BAB I
PENDAHULUAN

1.1. Latar Belakang

Banjir adalah fenomena alam yang sering melanda indonesia. Menurut Data Informasi Bencana Indonesia (DIBI), dalam kurun waktu tiga tahun terakhir, tercatat sebanyak 4580 kejadian banjir di Indonesia dan Jumlah tertinggi terjadi pada tahun 2020, mencapai 1531 kejadian, menjadi yang terbanyak dalam hampir satu dekade terakhir (Databoks, 2023). Penyebab banjir disebab kan oleh aliran air dan curah hujan yang tinggi di suatu daerah, namun banjir juga dapat terjadi karena kondisi lingkungan seperti hilangnya lahan terbuka hijau. (Dilla Evitasari et al., 2023).

Kota Samarinda merupakan ibu kota dari Provinsi Kalimantan Timur yang saat ini sedang dilanda permasalahan banjir yang cukup parah. Banjir yang sering terjadi akhir-akhir ini sangat mengganggu aktivitas warga. Sebagian besar wilayah kota Samarinda yang bermasalah dengan banjir berlokasi di DAS Karangmumus (320 km2). Selain itu terdapat dua sub sistem lain yang juga mempunyai masalah banjir yaitu DAS Karang Asam Besar (9,65 km2) dan DAS Karang Asam Kecil (16,25 km2). (Purwanto, 2020). Pada tahun 2020 banjir terjadi pada 10 kecamatan, 4 kelurahan menyebabkan sebanyak 27.000 jiwa terkena dampak banjir yang merugikan masyarakat (Ernawati et al., 2021).

Klasifikasi banjir berdasarkan penyebabnya dapat membantu memperbaiki perkiraan frekuensi banjir, mendukung deteksi serta penafsiran untuk perubahan kejadian dan tingkat keparahan banjir (Tarasova et al., 2019). Oleh karena itu, perlu diadakan evaluasi perbaikan akurasinya dengan metode klasifikasi data *mining*. *Data mining* merupakan proses yang dilakukan dengan penggabungan teknik analisis data untuk memperoleh pola penting pada suatu data (Tarigan et al., 2022). Penerapan teknik *data mining* memiliki relevansi yang luas, termasuk dalam konteks bencana alam seperti banjir. Penggunaan *data mining* memegang peranan penting dalam menghubungkan teknologi dan penelitian, serta dapat mengenali pola asosiasi, melakukan klasifikasi, dan berinteraksi dengan algoritma pengklasifikasi untuk mendapatkan hasil yang bervariasi dari hasil yang buruk hingga mendapatkan hasil yang baik (Mian & Ghabban, 2022).

Ketidakseimbangan kelas *(class imbalance)* terjadi ketika sebagian besar data condong pada satu label kelas. Hal ini dapat terjadi dalam kedua klasifikasi kelas dua dan multi-kelas. Algoritma pembelajaran mesin menganggap bahwa data didistribusikan secara rata, jadi ketika ada kelas yang tidak seimbang, mesin akan lebih bias pada kelas yang dominan dengan mengabaikan kelas minoritas, sehingga pada kelas mayoritas lebih cenderung menunjukkan nilai akurasi yang lebih baik. Hal ini disebabkan oleh fakta bahwa fungsi pembelajaran mesin secara konsisten berusaha untuk mengoptimasi kuantitas, seperti tingkat *error,* tanpa mempertimbangkan distribusi data (Yoga Siswa, 2023).

Data berdimensi tinggi atau *High Dimensional* merupakan data yang mempunyai banyak atribut yang dapat digunakan dalam proses analisis. Misalnya, mempunyai puluhan bahkan ratusan atribut, maka data tersebut dapat diklasifikasikan sebagai data berdimensi tinggi (Hakimah et al., 2022). Data berdimensi tinggi dalam kumpulan data menyebabkan beberapa masalah dalam *Machine Learning*. Pertama, sulit bagi model pembelajaran untuk mencapai performa optimal karena semakin banyak fitur yang digunakan, semakin sulit bagi model pembelajaran mesin untuk memodelkan masalah tersebut. Kedua, jumlah data yang besar ini dapat menyebabkan *overfitting* karena banyaknya konfigurasi karakteristiknya meskipun data yang kita miliki sedikit. Ketiga, data dengan dimensi yang besar susah untuk diproses secara komputasi (*computationally expensive*) baik dari segi memori maupun waktu (Ariyoga, 2022).

Algoritma *k-nearest neighbour* merupakan metode klasifikasi yang mengelompokkan data uji menjadi data latih berdasarkan jarak antara beberapa tetangga *(neighbor)* terdekat dari data uji tersebut, Algoritma *k-nearest neighbour* juga sering digunakan dalam penyelesaian data *mining* dalam klasifikasi (Sitepu & Manohar, 2022). Dari beberapa penelitian yang pernah dilakukan sebelumnya dalam klasifikasi data banjir, Algoritma *k-nearest neighbour* (KNN) tanpa seleksi fitur dinilai memiliki performa lebih unggul dalam klasifikasi data banjir dengan akurasi 94,91% dibandingkan dengan *Random Forest* 71,3 %, *Support Vector Machine* 52,71%*, Naive Bayes* 89,23%. (Gauhar et al., 2021; Hossain & Zeyad, 2023; Dilla Evitasari et al., 2023; Vafakhah et al., 2020; Daniel et al., 2023; Farhan & Setiaji, 2023).

Berdasarkan penelitian lain yang menerapkan algoritma KNN pada data banjir dengan dimensi data yang tinggi menunjukkan akurasi yang lebih rendah yaitu 88, 94%, Ditemukan adanya permasalahan pada penelitian dengan data *High Dimensional* yang dapat menurunkan akurasi(Cumel, David Zamri, Rahmaddeni, 2022). Kemudian, pada studi klasifikasi kesesuaian air, dimana algoritma KNN juga memiliki akurasi yang rendah, dan rendahnya akurasi KNN disebabkan karena bertemu dengan data berdimensi tinggi atau *high dimension* (Sopiatul Ulum et al., 2023). Pada dataset Banjir yang akan digunakan pada penelitian ini terdapat 19 fitur, dimana untuk jumlah fitur yang tinggi seringkali merujuk pada data berdimensi tinggi. Oleh karena itu, untuk mengatasi masalah tersebut dilakukan *feature selection* untuk mengidentifikasi fitur-fitur yang paling relevan, dengan tujuan untuk meningkatkan performa yang lebih baik dengan menggunakan *feature selection.*

Pendekatan yang digunakan dari penelitian sebelumnya dalam mengatasi dimensi tinggi menggunakan *feature selection Relief* terbukti bisa memberikan peningkatan akurasi sebesar 5-10%. Penelitian yang dilakukan memberikan peningkatan klasifikasi *Naive bayes* dan KNN dimana sebelum penerapan *feature selection* akurasi yang diperoleh sebesar 73,4% untuk *Naive bayes* dan 66,24% untuk KNN. Kemudian, setelah dilakukan penerapan *feature selection Relief* didapatkan peningkatan akurasi dimana akurasi *Naive bayes* menjadi 74,38% dan KNN menjadi 72,22% (Yahdin et al., 2021). Terdapat peningkatan yang signifikan pada akurasi algoritma KNN, yang sebelumnya memiliki akurasi sebesar 85,31% dan setelah penerapan *feature selection Relief* meningkat menjadi 95,63% (Yusra et al., 2021). Kemudian akurasi di atas 90% diperoleh setelah *feature selection Relief* dikombinasikan dengan KNN pada penelitian yang dilakukan oleh (Kemal Musthafa Rajabi et al., 2023; Abdulrazaq et al., 2021).

Kemudian dalam menghadapi ketidakseimbangan kelas pada dataset banjir dimana pada data yang diperoleh dari BMKG dan BPBD yang terjadi banjir berjumlah 49 data sedangkan yang tidak banjir berjumlah 841 data, maka pada penelitian ini juga akan menggunakan teknik *Synthetic Minority Over-samplingTechnique* (SMOTE) agar performa model yang dihasilkan dapat optimal. Berdasarkan penelitian sebelumnya, teknik SMOTE pernah digunakan dalam menangani ketidakseimbangan kelas pada dataset banjir dan dianggap dapat memberikan peningkatan akurasi terhadap model klasifikasi sebesar 0.21-10% (Nawi et al., 2020; Priscillia et al., 2022; Nursyahfitri et al., 2022; Razali et al., 2020; Dwi Astuti & Nova Lenti, 2021).

Berdasarkan beberapa penelitian sebelumnya yang pernah dilakukan, Optimasi *Particle Swarm Optimization* (PSO) pernah digunakan untuk mengklasifikasikan data banjir. penerapan PSO) menghasilkan hasil yang signifikan dalam memberikan peningkatan kinerja (optimasi) pada algoritma *Naïve Bayes* dan *K-Nearest Neighbor* (Yoga & Prihandoko, 2018). Penelitian lain yang menggunakan Optimasi PSO juga dapat mengoptimalkan peforma, dimana penerapan Optimasi tersebut dapat memberikan peningkatkan akurasi sebesar 3-11% (Dwiasnati & Yudo Devianto, 2022; Faldi et al., 2023;Arora et al., 2021). sehingga PSO akan di gunakan sebagai metode optimasi dalam penelitian pada data banjir.

Data tersebut setelah dilakukan Pengolahan data, terdapat permasalahan imbalanced data dan Data High Dimensional. Peneliti berencana akan menggunakan algoritma KNN dimana dari penelitan sebelumnya, knn mampu memberikan hasil akurasi yang baik dalam mengklasifkasi berbagai macam data, terutama ketika dikombinasikan dengan teknik pemilihan fitur dan penanganan ketidakseimbangan kelas. Namun, sebagian besar penelitian sebelumnya menggunakan data dengan dimensi yang berbeda dan teknik pemilihan fitur yang beragam. selain dilakukan pengolahan data, dilakukan juga sebuah analisa melalui penelitian sebelumnya, bahwa belum ada penelitian yang sama dengan menggunakan Kombinasi model KNN-PSO-*Relief* (KNN-PSORF) dan teknik Oversamping SMOTE dalam menangani *High Dimensional* dan *Imbalaced Data.* Diharapakan *Kombinasi model KNN, PSO, Relief, SMOTE* dapat memberikan peforma optimal dalam klasifikasi data banjir Kota Samarinda.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah untuk penelitian ini dapat dirumuskan sebagai berikut:

1. Fitur Apa saja yang memiliki pengaruh penting Pada algoritma *k-nearest neighbors* (KNN) dengan menggunakan optimasi PSO, seleksi fitur *Relief* dan *oversampling* SMOTE dalam meningkatkan akurasi pada dataset banjir Kota Samarinda?
2. Seberapa besar peningkatan akurasi yang didapat Algoritma *k-nearest neighbors* (KNN) dalam Klasifikasi data banjir Kota Samarinda dengan menggunakan PSO sebagai optimasi, seleksi fitur *Relief* dalam menangani high dimensional dan *oversampling* SMOTE dalam menangani *imbalanced data* ?

1.3. Tujuan Penelitian

Berdasarkan latar belakang masalah yang telah diuraikan, tujuan utama dari penelitian ini adalah:

1. Menentukan atribut yang berpengaruh pada algortima *k-nearest neighbour* (KNN) terhadap dataset banjir Kota Samarinda.
2. Mengevaluasi hasil kinerja algoritma *k-nearest neighbors* (KNN) yang dievaluasi menggunakan metode *validasi cross-fold k-fold* dan *matrix confusion.*

1.4. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Penulis:
   a. Penelitian ini akan memberikan pengalaman dan pengetahuan praktis dalam mengembangkan metode klasifikasi pada data banjir dengan mengatasi masalah akurasi pada data berdimensi tinggi menggunakan algoritma KNN.
   b. Penelitian ini memberikan pengalaman dalam menerapkan teknik seleksi fitur dan optimasi pada algoritma klasifikasi untuk menghasilkan hasil penelitian yang lebih baik.
2. Peniliti selanjutnya:
   a. Hasil Penelitian ini diharapkan bisa memberikan kontribusi pada pengembangan ilmu pengetahuan, khususnya dalam bidang teknik data mining dan pengolahan data berdimensi tinggi, serta memberikan wawasan baru dalam penggunaan algoritma KNN dalam konteks klasifikasi banjir.
   b. Diharapkan bisa bermanfaat dan menjadi referensi bagi penelitian selanjutnya dalam bidang yang sama atau terkait, serta menjadi landasan untuk pengembangan penelitian lebih lanjut dalam upaya meningkatkan prediksi dan pengendalian banjir di daerah-daerah lain.

1.5. Batasan Masalah

Agar ruang lingkup permasalahan yang dibuat tidak meluas, maka peneliti membatasi penelitian sebagai berikut :

a. Data yang digunakan dalam penelitian ini adalah dataset banjir Kota Samarinda yang diperoleh dari BPBD (Badan Penanggulangan Bencana Daerah) Kota samarinda pada tahun 2021-2023.
b. Algoritma klasifikasi yang akan digunakan dalam penelitian kali ini adalah *K-Nearest Neighbor (KNN*) dengan tambahan metode seleksi fiturnya berupa *Relief*, metode *optimasi Particle Swarm Optimization (PSO)*.
c. Fitur-fitur yang digunakan dalam klasifikasi pada dataset banjir dalam penelitian ini meliputi (i) Temperatur-minimum, (ii) Temperatur-maksimum, (iii) Temperatur, (iv) Kelembaban, (v) Curah-hujan, (vi) Lamanya-penyinaran-matahari, (vii) Kecepatan-angin, (viii) Arah-angin-maksimum, (ix) Kecepatan-angin-rata-rata, (x) Arah-angin-terbanyak dan (xi) Terjadi-banjir yang akan dijadikan kelas atau target dari klasifikasi.