PAPER NAME

ALL- An Enhanced K-Means Clustering Algorithm for Pattern Discovery in Big Data .pdf

AUTHOR

Dikpride Despa

WORD COUNT

3824 Words

CHARACTER COUNT

21098 Characters

PAGE COUNT

9 Pages

FILE SIZE

3.8MB

SUBMISSION DATE

Jan 10, 2023 12:44 PM GMT+7

REPORT DATE

Jan 10, 2023 12:45 PM GMT+7

23% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 14% Internet database
- 14% Publications database
- Crossref database
- Crossref Posted Content database
- 17% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Cited material

An Enhanced K-Means Clustering Algorithm for Pattern Discovery in Big Data Analysis of 3-Phase Electrical Quantities

Dikpride Despa[1], Gigih Forda Nama[2]

[1]*Department of Electrical Engineering*
[2]*Department of Informatics*
University of Lampung
**Corresponding author E-mail: gigih@eng.unila.ac.id*

Abstract

The Unila Internet of Things Research Group (UIRG) was developed online monitoring of power distribution system based on Internet of Things (IoT) technology on Department of Electrical Engineering University of Lampung (Unila), has been running for several months, this system monitored electrical quantities of 3-phase main distribution panel of H-building. The measurement system involve multiple sensors such current sensors and voltage sensors, the measurement data stored in to database server and shown the information in a real-time through a web-based application.
Main objective of this research was to capture, analyze, and identified the knowledge pattern of electrical quantities data measurements, using Cross-Industry Standard Process for Data Mining (CRISP-DM) data mining framework, for helping the stake holders to continuous improvement of the quality of electricity services, the initial research limited to total 7708 electrical quantities recorded data that save on database system, since 1 September - 31 October 2018, the dataset consist of 21 attribute electrical quantities such as; voltage, current, power factor values, energy consumption, frequency, on H building 3-Phase main panel control.
Rapidminer as leading application on knowledge discovery application was used to analyze the big data, K-Mean cluster algorithm implemented to identify the data pattern, the result indicated that 3-Phase load was unbalanced, and Phase-0 was the most utilized phase, based on from total 5 cluster analysis result.

Keywords: *Data Mining, Electrical Quantities, Rapidminer, CRISP-DM, K-Mean, Clustering, 3-Phase, Internet of Things (IoT), Big Data*.

1. Introduction

The Unila Internet of Things Research Group (UIRG) was developed online monitoring of power distribution system based on Internet of Things (IoT) technology on Department of Electrical Engineering University of Lampung (Unila), has been running for several months, this system monitored electrical quantities of 3-phase main distribution panel of H-building. The measurement system involve various sensors such current sensors and voltage sensors, while data processing conducted by smart embedded system, the measurement data stored in-to database server and shown the information in a real-time through a web-based application. This measurement system has several important features especially for real-time monitoring, robust data acquisition and logging, security, system reporting, so it will produce an important information that can be used for various purposes of future power analysis such estimation and planning.

Main objective of this research was to capture, analyze, and identified the knowledge pattern of electrical quantities data measurements, with its diversity variable, for helping the stake holders of Electrical Engineering Departments to continuous improvement of the quality of electricity services.

These analyze result are particularly useful for the strategic management of Electrical Engineering Departments, because they will know the data trend such as fluctuations in voltage, current, power factor values, energy consumption, frequency, as an evaluation material for making electricity policy in the future.

2. Literature Review

The important step in this research was to determine the best knowledge discovery application for analyze the huge amount data of electrical quantities, produced by the Internet of Things (IoT) system that already running for several months on Electrical Engineering Department Building University of Lampung, Indonesia. We made a decision after considering Gartner's magic quadrant recommendation related to Data Science and Machine Learning Platforms comparation. Gartner is an organization engaged in the field area of IT, Finance, HR, Customer Service, Legal and Compliance, Marketing, Sales, and Supply Chain, and often used as a reference in determining IT strategy for many company all around the world. According to the magic quadrant data in the "2018 Gartner Magic Quadrant for Data Science and Machine Learning Platforms", shows that RapidMiner still as a leader for the last fifth year in a row [1]. It founded on 2007 and claim that will brings artificial intelligence for enterprise organization through an open and extensible data science platform. Until today, there are 430,000 analytics professionals already joint and use RapidMiner

[2]. It provides deep and various modeling capabilities for automated end-to-end development, also has visual workflow designer, guided analytics, and supports automatic retraining of models, based on many platform data inter-connection.

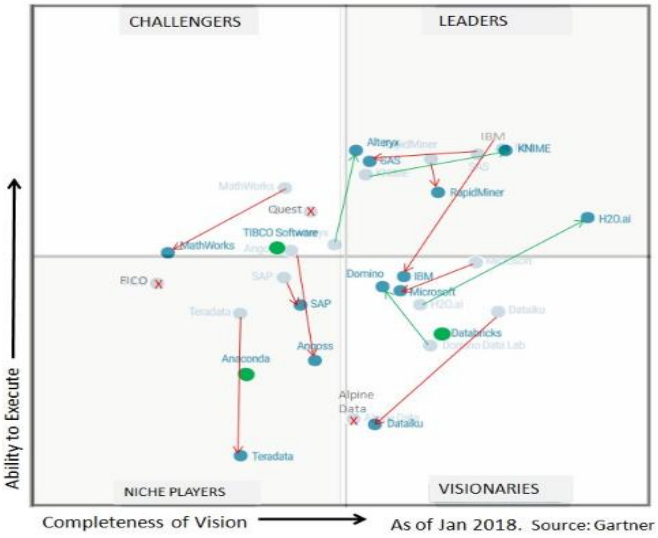


Fig. 1: Gartner Magic Quadrants for Data Science and Machine Learning Platforms compared, 2018 vs 2017 [1]

Figure 1 shows Gartner Magic Quadrants (MQ) for Data Science and Machine Learning Platforms compared, 2018 vs 2017, represent that those are (5) Leaders: RapidMiner, H2O.ai, KNIME, Alteryx, SAS, and (2) Challengers: TIBCO Software (new), MathWorks, (5) Visionaries: Databricks (new), IBM, Microsoft, Domino Data Lab, Dataiku (4) Niche Players: SAP, Angoss, Anaconda (new), Teradata, 3 new firms were added at 2017 those are: Anaconda, TIBCO Software, and Databricks. Three others shown on Magic Quadrant 2017 were dropped that are: Alpine Data, FICO, and Quest.

In the field area of Data Mining research, some research conducted on several works, such; Alduraibi et al, with their research using Rapidminer for predict the gold price movement using several algorithm Decission Tree, SVM, KNN, and linear regression [3], similar to the work done by Estrada developed models that automatically recognize postures by using a web camera with KNN, SVM, MLP [4]. Alhaj also built two classification models (Rule Induction and Random Forest) to predict the survivability of cancer patients of Gazastrip [5]. Cabral et al already made an analysis in field area of fraud detection system for electricity consumption based on data mining technique, they used Self-Organizing Maps (SOM) to gathering the data pattern. Geetha used Rapidminer for meteorogical application for modelling the rainfall prediction using decision trees algoritm, DT also discussed work [6], Rianto2019 et al focused on pattern discovery of users when they are doing online shopping [7]. And some prove of work using K-Means algorithm on clustering could be found on works [8][9][10] [11]

Clustering is a type of categorization inflicted rules on a group of objects. As a result, a cluster is an aggregation of objects. K-means clustering algorithm is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of *k* groups. K-means algorithm can be summarized as follows [12]:

1. Specify the number of clusters (K) to be created
2. Select randomly k objects from the data set as the initial cluster centers or means
3. Assigns each observation to their closest centroid, based on the Euclidean Distance between the object and the centroid, with formulation show on equation formula 1

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad ; i = 1,2,3 \ldots n \tag{1}$$

Where; *xi = coordinat object x on i*
yi = coordinat object y on i
n = dimention of data

4. For each of the k clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster.
5. Iteratively minimize the total within sum of square. That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached.

3. Methodology

The data mining frame work that used on this research was Cross-Industry Standard Process for Data Mining (CRISP-DM), these method could be found on several research followed in succession [13][14][15][16].

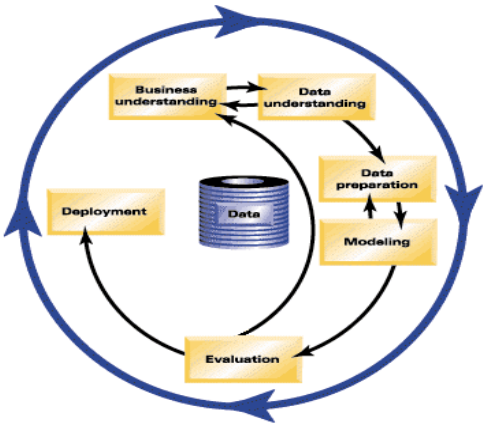


Fig. 2: Phases of CRISP-DM [17]

The CRISP-DM life cycle consists of 6 phases (business understanding, data understanding, data preparation, modeling, evaluation, and deployment), shown on Figure 2, with arrows indicating the most important and dependencies between each phases, while the sequence of the phases is not strict. This model is very flexible and can be easily customized. Instead of modeling, the work will focus on data exploration and visualization to identified knowledge pattern. It allows to create a data mining model that fits with particular needs.

4. Results and Dicsussion

Following are the steps carried out based on the CRISP-DM phase;

A. Business understanding

The first step in the process CRISP-DM is to construct a concrete primary business objective to specific data mining role. Online monitoring of power distribution system based on Internet of Things (IoT) [18][19] technology was deploy and implemented on Department of Electrical Engineering University of Lampung (Unila) for several months, monitored three-phase main distribution panel H-building The measurement system involve multiple sensors such current sensors and voltage sensors, while data processing conducted by smart embedded system, implementing the security model and using several open source program like works on [20][21], the measurement data stored in to the database server

and shown in a real-time through a web-based application. This measurement system has several important features especially for real-time monitoring, robust data acquisition and logging, system reporting, so it will produce an important information that can be used for various purposes of future power analysis such estimation and planning [22], The circuit of the sensor system hardware, and web application can be seen in Figure 3

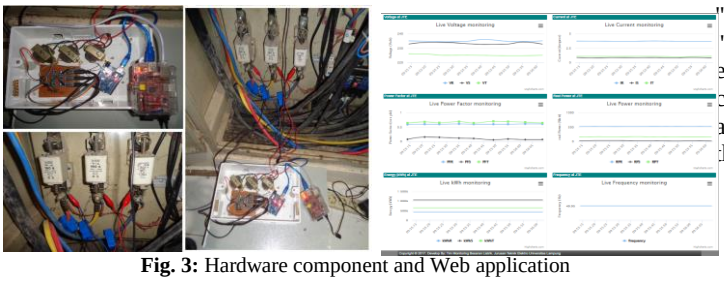


Fig. 3: Hardware component and Web application

1. Voltage transformator 220V:12V, 500 mA
2. Current sensors YHDC SCT 013
3. Resistor burden 33Ω
4. Voltage divider
5. Signal conditioner circuit

B. Data understanding

It is necessary to understand the electrical quantities monitoring data that have been recorded to database system. This research used time series data of monitoring that store on MySQL server platform from 1 September to 31 October 2018.

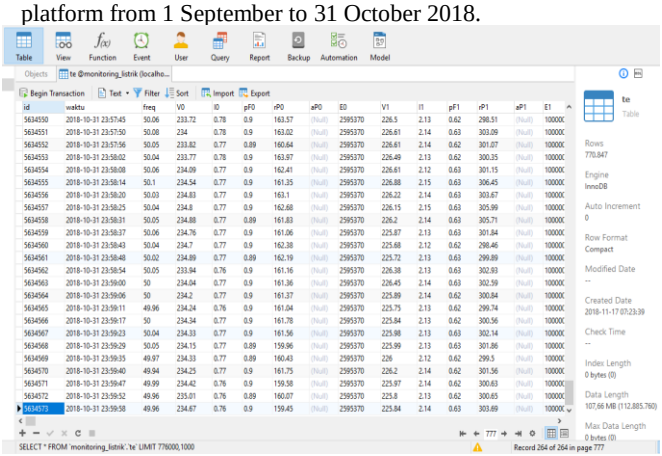


Fig. 4: Monitoring Data recorded on MySQL system

The data structure has 21 attributes those are; "id", "waktu", "freq", "V0", "I0", "pF0", "rP0", "aP0", "E0", "V1", "I1", "pF1", "rP1", "aP1", "E1", "V2", "I2", "pF2", "rP2", "aP2", "E2", with total 770847 recorded data for 2 month system monitoring run, shown on Fig. 4

C. Data Preparation

Data preparation is one of the most important and often time-consuming aspects of data mining. In fact, it is already takes 50-70% of research time and effort on data preparation. In the process of data preprocessing into desired form for statistical analysis and predictive models, we need to perform the following as below:

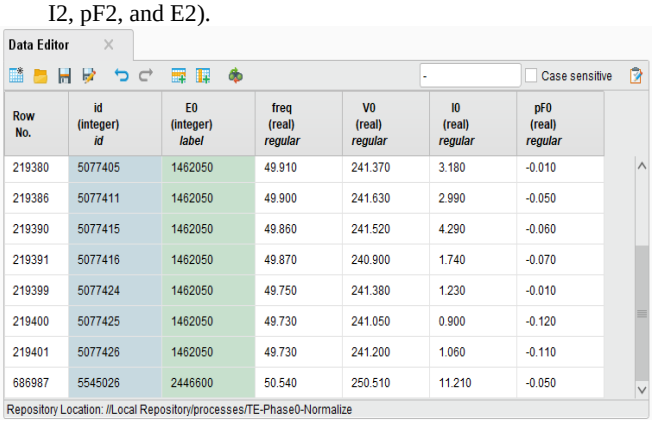
- Merging data sets and/or records
- Selecting a sample subset of data
- Aggregating records
- Deriving new attributes
- Sorting the data for modeling
- Removing or replacing blank or missing values
- Splitting into training and test data sets

Selecting items; The initial research will be limited to 770847 electrical quantities recorded data that save on database system since 1 September – 31 October 2018, so filters need to be set up to exclude the outlier data.

Selecting attributes. The monitoring database contain many information about electrical quantities data, so it is important to filter attributes such as grouping the data according to each Phase.

There were total 21 attribute data those are; "id", "waktu", "freq", "V0", "I0", "pF0", "rP0", "aP0", "E0", "V1", "I1", "pF1", "rP1", "aP1", "E1", "V2", "I2", "pF2", "rP2", "aP2", "E2", we should eliminate (6) attribute that are; rP0, aP0, rP1, aP1,rP2,aP2, because all data on those attribute are not filled, and leave only 16 attribute only, then we grouping the data on to 3 type, according to the exixting Phase,

- TE-Phase0-Normalized has attribute; (id, waktu, freq, V0, I0, pF0, and E0).
- TE-Phase1-Normalized has attribute; (id, waktu, freq, V1, I1, pF1, and E1).
- TE-Phase2-Normalized has attribute; (id, waktu, freq, V2, I2, pF2, and E2).



Data Editor

Case sensitive

Row No.	id (integer) id	E0 (integer) label	freq (real) regular	V0 (real) regular	I0 (real) regular	pF0 (real) regular
219380	5077405	1462050	49.910	241.370	3.180	-0.010
219386	5077411	1462050	49.900	241.630	2.990	-0.050
219390	5077415	1462050	49.860	241.520	4.290	-0.060
219391	5077416	1462050	49.870	240.900	1.740	-0.070
219399	5077424	1462050	49.750	241.380	1.230	-0.010
219400	5077425	1462050	49.730	241.050	0.900	-0.120
219401	5077426	1462050	49.730	241.200	1.060	-0.110
686987	5545026	2446600	50.540	250.510	11.210	-0.050

Repository Location: /Local Repository/processes/TE-Phase0-Normalize

Fig. 5: Grouping and set the attribute role (Phase-0 group)

Fig.5 show the attribute used for Phase-0 group, and set the role of each attribute, E0 was set to label, and id set to id, while the others attribute set to regular role. At this stage we also removing the missing value and outlier data on each attribute, and we found total 17 data within ID ; 5544331, 5543377, 5544191, 5544457, 5544913, 5544921, 5544961, 5544102, 5545028, 5544416, 5544919, 5544105, 5544438, 5544927, 5541075, 5544143, 5544913 was removed from the database.

D. Modelling

After preprocessing data to the desired structure, the CRISPDM process follows with the modeling phase. This phase is the main part of this research. All the models that we have designed and created worked with the same real data from the electrical quantities monitoring system. As part of the research these two types of experiments were designed:

- Modeling the Cluster by using K-Mean algorithm on Rapidminer.
- Identified The Cluster of Electrical Quantities data.

This section describes the performed and the application of the DM Clustering technique used to classify the data of the electrical quantitates data at Electrical Engineering Department building. K-Means algorithm has been executed on 21 original attributes that are part of the file. To select the best attributes, we reviewed and analyze the results, made a several iteration on data preprocessing and eliminate the outlier attribute and data. The best attribute classified by 3 type of category, 1). TE-Phase0-Normalized with 6 best attribute, 2). TE-Phase1-Normalize with 6 best attribute, 3). TE-Phase2-Normalize with also 6 attribute.

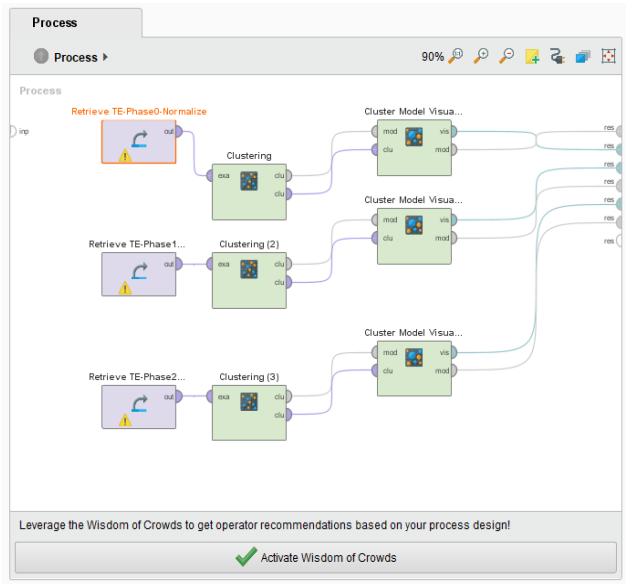


Fig. 6: Process of models and operator creation for clustering analysis on RapidMiner

Figs. 6 show the process for executing the clustering algorithm (K-Means) on RapidMiner, some operators was applied to identified and evaluate the centroids on each cluster. It has the facility to include operators to determine the color of points in each cluster at the time of making a ScatterPlot type chart. Some important parameter implementing on K-Means cluster operator was;

k = 5,
max runs = 10
measures type = Bregman Divergence
divergences = Square Eucledian Distance
max optimization steps = 100

1. DM Modeling result on Phase 0

After executing the previous model build for data on Phase 0 using K-Means algorithm, rapidminer generate the cluster model of data consist of 5 cluster that are; Cluster 0: 522730 items, Cluster 1: 77250 items, Cluster 2: 112741 items, Cluster 3: 7002 items, Cluster 4: 56541 items, with total number of items was: 776264, data visualization of cluster member shown on fig. 7.

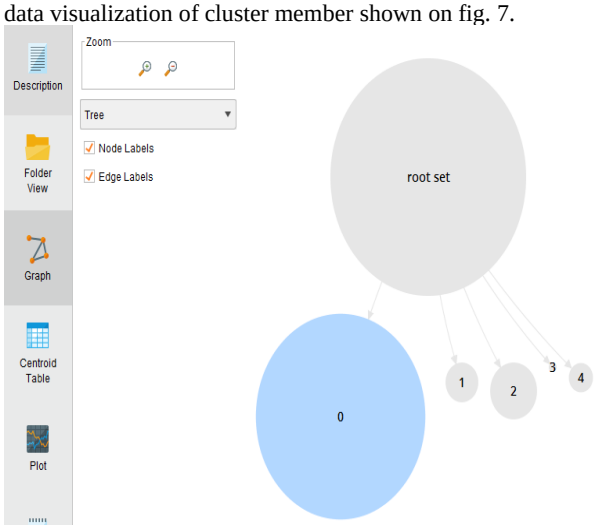
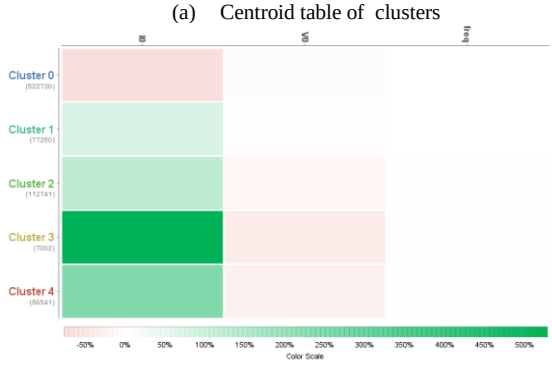


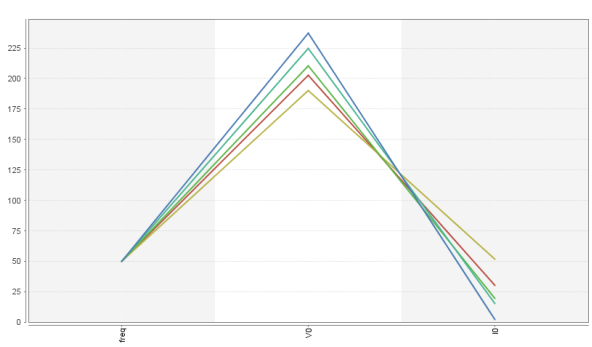
Fig. 7: Graph of data cluster on Phase 0

Cluster	freq	V0	I0
Cluster 0	49.948	237.071	2.045
Cluster 1	49.928	224.456	15.054
Cluster 2	49.908	210.482	19.536
Cluster 3	49.922	190.148	51.731
Cluster 4	49.916	202.784	29.856

(a) Centroid table of clusters



(b) Heat Map Cluster data visualization



(c) Centroid Chart of each cluster

Fig. 8: (a) (b) (c) Cluster Data Visualization of Phase-0

Fig. 8 (a) (b) (c) show cluster data visualization of Phase-0, figure 8 (a) shown the centroid table of Phase-0, in these results, rapidminer clusters data for 776264 record into 5 clusters based on the initial partition that was previous specified. Cluster 0 to Cluster 4 contains 3 attributes observations, those are for 3 type of attribute (Frequency (Hz), V0 (volt), I0 (ampere). We concluded that these final groupings are adequate for the data.

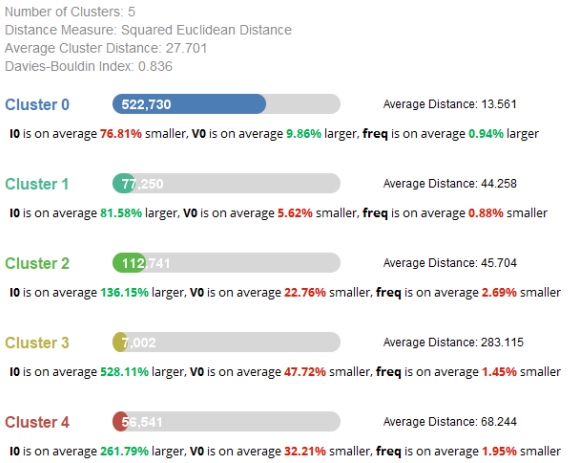


Fig. 9: Phase 0 data cluster overview

Fig. 9 show the data cluster overview with total Number of cluster=5, using Distance Measure Squared Euclidean Distance algorithm, with average cluster distance= 27.701 and Davies Bouldin index= 0.836, the following explanation detail is;

1. Cluster 0. Was the largest cluster with 522730 member data, has average distance 13.561 with I0 average= 76.81% smaller, V0 average=9.86% larger, freq average=0.94% larger.
2. Cluster 1. Has 77250 member data, and average distance=44.258 with I0 average=81.58% larger, V0 average 5.62% smaller, freq average=0.88 smaller.
3. Cluster 2. Has 112741 member data, and average distance=45.704 with I0 average=136.15% larger, V0 average 22.76% smaller, freq average=2.69 smaller.
4. Cluster 3. Has 7002 member data, and average distance=283.115 with I0 average=528.11% larger, V0 average 47.72% smaller, freq average=1.45% smaller.
5. Cluster 4. Has 54541 member data, and average distance=68.244 with I0 average=261.79% larger, V0 average 32.21% smaller, freq average=1.95% smaller.

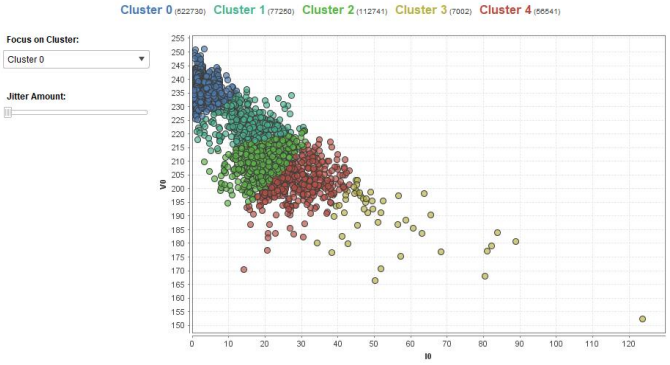


Fig. 10: Phase 0 data cluster scatter-plot visualization

Fig. 10 show Phase 0 data cluster scatter-plot visualization, from this chart can be concluded that the electrical quantities quality on this phase dominate with cluster 0 with centroid value is freq=49.948, V0=237.071, I0=2.045. While the data statistic of Phase 0 for 1 day shown on Fig 11

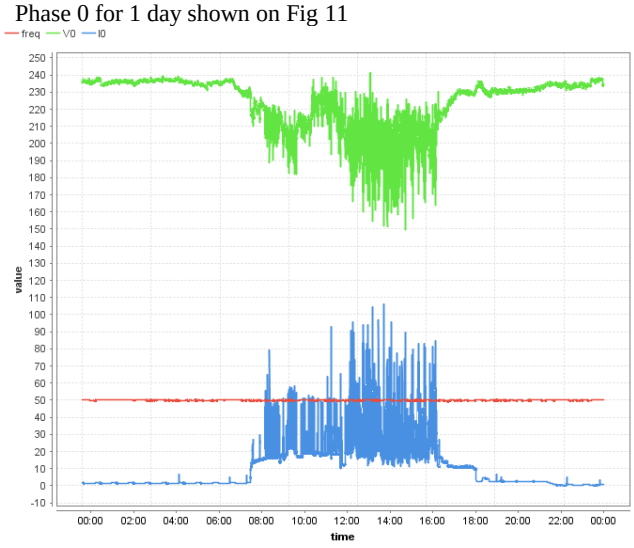


Fig. 11: Electrical Quantitites Measurement on Phase 0 for 1 day

Fig. 11 show electrical quantitites measurement on phase 0 for 1 day, start from 00.00 WIB until 23.59 WIB, from this chart can be concluded that the activity of electricity start significantly during working hours (08.00 WIB until 16.00 WIB), before working hours began, the voltage data shown very stable on range 220-230 V, with electric current below 2 Ampere, and entering the working hours, the Voltage just drop until to the lowest point at 150-160 V and electric current more than 60 A. Detail descriptive data statistik shown of table 1

Table 1: Descriptive Statistik of Electrical Quantities on Phase 0 for 1 day

	v0 (Volt)	i0 (Ampere)	pf0
Mean	224,711	10,106	0,768
Standard Error	0,113	0,096	0,002
Median	230,910	2,420	0,850
Mode	236,880	1,250	0,940
Standard Deviation	13,920	11,856	0,195
Sample Variance	193,762	140,555	0,038
Kurtosis	0,643	4,754	-0,028
Skewness	-1,172	1,715	-0,936
Range	91,430	105,660	0,750
Minimum	149,780	0,410	0,230
Maximum	241,210	106,070	0,980
Count	**15.152**	**15.152**	**15.152**

From the descriptive statistic on table 1, can conluded that the Mean of V0=224,711 Volt, I0=10,106 Ampere, pF0=0,76, While the Minimum value was V0=149,780 Volt, I0=0,410 Ampere, pF0=0,230, and the Maximum value was V0=241 Volt, I0=106,070 Ampere, pF0=0.980.

2. DM Modeling result on Phase 1

After executing the previous model build for data on Phase 1 using K-Means algorithm, rapidminer generate the cluster model of data consist of 5 cluster those are; Cluster 0: 529976 items, Cluster 1: 18090 items, Cluster 2: 122836 items, Cluster 3: 7 items, Cluster 4: 105355 items, with Total number of items: 776264, data visualization of cluster member shown on fig. 12.

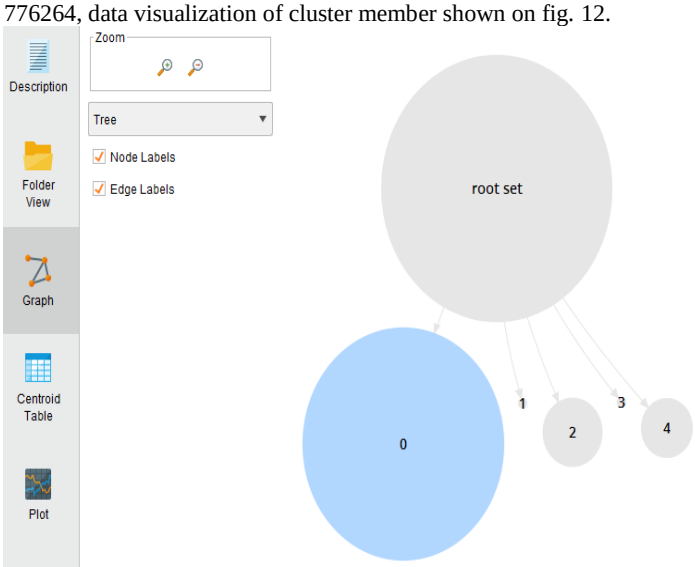
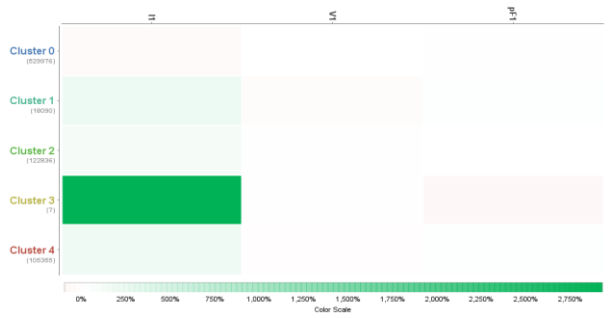


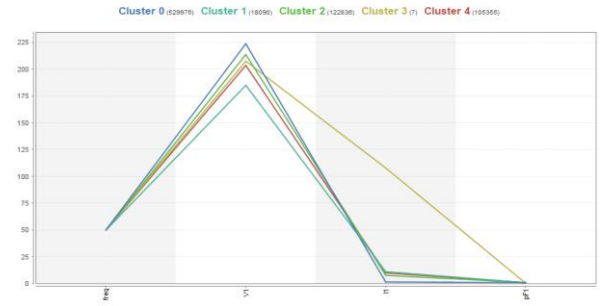
Fig. 12: Graph of data cluster on Phase 1

Cluster	freq	V1	I1	pF1
Cluster 0	49.949	223.584	1.225	0.705
Cluster 1	49.914	185.035	11.207	0.957
Cluster 2	49.925	213.792	7.803	0.909
Cluster 3	49.937	207.680	107.531	0.019
Cluster 4	49.904	203.589	9.980	0.957

(a) Centroid table of clusters



(a) Heat Map Cluster data visualization



(a) Centroid Chart of each cluster

Fig. 13: (a) (b) (c) Cluster Data Visualization of Phase-1

Fig. 13 (a) (b) (c) show cluster data visualization of Phase-1, figure 9 (a) shown the centroid table of Phase-1, in these results, rapidminer clusters data for 776264 record into 5 clusters based on the initial partition that was previous specified. Cluster 0 to Cluster 4 contains 3 attributes observations, those are for 3 type of attribute (Frequency (Hz), V0 (volt), I0 (ampere). We concluded that these final groupings are adequate for the data.

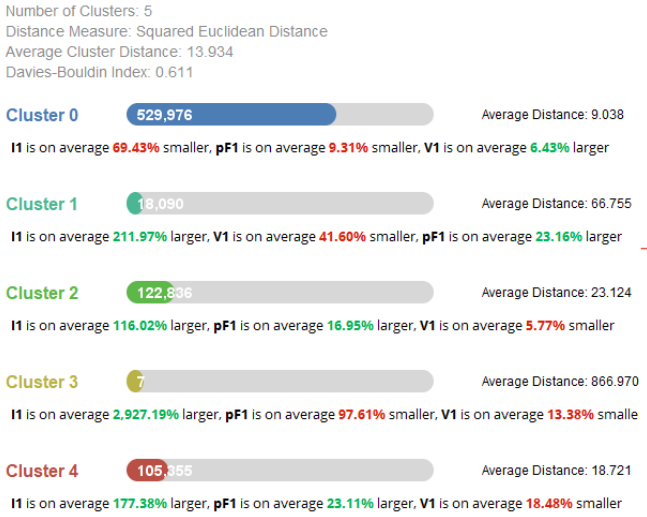
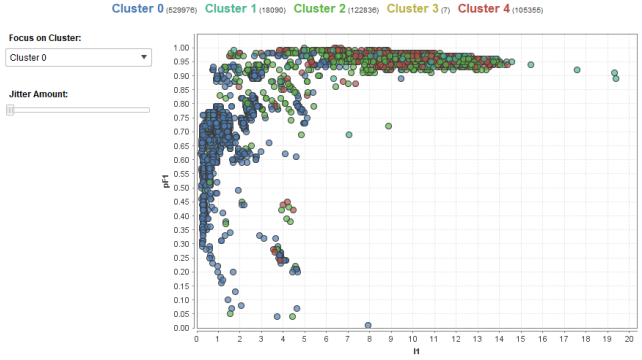


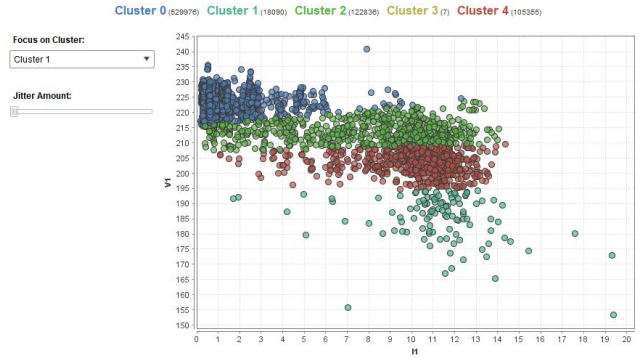
Fig. 14: Phase 1 data cluster overview

Fig. 14 show the data cluster overview with total Number of cluster=5, using Distance Measure Squared Euclidean Distance algorithm, with average cluster distance= 13.934 and Davies Bouldin index= 0.611, the following explanation detail is;

1. Cluster 0. Was the largest cluster with 529976 member data, has average distance 9.038 with I1 average= 69.43% smaller, pF1 average=9.31% smaller, V1 average=6.43% larger.
2. Cluster 1. Has 18090 member data, has average distance 66.755 with I1 average= 211.97% larger, pF1 average=23.16% larger, V1 average=41.60% smaller.
3. Cluster 2. Has 122836 member data, has average distance 23.124 with I1 average= 116.02% larger, pF1 average=16.95% larger, V1 average=5.77% smaller.
4. Cluster 3. Has 7 member data, has average distance 866.970 with I1 average= 2927%% larger, pF1 average=97.61% smaller, V1 average=13.38% smaller.
5. Cluster 4. Has 105355 member data, has average distance 18.721 with I1 average= 177.38%% larger, pF1 average=23.11% larger, V1 average=18.48% smaller.



(a) Scatter plot relation between pF1 and I1 on Phase 1



(b) Scatter plot relation between V1 and I1 on Phase 1

Fig. 15: Phase 1 data cluster scatter-plot visualization

Fig. 15 show Phase 1 data cluster scatter-plot visualization, from this chart can be concluded that the electrical quantities quality on this phase 1 dominate with cluster 0 with centroid value is freq=49.949, V1=223.584, I1=1.225, pF1=0.705. While the data statistic of Phase 0 for 1 day shown on Fig 16

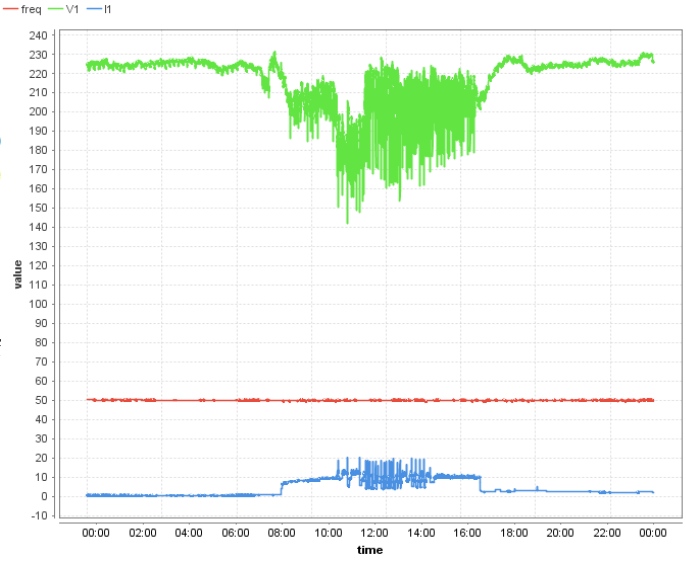


Fig. 16: Electrical Quantitites Measurement on Phase 1 for 1 day

Fig. 16 show electrical quantitites measurement on phase 1 for 1 day, start from 00.00 WIB until 23.59 WIB, from this chart can be concluded that the activity of electricity start significantly during working hours (08.00 WIB until 16.00 WIB), before working hours began, the voltage data shown very stable on range 220-230

V, with electric current below 2 Ampere, and entering the working hours, the Voltage just drop until to the lowest point at 190-200 V and electric current between 10-20 A. Detail descriptive data statistik shown of table 2

Table 2: Descriptive Statistik of Electrical Quantities on Phase 1 for 1 day

	V1 (Volt)	I1 (Ampere)	pf1
Mean	217,249	4,091	0,743
Standard Error	0,098	0,031	0,001
Median	223,370	2,560	0,770
Mode	225,750	0,570	0,560
Standard Deviation	12,124	3,877	0,172
Sample Variance	146,992	15,033	0,030
Kurtosis	2,838	-0,622	-1,581
Skewness	-1,678	0,848	0,070
Range	89,400	19,830	0,570
Minimum	141,900	0,540	0,420
Maximum	231,300	20,370	0,990
Count	**15.152**	**15.152**	**15.152**

From the descriptive statistic on table 2, can conluded that the Mean of V1=217,249 Volt, I1=4,091 Ampere, pF1=0,43, While the Minimum value was V1=141,900 Volt, I1=0,540 Ampere, pF1=0,420, and the Maximum value was V1=231 Volt, I1=20,370 Ampere, pF1=0.990.

3. DM Modeling result on Phase 2

After executing the previous model build for data on Phase 2 using K-Means algorithm, rapidminer generate the cluster model of data consist of 5, those are; Cluster 0: 291343 items, Cluster 1: 232413 items, Cluster 2: 88333 items Cluster 3: 49869 items, Cluster 4: 114306 items, with Total number of items: 776264, data visualization of cluster member shown on fig. 17.

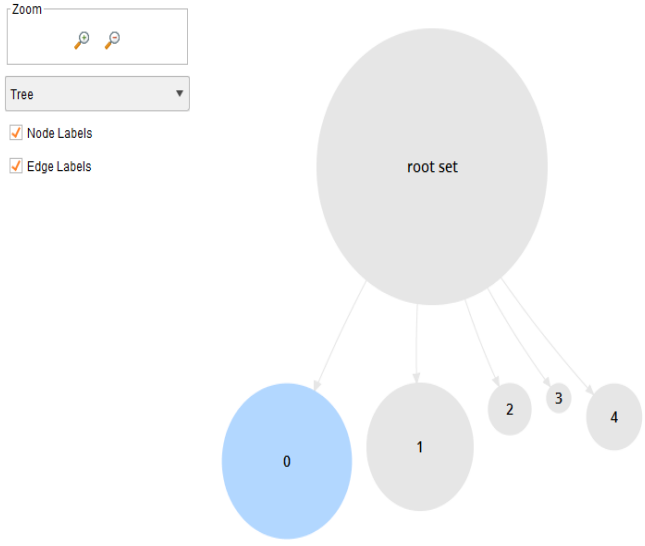
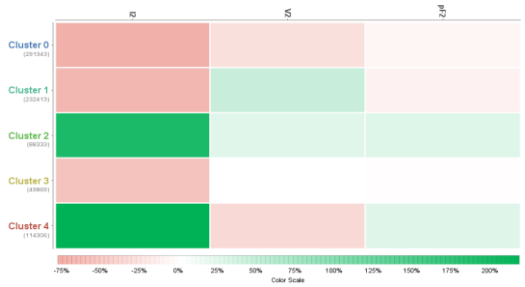


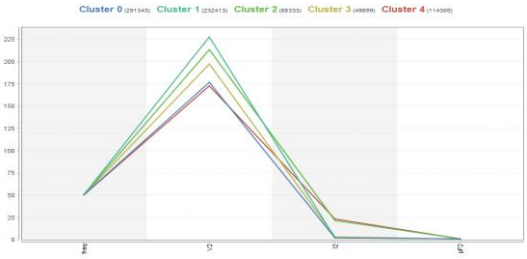
Fig. 17: Graph of data cluster on Phase 2

Cluster	freq	V2	I2	pF2
Cluster 0	49.935	176.780	1.769	0.697
Cluster 1	49.967	227.565	2.190	0.670
Cluster 2	49.913	213.701	21.706	0.982
Cluster 3	49.938	197.124	3.106	0.756
Cluster 4	49.906	172.552	23.410	0.976

(a) Centroid table of clusters



(b) Heat Map Cluster data visualization



(c) Centroid Chart of each cluster

Fig. 18: (a) (b) (c) Cluster Data Visualization of Phase-2

Fig. 18 (a) (b) (c) show cluster data visualization of Phase-2, figure 18 (a) shown the centroid table of Phase-2, in these results, rapidminer clusters data for 776264 record into 5 clusters based on the initial partition that was previous specified. Cluster 0 to Cluster 4 contains 3 attributes observations, those are for 3 type of attribute (Frequency (Hz), V0 (volt), I0 (ampere). We concluded that these final groupings are adequate for the data.

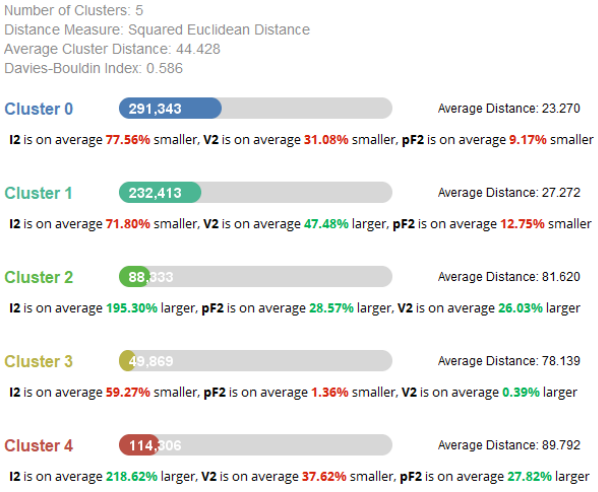
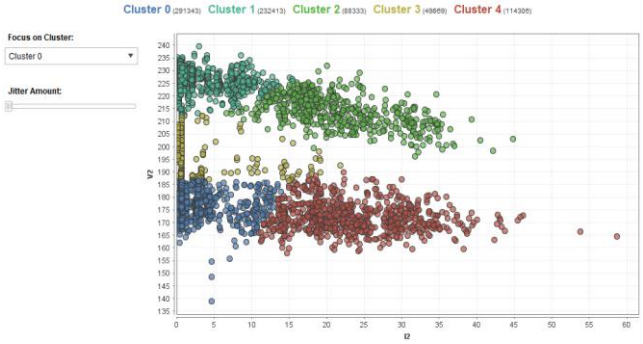


Fig. 19: Phase 2 data cluster overview

Fig. 19 show the data cluster overview with total Number of cluster=5, using Distance Measure Squared Euclidean Distance algorithm, with average cluster distance= 44.428 and Davies Bouldin index= 0.586, the following explanation detail is;

1. Cluster 0. Was the largest cluster with 291343 member data, has average distance 23.270 with I2 average= 77.56% smaller, V2 average=31.08% smaller, pF2 average=69.17% smaller.
2. Cluster 1. Has 232413 member data, has average distance 27.272 with I2 average= 71.80% smaller, V2 average=47.48% larger, pF2 average=12.75% smaller.
3. Cluster 2. Has 88333 member data, has average distance 81.620 with I2 average= 195.30% larger, V2 average=26.03% larger, pF2 average=28.57% larger.
4. Cluster 3. Has 49869 member data, has average distance 78.139 with I2 average= 59.27% smaller, V2 average=0.39% larger, pF2 average=1.36% smaller.

5. Cluster 4. Has 114306 member data, has average distance 89.792 with I2 average= 218.62% larger, V2 average=37.62% smaller, pF2 average=27.82% larger.



(a) Scatter plot relation between I2 and V2 on Phase 2

(b) Scatter plot relation between I2 and pF2 on Phase 2

Fig. 20: (a) (b) Phase-2 data cluster scatter-plot visualization

Fig. 20 show Phase 2 data cluster scatter-plot visualization, from this chart can be concluded that the electrical quantities quality on this phase 2 dominate with cluster 0 with centroid value is freq=49.935, V1=176.780, I1=1.769, pF1=0.697. While the data statistic of Phase 2 for 1 day shown on Fig 21

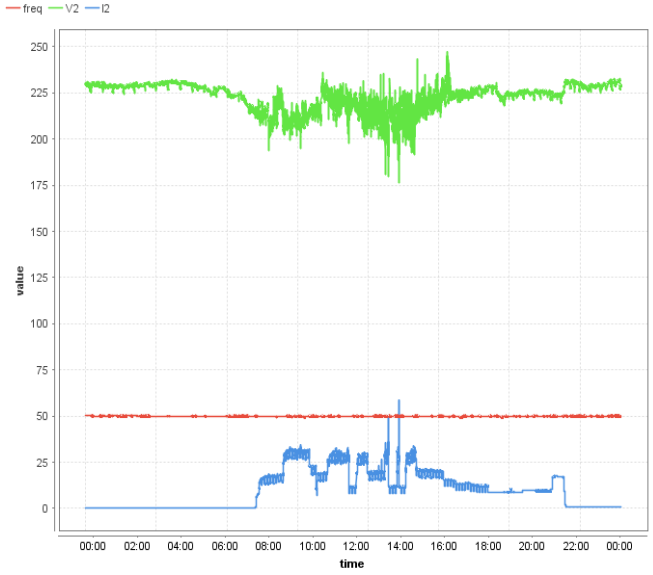


Fig. 21: Electrical Quantitites Measurement on Phase 2 for 1 day

Fig. 21 show electrical quantitites measurement on phase 1 for 1 day, start from 00.00 WIB until 23.59 WIB, from this chart can be concluded that the activity of electricity start significantly during working hours (08.00 WIB until 16.00 WIB), before working hours began, the voltage data shown very stable on range 220-225 V, with electric current below 2 Ampere, and entering the working hours, the Voltage just drop until to the lowest point at 200-210 V and electric current between 20-30 A. Detail descriptive data statistik shown of table 3

Table 3: Descriptive Statistik of Electrical Quantities on Phase 2 for 1 day

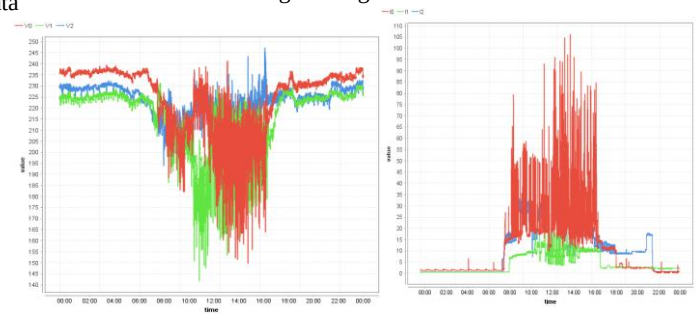
	V2 (Volt)	I2 (Ampere)	pF2
Mean	222,812	10,152	0,910
Standard Error	0,061	0,081	0,001
Median	225,020	8,900	0,980
Mode	229,050	0,530	0,990
Standard Deviation	7,449	9,977	0,116
Sample Variance	55,484	99,547	0,014
Kurtosis	0,432	-0,755	2,280
Skewness	-1,014	0,652	-1,776
Range	70,910	58,120	0,420
Minimum	176,230	0,510	0,580
Maximum	247,140	58,630	1,000
Count	**15.152**	**15.152**	**15.152**

From the descriptive statistic on table 3, can conluded that the Mean of V2=222,812 Volt, I2=10,152 Ampere, pF1=0,910, While the Minimum value was V2=176,230 Volt, I2=0,510 Ampere, pF1=0,580, and the Maximum value was V2=247,140 Volt, I2=56,630 Ampere, pF1=1.000.

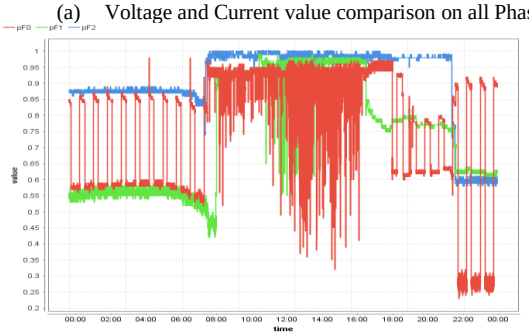
E. Evaluation

Rapidminer licensed for education was used on this research, it The Educational license of RapidMiner Studio provides unlimited data rows, a single logical processor, and includes premium features including RapidMiner Turbo Prep and Auto Model, the application run on environtment with Processor=Intel(R) Core (TM) i7-3632QM CPU @ 2.20 Ghz 2.20 Ghz, Installed memory= 16 GB, with Window 10 64-bit professional edition, 3 TB SSD storage. The application run with total 770847 record item dataset, and run with no problem.

The overall results of the electrical quantities measurement with data mining are fairly easy to communicate from a business perspective: the research produced what are hoped to be better electricity policy recommendations and an improved quality on Departement of Electrical Engineering.



(a) Voltage and Current value comparison on all Phase



(b) Power Factor value on all Phase

Fig. 22: Corelation data between V, I, Pf on All Phase for 1 day

Figure 22 shown the voltage, current, and power factor comparation on all Phase for 1 day, from the data can concluded that on Phase 1 has most stable voltage, while Phase 0 was the most high load electricity consumption indicated by a large current value, form whole graphic indicated that the load on each Phase was unbalanced.

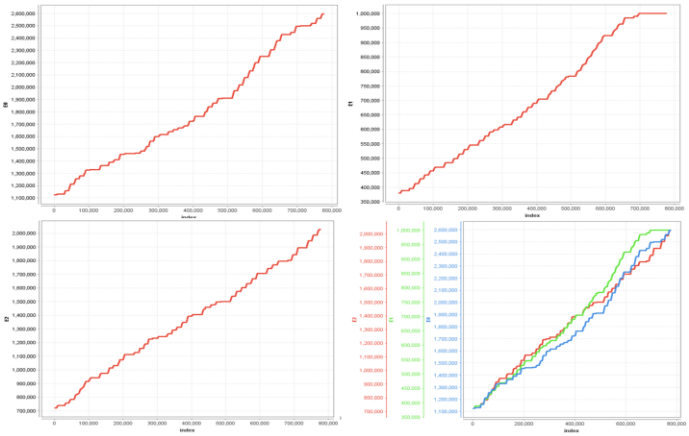


Fig. 23: Energy consumption comparation on all Phase

Figure 23 shown the Energy consumption comparation on all Phase during 2 month data monitoring (September-Oktober 2018), from the chart information, concluded that at Phase 0 was the most large energy consumption value, appropriate to the high current value in this Phase.

F. Deployment

At the CRISP-DM Deployment Phase, is the process of using the new insights of the electrical quantities data pattern founded during the research to make improvement within organization. Based on the data analysis, it shown that the voltage, current, and power factor comparation on all Phase was unbalanced, already reported to the stake-holders and recommended them to conduct a total evaluation of the use of electrical devices so that electricity loads on each phase can be balanced, especially the transfer of electrical loads in Phase 0 that are too large compared to Phases 1 and 2.

5. Acknowledgement

We Would like to thank to KEMENRISTEK DIKTI for a part of financial support of the research through the Hibah Terapan gran

6. Conclusion

Main objective of this research was to capture, analyze, and identified the knowledge pattern of electrical quantities data measurements, using Cross-Industry Standard Process for Data Mining (CRISP-DM) data mining framework, for helping the stake holders to continuous improvement of the quality of electricity services, the initial research limited to total 770847 electrical quantities recorded data that save on database system, since 1 September – 31 October 2018, the dataset consist of 21 attribute electrical quantities such as; voltage, current, POWer factor values, energy consumption, frequency, on H building 3-Phase main control. Based on the data analysis, it shown that the voltage, current, and power factor comparation on all Phase was unbalanced, already reported to the stake-holders and recommended them to conduct a total evaluation of the use of electrical devices so that electricity loads on each phase can be balanced, especially the transfer of electrical loads in Phase 0 that are too large compared to Phases 1 and 2.

References

[1] P.Gregor. Gartner 2018 Magic Quadrant for Advanced Analytics Platforms: who gained and who lost.

[2] Introduction To Rapidminer, https://rapidminer.com/us/, 2018.

[3] W. A. Al-Dhuraibi, J. Ali, Using Classification Techniques to Predict Gold Price Movement, 4th International Conference on Computer and Technology Applications, 2018.

[4] J. Estrada, L. Vea, Sitting Posture Recognition for Computer Users using Smartphones and a Web Camera, Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.

[5] M. A. Alhaj, A. Y. A. Maghari, Cancer Survivability Prediction using Random Forest and Rule Induction Algorithms, 8th International Conference on Information Technology (ICIT), 2017.

[6] A. Geetha, G. M. Nasira, Data Mining for Meteorological Applications: Decision Trees for Modeling Rainfall Prediction, IEEE International Conference on Computational Intelligence and Computing Research, 2014.

[7] Rianto, L. E. Nugroho, P. I. Santosa, Pattern Discovery of Indonesian Customers in an Online Shop: A Case of Fashion Online Shop, Proc. of 2016 3rd Int. Conf. on Information Tech., Computer, and Electrical Engineering (ICITACEE), Oct 19-21st, Semarang, Indonesia., 2016.

[8] M. R. Mahmud, M. A. Mamun, M. A. Hossain, M. P. Uddin, Comparative Analysis of K-Means and Bisecting K-Means Algorithms for Brain Tumor Detection, International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018.

[9] M. S. Rahim, T. Ahmed, An initial centroid selection method based on radial and angular coordinates for K-means algorithm, 20th International Conference of Computer and Information Technology (ICCIT), 2017.

[10] M. A. Altuncu, B. Türkoğlu, M. A. Çavuşlu, S. SahIn, Implementation of K-means algorithm on FGGA, 26th Signal Processing and Communications Applications Conference (SIU), 2018.

[11] P. Manivannan, P. Isakki Devi, Dengue fever prediction using K-means clustering algorithm, IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017.

[12] UC Business Analytics R Programming Guide - K-mean Algorithm. https://uc-r.github.io/kmeans_clustering, 2018.

[13] P. Kalgotra, R. Sharda, Progression analysis of signals: Extending CRISP-DM to stream analytics, IEEE International Conference on Big Data (Big Data), 2016.

[14] F. Chiheb, F. Boumahdi, H. Bouarfa, D. Boukraa, Predicting students performance using decision trees: Case of an Algerian University, International Conference on Mathematics and Information Technology (ICMIT), 2017

[15] L. C. Chinchilla, K. A. R. Ferreira, Analysis of the behavior of customers in the social networks using data mining techniques, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016.

[16] Z. Hou, Data Mining Method and Empirical Research for Extension Architecture Design, International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2018.

[17] C.Shearer. The CRISP-DM Model: The New Blueprint for Data Mining, Journal of Data Warehousing, Volume 5, Number 4, pag. 13-22, 2000.

[18] D. Despa, G. F. Nama, M. A. Muhammad, K. Anwar, The Implementation Internet of Things (IoT) Technology in Real Time Monitoring of Electrical Quantities, The 2nd International Conference on Mathematics, Science, Education and Technology, 5–6 October, Padang, West Sumatera, Indonesia, 2017.

[19] G. F. Nama, D. Despa, Mardiana, Real-time monitoring system of electrical quantities on ICT Centre building University of Lampung based on Embedded Single Board Computer BCM2835, International Conference on Informatics and Computing (ICIC), 2016.

[20] G. F. Nama, K. Muludi, Implementation of Two-Factor Authentication (2FA) to Enhance the Security of Academic Information System, Journal of Engineering and Applied Sciences 13 (8), 2209-2220, 2018.

[21] G. F. Nama, G. I. Suhada, A. Zaenudin, Smart System Monitoring of Gradient Soil Temperature at the Anak Krakatoa Volcano, Asian Journal of Information Technology 16 (2), 337-347, 2017.

[22] D. Despa, F. X. A. Setyawan, G. F. Nama, J. Delano, Artificial Neural Network Applications Use Measurements Of Electrical Quantities To Estimate Electric Power, International Conference on Engineering, Technologies, and Applied Sciences (ICETsAS 2018), 2018.

23% Overall Similarity

Top sources found in the following databases:

- 14% Internet database
- Crossref database
- 17% Submitted Works database
- 14% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Peter Michalik, Simona Polackova, Iveta Zolotova. "Analysis of data fro... Crossref	3%
2	mafiadoc.com Internet	3%
3	Diego Buenano Fernandez, Sergio Lujan-Mora. "Comparison of applica... Crossref	2%
4	University of Greenwich on 2022-06-30 Submitted works	2%
5	bookdown.org Internet	2%
6	docplayer.net Internet	2%
7	ijsrset.com Internet	1%
8	ibm.com Internet	1%

9	Universitas Brawijaya on 2017-10-11 Submitted works	<1%
10	sthda.com Internet	<1%
11	Snyder, Ryan. "Assessment of Physical and Technical Indicators in the ... Publication	<1%
12	edas.info Internet	<1%
13	muetjamshoro on 2020-10-26 Submitted works	<1%
14	Artoto Arkundato, Fiber Monado, Zaki Su'ud. "Effect of temperature on ... Crossref	<1%
15	Ajay Mandal, Jitendar Kumar Tiwari, Bandar AlMangour, N. Sathish, Su... Crossref	<1%
16	Gigih Forda Nama, Fadillah Halim Rasyidy, Raden Arum S P., Mardiana ... Crossref	<1%
17	Pritika Reddy, Kaylash Chaudhary, Bibhya Sharma, Ronil Chand. " Digit... Crossref	<1%
18	Fangyan Zhang, Hao Liang. "Data Mining Based Analysis Method for C... Crossref	<1%
19	Duan, R.C., F.H. Wang, J. Zhang, R.H. Huang, and X. Zhang. "Data minin... Crossref	<1%
20	N Iswandhani, M Muhajir. "K-means cluster analysis of tourist destinati... Crossref	<1%

21	University of Sheffield on 2018-05-23 Submitted works	<1%
22	Colorado State University, Global Campus on 2019-12-16 Submitted works	<1%
23	irjabs.com Internet	<1%
24	"Software Engineering Trends and Techniques in Intelligent Systems", ... Crossref	<1%