ANALISIS PERBANDINGAN CLUSTER K-MEANS DAN FUZZY C-MEANS HASIL UJIAN NASIONAL SMP

THESIS

Oleh:
ADI NURRACHMAN
NIM. 200605210012

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2021

THE COMPARATIVE ANALYSIS OF K-MEANS AND FUZZY C-MEANS CLUSTERS OF JUNIOR HIGH SCHOOLS' FINAL EXAMINATION

THESIS

By:
ADI NURRACHMAN
NIM. 200605210012

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2021

ANALISIS PERBANDINGAN CLUSTER K-MEANS DAN FUZZY C-MEANS HASIL UJIAN NASIONAL SMP

THESIS

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi salah satu persyaratan dalam
memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
ADI NURRACHMAN
NIM. 200605210012

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2021

ANALISIS PERBANDINGAN CLUSTER K-MEANS DAN FUZZY C-MEANS HASIL UJIAN NASIONAL SMP

THESIS

Diajukan kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi salah satu persyaratan dalam
memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
ADI NURRACHMAN
NIM. 200605210012

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2021

ANALISIS PERBANDINGAN CLUSTER K-MEANS DAN FUZZY C-MEANS HASIL UJIAN NASIONAL SMP

THESIS

Oleh:
ADI NURRACHMAN
NIM. 200605210012

Tesis diperiksa dan disetujui untuk diuji:

Tanggal: 29 Desember 2021

Pembimbing I	Pembimbing II
Dr. Muhammad Faisal, MT	Dr.Fachrul Kurniawan, M.MT.,IPM
NIP. 197405102005011007	NIP. 197710202009121001

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian
NIP. 197404242009011008

Dipindai dengan CamScanner

ANALISIS PERBANDINGAN CLUSTER K-MEANS DAN FUZZY C-MEANS HASIL UJIAN NASIONAL SMP

THESIS

Oleh:
ADI NURRACHMAN
NIM. 200605210012

Telah Dipertahankan di Depan Dewan Penguji Thesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal: 29 Desember 2021

Sususan Dewan Penguji — **Tanda Tangan**

Penguji Utama : Dr. M. Amin Hariyadi, M.T ( )
NIP. 196701182005011001

Ketua Penguji : Dr. Ririen Kusumawati, S.Si., M.Kom ( )
NIP 19720309 2005012002

Sekretaris Penguji : Dr. Muhammad Faisal, M.T ( )
NIP. 197405102005011007

Anggota Penguji : Dr.Fachrul Kurniawan, M.MT.,IPM ( )
NIP. 197710202009121001

Mengetahui dan Mengesahkan
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian
NIP. 197404242009011008

Dipindai dengan CamScanner

iv

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Adi Nurrachman
NIM	: 200605210012
Program Studi	: Magister Informatika
Fakultas	: Sains dan Teknologi

Menyatakan dengan sebenarnya bahwa Thesis yang saya tulis ini benar-banar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan Thesis ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 29 Desember 2021

Yang membuat pernyataan,

Adi Nurrachman

NIM. 200605210012

MOTTO

فَاِذَا فَرَغْتَ فَانْصَبْۙ - ٧ وَاِلٰى رَبِّكَ فَارْغَبْ ۧ - ٨

Maka apabila engkau telah selesai (dari sesuatu urusan), tetaplah bekerja keras (untuk urusan yang lain). Dan hanya kepada Tuhanmulah engkau berharap.

QS. Al Insyirah 7-8

HALAMAN PERSEMBAHAN

Tesis ini dipersembahkan untuk:

1. Kedua orang tuaku (Papa dan Mama) tercinta yang telah mencurahkan segala daya dan upayanya demi pendidikan anak-anaknya tersayang.
2. Bapak dan Ibu tersayang yang selalu mensupport penulis selama menjalani pendidikan.
3. Istri ku yang cantik dan sholeha, terimakasih untuk doa dan supportnya
4. Anak-anak ku yang pintar dan sholeh.
5. Teman-teman satu angkatan “mlebu bareng lulus bareng”

KATA PENGANTAR

Assalammu'alaikum warahmatullohi wabarakatuh,

Segala puji bagi Alloh yang hanya kepada-Nya kami memuji, memohon pertolongan, dan mohon keampunan. Kami berlindung kepadaNya dari kekejian diri dan kejahatan amalan kami. Barang siapa yang diberi petunjuk oleh Alloh maka tidak ada yang dapat menyesatkan, dan barang siapa yang tersesat dari jalanNya maka tidak ada yang dapat memberinya petunjuk. Dan saya bersaksi bahwa tiada sembahan yang berhak disembah melainkan Alloh saja, yang tiada sekutu bagiNya. Dan saya bersaksi bahwa Muhammad adalah hambaNya dan RasulNya.

Selanjutnya penulis haturkan ucapan terimakasih seiring *do'a* dan harapan *jazakumullah ahsanal jaza'* kepada semua pihak yang telah membantu terselesaikannya tesis ini. Untuk itu penulis sampaikan terima kasih dan penghargaan yang setinggi-tingginya kepada yang terhormat:

1. Rektor UIN Maulana Malik Ibrahim Malang, Prof. Dr. M. Zainuddin, MA.
2. Dekan Fakultas Sains dan Teknologi, Dr. Sri Harini, M.Si.
3. Kepala Program Studi Magister Informatika, Dr. Cahyo Crysdian.
4. Dosen Pembimbing I, Dr. Muhammad Faisal, MT. atas bimbingan, saran, kritik, dan koreksinya dalam penulisan tesis.
5. Dosen pembimbing II, Dr. Fachrul Kurniawan, M.MT., IPM.atas bimbingan, saran, kritik dan koreksinya dalam penulisan tesis.
6. Semua dosen Magister Informatika yang telah mencurahkan ilmu pengetahuan, wawasan dan inspirasi bagi penulis untuk meningkatkan kualitas akademik

7. Semua staf dan tenaga kependidikan Magister Informatika yang banyak memberikan kemudahan layanan akademik dan administrative selama penulis menyelesaikan studi.
8. Manajemen SMK Telkom Malang yang telah mengizinkan dan memudahkan penulis menempuh pendidikan lanjut serta guru-guru yang mensupport dengan saran dan masukannya.
9. Kedua orang tua, Papa dan Mama yang tidak henti-hentinya memberikan motivasi dan do'a kepada penulis.
10. Bapak dan Ibuk yang selalu memberikan bantuan dan dorongan moral serta perhatian selama penulis menempuh studi.
11. Istri dan anak-anak ku semua yang selalu menjadi inspirasi dalam menjalani hidup.
12. Teman-teman satu angkatan, yuk lulus bareng.

Penulis hanya bisa menyampaikan ucapan terimakasih dan berdo'a semoga amal shalih yang telah mereka semua lakukan, diberikan balasan yang berlipat ganda oleh Alloh *Subhanahu Wata 'Alaa*.

Malang, 28 Desember 2021

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

ABSTRAK

Nurrachman, Adi, 2021, *Analisis Perbandingan Cluster K-Means Dan Fuzzy C-Means Hasil Ujian Nasional SMP*, Program Magister Informatika Universitas Islam Negeri Maulana Malik Ibrahim, Pembimbing: (1) Dr. Muhammad Faisal, MT. (2) Dr. Fachrul Kurniawan, M.MT., IPM.

Kata Kunci : Unsupervised learning, Algoritma K-Means, Algoritma Fuzzy C-Means

Hasil penilaian Ujian Nasional dapat dipergunakan sebagai dasar pemetaan mutu suatu program sekolah atau satuan pendidikan dan sebagai bahan pertimbangan seleksi masuk ke jenjang pendidikan yang lebih tinggi, Dengan mengetahui posisi sekolah yang berprestasi, maka sekolah di jenjang berikutnya dapat membuat strategi dalam mendapatkan calon siswa terbaik. Penelitian ini bertujuan untuk menganalisis perbandingan antara algoritma K-Means dengan Fuzzy C-Means pada hasil Ujian Nasional tingkat SMP dimana hasil analisis dapat digunakan sebagai *guidance strategy* oleh SMK swasta untuk mendapatkan calon siswa terbaik dari sekolah unggulan. Penelitian ini menggunakan pendekatan kuantitatif dengan rancangan sistem *Knowledge Discovery in Databases*. Pengumpulan data dilakukan dengan mengambil data hasil Ujian Nasional tingkat Sekolah Menengah Pertama di Provinsi Jawa Timur tahun 2019. Teknik analisis data yang digunakan meliputi *data collection, data selection, data cleaning, data transformation, data mining,* dan *data evaluation*. Hasil penelitian dengan menggunakan metode algoritma K-Means didapatkan 3 *cluster* yang terdiri dari 285 sekolah pada *cluster* sekolah dengan hasil UN rendah, 198 sekolah dengan hasil UN sedang dan 112 sekolah dengan hasil UN tinggi. Saat menggunakan metode algoritma Fuzzy C-Means didapatkan data bahwa ada sebagian sekolah pada cluster sekolah dengan hasil UN sedang dapat termasuk pada cluster sekolah dengan hasil UN tinggi.

ABSTRACT

Nurrachman, Adi, 2021, *The Comparative Analysis of K-Means and Fuzzy C-Means Clusters of Junior High Schools Final Examination,* Master Program of Computer Science, Maulana Malik Ibrahim State Islamic University of Malang, Advisors: (1) Dr. Muhammad Faisal, MT. (2) Dr. Fachrul Kurniawan, M.MT., IPM.

keywords: Unsupervised learning, K-Means algorithm, Fuzzy C-Means algorithm

The results of the National Examination assessment can be used as a basis for mapping the quality of a school program or educational unit and as a consideration for selection to enter a higher level of education. This study aims to analyze the comparison between the K-Means algorithm and the Fuzzy C-Means on the results of the Junior High School National Examination where the results of the analysis can be used as a guidance strategy by private vocational schools to get the best prospective students from superior schools. This study uses a quantitative approach to the design of the Knowledge Discovery in Databases system. Data collection was carried out by taking data from the National Examination results for Junior High School in East Java Province in 2019. The data analysis techniques used included data collection, data selection, data cleaning, data transformation, data mining, and data evaluation. The results of the study using the K-Means algorithm method obtained 3 clusters consisting of 285 schools in the school cluster with low UN results, 198 schools with moderate UN results and 112 schools with high UN results. When using the Fuzzy C-Means algorithm method, it is found that there are some schools in the school cluster with moderate UN results that can be included in the school cluster with high UN results..

نبذة مختصرة

نورراكمان، عدي ، 2021 ، تحليل مقارن لمجموعات K-Means و Fuzzy C-Means من نتائج الامتحان الوطني للمدرسة الإعدادية، ماجستير في علوم الكمبيوتر ، مولانا مالك إبراهيم جامعة ولاية مالانج الإسلامية ، المشرف: (1) د. محمد فيصل (2) د. فخر القرنيوان.

الكلمات الرئيسية: التعلم غير الخاضع للإشراف ، خوارزمية K-Means ، خوارزمية Fuzzy C-Means

تم استخدام نتائج الامتحان الوطني كأساس لرسم خرائط جودة البرامج و / أو الوحدات التعليمية والنظر في الاختيار للمستوى التالي من التعليم.تهدف هذه الدراسة إلى تحليل المقارنة بين خوارزميات K-Means و Fuzzy C-Means على نتائج الاختبار الوطني للمدرسة الإعدادية حيث يمكن استخدام نتائج التحليل كاستراتيجية إرشادية من قبل المدارس المهنية الخاصة للحصول على أفضل التوقعات. طلاب المدارس العليا.تستخدم هذه الدراسة المنهج الكمي لتصميم نظام اكتشاف المعرفة في قواعد البيانات. تم إجراء جمع البيانات من خلال أخذ البيانات من نتائج الفحص الوطني لـ SMP في مقاطعة جاوة الشرقية في عام 2019. تضمنت تقنيات تحليل البيانات المستخدمة جمع البيانات واختيار البيانات وتنظيفها وتحويل البيانات واستخراج البيانات والتقييم.حصلت نتائج الدراسة باستخدام طريقة خوارزمية K-Means على 3 مجموعات تتكون من 285 مدرسة في مجموعة المدارس المتوسطة ، و 198 مدرسة في مجموعة المدارس المتوسطة ، و 112 مدرسة في مجموعة المدارس العليا. عند استخدام طريقة خوارزمية Fuzzy C-Means ، وجد أنه يمكن تضمين بعض المدارس في مجموعة المدارس المتوسطة في مجموعة المدارس العليا.

BAB I

PENDAHULUAN

A. Latar Belakang Masalah

Perkembangan suatu bangsa dalam hal sumber daya manusia (SDM) dapat diukur terhadap perkembangan ilmu pengetahuan pendidikan nya. Pendidikan adalah suatu proses pembelajaran dengan tujuan untuk mengembangkan keterampilan dan pengetahuan pada diri peserta didik sesuai dengan bakat dan minat baik itu bersifat kepribadian, kecerdasan spiritual keagamaan dan sosial (Juliya & Herlambang, 2021). Pendidikan diartikan pula sebagai upaya penuh kesadaran yang sistematis dalam menggapai kehidupan yang lebih baik. Pendidikan merupakan pelajaran dan pengalaman berharga bagi peserta didik yang menjadikan manusia mampu berpikir lebih kritis dan memiliki karakter yang di inginkan.

Sebelum dihapusnya Ujian Nasional oleh pada tahun 2020 oleh Kementerian Pendidikan dengan dikeluarkannya Surat Edaran Nomor 1 Tahun 2020 mengenai kebijakan merdeka belajar pada satuan pendidikan dalam penentuan kelulusan peserta didik, maka peserta didik pada Sekolah Menengah Pertama yang hendak melanjutkan pendidikan ke jenjang berikutnya yang lebih tinggi, wajib mengikuti Ujian Nasional yang dilaksanakan secara serentak waktunya di seluruh wilayah Indonesia.

Mengacu pada Peraturan Pemerintah No. 13 Tahun 2015 Pasal 68 disebutkan bahwa penilaian hasil Ujian Nasional dapat dipergunakan sebagai dasar pemetaan mutu program sekolah pada satuan pendidikan, sebagai bahan

pertimbangan seleksi masuk jenjang pendidikan yang lebih tinggi berikutnya, pembinaan sekolah serta pemberian bantuan dengan tujuan memperbaiki mutu pendidikan pada satuan pendidikan.

Peningkatan mutu pendidikan erat kaitannya dengan tuntunan agama Islam agar tidak meninggalkan keturunan yang lemah (aqidah, ibadah, ilmu dan ekonomi) seperti yang disebut didalam Al-Quran surah An-Nisa ayat 9:

وَلْيَخْشَ الَّذِينَ لَوْ تَرَكُوا مِنْ خَلْفِهِمْ ذُرِّيَّةً ضِعَافًا خَافُوا عَلَيْهِمْ فَلْيَتَّقُوا اللَّـهَ وَلْيَقُولُوا قَوْلًا سَدِيدًا

Artinya: *"Dan hendaklah takut kepada Alloh orang-orang yang seandainya meninggalkan dibelakang mereka anak-anak yang lemah, yang mereka khawatir terhadap (kesejahteraan) mereka. Oleh sebab itu hendaklah mereka bertakwa kepada Alloh dan hendaklah mereka mengucapkan perkataan yang benar".*

Hasil Ujian Nasional dapat digunakan dalam pemetaan mutu program sekolah pada satuan pendidikan dan salah satu pertimbangan untuk seleksi masuk ke jenjang pendidikan lebih tinggi berikutnya. Berangkat dari hal tersebut, penulis memandang pentingnya dilakukan klusterisasi atau pengelompokkan sekolah berdasarkan hasil Ujian Nasional.

Hal ini dipandang penting karena dapat memberikan pengetahuan bagi pemangku kebijakan terkait dengan strategi yang semestinya diterapkan ke masing-masing sekolah sesuai dengan pengelompokkannya agar mutu pendidikan semakin baik.

Berdasarkan data set yang diambil dari hasil UN Provinsi Jawa Timur, peneliti akan mengelompokkan sekolah-sekolah menggunakan metode *clustering*. *Clustering* merupakan salah satu metode dalam *data mining* yang mempunyai sifat

unsupervised learning yang bertujuan untuk mempartisi data kedalam sebuah atau banyak *cluster* atau kelompok.

Clustering menggunakan metode K-Means merupakan metode yang umum digunakan, dikarenakan memiliki kemampuan untuk mengelompokkan data dalam jumlah yang besar dengan waktu perhitungan menggunakan komputer yang cepat dan efisien (Arai & Ridho Barakbah, 2007).

Clustering Fuzzy C-Means merupakan sebuah metode pengelompokkan data yang mengizinkan satu objek data menjadi anggota pada dua cluster atau lebih. Sehingga bila dibandingkan dengan metode *clustering* K-Means kegagalan konvergen menggunakan metode *clustering* Fuzzy C-Means lebih kecil (Yudi Agusta, 2007).

Dengan pertimbangan tersebut, maka penulis memilih metode *clustering* K-Means dan Fuzzy C-Means sebagai metode *clustering* yang akan dibandingkan secara teoritis dan aplikasi untuk mengelompokkan sekolah-sekolah berdasarkan hasil Ujian Nasional SMP Provinsi Jawa Timur.

Untuk mengolah data hasil UN tersebut, peneliti menggunakan Bahasa pemrograman Python. Python merupakan Bahasa pemrogram *opensource* yang mudah digunakan untuk memproses *machine learning algorithm*, dengan menggunakan library sckit learn (Hackeling, 2014).

B. Pernyataan Masalah

1. Apakah metode K-Means dan Fuzzy C-Means dapat dipergunakan buat mengelompokkan SMP di Provinsi Jawa Timur berdasarkan hasil Ujian Nasional?
2. Bagaimana hasil perbandingan antara metode K-Means dengan metode Fuzzy C-Means guna mengelompokkan SMP di Provinsi Jawa Timur berdasarkan hasil Ujian Nasional?
3. Bagaimana seharusnya strategi yang dibuat oleh sekolah terhadap hasil penelitian ini.

C. Tujuan Penelitian

Tujuan penelitian ini adalah untuk mengetahui apakah algoritma K-Means dan Fuzzy C-Means dapat digunakan untuk mengelompokkan SMP berdasarkan hasil Ujian Nasional dan menganalisis perbandingan antara algoritma K-Means dengan Fuzzy C-Means pada hasil Ujian Nasional tingkat SMP untuk kemudian hasil analisis dapat digunakan oleh sekolah untuk memetakan mutu pendidikan dan strategi *marketing* PPDB.

D. Manfaat Penelitian

1. Membandingkan hasil pembuktian pengelompokkan antara alogoritma K-Means dengan Fuzzy C-Means.

2. Hasil penelitian dapat dipergunakan oleh SMK sebagai *guidance strategy* untuk melakukan promosi PPDB dengan harapan mendapatkan calon peserta didik terbaik guna peningkatan mutu pendidikan di sekolah tersebut.

E. Ruang Lingkup Penelitian

Menggunakan data sekunder yang didapatkan dari laporan hasil Ujian Nasional tingkat SMP di Provinsi Jawa Timur pada tahun 2019 penelitian ini menggunakan pendekatan kuantitatif.

Metode untuk *data clustering* yang digunakan pada penelitian ini adalah metode algoritma K-Means dan Fuzzy C-Means.

F. Penelitian Terdahulu

Dalam penulisan tesis, penulis terlebih dahulu menggali informasi dari jurnal penelitian sejenis yang ada sebelumnya sebagai bahan perbandingan, untuk mencari tahu keperbaruan dan mengetahui kekurangan dan kelebihan yang sudah ada.

Penelitian yang dilakukan oleh (Aditya et al., 2020) bertujuan objek data yang memiliki persamaan antara satu data dengan data lainnya akan di kelompokkan. Algoritma K-Means akan mengelompokkan suatu data kedalam satu kelompok atau lebih dan termasuk dalam metode *clustering* non hirarki. *Clustering* data didalam penelitian ini dilakukan dengan memakai metode algoritma K-Means dengan dataset yang dipakai diambil dari hasil Ujian Nasional Sekolah Menengah Pertama pada tahun 2018/2019 yang diperoleh dari website resmi Kementrian

Pendidikan dan Kebudayaan Republik Indonesia. Didapatkan hasil perhitungan tiga kelompok, dengan kelompok pertama merupakan *cluster* dengan nilai ujian nasional baik, kelompok kedua merupakan *cluster* dengan nilai ujian nasional kurang dan kelompok ketiga merupakan *cluster* dengan nilai ujian nasional sedang. Pada kelompok pertama terdapat 14 provinsi, kelompok kedua terdapat 5 provinsi, dan kelompok ketiga terdapat 15 provinsi.. Hasil evaluasi menggunakan metode K-Means dengan jumlah *cluster* 3 menghasilkan nilai evaluasi Dunn 0.246, Silhouette 0.464 dan Connectivity 11.916.

Penelitian yang dilakukan oleh (Suputra, 2021) dengan algoritma K-Means bertujuan untuk pengelompokkan kualitas pendidikan pada tingkat SMA/MA berdasarkan hasil ujian nasional SMA/MA provinsi Jawa Timur pada tahun ajaran 2018/2019. Metode K-Means melakukan pengelompokan objek ke dalam satu atau beberapa kelompok yang termasuk dalam metode *clustering* non hirarki. Algoritma pada K-Means ini me minimumkan perbedaan objek data di dalam *cluster* yang sama dan me maksimumkan perbedaan objek data pada *cluster* yang berbeda. Data yang berisi rerata variabel hasil ujian nasional didapatkan melalui website resmi Pusat Penilaian Pendidikan dan Kebudayaan Kementrian Pendidikan dan Kebudayaan Republik Indonesia yang selanjutnya akan di lakukan *data mining*. Prosedur penambangan data yang dilakukan yakni pembersihan data awal, integrasi data untuk validasi data, pemilihan data yang digunakan, transformasi data kedalam bentuk yang diinginkan, penambangan data, mengevaluasi pola yang ditemukan untuk kemudian mempresentasikan informasi. Hasil pengelompokkan yang diperoleh dalam penelitian ini ialah kategori kelompok

1 terdapat 2 provinsi, kategori kelompok 2 terdapat 10 provinsi 2, kategori kelompok 3 terdapat 11 provinsi dan kategori kelompok 4 terdapat 12 provinsi. Hasil evaluasi perhitungan menggunakan algoritma K-Means diperoleh nilai evaluasi Partition Coefficiens Index (PCI) sebesar 0,81.

Penelitian yang dilakukan oleh (Selviana & Mustakim, 2016) membandingkan algoritma menggunakan metode K-Means dengan Fuzzy C-Means. Pengukuran motivasi belajar dikategorikan menjadi empat yaitu, Satifaction, Confidence, Relevance, Attention. Penggunaan metode belajar seperti E-learning, pelaksanaan praktek teori diluar kelas dan praktikum di laboratorium digunakan sebagai strategi pembelajaran. Penelitian menunjukkan strategi pembelajaran praktikum di laboraturium lebih unggul dalam penerapan dibandingkan pembelajaran e-learning dan pelaksanaan praktek diluar kelas yang dibuktikan dengan hasil perhitungan e-learning 29,6%, praktek teori lapangan 34,9% dan praktikum di laboratorium 35,4% menggunakan algoritma K-Means dan algoritma Fuzzy C-Means. Berdasarkan hasil validasinya algoritma Fuzzy C-Means lebih baik dari algoritma K-Means dengan nilai 0,5098 berbanding 0,2896.

Penelitian yang dilakukan oleh (As'Sidiq & Mandala, 2020) bermaksud untuk membuat sebuah aplikasi untuk membantu universitas membagi mahasiswa baru Teknologi Informasi ke dalam kelas berdasarkan kompetensi dan pengalaman tentang Teknologi Informasi (TI) pada SMA. Percobaan dilakukan dengan mengumpulkan data siswa IT yang bukan semester kesatu. Data terdiri dari pengalaman tentang IT bidang pengetahuan dan nilai IPK. Hasil percobaan menunjukkan bahwa dari 50 sampel data yang dikumpulkan, aplikasi dengan benar

memprediksi 34 siswa berdasarkan rentang IPK pada kompetensi responden dengan IT dan bidang pengetahuan lain selama mereka belajar di SMA.

Penelitian yang dilakukan oleh (Utomo et al., 2020) bertujuan untuk menganalisis sentimen publik pengguna media sosial tentang Ujian Nasional di Indonesia. Data diambil dengan merayapi data melalui Twitter API. Data tersebut perlu diproses terlebih dahulu dan fitur diekstraksi menggunakan TF-IDF. Namun karena data teks di Twitter tidak terstruktur dan datanya sangat beragam (variety), maka tahap pengelompokan harus dilakukan terlebih dahulu. Teknik pengelompokan menggunakan K-Means *Clustering* pada Spark. Teknik Spark *Clustering* digunakan untuk mengatasi pengelompokan data pada jumlah data yang sangat besar dan kompleks. Dari proses *clustering* menggunakan Spark didapatkan bahwa proses pengelompokan menghasilkan 3 *cluster* dimana deteksi elbow ditemukan pada *cluster* ketiga jumlah *cluster* antara 2 sampai dengan 50. Hasil *clustering* yang berupa 3 kelompok besar diproses lebih lanjut (dengan teknik klasifikasi) untuk mendapatkan perbandingan sentimen positif atau negatif dari komentar pengguna media sosial tentang ujian nasional. Selanjutnya hasil tersebut menjadi rekomendasi dan pengetahuan baru tentang perilaku masyarakat terkait Ujian Nasional Berbasis Media Sosial.

Tabel 1. 1 Penelitian Terdahulu dan Orisinalitas Penelitian

No	Nama Peneliti, Tahun dan Sumber	Persamaan	Perbedaan
1	Agil Aditya, Ivan Jovian, Betha Nurina Sari, 2020, jurnal Media Informatika Budidarma	Algoritma: - K-Means Dataset: - Hasil UN SMP Rancangan sistem: - KDD	Algoritma: - Fuzzy C-Means Periode dataset: - Hasil UN SMP tahun 2018 Tools: R
2	IWA. Suputra, IM. Candiasa, IPP. Suryawan, 2021, Jurnal Matematika, Sains, dan Pembelajarannya	Algoritma: - K-Means Rancangan sistem: - KDD	Algoritma: - Fuzzy C-Means Dataset: - Hasil UN SMA/ MA
3	Nur Indah Selviana, Mustakim, 2016, Seminar Nasional Teknologi Informasi, Komunikasi dan Industri (SNTIKI)	Algoritma: - K-Means - Fuzzy C-Means	Dataset: - Data kuisioner Tools: - Matlab
4	Arfan As'Sidiq, Rila Mandala, 2016, Jurnal IT FOR SOCIETY	Algoritma: - K-Means Tools: - Python	Algoritma: - Fuzzy C-Means Dataset: - Data kuisioner
5	Utomo, Chandra Eko Wahyudi, Hariadi, Mochamad Sumpeno, Surya, 2020, Jurnal JAREE (Journal on Advanced Research in Electrical Engineering)	Algoritma: - K-Means	Algoritma: - Fuzzy C-Means Tools: Spark Dataset: - Analisis sentimen di twitter

Berdasarkan tabel 1.1 terdapat persamaan dan perbedaan antara penelitian yang dilakukan peneliti saat ini dengan penelitian terdahulu. Adapun persamaannya terkait dengan rancangan sistem yang digunakan dan perbedaanya terkait algoritma, dataset serta *tools* yang digunakan.

G. Definisi Operasional

Sebuah atribut atau nilai atau sifat dari suatu objek atau kegiatan yang memiliki variasi nilai tertentu harus di ditetapkan oleh peneliti untuk dipelajari dan ditarik kesimpulannya dalam suatu keterangan operasional variabel penelitian. Variabel dalam penelitian harus di definisikan atau dirumuskan untuk menghindari kesalahan dalam pengumpulan data. Definisi atau keterangan operasional variabel dalam penelitian ini ditetapkan sebagai berikut:

1. Ujian Nasional

Ujian nasional merupakan assesmen evaluasi pencapaian hasil pembelajaran pada satuan pendisikan di tingkat nasional yang ditetapkan oleh pemerintah untuk mengetahui mutu sebaran pendidikan.

2. Satuan Pendidikan

Satuan Pendidikan adalah kelompok, organisasi atau badan hukum yang menyelenggarakan layanan pendidikan baik menggunakan jalur formal maupun informal disetiap jenjang pendidikan.

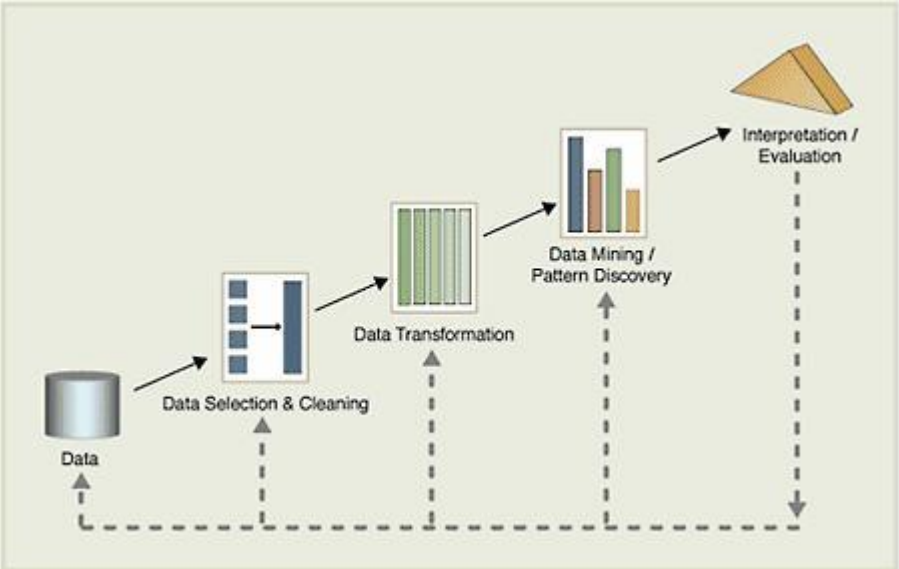
BAB II

KAJIAN PUSTAKA

A. Data Mining

Data mining merupakan metode untuk mencari, menemukan dan mendapatkan pola tertentu yang berbeda dari kumpulan *big data*, dimana data tersebut tersimpan dalam suatu *database*, *data warehouse*, atau *repositori* informasi lainnya.(Agarwal, 2014)

Untuk mendapatkan pola tertentu dari kumpulan *big data* diperlukan langkah-langkah atau proses yang disebut *Knowledge Discovery in Databases* (KDD) dimana *data mining* merupakan bagian tidak terpisahkan dari proses tersebut.



Gambar 2. 1 Tahapan proses KDD dalam menghasilkan knowledge

Berikut penjelasan langkah-langkah pada KDD:

a. Data Cleaning

   Digunakan untuk menghapus data yang tidak terpakai dan inkonsisten.

b. Data Integration

   Digunakan untuk menggabungkan data dari berbagai sumber data.

c. Data Selection

   Data yang digunakan akan diseleksi dan dikembalikan dari *database* sesuai tugas analisis.

d. Data Transformation

   Data akan diubah kedalam bentuk siap untuk diolah dengan melakukan analisis menggunakan operasi *summary* atau *aggregation*.

e. Data Mining

   Proses menggunakan metode intelijen yang sudah diterapkan untuk mengolah *data pattern.*

f. Pattern Evaluation

   Untuk menemukan dan identifikasi pola yang dapat menjelaskan pengetahuan dasar yang telah di olah.

g. Knowledge Presentation

   Teknik presentasi berupa visualisasi data yang digunakan untuk menyajikan hasil *knowledge* yang telah diolah kepada pengguna.

1. Data, Informasi dan Pengetahuan (Knowledge)

Data adalah hasil observasi lansung suatu peristiwa yang masih mentah. Informasi merupakan kumpulan data terstruktur dan hubungan antar data yang sudah diolah. Sedangkan pengetahuan merupakan sebuah model informasi yang dimodifikasi (Teskey, 1989).

2. Pengelompokkan Data Mining

Berdasarkan tugas yang dapat dilakukan, *data mining* dibagi menjadi beberapa kelompok (Larose & Larose, 2014), yaitu:

a. Deskripsi (Description)

Digunakan untuk menjelaskan dan menggambarkan pola kecenderungan yang terdapat dalam suatu data.

b. Estimasi (Estimation)

Pada estimasi nilai variabel target lebih kearah numerik dibandingkan variabel target klasifikasi yang kearah kategori. Lengkapnya nilai pada record dari variabel target sebagai nilai prediksi dapat digunakan untuk membangun model. Selanjutnya berdasarkan nilai variabel prediksi dapat dibuat estimasi nilai dari variabel target.

c. Prediksi (Prediction)

Teknik dan metode pada estimasi dan klasifikasi dapat digunakan pula pada prediksi. Nilai hasil prediksi yang telah di olah akan ada di masa depan.

d. Klasifikasi (Classification)

Dalam klasifikasi, pengguna dapat memilik target variabel kategori. Contohnya pengelompokkan jumlah penghasilan dapat dipisahkan menjadi tiga kategori, yaitu penghasilan tinggi, penghasilan menengah dan penghasilan rendah.

e. Pengklusteran (*Clustering*)

Pengklusteran merupakan proses mengelompokkan data, mengamati dan memperhatikan serta membentuk kelompok pada objek yang memiliki kemiripan. *Cluster* merupakan sekumpulan data yang memiliki kecenderungan data yang mirip antara satu data dengan data yang lainnya dan memiliki ketidakmiripan satu data dengan data yang lain. Berbeda dengan teknik klasifkasi yang memilik variable target, pada teknik *clustering* tidak memiliki variabel target. Algoritma *clustering* melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang mana kemiripan data dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan data dalam kelompok lain akan bernilai minimal, alih-alih melakukan klasifikasi, estimasi, atau prediksi nilai dari variabel target.

f. Asosiasi (Association)

Teknik asosiasi digunakan untuk menemukan atribut yang muncul dalam satu waktu. Penggunaan teknik asosiasi pada *data mining* disebut analisis keranjang belanja (Shopping Cart Analysis).

B. Clustering

Menurut (Gorunescu, 2011) ada dua tujuan penggunaan *clustering* yaitu *clustering* untuk pemahaman dan *clustering* dengan tujuan penerapan. *Cluster* yang terbentuk harus dapat mengungkapkan struktur data aslinya agar tujuan *clustering* untuk pemahaman tercapai. Penggunaan *clustering* dengan maksud pemahaman hanya sebagai awal proses *clustering* untuk selanjutnya dilakukan tugas yang lain

seperti summarization (menghitung rata-rata dan standar deviasi), pelabelan nama kelas dalam setiap kelompok yang akan di *cluster* atau dimanfaatkan sebagai data latih (training) pada klasifikasi. Jika penggunaan adalah tujuan *clustering* nya maka dilakukan pengolahan data untuk mendapatkan prototype *cluster* yang paling representative terhadap data dan memberikan penjelasan abstrak terhadap objek data dalam *cluster* dimana sebuah data terletak didalamnya.

Para ahli mengembangkan berbagai metode *clustering*, dimana setiap metode mempunyai kelebihan dan kekurangan masing-masing serta karakter yang berbeda. Persamaan data dalam *cluster*, keanggotaan tiap-tiap *cluster* dan perbedaan struktur *cluster* digunakan untuk membedakan *clustering*.

Menurut strukturnya, metode *clustering* dibagi menjadi dua kategori yaitu *hierarchical clustering* dan *partition clustering*. Pada *hierarchical clustering* memiliki aturan bahwa sebuah kelompok bisa terdiri hanya dari satu data tunggal, lalu dua atau lebih kelompok kecil dapat bergabung menjadi satu kelompok besar dan begitu seterusnya hingga semua data dapat bergabung kedalam satu kelompok besar. Sedangkan pada *partition clustering* memisahkan set data kedalam sejumlah kelompok yang tidak tumpang tindah (overlap) antara kelompok yang satu dengan kelompok yang lain, yang berarti setiap objek data hanya dapat menjadi anggota pada satu kelompok. Metode yang masuk pada kategori tersebut adalah K-Means dan Fuzzy C-Means.

Menurut keanggotaan, metode *clustering* dibagi menjadi dua, yaitu kategori eksklusif dan kategori tumpang tindih. Suatu metode disebut kategori ekslusif jika sebuah data hanya menjadi anggota pada satu kelompok dan tidak

menjadi anggota pada kelompok yang lain. Metode *clustering* K-Means dan DBSCAN termasuk dalam kategori ekslusif. Sedangkan apabila sebuah data dapat menjadi anggota dibeberapa kelompok maka termasuk kedalam kategori tumpang tindih, contohnya clustering menggunakan metode Fuzzy C-Means.

C. Clustering K-Means

Metode K-Means merupakan algoritma *clustering* yang sering digunakan dalam bidang *data mining* (Rutledge, 2009). Algoritma pada K-Means akan membagi himpunan *d* menjadi *k cluster* data. K-Means meng *cluster* semua titik data pada *d* sehingga titik data x_i menjadi satu-satunya *k* partisi. Artinya satu titik data hanya masuk kedalam satu *cluster*.

Berikut langkah-langkah perhitungan menggunakan algoritma K-Means (Tan, P.N., Steinbech, M., & Kumar, 2006) :

1. Peneliti menentukan *k* sebagai *centroid*, dimana *k* merupakan jumlah *cluster* yang diinginkan.

2. Kemudian setiap titik data akan dicari *centroid* terdekatnya.

3. *Cluster* akan terbentuk dari setiap himpunan titik data yang menjadi *centroid*.

4. Lakukan perhitungan kembali *centroid* dari setiap *cluster*.

5. Ulangi langkah ke 1- 4 hingga nilai *centroid* tidak berubah.

Jarak Euclidean dipakai guna mengukur kedekatan data pada metode *clustering* menggunakan algoritma K-Means. Tujuan penggunaan metode

algoritma K-Means adalah untuk meminimumkan jarak total Euclidean diantara setiap titik x_i, dan *cluster* terdekat yakni c_j.(Rutledge, 2009). Untuk menentukan Jarak Euclidean menggunakan persamaan berikut:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \{x_{ik} - x_{jk}\}^2} \quad (2.1)$$

Keterangan:

d_{ij} = jarak antara data ke-i dan data ke-j

n = dimensi data

x_{ik} = koordinat data ke-i pada dimensi k

x_{jk} = koordinat data ke-j pada dimensi k

Algoritma K-Means membagi data ke dalam *k*-buah *cluster* yang telah ditentukan (Han et al., 2012). Manhattan distance, Chebisev distance dan Euclidean distance merupakan beberapa cara yang dapat digunakan untuk menghitung jarak . Euclidean distance merupakan fungsi matriks jarak yang paling banyak digunakan dengan tingkat identifikasi kemiripan (*similarity*) lebih tinggi dibanding metode yang lain (Nishom, 2019).

Berdasarkan hal tersebut, maka formula Euclidean Distance yang akan digunakan untuk menentukan jarak pada penelitian ini.

D. Clustering Fuzzy C-Means

Ditemukan pertama kali oleh Dunn (1973) dan dikembangkan oleh Bezdek (1981) algoritma Fuzzy C-Means (FCM) sering digunakan untuk tujuan *Pattern Recognation*. Langkah pertama pengaplikasian algoritma Fuzzy C-Means adalah membuat perhitungan kelas yang akan dibuat basis klasifikasi terlebih dahulu. Kemudian dilakukan perhitungan atau iterasi sampai tiap kelompok mendapatkan anggota masing-masing. Langkah-langkah yang dilakukan dengan algoritma *fuzzy* akan menghasilkan perhitungan yang halus.

Hasil perhitungan yang halus adalah data objek pengamatan tidak mutlak hanya menjadi anggota di satu kelompok, namun memungkinkan dapat menjadi anggota pada kelompok lain dengan ukuran derajat keanggotaan yang berbeda. Suatu objek akan cenderung menjadi anggota pada kelompok tertentu dimana derajat keanggotaan objek tersebut dalam kelompok itu paling besar dibandingkan dengan derajat keanggotaan pada kelompok lain.

Anggaplah terdapat sejumlah data pada dataset X yang terdiri dari *n data* yang di notasikan dalam X = {x1, x2, …., xn}, dimana setiap data mempunya fitur *r* dimensi; xi1, xi2, …, xir, dinotasikan xi = {xi1, xi2, …, xir}. Ada sejumlah *cluster c* dengan *centroid* c1, c2, …, ck, dimana *k* adalah jumlah *cluster*. Masing-masing data mempunyai derajat keanggotaan pada setiap *cluster*, dinyatakan dengan *uij* dengan nilai antara 0 dan 1, *i* menyatakan data *xi*, dan *j* menyatakan *cluster cj*. Jumlah nilai derajat keanggotaan setiap data *xi* selalu sama dengan 1, yang di formulasikan pada persamaan berikut:

$$\sum_{j=1}^{k} uij = 1 \qquad (2.2)$$

Untuk *cluster cj,* setiap *cluster* berisi paling sedikit satu data dengan nilai keanggotaan tidak nol, namun tidak berisi derajat satu pada semua data. Cluster *cj* dapat diformulasikan sebagai berikut:

$$0 < \sum_{i=1}^{n} uij < n \tag{2.3}$$

Karakteristik pada perhitungan algoritma *fuzzy*, sebuah data dapat menjadi anggota dibeberapa kelompok yang dinyatakan melalui tingkat derajat keanggotaan pada setiap kelompok, maka di dalam Fuzzy C-Means masing-masing data juga menjadi anggota pada setiap *cluster* dengan derajat keanggotaan *uij*. Nilai derajat keanggotaan data *xi* pada *cluster cj*, dapat diformulasikan sebagai berikut:

$$u_{ij} = \frac{D(xi, cj)\frac{-2}{w-1}}{\sum_{i=1}^{k} D(xi, cj)\frac{-2}{w-1}} \tag{2.4}$$

Centroid cluster ke *j* menjadi parameter *cj*. *D()* merupakan jarak antara data dengan *centroid*, sedangkan *w* merupakan parameter bobot pangkat (weighting exponent) yang diperkenalkan dalam Fuzzy C-Means. Biasanya nilai *w* > 1 dan umumnya diberi nilai 2 dikarenakan *w* tidak memiliki nilai ketetapan.

Nilai keanggotaan tersebut disimpan dalam matrik *fuzzy pseudo-partition* berukuran Nxk, dengan kolom adalah nilai keanggotaan pada setiap cluster dan baris merupakan data.

$$\begin{pmatrix} u_{11}(x1) & u_{12}(x1) \dots \dots \dots & u_{1k}(x1) \\ u_{21}(x2) & u_{22}(x2) \dots \dots \dots & u_{2k}(x2) \\ u_{n1}(xn) & u_{n2}(xn) \dots \dots \dots & u_{nk}(xn) \end{pmatrix} \tag{2.5}$$

Persamaan berikut digunakan untuk menghitung *centroid* pada cluster *ci* pada fitur *j*:

$$C_{ij} = \frac{\sum_{i=1}^{n}(u_{il}) \;\; w_{xij}}{\sum_{i=1}^{n}(u_{il}) \;\; w} \qquad (2.6)$$

Parameter *n* adalah banyak data, sedangkan *w* adalah bobot pangkat dan *uil* adalah data *xi* ke cluster *ci*

Sedangkan fungsi objective menggunakan persamaan berikut:

$$J = \sum_{i=1}^{n} \sum_{l=1}^{k} (u_{il}) \;\; w_{D(x1,c1)} \qquad (2.7)$$

Proses perhitungan pada Fuzzy C-Means mempunyai kesamaan prinsip algoritma dengan perhitungan pada algoritma K-Means (Yudi Agusta, 2007), sehingga perhitungan jarak matriks anggota menggunakan formula Euclidean distance seperti halnya pada K-Means.

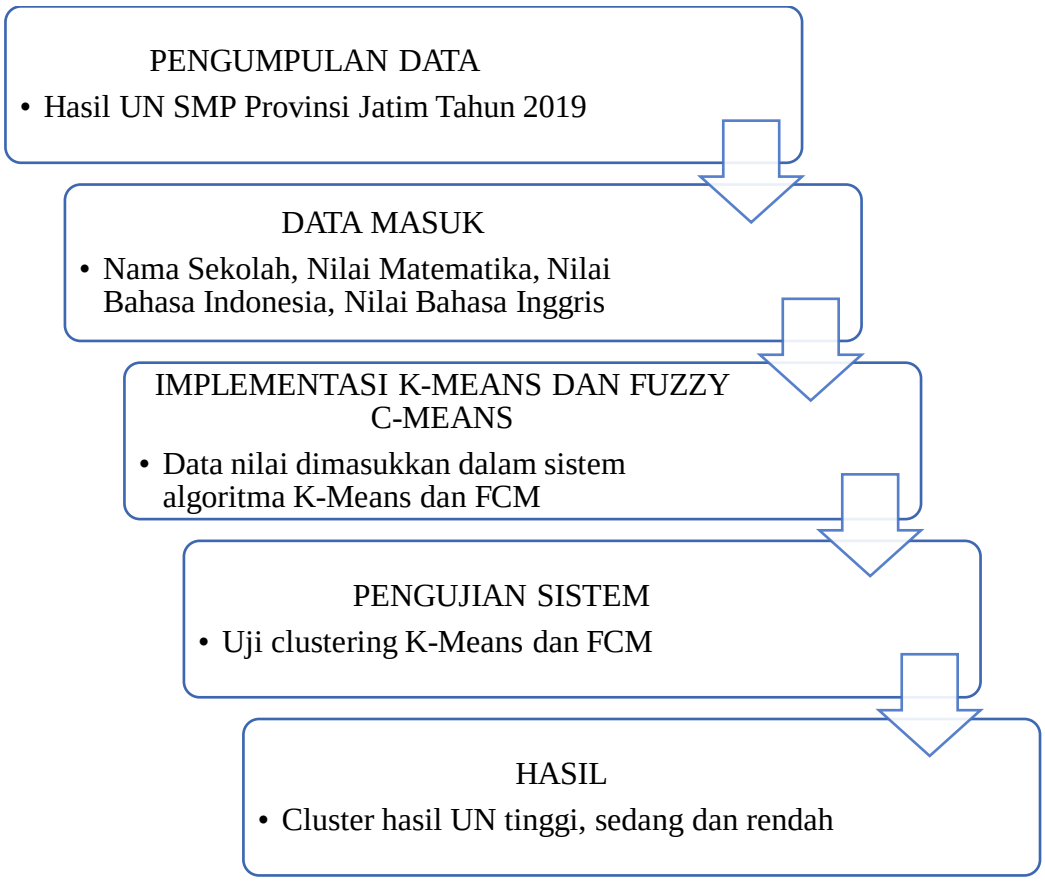
Berdasarkan pertimbangan tesebut, maka penulis memilih metode *clustering* K-Means dan Fuzzy C-Means sebagai metode yang akan dibandingkan secara teoritis dan aplikasi untuk mengelompokkan sekolah-sekolah berdasarkan hasil Ujian Nasional SMP Provinsi Jawa Timur.

BAB III

METODE PENELITIAN

A. Jenis Penelitian

Pendekatan kuantitaif dipergunakan dalam penelitian ini dalam bentuk pengolahan dataset. Data yang digunakan di dalam penelitan adalah hasil Ujian Nasional SMP yang didapatkan dari website resmi Kementrian Pendidikan dan Kebudayaan, yaitu https://hasilun.puspendik.kemdikbud.go.id/ merupakan data sekunder.



Gambar 3. 1 Diagram Blok Metode Penelitian

B. Variabel Penelitian

1. Tempat dan waktu penelitian

   Penelitian dilakukan di SMK Telkom Malang selama 6 bulan.

2. Bahan dan alat

   Bahan yang akan digunakan pada penelitian ini adalah:

   a. Laptop

   b. Akses untuk internet.

   c. Aplikasi pendukung berupa:

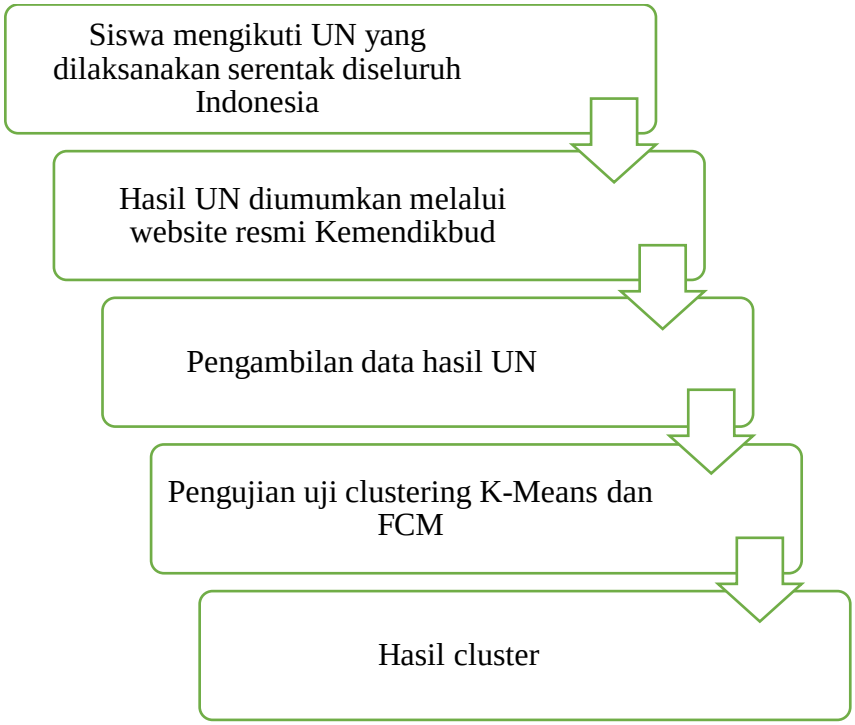
      - Jupyter Notebook: digunakan untuk menuliskan code program di python

      - Ms. Excel: digunakan untuk memyimpan data

C. Pengumpulan Data

Jenis instrumen yang relevan untuk digunakan dalam penelitian ini adalah pengambilan data set secara langsung melalui website Kementrian Pendidikan dan Kebudayaan, karena sesusai dengan pendekatan yang digunakan dan tujuan penelitian dalam penelitian ini.

Berikut proses pengumpulan data dimulai dari siswa mengikuti Ujian Nasional seperti pada gambar 3.2 Diagram alur pengumpulan data

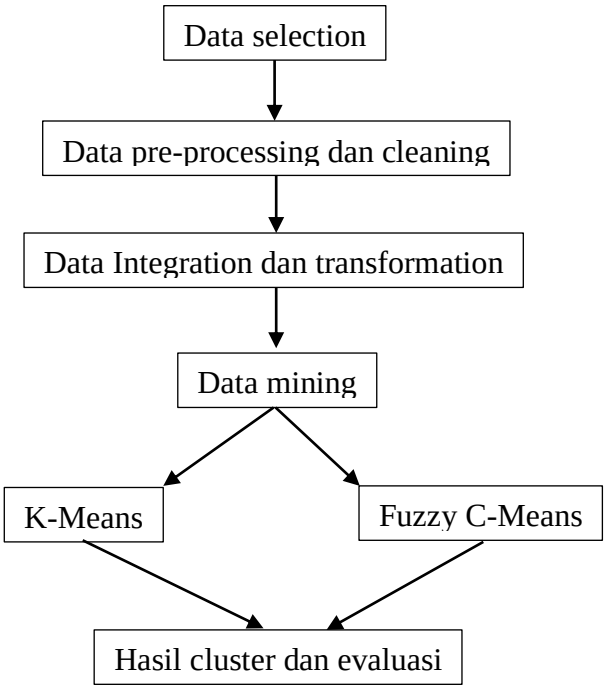


Gambar 3. 2 Diagram Alur Pengumpulan Data

1. Siswa diseluruh Indonesia mengikuti Ujian Nasional yang dilaksanakan serentak oleh Kemendikbud.
2. Hasil UN siswa diumumkan secara online melalui website Kemendikbud.
3. Peneliti mengambil data secara langsung dengan melakukan filter terlebih dahulu sebelum download data yang dibutuhkan.
4. Data kemudian di proses dan dianalisis menggunakan metode *clustering* K-Means dan Fuzzy C-Means.
5. Hasil analisis dapat digunakan untuk mengelompokkan sekolah-sekolah dengan hasil UN rendah, sedang dan tinggi.

D. Analisis Data

Tingkatan proses dari metode KDD (Knowledges Discovery in Databases) digunakan pada penelitian ini untuk melakukan *clustering* menggunakan algoritma K-Means dan Fuzzy C-Means. Mengenai tingkatan proses dari metode KDD yang digunakan adalah *data selection,* selanjutnya dilakukan *data pre-processing*, kemudian dilakukan *data integration,* kemudian data akan di ubah pada *data transformation,* selanjutnya dilakukan perhitungan pada *data mining*, dan *evaluation* (Agarwal, 2014). Berikut alur analisis data *clustering* K-Means dan Fuzzy C-Means hasil Ujian Nasional tingkat Sekolah Menengah Pertama.



Gambar 3. 3 Alur Analisis K-Means dan Fuzzy C-Means

1. Data selection

Dataset data pertama yang dipergunakan dalam penelitian ini merupakan dataset hasil Ujian Nasional tingkat SMP Provinsi Jawa Timur Tahun Ajaran 2018/2019 yang diambil dari Puspendik Kemendikbud. Dataset ini terdiri dari 10 atribut dan memiliki 38 *record*. Keterangan data dapat dilihat pada tabel 3.1

Tabel 3. 1 Keterangan awal dataset data pertama

Atribut	Tipe data	Keterangan
No	Integer	No Urut
Kode	Integer	Kode Kota/Kabupaten
Nama Kota/Kabupaten	Char	Nama Kota/Kabupaten
Jumlah Satuan Pendidikan	Integer	Banyaknya satuan Pendidikan di tiap kota/Kabupaten
Jumlah Peserta	Integer	Banyaknya peserta ujian di tiap Kota/Kabupaten
Rata-rata Nilai Bahasa Indonesia	Integer	Hasil nilai Bahasa Indonesia di tiap Kota/Kabupaten
Rata-rata Nilai Bahasa Inggris	Integer	Hasil nilai Bahasa Inggris di tiap Kota/Kabupaten
Rata-rata Nilai Matematika	Integer	Hasil nilai Matematika di tiap Kota/Kabupaten
Rata-rata Nilai IPA	Integer	Hasil nilai IPA di tiap Kota/Kabupaten
Rata-rata Nilai	Integer	Hasil rerata nilai keseluruhan

Dari dataset data pertama selanjutnya dilakukan proses seleksi untuk mencari Nama Kota/Kabupaten dengan rerata nilai diatas 61. Didapatkan hasil seperti tabel 3.2

Tabel 3. 2 Dataset hasil seleksi data pertama

No	Nama Kota/Kabupaten	Rerata Nilai
1	KOTA MALANG	65,41
2	KOTA MADIUN	63,84
3	KOTA SURABAYA	62,46
4	KOTA BLITAR	62,09
5	KOTA KEDIRI	61,64

Lanjutan tabel 3.2

No	Nama Kota/Kabupaten	Rerata Nilai
6	KOTA MOJOKERTO	61,36
7	KABUPATEN TULUNGAGUNG	61,03

Data kedua berikutnya yang dipergunakan merupakan dataset hasil Ujian Nasional tingkat SMP Provinsi Jawa Timur tahun 2018/2019 di tujuh kota yang memiliki nilai Rerata tertinggi. Data ini terdiri dari 11 atribut yang memiliki 600 *record*. Keterangan data dapat dilihat pada tabel 3.3

Tabel 3. 3 Keterangan awal dataset data kedua

Atribut	Tipe data	Keterangan
No	Integer	No Urut
Kode	Integer	Kode Kota/Kabupaten
Nama Satuan Pendidikan	Char	Nama Kota/Kabupaten
NPSN	Integer	Nomor Pokok Sekolah Nasional
Status	Char	Status Satuan Pendidikan
Jumlah Peserta	Integer	Banyaknya peserta ujian di tiap Kota/Kabupaten
Rata-rata Nilai Bahasa Indonesia	Integer	Hasil nilai Bahasa Indonesia di tiap Kota/Kabupaten
Rata-rata Nilai Bahasa Inggris	Integer	Hasil nilai Bahasa Inggris di tiap Kota/Kabupaten
Rata-rata Nilai Matematika	Integer	Hasil nilai Matematika di tiap Kota/Kabupaten
Rata-rata Nilai IPA	Integer	Hasil nilai IPA di tiap Kota/Kabupaten
Rata-rata Nilai	Integer	Hasil rata-rata nilai keseluruhan

Dari dataset kedua selanjutnya dilakukan proses penyeleksian data, dimana atribut-atribut yang tidak digunakan akan di hapus. Selanjutnya atribut yang dipergunakan adalah Nama Satuan Pendidikan, Rerata Bahasa Indonesia, Rerata Bahasa Inggris, Rerata Matematika, dan Rerata IPA. Maka akan terlihat perubahannya seperti tabel 3.4

Tabel 3. 4 Keterangan awal dataset data kedua

Atribut	Tipe data	Keterangan
Nama Satuan Pendidikan	Char	Nama Kota/Kabupaten
Rata-rata Nilai Bahasa Indonesia	Integer	Hasil nilai Bahasa Indonesia di tiap Kota/Kabupaten
Rata-rata Nilai Bahasa Inggris	Integer	Hasil nilai Bahasa Inggris di tiap Kota/Kabupaten
Rata-rata Nilai Matematika	Integer	Hasil nilai Matematika di tiap Kota/Kabupaten
Rata-rata Nilai IPA	Integer	Hasil nilai IPA di tiap Kota/Kabupaten

2. Data Pre-processing dan Cleaning

Normalisasi data dalam penelitian ini tidak dilakukan karena rata-rata nilai hasil UN berada pada rentang 0 – 100, sehingga tidak terdapat parameter nilai yang mendominasi di dalam perhitungan jarak antar data yang akan dilakukan pada tahap selanjutnya.

3. Data Integration dan Transformation

Setelah data pre-processing selesai dilakukan, maka langkah selanjutnya adalah mengumpulkan semua dataset ke dalam satu dataset yang disebut dengan *data integration*.

Tabel 3. 5 Dataset nilai Ujian Nasional

Nama Satuan Pendidikan	Rata-rata Nilai Bahasa Indonesia	Rata-rata Nilai Bahasa Inggris	Rata-rata Nilai Matematika	Rata-rata Nilai IPA
SMP NEGERI 1 TULUNGAGUNG	88,02	83,54	87,11	80,15
SMP KATOLIK SANTA MARIA	78,05	65,39	57,46	58,02
SMP NEGERI 1 MOJOKERTO	84,69	75,46	72,32	72,06

Lanjutan tabel 3.5

Nama Satuan Pendidikan	Rata-rata Nilai Bahasa Indonesia	Rata-rata Nilai Bahasa Inggris	Rata-rata Nilai Matematika	Rata-rata Nilai IPA
SMP NEGERI 24 MALANG	79,7	66,63	58,45	61,48
SMP NEGERI 27 MALANG	74,19	49,21	45,2	50,99
SMP NEGERI SATU ATAP MERJOSARI	73,74	52,95	45,14	51,73
SMP PLUS AL - KAUTSAR	80,31	69,73	57,54	60,42
…	…	…	…	…
SMP NEGERI 9 SURABAYA	82,74	72,1	75,5	72,6

Setelah data di gabungkan maka dilakukan proses *transformation,* yaitu mengganti atau merubah nama variabel yang ada agar memudahkan proses selanjutnya yaitu *data mining* seperti terlihat pada tabel 3.6

Tabel 3. 6 Pergantian nama variabel data

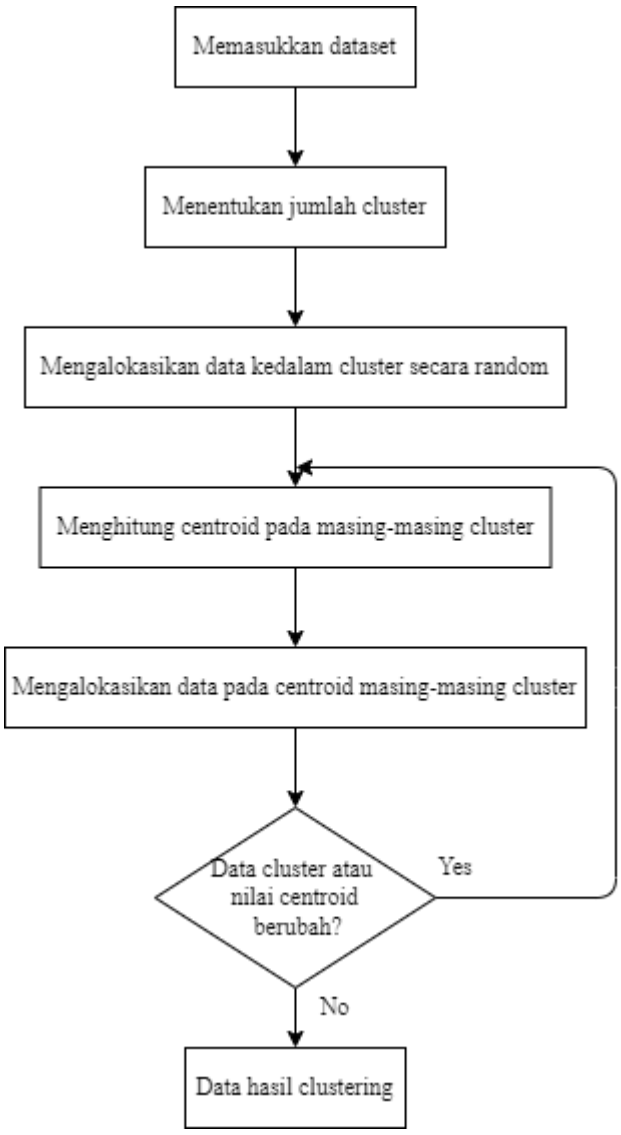
Nama variabel lama	Nama variabel baru
Nama Satuan Pendidikan	NAMA SEKOLAH
Rata-rata Nilai Bahasa Indonesia	IND
Rata-rata Nilai Bahasa Inggris	ING
Rata-rata Nilai Matematika	MAT
Rata-rata Nilai IPA	IPA

4. Data mining

Setelah dilakukan data transformasi, maka dilakukan proses penambangan (data mining) menggunakan teknik *clustering* dengan algoritma K-Means dan Fuzzy C-Means. Pada tahapan ini, proses data mining menggunakan Bahasa pemrograman python yang dijalankan pada Jupiter Notebook.

a. Implementasi K-Means

Jumlah *cluster* yang akan digunakan dalam penelitian ini ada tiga, selanjutnya menggunakan formula Euclidean Distance untuk menentukan jarak centroid dan dilakukan iterasi berulang-ulang hingga titik pusat *centroid* dari setiap *cluster* tidak berubah dan tidak lagi ditemukan data yang berpindah dari satu *cluster* ke *cluster* yang lain.



Gambar 3. 4 Alur pengolahan data pada K-Means

Berikut tahapan pengolahan data pada algoritma K-Means:

1. Memasukkan dataset hasil transformasi data yang dilakukan sebelumnya.

2. Metode Elbow digunakan untuk mencari nilai *K* optimum. Selain metode Silhouette, metode yang melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu garis untuk menentukan jumlah *cluster* terbaik adalah metode Elbow. Apabila nilai antara *cluster* pertama dengan nilai *cluster* kedua membentuk sudut dalam grafik atau nilainya mengalami penurunan paling besar maka jumlah nilai *cluster* tersebut yang tepat (Putu et al., 2015). Dari hasil perhitungan menggunakan metode Elbow didapatkan bahwa nilai *K* Optimum yakni 3. Sehingga jumlah *cluster* yang digunakan sebanyak 3 *cluster*.

3. Dilakukan proses standarisasi data karena apabila terdapat perbedaan yang cukup besar diantara variabel yang diteliti dapat mengakibatkan kesalahan perhitungan pada analisis *cluster* yang mengakibatkan hasilnya menjadi kurang valid, untuk selanjutnya data tersebut dialokasikan kedalam *cluster* secara random.

4. Menghitung centroid pada masing-masing cluster menggunakan jarak Euclidean menggunakan persamaan $d_{ij} = \sqrt{\sum_{k=1}^{n}\{x_{ik} - x_{jk}\}^2}$

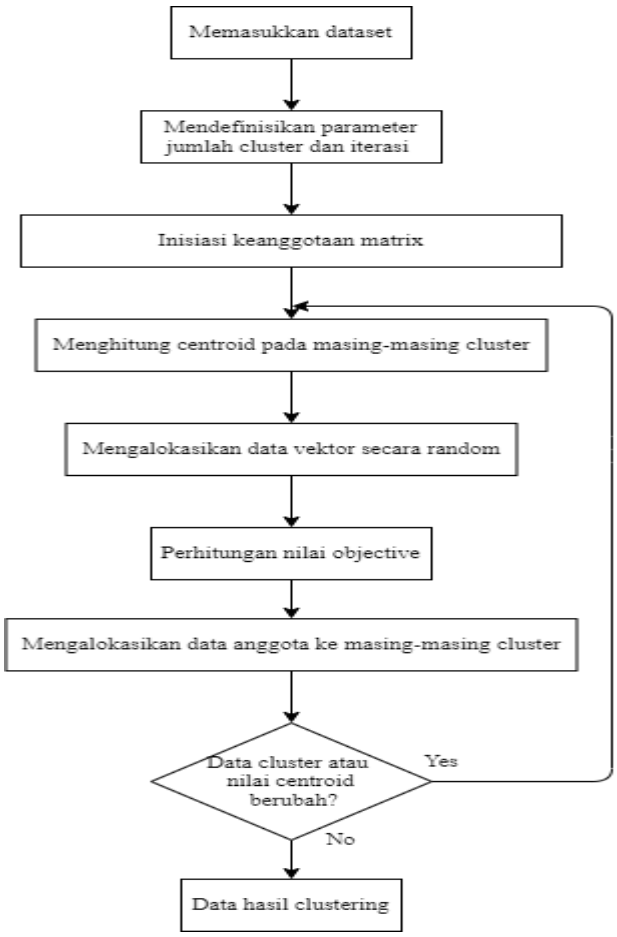
5. Mengalokasikan data pada centroid masing-masing cluster

6. Dilakukan iterasi berulang-ulang hingga titik pusat atau *centroid* dari setiap *cluster* tidak berubah dan tidak ada lagi data yang berpindah dari satu *cluster* ke *cluster* yang lain.

b. Implementasi Fuzzy C-Means

Tahap pertama pengaplikasian algoritma Fuzzy C-Means adalah membuat perhitungan kelas yang akan dijadikan basis klasifikasi terlebih dahulu. Selanjutnya dilakukan perhitungan atau iterasi sampai tiap kelompok mendapatkan anggotanya masing-masing. Tahapan yang dilakukan dengan algoritma *fuzzy* akan menghasilkan perhitungan yang halus.

Hasil perhitungan yang halus adalah data objek pengamatan tidak mutlak hanya menjadi anggota di satu kelompok, namun memungkinkan dapat menjadi anggota pada kelompok lain dengan ukuran derajat keanggotaan yang berbeda. Suatu objek akan cenderung menjadi anggota pada kelompok tertentu dimana derajat keanggotaan objek tersebut dalam kelompok itu paling besar dibandingkan dengan derajat keanggotaan pada kelompok lain.



Gambar 3. 5 Alur pengolahan data pada Fuzzy C-Means

Berikut tahapan pengolahan data pada algoritma Fuzzy C-Means:

1. Memasukkan data set hasil transformasi data yang dilakukan sebelumnya

2. Mendefinikan parameter yang akan digunakan dalam perhitungan. Untuk nilai K (jumlah cluster) ditentukan sebanyak tiga cluster, dengan percobaan atau iterasi sebanyak maksimal 100 kali dan parameter perhitungan Fuzzy diberi nilai 1.7

3. Nilai keanggotaan disimpan kedalam *matrix* fuzzy pseudo-partition dengan ukuran Nxk. dengan kolom adalah nilai keanggotaan pada setiap cluster dan baris merupakan data

```
[[0.70212766, 0.52238806, 0.4057971 ,0.43548387],
[0.68085106, 0.52238806, 0.52173913, 0.51612903],
[0.65957447, 0.46268657, 0.4057971 , 0.48387097],
...,
[0.85106383, 0.97014925, 0.85507246, 0.74193548],
[0.82978723, 0.97014925, 0.97101449, 0.85483871],
[0.80851064, 0.82089552, 0.82608696, 0.67741935]]
```

4. Setelah didapatkan anggota cluster dalam bentuk *matrix* selanjutkan mencari nilai titik pusat atau *centroid* dari masig-masing *cluster* menggunakan persamaan

$$C_{ij} = \frac{\sum_{i=1}^{n}(u_{il})\ w_{xij}}{\sum_{i=1}^{n}(u_{il})\ w}$$

dimana nilai *n* adalah jumlah data, sedangkan *w* adalah bobot pangkat dan *ui*l data x*i* ke cluster *ci*

5. Memasukkan data vector kedalam *cluster* secara random

6. Menghitung nilai objective untuk menguji validitas hasil *clustering* menggunakan algoritma Fuzzy C-Means, berikut index yang digunakan dalam pengujian:

a. Partition Coefficient Index (PCI) digunakan untuk menguji derajat keanggotaan tanpa melihat nilai vector data. Rentangan nilai pada PCI adalah [0,1] , kualitas cluster semakin baik bila nilai PCI semakin mendekati [1].

b. Partition Entropy Index (PEI) digunakan untuk mengevaluasi keteracakan data dalam cluster. Rentangan nilai pada PEI adalah [0,1], dimana kualitas cluster akan semakin baik bila nilai PEI semakin mendekati [0].

c. Menggunakan formula dari Fukuyama Sugeno Index (FSI) dimana mengukur fungsi objektif kohesi maupun separasi. Semakin kecil nilai FSI artinya kualitas cluster semakin baik.

7. Mengalokasikan data pada centroid masing-masing cluster

8. Dilakukan iterasi beberapa kali hingga tidak ada perubahan titik pusat atau centroid pada setiap cluster dan tidak lagi ditemukan data yang berpindah dari satu cluster ke cluster yang lain

BAB IV

HASIL PENELITIAN

A. Deskripsi Penelitian

Penelitian ini menggunakan Bahasa Pemrograman Python Ver 3 yang dijalankan pada software Jupyter Notebook. Data yang digunakan sudah melalui empat tahapan proses KDD (Knowledges Discovery in Databases), yaitu *data selection, data pre-processing, data integration, data transformation* untuk selanjutnya akan di olah menggunakan metode *clustering* algortima K-Means dan Fuzzy C-Means yang termasuk didalam metode unsupervised learning pada keilmuan *machine learning*, dan selanjutnya akan dilakukan *evaluation* terhadap hasil *clustering*.

B. Pengujian Hipotesis

1. Hasil analisis K-Means

Berikut langkah-langkah untuk mendapatkan hasil *clustering* menggunakan metode algoritma K-Means:

a. Memasukkan Libraries

Untuk menjalankan algoritma *clustering* K-Means pada python maka dibutuhkan beberapa library seperti pandas, numpy, seaborn, matplotlib dan scikit-learn.

```
In [1]: # Import package libraries

        import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt
        from sklearn.cluster import KMeans
        from sklearn.preprocessing import MinMaxScaler
```

Gambar 4. 1 Memasukkan library pada python

b. Memasukan dataset

Dataset yang digunakan pada penelitian ini berupa file *.csv yang selanjutnya diimport menggunakan library pandas. Dan pastikan bahwa hasil keluaran tidak mengalami kesalahan.

```
In [2]: # Input data

        sekolah = pd.read_csv("datasetsekolah.csv")
        sekolah.head(600)
```

Out[2]:

	NAMA SEKOLAH	IND	ING	MAT	IPA
0	SMP NEGERI 1 TULUNGAGUNG	88	84	87	80
1	SMP KATOLIK SANTA MARIA	78	65	57	58
2	SMP NEGERI 2 TULUNGAGUNG	81	70	69	72
3	SMP ISLAM AL AZHAAR TULUNGAGUNG	77	65	65	63
4	SMP NEGERI 6 TULUNGAGUNG	76	61	57	61
...	...	...	...	...	...
590	SMP NEGERI 58 SURABAYA	69	48	47	47
591	SMP NEGERI 59 SURABAYA	72	53	49	51
592	SMP KRISTEN ANAK PANAH SURABAYA	83	85	86	73
593	SMP LABSCHOOL UNESA SURABAYA	74	64	50	55
594	SMP IT UTSMAN BIN AFFAN SURABAYA	72	57	49	51

595 rows × 5 columns

Gambar 4. 2 Memasukkan data set pada python

c. Melihat informasi variabel pada data

Memastikan bahwa variabel pada data baik nama kolom maupun tipe data nya tidak salah.

```
In [3]: # Melihat informasi dari setiap variabel

        sekolah.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 595 entries, 0 to 594
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   NAMA SEKOLAH  595 non-null    object
 1   IND           595 non-null    int64 
 2   ING           595 non-null    int64 
 3   MAT           595 non-null    int64 
 4   IPA           595 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 23.4+ KB
```

Gambar 4. 3 Melihat informasi dari setiap variabel

d. Menentukan variabel yang akan di cluster

Variabel yang akan digunakan untuk *clustering* adalah nilai pada kolom IND, ING, MAT dan IPA. Sedangkan kolom NAMA SEKOLAH pada langkah ini akan dihapus.

```
In [4]: # Menentukan variabel yang akan di cluster
        # Dalam hal ini coloumn sekolah tidak digunakan
        sekolah = sekolah.drop(['NAMA SEKOLAH'], axis=1)
        sekolah_x = sekolah.iloc[:,0:4]
        sekolah_x.head(600)
```

Out[4]:

	IND	ING	MAT	IPA
0	88	84	87	80
1	78	65	57	58
2	81	70	69	72
3	77	65	65	63
4	76	61	57	61
...	...	...	...	...
590	69	48	47	47
591	72	53	49	51
592	83	85	86	73
593	74	64	50	55
594	72	57	49	51

595 rows × 4 columns

Gambar 4. 4 Menentukan variabel yang akan di cluster

e. Mengubah variabel ke dalam array

Agar nilai pada dataset dapat digunakan untuk menghitung nilai K, maka variabel yang sebelumnya berupa data frame perlu diubah ke dalam bentuk array.

```
In [5]: # Mengubah variabel yang sebelumnya berbentuk data frame menjadi array.
        x_array = np.array (sekolah_x)
        print (x_array)
```

```
[[88 84 87 80]
 [78 65 57 58]
 [81 70 69 72]
 ...
 [83 85 86 73]
 [74 64 50 55]
 [72 57 49 51]]
```

Gambar 4. 5 Mengubah variabel data frame menjadi array

f. Proses standarisasi data

Berdasarkan hasil pada gambar 4.5 ke empat variabel memiliki rentang nilai dengan variansi yang cukup besar. Proses standarisasi data dilakukan apabila terdapat perbedaan yang cukup besar diantara variabel yang diteliti yang dapat mengakibatkan perhitungan pada analisis cluster menjadi kurang valid.

```
In [6]: # Proses standarisasi agar scalling data berkisar 0-1
        scaler = MinMaxScaler()
        x_scaled = scaler.fit_transform(x_array)
        x_scaled

Out[6]: array([[0.91489362, 0.80597015, 0.84057971, 0.79032258],
               [0.70212766, 0.52238806, 0.4057971 , 0.43548387],
               [0.76595745, 0.59701493, 0.57971014, 0.66129032],
               ...,
               [0.80851064, 0.82089552, 0.82608696, 0.67741935],
               [0.61702128, 0.50746269, 0.30434783, 0.38709677],
               [0.57446809, 0.40298507, 0.28985507, 0.32258065]])
```

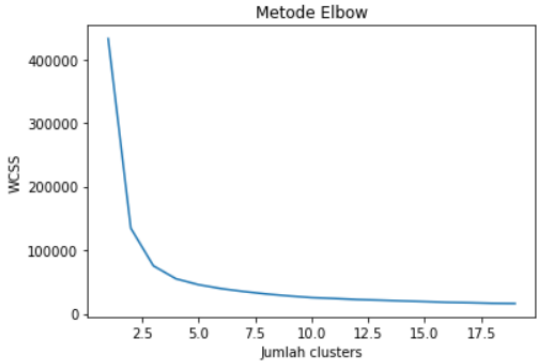
Gambar 4. 6 Proses standarisasi data

g. Menentukan nilai K dengan metode Elbow

Selain metode Silhouette, metode yang melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu garis untuk menentukan jumlah *cluster* terbaik adalah metode Elbow. Apabila nilai antara kelompok pertama dengan nilai kelompok kedua membentuk sudut dalam grafik atau nilainya mengalami penurunan paling besar maka jumlah nilai *cluster* tersebut yang tepat (Putu et al., 2015). Dari hasil perhitungan menggunakan metode Elbow didapatkan bahwa nilai *K* Optimum yakni 3. Maka dari itu jumlah *cluster* yang digunakan sebanyak 3 *cluster*.

```
In [7]: from sklearn.cluster import KMeans
wcss = []
for i in range(1,20):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(sekolah_x)
    wcss.append(kmeans.inertia_)
plt.plot(range(1,20), wcss)
plt.title('Metode Elbow')
plt.xlabel('Jumlah clusters')
plt.ylabel('WCSS')
plt.show()
```

```
C:\Users\adinu\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:881:
n Windows with MKL, when there are less chunks than available threads. You
_NUM_THREADS=3.
  warnings.warn(
```



Gambar 4. 7 Hasil nilai K berdasarkan metode Elbow

h. Menentukan jumlah cluster

Berdasarkan nilai K maka pengelompokkan sekolah berdasarkan nilai UNBK Provinsi Jawa Timur dapat dikelompokkan ke dalam 3 kelompok.

```
In [8]: # Menentukan dan mengkonfigurasi fungsi kmeans
kmeans = KMeans(n_clusters = 3, random_state=123)
# Menentukan kluster dari data
kmeans.fit(x_scaled)
```

```
Out[8]: KMeans(n_clusters=3, random_state=123)
```

Gambar 4. 8 Menentukan jumlah cluster

i. Mencari nilai centroid

Setelah data dicluster secara random, maka data tersebut perlu di alokasikan pada cluster masing-masing menggunakan centroid masing-masing cluster.

```
In [9]: # Mencari nilai pusat dari masing masing cluster
        print(kmeans.cluster_centers_)
```

```
[[0.642726   0.44760967 0.35967042 0.40379892]
 [0.35575063 0.22095815 0.15080572 0.18903677]
 [0.79920213 0.79704158 0.69746377 0.67151498]]
```

Gambar 4. 9 Mencari nilai centroid setiap cluster

j. Menampilkan hasil cluster

Dengan menggunakan library sckit-learn selanjutnya adalah menambahkan kolom “Cluster” yang akan menampilkan hasil *clustering*.

```
In [10]: # Menampilkan hasil kluster
         print(kmeans.labels_)
         # Menambahkan kolom "Cluster" dalam data frame sekolah
         sekolah["Cluster"] = kmeans.labels_
         print(sekolah)
```

```
[2 0 2 0 0 1 2 0 1 2 1 1 0 1 2 1 0 1 0 1 0 1 1 0 0 0 0 0 2 1 1 0 2 1 1 2 0
 0 0 2 0 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 0 1 1 1 1 0 0 1 1 1 0
 0 0 1 1 0 2 1 2 0 0 1 2 1 0 1 0 0 1 1 0 1 0 0 0 0 2 1 0 0 1 1 0 1 0 1 2 1
 1 0 1 0 2 1 1 1 0 1 1 0 1 1 0 1 0 0 1 0 2 0 2 2 0 0 0 0 0 0 2 1 1 1 2 0 1 0
 1 0 1 2 0 0 2 2 0 1 2 1 2 1 2 2 0 0 0 1 2 1 2 2 0 0 0 0 0 1 1 1 2 1 1 0 2 1
 1 0 0 1 0 1 0 2 1 0 0 1 0 1 1 1 1 0 1 1 1 0 0 1 1 0 1 1 1 0 1 1 1 1 2 1 0
 1 1 0 0 2 1 2 1 1 0 1 2 1 0 2 0 0 0 0 2 1 0 2 1 0 2 0 1 2 1 0 2 0 2 1 2 2
 0 0 0 0 0 0 0 0 2 1 1 0 1 1 2 0 0 2 0 2 2 1 1 1 2 0 2 1 1 1 1 2 1 2 1 0 0
 1 1 1 1 1 0 1 1 2 0 1 1 1 1 1 0 1 1 1 1 0 1 0 0 2 0 1 0 1 0 1 0 0 1 1 1 1
 2 0 1 1 2 1 0 0 0 1 1 0 1 1 1 2 2 2 2 2 2 1 1 1 0 1 1 1 0 0 1 0 2 1 2 1 0
 1 1 0 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 2 1 1 0 1 1 0 1 1 0 1
 1 0 0 0 1 0 0 1 1 2 1 1 0 0 1 1 1 0 1 1 1 0 1 0 1 1 1 0 1 2 1 1 1 0 1 0 0 1 2
 1 2 1 1 1 1 2 1 1 1 0 1 1 1 1 2 2 2 0 0 1 1 1 2 1 2 2 1 2 0 1 1 1 0 2 2 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 2 1 1 1 0 1 1 1 0 1 1 1 1 1 1 2
 1 2 0 0 1 1 1 1 0 0 1 0 1 1 2 0 1 2 2 0 2 1 1 0 0 0 1 1 0 2 0 0 0 2 2 0 1
 2 0 0 0 0 2 0 2 2 2 0 1 0 2 0 0 0 2 0 2 2 2 2 0 2 2 2 0 0 2 2 2 0 0 0 1 0
 2 0 0]
     IND  ING  MAT  IPA  Cluster
0     88   84   87   80        2
1     78   65   57   58        0
2     81   70   69   72        2
3     77   65   65   63        0
4     76   61   57   61        0
..   ...  ...  ...  ...      ...
590   69   48   47   47        1
591   72   53   49   51        0
592   83   85   86   73        2
593   74   64   50   55        0
594   72   57   49   51        0

[595 rows x 5 columns]
```

Gambar 4. 10 Menampilkan hasil clustering

k. Kelompok hasil *clustering*

Berdasarkan hasil *clustering*, didapatkan data kelompok sebagai berikut:

1) Kelompok *cluster* sekolah dengan nilai hasil Ujian Nasional sedang dengan label *cluster* 0 berjumlah 198 sekolah

```
In [12]: # Menghitung jumlah Cluster 0
         print(sekolah.loc[sekolah['Cluster'] == 0])
         (sekolah['Cluster']==0).sum()

             IND  ING  MAT  IPA  Cluster
         1    78   65   57   58        0
         3    77   65   65   63        0
         4    76   61   57   61        0
         7    76   58   53   58        0
         12   73   52   52   53        0
         ..   ...  ...  ...  ...      ...
         588  70   52   47   52        0
         589  71   52   44   50        0
         591  72   53   49   51        0
         593  74   64   50   55        0
         594  72   57   49   51        0

         [198 rows x 5 columns]

Out[12]: 198
```

Gambar 4. 11 Kelompok cluster 0

2) Kelompok *cluster* sekolah dengan nilai hasil Ujian Nasional rendah dengan label cluster 1 berjumlah 285 sekolah

```
In [13]: # Menghitung jumlah Cluster 1
         print(sekolah.loc[sekolah['Cluster'] == 1])
         (sekolah['Cluster']==1).sum()

             IND  ING  MAT  IPA  Cluster
         5    49   37   34   36        1
         8    68   48   46   47        1
         10   66   48   50   45        1
         11   60   46   40   42        1
         13   62   44   40   44        1
         ..   ...  ...  ...  ...      ...
         544  58   42   40   39        1
         545  54   43   36   39        1
         554  63   57   43   45        1
         566  71   48   43   48        1
         590  69   48   47   47        1

         [285 rows x 5 columns]

Out[13]: 285
```

Gambar 4. 12 Kelompok cluster 1

3) Kelompok cluster sekolah dengan nilai hasil Ujian Nasional tinggi dengan label cluster 2 berjumlah 112 sekolah.

```
In [14]: # Menghitung jumlah Cluster 2
         print(sekolah.loc[sekolah['Cluster'] == 2])
         (sekolah['Cluster']==2).sum()

              IND  ING  MAT  IPA  Cluster
         0     88   84   87   80        2
         2     81   70   69   72        2
         6     83   74   73   75        2
         9     83   67   73   68        2
         14    83   72   75   72        2
         ..   ...  ...  ...  ...      ...
         581   75   95   77   69        2
         584   78   89   68   76        2
         585   85   95   88   77        2
         586   84   95   96   84        2
         592   83   85   86   73        2

         [112 rows x 5 columns]

Out[14]: 112
```

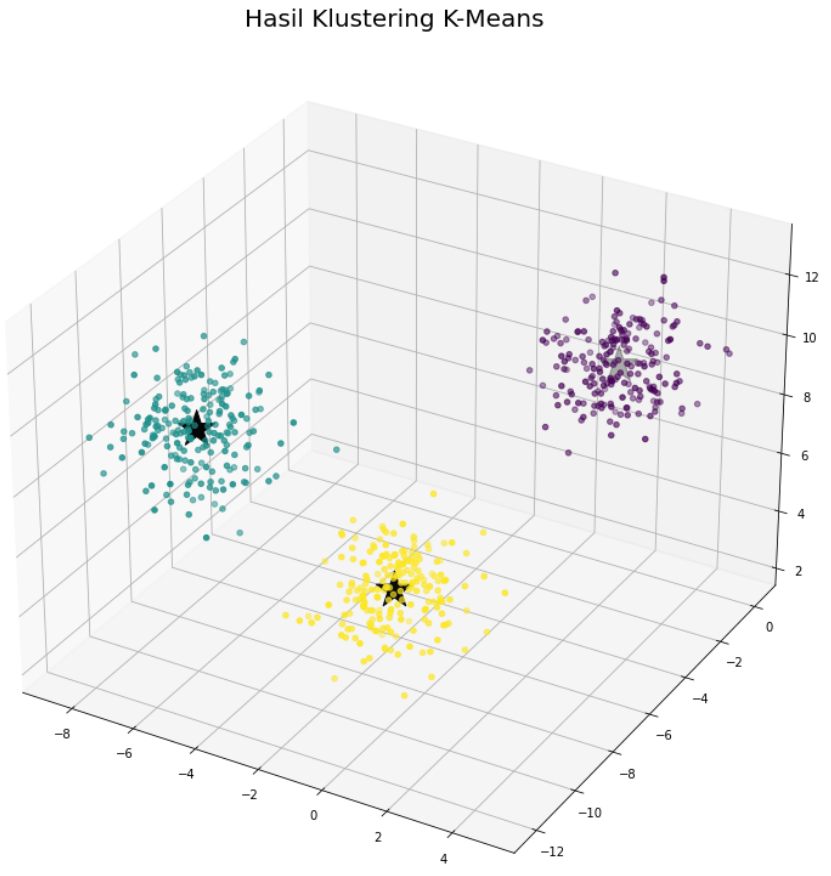
Gambar 4. 13 Kelompok cluster 2

l. Visualisasi hasil *clustering* K-Means

Untuk memvisualisasikan hasil *clustering* menggunakan plot 3D yang setiap kelompook diwakilkan dengan warna berbeda.

- Warna titik hijau menunjukkan cluster 0 berjumlah 198 sekolah
- Warna titik kuning menunjukkan cluster 1 berjumlah 285 sekolah
- Warna titik ungu menunjukkan cluster 2 berjumlah 112 sekolah
- Tanda bintang ★menunjukkan titik centroid masing-masing cluster.

```
In [22]: from mpl_toolkits.mplot3d import Axes3D
         from sklearn.datasets import make_blobs

         plt.rcParams['figure.figsize'] = (10,10)

         # Creating a sample dataset with 3 clusters
         X, y = make_blobs(n_samples=595, n_features=3, centers=3)
         # Initializing KMeans
         kmeans = KMeans(n_clusters=3)
         # Fitting with inputs
         kmeans = kmeans.fit(X)
         # Predicting the clusters
         labels = kmeans.predict(X)
         # Getting the cluster centers
         C = kmeans.cluster_centers_
         fig = plt.figure()
         ax = Axes3D(fig)
         ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=y)
         ax.scatter(C[:, 0], C[:, 1], C[:, 2], marker='*', c='#050505', s=1000)
         plt.title("Hasil Klustering K-Means", fontsize =20)

Out[22]: Text(0.5, 0.92, 'Hasil Klustering K-Means')
```



Gambar 4. 14 Visualisasi hasil clustering K-Means

m. Data hasil clustering K-Means

Berikut data hasil clustering kelompok cluster sekolah dengan nilai hasil Ujian Nasional tinggi dengan label cluster 2 berjumlah 112 sekolah.

Tabel 4. 1 Sekolah dengan hasil UN tinggi

NAMA SEKOLAH	IND	ING	MAT	IPA	CLUSTER
SMP NEGERI 4 KEDIRI	85	74	72	73	2
SMP AR - RISALAH KOTA KEDIRI	82	80	74	79	2
SMP NEGERI 1 MADIUN	89	85	84	83	2
SMP ST. BERNARDUS MADIUN	81	80	64	66	2
SMP NEGERI 2 MADIUN	87	80	74	73	2
SMP NEGERI 4 MADIUN	82	71	64	68	2
---	---	---	---	---	---
SMP KRISTEN LOGOS	90	97	87	88	2

Berikut data hasil *clustering* kelompok *cluster* sekolah dengan nilai hasil Ujian Nasional sedang dengan label cluster 0 berjumlah 198 sekolah.

Tabel 4. 2 Sekolah dengan hasil UN sedang

NAMA SEKOLAH	IND	ING	MAT	IPA	CLUSTER
SMP KATOLIK SANTA MARIA	78	65	57	58	0
SMP ISLAM AL AZHAAR TULUNGAGUNG	77	65	65	63	0
SMP NEGERI 6 TULUNGAGUNG	76	61	57	61	0
SMP NEGERI 4 TULUNGAGUNG	76	58	53	58	0
SMP NEGERI 1 NGANTRU	73	52	52	53	0
---	---	---	---	---	---
SMP IT UTSMAN BIN AFFAN SURABAYA	72	57	49	51	0

Berikut data hasil *clustering* kelompok *cluster* sekolah dengan nilai hasil Ujian Nasional rendah dengan label cluster 1 berjumlah 285 sekolah.

Tabel 4. 3 Sekolah dengan hasil UN rendah

NAMA SEKOLAH	IND	ING	MAT	IPA	CLUSTER
SMP ISLAM AL - AMIN	61	42	38	40	1
SMP WASKITA DHARMA	58	47	42	38	1
SMP PGRI 4	61	39	40	38	1
SMP KRISTEN ELIM	64	47	45	45	1
SMP BAHRUL MAGHFIROH	67	47	47	47	1
---	---	---	---	---	---
SMP MUHAMMADIYAH 1 MOJOKERTO	69	56	43	47	1

2. Hasil analisis Fuzzy C-Means

Berikut tahapan proses *clustering* menggunakan metode Fuzzy C-Means:

a. Memasukkan Libraries

Untuk menjalankan algoritma *clustering* Fuzzy C-Means pada python maka dibutuhkan beberapa library seperti pandas, numpy, random, operator, matplotlib dan scikit-learn.

```
In [1]: # IMPORT PACKAGE LIBRARIES
        import pandas as pd # reading all required header files
        import numpy as np
        import random
        import operator
        import math
        import matplotlib.pyplot as plt
        from scipy.stats import multivariate_normal     # for generating pdf
        from sklearn.preprocessing import MinMaxScaler
```

Gambar 4. 15 Memasukkan library pada python

b. Memasukan dataset

Dataset yang dipergunakan di dalam penelitian berupa file *.csv yang akan diimport menggunakan library pandas. Dan pastikan bahwa hasil keluaran tidak mengalami kesalahan.

```
In [2]: # INPUT DATA
        df_full = pd.read_csv("datasetcmeans.csv")
        df_full.head()
```

Out[2]:

	NAMA SEKOLAH	IND	ING	MAT	IPA	CLUSTER	Unnamed: 6
0	SMP KATOLIK SANTA MARIA	78	65	57	58	NOL	NaN
1	SMP ISLAM AL AZHAAR TULUNGAGUNG	77	65	65	63	NOL	NaN
2	SMP NEGERI 6 TULUNGAGUNG	76	61	57	61	NOL	NaN
3	SMP NEGERI 4 TULUNGAGUNG	76	58	53	58	NOL	NaN
4	SMP NEGERI 1 NGANTRU	73	52	52	53	NOL	NaN

Gambar 4. 16 Memasukkan data set pada python

c. Melihat informasi variabel pada data

Memastikan bahwa variabel pada data baik nama kolom maupun tipe data nya tidak salah.

```
In [3]: # Melihat informasi dari setiap variabel
        df_full.info
```

```
Out[3]: <bound method DataFrame.info of                          NAMA SEKOLAH  IND  ING  MAT  IPA CLUSTER  Unnamed: 6
        0                SMP KATOLIK SANTA MARIA   78   65   57   58     NOL         NaN
        1      SMP ISLAM AL AZHAAR TULUNGAGUNG   77   65   65   63     NOL         NaN
        2              SMP NEGERI 6 TULUNGAGUNG   76   61   57   61     NOL         NaN
        3              SMP NEGERI 4 TULUNGAGUNG   76   58   53   58     NOL         NaN
        4                  SMP NEGERI 1 NGANTRU   73   52   52   53     NOL         NaN
        ..                                  ...  ...  ...  ...  ...     ...         ...
        590     SMP SURABAYA CAMBRIDGE SCHOOL   75   95   77   69     DUA         NaN
        591                         SMP BELL   78   89   68   76     DUA         NaN
        592    SMP JALINAN ADISISWA CEMERLANG   85   95   88   77     DUA         NaN
        593              SMP LITTLE SUN SCHOOL   84   95   96   84     DUA         NaN
        594  SMP KRISTEN ANAK PANAH SURABAYA   83   85   86   73     DUA         NaN

        [595 rows x 7 columns]>
```

Gambar 4. 17 Melihat informasi dari setiap variabel

d. Menentukan variabel yang akan di cluster

Variabel yang akan digunakan untuk *clustering* adalah nilai pada kolom IND, ING, MAT dan IPA. Sedangkan kolom NAMA SEKOLAH dan CLUSTER pada langkah ini akan dihapus.

```
In [4]: # Menentukan variabel yang akan di cluster, dalam hal ini coloumn sekolah tidak digunakan
df_full = df_full.drop(['NAMA SEKOLAH', 'Unnamed: 6'], axis=1)

columns = list(df_full.columns)
features = columns[:len(columns)-1]
class_labels = list(df_full[columns[-1]])
df = df_full[features]
print (df)
```

```
     IND  ING  MAT  IPA
0     78   65   57   58
1     77   65   65   63
2     76   61   57   61
3     76   58   53   58
4     73   52   52   53
..   ...  ...  ...  ...
590   75   95   77   69
591   78   89   68   76
592   85   95   88   77
593   84   95   96   84
594   83   85   86   73

[595 rows x 4 columns]
```

Gambar 4. 18 Menentukan variabel yang akan di cluster

e. Mengubah variabel kedalam array

Agar nilai pada dataset dapat digunakan untuk menghitung nilai K, maka variabel yang sebelumnya berupa data frame perlu diubah ke dalam bentuk array. Selanjutnya dilakukan proses standarisasi data agar perhitungan analisis cluster valid.

```
In [5]: # PLOTTING DATA
sepal_df = df_full.iloc[:,0:4]
sepal_df = np.array(sepal_df) # Mengubah variabel yang sebelumnya berbentuk data frame menjadi array.
scaler = MinMaxScaler() # Proses standarisasi agar scalling data berkisar 0-1
x_scaled = scaler.fit_transform(sepal_df)
x_scaled
```

```
Out[5]: array([[0.70212766, 0.52238806, 0.4057971 , 0.43548387],
               [0.68085106, 0.52238806, 0.52173913, 0.51612903],
               [0.65957447, 0.46268657, 0.4057971 , 0.48387097],
               ...,
               [0.85106383, 0.97014925, 0.85507246, 0.74193548],
               [0.82978723, 0.97014925, 0.97101449, 0.85483871],
               [0.80851064, 0.82089552, 0.82608696, 0.67741935]])
```

Gambar 4. 19 Mengubah variabel data frame menjadi array

f. Mendefinisikan parameter

Sebelum dilakukan perhitungan analisis cluster, maka diperlukan penetapan nilai-nilai yang akan digunakan untuk perhitungan. Untuk nilai K (jumlah cluster) ditentukan sebanyak tiga cluster, dengan percobaan atau iterasi sebanyak maksimal 100 kali dan parameter perhitungan Fuzzy diberi nilai 1.7.

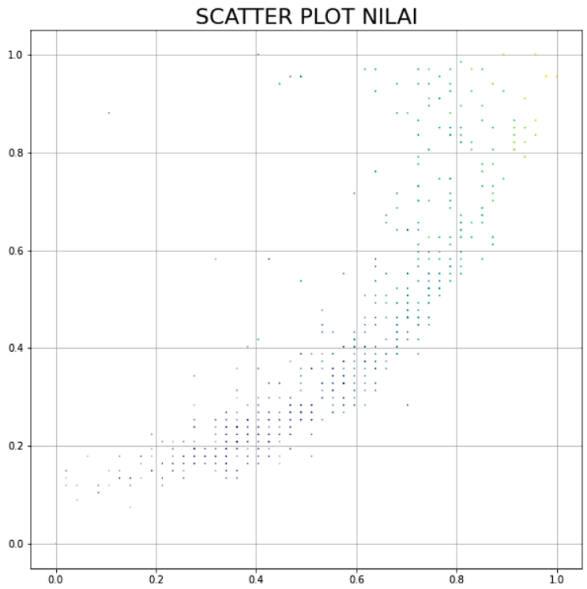
```
In [6]: # DEFINING PARAMETERS
        k = 3 # Number of Clusters
        MAX_ITER = 100 # Maximum number of iterations
        n = len(df) # Number of data points
        m = 1.7 # Fuzzy parameter - Select a value greater than 1 else it will be knn
```

Gambar 4. 20 Mendefiniskan parameter

g. Visualisasi Scatter plot awal

Berikut merupakan hasil visualisasi scatter plot awal dari data yang akan digunakan untuk perhitungan analisis cluster Fuzzy C-Means.

```
In [7]: #SCATTER PLOTS
        plt.figure(figsize=(10,10))
        plt.scatter(list(x_scaled[:,0]), list(x_scaled[:,1]), list(x_scaled[:,2]), list(x_scaled[:,3]), marker='o')
        plt.title('SCATTER PLOT NILAI', fontsize=22)
        plt.grid()
        plt.show()
```



Gambar 4. 21 Visualisasi dataset di awal

h. Menghitung akurasi

Perhitungan akurasi digunakan untuk mendapatkan nilai terbaik dari hasil prediksi dataset.

```
In [8]: #Calculating the accuracy
        def accuracy(cluster_labels, class_labels):
            correct_pred = 0
            #print(cluster_labels)
            seto = max(set(labels[0:198]), key=labels[0:198].count)
            vers = max(set(labels[198:483]), key=labels[198:483].count)
            virg = max(set(labels[483:]), key=labels[483:].count)

            for i in range(len(df)):
                if cluster_labels[i] == seto and class_labels[i] == 'NOL':
                    correct_pred = correct_pred + 1
                if cluster_labels[i] == vers and class_labels[i] == 'SATU' and vers!=seto:
                    correct_pred = correct_pred + 1
                if cluster_labels[i] == virg and class_labels[i] == 'DUA' and virg!=seto and virg!=vers:
                    correct_pred = correct_pred + 1

            accuracy = (correct_pred/len(df))*100
            return accuracy
```

Gambar 4. 22 Menghitung akurasi

i. Inisiasi keanggotaan matrix

Pengelompokkan dataset kedalam setiap cluster dilakukan dalam bentuk keanggotaan matrix.

```
In [9]: # initializing the membership matrix
        def initializeMembershipMatrix():
            membership_mat = []
            for i in range(n):
                random_num_list = [random.random() for i in range(k)]
                summation = sum(random_num_list)
                temp_list = [x/summation for x in random_num_list]

                flag = temp_list.index(max(temp_list))
                for j in range(0,len(temp_list)):
                    if(j == flag):
                        temp_list[j] = 1
                    else:
                        temp_list[j] = 0

                membership_mat.append(temp_list)
            return membership_mat

In [10]: membership_mat = initializeMembershipMatrix()
```

Gambar 4. 23 Inisasi anggota matrix

j. Mencari nilai centroid

Setelah didapatkan anggota cluster dalam bentuk matrix selanjutkan mencari nilai centroid dari masig-masing cluster.

```
In [11]: # calculating the cluster center
         def calculateClusterCenter(membership_mat):
             cluster_mem_val = list(zip(*membership_mat))
             cluster_centers = []
             for j in range(k):
                 x = list(cluster_mem_val[j])
                 xraised = [p ** m for p in x]
                 denominator = sum(xraised)
                 temp_num = []
                 for i in range(n):
                     data_point = list(x_scaled[i])
                     prod = [xraised[i] * val for val in data_point]
                     temp_num.append(prod)
                 numerator = map(sum, list(zip(*temp_num)))
                 center = [z/denominator for z in numerator]
                 cluster_centers.append(center)
             return cluster_centers

In [12]: #cluster_centers = calculateClusterCenter(membership_mat)
         calculateClusterCenter(membership_mat)

Out[12]: [[0.5245017810724936,
           0.3935300871209563,
           0.3128073972063742,
           0.33819880309443867],
          [0.5401697477037563,
           0.41285376396705,
           0.3319078165835117,
           0.36065573770491804],
          [0.5398239946530018,
           0.4090021098695007,
           0.32582138250246606,
           0.3564431683837193]]
```

Gambar 4. 24 Mencari nilai centroid

k. Update nilai keanggotaan

Diperlukan memperbarui nilai dari setiap keanggotan masing-masing cluster berdasarkan keanggotaan matrix dan nilai centroid yang telah didapatkan sebelumnya untuk kemudian dicari titik cluster nya.

```
In [13]: # Updating the membership value
def updateMembershipValue(membership_mat, cluster_centers):
    p = float(2/(m-1))
    for i in range(n):
        x = list(x_scaled[i])
        distances = [np.linalg.norm(np.array(list(map(operator.sub, x, cluster_centers[j])))) for j in range(k)]
        for j in range(k):
            den = sum([math.pow(float(distances[j]/distances[c]), p) for c in range(k)])
            membership_mat[i][j] = float(1/den)
    return membership_mat

In [14]: # getting the clusters
def getClusters(membership_mat):
    cluster_labels = list()
    for i in range(n):
        max_val, idx = max((val, idx) for (idx, val) in enumerate(membership_mat[i]))
        cluster_labels.append(idx)
    return cluster_labels
```

Gambar 4. 25 Update nilai keanggotaan

l. Iterasi secara random vector

Prores iterasi pada Fuzzy C-Means dapat menggunakan salah satu dari tiga metode iterasi. Pertama Fuzzy C-Means dengan pusat cluster di titik asal yaitu 0. Kedua Fuzzy C-Means dengan pusat cluster di lokasi acak dalam menggunakan metode distribusi Gaussian multivariat dengan mean-nol dan varians unit. Ketiga Fuzzy C-Means dengan pusat cluster pada vektor secara acak yang dipilih dari data.

```
In [15]: #Third iteration Random vectors from data
def fuzzyCMeansClustering():
    # Membership Matrix
    membership_mat = initializeMembershipMatrix()
    curr = 0
    acc=[]
    while curr < MAX_ITER:
        cluster_centers = calculateClusterCenter(membership_mat)
        membership_mat = updateMembershipValue(membership_mat, cluster_centers)
        cluster_labels = getClusters(membership_mat)

        acc.append(cluster_labels)

        if(curr == 0):
            print("Cluster Centers:")
            print(np.array(cluster_centers))
        curr += 1
    print("-----------------------------")
    print("Partition matrix:")
    print(np.array(membership_mat))
    #return cluster_labels, cluster_centers
    return cluster_labels, cluster_centers, acc
```

Gambar 4. 26 Iterasi random vector

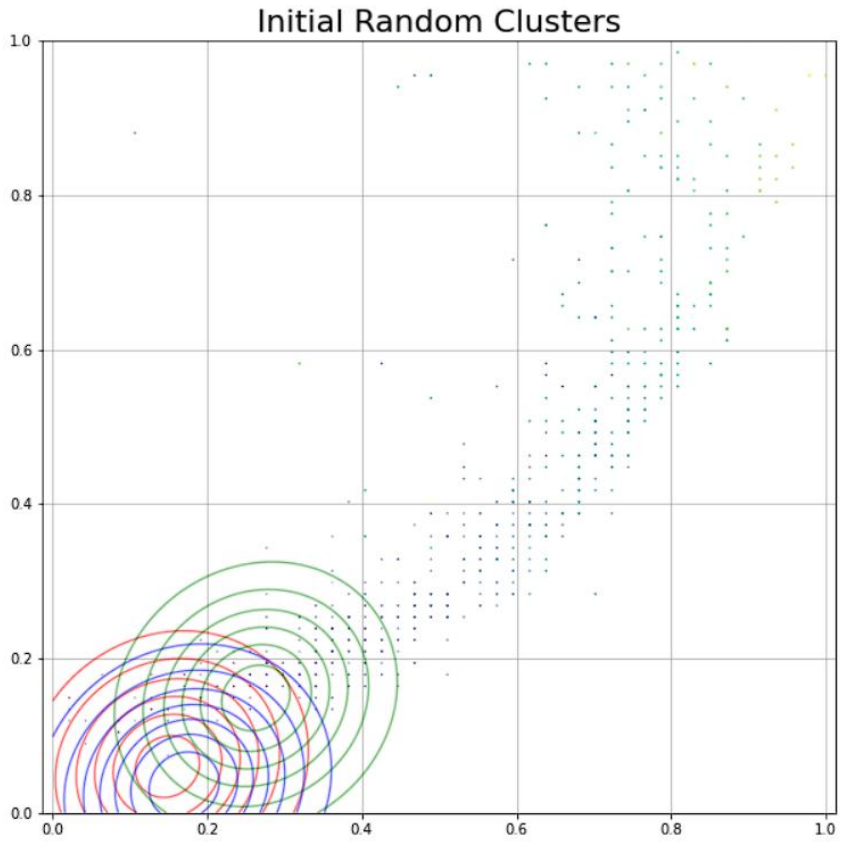
m. Mencari nilai mean dan standard deviasi

Dari hasil perhitungan didapatkan nilai mean sebesar 97.18, nilai standard deviasi sebesar 4.32 dengan tingkat akurasi sebesar 98.32.

```
In [18]: #calculating accuracy and std deviation 100 times
         acc_lis = np.array(acc_lis)
         print("mean=",np.mean(acc_lis))
         print("Std dev=",np.std(acc_lis))
         print("Accuracy = " + str(round(a, 2)))

         mean= 97.18823529411767
         Std dev= 4.322554136189892
         Accuracy = 98.32

In [19]: #final cluster centers
         print("Cluster center vectors:")
         print(np.array(centers))

         Cluster center vectors:
         [[0.34959217 0.21619861 0.1479333  0.18447831]
          [0.63801302 0.43887813 0.35157184 0.39716972]
          [0.80495749 0.79191917 0.69965504 0.67267899]]
```

Gambar 4. 27 Mencari nilai mean dan standard deviasi

n. Visualisasi awal random clusters



Gambar 4. 28 Initial random cluster

o. Mencari keanggotaan setiap cluster

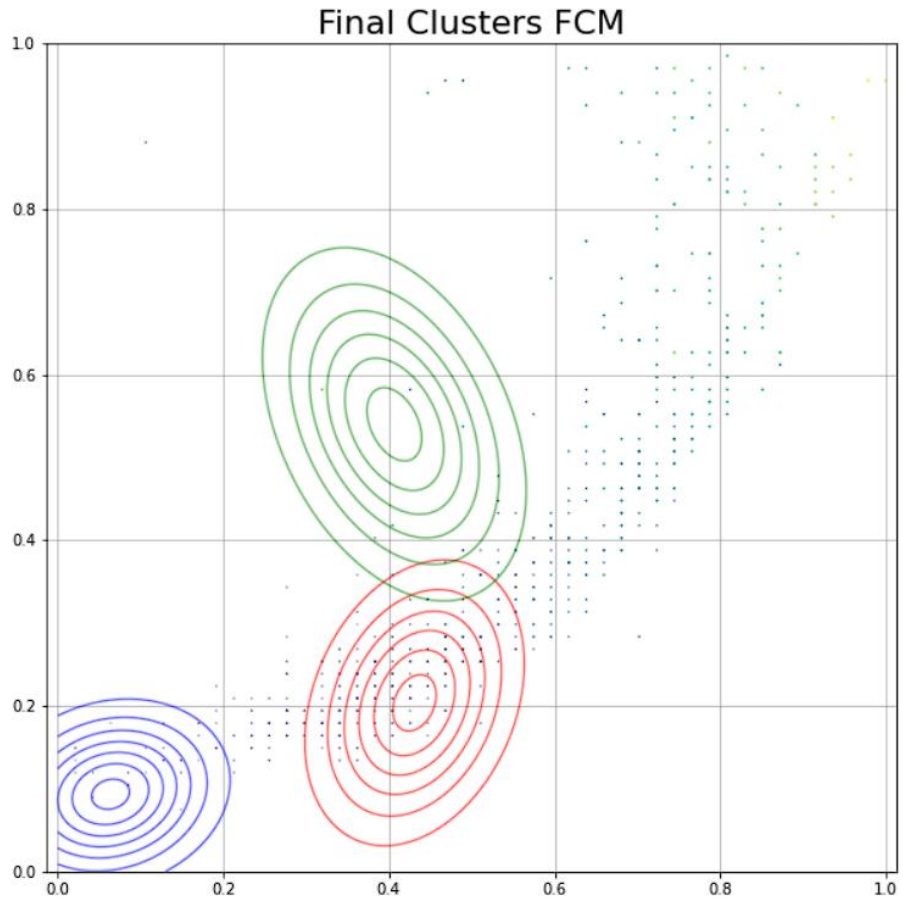
Langkah terakhir adalah mencari anggota dimasing-masing cluster yang kemungkinan dapat menjadi anggota di cluster yang yang berbeda.

```
In [22]: #finding mode
         seto = max(set(labels[0:198]), key=labels[0:198].count)
         vers = max(set(labels[198:483]), key=labels[198:483].count)
         virg = max(set(labels[483:]), key=labels[483:].count)

In [23]: #mean cluster
         s_mean_clus1 = np.array([centers[seto][0],centers[seto][1],centers[seto][2],centers[seto][3]])
         s_mean_clus2 = np.array([centers[vers][0],centers[vers][1],centers[vers][2],centers[vers][3]])
         s_mean_clus3 = np.array([centers[virg][0],centers[virg][1],centers[virg][2],centers[virg][3]])

In [24]: values = np.array(labels) #Label

         #search all 3 cluster
         searchval_seto = seto
         searchval_vers = vers
         searchval_virg = virg

         #index of all 3 cluster
         ii_seto = np.where(values == searchval_seto)[0]
         ii_vers = np.where(values == searchval_vers)[0]
         ii_virg = np.where(values == searchval_virg)[0]
         ind_seto = list(ii_seto)
         ind_vers = list(ii_vers)
         ind_virg = list(ii_virg)

In [25]: x_scaled
         sepal_df = pd.DataFrame(x_scaled)
         sepal_df

In [26]: seto_df = sepal_df[sepal_df.index.isin(ind_seto)]
         vers_df = sepal_df[sepal_df.index.isin(ind_vers)]
         virg_df = sepal_df[sepal_df.index.isin(ind_virg)]

In [27]: cov_seto = np.cov(np.transpose(np.array(seto_df)))
         cov_vers = np.cov(np.transpose(np.array(vers_df)))
         cov_virg = np.cov(np.transpose(np.array(virg_df)))

In [28]: sepal_df = np.array(sepal_df)

In [29]: x1 = np.linspace(0,1,595)
         x2 = np.linspace(0,1,595)
         X, Y = np.meshgrid(x1,x2)

         Z1 = multivariate_normal(s_mean_clus1, cov_seto)
         Z2 = multivariate_normal(s_mean_clus2, cov_vers)
         Z3 = multivariate_normal(s_mean_clus3, cov_virg)

         pos = np.empty(X.shape + (4,))              # a new array of given shape and type, without initializing entries
         pos[:, :, 0] = X; pos[:, :, 1] = Y

         plt.figure(figsize=(10,10))                                                    # creating the figure and assigning the size
         plt.scatter(sepal_df[:,0], sepal_df[:,1], sepal_df[:,2], sepal_df[:,3], marker='o')
         plt.contour(X, Y, Z1.pdf(pos), colors="r" ,alpha = 0.5)
         plt.contour(X, Y, Z2.pdf(pos), colors="b" ,alpha = 0.5)
         plt.contour(X, Y, Z3.pdf(pos), colors="g" ,alpha = 0.5)
         plt.axis('equal')                                                          # making both the axis equal
         #plt.xlabel('Sepal Length', fontsize=16)                                        # X-Axis
         #plt.ylabel('Sepal Width', fontsize=16)                                         # Y-Axis
         plt.title('Final Clusters FCM', fontsize=22)                                   # Title of the plot
         plt.grid()                                                                 # displaying gridlines
         plt.show()
```

Gambar 4. 29 Mencari keanggotaan setiap cluster

p. Visualisasi final cluster Fuzzy C-Means

Visualisasi hasil *clustering* menggunakan scatter plot terlihat bahwa ada beberapa anggota pada cluster sekolah medium yang berwarna merah dapat menjadi anggota pada cluster sekolah unggulan yang berwarna hijau.



Gambar 4. 30 Visualisasi hasil cluster Fuzzy C-Means

C. Analisis cluster dalam perspektif Al-Qur'an

Manusia merupakan satu dari beberapa makhluk yang diciptakan oleh Alloh Subhanahu Wata'ala. Diukur dari struktur tubuh dan organ, dalam kajian ilmu pengetahuan banyak peneliti menyatakan adanya kesamaan antara manusia

dan hewan. Dikarenakan manusia mempunyai akal yang harus digunakan dengan sebaik-baiknya, rasa bijaksana dari nurani yang dimiliki manusia, dan budi pekerti yang digunakan untuk saling menghormati dan norma kesopanan antar sesama maka manusia sangatlah berbeda dengan hewan dan makhluk lain ciptaan Alloh Subhanahu Wata'ala (Shihab, 2015).

Alloh Subhanahu wata'ala secara tersirat dan tersurat membagi konsep manusia kedalam beberapa *cluster*. Ada tiga kata yang umum digunakan Al-Qur'an untuk merujuk kepada arti manusia dengan segala bentuknya, yaitu *al-insan*, *al-basyar* dan *bani Adam* atau *zurriyat Adam* (Cahaya Kesuma, 2013).

1. Konsep Al Insan

Didalam Al-Qur'an sebutan *al-insan* yang berasal dari kata *al-uns*, disebutkan sebanyak 73 kali dan tersebar pada 43 surat. Kata *al-insan* dapat diartikan lemah lembut, harmonis, tampak dan pelupa secara etimologi. Berasal dari akar kata *naus*, *al-insan* mempunyai arti pergerakan atau dinamis. Apabila dirujuk dari asal katanya dapat dipahami apabila manusia pada dasarnya mempunyai potensi positif untuk tumbuh dan berkembang baik secara fisik maupun mental spiritual. Selain itu, manusia juga dibekali oleh Alloh dengan potensi lainnya, yang mampu untuk mendorong manusia ke arah tindakan, sikap, serta perilaku negatif dan merugikan.

Al-insan digunakan oleh Al-Qur'an untuk menunjukkan bahwa manusia sudah sempurna sebagai makhluk jasmani dan rohani. Hubungan aspek jasmani dan rohani tersebut dengan berbagai potensi yang dimilikinya menjadikan manusia sebagai makhluk ciptaan Alloh yang unik dan berbeda antara satu dengan yang lain,

istimewa dan sempurna sebagai pribadi dan makhluk sosial, sehingga mampu menerima tanggung jawab sebagai kholifah di muka bumi.

Perpaduan antara aspek jasmani dan rohani mendorong manusia untuk mengekspresikan arti dari *al-insan*, yaitu sebagai makhluk yang memiliki budaya, mampu berkomunikasi, dapat membedakan hal baik dan buruk, dan lain sebagainya.

Dengan kemampuan yang dimiliki, manusia mampu mengemban tanggung jawab yang diberikan Alloh dimuka bumi secara utuh dengan nuansa ilahiah dan hanif. Tanggung jawab itu antara lain dapat membentuk, mengembangkan diri dan komunitasnya sesuai dengan nilai-nilai insaniah kebaikan. Pribadi yang berintegritas akan terlihat pada nilai-nilai keimanan dan pengamalannya. Namun disayangkan, manusia seringkali lalai bahkan melupakan nilai-nilai insaniah yang dimilikinya dengan berbuat berbagai bentuk kerusakan di muka bumi.

2. Konsep Al Basyar

Di dalam Al-qur'an kata *al-basyar* disebutkan sebanyak 36 kali dan tersebar pada 26 surat. *Al-basyar* secara etimologi memiliki arti *mulamasah*, yakni bersentuhnya kulit laki-laki dengan perempuan. Bisa dimengerti bahwasanya manusia merupakan makhluk yang memiliki segala kebutuhan kemanusiaan yang terbatas, seperti makan minum, hubungan seksual, rasa aman, kebahagiaan, dan lain sebagainya. Penggunaan kata *al-basyar* pada Al-Qur'an ditujukan kepada seluruh manusia tanpa kecuali, baik yang beriman maupun tidak, termasuk kepada para rasul Nya (Tanjung, 2020).

Merujuk pada konsep *al-basyar*, manusia tidak berbeda dengan makhluk biologis lainnya, yang artinya dalam hidup manusia terikat kepada prinsip kehidupan biologis seperti halnya keinginan untuk berkembang biak, mengalami fase tumbuh kembang dari bayi menjadi dewasa dan perkembangan serta kematangan berfikir logis.

Manusia dalam pengertian *al-basyar* ini banyak dijelaskan dalam Al-Qur‟an, diantaranya dalam QS.Ibrahim ayat 10

قَالَتْ رُسُلُهُمْ اَفِى اللّٰهِ شَكٌّ فَاطِرِ السَّمٰوٰتِ وَالْاَرْضِۗ يَدْعُوْكُمْ لِيَغْفِرَ لَكُمْ مِّنْ ذُنُوْبِكُمْ وَيُؤَخِّرَكُمْ اِلٰٓى اَجَلٍ مُّسَمًّىۗ قَالُوْٓا اِنْ اَنْتُمْ اِلَّا بَشَرٌ مِّثْلُنَاۗ تُرِيْدُوْنَ اَنْ تَصُدُّوْنَا عَمَّا كَانَ يَعْبُدُ اٰبَآؤُنَا فَأْتُوْنَا بِسُلْطٰنٍ مُّبِيْنٍ - ١٠

Artinya: *Berkata rasul-rasul mereka "Apakah ada keragu-raguan terhadap Alloh, Pencipta langit dan bumi? Dia menyeru kamu untuk memberi ampunan kepadamu dari dosa-dosamu dan menangguhkan (siksaan)mu sampai masa yang ditentukan?" Mereka berkata: "Kamu tidak lain hanyalah manusia seperti kami juga. Kamu menghendaki untuk menghalang-halangi (membelokkan) kami dari apa yang selalu disembah nenek moyang kami, karena itu datangkanlah kepada kami, bukti yang nyata"*

3. Konsep Bani Adam

Penyebutan Bani Adam didalam Al-Qur’an lebih menekankan pada peringatan kepada umat manusia agar selalu bersyukur atas nikmat yang selalu diberikan oleh Alloh, seperti nikmat pemberian kemuliaan hidup, hasil alam baik di darat dan di laut, pemberian rizki maupun kedudukan di atas makhluk ciptaan Alloh lainnya seperti yang terkandung dalam QS. Al-Isra' ayat 70

وَلَقَدْ كَرَّمْنَا بَنِيْٓ اٰدَمَ وَحَمَلْنٰهُمْ فِى الْبَرِّ وَالْبَحْرِ وَرَزَقْنٰهُمْ مِّنَ الطَّيِّبٰتِ وَفَضَّلْنٰهُمْ عَلٰى كَثِيْرٍ مِّمَّنْ خَلَقْنَا تَفْضِيْلًا ع - ٧٠

Ikatan janji suci antara manusia dengan pencipta nya untuk tidak menyembah selain Alloh dan bersaksi bahwa Alloh adalah Tuhannya sesuai isi QS. Yaasiin ayat 60

اَلَمْ اَعْهَدْ اِلَيْكُمْ يٰبَنِيْٓ اٰدَمَ اَنْ لَّا تَعْبُدُوا الشَّيْطٰنَۚ اِنَّهٗ لَكُمْ عَدُوٌّ مُّبِيْنٌ - ٦٠

Selanjutnya yang sudah membagikan pakaian ketakwaan yang harus manusia pergunakan setiap kali manusia pergi ke tempat ibadah diseluruh penjuru bumi sesuai isi QS.al-A῾raaf ayat 31

يٰبَنِيْٓ اٰدَمَ خُذُوْا زِيْنَتَكُمْ عِنْدَ كُلِّ مَسْجِدٍ وَّكُلُوْا وَاشْرَبُوْا وَلَا تُسْرِفُوْاۚ اِنَّهٗ لَا يُحِبُّ الْمُسْرِفِيْنَ ۚ - ٣١

Menurut Thabathaba'i dalam kitab Samsul Nizar, penggunaan kata *bani adam* yang menunjuk pada pengertian manusia secara umum ini setidaknya ada tiga aspek yang perlu diketahui, yaitu:

1. Anjuran kepada manusia untuk berbudaya sesuai dengan ketentuan Alloh, seperti berpakaian dengan sopan guna manutup auratnya.
2. Mengingatkan pada seluruh keturunan Nabi Adam agar jangan terjerumus pada bujuk rayu setan yang mengajak kepada keingkaran dan mendustakan Alloh.
3. Manusia hendaknya memanfaatkan semua yang ada di alam semesta ini dalam rangka untuk beribadah dan menyembah Alloh Subhanahu wata'ala.

BAB V

PEMBAHASAN

Berdasarkan hasil pengujian hipotesis menggunakan metode algoritma K-Means dan metode algoritma Fuzzy C-Means menunjukkan bahwa kedua metode tersebut dapat digunakan untuk *clustering* hasil Ujian Nasional SMP di Provinsi Jawa Timur.

Metode algoritma K-Means membagi dataset kedalam tiga cluster secara tegas, yaitu cluster sekolah dengan nilai UN rendah, cluster sekolah dengan nilai UN sedang dan cluster sekolah dengan nilai UN tinggi dengan perhitungan jarak masing-masing cluster menggunakan Euclidian distance.

Pada cluster sekolah dengan nilai UN rendah didapatkan hasil 285 sekolah, pada cluster sekolah dengan nilai UN sedang didapatkan hasil 198 sekolah dan pada cluster sekolah dengan nilai UN tinggi didapatkan hasil 112 sekolah.

Sedangkan pada algoritma Fuzzy C-Means dataset dibagi kedalam tiga cluster secara lembut (soft) dimana beberapa anggota sekolah yang terdapat pada cluster sekolah dengan nilai UN sedang bisa menjadi anggota pada cluster sekolah dengan nilai UN tinggi. Begitu pula sebaliknya, dimana beberapa anggota sekolah yang terdapat pada cluster sekolah dengan nilai UN tingggi bisa menjadi anggota sekolah pada cluster sekolah dengan nilai UN sedang.

Perhitungan jarak pada Fuzzy C-Means menggunakan Euclidian distance pada jarak antar vector. Euclidean distance merupakan fungsi matriks jarak yang

paling banyak digunakan dengan tingkat identifikasi kemiripan (similarity) lebih tinggi dibanding metode yang lain.

Dengan demikian hasil hipotesis pada penelitian ini sesuai dengan penelitian-penelitian yang sudah ada sebelumnya yang menggunakan metode yang sama, yaitu metode K-Means dan Fuzzy C-Means. Walaupun terdapat perbedaan terhadap tools yang digunakan dan rentang periode dataset yang berbeda.

BAB VI

PENUTUP

A. Kesimpulan

Berdasarkan hasil penelitian, dapat dibuat kesimpulan sebagai berikut:

1. Metode algoritma K-Means dan algoritma Fuzzy C-Means dapat dipergunakan melakukan *clustering* pada hasil Ujian Nasional SMP di Provinsi Jawa Timur.

2. Pengalokasian anggota kedalam tiap-tiap *cluster*, bila metode K-Means menggunakan metode pengalokasian yang bersifat tegas (hard) maka metode Fuzzy C-Means menggunakan metode pengalokasian yang bersifat lunak (Soft).

3. Pada metode Fuzzy C-Means kemungkinan tingkat kegagalan konvergen lebih kecil dibandingkan K-Means.

4. Didapatkan hasil titik-titik pusat *cluster* pada metode K-Means yaitu:

[[0.642726 0.44760967 0.35967042 0.40379892]

[0.35575063 0.22095815 0.15080572 0.18903677]

[0.79920213 0.79704158 0.69746377 0.67151498]]

Hasil titik-titik pusat *cluster* pada metode Fuzzy C-Means yaitu:

[[0.34959217 0.21619861 0.1479333 0.18447831]

[0.63801302 0.43887813 0.35157184 0.39716972]

[0.80495749 0.79191917 0.69965504 0.67267899]]

5. Hasil *clustering* dapat digunakan sebagai basis pengetahuan bagi SMK dalam melakukan program PPDB.

B. Implikasi Teoritis

Data hasil Ujian Nasional SMP Provinsi Jawa Timur dapat dikelompokkan menjadi tiga *cluster* sekolah menggunakan metode algoritma K-Means maupun alogoritma Fuzzy C-Means.Pada algoritma K-Means pengalokasian anggota masing-masing *cluster* bersifat tegas (hard) sedangkan pada algoritma Fuzzy C-Means bersifat lembut (soft) dan sudah sesuai dengan teori yang telah ada.

C. Saran

Saran untuk pengembangan penelitian lebih lanjut sebagai berikut:

1. Penelitian selanjutnya dapat dikembangkan dengan menggunakan metode *cluster* hirarki seperti single linkage dan average linkage.
2. Hasil *clustering* dari penelitian ini dapat digunakan menjadi basis pengetahuan untuk mendukung keputusan maupun rekomendasi sekolah unggulan dalam program PPDB di Sekolah Menengah Kejuruan.

DAFTAR PUSTAKA

Aditya, A., Jovian, I., & Sari, B. N. (2020). Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019. *Jurnal Media Informatika Budidarma*, *4*(1), 51. https://doi.org/10.30865/mib.v4i1.1784

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Arai, K., & Ridho Barakbah, A. (2007). Hierarchical K-Means: an algorithm for centroids initialization for K-Means. *Rep. Fac. Sci. Engrg. Reports of the Faculty of Science and Engineering*, *36*(1), 25–31.

As'Sidiq, A., & Mandala, R. (2020). Implementation of K-Means Algorithm for Information Technology Freshman Class Division. *IT for Society*, *4*(1), 1–6. https://doi.org/10.33021/itfs.v4i1.1170

Cahaya Kesuma, G. W. (2013). Konsep Fitrah Manusia Perspektif Pendidikan Islam. *Ijtimaiyya*, *6*(2), 79–94. https://media.neliti.com/media/publications/69573-ID-konsep-fitrah-manusia-perspektif-pendidi.pdf

Gorunescu, F. (2011). Introduction to data mining. *Intelligent Systems Reference Library*, *12*, 1–43. https://doi.org/10.1007/978-3-642-19721-5_1

Hackeling, G. (2014). Mastering Machine Learning with scikit-learn. In *Book*. http://books.google.com/books?id=fZQeBQAAQBAJ&pgis=1

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. Elsevier Inc. https://doi.org/10.1016/C2009-0-61819-5

Juliya, M., & Herlambang, Y. T. (2021). ANALISIS PROBLEMATIKA

PEMBELAJARAN DARING DAN PENGARUHNYA TERHADAP MOTIVASI BELAJAR SISWA. *Genta Mulia : Jurnal Ilmiah Pendidikan*, *12*(1).

Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition. In *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition* (Vol. 9780470908). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118874059

Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *Jurnal Informatika: Jurnal Pengembangan IT*, *4*(1), 20–24. https://doi.org/10.30591/jpit.v4i1.1253

Putu, N., Merliana, E., & Santoso, A. J. (2015). *Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means*. 978–979.

Rutledge, J. (2009). The Top Ten Algorithms in Data Mining. *Journal of Quality Technology*, *41*(4), 441–441. https://doi.org/10.1080/00224065.2009.11917798

Selviana, N. I., & Mustakim. (2016). Analisis perbandingan K-Means dan Fuzzy C-Means untuk pemetaan motivasi balajar mahasiswa. *Seminar Nasional Teknologi Informasi, Komunikasi Dan Industri (SNTIKI) 8*, *01*(01), 95–105.

Shihab, M. Q. (2015). *Dia dimana-mana: "Tangan" Tuhan di balik setiap fenomena.* Penerbit Lentera Hati.

Suputra, W. A. (2021). Klasterisasi Hasil Ujian Nasional SMA/MA dengan Algoritma K-Means. *Wahana Matematika Dan Sains: Jurnal …*, *15*(1), 22–30. https://ejournal.undiksha.ac.id/index.php/JPM/article/view/25380

Tan, P.N., Steinbech, M., & Kumar, V. (2006). *Introduction to data mining. Boston: Pearsong Education, Ltd*.

Tanjung, M. (2020). Konsep Manusia Dalam Perspektif Filsafat Pendidikan Islam.

An Nadwah, *25*(1), 46–63. http://jurnal.uinsu.ac.id/index.php/nadwah/article/view/7480

Teskey, F. N. (1989). User models and world models for data, information, and knowledge. *Information Processing and Management*, *25*(1), 7–14. https://doi.org/10.1016/0306-4573(89)90087-3

Utomo, C. E. W., Hariadi, M., & Sumpeno, S. (2020). Clustering Data National Examinations based on Social Media Using K-Means Methods. *JAREE (Journal on Advanced Research in Electrical Engineering)*, *4*(2). https://doi.org/10.12962/J25796216.V4.I2.152

Yudi Agusta. (2007). K-Means – Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sistem Dan Informatika*, *3*(Februari), 47–60.