# Group ID - MSc in Data Analytics

Author: Jose Gabriel Pina Zambrano

e-mail: [2023206@student.cct.ie](mailto:2023206@student.cct.ie)

Student ID: 2023206

**Introduction**

First, I clarify that the selected database corresponds to the geographic distribution (specifying the surveyed locality), by age group (grouping in 5-year increments up to 85 years and considering all those above 85 years as a single group), and health status in Ireland for the year 2022.

I found it interesting to choose, for my study, a dataset that didn't focus solely on demographic or nationality information of the population. I thought that the topic of health could be important, as it is one of the subjects that, socially, generates divided opinions when making decisions, whether in investment or policy application. Although I know it's very difficult to reach a firm conclusion on anything with this research, I found it interesting to focus on this aspect of society in order to have a clearer opinion.

I would like to highlight that in my analysis, I made some modifications to the data base to make it more functional. For example, through the Exploratory Data Analysis process, I noticed that the database was incomplete because part of the population was not included in the census. Another case worth noting is that for towns with fewer than 1500 inhabitants, we do not have geographic location data by county. Therefore, I excluded them from my analysis, as my goal is to conduct a study that helps determine the most effective place (county) to allocate investments in public health, for instance.

**Methods and Results**

Regarding the handling of my database in Jupyter, the first step involved loading the libraries that I deemed necessary for data management, statistical treatment, and subsequent graphical visualization of information. Throughout the entire process, I added libraries as needed:

* pandas as pd: This allows reading, manipulating, and analyzing data present in the table (.CSV).
* seaborn as sns: This enables graphical representation of data.
* scipy import stats: This facilitates the necessary statistical analysis, for example, the Z-TEST. In my case, I used it to determine if the difference between men and women in the study could represent a significant variation to be considered.
* numpy as np: This helps treat data as an array if necessary, depending on the analysis I want to perform.
* matplotlib.pyplot as plt: This library allows me to graph the data obtained after modeling the information.
* sklearn.naive\_bayes import GaussianNB: Scikit-learn is a machine learning library that provides tools for various purposes. In this case, this module allows me to apply a machine learning method to predict data in the model I am generating.
* sklearn.model\_selection import train\_test\_split: This module allows the system to split information into subsets in machine learning algorithms, performing random tests and learning to make more accurate predictions over time.
* sklearn import metrics: This allows me to evaluate the accuracy of machine learning systems and other metrics if needed for evaluation.
* sklearn.tree import DecisionTreeClassifier: This module helps me apply the decision tree to my model as a second machine learning method.
* sklearn.tree import plot\_tree: It is the necessary module for plotting the decision tree that was generated.

After loading the libraries, my next step was to load the database and print it to understand its shape and main characteristics to determine the next steps.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Illustration 1Code to read and verify the shape of de df

After loading the database, I observed that it contained the total population for each city, each age group, and each health status. Additionally, there was an overall total for each of the fields. For that reason, my next step was to filter this information to initially understand the overall picture and work with a general value. This approach allows me to move from general to specific before reaching any conclusions.

Imagen que contiene Texto

Descripción generada automáticamente

Illustration 2. Filtering the information

This step was also carried out to avoid counting the population duplicates. When analyzing the information without any filtering, I noticed that the count was the sum of values for each city, including the already totaled values. Filtering allowed me to count how many times each county was repeated in the list. On the other hand, I summed the population for each county to determine the overall population without differentiating age groups or gender. Additionally, I observed that in the database, the count of counties in Ireland was 42 instead of the correct 26. This discrepancy was due to some cases, specifically 16 instances, where counties sharing a border were considered as a single additional unit. This occurred in cases of cities shared between two counties.

Once the general data was filtered and known, I applied the reverse method by filtering the same database to exclude these values and obtain only specific ones. These specific values include the differentiation of gender, age group, and location in each county of the population, as well as the health status of the population. This helped me determine which counties have the highest population density, which age groups are the most representative, and the number of men and women in the sample.

Using these values, I began to segment the information. Initially, I divided the distribution of men and women and represented it in a PIE chart. This choice was made because it is the best option for graphically displaying this distribution with only two variables. The chart provides a clear representation, minimizing the chance of errors in interpreting the values.

Gráfico

Descripción generada automáticamente con confianza media

Illustration 3. Filtering by SEX

Displaying values of 51.34% (1,547,145) against 48.66% (1,466,607), with such a small difference, I proceeded to apply one of the methods studied in the field of statistics: the Z-TEST. The same reasons for applying this test are why I ruled out using a T-TEST, and I'll explain them below:

Diagrama

Descripción generada automáticamente

Illustration 4. Decision tree to apply z-test

Based on the decision tree presented in class, one branch recommending the use of the T-TEST is already ruled out upon knowing the studied population. Furthermore, with a representative sample size exceeding 30 records, the T-TEST is also dismissed in favor of the Z-TEST. The reason for choosing this type of analysis is to explore whether the difference between males and females could be a factor in the variations observed in the comparison of age groups or the overall health status corresponding to each group.

Following the study, I set a null hypothesis, which, in my case, would be "The difference in gender is not significant for the study." Considering the standard value of ALPHA typically used in such tests is 0.05, the Z value must be greater than 1.96 or less than -1.96 to reject this hypothesis. Upon conducting the test, the results were as follows:

Texto

Descripción generada automáticamente

Illustration 5. z-test results

Therefore, -1.96 < -0.54 < 1.96, which means that the null hypothesis is accepted. Thus, I conclude that the difference in the number of men and women studied is not sufficiently significant to be considered a determining factor in the subsequent phases of this study. Consequently, I do not take it into account as a factor contributing, for example, to the deterioration of health in the studied population.

In my next step, I created a table to classify the data, going as specific and segmented as possible. Having the total information, I classified it in the following order: first, by the health status of the sample; then, by gender; subsequently, by age group; and finally, by county. Based on the obtained information, I created bar charts using the County variable to visualize the population in each county organized by age group. This allowed me to determine which age groups the population is concentrated in for each county, helping me choose the most representative sample in the subsequent steps, in case I need to focus the study on a specific variable related to age. Below is an example of the charts created:

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

From this point onward, in the subsequent steps, I proceeded to relate the variables to segment and clarify the information into different groups. For example, I segmented the data into different age groups to graph the distribution or fixed a risk level by relating age to health status. Because some of my variables are nominal and others ordinal, the logical step was to create indices for my variables to relate and graph them without issues. Initially, I encountered problems when trying to relate different strings.

The first variables I encoded were the age groups and health statuses in my database (this encoding will be available in an additional Excel file that I will add as a dictionary for my database).

Next, I used a bar chart to represent the number of inhabitants in each age group. Using a stacked bar chart, I segmented them by health status. This provided an overall view of the distribution of these groups. I also obtained the total number of people in the database for each age range, along with the percentage they represent of the total population studied. With this, I determined that 55.87% of the population is between 25 and 59 years old. Using the previously created charts, I could observe that this age group enjoys better health. Consequently, I regrouped the age groups to study this population as a unit:

Texto

Descripción generada automáticamente

Illustration 6. Coding of age groups

In addition to obtaining the number of individuals in each group, I also graphed the percentage they represent in the total population:

Gráfico, Gráfico circular

Descripción generada automáticamente

In my next step, I used the previously encoded information to graph the obtained data. I related the number of inhabitants in each age group, classified them by health status, and represented them in a bar chart. By visualizing the distribution in the categories, I observed that as the population's age increases, the distribution in health statuses becomes more dispersed. Thus, I can determine that age is an influential factor in health deterioration, which will be taken into account in the subsequent steps of the study.

After confirming through the charts that the group of individuals over 60 years old tends to have more varied health on average, I considered it important to know the geographical distribution of this group. I did this by determining the number of people over 60 in each county, creating the variable "seniors\_by\_county" and counting the corresponding values in the sample.

Now, considering that, in my opinion, the older the age, the worse the health, the higher the care needs, and the greater the risk of death, for this reason I found it appropriate to assign a risk level to each health status to determine where the population falls on these "risk of death" levels. To establish this relationship, I defined a function called "assign\_risk," which was applied to all observations in the database based in the health status regardless of age. Texto

Descripción generada automáticamente

My next step was to graph this risk to understand the distribution of the population based on this characteristic. With this, we can observe that the surveyed population in Ireland generally enjoys good health. The difference between the values for each risk level is quite broad, favoring the "Very good" health level.

Gráfico

Descripción generada automáticamente

Illustration 7. Population by General Health

Although in a previous step, I determined that gender is not a determining factor in health deterioration, I found it interesting to obtain data on how many males and females were in each health status group. To establish this relationship, similar to what I did earlier, I encoded the gender for each individual and related it to the encoded health status. Subsequently, I graphed the values.

After this, I started applying machine learning algorithms. For this purpose, I modified my data frame again, encoding the risk levels to have all the variables in the same format. I named this new data frame "df\_encoded."

Tabla

Descripción generada automáticamente

Once I had this table, I applied a decision tree to determine in which risk group the observations should be grouped based on "Encoded Age Group" and "Encoded General Health." Since I have three well-defined variables, I set that number as the levels of my tree. Additionally, I used 50% of the observations in my database as a sample. It is worth noting that when applying the tree, the accuracy is 100%, which indicates the possibility of overfitting. This means the tree may be too sensitive to errors or noise in cases where values are different or not well-encapsulated within my scales. Although the conclusion I reached is that the variables referenced in the tree are quite fixed and do not allow for interpretation. In other words, all possibilities and combinations between health and age always correspond to a single category.

Diagrama

Descripción generada automáticamente

Illustration 8. Decision tree

The explanation of the tree would be as follows:

For a sample of 50% of the observations (19,620) in my database, the system sets a threshold for decision-making. At the first level, this threshold is 3.5 since, based on health states, it is the midpoint between the two risk categories (High Risk and Medium Risk). We also have the corresponding values for each present value in the sample. Another value present in the box is the Gini index. Simply put, this index indicates the level of homogeneity or impurity of the node; the lower this value, the fewer categories the node is divided into.

In the second node of the tree, we can observe that the division was made, with values greater than or equal to 3.5 on one side (representing good health and very good health), with a Gini of 0.0, indicating that all values in that node belong to a single category, and the node cannot be divided again. On the other side, values that met the condition and were less than or equal to 3.5; in this second node, the Gini is 0.44, indicating that the node could be divided to generate new, more homogeneous nodes. So, setting a new threshold, in this case, <=1.5, which would again be the midpoint for changing the category between Medium Risk and Low Risk, the node is divided again, generating two new classifications with a Gini value of zero, indicating that the nodes are completely homogeneous and cannot be divided again. This shows a clear classification of the values present in the sample.

For my second machine learning method, I decided to use Naïve Bayes. Previously, I determined that age and health variables are related, as the sample becomes more dispersed in health states with increasing age. Although they are not directly related since it is not a constant, I consider them independent for this case to help with the classification into risk levels. For this method, I again took 50% of the dataset to apply the algorithm, assigning a risk level by relating the values of health level and age group. In the case of this method, the system's effectiveness is 100%, serving two purposes: first, to verify the purity of my database, and second, to support the results obtained in my decision tree.

Based on the results of my information analysis, considering the obtained graphs and the results of the machine learning algorithms, I dare to determine that the information in the database could be used, for example, to determine in which counties it would be worth investing in the healthcare sector, based on population risk levels and age ranges. Considering the life expectancy in Ireland, a plan could also be started to consider promoting a new demographic distribution by encouraging the movement of the population to areas with a higher average age.

To continue with the requirements as a method of discrete distribution, I applied the geometric distribution to calculate the probability of finding a person with very bad health in a specific county, in this step, in the most densely populated areas (Counties Dublin & Meath). Obtaining the following result:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Illustration 9. Geometric distribution results

This analysis helps me determine the possibility of finding a person with very bad health in the specified county. Theoretically, it measures the likelihood of success in a trial with constant probability variables. Understanding this value is important because it helps us grasp the magnitude of the number of people with bad health in the county. In this case, even though it is the most populated, it contributes to the low representation of people with bad health in the overall percentage.

For the project management framework, I utilized KDD (Knowledge Discovery in Databases) since it allowed me to process raw databases to extract elements essential for decision-making, especially regarding resource distribution in healthcare. The data was processed to convert it into useful information for decision-making in the specific area. In the context of my research, I followed the stages of this theoretical framework, which I will enumerate below, correlating them with my project to highlight the alignment with my research approach:

* The first step was selecting the data, meaning I made the decision on the database to use, explaining why I chose it, and determining the relevant data to work with. This involved deciding how to segment the data.
* My next step was processing the raw data to extract maximum value. This included actions such as summing or eliminating columns, such as encoded data or county names, which later helped with segmentation in the graphs. I also removed values for people whose census had not yet started and observations with totals to avoid data duplication during aggregation.
* The third step was data transformation, mainly focusing on standardizing the data. I addressed discrepancies in format, allowing me to relate variables. Many of my relationships involved comparing a string against a numerical value representing the total for a population group. I achieved this by creating an index and assigning a value to each group.
* In a subsequent stage, I began applying data mining by establishing relationships, using machine learning algorithms, and conducting statistical tests like the Z-test. This aimed to determine significant relationships, identify relevant values, or uncover patterns or trends that might reveal important knowledge within the database.
* As the final stage after each experiment, I obtained a graphical representation. This helped make the project more understandable when generating conclusions or justifying the execution of each test. Ultimately, it contributed to presenting the information in a less dense manner, making it easy to explain and user-friendly for the final audience if any form of presentation was required.

Regarding the learning algorithms I employed (Decision Tree and Naïve Bayes), I opted for supervised learning due to its ability to learn based on the relationship between data labels for inputs and outputs. In theory, based on my understanding of these methods, this would facilitate predicting data or assigning labels to new output data, provided I have the correct input labels for predicting information in the trained system.

Another key feature of these algorithms is their capability to classify information by assigning groups, such as health statuses or age ranges in my case. This classification task can be automated without the constant need for my intervention in the algorithm. A significant advantage, which I observed while applying them, is their simplicity and interpretability, making them very user-friendly for the final users of the projects.

Next, I will present a comparison between the algorithms I used to demonstrate the differences between them.

|  |  |  |
| --- | --- | --- |
| Caracteristicas  Algorithm | Decisions Tree | Naïve Bayes |
| Interpretation: | They are easily interpretable and provide clear decision rules. | It is also relatively easy to interpret |
| Training and Prediction: | May require more training time, but prediction is fast. | Fast in both training and prediction. |
| Scalability: | Can become complex on large datasets. | Generally scalable and efficient on large datasets. |
| Robustness to Noisy Data: | Can be sensitive to noise and outliers. | Typically, robust to noisy data. |
| Handling of Independent Features: | Can model complex relationships between features | Assumes conditional independence between features, which may not be realistic in some cases. |

Illustration 10. Decision tree VS Naive Bayes comparison

The main reason I selected these algorithms is for their simplicity in displaying results and ease of application. On the other hand, it's worth noting that, as I mentioned before, being supervised, they don't require constant manipulation or intervention on my part. Another noteworthy point, and in my opinion, one of the most important, is that they are fast and effective. Perhaps one of the peculiarities that made them suitable for me is that, in my dataset, most of the data is not subject to interpretation. In my understanding, being all "black or white" allows them to be 100% accurate, avoiding the uncertainties or gray areas that can be found in other types of databases.

Regarding the comparison between the results of the algorithms I used, I emphasize that in both cases, the effectiveness I achieved was 100%. As I interpret the data, as I mentioned before, without gray areas, I understand that this contributes to such high effectiveness because it's practically an assignment rather than a prediction. This proves very convenient when filling in the data. However, I also understand that, on the other hand, it makes my machine learning systems sensitive to variations and noises that may exist due to, for example, human intervention when completing the database because of any discrepancies or inconsistencies in the information.

**Conclusions**

from the data processing steps that brought me to this point in the research are as follows:

1. Data encoding stands out as one of the most useful tools when it comes to relating diverse types of data.
2. Choosing the appropriate data visualization is highly beneficial, not only for presenting the data but also for advancing research through accurate interpretation.
3. It is very importance to consider, as the research project progresses, which features are important and which ones I can stop studying because they are not relevant. However, I can only know this by having a clear focus and direction that I want to give to my research project, so it is very helpful to have a clear objective from the beginning.
4. While it is true that I am more accustomed to working with databases in programs like Excel or Power BI, this method of data processing proves to be more effective in terms of time and resources. Even though it may not be a routine practice for me as of today.

# Bibliografía

Severance, C. (2009). *Python for Everybody.*

*https://data.cso.ie/*. (n.d.). Retrieved from https://data.cso.ie/

*https://matplotlib.org/*. (n.d.). Retrieved from https://matplotlib.org/

Subhash, M. (2023, May 17). *https://www.analyticsvidhya.com/*. Retrieved from https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/

Waskom, M. (n.d.). *https://seaborn.pydata.org/*. Retrieved from https://seaborn.pydata.org/