BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Merokok adalah salah satu kebiasaan umum yang terdapat di berbagai negara, selain di negara maju, di negara berkembang juga sudah menjadi kebiasaan, terutama di negara Indonesia. Berdasarkan Riset yang dilakukan oleh Kementerian Kesehatan pada Hasil Riset Kesehatan Dasar (Riskesdas), menunjukkan prevalensi perokok di Indonesia usia 15 tahun ke atas, yaitu sebesar 34,2% pada Riskesdas 2007, 34,7% pada Riskesdas 2010, dan 36,3% pada Riskesdas 2013 [1]. *World Health Organization* (WHO), memperkirakan di tahun 2025 jumlah perokok di negara Indonesia akan semakin tinggi kurang lebih 45% berasal total populasi [2].

Terdapat sekitar 65,6 juta wanita dan 43 juta anak yang terkena paparan asap rokok maka termasuk ke dalam perokok pasif. Dikarenakan banyaknya perokok aktif di Indonesia yang merokok di lingkungan tertutup seperti di rumah (sekitar 91%), ini menjadi salah satu penyebab banyaknya perokok pasif [3]. Berdasarkan nilai asap rokok yang dihirup oleh perokok, perokok dapat dibedakan menjadi perokok aktif dan perokok pasif. Perokok aktif dapat dikategorikan berdasarkan banyaknya batang rokok yang dihisap per hari. Perokok pasif didefinisikan sebagai seseorang yang terkena asap rokok selama lebih dari 15 menit sehari lebih dari 1 hari dalam seminggu [4].

Berdasarkan hasil wawancara peneliti dengan Dokter Ratna Sudarti yang bertugas di Puskesmas Helvetia Medan Helvetia (wawancara dilakukan secara online). Dari beberapa atribut yang digunakan untuk mengklasifikan seseorang tersebut termasuk perokok aktif ataupun perokok pasif, ada beberapa atribut yang penting digunakan untuk melakukan pengklasifikasian tersebut. Diantaranya, jenis kelamin, umur, tinggi dan berat badan, Panjang lingkar pinggang, sistolik, relaksasi, *fasting blood sugar*, kolestrol, trigliserida, HDL (lemak baik), LDL (lemak jahat), hemoglobin, protein urin, kreatinin serum, y-GTP, *oral test*, karies gigi, karang gigi.

Merokok tidak hanya berdampak buruk bagi perokok itu sendiri, tetapi juga bagi orang-orang yang merokok di sekitarnya. Hal ini karena asap yang dihirup oleh perokok pasif belum melalui proses penyaringan dan karenanya mengandung kadar senyawa yang jauh lebih tinggi. Kondisi ini membuat pembakaran kurang sempurna dan melepaskan lebih banyak bahan kimia [5]. Selama ini orang-orang tidak menyadari kalau mereka tergolong kepada perokok pasif, sehingga tidak tau resiko yang telah diterima mereka dan mereka tanpa sadar sudah terkena penyakit juga, kemudian perokok aktif yang tidak toleran terhadap dirinya sendiri dan orang sekitarnya sehingga diperlukannya sebuah sistem prediksi menggunakan beberapa algoritma untuk mengklasifikan perokok aktif dan perokok pasif.

Klasifikasi merupakan salah satu metode data *mining* dari *supervised learning*, yang membutuhkan data latih yang telah diberi label sebagai kelas yang dipelajari untuk memperkirakan kelas suatu objek yang belum diketahui kelasnya. Klasifikasi terdiri dari beberapa algoritma yang dapat dipakai dalam proses klasifikasi yaitu *Decision Trees*, *Naive Bayes*, K-*Nearest Neighbors* (KNN) dan masih banyak lagi algoritma data *mining* lainnya. Sebelum memutuskan algoritma apa yang ingin digunakan dalam kasus klasifikasi, yang terbaik adalah mengetahui algoritma mana yang terbaik dan memiliki akurasi tinggi. Akurasi tinggi berpengaruh besar pada algoritma, karena jika algoritma memiliki akurasi tinggi dalam menyelesaikan suatu kasus klasifikasi, maka dapat diklasifikasikan sebagai keberhasilan klasifikasi, dan hasilnya akurat dan presisi [6].

Ada beberapa algoritma yang dapat dipergunakan dalam mengelompokkan perokok aktif dan perokok pasif diantaranya *Naïve Bayes*, C4.5 (*Decision Tree*), dan K-*Nearest Neighbor*. *Naïve Bayes* merupakan salah satu algoritma pembelajaran induktif yang cukup efektif dan efisien dalam pembelajaran mesin serta data *mining*. Performa *Naïve Bayes* yang kompetitif bahkan memakai proses pembagian terstruktur mengenai perkiraan independensi atribut (tidak ada korelasi antar atribut). Asumsi independensi atribut di data ini sebenarnya jarang terjadi, tetapi kinerja pembagian terstruktur mengenai *Naïve Bayes* reltif tinggi bahkan waktu asumsi independensi atribut dilanggar, seperti yang ditunjukkan dalam berbagai macam studi empiris [7].

Kemudian, *Decision Tree* (C4.5) mempunyai konsep secara umum dalam pembuatan Model *Training* dalam menentukan kelas atau nilai variabel target memakai *learning decision rules* yang dipelajari yang berasal dari data *training*. Keuntungan dari algoritma *Decision Tree* adalah mudah dijelaskan, algoritmanya mudah menangani interaksi fitur dan tidak menggunakan parameter. Namun kekurangan dari algoritma *Decision Tree* adalah tidak didukung penggunaan *online learning*, sehingga apabila terdapat data baru maka harus dibangun kembali pohon keputusannya [8].

Algoritma K-*Nearest Neighbor* (K-NN) adalah algoritma yang tidak membutuhkan pembuatan asumsi tentang distribusi data yang memuatnya. K-NN juga tidak membutuhkan data training sehingga mengakibatkan fase training lebih cepat dan fleksibel. Tetapi, kelemahan dari algoritma ini adalah ketika data mempunyai banyak *noise*, mengakibatkan prediksi yang diberikan kemungkinan memiliki banyak *noise* [9].

Dalam penelitian yang dilaksanakan oleh [10] dengan judul perbandingan algoritma C4.5 dan k-*Nearest Neighbor* pada klasifikasi penyakit *Disk Hernia* dan *Spondylolisthesis* dalam *Kolumna Vertebralis*. Dengan melakukan pengujian memakai data sebanyak 310 data pasien, Normal (100 pasien), *Spondylolisthesis* (150 pasien) dan *Disk Hernia* (60 pasien), Penelitian ini menghasilkan akurasi *classifier* C4.5 sebesar 89% dan K-NN sebesar 83%. Dengan lama waktu rata-rata untuk melakukan klasifikasi C4.5 0,00912297 detik dan klasifikasi K-NN 0,000212303 detik. Maka hasil dari penelitian ini menyatakan bahwa untuk mengklasifikasikan penyakit *Spondylolisthesis* dan *Disk Hernia* lebih baik dengan memakai algoritma *Decision Tree*- C4.5 namun untuk *running time* algoritma K-NN memiliki waktu klasifikasi lebih cepat.

Dalam penelitian [11] dengan judul perbandingan algoritma data *mining Naïve Bayes* dan *Bayes Network* untuk mengidentifikasi penyakit tiroid. Masalah yang terdapat pada penelitian ini adalah belum diketahuinya algoritma yang paling akurat dalam penentuan diagnosa untuk memprediksi penyakit tiroid. Dari penelitian ini menghasilkan kesimpulan bahwa algoritma yang mempunyai klasifikasi paling baik secara berurut adalah Bayes Network dengan akurasi sebesar

98,491% dan *Naïve Bayes* dengan akurasi dengan akurasi sebesar 91,803%, berdasarkan penilaian AUC 0,90 - 1,00 dengan demikian algoritma Bayes Network dapat memberikan solusi dalam permasalahan pengidentifikasian penyakit tiroid.

Pada penelitian ini menggunakan dataset yang diambil dari *Body signal of smoking* (*Kaggle*). *Kaggle* merupakan komunitas *online* yang berisikan para pegiat di bidang *data science* yang dapat berbagi ide, inspirasi atau bersaing dalam bidang *machine learning* dan ilmu lainnya. *Body signal of smoking* merupakan dataset yang berisi kumpulan data kesehatan biologi dasar yang bertujuan untuk mengetahui ada tidaknya kebiasaan merokok melalui *bio-signal*. Penelitian ini melakukan perbandingan memakai 3 algoritma untuk mengetahui manakah yang cocok digunakan untuk menyelesaikan kasus klasifikasi perokok aktif dan perokok pasif. Adapun 3 algoritma tersebut yaitu *Naïve Bayes*, C4.5(*Decision Tree*), K-*Nearest Neighbor*. Pada penelitian ini memakai aplikasi RapidMiner 9.1 yang akan menunjukkan akurasi masing – masing algoritma.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka dapat ditemukan permasalahan bahwa perokok aktif akan membahayakan tidak hanya pada dirinya sendiri tapi juga kepada orang lain yang tidak merokok atau perokok pasif. Untuk itu diperlukan suatu sistem prediksi dalam mengklasifikasikan perokok aktif dan perokok pasif menggunakan beberapa algoritma yaitu algoritma *Decision Tree* (C4.5), *K-Nearest Neighbor* dan *Naïve Bayes*.

1.3 Pertanyaan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan di atas. Peneliti merumuskan beberapa pertanyaan pada penelitian ini, diantaranya:

1. Bagaimana cara memprediksi perokok aktif dan perokok pasif dengan menggunakan dataset *Body Signal of Smoking (Kaggle)*?
2. Bagaimana cara mencari nilai akurasi terbaik perokok aktif dan perokok pasif dengan menggunakan algoritma *Decision Tree (C4.5), K-Nearest Neighbor dan Naïve Bayes*?

1.4 Tujuan Penelitian

Dari rumusan masalah yang telah diuraikan di atas penulis memeliki tujuan sebagai berikut:

1. Mengetahui implementasi algoritma *Decision Tree (C4.5), K-Nearest Neighbor dan Naïve Bayes* dengan menggunakan dataset *Body Signal of smoking (Kaggle)* dalam memprediksi perokok aktif dan perokok pasif.
2. Mengetahui performa algoritma terbaik antara 3 algoritma yang dibandingkan yaitu *Decision Tree* (C4.5), *K-Nearest Neighbor* dan *Naïve Bayes*.

1.5 Batasan Masalah

Berdasarkan rumusan masalah dan tujuan penelitian, maka batasan masalah penelitian sebagai berikut:

1. Mengfokuskan pada penggunaan dataset pada penelitian ini dalam bentuk yang diperoleh dari situs Kaggle yang bersumber dari data *Body signal of smoking (Kaggle)*.
2. *Tools* yang digunakan adalah *software* RapidMiner
3. Data yang digunakan sebanyak 55692 data
4. Atribut yang digunakan sebanyak 25 atribut

1.6 Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Memberikan pengetahuan mengenai implementasi akurasi terbaik *machine learning* menggunakan algoritma C4.5, *Naïve Bayes* dan KNN untuk memprediksi perokok aktif dan pasif dengan atribut *bio-signal*.
2. Mengetahui tingkat akurasi dari implementasi algoritma klasifikasi yaitu *Decision Tree* (C4.5), *K-Nearest Neighbor* dan *Naïve Bayes*.