International Journal of Network and Communication Research 8(1),1-22,2024

Online ISSN: 2054-636X (Online)

Print ISSN: 2054-6351(Print)

Website: https://www.eajournals.org/

Publication of the European Centre for Research Training and Development-UK

A Hybrid Machine Learning Model for Clustering and Prediction of Closing Price of Cryptocurrency

Mfoniso Asuquo[1] and Imeh Umoren[2*]

Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Nigeria

*****Corresponding author: imehumoren@aksu.edu.ng

doi: https://doi.org/10.37745/ijncr.16/vol8n1122 Published January 27 2024

Citation: Asuquo M. and Umoren I. (2024) A Hybrid Machine Learning Model for Clustering and Prediction of Closing Price of Cryptocurrency, *International Journal of Network and Communication Research,* 8(1),1-24

ABSTRACT: *The Present financial system operates on technological innovations with attention to cryptocurrencies and its related processes. Blockchain technology has increased exponentially since the release of Bitcoin in year 2008 as the first viable decentralized cryptocurrency. In recent years, research on machine learning, blockchain technology, and interaction has improved significantly. This work considers utilizing clustering and Machine Learning techniques to predict the closing price of cryptocurrencies. The study adopts K-means, Model Based Clustering Algorithms and Artificial Neural Network (ANN) approaches to implement clustering and time-series predictions. The capability of the Neural Network to provide one-day closing price prediction of Bitcoin was evaluated. The paper espouses a training ratio of 70:30 on the dataset deployed in the work for an increased accuracy due to large number of data-points. Basically, using K-Means and Model-Based Clustering algorithms indicated 74.5% accuracy based on the Elbowing method adopted for the determination of optimal number of clusters in the data set. Increasing the number of clusters to 10 in the data points demonstrates an accuracy of 88.4%. The Empirical findings reveal that, Mean Squared Error (MSE) score shows 1.20, the Mean Absolute Error (MAE) score illustrate 2.9 on the test dataset after evaluation of the prediction model. A comparative analysis shows the advantages of using K-Means, Model-Based Clustering algorithms and artificial neural network to provide trustworthy, automatic monitoring and clustering as well as prediction. This further reveals that it is feasible to produce an estimation for which price moving indicators can impact the actual coin closing price operations.*

KEYWORDS*:* blockchain technology, cryptocurrencies, machine learning (ml), model-based clustering algorithms, artificial neural network (ANN), k-means cluttering

INTRODUCTION

The advent of technology based economy, notably application of artificial intelligence (AI) result in widespread use of new financial tools. Individuals and financial organizations alike have considered machine learning, a sub-field of AI to create new approaches to drive information

system security and digital economy. Since the introduction of Blockchain technology, most published research works have focused on non-technical elements of the technology, such as legal difficulties and its involvement in criminal activity [10]. Given the novelty of Blockchain technology and the quick advancements in machine learning techniques, research on the intersection is still in its infancy compared to many other fields. The age of bank digitalization and Data Science stemmed with both prospects and challenges [11]. In [11], the paper assert that, the Internet has played a significant role in changing how we interact and communicate with other people as well as how we do business today. As a result of the Internet, electronic commerce has emerged, allowing businesses to effectively interact with their customers and other establishments inside and outside their industries. Various strategies have been created based on previous studies, including the use of AI to forecast the future price of cryptocurrency. Basically, prediction techniques may be divided into two groups: prediction-based techniques and cluster-based approaches. Prediction-based approaches utilize tools such as Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Decision Support System (DSS). Other techniques include Hidden Markov Model (HMM), Naive Bayes (NB), NN, Support Vector Regression (SVR), and SVM. Similarly, approaches based on filtering, fuzzy, k-means, and optimization are classified as clustering techniques [5]. Cryptocurrency prices are extremely volatile, and this complexity attracts researchers and statisticians who want to Figure out how to predict the future values of such cryptocurrrency prices. Despite the numerous study papers published in related areas to price prediction of cryptocurrency and stocks, many people still believe that cryptocurrency prices cannot be predicted. This is due to the large number of factors that influence coin and token values, many of which are dependent on other, possibly unknown factors and majorly due to its decentralized system. Furthermore, the price of a crypto currency is a time series that is inherently noisy, dynamic, nonlinear, intricate, nonparametric, and chaotic [23]. The most popular method for financial market forecasting is to utilize previous price change experiences to forecast future price changes. According to [8] "financial forecasting is an example of a signal processing task that is tough owing to short sample sizes, excessive noise, non-stationarity, and non-linearity." This is because the cryptocurrency price data is often coin values from real-time transactions, obtaining a larger data sample will require a longer time period. The cryptocurrency financial markets are not always static over extended periods of time; thus a longer length of time does not necessarily yield the best results. However, if the elements that influence crypto prices can be identified, it may be feasible to predict prices without depending so much on previous data. It is practically hard to comprehend all of the variables that affects crypto prices, this research explores a possible solution to the aforementioned problems.

LITERATURE REVIEW

The problems associated with cryptocurrencies are one of the essential issues that many studies have tried to overcome. The Machine learning techniques have been proposed in order to support the clustering and prediction of closing price of cryptocurrency. In these studies, the authors have attempted to highlight the significant of the approaches in finance been explored for decades. The

amount of research done on the effectiveness of machine learning approaches for predicting the price of cryptocurrency is minimal. In [21], *End-2-End* behavioural analysis of wormhole attacks on the transmitting networks layer of MANET was realized to detect the level of (security) wormhole attack on the network based on several parameters considered by determining the degree of severity of wormhole attack which may upset the Quality of Service (QOS) delivery. Again, [12] used Bayesian neural networks (BNN), as well as elements drawn from Bitcoin and block chain technology's principles. The authors validated Bitcoin's significant volatility and developed a successful price time series predictor. On modeling and forecasting the Bitcoin process, they perform an empirical research that compares the Bayesian neural network to other linear and non-linear benchmark models. Through their cross validation and bootstrap resampling methods, they were able to extract training and testing dataset from the original dataset. In their work, evaluation metrices including root mean square error and mean absolute percentage error were techniques adopted to assess the performances of all the trained models. Based on their experimental results, they demonstrated that BBN based model performed comparatively better than other models in effectively describing the bitcoin log price and price volatility. [14] looked at anticipating Bitcoin price variations from a classification-based method. They used three classification based algorithms: Binomial GLM, SVM and Random Forest. The binomial GLM classifier performed better than the other two with an accuracy score of 98.7% followed by Random Forest tree based algorithm with 95% and lastly, SVM achieving an accuracy of 27% in daily forecasting of the bitcoin price, while 10-minute time intervals had a 50-55 percent accuracy for Binomial and Radom Forest. Their method of model performance evaluation were standard techniques used in classification-based algorithms including, specificity score, sensitivity score, precision score and accuracy. Another study looked at the link between fundamental economic data, emotive analysis, and the price of Bitcoin. Support vector machines were used by [6] in their research, they discovered that the frequency of Wikipedia views and the network hash rate had a positive link with Bitcoin's price. Sentiment is increasingly being used as a predictor of bitcoin price due to the enormous quantity of data accessible. Nonparametric classification is described by [19], as a method for forecasting trends. They propose a latent source model, which [4] analyze. According to [19], a latent source is defined as a group of latent variables that provide features that are not immediately observable, rather than a strategy that highlights a system phenomenon from a set of variables. In particular, Bayesian regression is used to forecast Bitcoin price changes, resulting in a roughly twofold return on their initial investment in just 60 days [19]. The paper attained a price movement accuracy of 55% by adding the characteristics to SVM and ANN algorithms. The results obtained by [9] are hopeful and contradict the EHM hypothesis in certain ways. Sentiment analysis has been used in several research, and it is becoming increasingly essential in today's culture, which is heavily impacted by social media. [13] investigated online cryptocurrency forums in attempt to forecast price swings using the Granger casuality test, a well-known method for forecasting currency and stock value. The authors arrived to the conclusion that comments and views on cryptocurrencies on social media had an impact on their pricing. Recently, a number of publications and research have begun to experiment with the use of recurrent neural networks (RNN) on Bitcoin data. RNNs may be able to increase performance, according to evidence. [15]

attempted to forecast the trajectory of Bitcoin's price versus the US dollar in their study. On daily Bitcoin values, both RNN and LSTM (Long short-term memory) networks were evaluated. With a 52 percent accuracy over a 100-day span, the LSTM network produced the greatest results. On sample Bitcoin data, [20] investigated the performance of ARIMA (AutoRegressive Integrated Moving Average) and RNNs. They demonstrated that RNN outperforms non-dynamic ARIMA in terms of performance. In [22], the system used Model-Based Design (MBD) which consists of two steps: the creation of a conceptual model and the transformation of that model into mathematical representation. The Model formulation represent the process of converting the understanding of a natural system into mathematical form as a quantitative and graphical approach to solving challenges in ensuring that all aspects of data processing are implemented for good systems design.

Summary of Related Literature

From the previous section on the review of related literature, the following summary is drawn;

Table .1: Summary of Related Literature

Author	Work	Methodology Used	Contribution to Knowledge	Limitations
[12]	Bayesian neural networks (BNN), as well as elements drawn from Bitcoin and blockchain technology's principles	Bayesian neural networks (BNN)	They demonstrated that BBN based model performed comparatively better than other models in effectively describing the bitcoin log price and price volatility.	Lack of clustering-based model.
[14]	Anticipating Bitcoin price variations from a classification-based method	Binomial GLM, SVM and Random Forest	The binomial GLM classifier performed better than the other two with an accuracy score of 98.7%, Random Forest tree based algorithm with 95% and SVM achieved an accuracy of 27% in daily forecasting of the bitcoin price.	The accuracy of other models are dramatically low at 27% in daily forecasting of the price of bitcoin.
[15]	Forecasting the trajectory of Bitcoin's price versus the US dollar using RNN and LSTM	RNN and LSTM (Long short-term memory)	With a 52 percent accuracy over a 100-day span, the LSTM network produced the greatest results	The forecasting accuracy varies greatly between the models and cryptocurrencies used in the research work

[20]	Performance of ARIMA and RNNs	Performance Evaluation of ARIMA (AutoRegressive Integrated Moving Average) and RNNs	RNN outperforms non-dynamic ARIMA in terms of performance.	The forecasting accuracy of the models evaluated in the work was low compared to similar studies.
[1]	Cryptocurrency price prediction	using tweet volumes and sentiment analysis.	SMU Data Science	Lack of clustering-based model.
[2]	Distributed ledgers and operations: What operations management researchers should know about Blockchain technology.	Blockchain Technology	Manufacturing & Service Operations Management, .	Lack of clustering-based model.
[3]	Arguments against avoiding RMSE in the literature. Geoscientific model development.	Root mean square error (RMSE) or mean absolute error (MAE).	Geoscientific model development	Lack of clustering-based model.

SYSTEM FRAMEWORK

Proposed System Framework

The proposed system's structure is depicted in the Figure 1. The suggested framework is used as a guide to develop and implement a hybrid machine learning project for forecasting of future price of cryptocurrencies

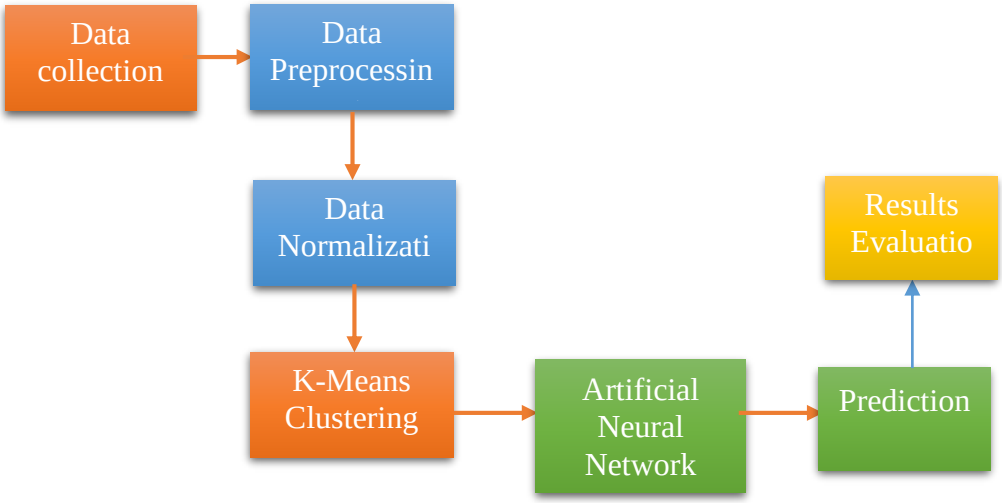

Fig. 1: Structure of the proposed system

MATERIALS AND METHODS

Quantitative research approaches are often used in machine learning based research. This study adopts quantitative research methodology. Quantitative research as an empirical. methods focus on the domain of collecting and analyzing data from many sources, including the use of mathematical, statistical, or computational methods to arrive at good research findings. The work involves machine learning (ML) approach for a systematic use of algorithms and statistical models to efficiently execute tasks without explicit instructions, relying instead on patterns and inference.

Research Design

This paper used an experimental research design to cover the planning, implementation, and analysis of experimentation of financial cryptocurrency market data. The methods used for collection and acquisition of both primary and secondary research data, as well as the use of data preprocessing techniques, model training, testing, and assessment.

Data Collection

The application of quantitative experimental research study measured and allowed the secondary data collection method to be used for collection of historical price data of bitcoin cryptocurrency. The researcher collected bitcoin dataset from yahoo finance aggregator for the period of 2019 to 2021 with time series of a minute prices of bitcoin against the United State Dollars (USD). The dataset contains no header and the time feature is coded using unix timestamp format. Cryptocurrency data was gathered or collected from cryptodatadownload, an on repository for all cryptocurrency ranging or spanning for four years' interval. Top five (5) coins were chosen

according to the latest coin ranking from coinmarketcap. A total of 3500 records were filtered out based on 4 years' interval (2019-2022) which was used for the clustering and prediction.

Data Preprocessing

Data Preprocessing was carried out as an important phase in the research for the model's validity and performance, as well as the model's outputs. Machine learning relies on data to make judgments and forecast future outcomes, given supervised or unsupervised learning, reinforce learning, classification, or regression type of machine and statistical learning, data preprocessing incorporates a set of operations for efficient machine learning research outcomes. Preprocessing data is an important mining technique for transforming raw data into a usable and efficient format Variables, traits, fields, attributes, and dimensions are all terms used to describe features. Table *2* shows how we categorize statistical input depending on measurement units.

Table 2: Field Data Attributes Categorization

Attributes	Category/units
Trade Count (Adj_Close)	Numeric
Opening Price (Open)	Numeric
Low Price (Low)	Numeric
High Price (High)	Numeric
Volume	Numeric
Closing Price (Close)	Numeric

Model Training and Testing

The Model Training and Testing phase considers cleaning and data normalizing, a decision on how much of the dataset should be set aside as a training set and how much should be set aside as a testing set was essential. Realistically, in machine learning, the training data set is the real dataset that is used to train the model to perform various actions. This is the actual data that the models learn as part of the continual development process, using various APIs and algorithms to enable the machine to work autonomously. One of the tasks in machine learning is to research and develop algorithms that can learn from and predict data. These algorithms work by creating a mathematical model from input data and producing data-driven predictions or judgments. The training set in this study was taken from statistical data gathered binnance data on cryptocurrency of five different coins. The data was subjected to preprocessing phases to remove certain noisy and redundant data. Our data was divided into a ratio of 0.7%(training) and 0.3% (test set) for which was used during the prediction process. The splitting ratio was 0.7%(training) and 0.3%(testing). After then, we carried out normalization of data which one of the most important part of clustering of data. Table 3 depict the normalized data.

Model Evaluation

A machine learning model's performance and accuracy are used to assess its integrity, validity, and robustness. The work required validation and analyzes of how well the model predicts the output variable (future prices) on data that were not utilized to train the model when we use a supervised learning strategy to train a predictive model. However, classification machine learning models utilized confusion matrix and AU-ROC graphs to evaluate model performance and accuracy. The R-squared coefficient, root means square, mean absolute error and mean square error techniques was adopted to evaluate the model performance and accuracy of the predicting model.

Clustering Algorithm for the proposed K-means

The proposed algorithm listed below can be implemented using any programming language of choice. In the next chapter, we implemented them using R programming language.

i. *We downloaded trading data in .csv format from coin market cap.*
ii. *The downloaded data was read from .csv file*
iii. *The data was summarized to identify the relevant properties of the input data.*
iv. *Data scaling (data normalization) was done*
v. *Initial cluster points were selected to be two randomly*
vi. *Plot the result of iv*
vii. *Apply elbow method in selection of cluster from 1: 10*
viii. *Plot the elbow graph*
ix. *Select the number of clusters that forms the elbow*
x. *Carryout out clustering again based on the selected number in viii.*
xi. *Plot the clusters in ix.*
xii. *Evaluate results using sum of squares*
xiii. *Also apply the data in another clustering technic i.e model-based cluster*
xiv. *Plot the results*

ANN Algorithm for the Proposed Model

i. Load in required libraries
ii. Read in the crypto data
iii. Check the structure of data
iv. Check for any inconsistencies in the data
v. Carryout data slicing in the ratio of 70%: 30%
vi. Normalized the independent variables
vii. Create the model
viii. Compile the model
ix. Fit in data for the model training
x. Evaluate the model on test set
xi. Evaluate the model using Mean Absolute Error
xii. Plot the model
xiii. Plot the predicted value against the actual

MODEL DESIGN

Using price information from the previous 20 days of the moving window, our study's objective is to predict closing prices for a day in the future. Additionally, the work utilizes each cryptocurrency's daily closing price as a data point for analysis. The price history in our dataset dates back to the day the cryptocurrency was first published on a coin-exchange platform. Indeed, all values have been obtained from open cryptocurrency exchanges and are all expressed in United State Dollars. Efficient model design often employed for time-series prediction was used in this work. In other to forecast future closing price of a Bitcoin, a unique model was created from basis. The model design consists of a multilayer perceptron (MLP), a type of feed-forward ANN with three (3) layers of nodes. Except for the input nodes, every node in an MLP is a neuron that functions by means of a nonlinear activation function. Each neuron's output is determined by the activation function for each combination of inputs.

Table 3: Normalized data

open	high	low	Volume.BTC	Volume	tradecount	close
1.909369	1.841285	1.977726	-0.3822412	-0.3363120	-0.4665936	-0.7111860
1.800672	1.870811	1.877671	2.3334093	2.1580125	-0.4648501	-0.7110975
1.796066	1.759571	1.850351	0.8740458	0.7650896	-0.4657184	-0.7124980
1.783655	1.740522	1.845076	0.7612471	0.6608633	-0.4657690	-0.7125574
1.840153	1.791794	1.826770	1.8320111	1.6076880	-0.4651711	-0.7127172
1.792201	1.795924	1.853617	1.9192325	1.6965725	-0.4651282	-0.7119893
1.942079	1.885796	1.832976	3.0940375	2.8081114	-0.4646693	-0.7126071
1.936731	1.917889	1.982062	1.9998198	1.8886535	-0.4651088	-0.7106762
1.766413	1.909171	1.833061	2.7513183	2.4983488	-0.4647927	-0.7107450
1.733920	1.711963	1.796131	0.9356317	0.7928944	-0.4657153	-0.7129392
1.704297	1.688006	1.772472	0.7999115	0.6654935	-0.4658209	-0.7133580
1.747043	1.696826	1.752274	2.7761109	2.3437741	-0.4646807	-0.7137397
1.772712	1.732452	1.728421	3.3790718	2.8905601	-0.4645138	-0.7131889
1.811608	1.934066	1.864112	5.0272497	4.6428089	-0.4632361	-0.7114441
2.109424	2.049908	1.887553	6.3885194	6.0513880	-0.4627516	-0.7123565
2.216877	2.141569	2.171562	2.2724789	2.3592572	-0.4647163	-0.7085202
2.256462	2.187665	2.251845	0.8417126	0.9511082	-0.4656865	-0.7071357
2.296797	2.227020	2.290629	2.3047215	2.5159114	-0.4649346	-0.7066256
2.532766	2.457186	2.305212	3.0322643	3.4345290	-0.4644835	-0.7061051
2.385128	2.473669	2.481075	2.2574913	2.7003062	-0.4648776	-0.7030660
2.449483	2.379943	2.472915	0.7251941	0.9209184	-0.4657600	-0.7050631
2.536391	2.462988	2.520232	1.5226706	1.8568321	-0.4653099	-0.7041389
2.497181	2.495837	2.574669	2.1883205	2.6700523	-0.4649638	-0.7030192
2.413643	2.429921	2.495759	2.0121766	2.4106914	-0.4650491	-0.7035244
2.587212	2.526621	2.482363	2.4382780	2.9230918	-0.4647723	-0.7046006
2.513988	2.515715	2.516431	2.4114170	2.8927815	-0.4647656	-0.7023644
2.512520	2.463410	2.576399	0.7161109	0.9514498	-0.4657991	-0.7033078

1 to 27 of 3,717 entries, 7 total columns

Model Formulation Using Artificial Neural Network (ANN)

The processing element is one of the most basic ANN architectures. It is based on [7] threshold logic unit (TLU), also characterized as the linear threshold unit, which is a somewhat different neuron (LTU). The inputs and outputs are integers (rather than binary on/off values), and each input connection has a weight. In equation (4.1), the TLU computes the weighted average of its inputs. The Threshold L`ogic Unit is illustrated in Fig. 2.

$$z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = x^T w \tag{1}$$

We can then apply a step function to the sum and the outputs the result:

$$h_w(x) = step(z) \tag{2}$$

where

$$z = x^T w \tag{3}$$

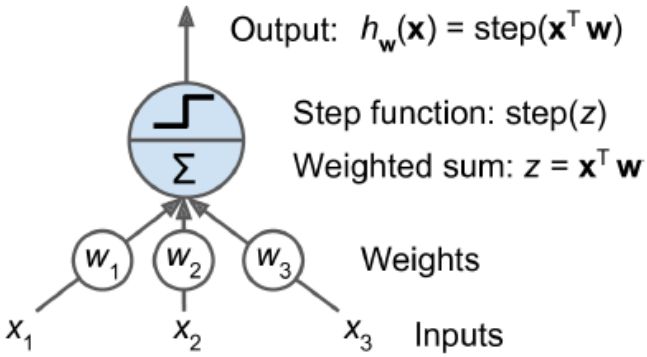

Fig. 2: Threshold L`ogic Unit; Adapted from [7]

$$h_{wb}(X) = \phi(XW + b) \tag{4}$$

Where;

1. X represents the matrix of input features, it has one row per instance, one column per feature.
2. The weight matrix W contains all the connection weights excerpts for the ones from the bias neuron. The layer comprises one row for each input neuron and one column for each artificial neuron.
3. All of the connection weights between the bias neuron and the synthetic neurons are contained in the bias vector b. It has one bias term per artificial neuron.
4. The function is known as the activation function, and it is a step function when the artificial neurons are TLUs.

Perceptron's, on the other hand, are trained using a rule that considers the network's error and reinforces connections that assist minimize the error. The Perceptron is presented with one training sample at such a time and makes recommendations for each one. It reinforces those weights of the connections from of the inputs which would have contributed to the right prediction for every output neuron that made an erroneous prediction. The following is the perceptron learning rule: eqn 4.5:

$$w_{ij}^{next\ step} = w_{ij} + \eta(y_j - \hat{y}_i)x_i \tag{5}$$

where,

1. w_{ij} is the connection weight between the i^{th} input neuron and the j^{th} output neuron
2. x_i is the i^{th} input value of the current training instance.
3. $\hat{y}_i$ is the output of the j^{th} output neuron for the current training instance.
4. y_i is the target output of the j^{th} output neuron for the current training instance.
5. η is the learning rate.

5.2 Model Formulation Using Clustering

Clustering is essentially an unsupervised learning technique. Unsupervised learning is a technique for extracting references from datasets that contain input data but no labeled responses. It is a method for identifying significant structure, explaining underlying processes, generating traits, and groups in a set of samples. Hence, clustering is the process of partitioning a population or set of data points into several groups so that data points in the same group are more similar to each other and dissimilar to data points in other groups. It is essentially a collection of objects based on their similarity and dissimilarity. When selecting a clustering algorithm, we need to about whether it will scale to your dataset. Machine learning datasets can contain millions of samples, but not all clustering algorithms scale well. The similarity between all pairs of instances is computed by many clustering algorithms.

Types of Clustering

Clustering can be done in several ways. A Comprehensive Survey of Clustering Algorithms has an exhaustive list. Each strategy is best suited for a specific data distribution. A brief explanation of four common techniques is provided here, with a focus on centroid-based clustering using k-means.

Centroid-based Clustering

In comparison to hierarchical clustering described below, centroid based clustering organizes data into nonhierarchical clusters. The most extensively used centroid-based clustering algorithm is k means. The efficiency of centroid based algorithms is limited by beginning conditions of outliners. Because k means is an efficient, effective and easy clustering algorithm, it is the subject of this course. Figure 3 shows a sample of centroid-based clustering.

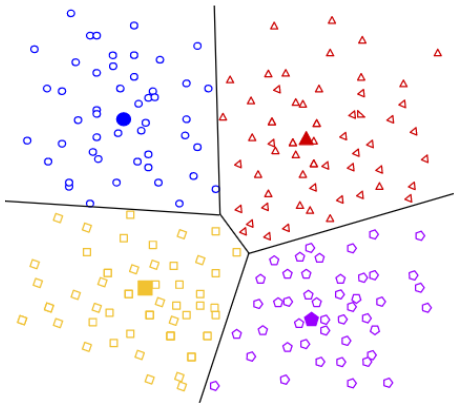

Fig. 3: Centroid-based clustering.

Density-based Clustering

Clusters of high example density are connected using density-based clustering. Therefore, has long as dense areas can be connected, arbitrary-shaped distributions are possible. These algorithms

struggle with data with data with a wide range of densities a dimension. Figure 4 illustrates density-based clustering.

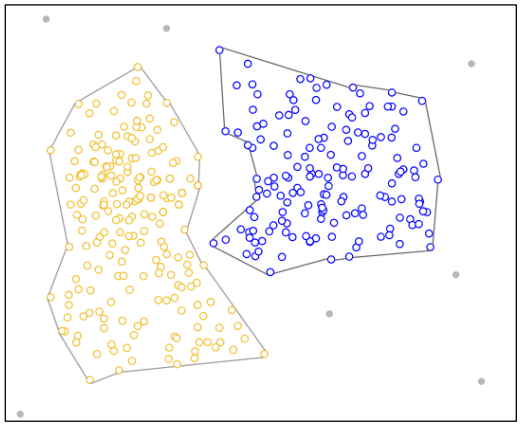

Fig. 4: Density-based clustering

Distribution-based Clustering

This clustering method presupposes that data is made up of distributions, such as Gaussian ones. In Figure 5, the data is clustered into three Gaussian distributions using a distribution-based technique. The likelihood that a place belongs to the distribution diminishes as distance from the center increases. The bands depict the likelihood drop. We must use a different algorithm if you don't understand the type the distribution of your data.

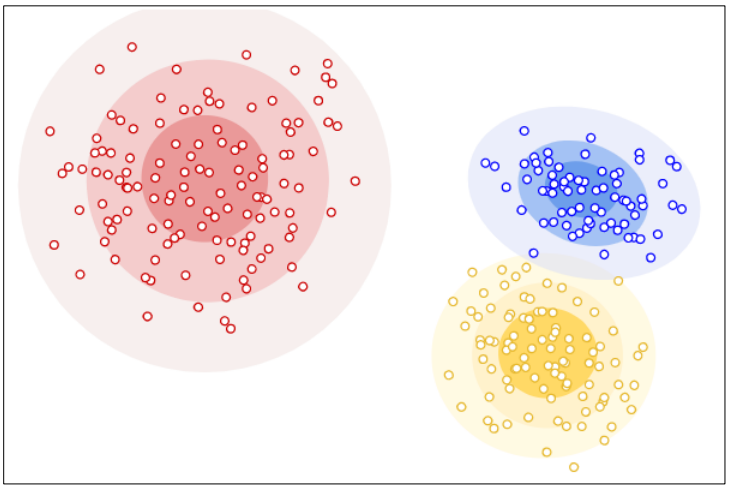

Fig. 5: Distribution-Based Clustering.

Hierarchical Clustering

A tree of clusters is created through hierarchical clustering. Not expectedly, hierarchical clustering is highly suited to hierarchical data, such as taxonomies Figure 6 Illustrate a sample hierarchical tree clustering Animal.

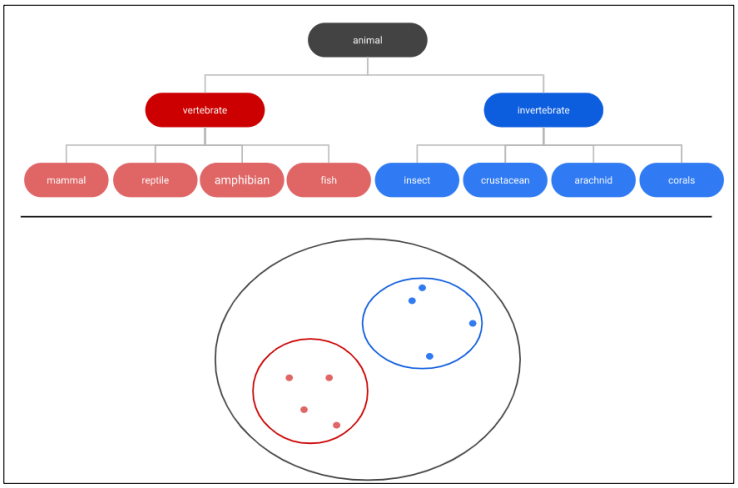

Fig. 6: A sample hierarchical tree clustering Animal

The centroid-based clustering algorithm which is the k means clustered is adopted for this research for the efficient clustering of crypto currency data gathered from yahoo finance. The K-means algorithm is presented below. Unsupervised Learning Algorithm K-Means Clustering divides the unlabeled dataset into different clusters. K specifies the number of pre-defined clusters that must be produced during the process; for example, if K=2, two clusters will be created, and if K=3, three clusters will be created, and so on. K-means provides a simple approach to cluster data into multiple groups and a quick way to determine the categories of groups in an unlabeled dataset without any training. In this centroid-based method, every cluster is associated with just a centroid. This method's primary purpose is to lower the sum of ranges between data points and the clusters to which they correspond.

The k-means clustering algorithm is typically used to achieve two objectives:

1. Determines the best value for K center points or centroids iteratively.

2. The nearest k-center is assigned to each data point. Data points that are near to a given k-center create a cluster.

As just an outcome, every cluster has datasets that are comparable but distinct from the others. The K-means Clustering Algorithm is illustrated in Figure 7.

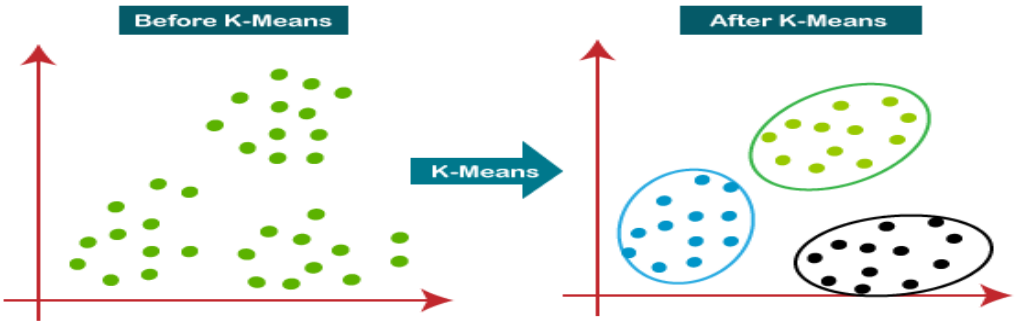

Fig. 7: K-means Clustering

Hence, the step by step algorithm is presented below as;

i. Step 1: To determine the number of clusters, choose the number K.
ii. Step 2: At random, choose K locations or centroids. (It's possible that it's not the same as the arriving dataset.)
iii. Step 3: Allocate every piece of data to the closest centroid, resulting in the preset K clusters.
iv. Step 4: Compute the variance and move the centroid of each cluster.
v. Step 5: Repeat the third stage, reallocating each piece of data to its new cluster centroid.
vi. Step-6: If there is a reassign, go to step 4; else, go to FINISH.
vii. Step 7: The model has now been finished.

Data Visualization

Examining data to determine what it means is known as exploratory data analysis (EDA). EDA is extremely important with today's data since it doesn't require a pre-specified hypothesis. In the past, conducting traditional data analysis required the analyst to have a basic grasp of the data in order to formulate clear questions to draw conclusions from the data and to know in advance which hypothesis to test. However, unless an iterative method is utilized to extract knowledge from data and the data's understanding is updated using EDA, it is hard to know what to look for because data is mined exponentially from numerous sources and in varied forms. Visual data exploration is at the heart of EDA (VDE). VDE is a technique for displaying data visually. Through presentation, the analyst might get anticipated or unexpected insights into the data, enabling the formulation of hypotheses. VDE aims to combine modern computational power with human perceptive skills in the information discovery process. The following visual data investigation was carried out on the project dataset: Figure. 8: indicates variable pairing

1. Variable Pairing

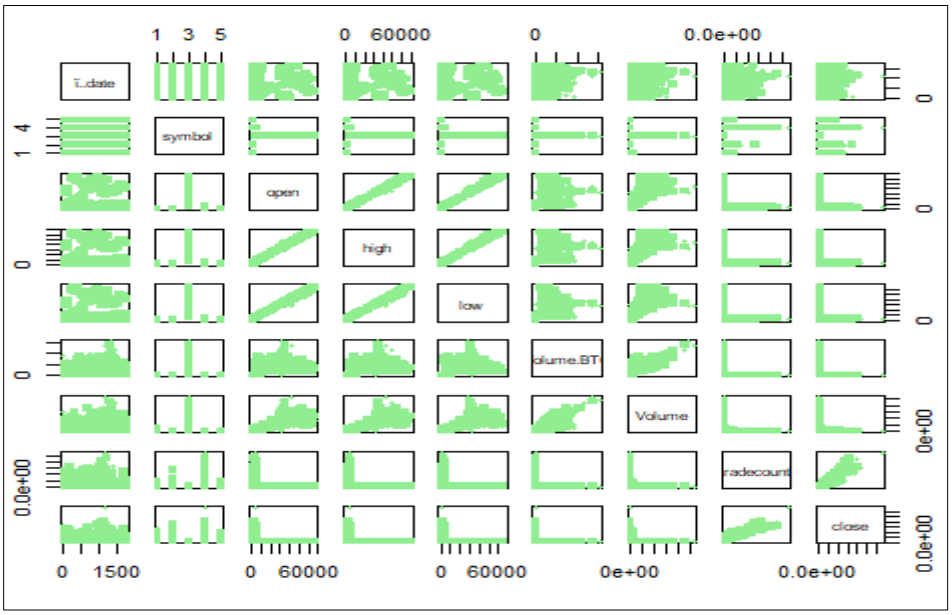

Fig. 8: Variable pairing

Classification Plot

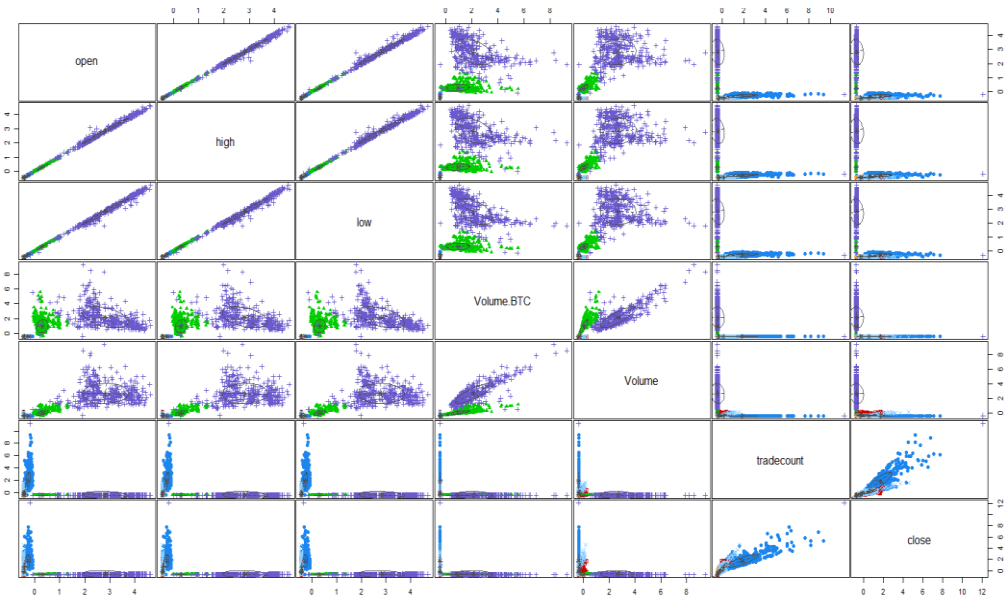

Fig. 9: Classification plot against Independent Variables

Evaluation Methodology

In this section, the methods of evaluating the cluster algorithm is presented.

Cluster Evaluation

We carried out performance evaluation on the k-means clustering using sum of squares. Figure 10a depicts the performance of our 4 cluster which gave 74.5% accuracy.

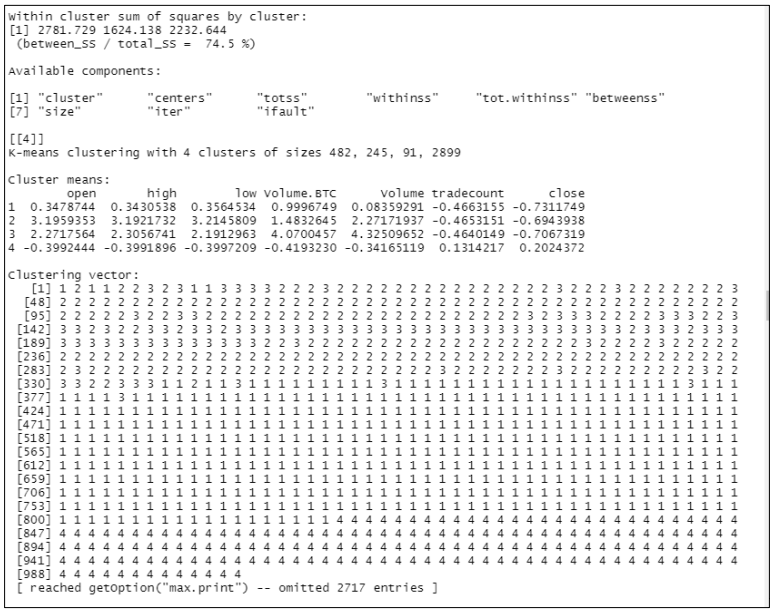

Fig. 10 a: cluster Evaluation

Furthermore, it was observed that on increasing the number of clusters to 10, the accuracy was high which gave an 88.4% which is depicted in Figure 10 b depicts cluster Evaluation

```
Clustering vector:
   [1]  2  1  2  2  1  1  1  1  1  2  2  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
  [32]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
  [63]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
  [94]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 [125]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 [156]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 [187]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 [218]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 [249]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 [280]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 [311]  1  1  1  2  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  2  1  2  2
 [342]  1  2  2  2  2  2  2  2  2  2  1  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 [373]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 [404]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 [435]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 [466]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 [497]  2  2  2  2  2  6  6  6  6  2  2  2  2  2  6  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 [528]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 [559]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 [590]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 [621]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  6
 [652]  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 [683]  2  2  2  2  2  2  2  2  6  2  2  2  2  2  2  2  2  6  2  6  6  6  6  6  6  2  6  2  2  2
 [714]  6  6  2  2  2  6  6  6  6  2  6  2  2  2  6  6  2  2  2  6  2  2  2  2  2  2  2  2  2  2  2
 [745]  6  2  2  6  6  2  6  6  2  2  2  6  2  6  6  2  2  6  6  2  2  2  2  2  2  2  6  2  6  6
 [776]  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6
 [807]  6  6  6  6  6  6  6  6  6  6  6  6  6  5 10 10  5  5  5  5 10 10  8  8  5 10  5 10 10  5
 [838]  7  5  5  7 10 10  7  9  9  8  8  5 10  5  5  5 10  5 10 10  5  5  5  5  5 10  3  5  5 10 10
 [869]  5 10  3 10  5  5  5  5 10 10  5  5  8  5  5  5  5  5  5  5  5  5 10  3  5  5  5  5  5 10 10
 [900]  5 10  5  5 10 10  3  5  5  5  5  5 10 10  5  5  5  8  8  8  5  8  9  5  8  8  5 10  5  5  5
 [931]  5  5 10  5  5  8  5  5  5 10  5  8  5  5  5  5 10 10  5  5  8  5  8  5  8  8  5  5  5  5 10
 [962] 10 10  5  5  5  5 10  5  8  5  5  5 10 10 10  5 10  5  5 10 10 10 10  5 10  5  5  5  5  5  5
 [993]  8  5  8 10  5  8  5  5
 [ reached getOption("max.print") -- omitted 2717 entries ]

Within cluster sum of squares by cluster:
 [1] 1879.14179  567.68890   33.35349   29.63689   74.08236  126.58960   32.84075   73.89864
 [9]  127.54078   75.67629
 (between_SS / total_SS =  88.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

Fig. 10 b: Cluster Evaluation

Furthermore, we, evaluate our training model based on the test set which we splited in order to see how or model fit well on test and predicted. The Predicted vs. Actual plot is illustrated in Figure 11.

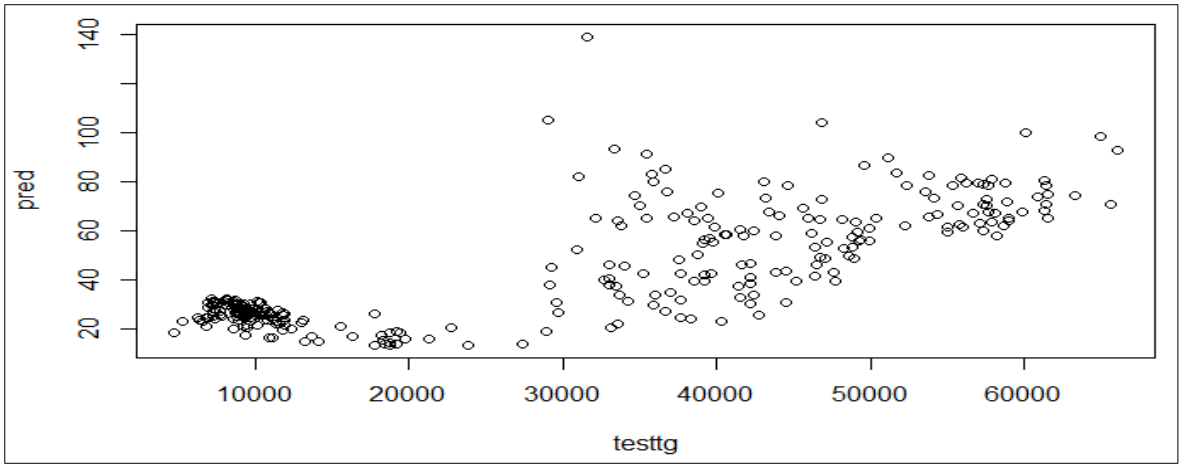

Fig. 11: Predicted vs. Actual

RESULTS AND DISCUSSION

When evaluating the clustering algorithm based on the k-means technique using sum of squares, the model (denoted using four clusters) as depicted in Figure 10, we obtained an accuracy score of 74.5%. The K-Means Clustering and Model-Based Clustering algorithms was evaluated using the Elbowing Method which was adopted for the determination of the optimal number of clusters in the dataset. When the number of clusters to 10 the accuracy gave us 88.4%. Before arriving on the final models provided in this research, neural network adopted in this research was built by optimizing model hyper parameters such as activation functions, loss function and optimizer. Finally, the evaluation of our prediction model was carried out using mean square error and mean absolute error, Mean Squared Error (MSE) was 1.20 while Mean Absolute Error (MAE) gave us 2.9 after evaluation on the test dataset.

CONCLUSION

In recent years, researchers and practitioners have been interested in machine learning and Blockchain technology. From self-driving vehicles to anonymous cryptocurrency-based payment systems that are widely utilized throughout the world, their applications in the real world are becoming increasingly evident to everyone. The purpose of this research is to give scholars and practitioners a better knowledge of the performance of typical time-series prediction models on cryptocurrencies. Typically, when researchers attempt to predict future prices of cryptocurrency, they employ sentiment analysis. Instead of mood, type of coin or popularity of the coin, the price movement indicators were employed as the main target variables in this project. The idea behind this was to use historical network data of each cryptocurrency to predicts the trend of future prices as well as determine the variable with significance influence on the cryptocurrency market. The results of the artificial neural network reveal that it is feasible to produce a estimation for which price moving indicators can effect coin price. Finally, visualization methods were employed to boost human capacities, improve computational decision-making, and make articulating the relationship between price variables and coin price. When there is limited insight on what questions to ask or hypothesis to be formulated ahead of time, data visualization was used in this project helped to make sense of the data. The use of data visualization, curve plotting, and diagrams facilitated the output of the clustering-based model. Correlated variables, clustering variable relationships and groups were visualized using modern visualization techniques. This further attributed to the effective and efficient design, development and implementation of this proposed system.

ACKNOWLEDGMENTS:

The authors would like to acknowledge the effort of anonymous reviewers and colleagues whose remarks have influenced the final shape of this article. We also acknowledge the services of Department of Computer Science, Faculty of Physical Sciences and ICT Directorate, Akwa Ibom

State University, Nigeria, for their enormous resources and contributions during this research work.

Author Contributions: Conceptualization, M. Asuquo, I. Umoren; methodology, I. Umoren; Software, M. Asuquo; validation; I. Umoren, formal analysis, M. Asuquo; investigation, M. Asuquo; resources, I. Umoren, M. Asuquo, data curation, M. Asuquo; writing original draft preparation, M. Asuquo, I. Umoren; writing review and editing, I. Umoren, M. Asuquo; visualization, M. Asuquo; Supervision, I. Umoren; project administration, I. Umoren, funding acquisition:

All authors have read and agreed to the published version of the manuscript.

Funding: This paper did not receive funding.
Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.
Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Science Review, 1(3), 1.
[2] Babich, V., & Hilary, G. (2020). OM Forum—Distributed ledgers and operations: What operations management researchers should know about blockchain technology. Manufacturing & Service Operations Management, 22(2), 223-240.
[3] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). Arguments against avoiding RMSE in the literature. Geoscientific model development, 7(3), 1247-1250.
[4] Chen, G. H., Nikolov, S., & Shah, D. (2013). A latent source model for nonparametric time series classification. Advances in neural information processing systems, 26.
[5] Dattatraya, K. N., & Rao, K. R. (2019). Hybrid based cluster head selection for maximizing network lifetime and energy efficiency in WSN. Journal of King Saud University-Computer and Information Sciences.
[6] Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D., & Giaglis, G. M. (2015). Using time-series and sentiment analysis to detect the determinants of bitcoin prices. Available at SSRN 2607167.
[7] Geron A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd. ed.). O'Reilly Media, Inc.

[8] Giles, C. L., Lawrence, S., & Tsoi, A. C. (2001). Noisy time series prediction using recurrent neural networks and grammatical inference. Machine learning, 44(1), 161-183.
[9] Greaves, A., & Au, B. (2015). Using the bitcoin transaction graph to predict the price of bitcoin. No data, 8, 416-443.
[10] Holub M. and Johnson J., "Bitcoin research across disciplines," The Information Society, vol. 34, no. 2, pp. 114-126, 2018.
[11] lmeh Umoren1, Saviour lnyang, Onukwugha Gilean (2021). Bayesian Network Algorithm for Predictive Modeling of Cyber Security for Efficient Bank Channels Digitalization. International Journal of Information Security, Privacy on Digital Forensics, Nigeria Computer Society (NCS), Vol. 5, No. 2. Nigeria. www.ncs.org.ng.
[12] Jang, H., & Lee, J. (2017). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. Ieee Access, 6, 5427-5437.
[13] Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S. J., & Kim, C. H. (2016). Predicting fluctuations in cryptocurrency transactions based on user comments and replies. PloS one, 11(8), e0161197.
[14] Madan, I., Saluja, S., & Zhao, A. (2015). Automated bitcoin trading via machine learning algorithms. URL: http://cs229. stanford. edu/proj2014/Isaac% 20Madan, 20.
[15] McNally, S., Roche, J., & Caton, S. (2018, March). Predicting the price of bitcoin using machine learning. In 2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP) (pp. 339-343). IEEE.
[16] Pratiti Mishra, Sumanjit Das, Mrs. Sumati Baral. An Efficient Model for Stock Price Prediction using Soft Computing Approach (2015). inproceedings{Mishra2015AnEM
[17] Rosadi, D., Tarno, Subanar & Suhartono (2013). Analysis of Financial Time Series Data Using Adaptive Neuro Fuzzy Inference System (ANFIS). International Journal of Computer Science Issues (IJCSI), 10(2 Part 1), 491.
[18] Sebastião, H., & Godinho, P. (2021). Forecasting and trading cryptocurrencies with machine learning under changing market conditions. Financial Innovation, 7(1), 1-30.
[19] Shah, D., & Zhang, K. (2014, September). Bayesian regression and Bitcoin. In 2014 52nd annual Allerton conference on communication, control, and computing (Allerton) (pp. 409-414). IEEE.
[20] Torres, D. G. and Qiu H. (2018). "Applying Recurrent Neural Networks for Multivariate Time Series Forecasting of Volatile Financial Data".
[21 Umoren, I., Okpongkpong, E., & Udoeka, I. (2022). A Fuzzy Knowledge-Based Approach for End-2-End Behavioural Analysis of Wormhole Attacks on Mobile Ad-Hoc Networks. *International Journal of Information Systems and Informatics*, *2*(4). https://doi.org/10.47747/ijisi.v2i4.581
[22] Victoria Essien, Imeh Umoren, Idongesit Umoh (2021). A Bio-Informatics System for Intelligent Classification of Severity Index of Hypertension. International Journal of Innovative Research in Sciences and Engineering Studies (IJIRSES). ISSN: 2583-1658 | Vol.: 1 Issue: 2. www.ijirses.com.

[23] Zhang, Y., & Wu, L. (2009). Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network. Expert systems with applications, 36(5), 8849-8854.