BAB II

LANDASAN TEORI

2.1 DATA MINING

Data Mining merupakan bagian dari *Knowledge Discovery in Database* (KDD) yang terdiri dari beberapa tahapan yaitu pemilihan data, *preprocessing, transformation, data mining* serta *interpretation evaluation.* Menurut Dewi Kartika Pane. [8] menyatakan bahwa *Data Mining* merupakan suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam basis data. *Data mining* juga merupakan proses yang menggunakan matematika, teknik statistik, kecerdasan buatan dan *machine learning* untuk mengidentifikasi dan mengekstraksi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar.

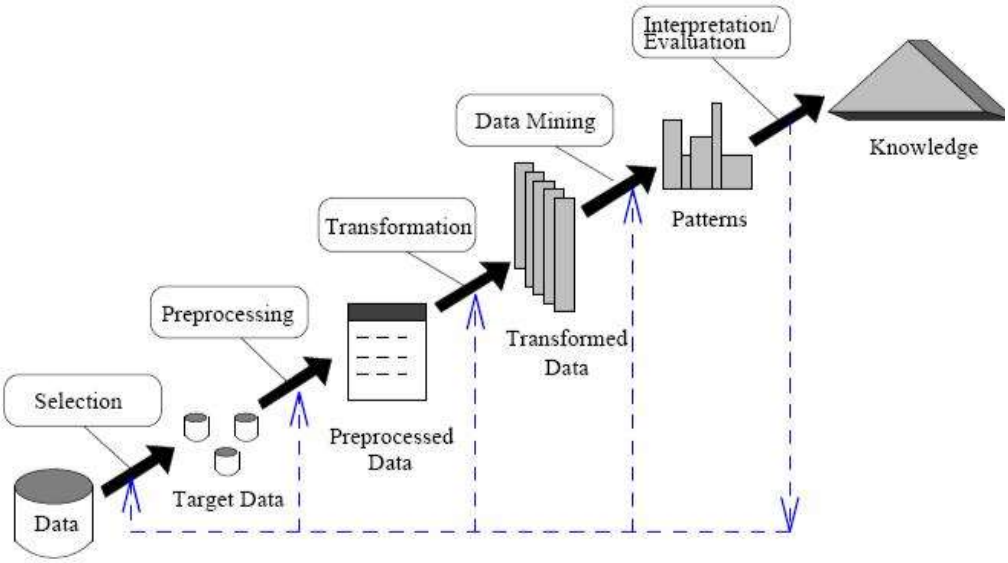
Menurut Jiawei Han dkk. [9] menyatakan bahwa *Data Mining* adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database*, data *warehouse*, atau penyimpanan informasi lainnya. *Data mining* berkaitan dengan bidang ilmu-ilmu lain, seperti *database system*, data *warehousing*, *statistik*, *machine learning*, *information retrieval*, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti *neural network*, pengenalan pola, spatial data *analysis*, *image database* dan *signal processing*.

Dari beberapa pendapat diatas mengenai *data mining*, maka dapat disimpulkan bahwa *data mining* adalah sebuah teknik untuk menemukan pola-pola

yang sebelumnya tidak diketahui dari berbagai macam bidang ilmu untuk menggali data yang besar dan mengesktrak data tersebut sehingga menjadi suatu informasi yang bermanfaat dikemudian hari.

2.1.1 Tahapan Proses Data Mining

Istilah *data mining* dan *Knowledge Discovery in Database* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi yang tersembunyi dalam suatu basis data yang besar. Salah satu tahapan dalam keseluruhan proses *Knowledge Discovery in Database* (KDD) adalah *data mining* secara detail prosesnya disajikan pada gambar 2.1



Gambar 2. 1 Proses KDD Data Mining[10]

Berdasarkan pada gambar 2.1 yang telah dicantumkan diatas, maka proses *data mining* dapat dijelaskan sebagai berikut :

a) Seleksi data (*selection*)

Pemilihan atau seleksi data yang dilakukan dari sekumpulan data. Data yang ada pada basis data seringkali tidak dipakai secara menyeluruh, oleh karena itu hanya data yang sesuai untuk proses analisis yang akan diambil dari basis data tersebut.

b) Pemilihan data (*Preprocessing*)

Pada tahap *preprocessing*, perlu dilakukan tahap perbersihan atau *cleaning* pada data. Pembersihan data mencakup pemilihan memeriksa data yang tidak konsisten, menghilangkan data *noise* dan data duplikasi. setelah data dibersihkan selanjutnya dilakukan pembagian data atau *split data* menjadi *data training* dan *data testing* yang kemudian akan di kelola menggunakan metode mining.

c) Transformasi data (*data transformation*)

Data transformasi sesuai jenis atau pola informasi yang akan dicari sehingga sesuai untuk proses *data mining*. Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*. Beberapa metode *data mining* membutuhkan format data yang khusus sebelum diaplikasikan.

d) *Data Mining*

Data Mining adalah proses pencarian pola atau informasi yang menarik dalam data yang terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode dan algoritma dalam Data Mining sangat bervariasi. Pemilihan

metode dan algortima yang akan digunakan sangat bergantung pada tujuan proses KDD secara menyeluruh.

e) *Interpretation/evaluation*

Proses menerjemahkan pola informasi atau data yang telah didapat kedalam bentuk yang lebih mudah dimengerti oleh *end-user* dan pihak-pihak yang berkepentingan.

2.1.2 Pengelompokan Data Mining

Menurut Masripah. [11] "dalam *data mining* dibagi beberapa kelompok berdasarkan tugas yang dapat dilakukan" yaitu :

a. Deskripsi

Terkadang peneliti dan analisis secara sederhana ingin mencoba mencari cari untuk menggambarkan pola dan dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecendrungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecendrungan.

b. Estimasi

Estimasi adalah untuk memperkirakan suatu hal yang ada datanya. Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih kearah numerik daripada kearah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi.

Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

c. Prediksi

Prediksi adalah memperkirakan hasil dari hal yang belum diketahui. Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.

d. Klasifikasi

Melakukan pengelompokan objek berdasarkan kelompok yang sudah ada. Berbeda dengan klastering, klasifikasi ini memerlukan data pelatihan yang sudah diberi label kelompok atau kelas.

e. Klasterisasi (*Clustering*)

Data yang dikelompokan disebut objek atau catatan yang memiliki kemiripan atribut kemudian dikelompokan pada kelompok yang berbeda. *Cluster* adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record* dalam *cluster* lain.

f. Asosiasi

Melakukan asosiasi antar objek dalam suatu set data, biasanya data transaksional. Asosiasi dilakukan dengan menghitung berapa kali dalam suatu set data suatu transasksi yang mangandung dua item atau lebih yang berhubungan. Sering disebut *Market Basket Analytics.*

2.2 KLASIFIKASI

Klasifikasi merupakan salah satu pembelajaran yang paling umum di *data mining*. Berikut beberapa definisi klasifikasi yaitu :

Menurut Eko Prasetyo [12] mengatakan bahwa :

> "Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu (1) pembangunan model sebagai prototipe untuk disimpan sebagai memori dan (2) penggunaan model tersebut untuk melakukan pengenalan klasifikasi atau prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya".

Menurut Husin et al. [13] menyatakan bahwa "Klasifikasi adalah proses pengelompokan atau pengidentifikassian objek ke dalam sebuah kategori, kelas atau kelompok berdasarkan prosedur yang telah didefinisikan".

Menurut Nelly et al. [14] menyatakan bahwa "Klasifikasi didefinisikan sebagai bentuk analisis data dan proses pengelompokkan objek yang memiliki karakteristik atau ciri yang sama ke dalam beberapa kelas".

Kelas dalam klasifikasi merupakan atribut dalam satu set data yang paling unik yang merupakan variabel bebas dalam statistik. Klasifikasi data terdiri dari dua proses yaitu tahap pembelajaran dan tahap pengklasifikasian. Tahap pembelajaran merupakan tahapan dalam pembentukan model klasifikasi, sedangkan tahap pengklasifikasian merupakan tahapan penggunaan model klasifikasi untuk memprediksi label kelas dari suatu data. Contoh sederhana dari teknik data mining adalah pengklasifikasian hewan berdasarkan atribut jumlah kaki, habitat dan orga pernafasan yang akan diklasifikasikan ke dalam dua label kelas yaitu unggas dan

ikan. Label kelas unggas adalah data yang memiliki jumlah kaki dua, habitatnya di darat dan organ pernafasnnya menggunakan paru-paru, sedangkan label kelas sikan adalah data yang memiliki jumlah kaki nol (tidak memiliki kaki), habitat di air dan organ pernafasannya menggunakan insang. Banyak algoritma yang dapat digunakan dalam pengklasifikasian data, namun pada penelitian ini hanya menggunakan algoritma *naïve bayes*.

Beberapa algortima yang dapat digunakan dalam klasifikasi *data mining* adalah sebagai berikut :

a. *Neural Network*

Neural Network (jaringan Saraf Tiruan) adalah prosesor tersebar yang sangat besar dan memiliki kecendrungan untuk menyimpan pengetahuan yang sangat besar dan memiliki kecendrungan untuk menyimpan pengetahuan yang bersifat pengalaman dan membuatnay siap untuk digunakan.

b. *Decision Tree*

Decision Tree sendiri merupakan metode klasifikasi dan prediksi yang sangat kuat dan banyak diminati. Dalam *decision tree* ini data yang berupa fakta dirubah menjadi sebuah pohon keputusan yang berisi aturan dan tentunya dapat lebih mudah dipahami dengan bahasa alami. Model pohon keputusan banyak digunakan pada kasus data dengat *output* yang bernilai diskrit. Walaupan tidak menutup kemungkinan dapat juga digunakan untuk kasus data dengan atribut *numeric*.

c. *Naïve Bayes*

Naïve Bayes merupakan sebuah model klasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. *Naïve Bayes* didasarkan pada *teorema bayes* yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network.*

d. *K-Nearest Neighbor*

K-Nearest Neighbor adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Ketepatan algoritma *K-Nearest Neighbor* ini sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevasinya terhadap klasifikasi.

e. *Logistic Regression*

Regresi logostik (*Logistic Regression*) adalah bagian dari analisis regresi yang digunakan ketika variabel dependen atau respon merupakan variabel dikotomi.

Jadi dapat disimpulkan bahwa klasifikasi merupakan kegiatan yang digunakan untuk mengelompokkan data kedalam kelas-kelas tertentu dalam pengerjaan klasifikasi dapat menggunakan beberapa metode algortima seperti *Neural Network, Decision Tree, Naïve Bayes, K-Nearest Neighbor* dan *Logistic Regression*.

2.3 *NAÏVE BAYES*

Naive Bayes Classifer merupakan salah satu metode *machine learning* yang memanfaatkan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuan Inggris *Thomas Bayes* yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman sebelumnya.

Menurut Saleh. [15] menyatakan bahwa "*Naive Bayes* merupakan sebuah pengklasifikasian probalistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema *bayes* dan mengansumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. *Naive Bayes* juga didefinisikan sebagai pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuan inggis *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya".

Menurut Utomo dan Harsanto. [16] menyatakan bahwa "*Naïve Bayes* adalah sebuah teori kondisi probabilitas yang memperhitungkan probabilitas suatu kejadian (*hipotesis*) yang bergantung pada bukti dari kejadian sebelumnya".

Menurut A. A. Zainal mengatakan bahwa :

> "*Naïve Bayes Classifier* merupakan pengklasifikasi probabilitas sederhana berdasarkan pada teorema *bayes. Teorema bayes* dikombinasikan dengan *Naive* yang berarti setiap atribut atau variabel bersifat bebas (independent). *Naïve Bayes Classifier* dapat dilatih dengan efisien dalam pembelajaran terawasi (*supervised learning*)".

Dari beberapa pendapat para ahli diatas, dapat disimpulkan bahwa *Naïve Bayes* merupakan salah satu metode algoritma yang memiliki independensi yang

kuat (naif) yang digunakan untuk memprediksi kemungkinan dimasa yang akan datang melalui pengalaman dimasa lampau.

Prediksi *Bayes* didasarkan pada teorema *Bayes* dengan formula umum sebagai berikut :

$$\mathbf{P\ (Ci|\ X)} = \frac{\mathbf{P(X\ |\ Ci)\ P\ (C)}}{\mathbf{P(X)}} = \text{……..……} \quad (2.1)$$

Keterangan :

X	:	Kriteria suatu kasus berdasarkan masukan
C*i*	:	Kelas solusi pola ke-I, dimana I adalah jumlah label kelas.
P(Ci | *X*)	:	Probabilitas kemunculan label kelas Ci dengan kriteria masukan X.
P (*X* | C*i*)	:	Probabilitas kriteria masukkan *X* dengan label kelas Ci.
P(*X*)	:	Probabilitas labeb C*i*.

2.4 PENILAIAN KINERJA

Suatu organisasi perusahaan didirikan karena memiliki tujuan tertentu yang ingin dan harus dicapai. Salah satu upaya yang dilakukan perusahaan dalam rangka mencapai tujuan instansi secara keseluruhan adalah *performance appraisal* yang disebut juga penilaian prestasi kerja, penilaian pelaksanaan pekerjaan, penilaian kondite dan sebagainya. Penilaian kinerja karyawan sering diartikan sebagai pencapaian tugas, diamana karyawan yang bekerja harus sesuai dengan program kerja instansi untuk menunjukkan tingkat kinerja dalam mencapai visi, misi dan tujuan instansi tersebut.

Menurut Hanggraeni. [17] menyatakan bahwa “Penilaian kinerja adalah sebuah proses dimana perusahaan melakukan evaluasi dan penilaian kinerja individu setiap perkerjaannya”.

Menurut Sedarmayati. [18] menyatakan bahwa “Penilaian kinerja karyawan adalah sistem formal untuk memeriksa atau mengkaji dan mengevaluasi secara berkala kinerja seseorang”.

Menurut Sofyandi. [19] menyatakan bahwa “Penilaian kinerja adalah penilaian tentang prestasi kerja karyawan. Dalam persaingan global, perusahaan-perusahaan menuntut karyawannya untuk memiliki kinerja yang tinggi. Seiring dengan hal tersebut, karyawan membutuhkan *feedback* (umpan balik) atas kinerja mereka sebagai pedoman perilaku dan keputusan dimasa yang akan mendatang. Penilaian kinerja mencakup aspek kualitatif maupun kuantitatif dari pelaksanaan kerja”.

Dari beberapa definisi menurut para ahli diatas, maka dapat disimpulkan bahwa penilaian kinerja adalah sistem formal yang digunakan untuk mengevaluasi kinerja karyawan agar sesuai dengan standar kerja yang telah ditetapkan oleh perusahaan. Selain itu, juga untuk menentukan kebutuhan pelatihan kerja secara tepat, memberikan tanggung jawab yang sesuai kepada karyawan sehingga dapat melaksanakan pekerjaan yang lebih baik di masa mendatang sebagai dasar untuk menentukan kebijakan dalam hal promosi atau penentuan imbalan.

2.5 TOOLS DATA MINING

Tools Data Mining adalah sebuah *software* yang digunakan untuk mempermudah seorang peneliti, akademis dan pihak manapun dalam hal mengolah sebuah data [20]. Berikut beberapa *tools* dari *data mining* yang akan digunakan pada penelitian ini :

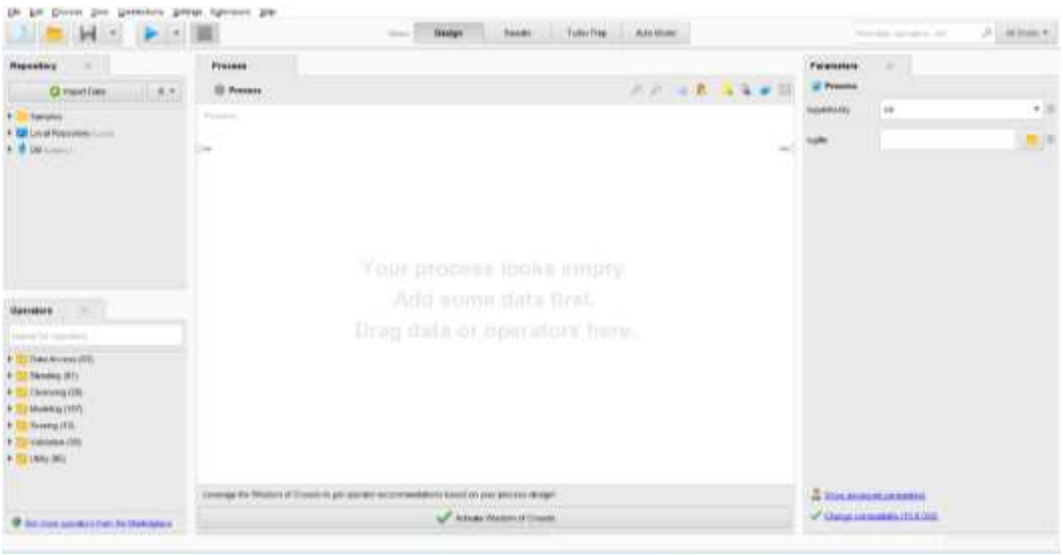
1. WEKA (*Waikato Environment For Knowledge Analysis*)

WEKA merupakan *software* terintegrasi yang berisi implementasi dari metode-metode *data mining*. *WEKA* dikembangkan oleh Universitas Wakaito, Selandia Baru menggunakan bahasa pemrograman Java. Dengan mengadopsi konsep *open source software*, menjadikan *WEKA* dapat digunakan dan dimodifikasi siapapun secara gratis. Perangkat sistem operasi *WEKA* dipergunakan dalam pembuatan aplikasi penerapan *data mining*, karena sarana akses data yang lebih cepat dan terdapat beberapa kelengkapan mengenai *tools* teknik *data mining* [21].

Gambar 2. 2 Tampilan Halaman Utama *WEKA*

2. *Rapid Miner*

Rapid Miner merupakan *software* yang bersifat *open source*. *Rapid Miner* adalah sebuah solusi untuk melakukan analisis prediksi. *Rapid Miner* menggunakan berbagai teknik deslriptif dan prediksi dalam memberikan wawasan kepada pengguna sehigga dapat membuat suatu keputusan yang baik [20].



Gambar 2. 3 Tampilan Halaman Utama *Rapid Miner*

2.6 PENELITIAN TERKAIT

Penulis memasukkan beberapa kajian dari penulis-penulis yang telah melakukan penelitian sejenis untuk menjadi bahan perbandingan ataupun acuan bagi penulis. Beberapa penelitian telah banyak dilakukan dengan menggunakan teknik *data mining* untuk menggali berbagai informasi dari sebuah *database*. Penelitian-penelitian tersebut membahas tentang topik yang terkait dengan penelitian penulis, yaitu penelitian mengenai algoritma yang akan digunakan oleh penulis :

Tabel 2. 1 Penelitian Terkait

No	Penulisan dan Tahun	Masalah	Hasil	Perbedaan
1	Luthfi Indriyani, Weko Susanto (Vol. 6 No 02, September 2019) [22]	Pada laporan pertanggung jawaban terdapat penurunan pemberian (piutang) kepada anggota koperasi sebesar 17.319.802.163, pada tahun 2015 pemberian piutang 127.866.969.180 menjadi 110.547.167.017 pada tahun 2016 di seluruh wilayah dan salah satu faktor penurunan kinerja disebabkan kredit bermasalah atau tidak tertagih sehingga modal utama mengendap pada anggota yang tidak tertagih di Koperasi Keluarga Guru Jakarta pada piutang 2015 dan 2016	Pengujian pada data rekapitulasi peminjam yang berjumlah 450 data dari Koperasi Keluarga Guru Jakarta dengan proses mining algoritma naïve bayes menghasilkan tingkat akurasi 84,00%, dimana dalam pengujian model data, keseluruhan data set digunakan sebagai data training.	Tidak menggunakan *tools Data Mining* untuk meng implementasikan metode *naïve bayes*.
2	Ulfa Pauziah	Di dalam dunia pekerjaan adanya	Dari hasil perhitungan	Atribut dan proses penilaian yang

	Erna (Vol. 1 No 1, 2017) [3]	karyawan terbaik menjadi tolak ukur kemajuan dari perusahaan itu sendiri. Dalam penentuan biasanya dengan melihat kinerja karyawan tersebut misal dari kerajinan, kedisiplinan dan juga prestasi lainnya. Dengan cara seperti ini agak kurang efektif dan akurat.	algoritma *naïve bayes* meng-gunakan *tools* WEKA didapat hasil 98,5714% dapat membantu pengambilan keputusan karyawan terbaik.	berbeda serta jumlah data karyawan berbeda.
3	Viny Novika Sari (Vol. 2 No 1, Maret 2020) [5]	Pada PT. PELITA WIRA SEJAHTERA penilaian terkadang dilakukan secara subjektif dan keterbatasan dalam mengontrol setiap karyawan yang bekerja. Oleh karena itu penulis melakukan analisis *data mining* pada data-data penilaian karyawan tersebut agar dapat mengetahui mana karyawan yang memiliki kinerja yang sangat baik, baik,	Presentai akurasi terbesar diperoleh dengan meng-gunakan *Use Training Set Correctly* sebasar 95,302%, meng-gunakan 5-*Fold Cross Validation Correctly* sebesar 93,9597% dan 10-*Fold Cross Validation Correctly* sebesar 93,9597%.	Objek dan jumlah data karyawan pada penelitian yang digunakan.

		cukup, dan kurang. Penulis meng- gunakan data-data karyawan sebanyak 149 data yang kemudian disajikan kedalam format arff.		
4	Sri W. Utami, Ahmad A. Supianto, Fitra A. Bcahtiar (Vol. 3 No. 6, Juni 2019) [23]	Pengajaran yang baik dapat membantu mahasiswa dalam mencapai hasil yang maksimal. Untuk me- ningkatkan kualitas pem- belajaran dan stan- darisasi akademik perlu dilakukan evaluasi sehingga dapat menghasilkan mahasiswa-mahasiswa yang berkualitas. Oleh karena itu, Jurusan Sistem informasi selalu melakukan evaluasi terhadap kinerja menggunakan kuisioner yang diisikan oleh mahasiswa disetiap akhir semester. Hasil kolom saran dapat dilakukan analisis sentimen untuk mengetahui saran	Hasil pengujian terhadap 4 parameter men- ghasilkan akurasi sebesar 80,1%, *precision* 80,3%, *recall* 80,3%. Hasil dari *Usability testing* diperoleh nilai rata-rata kedalam kategori *Acceptance* dan pada rating "Good".	Persentasi akurasi pada penelitian ini dibawah 90%

		tersebut bernilai positif, negatif atau netral.		
5	Ikhsan Romli (Vol. 1 No. 2, November 2020) [24]	PT. Berkat Sinar Sentosa merupakan perusahaan yang bergerak di bidang penyediaan jasa security, office boy dan driver. Namun sangat disayangkan karena perusahaan masih manual dalam mengevaluasi penilaian kinerja security, sehingga pengambilan keputusan dalam menentukan security belum bisa dikatakan kompeten dan tidak kompeten. menjadi efektif dan efisien.	Data yang digunakan dalam penelitian ini menggunakan 63 data latih tanpa menggunakan data uji, namun masih berupa data sampel uji sebagai acuan perhitungan naïve bayes. Tes ini dilakukan dengan menggunakan algoritma data mining Naïve Bayes dan Rapid Miner Framework. Dari perhitungan menggunakan algoritma Naïve Bayes diperoleh akurasi sebesar 85,71%, presisi sebesar 88,24% dan recall sebesar 85,71%.	Penelitian ini hanya memaparkan hasil persentasi menggunakan *tools Rapid Miner*.

Dari beberapa penelitian pada tabel 2.1, dapat disimpulkan bahwa metode *Naïve Bayes* memiliki akurasi yang baik dalam pengkalsifikasian. Maka dari itu dalam penelitian ini penulis menggunakan metode *Naïve Bayes* untuk klasifikasi data evaluasi penilaian kinerja karyawan PT. Rimba Hutani Mas berdasarkan kategori yang telah ditentukan oleh penulis. Penelitian-penelitian yang telah dilakukan sebelumnya ini akan menjadi acuan penulis dalam melakukan penelitian.

Berdasarkan dari pemaparan penelitian sebelumnya, dapat dilihat persamaan adalah memiliki tujuan yang sama yaitu mengenai evaluasi penilaian kinerja karyawan serta metode yang diguakan untuk mengetahui tingkatan kinerja karyawan selama bekerja di instansi tersebut. Sedangkan perbedaan penelitian ini dengan penelitian sebelumnya yaitu penulis mengolah data hasil evaluasi kinerja karyawan PT. Rimba Hutani Mas dengan cara mengklasifikasikan data karyawan di tahun 2022 dan atribut dalam pengklasifikasian data pada penelitian ini berbeda dari atribut penelitian sebelumnya. Hasil diperoleh bertujuan untuk mengetahui kinerja karyawan selama karyawan bekerja di perusahaan PT. Rimba Hutani Mas.