BAB I
PENDAHULUAN

1.1 Latar Belakang

Saat ini, banyak orang yang mulai melakukan kegiatan penambangan data. Kegiatan tersebut biasanya disebut dengan data mining. Data mining sendiri merupakan proses penambangan data atau penggalian informasi yang berguna dan berharga yang didapatkan dari sekumpulan data yang besar atau database. Selain itu, data mining biasanya digunakan untuk menemukan pola-pola yang tidak diketahui untuk menghasilkan sebuah pengetahuan baru atau sebuah kesimpulan dari sekumpulan data tersebut [1].

Dalam konteks data mining, salah satu bidang yang terkait secara erat adalah machine learning. Machine learning merupakan proses pembelajaran komputer secara otomatis tanpa harus diprogram secara explisit. Dalam mendapatkan kecerdasan secara otomatis, machine learning melalui dua proses yaitu proses training dan proses testing. Pada saat ini machine learning terbagi menjadi tiga kategori yaitu supervised learning, unsupervised learning, dan reinforcement learning [2]. Pada supervised learning mempunyai berbagai cabang pembahasan seperti regresi dan klasifikasi

Pada penelitian ini, penulis berfokus pada proses klasifikasi yang menggunakan data pengecekan kesehatan rutin para pekerja orang korea selatan. Data tersebut memiliki variable dependen / varibel yang akan diprediksi berupa perokok dan tidak perokok. Dari data tersebut didapatkan beberapa hasil lab para pekerja yang dapat digunakan untuk proses klasifikasi antara perokok dan bukan perokok. Adapun beberapa fitur yang terdapat pada data tersebut seperti jenis kelamin, gula darah, karang gigi dan lain – lain.

Kemudian ˙beberapa hal yang dapat mempengaruhi performa dari model klasifikasi machine learning adalah imbalance class dan pemilihan fitur. Pembuatan model klasifikasi pada imbalance class akan cenderung mengabaikan kelas dengan jumlah sampel yang sedikit sehingga dapat berdampak buruk terhadap performa

model [3]. Penggunaan fitur yang tidak mempunyai relasi pada variable dependen juga akan sangat mempengaruhi performa model. Fitur yang tidak terpakai harus diseleksi menggunakan algoritma pemilihan fitur yang tepat [4]. Pada penelitian ini, penulis menggunakan teknik SMOTE untuk mengatasi imbalance class. Sedangkan pada masalah pemilihan fitur, penulis menggunakan metode forward selection dan pearson correlation.

Selain itu, performa model machine learning juga sangat dipengaruhi dari penggunaan algoritma pembuatan model yang tepat. Pada penelitian ini, penulis menggunakan dua algoritma klasifikasi yaitu Support Vector Machine dan K-nearest neighbors. SVM merupakan algoritma klasifikasi dengan prinsip mencari hyperline dengan margin terbesar[5]. Sedangkan KNN merupakan metode klasifikasi dengan cara mencari kedekatan data dengan data lainnya [6]. Penggunaan kedua metode tersebut karena pada penelitian sebelumnya yang membahas tentang klasifikasi pemodelan gaji dengan binary klasifikasi mendapatkan nilai akurasi diatas 80% [7].

Penelitian ini bertujuan untuk mengukur performa model yang telah dikembangkan menggunakan algoritma SVM dan KNN pada dataset penentuan status perokok pasif. Dengan menerapkan SMOTE, fitur selection dan penggunaan parameter klasifikasi yang tepat, diharapkan penelitian ini dapat memberikan hasil peforma model yang bagus dan berkontribusi menambah wawasan tentang faktor – faktor yang mempengaruhi klasifikasi pada status perokok. Adapun untuk melakukan pengujian model, penulis menggunakan confunsion matrix yang dapat menampilkan tabel dari jumlah klasifikasi pada data uji yang benar dan data uji yang salah [8].

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan diatas, maka rumusan masalah dalam penelitian ini sebagai berikut:.

1. Bagaimana menerapkan SMOTE dan fitur selection pada dataset status perokok?

2. Bagaimana menerapkan metode Support Vector Machine dan K-nearest neighbors untuk pembuatan model pada penentuan status perokok?
3. Berapa hasil evaluasi menggunakan confuntion matrix pada performa model yang telah dibuat?

1.3 Batasan Masalah

Dalam penelitian ini, penulis menetapkan batasan ruang lingkup agar terhindar dari kemungkinan kesalahan dalam pelaksanaan penelitian. Batasan masalah yang ditetapkan dalam proses penelitian ini meliputi:

1. Dataset yang digunakan berupa data numerik yang didapatkan dari pemeriksaan kesehatan dasar para pekerja JKN di Negara Korea.
2. Peneliti menggunakan teknik SMOTE untuk mengatasi imbalance class.
3. Pada masalah pemilihan fitur, peneliti menggunakan metode Forward selection dan Pearson Correlation.
4. Peneliti berfokus pada pembuatan model menggunakan algoritma Support Vector Machine dan K-nearest neighbors.
5. Tingkat akurasi dan performa dari model diukur menggunakan metode confunsion matrix.

1.4 Tujuan Penelitian

Tujuan penelitian yang dilakukan oleh penulis adalah untuk mengetahui hasil evaluasi menggunakan confuntion matrix dari perbandingan performa model yang didapatkan pada proses SMOTE, pemilihan fitur, dan perbandingan performa model pada penggunaan dua algoritma yang berbeda yaitu Support Vector Machine dan K-nearest neighbors pada dataset status perokok.

1.5 Manfaat Penelitian

Dalam penelitian ini mempunyai beberapa manfaat yang dapat digunakan, baik secara teori maupun praktisi antara lain:

1.5.1 Manfaat secara teori

Penelitian ini dapat digunakan sebagai acuan dan untuk menambah wawasan tentang faktor – faktor yang mempengaruhi performa

model machine learning seperti imbalance class, pemilihan fitur dan penggunaan algoritma yang tepat.

1.5.2 Manfaat secara praktisi

a) Peneliti

Bagi peneliti bisa menggunakan model machine learning yang telah dibuat dikembangkan lebih lanjut dengan meningkatkan performa model dan dibuat menjadi aplikasi yang dapat digunakan.

b) Pembaca

Bagi pembaca, penelitian ini dapat digunakan sebagai wawasan dalam mempelajari machine learning terutama pada proses – proses yang ada di dalamnya dan penggunaan algoritma.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN : Bab ini menjelaskan tentang latar belakang, rumusan masalah, batasan masalah, tujuan, manfaat penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA : Bab ini berisi tentang beberapa jurnal yang telah dikakukan sebelumnya serta beberapa dasar teori yang digunakan pada penelitian ini.

BAB III METODE PENELITIAN : Bab ini menjelaskan alat dan bahan yang digunakan pada penelitian ini.

BAB IV HASIL DAN PEMBAHASAN : Bab ini melakukan pembahasan secara code, teknik SMOTE, pemilihan fitur, implementasi algoritma, dan hasil evaluasi yang didapatkan oleh performa model.

BAB V PENUTUP : berisi kesimpulan dan saran yang didapatkan selama penelitian berlansung.