BAB 2

TINJAUAN PUSTAKA

2.1 Banjir

Di Indonesia, banjir merupakan jenis bencana alam yang paling sering terjadi. Banjir adalah debit aliran air sungai yang jauh lebih besar dari biasanya akibat hujan yang turun di suatu tempat tertentu secara terus menerus. dengan memperhatikan curah hujan dan pergerakan air dapat membantu mengantisipasi datangnya banjir (Taryana et al., 2022).

Penelitian lainnya yang dilakukan (Sulaiman et al., 2020) penyebab banjir di Kota Samarinda terjadi akibat berlebihnya limpasan permukaan dan tidak tertampungnya limpasan tersebut dalam badan sungai sehingga air meluap.

Penelitian oleh (Savitri Febiyani & Prita Wardhani, 2019) telah melakukan penelitian dengan fokus pada pengamatan terhadap jumlah kejadian banjir, jumlah kejadian kekeringan, jumlah kejadian angin puting beliung, jumlah kejadian gelombang pasang, jumlah curah hujan, jumlah hari hujan, kecepatan angin, kelembaban, suhu rata-rata, tekanan udara, penyinaran matahari. Penelitian ini menunjukkan pengaruh signifikan dari jumlah kejadian angin puting beliung, jumlah hari hujan, dan penyinaran matahari terhadap jumlah kejadian banjir.

2.2 Data Mining

Data mining adalah proses menggunakan pendekatan atau metode tertentu untuk mencari pola atau informasi yang berguna dalam sekumpulan data. Metode, teknik, dan algoritma data mining bisa sangat berbeda. Tujuan dan seluruh proses *Knowledge Discovery in Databases* (KDD) benar-benar menentukan pendekatan atau algoritme mana yang terbaik (Fatmawati & Windarto, 2018). Sedangkan pengertian data mining menurut (Laila Sari et al., 2022) data mining adalah proses yang mengekstraksi dan mengidentifikasi informasi yang dapat digunakan dan pengetahuan terkait dari beberapa kumpulan data besar menggunakan teknik statistik, matematika, kecerdasan buatan, dan pembelajaran

mesin.Berdasarkan beberapa definisi-definisi yang telah dijabarkan di atas, dapat disimpulkan bahwa data mining merupakan proses melakukan pengambilan data secara besar untuk ekstraksi dan identifikasi informasi bermanfaat yang terdapat pada sekumpulan data dengan menggunakan teknik-teknik pembelajaran komputer atau machine learning. Menurut (Santosa & Umam, 2018) adapun tugas – tugas yang biasa dilakukan data mining antara lain:

1) Klastering
2) Klasifikasi
3) Regresi atau estimasi
4) Asosiasi

2.2.1 Klasifikasi

Klaifikasi adalah pengelompokan suatu objek-objek berdasarkan kelompok yang sudah ada, yang membutuhkan data pelatihan kelompok atau kelas yang berlabel. Algoritma klasifikasi yang sering digunakan antara lain K-Nearest Neighbor dan C4.5. (Santosa & Umam, 2018) Klasifikasi merupakan istilah yang didapat dari bahasa Belanda, yakni classificatie, yang berasal dari bahasa Prancis classification. Definisi ini mengacu pada teknik untuk sistemati data berdasarkan kategori atau aturan tertentu dengan cara yang terstruktur (Nasrullah & Harun, 2023)

2.2.2 *K-Nearest Neighbor*

Algoritma *k-Nearest Neighbour* (k-NN) adalah algoritma klasifikasi sederhana yang banyak digunakan. Algoritma ini termasuk dalam kategori lazy learner karena tidak belajar dari data latih, melainkan langsung menggunakan tetangga terdekat dalam klasifikasinya. Tujuan penggunaan k-NN adalah memprediksi kelompok objek baru berdasarkan tetangga terdekatnya. Algoritma ini menggunakan representasi data dalam bentuk ruang vector untuk mengidentifikasi jarak antara dua titik yaitu pada data train(x) dan titik pada data testing (y) (Taghfirul Azhima Yoga Siswa, S.Kom., 2023). Adapun tahapan kerja algoritma kNN adalah sebagai (Darmayanti et al., 2021)

1. Memilih nilai` k, yaitu jumlah tetangga terdekat yang akan digunakan dalam klasifikasi.
2. Melakukan perhitungan Euclidian distance antara objek yang ingin diklasifikasikan dengan setiap objek dalam data training.
3. Urutkan hasil perhitungan Euclidian distance dari yang tertinggi hingga terendah.
4. Melakukan pengumpulan kategori Y (Klasifikasi Nearest Neighbor dari suatu objek berdasarkan nilai k)
5. Membuat keputusan nearest neighbor dari suatu objek dengan cara mengambil hasil nilai yang terbanyak/sering muncul objek tersebut

Euclidean distance merupakan salah satu cara untuk menghitung seberapa jauh jarak dua titik dalam ruang Euclidean, yang dapat berupa dua dimensi, tiga dimensi, atau bahkan lebih kompleks. Untuk perhitungan rumus jarak Euclidean Distance dari *K-Nearest Neighbor* dapat dijelaskan pada persamaan (Susanto et al., 2018) adalah sebagai berikut

$$Euclidien\ distance = \sqrt{\sum_{i=1}^{p} (\alpha_{k-} b_k)^2} \quad (2.\ 1)$$

Sumber : (Jatmiko Indriyanto, 2021)

Keterangan :

α_K : Sampel data

b_k : Data uji atau Data testing

P : Dimensi data

i : variable data

2.2.3 Data Preprocessing

Data preprocessing adalah tahapan awal dari data mining untuk menghasilkan analisis yang lebih akurat dalam pemakaian teknik-teknik machine learning (Santosa & Umam, 2018).

Data preprocessing merupakan tahap pertama dalam pengolahan data yang melibatkan serangkaian langkah untuk menyiapkan data yang relevan dan sesuai guna perhitungan. Tujuan utamanya adalah melakukan pembersihan data dengan menghilangkan data yang duplikat, tidak konsisten, dan kosong. Selanjutnya, dilakukan pemilihan field yang penting untuk digunakan dalam pemodelan. Tahap terakhir adalah transformasi data menjadi format binominal agar dapat dianalisis dengan lebih efektif (Sartika & Saluza, 2022). Terdapat beberapa teknik dalam data preprocessing anatara lain : data cleaning,data transformation ,data integration (Taghfirul Azhima Yoga Siswa, S.Kom., 2023)

a) Data Cleaning yang bertujuan untuk menghapus noise dan memperbaiki ketidak konsistenan data
b) Data Transformation dapat diterapkan untuk mengubah skala data menjadi rentang yang lebih kecil, misalnya dari 0 hingga 10. Transformasi data ini memiliki manfaat meningkatkan akurasi dan efisiensi algoritma data mining yang melibatkan pengukuran jarak
c) Data Balancing

   Data balancing merupakan kondisi data tidak seimbang ketika kelas-kelas dalam masalah klasfikasi tidak diwakili secara sama untuk mengatasi masalah ini pada tahapan modeling, proses oversampling acak dapat digunakan. Dalam metode ini, baris data dengan kelas terkecil akan diduplikasi secara acak hingga jumlah kelas terbanyak tersedia (Sitorus et al., 2020). Namun, pada metode ini dapat menyebabkan overfitting karena dilakukannya duplikasi data yang sudah ada, yang mengakibatkan pengklasifikasi terpapar pada informasi yang sama.untuk mengatasi masalah tersebut, digunakan metode synthetic minority oversampling technique (SMOTE). Metode ini merupakan salah satu metode oversampling yang paling populer yang digunakan. SMOTE dilakukan dengan menambahkan data sintetis pada kelas minoritas (WIJAYANTI et al., 2021). Dalam implementasinya, modul `imblearn.over_sampling` di Python dan fungsi 'SMOTE (sampling_strategy)' digunakan. Metode ini membantu dalam menciptakan variasi data sintetis

sehingga mengurangi kemungkinan overfitting dan memberikan representasi yang lebih baik untuk kelas minoritas.

d) Data Integration yang menggabungkan data dari berbagai sumber menjadi terkait dalam penyimpanan data

2.2.4 K-*Fold Cross-validation*

K-Fold Cross-Validation, merupakan suatu metode dimana suatu kumpulan data dibagi menjadi dua bagian yaitu data latih dan data uji untuk k kelompok, dimana jumlah data latih dan data uji pada masing-masing kelompok adalah sama. Sehingga dilakukan pengulangan sebanyak-banyaknya untuk melakukan pelatihan dan evaluasi (Id, 2021).

2.2.5 *Confusion* Matrix

Confusion matrix merupakan ukuran N x N yang dapat digunakan untuk masalah klasifikasi, dimana N adalah jumlah kelas yang diharapkan di prediksi (Id, 2021). Ada empat kolom dalam confusion matrix yaitu (Id, 2021):

Tabel 2. 1 confusion matrix

Class	Actual = Yes	Actual = No
Predicted = Yes	TP	FP
Predicted = No	FN	TN

Sumber : (Id, 2021)

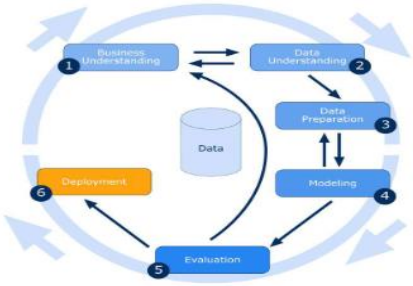
1) TP (True Positive) merupakan jumlah titik data berlabel yang nilainya teridentifikasi dengan benar.
2) TP (True Negative) merupakan jumlah titik data berlabel no yang nilainya tidak teridentifikasi dengan benar.
3) FP (False Positives) merupakan banyaknya titik data yang berlabel “ya” dengan suatu nilai sebenarnya diidentifikasi secara tidak benar.
4) FN (False Negative) merupakan jumlah titik data berlabel no yang nilai aktualnya teridentifikasi dengan benar.

Nilai akurasi pada model dapat dihitung dengan menggunakan rumus di bawah (Pratiwi et al., 2021):

$$Accuraccy = \frac{TP + TN}{TP + TN + FP + FN} \text{ X } 100\% \qquad (2.2)$$

2.2.6 CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*) merupakan suatu standarisasi pengolahan data mining yang telah dikembangkan dimana data yang ada akan melewati setiap fase terstruktur secara jelas dan efisien dipenuhi oleh data, yang bertujuan untuk menyelesaikan masalah yang umum dalam bisnis dan penelitian (Hasanah et al., 2021). CRISP DM memiliki enam tahapan,yaitu seperti pada gambar (Hasanah et al., 2021):



Gambar 2. 1Alur CRISP-DM

Sumber: (Hasanah et al., 2021)

1) Business Understanding (Pemahaman Bisnis)

   Pada tahapan ini, terdapat beberapa kegiatan yang dilakukan seperti memahami keperluan dan tujuan dari sudut pandang bisnis selanjutnya mengartikan pengetahuan ke dalam bentuk pendefinisian masalah pada data mining dan kemudian menentukan rencana serta strategi yang sesuai untuk mencapai tujuan pada data mining

2) Understanding (Memahami data)

   tahapan ini dimulai dengan menngumpulkan data, mendeskripsian data, serta mengevaluasi kualitas data

3) Data Preparation (pengolahan data)

   Dalam tahapan ini yang dilakukan adalah membangun dataset akhir yang

terdiri dari data mentah.Beberapa kegiatan yang dilakukan antara lain membersihkan data (Data Cleaning),mentransformasi data (Data Transformation) untuk dijadikan masukan dalam tahap pemodelan,Data balancing untuk mengurangi menyimpan data dan analisi data,serta data integration untuk metode mengintegrasikan data

4) Modeling (Pemodelan)

   Pada tahap ini, Machine Learning digunakan langsung untuk menentukan teknik, alat, dan algoritma data mining

5) Evaluation (Evaluasi)

   Tahap ini dilakukan dengan memeriksa seberapa baik pola yang dihasilkan oleh algoritma dapat berperforma. Untuk membandingkan kinerja algoritma, digunakan parameter evaluasi dalam bentuk Confusion Matrix dengan mengacu pada akurasi, presisi, dan recall.

2.2.7. Gain Ratio

Gain ratio adalah seleksi fitur yang banyak digunakan oleh peneliti karena handal dan mampu berjalan pada dimensi data yang tinggi. Tahap seleksi fitur sebenarnya melibatkan perhitungan kepentingan dari semua atribut data yang ada.atribut dengan tingkat kepentingan tinggi yang akan digunakan, sementara atribut dengan tingkat kepentingan rendah tidak akan digunakan dalam tahap berikutnya, yaitu klasifikasi(Kurniawan et al., 2018).Berikut adalah langkah-langkah dan rumus dalam perhitungan Gain Ratio (Yuliska & Syaliman, 2020):

- Langkah pertama adalah menghitung nilai entropy dari setiap atribut dalam dataset. Entropy adalah mengukur tingkat ketidakteraturan atau kebingungan dalam himpunan data dengan menggukan persamaan :

$$Entropy\ (S)\ = \sum_{i}^{n} \quad = 1\ -\ pi\ *\ pi$$

(2. 2)

- Selanjutnya adalah menghitung Information Gain (IG) dari setiap atribut dalam dataset. Information Gain mengukur seberapa banyak informasi yang

diperoleh dengan memilih atribut tersebut untuk memisahkan data. Dengan menggunakan persamaan :

$$Informatio\ Gain\ (S,A)\ =\ Entropy(S)\ -\ \sum_i^n\ \ =1\ \frac{|S_i|}{|S|}\ \times\ \ \ \ Entropy\ (Si)$$

(2. 3)

- Selanjutnya, perlu menghitung Split Information (SI) dari setiap atribut. Split Information mengukur seberapa homogen (atau heterogen) data setelah dibagi menggunakan atribut tersebut. dengan menggukan persamaan :

$$SplitInfo_A(D)\ =\ \sum_j^v\ \ =1\ \frac{|Dj|}{|D|}\ \times log\ log\ 2\ \left(\frac{Dj}{D}\right)$$

(2. 4)

- Setelah kita memiliki nilai Information Gain dan Split Information untuk setiap atribut, langkah terakhir adalah menghitung Gain Ratio (GR). Gain Ratio membantu memilih atribut terbaik dengan mempertimbangkan Gain dan Split Information secara bersamaan. Dengan menggunakan persamaan sebagai berikut

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{Split\ Info\ (A)}$$

(2.5*)*

2.3 Penelitian Terdahulu

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.