**CA1 Project**

(\*Word count: 1327)

*\*Excluding diagrams, code, references and titles.*

By

Leopoldo Rojo Romero (Student ID: 2023355)

Machine Learning for Business

Dr. Muhammad Iqbal

CCT College

Dublin, Ireland

**Table of contents**

[**Introduction** 2](#_Toc164890120)

[**Data sources** 2](#_Toc164890121)

[**Ethical considerations** 3](#_Toc164890122)

[**Clustering** 3](#_Toc164890123)

[**Time Series** 7](#_Toc164890124)

[**Github** 11](#_Toc164890125)

[**References** 11](#_Toc164890126)

# **Introduction**

In the present analysis we will approach a clustering scenario and a time series scenario.

**For our clustering analysis**, we will take a look to a dataset that has obesity levels in individuals from the countries of México, Perú and Colombia, in this case, we will take into account the Height and Weight of each individual to attempt to identify how the obesity levels are grouped.

This topic that we have chosen has a huge relevance due to the fact that in the last 50 years, overweight and obesity rates have triplicated, affecting 62.5% of the population in Latin America (United Nations, 2023).

Latin America and the Caribbean has the highest healthy diet costs worldwide, according to Panorama 2023, the average cost of a healthy diet globally is US$3.66 per person per day, whereas in Latin America and the Caribbean is US$4.08 per person per day, in comparison with Europe where the cost is US$3.22 per person per day (United Nations, 2023).

**For our time series analysis**, we will be taking a look to a stock dataset, to the closing stock prices, in this case from Lockheed Martin Corporation (LMT), a security and aerospace company, engages in the research design, development, manufacture, integration, and sustainment of technology systems, products and services worldwide.

Having knowledge in regards to stocks of solid companies like LMT is becoming key to everyone’s financial future since we had an inflation crisis in 2021/22 where almost every economy experienced inflation rates far higher than two percent (Statista, n.d.), by investing in these type of companies we can tackle inflation, this is because investing in the stock market have the following potential benefits:

* Potential capital gains from owning a stock that grows in value over time
* Potential income from dividends paid by the company
* Lower tax rates on long term capital gains

(The Basics of Investing In Stocks, n.d.).

# **Data sources**

We will use publicly available datasets, these datasets can be found in the following links:

Clustering

<https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition>

Time series

<https://finance.yahoo.com/quote/LMT/history>

# **Ethical considerations**

We do not seek to make any health / medical conclusions based on our clustering analysis and we do not seek to make any financial recommendations based on our time series analysis.

The following analysis have the only educational purposes.

# **Clustering**

Clustering techniques can be used for many purposes including customer segmentation, image classification, social network analysis and anomaly detection (Yazar, 2023).

Our analysis could be considered “customer segmentation”, since we are attempting to identify certain groups of individuals based on obesity levels.

Let’s have a first look at our data, stored in the “clu” variable:

Pantalla de computadora

Descripción generada automáticamente con confianza media

Figure 1: Clustering dataset

Relevant dataset information:

Height: Will be our Y axis value.

Weight: Will be our X axis value.

NObeyesdad: Obesity Level, we will use these levels to find our clusters.

**Obesity levels:** Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

77% of the data was generated synthetically using the Weka tool and the SMOTE filter, 23% of the data was collected directly from users through a web platform.

This dataset has 2111 observations and 17 features (including the target feature), before going further, we need to search for any null values in the data because it could cause errors and noise.

Texto

Descripción generada automáticamente

Figure 2: Checking for null values in 'clu' dataset

As we can see from the above “isnull() method, our data does not have any null values.

Let’s visualise our data to get a better understanding of how the obesity levels are distributed along X (Weight) and Y (Height).

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 3: Visualising 'clu' dataset

From the above plot, we can observe:

* Clearly defined obesity levels groups
* The more height, the more weight needed to be in the following obesity level
* The highest obesity level is predominated by females

Following the visualisation of our data, we can select our clustering algorithms, to be able to compare 2 methods with the data that we have, we will use one algorithm that is based on centroids and one that is based on density.

**K-Means:** Works by assigning data points to clusters based on the shortest distance to the centroids or centre of the cluster, its main goal is to minimise the sum of distances between data points and their respective clusters (REGUNATH, 2022).

**DBSCAN:** It determines clusters based on how dense regions are, it can find irregular shaped clusters and outliers quite well (REGUNATH, 2022).

To measure how well these methods do within our data, we will use the Silhouette score, this score goes from -1 to 1, being 1 a clearly defined and separate clusters, 0 means clusters are indifferent or that the distance between clusters is not significant, -1 means clusters are assigned in the wrong way (Bhardwaj, 2020).

Since we already know that we have 7 obesity lebels (clusters), we can put our Height and Weight features directly into the K-Means algorithm, these are the results:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 4: K-Means clusters

Silhouette score: 0.5669

For the DBSCAN algorithm, we will search for the maximum distance between two samples for one to be considered as in the neighbourhood, then, we will create a function to iterate different parameters to find the best ones, measuring them with the Silhouette score.

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

Figure 5: Searching for our epsilon value

As shown above, we can notice that the epsilon for our data would be between 0.1 and 0.5.

Texto

Descripción generada automáticamente

Figure 6: Searching for the best Silhouette score for DBSCAN

From the above function, we can see that the highest Silhouette score (0.4840) is with the eps:0.2 with min samples:4, now with can pass these parameters through the DBSCAN algorithm.

Comparing the K-Means and DBSCAN Silhouette scores, we can expect that DBSCAN will not be doing as well as the K-Means algorithm.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure 7: DBSCAN clusters

As we can observe, based on the Silhouette score, we can say that our assumption that K-Means works better for our data is correct.

# **Time Series**

Is a way of studying the characteristics of the response variable concerning time as the independent variable. To estimate the target variable in forecasting, we use the time variable as the reference point (Pandian, 2021).

Let’s have a first look at our data, stored in the “time\_s” variable:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figure 8: Time series dataset

This dataset has 473 observations and 7 features (including the target feature), before going further, we need to search for any null values in the data because it could cause errors and noise.

Una captura de pantalla de un celular

Descripción generada automáticamente con confianza media

Figure 9: Checking for null values in time\_s dataset

Since we do not have any null values in our dataset, we can proceed by setting our ‘Date’ feature as index.

Captura de pantalla de un videojuego

Descripción generada automáticamente

Figure 10: Setting our index

Now we need to visualise the Autocorrelation and Partial Autocorrelation of our data.

Autocorrelation (ACF): Calculation of the correlation of the time series observations with values of the same series, but at previous times.

Partial Autocorrelation (PACF): Summarizes the relationship between an observation in a time series with observations at previous time steps, but the relationships of intervening observations are removed.

(Marv, 2021)

Gráfico

Descripción generada automáticamente

Figure 11: ACF and PACF plots

From these plots, we are looking at the points outside the blue shadow, that means that they have statistical significance, that indicates that an observed relationship is unlikely to be due to chance.

Since our Autocorrelation plot have more than 20 lags with positive correlation above 0.5, we can expect that we will need a differencing above 1.

Let’s see if our data is stationary or not by performing a Dickey Fuller test.

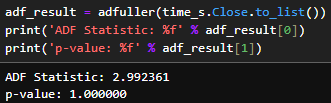


Figure 12: Checking if our data is stationary or not

Since we got a p-value of 1, we can conclude that our data is not stationary, therefore, we will use the differencing method to make it stationary, as we have mentioned before with our autocorrelation plot.

In this case, we will use the ARIMA algorithm, this algorithm is used to understand past data or predict future data in a series.

To search for the best values for our ARIMA algorithm, we will create a function to iterate multiple values through the algorithm, we are looking for the lowest Akaike’s Information Criterion (AIC).

Texto

Descripción generada automáticamente

Figure 13: Function to search for the best AIC values for ARIMA

After getting our best AIC value, we can make a forecast with the ARIMA algorithm, let’s visualise the results.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 14: ARIMA forecast

As we can observe, the ARIMA algorithm works efficiently with our dataset and we can see that by looking at the confidence intervals.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Figure 15: Confidence intervals

In the case of the 95% confidence interval, suggests that there is a high likelihood that the real observation will be within the range of 432.637 and 474.824.

# **Github**

<https://github.com/LeopoldoCCT/ML_ForBusiness>

# **References**

Bhardwaj, A. (2020). *Silhouette Coefficient : Validating clustering techniques*. [online] Medium. Available at: <https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c>.

Marv (2021). *What’s The Difference Between Autocorrelation & Partial Autocorrelation For Time Series Analysis?* [online] Medium. Available at: <https://emel333.medium.com/interpreting-autocorrelation-partial-autocorrelation-plots-for-time-series-analysis-23f87b102c64>.

Pandian, S. (2021). *A Comprehensive guide to Time Series Analysis*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/>.

REGUNATH, G. (2022). *10 Incredibly Useful Clustering Algorithms*. [online] Advancing Analytics. Available at: <https://www.advancinganalytics.co.uk/blog/2022/6/13/10-incredibly-useful-clustering-algorithms-you-need-to-know>.

Statista. (n.d.). *The Inflation Crisis*. [online] Available at: <https://www.statista.com/study/116987/the-inflation-crisis/>.

The Basics of Investing In Stocks. (n.d.). *The Basics of Investing In Stocks*. [online] Available at: <https://dfi.wa.gov/financial-education/information/basics-investing-stocks#:~:text=The%20potential%20benefits%20of%20investing>.

United Nations (2023). *New UN report: 43.2 million people suffer from hunger in Latin America and the Caribbean, and the region has higher levels of overweight and obesity than the global estimate - PAHO/WHO | Pan American Health Organization*. [online] www.paho.org. Available at: <https://www.paho.org/en/news/9-11-2023-new-report-432-million-people-suffer-hunger-latin-america-and-caribbean-and-region#:~:text=%22In%20the%20last%2050%20years>.

Yazar, O. (2023). *Complete guide to perform clustering analysis on python*. [online] Medium. Available at: <https://medium.com/@yazar.or/complete-guide-to-perform-clustering-analysis-on-python-8597167eb4a4> [Accessed 24 Apr. 2024].