. ***Which year had the highest mortality rate for each disease?***



The diseases of the circulatory system had the highest mortality rate over the 10 years while the External causes of injury and poisoning had the lowest mortality rate over the 10 years.

1. ***What is the mortality trend of each disease over the years?***



* The mortality rate of the diseases of the circulatory system and neoplasms was very high compared to diseases of the respiratory system and external causes of injury and poisoning.
* The mortality rate of the disease of the circulatory system decreased over the years.
* The mortality rate of neoplasms increased and then decreased over the years.
* The mortality rate of diseases of the respiratory system decreased and then increased over the years.
* The mortality rate of external causes of injury and poisoning decreased over the years

### **2.3.5 Programming in Data Preparation and Visualization**

#### **2.3.5.1 Python Libraries**

Several Python libraries were used for Data Preparation and visualization for this study. The pandas, NumPy, matplotlib, and seaborn were used. The pandas were used for EDA, data preprocessing, and data cleaning processes. Matplotlib and seaborn libraries were used for data visualization for both EDA and visualization after data cleaning.

#### **2.3.5.2 Data Structures**

The study employed various Python programming data structures. Lists, Dictionaries, and Data frames were used to perform various data preparation and visualizations.

#### **2.3.5.3 Programming Paradigms**

The imperative programming paradigm was mostly used and preferred during data preparation and visualization because it is a sequence of directions for performing actions. (Vujosevic Janicic *et al.*, 2008).When performing EDA: importing the dataset, checking the dataset shape, checking for missing data, and checking for the data types, a sequence of directions was coded so that the above actions could be performed. Using the imperative paradigm facilitated the creation of statements that allowed the performance of various data visualization and preparation techniques giving a better understanding of the data. For loops were also used, which are a part of imperative programming to generate various plots such as heatmaps.

## Statistics

Descriptive and predictive statistics were used in this study. Descriptive was used to understand the past data while predictive analysis was used to make predictions of the mortality rate for the four diseases.

### **2.4.1 Descriptive Statistics**

Descriptive statistics included calculating the measures of central tendencies: (mean, median, and frequencies) and the measures of dispersion (Variance, standard deviation, kurtosis) for the mortality rate of each disease.

1. **Mortality rate by ICD Diagnostic Group**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ICD-10 Diagnostic Group** | **Mortality rate (per 1000)** | | | | |
| **n (Mean)** | **Median (SD)** | **Variance** | **IQR** | **Min-Max** |
| Diseases of the circulatory system | 420(2.12) | 2.11(0.35) | 0.13 | 1.93-2.30 | 1.00-3.92 |
| Diseases of the respiratory system | 420(0.80) | 0.79(0.14) | 0.02 | 0.71-0.87 | 0.35-1.48 |
| External Causes of Injury and Poisoning | 420(0.36) | 0.35(0.10) | 0.01 | 0.29-0.41 | 0.10-0.75 |
| Neoplasms | 420(1.92) | 1.91(0.23) | 0.06 | 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.
* The external causes of injury and poisoning had the lowest mean mortality rate compared to the other diseases.
* The measures of central tendency and dispersion were also calculated for the mortality rate of each disease by area of residence and the mortality rate of each disease by year. Ref: python JupyterNotebook line 43-45.
* The year 2008 had the highest mortality rates

### **2.4.2 Coefficient of variation**

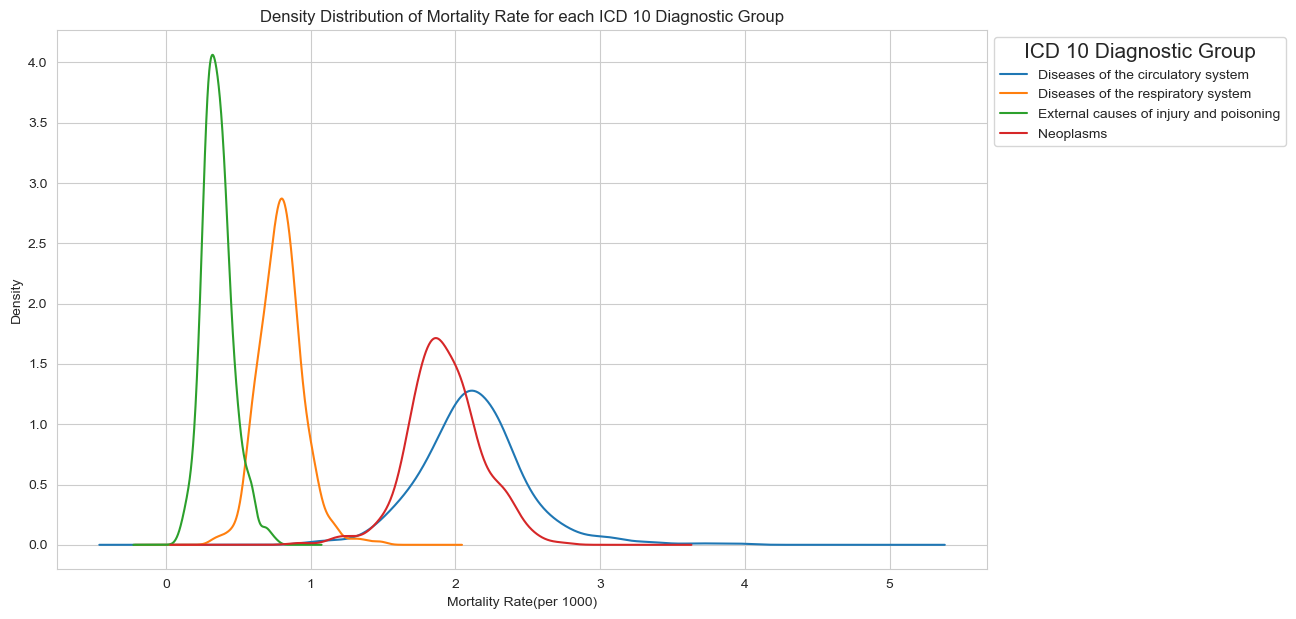
The coefficient of variation (CV) is the relative measure of variability for the size of the standard deviation to its mean. (*Coefficient of Variation in Statistics - Statistics By Jim*, 2020). CV was calculated for the mortality rate of each disease.

|  |  |
| --- | --- |
| **Disease** | **CV (%)** |
| Disease of the circulatory system | 13.66 |
| Disease of the respiratory system | 6.71 |
| External cause of injury and poisoning | 25.60 |
| Neoplasm | 10.14 |

The CV of the four diseases is less than 100% this means that the SD is less than the mean, meaning there is low variability in the dataset and this is acceptable.

### **2.4.3 Density Probability Distribution**

A density distribution and histogram were plotted to check the spread of the data. The density plots showed that external causes of injury and poisoning had a high peak while diseases of the circulatory system have a flat-topped curve.

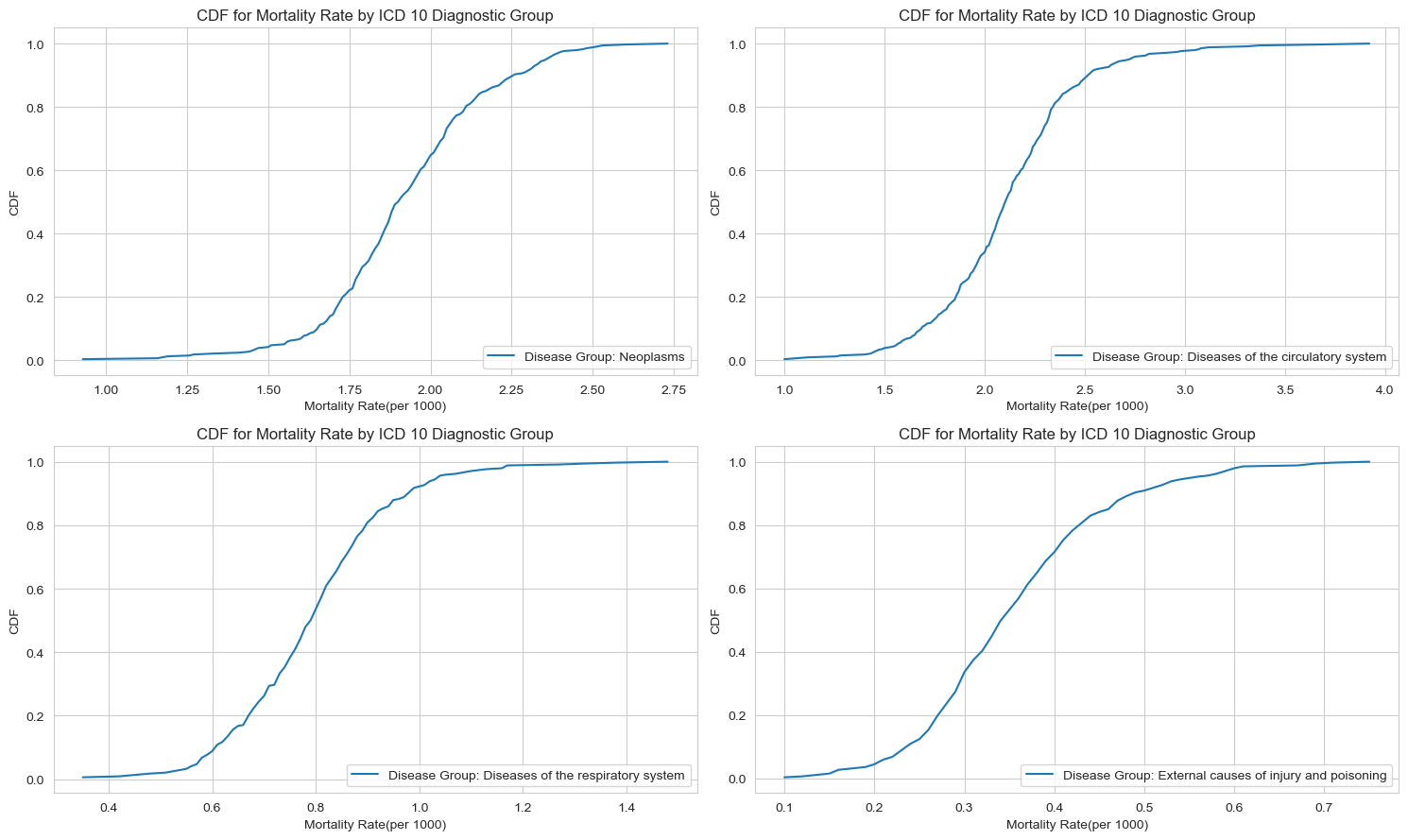


### **2.4.4 Skewness and Kurtosis**

Kurtosis is a measure of the peakness of a distribution while Skewness is a measure of the asymmetry of a distribution. A distribution is normally distributed if the skewness value is 0. A positively skewed data is one where the right-side tail is longer than the left and the bulk of the values lie left of the mean, while negatively skewed data is the inverse curve. (Kim, 2013). The kurtosis of the diseases was positive kurtosis indicating heavy tails. Three diseases were positively skewed while one was negatively skewed.

|  |  |  |
| --- | --- | --- |
| **Disease** | **Skewness** | **Kurtosis** |
| Disease of the circulatory system | 0.70 | 2.98 |
| Disease of the respiratory system | 0.55 | 1.88 |
| External cause of injury and poisoning | 0.67 | 0.95 |
| Neoplasm | -0.09 | 1.05 |

### **2.4.5 Cumulative Density Function**

****A CDF was plotted to visualize the shape of the distribution of the mortality rate for each disease

### **2.4.6 Normality Tests (Normal Distribution)**

The normality test was done using Q-Q plots and the Kolmogorov-Smirnov test. The mortality rate variable for each disease was tested.

#### **2.4.6.1 Q-Q plots**

The Q-Q plots for the mortality rate of each disease showed that the mortality rate for each disease is not normally distributed because the observations did not lie on the straight line.

#### **2.4.6.2 Kolmogorov-Smirnov Test**

Kolmogorov-Smirnov test was used to test for normality using p-values. It was used because it is better for sample sizes greater than or equal to 50 (Mishra *et al.*, 2019).

Set the hypothesis to test for normality:

H0: The Mortality rate variable for each disease follows a normal distribution

H1: The mortality rate variable for each disease does not follow a normal distribution

|  |  |  |  |
| --- | --- | --- | --- |
| **Disease** | **Test-Statistic** | **P-Value** | **Conclusion** |
| Disease of the circulatory system | 0.91 | 0.0 | The Mortality rate variable does not follow a normal distribution (reject H0) |
| Disease of the respiratory system | 0.68 | 5.05e-159 | The Mortality rate variable does not follow a normal distribution (reject H0) |
| External cause of injury and poisoning | 0.55 | 1.13e-96 | The Mortality rate variable does not follow a normal distribution (reject H0) |
| Neoplasm | 0.90 | 0.0 | The Mortality rate variable does not follow a normal distribution (reject H0) |

### **2.4.7 Programming for Statistics**

#### **2.4.7.1 Python Libraries**

Several Python libraries were used to perform different statistical procedures. The pandas, NumPy, matplotlib, and seaborn were used for various aspects of statistical analysis. The pandas were used to calculate measures of central tendency and dispersion. The pandas were preferred because they contained the grouping functions that were necessary for the analysis. Ref the jupyter notebook line 42. Matplotlib and seaborn libraries were used for data visualization of these measures.

#### **2.4.7.2 Data Structures**

The study employed various Python programming data structures, e.g., Dictionaries and lists were used to perform various statistical processes. Dictionaries were used to label the various visualizations. Lists were used when calculating the kurtosis and skewness values

#### **2.4.7.3 Programming Paradigms**

Imperative programming paradigm was mostly used and preferred during the statistical analysis process because it is a sequence of directions for performing actions. (Vujosevic Janicic *et al.*, 2008). When calculating measures of central tendencies and dispersion were coded using imperative programming so that the above actions could be performed. For loops were also used to facilitate plotting the cdf curves, Q-Q plots, and performing the Kolmogorov test.

Functional programming was also used in calculating measures of central tendencies because the codes involved calculating the quantiles using the agg function and the lambda function.

## 2.5 Machine Learning

Machine learning involved the consideration of three project management frameworks: CRISP-DM, KDD, and SEMMA. Knowledge discovery in databases (KDD) is the process of selecting the target data, pre-processing, transforming, data mining, and interpreting it. The sample, explore, modify, model, and assess (SEMMA) is the process of conducting a data mining project by sampling the data, exploring the data, modifying the data, modeling the data, and assessing the data by evaluating the results. The Cross-industry standard process for data mining (CRISP-DM) is a process that uses business understanding, data understanding, data preparation, data modeling, evaluation, and deployment of the results. (Martins, Pesado and García-Martínez, 2016).

The project management framework used for data science projects is the CRISP-DM. The CRISP-DM framework was applied in a study on Data Quality Improvement to Support Machine Learning of Stunting Prediction in Infants and Toddlers.(Purbasari *et al.*, 2021)

Supervised machine learning techniques were used because our data was labeled data. Secondly, the majority of work done or research done using mortality rate data leaned towards supervised learning as the best approach for modeling mortality.(Saroj *et al.*, 2022), (Deprez, Shevchenko and Wüthrich, 2017; Krittanawong *et al.*, 2021; ‘11-23-22\_Mortality-trend-prediction-using-ML’, 2022)

**The machine learning process involved the following steps: -**

1. Splitting the dataset.
2. Supervised ML Model selection.
3. Making predictions

### **2.5.1 Data Splitting**

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 as shown below: -

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

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

Data Splitting in training and test sets was very important because it helped find the most efficient set of model parameters that had the correct balance between the model complexity and the model’s generalization capabilities.(Eliane Birba, 2020).

### **2.5.2 Supervised Machine Learning Algorithms Selection and model building**

Regression algorithms were selected because the target variable/ independent variable i.e., the Mortality rate was continuous.(Shetty *et al.*, 2022). The following Regression models were applied: -

* ***K-Nearest Neighbors regression-*** is a regression model that makes predictions about a continuous target variable by identifying the K observations nearest to the new point we want to predict. (*Chapter 7 Regression I: K-nearest neighbors | Data Science*, 2023).
* ***Decision Tree Regression –*** is an algorithm that makes predictions of a continuous variable by forming decision trees by asking a series of questions and creating decision rules according to the dataset structure that constitutes the problem.(KOCARIK GACAR and DEVECİ KOCAKOÇ, 2020)
* ***Random Forest Regression-*** is an algorithm that uses a collection of tree predictors to make predictions about a continuous target variable. (Segal, 2003)
* ***Linear Regression-*** is an algorithm that finds if there is a relationship and dependency between variables. (Nasteski, 2017)
* ***Ridge Regression –*** is an advanced version of multiple linear regression that is used to tune the model to help reduce the complexities of the model decreasing overfitting problems. (Venkatesh, Mishra and Manimozhi, 2023)
* ***Lasso Regression-*** is a regularization technique that uses shrinkage for more accurate predictions
* ***Support Vector Machine-*** is an algorithm that maximizes the distance between the separating hyperplane and then trains the samples that are close to that hyperplane. (*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020)
* ***gridsearchCV-*** is hyperparameter tuning that involves picking out a grid of hyperparameter values evaluating them, then returning the one that is the best. It was applied for Ridge, Lasso, and support vector machines.

### **2.5.3 Model Evaluation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ML Algorithm** | **R2** | **Others features** | **Ranking** | **Characteristics** |
| **Test set** | * X (1360 observations 39 features) * Y (1360 observations 1 feature) * X\_train (952 observations 39 features) * Y\_train (952 observations 1 feature) * X\_test (408 observations 39 features) * Y\_test (408 observations 1 feature) |
| K-Nearest Neighbors regression | 0.899 |  | 6 |
| Decision Tree Regression | 0.906 |  | 5 |
| Random Forest Regression | 0.944 |  | 2 |
| Multiple Linear Regression | 0.936 |  | 3 |
| Ridge Regression | 0.936 |  | 3 |
| Ridge Regression(alpha=10) | 0.933 |  | 4 |
| Ridge Regression(alpha=0.1) | 0.936 |  | 3 |
| Ridge Regression (gridSearchCV) | 0.936 |  | 3 |
| Lasso Regression | -0.01 | features=0 |  |
| Lasso Regression(alpha=0.01) | 0.91 | features=6 | 4 |
| Lasso Regression (alpha=0.0001) | 0.94 | features= 38 | 2 |
| Lasso Regression (gridSearchCV) | 0.91 |  | 4 |
| Support vector Machine | -0.05 |  |  |
| Support Vector Machine (gridsearchCV) | 0.947 |  | 1 |

The coefficient of determination R2 was used to evaluate the model’s performance. It takes the range of values from (-infinity, 1], according to the mutual relation between the prediction model and the ground truth. (Chicco, Warrens and Jurman, 2021). It was preferred to SMAPE, MAPE, MAE, MSE, and RMSE, because it was the most informative rate in many model evaluation cases. (Chicco, Warrens and Jurman, 2021)

A negative R-squared indicated that the model performed poorly, therefore Lasso Regression(alpha=0.1) and support vector machine (c=1.0 and verbose= False) were not good models for predicting mortality rate. (Chicco, Warrens and Jurman, 2021).

All the other models were good in making predictions about the mortality rate for the four diseases. When ranking the best to the last, the best model for mortality rate prediction was the support vector machine (gridSearchCV).

### **2.5.4 Programming in Machine Learning**

#### **2.5.4.1 Python Libraries**

Libraries used for ML in this study included the scikit-learn library. It was used because it had all the regression models’ functions. Matplotlib library was also used for visualization.

#### **2.5.4.2 Data Structures**

The study employed various Python programming data structures. Lists, Dictionaries, and Data frames were used. Dictionaries were used when printing the R-squared values for the test dataset.

#### **2.5.4.3 Programming Paradigms**

Imperative programming paradigm was mostly used and preferred during the machine learning process because it is a sequence of directions for performing actions. (Vujosevic Janicic *et al.*, 2008). The codes were executed in a sequence.

# **REFERENCES**

1. ‘11-23-22\_Mortality-trend-prediction-using-ML’ (2022).
2. *22 - Injury, poisoning or certain other consequences of external causes - ICD-11 MMS* (2023). Available at: https://www.findacode.com/icd-11/chapter-22.html (Accessed: 4 November 2023).
3. Aguinis, H., Gottfredson, R.K. and Joo, H. (2013) ‘Best-Practice Recommendations for Defining, Identifying, and Handling Outliers’, *Organizational Research Methods*. SAGE Publications Inc., pp. 270–301. Available at: https://doi.org/10.1177/1094428112470848.
4. *Chapter 7 Regression I: K-nearest neighbors | Data Science* (2023). Available at: https://datasciencebook.ca/regression1.html#multivariable-knn-regression (Accessed: 11 November 2023).
5. Chicco, D., Warrens, M.J. and Jurman, G. (2021) ‘The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation’, *PeerJ Computer Science*, 7, pp. 1–24. Available at: https://doi.org/10.7717/PEERJ-CS.623.
6. *Circulatory System Diseases: Risk Factors & Symptoms* (no date). Available at: https://my.clevelandclinic.org/health/articles/23489-circulatory-system-diseases (Accessed: 4 November 2023).
7. *Coefficient of Variation in Statistics - Statistics By Jim* (2020). Available at: https://statisticsbyjim.com/basics/coefficient-variation/ (Accessed: 8 November 2023).
8. Cousineau, D. and Chartier, S. (2010) ‘Outliers detection and treatment: a review.’, *International Journal of Psychological Research*, 3(1), pp. 58–67. Available at: https://doi.org/10.21500/20112084.844.
9. *cso ireland regions - Google Search* (no date). Available at: https://www.google.com/search?q=cso+ireland+regions&oq=&gs\_lcrp=EgZjaHJvbWUqCQgAECMYJxjqAjIJCAAQIxgnGOoCMgkIARAjGCcY6gIyCQgCECMYJxjqAjIJCAMQIxgnGOoCMgkIBBAjGCcY6gIyCQgFECMYJxjqAjIJCAYQIxgnGOoCMgkIBxAjGCcY6gLSAQkxMDUwajBqMTWoAgiwAgE&sourceid=chrome&ie=UTF-8 (Accessed: 7 November 2023).
10. Dahouda, M.K. and Joe, I. (2021) ‘A Deep-Learned Embedding Technique for Categorical Features Encoding’, *IEEE Access*, 9, pp. 114381–114391. Available at: https://doi.org/10.1109/ACCESS.2021.3104357.
11. Deprez, P., Shevchenko, P. V. and Wüthrich, M. V. (2017) ‘Machine learning techniques for mortality modeling’, *European Actuarial Journal*, 7(2), pp. 337–352. Available at: https://doi.org/10.1007/S13385-017-0152-4/FIGURES/8.
12. Eliane Birba, D. (2020) *A Comparative study of data splitting algorithms for machine learning model selection*, *DEGREE PROJECT IN COMPUTER SCIENCE AND ENGINEERING*.
13. *head () and tail () Functions Explained with Examples and Codes* (2023). Available at: https://www.analyticsvidhya.com/blog/2023/07/head-and-tail-functions/ (Accessed: 8 November 2023).
14. *ICD-10 - Wikipedia* (no date). Available at: https://en.wikipedia.org/wiki/ICD-10 (Accessed: 4 November 2023).
15. *ICD-10 Version:2019* (no date). Available at: https://icd.who.int/browse10/2019/en#/II (Accessed: 4 November 2023).
16. Jakobsen, J.C. *et al.* (2017) ‘When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts’, *BMC Medical Research Methodology*, 17(1). Available at: https://doi.org/10.1186/s12874-017-0442-1.
17. Kang, H. (2013) ‘The prevention and handling of the missing data’, *Korean Journal of Anesthesiology*, pp. 402–406. Available at: https://doi.org/10.4097/kjae.2013.64.5.402.
18. Kim, H.-Y. (2013) ‘Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis’, *Restorative Dentistry & Endodontics*, 38(1), p. 52. Available at: https://doi.org/10.5395/rde.2013.38.1.52.
19. KOCARIK GACAR, B. and DEVECİ KOCAKOÇ, İ. (2020) ‘Regresyon Analizleri mi Karar Ağaçları mı?’, *Celal Bayar Üniversitesi Sosyal Bilimler Dergisi*, pp. 251–260. Available at: https://doi.org/10.18026/cbayarsos.796172.
20. Krittanawong, C. *et al.* (2021) ‘Machine learning and deep learning to predict mortality in patients with spontaneous coronary artery dissection’, *Scientific reports*, 11(1), p. 8992. Available at: https://doi.org/10.1038/s41598-021-88172-0.
21. Martins, S., Pesado, P. and García-Martínez, R. (2016) ‘Information mining projects management process’, in *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*. Knowledge Systems Institute Graduate School, pp. 504–509. Available at: https://doi.org/10.18293/SEKE2016-009.
22. Mishra, P. *et al.* (2019) ‘Descriptive statistics and normality tests for statistical data’, *Annals of Cardiac Anaesthesia*, 22(1), pp. 67–72. Available at: https://doi.org/10.4103/aca.ACA\_157\_18.
23. Nasteski, V. (2017) ‘An overview of the supervised machine learning methods’, *HORIZONS.B*, 4, pp. 51–62. Available at: https://doi.org/10.20544/horizons.b.04.1.17.p05.
24. *New research reveals where and how people die in Ireland - News & Events | Trinity College Dublin* (2021). Available at: https://www.tcd.ie/news\_events/articles/new-research-reveals-where-and-how-people-die-in-ireland/ (Accessed: 4 November 2023).
25. *OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059* (2020). Available at: https://studylib.net/doc/25722275/oreilly.media.machine.learning.and.data.science.blueprint... (Accessed: 7 November 2023).
26. *Population Changes - CSO - Central Statistics Office* (no date). Available at: https://www.cso.ie/en/releasesandpublications/ep/p-cpsr/censusofpopulation2022-summaryresults/populationchanges/ (Accessed: 9 October 2023).
27. Purbasari, A. *et al.* (2021) ‘CRISP-DM for Data Quality Improvement to Support Machine Learning of Stunting Prediction in Infants and Toddlers’, *Proceedings - 2021 8th International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2021* [Preprint]. Available at: https://doi.org/10.1109/ICAICTA53211.2021.9640294.
28. *Respiratory disease | Definition, Causes, & Major Types | Britannica* (2023). Available at: https://www.britannica.com/science/respiratory-disease (Accessed: 4 November 2023).
29. Saroj, R.K. *et al.* (2022) ‘Machine Learning Algorithms for understanding the determinants of under-five Mortality’, *BioData Mining*, 15(1). Available at: https://doi.org/10.1186/s13040-022-00308-8.
30. *Secondary Analysis of Electronic Health Records* (2016) *Secondary Analysis of Electronic Health Records*. Springer International Publishing. Available at: https://doi.org/10.1007/978-3-319-43742-2.
31. Segal, M.R. (2003) *UCSF Recent Work Title Machine Learning Benchmarks and Random Forest Regression Publication Date Machine Learning Benchmarks and Random Forest Regression*.
32. Shetty, S.H. *et al.* (2022) ‘Supervised Machine Learning: Algorithms and Applications’, *Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools, and Applications*, pp. 1–16. Available at: https://doi.org/10.1002/9781119821908.CH1.
33. Venkatesh, K.A., Mishra, D. and Manimozhi, T. (2023) ‘Model selection and regularization’, *Statistical Modeling in Machine Learning: Concepts and Applications*, pp. 159–178. Available at: https://doi.org/10.1016/B978-0-323-91776-6.24001-3.
34. Vujosevic Janicic, M. *et al.* (2008) *The role of programming paradigms in the first programming courses*. Available at: https://www.researchgate.net/publication/265630430.