**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning for Business |
| **Assessment Title:** | CA1 Project |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Daniela Mariano Barreto |
| **Student Number:** | 2023278 |
| **Assessment Due Date:** | 24/04/2024 |
| **Date of Submission:** | 21/04/2024 |

**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. |

Table of Contents

[**Introduction** 3](#_Toc164510232)

[Motivation 3](#_Toc164510233)

[Problem domain and objectives 3](#_Toc164510234)

[Data description 3](#_Toc164510235)

[Word count 4](#_Toc164510236)

[**Clustering Algorithms** 4](#_Toc164510237)

[DBSCAN 5](#_Toc164510238)

[OPTICS 7](#_Toc164510239)

[**Time Series Analysis** 8](#_Toc164510240)

[ARIMA 8](#_Toc164510241)

[**Evaluation** 9](#_Toc164510242)

[**Conclusion** 9](#_Toc164510243)

[**References** 9](#_Toc164510244)

# **Introduction**

A global trend in pedagogical approach, known as Massive Open Online Courses (MOOCs), serves as an alternative and supplement to traditional models of learning by utilizing online platforms (Sharma, 2013, p.19). It provides high-quality educational content that students worldwide can access more quickly and easily (Thuy et al., 2023, p.1-2).

Since 2012, numerous platforms have emerged for online education, among which Coursera stands out as the most popular due to its variety of courses and strong partnerships with prestigious higher educational institutions like MIT, Harvard, and Stanford (Sharma, 2013, p.19; Zotova et al., 2021, p.167).

## Motivation

During the pandemic period, MOOCs experienced a significant increase (Serravallo, 2020, p.1). This market is estimated to reach USD 22.8 billion this year, with an expectation of USD 119 billion by 2029 (www.mordorintelligence.com, n.d.), making it a lucrative investment opportunity. According to this source, the Coursera platform experienced a 640% increase during the pandemic compared to the previous period.

## Problem domain and objectives

With the modernization of the educational system with MOOCs, platforms such as Coursera developed and sought to serve the most diverse areas of knowledge, signalling promising investment prospects.

In this context, this project seeks to compare two clustering algorithms (DBSCAN and OPTICS) to evaluate Coursera's course based on student rates and approximate duration required to complete them, which could enable recommendations on course duration. Analysis of stock market for Coursera using the ARIMA algorithm will also be covered.

## Data description

The dataset used is from the Kaggle repository (Elvin, 2024, p.1) and contains records of various courses available on the Coursera platform. The time series data is from the stock market of Coursera and was gathered from Yahoo Finance. It can also provide information about the actual financial situation of this educational platform.

Time series data: [Coursera, Inc. (COUR) Stock Price, News, Quote & History - Yahoo Finance](https://finance.yahoo.com/quote/COUR/)

A close-up of a document

Description automatically generated

**Table 1:** Data dictionary based on the repository information.

## Word count

Introduction:

Clustering algorithms:

Time series analysis:

Evaluation:

Conclusion:

**Total:**

# **Clustering Algorithms**

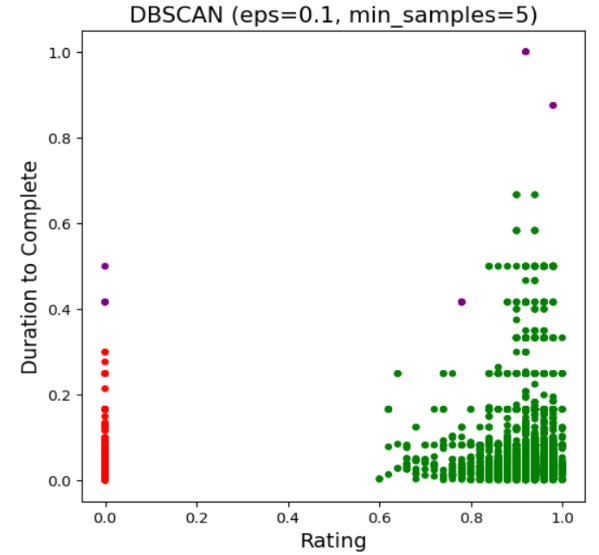
Unsupervised learning algorithms predict without a target label, relying solely on features (Müller and Guido, 2017, p.131). Clustering algorithms group objects based on similarity criteria, useful in business for recommendations and segmentation (Navlani, Fandango and Idris, 2021, p.325).

The Silhouette score and Davies-Bouldin index (DBI) assess clustering quality (Navlani, Fandango, and Idris, 2021, p.350). DBI considers compactness and separation, with lower scores indicating better clusters. Silhouette score measures cluster separation, with higher values indicating better results (Navlani, Fandango, and Idris, 2021, p.351).

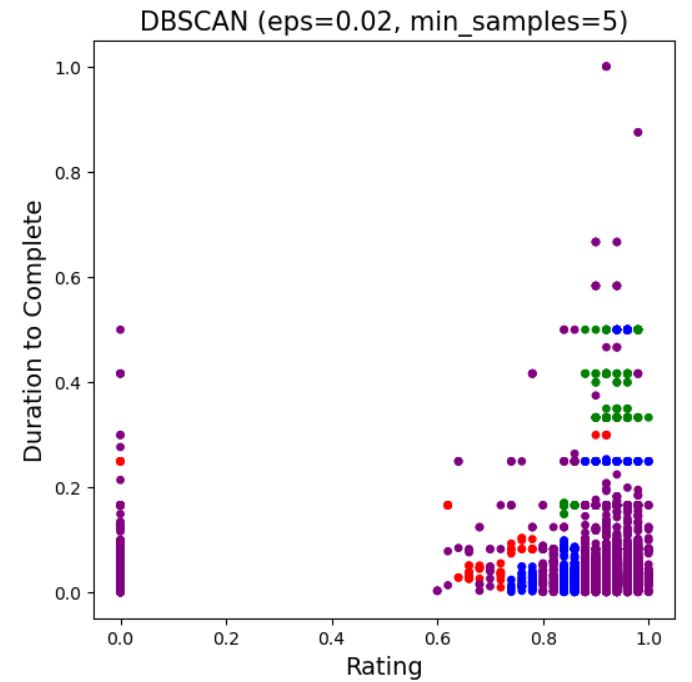
The segmentation was performed using the Density Based Spatial Clustering of Application with Noise (DBSCAN) and the Ordering Points To Identify the Clustering Structure (OPTICS) algorithms.

## DBSCAN

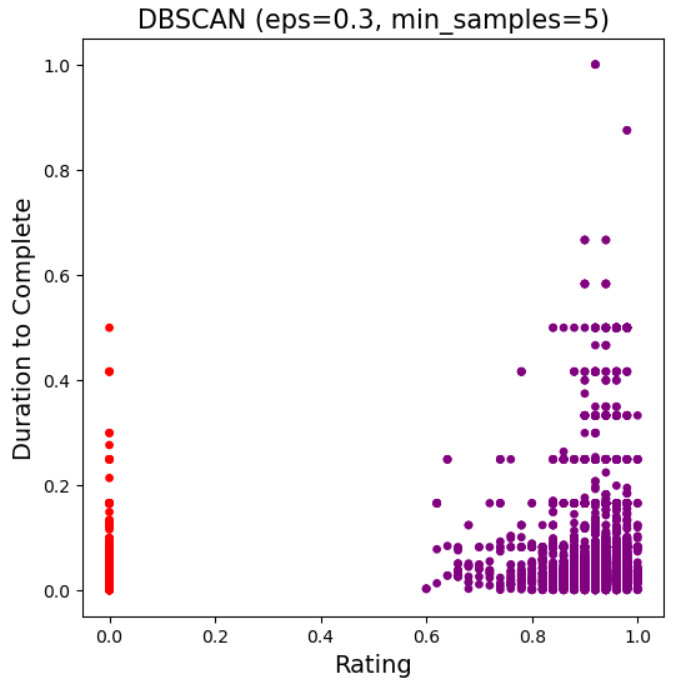
The DBSCAN can be suitable for this dataset because it can handle relatively large datasets without pre-setting the number of clusters, capture complex and dense shapes, and identify points that are not part of any cluster (noise) (Müller and Guido, 2017, p.187), even though determining parameters might be challenging.



b



a



c

**Figure 2:** DBSCAN clustering results with different epsilon values. a) Silhouette score equal 0.2790 and Davies-Bouldin index (DBI) equal 1.3923. b) Silhouette score equal 0.8810 and BDI equal 0.7302. c) Silhouette score equal 0.9061 and DBI equal 0.0978.

2 clusters + noise

30 clusters + noise

2 clusters + no noise

## OPTICS

This algorithm tends to be beneficial to this project since it can provide insights into the distribution and correlation within the dataset by creating an ordering of the data based on its density, storing the core distance and a reachability distance for each object, and helping to identify clusters of varying densities in the data (Ankerst et al., 1999, pp. 49-60).

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

# **Time Series Analysis**

a

What insights can you derive from the initial exploration of the time series data based on the provided topics? Describe any trends, seasonality, or anomalies observed. How did you determine the appropriate parameters (p, d, q) for the ARIMA model. Evaluate the performance of the ARIMA model in forecasting future values, highlighting any strengths and limitations based on your chosen dataset.

b

## ARIMA

As the series is non-stationary, Auto-Regressive Integrated Moving-Average (ARIMA) model will be used. The term Integrated means the model will transform the series into a stationary one.

# **Evaluation**

# **Conclusion**

# **References**

Thuy, T., Thanh Tu Tran, Gia Khuong An and Nguyen, P. (2023). Students’ Perception Towards Learning Massive Open Online Courses on Coursera Platform: Benefits and Barriers. *International Journal of Emerging Technologies in Learning (ijet)*, 18(14), pp.4–23. doi:https://doi.org/10.3991/ijet.v18i14.39903.

Sharma, R.C. (2013). How to Offer a Massive Open Online Course (page 19-20). *EduComm Asia*. [online] Available at: https://www.academia.edu/5505040/How\_to\_Offer\_a\_Massive\_Open\_Online\_Course\_page\_19\_20\_. [Accessed 16 Mar. 2024].

Zotova, M., Likhouzova, T., Shegai, L. and Korobeynikova, E. (2021). *The Use of MOOCS in Online Engineering Education*. International Journal of Engineering Pedagogy (iJEP), [online] 11(3), p.157. doi:https://doi.org/10.3991/ijep.v11i3.20411. [Accessed 16 Mar. 2024].

Serravallo, J. (2020). Connecting with students online : strategies for remote teaching & learning. Portsmouth, New Hampshire: Heinemann.

www.mordorintelligence.com. (n.d.). Massive Open Online Course (MOOC) Market | Growth, Trends, and Forecasts (2020 - 2025). [online] Available at: <https://www.mordorintelligence.com/industry-reports/massive-open-online-course-mooc-market>. [Accessed 17 Mar. 2024].

Navlani, A., Fandango, A. and Idris, I. (2021). *Python Data Analysis*. Packt Publishing Ltd.

Müller, A. C. and Guido, S. (2017). Introduction to machine learning with Python: a guide for data scientists. 1st ed. United States of America. O’reilly Media.

Elvin, R. (2024). Coursera Dataset. [online] Available at: https://www.kaggle.com/datasets/elvinrustam/coursera-dataset. [Accessed 12 Feb. 2024].

Ankerst, M., Breunig, M.M., Kriegel, H.-P. and Sander, J. (1999). *OPTICS: ordering points to identify the clustering structure*. ACM SIGMOD Record, 28(2), pp.49–60. doi:https://doi.org/10.1145/304181.304187.