BAB I
PENDAHULUAN

1.1 Latar Belakang

Data mining merupakan proses menggunkan statistik,matematika,kecerdasan buatan dan *juga machine learning* untuk mengidentifikasi masalah yang ada pada sebuah data sehingga menghasilkan informasi yang bermanfaat.Berdasarkan fungsinya data mining di kelompokan menjadi,deskripsi, estimasi, klasifikasi, *clustering*, dan asosiasi(Admojo 2020).

Algoritma klasifikasi merupakan sebuah proses mencari kumpulan pola maupun fungsi untuk memdeskripsiakan dan memisahkan kelas data satu dengan lainya.Yang berguna untuk mengelompokan data tersebut pada kategori yang telah ditentukan. Klasifikasi adalah bentuk analisi data yang mengekstrak model kelas data. Klasifikasi termasuk dalam *supervised learning* karena menggunakan karena menggunakan data yang sudah dianalisis untuk di uji dan di klasifikasikan. Proses klasifikasi sendiri terdiri dari pemebelajaran dan klasifikasi.Pada tahap pemebelajaran ada data latih dan data uji untuk memastikan *rule* dari akurasi.Teknik klasifikasi dibagi menjadi lima yaitu berbasis statistik, berbasis jarak, berbasis pohon keputusan, berbasis jaringan syaraf dan berbasis *rule*.

K-NN merupakan salah satu metode data mining yang terbaik serta banyak digunakan dalam penelitian.Algoritma K-NN dikenalkan oleh Fix dan Hodges pada tahun 1951.Algoritma K-NN merupakan algoritma yang sederahana dan sering digunakan untuk meneglompokan data supervised. K-NN mengklasifikasikan objek berdasarkan data latih berdasarkan jarak yang paling dekat dengan objek tersebut. Jarak dekat dan jauhnya objek dengan data latih dapat dihitung menggunkan eulidiance(Adisasmita et al. 2019).Namun K-NN meiliki kekurangan yaitu dalam perhitungan algoritmanya K-NN harus melakukan perhitungan jarak pada setiap pada setiap *query instance* secara bersama-sama sehingga hasil pada perhitungan algoritma K-NN menggunakan *dataset* berdiemensi tinggi tanpa penambahan method lain akurasinya rendah.

Namun dalam beberapa penelitian K-NN sering mendapatkan akurasi yang rendah dari algoritma yang lain seperti dalam penelitian yang dilakukan oleh

Muhammad Rangga Aziz Nasution dan Mardhiya Hayaty(Rangga, Nasution, and Hayaty 2019) dengan menggunakan perbandingan algoritma K-NN dan SVM pada dataset analisis sentimen twitter.Dari hasil penelitian ini SVM memiliki akurasi yang lebih tinggi dari K-NN namun K-NN memiliki waktu proses yang lebih cepat dari SVM.

Penelitian lain juga dilakukan oleh Mohammad farid naufal(Perbandingan, Svm, and Untuk 2021) yang membandingkan algoritma SVM, K-NN dan CNN untuk mengklasifikasikan citra cuaca. Berdasarkan penelitian yang dilakukan didapatkan hasil bahwa CNN memiliki performa paling baik diantara algoritma lainya pada dataset citra cuaca.

Berdasarkan beberapa penelitian *feature selection* dapat meningkatkan kinerja algoritma seperti penlitian oleh imam santoso, dkk(Sistem, Analisis, and Pemilihan 2021) yang menggunkan *feature selection* pada algoritma *support vector machine*(SVM).Penggunaan *feature selection* pada algoritma ini mampu meningkatkan kinerja SVM dari 66.49% menjadi 81.18%.

Selain itu penggunaan *feature selection* pada algoritma lain dilakukan oleh Mula Agung Barata, dkk(Tea, Using, and Nose 2023) yang menggunakan algoritma C4.5 dengan penambahan *feature selection chi-squre.* Dalam penelitian ini *feature selection* juga mampu meningkatkan akurasi pada algoritma C4.5 yaitu dari 93.51% menjadi 94.27%.

Berdasarkan latar belakang tersebut, maka peneliti menggunakan algoritma K-NN dengan *feature selection* untuk klasifikasi data berdimensi tinggi.Dengan adanya penelitian ini diharapkan mampu memberikan pandangan tentang pemilihan algoritma yang tepat untuk dataset berdimensi tinggi.

1.2 Rumusan Masalah

Berdasarkan pemaparan latar belakang diatas, maka permasalahan yang akan di bahas dalam penelitian ini adalah :

1. Bagaimana mengembangkan sistem untuk mengklasifikasi data berdimensi tinggi dengan algoritma K-NN menggunakan penambahan *feature selection*?
2. Bagaimana meningkatkan akurasi algoritna K-NN dengan penambahan *feature selection* pada algoritma?

1.3 Tujuan Penelitian

1. Untuk mengembangkan sistem yang dapat mengklasifikasi data berdimensi tinggi dengan algoritma K-NN menggunakan penambahan *feature selection* pada algoritma.
2. Untuk meningkatkan akurasi pada K-NN dengan *feature selection.*

1.4 Batas Penelitian

a. Menggunakan algoritma k-nn
b. Menggunakan seleksi fitur *forward selection, backward elimination* dan *chi-square*
c. Menggunakan 2 dataset berbeda yang memeiliki dimensi tinggi.
d. Hasil akhir berupa akurasi yang di dapatkan dari pengujian dataset.
e. Menggunakan *software rapidminer* untuk menentukan akurasi algoritma.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah :

Adapun manfaat yang diharapkan oleh peneliti sesuai penelitian yang dibuat antara lain :

1. Manfaat teoritis

   Manfaat teoris dalam penelitian ini antara lain :

   a. Mendukung pengembangan teori tentang klasifikasi data berdimensi tinggi meggunakan algoritma K-NN dengan improvisasi dengan method *forward selection, backward elimination* dan *chi-square*
   b. Menambah dan memeprkuat pengetahuan tentang meningkatkan kinerja algoritma.
   c. Memberikan solusi untuk pengklasifikasian data bedimensi tinggi.

2. Manfaat Praktis

   Manfaat praktis yang diperoleh dari penelitian ini antara lain :

   1. Membantu memilih algoritma yang cocok untuk data berdimensi tinggi.
   2. Membantu memberikan solusi untuk meningkatkan kinerja algoritma K-NN