

## Behavior Analysis of Mobile Phone User Customers

Nopphorn Somrit<sup>1</sup> and Chumpol Bunkhumpornpat<sup>2</sup>

<sup>1</sup> Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand

<sup>2</sup> Electrical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand  
nopphorn\_s@cmu.ac.th

**Abstract.** This independent study is to develop a system to analyze the behavior of customers who use mobile phones by using the Clustering model as a tool to group customers according to their behavior. The researcher conducted a comparative study of appropriate behavioral grouping. It was divided into two sub-studies to study clustering by constructing two basic models: K-means and DBSCAN from those two basic models to find the most suitable clustering method for the data set characteristics. This comparison of the average accuracy of classification of customers in each group from all five models: Random Forest, Decision Tree, SVM, Naïve Bayes, and KNN. Performance measurement of the system developed in this study. It is a comparison of accuracy found that K-means clustering has better customer classification efficiency than DBSCAN. From the experimental results, it can be said that the K-means model has a higher mean accuracy of five classification models than DBSCAN.

**Keywords:** Clustering, Classification

### 1. Introduction

Wearesocial's annual digital behavior survey found that digital usage has found interest in digital usage among people across the world. Maybe because of the COVID-19 period. forced to work Sales and meetings via digital are tightly packed. And people want to go back to their pre-pandemic lifestyle. But the use of social media, and buying Ads by marketers, brands or pages continues to grow as before. because people are familiar with and want to reinforce brand visibility more Looking specifically at mobile behavior in 2022, there are over 5.31 billion mobile phone users globally or 67.1% of the world's population. which increased by +1.8% or more than 95 million people compared to the previous year interestingly, the number of mobile phones with active connections is higher than 8.28 billion. which grew from the previous year to 2.9%, or a figure of 233 million.

Therefore, big data is difficult to identify and analyze customer types. It can be in an unstructured or semi-structured format. This type of data is called big data. Research shows that big data has many uses. For example: Use data to attract and retain customers. Customers are the most important asset you should pay attention to. No enterprise can succeed without a strong customer base. However, even

if you have a strong customer base, however, if you neglect to study what customers really want, it's easy to put forward what they don't want. Eventually, you will lose customers, which will hinder your path to success. Use big data Help your enterprise better observe customer patterns and trends by easily collecting all customer information needed. This means that it is easier to understand customers in the digital era. Through data analysis and customer behavior observation mechanisms, enterprises can obtain in-depth customer behavior data. This is crucial to maintain the customer base of the enterprise. Understanding customer insight will enable the enterprise to meet customer needs. This is the most basic step to achieving the goal of customer care and is also the key to establishing enterprise or brand loyalty.

## **2. Literature Review**

### **2.1. Clustering method**

DBSCAN (Density-Based Spatial Clustering of Applications with Noise: DBSCAN groups data points together based on the distance metric. It follows the criterion for a minimum number of data points. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers. It takes two parameters – eps and minimum points. Eps indicates how close the data points should be to be considered neighbors. The criterion for minimum points should be completed to consider that region as a dense region.

K-Means Clustering: K-Means clustering is one of the most widely used algorithms. It partitions the data points into k clusters based on the distance metric used for the clustering. The value of 'k' is to be defined by the user. The distance is calculated between the data points and the centroids of the clusters. The data point which is closest to the centroid of the cluster gets assigned to that cluster. After an iteration, it computes the centroids of those clusters again and the process continues until a pre-defined number of iterations are completed or when the centroids of the clusters do not change after an iteration. It is a very computationally expensive algorithm as it computes the distance of every data point with the centroids of all the clusters at each iteration. This makes it difficult for implementing the same for huge data sets. [1] [2]

### **2.2. Classification method**

Classification is a supervised machine learning approach, in which the algorithm learns from the data input provided to it and then uses this learning to classify new observations. In other words, the training dataset is employed to obtain better boundary conditions that can be used to determine each target class; once such boundary conditions are determined, the next task is to predict the target class.

Binary classifiers work with only two classes or possible outcomes (for example positive or negative sentiment; whether the lender will pay the loan or not; etc), and Multiclass classifiers work

with multiple classes (ex: to which country a flag belongs, whether an image is an apple or banana or orange; etc). Multiclass assumes that each sample is assigned to one and only one label.

**Naive Bayes:** The algorithm is a simple algorithm to implement and usually represents a reasonable method to kickstart classification efforts. It can easily scale to larger datasets (takes linear time versus iterative approximation, as used for many other types of classifiers, which is more expensive in terms of computation resources) and requires a small amount of training data. However, Naive Bayes can suffer from a problem known as a 'zero probability problem', when the conditional probability is zero for a particular attribute, failing to provide a valid prediction. One solution is to leverage a smoothing procedure (ex: Laplace method).

**Decision Trees:** Decision Trees are in general simple to understand and visualize, requiring little data prep. This method can also handle both numerical and categorical data. On the other hand, complex trees do not generalize well ("overfitting"), and decision trees can be somewhat unstable because small variations in the data might result in a completely different tree being generated.

**Random Forest:** essentially a "meta-estimator" that fits a number of decision trees on various sub-samples of datasets and uses an average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is the same as the original input sample size but the samples are drawn with replacement. Random Forests tend to exhibit a higher degree of robustness to overfitting (>robustness to noise in data), with efficient execution time even in larger datasets. They are more sensible however to unbalanced datasets, being also a bit more complex to interpret and requiring more computational resources.

**kNN:** kNN (k Nearest Neighbors) is also often used for classification problems. kNN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). It has been used in statistical estimation and pattern recognition already at the beginning of 1970s as a non-parametric technique.

**SVM:** Support Vector Machine is a supervised classification algorithm where we draw a line between two different categories to differentiate between them. SVM is also known as the support vector network. [3] [4]

### **3. Data and Methodology**

#### **3.1. Data**

The data source is the Kaggle platform, which is a telecommunication company's dataset with detailed data. The dataset contains 3,333 records and a total of 21 features. 21 features are Vmail Message, Day Mins, Eve Mins, Night Mins, Intl Mins, CustServ Calls, Day Calls, Day Charge, Eve

Calls, Eve Charge, Night Calls, Night Charge, Intl Calls, Intl Charge, Area Code, Phone, Account Length, Churn, Intl Plan, Vmail Plan, State. (Figure 1).

Row ID	VMail M...	Day Mins	Eve Mins	Night Mins	Intl Mins	ClustCo...	Day Calls	Day Ch...	Eve Calls	Eve Ch...	Night C...	Night C...	Intl Calls	Intl Ch...	Area C...	Phone	Account...
Row1	25	265.1	197.4	244.7	10	1	110	45.07	99	16.78	91	11.01	3	2.7	415	382-4657	128
Row2	26	181.6	195.5	254.4	13.7	1	123	27.47	103	16.62	103	11.45	3	3.7	415	371-7291	107
Row3	0	293.4	121.2	162.6	10.2	0	114	49.38	110	10.3	104	7.35	5	3.26	415	398-1921	137
Row4	0	299.4	61.9	196.9	6.6	2	71	50.9	88	5.26	89	8.86	7	1.78	408	175-9999	84
Row5	0	166.7	148.3	188.9	10.1	0	113	28.34	122	12.61	121	8.41	3	2.73	415	330-6626	75
Row6	0	221.4	220.6	203.9	6.3	0	98	37.98	101	18.76	118	9.18	6	1.7	510	391-8027	118
Row7	24	218.2	348.5	212.6	7.5	3	88	37.09	108	26.42	118	9.57	7	2.03	510	355-9993	121
Row8	0	117	163.1	211.8	7.1	0	99	26.69	94	8.76	96	9.53	6	1.92	415	229-9011	147
Row9	0	184.5	251.6	215.8	8.7	1	97	31.27	80	26.89	90	9.71	4	2.35	408	336-4718	117
Row10	37	258.6	222	326.4	11.2	0	84	43.96	111	18.87	97	14.69	5	3.02	415	330-8173	141
Row11	0	129.1	238.5	208.8	12.7	4	137	21.95	83	19.42	111	9.4	6	3.43	415	329-6623	95
Row12	0	187.7	163.4	196	9.1	0	127	31.91	148	13.89	94	8.82	5	2.46	415	344-9403	74
Row13	0	138.8	104.9	141.1	11.2	1	96	21.9	71	8.92	128	6.35	2	3.02	408	363-1107	168
Row14	0	126.6	247.6	192.3	12.3	3	88	26.62	79	21.03	115	8.65	5	3.32	510	394-8006	95
Row15	0	120.7	307.2	203	13.1	4	70	30.52	76	26.11	99	9.14	6	3.54	415	366-9238	62
Row16	0	332.9	317.8	180.6	5.4	4	67	36.59	97	27.01	128	7.23	9	1.46	415	351-7289	161
Row17	27	196.4	281.9	89.3	13.8	1	139	31.39	80	23.88	75	4.02	4	3.73	408	359-8884	85
Row18	0	190.7	218.2	129.6	8.1	3	114	32.42	111	18.55	121	5.83	3	2.19	510	386-2923	93
Row19	33	189.7	212.8	165.7	10	1	66	32.25	85	18.09	108	7.46	5	2.7	510	396-2992	76
Row20	0	224.4	129.5	122.8	1.3	1	90	28.15	88	13.56	74	6.68	2	3.51	415	373-2762	73
Row21	0	155.1	239.7	208.8	10.6	0	117	26.37	93	20.37	133	8.4	4	2.86	415	396-5800	147
Row22	0	61.4	169.9	209.6	1.7	3	89	10.61	121	14.44	84	9.43	6	1.54	408	393-7984	77
Row23	0	183	72.9	181.8	9.5	0	112	31.11	99	6.7	78	8.18	19	4.15	415	398-1958	120
Row24	0	120.4	137.3	189.6	7.7	2	103	18.77	102	11.67	105	8.53	6	2.08	415	350-2565	111
Row25	0	81.1	246.2	227	10.3	0	86	13.79	72	20.84	115	10.67	2	2.78	510	343-4696	132
Row26	0	124.3	279.1	250.7	15.5	3	76	31.31	112	21.55	115	11.28	5	4.19	415	321-3698	174
Row27	39	213	191.1	182.7	9.5	0	115	36.21	112	16.24	115	8.22	3	2.57	408	357-3817	57
Row28	0	134.3	155.5	102.1	14.7	2	73	22.83	100	13.22	68	4.59	4	3.97	408	418-4412	54
Row29	0	190	258.2	181.5	6.3	0	109	32.3	84	21.95	102	8.17	6	1.7	415	353-9030	20
Row30	0	129.3	215.1	178.7	11.1	1	117	20.28	109	18.28	90	8.04	1	3	510	410-7789	49
Row31	0	84.8	126.7	250.5	14.2	2	85	14.42	83	11.62	148	11.27	6	3.83	415	416-4208	143
Row32	0	226.1	201.5	246.2	10.3	1	105	38.44	107	17.13	98	11.08	5	2.78	510	370-3359	75
Row33	0	212	31.2	293.3	12.6	3	121	26.04	115	2.65	78	13.2	10	3.4	408	383-1121	172
Row34	25	248.6	242.4	280.2	11.8	1	118	42.43	119	21.45	90	12.61	3	3.19	408	360-1595	12
Row35	37	176.8	195	213.5	8.3	0	94	30.06	75	16.58	116	9.61	4	2.24	408	395-2854	57
Row36	30	220	217.3	152.8	14.7	3	80	27.4	102	18.47	71	6.88	6	3.97	415	362-1407	72
Row37	0	146.3	162.5	129.3	14.5	0	138	24.87	80	13.81	109	5.82	6	3.92	408	241-6764	36
Row38	33	130.8	223.7	227.8	10	1	64	22.24	116	19.01	108	10.25	5	2.7	415	353-3305	78
Row39	33	203.9	187.6	191.7	10.5	3	106	24.66	99	15.95	107	4.39	6	2.84	415	402-1381	136
Row40	0	126.3	271.8	188.3	11.1	1	94	23.87	92	23.1	108	8.47	9	3	408	332-8911	149
Row41	0	126.3	166.8	187.8	9.4	3	102	21.47	85	14.18	135	8.45	2	2.54	408	372-9976	98
Row42	41	173.1	203.9	122.2	14.6	0	85	20.43	107	17.33	78	5.5	15	3.94	408	383-6029	135
Row43	0	124.8	282.2	313.5	10	2	82	31.22	98	23.99	78	14.02	4	2.7	510	393-7898	34
Row44	0	85.8	165.3	178.5	9.2	3	77	14.59	110	14.05	92	8.03	4	2.48	415	390-7274	160
Row45	0	154	225.8	265.3	15.5	1	67	26.38	118	19.39	86	11.94	3	0.95	510	352-1237	84
Row46	38	120.9	213	163.1	8.5	2	87	20.15	92	18.11	116	7.24	5	2.1	408	353-3661	99

Fig. 1. Data Display

Data cleaning, data cleansing, or data scrubbing is the act of first identifying any issues or bad data, then systematically correcting these issues. If the data is unfixable, you will need to remove the bad elements to properly clean the data. Unclean data normally comes as a result of human error, scraping data, or combining data from multiple sources. Multichannel data is now the norm, so inconsistencies across different data sets are to be expected. In this dataset, there is little data cleaning. changing the type of data in each column appropriately and normalizing This keeps the data at the same distance and makes the model work more efficiently.

### 3.2. Methodology

To analyze customer phone usage behavior, want to bring different types of user behavior of customers to study by grouping the data by using the past data to create a clustering model for K-means and DBSCAN, then measure the results by using the results of each model to classify into 5 models with SVM, Naïve Bayes, Random Forest, KNN, and Decision Tree. The final step is to average the classification accuracy of each model to conclude that the analyzed data are suitable for stratification.

### 3.3. Data Preparation

First of all, in data preparation, it is necessary to check the data type to be studied. If incorrect data types are found, first change the data type and filter the remaining data as unnecessary data, so as to be more convenient and quick in the experiment. Then, it is necessary to check for any missing values in the dataset to be studied in order to effectively analyze the model, so as to analyze the model effectively. Before creating the model, the range of data is normalized to the same range [0-1] through normalization. (Figure 2).

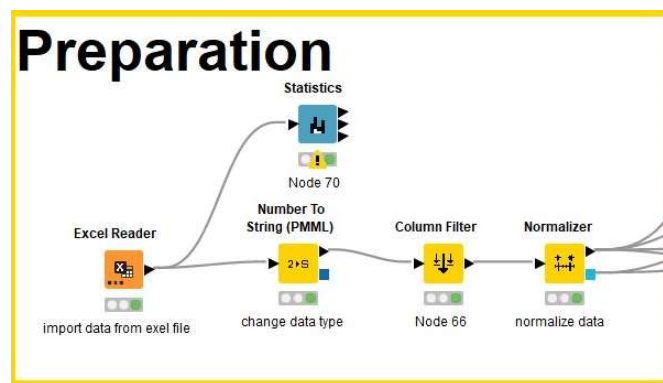


Fig. 2. Preparation flow

### 3.4. Find the optimal number of k with the Silhouette method

Before cleaned data can be grouped with a clustering model, the optimal K value must be determined. In this research, the Silhouette method was used. To determine the K value by this method, PCA was required (Figure 3) to reduce the dimension of the data to a two-dimensional form first. Then find the K value and display the result with a line graph. (Figure 4).

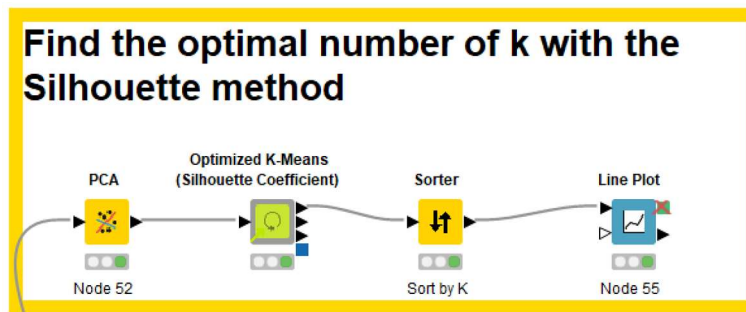
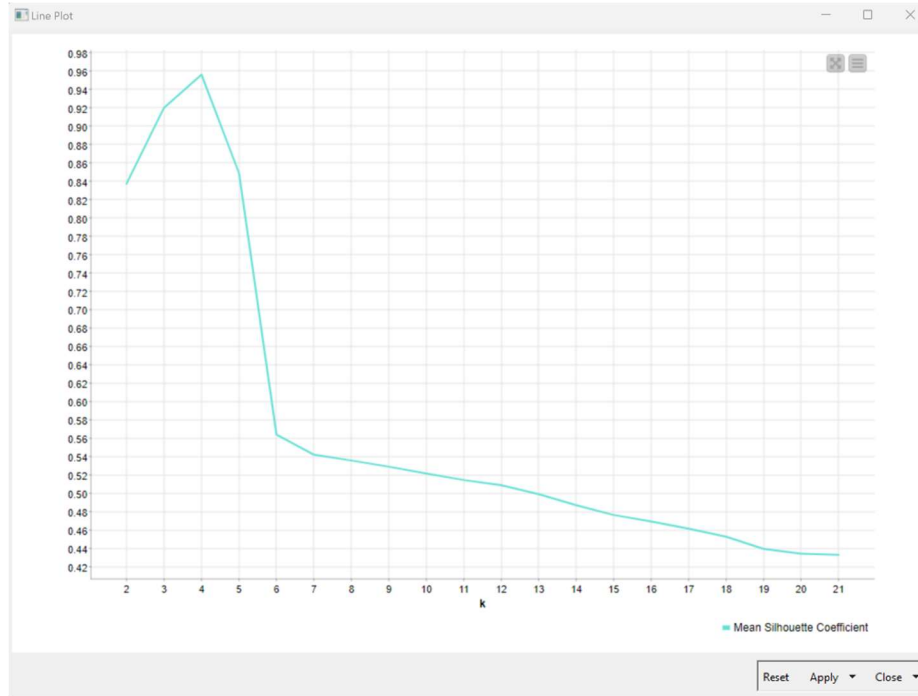


Fig. 2. Silhouette method



**Fig. 4.** Silhouette method

From this method, when plotting the results on a line graph, it was found that the appropriate K value for this data set is  $K = 4$ . Therefore, in the next process, K is set at 4. [5] [2]

### 3.5. Clustering

After finding the number of suitable K values for the dataset equal to four Further groupings are achieved by taking the data after normalization. used to create a clustering model. Here we will compare two clustering models, K-means and DBSCAN, which are used only because of their Hierarchical fit to the data with a small number and easy to read tree diagram. (Figure 5)

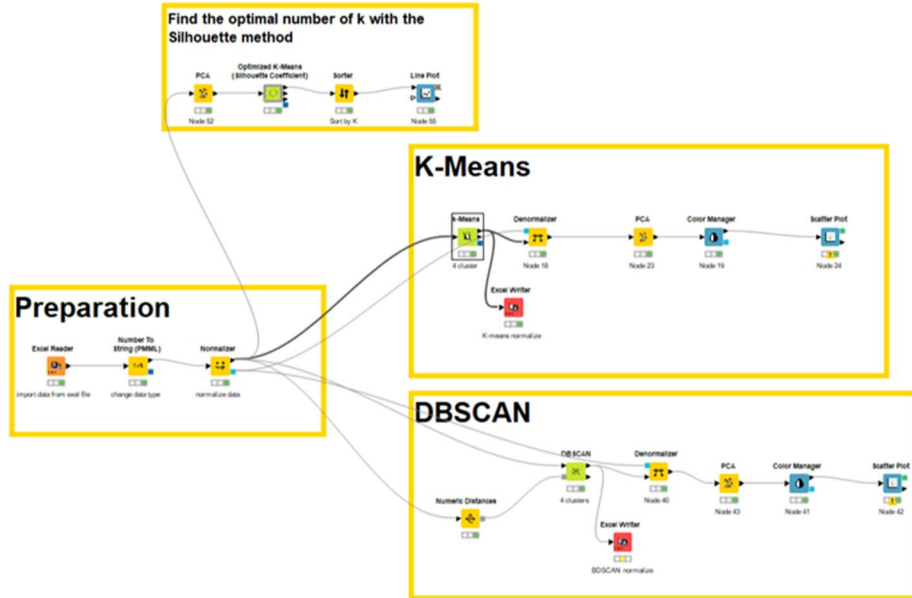


Fig. 5. Clustering workflow

When the usage behavior is in the range of 0 to 1 from the experiment above Next, use the above data to create a model that we set. And after creating K-means and DBSCAN, K-means must define the data to be grouped and the number of clusters (K value). (Figure 6) DBSCAN has to be defined data to be grouped the same as K-means and also needs to configure Epsilon which in this case is set to 1.0. (Figure 7)

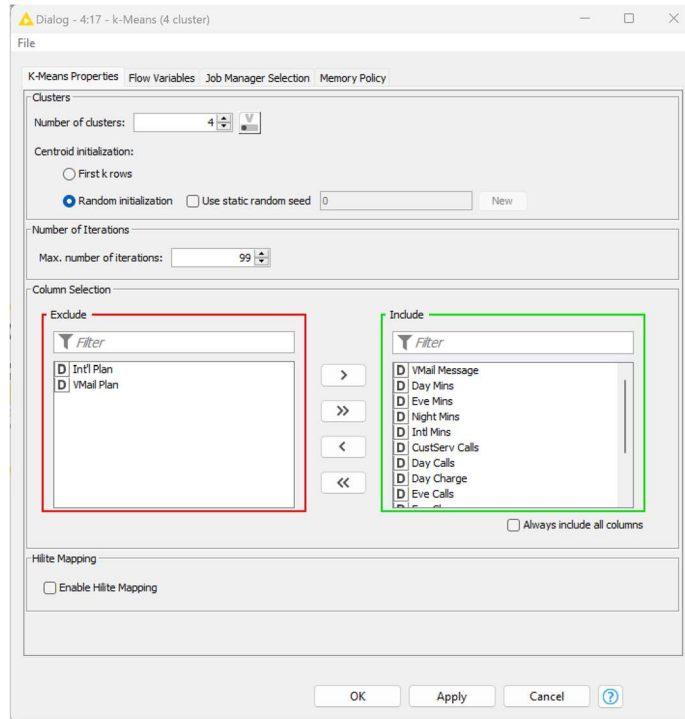


Fig. 6. Configuration of K-means

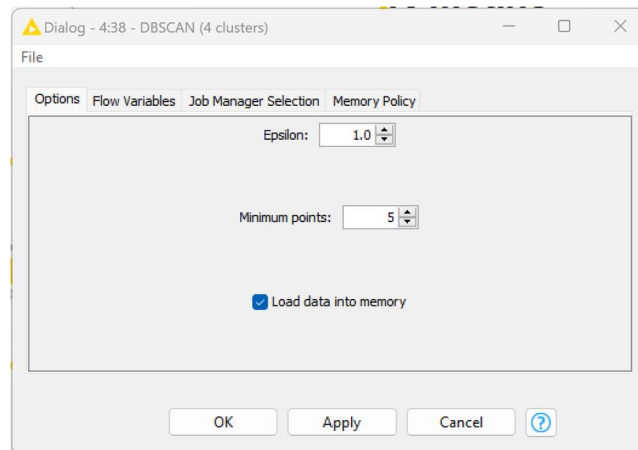
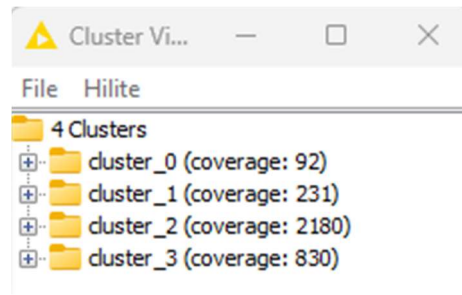


Fig. 7. Configuration of DBSCAN



After executing both models, the model divides the data into four clusters. The result of both models is that K-means has cluster 0 = 92 data, cluster 1 = 231 data, cluster 2 = 2180 data, and cluster 3 = 830 data. (Figure 8)



**Fig. 8.** Number in each cluster of K-means

The DBSCAN model can be divided into groups as cluster 0 = 92 data, cluster 1 = 2180 data, cluster 2 = 231 data, cluster 3 = 830 data, and noise = 0. (Figure 9)

Row ID	Count
Noise	0
Cluster_0	92
Cluster_1	2180
Cluster_2	231
Cluster_3	830

**Fig. 9.** Number in each cluster of DBSCAN

### 3.6. Classification

After the analysis results of each clustering model are released. The next step will be to determine which model results are accurate and appropriate for the data set used in this study. The way to measure the accuracy of the stratification is to bring out the results for analysis by the method. classification It will use the cluster results that come out as labels. And the value of customer usage behavior is used to make predictions. A total of five models, namely Random Forest, Decision Tree, SVM, Naive Bayes, and KNN, were used to make predictions, after which the accuracy of all five models was averaged to determine that the highest mean was obtained from K-means or DBSCAN. (Figure 10,11) [6]. In order to make the distribution of the dataset similar, a 10-fold validation technique was used to divide the data into 10 equal parts to create and test the model. Calculate average accuracy and errors before using the model to predict the test set. [7] [8]

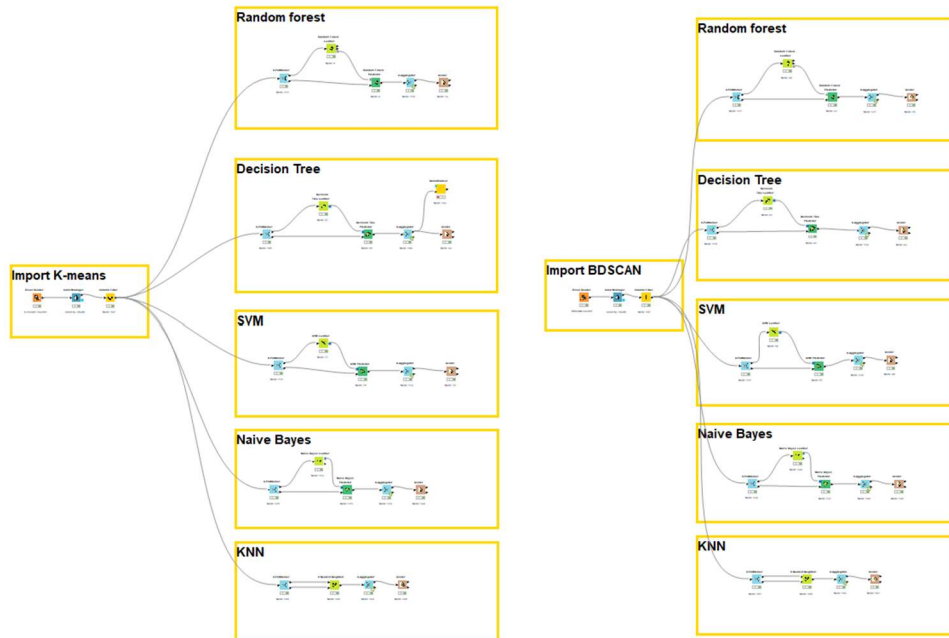
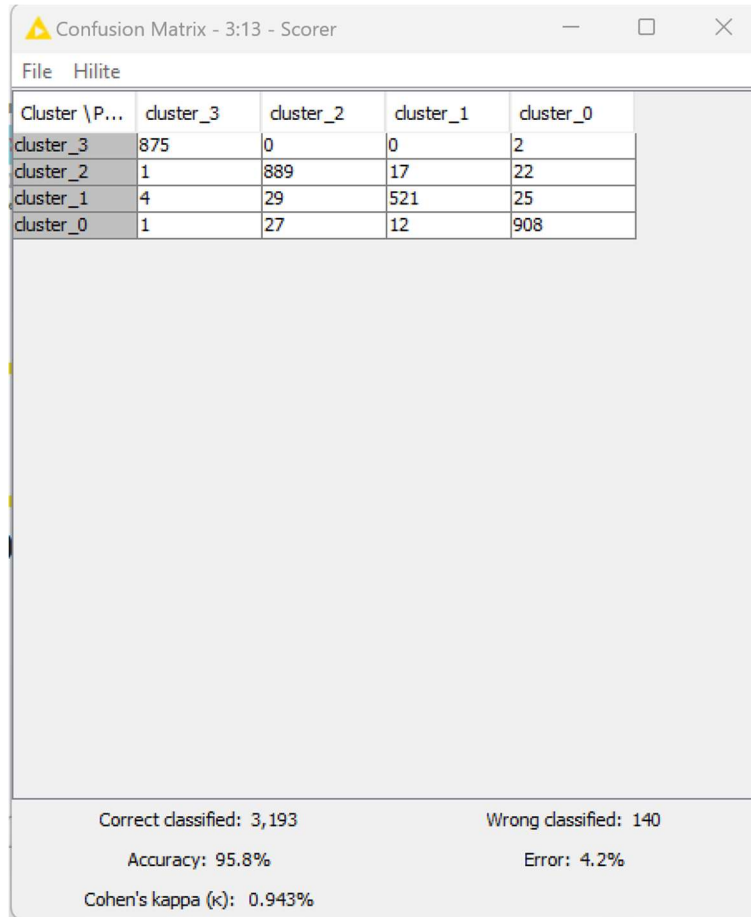


Fig. 10. Classification workflow



The screenshot shows a window titled "Confusion Matrix - 3:13 - Scorer" with a menu bar containing "File" and "Hilite". The main content is a confusion matrix table with the following data:

Cluster \ P...	cluster_3	cluster_2	cluster_1	cluster_0
cluster_3	875	0	0	2
cluster_2	1	889	17	22
cluster_1	4	29	521	25
cluster_0	1	27	12	908

Below the table, the following statistics are displayed:

- Correct classified: 3,193
- Wrong classified: 140
- Accuracy: 95.8%
- Error: 4.2%
- Cohen's kappa ( $\kappa$ ): 0.943%

**Fig. 11.** Confusion matrix of classification method

## 4. Results

### 4.1. Average accuracy of classification

From the confusion matrix, K-means model has a more accurate model than DBSCAN, that is KNN model. DBSCAN has two models with higher accuracy, Random Forest and Decision Tree. The remaining models are SVN and Naïve Bayes, and both K-means model and DBSCAN have the same accuracy. From the average accuracy of five Classification models, K-means had an average accuracy

of 89.523% and DBSCAN had an average accuracy of 39.397%. as shown in the table below. (Figure 12) [4] [6]

	K-means	DBSCAN
Random Forest	90.159%	90.219%
Decision Tree	87.939%	88.659%
SVM	90.159%	90.159%
Naïve Bayes	89.829%	89.829%
KNN	89.529%	88.119%
Average	89.523%	89.397%

**Fig. 12.** Average accuracy

## 5. Discussion and Conclusion

Based on this independent study, we created a mobile customer behavior research model with a total of 3333 datasets. Group the model into 4 groups and compare them with K-means and DBSCAN. Two models were found to divide the data into the same numbers: 2180, 830, 92, and 231. The data is defined as cluster0, cluster1, cluster2, and cluster 3.

The accuracy is evaluated by five classification models: Random Forest, decision tree, SVM, Naif Bay, and KNN. 10-fold cross validation to prevent excessive data installation and divide the data into 3000 training and testing data. 333 data items show that the accuracy of all models is similar, and on average, the K-means data of all five models The average accuracy is higher than the DBSCAN model, with accuracy of 89.523% and 89.397%, respectively.

In the utilization of computational resources during model training, the researchers only compared the Batch parameter because the Epoch parameter determines the maximum number of training rounds. However, the Batch parameter determines the number of iterations used to train the model within one Epoch. For example, if there are 200 images and the Batch is set to 2, there will be 100 iterations, divided into 4 parts.

From the accuracy results of the classification model, it can be concluded that K-means is the most accurate and suitable model for clustering the mobile phone usage behavior of customers in this data set [8]. If the company uses the results of this grouping, it must be able to describe the characteristics of each group in order to be able to use it further. By describing the characteristics or classification conditions of each group can be explained from the tree from the Decision Tree model as shown in Figure 5.1. The more factors used in the analysis will result in more complex classification conditions as well.

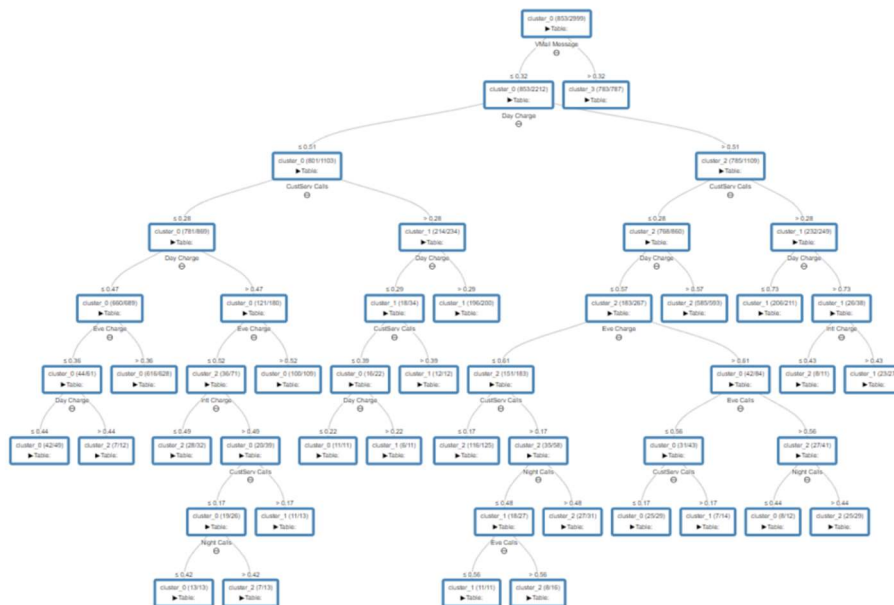


Fig. 13. Decision Tree

## References

- [1] D. P. R. M. Trupti M. Kodinariya, "Review on determining number of Cluster in K-Means Clustering," 6 November 2013.
- [2] K. D. M. D. Thanh N. Tran, "Revised DBSCAN algorithm to cluster data with dense adjacent clusters," November 2013.
- [3] C. Z. a. M. O. Tao Li, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," April 2004.
- [4] J. W. a. A. Mingkhwan, "A Comparative Efficiency of Correlation Plot Data Classification," August 2011.
- [5] P. J. a. S. S. Natthawan Phonchan, "Clustering Efficiency Comparison of Outliers Data in Data Mining," June 2020.
- [6] A. P. a. J. A. L. M. Juan Diego Rodri'guez, "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation," March 2010.
- [7] T. M. K. a. D. P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," November 2013.
- [8] A. K. J. L. C. E. S. Joshua M. Dudik, "A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals," January 2015.
- [9] Y. L. C. C. L. a. X. T. Chen Huang, "Learning Deep Representation for Imbalanced Classification," 2016.
- [10] M. Y. Kiang, "A comparative assessment of classification methods," 1 May 2002.