BAB II

LANDASAN TEORI

2.1 Teori yang digunakan

Teori teori pendukung yang digunakan dalam penelitian ini adalah sebagai berikut:

2.1.1. Data Mining

Data mining adalah proses menganalisis sejumlah besar data untuk mengidentifikasi pola yang bermakna dan mendeteksi hubungan, yang dapat mengarah pada prediksi tren masa depan dan pengambilan keputusan yang tepat. Aplikasi data mining sangat penting dalam beberapa bidang industri seperti bisnis pemasaran, perbankan maupun kedokteran [16]. *Data mining* merupakan cara untuk mengekstraksi pola dan pengetahuan yang menarik dari sejumlah besar informasi. Penambangan data memiliki proses seperti kombinasi dari memilih, menganalisis, merencanakan, menafsirkan, dan mengevaluasi hasil [17]. *Data mining* didefinisikan sebagai praktik dalam memeriksa basis data besar yang sudah ada sebelumnya untuk menghasilkan informasi baru [18]. *Data mining* sendiri mengacu pada proses mengekstraksi informasi atau pengetahuan yang ada yang berasal dari proses *Knowledge Data Discovery (KDD)* di mana algoritma diterapkan untuk mengekstraksi pengetahuan [19]. Penelitian ini, menyajikan gambaran aplikasi data mining dalam bidang kedokteran untuk memberikan pandangan yang jelas tentang tantangan dan pekerjaan sebelumnya di bidang ini bagi para peneliti [16].

2.1.1.1 Teknik *Data Mining*

Teknik *data mining* adalah proses mengidentifikasi pola dan tren data untuk mendapatkan informasi yang berguna dalam kumpulan data yang sangat besar sehingga dapat menilai atau memutuskan [18]. Teknik *data mining* proses menganalisis volume data untuk menemukan pola, menemukan tren, dan mendapatkan wawasan tentang bagaimana data tersebut dapat digunakan [20]. Proses dalam teknik data mining

diklasifikasikan menjadi deskriptif dan prediktif. Berikut ini adalah penjelasan masing masing teknik *data mining* [21];

A. *Classification*

Klasifikasi merupakan proses menemukan model yang menggambarkan dan membedakan data kelas dan konsep. Ini juga merupakan proses memetakan atau mengklasifikasikan data ke dalam salah satu dari beberapa kelas yang telah ditentukan [21].

B. *Regression*

Analisis regresi adalah proses ekstraksi data di mana hubungan antar variabel didefinisikan dan dianalisis. Ini digunakan untuk mengevaluasi kemungkinan variabel tertentu, karena ada variabel lain [21].

C. *Clustering*

Clustering merupakan salah satu bentuk ekstraksi data untuk mengklasifikasikan data yang terkait. Metode ini membantu untuk mempertimbangkan kesenjangan antara data dan kesamaan [21].

D. *Association rule*

Teknik ekstraksi data ini membantu menemukan hubungan antara dua objek atau lebih. Ini juga dikenal sebagai teknik relasi karena menggunakan hubungan antara item data [21].

E. *Outer Detection*

Teknik *data mining* ini mengacu pada pengamatan elemen data yang tidak sesuai dengan diprediksi pola perilaku dalam pengumpulan data [21].

F. *Sequential Patterns*

Teknik *data mining* ini membantu mendeteksi atau menemukan pola atau tren serupa dalam transaksi data selama jangka waktu tertentu [21].

G. *Prediction*

Prediksi menggunakan banyak teknik lain untuk penambangan data. Ini menganalisis peristiwa atau keadaan masa lalu dengan urutan yang benar untuk memprediksi kejadian masa depan [21].

2.1.2. Algoritma Naïve Bayes

Naive Bayes adalah algoritma yang dapat mengklasifikasikan variabel tertentu dengan menggunakan teknik probabilistik dan statistik [22]. *Naive Bayes* memodelkan atribut numerik menggunakan distribusi normal. *Naïve Bayes* dapat menggunakan estimator kepadatan kernel. Ini meningkatkan kinerja ketika asumsi normalitas benar. *Naïve Bayes Updateable* adalah versi inkremental, memproses satu permintaan dalam satu waktu. Estimator kernel dapat digunakan dalam versi ini, tetapi tidak diskritisasi [23]. Algoritma ini biasa dikenal sebagai algoritma dengan perhitungan yang sederhana, apabila digunakan sebagai klasifikasi maka algoritma ini disebut *Naïve bayes classifier.* Persamaan matematika untuk perhitungan akurasi terdapat pada persamaan (1).

$$p\ (H|E) : \frac{p(H|E)\ x\ p(H)}{p(E)} \qquad (1)$$

Dimana penjelasan dari rumus diatas yaitu [24]:

(a) p(H|E) berarti probabilitas bahwa hipotesis H dapat terjadi ketika bukti E hadir;
(b) p(E|H) menunjukkan probabilitas terjadinya bukti E jika hipotesis H hadir;
(c) p(H) menunjukkan probabilitas terlepas dari hipotesis H tidak ada bukti;
(d) p(E) mewakili probabilitas meskipun ada bukti E.

2.1.3. Split Data

Split data berarti membagi data secara acak menjadi dua bagian, satu sebagai data latih dan satu lagi sebagai data uji. Dengan validasi terpisah, uji coba pelatihan dilakukan berdasarkan rasio distribusi yang telah ditentukan, setelah itu rasio distribusi data pelatihan lainnya dianggap sebagai data uji [25]. Untuk penanganan data yang tidak seimbang, pemilihan proporsi rasio harus dilakukan dengan hati-hati untuk memastikan bahwa kelas minoritas tidak terlalu sedikit dalam training set atau testing set sehingga model tidak memiliki representasi yang cukup untuk kelas tersebut. Jika dataset tidak seimbang, proporsi pembagian dataset yang direkomendasikan mungkin bervariasi dari 70:30, 60:40 atau bahkan 50:50 [26].

2.1.4. Kurva ROC

Kurva ROC adalah visualisasi yang mengevaluasi hasil prediksi data menggunakan 2 kelas sebagai keputusan, TP (*True Positive*) adalah pada sumbu Y, sedangkan FP (*False Positive*) adalah pada sumbu X. Ini merupakan salah satu *tools* dalam RapidMiner yang digunakan untuk membandingkan kinerja beberapa algoritma dalam klasifikasi kasus dalam penelitian ini, rentang nilai ROCs adalah 0-1 untuk perbandingan [24].

Tabel 2.1. Rentang nilai ROC [24]

Rentang Akurasi	Kualitas
0.90 – 1.00	*Excellent*
0.80 – 0.90	*Good*
0.70 – 0.80	*Fair*
0.60 – 0.70	*Poor*
0.50 – 0.60	*Failure*

2.1.5. Confusion Matrix

Confusion matrix sendiri adalah merupakan metode dengan bentuk atau berupa table yang biasa digunakan dalam perhitungan akurasi data mining [27] Mereka digunakan untuk mengevaluasi kinerja model

klasifikasi dan untuk menentukan kualitas prediksi yang dibuat oleh model. Terdapat 3 metode pengujian dalam *confusion matrix* yaitu [24]:

A. Akurasi

Akurasi merupakan salah satu dari metode untuk menguji suatu algoritma berdasarkan kedekatan antara nilai prediksi dan nilai sebenarnya. Persamaan matematika untuk perhitungan akurasi terdapat pada persamaan (2) [24]:

$$Accuracy : \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (2)$$

B. Sensitivitas

Sensitivitas atau *recall* adalah metode untuk menguji algoritma dengan membandingkan data benar yang diperoleh sistem dengan data benar yang dipulihkan atau ditinggalkan system. Persamaan matematika untuk perhitungan sensitivitas terdapat pada persamaan (3) [24]:

$$Recall : \frac{TP}{TP+FN} * 100\% \qquad (3)$$

C. Presisi

Presisi merupakan salah satu metode pengujian algoritma yang membandingkan data yang benar dari sistem dengan total data yang dikumpulkan oleh sistem (benar dan salah). Persamaan matematika untuk perhitungan presisi terdapat pada persamaan (4) [24]:

$$Precision : \frac{TP}{TP+FP} * 100\% \quad (4)$$

2.1.6. SMOTE Upsampling

Synthetic Minority Oversampling Technique (SMOTE) merupakan salah satu teknik *oversampling*, yaitu teknik pengambilan sampel yang memperbanyak jumlah data kelas positif dengan mengalikan secara acak jumlah data kelas positif sehingga jumlahnya sama dengan kelas data positif

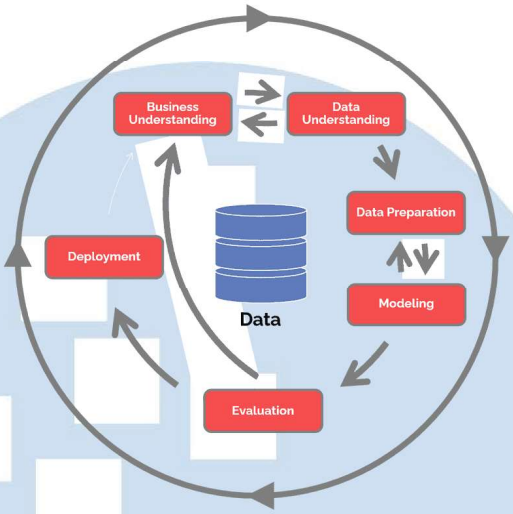
[28]. Pendekatan ini bekerja dengan membuat data "sintetik", yaitu data yang disalin dari data jarang. Metode SMOTE bekerja dengan cara mencari k tetangga terdekat (tetangga dari data). Teknik ini membuat data pelatihan tambahan dengan melakukan operasi tertentu pada data asli [29]. Menggunakan SMOTE untuk data yang tidak seimbang penting karena dapat membantu menyeimbangkan sensitivitas dan spesifisitas model, yang menghasilkan akurasi prediksi yang lebih baik [30]. Dengan demikian, salah satu kelebihan metode ini adalah tidak menyebabkan hilangnya data karena data tidak tereduksi seperti yang dilakukan pada metode downsampling [31].

2.2 Framework Data Mining

Terdapat tiga *framework* yang yang dapat digunakan sebagai pemandu dalam pelaksanaan proses data mining. Ketiga *framework data mining* yang paling populer adalah Model *Knowledge Discovery in Database* (KDD), CRISP-DM dan SEMMA [32].

2.2.1. CRISP-DM

*Cross-Industry Standard Process for Data Minin*g atau biasa disingkat CRISP-DM ini adalah model proses penambangan data yang banyak digunakan yang menyediakan pendekatan terstruktur untuk merencanakan, melaksanakan, dan mengevaluasi proyek penambangan data. Model CRISP-DM fleksibel dan dapat disesuaikan dengan berbagai jenis proyek dan industri penambangan data. Model ini terdiri dari enam fase: *business understanding*, *data understanding, data preparation, modelling, evaluation,* dan *deployment* [33].

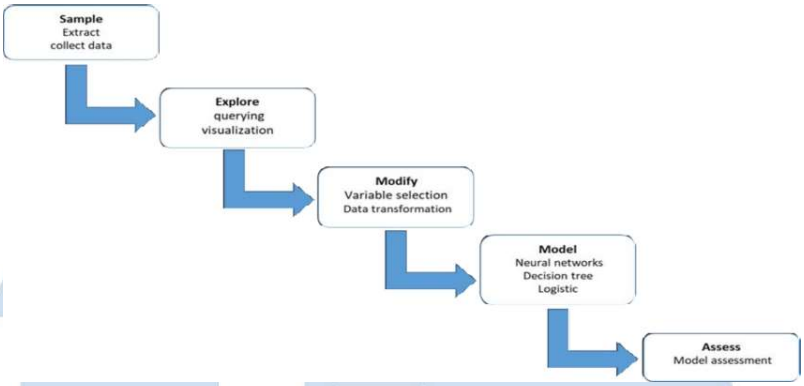


Gambar 2.1 Cross-Industry Standard Process for Data Mining [34]

Gambar 2.1 merupakan gambar proses penelitian pada Crisp-DM, dimulai dari b*usiness understanding* - tahap pemahaman domain (penelitian). *Data understanding* - tahap pemahaman data adalah langkah pengumpulan data mentah, eksplorasi data untuk mengidentifikasi data penggunaan. *Data preparation* - tahap persiapan data, fase ini sering disebut padat karya. *Modelling* - tahap pemodelan merupakan tahapan penentuan teknik *data mining* yang akan digunakan, *tools data mining*, algoritma *data mining*, parameter dengan nilai optimal. *Evaluation* - tahap ini adalah tahap interpretasi. hasil data mining disajikan dalam proses pemodelan dari langkah sebelumnya. *Deployment* - tahap implementasi adalah tahap di mana laporan atau presentasi data yang diperoleh selama evaluasi proses *data mining* disiapkan [32].

2.2.2. SEMMA

SEMMA (*Sample, Explore, Modify, Model, Assess*), SAS Institute yang mengembangkan model tersebut, menjelaskan bahwa bukan metode *data mining*, melainkan seperangkat alat untuk melaksanakan tugas inti dari data mining [35]. Metode ini terdiri atas tahap yang merupakan singkatan SEMMA itu sendiri.



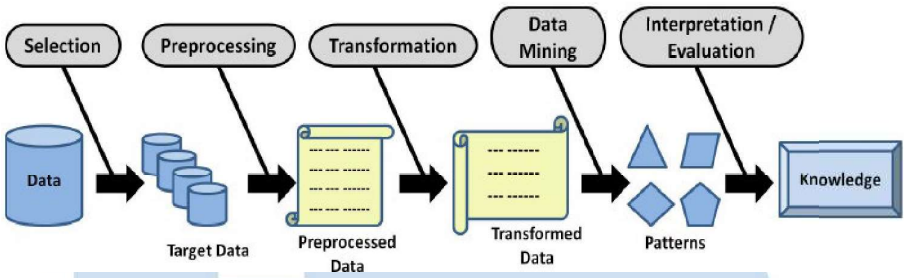
Gambar 2.2 Proses pada framework SEMMA [36]

Sample - Langkah ini terdiri dari pengambilan sampel data melalui ekstraksi data ukurannya cukup besar untuk memuat informasi penting. *Explore* - Fase itu terdiri dari memeriksa data untuk mencari tren yang tidak terduga dan anomali yang dalam untuk pemahaman dan gagasan. *Modify* - Langkah mengedit data membuat, memilih dan memodifikasi variabel untuk memusatkan proses pemilihan model. Model - Langkah ini terdiri dari pemodelan data menggunakan perangkat lunak pencarian secara otomatis menggabungkan data yang andal untuk memprediksi hasil ini diinginkan. *Asses* - Langkah ini terdiri dari mengevaluasi data dengan menilai kegunaannya dan keandalan hasil proses *data mining* [32].

2.2.2. Knowledge Discovery in Database

Knowledge discovery in database atau KDD, umumnya dikenal sebagai penambangan data, adalah proses penemuan pola dan pemodelan prediktif dalam basis data besar. KDD memanfaatkan secara ekstensif metode penambangan data, proses otomatis, dan algoritma yang memungkinkan pengenalan pola [37]. Hasil proses KDD adalah untuk mengekstrak pengetahuan dari data dalam konteks basis data yang besar. Ini merupakan teknik data mining yang banyak digunakan karena didalamnya terdapat proses yang meliputi persiapan dan pemilihan data, pembersihan data, penggabungan pengetahuan sebelumnya tentang kumpulan data dan

interpretasi yang akurat solusi dari hasil pengamatan [38].



Gambar 2.3 ***Knowledge discovery in database*** [39]

Gambar 2.3 merupakan gambaran daripada proses KDD, yang kemudian proses ini akan dijelaskan atau diterapkan pada *software data mining* yang digunakan yaitu *rapidminer.* Penjelasan setiap langkah pada gambar diatas adalah sebagai berikut [40]:

A. *Data selection*

Data selection merupakan proses pertama yang perlu dilakukan yaitu pemilihan atau seleksi data sebelum proses KDD dimulai. Hasil data yang telah diseleksi kemudian digunakan untuk proses data mining, data tersebut disimpan dalam suatu berkas yang terpisah dari basis data operasional [40]. Operator yang digunakan adalah operator *select attributes*, dimana hanya atribut yang dipilih yang dikirim ke *port output*. Sisanya dihapus dari *ExampleSet* [41].

B. *Preprocessing*

Preprocessing merupakan tahap dimana data dipersiapkan dan melalui proses *cleansing,* proses ini antara lain menghapus duplikat data, periksa data yang tidak konsisten, dan perbaiki kesalahan data. Proses pengayaan juga dilakukan. Artinya, proses "memperkaya" data yang ada dengan data atau informasi lain yang dibutuhkan dalam kaitannya dengan *knowledge discovery in database* (KDD) dalam basis data [40]. Operator yang digunakan, diantaranya:

a) Operator Trim

Operator trim membuat atribut baru dari atribut nominal yang dipilih dengan menghapus spasi awal dan akhir dari nilai nominal [41].

b) Operator Replace Missing Value

Operator ini mengganti nilai yang hilang dalam Contoh atribut yang dipilih dengan pengganti yang ditentukan. Nilai yang hilang dapat diganti dengan nilai minimum, maksimum, atau rata-rata dari atribut tersebut [41]. Dengan mengganti data yang hilang dengan nilai yang masuk akal berdasarkan data yang diamati, pendekatan ini memungkinkan model analisis dapat meningkatkan akurasi dan presisi [42].

C. *Transformation*

Transformation merupakan proses mengubah data yang dipilih menjadi sesuatu yang cocok untuk proses penambangan data. Dalam *Knowledge Discovery in Database* (KDD) proses ini merupakan proses yang penting juga kreatif tetapi sangat bergantung pada jenis atau pola informasi yang ingin dicari dalam data [40]. Operator yang digunakan adalah *generates attributes*, Operator ini membangun atribut yang ditentukan pengguna baru menggunakan ekspresi matematika [41].

D. *Data Mining*

Data mining adalah proses pencarian pola dan informasi yang menarik pada data terpilih dengan menggunakan teknik dan metode tertentu. Baik teknik, metode, ataupun algoritma dalam data mining sangatlah bervariasi. Memilih metode atau algoritma yang tepat sangat tergantung pada tujuan keseluruhan dan proses penemuan pengetahuan dalam basis data (KDD) [40]. Operator yang digunakan, diantaranya:

a) Operator Set Role

Operator ini digunakan untuk mengubah peran dari satu atau lebih Atribut. Peran Atribut menjelaskan bagaimana Operator lain menangani Atribut ini. Peran *defaultnya regular*, peran lain

diklasifikasikan sebagai khusus. Berbagai jenis peran dijelaskan di bawah di bagian parameter [41].

b) Operator Naïve Bayes (kernel)

Operator ini menghasilkan model klasifikasi kernel *naïve bayes* menggunakan kepadatan kernel yang diperkirakan [41].

c) Operator Split Data

Operator ini menghasilkan jumlah himpunan bagian yang diinginkan dari *ExampleSet* yang diberikan. *ExampleSet* dipartisi menjadi subset sesuai dengan ukuran relatif yang ditentukan [41].

d) Operator Apply Model

Operator ini menerapkan model pada ExampleSet. Model pertama kali dilatih pada *ExampleSet* oleh operator lain, yang seringkali merupakan algoritma pembelajaran [41].

e) Operator SMOTE Upsampling

Operator ini menerapkan Teknik Over-sampling Minoritas Sintetis [41].

f) Operator Performance (Binomial Classification)

Operator ini digunakan untuk mengevaluasi kekuatan dan kelemahan klasifikasi biner secara statistik, setelah model terlatih diterapkan pada data berlabel [41].

E. *Interpretation / Evaluation*

Pola informasi yang dihasilkan dari proses data mining harus ditampilkan dalam format yang mudah dipahami dan dikemas secara menarik. Tahapan ini merupakan bagian dari proses penemuan pengetahuan di database (KDD) yang disebut dengan Interpretasi. Fase ini harus melihat apakah ada pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya [40].

2.3 Software Data Mining

Software atau *tools* yang digunakan pada *data mining* beragam, berikut merupakan 3 penjelasan dan perbandingan terhadap *software data mining* yang populer untuk digunakan, yaitu:

2.3.1. Software KNIME

KNIME merupakan, yang memungkinkan visual yang mudah perakitan dan eksekusi interaktif dari pipa data. Ini adalah sebuah alat analisis yang sangat kuat untuk mengekstraksi pengetahuan baru dari data yang tersedia. KNIME dirancang sebagai platform pengajaran dan penelitian, yang memungkinkan integrasi berbagai algoritma dan alat dibentuk node baru [43]. KNIME seperti kebanyakan data mining alat adalah alat grafis berisi lebih dari 1000 node yang terhubung satu sama lain dan melakukan algoritma penambangan data [44].

2.3.2. Software WEKA

WEKA adalah kumpulan algoritma pembelajaran mesin untuk tugas penambangan data. Algoritma dapat diterapkan langsung ke kumpulan data atau dipanggil dari kode Java sendiri. WEKA sangat dekat dengan KNIME karena banyak fitur bawaannya yang tidak memerlukan pengetahuan pemrograman atau pengkodean [43]. Weka adalah alat yang kuat untuk penambangan data dan pembelajaran mesin yang termasuk sempurna kumpulan alat data preprocessing dan algoritma pembelajaran mesin. Apalagi WEKA bisa terapkan beberapa pelajar ke data dan bandingkan dan evaluasi kinerja mereka untuk memilih pelajar terbaik untuk prediksi [44].

2.3.3. Software RapidMiner

RapidMiner merupakan *software* untuk penambangan data yang dapat digunakan sebagai kerangka kerja yang berdiri sendiri untuk analisis data atau disematkan ke dalam perangkat lunak lain sebagai alat penambangan data [43]. RapidMiner adalah alat yang ampuh untuk integrasi data, *Extract Transform Load* (ETL), dan analisis data, *RapidMiner* memiliki *Graphical*

User Interface (GUI) yang sangat efektif untuk desain proses analitik. Ini berisi berbagai repositori untuk proses, operator, data dan membantu dalam manajemen metadata. Ini membantu dalam memperbaiki *bug* dan deteksi kesalahan [45].

Tabel 2.2 Perbandingan *Tools Data Mining* [45][43]

Performance	RapidMiner	KNIME	WEKA
Description	*Software* yang menyediakan terintegrasi lingkungan untuk mesin pembelajaran, penggalian data, penambangan teks, prediktif analitik dan bisnis analitik	*Open source data analytics*, *reporting and integration platform*	*Popular suite* dari pembelajaran mesin perangkat lunak
Programming Language	JAVA	JAVA	JAVA
Launch date	2001	2006	2002
Development Team	Rapid-I Foundation	Silicon Valley software company	University Of Waikato
Kelebihan	- *Free Community Edition Commercial Enterprise Edition* - Meliputi analisis statistik dan prediktif yang mudah diimplementasikan ke dalam sistem - Fitur paling algoritmik -Memiliki antarmuka pengguna yang lebih menarik dan grafis (GUI / *Graphical User Interface*). - mampu melakukan operasi parametrik dalam pembelajaran mesin/metode statistik - mampu memvalidasi model	- plug-in yang mudah digunakan - menyediakan proses aliran data dengan menyeret dan menjatuhkan node baru - meng - integrasikan semua modul analisis - memiliki kemampuan untuk berinteraksi dengan program yang memungkinkan visualisasi dan analisis data.	- Software WEKA memiliki lisensi open source, sehingga gratis - Dapat berjalan dibeberapa platform, sehingga portabel - Memiliki antarmuka pengguna grafis yang mudah dipahami untuk pengguna biasa - Dapat diimplementas

Performance	RapidMiner	KNIME	WEKA
	dengan set validasi independen dengan validasi silang.		ikan dalam bahasa pemrograman Java - Tingkat pengguna Weka bisa dari pemula bahkan untuk ahli berkat banyak fungsi bawaan.
Kekurangan	- Lisensi Komunitas RapidMiner diperlukan untuk menggunakan aplikasi. - Memiliki lebih banyak pemrograman/pengkodean.	- Tidak dapat menginput banyak data untuk diproses (overload) - Tidak dapat menyimpan parameter yang dapat diterapkan pada dataset mendatang. - Tidak secara otomatis melakukan fungsi parameter pembelajaran mesin / metode statistik - Proses tidak dapat disimpan selama validasi model dengan validasi silang, sehingga diperlukan rekonstruksi model.	- Opsi yang kurang cocok untuk alur kerja kompleks yang besar - kemampuan partisi terbatas untuk dataset - hanya memiliki metode pengukuran kesalahan yang terbatas - tidak memiliki metode pembungkus untuk pemilihan descriptor - Tidak memiliki fasilitas otomatis untuk optimasi parameter metode *machine learning*.

Mengacu pada Tabel 2.2 pemilihan penggunan *software RapidMiner* dikarenakan, *software* ini memungkinkan penggunanya untuk melakukan pemrosesan data dalam jumlah yang banyak atau alur yang kompleks, didukung oleh fitur yang lebih mudah diimplementasikan ke dalam sistem,

dan kemudahan dalam memperbaiki kesalahan karena notifikasi kesalahan yang jelas dan terarah untuk segera diperbaiki beserta opsi pembetulannya.

2.4 Penelitian Terdahulu

Berikut adalah perbandingan penelitian terdahulu yang berisikan rangkuman dari bahasan mengenai penanganan *imbalance data* pada penggunaan algoritma *naïve bayes*:

Tabel 2.3 Penelitian Terdahulu

Penulis	Jurnal	Objek	Dataset	Metode	Hasil	Saran
Sajana Tiruveedhula, Mandarama narasingarao.	International Journal of Engineering & Technology, 7 (2.7) (2018) 786-790	Ketidakseimbangan pada data penyakit malaria	- 165 data pasien dari bangsal medis Narasaraopet - 15 atribut	- *Naïve bayes* - SMOTE - *Confusion matrix*	SMOTE terbukti bekerja dengan baik dan WEKA memiliki akurasi tertinggi sebesar 88,5% dibandingkan akurasi menggunakan bahasa pemrograman R sebesar 87,5%	Penanganan sampel kelas minoritas dan klasifikasinya menjadi isu panas di bidang medis. Penggunaan data medis lainnya diharapkan dapat dilakukan pada penelitian selanjutnya.
Yoga Pristyanto	Jurnal Teknoinfo, Vol. 13, No. 1, 2019	Peningkatan kinerja algoritma pada data yang tidak seimbang	- Data berasal dari UCI Machine Learning Repository. - Dataset memiliki 403 instance, dengan 5 atribut fitur dan 1 kelas atribut	- *Support vector machine* - *Naïve bayes* - *Decision tree* - *Adaptive boosting* - *Confusion matrix*	Kinerja dihasilkan lebih baik dengan menggunakan *adaptive boosting*. Pada *naïve bayes* didapatkan hasil akurasi dan sensitivitas sebesar 91.98, spesifitas sebesar	Dapat dilakukan penelitian khususnya pada metode pengujian yang menggunakan metode umum lainnya.

Penulis	Jurnal	Objek	Dataset	Metode	Hasil	Saran
					96.49, dan G-mean sebesar 94.21.	
Afifah Ratna Safitri, Much Aziz Muslim	Journal of Soft Computing Exploration , Vol. 1, No. 1, September 2020	Peningkatan Akurasi Pengklasifikasi Naive Bayes untuk Penentuan Pelanggan Churn Menggunakan SMOTE dan Algoritma Genetika	- Data berasal dari UCI Machine Learning Repository - terdiri atas 1000 *instance* dan 20 atribut. - 13 tipe atribut nominal dan 7 numerik.	- *Naïve Bayes* - *Confussion Matrix* - *SMOTE*	Hasil akurasi yang didapatkan pada penerapan algoritma Naive Bayes tanpa proses pre-processing yaitu sebesar 73%. Sedangkan akurasi rata-rata dari sepuluh eksekusi yang diperoleh dengan menggunakan algoritma SMOTE pada algoritma Naive Bayes adalah 74,918% dan hasil rata-rata akurasi dari sepuluh eksekusi yang diperoleh dengan menggunakan algoritma SMOTE dan pemilihan atribut dari	- Oleh karena itu perlu dilakukan penyeimbangan kelas dengan membuat data baru pada kelas churn. Penggunaan data dengan *instance* lebih banyak patut dipertimbangkan untuk dicoba.

Penulis	Jurnal	Objek	Dataset	Metode	Hasil	Saran
					Algoritma Genetika dari algoritma Naive Bayes adalah 80,948%.	
Reza Dwi Fitriani, Hasbi Yasin, Tarno	JURNAL GAUSSIAN, Volume 10, Nomor 1, Tahun 2021	Ketidaksembangan data pada studi kasus peserta KB IUD di kabupaten kendal	- Data yang digunakan dalam penelitian ini adalah informasi status kontrasepsi dalam kandungan (IUD) peserta KB di Kabupaten Kendal tahun 2018. - Data diambil dari responden yang memasang IUD sejak tahun 2015, berjumlah 250 pasien dan memiliki 15 variabel.	- *Naïve bayes* - *Feature Selection* - *Confusion matrix* - *Cross Validation* - *Random Oversampling*	Menghasilkan nilai g-mean naïve bayes kurang dari 60%. G-mean dari ROS-naïve bayes adalah 96,6%, dengan nilai akurasi 0,938, sensitivitas 1,0, dan spesifisitas 0,934. Dapat disimpulkan bahwa pada penelitian ini metode ROS-naïve bayes bekerja dengan baik	Jika peneliti menginginkan pelajaran prediksi kegagalan IUD yang lebih akurat di Kabupaten Kendal sebaiknya menggunakan metode ROS-Naive Bayesian. Alternatif lain adalah dengan menggunakan algoritma atau metode optimasi lain.
Noviyanti Santoso, Wahyu Wibowo, Hilda Himawati	Indonesian Journal of Electrical Engineering and Computer Science Vol. 13, No. 1, January 2019	Integrasi SMOTE untuk ketidaksembangan kelas	Penelitian ini menggunakan dataset Bank Marketing dari UCI Machine Learning. Dari 45210 instance diambil	- *Naïve bayes* - *Support vector machine* - *Random forest* - *SMOTE* - *Confusion matrix*	Penggunaan SMOTE pada algoritma *naïve bayes* menghasilkan akurasi sebesar 83.68%, dan berhasil	Untuk penelitian selanjutnya, saran-saran berikut dapat dipertimbangkan; menggabungkan metode

Penulis	Jurnal	Objek	Dataset	Metode	Hasil	Saran
			sampel sebanyak 10% secara acak sehingga jumlah instance yang digunakan adalah 4521.	- *ROC Curve*	meningkatkannilai AUC 76.19%, nilai f-measure 60.26%, dan nilai g-means 75.04%.	*resampling* lain seperti tautan *tomek links* dan *random under-sampling*; Mengguna kan lebih banyak sampel dataset yang tidak seimbang dengan distribusi kelas yang berbeda.
Nina Sulistiyowati, Mohamad Jajuli	Jurnal Nuansa Informatika, Vol 14 No 1, Januari 2020	Menangani data tidak seimbang dengan integrasi SMOTE	- Data diperoleh dari Guru Rawamerta data nasabah kredit pada tahun 2015-2017, dengan total data 878. - Berjumlah 7 atribut.	- *Naïve bayes* -*Confusion matrix* - *G-mean* - SMOTE	Dua skenario penelitian dilakukan, skenario pertama melakukan klasifikasi naive bayes pada dataset asli, sedangkan skenario kedua melakukan klasifikasi naive bayes pada dataset yang dibuat dengan metode SMOTE. secara keseluruhan metode smote mampu menangani permasalahan imbalance data	Pengujian dapat dilakukan dengan beberapa algoritma klasifikasi dikombinasikan dengan SMOTE dan dapat diuji coba dengan variasi nilai *k-fold cross validation*

Penulis	Jurnal	Objek	Dataset	Metode	Hasil	Saran
					dengan tingkat akurasi 94.015% dan G-mean 0.948%	

Tabel 2.3 menunjukkan penelitian terdahulu yang menjadi acuan dalam penelitian ini, dimana 4 penelitian terdahulu [9],[13],[15],[46] yang menunjukkan bahwa metode pengujian untuk menangani *imbalance data* seperti *adaptive booster*, SMOTE, dan *random over sampling*, yang digunakan terbukti baik dalam meningkatkan kinerja dan mengatasi ketidakseimbangan data, dengan rata rata kinerja diatas 80%. Pada penelitian terdahulu terlihat bahwa, dari 2 lain yang menggunakan penelitian menggunakan metode SMOTE [14] dan [47] keduanya menunjukan peningkatan yang signifikan pada metrik lain selain akurasi. Penelitian [14] menunjukan peningkatan pada metrik AUC, f-measure, g-means, sedangkan penelitian [47] peningkatan pada g-mean.