# BAB I

# PENDAHULUAN

## 1.1. Latar Belakang

Klasifikasi dengan berbagai algoritma machine learning, yaitu algoritma yang mempelajari pengelompokan data berdasarkan pengolahan data yang telah ada sebelumnya, bertujuan untuk mendapatkan target kelas yang akurat. Namun pada kenyataannya muncul permasalahan dalam proses klasifikasi tersebut ketika salah satu kelas memiliki jumlah yang jauh lebih kecil pada training dataset-nya, yaitu kumpulan data yang dijadikan bahan pengolahan data. Permasalahan tersebut disebut juga dengan imbalance dataset problem. Kondisi tersebut akan berpengaruh terhadap klasifikasi data yang akan dilakukan untuk menentukan kelas suatu data. Jika kondisi data tersebut imbalance maka kecenderungan kelas data tidak stabil karena data akan lebih condong ke bagian data yang memiliki komposisi data lebih besar (*Majority class*)[1].

Dengan adanya ketidakseimbangan ini membuat algoritma klasifikasi seperti *naïve bayes, k-nn, C4.5, SVM* dan lain-lain menghasilkan performa yang buruk dalam mengklasifikasikan data. Sehingga klasifikasi pada kelas mayoritas menghasilkan akurasi yang sangat tinggi sedangkan klasifikasi pada kelas minoritas menghasilkan akurasi yang sangat rendah Tentu saja hal ini tidak bagus, terlebih lagi apabila hasil klasifikasi kelas minoritas itu sangat diperlukan informasinya. Karena pada kasus tertentu, kelas minoritas perlu untuk diidentifikasi secara tepat.

Untuk mengatasi masalah tersebut, maka dilakukan proses pendekatan level data dengan cara melakukan *resampling* pada data. terdapat dua jenis teknik *resampling* data yaitu *under sampling* dan *over sampling*. *Over sampling* dipilih karena jumlah data yang digunakan tidak terlalu banyak. Terdapat algoritma-algoritma *over sampling* seperti *Random sampling* dan *Synthetic Minority Over Sampling Method* (Riswanto, 2007).

Salah satu taknik oversampling yaitu teknik *Synthetic Minority Oversampling Technique* (SMOTE). Teknik SMOTE bekerja dengan cara mensintesis data buatan berdasarkan sampel yang diambil dari data kelas minoritas. Data asli dan data yang telah di-*resampling* akan dilakukan proses klasifikasi. Algoritma klasifikasi yang akan digunakan adalah Algoritma *K-Nearest Neighbor* (KNN) (Chawla, 2002).

Dengan dilakukannya penelitian ini diharapkan dapat mengetahui dampak dari penggunaan teknik resampling menggunakan algoritma *Synthetic Minority Oversampling Technique* (SMOTE) terhadap tingkat akurasi dan diharapkan dapat meningkatkan performanya pada klasifikasi data menggunakan algoritma *K-Nearest Neighbor* (KNN).

## 1.2. Rumusan Masalah

Berdasarkan latar belakang diatas maka, rumusan masalah dari penelitian ini adalah apakah algoritma *Synthetic Minority Oversampling Technique* (SMOTE) dapat memperbaiki klasifikasi data pada kelas minoritas dan

meningkatkan performa algoritma klasifikasi pada dataset yang mengalami *imbalance class*?

### 1.3. Batasan Masalah

Untuk menghidari luasnya pembahasan dan guna memberikan fokus masalah pada kajian skripsi ini, maka masalah yang dibatasi dalam pembahasan skripsi ini meliputi:

1. Metode sampling yang akan digunakan adalah Teknik *Oversampling* SMOTE.
2. Mengguanakan algoritma *K-Nearest Neighbor* (KNN) untuk klasikasi data.
3. Data set yang digunakan adalah data yang bersifat imbalance dan multilabel.
4. Lingkup dataset yang digunakan pada penelitian ini adalah data yang dipublikasikan oleh *UCI Machine Learning* melalui situs resmi https://archive.ics.uci.edu

### 1.4. Tujuan Penelitian

Berdasarkan rumusan masalah diatas, maka tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Menangani imbalance class menggunakan Teknik *Oversampling* SMOTE.
2. Mengetahui performa algoritma klasifikasi pada data tidak seimbang dan seimbang.
3. Membandingkan dan mengevaluasi dampak dari penggunaan algoritma SMOTE teknik terhadap pada performa algoritma *K-Nearest Neighbor* (KNN) dalam penanganan masalah *imbalance class*.

## 1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

### 1.5.1. Bagi Peneliti

Untuk mengetahui tingkat performa algoritma *K-Nearest Neighbor* (KNN) pada klasifikasi data yang telah di resampling menggunakan algoritma SMOTE. Penelitian ini sebagai salah satu media untuk menerapkan ilmu yang diperoleh selama kuliah. Selain itu juga, diharapkan penelitian ini dapat bermanfaat bagi pembaca sebagai sumber informasi bagi pihak-pihak yang berkepentingan dan sebagai rujukan atau landasan bagi peneliti selanjutnya.

### 1.5.2. Bagi Akademik

Penelitian ini diharapkan dapat memebrikan kontribusi terhadap akademik sebagai tambahan referesnsi dalam penelitian sejenis dimasa yang akan datang.

## 1.6. Metodologi Penelitian

### 1.6.1. Pengumpulan Dataset

Dataset yang digunakan adalah data *Car Evaluation* yang berasal dari model keputusan hirarkis yang dikembangkan oleh Bohanec, V. Rajkovic untuk pengambilan keputusan yang dipublikasikan oleh *UCI Machine Learning* melalui situs resminya https://archive.ics.uci.edu.

### 1.6.2. Preprocessing

Pada tahap ini akan dilakukan *fitting* atau memisahkan data atribut target dengan atrdan melakukan diskritisasi pada beberapa *field* atau *variable*.

### 1.6.3. Klasifikasi

Dataset yang telah di klasifikasi menggunakan algoritma *K-Nearest Neighbor* (KNN) akan dihitung performa atau akurasi yang dihasilkan dan dilakukan resample data menggunakan algoritma SMOTE ( *Synthetic Minority Over Sampling Method*) untuk mengetahui perbandingan performa dari kudua algortima.

#### 1.6.4. Evaluasi

Setiap proses klasifikasi yang diukur performanya akan dibandingkan hasilnya antara klasifikasi pada dataset sebelum dan sesudah di-*oversampling*. hal ini akan dilakukan untuk melihat pengaruh implementasi algoritma SMOTE pada performa klasifikasi algoritma *K-Nearest Neighbor* (KNN).

Metode evaluasi yang digunakan adalah menggunakan bahasa pemrograman *python* dengan memanfaatkan *library* yang tersedia, kemudian diukur secara program dan manual performa yang dihasilkan menggunakan tabel *Confusion Matrix*.

### 1.7. Sistematika Penulisan

#### 1.7.1. BAB I Pendahuluan

Bab I berisi pemaparan pendahuluan, pengenalan latar belakang, rumusan masalah, batasan masalah, tujuan dan manfaat, dn sistematika penulisan.

#### 1.7.2. BAB II Landasan Teori

Bab ini menjelaskan seluruh teori yang digunakan dalam penelitian ini meliputi klasifikasi, algoritma *Naive Bayes*, *Imbalance Class*, algoritma *Synthetic Minority Oversampling Technique* (SMOTE) dan *Confusion matrix*.

### 1.7.3. BAB III Metodologi Penelitian

Bab ini menjelaskan proses penelitian secara garis besar yang dimulai dari pencarian dataset yang akan digunakan, melakukan proses *resampling/oversampling*, melakukan proses klasifikasi, dan mengevaluasi performa dari algoritma klasifikasi.

### 1.7.4. BAB IV Hasil Implentasi dan Pembahasan

Berisi pembahasan dari hasil implementasi yang telah dilakukan pada bab sebelumnya serta menganalisis hasilnya.

### 1.7.5. BAB V Penutup

Pada bab ini akan memaparkan kesimpulan dari hasil penelitian serta menjawab rumusan masalah dan memberikan saran untuk peneltian selanjutnya.

### 1.7.6. Daftar Pustaka

Mencantumkan semua referensi literatur yang mendukung proses penelitian ini.

### 1.7.7. Lampiran

Dokumen tambahan atau pendukung lainnya yang dilampirkan.