Institut Riset dan Publikasi Indonesia (IRPI)
SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat
Homepage: https://journal.irpi.or.id/index.php/sentimas
Prosiding Seminar Nasional/ ISSN(E): 2830-3083
25 Agustus 2022/ Pages: 154-161

K-Means and Decision Tree Algorithm for Prediction of Postgraduate Students Admission in University of Indonesia

Algoritma K-Means dan Decision Tree untuk Prediksi Penerimaan Calon Mahasiswa Pascasarjana pada Universitas Indonesia

Bima Ardiansyah[1], Irwan Daulay[2], Ricky Hutagaol[3], Rahmaddeni[4*]

[1,2,3,4] Program Studi Teknik Informatika, STMIK AMIK Riau, Indonesia

E-mail: [1]2110031802141@sar.ac.id,[2]2110031802140@sar.ac.id,
[3]2110031802145@sar.ac.id, [4]rahmaddeni@sar.ac.id

Corresponding Author: Rahmaddeni

Abstract

Postgraduate is the level of education after taking a bachelor's or bachelor's degree either at the master or doctoral level. Indonesia is one of the countries that has a Postgraduate program. There is quite a lot of demand for Master degree in Indonesia because it is to improve skills in certain fields. At this time, the determination of passing the exam to continue to the postgraduate level is still using a selection with TOEFL score parameters, university test scores and manual selection, so that the graduation results have not shown such a strict and perfect selection, especially at favorite universities in Indonesia. Therefore, with this research, a contribution or breakthrough is made in data mining applications using the K-means Algorithm and Decision Tree so that it is more automatic and later will be added several parameters such as research experience and GPA scores so that before prospective undergraduate students want to continue their education at the Masters or Doctoral level. Students must have some researches related to the field to be taken and consideration of GPA scores. It will produce Masters and Doctoral candidates who can create jobs and build new innovations in their environment.

Keywords: Data Mining, Decision Tree, K-Means, New Student Admissions

Abstrak

Pascasarjana adalah tingkatan Pendidikan setelah menempuh sarjana atau S1 baik pada tingkat Magister atau Doctoral. Indonesia adalah salah satu negara yang memiliki program Pascasarjana. Peminatan untuk melanjutkan S2 di Indonesia cukup banyak karena untuk meningkatkan *skill* terhadap bidang tertentu. Pada saat ini penentuan kelulusan ujian untuk melanjut ke tingkatan ke Pascasarjana masih menggunakan seleksi dengan parameter nilai TOEFL, nilai ujian dari universitas dan seleksi manual, sehingga hasil kelulusan belum menunjukkan seleksi yang begitu ketat dan sempurna terutama di universitas favorit di Indonesia. Dengan penelitian ini dibuat suatu kontribusi atau trobosan pada data mining menggunakan algoritma K-means dan Decision Tree sehingga lebih otomatis dan nantinya akan di tambahkan beberapa parameter seperti pengalaman penelitian dan Nilai IPK sebelum calon mahasiswa S1 ingin melanjutkan Pendidikan tingkat Magister atau Doctoral. Mahasiswa harus memiliki beberapa penelitian terkait bidang yang akan diambil dan pertimbangan terhadap Nilai IPK. Sehingga akan menghasilkan calon Magister dan Doctoral yang dapat menciptakan lapangan pekerjaan dan membangun inovasi baru di lingkungannya.

Kata Kunci: *Data Mining, Decision Tree, K-Means, Penerimaan Calon Mahasiswa Baru*

1. PENDAHULUAN

Seleksi masuk Pascasarjana pada Universitas di Indonesia merupakan sebuah proses seleksi untuk menerima calon mahasiswa yang telah selesai menempuh tingakat sarjana strata 1. Seleksi masuk perguruan tinggi tingkat Pascasarjana dilakukan baik di swasta maupun nasional untuk mendapatkan calon mahasiswa yang memiliki kualitas yang baik. Data statistik Indonesia tahun 2021 menunjukkan bahwa jumlah perguran

tinggi yang tersebar di seluruh Indonesia sebanyak 3.115 perguruan tinggi yang sebagian besar didominasi oleh perguruan tinggi swasta sekitar 90%. Jumlah perguruan tinggi swasta yaitu 2,990 dan perguruan tinggi negeri 125 universitas[1].

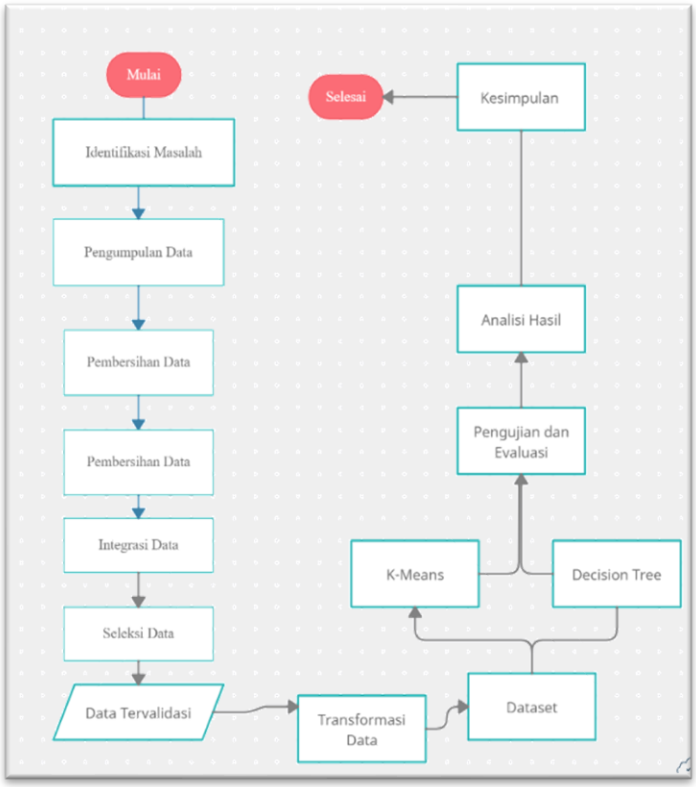
Universitas nasional yang melakukan proses seleksi cukup ketat adalah Universitas Indonesia. Universitas Indonesia merupakan salah satu universitas yang banyak diminati oleh calon mahasiswa baru di Indonesia. Pada proses seleksi dilakukan dengan beberapa tahapan yang di gunakan sebagai dasar untuk menentukan mahasiswa lulus atau tidak lulus.

Berdasarkan penelitian terdahulu yang dilakukan oleh Yobioktabera pada 2021 yang berkaitan dengan prediksi calon mahasiswa baru Fakultas Kedokteran menggunakan algortima *K-nearest neighbors (KNN)* dapat digunakan dalam memprediksi penerimaan calon mahsiswa kedokteran dengan mempertimbangkan beberapa faktor Variable Independent dan Variabel Dependent. Diperoleh hasil RMSE terbaik dengan data training berjumlah 70 data dan diperoleh akurasi sebesar 76,1 %[2]. Sedangkan berdasarkan penelitian terdahulu dari Yunita pada 2018 berkaitan dengan penerapan data mining dalam menggunakan algoritma K-means clustering pada penerimaan mahasiswa baru dengan menggunakan beberapa atribut seperti Asal Sekolah, Program Studi dan Nilai UAS. Hasil penelitian ini digunakan untuk dasar pengambilan keputusan dalam menentukan strategi mempromosikan masing-masing program studi sehingga membentuk 3 cluster berdasarkan program studi yang banyak diminati di Universitas Islam Indragiri[3].

Berdasarkan latar belakang penelitian sebelumnya, maka penelitian ini menggunakan metode K-Means dan Decision Tree dengan membandingkan akurasi dari setiap metode. Data yang digunakan dalam penelitian ini adalah data yang bersifat publik. Atribut yang digunakan adalah binominal, label, serta polynominal untuk prediksi penerimaan calon mahasiswa baru Pascasarjana di Indonesia.

2. METODE PENELITIAN

Tahapan penelitian dilakukan dengan melakukan identifikasi pada masalah, selanjutnya melakukan pengumpulan data, pembersihan data, integrasi data, dan seleksi data sehingga diperoleh data yang valid. Metode penelitian dapat dilihat pada gambar 1.



Gambar 1. Metode Penelitian

Dataset akan diolah dengan menggunakan Algortima K-Means dan Decision Tree yang menghasilkan prediksi yang valid. Selanjutnya akan diuji, dievaluasi dan dianalisis, berikutnya hasil analisis tersebut akan digunakan untuk menarik kesimpulan dari penelitian ini.

2.1 Sumber Data

Sumber data yang digunakan dalam penelitian ini bersumber dari *kaggle.com* dengan jumlah data sebesar 500 data. Variabel yang digunakan dari sumber data yang diperoleh yakni: Nilai TOEFL, Nilai Ujian, Nilai CGPA, dan Jumlah Penelitian.

2.2 Data Mining

Data Mining adalah metode yang digunakan untuk pengolahan data bersekala besar. Oleh sebab itu maka data mining menjadi suatu peranan yang sangat penting dalam proses pengolahan data. Data mining dapat diartikan sebagai serangkaian kegiatan untuk memperoleh data dalam jumlah yang besar dan dapat disimpan di dalam *database*, *data warehouse* atau penyimpanan informasi lainnya. Beberapa teknik data mining adalah data analisis, signal processing, neural network dan pengenalan pola[4].

2.3 Clustering

Algoritma Clustering dalam data mining adalah teknik pengelompokan elemen data yang memungkinkan untuk membangun/menghubungkan objek data yang serupa[5].

2.4 K-Means

K-Means Clustering adalah algoritma pembelajaran tanpa pengawasan yang digunakan untuk memecahkan masalah pengelompokan dalam pembelajaran mesin atau ilmu data[6][7]. Clustering merupakan algoritma yang mengelompokkan kumpulan data yang tidak berlabel ke dalam klaster yang berbeda jumlah cluster yang telah ditentukan sebelumnya yang perlu dibuat dalam proses, seolah-olah *k=2*, akan ada dua cluster, dan untuk *k=3*, akan ada tiga cluster, dan seterusnya. Ini memungkinkan untuk mengelompokkan data ke dalam grup yang berbeda dan cara yang mudah untuk menemukan kategori grup dalam kumpulan data yang tidak berlabel sendiri tanpa perlu pelatihan apa pun. Algoritma K-Means merupakan algoritma yang berbasis centroid, di mana setiap cluster dikaitkan dengan centroid. Tujuan utama dari algoritma ini adalah untuk meminimalkan jumlah jarak antara titik data dan cluster yang sesuai. Algoritma K-Means dikalkulasikan menggunakan beberapa langkah[8].

1. Menetukan nilai *k* dengan pertimbangan teoritis dan konseptual untuk memperoleh berapa banyak kluster dalam data penelitian tersebut.
2. Menerapkan *k* centroid sebagai titik pusat cluster awal secara acak dari objek yang tersedia. Selanjutnya menghitung centroid cluster ke-i yang ditunjukkan oleh rumus (1).

$$\boldsymbol{v} = \frac{\sum_{i}^{n}=1^{xi}}{\boldsymbol{n}} \qquad \textbf{; i=1, 2, 3, …n} \tag{1}$$

Keterangan:

v	*: centroid pada cluster*
xi	*: objek ke-i*
n	*: banyaknya objek/jumlah objek dari anggota cluster*

3. Menghitung jarak objek pada setiap centroid dari masing-masing cluster dengan menggunakan *Euclidian Distance*.
4. Menempatkan objek ke dalam centroid yang memiliki jarak terdekat. Pada bagian pengalokasian objek ke setiap cluster, iterasi dilakukan dengan menggunakan cara hard K-Means.
5. Melakukan iterasi untuk posisi centroid yang dilakukan menggunakan persamaan (1).
6. Mengulangi langkah 3, apabila pada posisi centroid baru menghasilkan perbedaan. Untuk pengecekan metode konvergensi berfungsi untuk membandingkan matriks *assignment group* pada matriks *group assignment* sebelumnya dan yang sedang berjalan. Apabila menghasilkan hasil yang sama maka algoritma K-Means sudah konvergen, apabila hasilnya berbeda maka belum konvergen.

Pada metode K-means pengolahan data dalam perhitungan adalah data numerik. Data diluar numerik dapat di aplikasikan tetapi harus dilakukan pengkodean terlebih dahulu yang nantinya dapat mempermudah menghitung kesamaan karakter yang ada pada objek. Jenis yang dihitung di setiap objek adalah kedekatan jaraknya terhadpa karakter yang dimiliki dengan pusat cluster yang sudah ada sebelumnya. Jika jumlah cluster telah ditentukan, maka dipilihlah 3 objek yang memiliki kedekatan terhadap jarak pada semua cluster[6].

2.5 Decision Tree

Decision Tree merupakan metode pengelompokan yang paling umum digunakan dalam data mining, karena sangat mudah dimengerti oleh manusia. Decision Tree adalah model prediksi yang menggunakan struktur hirarki dan pohon. Pada konsep pohon keputusan di defenisikan sebagai pengganti sebuah data

menjadi sebuah aturan keputusan untuk mengubah data menjadi Decision Tree dan aturan-aturan keputusan. Kegunaan dalam pemilihan Decision Tree sebagai metode data mining adalah untuk membagi sebuah proses pengambilan keputusan yang sebelumnya sangat kompleks menjadi lebih sederhana, sehingga lebih cepat dalam pengambilan dalam suatu kesimpulan atau solusi dalam sebuah permasalahan[9].

$$d(x,y) = \| x - y \| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \; ; \; i = 1,2,3,....,n \quad (2)$$

Keterangan:

xi *: objek x ke-i*
yi *: daya y ke-i*
n *: banyaknya objek*

3. HASIL DAN PEMBAHASAN

Sumber data dalam penelitian ini di ambil dari situs *kaggle.com* (Graduate Admission Prediction) tahun 2022 dengan rincian yang ditunjukkan oleh gambar 1.

Tabel 1. Rincian data dan parameter

Data	Variabel	
	Variabel Independent	Variable Dependent
500 pendaftar/calon mahasiswa Pascasarjana *Data Training*, dikutip dari sisa *data training* sebanyak 30 data.	Nilai Ujian : Skor pada saat Test dari Universitas Indonesia Nilai TOEFL : Hasil penilaian TOEFL Pengalaman Penelitian: Pengalam penilitian yang dimiliki oleh calon mahasiswa. CGPA	Lulus diartikan dengan nilai 1 Tidak Lulus diartikan dengan nilai 0

3.1 Pengujian dengan algoritma K-Means

Kinerja algoritma K-means yang dilakukan pengujian dengan menggunakan *tool RapidMiner* dengan atribut yang digunakan yaitu GRE Score, TOEFL Score, CGPA, Research, dan Graduated. Hasil pengolahan dengan RapidMiner ditunjukkan pada tabel 2.

Tabel 2. Hasil pengolahan dengan RapidMiner

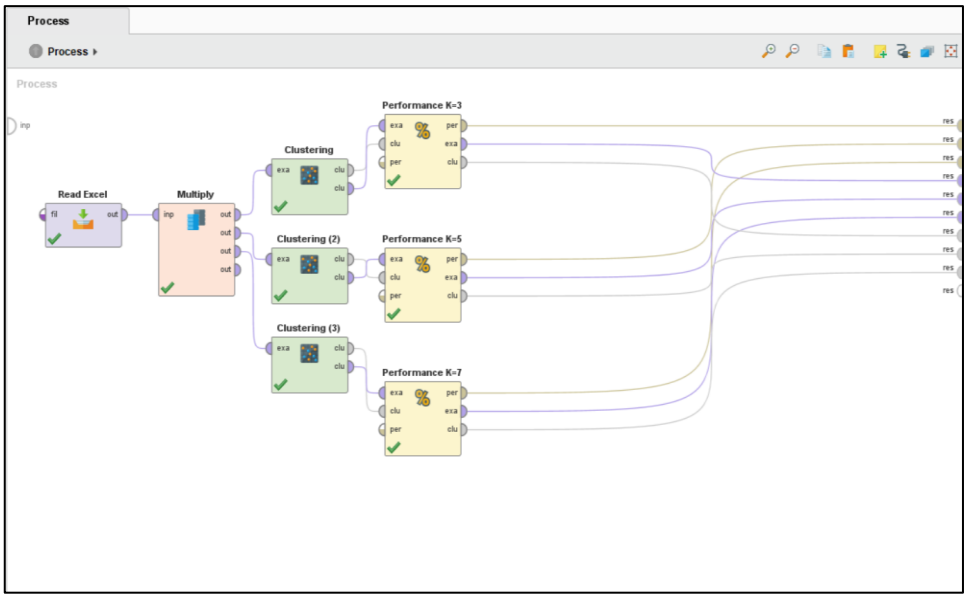
No	GRE Score	TOEFL Score	CGPA	Research	Graduated
1	337	118	9,65	1	Lulus
2	324	107	8,87	1	Lulus
3	316	104	8	1	Lulus
4	322	110	8,67	1	Lulus
5	314	103	8,21	0	Lulus
6	330	115	9,34	1	Lulus
7	321	109	8,2	1	Lulus
8	308	101	7,9	0	Lulus
9	302	102	8	0	Lulus
10	323	108	8,6	0	Tidak Lulus
11	325	106	8,4	1	Lulus
12	327	111	9	1	Lulus
13	328	112	9,1	1	Lulus
14	307	109	8	1	Lulus
15	311	104	8,2	1	Lulus
16	314	105	8,3	0	Lulus
17	317	107	8,7	0	Lulus
18	319	106	8	1	Lulus
19	318	110	8,8	0	Lulus
20	303	102	8,5	0	Lulus
21	312	107	7,9	1	Lulus
22	325	114	8,4	0	Lulus
23	328	116	9,5	1	Lulus
24	334	119	9,7	1	Lulus
25	336	119	9,8	1	Lulus
26	340	120	9,6	1	Lulus

No	GRE Score	TOEFL Score	CGPA	Research	Graduated
27	322	109	8,8	0	Lulus
28	298	98	7,5	1	Tidak Lulus
29	295	93	7,2	0	Tidak Lulus
30	310	99	7,3	0	Lulus

Pengujian menggunakan metode K-Means yang dilakukan memiliki performance dengan *k=5* serta hasil *Davies Bouldin* = -0.970. Pengujian dilakukan dalam 2 tahapan.

1. Performance *k*=5 → 0.871
2. Performance *k*=7 → 0.978

Dapat disimpulkan bahwa Davies Bouldin pada performance *k*=5 memiliki angka yang paling kecil yaitu 0.871, dimana prinsip dasar dari clustering *Davies Bouldin* adalah jika angka yang dihasilkan lebih kecil maka akurasi semakin baik. Perhitungan performance dengan *k*=5 memiliki tingkat akurasi yang lebih baik dan dapat digunakan untuk mengitung prediksi penerimaan calon mahasiswa baru Pascasarjana di Universitas Indonesia. Model prediksi dan performance vector algoritma K-Means dapat dilihat pada gambar 2, 3 dan 4.



Gambar 2. Model Prediksi pada Tool RapidMiner

Gambar 3. Performance Vector (Performance *k*=5)

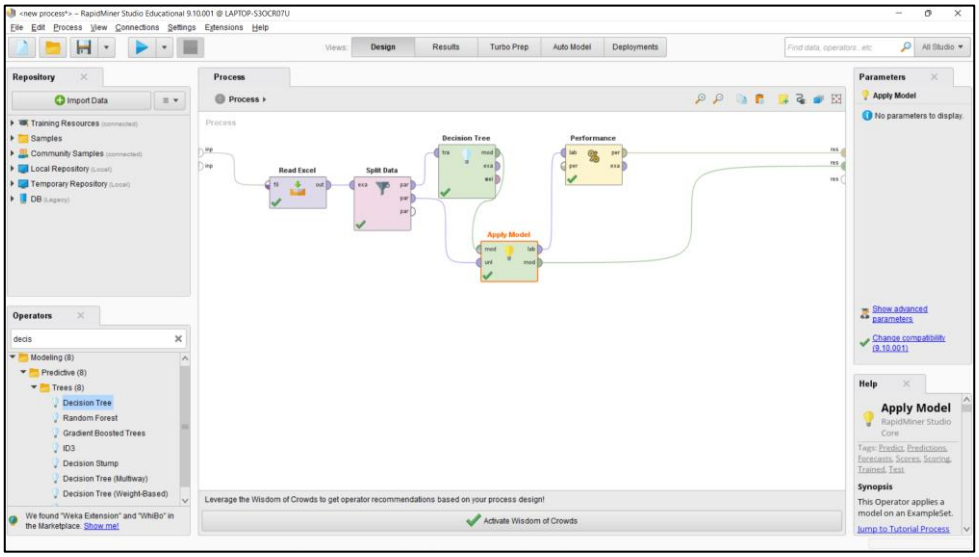
Gambar 4. Performance Vector (Performance *k*=7)

3.2 Pengujian dengan algoritma Decision Tree

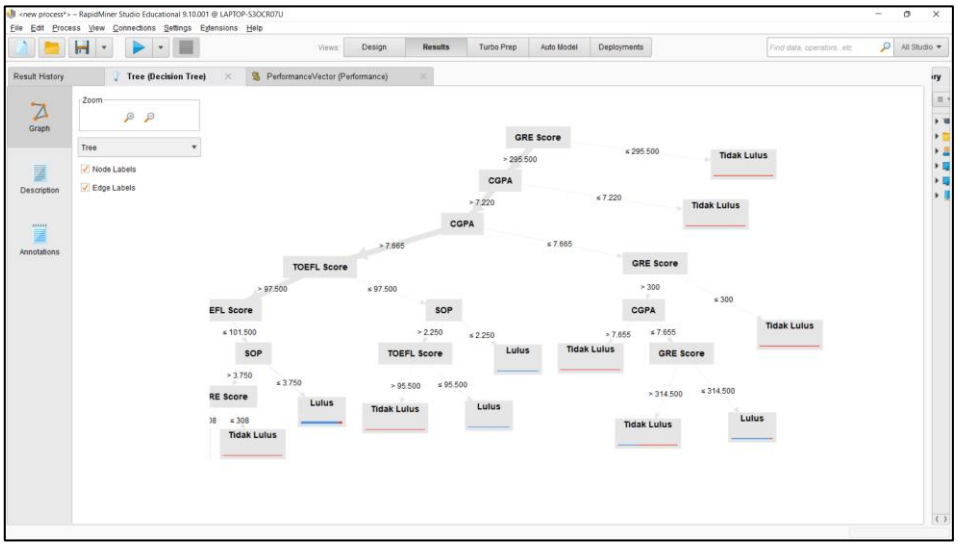
Untuk mengetahui kinerja algoritma Decision Tree dilakukan dengan menggunakan *tool* RapidMiner. Susunan diagram Decision Tree pada penelitian ini menggunakan Split data, Decision tree, Apply model, dan Performance. Dari pengujian dengan metode Decision Tree untuk score GRE > 295, CGPA > 7.2, score TOEFL > 101 dapat disimpulkan bahwa calon mahasiswa Pascasarjana Lulus.

Jika score GRE <295, CGA < 7.2, score TOEF < 101 dapat disimpulkan bahwa calon mahasiswa Pascasarjana Tidak Lulus.

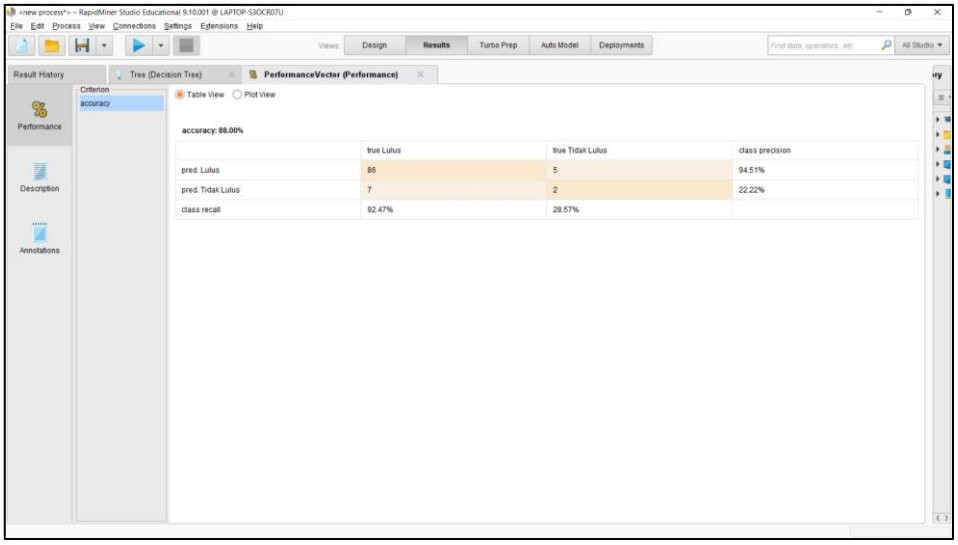
Hasil metode Decision Tree diperoleh akurasi prediksi tingkat kelulusan sebesar 88%, sehingga dapat disimpulkan bahwa metode Decisiom Tree dapat digunakan sebagai metode dalam menentukan penerimaan mahasiswa Pascasarjana. Seluruh hasil percobaan dengan metode Decision Tree dapat dilihat pada gambar 5, 6 dan 7.



Gambar 5. Model Prediksi Metode Decision Tree



Gambar 6.. Bentuk Pohon Keputusan Algoritma Decision Tree



Gambar 7. Performance Vector Decision Tree

4. KESIMPULAN

Dari pembahasan dapat disimpulkan bahwa metode K-means dan Decision Tree dapat digunakan untuk memprediksi penerimaan calon mahasiswa Pascasarjana pada Universitas Indonesia. Algoritma K-Means memiliki tingkat akurasi yaitu 87% dan Decision Tree 88 %. Dari hasil perbandingan terlihat bahwa algoritma Decision Tree memiliki nilai yang paling baik dibandingkan dengan algoritma K-Means dalam hal menetukan kelulusan penerimaan mahasiswa Pascasarjana pada Universitas Indonesia.

REFERENSI

[1] D. Z. Abidin, S. Nurmaini, and R. F. Malik, "Penerapan Metode K-Nearest Neighbor dalam Memprediksi Masa Studi Mahasiswa (Studi Kasus : Mahasiswa STIKOM Dinamika Bangsa)," *Pros. Annu. Res. Semin.*, vol. 3, no. 1, pp. 133–138, 2017.

[2] A. Yobioktabera, "Penerapan Data Mining Untuk Memprediksi Penerimaan Calon Mahasiswa Baru Fakultas Kedokteran Menggunakan Algoritma K-NN," *JTET (Jurnal Tek. Elektro Ter.*, pp. 16–19, 2021.

[3] F. Yunita, "Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru (STUDI KASUS : UNIVERSITAS ISLAM INDRAGIRI)," *Sistemasi*, vol. 7, no. 3, pp. 238–249, 2018.

[4] L. R. Angga Ginanjar Mabrur, "Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit," *J. Komput. dan Inform.*, vol. 1, no. 1, pp. 53–57, 2012.

[5] V. Hananto, "Analisis Penentuan Metode Data Mining Untuk Prediksi Kelulusan Mahasiswa Sebagai

Penunjang Angka Efisiensi Edukasi," *J. Ilm. SCROLL*, vol. 5, no. 1, pp. 1–11, 2017, [Online]. Available: http://repository.dinamika.ac.id/id/eprint/2122/4/Panalisis_Penentuan_Metode_Data_Mining_Utk_Prediksi_Kelulusan_Mahasiswa.pdf.

[6] A. Trimanto, F. Faqih, I. M. Irfani, and S. Timur, "Penerapan Data Mining Untuk Evaluasi Status Kelulusan Mahasiswa Fakultas Teknologi Pertanian Tahun 2015 Menggunakan Algoritma Naïve Bayes Classifier," *J. Ilm. Ilmu Komput.*, 2015.

[7] M. S. Mustafa, M. R. Ramadhan, and A. P. Thenata, "Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Creat. Inf. Technol. J.*, vol. 4, no. 2, p. 151, 2018, doi: 10.24076/citec.2017v4i2.106.

[8] A. Syahrin, "Implementasi algoritma k-means untuk klasterisasi mahasiswa berdasarkan prediksi waktu kelulusan skripsi," *UPN "Veteran" Jatim*, vol. 1–23, 2013.

[9] M. N. Yatimah, "Implementasi Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa STIMIK ESQ Menggunakan Decision Tree C4.5," *JUMANJI (Jurnal Masy. Inform. Unjani)*, vol. 5, no. 2, p. 89, 2021, doi: 10.26874/jumanji.v5i2.95.