A logo for college computing

Description automatically generated

|  |  |
| --- | --- |
| Student Full Name: Jeisson Steven Preciado Wilches |  |
| Student Number: 2024326 |  |
| Module Title: Strategic Thinking |  |
| Assessment Title: CA 1 |  |
| Assessment Due Date: 18/05/25 |  |
| Date of Submission: 18/05/25 |  |

**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

Table of Contents

[Introduction 3](#_Toc197121379)

# Introduction

This project leverages machine learning and financial analysis from an Irish company, after looking a database to work it, I have selected their stock prices where I applied models such as Lineal Regression, K-Means Clustering, KNN, etc. The company that I have analysed is called Medtronic, a global enterprise in medical technology, services, and solutions, dedicated to improving healthcare for customers around the world. Medtronic has pioneered innovations in areas like cardiovascular therapies, diabetes, neurological problems and minimally invasive procedures. Essentially, the mission of the company is responsible for resolves pain, restoring health focusing on welfare by using biomedical engineering.

In this analysis, we are going to examine Medtronic’s stock performance using data science techniques as I mentioned before. The dataset comprises historical stock price for the last 5 years, capturing key features such as opening and closing prices, highs and lows, and so on. Using these methods, my aim is to find patterns that help me to have a better understanding of the stock prices, create forecasting and through different techniques get multiple angles to take a decision that benefit the company and stakeholders.

Finaly, examining this company stocks will help us not only to predict market behaviour but also to have a contribution of a deeper understanding of the company stocks during the giving period. This study will serve as a case study applying different machine learning models and time series analysis, also explaining answers and point of views, data visualization will be essential for a better comprehension of the dataset to present the findings.

# Objectives

Present the use of machine learning models to predict stocks prices based on historical trends, struggling with volatile markets, and identifies different market regimes. On the other hand, analyse time series to identifies trends & seasonality, handle with no seasonality trends, and the use of ARIMA due is a statistical model useful for forecasting of the variables from the data that contains, in this case, the Medtronic’ stock prices.

* General Objectives
  + Understand the dataset and the use of models on it.
  + The use of this techniques in real life.
  + Examine the stock prices of an Irish company.
  + Inspect the results obtained and give a deep explanation.
* Specific Objectives
  + Analyse the results through model performance metrics, robustness checks, and comparative analysis.
  + Clean and normalize the dataset and perform it through visualizations.
  + Determine optimal parameters to execute before evaluating the models.
  + Describe the results and their applicability in similar cases, particularly in relation to forecasting, smooth noisy data and identifying correlations.

# Problem Definition

Context: A través de una búsqueda y selección de una empresa irlandesa, se investiga la compañía para generar un contexto general de lo que hace la empresa, sus productos y/o servicios y misión. Luego, para ámbito de este proyecto, se consigue una base de datos la cual sean aptas para las aplicaciones de modelos de machine learning, que en esta ocasión será el precio de mercado de la empresa en los últimos cinco años.

Impact: Este proyecto da conocer el uso de las herramientas de machine learning y los análisis de los resultados que son importantes para las empresas interesadas en inversiones, ya que el impacto mediante el uso de herramientas de análisis de datos y de machine learning, les va a permitir tener un contexto mas amplio sobre las acciones y determinar patrones a lo largo de los años, anomalías, etc. Por otro lado, el impacto que esta base de datos juntos con los modelos que se aplicaron en el proyecto es que mejorara las habilidades en el uso de machine learning y análisis de tiempo para bases de datos especifica, no obstante, al tener conocimiento de los modelos y la aplicación en algunas bases de datos, estas habilidades adquiridas durante el desarrollo del trabajo nos permitirán tener un contexto mas amplio de los datos que manejamos y llegaremos a manejar.

Importance: La importancia de este proyecto es tener en cuenta las herramientas que utilizaremos y describiremos no solo como parte de este proyecto, sino que también, como un herramienta que podemos llegar a utilizar en las empresas, esto se debe a que hay compañías que trabajan con base de datos, en este caso de precio en el mercado, que necesitan saber que hacer antes de tomar una decisión, es por esto por lo que la aplicación de machine learning podría ser fundamental en instituciones financieras como fondo de inversiones, bancos, consultoras, y otro tipos de sectores que se puede beneficiar de los resultados del estudio de los datos.

# Business Description & Methods

In this initial step, the objective is to convert raw stock market data into a clean, structured format suitable for modelling. Despite, the original data did not have missing values, there were some observations in the data that was changed from excel file. On the other hand, I have created a simple description of the data to have a clear idea of the data’s observation and feature. A forecasting condition was applied where a time-based split was coded in Python before using a machine learning model. In this split, 80% of the data is allocated for training and 20% for testing.

In this part of the project, metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), along with cross-validation techniques like the Davies-Bouldin Index and Silhouette Score Method, were applied. Additionally, Time Series validation was utilized to assess performance across multiple time windows, ensuring stability against market regime changes. After analysing patterns, I monitored the prediction intervals in this project to quantify the uncertainty in the data set of Medtronic's stock prices. This is crucial as a goal if I am thinking to work in the financial sector as it allows for rebalancing based on predicted returns, optimizing client portfolios, and managing risk assessment scenarios based on predictions.

# Technologies & Accomplishment

En este trabajo se presenta una descripción detallada de una empresa irlandesa especializada en productos para la industria de la salud, así como su comportamiento en el mercado bursátil. Tras descargar la base de datos que contiene los precios de las acciones de los últimos cinco años, junto con la información sobre apertura, cierre, cierre ajustado, máximo, mínimo y volumen de las acciones, se procede a realizar un análisis mediante series temporales.

Este enfoque es fundamental para manejar bases de datos que contienen información que puede llegar a ser estacionaria, que contenga ruido, mirar el comportamiento de las tendencias a lo largo de las semanas buscando escenarios que se repitan, creando un patrón que permita saber cuándo es común vender o comprar. Además de obtener pronósticos mediante el uso del modelo ARIMA, esto es relevante en el ámbito profesional porque facilita la toma de decisiones informadas en el sector financiero, gestionando riesgos y optimizando estrategias.

Por otro lado, este proyecto tiene como objetivo demostrar la aplicación del machine learning, una herramienta sumamente eficiente para el manejo de grandes volúmenes de datos. Uno de los modelos de machine learning implementados en este proyecto fue la "regresión lineal". Se uso este modelo ya que nos ayuda a identificar las subidas y bajadas de las tendencias al largo del periodo, también nos intenta pronosticar el precio de las acciones basado en las tendencias que se muestran en los datos, y finalmente, si aplicamos estos modelos a otras bases de datos que contengan los precios de mercado de varias empresas, no ayudara a entender a la contribución de riesgo o rendimiento en el portafolio.

Otro modelo de machine learning utilizado fue el "K-Means Clustering", el cual facilitó la agrupación de las acciones de la empresa estudiada, simplificando el análisis del movimiento de los precios y la volatilidad. Este enfoque permitió identificar qué acciones presentaban comportamientos similares. Además, es un modelo eficaz para detectar anomalías o valores atípicos, los cuales son fácilmente identificables en las gráficas.

Por último, se empleó el modelo "K-nearest neighbors". Este modelo permite identificar patrones en los movimientos de precios del mercado, lo cual facilita la identificación de oportunidades de compra o venta. Asimismo, ayuda a realizar pronósticos basados en la información obtenida al detectar dichos patrones. Adicionalmente, el modelo al analizar los datos es capaz de detectar anomalías o valores atípicos. Estas desviaciones pueden deberse a factores externos, como decisiones en el comercio internacional que afecten directa o indirectamente a la empresa. Asimismo, puede haber eventos internos en la empresa que provoquen fluctuaciones en las acciones. De esta manera, el modelo permite identificar y entender el comportamiento de las acciones durante esos periodos específicos.

Como extra he aplique “simple exponential smoothing” en donde la idea es que este modelo pueda estudiar el comportamiento del precio de la acción de la empresa Medtronic. La razón para la aplicación de este modelo es que según la gráfica que nos muestra el movimiento del precio de las acciones a lo largo de los últimos 5 años, es una gráfica que nos muestra series temporales sin estacionalidad. La razón por la cual se aplico para esta base de datos es para poder pronosticas a corto plazo el movimiento del precio en el mercado de esta empresa, especialmente cuando no hay una tendencia clara. Otro aspecto muy importante es que este modelo nos ayuda a eliminar ruido que no le aporten a nuestro objetivo de realizar pronósticos, en otras palabras, suaviza los datos lo que hace que tenga en cuenta los datos mas recientes para identificar el comportamiento y luego poder hacer los pronósticos. Esto sin duda alguna, tiene una aplicación valiosa en el mundo profesional como lo puede ser en finanzas, retail, operaciones y en la economía.

# Challenges encountered

Durante la ejecución del trabajo, uno de los principales desafíos fue encontrar una base de datos que cumpliera con los requisitos necesarios para aplicar modelos de machine learning y análisis de tiempo. En internet existen numerosas bases de datos que presentan observaciones y características sobre diversos temas relacionados con finanzas, mercadeo, entre otros. Por esta razón, seleccioné una base de datos completamente numérica que fuera fácil de utilizar para la aplicación de modelos y que, al mismo tiempo, resultara útil no solo para este proyecto, sino también en el ámbito laboral.

Además, otro desafío considerable fue asegurarse de que la base de datos estuviera actualizada y cumpliera con los estándares necesarios para obtener resultados precisos y relevantes. Al analizar las diferentes opciones disponibles, se tomó en cuenta no solo la facilidad de uso sino también la calidad de los datos y su aplicabilidad en diferentes escenarios laborales. Al final, la base de datos elegida ofrecía un equilibrio entre datos recientes y la aplicación que le podía dar, permitiendo implementar diversos modelos y técnicas de machine learning eficientemente, lo cual resultó en análisis más profundos y útiles en contextos prácticos.

Por otro lado, también enfrente problemas con los modelos de machine learning al momento de ejecutarlos ya que algunos de los códigos fueron tomados directamente de otras clases tome como estudiante en el último año universitario. Estos códigos presentaban algunas condiciones específicas para una base de datos determinada, lo que hace que para mi base de datos en este proyecto no funcionara. No obstante, se cambiaron algunos códigos con el fin de que cumplieran con las condiciones de los códigos, para que finalmente pudiera dar respuesta sin ningún problema.

Estos problemas no fueron fáciles de resolver. Hubo momentos en los que me preocupaba que las soluciones que había implementado no fueran suficientes y que necesitaría comenzar desde cero. Primero, tuve que analizar detenidamente cada línea de código para entender sus funcionalidades y cómo estaban relacionadas con la base de datos original. Luego, realicé modificaciones en algunos códigos para cumplir con las condiciones necesarias, ajustándolos para que funcionaran correctamente con mi propia base de datos. Además, tuve que asegurarme de que todas las dependencias y bibliotecas estuvieran actualizadas y compatibles.

El proceso requería mucha prueba y error, así como tiempo dedicado a investigar soluciones en internet y ejemplos de bases de datos de otras clases de la universidad. Trabajé de manera diligente y exhausta para identificar los errores y ajustar los algoritmos según fuese necesario. Finalmente, después de varios intentos y revisiones, logré que los modelos de machine learning funcionaran sin inconvenientes. Obtener resultados precisos fue un gran alivio y perseverancia que puse en este proyecto.

# Results & Analysis

A continuación, se presentan los resultados de los modelos aplicados en el proyecto, utilizando Python y Jupyter Notebook. El proyecto tenía como objetivo analizar y predecir patrones a partir de un conjunto de datos extenso. Para lograr esto, se implementaron varios modelos de machine learning, incluyendo regresión lineal, árboles de decisión y redes neuronales. Cada modelo fue entrenado y evaluado usando diferentes métricas para asegurar su precisión y eficacia. Además, se realizaron visualizaciones de los datos y los resultados obtenidos mediante bibliotecas como Matplotlib y Seaborn, permitiendo una mejor comprensión de las tendencias y comportamientos observados en el conjunto de datos.

## Data Base

A continuación, se presenta un gráfico que muestra el comportamiento del precio de las acciones de Medtronic durante los últimos cinco años. Este gráfico proporciona una visión detallada de las fluctuaciones en el valor de las acciones, destacando periodos de crecimiento, estabilidad y declive. Además, permite a los inversores analizar las tendencias históricas y tomar decisiones informadas sobre futuras inversiones en la empresa.

A graph showing a line graph

AI-generated content may be incorrect.

Luego, observamos los “features” de esta base de datos obtenida de “Yahoo Finance”, estos datos son:

* Date: Las fechas en que se realizaron las operaciones en el mercado en Estados Unidos.
* Open: El valor de la acción con la que abrió.
* High: El valor máximo de la acción a la que llego en ese día.
* Low: El valor mínimo de la acción a la que llego en ese día.
* Close: El valor de la acción con la que cerro.
* Adj close: El valor cerrado ajustado de la acción, es cuando se descuenta el valor del cierre para el pago de dividendos a los accionistas, splits y/o distribuciones de capital.
* Volumen: El numero total de acciones negociadas en el día.

A screenshot of a computer code

AI-generated content may be incorrect.

## Time Series Analysis

Continuando con los resultados de los modelos aplicados a la base de datos. Yo he realizado la aplicación de análisis de serie en el tiempo, no sin antes de entrenar y testear la base de datos que cuenta con 6 “features” y 1257 “obsevation”. Aquí tuve en cuenta el uso del modelo “Baselines” que me permite medir el mínimo rendimiento, antes de aplicar “ARIMA model”, junto con el uso de MAPE (Mean Absolute Percentage Error). Luego, realice cuatro predicciones:

1. Historical
2. Last Year
3. Last Know Value
4. Naive Seasonal Forecast

A graph with blue rectangular bars

AI-generated content may be incorrect.

The result was that three of them were under 10%, which is good because that mean that the TSM in our database has less volatility or complex patterns, being “last\_year\_mean” the best baseline model with less than 6.5%. So, after getting this result the next step is starting to apply ARIMA model to reduce noisy in the data, get tendencies and identify some patterns, having a better understanding and precision than just the Baselines model.

Once I have applied ARIMA model, this was the first view.

A graph of a bar graph

AI-generated content may be incorrect.

Consequently, based on the “Autocorrelation” image, it is evident that the series is not stationary. Furthermore, this conclusion is corroborated by the P-Value and ADF Statistic, with values of 0.669774 and -1.208857, respectively.

In this part, now we must transform into a stationary station using (1,1,0), which is for trended data and is workable for data about stock prices, like this project.

A graph of different values

AI-generated content may be incorrect.

After applying ARIMA model (1,1,0), this is the result from the Medtronic stocks.

A comparison of a graph

AI-generated content may be incorrect.

In this case, according from the “Autocorrelation” image we see that now our series is stationary. I have calculated the p-value and ADF statistics again to confirm that is stationary and the result was 0.000 and -19.70 respectively. Next, I have coded the Akaike information criterion (AIC) on python to compare the quality of different ARIMA models in terms of best balance of fit and determine which one is the best, and the result from the smallest value was (1,1,1) 4421.30.

Below, I have compared the “actual” price market with the ARIMA (1,1,1) forecast.

A graph with red and blue lines

AI-generated content may be incorrect.

Afterwards, I have used the ARIMA model but with confidence intervals that are particularly valuable for risk assessment and decision-making. The 80% interval might be suitable for routine planning where some risk is acceptable, while the 99% interval would be appropriate for conservative scenarios where being wrong could have severe consequences.

A graph showing the growth of the stock market

AI-generated content may be incorrect.

## Linear Regression

Here, I have started to work with the applying of machine learning models to the data base that I have been working on. This is because, this model will allow me to predict trends in short periods of time, also, will show me the best baseline model using other complex models

After made some examples finding which would be the best option for being a dependent and independent variable, I have chosen as an independent variable (x) “Open” and as a dependent variable (y) “Close”. The reason why I have selected these two variables is because I want to focus on intraday prediction from close price from the open price from this data I am working on.

We can see that there is a positive correlation between these two variables. After training and testing these values, I have got the CV mean and the standard deviation, which values are 0.995 and 0.0009 respectively. CV mean has a high value which could mean overfitting and leakage because of my dataset, this is a financial data where the markets are in constantly changing, so that is the reason for the outcome’s values.

A graph with blue dots

AI-generated content may be incorrect.

Next, we must check cross validation, so train and test the data and I have specified that I am using linear regression in thar cross validation process. Also, I applied k-fold cv=10 to split my dataset into the process of cross validation to reduce overfitting risk. However, in this case, where my dataset is about stock prices, I must use another cross-validation process. In this case, time series split is the ideal method for this project. First, I have selected ridge regression as a model, after I have defined the hyperparameter through the regularization strength where higher values are more regularized. Next, I have combined both time series split and grid search to prevent data leakage and an optimal hyperparameter using alpha.

## K-Means Clustering

This is the next machines learning model that I have decided to develop for my project. This will help to identify same groups in stocks with similar behaviours. It is important if we are talking about risk assessment, trading strategies or portfolio diversification using the centroids to take a decision.

A graph with red and blue dots

AI-generated content may be incorrect.

After getting the result, we can appreciate the positive distribution and the centroids that will help us to have a better understanding about for which group is going each stock price. For example, in this visualization I can use centroids as an equilibrium point as risk categories, if the price is close to the centroid I can hold or ignore it, but it is far from the centroid I must start to sell or just be aware. Also, there will be some opportunities to sell when prices approach their cluster centroids (close to the centroids) and buy when stocks are oversold (far from centroids) due to would have a bounce.

On the other hand, Both cross validation and hyperparameter tuning I have used Silhouette Score and Davies-Bouldin Index. A graph of different types of data

AI-generated content may be incorrect.

Where Davies suggest me to take the smallest values, which is two, and Silhouette to take the highest value, which is two as well. That means that for these data set is recommended take 2 as several clusters.

## K-Nearest Neighbours

I used the K-Nearest Neighbours model on my dataset to identify patterns for short-term forecasting. It helps analyse repeating trends and reactions to similar market conditions and captures nonlinear relationships without complex modelling. This model is particularly useful because classifies data points based on their proximity to other points in the dataset, Allowing robust pattern recognition even in volatile or unpredictable scenarios. By comparing new data against historical records, the K-Nearest Neighbours model can predict future movements with significant accuracy. Additionally, it doesn't require the assumptions of linearity, making it valuable for diverse and complex datasets where traditional models might fail.

After training and testing, standardize my data, and execute the machine learning model, the result was:

* R2 score: 0.9865  
  This result is close to 1, which means that in this occasion this model captures almost all variability of my data set about stock prices. Nevertheless, this could also mean that there is overfitting.
* MAE: 1.3231 & MSE: 3.314  
  Which the average of error in my model prediction, in this case my model’s prediction is $1.32 away from the actual prices. Meanwhile, MSE means that my model’s squared error average is 3.314, focusing on large errors than small errors.

A graph with blue dots and red lines

AI-generated content may be incorrect.

The blue dots represent the model's predictions, while the red line indicates perfect predictions. In this visualization, each blue dot corresponds to a specific prediction made by the model for a given data point. The closer these blue dots are to the red line, the more accurate the model's predictions are. Deviations from the red line show errors or inaccuracies in the model's predictions, providing insights into areas where the model may need further refinement or improvement. This comparison helps assess the model’s performance and guides adjustments to enhance the predictive accuracy.

On the other hand, as we have been doing through this project, it is time to apply hyperparameter tuning and cross-validation for each machine learning model. In this case for this machine learning model (KNN) I did not apply cross validation, this is because the previous machine learning model explained I coded “Time Series Split”, that means that for this exercise will have the same name “tscv” and it was split into 5, increasing training. However, in the case of hyperparameter I have used another method for k-nearest neighbours, trying to find the best value for the ML model, which is the number of similar data point that the model considers doing any kind of predictions. After analysing the results from the parameter grid, this determine that the optimal parameter is 10. This indicates that using 10 neighbours will achieve a balance that yields predictions with less errors and better generalization for the dataset, allowing the model to consider sufficient data points and avoid outliers.

Simple Exponential Smoothing

Finally, after applying Time Series Model and ARIMA model to transform into a stationary series I have decided to continue applying a basic series forecasting technique, simple exponential smoothing. This will help me to predict futures values calculating past observations and averaging from my dataset. Like this data set is about stock prices from Medtronic, I consider that is interesting to understand the use of SES in this dataset, in this case, SES will help me to find opportunities in the stock market when prices deviate from the smoothed trend, this smoothed will remove the noise and focus in the short-term trends.

A graph of a stock market

AI-generated content may be incorrect.