

# TARU

TUNKU ABDUL RAHMAN  
UNIVERSITY COLLEGE


**TUNKU ABDUL RAHMAN UNIVERSITY COLLEGE**

**FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY (FOCS) ACADEMIC**

**YEAR 2021/2022**

**AACS2383 INTRODUCTION TO DATA MINING**

**ASSIGNMENT**

<b>COURSE NAME</b>	:	<b>Introduction to Data Mining</b>
<b>COURSE CODE</b>	:	<b>AACS2383</b>
<b>PROGRAMME</b>	:	<b>Diploma in Computer Science (Data Science) (DDS) Diploma in Computer Science (DCS)</b>
<b>ASSIGNMENT TITLE</b>	:	<b>Data Analysis</b>
<b>SESSION</b>	:	<b>202105</b>
<b>SUBMISSION DEADLINE</b>	:	<b>14 September 2021 (Week 13)</b>
<b>TOTAL MARK</b>	:	<b>100%</b>
<b>WEIGHTAGE TO FINAL MARK</b>	:	<b>70%</b>
<b>STUDENT NAME</b>	:	<b>TAN KANG HONG</b>
<b>STUDENT ID</b>	:	<b>2002959</b>
<b>STUDENT SIGNATURE</b>	:	

<b>Assignment</b>	<b>Marks</b>
<b>TOTAL (100%)</b>	

## ASSIGNMENT RUBRICS

Your written assignment and presentation will be assessed against the following criteria.

CLO/ Attribute	Attribute/ Subattribute	Weight (%)	Attainment Scale					Student Attainment		
			1 Very Weak	2 Weak	3 Fair	4 Good	5 Very Good	Levels	Marks	CLO Marks
CLO2 (A2)	Introduction on application	10%	Not able to explain an introduction, even with assistance.	Able to partially explain an introduction with maximum assistance.	Able to explain an introduction with minimum assistance.	Independently able to explain an introduction clearly without assistance.	Able to explain the introduction very clearly and accurately.			
	Analysis	20%	Not able to organise and analyse gathered information or data and fails to define the factors that contribute to the problem/issue or explain the root of the problem.	Finds difficulty in organizing and analysing gathered information or data and finds difficulty in explaining the factors that neither contribute to the problem/issue nor explains the root of the problem.	Able to organise and analyse gathered information or data but does not clearly describe the factors that contribute to the problem/issue or clearly explain the root of the problem.	Able to organise and analyse gathered information or data, clearly describe some factors that contribute to the problem/issue or explain the possible roots of the problem.	Able to organise and analyse gathered information or data, clearly describe the factors that contribute to the problem/issue or explain the root of the problem.			
	Application/ software usage	20%	Not able to apply any new idea or knowledge to a given problem.	Limited ability to apply a new idea or knowledge to solve problems.	Able to apply new idea or knowledge to a given problem with assistance from lecturer or student.	Able to apply a new idea or knowledge to a given problem independently.	Able to apply a new idea or knowledge to a given problem and able to propose alternative applications to solve problems.			
	Synthesis and Evaluation	10%	Fails to gather information for synthesis and evaluation.	Has difficulty in gathering, synthesising, and evaluating information.	Able to gather relevant information, synthesise and evaluate the information and offers simple, unsupported conclusions.	Able to gather and thinks about information, synthesise, able to offer responsible interpretations; provides sufficient evidence to form a whole idea, new solution, or system.	Able to gather and evaluates information, chooses a clear interpretation, and provides sufficient evidence (quality and quantity) to form a whole idea or new solution.			

	Decision making	10%	Not able to make decisions based on comparison and contrast between information, ideas, and solutions even with assistance.	Able to make decisions based on comparison and contrast between information, ideas, and available solutions with some assistance.	Able to make decisions based on comparison and contrast between information, ideas, and available solutions.	Able to make good decisions based on comparison and contrast between information, ideas, and available solutions.	Able to make excellent decisions based on comparison and contrast between information, identify problems and available solutions.			
--	-----------------	-----	---	---	--	---	---	--	--	--

CLO3 (A2)	Work Ethics	10%	Practise inappropriate working culture such as bad behaviour, no punctuality as well as not being efficient, productive, and ethical at work in all situations.	Practise less appropriate working culture such as inconsistent behaviour, less punctuality as well as being less efficient, productive, and ethical at work in many situations.	Practise good working culture such as good moral, timeliness as well as being efficient, productive, and ethical at work in general.	Practise good working culture such as good moral, timeliness as well as being efficient, productive, and ethical at work in most situations.	Always practise excellent working culture such as good moral, timeliness as well as being efficient, productive, and ethical at work in all situations.			
	Work Responsibility	10%	Does not perform assigned tasks within by the scope of work even with close supervision.	Perform assigned tasks within by the scope of work with close supervision.	Perform assigned tasks within by the scope of work and meets expectation.	Perform assigned tasks within by the scope of work and exceeds expectation.	Perform assigned tasks beyond the scope of work and beyond expectation.			
	Integrity	10%	Perform a task with lack of trust, honesty, sincerity, and transparency.	Perform a task with limited trust, honesty, sincerity, and transparency.	Perform a task with acceptable trust, honesty, sincerity, and transparency.	Perform a task with trust, honesty, sincerity and transparent in most situations.	Always perform a task with trust, honesty, sincerity and transparent in any situation.			
	<b>Total</b>	100%								

**Remarks:**

## **Table of Contents**

<b>Title</b>	<b>Page</b>
<b>Table of Contents</b>	<b>4</b>
<b>Chapter 1: Application area and goals (Business Understanding)</b>	<b>5</b>
<b>Chapter 2: Structure and size of the data set (Data Understanding)</b>	<b>6 – 8</b>
<b>Chapter 3: Pre-processing</b>	<b>9</b>
<b>Chapter 4: Data Mining</b>	<b>10</b>
4.1 Machine Learning approaches	
4.1.1 Classification (Naïve Bayes & K-NN)	10
4.1.2 Clustering (K-means & DBSCAN)	10
4.2 Evaluation	10
<b>Chapter 5: Result and Discussion</b>	<b>11</b>
5.1 Classification	
5.1.1 Naïve Bayes (Bayesian classifier)	11 - 12
5.1.2 K-NN (Distance-based classifier)	12 - 14
5.2 Clustering	
5.2.1 K-means	14 – 16
5.2.2 DBSCAN	16 – 20
5.3 Conclusion	21
<b>References</b>	<b>22</b>

## **Chapter 1: Application are and goals (Business Understanding)**

Nowadays, obesity issue always become person attend and serious. It is an uncommon and abrupt increase in body fat. How people become obesity? The mostly caused by lazy, not patience, overeating and lack of activity. In South America country, Obesity issue become more serious especially in Mexico, Peru and Colombia. Their government find the company in charge the program of obesity. In South America, Tan company has a history of nearly 20 years in the area of mass weight loss. Our company will help those country to solve their problem. First, we can build infrastructure on a large scale for people to use and provided the free bicycles in the garden, that will be use by physical activity frequency (FAF). Secondly, Legislation making calorie labelling on menus for food and drinks in cafes, restaurants, bars and takeaways compulsory for all businesses with more than 250 employees, that will be using for calories consumption monitoring (SCC). Then, the government can restrict imports of high-calorie products in order to reduce the usage and purchase rates in the country, that will help people decrease the high caloric food (FAVC). And also, we can provide the advertising of food high in fat, sugar or salt on television and online before 10pm, when children are most likely to see them, it will help people decrease number of main meals (NCP). The government stipulates that every Saturday and Sunday is a car-free day, and the plan is implemented in some prosperous streets. First, you can lose weight by walking more, and second, you can reduce carbon dioxide, that can be transportation used (MTRANS). In terms of education, we use schools to convey health information and the symptoms of obesity, it is proposed to hold intramural sports meets, state sports meets and national sports meets every year, and also time using technology devices (TUE). The data contains numerical data and continuous data, so it can be used for analysis based on algorithms of classification and clustering. The purpose of the government's move is to hope that the people can maintain physical and mental health and greatly reduce the national obesity rate. To provide a healthy and perfect city for the family and the country.

The primary goal and objective of this project to achieve accuracy of higher than 80%. This is because the primary goal can help people to keep fit and reduce the obesity. In this project, the main purpose and objective is to know the number of individuals that are in different obesity levels such as Overweight Level 1, Overweight Level 2, Obesity Type and more. The mission of the assignment is the people who live in Mexico, Peru and Colombia. This purpose is to help relevant countries carry out weight-loss plans, and the target user is the people of the whole country. The countries my mention is really having high level obesity, there are only a few ways to measure a quantitative assessment, such as a BMI. So, there are only a few types of levels, such as underweight, overweight and obesity.

## Chapter 2: Structure and Size of the data set (Data Understanding)

In the dataset included data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia. Based on their eating habits and physical condition. The data contains 17 attributes and 2111 records, the records, the records are labelled with the class variable Obesity (obesity level), that allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

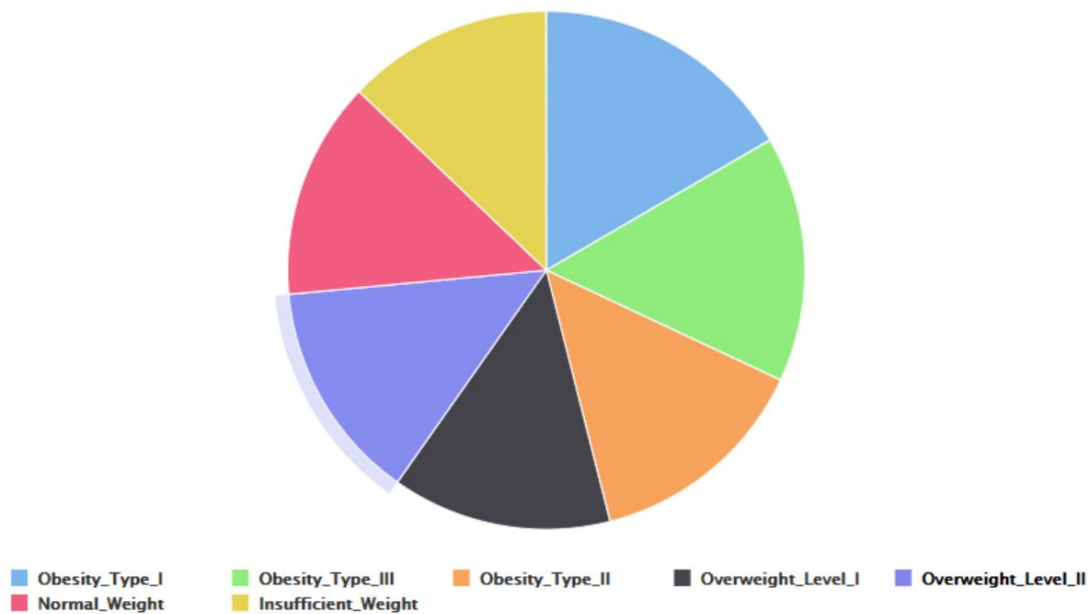
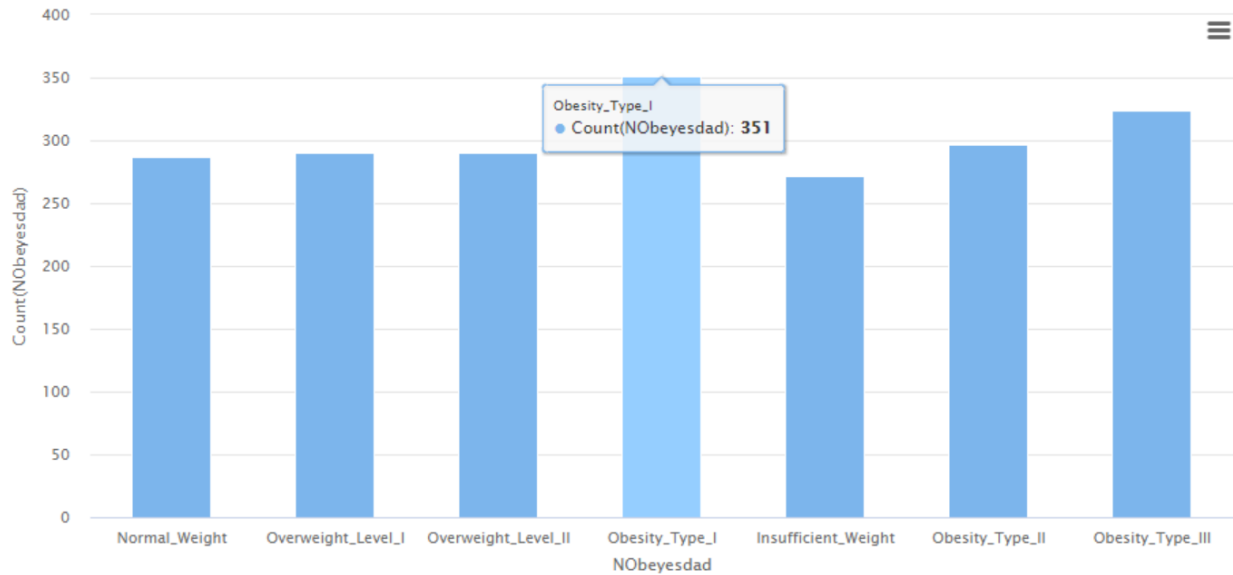
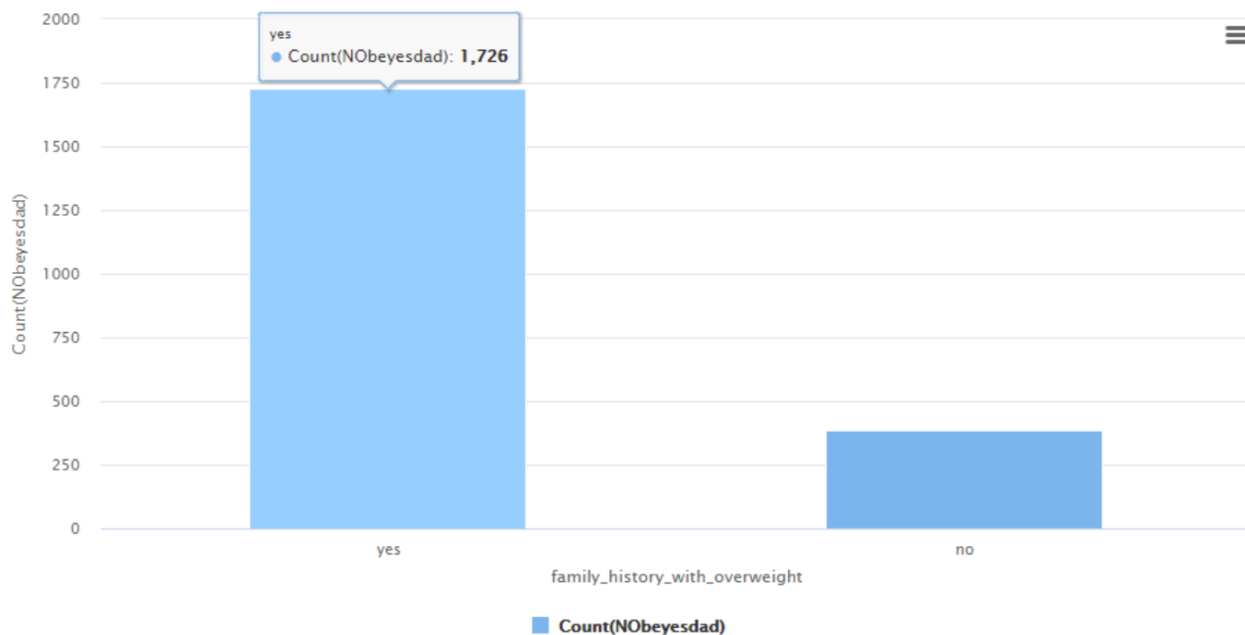


Image 1 The value of Obesity

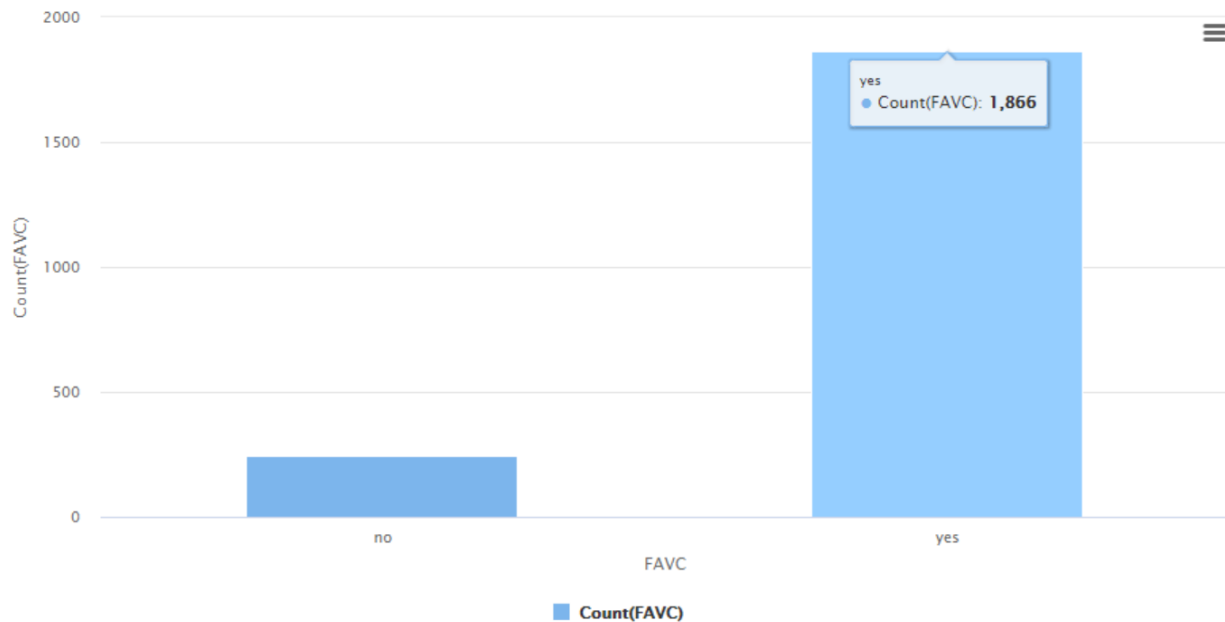


From the graph, it can be seen that obesity type I is the most serious issue in NObesydad. As we all know, it is easiest to gain weight. From here, we clearly know that the country obesity rate is at high risk. So, it is common for the country to completely lock n the high obesity rate. People are not fully aware of the crises brought about by obesity. Compared with other types, they have a relatively high rate of obesity, which is also a sign of obesity. And this is why I will choose this attribute for my data set.



We also did some research in this country. As we can see, most people are being family history with overweight. We will use the predicted conclusions to use medical resources to help people with obesity issues as soon as possible. In addition, we will use predictions to understand whether obesity inheritance is related to usual eating habits. We will provide nutrition packages for people with obesity genetics to understand whether long-term consumption of nutritious meals can help them lose weight.

Compared with other groups, they have a relatively high rate of obesity, which is also a sign of obesity. And this is why I will choose this attribute for my data set.



As we can see, most of the people are like to eat high caloric food (FAVC), this is why obesity in this country will increasing. Most people choose high-calorie foods and various businesses provide high-calorie foods frequently, which greatly increases the obesity rate in the country. Through the predicted results, restricting high-calorie foods and exercising more can effectively reduce fat and lose weight. Compared with other groups, they have a relatively high rate of obesity, which is also a sign of obesity. And this is why I will choose this attribute for my data set.



## Chapter 3: Pre-processing

During pre-processing, we have selected 8 out of 17 attributes, for analyzation and estimation for the obesity level of individuals. This is to avoid complex data that might affect the performance during data analysis. Furthermore, we also do some filters to the dataset before our data mining process, as a result we will analyses the data according to the specific conditions.

Let us filtered individuals who are less than or equal to 35 years old, this will used for predict the probability of individuals who have obesity during their active years, and test whether that will affect their lifestyle. To reduce the scope of individuals, we also needed filtered out who is their family does not suffer from overweight. Also, we provide a new technology to help us to quickly solve the scope, that is rapid miner.

The attribute is shown during my analysis

Attribute	Value	High Score
Age	Numeric value	18 - 23
CAEC	No, Sometimes, Frequently & Always	Sometimes
CALC	I do not drink, Sometimes, Frequently & Always	Sometimes
CH20	Less than a liter, Between 1 and 2L & more than 2L	2L – 2.2L
FCVC	Never, Sometimes & Always	2 – 2.2
MTRANS	Automobile, Motorbike, Bike, Public Transportation & Walking	Public Transportation
NCP	Between 1 y 2, Three & More than three	2 - 3
NObeyesdad	Insufficient Weight, Normal Weight, Obesity Type I, Obesity Type II, Obesity Type III, Overweight Level I, Overweight Level II	Obesity Type I
FAF	I do not have, 1 or 2 days, 2 or 3 days & 4 or 5 days	0.9 – 1.2
Family_History_With_Overweight	Yes , No	Yes
FAVC	Yes , No	Yes
Gender	Male (M), Female (F)	Male
Height	Numeric value in meters	1.715 – 1.768
SCC	Yes , No	No
SMOKE	Yes , No	No
TUE	0 – 2 hours, 3 – 5 hours & more than 5 hours	0 – 2 hours
Weight	Numeric value in kilograms	79.2 – 92.6

# **Chapter 4: Data Mining**

## **4.1 Machine Learning Approaches**

### 4.1.1 Classification

The Naïve Bayes and K-Nearest Neighbor (K-NN) were the approaches used for the classification of dataset. When Naïve Bayes has a sound management method, it is suitable for large dataset. It tends to be fast, and it also has a good accuracy rate. In addition, Naïve Bayes cannot be affected by the problem of several data sets, such as high-dimensional data and a large number of feature combinations.

K-NN puts all the emphasis on the dataset, because it calculates every step of the process. Therefore, although it takes slightly more time than Naïve Bayes, it provides a higher accuracy rate. The accuracy of K-NN can be adjusted using the value of k. Even though the result is not satisfactory, we can still adjust, so the result we can get is satisfaction.

### 4.1.2 Clustering

Many real-world data are unlabeled and do not have any specific categories. The advantage of using algorithms like K-means clustering is that we usually don't know how the instances in the data set should be grouped. For example, consider the problem of trying to group cluster based on similar viewing behavior. As we know, some clusters are different, so we would not know whether which is cluster, so K-means can help us to build up and solve the problem.

For DBSCAN, DBSCAN is a density-based clustering algorithm that forms clusters of dense regions of data points ignoring the low-density areas (considering them as noise). But it will help us solve a lot problem during data analysis. And most important is it can identify outliers easily. Clusters can take any irregular shape unlike K-Means where clusters are more or less spherical.

## **4.2 Evaluation**

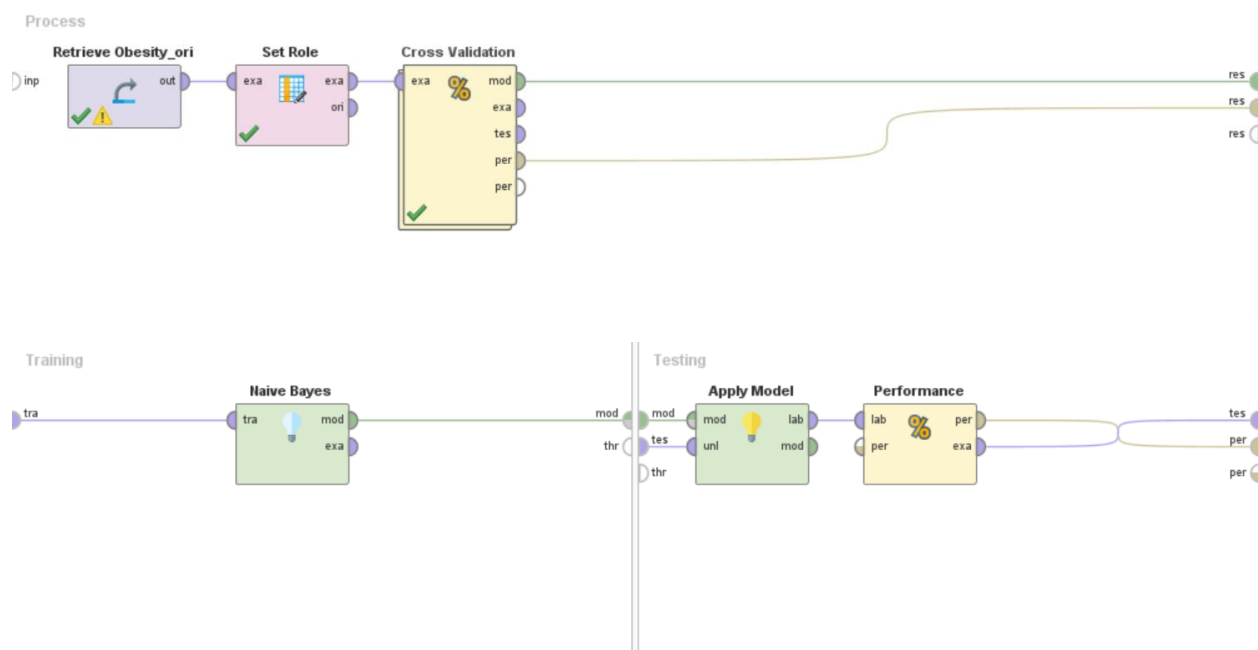
For Evaluation part, I will display in Output there, because can easier to mention and understanding. In Output part, all of the evaluation explain I will mention at the top of them.

# Chapter 5: Result and Discussion

## 5.1 Classification

### 5.1.1 Classification (Naïve Bayes)

The Naïve Bayes Model design that is applied on the Obesity dataset is depicted below.



We tried the Naïve Bayes with the correction enabled, as well as the K-NN with k value of 5. The target role for this test is the NObeyesdad, family\_history\_with\_overweight and FAVC attribute.

accuracy: 64.76% +/- 2.93% (micro average: 64.76%)

	true Normal_...	true Overweigh...	true Overweigh...	true Obesity_T...	true Insufficien...	true Obesity_T...	true Obesity_T...	class precision
pred. Normal_...	150	34	39	1	27	1	0	59.52%
pred. Overweig...	34	143	16	23	6	0	0	64.41%
pred. Overweig...	19	62	129	61	0	0	0	47.60%
pred. Obesity_...	0	24	75	179	0	44	1	55.42%
pred. Insufficie...	75	9	0	0	239	0	0	73.99%
pred. Obesity_...	0	0	7	53	0	204	0	77.27%
pred. Obesity_...	9	18	24	34	0	48	323	70.83%
class recall	52.26%	49.31%	44.48%	51.00%	87.87%	68.69%	99.69%	

Figure 5.1.1.1 NObeyesdad

accuracy: 83.14% +/- 3.21% (micro average: 83.14%)

	true yes	true no	class precision
pred. yes	1445	75	95.07%
pred. no	281	310	52.45%
class recall	83.72%	80.52%	

Figure 5.1.1.2 family\_history\_with\_overweight

accuracy: 82.24% +/- 1.87% (micro average: 82.24%)

	true no	true yes	class precision
pred. no	140	270	34.15%
pred. yes	105	1596	93.83%
class recall	57.14%	85.53%	

Figure 5.1.1.3 FAVC

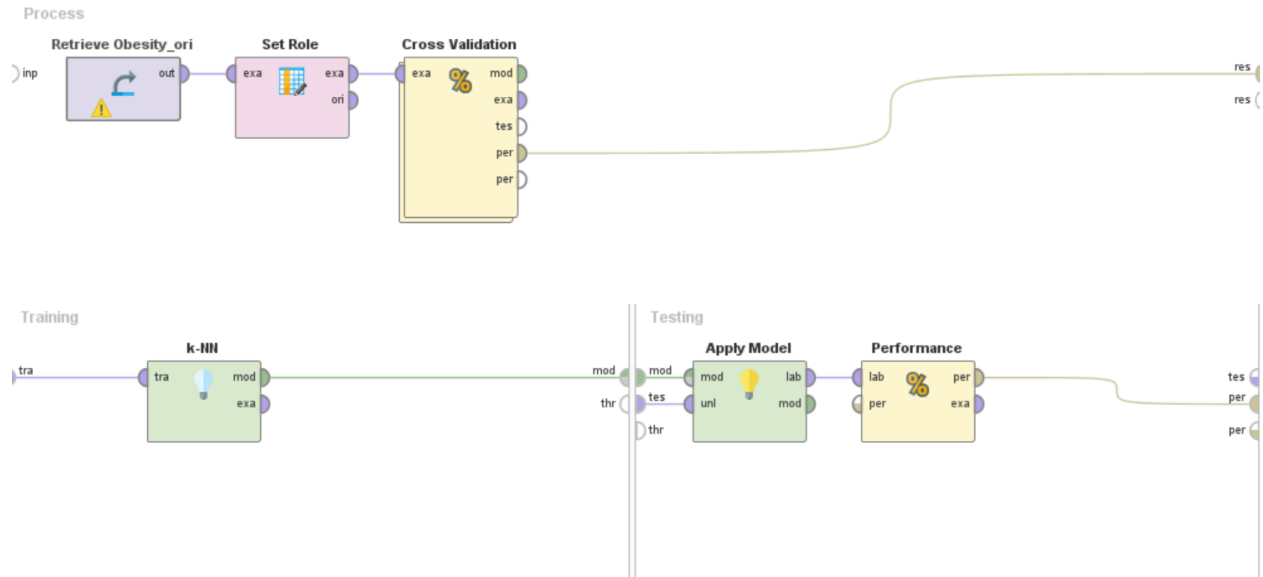
### **Naïve Bayes (Evaluation)**

Attribute	Accuracy
NObeyesdad	64.76% +/- 2.93% (micro average: 64.76%)
family_history_with_overweight	83.14% +/- 3.21% (micro average: 83.14%)
FAVC	82.24% +/- 1.87% (micro average: 82.24%)

Combining the above average, the three performances of Naïve Bayes did successful pass the 64% accuracy of the goal mark. As we can see, the Figure 5.1.1.1 is the lower in the performances of Naïve Bayes, it is because the most of the people are obesity level III, this is a serious issue. The Figure 5.1.1.2 is the higher in the performances of Naïve Bayes, this is possibly due to the people are increase their obesity, some of them are family history with overweight. The Figure 5.1.1.3 is the second higher in the performances of Naïve Bayes, it is because people are doing exercise and also decrease some high caloric food. It is easy and fast to predict the class of the test data set. It also performs well in multi-class prediction. When assumption of independence holds, a Naïve Bayes classifier perform better compare to other models like logistic regression and you need less training data. It performs well in case of categorical input variables compared to numerical variable. For numerical variable, normal distribution is assumed.

## **5.1.2 Classification (K-NN)**

The K-NN Model design that is applied on the Obesity dataset is depicted below.



We tried the K-NN with the correction enabled, as well as the K-NN with k value of 5 to 3. The target role for this test is the NObeyesdad, family\_history\_with\_overweight and FAVC attribute.

accuracy: 88.96% +/- 2.46% (micro average: 88.96%)

	true Normal_...	true Overweigh...	true Overweigh...	true Obesity_T...	true Insuffici...	true Obesity_T...	true Obesity_T...	class precision
pred. Normal_...	158	8	6	0	8	0	0	87.78%
pred. Overweig...	50	262	6	0	0	0	0	82.39%
pred. Overweig...	20	12	254	13	0	1	0	84.67%
pred. Obesity_...	5	5	20	327	0	3	0	90.83%
pred. Insufficie...	54	3	0	0	264	0	0	82.24%
pred. Obesity_...	0	0	4	7	0	291	2	95.72%
pred. Obesity_...	0	0	0	4	0	2	322	98.17%
class recall	55.05%	90.34%	87.59%	93.16%	97.06%	97.98%	99.38%	

Figure 5.1.2.1 NObeyesdad , K value - 5

accuracy: 90.34% +/- 2.49% (micro average: 90.34%)

	true yes	true no	class precision
pred. yes	1658	136	92.42%
pred. no	68	249	78.55%
class recall	96.06%	64.68%	

Figure 5.1.2.2 family\_history\_with\_overweight, K value - 4

accuracy: 91.43% +/- 1.44% (micro average: 91.43%)

	true no	true yes	class precision
pred. no	115	51	69.28%
pred. yes	130	1815	93.32%
class recall	46.94%	97.27%	

Figure 5.1.2.3 FAVC, K value - 3

### **K-NN (Evaluation)**

Attribute	Value of K	Accuracy
NObeyesdad	5	88.96% +/- -2.46% (micro average: 88.96%)
family_history_with_overweight	4	90.34% +/- -2.49% (micro average: 90.34%)
FAVC	3	91.43% +/- -1.44% (micro average: 91.43%)

Combining the above average, the three performances of K-NN did successful pass the 80% accuracy of the goal mark. As we can see, the Figure 5.1.2.1 is the lower in the performances of K-NN, K value is 5 and it is because most people are the type of obesity level III. The Figure 5.1.2.2 is the most stable in the performances of K-NN, K value is 4 and this is possibly due to family history with overweight. The Figure 5.1.2.3 is the higher in the performances of K-NN, K value is 3 and it is based on our plan is work, most people are decrease to eat more calories food. The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, you can use the KNN algorithm for applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure. Therefore, the KNN algorithm is suitable for applications for which sufficient domain knowledge is available. This knowledge supports the selection of an appropriate measure. The KNN algorithm is a type of lazy learning, where the computation for the generation of the predictions is deferred until classification. Although this method increases the costs of computation compared to other algorithms, KNN is still the better choice for applications where predictions are not requested frequently but where accuracy is important.

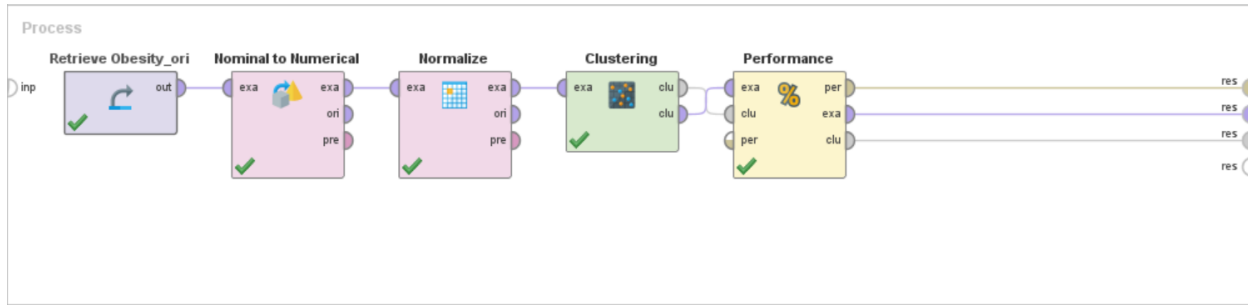
### **Conclusion of classification (Evaluation)**

As shown in the Figure above, there are two classifications. It is Naïve Bayes and K-NN. From Figure, we can conclude that the accuracy of Naïve Bayes is lower than K-NN. For accuracy, we know K-NN is did a great job, it is because it makes the expected job performance relatively outstanding. We found that if Naïve Bayes is taken out, it will effectively improve the accuracy of K-NN. Of course, if the relevant k value is not suitable for the data, K-NN will also reduce its accuracy.

## **5.2 Clustering**

### **5.2.1 Clustering (K-mean)**

We run the dataset with the k-value of 3 to 5 to compare the differences of its distribution of the clusters. We set both of the runs have maximum number of 10.



We tried the K-mean with the correction enabled, as well as the K-mean with k value of 5. The cluster model is regarding how many values, it is 4 clusters. About performance vector, the average of cluster distance is -28.966. The higher mean in this performance vector is cluster\_0 (-59.837).

### Cluster Model

```

Cluster 0: 189 items
Cluster 1: 96 items
Cluster 2: 404 items
Cluster 3: 1059 items
Cluster 4: 363 items
Total number of items: 2111
  
```

### PerformanceVector

```

PerformanceVector:
Avg. within centroid distance: -28.966
Avg. within centroid distance_cluster_0: -59.837
Avg. within centroid distance_cluster_1: -53.694
Avg. within centroid distance_cluster_2: -24.338
Avg. within centroid distance_cluster_3: -19.473
Avg. within centroid distance_cluster_4: -39.195
Davies Bouldin: -2.289
  
```

Figure 5.2.1.1 K-value of 5

We tried the K-mean with the correction enabled, as well as the K-mean with k value of 4. The cluster model is regarding how many values, it is 3 clusters. About performance vector, the average of cluster distance is -29.861. The higher mean in this performance vector is cluster\_2 (-53.690).

### Cluster Model

```

Cluster 0: 323 items
Cluster 1: 44 items
Cluster 2: 595 items
Cluster 3: 1149 items
Total number of items: 2111
  
```

### PerformanceVector

```

PerformanceVector:
Avg. within centroid distance: -29.861
Avg. within centroid distance_cluster_0: -3.745
Avg. within centroid distance_cluster_1: -48.275
Avg. within centroid distance_cluster_2: -53.690
Avg. within centroid distance_cluster_3: -24.157
Davies Bouldin: -2.041
  
```

Figure 5.2.1.2 K-value of 4

We tried the K-mean with the correction enabled, as well as the K-mean with k value of 3. The cluster model is regarding how many values, it is 2 clusters. About performance vector, the average of cluster distance is -31.916. The higher mean in this performance vector is cluster\_1 (-56.115).

## Cluster Model

```
Cluster 0: 694 items
Cluster 1: 532 items
Cluster 2: 885 items
Total number of items: 2111
```

## PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -31.916
Avg. within centroid distance_cluster_0: -20.775
Avg. within centroid distance_cluster_1: -56.115
Avg. within centroid distance_cluster_2: -26.107
Davies Bouldin: -2.468
```

Figure 5.2.1.3 K-value of 3

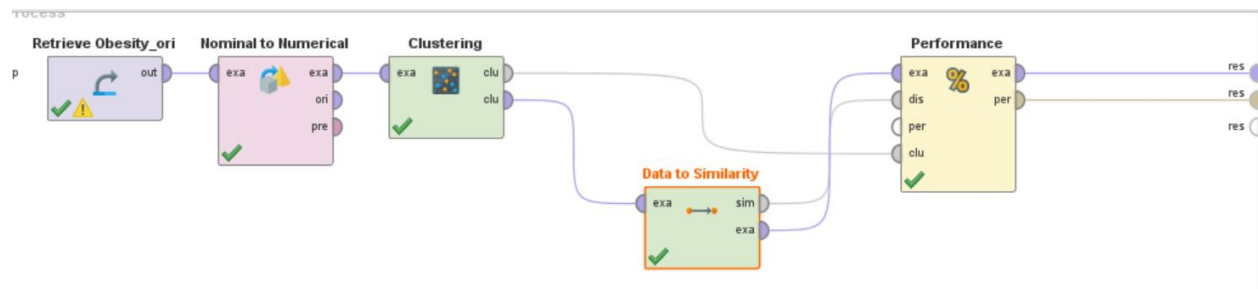
### K-mean (Evaluation)

Value of K	Accuracy Within Centroid Distance
5	-28.966
4	-29.861
3	-31.916

As shown in the figure above, they are cluster model and performance vector. The K-value of Figure 5.2.1.1 is 5. The K-value of Figure 5.2.1.2 is 4 and the K-value of Figure 5.2.1.3 is 3. In Figure, we know when the k value is decreasing the Davies Bouldin will also be decrease. As we can know, the K-means model with 2 clusters has the lowest score in performance vector, which is 2.468. So, we know the model has highest inter-cluster distances and lowest intra-cluster distances. The Highest performance vector belong to K-means is 5 clusters.

### 5.2.2 Clustering (DBScan)

We run the dataset with the Epsilon of 0.1 to 2.5 to compare the differences of its distribution of the clusters. We set both of the runs have minimum number of 5.



We tried the DBScan with the correction enabled, as well as the DBScan with Epsilon of 1.0 and min point = 15. The average of cluster distance is -60766.063. The higher mean in this performance vector is cluster\_0 (-64491.611).



## PerformanceVector

```
PerformanceVector:  
Avg. within cluster distance: -60766.063  
Avg. within cluster distance for cluster 0: -64491.611  
Avg. within cluster distance for cluster 1: -1.045  
Avg. within cluster distance for cluster 2: -46.332  
Avg. within cluster distance for cluster 3: -5.710  
Avg. within cluster distance for cluster 4: -4.114
```

Figure 5.2.2.1 Epsilon of 1.0

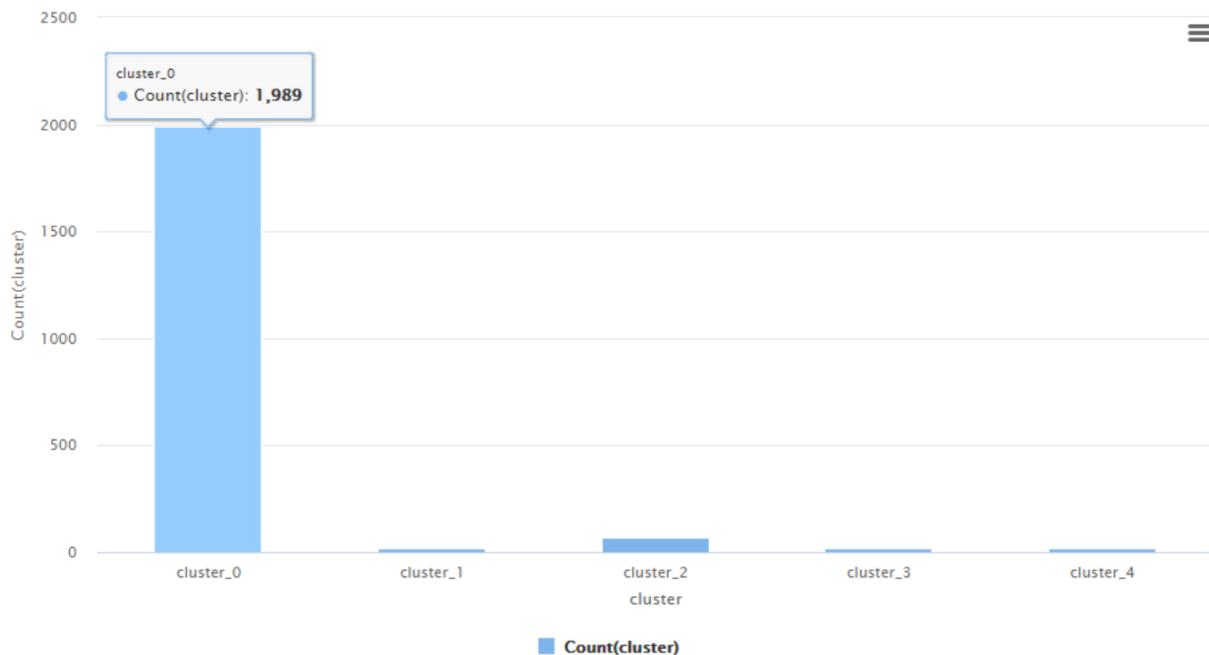


Diagram 5.2.2.1

We tried the DBScan with the correction enabled, as well as the DBScan with Epsilon of 2.5 and min point = 5. The average of cluster distance is -3351.654. The higher mean in this performance vector is cluster\_0 (-11056.180).

## PerformanceVector

```
PerformanceVector:
Avg. within cluster distance: -3351.654
Avg. within cluster distance for cluster 0: -11056.180
Avg. within cluster distance for cluster 1: -343.867
Avg. within cluster distance for cluster 2: -1324.809
Avg. within cluster distance for cluster 3: -881.847
Avg. within cluster distance for cluster 4: -192.503
Avg. within cluster distance for cluster 5: -106.534
Avg. within cluster distance for cluster 6: -1040.543
Avg. within cluster distance for cluster 7: -33.231
Avg. within cluster distance for cluster 8: -311.854
Avg. within cluster distance for cluster 9: -230.704
Avg. within cluster distance for cluster 10: -207.874
Avg. within cluster distance for cluster 11: -17.531
Avg. within cluster distance for cluster 12: -42.530
Avg. within cluster distance for cluster 13: -59.879
Avg. within cluster distance for cluster 14: -65.271
Avg. within cluster distance for cluster 15: -22.918
Avg. within cluster distance for cluster 16: -3718.296
Avg. within cluster distance for cluster 17: -173.469
Avg. within cluster distance for cluster 18: -16.138
Avg. within cluster distance for cluster 19: -22.380
Avg. within cluster distance for cluster 20: -29.954
Avg. within cluster distance for cluster 21: -10.219
Avg. within cluster distance for cluster 22: -99.606
Avg. within cluster distance for cluster 23: -35.891
Avg. within cluster distance for cluster 24: -68.492
Avg. within cluster distance for cluster 25: -32.475
Avg. within cluster distance for cluster 26: -10.692
Avg. within cluster distance for cluster 27: -14.452
Avg. within cluster distance for cluster 28: -9.209
Avg. within cluster distance for cluster 29: -8.349
Avg. within cluster distance for cluster 30: -20.148
Avg. within cluster distance for cluster 31: -15.870
Avg. within cluster distance for cluster 32: -120.390
Avg. within cluster distance for cluster 33: -42.520
Avg. within cluster distance for cluster 34: -10.773
Avg. within cluster distance for cluster 35: -7.484
Avg. within cluster distance for cluster 36: -9.759
Avg. within cluster distance for cluster 37: -60.060
```

Figure 5.2.2.2 Epsilon of 2.5

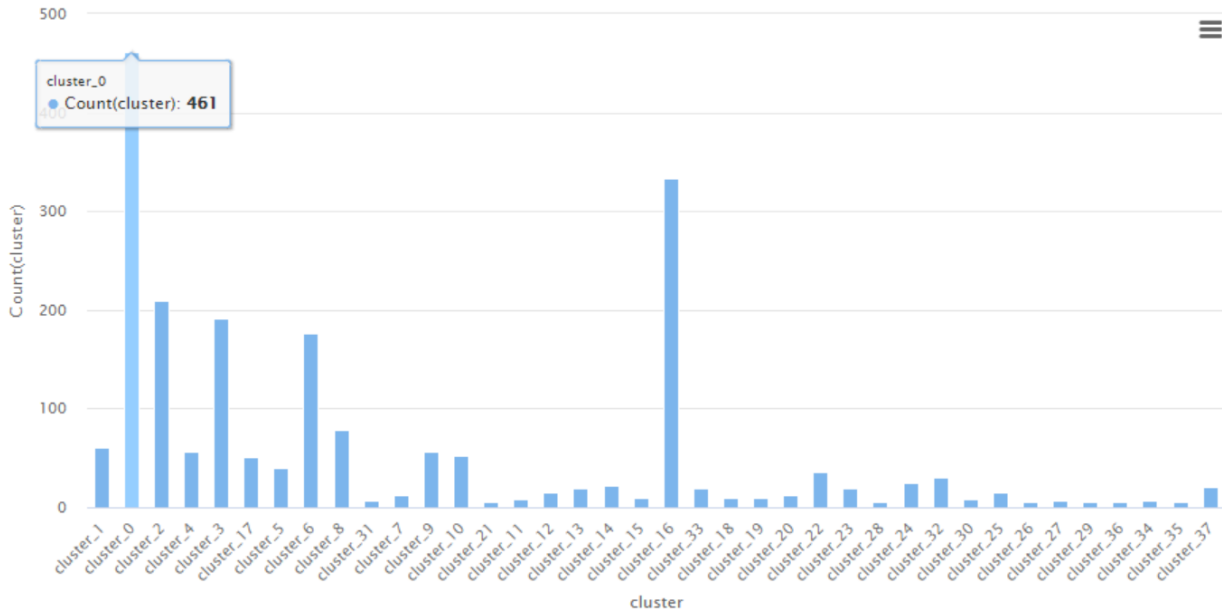


Diagram 5.2.2.2

We tried the DBScan with the correction enabled, as well as the DBScan with Epsilon of 0.1 and min point = 10. The average of cluster distance is -64234.597. The higher mean in this performance vector is cluster\_0 ( -66176.206).

## PerformanceVector

```
PerformanceVector:
Avg. within cluster distance: -64234.597
Avg. within cluster distance for cluster 0: -66178.206
Avg. within cluster distance for cluster 1: -0.034
Avg. within cluster distance for cluster 2: -2.430
Avg. within cluster distance for cluster 3: -0.709
```

Figure 5.2.2.3 Epsilon of 0.1

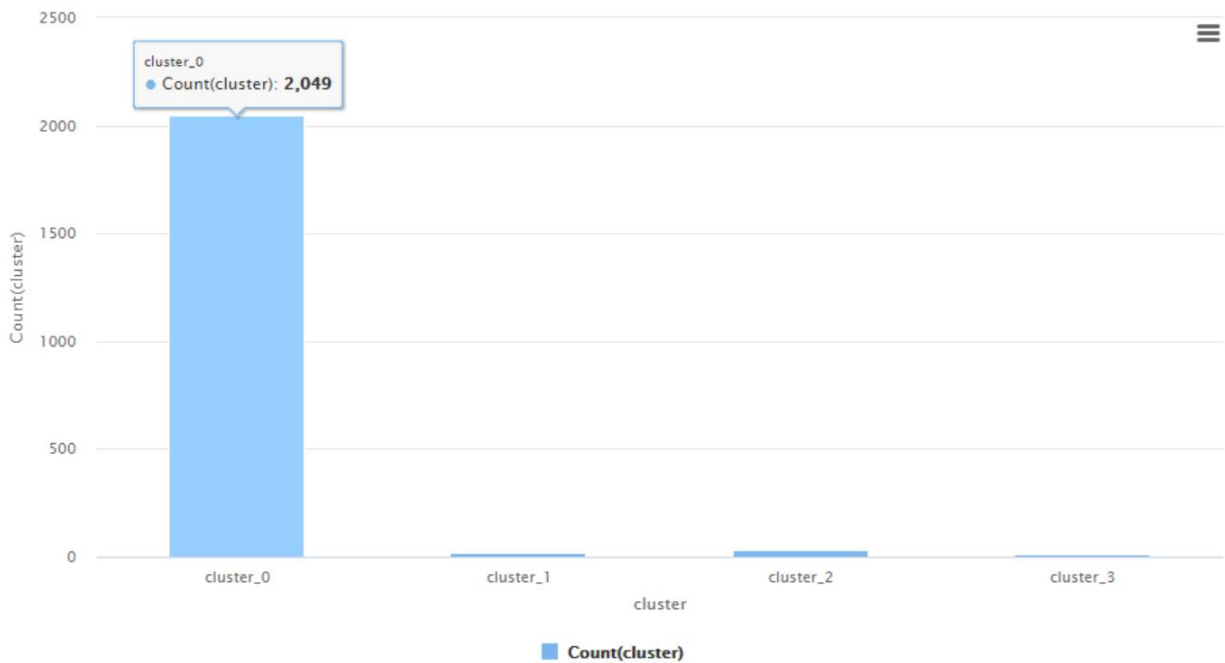


Diagram 5.2.2.3

### **DBScan (Evaluation)**

Value of Epsilon	Minimum Value	Number of Cluster	Accuracy Within Centroid Distance
1.0	15	4	-60766.063
2.5	5	37	-3351.654
0.1	10	4	-64234.597

As shown in the figure above, they are performance vector. The Epsilon of Figure 5.2.2.1 is 1.0. The Epsilon of Figure 5.2.2.2 is 2.5 and the Epsilon of Figure 5.2.2.3 is 0.1. As we know, when the epsilon is increase, the cluster also will be increase. And accuracy increase, the cluster will be decreasing. In Figure, the higher will be Figure 5.2.2.3 in the epsilon of 0.1 and the min value is 10. The lowest will be the Figure 5.2.2.2 in the epsilon of 2.5 and the min value is 5. At the same time, we found that the size of

epsilon will not affect the average. For example, Figure 5.2.2.3 is a good example to us to understanding as well.

### **Conclusion of clustering (Evaluation)**

From the figure above, there are two performance vectors. It is K-mean and DBSCAN. From Figure, we can conclude that both models produce high distinct result. They are not comparable. As we can see in Figure, K-means tend to the data analysis given by obesity (original dataset). For DBScan, we can know that are make a new relationship in features. At the moment, we should choose to able to compare with the original data set instead of choosing a new pool. So, I think choice K-mean is better than DBSCAN, it is because the system can respond quickly and provide more performance vector.

## **5.3 Conclusions**

In conclusion, we have gone through data mining tasks, that are classification and clustering. In order to establish a predictive model with high accuracy, the model can quickly respond and detect people related to obesity. With the support of the government, this plan received a great response and received more detailed information sources. Under the leadership of the government and the company, we can analyze the problems and quickly launch actions to solve the health problems caused by obesity.

Based on Classification, they are Naïve Bayes and K-NN. In data analysis, we can provide K-NN is better than Naïve Bayes. It is because accuracy of the K-NN is the most stable and high in the Figure that I display at the top. The accuracy of Naïve Bayes is lower than K-NN. Regarding Naïve Bayes, total accuracy is 76.71%. For K-NN, the total accuracy is 90.24%. From the related Figure, the achieved accuracy rate is as high as 90.24%, it is sufficient to meet the project of the goal.

In clustering algorithms, they are K-mean and DBSCAN. In performance vector, we can provide K-mean is better than DBSCAN. It is because accuracy of the K-mean can be control and produce by ourself. That mean, I can choose how many clusters for analysis. In the end, we found that using K-mean can reduce some unnecessary things and then achieve precise efficiency. For DBSCAN, it has the lower average intra-cluster distance. We cannot control the cluster and make easy to understanding.

As classification and clustering, it provides better analysis and accurate data set, help us reduce some unnecessary troubles. Classification can effectively predict the degree of obesity. Clustering assists us in finding hidden data through a centralized method. From the above analysis, I believe that when the data set meets predictions, classification and clustering can play a role. In the case of complementarity, the related accuracy is greatly improved, thereby reducing errors.

## **Chapter 6 : Reference**

Chapter 1: Business Understanding

<https://www.nutrition.org.uk/nutritioninthenews/new-reports/obesity-strategy.html> [Accessed 1 September 2021]

Chapter 4: K-mean

<https://automaticaddison.com/advantages-of-k-means-clustering/> [Accessed 2 September 2021]

Chapter 4: DBScan

<https://towardsdatascience.com/k-means-vs-dbscan-clustering-49f8e627de27> [Accessed 4 September 2021]

Chapter 5.1.1 Naïve Bayes

<https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf> [Accessed 7 September 2021]

Chapter 5.1.2 K-NN

<https://www.ibm.com/docs/en/ias?topic=knn-usage> [Accessed 7 September 2021]

Chapter 5.3 Conclusion

<https://techdifferences.com/difference-between-classification-and-clustering.html> [Accessed 12 September 2021]