National College of Ireland

Configuration Manual

MSc Research Project

MSc In Data Analytics

Leela Sri Venkata Sai Naryana Vakacharla

Student ID: X23246383

School of Computing

National College of Ireland

Supervisor: Bharath Agarwal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Leela Sri Venkata Sai Narayana Vakacharla
Student ID:	X23246383
Programme:	MSc In Data Analytics **Year:** 2025-2026
Module:	Research Practicum
Lecturer:	Bharat Agarwal
Submission Due Date:	28/01/2026
Project Title:	Cluster-Based Customer Segmentation for Targeted Marketing in Retail Using Unsupervised Learning Algorithms
Word Count:	581 **Page Count:** ……..……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Leela Vakacharla

Date: 28/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Forename Surname

Student ID:

1 System Requirement and Software Configurations

The analytical environment used to develop this project was written in Python and was aimed at facilitating data preprocessing, clustering, and evaluation processes. Its implementation is based on Python 3.10+ and common libraries of data science, such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-Learn. These were needed to process retail transaction data, form RFM (Recency, Frequency, Monetary) variables, form visualisation, and run clustering algorithms (K-Means, DBSCAN, Hierarchical Clustering and Gaussian Mixture Models). These dependencies were managed by invoking pip package manager and provided a consistent version control that implies reproducibility. The coding interface was developed in Jupyter Notebook to execute the cells interactively and provide the ability of running data cleaning, data exploration, feature scaling and model training in an iterative fashion. Another configuration of the system involved warnings merits that would have been enabled on cleaner output, PCA is imposed to reduce the dimensionality, and sufficient memory is allocated to process loading and the transformation of the data. It was completely compatible with both windows and the macos operating systems as long as the machine had 8GB of RAM or more since it could effectively handle clustering algorithms and visualisation processes.

2 Preparation and Data Processing

The data utilised in this project comes out of contentedly accessible Kaggle data Customer Segmentation: Clustering. When the data set was loaded in the environment, various preprocessing interventions were performed in order to prepare the data to be used in unsupervised learning. df.info, df.describe and head/ tail previews have been used to initially inspect data and ensure that the data is of the correct format, has no missing points, and that the ranges of the variables are within acceptable limits. To accomplish the cleaning process, it was necessary to eliminate duplicates, correct the inconsistent date formats as well as fill in the absent values especially the Income field with the median. Outlier management was carried

out with the IQR approach on Recency, Frequency and Monetary with an end result of capping the extreme values of both 1 st and 99 th percentile with stabilising clustering. The RFM variables were then modelled by determining the timeliness of a customer buying, frequency of transaction, and the amount that was spent. To avoid the bias related to the scale of algorithms based on distance, these features were normalised by MinMaxScaler to a range between 0 and 1. To better visualise PCA was used to transform three RFM features into two principal components in the treatment of the evaluation score.

3 Model Implementation and Environment Set-Up

To carry out the implementation part, four algorithms involving clustering were performed, including K-Means, DBSCAN, Hierarchical Clustering, and GMM, with the help of Scikit-Learn clustering modules. The reusable Python functions were used to set the number of clusters, hyperparameters, and evaluation metrics differently and uniquely to each model. Elbow Method was used to choose the K-Means value of k=4 and the parameters of DBSCAN (eps and min samples) were experimented with. Agglomerative linkage was operated in hierarchical clustering whereas probabilistic soft clustering was utilised in GMM to identify overlapping patterns. Silhouette Score, Davies-Bouldin Index and Calinski-Harabasz Index were used to measure model performance creating a table of comparison between algorithms which ranked by cohesion, compactness and separation. The effectiveness of each of the models could be qualitatively checked by visualising clusters in the space of PCA-transformed data. Every step of configuration was such that it would be reproducible, transparent and conform to the dissertation analytical flow.