

## **ANALISIS PEMETAAN TINGKAT PENGANGGURAN DI PULAU JAWA DAN BALI DENGAN METODE *K-MEANS***

<sup>1</sup>Falentino Sembiring <sup>2</sup>Saeful Bahri Rizqi <sup>3</sup>M Abdul Aziz, <sup>4</sup>Dado Firmansyah

<sup>1,2,3,4</sup>Program Studi Sistem Informasi

<sup>1,2,3,4</sup>Sekolah Tinggi Teknologi Nusa Putra

<sup>1,2,3,4</sup>Jl. Raya Cibat Cisaat No.21, Cibolang Kaler, Kec. Cisaat, Sukabumi

e-mail: <sup>1</sup>falentino.sembiring@nusaputra.ac.id, <sup>2</sup>saeful.bahri\_si17@nusaputra.ac.id,

<sup>3</sup>rizqi.m.abdul\_si17@nusaputra.ac.id, <sup>4</sup>dado.firmansyah\_si17@nusaputra.ac.id

Korespondensi: <sup>2</sup>saeful.bahri\_si17@nusaputra.ac.id

### **ABSTRAK**

Tingkat pengangguran di Indonesia dari tahun ke tahun mengalami peningkatan dan penurunan sesuai dengan pertumbuhan penduduk. Tingkat pengangguran tertinggi di Indonesia selama tahun 2014 sampai 2019 terdapat di pulau Jawa yakni pada provinsi Banten. Kemudian untuk tingkat pengangguran terendah selama tahun 2014 sampai 2019 terdapat pada provinsi Bali. Penelitian ini menggunakan metode *K-Mean Clustering* dan implementasi program dengan bahasa pemrograman python. Dengan adanya penelitian ini diharapkan bisa memberikan solusi untuk pemetaan tingkat pengangguran di pulau Jawa dan Bali. Penelitian ini juga memberikan manfaat bagi pemerintah Indonesia agar mampu mengantisipasi tingkat pengangguran terdidik maupun tidak terdidik Hasil klasterisasi tingkat pengangguran pada pengujian *coding* bahasa pemrograman python yaitu terdapat 1 wilayah tingkat pengangguran rendah (C1), 3 wilayah tingkat pengangguran sedang (C2) dan 3 wilayah lagi dikategorikan sebagai tingkat pengangguran tinggi (C3). Dengan demikian, hasil penelitian ini diharapkan dapat membantu menjawab permasalahan tingkat pengangguran di pulau Jawa dan Bali dan juga bisa digunakan untuk mengklasterisasikan tingkat pengangguran rendah, sedang dan tinggi.

**Kata kunci : Algoritma *K-Means Clustering*, Python, Tingkat pengangguran**

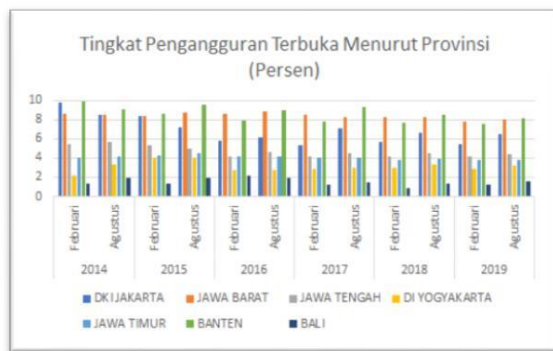
### **ABSTRACT**

*The unemployment rate in Indonesia from year to year has increased and decreased in accordance with population growth. The highest unemployment rate in Indonesia during 2014 to 2019 was on the island of Java, namely in the province of Banten. Then the lowest unemployment rate during 2014 to 2019 was in the province of Bali. This study uses the K-Mean Clustering method and program implementation with the python programming language. This research is expected to provide a solution for mapping the unemployment rate in Java and Bali. This research also provides benefits for the Indonesian government to be able to anticipate the level of educated and uneducated unemployment. The results of the clustering of the unemployment rate in the Python programming language coding test are 1 low unemployment rate area (C1), 3 medium unemployment rate regions (C2) and 3 more regions. categorized as high unemployment rate (C3). Thus, the results of this study are expected to help answer the problem of unemployment rates in Java and Bali and can also be used to classify low, medium and high unemployment rates.*

**Keywords : *K-Means Clustering Algorithm*, Python, Unemployment rate**

## I. PENDAHULUAN

Tingkat pengangguran di Indonesia dari tahun ke tahun mengalami peningkatan dan penurunan sesuai dengan pertumbuhan penduduk terutama pengangguran yang ada di pulau Jawa dan Bali. Data BPS menunjukkan bahwa selama 2004 hingga 2014 rata-rata pengangguran mempunyai laju penurunan mencapai 5%. Tingkat pengangguran tertinggi terjadi pada tahun 2005 yang mencapai 11,2% sementara tingkat pengangguran terendah dicapai pada tahun 2014 yang mencapai 5,9% atau telah turun sebanyak 47% sejak menyentuh angka pengangguran tertinggi pada tahun 2005 [1].



Gambar 1. Grafik tingkat pengangguran

Dari gambar 1 diatas dapat di jelaskan bahwa tingkat pengangguran yang terdapat di pulau Jawa dan Bali mengalami peningkatan dan penurunan dari setiap tahunnya, tingkat pengangguran tertinggi di Indonesia selama tahun 2014 sampai 2019 terdapat di pulau Jawa yakni pada provinsi Banten pada tahun 2014 bulan februari sebesar 9,87%, namun pada bulan agustus provinsi Banten terjadi penurunan menjadi 9,07%. Kemudian untuk tingkat pengangguran terendah selama tahun 2014 sampai 2019 terdapat pada provinsi Bali pada tahun 2018 bulan februari sebesar 0,88%, namun provinsi Bali terjadi peningkatan pada bulan agustus menjadi 1,40%.

Penelitian ini menggunakan metode K-Means yang memberikan solusi untuk pemetaan tingkat pengangguran di pulau Jawa dan Bali, dimana akan menentukan pusat pengangguran yang terjadi di daerah tertentu yang menjadi acuan, serta sebaran klaster pengangguran. Penelitian ini jugamenampilkan suatu wilayah tingkat pengangguran rendah di *cluster* pertama (C1), Untuk wilayah tingkat pengangguran sedang di *cluster* kedua (C2) dan wilayah dengan tingkat

pengangguran yang tinggi terdapat pada *cluster* ketiga (C3).

Penelitian ini memberikan manfaat bagi pemerintah Indonesia agar mampu mengantisipasi tingkat pengangguran terdidik maupun tidak terdidik, sehingga dengan ini di harapkan, dengan memperbanyak lapangan pekerjaan di berbagai sektor serta memberikan pelatihan-pelatihan bagi para calon pekerja guna, membekali calon pekerja untuk menjadi wirasuwasta. Selain itu dapat pula membatasi Tka/tenaga kerja asing, dengan cara hanya membatasi pekerja yang hanya dibutuhkan bagi sektor perusahaan tertentu. Kondisi ini tentunya akan mengurangi secara significant tingkat kriminalitas di pulau Jawa dan Bali.

Pada penelitian ini akan digunakan suatu pengolahan data dari tingkat pengangguran di pulau Jawa dan Bali menggunakan pengujian secara ilmiah. Pada pengujian data dalam penelitian ini akan lakukan menggunakan pengujian bahasa pemrograman python. Selain itu akan dilakukan juga penerapan sebuah algoritma data mining, yaitu *K-Means Clustering*.

Hasil penelitian ini diharapkan dapat membantu menjawab permasalahan tingkat pengangguran di pulau Jawa dan Bali dan juga bisa digunakan untuk mengklasterisasikan tingkat pengangguran rendah, sedang dan tinggi.

## II. TINJAUAN PUSTAKA

### 2.1 Data Mining

Data Mining Data Mining merupakan proses ekstraksi data menjadi informasi yang sebelumnya belum tersampaikan, dengan teknik yang tepat proses data mining akan memberikan hasil yang optimal [2]. Adapun pengelompokan data mining dapat dibagi menjadi empat kelompok, diantaranya yaitu model prediksi (*prediction modelling*), analisis kelompok (*Cluster analysis*), analisis asosiasi (*association analysis*) dan deteksi anomaly (*anomaly detection*). Riset ini salah satunya menggunakan algoritma *K-Means Clustering* yang termasuk pada tipe peran *clustering* pada data mining.

## 2.2 Clustering

Clustering merupakan salah satu metode Data Mining yang bersifat tanpa arahan (unsupervised), maksudnya metode ini diterapkan tanpa adanya latihan (training) dan tanpa ada guru (teacher) serta tidak memerlukan target output [3].

## 2.3 Konsep Pengangguran

Pengangguran adalah seseorang yang tergolong angkatan kerja dan ingin mendapat pekerjaan tetapi belum dapat memperolehnya. Masalah pengangguran yang menyebabkan tingkat pendapatan nasional dan tingkat kemakmuran masyarakat tidak mencapai potensi maksimal yaitu masalah pokok makro ekonomi yang paling utama [4].

## 2.5 Algoritma K-Means Clustering

Algoritma k-means adalah algoritma yang mempartisi data ke dalam cluster – cluster sehingga data yang memiliki kemiripan berada pada satu cluster yang sama dan data yang memiliki ketidaksamaan berada pada cluster yang lain [5].

Dalam algoritma K-Means, setiap data harus termasuk ke cluster tertentu pada suatu tahapan proses, pada tahapan proses berikutnya dapat berpindah ke cluster yang lain [6].

## III. METODOLOGI PENELITIAN

### 3.1 Pengumpulan Data

Pengelompokan menggunakan algoritma K-Means adalah sebagai berikut :

1. Jumlah  $k$  (*cluster*) dan *centriod* awal di tentukan secara acak.
2. Jarak data dan pusat *cluster* dihitung menggunakan *Euclidean Distance* yang di rumuskan sebagai berikut :

$$D(i,j) = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + \dots + (Xki - Xkj)^2}$$

Keterangan :

$D(i,j)$  = Jarak data ke  $i$  ke pusat *cluster*

$j$   $Xki$  = Data ke  $i$  pada atribut data ke  $k$

$Xkj$  = Titik pusat ke  $j$  pada atribut ke  $k$

3. Mengelompokan data kedalam *cluster* yang dengan jarak paling dekat.

4. Mengitung Pusat *cluster* yang baru dengan mencari nilai rata-rata pada kelompok data.

5. Pusat *cluster* akan ditentukan bila semua data telah ditetapkan dalam *cluster* terdekat.

Proses penentuan pusat *cluster* dan penempatan data dalam *cluster* di ulangi sampai nilai *centroid* tidak berubah lagi.

## IV. HASIL DAN PEMBAHASAN

### 4.1 Pengolahan Data

1. Data Sampel

Data sampel yang digunakan yaitu data tingkat pengangguran di pulau jawa dan bali pada tahun 2014 sampai 2019. Jumlah data yang didapatkan adalah 7 data, yang mana data ini adalah data dari setiap provinsi selama 6 (Enam) bulan sekali.

2. *Preprocessing*

Setelah didapatkan data tingkat pengangguran di pulau Jawa dan Bali dengan kondisi yang perlu dilakukan penyeleksian ulang, maka sesuai dengan hasil kajian pustaka diperlukan adanya tahap *preprocessing* data sebelum dilakukan tahap pengimplementasian pada sistem analisis. Adapun tahapan *preprocessing* data yang dilakukan sebagai berikut ini:

- a. Seleksi Atribut

Dari beragamnya atribut pada data awal, setelah dilakukan penyeleksian atribut maka yang digunakan pada penelitian ini ada 3 (Tiga) atribut yaitu, data provinsi di Pulau Jawa dan Bali, data pengangguran pada bulan Februari dan Agustus.

- b. Pembersihan Data (*Data Cleaning*)

Setelah 3 (Tiga) atribut ditentukan, maka dilakukan juga pembersihan data yang tidak sesuai sebelum proses pengolahan data lebih lanjut. Seperti data yang salah input karna kekeliruan (*missing value*) atau juga data yang sama sekali tidak berarti (*noise data*).

- c. Transformasi Data

Dataset dari hasil proses seleksi atribut dan pembersihan data sebelumnya, selanjutnya dilakukan perubahan data yang memiliki tipe data yang tidak dapat diolah secara matematis menjadi data yang dapat diolah

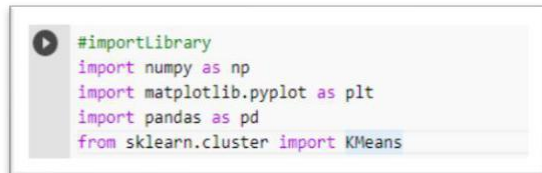
### 4.2 K-Means Clustering

1. Langkah awal, dilakukan *upload* dataset seperti pada gambar 3 dibawah ini.



Gambar 3. Proses Upload File Data

2. Kemudian dilakukan *import library* sesuai kebutuhan proses pengujian sistem.



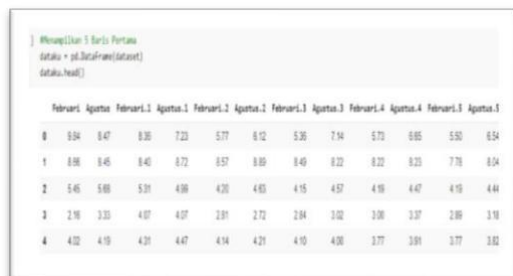
Gambar 4. Proses Import Library

3. Setelah dilakukan *upload file* dataset dilakukan penyimpanan dataset tersebut ke dalam suatu variabel. Kemudian dilakukan pemanggilan dataset yg sudah disimpan ke dalam variabel seperti pada gambar 5 dibawah ini :



Gambar 5. Penyimpanan dan Pemanggilan Dataset

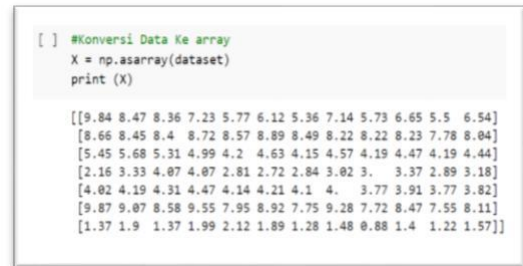
4. Tahap selanjutnya menampilkan terlebih dahulu 5 baris pertama untuk memastikan dataset yang digunakan sesuai dengan dataset yang diperlukan.



Gambar 6. (Lima) Dataset Baris Pertama

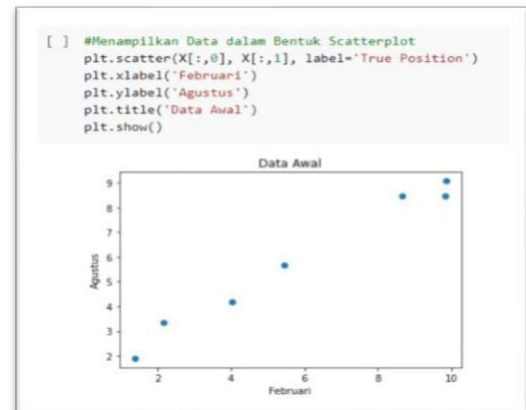
5. Setelah dataset sudah sesuai untuk dilakukan implementasi terhadap sistem, selanjutnya dilakukan konversi data terlebih dahulu ke dalam bentuk array dengan bantuan *library* Numpy agar dataset tersebut bisa diolah

berdasarkan ketentuan sistem yang berlaku. Seperti yang disajikan pada gambar 7 dibawah ini.



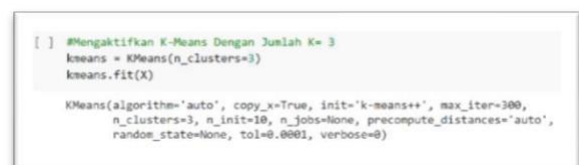
Gambar 7. Proses Konversi Data

6. Agar memudahkan dalam menerjemahkan dataset yang akan diolah, kemudian dilakukan visualisasi dalam bentuk scatterplot dengan bantuan *library* Matplotlib seperti pada gambar yang bisa di lihat pada gambar 8 dibawah ini :



Gambar 8. Proses Menampilkan Data Dalam Bentuk Scatterplot

7. Tahap selanjutnya yaitu mengaktifkan algoritma *K-Means Clustering* dengan bantuan Scikit Learn dan dilakukan juga penentuan jumlah *cluster* . Seperti yang terlihat pada gambar 9 di bawah ini.



Gambar 9. Proses Penerapan Algoritma K-Means

8. Tahap selanjutnya melihat hasil nilai *centroid* yang pada setiap *cluster* yang dilakukan pada tahap sebelumnya. Terlihat hasilnya pada gambar 10 berikut :

```
[ ] #Menampilkan Nilai Centroid
print(kmeans.cluster_centers_)

[[1.37      1.9      1.37      1.99      2.12      1.89
  1.28      1.48      0.88      1.4      1.22      1.57
  9.45666667 8.66333333 8.44666667 8.5      7.43      7.97666667
  7.2      8.21333333 7.22333333 7.78333333 6.94333333 7.56333333
  3.87666667 4.4      4.56333333 4.51      3.71666667 3.85333333
  3.69666667 3.86333333 3.65333333 3.91666667 3.61666667 3.81333333]]
```

Gambar 10. Menampilkan Hasil Nilai *Centroid*

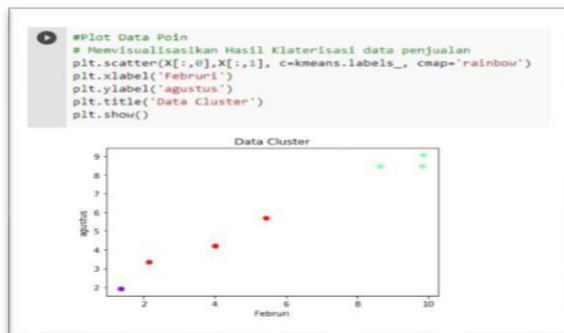
9. Selain melihat hasil nilai *centroid* perlu juga melihat hasil pelabelan data hasil pengolahan yang dilakukan sebelumnya. Proses dan hasilnya bisa dilihat pada gambar di bawah :

```
[ ] #Menampilkan Label data poin
print(kmeans.labels_)

[1 1 2 2 2 1 0]
```

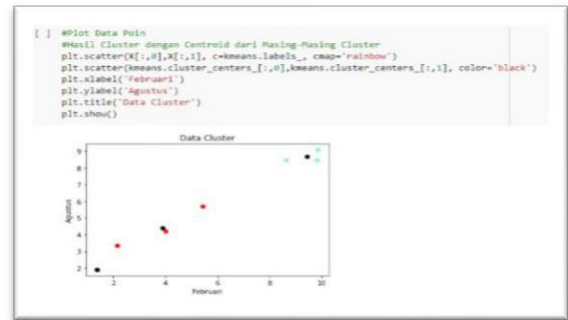
Gambar 11. Tampilan Label Data Point

10. Dari hasil yang didapatkan pada tahap sebelumnya, kemudian dibuatkan warna yang berbeda untuk setiap *cluster* nya. Proses dan hasil visualisasi scatterplot bisa dilihat pada gambar dibawah 12 ini.



Gambar 13. Proses dan Hasil Klasterisasi Pengolahan Data

11. Setelah melihat visualisasi data hasil klasterisasi dari pengolahan data sebelumnya, selanjutnya perlu ditampilkan juga hasil klasterisasi dan juga *centroid* dari masing-masing *cluster* yang dihasilkan. Seperti yang terlihat pada gambar yang di sajikan di bawah 14 ini.



Gambar 14. Tampilan Hasil Klasterisasi dan *centroid* Pengolahan Data

12. Langkah selanjutnya yaitu melakukan penambahan data dari *cluster* yang di dapatkan dengan dataset awal, Berikut proses dan hasil dari penambahan cluster kedalam dataset.

```

]] Menambahkan ke dalam cluster pada dataset
dataset['Cluster'] = kmeans.labels_
dataset

```

	Februari	Agustus	Februari.1	Agustus.1	Februari.2	Agustus.2	Februari.3	Agustus.3	Februari.4	Agustus.4	Februari.5	Agustus.5	Cluster
0	9.84	8.47	8.36	7.23	5.77	6.12	5.36	7.14	5.73	6.65	5.59	6.54	1
1	6.99	6.45	6.40	6.72	6.57	6.99	6.48	6.22	6.23	7.79	6.94	1	
2	5.45	5.58	5.31	4.99	4.25	4.53	4.15	4.57	4.19	4.47	4.19	4.44	2
3	2.16	3.33	4.07	4.07	2.81	2.72	2.84	3.02	3.00	3.37	2.84	3.18	2
4	4.02	4.19	4.21	4.47	4.14	4.21	4.10	4.06	3.77	3.91	3.77	3.82	2
5	9.07	9.07	8.58	9.55	7.95	8.52	7.75	9.28	7.72	8.47	7.55	8.11	1
6	1.37	1.96	1.37	1.96	2.12	1.99	1.28	1.40	0.88	1.40	1.22	1.57	0

Gambar 15. Hasil Penambahan *Cluster* Terhadap Dataset Awal

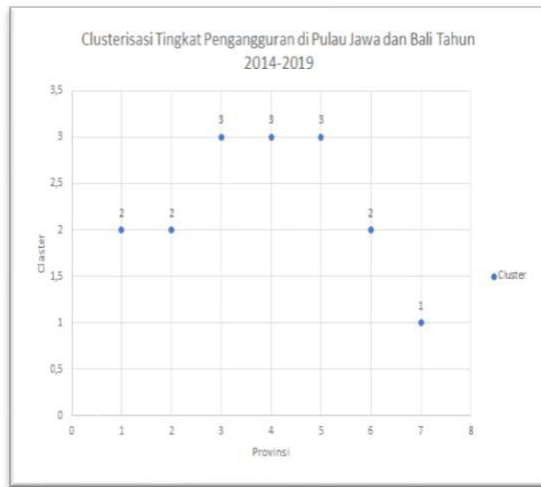
13. Bagian terakhir yaitu dilakukan pengunduhan data hasil dari penggabungan dataset awal dengan *cluster* yang didapatkan dari hasil pengolahan data yang dilakukan pada implementasi bahasa pemrograman python ini. Sehingga data ini yang akan didapatkan informasi lebih lengkap dan bermanfaat. Berikut tahap yang dilakukan pada saat pengunduhan data tersebut.

```
[ ] #Download Hasil Data
dataset.to_csv('hasildatanggur.csv', index=False)
```

Gambar 16. Proses Pengunduhan Hasil Dataset



Dari hasil penelitian yang diuraikan diatas maka interpretasi hasil klaster akhir yaitu sebagai berikut.



Gambar 17. Grafik Hasil Klasterisasi Berdasarkan Pengolahan Python

## V. PENUTUP

### 5.1 Kesimpulan

Penggunaan K-Means Clustering menggunakan beberapa tahapan diantaranya Jumlah  $k$  (*cluster*) dan *centriod* awal di tentukan secara acak, Jarak data dan pusat *cluster*, Mengelompokan data kedalam *cluster* yang dengan jarak paling dekat, Mengitung Pusat *cluster* yang baru dengan mencari nilai rata-rata pada kelompok data, Pusat *cluster* akan ditentukan bila semua data telah ditetapkan dalam *cluster* terdekat, Proses penentuan pusat *cluster* dan penempatan data dalam *cluster* di ulangi sampai nilai *centroid* tidak berubah lagi. Hasil klasterisasi terhadap pengujian data penelitian bahasa pemrograman python berhasil diselesaikan dengan hasil seperti pada gambar 2.11, yaitu warna biru menunjukan kelompok wilayah yang termasuk *cluster* 1 (C1) dengan jumlah 1 wilayah, kemudian kelompok berwarna merah termasuk *cluster* 2 (C2) dengan jumlah 3 wilayah, dan untuk warna hijau termasuk *cluster* 3 (C3) dengan jumlah 3 wilayah.

### 5.2. Saran

Penentuan metode pengukuran jarak data sebaiknya disarankan untuk menggunakan  $K$ -

*Means Clustering* terhadap pengujian data penelitian menggunakan bahasa pemrograman python, yaitu *cluster* pertama (C1) menyatakan bahwa terdapat 1 wilayah di kategori tingkat pengangguran rendah, *cluster* kedua (C2) menyatakan bahwa terdapat 3 wilayah di kategori tingkat pengangguran sedang, dan *cluster* ketiga (C3) menyatakan bahwa terdapat 3 wilayah dikategorikan sebagai tingkat pengangguran tinggi (C2) . Dari hasil tiga *cluster* tersebut masih bisa dilakukan penyempitan kategori dengan menambahkan *cluster-cluster* baru sehingga akan didapatkan kategori yang lebih spesifik.

## DAFTAR PUSTAKA

- [1] D. Setiawan, Muchtar, H. Muafiqie. "Faktor-Faktor Determinan yang Berpengaruh Pada Tingkat Pengangguran Di Indonesia Periode 2000 2016", Journal of public power, Vol.1,No.1 2017.
- [2] G.Abdillah, F.A. Putra, F. Renaldi. "Penerapan Data Mining Pemakaian Air Pelanggan untuk Menentukan Klasifikasi Potensi Pemakaian Air Pelanggan Baru di PDAM Tirta Raharja Menggunakan Algoritma K-Means", Seminar Nasional Teknologi Informasi dan Komunikasi 2016 (SENTIKA 2016), Yogyakarta, 18-19 Maret 2016.
- [3] S. T. Siska, "Analisa dan Penerapan Data Mining untuk Menentukan Kubikasi Air Terjual Berdasarkan Pengelompokan Pelanggan Menggunakan Algoritma K-Means Clustering", Jurnal Teknologi Informasi & Pendidikan, VOL. 9 NO. 1 April 2016.
- [4] Muhdar HM. "Potret Ketenagakerjaan, Pengangguran, dan kemiskinan di Indonesia: Masalah dan Solusi", Al-Buhuts, Volume 11 Nomor 1 Juni 2015, Halaman 42-66.
- [5] N. Rohmawati, S. Defiyanti, M. Jajuli, "Implementasi Algoritma K-Means dalam Pengklasteran Mahasiswa Pelamar Beasiswa", Jurnal Ilmiah Teknologi Informasi Terapan, Volume I, No 2, 30 April 2015.
- [6] S. Ghosh, S. K. Dubey. "Comparative Analysis of K-Means and Fuzzy C-Means Algorithm", India, 2013.