National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	Bolormaa Mendbayar		
Student ID:	x23176725		
Programme:	MSc in Data Analytics	**Year:**	2024
Module:	Data Mining and Machine Learning 2		
Lecturer:	Michael Bradford		
Submission Due Date:	19/05/2024		
Project Title:	Q1: Prediction of the value of the Ethereum crypto-currency Q2: Review of Parsimonious Bayesian model-based clustering with dissimilarities		
Word Count:	Q1: 2717, Q2: 1448		

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 references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Bolormaa Mendbayar

Date: 19/05/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. You must ensure that you retain a HARD COPY of ALL projects, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. Late submissions will incur penalties.
5. All projects must be submitted and passed in order to successfully complete the year. Any project/assignment not submitted will be marked as a fail.

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

Prediction of the value of the Ethereum crypto-currency

Ethereum (ETH), a major cryptocurrency, is a decentralized blockchain platform that establishes a peer-to-peer network that securely executes and verifies application code, called smart contracts. Predicting future ETH value is a complex challenge because it's constantly changing and depends on many things such as market and technological development. This proposal describes a method for using machine learning to predict the value of Ethereum.

1. Exploratory Data Analysis / Data Cleaning

Technical indicators commonly used in predicting cryptocurrency values include historical price data, candlestick patterns, trading volume, and moving averages, as outlined by Bandara *et al.* (2015), Christina and Umbara (2015), and Udagawa (2018).

For this project, data will be sourced from reputable platforms such as CoinMarketCap for cryptocurrencies and Yahoo Finance for traditional assets. The dataset will encompass price data from May 15th, 2019, to May 16th, 2024, with each row representing a daily timestamp data. Daily data is chosen primarily due to the significant volatility observed during daily intervals, making it possible to capture price fluctuations effectively. Moreover, the decision to use daily data aligns with recommendations from prior research, as many studies focus on weekly returns. Additionally, daily data collection enables a larger number of observations to be gathered within a limited time frame.

The most relevant variables to ETH, including Bitcoin (BTC), Tether (USDT), and the S&P 500, will be considered. The dataset will contain 40 variables encompassing various factors related to ETH, BTC, USDT, and the S&P 500.

ETH	**BTC**	**USDT**	**S&P 500**
Date	Date	Date	Date
Open and Close prices	Open and Close prices	Open and Close prices	Open and Close prices
Low and High prices	Low and High prices	Low and High prices	Low and High prices
Trading Volume	Trading Volume	Trading Volume	Trading Volume
Market Cap (USD)	Market Cap (USD)	Market Cap (USD)	Market Cap (USD)
Total Addresses	Total Addresses	Total Addresses	
Supply last active 3 years ago	Supply last active 3 years ago	Supply last active 3 years ago	
Supply last active 3y-5y	Supply last active 3y-5y	Supply last active 3y-5y	
Supply last active 5y-7y	Supply last active 5y-7y	Supply last active 5y-7y	

Table 1. All variables

Variable explanation:

Open and Close Prices: Starting and ending prices within a day's trading period.

Low and High Prices: Lowest and highest price levels during the day's trading period, reflecting price volatility.

Trading Volume: Total amount of each cryptocurrency or index exchanged during the day, indicating market activity and liquidity.

Market Cap: Total value of all circulating coins or assets, calculated by multiplying the current price per coin or unit by the total supply, providing a measure of overall market size and valuation.

Total Addresses: Total number of unique addresses that have participated in transactions on the blockchain network for cryptocurrencies, or the number of participants for the S&P 500, reflecting the level of network activity and adoption or investor interest.

Supply: Total quantity of coins in circulation at a given time, excluding permanently lost or removed coins, influencing scarcity and perceived value in the market. Market dynamics by indicating the distribution over different time horizons will help to understand long-term holder behavior, market sentiment, and historical trends, contributing to more informed predictions about the coin's price movements and market performance.

The study requires several crucial steps to ensure the analysis is reliable and effective.

- Descriptive Statistics and Data Visualization: As part of the initial phase of the analysis, conduct descriptive statistics to understand the tendencies of the ETH price data. Following the methodology outlined by Kim *et al.* (2021), compute basic statistics including mean, median, standard deviation, and range for the ETH price data. These statistics serve as a foundation for further analysis and model development. To identify any trends, seasonality, or outliers in the data, create visualizations such as Time series plots, box plots, and histograms to explore the distribution of ETH prices over time.
- Time Series Decomposition: Decompose the time series data into its trend, seasonal, and residual components using techniques like seasonal decomposition of time series (STL) or moving averages. This provides insights into the underlying patterns and cycles in ETH price data.
- Handling Missing Values and Data errors: The majority of the papers used on-chain data had no missing values. In this case will check the missing values, if missing values are present, will build K Nearest Neighbour (KNN), a method recommended by Kumar, Pv, and Jackson (2023), which involves computing the average from nearby existing data points. Unlike simply dropping missing values, KNN imputation allows to make the most of the available information by leveraging similarities between neighboring data points. This approach leverages the information from neighboring data instances to impute missing values accurately. Additionally, identify and correct any errors or inconsistencies in the dataset, such as duplicate entries or data formatting issues, to ensure the dataset is clean and ready for further analysis.
- Outlier Detection and Treatment: Detect and treat outliers in the dataset using the Z-score analysis method. This method is selected due to its simplicity and effectiveness in providing standardized measures to identify data points that deviate significantly from the mean, ensuring robustness in the analysis.
- Normalization: Normalize each feature using the Z-score normalization formula as mentioned in Kim *et al.* (2021). This step is essential to ensure that all variables are on a similar scale, facilitating fair comparison and analysis of the data. Z-score normalization is often the preferred method for time series data due to its ability to handle different scales and stabilize variance. While other normalization techniques have their merits, they may not be as effective in dealing with the unique challenges presented by time series data, such as trends, seasonality, and varying scales.

- Sliding Window Method: Employ a sliding window method as outlined in Kim *et al.* (2021) with a window size of n=7 to prepare the ETH price data. This involves dividing the data into weekly samples, allowing for the analysis of trends and patterns over time.

2. Dimensionality Reduction/ Feature Selection

To develop an effective predictive model for ETH price, it's crucial to address the high dimensionality of the dataset and select the most relevant features. This section outlines the methodology for dimensionality reduction and feature selection, focusing on techniques to streamline the analysis and enhance model performance.

To determine the most influential features for predicting ETH's value, the cross-correlation function (CCF) method will be employed, as highlighted in Kim *et al.* (2021). They utilized a comprehensive collection of 254 variables and considered time lags ranging from -3 to 3, resulting in the identification of the top 42 variables exhibiting the highest correlation with BTC price. Cryptocurrency markets often exhibit lagged effects, where the impact of certain variables on ETH price may manifest with a delay. CCF allows for the detection of such lagged effects by examining correlations at different time lags, thereby enhancing the predictive capacity of the model.

While alternative feature selection methods exist, such as mutual information, principal component analysis (PCA), machine learning-based approaches, and Random Forest modeling, CCF is deemed most appropriate for this study. This decision is based on its ability to capture time-dependent relationships and identify lagged effects, which are inherent characteristics of cryptocurrency markets. Furthermore, CCF offers simplicity, interpretability, and widespread applicability in time-series analysis, making it well-suited for the analysis.

Mutual information, PCA, machine learning-based approaches, and Random Forest modeling offer distinct advantages in certain contexts. However, CCF is chosen over these alternatives due to its specific relevance to the dynamics of cryptocurrency markets. Mutual information, for instance, may overlook temporal dependencies, while PCA may not effectively capture lagged effects. Machine learning-based approaches, while powerful, may require substantial computational resources and are often less interpretable than CCF.

The insights gained from feature selection using CCF will directly inform the development of the predictive model for ETH price. By identifying the most relevant features, the aim is to improve the model's accuracy and robustness in forecasting ETH price movements.

3. Feature Engineering / Feature Extraction

Feature engineering is crucial for model development, incorporating domain knowledge and extracting meaningful patterns from raw data. By averaging the high and low prices across various coins and assets daily, can gain a comprehensive perspective of price dynamics across different cryptocurrencies and traditional assets. Analyzing the variations in these daily average prices can offer valuable insights into predicting the price of ETH.

There are many ways to analyze price movements, but averaging highs and lows is a simple and effective method for this situation. It captures the overall price trends and makes it easy to see how ETH's price relates to other assets. Although other methods exist, like considering past prices (lagged variables) or using moving averages, this approach is a good starting point for this specific case.

4. Choice of modeling techniques

This section will discuss the methodology employed in previous studies for cryptocurrency price prediction, with a focus on the choice of modeling techniques.

Most studies in cryptocurrency price prediction, including Zoumpekas *et al.* (2023) and Kim *et al*. (2021), have leveraged LSTM (Long Short-Term Memory) networks as a primary modeling choice. LSTM is preferred for several reasons:

Cryptocurrency price data is essentially sequential, with each data point depending on previous observations. LSTM networks are well-suited for handling such sequential data, making them ideal for time-series forecasting tasks like cryptocurrency price prediction. Traditional time series models may struggle to capture the complex and non-linear patterns present in cryptocurrency prices.

LSTM is a type of recurrent neural network (RNN) architecture specifically designed to capture long-term dependencies in sequential data. Unlike time series models, LSTM networks possess memory cells that can maintain information over long sequences, allowing them to capture temporal dependencies effectively (Van Houdt *et al*., 2020).

In the study by Zoumpekas *et al*. (2023), LSTM models demonstrated superior performance in terms of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) compared to other recurrent neural network architectures such as GRU. This suggests that the additional complexity of LSTM networks, including the presence of a cell state and multiple gates, contributes to their ability to capture the dynamics of cryptocurrency prices more effectively.

Similarly, Kim *et al*. (2021) achieved notable accuracy metrics including Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), MAE, and RMSE using LSTM modules within their Self-Attention-Based Multiple LSTM framework for predicting ETH and BTC prices.

While time series models like ARIMA (AutoRegressive Integrated Moving Average) have been widely used for forecasting, they may not be suitable for capturing the complex and non-linear patterns exhibited by cryptocurrency prices. LSTM networks, with their ability to capture both short-term and long-term dependencies, offer a more flexible and powerful approach for cryptocurrency price prediction.

5. Hyperparameter optimisation

Hyperparameters play a crucial role in determining the performance of machine learning models, as they are set before the learning process begins. To optimize these hyperparameters, the grid search technique will be employed. Grid search involves defining a grid of hyperparameter values and exhaustively searching through all combinations to find the optimal set of hyperparameters. This method, chosen for its simplicity and efficacy, involves defining a grid of hyperparameter values and systematically evaluating all combinations to identify the most optimal set.

Following the approach outlined by Ortu *et al.* (2022), the hyperparameters of four deep learning algorithms fine-tuned: Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), Multi-Attention Long Short-Term Memory Fully Convolutional Network (MALSTM-FCN), and Convolutional Neural Network (CNN). Specifically, hyperparameters such as learning rate, number of LSTM units, dropout size, and batch size were selected for this optimization.

They defined specific searching intervals for each hyperparameter and iterated through all combinations to identify the configuration that yielded the best performance in terms of selected prediction error metrics. To define ranges for each hyperparameter, common values used in similar studies and an understanding of the ETH price prediction problem will be considered. For instance, the learning rate might range from 0.001 to 0.1, and the number of LSTM units from 50 to 200.

6. Model evaluation

The model evaluation phase of the project involves assessing the performance and reliability of the predictive model developed for ETH price prediction. To achieve this, the dataset will be split into training

and testing sets using a time-based splitting strategy to preserve temporal order, and cross-validation techniques such as k-fold cross-validation will be implemented. This technique ensures that the model learns from historical patterns and generalizes unseen future data effectively, aligning with the dynamic nature of cryptocurrency markets.

Since the target variable is numeric, various evaluation metrics are selected to validate the results. Evaluation metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) will be calculated to evaluate the model's accuracy as outlined previous paper by (Zoumpekas, Houstis, and Vavalis, 2023) and Kim *et al.* (2021). This comprehensive approach will assess various aspects of the model's performance, encompassing both the extent and proportionality of its prediction errors.

7. Scalability issues

In this project, scalability issues may arise in various stages of the analysis. Efficient data collection and storage strategies will be necessary to manage the growing volume of historical price data and additional variables from multiple platforms.

As computational demands increase with complex algorithms like LSTM, access to powerful computing resources, such as high-performance clusters or cloud solutions, becomes essential. Techniques like CCF for feature selection and model training with hyperparameter optimization may face scalability challenges due to the growing dataset size, potentially leading to longer processing times.

Ensuring the predictive model's scalability involves continuous adaptation to evolving market dynamics and data characteristics, which may require frequent updates and maintenance to sustain its effectiveness over time. These scalability considerations will be vital for designing a robust and sustainable framework for ETH price prediction.

8. Ethical implications

In conducting this project, it is essential to address ethical considerations to ensure responsible data usage and minimize potential risks associated with cryptocurrency price prediction. While the dataset sourced from reputable platforms such as Yahoo Finance and CoinMarketCap does not involve personal data, several ethical implications can be considered.

It's important to maintain transparency and accountability in how to collect and analyze the data to avoid biases and inaccuracies. The predictions made by the model could impact financial markets and influence investor decisions, so it's essential to provide clear information about the model's limitations to prevent people from making uninformed investment choices. Additionally, need to consider the broader societal effects of cryptocurrency hypothesizing, like wealth inequality and risky investment behaviors. By focusing on transparency, accountability, and societal impact, this research aims to promote responsible practices in cryptocurrency prediction.

REFERENCES

Bandara, M. N., Ranasinghe, R. M., Arachchi, R. W. M., Somathilaka, C. G., Perera, S., & Wimalasuriya, D. C. (2015). A complex event processing toolkit for detecting technical chart patterns. *In Proceedings of the IEEE International Parallel and Distributed Processing Symposium Workshop* (pp. 547–556). doi: 10.1109/IPDPSW.2015.83.

Christina, C., & Umbara, R. F. (2015). Gold price prediction using type2 neuro-fuzzy modeling and ARIMA. *In Proceedings of the 3rd International Conference on Information and Communication Technology (ICoICT)* (pp. 272–277). doi: 10.1109/ICoICT.2015.7231435.

Hansson, P. (2022). The Underlying Factors of Ethereum Price Stability: *An Investigation on What Underlying Factors Influence the Volatility of the Returns of Ethereum.* Master's Thesis, Jönköping University.

Kim, G., Shin, D. H., Choi, J. G., & Lim, S. (2021). A deep learning-based cryptocurrency price prediction model that uses on-chain data. *IEEE Access*, 10, pp. 56232–56248.

Kumar, S. A., Pv, G., & Jackson, B. (2023) "Machine Learning-Based Timeseries Analysis for Cryptocurrency Price Prediction: A Systematic Review and Research," in 2023 *International Conference on Networking and Communications (ICNWC)*, Chennai, India, pp. 1-5. doi: 10.1109/ICNWC57852.2023.10127439.

Ortu, M., Uras, N., Conversano, C., Bartolucci, S., & Destefanis, G. (2022). On technical trading and social media indicators for cryptocurrency price classification through deep learning. *Expert Systems with Applications*, 198, 116804.

Udagawa, Y. (2018). Predicting stock price trend using candlestick chart blending technique. *In Proceedings of the IEEE International Conference on Big Data (Big Data)* (pp. 4162–4168). doi: 10.1109/BigData.2018.8622402.

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53, 5929–5955. Available at: https://doi.org/10.1007/s10462-020-09838-1.

Zoumpekas, T., Houstis, E., & Vavalis, M. (2023). ETH analysis and predictions utilizing deep learning. *Journal of Cryptocurrency Research*, 5(2), 123-135.

Review of Parsimonious Bayesian model-based clustering with dissimilarities

1. Structure & Title

The title is clear and accurately reflects the focus on Bayesian Model-Based Clustering (BMCD) and its application to dissimilarities. However, it lacks detail regarding the specific contributions or innovations introduced in the paper. The title should be more descriptive to highlight the paper's unique contributions and practical applications. For example, Enhanced Bayesian Model-Based Clustering for Dissimilarities: New Models, Selection Criteria, and Computational Strategies.

The structure follows a logical sequence, starting with an introduction, followed by methodology, simulation studies, practical application, and conclusion. The paper covers all necessary aspects, from introducing the topic to providing validation and application.

2. Abstract

The abstract gives a brief overview of the paper's goal which is to expand the BMCD method by suggesting alternative models, model selection criteria, and strategies for reducing computing time. It explains that when only distances between data points are available, standard model-based clustering cannot be used, so this method offers a solution. While it effectively outlines the focus and contributions of the paper, it could provide more specific details on the proposed extensions to enhance clarity and impact.

The abstract could be a bit clearer on exactly what improvements are being made and why they matter. It is concise but could be a bit more detailed about the key findings and proposed improvements to the BMCD framework without exceeding an appropriate length.

3. Introduction

The introduction is effective in explaining the basics of clustering and why it's important for finding patterns in data. It clearly states that the Gaussian Mixture Model (GMM) is a popular method but highlights its limitation when only distances between objects are available. The introduction mentions alternative methods and then introduces BMCD as a solution, explaining its benefits, including handling measurement errors.

However, the introduction could be clearer about the specific problem it addresses and why BMCD is significant. It would help to organize the introduction better by clearly separating each main point, such as the limitations of GMM, the introduction of BMCD, and the new models proposed in the paper. Including more specific citations for the benefits of BMCD and the new models would also strengthen the introduction. Overall, it sets up the rest of the paper well but could be more precise and better organized.

4. Graphical Abstracts and/or Highlights

The graphical abstract effectively highlights the core contributions of the paper, clearly representing the BMCD method and its proposed extensions, such as comparing diagnostics for different weight plots, clustering partitions from Mclust 1 and BMCD, simulation study plots, and BIC values for 42 candidate models. The use of color and design makes it easily understandable, engaging, and appealing to readers.

However, including diagrams that compare the performance of the new BMCD models against traditional methods in various scenarios would be beneficial, particularly to highlight their effectiveness in high-dimensional spaces and their ability to handle measurement errors. Additionally, while highlights are not

extensively used in the paper, emphasizing the development of new models, alternative selection criteria, reduced computational time, and successful real-world applications would enhance the paper's clarity and impact.

5. Methodology

The methodology section gives a detailed explanation of how the research was conducted, focusing on steps like Gibbs sampling, prior distributions, and model parameters. Replicating the experiments might be tough due to the complexity of the method, especially with techniques like Gibbs sampling and Bayesian methods. However, the paper explains the algorithms and steps involved in the BMCD method well, which could help others reproduce the experiments. Using software like Mclust for fitting models can make the process easier. However, including specific details about the computational environment, random seed settings, and any preprocessing steps would further enhance reproducibility.

Although the paper explains the algorithms clearly, it might be useful to start with a review section summarizing relevant previous studies and explaining how this research fits into fields like psychology, genomics, market research, and sociology. The authors thoroughly reference the BMCD method by Oh and Raftery (2007), showing their understanding of the challenges and applications in related fields. To make the transition smoother between the review and methodology sections, it could be helpful to connect the ideas more clearly.

The paper accurately explains the BMCD method, including how it estimates object configurations and handles errors. However, it could be clearer about the assumptions behind these new models and how they compare to existing ones. Some parts, like Gibbs sampling and dealing with label switching, might need clearer explanations for better understanding. This could include discussing the advantages and limitations of the BMCD method compared to traditional approaches.

To make the paper more useful, additional examples, real-world applications, and extended simulation results could be included. These would help demonstrate how well the proposed models work in different situations.

6. Results

The manuscript presents a novel approach to Bayesian model-based clustering using dissimilarities, but several areas need improvement to enhance clarity and ensure robustness.

The figures provide detailed explanations and effectively illustrate clustering results, with comprehensive captions explaining the purpose and significance. Adding clear labels and visual aids like bar graphs or heatmaps would make the data more accessible. For example, a bar graph could visually highlight differences in performance metrics among models.

The manuscript's interpretations, particularly regarding the Wisconsin Breast Cancer dataset, need stronger statistical support. For example, the results section mentions the model's performance but does not provide confidence intervals or p-values to quantify the significance of the findings. The discussion should include a more thorough analysis of potential biases and limitations of the Bayesian clustering method. This would provide a balanced view and acknowledge areas where the method may not perform well.

Including a flowchart or schematic of the methodology would help readers understand the step-by-step process of the Bayesian clustering method.

To validate the robustness of the method, it should be tested on a variety of datasets from different fields. This would demonstrate the generalizability of the approach. Conducting comparisons with other clustering

methods would provide a solid foundation for the method's performance. This would show its strengths and weaknesses relative to existing techniques. Performing a sensitivity analysis on the model parameters would help understand how changes in parameters affect the results. This is crucial for ensuring the model's reliability and computational efficiency.

Addressing the computational complexity of the Bayesian clustering method, especially for high-dimensional data, is essential. The manuscript should discuss strategies to reduce computing time. Providing detailed guidelines and examples for the software implementation of the method would make it more accessible to users. This includes clear documentation and user manuals. Addressing these points will ensure the manuscript's findings are clear, reliable, and applicable, thereby justifying its publication.

7. Conclusion/Discussion

The conclusion highlights improvements to Bayesian model-based clustering, including five new models and better computational methods. These improvements are essential for working with high-dimensional data. The practical use of BMCD on the Wisconsin Diagnostic Breast Cancer data and its success in simulations show its value. However, the conclusion could explain more clearly the limitations in computational efficiency and specify the data types where BMCD works best to avoid making overly broad claims. Statements about BMCD being better than other methods should be more specific about the conditions under which this is true. Redundant details, like repeated mentions of the GitHub repository and computational strategies, should be removed. Methodology summaries and specific results should be included in the abstract instead of the conclusion.

8. Language

The article is well-written, with just a few small grammar mistakes that don't get in the way of understanding. The pictures and charts in the article are important for showing the research findings and backing up the story well. For example, the pictures comparing the BMCD model to true clustering partitions are crucial for showing how accurate and effective the proposed models are. The charts make the data easy to understand and are all consistent, with clear scales and bars that are the same width. Figure 9, clearly compares how well the BMCD model fits compared to true clustering, making it easy to see how well the model performs. The tables give countless details from both simulations and real-world examples, helping to understand how effective the BMCD algorithm is and how it's better than older methods.

9. Previous Research

The paper appropriately references earlier research, particularly the work by Oh and Raftery. It cites their 2007 paper introducing BMCD and their 2001 paper on Bayesian multidimensional scaling (BMDS). Other important works related to clustering techniques and their applications in various fields are mentioned. The article includes key methods like Gaussian mixture models and model selection criteria like BIC and ICL, building on existing research to propose new improvements to the BMCD method.