BAB 2
TINJAUAN PUSTAKA

2.1. Banjir

Banjir merupakan peristiwa yang disebabkan oleh akumulasi menumpuknya air jatuh yang tidak dapat ditampung oleh tanah. Fenomena ini terjadi karena air yang jatuh ke dataran tidak memiliki daerah resapan. Banjir didefinisikan sebagai situasi di mana suatu daerah terendam oleh sejumlah besar air. Banjir juga dapat diprediksi dengan memperhatikan curah hujan dan aliran air (Miriam Budiardjo, 2020).

Berdasarkan penelitian (Sulistyo & Pranoto, 2020) menemukan bahwa faktor penyebab banjir diantaranya yaitu topografi wilayah daerah dataran rendah tipe cekungan, dan adanya sedimen serta sampah di dasar saluran drainase sehingga menghambat aliran air. Penelitian lainnya yang dilakukan (Sundari, 2020) limpasan permukaan yang berlebihan menyebabkan banjir di Kota Samarinda karena tidak ada yang menampung limpasan tersebut dalam badan sungai sehingga air meluap. Banyak faktor penyebab terjadinya banjir. Tetapi biasanya penyebab banjir dapat dikelompokkan menjadi 2 kategori, yaitu banjir yang terjadi karena alam dan banjir yang terjadi karena tindakan manusia.

a) Faktor alam

   Faktor alam seperti curah hujan yang tinggi, pengaruh fisiografi, topografi setempat, kapasitas sungai, pengaruh air pasang dan lain-lain.

b) Faktor Manusia

   Faktor manusia biasanya terjadi akibat pertumbuhan penduduk, kawasan kumuh sampah, kerusakan bangunan pengendali air serta rusaknya hutan. Banjir merupakan salah satu bencana hidrometeorologi, merupakan bencana yang diakibatkan oleh parameter meteorologi seperti curah hujan, kelembapan, temperatur, dan angin (Simbolon et al., 2022).

2.2. Data Mining

Data mining adalah salah satu bidang ilmu yang digunakan untuk mengatasi masalah pengambilan informasi dari database besar dengan teknik dari statistik, pembelajaran mesin, visualisasi data, pengenalan pola dan *database* (Yuli Mardi, 2019). Secara umum, proses data mining dapat dikelompokkan menjadi dua kategori, yaitu deskriptif dan prediktif. Kategori deskriptif dalam data mining digunakan untuk memperoleh pemahaman yang lebih mendalam tentang data yang sedang diamati, atau dengan kata lain, bagaimana mengidentifikasi karakteristik dari data tersebut. Di sisi lain, kategori prediktif dalam data mining merupakan proses yang digunakan untuk mengembangkan model pengetahuan guna melakukan prediksi atau ramalan (Siswa, 2023). Pada Data Mining sendiri terdapat 7 metode yaitu Clustering, Asosiasi, Regresi, Forecasting, Sequence, Devition Analysis, dan Klasifikasi (Marisa et al., 2021).

2.2.1. Klasifikasi

Klasifikasi merupakan suatu proses dimana model atau fungsi dibuat untuk mengidentifikasi dan membedakan berbagai kelas data atau konsep tertentu, dengan tujuan dapat digunakan untuk meramalkan kelas dari objek yang tidak memiliki label kelasnya (Kamber, 2006).

Klasifikasi merupakan teknik membangun suatu model yang bisa mengklasifikasikan suatu obek berdasarkan atribut-atributnya. Kelas target sudah tersedia dalam data sebelumnya, sehingga fokusnya adalah bagaimana mempelajari data yang ada agar klasifikator bisa mengklasifikasikan sendiri (Jollyta et al., 2021).

2.2.2. CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) merupakan proses model standar yang sering digunakan dalam data mining (Schröer et al., 2021):

a) *Business Understanding*

   Tahapan ini bertujuan memahami situasi serta permasalahan yang ada pada objek untuk menentukan tujuan.

b) *Data Understanding*

Tahapan ini bertujuan untuk mengumpulkan dan mengeksplorasi data untuk melihat kualitas data. Eksplorasi data dilakukan melalui pengecekan nilai yang hilang, data yang berlebihan dan outliers dalam data yang akan digunakan.

c) *Data Preparation*

Dilakukan pembersihan data dari *missing value* pada data yang akan digunakan. Setelah itu dilakukan *data selection* yaitu menghapus atribut yang tidak diperlukan untuk tahap pembuatan model klasifikasi. Kemudian dilakukan *data transformation* yaitu mengubah data nominal menjadi data numerik dan melakukan normalisasi pada atribut tertentu.

d) *Modelling*

Pada tahap ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal.

e) *Evaluation*

Pada tahap ini dilakukan evaluasi setiap model skenario yang telah dibuat. Menggunakan *confusion matrix* dapat diketahui nilai true *positive, true negative, false positive* dan *false negative* pada suatu model. Hasil evaluasi setiap model skenario berupa nilai *accuracy, precision, recall* dan *f1-score*.

f) *Deployment*

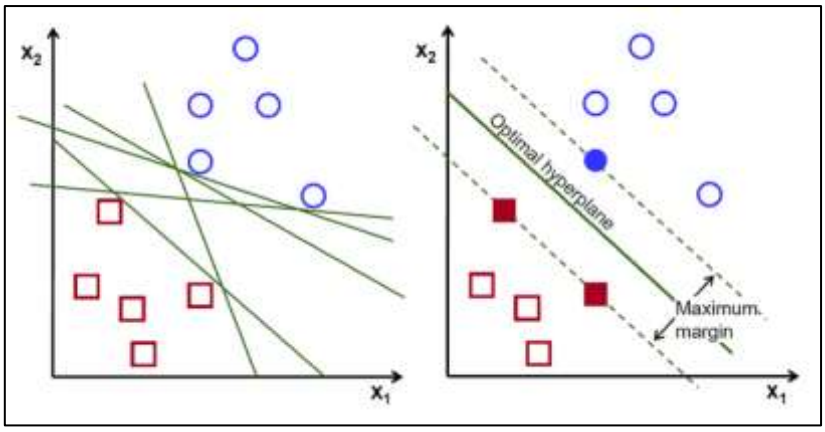
Tahap terakhir dalam model CRISP-DM adalah Deployment. Perencanaan untuk Deployment dimulai selama Business Understanding dan harus menggabungkan tidak hanya bagaimana untuk menghasilkan nilai model, namun juga bagaimana mengkonversi skor keputusan, dan bagaimana menggabungkan keputusan pada sistem operasional.

2.2.3. *Support Vector Machine*

Support Vector Machine adalah metode machine learning yang bekerja berdasarkan prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan *hyperlane* terbaik yang memisahkan dua buah kelas pada *input space*. SVM merupakan sistem pembelajaran yang menggunakan ruang hipotesis

berupa fungsi-fungsi *linear* dalam sebuah ruang fitur berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan learning bias yang berasal dari teori pembelajaran statistik. Tujuan dari algoritma SVM adalah untuk menemukan hyperlane dalam ruang N-dimensi (N- jumlah fitur) yang secara jelas mengklasifikasikan titik data (Dharmawan, 2021).

SVM adalah salah satu metode klasifikasi yang paling umum digunakan oleh banyak peneliti saat ini karena menawarkan kemampuan yang sangat baik dalam banyak aplikasi. Ide dasar SVM adalah memaksimalkan batas *hyperplane*, yang diilustrasikan seperti gambar berikut:



Gambar 2. 1 (a) hyperlane non optimal (b) hyperlane optimal

Gambar (a) menunjukkan sejumlah pilihan hyperplane dimungkinkan untuk kumpulan data, sedangkan gambar (b) adalah *hyperplane* dengan margin yang paling maksimal. Meskipun sebenarnya pada gambar (a) bisa juga menggunakan *hyperplane* sembarang, namun hyperplane dengan margin yang maksimal akan memberikan generalisasi yang lebih baik pada metode klasifikasi. Konsep klasifikasi menggunakan SVM dapat dijelaskan secara sederhana sebagai upaya untuk mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah kelas data pada input space (Nugroho et al., 2023). Dalam memperoleh *hyperplane* pada SVM, dapat menggunakan persamaan:

$$(w \,.\, x_i) + b = 0 \qquad (2.1)$$

Di dalam data xi, yang termasuk pada kelas -1 dapat dirumuskan seperti

persamaan:

$$(w \,.\, x_i) + b \leq 1, y_i = -1 \qquad (2.2)$$

Sedangkan data data xi, yang termasuk pada kelas +1 dapat dirumuskan seperti persamaan:

$$(w \,.\, x_i) + b \geq 1, y_i = 1 \qquad (2.3)$$

Keterangan:

Xi = data ke -i

W = niai bobot support vector yang tegak lurus dengan hyperplane

b = niai bias

$Yi = kelas\ data\ ke - i$

Dalam proses klasifikasi dengan SVM biasanya ditemui kondisi dimana *kernel linear* bekerja tidak optimal yang mengakibatkan hasil klasifikasi terhadap data menjadi buruk. Hal tersebut dapat diatasi dengan menggunakan *kernel non-linear* dengan memanfaatkan *kernel trick.* Dengan memanfaatkan *kernel trick*, akan dilakukan *mapping data input* ke *feature space* yang dimensinya lebih tinggi sehingga membuat data input yang dihasilkan akan terpisah secara *linear* dan membentuk *hyperplane* yang optimal (Shandra et al., 2019). Berikut persamaan dari setiap kernel SVM:

Type of SVM	Mercer Kernel
Gaussian or Radial Basis Function (RBF)	$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$
Linear	$K(x_1, x_2) = x_1^{\mathsf{T}} x_2$
Polynomial	$K(x_1, x_2) = \left(x_1^{\mathsf{T}} x_2 + 1\right)^{\rho}$
Sigmoid	$K(x_1, x_2) = \tanh\left(\beta_0 x_1^{\mathsf{T}} x_2 + \beta_1\right)$

Gambar 2. 2 Persamaan kernel SVM

Pada penelitian (Mase et al., 2018) kernel RBF merupakan kernel yang efektif dalam menangani data kompleks juga memberikan hasil yang baik dalam klasifikasi data. Maka akan diterapkan kernel RBF pada penelitian ini.

2.2.4. *Genetic Algorithm*

Genetic Algorithm meniru kejadian alam untuk menyelesaikan suatu masalah yaitu dengan menggabungkan teori reproduksi, seleksi alam, dan teori evolusi Darwin (Ramdhani et al., 2022) *Genetic Algorithm* sebuah paradigma perangkat lunak yang relatif baru dan populer (Oktarina & Hajjah, 2019). Fase utamanya evolusi adalah *selection, crossover, mutation. Genetic Algorithm* melakukan proses pencarian nilai optimal pada beberapa titik secara bersamaan dalam satu generasi (Khotimah, 2020). Metode *Genetic Algorithm* ini dapat diterapkan pada sistem data mining yang mengklasifikasikan data untuk mendapatkan informasi yang berguna dalam data mining (Yusuf Bakhtiar, 2020). Ada beberapa langkah yang diperlukan di dalam GA yaitu:

a) Inisialisasi Populasi Awal

   Misalkan kita ingin mencari solusi terbaik untuk sebuah permasalahan dengan menggunakan angka-angka dari 0 hingga 9. Kita memulai dengan membuat populasi awal yang terdiri dari beberapa individu secara acak. Misalnya, kita membuat populasi awal sebanyak 10 individu dengan panjang kromosom (jumlah angka dalam setiap individu) sebanyak 5. Contoh: [3, 1, 4, 2, 7], [0, 9, 6, 8, 5], [2, 4, 6, 3, 1], dst.

b) Evaluasi Individu

   Setiap individu dalam populasi dievaluasi menggunakan suatu fungsi evaluasi yang biasa disebut sebagai "fitness function" untuk menentukan seberapa baik individu tersebut. Dalam contoh ini, misalnya kita menggunakan fungsi evaluasi sederhana yang menjumlahkan semua angka dalam individu. Semakin besar jumlah total angka dalam individu, semakin baik hasilnya.

c) Seleksi Orangtua

   Individu yang memiliki kinerja (fitness) lebih baik memiliki peluang yang lebih besar untuk dipilih sebagai orangtua. Misalnya, kita menggunakan metode seleksi turnamen sederhana, di mana dua individu dipilih secara acak dan individu dengan fitness yang lebih tinggi dipilih sebagai orangtua.

d) Rekombinasi (*Crossover*)

Pasangan orangtua dipilih untuk melakukan rekombinasi (*crossover*) genetik. Misalnya, kita menggunakan crossover satu titik, di mana kita memilih titik pemutusan acak dalam kromosom dan menukar segmen kromosom antara kedua orangtua. Contoh: Orangtua 1: [3, 1, 4, 2, 7], Orangtua 2: [0, 9, 6, 8, 5]. Pemutusan titik: 3. Hasil crossover: Anak 1: [3, 1, 4, 8, 5], Anak 2: [0, 9, 6, 2, 7]

e) Mutasi

Untuk memperkenalkan variasi genetik baru, kita menerapkan mutasi pada beberapa individu secara acak. Misalnya, kita memilih individu secara acak dan mengganti salah satu angka dalam kromosomnya dengan angka acak. Contoh: Individu terpilih: [0, 9, 6, 2, 7]. Mutasi pada posisi ke-2. Hasil mutasi: [0, 3, 6, 2, 7].

f) Generasi Baru

Setelah langkah rekombinasi dan mutasi selesai, kita membentuk generasi baru yang terdiri dari individu-individu hasil rekombinasi, mutasi, dan beberapa individu terbaik dari generasi sebelumnya.

g) Evaluasi generasi baru

Langkah-langkah 2-6 diulang untuk generasi baru, yaitu evaluasi individu, seleksi orangtua, rekombinasi, dan mutasi.

2.2.5. Data *Pre-Processing*

Data Preprocessing merupakan teknik pengembangan data awal untuk mengubah data mentah menjadi format dan informasi yang lebih efisien dan berguna (Said et al., 2022). *Preprocessing* data harus dilakukan dalam proses *data mining*, karena tidak semua data atau atribut dalam data digunakan pada proses *data mining*. Terdapat beberapa teknik dalam data *prepocessing* antara lain: *data cleaning, data transformation, data balancing,* dan *data integration* (Marisa et al., 2021)

a. *Data Cleaning* merupakan proses memperbaiki atau menghapus data yang buruk, rusak, cacat, berlebihan, atau tidak lengkap dalam kumpulan data.

b. *Data Transformation* adalah proses mengubah data dari satu format atau struktur ke format lainnya. *Data transformation* digunakan untuk megubah data menjadi bentuk yang sesuai dalam proses *data mining*.
c. *Data Balancing* bertujuan untuk mengatasi ketidakseimbangan kelas dalam dataset. Ketidakseimbangan kelas terjadi ketika jumlah sampel pada kelas minoritas jauh lebih sedikit daripada jumlah sampel pada kelas mayoritas.
d. *Data Integration* adalah metode mengintegrasikan data dari sumber yang berbeda ke dalam satu kumpulan data dengan tujuan akhir menyediakan akses dan pengiriman data yang konsisten kepada pengguna di berbagai spektrum subjek dan tipe struktur, dan memenuhi kebutuhan informasi semua aplikasi dan proses bisnis. Proses integras data adalah salah satu komponen kunci dalam keselurahan proses pengelolaan data.

2.3. *K-Fold Cross Validation*

K-Fold Cross Validation merupakan salah satu jenis uji *cross validation* yang berfungsi untuk mengevaluasi kinerja proses dari suatu metode algoritma dengan cara membagi sampel data secara acak dan mengelompokkan data berdasarkan nilai K *k-fold*. Kemudian salah satu kelompok k-*fold* digunakan sebagai data uji dan sisa kelompok lainnya digunakan sebagai data latih (Tangguh Admojo, 2020). K-*Fold* adalah jenis *cross validation* yang banyak digunakan pada beberapa penelitian di bidang *data mining*.

2.4. *Confusion Matrix*

Confusion Matrix adalah matriks yang merepresentasikan hasil klasifikasi dalam suatu dataset. Confusion *Matrix* dapat menentukan kinerja dari klasifikasi, akurasi, presisi, *recall* (Belavkin et al., 2022). *Confusion Matrix* juga salah satu metode yang digunakan untuk mengevaluasi metode-metode klasisfikasi. Ilustrasi untuk tabel *confusion matrix* dapat dilihat pada Tabel 2.1.

Tabel 2. 1 Nilai Confusion Matrix

		Actual	
		True	***False***
Predicted	*Positive*	*True Positive (TP)*	*False Positive (FP)*
	Negative	*False Negative (FN)*	*True Negative (TN)*

Keterangan:

TP (True Positive) = Data positif yang terklasifikasi secara benar,

TN (True Negative) = Data negatif yang terklasifikasi secara benar,

FP (False Positive) = Data negatif yang terklasifikasi menjadi positif,

FN (False Negative) = Data positif yang terklasifikasi menjadi negatif.

Nilai dari Confusion Matrix tersebut nantinya akan digunakan untuk menghitung accuracy sebagai hasil nilai evaluasi dari suatu model.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (2.8)$$

2.5. Penelitian Terkait

Penelitian ini meninjau penelitian-penelitian terdahulu untuk menjadi salah satu acuan peneliti dalam melakukan penelitian, sehingga peneliti dapat memperkaya teori yang digunakan dalam mengkaji penelitian yang dilakukan. Berikut merupakan penelitian terdahulu berupa beberapa jurnal terkait dengan penelitian yang dilakukan penulis.

Tabel 2. 2 Penelitian Terdahulu

No	Peneliti	Judul	Hasil
1	(Siswa & Prihandoko, 2018)	Perbandingan Kinerja Algoritma C4. 5, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Dan Support Vector Machines Untuk Mendeteksi Penyakit Kanker Payudara	Hasil kinerja terbaik yang diuji menggunakan didapatkan bahwa algoritma Logistic Regression dan Support Vector Machines memiliki nilai akurasi tertinggi yang sama nilainya yaitu sebesar 0,968.
2	(Monika & Furqon, 2018)	Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak	Pengujian pada penelitian ini menggunakan jenis pengujian K-fold Cross Validation, dimana fold yang digunakan sebanyak 8 fold. Nilai akurasi terbaik yang dihasilkan pada penelitian ini adalah sebesar 63,11%.
3	(Istiadi & Rahman, 2020)	OPTIMISASI PARAMETER *SUPPORT VECTOR MACHINE* BERBASIS ALGORITMA GENETIKA PADA KLASIFIKASI TEKS PENGADUAN MASYARAKAT	Berdasarkan hasil pengujian menunjukkan AG mampu menghasilkan nilai-nilai parameter untuk SVM berdasarkan masing-masing kernelnya. Pengujian menunjukkan variasi jumlah data latih terhadap data uji berpengaruh nilai akurasi pada masing-masing kernel. Hasil kinerja terbaik terjadi pada *kernel linear* dengan nilai akurasi sebesar 85,37% pada rasio data latih terhadap data uji sebesar 80% : 20%.

Tabel 2. 3 Penelitian Terdahulu (Lanjutan)

No	Peneliti	Judul	Hasil
4	(Yusuf Bakhtiar, 2020)	KLASIFIKASI PENELITIAN DOSEN MENGGUNAKAN *NAÏVE BAYES CLASSIFIER* DAN ALGORITMA GENETIKA	Hasil penelitian mendapatkan kenaikan 26.06 % menjadi 78.61 % pada nilai akurasinya. Dari hasil penelitian ini bisa disimpulkanbahwa algoritma genetika mampu menaikan akurasi pada *Naive Bayes* dengan mengoptimalkan proses seleksifitur berupa kata yang ada pada abstrak tiap penelitian.
5	(Ramadhan & Khoirunnisa, 2021)	Klasifikasi Data Malaria Menggunakan Metode *Support Vector Machine*	Pada penelitian ini SVM mampu menghasilkan akurasi tertinggi 92.3% dimana hasil akurasi model SVMlebih unggul dengan gap akurasi rata-rata 25%.
6	(Normah et al., 2022)	Komparasi Algoritma *Support Vector Machine* Dan *Naïve Bayes* Dengan Algoritma Genetika Pada Analisis Sentimen Calon Gubernur Jabar 2018-2023	Pada penelitian tersebut model algoritma SVM dengan berbasis *Genetic Algorithm* mampu menjadi model terbaik dalam penelitian dan dapat memberikan hasil terbaik dengan menghasilkan rata-rata akurasi 93,03%.

Menurut referensi terkait, maka didapatkan perbedaan penelitian dari yang terdahulu sebagai dasar penelitian ini yaitu penelitian ini didasarkan pada algoritma dan atribut berbeda yang digunakan oleh penelitian sebelumnya dan dalam penelitian ini menggunakan algoritma SVM dalam mengklasifikasikan data banjir menggunakan GA pada seleksi fiturnya.