Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik

Yogasetya Suhanda[1)*)], Ike Kurniati[2)], Siti Norma[3)]

[1)2)3)] Program Studi Sistem Informasi, Institut Teknologi dan Bisnis Swadharma
[*)]Correspondence Author: yogasetyas@swadharma.ac.id, Jakarta, Indonesia
DOI: https://doi.org/10.37012/jtik.v6i2.299

Abstrak

Segmentasi adalah usaha untuk membagi suatu populasi menjadi kelompok-kelompok yang dapat dibedakan satu sama lain. Segmentasi juga dapat digunakan untuk mendapatkan informasi yang berguna mengenai karakteristik mahasiswa dan dapat dijadikan bahan masukan untuk menyusun program-program akademik yang lebih baik. Permasalahan yang ada saat ini adalah banyak mahasiswa yang lulus tidak tepat waktu. Penyebab kegagalan mahasiswa dalam kelulusan diantaranya indeks presentasi yang rendah, kurangnya interaksi mahasiswa dengan dosen di kelas saat mata kuliah berlangsung, absensi ataupun dari faktor lain. Penelitian ini menggunakan metode CRISP-DM (Cross-Industry Standard Process Model for Data Mining) dengan algoritma K-Means clustering untuk menghasilkan clustering mahasiswa berdasarkan kemampuan akademik. Data yang diolah merupakan data mahasiswa dari tahun 1992 – 2019 dengan total data mencapai 253.886 data. Hasil pemodelan menghasilkan sistem dashboard yang menampilkan hasil clustering mahasiswa berdasarkan program studi, persentase nilai mahasiswa berdasarkan cluster, perolehan nilai mahasiswa berdasarkan jenis kelamin pada masing-masing cluster dan informasi IPK Mahasiswa berdasarkan cluster pada periode tahun 2009 -2018. Hasil penelitian diharapkan dapat digunakan untuk oleh manajemen ITBS untuk mendukung pengambilan keputusan strategi khususnya di bidang akademik.

Kata Kunci: Segmentasi, Crisp-dm, K-means

Abstract

Segmentation is an attempt to divide a population into groups that can be distinguished from one another. Segmentation can also be used to obtain useful information about student characteristics and can be used as input for developing better academic programs. The problem that exists today is that many students do not graduate on time. The causes of student failure in graduating include low presentation index, lack of student interaction with lecturers in class during the course, attendance, or other factors. This research using the CRISP-DM (Cross-Industry Standard Process Model for Data Mining) method with the K-Means clustering algorithm to produce student clustering based on academic ability. The data processed is student data from 1992 - 2019 with a total data of 253,886 data. The modeling results produce a dashboard system that displays the results of student clustering based on the study program, the percentage of student scores based on clusters, the acquisition of student grades based on gender in each cluster, and student GPA information based on clusters in the period 2009-2018. The research results are expected to be used by ITBS management to support strategic decision making, especially in the academic field..

Keywords: *Segmentation, Crisp-dm, K-means*

PENDAHULUAN

Segmentasi adalah usaha untuk membagi suatu populasi menjadi kelompok-kelompok yang dapat dibedakan satu sama lain. Segmentasi tidak hanya digunakan dalam kegiatan bisnis saja tetapi telah berkembang ke berbagai bidang termasuk pendidikan. Dalam bidang pendidikan, segmentasi dapat digunakan untuk mendapatkan informasi yang berguna mengenai karakteristik mahasiswa dan dapat dijadikan bahan masukan untuk menyusun program-program akademik yang lebih baik.

Permasalahan saat ini yaitu jumlah mahasiswa yang lulus dan mahasiswa baru yang masuk setiap tahunnya tidak sebanding, karena banyak mahasiswa yang lulus tidak tepat waktu. Penyebab kegagalan mahasiswa dalam kelulusan diantaranya indeks presentasi yang rendah, kurangnya interaksi mahasiswa dengan dosen di kelas saat mata kuliah berlangsung, absensi ataupun faktor lain. Untuk itu diperlukan suatu metode untuk mengetahui kemampuan mahasiswa yang cepat tanggap ataupun yang lambat dalam pemahaman materi yang diberikan dosen agar dapat membantu dosen untuk lebih mengenal situasi mahasiswa dan dapat dijadikan sebagai pengetahuan dini untuk mengambil tindakan preventif seperti meningkatkan pengawasan serta wawasan kepada mahasiswa saat proses belajar mengajar dengan tujuan meningkatkan prestasi mahasiswa.

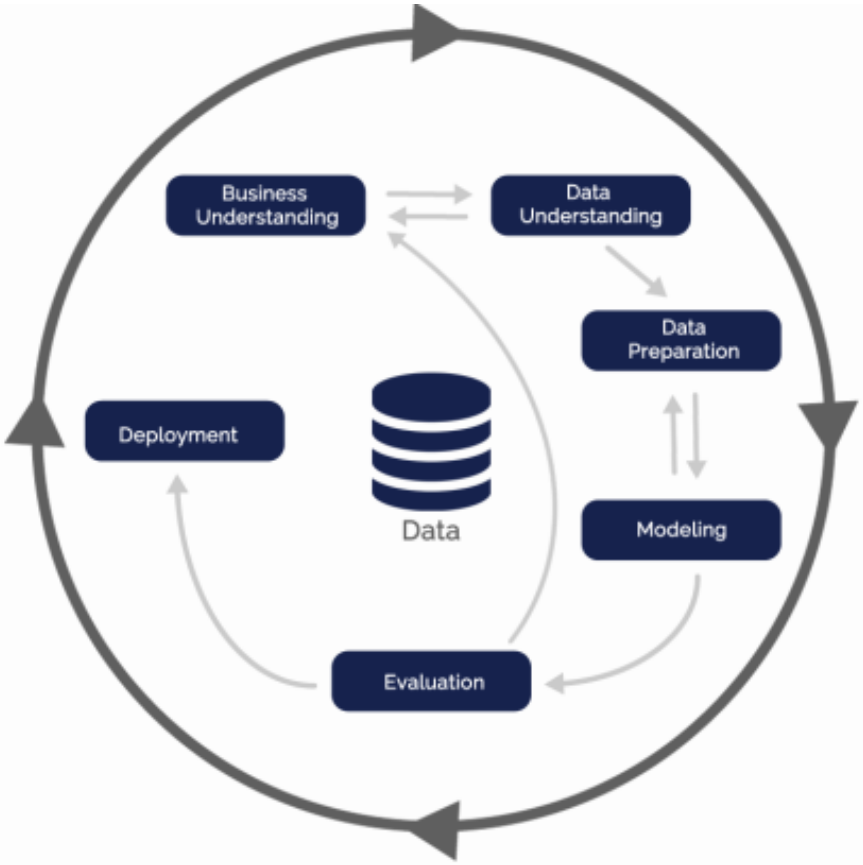
Data mining adalah proses menemukan pola yang menarik dan pengetahuan dari data yang berjumlah besar. Secara garis besar data mining dikelompokkan menjadi 2 kategori utama yaitu predictive dan descriptive. (Han, Kamber, & Pei, 2012). Metode yang banyak diterapkan dalam data mining adalah CRISP-DM. CRISP-DM (Cross-Industry Standard Process Model for Data Mining)menjelaskan tentang proses data mining dalam enam tahap yaitu (1) Business Understanding; (2) Data Understanding; (3) Data Preparation; (4) Modeling; (5) Evaluation; (6) Deployment (Chapman et al., 2000).

K-means merupakan salah satu metode data clustering non hirarki untuk clustering data yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster/kelompok berdasarkan atribut menjadi k partisi, dimana k < n . Algoritma k-means merupakan algoritma yang membutuhkan parameter input sebanyak k dan membagi sekumpulan n objek kedalam k cluster sehingga tingkat kemiripan antar anggota dalam satu cluster tinggi sedangkan tingkat kemiripan dengan anggota pada cluster lain sangat rendah. Kemiripan anggota terhadap cluster diukur dengan kedekatan objek terhadap nilai mean pada cluster atau dapat disebut sebagai centroid cluster atau pusat massa (Widyawati, 2010).

Penerapan algoritma K-Means dalam pengelompokan data mahasiswa sudah dilakukan sebelumnya oleh (Asroni & Adrian, 2015) dan (Poerwanto & Fa'rifah, 2016). Hal ini yang mendorong penelitian untuk melakukan segmentasi kualitas mahasiswa menggunakan metode CRISP-DM dengan algoritma K-Means Clustering untuk menghasilkan suatu sistem dashboard yang akan menampilkan data akademik mahasiswa berdasarkan cluster yang dipilih pengguna.

METODE

Pelaksanaan penelitian dilakukan dengan menggunakan metode Cross Industry Standard Process Model for Data Mining (CRISP-DM). Berikut gambaran alur proses yang terjadi pada CRISP-DM:



Gambar 1. Alur CRISP DM

Pengumpulan data dengan cara extract data melalui database Mahasiswa yang tersimpan pada Sistem Informasi Akademik. Aplikasi yang digunakan untuk mengolah data adalah SPSS Modeler Penerapan metode CRIPS-DM dalam segmentasi mahasiswa ITB Swadharma sebagai berikut :

1. Business Understanding

   Pemahaman masalah penelitian mengacu pada segmentasi mahasiswa berdasarkan kualitas untuk tindakan preventif mahasiswa di ITB Swadharma. Pada tahapan ini diperlukan pemahaman tentang pentingnya pemanfaatan data Mahasiswa, agar dapat digunakan untuk mengetahui segmentasi Mahasiswa berdasarkan kualitas nilai yang dicapai dan merancang tindakan preventif yang akan diterapkan kepada mahasiswa dengan tujuan untuk meningkatkan kualitas mahasiswa berdasarkan nilai.

2. Data Understanding

Pemahaman data mengacu pada database Mahasiswa. Tahap memahami format data secara permukaan (format form dan report) dan secara lebih mendalam (bentuk fisik data). Berikut data yang terdapat dalam database antara lain:

a. Data Akademik Mahasiswa. Tabel terdapat 18 field, namun hanya 5 field yang digunakan pada pemrosesan data di SPSS modeler antara lain : Nama, Jurusan, Kelas, IPK, Semester.
b. Data Pribadi Mahasiswa. Tabel terdapat 26 field, namun hanya 2 field yang digunakan pada pemrosesan data di SPSS modeler yaitu Nim dan Jenis_kelamin.
c. Tabel KRS. Berisi tentang field-field KRS yang sudah diambil oleh mahasiswa. Terdapat 15 field, namun hanya 12 field yang digunakan pada pemrosesan data di SPSS modeler yaitu : Tahun, Nim, Semester, Kelas, Mata_Kuliah, SKS, Absen, Tugas, UTS, UAS_upm, Nilai_akhir, nilai.
d. Tabel Master Mahasiswa. Terdapat 16 field, namun hanya 5 field yang digunakan pada pemrosesan data di SPSS modeler yaitu : Nama, Jurusan, Kelas, IPK, Semester.
e. Tabel Data Preparation. Tabel dari hasil penggabungan 3 tabel yaitu, tabel data akademik, tabel mahasiswa pribadi dan tabel KRS, Tabel memiliki 17 field yang digunakan pada pemrosesan data di SPSS modeler yaitu : Nim, Nama, Jurusan, Kelas, IPK, Semester, Jenis_kelamin, Tahun, Semester, Mata_kuliah, SKS, Absen, Tugas, UTS, UAS_upm, Nilai_akhir, Nilai
f. Tabel Data Preparation 1. Merupakan hasil dari tabel data preparation dengan penambahan 1 field yaitu field jur.

3. Data Preparation

Data preparation adalah tahapan untuk memperbaiki masalah yang terdapat pada data sebelum data masuk ke tahap modeling sehingga menghasilkan modeling yang bagus. Langkah pertama yang dilakukan pada proses data preparation adalah proses pengumpulan data yang diperoleh dari database Mahasiswa ITB Swadharma, kemudian data yang diperoleh dikoneksikan ke database. Penjelasannya sebagai berikut:

a. Proses cleansing yaitu proses analisa kualitas dari suatu data dengan cara mengubah, mengoreksi atau menghapus data-data yang tidak sesuai dengan kebutuhan penelitian.
b. Penambahan kolom diperlukan jika pada basis data terdapat ketidak seragaman isi pada kolom, misal pada kolom jurusan tertulis SI-S1, TI-S1, MI-S1 dan MI-DIII

sehingga perlu dilakukan penyeragaman isi kolom yang memuat nama jurusan sehingga sama secara penulisan yaitu SI, TI dan MI.

c. Missing Value

Missing value adalah mencari nilai yang hilang atau data yang kosong untuk kemudian diisi dengan nilai nol, nilai rata-rata atau NULL. Proses missing value dilakukan karena banyak kolom yang kosong atau banyak data yang tidak terinput sehingga perlu mengisi kolom yang kosong tersebut sesuai dengan kebutuhan, karena jika terdapat data yang kosong pada tabel maka tidak bisa diproses pada SPSS modeller.

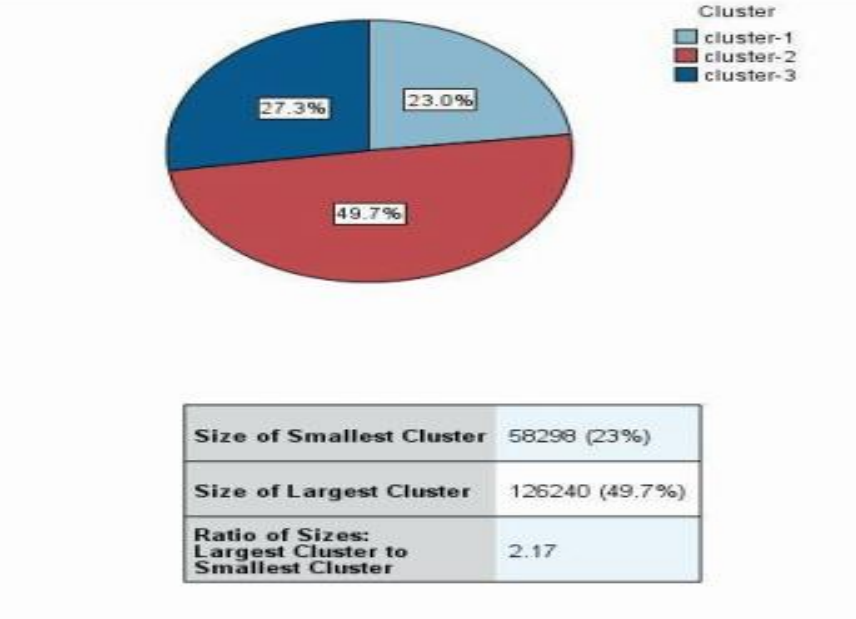
d. Merge

Proses menggabungkan 2 atau lebih data dengan tujuan untuk mendapatkan kolom-kolom pada tabel lain dan digabungkan menjadi satu tabel, pada penelitian ini proses merge dilakukan dengan menggabungkan Data Mahasiswa Akademik dengan Mahasiswa Pribadi dan data KRS, proses penggabungan tabel tersebut dilakukan dengan bantuan SPSS modeller serta menggunakan metode left join.

4. Data Modeling

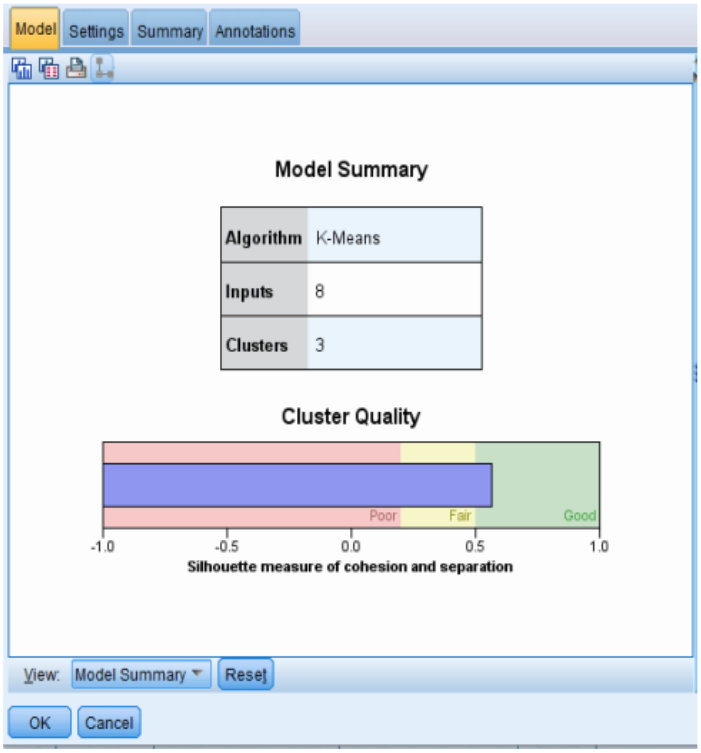
Data Modeling adalah tahapan untuk membuat model prediktif, yaitu untuk mengelompokan kualitas mahasiswa berdasarkan nilai Mahasiswa di ITB Swadharma. Pada tahap ini dapat menggunakan statistika dan *machine learning* untuk mendapatkan insight yang berguna dari data untuk mencapai tujuan penelitian.

Untuk melakukan clustering data mahasiswa pada ITB Swadaharma menggunakan algoriotma K-Means Cluster, berikut adalah hasil summary yang terbentuk pada SPSS modeler dengan menggunakan algoritma K-Means cluster :



Gambar 2. Feature Importance

Dari hasil diagram diatas dapat disimpulkan bahwa setelah dimodelkan dengan menggunakan algoritma K-means terbentuk tiga cluster yaitu cluster satu merupakan hasil pengolahan data dari program studi Sistem Informasi dengan persentase sebesar 23,0 %, pada cluster dua merupakan hasil pengolahan data dari program studi Manajemen Informatika dengan persentase sebesar 49,7 % dan pada cluster tiga merupakan hasil pengolahan data dari program studi Teknik Informatika dengan persentase sebesar 27,3 %. Untuk menentukan kualitas dari suatu perhitungan cluster menggunakan metode perhitungan Silhoutte, berikut hasil pengukuran kualitas cluster dengen metode Silhoutte :



Gambar 3. hasil perhitungan Silhoutte

Nilai Silhoutte adalah ukuran digunakan untuk mengukur kualitas cluster yang dihasilkan pada proses modeling, nilai Silhoutte berkisar dari −1 hingga +1, di mana terdapat 3 kategori pada perhitungan Silhoutte, jika nilai berkisar antara -1 sampai dengan 0,25 maka dikategorikan poor/rendah, jika nilai berkisar antara 0,25 sampai dengan 0,5 maka dikategorikan Fair/sedang dan jika nilai berkisar antara 0,5 sampai dengan 1 maka dikategorikan good/bagus, pada hasil penghitungan cluster quality dengan motede silhoutte diatas didapatkan nilai sebesar 0,6 sehingga dapat dikategorikan ke dalam good cluster.

HASIL DAN PEMBAHASAN

Pada proses modeling dengan menggunakan algoritma K-means menghasilkan tiga cluster, yaitu cluster 1, cluster 2 dan cluster 3. Dari total 253.886 data yang telah diolah dengan menggunakan SPSS modeler menghasilkan prosentase untuk masing masing cluster, untuk

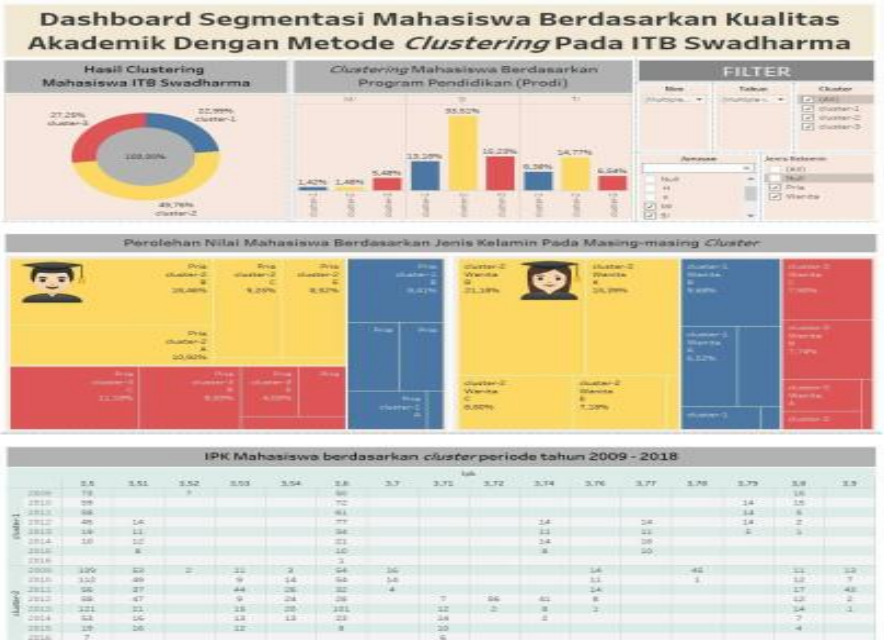
prosestase cluster 1 sebesar 22, 97 %, kemudian prosentase cluster 2 sebesar 49,74 dan prosebtase cluster 3 sebesar 27,29 % dari total keseluruhan data. Berikut hasil output data yang diperoleh setelah dilakukan modeling dengan algoritma K-means:

Tabel 1. Output data hasil modeling menggunakan algoritma K-Means

nim	nama	kelas	ipk	semester	jenis_kela	tahun	semester	mata_kul	sks	absen	tugas	uts	uas_upm	nilai_akhi	nilai	jur	$KM-K-Means
NULL	10110008		1 NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	SI	cluster-1
100001	NULL		2 2.2		7 Wanita	2003	30	Pengenali		2 0.0	0.0	0.0	0.0	60.0	C	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2003	7	Kerja Prak		2 0.0	0.0	0.0	0.0	0.0	E	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2002	5	Sistem Ba		2 0.0	0.0	0.0	0.0	64.0	C	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2004	50	Perancang		4 0.0	0.0	0.0	0.0	70.0	B	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2002	6	Sistem Op		4 0.0	0.0	0.0	0.0	0.0	E	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2002	6	Komputer		4 0.0	0.0	0.0	0.0	61.0	C	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2002	5	Perancang		4 0.0	0.0	0.0	0.0	56.0	C	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2001	40	Paket Proj		4 0.0	0.0	0.0	0.0	60.0	C	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2004	60	Sistem Op		4 0.0	0.0	0.0	0.0	75.0	B	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2001	3	Pendidika		2 0.0	0.0	0.0	0.0	58.0	C	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2003	7	Komputer		4 0.0	0.0	0.0	0.0	61.0	C	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2000	2	Aljabar Li		2 0.0	0.0	0.0	0.0	70.0	B	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2003	40	Sistem Ba		2 0.0	0.0	0.0	0.0	60.0	C	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2001	3	Statistik D		4 0.0	0.0	0.0	0.0	80.0	A	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2000	1	Pendidika		2 0.0	0.0	0.0	0.0	80.0	A	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2000	1	Paket Proj		4 0.0	0.0	0.0	0.0	0.0	E	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2000	2	Pendidika		2 0.0	0.0	0.0	0.0	70.0	B	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2002	5	Sistem Be		2 0.0	0.0	0.0	0.0	71.0	B	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2000	1	Pemrogra		4 0.0	0.0	0.0	0.0	68.0	B	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2004	60	Teknik Ris		4 0.0	0.0	0.0	0.0	65.0	C	MI	cluster-2
100001	NULL		2 2.2		7 Wanita	2002	5	Komputer		4 0.0	0.0	0.0	0.0	70.0	B	MI	cluster-2

tb_dp_nilai_cluster (2)

Agar data segmentasi Mahasiswa berdasarkan kualitas akademik yang telah dimodelkan tersebut mudah untuk dipahami, perlu dilakukan visualisasi data ke dalam sebuah dashboard diagram atau tampilan grafik, dengan memanfaatkan tableau aplication berdasarkan kebutuhan informasi yang akan ditampilkan.



Gambar 4. Dashboard hasil segmentasi mahasiswa

Pada dashboard diatas ditampilkan keseluruhan informasi yang didapat dari hasil modeling data seperti hasil clustering mahasiswa, clustering mahasiswa berdasarkan program studi, persentase nilai mahasiswa berdasarkan cluster, perolehan nilai mahasiswa berdasarkan jenis kelamin pada masing-masing cluster dan informasi IPK Mahasiswa berdasarkan cluster pada periode tahun 2009 -2018.

Dari hasil modeling data denngan metode clustering dengan algoritma K-Means didapatkan informasi sebagai berikut:

1. Terdapat tiga cluster dari hasil modeling data mahasiswa dari tahun 1992-2018, yaitu cluster 1 sebanyak 22,99% , cluster 2 sebanyak 49,76% dan cluster 3 sebanyaj 27,25%, cluster 2 mendominasi jumlah cluster mahasiswa pada ITB Swadharma periode tahun 1992 – 2018.
2. Jumlah persentase cluster pada program studi Sistem Informasi (SI) hampir separuh dari total persentase cluster pada semua program studi yaitu sebesar 47,92%, dengan persentase tertinggi pada cluster 2 yaitu 33,51%.
3. Jumlah persentase cluster pada program studi Teknik Informasi (TI) sebesar 29,69%, dengan persentase tertinggi pada cluster 2 yaitu sebesar 14,77%.
4. Jumlah persentase cluster pada program studi Manajemen Informatika (MI) sebesar 8,38%, dengan persentase tertinggi pada cluster 3 yaitu sebesar 5,48%.
5. Jumlah persentase pada cluster 2 menjadi yang tertinggi pada program studi Sistem Informasi dan Teknik Informasi.
6. Pada persentase jumlah perolehan nilai mahasiswa dengan jenis kelamin pria, perolehan nilai A tertinggi pada cluster 2 yaitu sebesar 10,92%, perolehan nilai B tertinggi pada cluster 2 yaitu sebesar 18,46%, perolehan nilai C tertinggi pada cluster 3 sebesar 11,18%, perolehan nilai D tertinggi pada cluster 3 yaitu sebesar 1,87% dan perolehan nilai E tertinggi pada cluster 2 yaitu sebesar 8,52%.
7. Pada persentase jumlah perolehan nilai mahasiswa dengan jenis kelamin wanita, cluster 2 mendominasi semua jumlah persentase nilai untuk perolehan nilai A sebesar 15,39 %, perolehan nilai B sebesar 21,18 %, perolehan nilai C sebesar 8,80 %, perolehan nilai D sebesar 0,96% dan perolehan nilai E sebesar 8,52%.
8. Untuk perolehan IPK pada setiap cluster didominasi oleh cluster 2 dengan sebaran datanya yang mereata hampir disemua nilai IPK dan setiap tahun.
9. Cluster 2 mendominasi total persentase dibandingkan dengan cluster lain, baik secara jumlah pada setiap program studi, pada perolehan nilai akademik dan pada nilaik IPK yang telah dikelompokan antara 3,5 – 4,00 pada periode tahun 2009 – 2018.

KESIMPULAN DAN REKOMENDASI

Berdasarkan hasil penelitian dari penerapan data mining pada segmentasi mahasiswa berdasarkan kualitas akademik dengan menggunakan algoritma K-Means clustering pada ITB Swadharma Jakarta, dapat diambil kesimpulan bahwa :

1. Data yang diolah adalah data yang didapat sejak 1992 hingga 2018 dengan data jumlah mahasiswa 253.886. Data yang sudah diolah menghasilkan 3 cluster, dengan persentase untuk cluster 1 sebesar 22, 97 %, kemudian persentase untuk cluster 2 sebesar 49,74%, dan persentase cluster 3 sebesar 27,29 % dari total keseluruhan data.
2. Signifikan faktor yang mempengaruhi hasil segmentasi mahasiswa antara lain : Nim, nama, kelas, IPK, semester, tahun, SKS.
3. Hasil summary model yang terbentuk pada proses segmentasi mahasiswa dengan metode cluster menghasilkan cluster quality sebesar 0,6 dengan metode perhitungan nilai Silhoutte, hasil perhitungan tersebut termasuk pada kateogori good cluster karena berkisar antara 0,5 sampai dengan 1.

REFERENSI

Asroni, & Adrian, R. (2015). Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang. *Jurnal Ilmiah Semesta Teknika*, *18*(1), 76–82.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. USA: SPSS Inc.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (Third). Waltham, MA: Morgan Kaufmann.

Poerwanto, B., & Fa'rifah, R. . (2016). Analisis Cluster K-Means Dalam Pengelompokan Kemampuan Mahasiswa. *Indonesian Journal Of Fundamental Sciences*, *2*(2), 92–96.

Widyawati, N. (2010). *Perbandingan Clustering Based On Frequent Word Sequence (FWS) dan K-Means Untuk Pengelompokan Dokumen Berbahasa Indonesia*. Bandung.