

# BAB I PENDAHULUAN

## 1.1 Latar Belakang Masalah

*K-Nearest neighbors* atau yang biasa disebut (K-NN) merupakan algoritma *supervised learning* dalam *machine learning*. K-NN biasanya digunakan untuk klasifikasi dan regresi, dalam K-NN klasifikasi menghasilkan sebuah kelas sedangkan K-NN regresi menghasilkan nilai objek. K-NN melakukan klasifikasi data baru dengan cara mengelompokkan data terdekatnya berdasarkan sejumlah *K-Neighbors* untuk di bandingkan dengan data terbaru berdasarkan kesamaan data. Pada penelitian ini digunakan algoritma K-NN karena dataset yang digunakan sudah berlabel dengan baik dan dataset bebas dari *noise*.

Sebelum data bisa digunakan sebagai klasifikasi data harus di proses terlebih dahulu dan proses ini disebut *pre-processing* ada banyak cara teknik *pre-processing* data diantaranya adalah normalisasi data dan *balancing data*, normalisasi data adalah proses untuk menyamakan rentang nilai dari yang berbeda menjadi sama yaitu 0 sampai 1 sedangkan *balancing data* adalah proses menyamakan data kelas yang tidakimbang menjadiimbang berdasarkan data terbesar atau terendah.

data yang akan digunakan untuk penelitian ini adalah data kualitas *red wine* yang di dapatkan dari *website* Kaggle.com dan data ini bersifat publik. Data set ini memiliki 1599 data dan memiliki 12 atribut, untuk atribut yang digunakan untuk klasifikasi bernama *quality* yang berisi nilai 1-10 semakin besar nilainya semakin tinggi kualitas *wine* tersebut, dataset ini digunakan hanya sebagai data untuk tes saja dikarenakan data sudah banyak digunakan dan agar hasil dari penelitian ini bisa di bandingkan dengan penelitian sebelumnya.

Berdasarkan latar belakang tersebut peneliti ingin mencari tahu pengaruh proses *pre-processing* data antara *balancing data* dan normalisasi data pada performa *K-Nearest Neighbors* dalam mengklasifikasi data dan menentukan jumlah *K-Neighbors* yang di perlukan untuk mencapai akurasi tinggi.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang di uraikan di atas, maka rumusan masalah pada penelitian ini adalah:

1. Berapa jumlah  $K$  pada  $K$ -Neighbor yang di perlukan agar mendapatkan akurasi yang tinggi?
2. Apa pengaruh *balancing* data pada akurasi klasifikasi?
3. Apa pengaruh normalisasi data pada akurasi klasifikasi?
4. Apa pengaruh kombinasi *balancing* data dan normalisasi data pada akurasi klasifikasi?

## 1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah:

1. Algoritma yang di gunakan hanya  $K$ -Nearest Neighbors
2. Metode normalisasi data hanya *min-max*
3. Metode *balancing* data hanya SMOTE
4. Bahasa pemrograman yang digunakan adalah Python
5. *Dataset Wine Quality* diambil dari website UCI machine learning
6. Data Red Wine pada *dataset Wine Quality* hanya 1599 data

## 1.4 Tujuan Penelitian

Berdasarkan dari latar belakang masalah dan rumusan masalah tujuan yang ingin di raih peneliti dalam pembuatan skripsi ini adalah apa pengaruh normalisasi data dan *balancing* data pada performa  $K$ -Nearest Neighbors dalam mengklasifikasi kualitas *wine* dan mengetahui berapa banyak  $K$ -Neighbor yang di butuhkan untuk mencapai akurasi tertinggi.

### 1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Dapat mengetahui pengaruh *balancing* data terhadap hasil akurasi
2. Dapat mengetahui pengaruh normalisasi data terhadap hasil akurasi
3. Dapat dijadikan acuan untuk penelitian di masa depan terkait *dataset* ini
4. Hasil penelitian ini dapat menjadi acuan untuk penelitian lain di masa depan yang menggunakan algoritma *K-Nearest Neighbors* (KNN)

### 1.6 Sistematika Penulisan

Sistematika penulisan laporan ini adalah sebagai berikut:

BAB I PENDAHULUAN, berisi Latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan

BAB II TINJAUAN PUSTAKA, berisi teori terori yang digunakan pada penelitian ini dan refrensi dari penelitian sebelumnya

BAB III METODE PENELITIAN, berisi tentang alur penelitian serta alat dan bahan yang digunakan

BAB IV HASIL DAN PEMBAHASAN, bab ini berisi proses tahapan pemrosesan data dan pembahasan hasil

BAB V PENUTUP, bab ini berisi tentang kesimpulan dari penelitian ini dan saran untuk pengembangan selanjutnya.