BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Penelitian Terdahulu

Pada bagian ini akan memberikan tinjauan literatur terhadap berbagai penelitian yang menggunakan teknik data mining, metode klasifikasi, dan sistem pendukung keputusan lainnya untuk memprediksi Penyakit Ginjal Kronis (PGK), guna membantu pembaca memahami hubungan antara penelitian yang dilakukan dengan penyakit ginjal kronik. penelitian sebelumnya. Studi-studi ini meliputi:

Tabel 2.1 Tinjauan Pustaka

No	Judul	Penulis	Dataset	Metode	Hasil
1	Implementasi Algoritma *Naïve Bayes* Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik	Qurotul A'yuniyah, Ena Tasia, Nanda Nazira, Pangeran Fadillah Pratama, Muhammad Ridho Anugrah, Jeni Adhiva, Mustakim 2022	UCI *Machine Learning Repository* 400 *record* 22 atribut	Algoritma NBC	NBC 96.43%.
2	Optimasi *Random Forest* Untuk Diagnosis Penyakit Ginjal Kronik Dengan Menggunakan *Particle Swarm Optimization*	Sheva NaufalRifqi 2022	UCI *Machine Learning Repository* 25 atribut	Algoritma *Random Forest* dan Particle Swarn Optimization	*Particle Swarn Optimization +Random Forest* 99.167%. *Random Forest* 98%
3	Precise transformer fault diagnosis via *Random*	Rahman Azis Prasojo, Muhammad Akmal A.	-	*Random Forest* model *enhanced by synthetic*	implementation of *SMOTE improves the*

No	Judul	Penulis	Dataset	Metode	Hasil
	Forest model *enhanced by synthetic minority over-sampling technique*	Putra, Ekojono, MeytiEka Apriyani, Anugrah Nur Rahmanto, Sherif S.M. Ghoneim, Karar Mahmoud, Matti Lehtonen, Mohamed M.F.Darwish 2023		*minority over sampling technique*	*classification accuracy from 0.897 To 0.944 for DPM1 based RF model. This trend also happens to DPM2 based model across allthe performance evaluation parameters. The proposed models perform satisfactorily in diagnosing faults for the evaluation dataset, with a total Accuracy of 96.2% for DPM1 and 96.5% for DPM2.*
4	Perbandingan metode data mining svm dan nn untuk klasifikasi penyakit ginjal kronis	Hilda Amalia 2018	*https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease* 400 *record* 24 atribut 1 label	*Supper vector Machine* dan *Neural network.*	*neural network* 93.37% dan SVM 95.16%
5	Algoritma *K-Nearest Neighbor* Berbasis *Particle Swarm Optimization* Untuk Prediksi Penyakit Ginjal	KronikWarid Yunus 2018	Dataset CKD 25 atribut 400 *record* 1 label	Algoritma*K-Nearest Neighbor* Berbasis *Particle Swarm Optimization*	*K-Nearest Neighbor* pada parameter K = 1, dengan tingkat akurasi tertinggi yaitu 78.75%, Knn + PSO

No	Judul	Penulis	Dataset	Metode	Hasil
					97.25%.
6	Implementasi *Fuzzy Decision Tree* UntukPrediksi Gagal Ginjal Kronis	Fitri Sofia Nur Khamidah, Dian Hapsari, Hendro Nugroho 2018	*Repository* UCI *Machine Learning* 25 atribut 400 *record* 1 label	*Fuzzy Decision Tree*	*Fuzzy Decision Tree* 98.28%.
7	Diagnosis of Breast Cancer Using *Random Forest*s.	Manas Minnoora, Veeky Bathsb. 2023	*The UC Irvine Machine Learning Repository*'s Wisconsin Breast Cancer Diagnostic dataset 11 atribut	*Support Vector Machine* (SVM), *Decision Tree, Multilayer Perceptron, Random Forest and K-Nearest Neighbors*	SVM 97.90% , *Decision Tree, Multilayer Perceptron* 95.80%, *Random Forest* 99.30%, dan *K-Nearest Neighbors* 93.01%
8	peningkatan akurasi klasifikasi algoritma c4.5 menggunakan teknik *bagging* pada diagnosis penyakit jantung	Erwin Prasetyo, Budi Prasetiyo 2020	*http://archive.ics.uci.edu/ml/datasets/heart+ Disease* 303 instance 75 atribut 1 kelas	Algoritma C4.5 dan tehnik *bagging*	C4.5 72,98%. *Bagging* + C4.5 81,84%.
9	Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi *Random Forest*	Widya Apriliah, Ilham Kurniawan, Muhamad Baydhowi, Tri Haryati 2020	*Dataset Diabetes Hospital in Sylhet*, Bangladesh yang diambil dari UCI *repository* 520 data 17 *atribut*	*Suport Vector Machine, Naive Bayes* dan*Random Forest*	SVM 94,80% *Random Forest* 97,88%. *Naive Bayes* 91,92%

Berdasarkan permasalahan dari hasil literature review pada tabel 2.1 akan dilakukan penelitian PGK menggunakan dua metode untuk menghasilkan nilai

akurasi terbaik pada klasifikasi PGK yaitu metode *Bagging* dan *Random Forest*, dengan dataset berjumlah 400 *records*.

2.2 Data Mining

Proses penerapan teknik atau metodologi tertentu pada data terpilih untuk menemukan pola atau informasi menarik dikenal dengan istilah data mining. Teknik, metode, atau algoritma penambangan data sangat berbeda satu sama lain. Pemilihan teknik atau algoritma terbaik sangat bergantung pada tujuan dan keseluruhan proses Penemuan Pengetahuan dalam *Database* (KDD) [8].

Komponen penting dari penambangan data meliputi [9]:

1. Proses otomatis menggunakan data yang sudah ada disebut data mining.
2. Banyaknya data yang harus diolah.
3. Menemukan hubungan atau pola yang dapat menghasilkan indikator yang berguna adalah tujuan dari data mining.

Beberapa model atau mode mempunyai fungsi dalam perkembangan teknologi data mining, seperti [10]:

1. *Association Rule Mining*: Dengan menggunakan pendekatan ini, seseorang dapat menentukan hubungan antara objek dalam kumpulan data. Banyak aplikasi, termasuk saran produk dan analisis belanja pengguna, dapat menggunakan konsep ini.
2. *Classification*: Dengan menggunakan paradigma ini, data dapat dikategorikan menurut sifat-sifatnya. Pendekatan ini dapat digunakan, misalnya, untuk mendeteksi penipuan dalam transaksi keuangan dan mengklasifikasikan email sebagai spam atau bukan.
3. *Clustering*: Berdasarkan kemiripan ciri-cirinya, model ini digunakan untuk mengklasifikasikan data serupa ke dalam kelompok yang berbeda. Model ini dapat digunakan, misalnya, dalam segmentasi pasar untuk memahami preferensi klien dan analisis jaringan sosial untuk membedakan berbagai komunitas.
4. *Regression*: Dengan menggunakan nilai satu atau lebih variabel independen, model ini digunakan untuk memprediksi nilai variabel dependen. Model ini

dapat digunakan misalnya untuk mengantisipasi harga rumah berdasarkan fitur tertentu atau penjualan berdasarkan harga, promosi, dan faktor lainnya, serta komunitas.

5. *Outlier Detection*: Dalam suatu dataset, model ini digunakan untuk menemukan data yang anomali atau tidak terduga. Pendekatan ini dapat digunakan, misalnya, untuk mendeteksi penipuan dalam transaksi keuangan dan masalah dalam sistem produksi.
6. *Time Series Analysis*: Data terkait waktu dianalisis menggunakan pendekatan ini. Model ini dapat digunakan, misalnya, untuk meramalkan harga saham dan permintaan barang tertentu.

2.3 Prediksi

Memperkirakan atau memproyeksikan kejadian di masa depan adalah definisi prediksi [11]. Peramalan dan prediksi adalah hal yang sama. Yang dimaksud dengan "ramalan" adalah "memperkirakan apa yang akan terjadi di masa yang akan datang berdasarkan fakta atau bukti yang ada", menurut Kamus Besar Bahasa Indonesia [12]. Proses menghasilkan informasi yang akurat tentang situasi sosial di masa depan melalui penggunaan data yang tersedia mengenai suatu masalah kebijakan disebut peramalan. Ada tiga jenis utama perkiraan: perkiraan, prediksi, dan proyeksi [12]:

1. Proyeksi merupakan perkiraan yang diperoleh dari ekstrapolasi pola historis dan prospektif. Proyeksi yang berasal dari argumen yang berasal dari metodologi tertentu dan kasus paralel menimbulkan masalah yang tegas.
2. Prediksi adalah ramalan yang didukung oleh asumsi teoritis yang jelas. Asumsi ini dapat berbentuk sebuah analogi, sebuah dalil teoretis seperti disintegrasi masyarakat sipil yang disebabkan oleh ketidaksesuaian antara ekspektasi dan kemampuan atau sebuah hukum teoretis seperti hukum nilai uang.
3. Perkiraan, kadang-kadang disebut hipotesis, adalah perkiraan mengenai keadaan masyarakat di masa depan yang didasarkan pada pendapat atau penilaian para ahli. Tujuan dari peramalan kebijakan adalah untuk

memastikan kejadian di masa depan yang akan berdampak pada pelaksanaan kebijakan dan hasilnya.

Berdasarkan penjadwalan operasi yang akan datang, peramalan dipisahkan menjadi tiga bidang, yaitu [13]:

1. Peramalan ekonomi, juga dikenal sebagai peramalan keuangan, prakiraan antara lain inflasi, jumlah uang beredar, dan jumlah uang yang dibutuhkan untuk membangun rumah baru, yang membantu menjelaskan siklus bisnis.
2. Peramalan teknologi memperhitungkan kemungkinan bahwa produk baru dengan fitur menarik akan segera diperkenalkan, sehingga memerlukan pembelian mesin dan pabrik baru.
3. Proyeksi permintaan suatu perusahaan terhadap produk atau jasanya dikenal sebagai peramalan permintaan.

Ramalan langkah demi langkah yang akurat dibuat dengan mengikuti pedoman pra-perencanaan yang baik. Terdapat sembilan prosedur yang perlu diikuti untuk menjamin sistem beroperasi secara efektif, antara lain [13]:

1. Tetapkan tujuan peramalan.
2. Memutuskan komponen permintaan independen mana yang akan diproyeksikan.
3. Pilih perkiraan jangka waktu (panjang, sedang, atau pendek).
4. Pilih model prediksi.
5. Memperoleh informasi yang diperlukan untuk melakukan peramalan.
6. Verifikasi model yang digunakan untuk memprediksi.
7. Buat proyeksi.
8. Mempraktikkan proyeksi hasil.
9. Verifikasi keakuratan temuan peramalan.

Teknik peramalan analisis time series menampilkan keterkaitan antara variabel terikat (yang dicari) dengan variabel bebas (yang mempengaruhinya). Waktu merupakan variabel yang dicari dalam analisis deret waktu. Teknik peramalan ini meliputi [12]:

1. Metode *Smooting* adalah jenis peramalan jangka pendek, termasuk inventarisasi dan perencanaan keuangan. Penggunaan teknik ini bertujuan

untuk mengurangi kelainan data sebelumnya seperti musiman.

2. Metode Box Jenkins, yang menggunakan deret waktu dan model matematika untuk peramalan jangka pendek.
3. Metode *proyeksi trend* dengan *regresi*, adalah teknik jangka pendek dan jangka panjang yang menghasilkan garis *tren* persamaan menggunakan pendekatan proyeksi *tren* regresi.

2.4 Klasifikasi

Dalam data komunitas pertambangan, urutan klasifikasi sangat penting. Klasifikasi adalah metode prediksi data mining yang menggunakan hasil yang diketahui dari berbagai kumpulan data untuk memprediksi nilai data. Banyak algoritma klasifikasi mengalami kehilangan informasi ketika berhadapan dengan data yang tidak seimbang, seperti ketika distribusi sampel di seluruh kelas sangat tidak seimbang, yang menyebabkan algoritma tersebut kesulitan dengan masalah akurasi [9].

Berikut adalah beberapa contoh pekerjaan kategorisasi dalam bisnis dan penelitian [14]:

1. Tetapkan status penipuan dari transaksi kartu kredit tertentu.
2. Menugaskan siswa baru pada jalur khusus yang berhubungan dengan kebutuhan khusus.
3. Menentukan adanya penyakit.
4. Cari tahu apakah orang mati menulis surat wasiat itu dengan jujur atau ada orang lain yang menulisnya dengan curang.
5. Menilai apakah tindakan moneter atau individu tertentu menunjukkan adanya potensi.
6. Tentukan apakah aktivitas keuangan atau pribadi tertentu mengarah pada potensi bahaya teroris.

Pengklasifikasi berguna di banyak domain, termasuk identifikasi pola, penambangan data, dan pembelajaran mesin. Agar pengklasifikasi dapat memprediksi dengan benar kelas atau kategori data yang baru dibuat yang belum pernah dilihat sebelumnya, pengklasifikasi harus terlebih dahulu mengidentifikasi

pola dan hubungan dalam kumpulan data pelatihan. Di antara metode klasifikasi tersebut adalah *Decision Trees*, *Random Forest*s, *Support Vector Machines* (SVM), *Naive Bayes*, dan *Neural Networks* [15].

2.5 Penyakit Ginjal Kronis (PGK)

Sepasang organ penting yang disebut ginjal terletak di bagian belakang tubuh kanan dan kiri. Setiap ginjal memiliki ukuran sekitar sepuluh hingga lima belas sentimeter dan berat sekitar seratus gram [4]. Ginjal melakukan banyak tugas penting untuk tubuh yaitu [4]:

1. Menghilangkan cairan asing dari darah dan produk metabolisme tubuh..
2. Menjaga keseimbangan cairan dan elektrolit, termasuk kalium dan natrium.
3. Mengontrol tekanan darah dan mendorong pembentukan sel darah merah.

Salah satu kondisi yang berdampak pada fungsi ginjal adalah penyakit ginjal. Penyakit yang paling banyak diderita di dunia, penyakit ginjal kronis (PGK) ditandai dengan prognosis yang buruk, meningkatnya biaya pengobatan, peningkatan prevalensi, dan peningkatan frekuensi gagal ginjal [3]. Pada suatu saat dalam hidup mereka, sekitar 10% orang di seluruh dunia menderita penyakit ginjal kronis (PGK), dan prevalensi PGK meningkat seiring dengan jumlah orang lanjut usia yang juga menderita hipertensi dan diabetes melitus [1]. Penderita penyakit ginjal kronis (PGK) terus meningkat, terutama disebabkan oleh faktor gaya hidup seperti diabetes, tekanan darah tinggi, dan kebiasaan makan yang buruk. Pasien dengan penyakit ginjal kronis (PGK) mungkin tidak mengalami gejala atau konsekuensi pada awalnya, namun penyakit ini dapat berlanjut menjadi gagal ginjal [4].

Deteksi dini penyakit ginjal memungkinkan pengobatan, pencegahan, dan pilihan terapi yang lebih baik [16]. Penurunan fungsi jaringan ginjal secara perlahan hingga massa ginjal yang tersisa tidak mampu menopang lingkungan internal tubuh adalah definisi lain dari gagal ginjal kronis. Gagal ginjal kronik juga diartikan sebagai kegagalan fungsi ginjal terutama pada unit nefron yang terjadi secara bertahap karena sebab yang menetap dan bertahan lama, sehingga menyebabkan penumpukan sisa metabolit atau toksin uremik yang menyebabkan

ginjal tidak lagi berfungsi dengan baik [1].

Ini adalah gejala yang berhubungan dengan penurunan fungsi ginjal, yang antara lain dapat menyebabkan gagal ginjal [16]:

1. Penumpukan Limbah Dalam Darah : Hal ini ditandai dengan rasa lelah, nyeri di sekujur tubuh, gatal-gatal, kram, mudah lupa, susah tidur, mual, tidak nafsu makan, dan menurunnya daya tahan tubuh terhadap infeksi.
2. Kesulitan keseimbangan cairan dengan pengumpulan cairan di wajah dan pergelangan kaki. Sebaliknya, mulut kering, mata sangat cekung, dan hampir tidak ada lendir di mulut merupakan tanda-tanda keluarnya cairan.
3. Gangguan hormon, Ginjal mungkin memproduksi lebih banyak atau lebih sedikit hormon jika kemampuannya berfungsi berkurang. Akibatnya, hormon yang berhubungan dengan tekanan darah meningkat sedangkan hormon yang berhubungan dengan proses tubuh lainnya menurun. Tubuh mengalami kelelahan, kehilangan darah, dan tulang rapuh akibat hal ini.
4. Keracunan dan trauma, seperti serangan langsung dan kuat ke ginjal. Obat-obatan tertentu, termasuk obat yang dijual bebas, dapat merusak ginjal jika dikonsumsi berulang kali dalam jangka waktu yang lama. Produk yang mengandung asetaminofen, aspirin, dan obat lain seperti ibuprofen diketahui paling berbahaya bagi ginjal. Jika kita sering menggunakan obat pereda nyeri, sebaiknya kita menemui dokter untuk memastikan obat tersebut tidak merusak ginjal kita.

2.6 *Bagging*

Diusulkan oleh Breiman, *bagging*, juga dikenal sebagai *boostrap agreggating*, adalah metode klasik untuk pembuatan *ansambel*. [17]. Meskipun masalah regresi data mungkin juga mendapat manfaat dari penggunaannya, masalah klasifikasi adalah tujuan awalnya [18]. Ini ditunjukkan dengan mengambil beberapa sampel dari kumpulan data yang sama dengan penyesuaian kembali melalui teknik *bootstrap*. Hal ini berguna untuk menghasilkan prediksi agregat karena memungkinkan pembuatan beberapa pohon berbeda untuk perkiraan yang sama [19].

Prinsip dasar metode *bagging* adalah membuat kumpulan data baru dengan mengambil sampel ulang kumpulan data asli secara acak dan mengembalikannya. Dengan menggunakan sampel acak berukuran N dengan penggantian dari data pelatihan (sampel *bootstrap* Sk dari Dk), kumpulan data baru Dk berukuran sama dengan data pelatihan $|D|$. Pohon klasifikasi dengan berbagai versi kemudian dibuat dengan dataset baru. Perkiraan akhir kemudian dihasilkan dengan menggabungkan pohon klasifikasi dari setiap versi [20].

Perkiraan akhir dari metode ini dapat dihasilkan dengan melakukan voting atau rata-rata untuk tantangan yang terkait dengan regresi dan klasifikasi. Hal ini memungkinkan untuk mengatur beberapa sampel menjadi sama [18].Tujuannya adalah untuk menghasilkan subset data menggunakan variabel pengganti dari set pelatihan yang dipilih secara acak. Intinya, proses pembelajaran dilatih menggunakan setiap subset kumpulan data. Sebagai hasilnya, kita memiliki kumpulan model yang berbeda. Dengan menggunakan rata-rata dari semua prediksi dari pembelajar dasar yang berbeda, hasilnya lebih dapat diandalkan daripada hanya menggunakan satu pembelajar dasar [21]. Manfaat II-11 untuk pembuatan batch adalah mengurangi kesalahan pada prediktor dasar, yang mungkin tidak stabil sebelum terjadi gangguan tertentu, dan memberikan perkiraan kinerja prediktifnya sendiri, yang terhambat oleh set pengujian atau perkiraan validasi silang [22].

Dalam metode *bagging*, ada dua tahapan. Tahap pertama adalah *bootstrap*, dan tahap kedua adalah *aggregating* [20]. Tahap bootstrap dilakukan dengan mengambil sampel dari data yang dimiliki, yang dikenal sebagai *resampling*. Tahap kedua dari metode *bagging* adalah *aggregating*. *Aggregasi* adalah proses menggabungkan berbagai nilai prediksi menjadi satu nilai prediksi [23]. Khususnya untuk klasifikasi, metode ini mengumpulkan suara terbanyak, atau mayoritas suara. Untuk *regresi*, rata-rata digunakan.

Menurut Erwin Prasetyo [6] tahapan *bagging* dapat diperhatikan sebagai sebagai berikut:

1. Tahapan *Bootstrap*
    a. Ambil sampel secara acak sebanyak n dari data latih.

b. Susun pohon terbaik berdasarkan data latihan tersebut setelah sampel diambil.
c. Dapatkan buah pohon klasifikasi B dengan mengulangi langkah a-b sebanyak B kali.

2. Tahapan *Aggregating*

   Untuk memprediksi atau memperkirakan gabungan buah pohon klasifikasi B, tujuan *aggregating* identik dengan tujuan mayoritas suara.

2.7 Algoritma *Random Forest*

Beberapa pohon keputusan dibuat menggunakan teknik *Random Forest* (RF), dimana setiap pohon digabungkan dan berfungsi sebagai model *ansambel*. Setiap pohon keputusan memiliki prediksi kelas dan pilihan yang disusun berdasarkan hasil tertinggi [24].

Ada beberapa proses yang terlibat dalam penggunaan pendekatan *Random Forest*, yaitu [25]:

1. Tahap pengambilan gambar acak dengan pemulihan dan dari data pelatihan disebut *bootstrapping*.
2. *Subsetting* acak, dalam tahap ini, pohon dengan variabel berbeda disiapkan menggunakan proses diskon acak terbaik (m < d) bergantung pada data yang tersedia.
3. Sampai k pohon dihasilkan secara acak, ulangi langkah a dan b sebanyak k kali.
4. Menyelesaikan estimasi gabungan berdasarkan k pohon adalah langkah terakhir. Hal ini dapat diterapkan pada regresi terhadap kasus rata-rata atau kasus yang diklasifikasikan berdasarkan suara terbanyak.

Tujuan dari teknik ini adalah untuk membangun pohon keputusan yang terdiri dari *root node*, *internal node*, dan *leaf node* dengan menggunakan data dan atribut secara acak. *Root node* adalah simpul di atas pohon keputusan, dan *internal node* adalah simpul percabangan yang memiliki satu input dan minimal dua *output*. *Leaf node* atau *terminal node* adalah simpul terakhir, yang hanya memiliki satu input dan tidak memiliki *output* [26].

Perhitungan nilai *entropy* menggunakan rumus pada persamaan 1, dan nilai *information gain* pada rumus persamaan 2.

$$\boldsymbol{Entropy}\ (\gamma) = -\sum_{i} \mathbf{p(c|\gamma)}\ \mathbf{Log_2 p(c|\gamma)} \ldots\ldots\ldots\ldots\mathbf{(1)}$$

Keterangan:

Y merupakan himpunan kasus.

p(c|Y) adalah *proporsi* nilai *Y* pada kelas *c*.

$$\boldsymbol{Information\ Gain}\ \mathbf{(Y, a)} =$$

$$\boldsymbol{Entropy}\ (\gamma) - \sum_{\mathbf{veValues}\ (a)} - |\gamma v|_{\mathbf{Entropy}\ (\gamma v)} \ldots\ldots\ldots\ldots\mathbf{(2)}$$

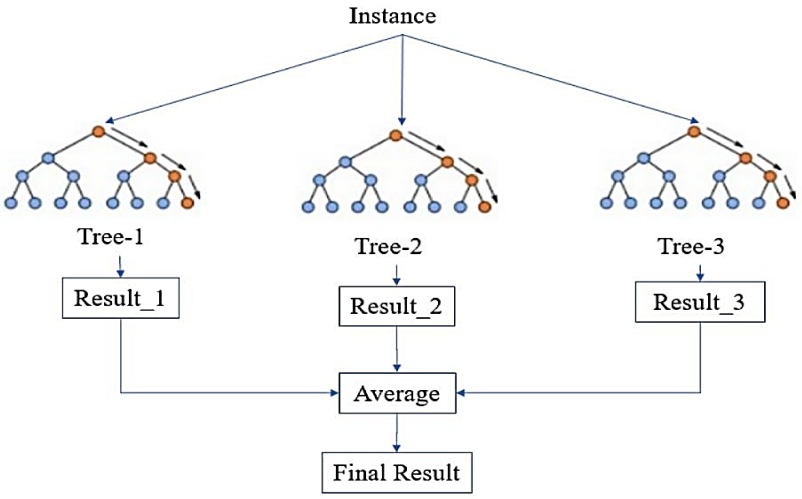
Keterangan:

Values (a) : Semua nilai yang mungkin pada himpunan kasus a.

Y_V : Subkelas dari Y dengan kelas v yang berkaitan dengan kelas a.

Y_a : Merupakan semua nilai yang sama dengan a.

Struktur dasar dari *Random Forest* ditunjukkan pada Gambar 2.1 [17].



Gambar 2.1 Struktur Sederhana Dari *Random Forest*

Setiap model sub-pohon melakukan pengambilan sampel acak dengan penggantian dari data pelatihan dan akhirnya menghasilkan hasil rata-rata dari semua sub-model [15]. Setiap sub-model dijalankan secara *pararel* tanpa adanya ketergantungan. Selain membangun setiap pohon menggunakan subset data yang berbeda, *Random Forest* berbeda dalam cara pembangunan pohon-pohon tersebut [20]. Pada pohon keputusan standar, setiap simpul bercabang menggunakan

keputusan optimum untuk pembagian di antara semua *variabel*, sehingga mengurangi *entropi* akibat pembagian himpunan data yang diwakili oleh simpul induk. Pada *Random Forest*, titik pembagian setiap simpul dipilih secara acak dari titik pembagian terbaik di antara *subset prediktor* [2 7] . Dengan demikian, *Random Forest* menghindari *overfitting*, yang umum terjadi pada pohon keputusan tunggal yang dalam [5].

2.8 Algoritma *Naive Bayes*

Thomas Baye adalah seorang ilmuwan Inggris, merancang algoritma *Naive Bayes*. Pengklasifikasi ini menggunakan teknik statistik dan probabilistik [28]. Dalam aplikasi pembelajaran mesin, model *Naive Bayes* banyak digunakan karena memungkinkan setiap karakteristik memiliki pengaruh yang sama terhadap hasil. Karena efisiensi pemrosesannya yang proporsional, pendekatan *Naive Bayes* menarik dan praktis dalam berbagai aplikasi. Kemungkinan *prior*, *posterior*, dan *kondisional* kelas adalah tiga komponen utama pengklasifikasi *Naive Bayes*. [29]. Rumus Teorema Bayes ditunjukkan pada Persamaan [30]

$$\mathbf{P(H/X)} = \frac{\mathbf{P(X/H)\ P(H)}}{\mathbf{P(H)}}$$

Keterangan:

X : Kelas data yang belum diketahui

H : *Hipotesa* data X adalah kelas spesifik

P(H|X) : Kemungkinan *Hipotesa* H berdasarkan keadaan X (*posteriori prob*)

P(H) : Kemungkinan *Hipotesa* H (*prior prob*)

P(X|H) : Kemungkinan X berdasarkan keadaan tersebut

P(X) : Kemungkinan dari X

2.9 *K-Nearest Neighbor* (K-NN)

Untuk mengklasifikasikan objek yang sangat dekat, pendekatan supervisi *K-Nearest Neighbor* (K-NN) bekerja dengan menghitung jarak terkecil antara informasi yang akan dinilai dengan k tetangga pada data pelatihan [31]. Selanjutnya, rumus diterapkan untuk menentukan tingkat kemiripan antara

Database vektor dan dataset pelatihan yang dikategorikan. Teorema K-NN memungkinkan estimasi jarak universal berikut:

$$\mathbf{d_i} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_{ij}} - \mathbf{p_j})^2}$$

Keterangan

d_i : Jarak sampel

X_{ij} : Data sampel pengetahuan

p_j : Data input var ke-j

n : Jumlah sampel

Berikut tata cara penggunaan metode K-NN [31]:

1. Tentukan jumlah tetangga terdekat, atau parameter k.
2. Tentukan kuadrat jarak Euclidean objek terhadap data instruksi yang diberikan.
3. Urutkan hasil nomor 2 secara berurutan dari nilai tinggi ke nilai rendah, atau dalam urutan menaik.
4. Kumpulkan klasifikasi tetangga terdekat (kategori Y) berdasarkan nilai k.
5. Kategori objek dapat diantisipasi dengan menggunakan kategori tetangga mayoritas.

2.10 Seleksi Fitur

Biasanya, pemilihan fitur digunakan untuk menghilangkan fitur yang tidak diperlukan, meminimalkan dimensi, meningkatkan kinerja algoritma pengklasifikasi hingga 12 kali lipat, dan memilih fitur terbaik. Subset dari kumpulan fitur digunakan dalam metode ini, dan hasilnya adalah keluaran yang sebanding dengan keseluruhan kumpulan fitur [32]. Pemilihan fitur merupakan tahapan penting dalam proses klasifikasi karena karakteristik yang dipilih mempunyai pengaruh besar terhadap keakuratan klasifikasi. Mengurangi fitur yang tidak relevan merupakan langkah penting dalam klasifikasi kumpulan data dengan banyak fitur [33]. Efisiensi klasifikasi yang dibangun dapat sangat dipengaruhi oleh pemilihan fitur, yang terkadang juga dapat meningkatkan

keakuratan klasifikasi berikutnya [32]. Untuk mencapai hasil terbaik, para peneliti telah membandingkan pemilihan fitur dan algoritma klasifikasi [9].

2.11 *Particel Sward Optimazion* (PSO)

PSO adalah singkatan dari *Particle Swarm Optimization*, [15], dimodelkan pada perilaku serangga yang berkerumun termasuk burung, rayap, semut, dan lebah. Algoritma PSO meniru interaksi sosial hewan-hewan ini. Perilaku sosial mencakup setiap tindakan yang dilakukan oleh seorang individu serta pengaruh anggota kelompok lainnya. Misalnya, kata "partikel" menggambarkan sekelompok burung. Dengan kecerdasannya masing-masing, setiap partikel atau individu bertindak secara terdistribusi, dan kecerdasannya juga memengaruhi perilaku kelompok agregat. Hasilnya, tidak peduli seberapa jauh mereka dari kelompok tersebut, anggota kelompok lainnya dapat dengan cepat mengikuti jika satu partikel atau burung menemukan jalur yang benar atau pendek menuju sumber makanan. Kawanan dianggap berukuran pasti atau tetap dalam optimasi multivariat, dengan setiap partikel dimulai dari lokasi acak dalam ruang multidimensi. Dipercaya bahwa setiap partikel memiliki dua karakteristik: lokasi dan kecepatan. Setiap partikel dalam ruang tertentu mengingat lokasi optimalnya yang muncul atau ditemukan sehubungan dengan nilai fungsi tujuan atau sumber makanan. Setelah memberikan informasi atau lokasi yang diinginkan kepada partikel lain, setiap partikel mengubah posisi dan kecepatannya sendiri sesuai dengan posisi informasi pilihan partikel lain. Perilaku burung dalam kawanan, misalnya.

Burung pada dasarnya buta huruf dan menganut perilaku tertentu, seperti berikut:

1. Jalur terbang seekor burung pada umumnya mengikuti spesiesnya.
2. Untuk menjaga jarak yang masuk akal antara masing-masing burung dalam kawanan, burung-burung akan diatur berdasarkan lokasi rata-ratanya.

Akibatnya, perilaku sekawanan burung akan bergantung pada kombinasi tiga faktor dasar berikut:

1. Kohesi, atau kemampuan terbang bersama.

2. Separasi, jangan terlalu dekat.
3. Penyesuaian, tau menuju ke arah umum yang sama.

Hasilnya, PSO dikembangkan menggunakan model di bawah ini:

1. Meski tidak secara langsung, burung lain mengikuti jalur makanannya.
2. Satu unsur tergantung pada mentalitas masing-masing burung. Kenangan ini berasal dari masa lalu.

Model ini berulang kali disimulasikan dalam ruang dimensi tertentu, dengan setiap iterasi posisi partikel secara progresif mengarah ke hasil yang diinginkan minimalkan atau maksimalkan fungsi. Hal ini terus berlanjut hingga kriteria penguatan lain tersedia, atau jumlah iterasi maksimum tercapai.

2.12 *Cross Validation*

Cross validation adalah salah satu metrik untuk mengukur hasil algoritma klasifikasi. Sementara itu, validasi kolom K adalah salah satu metode untuk mengetahui tingkat kesuksesan rata-rata sistem klasifikasi. *K-fold validation* akan mengacak sebuah *dataset* secara silang, yang memungkinkan sistem untuk diuji pada berbagai dataset yang sudah diacak sebelumnya. Tujuannya adalah untuk mencegah data mendominasi pembelajaran model klasifikasi. Pembahagian data menjadi *n-fold* yang diinginkan akan digunakan untuk validasi *k-fold*. Misalnya, jika data dibagi menjadi 5 akan menghasilkan 5 partisi data dengan ukuran yang sama, misalnya D1, D2, dan D3. Setelah itu, proses pengujian dan pelatihan dilakukan sebanyak jumlah *fold*. Data partisi n akan menjadi *dataset* uji dan *dataset* pelatihan pada setiap *iterasi* ke-i. Tabel *Confusion Matrix* mengandung empat kombinasi nilai aktual dan prediksi. Berdasarkan Tabel.2 memberikan penjelasan mengenai empat istilah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) yang mewakili hasil proses klasifikasi dalam *matriks konfusi*.

Tabel 2.2 Pengujian *Confusion Matrix*

Classification	***Predicted Class***	
	True	***False***
Actual: True	*True Positif (TP)*	*False Negatif (FN)*
Actual: False	*False Positif (FP)*	*True Negatif (TN)*

Setiap *iterasi* dihitung *Accuracy, Presisi, dan Recall* menggunakan rumus berikut [34]:

2.12.1 Akurasi

Akurasi adalah salah satu ukuran yang digunakan untuk mendorong model klasifikasi, Sederhananya, persentase prediksi model kami yang menjadi kenyataan. Seperti yang ditunjukkan di bawah, akurasi juga dapat dihitung dalam bentuk positif dan negatif [9]:

$$\textbf{Akurasi} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}}$$

Keterangan:

TP : *True Positif*

TN : *True Negatif*

FP : *False Positif*

FN : *False Negatif*

Besarnya akurasi klasifikasi ditunjukkan dengan skor TP (*True Positive*) dan TN (*True Negative*). Secara umum keakuratan klasifikasi semakin tinggi dengan nilai TP dan TN yang lebih besar. *False Positive* (FP) terjadi ketika label prediksi keluaran positif tetapi nilai sebenarnya salah. *False Negative* (FN) terjadi ketika label prediksi keluaran salah tetapi hasil sebenarnya benar [9].

2.12.2 *Precision*

Rasio hal-hal terkait yang dipilih dengan semua item yang dipilih dalam *Matrix Confusion* dikenal sebagai akurasi. Lebih jauh lagi, akurasi adalah tingkat kesesuaian antara data yang diharapkan pengguna dan respons sistem [9]**.**

$$\boldsymbol{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}}$$

Keterangan:

TP = *True Positif*

FP = *False Positif*

2.12.3 *Recall*

Kemungkinan item yang bersangkutan akan terpilih disebut *recall*. Nilai ini dihitung dengan membagi jumlah total rekomendasi yang relevan baik yang dipilih maupun yang tidak dipilih dengan jumlah rekomendasi relevan yang telah dipilih pengguna [9].

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

Keterangan:

TP = *True Positif*

FN = *False Negatif*

2.12.4 Kurva ROC dan AUC

Pengukuran kinerja machine learning sangat penting. Oleh karena itu, kita dapat mempercayai kurva AUC dan ROC untuk tugas klasifikasi. Kurva AUC (*Area Under the Curve*) dan ROC (*Receiver Operating Characteristics*) digunakan untuk mengevaluasi atau memvisualisasikan kinerja masalah klasifikasi multikelas. Salah satu ukuran penilaian paling penting untuk menunjukkan seberapa efektif kinerja setiap model kategorisasi. Ini juga dikenal sebagai *Area Under Receiver Operating Characteristic*, atau AUROC [9].

Berikut adalah standar tabel kategori pengklasifikasian berdasarkan nilai AUC pada Tabel 2.3 [35]

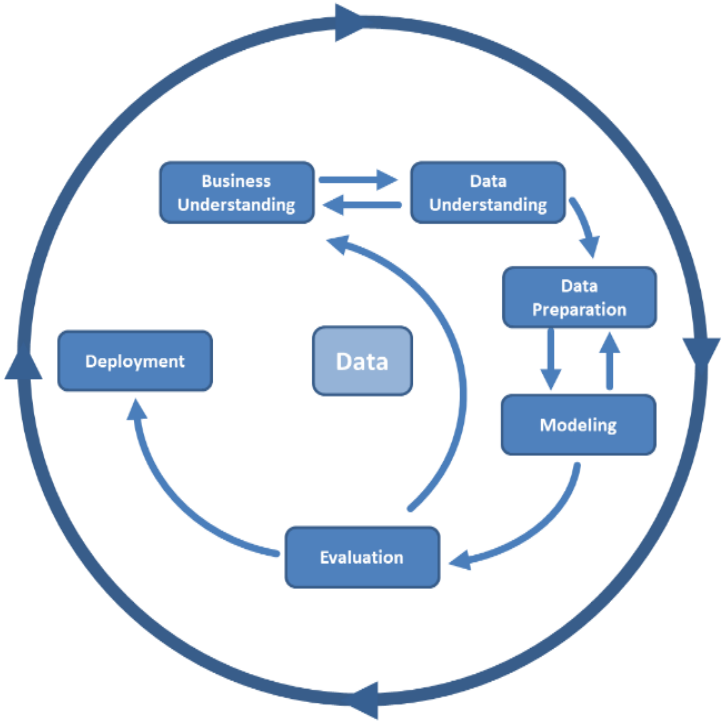
Tabel 2.3 Kategori Pengklasifikasian Berdasarkan Nilai AUC

Nilai AUC	Kategori Pengklasifikasian
0.90 - 1.00	*Excellent*
0.80 - 0.90	*Good*
0.70 - 0.80	*Fair*
0.60 - 0.70	*Poor*
0.50 - 0.60	*Fail*

2.13 CRISP-DM

Pada tahun 1996, analis dari berbagai sektor, termasuk Daimler Chrysler, SPSS, dan NCR, mengembangkan Proses Standar Lintas Industri untuk

Penambangan Data, atau CRISP-DM. Proses penambangan data didefinisikan oleh CRISP-DM sebagai teknik umum untuk menyelesaikan masalah bagi organisasi bisnis atau penelitian [36]. Siklus hidup proyek penambangan data di CRISP-DM dipecah menjadi enam tahap. Gambar 2.2 di bawah ini memberikan gambaran siklus hidup CRISP-DM secara umum [37].



Gambar 2.2 Siklus Hidup Dalam CRISP_DM

Gambar 2.2 menggambarkan bahwa siklus hidup CRISP-DM memiliki enam fase berbeda, yaitu sebagai berikut [36]:

1. *Business Understanding Phase* atau Fase Pemahaman Bisnis
    a. Tetapkan tuntutan dan tujuan proyek dalam parameter unit bisnis atau penelitian secara keseluruhan.
    b. Mengubah tujuan dan batasan menjadi definisi dan rumus masalah data mining.
    c. Membangun strategi awal untuk mencapai tujuan.
2. Data *Understanding Phase* atau Fase Pemahaman Data
    a. Mengumpulkan data.
    b. Mengevaluasi kualitas.

c. Pelajari lebih lanjut tentang informasi dan keahlian yang disertakan dalam data.
d. Pilih beberapa kumpulan data yang menurut Anda mungkin menyertakan pola masalah, jika Anda mau.

3. Data *Preparation Phase* atau Fase Pengolahan Data
   a. Karena ini adalah kumpulan data yang diperlukan untuk seluruh langkah selanjutnya, mulailah menyiapkan data yang dapat diakses. Ini adalah fase menantang yang perlu diselesaikan dengan hati-hati.
   b. Mengubah beberapa variabel jika diperlukan.
   c. Siapkan data awal untuk perangkat pemodelan.
4. *Modeling Phase* atau Fase Pemodelan
   a. Memilih dan menerapkan metode pemodelan yang tepat.
   b. Mengatur aturan model untuk mengoptimalkan hasil.
   c. Perlu diingat bahwa beberapa metode mungkin digunakan untuk masalah data mining yang sama.
   d. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk membuat data sesuai dengan spesifikasi yang diperlukan oleh teknik data mining tertentu.
5. *Evaluation Phase* atau Fase Evaluasi
   a. Sebelum diterapkan, nilai kemanjuran dan kualitas satu atau lebih model yang digunakan selama proses pemodelan.
   b. Menentukan apakah model memenuhi tujuan pada fase awal.
   c. Periksa untuk melihat apakah ada masalah penelitian atau bisnis yang tidak dapat diselesaikan.
   d. membuat keputusan tentang penggunaan hasil data mining.
6. *Deployment Phase* atau Fase Penyebaran
   a. Menggunakan model yang telah dibuat.
   b. Contoh penerapan dasar adalah menghasilkan laporan yang menunjukkan penyelesaian proyek.