PEMBANGUNAN SISTEM BERKELOMPOK BOTNET UNTUK PEMANTAUAN

AKMAL NABILAH ASWAMI FADILLAH

AFZAN ADAM

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Pada masa kini, ancaman siber semakin canggih dan serangan semakin rumit malah berevolusi menggunakan teknologi bagi melakukan serangan siber. Antara serangan siber yang bahaya adalah botnet dan serangan ini banyak dilaporkan oleh syarikat-syarikat keselamatan siber dan media. Botnet pada asasnya adalah koleksi peranti yang berkomunikasi melalui internet seperti komputer peribadi, pelayan, peranti mudah alih dan *Internet of Things* (IoT). Namun komunikasi ini disebabkan oleh jangkitan perisian merbahaya dan bukan disebabkan oleh pengguna. Ini adalah kerana serangan siber botnet ini dapat menjangkiti ramai pengguna dengan kadar yang cepat. Justeru, pentingnya ada sebuah sistem bagi memantau aktiviti botnet. Cadangan penyelesaian kepada masalah ini adalah dengan melaksanakan pembangunan sistem pengelompokan botnet untuk pemantauan. Pengelompokan ialah salah satu pendekatan yang boleh membantu untuk mengkategorikan perisian merbahaya. Melalui penyelesaian ini, pengamal keselamatan siber boleh menggunakan kaedah ini untuk membantu mereka membuat pemantauan *pattern* serangan yang telah berlaku sebelum ini agar pengamal keselamatan siber boleh mengambil tindakan dengan segera jika berlaku lagi serangan botnet tersebut. Penyelesaian ini juga akan dapat membantu organisasi kecil dengan kapasiti manusia dan teknologi yang terhad. Kaedah pengelompokan menggunakan model kmeans dan pemilihan nombor kluster adalah berdasarkan kaedah skor siluet. Berdasarkan hasil kajian didapati, nombor kluster yang sesuai untuk melakukan pengelompokan bagi data botnet tersebut adalah k=3.

1. PENGENALAN

Botnet pada asasnya adalah koleksi peranti yang berkomunikasi melalui internet seperti komputer peribadi, pelayan, peranti mudah alih dan *Internet of Things* (IoT). Namun komunikasi ini disebabkan oleh jangkitan perisian merbahaya dan bukan disebabkan oleh pengguna. Ini adalah kerana serangan siber botnet ini dapat menjangkiti ramai pengguna dengan kadar yang cepat. Justeru, pentingnya ada sebuah sistem bagi memantau aktiviti botnet. Botnet biasanya digunakan untuk menghantar e-mel spam dan terlibat dalam penipuan klik. Botnet berasal daripada perkataan robot dan rangkaian. Bot ialah sejenis aplikasi perisian atau skrip yang melaksanakan tugas automatik pada arahan. Ia adalah peranti yang dijangkiti oleh kod berniat jahat yang menjadi sebahagian daripada rangkaian atau jaring yang membolehkan penyerang mengawal komputer yang dijangkiti dan melakukan tugas berniat jahat. Botnet juga boleh digambarkan sebagai tentera zombi kerana selepas melancarkan serangan ke atas mangsa, peranti mangsa secara amnya tidak menyedari bahawa mesin mereka secara tidak sengaja melakukan tindakan berniat jahat. Pelayan arahan dan kawalan ialah komputer yang dikawal oleh penyerang yang digunakan untuk menghantar arahan untuk menjejaskan sistem menggunakan perisian merbahaya dan menerima data yang dicuri daripada rangkaian sasaran. Ini boleh dikatakan bahawa pelayan arahan (C&C) berfungsi sebagai peranan utama kepada mesin yang terjejas dalam botnet.

2. PENYATAAN MASALAH

Pada masa kini, ancaman siber semakin canggih dan serangan semakin rumit malah berevolusi menggunakan teknologi bagi melakukan serangan siber. Antara serangan siber yang bahaya adalah botnet dan serangan ini banyak dilaporkan oleh syarikat-syarikat keselamatan siber dan media. Teknologi sepatutnya membantu manusia dalam menyelesaikan masalah untuk melaksanakan tugas sukar yang manusia tidak mampu, tetapi pertumbuhan era internet dan sektor korporat yang berurusan komunikasi dalam talian telah memperkenalkan ancaman internet yang sering menarik kepada penjenayah siber. Ancaman Berterusan Lanjutan (Advanced Persistent Threat) ialah salah satu ancaman keselamatan siber yang paling serius. APT ialah tingkah laku berniat jahat atau anomali yang mengatasi sekatan keselamatan dengan utama mengintip, mencuri dan mengendalikan maklumat peribadi sensitif yang dimiliki sama ada individu atau organisasi. Ancaman ini biasanya disasarkan dan agak berbahaya. Daripada semua APT, botnet adalah salah satu serangan yang jelas dan tersembunyi untuk melakukan

jenayah siber. Serangan botnet ialah salah satu ancaman keselamatan siber yang paling serius kerana serangan botnet ialah serangan siber berskala besar yang menjalankan peranti yang dijangkiti perisian merbahaya yang boleh mengawal dari jauh. Mereka adalah risiko infrastruktur kritikal.

3. OBJEKTIF KAJIAN

Objektif projek ini adalah untuk menjalankan kaedah pengelompokan pada data yang diberikan dengan mengumpulkan taburan jangkitan di setiap negeri dan mencari negeri mana yang mempunyai bilangan jangkitan botnet tertinggi supaya pengamal keselamatan siber boleh menggunakannya untuk pemantauan.

4. METHOD KAJIAN

Kajian ini dibangunkan menggunakan *waterfall* model. Ini adalah kerana ia sangat mudah untuk difahami dan digunakan. Dalam konsep model air terjun, setiap fasa perlu diselesaikan sebelum fasa seterusnya boleh dimulakan supaya tidak berlaku fasa pertindihan.

4.1 Fasa Perancangan

Fasa ini penting untuk melakukan perancangan projek. Ia amat penting untuk membimbing pelaksanaan dan mengawal fasa projek. Ia akan memberi penerangan yang jelas tentang penyataan masalah, objektif, skop, kekangan, metodologi dan jadual pelaksaan.

4.2 Fasa Analisis

Fasa ini akan menyediakan spesifikasi keperluan yang diperlukan untuk pembangunan sistem pengelompokan seperti pengguna, keperluan perkakasan dan perisian.

4.3 Fasa Reka Bentuk

Fasa ini akan menentukan keperluan reka bentuk bagi menggambarkan proses pembangunan yang akan dilaksanakan untuk pembangunan sistem

4.4 Fasa Implementasi

Fasa ini akan merangkumi tentang proses pembangunan sistem mengunakan teknologi yang dipilih berdasarkan dokumen Keperluan dan Spesifikasi Reka Bentuk dimana segala keperluan perkakasan dan perisian yang disenaraikan akan digunakan dalam membangunkan sistem.

5. HASIL KAJIAN

Perisian yang digunakan dalam membangunkan sistem pengelompokan botnet ini adalah Jupyter Notebook. Jupyter Notebook ini akan menjadi perisian utama penulisan kod bagi membangunkan sistem tersebut.

Pemilihan Dataset

Dataset yang dipilih untuk melakukan pengelompokan adalah berdasarkan serangan botnet yang telah berlaku pada tahun 2016. Dataset ini merangkumi 209 data dan 29 attribut.

THE INSIGHT DATA

	SinkHoleMessage	TimeStamp	Malware	Destination	Port1	DataCenter	Source	Port2	unknownVariable1	Country	State	District	PostCode
0	SinkHoleMessage	130966847516333733	Conficker	60.52.97.106	63515	AS4788	104.244.14.252	80	NaN	MY	7.0	Ipoh	31350.0
1	SinkHoleMessage	130966847516333733	Conficker	203.106.161.241	3691	AS4788	104.244.14.252	80	NaN	MY	12.0	Petaling Jaya	46400.0
2	SinkHoleMessage	130966847516646248	Conficker	175.144.228.79	2441	AS4788	104.244.14.252	80	NaN	MY	1.0	Johor Bahru	81100.0
3	SinkHoleMessage	130966847516802523	Conficker	202.171.45.80	11820	AS23659	104.244.14.253	80	NaN	MY	12.0	Selangor	45400.0
4	SinkHoleMessage	130966847517271302	Conficker	219.92.59.77	2948	AS4788	104.244.14.253	80	NaN	MY	1.0	Johor Bahru	80500.0
5	SinkHoleMessage	130966847517271302	Conficker	60.52.64.122	62092	AS4788	104.244.14.253	80	NaN	MY	14.0	Kuala Lumpur	50350.0
6	SinkHoleMessage	130966847517271302	Conficker	210.195.216.177	3583	AS4788	104.244.14.252	80	NaN	MY	12.0	Petaling Jaya	47400.0

Rajah Pemilihan Dataset

Baca Data

Dimulakan dengan membaca data berbentuk JSON dan kemudian mendapatkan *attribute* dan *value* yang ada pada data tersebut. Disini *library* Pyspark digunakan bagi membaca dan memproseskan data kerana ia boleh memproses sejumlah besar data dengan lebih cepat.

```
READ THE DATA
root
 |-- Country: string (nullable = true)
 |-- DataCenter: string (nullable = true)
 |-- Destination: string (nullable = true)
 |-- District: string (nullable = true)
 |-- Latitude: string (nullable = true)
 |-- Longitude: string (nullable = true)
 |-- Malware: string (nullable = true)
 |-- Port1: string (nullable = true)
 |-- Port2: string (nullable = true)
 |-- PostCode: string (nullable = true)
 |-- SinkHoleMessage: string (nullable = true)
 |-- Source: string (nullable = true)
 |-- State: string (nullable = true)
 |-- TimeStamp: string (nullable = true)
 |-- unknownVariable1: string (nullable = true)
 |-- unknownVariable10: string (nullable = true)
 |-- unknownVariable11: string (nullable = true)
 |-- unknownVariable12: string (nullable = true)
 |-- unknownVariable13: string (nullable = true)
 |-- unknownVariable14: string (nullable = true)
 |-- unknownVariable15: string (nullable = true)
 |-- unknownVariable2: string (nullable = true)
 |-- unknownVariable3: string (nullable = true)
 |-- unknownVariable4: string (nullable = true)
 |-- unknownVariable5: string (nullable = true)
 |-- unknownVariable6: string (nullable = true)
 |-- unknownVariable7: string (nullable = true)
 |-- unknownVariable8: string (nullable = true)
 |-- unknownVariable9: string (nullable = true)

(209, 29)
```

Rajah 1 Baca Data

Pemilihan Atribut

Kemudian, pilih atribut bagi melakukan pengelompokan. Atribut “Severity”, “State”, “District” dan “Malware” dipilih kerana ia menepati dengan skop yang hendak dikaji iaitu melakukan pengelompokan terhadap negeri yang telah menjangkiti dengan perisian merbahaya.

```
SELECT ATTRIBUTE
+--------+-----+-------------+---------+
|severity|state|     district|  malware|
+--------+-----+-------------+---------+
|    High|   07|         Ipoh|Conficker|
|    High|   12|Petaling Jaya|Conficker|
|    High|   01|  Johor Bahru|Conficker|
|    High|   12|     Selangor|Conficker|
|    High|   01|  Johor Bahru|Conficker|
|    High|   14| Kuala Lumpur|Conficker|
|    High|   12|Petaling Jaya|Conficker|
|    High|   01|   Batu Pahat|Conficker|
|    High|   12|Petaling Jaya|Conficker|
|    High|   14| Kuala Lumpur|Conficker|
+--------+-----+-------------+---------+
only showing top 10 rows
```

Rajah 2 Pemilihan Atribut

Bagi Atribut “Severity”, pada asalnya atribut tersebut dinamakan “UnknownVariable15” dan kemudian ditukarkan ke “Severity” supaya *value* dalam atribut tersebut dapat difahami

```
df= df.selectExpr("unknownVariable15 as severity",
```

Rajah 3 Atribut Severity

Normalisasikan Data

Seterusnya, normalisasikan data yang dipilih kepada nombor kerana pengelompokan kmeans hanya menerima data dalam bentuk nombor. Ini akan memberikan nombor ID kepada data yang berbentuk *string type* sahaja.

```
+-----------+--------+
|severity_id|severity|
+-----------+--------+
|          0|    High|
+-----------+--------+

+----------+---------+
|malware_id|  malware|
+----------+---------+
|         0|Conficker|
+----------+---------+

+-----------+---------------+
|district_id|       district|
+-----------+---------------+
|          0|    Kuala Lumpur|
|          1|   Petaling Jaya|
|          2|     Johor Bahru|
|          3|      Batu Pahat|
|          4|            Ipoh|
|          5|       Shah Alam|
|          6|          Melaka|
|          7|     Bayan Lepas|
|          8|         Kuching|
|          9|   Kota Kinabalu|
|         10|         Kuantan|
|         11|          Penang|
|         12|  Seri Kembangan|
|         13|          Cheras|
|         14|           Klang|
|         15|         Puchong|
|         16|           Bharu|
|         17|Kuala Terengganu|
|         18|           Kulim|
|         19|           Lunas|
+-----------+---------------+
only showing top 20 rows
```

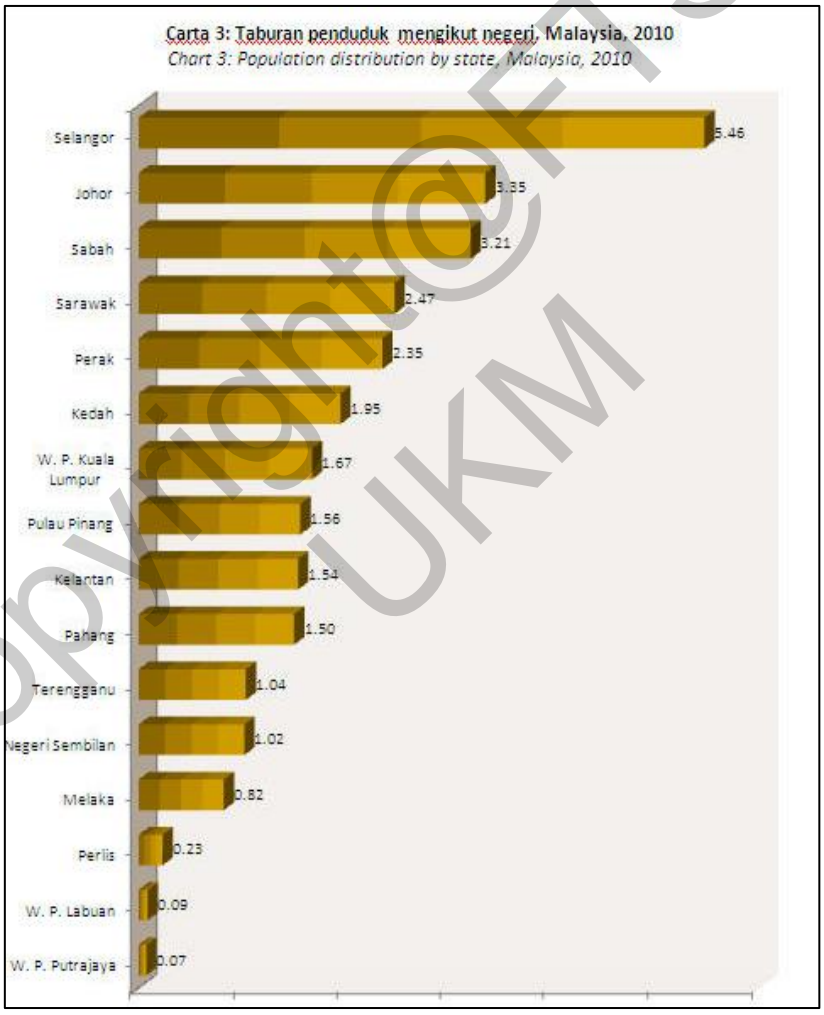
Rajah 4 Normalisasi Data

Filter Data

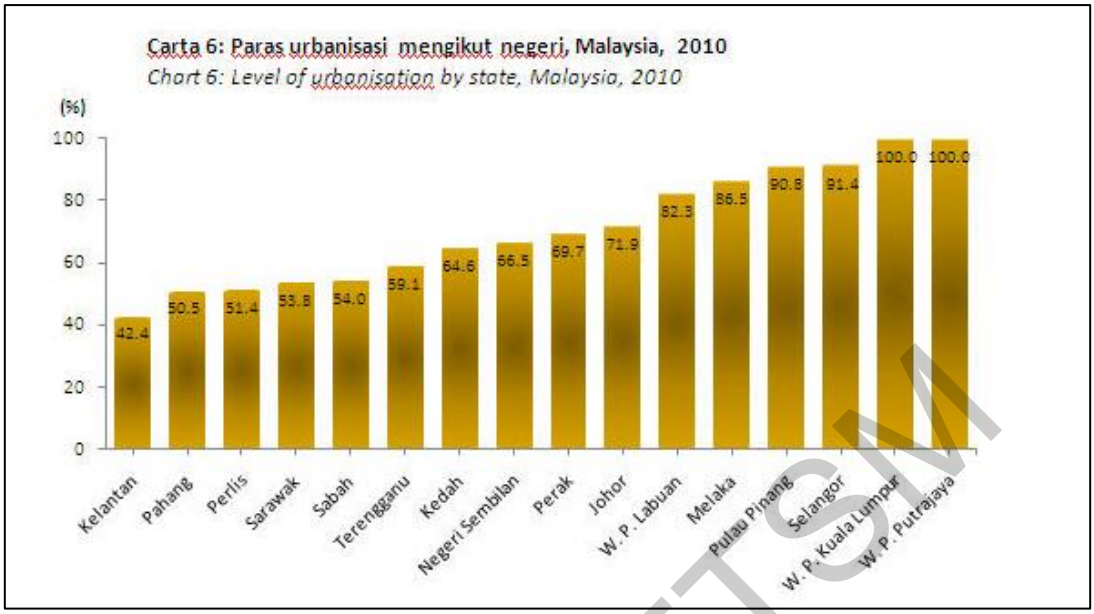
Kemudian, bagi state atribut, data tersebut akan di filter kepada 3 negeri sahaja iaitu Johor, Kuala Lumpur dan Selangor. Data yang di *filter* adalah berdasarkan faktor taburan penduduk mengikut negeri dan urbanisasi negeri.

```
#filter only 3 states
df = df.filter(F.col('state').isin(['1','12','14']))
```

Rajah 5 Filter Data



Rajah 6 Taburan Penduduk



Rajah 7 Urbanisasi Penduduk

Value untuk atribut "Severity" dan "Malware" digabungkan menjadi "severity_malware". Begitu juga untuk atribut "District" dan "State" digabungkan menjadi "district_state". Ini adalah untuk memplot graf X dan Y bagi tujan pengelompokan.

ADD THE ATTRIBUTE VALUE TOGETHER

severity	state	district	malware	severity_id	malware_id	district_id	severity_malware	district_state
High	12	Petaling Jaya	Conficker	0	0	1	0.0	1.12
High	1	Johor Bahru	Conficker	0	0	2	0.0	2.1
High	12	Selangor	Conficker	0	0	25	0.0	25.12
High	1	Johor Bahru	Conficker	0	0	2	0.0	2.1
High	14	Kuala Lumpur	Conficker	0	0	0	0.0	0.14
High	12	Petaling Jaya	Conficker	0	0	1	0.0	1.12
High	1	Batu Pahat	Conficker	0	0	3	0.0	3.1
High	12	Petaling Jaya	Conficker	0	0	1	0.0	1.12
High	14	Kuala Lumpur	Conficker	0	0	0	0.0	0.14
High	14	Kuala Lumpur	Conficker	0	0	0	0.0	0.14
High	14	Kuala Lumpur	Conficker	0	0	0	0.0	0.14
High	14	Kuala Lumpur	Conficker	0	0	0	0.0	0.14
High	14	Kuala Lumpur	Conficker	0	0	0	0.0	0.14
High	14	Kuala Lumpur	Conficker	0	0	0	0.0	0.14
High	1	Johor Bahru	Conficker	0	0	2	0.0	2.1
High	12	Petaling Jaya	Conficker	0	0	1	0.0	1.12
High	12	Petaling Jaya	Conficker	0	0	1	0.0	1.12
High	14	Kuala Lumpur	Conficker	0	0	0	0.0	0.14
High	1	Batu Pahat	Conficker	0	0	3	0.0	3.1
High	12	Petaling Jaya	Conficker	0	0	1	0.0	1.12

severity_malware	district_state
0.0	1.12
0.0	2.1
0.0	25.12
0.0	2.1
0.0	0.14
0.0	1.12
0.0	3.1
0.0	1.12
0.0	0.14
0.0	0.14

Rajah 8 Add Attribute

Semak Nilai Null dan Buang Data Sama

Selain itu, semak atribut yang mempunyai nilai null dan buang data yang mempunyai nilai yang sama.

```
CHECK FOR NULL VALUE
+----------------+--------------+
|severity_malware|district_state|
+----------------+--------------+
|               0|             0|
+----------------+--------------+
```

Rajah 9 Null Value

```
#full data removed duplicates (154, 2)
df = df.dropDuplicates(['severity_malware', 'district_state'])
print((df.count(), len(df.columns)))
```

```
REMOVE NOISE DATA
(17, 2)
```

Rajah 10 Buang Data Sama

Latih Model

Disini, model kmeans dipilih dan dilatih keatas data dan atribut yang telah dipilih. Pengelompokan kmeans dinilai berdasarkan skor siluet

```
vecAssembler = VectorAssembler(inputCols=['severity_malware','district_state'], outputCol="features").setHandleInvalid("ski

df = vecAssembler.transform(df)
# Trains a k-means model.
kmeans = KMeans().setK(n).setSeed(1)
model = kmeans.fit(df)

# Make predictions
predictions = model.transform(df)

# Evaluate clustering by computing Silhouette score
evaluator = ClusteringEvaluator()
```

Rajah 11 Latih Model

Menilai Skor Siluet

Skor Siluet dinilai dengan berkisar antara -1 hingga 1 dimana:

1: Bermakna gugusan adalah berjauhan antara satu sama lain dan dibezakan dengan jelas.

0: Bermakna gugusan adalah acuh tak acuh, atau kita boleh katakan bahawa jarak antara gugusan adalah tidak ketara.

-1: Bermakna gugusan ditetapkan dengan cara yang salah.

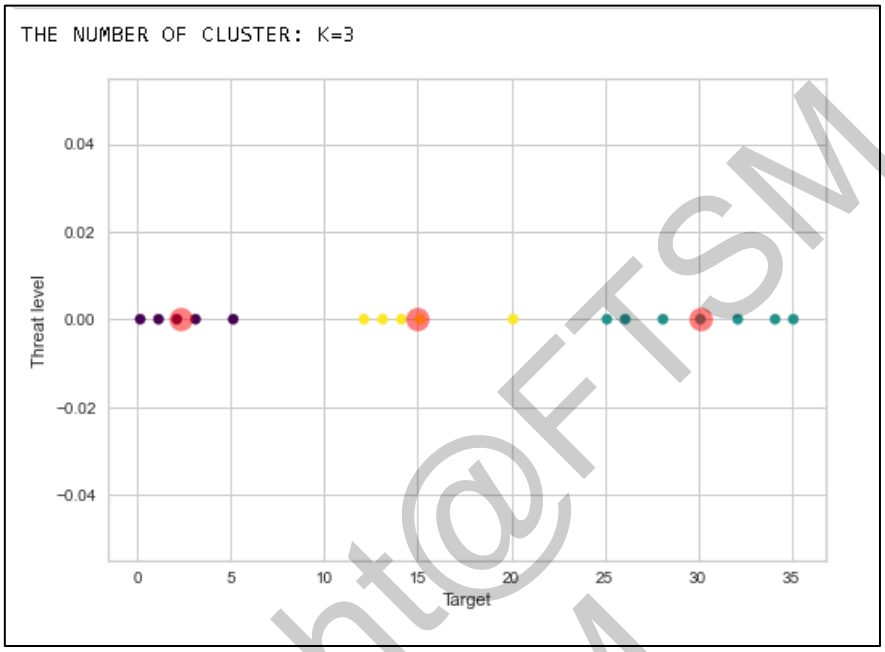
Seperti dilihat K=3 adalah terpaling dekat dengan 1 antara dengan yang lain. Ini menunjukkan K=3 merupakan pengelompokan yang terbaik.

```
EVALUATE CLUSTERING USING SILHOUETTE SCORE
Silhouette with squared euclidean distance = 0.9119413991204124
Cluster Centers:
[0.         1.10983333]
[ 0.     19.9075]
Silhouette with squared euclidean distance = 0.9574465217730224
Cluster Centers:
[0.         1.10983333]
[ 0.         29.11333333]
[ 0.    14.384]
Silhouette with squared euclidean distance = 0.8338109296628257
Cluster Centers:
[0.         0.48086957]
[ 0.         29.11333333]
[ 0.    14.384]
[0.         3.17642857]
Silhouette with squared euclidean distance = 0.8566792883266361
Cluster Centers:
[0.         0.66769231]
[ 0.    32.865]
[ 0.    14.384]
[ 0.    26.112]
[0.         3.98375]
```

Rajah 12 Skor Siluet

Output/Hasil

Oleh itu, nilai siluet skor tersebut digunakan untuk visualisasi pengelompokan. Berdasarkan pengelompokan tersebut dilihat terdapat 3 kluster yang boleh dinamakan “Kluster A”, “Kluster B” dan “Kluster C”.



Rajah 13 Pengelompokan K3

6. KESIMPULAN

Kesimpulan, ia merangkumi semua aspek yang diperlukan dalam membangunkan sistem seperti perancangan projek, kajian susastera, spesifikasi keperluan dan spesifikasi reka bentuk dan implementasi. Proses-proses pembangunan sistem pengelompokan dapat dilihat sehingga ia mencapai hasil visualisasi pengelompokan.

7. RUJUKAN

Hubert, O. 2017. *A Concept of Clustering-Based Method for Botnet Detection*. (PDF) A Concept of Clustering-Based Method for Botnet Detection (researchgate.net) [August 2017].

Search Security Tech Target. 2021. *Definition Botnet*. https://searchsecurity.techtarget.com [March 2021].

Miller, S. 2016. *The Role of Machine Learning in Botnet Detection.* https://www.researchgate.net/publication/313809055_The_Role_of_Machine_Learning_in_Botnet_Detection [December 2016].

Chowdhury, J. 2017. *Botnet detection using graph-based feature clustering*. https://link.springer.com/content/pdf/10.1186/s40537-017-0074-7.pdf [12 May 2017].

Krishna, V. 2020. *A Study on Advanced Botnets Detection in Various Computing Systems Using Machine Learning Techniques.* https://eprajournals.com/jpanel/upload/914pm_31.EPRA%20JOURNALS-5902.pdf [December 2020].

Guofei, G. BotMiner: *Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection*. https://www.usenix.org/legacy/event/sec08/tech/full_papers/gu/gu.pdf

Data Science at Home. 2019. *Waterfall or Agile? The best methodology for AI and machine learning*. https://datascienceathome.com/waterfall-or-agile-the-best-methodology-for-ai-and-machine-learning/ [14 August 2019].

Tanner, G. 2019. *Introduction to Machine Learning Model Interpretation*. https://gilberttanner.com/blog/introduction-to-machine-learning-model-interpretation [13 May 2019].

IBM Cloud Education. 2019. *What is Unsupervised Learning?* https://www.ibm.com/cloud/learn/unsupervised-learning [21 September 2020].

Dubey, A. 2017. *A Systematic Review on K-Means Clustering Techniques* https://gilberttanner.com/blog/introduction-to-machine-learning-model-interpretation [6 June 2017].