Prosiding Seminar Nasional Matematika dan Pendidikan Matematika (SESIOMADIKA) 2017 |
ISBN: 978-602-60550-1-9
Statistika, hal. 19-27

IMPLEMENTASI METODE IMPUTASI *MEAN* DAN *EXPECTATION MAXIMISATION* TERHADAP HASIL *CLUSTERING K-MEANS* MAHASISWA PELAMAR BEASISWA FAKULTAS ILMU KOMPUTER UNIVERSITAS SINGAPERBANGSA KARAWANG

MOHAMAD JAJULI[1], **OMAN KOMARUDIN**[2],

[1]Universitas Singaperbangsa Karawang, mohamad.jajuli@unsika.ac.id
[2]Universitas Singaperbangsa Karawang, oman.komarudin@unsika.ac.id

Abstrak. Biaya pendidikan yang semakin hari semakin mahal menjadikan masyarakat Indonesia tidak semua dapat menempuh pendidikan sampai tingkat universitas padahal memperoleh pendidikan merupakan hak setiap warga negara Indonesia sesuai dengan Undang Undang Dasar Negara Republik Indonesia Tahun 1945 pasal 31 ayat (1). Pemerintah bertanggung jawab memberikan bantuan kepada siswa yang berbakat dan berprestasi dari kalangan ekonomi kurang mampu dalam bentuk beasiswa agar dapat melanjutkan pendidikan ke jenjang yang lebih tinggi. Di setiap lembaga pendidikan khususnya perguruan tinggi banyak sekali beasiswa yang tersedia. Untuk memperoleh beasiswa tersebut tentunya harus sesuai dengan kriteria-kriteria yang telah ditetapkan. *Data mining* merupakan suatu teknik penggalian informasi dari data – data yang berukuran besar. Teknik ini dapat membantu dalam memprediksi calon penerima beasiswa dengan algoritma tertentu yang dapat mengklasifikasi mahasiswa yang berhak menerima beasiswa, mahasiswa yang di pertimbangkan dan mahasiswa yang tidak berhak menerima beasiswa yaitu adalah analisis *Cluster*. Pada beberapa kasus analisis *Cluster* terdapat *missing data* dalam dataset yang digunakan. Salah satu metode yang dapat digunakan untuk penanganan missing data yaitu teknik imputasi (*imputation technique*). Penelitian ini mencoba mengimplementasikan metode imputasi *Mean* dan *Expectation Maximisation* untuk penanganan *missing data* pada kasus penerimaan beasiswa di Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang. Data yang digunakan dalam penelitian ini adalah data mahasiswa yang mengajukan beasiswa kepada Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang angkatan 2008-2011. Hasil clustering K-Means sebanyak 2 *cluster* baik imputasi *Mean* dan *Expectation Maximisation* memiliki sebaran data yang sama sedangkan hasil *clustering* sebanyak 3 *cluster* memiliki sebaran data yang berbeda. Untuk imputasi *Mean* menghasilkan *cluster* 1 sebanyak 75% data, *cluster* 2 sebanyak 13,9% data, dan *cluster* 3 sebanyak 11,1% data. Untuk imputasi *Expectation Maximisation* yang terbentuk sebanyak 3 *cluster* dengan sebaran data di *cluster* 1 sebanyak 13,9%% data, *cluster* 2 sebanyak 75% data, dan *cluster* 3 sebanyak 11,1% data. Evaluasi dari clustering K-Means menunjukkan *Clustering K-Means* sebanyak 3 *cluster* yang terbentuk untuk penanganan missing data menggunakan metode imputasi *Mean* lebih baik dibandingkan *Expectation Maximisation* karena nilai *Cubic Clustering Criterion* dan *Pseudo* F untuk *Mean* lebih tinggi dibandingkan *Expectation Maximisation*.

Kata kunci: Imputasi, *Mean, Expectation Maximisation*, K-Means.

1. Pendahuluan

Seiring perkembangan zaman dan semakin majunya teknologi informatika sekarang ini menuntut masyarakat Indonesia untuk mempunyai pendidikan yang tinggi. Ada beberapa faktor yang menghambat masyarakat Indonesia untuk mendapatkan pendidikan yang lebih tinggi lagi, salah satunya adalah faktor biaya. Biaya pendidikan yang semakin hari semakin mahal menjadikan masyarakat Indonesia tidak semua dapat menempuh pendidikan sampai tingkat universitas padahal memperoleh pendidikan merupakan hak setiap warga negara Indonesia sesuai dengan Undang Undang Dasar Negara Republik Indonesia Tahun 1945 pasal 31 ayat (1). Pemerintah bertanggung jawab kepada masyarakat dalam memberikan pendidikan yang layak sesuai yang tertuang dalam Undang Undang Dasar Negara Republik Indonesia Tahun 1945 pasal 33 ayat (3). Pemerintah juga bertanggung jawab memberikan bantuan kepada siswa yang berbakat dan berprestasi dari kalangan ekonomi kurang mampu dalam bentuk beasiswa agar dapat melanjutkan pendidikan ke jenjang yang lebih tinggi. Dengan adanya beasiswa ini diharapkan siswa dapat menyelesaikan pendidikannya tanpa ada gangguan terutama yang berhubungan dengan keuangan hingga tuntas atau lulus di jenjang pendidikan (Zuwida dkk, 2014:390). Di setiap lembaga pendidikan khususnya perguruan tinggi banyak sekali beasiswa yang tersedia. Untuk memperoleh beasiswa tersebut tentunya harus sesuai dengan kriteria-kriteria yang telah ditetapkan, seperti jumlah penghasilan orang tua, jumlah tanggungan orang tua, jumlah saudara kandung, nilai rata-rata, dan persentase kehadiran siswa (Gunawan dkk, 2013:89). Dengan semakin banyaknya kriteria-kriteria yang ditetapkan dan banyak mahasiswa pendaftar beasiswa membuat kesulitan di dalam menentukan penerima beasiswa.

Data mining merupakan suatu teknik penggalian informasi dari data-data yang berukuran besar. Teknik ini dapat membantu dalam memprediksi calon penerima beasiswa dengan algoritma tertentu yang dapat mengklasifikasi mahasiswa dan salah satu algoritma dalam analisis Cluster yang bisa digunakan adalah K-Means. Pada penelitian yang dilakukan oleh Nurul Rohmawati W, Sofi Defiyanti, Mohamad Jajuli (2015) tentang implementasi algoritma *K-Means* dalam pengklasteran mahasiswa pelamar beasiswa mahasiswa Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang menjadi 3 *cluster* yaitu *cluster* 1 yang berhak menerima beasiswa, *cluster* 2 mahasiswa yang di pertimbangkan menerima beasiswa, dan *cluster* 3 mahasiswa yang tidak berhak menerima beasiswa. Hasil dari algoritam *K-Means* di cluster 1 terdapat 52,78% data, cluster 2 terdapat 22,22%, dan cluster 3 terdapat 25% data. Di penelitian yang dilakukan Nurul Rohmawati W terdapat data missing sebanyak 33,3% dari total data yang akan digunakan sehingga dapat memberikan pengaruh dari hasil clusteringnya.

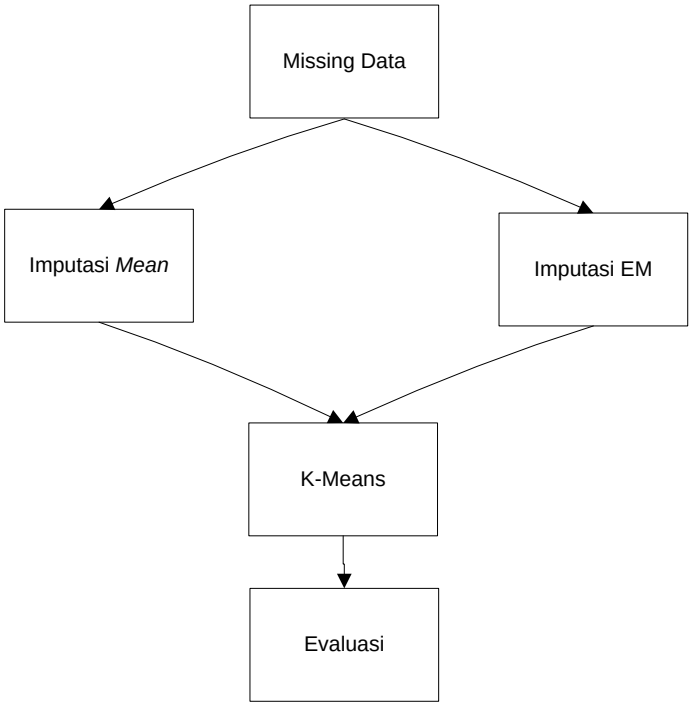
Missing data merupakan kelemahan umum pada banyak skenario klasifikasi pola (Laencina dkk, 2009) dan salah satu masalah yang dapat mempengaruhi hasil dari sistem prediksi data yang efektif (Malarvizhi dkk, 2012). Salah satu metode yang dapat digunakan untuk penanganan missing data yaitu teknik imputasi (*imputation technique*). Teknik imputasi adalah metode penanganan *missing data* berdasarkan informasi yang tersedia pada dataset yang bertujuan memprediksi nilai yang valid sebagai pengganti nilai yang hilang. Beberapa penelitian sudah pernah dilakukan tentang teknik imputasi, salah satunya penelitian yang dilakukan oleh Rahmawati dan Yohanes Eki Aprilawan. Penelitian Rahmawati (2012) tentang perbandingan metode regresi dan *expectation maximization* (EM) dalam mengisi data missing.

Hasil penelitian ini menunjukkan EM metode lebih baik dibandingkan dengan metode regresi dalam memperkirakan data yang hilang. Penelitian yang dilakukan oleh Yohanes Eki Aprilawan (2015) tentang teknik imputasi missing values pada data mining: studi kasus pada data hepatitis yang menyatakan bahwa metode imputasi yang paling efektif adalah metode imputasi *mean* dan modus, metode yang cukup efektif adalah *K-Nearest Neighbor*, dan metode yang kurang efektif adalah metode *Singular-*

Value Decomposition. Kefektifan ini diukur dengan menggunakan akurasi dari *class* yang di prediksi. Penelitian lain dilakukan oleh Sudirman tentang analisis perbandingan metode imputasi missing values global dan concept method pada data supervised yang menyimpulkan bahwa teknik penanganan missing value bergantung dari karakteristik data, beberapa jumlah atribut, record, dan jumlah missing value. Hasil akurasi klasifikasi yang diperoleh juga berbeda-beda dan memberikan hasil klasifikasi yang berbeda pula. Hal ini berarti teknik penanganan missing value dan algoritma klasifikasi yang digunakan menentukan besarnya akurasi yang diperoleh. Penelitian ini mencoba mengimplementasikan metode imputasi *Mean* dan *Expectation Maximisation* untuk penanganan *missing data* pada kasus penerimaan beasiswa di Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang dan mengevaluasi hasil *cluster* K-Meansnya.

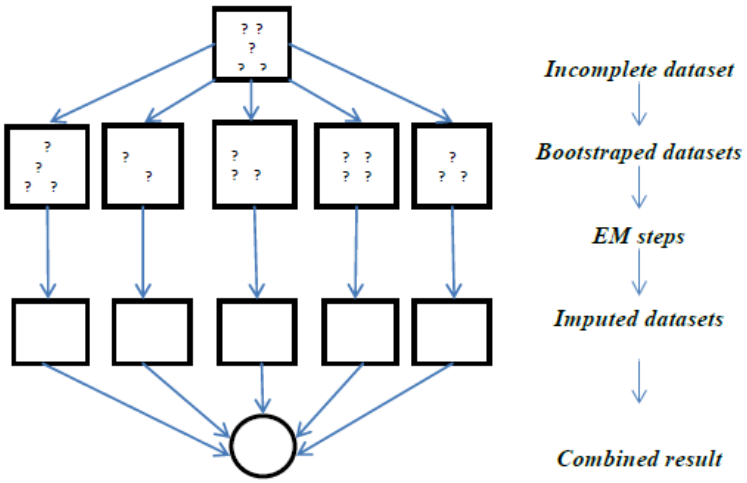
2. Metode

Data yang digunakan dalam penelitian ini adalah data mahasiswa pelamar beasiswa di Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang angkatan 2008-2011. Data ini terdiri dari 3 atribut yaitu data nilai IPK mahasiswa, penghasilan orang tua dan jumlah tanggungan orang tua yang diambil di bagian akademik Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang. Langkah-langkah yang ada dalam penelitian ini dapat dilihat pada Gambar 1.



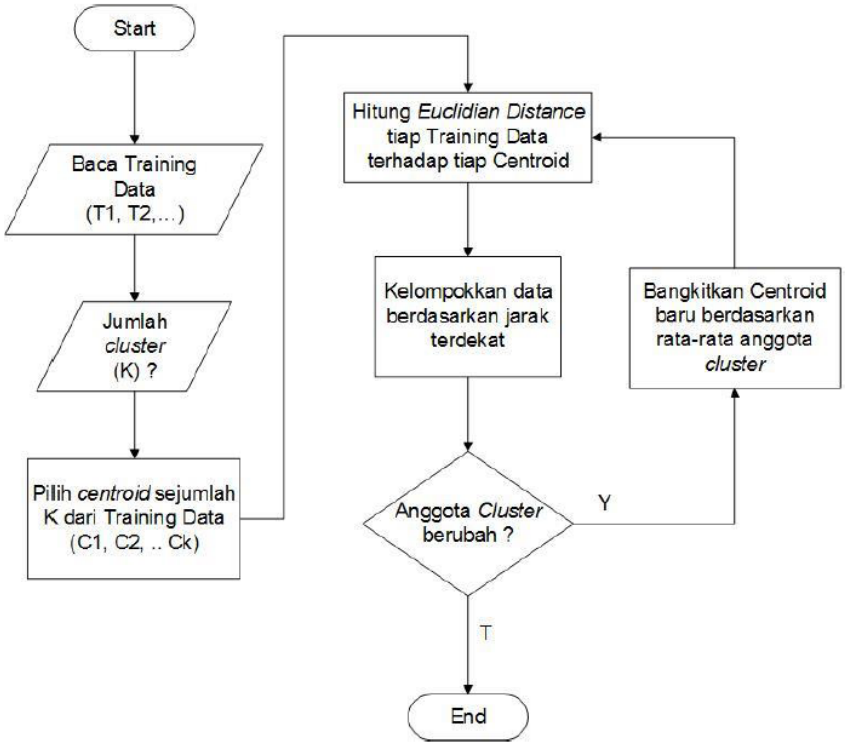
Gambar 1. Diagram Alir Penelitian.

Tahapan pertama yang akan dilakukan adalah mengidentifikasi *missing* data dari masing-masing atribut, Setelah itu dilakukan proses imputasi. Metode imputasi yang digunakan adalah imputasi *Mean* dan *Expectation Maximisation*. Untuk proses imputasi *Mean* adalah mengganti nilai *missing* data dengan nilai rata-rata pada data sedangkan proses imputasi *Expectation Maximisation* dapat dilihat pada Gambar 2.



Gambar 2. Langkah Imputasi *Expectation Maximisation.*

Setelah *missing* data sudah terisi dari masing-masing metode imputasi, selanjutnya dilakukan proses algoritma K-Means. Proses algoritma K-Means dapat dilihat pada Gambar 3.



Gambar 3. Langkah Algoritma *K-Means.*

Tahap akhir dari diagram alir penelitian ini yaitu evaluasi hasil *cluster* K-Means menggunakan *Cubic Clustering Criterion* (CCC) dan *Pseudo* F. Nilai CCC merupakan perbandingan koefisien nilai pengamatan dari R^2 dengan pendekatan nilai harapan dari R^2. CCC dapat dihitung menggunakan rumus sebagai berikut:

$$CCC = ln\left[\frac{1-E(R^2)}{1-R^2}\right]\frac{\left(\frac{np^*}{2}\right)^{1/2}}{(0.001+E(R^2))^2}$$

Keterangan:
R^2 : Keragaman yang dapat dijelaskan cluster
$E(R^2)$: Nilai harapan dari R^2
n : Jumlah pengamatan
p*<p, p: Jumlah peubah

Nilai CCC lebih dari 2 atau 3 mengindikasikan bahwa *cluster* yang terbentuk bagus, nilai CCC antara 0 dan 2 menunjukkan bahwa *cluster* yang terbentuk potensial, sedangkan nilai CCC negatif yang besar menunjukkan adanya pencilan. *Pseudo* F merupakan fungsi dari jumlah *cluster* yang dihasilkan dalam setiap langkah pengclusteran. Fungsi dari pseudo F adalah sebagai berikut:

$$F = \frac{SSB/(g-1)}{SSW/(n-g)}$$

Keterangan:
SSB : Jumlah kuadrat jarak antar *cluster*
SSW : Jumlah kuadrat jarak dalam *cluster* l
g : Banyak *cluster* yang dihasilkan
n : Banyak pengamatan

Nilai pseudo F maksimum menunjukkan banyaknya *cluster* yang optimal.

2. Hasil dan Pembahasan

Data mahasiswa pelamar beasiswa di Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang angkatan 2008-2011 yang terdapat missing data ada di atribut penghasilan orang tua. Untuk lebih jelas dapat dilihat pada Tabel 1.

Tabel 1. Beberapa Data Mahasiswa Pelamar Beasiswa

NPM	IPK	Jumlah Tanggungan Orang Tua	Penghasilan Orang Tua	
09 – 125	3.21	2	Rp	-
10 – 109	3.23	1	Rp	-
09 – 053	3.00	3	Rp	-
08 – 025	3.73	1	Rp	-
09 – 015	3.19	4	Rp	2,000,000
09 – 009	2.47	2	Rp	-
09 – 104	2.38	2	Rp	-
08 – 018	3.35	2	Rp	2,200,000
08 – 008	3.30	4	Rp	2,000,000
08 – 021	3.02	5	Rp	2,065,200
...	...	...		...

Missing data di atribut penghasilan orang tua ada sebesar 33.3%. Dalam penelitian, kehilangan 20-30% data bisa mempengaruhi hasil analisis dari penelitian (Little & Rubin, 2002). Bila hal ini dibiarkan, kemudian dianalisis tentu akan menimbulkan bias yang cukup besar dan hasil analisis yang bisa menyesatkan. Penanganan *missing data* yang pertama menggunakan metode imputasi *Mean* dengan data yang sudah di isi missing value dapat dilihat pada tabel 2.

Tabel 2. Data Menggunakan Metode Imputasi *Mean*

NPM	IPK	Jumlah Tanggungan Orang Tua	Penghasilan Orang Tua
09 – 125	3.21	2	Rp 2,021,575
10 – 109	3.23	1	Rp 2,021,575
09 – 053	3.00	3	Rp 2,021,575
08 – 025	3.73	1	Rp 2,021,575
09 – 015	3.19	4	Rp 2,000,000
09 – 009	2.47	2	Rp 2,021,575
09 – 104	2.38	2	Rp 2,021,575
08 – 018	3.35	2	Rp 2,200,000
08 – 008	3.30	4	Rp 2,000,000
08 – 021	3.02	5	Rp 2,065,200
...	...	...	...

Penanganan *missing data* yang kedua menggunakan metode imputasi *Expectation Maximisation* dengan hasil yang dapat dilihat pada tabel 3.

Tabel 3. Data Menggunakan Metode Imputasi *Expectation Maximisation*

NPM	IPK	Jumlah Tanggungan Orang Tua	Penghasilan Orang Tua
09 – 125	3.21	2	Rp 2,106,109
10 – 109	3.23	1	Rp 2,231,184
09 – 053	3.00	3	Rp 1,933,922
08 – 025	3.73	1	Rp 2,355,161
09 – 015	3.19	4	Rp 1,860,917
09 – 009	2.47	2	Rp 1,922,622
09 – 104	2.38	2	Rp 1,900,306
08 – 018	3.35	2	Rp 2,200,000
08 – 008	3.30	4	Rp 2,000,000
08 – 021	3.02	5	Rp 2,065,200
...	...	...	...

Setelah missing value sudah di isi dengan metode imputasi *Mean* dan *Expectation Maximisation* maka tahapan berikutnya adalah mengelompokan data pelamar beasiswa menggunakan algoritma *K-Means* dengan jumlah *cluster* (*k*) yang dibentuk pada penelitian ini adalah 2 dan 3 *cluster*. Hasil *clustering* K-Means dapat dilihat di Tabel 4.

Tabel 4. Hasil *Clustering K-Means*

Cluster	Imputasi *Mean*	Imputasi *Expectation Maximisation*
1	88,9%	88,9%
2	11,1%	11,1%

Cluster	Imputasi *Mean*	Imputasi *Expectation Maximisation*
1	75,0%	13,9%
2	13,9%	75,0%
3	11,1%	11,1%

Hasil *clustering* sebanyak 2 *cluster* baik imputasi *Mean* dan *Expectation Maximisation* memiliki sebaran data yang sama sedangkan hasil *clustering* sebanyak 3 *cluster* memiliki sebaran data yang berbeda. Untuk imputasi *Mean* paling banyak sebaran data di cluster 1 sebesar 75,0% sedangkan di imputasi *Expectation Maximisation* sebaran data paling banyak di cluster 2 sebesar 75,0%. Jika menggunakan hasil *clustering Mean* maka sebesar 75% mahasiswa pelamar beasiswa berhak menerima beasiswa sedangkan jika menggunakan hasil clustering *Expectation Maximisation* maka yang sebesar 75% mahasiswa pelamar beasiswa dipertimbangkan menerima beasiswa. Oleh karena itu diperlukan penentuan banyak cluster yang paling tepat menggunakan *Cubic Clustering Criterion* (CCC) dan nilai statistik pseudo F yaitu melihat rasio keseragaman antar *cluster* dan keragaman dalam *cluster*. Nilai CCC dan *pseudo* F disajikan pada Tabel 5.

Tabel 5. Nilai CCC dan *Pseudo* F

Cluster	Imputasi *Mean*		Imputasi *Expectation Maximisation*	
	CCC	*Pseudo* F	CCC	*Pseudo* F
2	-3.294	48.40	-3.953	39.61
3	4.704	444.39	3.047	309.43

Tabel 5 menunjukkan bahwa metode imputasi *Mean* dengan 3 *cluster* memiliki nilai CCC dan *pseudo* F yang lebih tinggi dibandingkan imputasi *Expectation Maximisation*. Hal ini menandakan bahwa untuk hasil *cluster* K-Means dengan 3 *cluster* yang dibuat pada kasus mahasiswa pelamar beasiswa Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang metode imputasi *Mean* lebih baik di dalam penanganan missing data dibandingkan *Expectation Maximisation*.

3. Kesimpulan

Hasil *Clustering K-Means* yang terbentuk sebanyak 3 *cluster* untuk penanganan missing data menggunakan metode imputasi *Mean* dan *Expectation Maximisation* memiliki sebaran data yang berbeda-beda di *cluster* 1 dan 2, sedangkan di *cluster* 3 memiliki sebaran data yang sama di kedua metode imputasi sebanyak 11,1%. Hasil evaluasi dari clustering K-Means menunjukkan *Clustering K-Means* yang terbentuk sebanyak 3 *cluster* menggunakan metode

imputasi *Mean* lebih baik dibandingkan *Expectation Maximisation* karena nilai CCC dan *pseudo* F-nya yang lebih tinggi.

Pernyataan terimakasih. Ucapan Terima kasih disampaikan kepada Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang dan LPPM Universitas Singaperbangsa Karawang atas support yang sudah diberikan sehingga terselesaikannya penelitian ini.

Referensi

[1] Acuna, dkk. (2004). The Treatment of Missing Values and It's Effect on Classifier Accuracy, Classification, Clustering, and Data Mining Ahallications. *Springer Berlin Heidelberg*: 12-15.
[2] Azwar Rizal A., Handayani T., Isye A. (2013). Perbandingan Performa Antara Imputasi Metode Konvensional dan Imputasi dengan Algoritma Mutual Nearest Neighbor. *Jurnal Teknik POMITS,* 2(1): A73-A76.
[3] Gunawan, dkk. (2013). Pengembangan Sistem Penunjang Keputusan Penentuan Pemberian Beasiswa Tingkat Sekolah. *Jurnal SIFO,* 14(2): 89-98.
[4] Hastuti, N. F., Saptono, R., Suryani, E. (2012). Pemanfaatan Metode K-Means Clustering Dalam Penentuan Penerima Beasiswa. *Jurnal Informatika*: 2.
[5] Huda, N. M. (2010). *Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Studi Kasus di Fakultas MIPA Universitas Diponegoro)*. [Skripsi]. Hal:10.
[6] Laencina, G., dkk. (2009). K-Nearest Neighbour with Mutual Information for Simultaneous Classification and Missing Data Imputation. *Nerocomputing,* 72: 1483-1493.
[7] Lutfhi, E. T., & Kusrini. (2009). *Algortima Data Mining.* Yogyakarta: ANDI.
[8] Malarvizhi, T. (2012). K-mn Classifier Performs Better than K-Means Clustering in Missing Value Imputation. *IOSR Journal of Computer Engineering,* 6(5): 12-15.
[9] Montgomery, D. C., and E. A. Peck. (1992). *Introduction to Linear Regression Analysis*. 2nd Ed. New York: John Wiley & Sons.
[10] Nango, D. N. (2012). *Penerapan Algoritma K-Means Untuk Clustering Data Anggaran Pendapatan Belanja Daerah di Kabupaten XYZ*. [Skripsi]. Hal: 11-12.
[11] Nurul Rohmawati W., Sofi Defiyanti, Mohamad Jajuli. (2015). Implementasi Algoritma K_Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa. *Jurnal Ilmiah Teknologi Informasi Terapan,* 1(2): 62-68.
[12] Rahmawati. (2012). Perbandingan Metode Regresi dan Expectation Maximization (EM) Dalam Mengisi Data Missing. *Jurnal Penelitian Akademi Kesehatan Rajekwesi Bojonegoro,* 6(3): 7-10.
[13] Sarwono, Y. T. (2012). Aplikasi Model Jaringan Syaraf Tiruan Dengan Radial Basis Function Untuk Mendeteksi Kelainan Otak (Stroke Infark). *Jurnal Sistem Informasi*: 3-4.
[14] Sofi Defiyanti, Mohamad Jajuli, Nurul Rohmawati W. (2017). Optimalisasi K-Medoid Dalam Pengklasteran Mahasiswa Pelamar Beasiswa Dengan Cubic Clustering Criterion. *Jurnal Teknologi dan Sistem Informasi (TEKNOSI),* 3(4): 211-218.
[15] Sudirman. (2012). Analisis Perbandingan Metode Imputasi Missing Values Global dan Concept Method Pada Data Supervised. *Jurnal Ilmiah Fakultas Ilmu Komputer Universitas Mercubuana (FIFO),* 4(2): 135-142.
[16] Triyani, Hendrawati. 2015. Kajian Metode Imputasi Dalam Menangani Missing Data. *Prosiding Seminar Nasional Matematika dan Pendidikan matematika UMS*: 637-642.
[17] Yusuf, A. 2013. *Prediksi Nilai dengan Metode Spectral Clustering dan Clusterwise Regression.* Surabaya: Teknik Informatika ITS.

[18] Zuwida, N., dkk. 2014. Tinjauan Pemanfaatan Beasiswa Bantuan Khusus Murid (BKM) Pada Siswa SMK Negeri 1 Pariaman. *Jurnal CIVED* 2(2): 389-394.