BAB II

TINJAUAN PUSTAKA

Pada bab ini akan dibahas tentang tinjauan pustaka dan alus penelitian yang akan dipakai dalam analisis prediksi kelulusan *course* pada e-learning menggunakan model klasifikasi. Adapun dasar Teori yang akan dibahas adalah, model lasifikasi (*classification*), metode *Decision Tree*, metode *Random Forest* dan penelitian sebelumnya yang berkaitan dengan metode klasifikasi dibidang pendidikan.

2.1 Kajian Pustaka

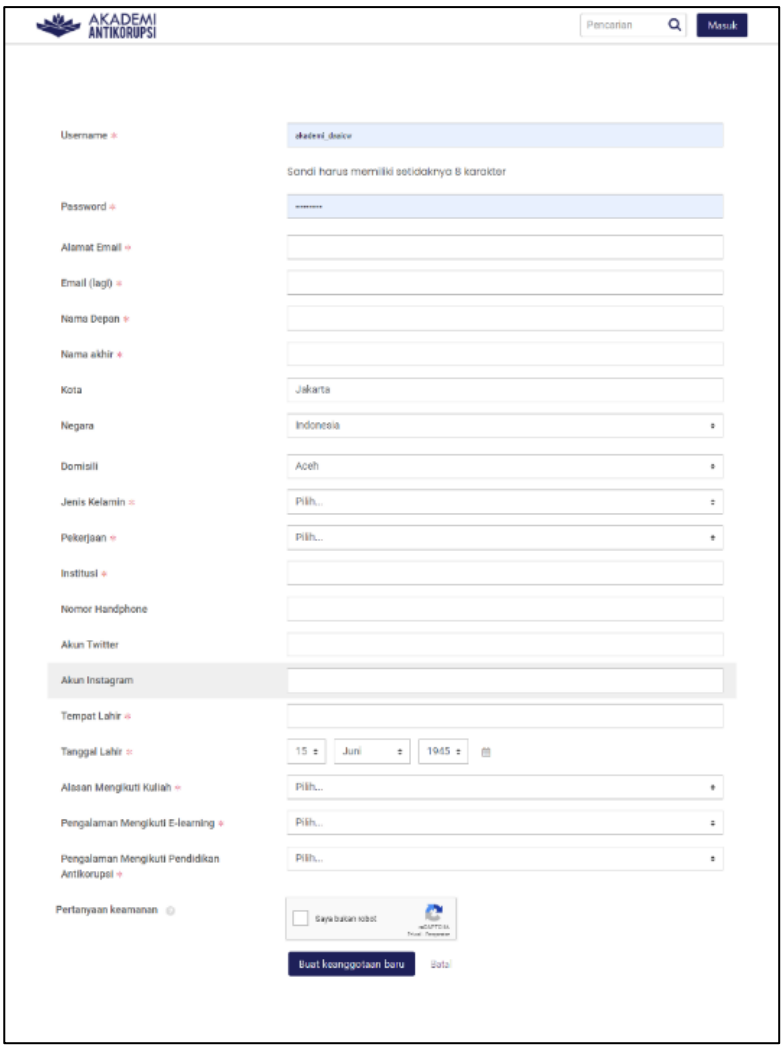
2.1.1 E-Learning Akademi Anti Korupsi

Gambar 2.1. Halaman Utama E-Learning Akademi Anti Korupsi

Akademi Anti Korupsi merupakan platform belajar dengan basis digital yang mudah diakses oleh semua masyarakat sehingga dapat belajar mengenai korupsi dan cara–cara memeranginya. Akademi anti korupsi dirancang guna mengatasi dua persoalan, yakni daya jangkau serta sumber daya uang melalui e-learning. Sehingga peserta akademi anti korupsi akan lebih luas dibandingkan melalui offline (Indonesia Corruption Watch, 2019). Akademi Anti Korupsi ini

merupakan salah satu upaya membangun konsep pendidikan pencegahan dan pemberantasan korupsi (CNNIndonesia, 2018).

Akademi Anti Korupsi sudah berjalan sejak tahun 2018. Program tersebut merupakan salah satu upaya untuk membangun konsep pendidikan dan pencegahan korupsi. Program Akademi Anti Korupsi bertujuan untuk memberikan gambaran dasar mengenai korupsi serta cara pemberantasan nya. Kemudian, berisikan materi yang memberikan pengetahuan dan pemahaman untuk dapat digunakan melawan korupsi. Pada peluncuran awal Akademi Anti Korupsi terdapat enam mata kuliah yang dapat diakses secara virtual (CNNIndonesia, 2018).



Gambar 2.2. Form Registrasi E-Learning Akademi Antikorupsi

Platform Akademi Anti Korupsi menggunakan website sebagai penunjang pelaksanaannya. Langkah pertama yang dilakukan peserta untuk mengakses Akademi Anti Korupsi adalah dengan melengkapi form registrasi peserta pada halaman akademi.antikorupsi.org berupa jenis kelamin, pekerjaan, pengalaman pendidikan Anti Korupsi, umur, institusi, domisili, pengalaman e-learning dan alasan mengikuti e-learning Anti Korupsi. Berdasarkan hal tersebut dapat digunakan untuk pemrosesan data untuk mengetahui klasifikasi data kelulusan menggunakan data yang dimiliki yaitu data yang diisi saat pendaftaran. Jadi kelulusan course seseorang dipengaruhi oleh faktor jenis kelamin, pekerjaan, pengalaman pendidikan Anti Korupsi, umur, institusi, domisili, pengalaman e-learning dan alasan mengikuti e-learning Anti Korupsi. Sehingga untuk mengetahui kelulusan course dapat dikelompokkan menjadi tidak lulus, lulus dengan mengambil 1 *course*, dan lulus dengan mengambil 2 *course* atau lebih.

2.1.2 Data Mining

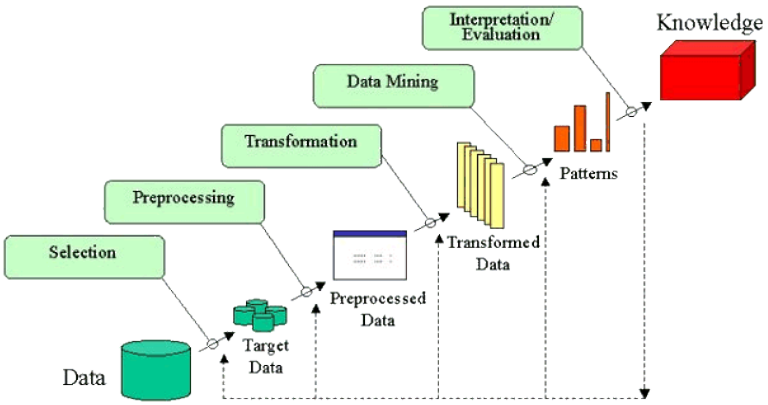
Penjelasan mengenai data mining menurut (Vercellis, 2009) adalah suatu pekerjaan yang mendeskripsikan suatu proses analisa yang mana suatu pekerjaan tersebut terjadi secara berulang dan dilakukan pada database yang cukup besar, dengan tujuan untuk mendapatkan informasi dan pengetahuan yang berhubungan dengan pengambilan keputusan serta pemecahan masalah. Sedangkan menurut (Han & Kamber, 2001) Data mining merupakan proses untuk mencari data dan suatu informasi dalam jumlah besar dari suatu database atau tempat penyimpanan informasi lainnya. Biasanya proses data mining juga menemukan pola yang unik dari suatu informasi. Pola yang ditemukan pada proses data mining harus memiliki arti dan juga memberikan keuntungan bagi penggunanya nya.

Dari beberapa penjelasan mengenai data mining sebelumnya, maka dapat disimpulkan yaitu arti atau penjelasan mengenai data mining ialah suatu metode atau cara untuk menggali dan mendapatkan informasi berharga dan tersembunyi dari suatu database yang besar. Sehingga dari proses tersebut ditemukan lah pola yang unik, yang mana pola tersebut tidak diketahui sebelumnya.

Data mining memiliki beragam metode yang bisa digunakan, diantaranya adalah metode KDD (Knowledge Discovery in Database), SEMMA (Sample, Emplore, Modify, Model, dan Assess) dan CRISP-DM (Cross Industry Standard Process for Data Mining). Setiap proses memiliki metode yang berbeda-beda dalam pencarian informasi penting yang ada di dalam suatu data sebagai berikut:

1. KDD (Knowledge Discovery in Database)

KDD (Knowledge Discovery in Database) adalah salah satu metode yang bisa digunakan dalam melakukan data mining. (Fayyad et al., 1996) mendefinisikan KDD sebagai proses dari menggunakan metode data mining untuk mencari informasi-informasi yang berharga, pola yang ada di dalam data, yang melibatkan algoritma untuk mengidentifikasi pola pada data. Siklus proses KDD dapat dilihat pada Gambar 2.3 berikut:

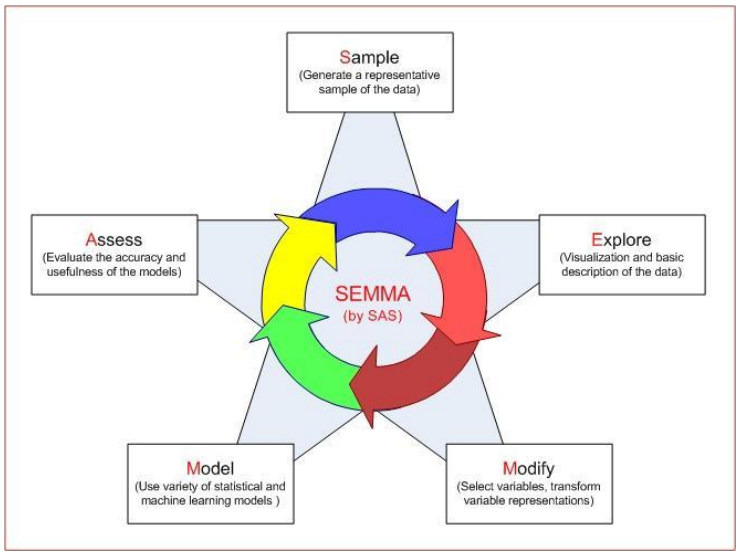


Gambar 2.3. Siklus Proses KDD (Quantum, 2019)

Terdapat lima tahap dalam proses KDD yaitu seleksi membuat kumpulan data target, atau berfokus pada subset variabel atau sampel data yang memerlukan eksplorasi lebih lanjut. Kemudian pra-pemrosesan yaitu untuk mendapatkan data yang konsisten. Selanjutnya, transformasi data menggunakan pengurangan dimensi. Dan data mining yaitu mencari pola yang menarik dalam bentuk representasi tertentu yang bergantung pada tujuan Data Mining (misalnya prediksi). Serta interpretasi/evaluasi yaitu melakukan interpretasi dan evaluasi pola yang sudah dimining.

2. SEMMA

SEMMA adalah singkatan dari Sample, Explore, Modify, Model and Access. Siklus proses SEMMA dapat dilihat pada Gambar 2 berikut:

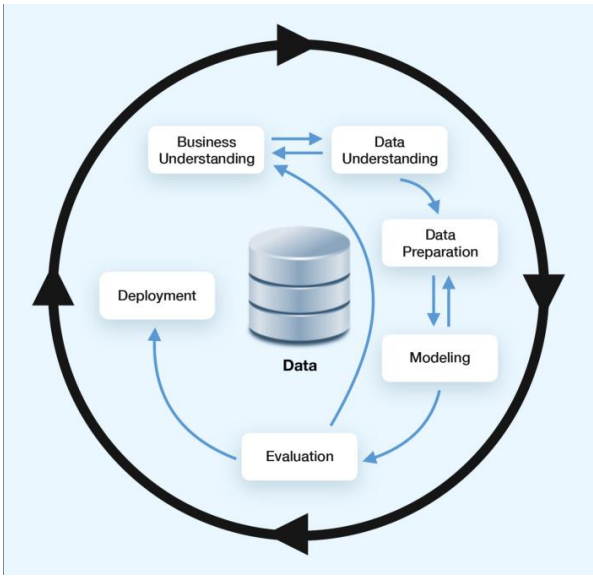


Gambar 2.4. Siklus proses SEMMA (Quantum, 2019)

Proses data mining SEMMA memiliki lima proses tahapan yaitu Sample, Explore, Modify, Model, dan Assess (Mariscal et al., 2010). Sampel yaitu mengambil sample data untuk mengekstrak informasi penting dari data tersebut kemudian dimanipulasi dengan cepat. Eksplorasi data dapat membantu dalam memperoleh pemahaman dan ide serta menyempurnakan proses penemuan dengan mencari tren. Modify yaitu tahap modifikasi data berfokus pada pembuatan, pemilihan dan transformasi variabel untuk memfokuskan proses pemilihan model. Tahap ini juga dapat mencari *missing value* dan mengurangi jumlah variabel. Model yaitu memodelkan data menggunakan berbagai *software*. Akses yaitu tahap terakhir berfokus pada evaluasi keandalan dan kegunaan penemuan dari proses data mining dan mengevaluasi sebaik mana model bekerja.

3. CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP-DM (Cross Industry Standard Process for Data Mining) menyediakan standar proses untuk data mining yang dapat diterapkan ke dalam strategi pemecahan masalah umum pada bisnis atau pada unit penelitian (Utari et al., 2020). Siklus proses CRISP-DM dapat dilihat pada Gambar 3 berikut:



Gambar 2.5. Siklus proses CRISP-DM (Quantum, 2019)

Terdapat enam tahap siklus pengembangan data mining (Daderman & Rosander, 2018) yaitu sebagai berikut :

1. *Business Understanding* (Pemahaman Bisnis)

   Tahap pertama adalah memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menterjemahkan pengetahuan ini ke dalam pendefinisian masalah dalam data mining. Selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut.

2. Data *Understanding* (Pemahaman Data)

   Tahap ini dimulai dengan pengumpulan data yang kemudian akan dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi adanya bagian

yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3. Data *Preparation* (Persiapan Data)

   Tahap ini meliputi kegiatan untuk membangun kumpulan data akhir (data yang akan diproses pada tahap pemodelan) dari data mentah. Pada tahap ini juga mencakup pemilihan tabel, record, dan atribut-atribut data, termasuk proses pembersihan dan transformasi data untuk kemudian dijadikan masukan dalam tahap pemodelan.

4. *Modeling* (Pemodelan)

   Dalam tahap ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan.

5. *Evaluation* (Evaluasi)

   Pada tahap ini, model sudah terbentuk dan diharapkan memiliki kualitas baik jika dilihat dari sudut pandang analisis data. Pada tahap ini akan dilakukan evaluasi terhadap apakah model dapat mencapai tujuan yang ditetapkan pada fase awal (pemahaman data).

6. Deployment (Pengembangan)

   Pada tahap terakhir, pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Tahap pengembangan dapat berupa pembuatan laporan sederhana atau mengimplementasikan proses data mining yang berulang dalam perusahaan.

Dari tiga metode data mining tersebut, (Daderman & Rosander, 2018) menyatakan bahwa CRISP-DM merupakan metode yang berasal dari perspektif bisnis, metode iterative dan mudah diimplementasikan ke dalam proses data mining. CRISP-DM juga terstruktur dan didefinisikan dengan baik. CRISP-DM adalah salah satu kerangka kerja yang paling banyak digunakan. Oleh karena itu, pada penelitian ini metode data mining yang akan digunakan adalah metode CRISP-DM.

Dalam melakukan analisis data mining, terdapat dua pendekatan berdasarkan tugas dan tujuan analisis, yaitu *supervised learning* dan *unsupervised learning* (Vercellis, 2009). *Supervised learning* merupakan sebuah proses pengelompokan data yang telah memiliki label dan akan dimasukkan/dikelompokkan berdasarkan labelnya, juga algoritma yang terdapat pada *supervised* bertujuan untuk memperkirakan atau memprediksi fungsi pada bidang pemetaan sehingga ketika ada variable input (X) maka dapat memprediksi variable output (Y). Sedangkan *unsupervised learning* merupakan sebuah proses pengelompokan data yang tidak diberi label, tipe algoritma yang memiliki variable input (X) dan tetapi tidak memiliki variable output yang sesuai. Tujuan dari unsupervised adalah untuk memodelkan struktur data agar dapat mempelajari data-data tersebut lebih lanjut lagi, mengidentifikasikan pola – pola dalam sekumpulan data yang pada umumnya tidak diklasifikasikan.

Terdapat beberapa jenis metode analisis data mining dengan *supervised learning* sebagai berikut (Vercellis, 2009):

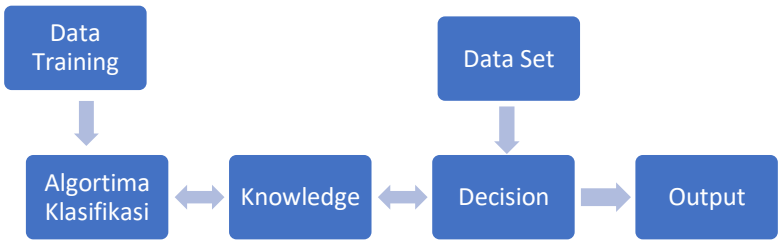
1. *Classification* (Klasifikasi), merupakan proses penemuan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui.
2. *Regression* (Regresi), merupakan metode analisis yang memakai hasil observasi masa lalu utk memprediksi nilai atribut target berdasarkan atribut penjelas hasil observasi di masa depan. Klasifikasi dapat dijadikan regresi, dan sebaliknya
3. *Characterization and Discrimination* (Karakterisasi dan diskriminasi), merupakan metode analisis untuk membandingkan nilai distribusi dari atribut-atribut yg ada di dalam suatu kelas, dan mendeteksi perbedaan antara suatu kelas dengan kelas lain melalui perbandingan distribusi nilai.
4. *Time series*, merupakan metode analisis untuk menginvestigasi data yang memiliki dinamika waktu dan bertujuan untuk memprediksi nilai atribut target dari satu periode mendatang atau lebih.

Metode analisis data mining dengan *unsupervised learning* sebagai berikut (Vercellis, 2009):

1. *Association* (Asosiasi), dinamakan juga analisis keranjang pasar dimana fungsi ini mengidentifikasi item-item produk yang kemungkinan dibeli konsumen bersamaan dengan produk lain.
2. *Clustering* (Pengelompokan), yaitu merupakan metode analisis yang bertujuan untuk melakukan segmentasi populasi yang heterogen menjadi sejumlah kelompok yang beranggotakan observasi dengan karakteristik yang homogen.
3. *Description and visualization* (Deskripsi dan Visualisasi), merupakan metode analisis untuk memberi gambaran secara ringkas bagi sekumpulan data yang jumlahnya sangat besar sehingga dapat memberikan penjelasan tentang pola yang tersembunyi di dalam dataset dan mengarah ke pemahaman yang lebih baik tentang fenomena dari dataset.

2.1.3 Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. (Akinshilo & Agboola, n.d.). Gambar 2.6 menunjukkan konfigurasi sistem klasifikasi secara umum.

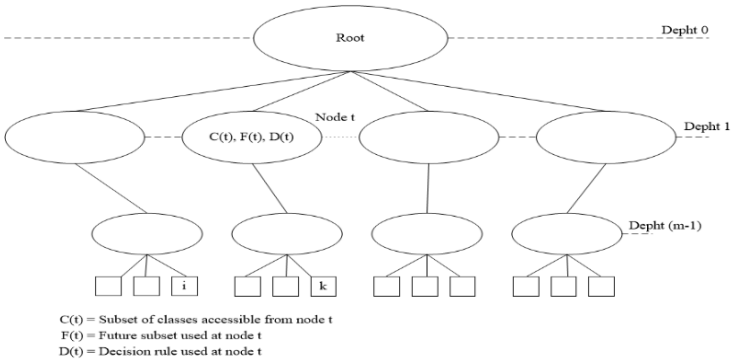


Gambar 2.6. Konfigurasi Sistem Klasifikasi (Sun & Sun, 2007)

Data *training* merupakan data yang berisi fitur atau ciri yang dijadikan sebagai dasar pengelompokan suatu obyek. Dengan memanfaatkan beberapa teknik pemodelan klasifikasi maka informasi yang ada pada data tersebut dapat diketahui. Model klasifikasi yang telah dibuat selanjutnya disimpan dan dijadikan sebagai dasar klasifikasi untuk prediksi obyek atau data baru yang belum diketahui sebelumnya. Beberapa contoh metode klasifikasi yang sudah dikembangkan antara lain:

2.1.4 Metode Decision Tree

Decision Tree Salah satu metode machine learning untuk *classification* dan *prediction* (regression) (Riyadi Yanto et al., 2020). Banyak dimanfaatkan pada knowledge system, seperti database. Decision tree berbentuk sebuah pohon keputusan seperti yang ditunjukkan pada Gambar 2.7. Metode ini memecah kumpulan data menjadi himpunan bagian yang lebih kecil dan pada saat yang sama pohon keputusan terkait dikembangkan secara bertahap. Hasil akhirnya adalah pohon dengan simpul keputusan (decision node) dan simpul daun (leaf node). Node keputusan memiliki dua atau lebih cabang dan node daun mewakili klasifikasi atau keputusan. Node keputusan teratas dalam pohon yang sesuai dengan predictor terbaik disebut simpul akar. Pohon keputusan dapat menangani data kategorikal maupun numerik.



Gambar 2.7. Diagram *Decision Tree* (Gorea & Buraga, 2006)

2.1.5 Algoritma C5.0

Algoritma C5.0 adalah salah satu algoritma data mining yang khususnya diterapkan pada *Decision Tree*. C5.0 merupakan penyempurnaan algoritma sebelumnya yang dibentuk oleh Ross Quinlan pada tahun 1987, yaitu ID3 dan C4.5. Dalam algoritma ini pemilihan atribut diproses menggunakan gain ratio. Dalam memilih atribut untuk memisah objek dalam beberapa kelas harus dipilih atribut yang menghasilkan gain ratio paling besar. Atribut dengan nilai gain ratio terbesar akan dipilih sebagai parent bagi node selanjutnya. C5.0 menghasilkan tree dengan jumlah cabang per node bervariasi (Dunham, 2003). Perbedaan yang dapat dilihat secara fisik yaitu tree yang akan dihasilkan oleh Algoritma C5.0 akan lebih ringkas jika dibandingkan dengan Algoritma C4.5. Karena itulah pembangunan tree pada algoritma ini lebih cepat dibandingkan Algoritma C4.5. Langkah kerja pembangunan tree pada Algoritma C5.0 mirip dengan pembangunan Algoritma C4.5. Kemiripan tersebut meliputi perhitungan kemunculan kejadian, perhitungan entropy dan information gain. Jika pada Algoritma C4.5 berhenti sampai perhitungan information gain, maka pada Algoritma C5.0 akan melanjutkan nya dengan perhitungan gain ratio dengan menggunakan information gain dan entropy yang telah ada.

Ukuran information gain dan gain ratio digunakan untuk memilih atribut uji pada setiap node di dalam tree. Ukuran ini digunakan untuk memilih atribut atau node pada tree. Persamaan yang digunakan untuk information gain sebagai berikut (Kantardzic, 2003):

$$Entropy(S) = \sum_{i=1}^{n}(-pi) * log_2(pi) \quad (2.1)$$

Keterangan:

E : Entropy

S : Himpunan Kasus n : Jumlah Partisi *S*

p_i : Jumlah Sampel Untuk Kelas i / Proporsi Dari Si Terhadap S

S adalah sebuah himpunan yang terdiri dari s data objek. Diketahui atribut class adalah m dimana mendefinisikan kelas-kelas di dalamnya, Ci (for i= 1, …, m), Si adalah jumlah objek pada S dalam class Ci. Untuk mengklasifikasikan objek yang digunakan maka diperlukan informasi dengan menggunakan aturan seperti di atas. Dimana adalah proporsi kelas dalam output seperti pada kelas Ci dan diestimasikan dengan Si / S. Atribut A memiliki nilai tertentu {a1, a2, …, av}. Atribut A dapat digunakan pada partisi S ke dalam v subset, {S1, S2, …, Sv}, dimana Sj berisi objek pada S yang bernilai aj pada A. Jika A dipilih sebagai atribut uji (sebagai contoh atribut terbaik untuk split), maka subset ini akan berhubungan pada cabang dari node himpunan S. Sij adalah jumlah objek pada class Ci dalam sebuah subset Sj. Untuk mendapatkan nilai information gain selanjutnya digunakan persamaan dibawah ini :

$$Information\ Gain\ (S, A) = \ Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (S_i) \qquad (2.2)$$

Keterangan:
G : Gain
S: Himpunan Kasus A : Atribut
n: Jumlah Partisi Atribut A
$|S_i|$: Jumlah Sampel Pada Partisi ke –i
|S| : Jumlah Sampel Dalam S

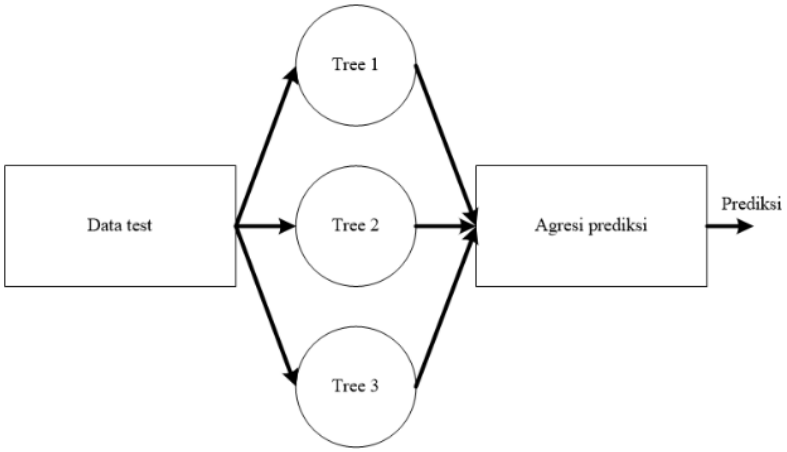
Perhitungan gain ratio untuk Algoritma C5.0 akan berjalan setelah perhitungan information gain diatas dilakukan. Perhitungan gain ratio selanjutnya menggunakan persamaan dibawah ini :

$$Gain\ Ratio = \frac{Information\ Gain\ (S,A)}{\sum_{i=1}^{n} Entropy\ (S_i)} \qquad (2.3)$$

Dengan adanya perhitungan gain ratio inilah yang menjadikan pembangunan tree pada C5.0 lebih ringkas dibanding tree pada Algoritma C4.5.

2.1.6 Metode Random Forest

*Random Forest*s (RF) atau *random decision forests* merupakan algoritma model klasifikasi dan regresi yang dikembangkan berdasarkan konsep *Decision Tree*, seperti ditunjukkan pada gambar 2.8. *Random Forest* bekerja dengan membangun lebih dari 1 *Decision Tree* secara acak pada saat *training*. Perbedaan *Decision Tree* dengan *Random Forest* adalah pada *Decision Tree* dilakukan *training* menggunakan sampel individu, sedangkan *Random Forest* pemilihan atribut diambil secara acak saat sebuah node akan pecah. Setiap tree diberi sampel data *training* dengan menggunakan metode bagging.



Gambar 2.8. Diagram *Random Forest*

2.1.7 Evaluasi Model

Setelah melakukan proses klasifikasi maka dilakukan evaluasi dengan mengukur performa (Sumitra et al., 2019). Pengukuran performa dilakukan untuk melihat hasil yang didapatkan dari klasifikasi. Terdapat beberapa cara untuk mengukur performa sebagai berikut (Wiyono et al., 2020):

1. Split Validation

Split validation adalah teknik validasi yang membagi data menjadi dua bagian secara acak, sebagian sebagai data *training* dan sebagian lainnya sebagai data *testing*. Pada pengujian *split validation* dilakukan evaluasi untuk sebagian

data saja, *split validation* 70% berarti ada sebanyak 70% ukuran sampel digunakan sebagai data *training* sedangkan sisanya 30% ukuran sampel digunakan untuk data *testing*.

2. Confusion Matrix

Untuk melakukan evaluasi terhadap model klasifikasi berdasarkan perhitungan objek *testing* mana yang diprediksi benar dan tidak benar. Perhitungan ini di tabulasi kan ke dalam tabel yang disebut *Confusion Matrix* (Wiyono et al., 2020), (P et al., 2019). Beberapa cara yang sering digunakan adalah dengan menghitung *Accuracy*, *Precision*, dan *Recall*.

1. *Accuracy* merupakan adalah ukuran untuk mengukur ketepatan prediksi pengklasifikasian pada kelas tertentu. Rumus untuk menghitung akurasi klasifikasi sesuai persamaan (2.7).

   $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.4)$$

2. *Precision* merupakan ukuran untuk mengukur ketepatan presisi pengklasifikasian pada kelas tertentu. Rumus untuk menghitung presisi klasifikasi sesuai persamaan (2.8).

   $$Precision = \frac{TP}{TP + FP} \quad (2.5)$$

3. *Recall* adalah ukuran untuk mengukur berapa banyak data dari kelas tertentu yang dapat diprediksikan secara benar. Rumus untuk menghitung *recall* klasifikasi sesuai persamaan (2.9)

   $$Recall = \frac{TN}{TN + FP} \quad (2.6)$$

4. Skor F1 adalah perbandingan rata-rata presisi dan *Recall* yang dibobotkan. Skor F1 dapat dikatakan terbaik jika ada semacam keseimbangan antara presisi dan *Recall* dalam sistem. Nilai dari skor F1 dapat dilihat pada persamaan (2.10).

   $$Skor\ F1 = 2\ x \frac{Recall\ x\ Precision}{Recall\ x\ Precision} \quad (2.7)$$

TP adalah True Positive, FP adalah False Positive, TN adalah True Negative, dan FN adalah False Negative.(Ram & Vishwakarma, 2021)

3. Kurva ROC (Receiver Operation Characteristic)

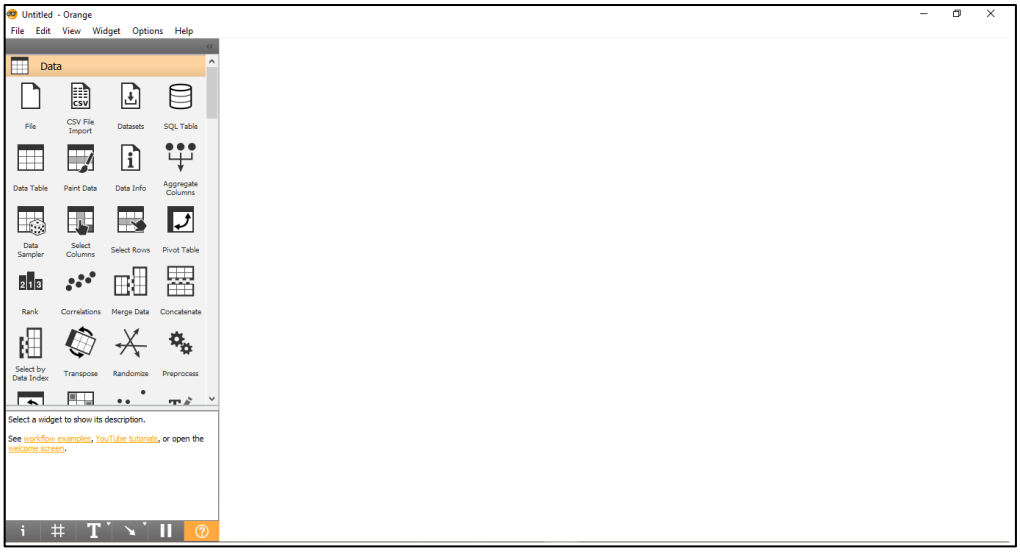
Kurva ROC menunjukkan visualisasi dari akurasi model dan perbandingan perbedaan antar model klasifikasi (Ram & Vishwakarma, 2021), (Riyadi Yanto et al., 2020). Kurva ROC mengekspresikan *Confusion Matrix*. ROC adalah grafik dua dimensi dengan false positive sebagai garis horizontal dan true positive untuk mengukur perbedaan per formasi metode yang digunakan (Gorunescu, 2011). Kurva ROC merupakan teknik untuk memvisualisasikan dan menguji kinerja pengklasifikasian. Model klasifikasi yang lebih baik yang menunjukkan ROC lebih besar.

Selain ROC, terdapat pula kurva AUC (area under curve) yaitu kurva yang berada di bawah area kurva ROC (Ram & Vishwakarma, 2021). Semakin tinggi AUC maka semakin baik. Model yang baik mempunyai AUC di dekat angka 1 sedangkan yang mendekati angka 0 maka model nya tidak baik. Untuk keakurasian nilai AUC dalam klasifikasi data mining dibagi menjadi lima kelompok (Gorunescu, 2011), yaitu:

Tabel 2.1. Klasifikasi Nilai AUC

Nilai	Keterangan
0.90 – 1.00	Klasifikasi Sangat Baik (*Excellent Classification*)
0.80 – 0.90	Klasifikasi Baik (*Good Classification*)
0.70 – 0.80	Klasifikasi Cukup (*Fair Classification*)
0.60 – 0.70	Klasifikasi Buruk (*Poor Classification*)
0.50 – 0.60	Klasifikasi Salah (*Failure*)

2.1.8 Software Orange



Gambar 2.9. Tampilan Awal *Software* Orange

Pada tahun 2009 pertama kalinya orange dirilis oleh GNU General Public License. Orange dapat digunakan pada beberapa sistem operasi yang berbeda seperti Microsoft Windows dan iOS. Adapun bahasa yang digunakan orange adalah phyton C++ dan bahasa C (Wiguna & Rifai, 2021). Tipe fitur yang disediakan oleh orange adalah machine learning, data mining, dan data visualization. Beberapa keuntungan dari orange adalah visualisasi data yang interaktif, user friendly bagi para pemula, memiliki debugger yang lebih baik, skrip yang digunakan sederhana, bersifat open source sehingga bisa diakses dengan mudah dan murah (Ishak et al., 2020). Sedangkan yang menjadi kelemahan dari orange adalah instalasi besar, pembelajaran mesin mempuyai kemampuan reporting yang terbatas, orange juga lemah dalam pengolahan data statistik. Orange merupakan perangkat lunak penambangan data atau teknologi pembelajaran mesin open source. Orange dapat digunakan untuk analisis dan visualisasi data eksploratif. Orange menyediakan platform untuk pemilihan eksperimen, Pemodelan *predictive*, dan sistem rekomendasi dan dapat digunakan untuk penelitian genomic, biomedis, Bio informatika, dan pengajaran. Orange mempermudah pemakai bermain dengan data open source serta melaksanakan proses data analytics secara intuitif.

2.2 Alur Penelitian (*Roadmap*)

2.2.1 Penelitian Terdahulu

Dalam penulisan tesis ini, penulis merujuk pada beberapa penelitian terdahulu mengenai prediksi kelulusan seseorang dengan model klasifikasi menggunakan metode *Decision Tree* dengan algoritma C5.0 dan metode *Random Forest* yang dapat dilihat pada Tabel 2.2 berikut ini:

Tabel 2.2. Penelitian yang berhubungan dengan Model Klasifikasi

No	Nama, Tahun	Judul	Hasil Penelitian	Persamaan	Perbedaan	
					Penelitian Terdahulu	Penelitian ini
1.	Meylani Utari, Budi Warsito, & Retno Kusumaningrum (2020)	Implementation of Data Mining for Drop-Out Prediction using *Random Forest* Method	*Random Forest* memiliki tingkat akurasi 93,43%.	- Sama menggunakan metode *Random Forest* - Sama menggunakan model CRISP-DM	- Metode yang digunakan hanya *Random Forest* - Menggunakan atribut nilai IPK - Data *missing value* dihilangkan - Pendistribusian dataset adalah cross-validation - *Tool* yang digunakan Rapid Minder	- Membandingkan 2 metode: *Decision Tree* dan *Random Forest* - Menggunakan atribut data pendaftaran - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan. - Pendistribusian dataset adalah *split validation* - *Tool* yang digunakan Orange
2.	Edy Budiman, Haviluddin, Nataniel Dengan, dkk (2018)	Performance of *Decision Tree* C4.5 Algorithm in Student Academic Evaluation	Algoritma C4.5 memiliki akurasi (AC) sebesar 78,57% dengan true positive rate (TP) sebesar 76,72% dengan menggunakan data *testing* 90% memiliki nilai 3.akurasi kinerja terbaik.	- Sama melakukan prediksi kelulusan - Sama menggunakan metode *Decision Tree* - Pendistribusian dataset adalah *split validation*	- Metode yang digunakan hanya *Decision Tree* menggunakan algoritma c4.5 - Menggunakan atribut nilai IPK - Data *missing value* dihilangkan - Tidak membahas metode penelitian - Menggunakan Model KDD - *Tool* yang digunakan WEKA	- Membandingkan 2 metode: *Decision Tree* menggunakan algoritma C5.0 dengan *Random Forest* - Menggunakan atribut data pendaftaran - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan - Menggunakan model CRISP-DM - *Tool* yang digunakan Orange
3.	Natanael Benediktus Raymond, Sunardi Oetama (2020)	The *Decision Tree* C5.0 *Classification* Algorithm for Predicting Student Academic Performance	Algoritma C.50 dibagi menjadi data latih sebesar 75% dan data uji sebesar 25% diperoleh akurasi sebesar 71,667% dalam memprediksi prestasi akademik mahasiswa.	- Sama menggunakan algoritma C5.0	- Prediksi prestasi mahasiswa - Menggunakan atribut data keaktifan siswa - Hanya menggunakan algoritma C5.0	- Prediksi kelulusan seseorang - Menggunakan atribut data pendaftaran - Membandingkan algoritma C5.0 dengan *Random Forest* - Menggunakan model

No	Nama, Tahun	Judul	Hasil Penelitian	Persamaan	Perbedaan	
					Penelitian Terdahulu	Penelitian ini
					- Tidak membahas metode penelitian - Ppendistribusian dataset adalah random - Tidak membahas *missing value* - Tidak membahas *Tool* yang digunakan WEKA.	CRIPS-DM - Pendistribusian dataset adalah *split validation* - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan. - *Tool* yang digunakan Orange
4.	W Baswardon, D Kurniadi, A Mulyani and D M Arifin (2019)	Comparative *Analysis* of *Decision Tree* algorithms: Random forest and C4.5 for airlines customer satisfaction *Classification*	Algoritma *Random Forest* memiliki hasil yang lebih baik daripada algoritma C4.5 untuk digunakan dalam klasifikasi kepuasan pelanggan maskapai penerbangan.	- Sama menggunakan metode *Random Forest* dan *Decision Tree* - Pendistribusian dataset adalah *split validation*	- Prediksi kepuasan pelanggan - Membandingkan algoritma C4.5 dengan *Random Forest* - Menggunakan model KDD - Data *missing value* dihilangkan - *Tool* yang digunakan Rapid Minder	- Prediksi kelulusan seseorang - Membandingkan algoritma C5.0 dengan *Random Forest* - Menggunakan model CRISP-DM - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan. - *Tool* yang digunakan Orange
5.	Dela Youlina Putri (2018).	*Analysis* of Students Graduation Target Based on Academic Data Record Using Algoritma C4.5 Algorithm Case Study: Information Systems Students of Telkom University.	Model prediksi Algoritma C4.5 memiliki nilai akurasi sebesar 82,24% dan menyatakan bahwa faktor yang paling berpengaruh dalam memprediksi kelulusan mahasiswa adalah IPK pada tahun kedua.	- Sama melakukan prediksi kelulusan - Sama menggunakan metode *Decision Tree*	- Hanya Menggunakan algoritma C4.5 - Menggunakan atribut nilai IPK - Menggunakan model KDD - Data *missing value* dihilangkan - Ppendistribusian dataset adalah random - *Tool* yang digunakan Rapid Minder	- Membandingkan algoritma C5.0 dengan *Random Forest* - Menggunakan atribut data pendaftaran - Menggunakan model CRISP-DM - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan. - Pendistribusian dataset menggunakan *split validation* - *Tool* yang digunakan Orange
6.	Mehta Smruti (2019).	Predicting Students' Performance using J48 *Decision Tree*.	Dalam proses pengujian, akurasi nya lebih dari 90%, artinya sangat cocok untuk mengklasifikasikan kumpulan data yang besar.	- Sama melakukan prediksi kelulusan - Sama menggunakan metode *Decision Tree*	- Hanya menggunakan algoritma J4.8 - Menggunakan atribut nilai IPK - Tidak membahas metode penelitian - Data *missing value* dihilangkan - Pendistribusian dataset adalah	- Membandingkan algoritma C5.0 dengan *Random Forest* - Menggunakan atribut data pendaftaran - Menggunakan model CRISP-DM - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan. - Pendistribusian dataset menggunakan *split validation*

No	Nama, Tahun	Judul	Hasil Penelitian	Persamaan	Perbedaan	
					Penelitian Terdahulu	Penelitian ini
					cross-validation - *Tool* yang digunakan WEKA	- *Tool* yang digunakan Orange
7.	Musaddiq Al Karim (2022).	Evaluating the Performance of ID3 Method to Analyze and Predict Students' Performance in Online Platforms.	Metode validasi silang 10 kali lipat mengungguli persentase split hampir 3 persen, di mana metode validasi ini mencapai akurasi hampir 77,86 persen.	- Sama melakukan prediksi kelulusan - Sama menggunakan metode *Decision Tree*	- Hanya Menggunakan algoritma ID3 - Tidak membahas metode penelitian - Data *missing value* dihilangkan - Ppendistribusian dataset adalah cross-validation	- Membandingkan algoritma C5.0 dengan *Random Forest* - Menggunakan model CRISP-DM - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan. - Pendistribusian dataset menggunakan *split validation*
8.	Radwan Qasrawi (2020)	Data Mining Techniques in Identifying Factors Associated with Schoolchildren Science and Arts Academic Achievement.	Hasilnya menunjukkan akurasi model CHAD 63% dalam mengklasifikasikan dan memprediksi faktor-faktor yang terkait dengan prestasi Seni dan Sains anak sekolah.	- Memprediksi prestasi akademik - Sama menggunakan metode *Decision Tree*	- Hanya Menggunakan algoritma CHAID - Tidak membahas metode penelitian - Tidak membahas data *missing value* - Tidak membahas ppendistribusian dataset	- Membandingkan algoritma C5.0 dengan *Random Forest* - Menggunakan model CRISP-DM - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan. - Pendistribusian dataset menggunakan *split validation*
9.	Dragutin Petkovic (2016)	Using the *Random Forest Classifier* to Assess and Predict Student Learning of *Software* Engineering Teamwork	Hasil akurasi metode *Random Forest* sekitar 70%.	- Sama menggunakan metode *Random Forest* - Pendistribusian dataset menggunakan *split validation*	- Prediksi awal efektivitas pembelajaran siswa dalam kerja tim rekayasa perangkat lunak. - Tidak membahas metode penelitian - Menghilangkan data *missing value*	- Membandingkan algoritma C5.0 dengan *Random Forest* - Menggunakan model CRISP-DM - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan.
10.	Saba Ba*Tool* (2021)	A *Random Forest* Students' Performance Prediction (RFSPP) Model Based on Students' Demographic Features	*Random Forest* dengan tiga dataset yang berbeda memberikan F-measure sebesar 81,20%, 95,10%, dan 84,16%.	- Sama menggunakan metode *Random Forest* - Sama melakukan prediksi kelulusan - Pendistribusian dataset menggunakan *split validation*	- Hanya Menggunakan metode *Random Forest* - Tidak membahas metode penelitian - Menghilangkan data *missing value*	- Membandingkan Algoritma C5.0 dengan *Random Forest* - Menggunakan model CRISP-DM - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan.
11.	Gerhana (2019)	Comparison of naive Bayes *classifier* and C4.5	Keakuratan algoritma Naive Bayes *Classifier* sekitar 88% sedikit lebih	- Sama menggunakan metode	- Membandingkan *Decision Tree* dan Naïve Bayes	- Membandingkan Algoritma C5.0 dengan *Random Forest*

No	Nama, Tahun	Judul	Hasil Penelitian	Persamaan	Perbedaan	
					Penelitian Terdahulu	Penelitian ini
		algorithms in predicting student study period	baik daripada C4.5 algoritma yang memiliki akurasi sekitar 87%.	*Decision Tree*	- Tidak membahas metode penelitian - Tidak membahas data *missing value* - Tidak membahas pendistribusian dataset	- Menggunakan model CRISP-DM - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan. - Pendistribusian dataset menggunakan *split validation*
12.	Anirudh Hebbar P (2019)	*Decision Tree* and *Random Forest* Based *Classification* Model to Predict Diabetes	Hasil akurasi 72% untuk *Decision Tree* dan 76,5% untuk *Random Forest*.	- Sama membandingkan *Decision Tree* dan metode *Random Forest*	- Memprediksi penyakit diabetes - Membandingkan algoritma C4.5 dan *Random Forest* - Tidak membahas metode penelitian - Menghilangkan data *missing value* - Tidak membahas pendistribusian dataset	- Memprediksi kelulusan seseorang - Membandingkan Algoritma C5.0 dengan *Random Forest* - Menggunakan model CRISP-DM - Membandingkan data *missing value* yang dihilangkan dengan yang tidak dihilangkan. - Pendistribusian dataset menggunakan *split validation*

2.2.2 Kerangka Pemikiran

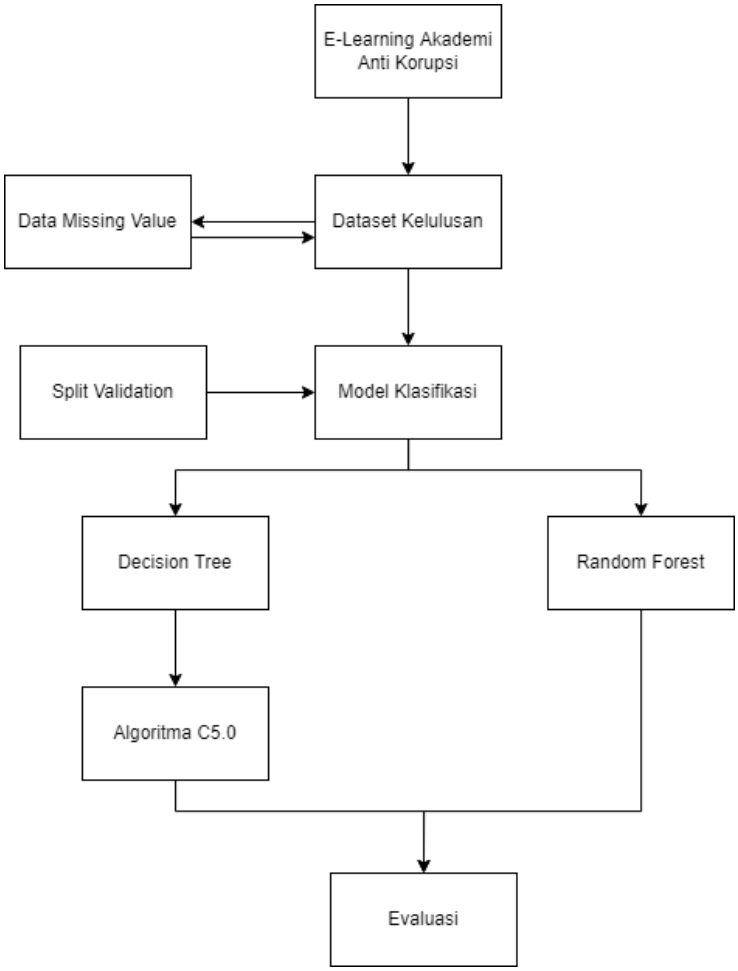
Penelitian-penelitian terdahulu yang berhubungan dengan analisis faktor-faktor yang mempengaruhi kelulusan seseorang menggunakan data mining telah banyak dilakukan. Kebanyakan dari penelitian-penelitian tersebut membahas teori-teori data mining untuk melakukan prediksi. Tabel 2.2 memperlihatkan beberapa penelitian yang berhubungan dengan faktor kelulusan seseorang. Banyak yang menggunakan Nilai IPK sebagai atribut yang digunakan. Sedikit sekali yang menggunakan atribut data pendaftaran peserta. Kemudian terlihat bahwa sebagian besar data mining digunakan pada bidang pendidikan seperti, perguruan tinggi dan sekolah menengah atas. Sedikit yang memperhatikan penerapan data mining untuk bidang pendidikan yaitu E-Learning. Di samping itu belum terdapat penerapan metode klasifikasi yang membandingkan metode *Decision Tree* menggunakan algoritma C5.0 dengan *Random Forest* pada E-Learning. Oleh karena itu, peneliti akan membandingkan metode keakurasian metode *Decision Tree* menggunakan algoritma C5.0 dengan *Random Forest* untuk

memprediksi kelulusan berdasarkan atribut pendaftaran peserta Akademi Anti Korupsi.

Pada proses persiapan data, dapat diketahui bahwa terdapat perbedaan dalam menangani data *missing value*, sebagian besar penelitian menghilangkan data *missing value* dan sebagian lain ada yang tidak menghilangkan data *missing value*, serta ada juga yang tidak membahas sama sekali mengenai data *missing value*. Penanganan data *missing value* ini perlu diperhatikan karena akan mempengaruhi hasil akurasi dari model klasifikasi Oleh karena itu, penelitian ini berusaha untuk membandingkan proses klasifikasi dengan menghilangkan data *missing value* dan tidak menghilangkan data *missing value*.

Pada proses penerapan metode klasifikasi, dapat diketahui bahwa terdapat perbedaan dalam pendistribusian dataset, sebagian besar penelitian menggunakan *cross validation* dan sebagian lain ada yang menggunakan *split validation*, serta ada juga yang menggunakan pendistribusian dataset secara random. Pendistribusian dataset ini perlu diperhatikan karena akan mempengaruhi hasil akurasi dari model klasifikasi Oleh karena itu, penelitian ini akan mengetahui konsistensi perhitungan masing-masing model klasifikasi pada tingkatan jumlah data *training* sehingga akan menerapkan pendistribusian dataset secara *split validation*.

Berdasarkan penjelasan diatas, maka dapat digambarkan kerangka penelitian yang akan dilakukan pada gambar 2.10 berikut ini:



Gambar 2.10. Kerangka Penelitian