

PENERAPAN ALGORITMA C4.5 UNTUK MEMPREDIKSI KELULUSAN MAHASISWA DI INSTITUT TEKNOLOGI TELKOM SURABAYA

Edriand Imens Raygrandi^{*1}, Rokhmatul Insani², Mochammad Nizar Palefi³

^{1,2,3}Institut Teknologi Telkom Surabaya

Email: ¹eraygrandi@student.ittelkom-sby.ac.id, ²insani@ittelkom-sby.ac.id, ³nizar@ittelkom-sby.ac.id

^{*}Penulis Korespondensi

(Naskah masuk: dd mmm yyyy, diterima untuk diterbitkan: dd mmm yyyy)

Abstrak

Akreditasi pada sebuah kampus memiliki banyak faktor perhitungan dan salah satunya adalah ketepatan kelulusan pada angkatan tersebut. Namun, untuk lulus tepat waktu memiliki banyak faktor yang menjadi tolak ukur dalam hal tersebut mau dari faktor *internal* mahasiswanya ataupun faktor *eksternal* disekitarnya. Hal tersebut menjadikan lulus tepat waktu susah untuk diketahui. Oleh sebab itu, pada penelitian ini akan dilakukan prediksi dalam ketepatan waktu lulus angkatan 2020 di IT Telkom Surabaya. Dengan melakukan penerapan *Data Mining* diharapkan dapat membantu dalam memprediksi kelulusan mahasiswa. Algoritma *Decision Tree* C4.5 dipilih untuk membantu dalam melakukan pengolahan data yang sudah disiapkan dari angkatan 2018. Data *training* dan *testing* yang digunakan merupakan data dari angkatan 2018, data-data tersebut akan diolah menggunakan sistem Cross-Industry Standart Process for Data Mining (CRISP-DM) dengan bantuan aplikasi RapidMiner untuk mengolah data hingga menjadi data hasil. Hasil yang dihasilkan pada penelitian ini adalah persentase prediksi kelulusan pada setiap program studi angkatan 2020 dengan hasil sebanyak 594 mahasiswa dapat lulus tepat waktu dan sebanyak 14 mahasiswa diprediksi tidak lulus tepat waktu, untuk tingkat akurasi yang dihasilkan dapat terbilang tinggi yaitu sebesar 82,27%, tingkat presisi 75,17%, dan juga tingkat *recall* sebesar 70,83%. Peneliti juga melakukan perbandingan antar algoritma *Decision Tree* C4.5 dengan Naïve Bayes dan KNN, naïve bayes memiliki hasil akurasi sebesar 91,55%, tingkat presisi 96,67%, dan juga tingkat *recall* sebesar 74,17%. Untuk KNN memiliki hasil akurasi 86,09%, tingkat presisi 85,83%, dan juga tingkat *recall* sebesar 64,17%. Hal tersebut dapat memastikan bahwa algoritma *Decision Tree* C4.5 dapat melakukan prediksi kelulusan mahasiswa.

Kata kunci: *Data Mining, Algoritma Decision Tree C4.5, Prediksi Kelulusan Mahasiswa, RapidMiner, Cross-Industry Standart Process for Data Mining*

APPLICATION OF THE C4.5 ALGORITHM TO PREDICT THE GRADUATION OF UNDERGRADUATE STUDENTS AT TELKOM INSTITUTE OF TECHNOLOGY SURABAYA

Abstract

Accreditation on a campus has many calculation factors and one of them is the accuracy of graduation in that class. However, to graduate on time has many factors that become benchmarks in this case, both from the internal factors of the students or the external factors around them. This makes graduating on time difficult to know. Therefore, in this study predictions will be made on the timeliness of graduating class of 2020 at IT Telkom Surabaya. By implementing *Data Mining* it is hoped that it can help in predicting student graduation. The C4.5 *Decision Tree* algorithm was chosen to assist in processing data that has been prepared from the 2018 batch. The training and testing data used is data from the 2018 batch, these data will be processed using the Cross-Industry Standard Process for Data Mining (CRISP) system. -DM) with the help of the RapidMiner application to process data to become result data. The results produced in this study are the percentage of predicted graduation in each class of 2020 study program with the result 594 students are being able to graduate on time and 14 students aren't being able to graduate on time, the level of accuracy produced can be said high, it is 82,27%, precision 75,17%, and also recall 70,83%. The researcher also made a comparison between the *Decision Tree* C4.5 algorithms with Naïve Bayes and KNN, naïve Bayes has an accuracy of 91,55%, precision 96,67%, and also recall 74,17%. KNN has an accuracy of 86,09%, precision 85,83%, and also recall 64,17%. This can ensure that the *Decision Tree* C4.5 algorithms can predict student graduation.

Keywords: *Data Mining, Decision Tree Algorithm C4.5, Prediction of Student Graduation, RapidMiner, Cross-Industry Standart Process for Data Mining.*

1. PENDAHULUAN

Institut Teknologi Telkom Surabaya atau biasa dikenal dengan IT Telkom Surabaya adalah salah satu perguruan tinggi IT yang berada di kota Surabaya, IT Telkom Surabaya sendiri merupakan salah satu perguruan tinggi dibawah naungan Telkom Group khususnya bagian Yayasan Pendidikan Telkom. IT Telkom Surabaya hingga saat ini memiliki 11 program studi yang dibagi menjadi 2 fakultas yang berbeda, yaitu: Fakultas Teknologi Elektro dan Industri Cerdas dan Fakultas Teknologi Informasi dan Bisnis (SURABAYA, 2022). IT Telkom Surabaya memiliki akreditasi Baik sejak 2020. Aspek kelulusan pada sebuah universitas ataupun institusi merupakan salah satu aspek penting dalam peninjauan akreditasi (GAOL, et al., 2021). Jika pada sebuah kampus terdapat banyak mahasiswanya lulus dengan tenggat waktu yang melebihi waktu seharusnya maka itu akan berpengaruh dalam penilaian akreditasi kampus tersebut (PRASETYO, 2018)

Studi Sarjana sendiri memiliki tenggat waktu umumnya 4 tahun dengan 8 semester atau mahasiswa dapat melakukan lulus cepat dengan waktu 3,5 tahun dan 7 semester untuk mencapai kelulusan, jika mahasiswa lulus dalam tenggat waktu tersebut maka mereka akan dibilang lulus tepat waktu namun jika mereka melebihi waktu tersebut maka mereka akan dinyatakan lulus tidak tepat waktu (AGWIL, et al., 2020). Namun yang jadi permasalahan disini adalah waktu kelulusan setiap mahasiswa tidak menentu, dimana ini dapat dipengaruhi dengan banyak hal mulai dari aspek *internal* mahasiswa tersebut hingga aspek-aspek *eksternal*-nya seperti lingkungannya. Maka dari itu pentingnya untuk mengetahui waktu kelulusan seorang mahasiswa dan juga mengetahui tingkat hasil akurasi, meskipun hanya prediksi semata namun hal tersebut tentunya dapat membantu kampus khususnya IT Telkom Surabaya dalam mendidik mahasiswanya agar dapat lulus tepat waktu sesuai dengan tingkatan yang diambil yaitu sarjana dengan waktu 3,5 hingga 4 tahun.

Banyak cara untuk dapat memprediksi kelulusan mahasiswa dan salah satunya yang sering digunakan adalah *Data Mining*, karena *data mining* akan mencari informasi tersembunyi dari tumpukan data yang ada yang dirasa tidak memiliki fungsi lainnya (DARMAWAN, et al., 2018; NATASUWARNA, 2019). Pada *data mining* sendiri memiliki alur yang harus dipenuhi sebelum masuk kedalam pengolahan data dengan atau tidak dengan bantuan aplikasi pemrosesan, salah satu alur yang harus dipenuhi yaitu menentukan algoritma yang akan digunakan karena dengan menggunakan algoritma yang tepat pada data yang tepat maka akan menghasilkan informasi yang lebih akurat dan tepat. Algoritma yang dapat membantu dalam melakukan prediksi kelulusan adalah: Decision Tree ID3,

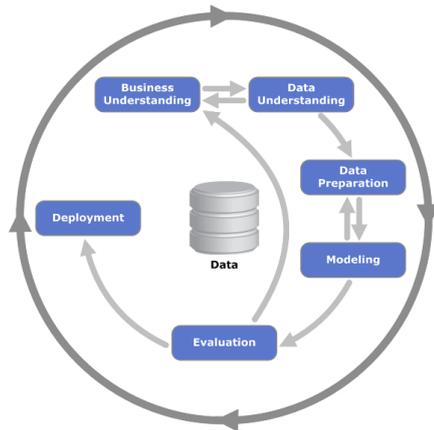
Decision Tree C4.5, Artificial Neural Networks (ANN), K-Nearest Neighbour (KNN), Naïve Bayes, dan beberapa algoritma lainnya (GAOL, et al., 2021). Setiap algoritma tersebut memiliki fungsi dan cara yang berbeda-beda untuk mengolah datanya demi menghasilkan informasi yang diinginkan, seperti contoh pada algoritma ID3 dan C4.5 haruslah memiliki atribut *label* untuk dijadikan sebagai hasil *output* dari algoritma tersebut dan tiap-tiap algoritma memiliki kententuannya masing-masing yang haruslah dipenuhi oleh pengguna.

Algoritma yang digunakan pada penelitian ini memiliki fokus pada algoritma Decision Tree C4.5 karena metode ini biasanya digunakan untuk dapat mengklasifikasikan data dengan atribut numerik. Sudah banyak penelitian lainnya yang serupa untuk mencari prediksi kelulusan dengan menggunakan algoritma yang sama, salah satunya adalah yang dilakukan oleh (GAOL, et al., 2021) dengan menggunakan metode yang sama memberikan akurasi sebesar 90,00% dengan tingkat presisi 91,38% dan tingkat recall 98,15%, adapun juga penelitian lainnya yaitu oleh (KAMIL & CHOLIL, 2020) dengan menggunakan perbandingan metode algoritma C4.5 dengan Naïve Bayes dan hasilnya algoritma C4.5 memiliki tingkat *confusion matrix* yang lebih tinggi dari naïve bayes, dan yang terakhir merupakan penelitian oleh (NURHASAN, et al., 2018) dengan membandingkan 3 metode yaitu C4.5, Naïve bayes, dan juga KNN dengan hasil akurasi tertinggi adalah Naïve bayes.

Dengan menggunakan aplikasi RapidMiner untuk melakukan perhitungan dapat menghasilkan pohon keputusan (*Decision Tree*) dan juga hasil penentuan label untuk perhitungan *confusion matrix*. Pada *confusion matrix* sendiri memiliki 3 hasil utama, yaitu: akurasi, *recall*, dan presisi yang tersedia dengan bantuan operator *performance* pada RapidMiner dan pada penelitian ini peneliti juga membandingkan 3 algoritma yang berbeda untuk melihat tingkat akurasi dari masing-masing algoritma yaitu Decision Tree C4.5, Naïve Bayes, dan *K-Nearest Neighbour* (KNN). Data yang akan digunakan pada penelitian ini adalah data yang berasal dari IT Telkom Surabaya, data angkatan 2018 akan menjadi data *training* dan data *testing* dimana data angkatan 2020 akan menjadi data perhitungan prediksi. Dari perhitungan data *training* dan data *testing* akan menghasilkan rule yang akan menjadi landasan dalam perhitungan prediksi kelulusan angkatan 2020 di IT Telkom Surabaya

2. METODE PENELITIAN

Setelah mengetahui latar belakang permasalahan yang terjadi pada penelitian ini, akan dilakukan dengan menggunakan metode Cross-Industry Standard Process for Data Mining.



Gambar 1. Cross-Industry Standard Process for Data Mining

2.1. Metode yang Digunakan

Cross-Industry Standard Process for Data Mining (CRISP-DM) adalah sebuah metode dimana menyediakan sebuah standart yang baku didalam data mining yang akan diterapkan pada strategi untuk pemecahan sebuah permasalahan umum pada sebuah bisnis ataupun pada sebuah unit penelitian dimana terdiri dari beberapa tahapan fase (SHEDRIKO & FIRDAUS, 2022), pada CRISP-DM sendiri memiliki beberapa tahapan yang harus diikuti, sebagai berikut (MAURITIUS & BINSAR, 2020; KADE S., 2022):

- a. **Bussiness Understanding**
Business Understanding merupakan sebuah fase dimana masalah bisnis akan didefinisikan secara sederhana dan tepat sasaran. Pada penelitian ini tahap *business understanding* yang dilakukan adalah untuk mempelajari tentang permasalahan yang terjadi didalam studi kasus dan mencari hipotesa untuk menyelesaikannya
- b. **Data Understanding**
Data Understanding merupakan sebuah proses yang dapat mempertemukan antara data yang dimiliki dengan data yang akan dibutuhkan untuk dilakukannya penelitian. Banyak faktor yang dapat mempengaruhi dalam penentuan algoritma dan tujuan dari *data mining* itu sendiri, salah satunya adalah faktor dari jenis data itu sendiri.
- c. **Data Preparation**
Data Preparation merupakan sebuah proses pengolahan data sebelum data digunakan dalam *data mining* itu sendiri. Pada fase ini, kejelasan dari kumpulan data sangatlah dibutuhkan. Ketelitian jugalah dibutuhkan dalam melakukan pembersihan data dan pengecekan ulang pada kumpulan data yang ada agar data benar-benar siap diolah pada proses *data mining*. Berikut adalah hal-hal yang umumnya dilakukan pada fase ini:
 - Melakukan pemeriksaan ulang pada keaslian data

- Mengelola data outlier
 - Memberlakukan *data missing* dan data inkonsistensi
- d. **Modelling**
Modelling pada dasarnya digunakan untuk membuat model secara prediktif ataupun deskriptif yang dibuat oleh bantuan sistem dimana hal tersebut membantu peneliti dalam menentukan beberapa hal, yaitu: alat, teknik, dan algoritma pada pengerjaan data mining. Jika terdapat keraguan akan hasil dari model maka peneliti dapat kembali ke fase persiapan data untuk memastikan ulang bahwa data sudah siap diolah.
 - e. **Evaluasi**
Evaluasi merupakan fase validasi dari model yang sudah dibuat pada fase sebelumnya yaitu modelling dengan dasar parameter yang relevan.
 - f. **Deployment**
Deployment merupakan fase yang paling penting didalam CRISP-DM. Pada fase ini setiap fase-fase sebelumnya haruslah sudah terpenuhi dahulu agar dapat mempertimbangkan apakah model tersebut dapat digunakan atau tidak karena jika model tidak cocok maka akan dilakukannya pergantian ke model lain yang lebih cocok pada pengolahan *data mining*.

2.2. Bahan dan Alat yang Digunakan

- a. **Bahan Penelitian**
Bahan penelitian yang digunakan pada penelitian ini adalah mahasiswa angkatan 2018 sebanyak 135 mahasiswa dan angkatan 2020 sebanyak 608 mahasiswa dimana data angkatan tersebut berasal dari Institut Teknologi Telkom Surabaya. Pada penelitian ini sendiri akan menggunakan beberapa atribut yang dirasa mempengaruhi tingkat kelulusan dari mahasiswa IT Telkom Surabaya sendiri yaitu: NIM sebagai ID, Presensi, IP dan SKS sem 1-4, IPK, Status Mahasiswa sebagai *label*.
- b. **Alat**
Alat yang digunakan pada penelitian ini adalah komputer dari user dengan spesifikasi prosesor I7-7700K, RAM 16GB, VGA Nvidia 1660S, dan hardisk internal sebesar 1TB. Untuk aplikasi yang digunakan dalam penelitian ini adalah RapidMiner versi 9.10.010 untuk melakukan pengolahan data dan juga Microsoft Excel 2016 untuk melakukan perhitungan rumus prediksi dan juga persiapan data.

3. HASIL DAN PEMBAHASAN

3.1. Analisa Alur Pengolahan Data

Mengikuti dari urutan pengerjaan dengan model CRISP-DM maka fase awalnya adalah *bussiness understanding* yang terjadi, yaitu dimana permasalahannya adalah ketidakpastian dalam waktu kelulusan maka perlu dilakukannya prediksi tingkat kelulusan agar terlihat mahasiswa mana yang sekiranya akan lulus terlambat.

Dilanjutkan dengan fase kedua yaitu *data understanding* dimana data yang akan diambil dari pihak IT Telkom Surabaya adalah data angkatan 2018 dan 2020 dengan total 22 atribut. Data-data yang sudah ditentukan tersebut hanya diperuntukkan untuk mahasiswa reguler bukan alih jenjang maupun kelas malam.

Fase selanjutnya adalah *data preparation* dimana atribut-atribut yang sudah diminta dari IT Telkom Surabaya akan diolah dengan men-sortir dan menyesuaikan dengan data angkatan 2020 yang akan menjadi data uji. Atribut numerik yang ada pada atribut akan sortir dengan menggunakan bantuan operator PCA dan juga dilakukan pemulihan *missing value* dengan menggunakan *Replace Missing Value* pada RapidMiner hingga mendapatkan atribut sebanyak 11 seperti yang sudah jelaskan pada sub bab 2.2.

Fase selanjutnya adalah fase *modelling* dimana fase ini peneliti akan menentukan algoritma dan aplikasi yang akan digunakan dalam melakukan penelitiannya dan pada penelitian ini peneliti menggunakan aplikasi RapidMiner dengan bantuan algoritma *Decision Tree C4.5*. Pada fase ini juga data akan mulai diolah dengan menggunakan beberapa *operator* pada RapidMiner untuk membantu dalam menemukan hasil akhirnya.

Fase keempat yaitu fase Evaluasi, fase ini akan mengevaluasi hasil dari fase *modelling* dimana jika terdapat kurang puasnya dengan hasil yang diberikan maka peneliti akan melakukan pengembalian fase ke fase pertama yaitu fase *bussiness understanding* karena peneliti harus menimbang ulang mulai dari tingkat permasalahan yang akan diselesaikan hingga balik ke fase *modelling* dan evaluasi. Pada fase ini jika hasil yang dikeluarkan sesuai dengan yang diharapkan oleh peneliti maka peneliti dapat melanjutkannya dengan melakukan perhitungan *confusion matrix* untuk melakukan uji tingkat akurasi, *recall*, dan presisi dari hasil yang sudah diberikan dengan bantuan *Performance* pada RapidMiner.

Fase terakhir adalah fase *deployment*, pada fase ini data yang sudah terbentuk akan dipresentasikan dengan bagus agar lebih mudah untuk dibaca dan dianalisa jika terdapat adanya kesalahan atau tidak sesuai dengan keinginan peneliti maka peneliti haruslah melakukan *bussiness understanding* lagi untuk memastikan bahwa apakah yang salah dari fase pertama ataukan hanya perlu mengubah *modelling*nya saja.

3.2. Bussiness Understanding

Pada tahap ini peneliti akan melakukan riset mengenai permasalahan yang akan diangkat ataupun diolah untuk diselesaikan. Peneliti memutuskan untuk menyelesaikan permasalahan mengenai ketidakpastian waktu dalam kelulusan mahasiswa pada Institut Teknologi Telkom Surabaya dan akan dilakukannya prediksi untuk mengetahui sekiranya waktu kelulusan mahasiswa IT Telkom Surabaya khususnya angkatan 2020 akan tepat waktu atau terlambat.

3.3. Data Understanding

Pada tahap ini peneliti akan mengambil data dari pihak IT Telkom Surabaya, dimana data-data yang akan diambil merupakan data angkatan 2018 yang sudah lulus maupun masih aktif di IT Telkom Surabaya seperti pada gambar 2 dan juga data angkatan 2020 yang akan menjadi data perhitungan prediksi kelulusan mahasiswa dengan menggunakan hasil RapidMiner dari data angkatan 2018.

Gambar 2. Format Data Awal

3.4. Data Preparation

Pada tahap ini peneliti akan melakukan pembersihan data, mulai dari adanya *missing value* akan digantikan dengan *value* lainnya menggunakan operator *Replace Missing Value* pada RapidMiner, dilanjutkan dengan men-sortir atribut yang akan digunakan dengan menggunakan operator *Principal Component Analysis*, dan juga penghapusan data alih jenjang pada angkatan 2020 untuk menyesuaikan data yang ada. Dengan hasil seperti pada gambar 3 setelah dilakukannya penyesuaian-penyesuaian pada data-data yang akan diolah kedepannya.

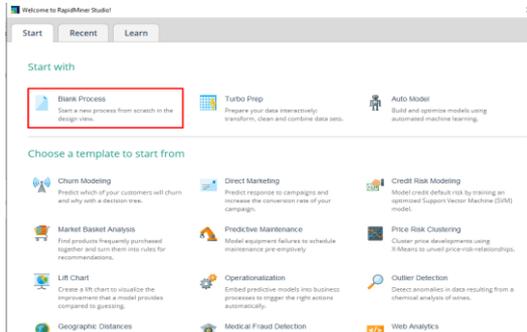
Gambar 3. Format Data Olah

3.5. Modelling

Pada tahap ini peneliti menentukan untuk menggunakan bantuan data mining untuk menyelesaikan masalah dengan pengolahan algoritma *Decision Tree C4.5* dengan aplikasi RapidMiner dalam mengolah data proses dan Microsoft Excel dalam mengolah pre-proses dan

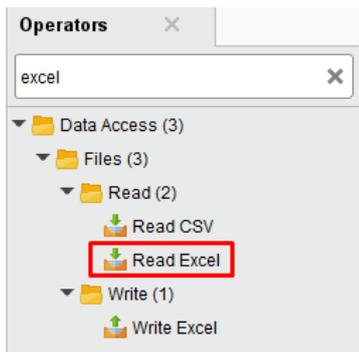
pasca proses dari RapidMiner. Berikut merupakan urutan singkat didalam penggunaan RapidMiner pada penelitian ini:

- 1) Buka RapidMiner lalu pilih *Blank Process* seperti pada gambar 4.



Gambar 4. Tampilan Awal RapidMiner

- 2) Pada kolom *operator* dibagian bawah kiri carilah *read excel* seperti pada gambar 5 atau sesuaikan dengan data yang sudah disiapkan, lalu *drag* ke dalam *canvas* yang tersedia atau tekan 2 kali pada *operator*-nya.



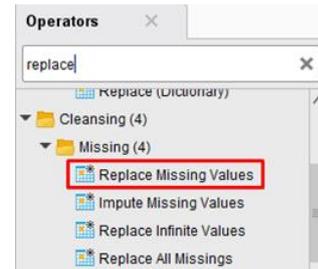
Gambar 5. Pemilihan Operator Pembaca Data

- 3) Pada *operator read excel* di *canvas* klik 2 kali lalu pilih data *training* yang kita siapkan dan sesuaikan dengan masing-masing *role* nya seperti pada gambar 6.

NIM	Presensi	IPK	IP sem 1	IP sem 2	IP sem 3
polynomial/ id	real	real	real	real	real
1201200xx	98.110	3.630	3.850	3.610	3.760
1201200xx	97.530	3.870	3.880	3.920	3.930
1201200xx	99.720	3.810	3.740	3.830	3.930
1201200xx	95.060	3.680	3.880	3.720	3.700
1201200xx	98.670	3.590	3.790	3.330	3.850
1201200xx	97.250	3.350	3.740	3.000	3.130
1201200xx	91.480	3.680	3.500	3.610	3.850
1201202xx	87.550	2.240	3.090	1.560	1.780
1201202xx	94.150	3.460	3.680	3.580	3.520
1201202xx	91.000	3.440	3.590	3.080	3.200
1103200xx	84.920	2.470	3.290	3.250	2.430
4444000xx	98.070	3.630	3.780	3.940	3.730

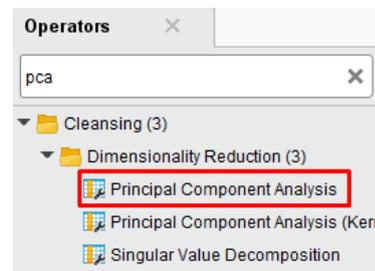
Gambar 6. Penentuan Tipe Data

- 4) Dilanjutkan dengan menambahkan *operator replace missing value* seperti pada gambar 7 dimana berfungsi untuk menghilangkan nilai-nilai yang kosong pada data excelnya dimana *setting* pada *operator* ini menggunakan *default average*.



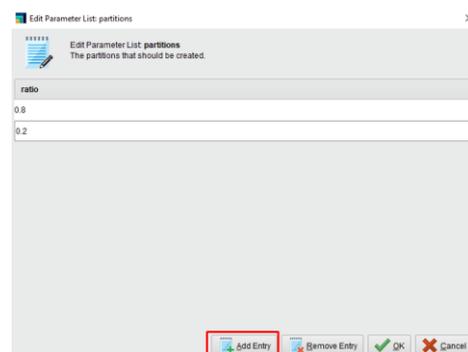
Gambar 7. Operator Replace Missing Value

- 5) Data yang sudah tidak terdapat *missing value* akan dilanjutkan dengan sortir atribut yang dibutuhkan pada pengolahan data penelitian ini dengan menggunakan *operator principal component analysis* seperti pada gambar 8 dan *setting* yang digunakan pada *operator* ini adalah *fixed number* dengan komponen yang menyesuaikan dari penelitian.



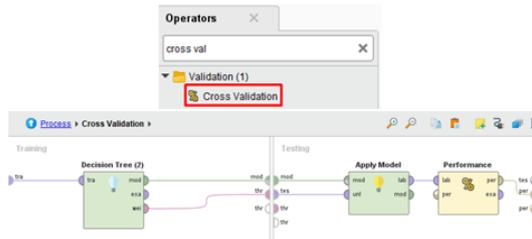
Gambar 8. Operator Principal Component Analysis

- 6) Ketika data sudah siap diolah maka langkah berikutnya adalah masukkan *operator split data* untuk membuat data yang ada terbagi menjadi 2 bagian yaitu data *training* dan data *testing* dengan rasio 0.8/0.2 seperti pada gambar 9.



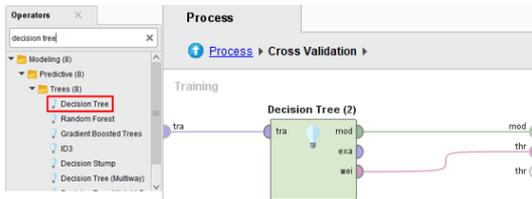
Gambar 9. Perbandingan Rasion Data

- 7) Setelah data terbagi maka akan dilanjutkan dengan memasukkan *operator cross validation* dengan isi *operator* penghitung utama, *apply model*, dan *performance* seperti pada gambar 10.



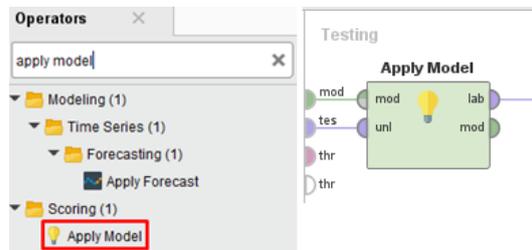
Gambar 10. Operator Cross Validation

- 8) Didalam *operator cross validation* isi proses *training* dengan algoritma penghitung utama seperti pada gambar 11, pada penelitian ini menggunakan *Decision Tree* dimana *setting* yang digunakan pada operator *Decision Tree* dapat disesuaikan dengan kebutuhan penelitian masing-masing.



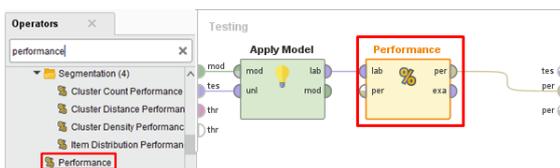
Gambar 11. Pemilihan Operator Penghitung Utama

- 9) Cari operator *apply model* untuk melakukan penerapan perhitungan dari data yang ada pada *split data* dengan menggunakan *Decision Tree*, drag operator *apply model* kedalam kolom sebelah kanan yaitu kolom *testing* seperti pada gambar 12.



Gambar 12. Operator Apply Model

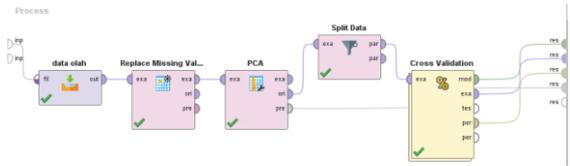
- 10) Cari operator *performance* pada folder *validation* -> *performance* seperti pada gambar 13 atau juga bisa menggunakan *performance* (classification) pada folder *predictive* didalam *performance* untuk menampilkan hasil tingkat akurasi dibagian *result*.



Gambar 13. Operator Performance

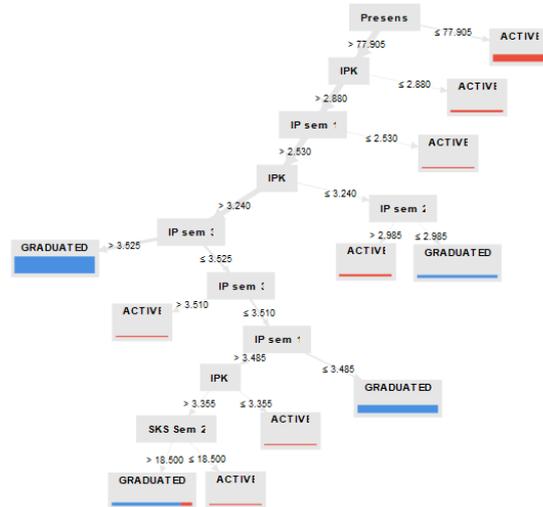
- 11) Gabungkan setiap *line* yang sudah dibuat pada *canvas* seperti pada gambar 14 lalu klik

tombol *play* atau bisa menggunakan *shortcut f11* pada keyboard.



Gambar 14. Penggabungan Setiap Operator

- 12) Setelah melakukan *running* maka hasil akan terlihat pada gambar 15, peneliti hanya harus menyesuaikan dengan kebutuhan penelitiannya atribut-atribut mana yang harus dikurangi ataupun *setting* pada *operator-operator* yang sekiranya harus diubah demi meningkatkan nilai akurasi pada hasil *performance*.



Gambar 15. Hasil Decision Tree C4.5

3.6. Evaluasi

Pada tahap ini peneliti akan melakukan evaluasi pada hasil dari algoritma *Decision Tree* yang dihasilkan pada tahap *modelling* sebelumnya. Dimana hasil tersebut akan dijadikan acuan untuk menghitung prediksi kelulusan mahasiswa angkatan 2020.

- a. Rule Decision Tree C4.5
Setelah melakukan perhitungan dan pengecekan data tanpa adanya kesalahan maka peneliti dapat melihat hasil rule Decision Tree pada tabel description yang dapat dilihat pada gambar 16 dimana disana menjelaskan secara detail alur dari tampilan *Decision Tree* yang terbentuk sebelumnya.

Tree

```

Presensi > 77.905
| IPK > 2.880
| | IP sem 1 > 2.530
| | | IPK > 3.240
| | | | IP sem 3 > 3.525: GRADUATED (GRADUATED=45, ACTIVE=0)
| | | | IP sem 3 ≤ 3.525
| | | | | IP sem 3 > 3.510: ACTIVE (GRADUATED=0, ACTIVE=1)
| | | | | IP sem 3 ≤ 3.510
| | | | | | IP sem 1 > 3.485
| | | | | | | IPK > 3.355
| | | | | | | | SKS Sem 2 > 18.500: GRADUATED (GRADUATED=6, ACTIVE=1)
| | | | | | | | SKS Sem 2 ≤ 18.500: ACTIVE (GRADUATED=0, ACTIVE=1)
| | | | | | | | IPK ≤ 3.355: ACTIVE (GRADUATED=0, ACTIVE=1)
| | | | | | | | IP sem 1 ≤ 3.485: GRADUATED (GRADUATED=20, ACTIVE=0)
| | | | | | | | IPK ≤ 3.240
| | | | | | | | IP sem 2 > 2.985: ACTIVE (GRADUATED=0, ACTIVE=4)
| | | | | | | | IP sem 2 ≤ 2.985: GRADUATED (GRADUATED=6, ACTIVE=0)
| | | | | | | | IP sem 1 ≤ 2.530: ACTIVE (GRADUATED=0, ACTIVE=1)
| | | | | | | | IPK ≤ 2.880: ACTIVE (GRADUATED=0, ACTIVE=4)
Presensi ≤ 77.905: ACTIVE (GRADUATED=0, ACTIVE=18)
    
```

Gambar 16. Hasil Rule Decision Tree C4.5

b. Hasil Confusion Matrix

Pada hasil *operator performance* RapidMiner pada gambar 17, peneliti dapat melihat bahwa hasil pengujian data tersebut tingkat akurasi sebesar 82,27%, tingkat presisi 75,17%, dan juga tingkat *recall* sebesar 70,83%. Dengan besarnya hasil tersebut maka data yang diolah dapat membuat peneliti melanjutkan ke tahap berikutnya.

accuracy: 82.27% +/- 8.41% (micro average: 82.41%)			
	true GRADUATED	true ACTIVE	class precision
pred GRADUATED	67	9	88.16%
pred ACTIVE	10	22	68.75%
class recall	87.01%	70.97%	

Gambar 17. Hasil Confusion Matrix

c. Penggunaan Rumus dari Hasil Rule Decision Tree

Setelah mengetahui tingkat *confusion matrix*, dapat membuat peneliti melanjutkan ke tahap selanjutnya yaitu melakukan pembuatan rumus dengan menggunakan logika pseudocode pada excel sesuai dengan rule hasil *Decision Tree* sebelumnya yang dapat dilihat pada gambar 18. Dimana rule tersebut akan menjadi acuan dalam menentukan prediksi kelulusan mahasiswa angkatan 2020

Uji Prediksi Mahasiswa ITTelkom Surabaya Angkatan 2020

deklarasi:

Presensi, IPK, IP sem 1, IP sem 2, IP sem 3: Real;

algoritma:

```

IF presensi > 77.90 AND IPK > 3,24 AND IP sem 1 > 3,48 AND
IP sem 2 > 2,98 AND IP sem 3 > 3,51
output('Tepat Waktu')
    
```

```

else IF presensi < 77.90 AND IPK < 3,24 AND IP sem 1 < 3,48
AND IP sem 2 < 2,98 AND IP sem 3 < 3,51
output('Tidak Tepat Waktu')
    
```

END IF

Gambar 18. Logika Pseudocode untuk Menghitung Hasil

3.7. Deployment

Data yang sudah diolah pada fase modelling dan juga Evaluasi maka pada tahap ini data akan dipresentasikan untuk dijelaskan kejelasan hasil akhir dari pengolahan data yang dilakukan oleh peneliti.

a. Hasil Prediksi

Pada tabel 1 merupakan hasil dari prediksi dari kelulusan mahasiswa ITTelkom Surabaya angkatan 2020 setiap prodi dengan jumlah 608 mahasiswa, 594 diprediksi lulus tepat waktu dan 14 diprediksi tidak lulus tepat waktu.

Tabel 1. Hasil Persentase Prediksi

Program Studi	Persentase Prediksi Kelulusan
Rekayasa Perangkat Lunak	100%
Sistem Informasi	99,53%
Teknik Elektro	92,31%
Teknik Industri	95,88%
Teknik Komputer	98,28%
Teknik Telekomunikasi	98,11%
Teknologi Informasi	95,51%

b. Perbandingan Confusion Matrix

Dengan menggunakan atribut yang sama untuk data testing dan training adalah data angkatan 2018, dilakukannya beberapa pengujian dengan algoritma lain yaitu: Naïve bayes dan K-Nearest Neighbour (KNN) untuk mengetahui akurasi mana yang tertinggi diantara 3 algoritma tersebut. Pada gambar 19 merupakan hasil dari pengujian tingkat akurasi antara algoritma C4.5, Naïve bayes, dan KNN.

accuracy: 82.27% +/- 8.41% (micro average: 82.41%) Hasil Akurasi Decision Tree C4.5			
	true GRADUATED	true ACTIVE	class precision
pred GRADUATED	67	9	88.16%
pred ACTIVE	10	22	68.75%
class recall	87.01%	70.97%	

accuracy: 91.55% +/- 7.04% (micro average: 91.67%) Hasil Akurasi Naive Bayes			
	true GRADUATED	true ACTIVE	class precision
pred GRADUATED	76	8	90.48%
pred ACTIVE	1	23	95.83%
class recall	98.70%	74.19%	

accuracy: 86.09% +/- 7.81% (micro average: 86.11%) Hasil Akurasi KNN			
	true GRADUATED	true ACTIVE	class precision
pred GRADUATED	73	11	86.90%
pred ACTIVE	4	20	83.33%
class recall	94.81%	64.52%	

Gambar 19. Perbandingan Confusion Matrix

Dari hasil yang sudah terlihat ketika membandingkan hasil akurasi pada ketiga algoritma tersebut bahwa *Decision Tree* C4.5 memiliki tingkat akurasi sebesar 82,27%, tingkat presisi 75,17%, dan juga tingkat *recall* sebesar 70,83%. Untuk Naïve Bayes sebesar 91,55%, tingkat presisi 96,67%, dan juga tingkat *recall* sebesar 74,17%. Dan untuk KNN sebesar 86,09%, tingkat presisi 85,83%, dan juga tingkat *recall* sebesar 64,17%.

4. KESIMPULAN DAN SARAN

Kesimpulan yang dapat ditarik dari penelitian ini adalah:

- 1) Setelah dilakukannya perhitungan prediksi pada data angkatan 2020 dengan menggunakan Data Mining dan algoritma *Decision Tree* C4.5 didapati bahwa hasil akhir prediksi dari 608 mahasiswa, sebanyak 594 mahasiswa diprediksi akan lulus tepat waktu dan 14 mahasiswa diprediksi tidak akan lulus tepat waktu dari Institut Teknologi Telkom Surabaya.
- 2) Penggunaan algoritma C4.5 memiliki hasil akhir *confusion matrix* yang cukup tinggi namun jika disandingkan dengan Naïve Bayes dan KNN, kedua algoritma tersebut masih memiliki tingkat akurasi *confusion matrix* yang lebih tinggi dari *Decision Tree* C4.5 dengan hasil perbandingan tingkat akurasi *Decision Tree* C4.5, Naïve Bayes, dan KNN secara berurutan adalah 82,27%, 91,55%, dan 86,09%, untuk tingkat presisi secara berurutan adalah 75,17%, 96,67%, dan 85,83%, dan tingkat *recall* secara berurutan adalah 70,83%, 74,17%, dan 64,7%. Hal tersebut juga memastikan bahwa algoritma C4.5 dapat memprediksi kelulusan mahasiswa.

Dari kesimpulan diatas yang telah diambil, maka dapat dikemukakan saran-saran yang dapat membantu untuk dilakukannya pengembangan penelitian ini selanjutnya:

1. Lakukan dengan atribut yang lebih banyak dan data dalam atribut memiliki perbedaan yang mencolok agar dapat diolah lebih mudah dan tingkat akurasi semakin tinggi.
2. Lakukan penelitian dengan menggunakan metode lainnya yang dimana dapat memprediksi lebih rinci lagi dengan atribut-atribut yang sudah disempurnakan ataupun ditambah lebih banyak lagi agar faktor penentu lebih akurat dan logis.

5. DAFTAR PUSTAKA

- AGWIL, W., FRANSISKA, H. & HIDAYATI, N., 2020. Analisis Ketepatan Waktu Lulus Mahasiswa Dengan Menggunakan Bagging Cart. *Jurnal Pendidikan Matematika dan Matematika*, VI(2), pp. 155-166.
- DARMAWAN, A., KUSTIAN, N. & RAHAYU, W., 2018. Implementasi Data Mining Menggunakan Model Svm Untuk Prediksi Kepuasan Pengunjung Taman Tabebuaya. *Jurnal Satuan Tulisan Riset dan Inovasi Teknologi*, II(3), pp. 299-307.
- GAOL, L. Y. L., M, S. & SUHENDRO, D., 2021. Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi Dengan Menggunakan Algoritma C4.5. *Jurnal Penerapan Kecerdasan Buatan*, ii(2), pp. 97-106.
- KADE S., I., 2022. *CRISP DM Sebagai Salah Satu Standard untuk Menghasilkan Data Driven Decision Making yang Berkualitas*, Jakarta: Kementerian Keuangan Republik Indonesia.
- KAMIL, M. & CHOLIL, W., 2020. ANALISIS PERBANDINGAN ALGORITMA C4.5 DAN NAIVE BAYES PADA LULUSAN TEPAT WAKTU MAHASISWA DI UNIVERSITAS ISLAM NEGERI RADEN FATAH PALEMBANG. *Jurnal Informatik*, Volume 7, pp. 97-106.
- MAURITIUS, T. & BINSAR, F., 2020. *Cross-Industry Standard Process for Data Mining (CRISP-DM)*, Jakarta: BINUS.
- NATASUWARNA, A., 2019. Tantangan Menghadapi Era Revolusi 4.0 - Big Data dan Data Mining. *Prosiding SINDIMAS*, 29 July.p. 23.
- NURHASAN, F., HIKMAH, N. & UTAMI, D. Y., 2018. PERBANDINGAN ALGORITMA C4.5, KNN, DAN NAIVE BAYES UNTUK PENENTUAN MODEL KLASIFIKASI PENANGGUNG JAWAB BSI ENTREPRENEUR CENTER. *Jurnal PILAR Nusa Mandiri*, Volume 14, pp. 169-174.
- PRASETYO, E. R., 2018. Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Decision Tree Dan Artificial Neural Network. *Jurnal Ilmiah Matrik*, xx(1), pp. 21-30.
- SHEDRIKO & FIRDAUS, M., 2022. *Penentuan Klasifikasi dengan CRISP-DM dalam memprediksi kelulusan mahasiswa pada suatu mata kuliah*. Jakarta, s.n.
- SURABAYA, I., 2022. *Institut Teknologi Telkom Surabaya*. [Online] Available at: <https://ittelkom-sby.ac.id/profil-singkat-itts/> [Diakses 02 February 2023].