BAB I
PENDAHULUAN

1.1 Latar Belakang

Hati merupakan organ terpenting bagi manusia, Hati juga memiliki fungsi untuk membentuk dan mengeluarkan empedu, selain itu hati juga merupakan alat untuk penyimpanan glikogen, sintesis urea, metabolisme kolesterol dan lemak, serta detoksifikasi. Masyarakat di indonesia termasuk kelompok beresiko untuk tertular hepatitis A dan hepatitis E. Kementrian kesehatan mendapatkan laporan bahwa setiap tahunnya selalu terjadi kejadian luar biasa (Kemenkes RI, 2014).

Hepatitis merupakan jenis penyakit endemik di beberapa negara berkembang, termasuk Indonesia. Penyakit ini disebabkan oleh infeksi jamur, bakteri, virus, obat-obatan, konsumsi alkohol, lemak berlebihan, atau penyakit *autoimun*. Ada lima jenis hepatitis, dari yang ringan hingga kronis, yaitu hepatitis A, hepatitis B, hepatitis C, hepatitis D, dan hepatitis E. Menurut data Riset Kesehatan Dasar (Riskesdas) 2013, menunjukkan bahwa 10 penduduk dari setiap 100 orang di Indonesia terinfeksi virus hepatitis C atau hepatitis B. Diperkirakan 28 juta orang terinfeksi hepatitis, 14 juta diantaranya menjadi hepatitis kronis, dan 1,4 juta dari hepatitis kronis tersebut terkena kanker hati. Pada tahun 2013, Indonesia memiliki prevalensi hepatitis B tertinggi kedua di Asia Tenggara. Hepatitis kronis, seperti hepatitis B, hepatitis C, dan hepatitis D, dapat berubah menjadi akut dan bisa menyebabkan sirosis dan kanker hati, pasien yang sudah dinyatakan mengidap hepatitis kronis bisa beresiko pada kematian (Khomsah, 2018).

Data mining adalah proses menemukan korelasi, pola, dan train baru yang bermakna dengan memilah-milah sejumlah besar data yang disimpan dalam repositori, menggunakan teknologi pengenalan pola serta teknik statistik dan matematika (Larose dan Larose 2014). Ada beberapa teknik yang diterapkan dalam data mining, salah satunya teknik klasifikasi.

Klasifikasi adalah sebuah teknik pengelompokkan data ke dalam beberapa kategori yang sudah ditentukan. Dalam klasifikasi, data yang diperoleh terlebih dahulu dilakukan pengolahan dengan menggunakan variabel yang ada untuk

menentukan data tersebut termasuk kategori yang mana (Sulastri, Hadiono, dan Anwar 2020).

Algoritma K-Nearest Neighbor merupakan algoritma untuk menentukan klasifikasi berdasarkan mayoritas dari kategori k-tetangga terdekat. Tujuan dari algoritma ini untuk mengklasifikasikan objek baru berdasarkan atribut dan sampel dari data latih (Ismail 2018).

Berdasarkan dataset atau data latih penelitian sering terjadi permasalahan-permasalahan diantaranya *imbalanced* data atau ketidakseimbangan data. Dataset dianggap tidak seimbang jika salah satu kelasnya memiliki dominan yang lebih besar dibandingkan kelas lainnya (Ali et al., 2015). Beberapa teknik yang digunakan untuk mengatasi data tidak seimbang salah satunya menggunakan teknik *Synthetic Minority Oversampling Technique* (SMOTE). Teknik *Synthetic Minority Oversampling Technique* (SMOTE) adalah salah satu metode yang digunakan untuk menangani kasus ketidakseimbangan data pada dataset (Hidayah et al., 2021).

Data yang tidak seimbang (data *imbalanced*) merupakan salah satu masalah terbesar yang terjadi pada dataset. Suatu dataset dianggap tidak seimbang jika salah satu kelasnya memiliki dominasi yang sangat tinggi dibandingkan dengan kelas lainnya, untuk mengurangi ketidakseimbangan data ada dua teknik pengambilan sampel utama yang digunakan, yaitu random *oversampling* (ROS) dan random *undersampling* (RUS). ROS secara tidak sengaja menyalin data kelas minoritas. ROS dapat menjadi pilihan yang baik jika tidak memiliki banyak data, tetapi dapat menyebabkan konfigurasi berlebih karena metode ini membuat salinan persis dari data kelas minoritas. Pada saat yang sama, RUS meninggalkan data (yang berasal dari kelas mayoritas) secara acak untuk mengubah distribusi kelas. Kelemahan dari RUS adalah dapat menyebabkan instalasi yang kurang karena menghapus informasi yang berpotensi berharga (Arifiyanti & Wahyuni, 2020). Luasnya penggunaan teknik data mining dalam dunia kesehatan dapat dilihat dari penelitian terdahulu sebagai bahan pertimbangan oleh peneliti.

Pada penelitian sebelumnya telah banyak digunakan metode untuk klasifikasi penyakit hepatitis dengan teknik data mining, diantaranya penelitian yang dilakukan oleh (Sulastri, Hadiono, dan Anwar 2020) yaitu prediksi penyakit hepatitis berdasarkan 155 data dan 20 atribut menggunakan metode *K-Nearest Neighbor*, *Naïve bayes* dan *Neural Network*. Algoritma *Naïve Bayes* tingkat akurasi yaitu 76.92%, tingkat error 23.01%, algoritma *Neural Network* tingkat akurasi yaitu 82,97%, tingkat error 17.03%, dan algoritma *K-Nearest Neighbor* tingkat akurasi yaitu 93%, tingkat error 7%. *K-Nearest Neighbor* termasuk tingkat akurasi terbaik dibandingkan *Naïve Bayes* dan *Neural Network*.

Penelitian yang dilakukan (Syukron et al., 2020) yaitu perbandingan metode *SMOTE Random Forest* dan *SMOTE XGBOOST* untuk klasifikasi tingkat penyakit hepatitis C pada *imbalance* class data berdasarkan 1385 data dan 24 variabel bebas. Metode *SMOTE Random Forest* tingkat akurasi yaitu 80,97% dan recall 75,55%, metode *SMOTE XGBOOST* tingkat akurasi yaitu 78,63% dan recall 76,82%. Metode *SMOTE Random Forest* merupakan tingkat akurasi tertinggi dibandingkan *SMOTE XGBOOST*.

Berikutnya, di tahun 2021 (Hidayah et al., 2021) meneliti klasifikasi data pasien penderita gagal jantung berdasarkan 299 data menggunakan algoritma *K-Nearest Neighbor* menerapkan teknik *SMOTE*. Untuk algoritma *K-Nearest Neighbor* tanpa *SMOTE* tingkat akurasi yaitu 71,59%, algoritma *K-Nearest Neighbor* menggunakan *SMOTE* tingkat akurasi yaitu 80,14%.

Berdasarkan penelitian yang dilakukan sebelumnya yakni pada penyakit hepatitis dan gagal jantung dengan menggunakan metode *K-Nearest Neighbor* menghasilkan akurasi yang cukup tinggi dibandingkan metode yang lain. Oleh karena itu pada penelitian ini penulis tertarik untuk meniliti dengan judul Pengaruh teknik *SMOTE* terhadap prediksi harapan hidup penderita penyakit hepatitis menggunakan metode *K-Nearest Neighbor*. Dataset penyakit hepatitis yang di ambil dari *KAGGLE* sebanyak 142 data dan tools yang digunakan menggunakan *phyton*.

1.2 Rumusan Penelitian

Berdasarkan latar belakang yang telah dikemukakan di atas, maka permasalahan dalam penelitian ini sebagai berikut:

1. Berapa nilai tertinggi dari presisi, akurasi dan recall yang diperoleh metode *K-Nearest Neighbor* dalam harapan hidup penderita penyakit hepatitis ?
2. Berapa nilai tertinggi dari presisi, akurasi dan recall yang diperoleh metode *K-Nearest Neighbor* dalam harapan hidup penderita penyakit hepatitis setelah melalui proses *SMOTE* ?

1.3 Tujuan Penelitian

Dari rangkaian permasalahan yang dijelaskan pada rumusan penelitian berikut tujuan yang ingin diperoleh dalam penelitian ini yaitu:

1. Mengetahui nilai tertinggi dari presisi, akurasi dan recall yang diperoleh metode *K-Nearest Neighbor* dalam harapan hidup penderita penyakit hepatitis.
2. Mengetahui nilai tertinggi dari presisi, akurasi dan recall yang diperoleh metode *K-Nearest Neighbor* dalam harapan hidup penderita penyakit hepatitis setelah melalui proses *SMOTE*.

1.4 Manfaat Penelitian

Hasil dalam penelitian ini diharapkan bermanfaat terhadap penulis, pembaca dan masyarakat luas. Berikut manfaat yang diharapkan dalam penelitian ini:

1. Dapat digunakan sebagai bahan referensi untuk penelitian selanjutnya.
2. Dapat menyeimbangkan data.
3. Pengembangan data mining dalam dunia pendidikan khususnya pada bidang kesehatan.

1.5 Batasan Penelitian

1. Data yang digunakan pada penelitian ini yaitu data penyakit hepatitis pada tahun 2019 yang disediakan *KAGGLE* oleh HariniR dengan alamat web : https://www.kaggle.com/harinir/hepatitis sebanyak 142 record data penyakit hepatitis, 116 pasien hidup dan 26 pasien meninggal.

2. Atribut yang digunakan terdiri dari 20 atribut, yaitu *age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver_big, liver_firm, spleen_palpable, spiders, ascites, varices, bilirubin, alk_phosphate, sgot, albumin, protime, histology, class.*
3. Tools yang digunakan adalah *Jupyter Notebook*.
4. Bahasa pemrograman yang digunakan adalah *Phyton*.