BAB 2
TINJAUAN PUSTAKA

2.1 Penelitian Perbandingan Algoritma Klasifikasi

Beberapa penelitian terkait perbandingan algoritma klasifikasi pada bidang medis telah dilakukan sebelumnya. Beberapa hasil penelitian terkait ditampilkan pada Tabel 2.1.

Tabel 2.1. Penelitian perbandingan algoritma klasifikasi pada bidang medis

No	Peneliti dan Tahun	Judul	Algoritma *(methods)*	Hasil *(results)*
1	(Jasri et al., 2022)	Penerapan Data Mining untuk Klasifikasi Penyakit Demam Berdarah Dengue (DBD) Dengan Metode Naïve Bayes (Studi Kasus Puskesmas Taman Krocok)	Naïve Bayes	Setelah serangkaian proses diatas dilakukan maka didapat tingkat akurasi untuk model klasifikasi Naïve Bayes didapat 92%
2	(Sastrawan et al., 2019)	Perbandingan Kinerja Algoritma Dempster Shafer dan Fuzzy-Naive Bayes Dalam Klasifikasi Penyakit Demam Berdarah dan Tifus	Algoritma Dempster Shafer dan Fuzzy-Naive Bayes	hasil klasifikasi digunakan confution matriks. Dengan k = 4 pada metode k-fold cross validation, dan perhitungan akurasi, presisi, dan recall pada confusion matriks didapat bahwa nilai akurasi menggunakan metode Dempster Shafer lebih besar daripada metode Fuzzy Naive Bayes

No	Peneliti dan Tahun	Judul	Algoritma *(methods)*	Hasil *(results)*
3	(Rohman et al., 2020)	Komparasi Algoritma C4.5 Berbasis PSO dan GA Untuk Diagnosa Penyakit Stroke	Algoritma C4.5, PSO dan Algoritma Genetika	C45 merupakan Algoritma yang paling banyak digunakan, dalam kasus ini akurasi dari algoritma C4.5 sebesar 99.07%. Selanjutnya Algoritma C4.5 dioptimasi dengan menggunakan Particle Swarm Optimization sehingga memperoleh akurasi sebesar 99.28% dan Algoritma C4.5 juga dioptimasi dengan menggunakan Genetic Algorithm sehingga memperoleh akurasi sebesar 99.38%
4	(Samosir et al., 2021)	Komparasi Algoritma Random Forest, Naïve Bayes dan K- Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung Amril	Algoritma Random Forest, Naïve Bayes dan K-Nearest Neighbor	Berdasarkan hasil perbandingan terhadap 304 dataset penyakit jantung, algoritma Naïve Bayes lebih baik dan optimal dibanding dengan Algoritma, K-Nearest Neighbor dan Random Forest untuk mengklasifikasikan penyakit jantung. Hasil klasifikasi dengan algoritma Naïve Bayes memiliki rerata hasil akurasi sebesar 0,91 AUC, 0,84 CA, 0,84 F1, 0,839 Precision dan 0,84 Recall
5	(Handayani et al., 2021)	Komparasi Algoritma C4.5 dan Naïve Bayes dalam Penentuan Status	Algoritma C4.5 dan Naïve Bayes	Dari hasil penelitian ini, hemoglobin adalah variabel paling menentukan kelayakan donor darah kemudian tekanan darah.

No	Peneliti dan Tahun	Judul	Algoritma *(methods)*	Hasil *(results)*
		Kelayakan Donor Darah		Algoritma terbaik dalam kasus ini adalah Naïve Bayes dengan akurasi 93,26%, sedangkan tingkat akurasi C4.5 93,22%. Naïve Bayes termasuk dalam predikat good classsification dengan AUC sebesar 0.833, sedangkan C4.5 termasuk dalam predikat fair classsification dengan AUC sebesar 0.758. Dari hasil uji beda t-test diperoleh hasil 0.841 yang menyatakan bahwa tidak ada perbedaan signifikan dalam penentuan klasifikasi status kelayakan donor darah untuk kedua algoritma
7	(Subarkah et al., 2021)	Komparasi Akurasi Algoritme CART dan Neural Network Untuk Diagnosis Penyakit Diabetic Retinopathy Comparison	Algoritme CART dan Neural Network	Hasil pengujian yang didapatkan dengan algoritme CART didapati nilai akurasi yaitu 63.4231% precision 0.64%, recall 0.634%, dan f-measure 0.634% sedangkan pada algoritme Neural Network didapatkan hasil nilai akurasi 72.285% , precision 0.723%, recall 0.723%, dan F- Measure 0.723%. Dari hasil tersebut dapat disimpulkan bahwa algoritme Neural Network lebih tepat guna mendiagnosis

No	Peneliti dan Tahun	Judul	Algoritma *(methods)*	Hasil *(results)*
				penyakit diabetes retinopathy
7	(Alhabib et al., 2022)	Komparasi Metode Deep Learning, Naïve Bayes dan Random Forest untuk Prediksi Penyakit Jantung Ivana	Deep Learning, Naïve Bayes dan Random Forest	Dataset diambil dari laman *Kaggle* dengan judul heart attack analysis dan prediction dataset. Akurasi tertinggi yang dapat dicapai dengan menggunakan algoritma deep learning, yang menghasilkan akurasi sebesar 83,49%
8	(Andryan et al., 2022)	Komparasi Kinerja Algoritma Xgboost Dan Algoritma Support Vector Machine (Svm) Untuk Diagnosis Penyakit Kanker Payudara	Algoritma Xgboost Dan Algoritma Support Vector Machine (SVM)	Hasil kinerja yang didapat setelah melakukan penelitian menggunakan kedua algoritma adalah Xgboost yang memiliki akurasi terbaik sebesar 95.12% dan nilai ROC AUC sebesar 0.99 dan algoritma SVM memiliki akurasi terendah sebesar 90.24% dan nilai ROC AUC sebesar 0.98
9	(Napiah et al., 2022)	Komparasi Algoritma Untuk Klasifikasi Penyakit ISPA (Infeksi Saluran Pernapasan Akut) Musriatun	Naïve Bayes, K-Nearest Neighbour, Support vector machine (SVM)	Hasil dari penelitian menggunakan *machine learning* yang digunakan Naïve Bayes, hasil akurasi yang diperoleh sebesar 98% dengan Kappa Score 95%, K-Nearest Neighbour hasil akurasi yang diperoleh sebesar 94% dengan Kappa Score 87%, sedangkan dengan menggunakan SVM diperoleh akurasi

No	Peneliti dan Tahun	Judul	Algoritma *(methods)*	Hasil *(results)*
				sebesar 99% dengan Kappa Score 97%
11.	(Dharmawan, 2022)	Komparasi Algoritma Klasifikasi Svm-Pso dan C4.5-Pso Dalam Prediksi Penyakit Jantung	Klasifikasi SVM - Pso dan C4.5-PSO	hasil pengujiannya akan di bandingkan dengan algoritma klasifikasi C4.5 sebagai pembanding algoritma mana yang lebih baik dalam memprediksi dari sebuah dataset. C4.5 digunakan juga untuk memberikan hasil klasifikasi yang di gabungan bersama Particle Swarm Optimization (PSO). Dari eksperimen yang di lakukan algoritma SVM-PSO mendapatkan nilai Accuracy 84.81% dan nilai AUCnya 0.898 sedangkan Algoritma C4.5-PSO mendapatkan nilai Accuracy 80.00% dan nilai AUCnya 0.787.

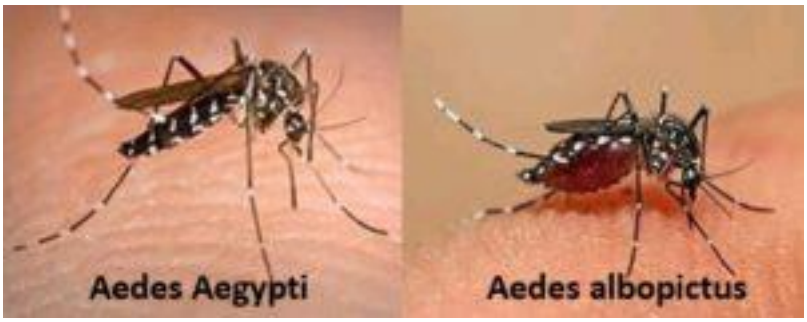
Berdasarkan hasil studi pustaka yang telah dilakukan pada Tabel 2.1., penelitian terkait klasifikasi untuk bidang medis cukup banyak dan secara umum penerapan pengelolaan data menggunakan algoritma klasifikasi berhasil membantu pada medis untuk melakukan diagnosa. Seperti penelitian (Jasri et al., 2022) menerapkan data mining untuk klasifikasi Demam Berdarah Dengue (DBD). Proses klasisfikasi menggunakan satu algoritma yaitu algoritma Naïve Bayes yang berperan sebagai seleksi fitur pada data yang dipeoleh dari Puskesmas Taman Krocok. Dari penelitian ini, penulis merasa ingin mengembangkan data yang ada dengan membandingkan (komparasi) dengan berbagai algoritma klasifikasi yang lainnya.

2.2 Demam Berdarah

Demam Berdarah Dengue (DBD) merupakan penyakit infeksi yang disebabkan oleh virus dengue (DENV). Virus DENV adalah virus RNA *single-stranded* yang terdiri dari empat *serotipe* yang berbeda yaitu DENV-1, DENV-2, DENV-3 dan DENV-4 termasuk dalam Genus Flavivirus, Family Flaviviridae. Penyakit ini ditandai dengan demam bifasik, leukopenia, limfadenopati, mialgia atau artralgia dan ruam (Isna & Sjamsul, 2021).

2.2.1 Nyamuk Penyebab DBD

Penyebab DBD yaitu hewan golongan nyamuk bernama Aedes albopictus dan Aedes aegypti. Nyamuk tersebut ditampilkan pada Gambar 2.1.



Gambar 2. 1 Nyamuk *Aedes albopictus* dan *Aedes aegypti*

Vasculopathy pada DBD ditandai dengan kebocoran pembuluh kapiler dan kelainan regulasi *hematologis* dan pada DSS terjadi *shock hipovolemik*. Penyakit Demam Berdarah Dengue (DBD) ditandai dengan demam tinggi mendadak tanpa sebab yang jelas, berlangsung terus menerus selama 2-7 hari, manifestasi perdarahan (*petechie, purpura*, perdarahan konjungtiva, epistaksis, perdarahan mukosa, perdarahan gusi, hematemesis, melena, hematuri) termasuk uji tourniquet (Rumple Leede) positif, trombositopeni (jumlah trombosit ≤ 100.000/l, hemokonsentrasi (peningkatan hemotokrit ≥ 20%) disertai atau tanpa pembesaran hati *(hepatomegali)* (Halstead, 2007). Banyak pasien yang meninggal dunia disebabkan penyakit ini. Tabel 2.2. menampilkan jumlah penderita dan kematian DBD di Indonesia (Isna & Sjamsul, 2021)

Tabel 2.2. Jumlah penderita, kematian, CFR dan IR DBD 2014-2019

No	Data	2014	2015	2016	2017	2018	2019
1	Penderita DBD	100.347	129.650	204.171	68.407	53.075	13.683
2	Kematian DBD	907	1.071	1.598	493	344	133
3	CFR DBD	0,9	0,83	0,78	0,72	0,65	0,94
4	IR DBD	39,83	50,75	78,85	26,10	20,01	5,08

Keterangan : (a) *Crude Fatality Rate* (CFR), (b) *Inciden Rate* (IR)

2.2.2 Diagnosis Demam Berdarah Dengue (DBD)

Diagnosis medis merupakan penentuan kondisi kesehatan yang sedang dialami oleh seseorang sebagai dasar pengambilan keputusan medis untuk prognosis dan pengobatan. Diagnosis dilakukan untuk menjelaskan gejala dan tanda klinis yang dialami oleh seorang pasien, serta membedakannya dengan kondisi lain yang serupa. Penegakan diagnosis diawali dengan mengumpulkan informasi melalui anamnesis yang dilanjutkan dengan pemeriksaan fisik terhadap pasien. Umumnya dokter melakukan diagnosis pasien DBD dengan melakukan pemeriksaan pasien dari melihat gejala-gejala yang timbul dengan melakukan kecocokan antara gejala DBD dengan gejala pada penyakit yang lain (Siswanto & Usnawati, 2019).

Tanda-tanda gejala DBD biasanya muncul sebelum DBD menjadi parah dan pasien direkomendasikan segera melakukan uji klinis laboratorium. Tes *tourniquet* berguna apabila tes laboratorium tidak dapat dilakukan. Untuk melakukan *test tourniquet*, seorang dokter harus melakukan pengukuran tekanan darah di lengan selama 5 menit. Selanjutnya Dokter akan melihat jumlah bintik-bintik merah kecil di kulit pasien. Jumlah bintik yang semakin banyak berarti bahwa pasien tersebut mungkin menderita demam dengue (Asidik et al., 2021)

Membedakan DBD dengan penyakit lain seperti malaria, tipus dan chikungunya dirasakan masih sulit dilakukan. Chikungunya adalah infeksi virus yang mirip dan memiliki banyak gejala yang sama dengan dengue, dan terjadi di wilayah yang sama di dunia. Dengue juga dapat memiliki gejala yang sama seperti penyakit lainnya, seperti leptospirosis, demam tifoid, dan penyakit meningokokus. Umumnya sebelum seseorang terdiagnosis dengue, Dokter atau petugas kesehatan

yang menanganinya akan melakukan tes untuk memastikan penyakit yang dialami pasien. Jika seseorang menderita dengue, perubahan paling awal yang dapat dilihat pada tes laboratorium adalah jumlah sel darah putih yang sedikit, trombosit dan demam tinggi. Jumlah trombosit yang sedikit dan asidosis metabolik juga merupakan tanda-tanda dengue. Beberapa cara melakukan diagnosis (Kemenkes RI, 2020; WHO, 2009)

- **Diganosis klinis.** Ditandai demam akut, muncul bintik-binik merah, kondisi tubuh lemas, sakit kepala berat, nyeri otot, trombositopenia, perdarahan ringan-berat, kebocoran plasma hemokonsentrasi, efusi pleura, hipoalbuminemia.
- **Diagnosis laboratorium.** Diagnosis ini dilakukan melalui pemeriksaan hematologi rutin, uji virology, dan uji serologi.

2.2.3 Gejala Demam Berdarah *Dengue* (DBD)

Gejala merupakan pengindikasian keberadaan sesuatu penyakit atau gangguan kesehatan yang tidak diinginkan, berbentuk tanda-tanda atau ciri-ciri penyakit dan dapat dirasakan, seperti misalnya perasaan mual atau pusing. Akan tetapi, ada hal yang tidak tercakup dalam pengertian istilah ini seperti halusinasi atau delusi, karena cara melakukan indikasi ini berdasarkan pada diri pelaku sering tanpa sadar, dan bukan hasil dari pengamatan yang dilakukan berdasarkan pemeriksaan kedokteran (Ulfi, 2018). Sekitar 80% dari pasien atau 8 dari 10 pasien yang terinfeksi virus dengue tidak menunjukkan gejala, atau hanya menunjukkan gejala ringan seperti demam biasa. Sekitar 5% dari orang yang terinfeksi sebanyak 5 dari 100 yang akan mengalami infeksi berat (Siswanto & Usnawati, 2019). Gejala DBD yang umum dialami pasien adalah suhu tubuh meningkat, muncul bintik-bintik merah, sakit kepala hebat, nyeri otot dan mengurangnya trombosit. Gejala ini akan muncul antara 3 dan 14 hari setelah seseorang terpapar virus dengue, oleh karena itu jika seseorang baru kembali dari wilayah yang memiliki banyak kasus dengue, kemudian seseorang tersebut menderita demam atau gejala lainnya setelah lebih dari 14 hari, kemungkinan penyakitnya tersebut adalah dengue (Asidik et al., 2021)

Kasus DBD juga banyak menyerang anak-anak. Apabila anak-anak terkena demam dengue, biasanya gejala yang muncul sama dengan gejala pilek atau *gastro enteritis*, maka harus segera dilakukan pemeriksaan ke layanan kesehatan terdekat, jika orang tua tidak melakukan maka anak-anak dapat mengalami masalah yang parah bahkan menimbulkan kematian (Hadinegoro et al., 2012)

2.2.4 Risiko Penyakit DBD pada Manusia

Dibandingkan dengan orang dewasa, bayi dan anak kecil yang menderita dengue lebih berisiko mengalami infeksi yang serius. Anak- anak cenderung berisiko mengalami sakit berat apabila mereka tergolong anak-anak yang kekurangan gizi. Perempuan lebih cenderung terserang sakit yang lebih parah daripada laki-laki. Dengue bisa mengancam jiwa pada pasien dengan penyakit kronis (jangka panjang), seperti Diabetes dan Asma (Bhatt et al., 2013)

2.3 Data Mining

Data mining merupakan proses menelusuri pengetahuan terbaru, pola dan tren yang dipilih dari jumlah data yang besar dan disimpan dalam *repositori* atau tempat penyimpanan dengan menggunakan teknik pengenalan pola serta statistik dan teknik matematika . Data mining adalah teknik untuk menemukan pola tertentu dari sekumpulan data berjumlah besar. Dalam sebuah *database,* pasti memuat data dalam jumlah yang sangat banyak (Han et al., 2014)

Data mining berguna untuk mencari dan ‘menambang’ pola-pola unik dalam data yang ada pada *database* tersebut. Teknik ilmu komputer ini biasa dipakai pada proses pencarian *knowledge*. Biasanya, metode data mining diterapkan dalam bidang *machine learning* dan statistika. Hal ini berawal dari semakin meningkatnya kompleksitas kerja komputer. Teknik data mining kemudian digunakan sebagai proses pengumpulan dan seleksi data yang lebih praktis. Inilah letak keuntungan data mining yang kemudian diterapkan pada bidang pekerjaan lain selain komputer.

2.3.1 Teknik Data Mining

Teknik data mining adalah metode yang dapat diterapkan pada berbagai bidang. Oleh sebab itu, metode ini perlu disesuaikan dengan permasalahan atau kebutuhan penggunanya (Anggarwal, 2015; Han et al., 2014). Ada tujuh klasifikasi data mining yang dibedakan berdasarkan cara kerjanya, seperti berikut :

1. ***Tracking Patterns/Sequencing*.** Teknik data mining yang pertama adalah melacak pola atau urutan peristiwa. Teknik ini berfungsi untuk menemukan suatu pola pada serangkaian kejadian *(sequence)* yang berurutan. Teknik *tracking patterns* dapat mendeteksi sesuatu pada interval tertentu, seperti lonjakan permintaan produk ketika akhir pekan atau jumlah orang yang mengunjungi situs anda saat cuaca tertentu
2. ***Classification.*** Teknik *classification* memerlukan teknik data mining yang lebih kompleks karena menuntut Anda untuk mengumpulkan seluruh data dari kelas atau kategori tertentu. Teknik ini merupakan yang paling umum digunakan. Anda dapat mengaplikasikannya untuk mengelompokkan data berdasarkan label yang Anda inginkan. Misalnya, berdasarkan informasi finansial dan transaksi, Anda dapat mengelompokkan pelanggan menjadi *low, medium,* atau *high credit risks.*
3. ***Association.*** Klasifikasi data mining berikutnya yaitu *association, market basket* analisis yang berhubungan dengan pemasaran produk. Analisis keranjang bertujuan untuk mengetahui atau mengidentifikasi produk yang sering dibeli bersamaan oleh pelanggan. Misalnya, ketika membeli makanan ringan seperti kentang tertentu, pelanggan juga membeli minuman soda kemasan. Dengan mengetahui kebiasaan pelanggan seperti ini, maka perusahaan juga dapat melabeli produk tertentu sebagai "*people also bought this"* pada *marketplace.*
4. ***Outlier Detection.*** Teknik ini bertujuan untuk mengidentifikasi ketika terjadi anomali pada pola data. Misalnya, ketika produk Anda biasanya selalu dibeli oleh pelanggan berjenis kelamin laki-laki, namun pada suatu minggu di bulan Februari, tiba-tiba terjadi lonjakan pembelian yang dilakukan oleh pelanggan perempuan. Teknik *outlier detection* berperan

untuk menganalisis lonjakan tersebut serta penyebabnya, sehingga Anda dapat memutuskan langkah penjualan selanjutnya.

5. ***Clustering.*** Teknik *clustering* hampir mirip dengan *classification,* namun memerlukan label atau grup data yang lebih banyak berdasarkan pola kesamaan. Misalnya, Anda ingin mengelompokkan demografi *audiens* yang berbeda-beda menjadi beberapa grup berdasarkan latar belakang, finansial, atau jumlah pembeliannya ketika berbelanja di toko Anda.
6. ***Regression.*** Klasifikasi data mining berikutnya, *regression*, bertujuan untuk mencari pola nilai numeriknya alih-alih kelasnya. Hasil dari teknik ini adalah sebuah fungsi sebagai penentu yang didasarkan pada nilai dari *input.* Misalnya, Anda dapat menggunakannya untuk menentukan harga produk berdasarkan faktor lain seperti ketersediaan, permintaan pelanggan, dan kompetitor.
7. ***Forecasting/Prediction.*** Teknik data mining yang terakhir bisa dikatakan paling *valuable* karena bertujuan untuk memprediksi nilai yang akan dicapai pada periode tertentu. Dengan teknik *predicition*, noise data dan nilai pada periode sebelumnya dijadikan acuan atau dasar dari prediksi. Misalnya, berdasarkan data pembelian bulan lalu, Anda dapat memprediksi kira-kira pelanggan akan melakukan pembelian seperti apa di bulan depan

2.3.2 Algoritma Klasifikasi Data Mining

Penelitian ini menggunakan teknik data mining klasifikasi. Teknik klasifikasi merupakan suatu pengelompokan data untuk memprediksi nilai dari sekelompok atribut dalam menggambarkan dan membedakan kelas label atau target yang bertujuan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Anggarwal, 2015). Performa algoritma data mining pada banyak kasus tergantung kepada kualitas *dataset* yang digunakan, karena data *training* yang berkualitas rendah dapat menyebabkan klasifikasi yang lemah. Beberapa algoritma klasifikasi data mining :

1. **Algoritma Decision Tree (D3)** adalah jenis algoritma klasifikasi yang strukturnya mirip seperti sebuah pohon yang memiliki akar, ranting, dan

daun. Simpul akar *(internal node)* mewakili fitur pada dataset, simpul ranting *(branch node)* mewakili aturan keputusan *(decision rule)*, dan tiap-tiap simpul daun *(leaf node)* mewakili hasil keluaran. Itulah kenapa algoritma ini disebut Decision tree atau pohon keputusan

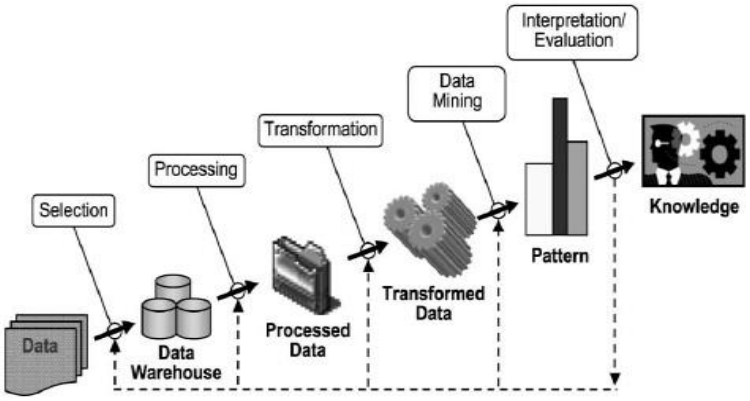
2. **Algoritma K-Nearest Neighbor (KNN)** adalah algoritma machine learning yang bersifat non-parametric dan lazy learning. Metode yang bersifat non-parametric memiliki makna bahwa metode tersebut tidak membuat asumsi apa pun tentang distribusi data yang mendasarinya. Dengan kata lain, tidak ada jumlah parameter atau estimasi parameter yang tetap dalam model, terlepas data tersebut berukuran kecil ataupun besar. Algoritma non-parametric seperti KNN menggunakan sejumlah parameter yang fleksibel, dan jumlah parameter seringkali bertambah seiring data yang semakin banyak. Algoritma non-parametric secara komputasi lebih lambat, tetapi membuat lebih sedikit asumsi tentang data
3. **Algoritma Naive Bayes.** Algoritma ini didasarkan pada teorema Bayes. Algoritma ini terutama digunakan ketika dimensi *input* tinggi. Pengklasifikasi dapat dengan mudah menghitung kemungkinan keluaran berikutnya. Ini adalah salah satu algoritma yang paling nyaman karena mudah dibangun dan tidak ada skema estimasi parameter yang rumit
4. **Algoritma Random Forest** adalah algoritma machine learning yang fleksibel dan mudah digunakan yang menghasilkan, bahkan tanpa menggunakan banyak parameter sehingga relatif menghasilkan hasil. Ini juga merupakan salah satu algoritma yang paling banyak digunakan, karena kesederhanaan dan keragamannya (dapat digunakan untuk tugas klasifikasi dan regresi). Random Forest merupakan algoritma pembelajaran yang supervised. *"Forest"* yang dibangunnya adalah kumpulan pohon keputusan, biasanya dilatih dengan metode *"bagging"*. Ide umum dari metode *bagging* adalah kombinasi model pembelajaran meningkatkan hasil keseluruhan
5. **Algoritma Regresi logistik** merupakan jenis *supervised learning* yang biasa digunakan untuk membuat sebuah model prediksi yang sama halnya

dengan regresi linear. Bedanya ada pada nilai variabel yang biasanya berupa nilai adalah ya/tidak, benar/salah, ataupun dalam bentuk bilangan biner 0/1

6. **Algoritma Gradien Boosted Tree** termasuk ke dalam algoritma klasifikasi yang menggunakan peningkatan akurasi prediktor. Beberapa perbedaan algoritma gradient boost dengan adaboost adalah gradient boost membangun tree 8 sampai 32 daun, sedangkan adaboost membangun stumps dengan 2 (dua) daun. Perbedaan kedua adalah gradient boost menggunakan boosting untuk proses pengoptimalan dengan menggunakan *loss function* untuk meminimalisir kesalahan. Algoritma ini disebut algoritma gradient boost karena terinspirasi dari penurunan gradien. Perbedaan yang terakhir adalah tree digunakan untuk memprediksi sisa sampel (hasil prediksi dikurangi aktual). Intinya, cara kerja algoritma gradient boost adalah membangun satu tree untuk menyesuaikan data, lalu tree berikutnya dibangun untuk mengurangi residual *(error)*
7. **Algoritma Support Vector Machine (SVM)** merupakan salah satu algoritma machine learning dengan pendekatan berbasis supervised learning yang dapat digunakan untuk masalah klasifikasi dan regresi. SVM adalah metode pada machine learning yang dapat digunakan untuk menganalisis data dan mengurutkannya ke dalam salah satu dari dua kategori. SVM telah digunakan dalam klasifikasi teks, hiperteks dan gambar

2.3.3 Tahapan Data Mining

Tahapan yang dilakukan pada proses data mining diawali dari seleksi data dari data sumber ke data target, tahap *pre-processing* untuk memperbaiki kualitas data, transformasi, data mining serta tahap interpretasi dan evaluasi yang menghasilkan *output* berupa pengetahuan baru yang diharapkan memberikan kontribusi yang lebih baik (Han et al., 2014). Secara detail dijelaskan sebagai berikut:



Gambar 2. 2 Tahapan data mining

Berikut ini adalah penjelasan tahapan data mining berdasarkan Gambar 2.2 :

1. *Data selection* atau pemilihan (seleksi) data. Sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dimulai. Data hasil seleksi yang digunakan untuk proses data *mining*, yang disimpan dalam suatu berkas, terpisah dari basis data operasional
2. *Pre-processing /cleaning.* Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data. *Preprocessing* adalah salah satu langkah terpenting dalam pemrosesan data. Data mungkin memiliki berbagai masalah yang dapat mempengaruhi hasil pengolahan data. *Preprocessing* merupakan salah satu langkah untuk menghilangkan berbagai masalah yang dapat mempengaruhi hasil pengolahan data. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data. pada fase ini dilakukan analisis terhadap data yang telah diproses sebelumnya dan fase klasifikasi dipindahkan. Langkah-langkah yang digunakan untuk proses *pretreatment* pada penelitian ini adalah sebagai berikut:
    a. ***Training Data*** Proses ini merupakan salah satu tahapan untuk CNN melakukan proses training agar mendapatkan nilai akurasi yang tinggi.
    b. ***Testing Data Proses*** ini digunakan untuk mengetahui performa dari algoritma yang dilatih dimana ketika ditemukan data baru yang sebelumnya belum pernah di *training*
3. *Transformation* adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Selanjutnya proses

coding yang berfungsi untuk menggali atau memproses secara kreatif dan sesuai jenis atau pola informasi yang akan dicari

4. Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses secara keseluruhan.
5. Klasifikasi merupakan suatu pengelompokan data untuk memprediksi nilai dari sekelompok atribut dalam menggambarkan dan membedakan kelas label atau target yang bertujuan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui
6. *Interpretation/evalution*. Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya

2.4 Himpunan Data *(dataset)*

Himpunan data merupakan sebuah kumpulan data yang sangat banyak berasal dari informasi masa-masa lampau dan dikelola menjadi sebuah informasi untuk melakukan teknik dari ilmu data mining (Torell et al., 2019). Set data (data set/himpunan data) merupakan kumpulan objek dan atributnya. Nama lain dari objek yang sering digunakan diantaranya *record, point, vector, pattern, event, observation, case, sample, instance*, entitas. Objek digambarkan dengan sejumlah atribut yang menerangkan sifat atau karakteristik dari objek tersebut. Atribut juga sering disebut variabel, *field,* fitur, atau dimensi. Atribut adalah sifat/properti/karakteristik objek yang nilainya bisa bermacam-macam dari satu objek dengan objek lainnya, dari satu waktu ke waktu yang lainnya (Renear et al., 2010).

Sebagai contoh seorang pasien merupakan objek, dimana objek pasien tersebut memiliki beberapa atribut seperti nama, usia, jenis kelamin dan lain-lain. Setiap pelanggan memungkinkan memiliki nilai atribut yang berbeda dengan

pelanggan lainnya, serta memungkinkan perubahan nilai atribut dari waktu ke waktu. Dataset terdiri dari 2 (dua) bagian data yaitu *Private* dan *Public.*

- ***Private Dataset,*** adalah data set yang dapat diambil dari sebuah organisasi yang akan kita lakukan sebagai objek penelitian misalnya seperti data bank, rumah sakit, universitas, perusahaan dan lain sebagainya
- ***Public Dataset,*** adalah data set yang bisa kita ambil dari repository publik yang disepakati oleh ulama-ulama peneliti data mining, misalnya seperti UCI Repository, Satu Data Puskesmas, dan lain-lain

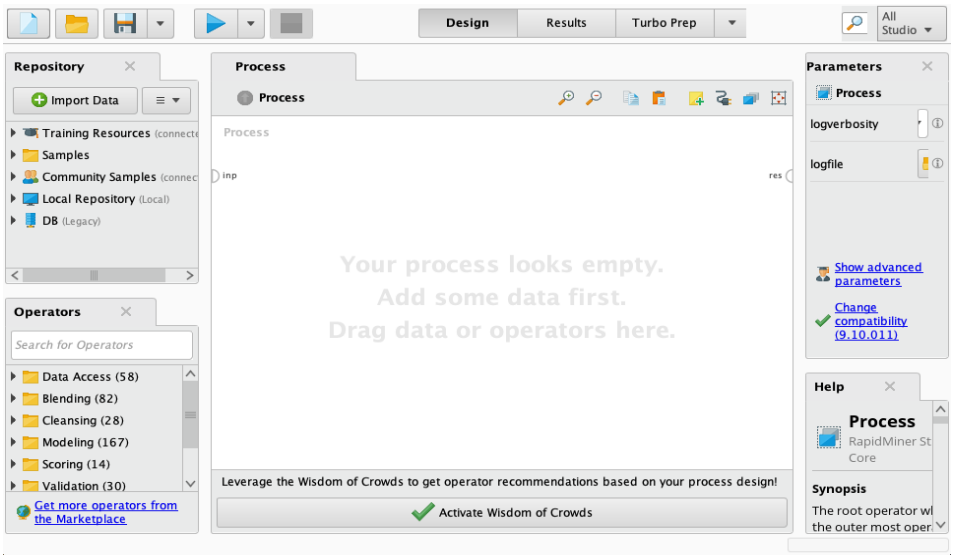
2.5 Rapid Miner Studio

Penelitian ini menggunakan RapidMiner Studio untuk mengelola data dan perbandingan algoritma. RapidMiner merupakan perangkat lunak yang bersifat terbuka *(open source)*. RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, *text mining* dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik (Fischer et al., 2002). RapidMiner memiliki sekitar 500 operator data mining, termasuk operator untuk *input, output, data preprocessing* dan visualisasi. RapidMiner merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan menggunakan bahasa java sehingga dapat bekerja di semua sistem operasi.

RapidMiner sebelumnya bernama **YALE** (*Yet Another Learning Environment),* dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh RalfKlinkenberg, Ingo Mierswa, dan Simon Fischer di *Artificial Intelligence Unit* dari University of Dortmund. RapidMiner didistribusikan di bawah lisensi AGPL *(GNU Affero General Public License)* versi 3. Hingga saat ini telah ribuan aplikasi yang dikembangkan menggunakan RapidMiner di lebih dari 40 negara. RapidMiner sebagai *software open source* untuk data mining tidak perlu diragukan lagi karena *software* ini sudah terkemuka di dunia (Klopper et al., 2016)

RapidMiner menempati peringkat pertama sebagai *Software* data mining pada polling oleh KDnuggets, sebuah portal data mining pada 2010-2011.

RapidMiner menyediakan GUI *(Graphic User Interface)* untuk merancang sebuah *pipeline* analitis. GUI ini akan menghasilkan *file* XML *(Extensible Markup Language)* yang mendefinisikan proses analitis keinginan pengguna untuk diterapkan ke data. *File* ini kemudian dibaca oleh RapidMiner untuk menjalankan analis secara otomatis. Tampilan RapidMiner Studio dapat dilihat pada Gambar 2.3.



Gambar 2. 3 RapidMiner Studio Versi 9.10.011

2.6 Confusion Matrix

Confusion Matrix adalah pengukuran performa untuk masalah klasifikasi machine learning dimana keluaran dapat berupa dua kelas atau lebih. Confusion Matrix adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan nilai actual (Ohsaki et al., 2017). Ada 4 (empat) istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix yaitu *True Positif (FP), True Negatif (TN), False Positif (FP), dan False Negatif (FN)* seperti Tabel berikut:

Tabel 2.3. Contoh Confusion Matrix

n= 175	Aktual: Positif (1)	Aktual: Positif (0)
Prediksi : Positif (1)	TP=125	FP=20
Prediksi : Negatif (0)	FN=25	TN=5
	145	25

Keterangan :

1) *True Positive (TP)* : Interpretasi: Anda memprediksi positif dan itu benar.
2) *True Negative (TN)* : Interpretasi: Anda memprediksi negatif dan itu benar.

3) *False Positive (FP):* (Kesalahan Tipe 1) Interpretasi: Anda memprediksi positif dan itu salah
4) *False Negative (FN):* (Kesalahan Tipe 2, kesalahan tipe 2 ini sangat berbahaya) Interpretasi: Anda memprediksi negatif dan itu salah.

Untuk menghitung nilai accuracy, precision, dan recall menggunakan confusion matrix maka digunakan formula berikut :

$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}$, $Precision = \frac{TP}{(TP+FP)}$, $Recall = \frac{TP}{(TP+FN)}$,

2.7 F1-Score dan F-Measure

Kesenjangan antara Precision dan Recall karena adanya tradeoff di antara keduanya. Ketika Recall sangat tinggi, Precision akan sangat rendah, begitu juga sebaliknya. Meskipun terdapat situasi yang ideal, dimana data dapat dipisahkan dengan sempurna dengan skor 1.0 diantara keduanya. Namun, sangat jarang atau bahkan tidak pernah terjadi oleh karena itu terdapat penilaian lain seperti F1-Score dan F-Measure yang dapat dipertimbangkan.

F1-Score adalah *harmonic mean* dari precision dan recall. F-Measure merupakan salah satu perhitungan evaluasi dalam informasi temu kembali yang mengkombinasikan recall dan precision. Nilai recall dan precision pada suatu keadaan dapat memiliki bobot yang berbeda. Ukuran yang menampilkan timbal balik antara Recall dan Precision adalah F-Measure yang merupakan bobot harmonic mean dan recall dan precision. Perhitungan F1-Score dan F-Measure dihitung berdasarkan persamaan berikut:

$$f - score = \frac{2TP}{2TP + FP + FN}, F - Measure = \frac{2 * Recall * Precision}{Recall + Precision}$$