BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Diabetes adalah penyakit yang disebabkan oleh kadar gula darah yang tidak terkontrol dalam tubuh, yang mencegah pankreas memproduksi insulin yang cukup. Insulin sendiri merupakan hormon yang bertugas untuk membawa glukosa kepada sel-sel tubuh sebagai bahan bakar yang diperlukan oleh sel tersebut (Howsalya Devi et al., 2020). Diabetes telah menjadi penyakit yang mampu melumpuhkan bahkan mengancam jiwa penderitanya. Semakin lama seseorang mengidap penyakit diabetes akan menyebabkan risiko terjadi komplikasi semakin tinggi. Komplikasi pada penyakit diabetes dapat menimbulkan penyakit kardiovaskular, stroke, jantung, kerusakan pada pembuluh darah, penglihatan, pendengaran, kulit, ginjal, kaki, dan dapat menyebabkan depresi (Mayo Clinic, 2018). Seluruh dunia telah menganggap diabetes sebagai masalah utama pada kesehatan.

Indonesia menduduki peringkat ke-2 di Kawasan Pasifik Barat dan ke-7 diantara 10 negara paling terdampak dengan 10,7 juta orang terkena dampak dari 17,2 juta orang dewasa pada tahun 2019 (Atlas, 2019). Dengan angka tersebut membuat Indonesia sebagai negara di Asia Tenggara yang memiliki kasus diabetes terbanyak.

Salah satu tindakan pencegahan untuk menghindari bahaya dari penyakit diabetes adalah dengan melakukan prediksi dini yang dapat memprediksi penyakit diabetes. Dengan mengetahui penyakit diabetes sejak dini dapat membantu menghindari dan mengobati penyakit yang disebabkan oleh diabetes. Metode yang dapat digunakan untuk melakukan prediksi apakah seseorang menderita diabetes adalah dengan memanfaatkan *data mining*.

Data mining adalah metode pemrosesan data yang digunakan untuk menemukan informasi agar dapat menyelesaikan suatu masalah dengan melakukan analisis terhadap kumpulan data. Selain analisis, *data mining* juga berfungsi untuk mencari pola yang dapat memberikan informasi berdasarkan data yang diberikan

(Pristyanto, 2019). Beberapa bidang ilmu yang telah berhasil menerapkan *data mining* adalah pemasaran, bisnis, bioinformatika, pendidikan, kedokteran dan lain sebagainya (Singh & Singh, 2020). Beberapa metode pemrosesan data yang biasa digunakan dalam *data mining* adalah regresi, klasifikasi, *clustering*, asosiasi dan masih banyak lagi metode lainnya (Yunial, 2020). Dari metode-metode pemrosesan data tersebut, metode yang digunakan untuk medeteksi penyakit diabetes adalah klasifikasi. Klasifikasi adalah metode yang digunakan untuk menemukan suatu pola yang mampu mendeskripsikan dan membedakan kelas-kelas dalam suatu kumpulan data (Pristyanto, 2019). Salah satu algoritma klasifikasi yang umum digunakan untuk mendeteksi diabetes adalah *support vector machine* (Manurung, 2018).

Algoritma klasifikasi *Support vector machine* (SVM) adalah algoritma yang mampu menemukan *hyperplane* optimal yang dapat memisahkan setiap kelas pada data (Gazni, 2020). Namun, pengklasifikasian data dengan distribusi kelas yang tidak seimbang dapat menyebabkan SVM menghasilkan model klasifikasi yang buruk (Amelia et al., 2018). Kekurangan lain yang dimiliki oleh SVM adalah pemilihan *hyperparameter* optimal yang sulit (Wahyuni, 2019).

Untuk dapat mengatasi kekurangan SVM saat melakukan klasifikasi pada data yang tidak seimbang maka digunakan *borderline*-SMOTE untuk melakukan *oversampling* pada data. Metode *oversampling* berfungsi untuk meningkatkan jumlah data pada kelas minor agar distribusi kelas menjadi seimbang (Muthahari, 2018). Sayangnya, metode *oversampling* memiliki kekurangan karena dapat menyebabkan *overfitting*. Hal ini terjadi karena data yang dihasilkan dari metode *oversampling* terlalu fokus pada *training data* dari *dataset* tertentu sehingga tidak dapat memprediksi dengan tepat apabila diberikan *dataset* lain yang serupa (Saputra et al., 2021).

Masalah optimasi *hyperparameter* SVM dan *overfitting* yang dihasilkan dari metode *oversampling* dapat diselesaikan dengan menggunakan *grid search*. *Grid search* akan mencari *hyperparameter* optimal dengan melakukan pencarian lengkap terhadap subset ruang *hyperparameter* yang telah ditentukan (Sulistiana, 2020). Kemudian untuk meningkatkan hasil klasifikasi SVM menjadi lebih baik dan menghindari *overfitting* serta mengurangi variansi maka akan digunakan

algortitma *bagging*. *Bagging* bekerja dengan cara mengkombinasikan model klasifikasi dari *dataset training* yang telah di-*sampling* secara acak (A. Nugroho & Religia, 2021).

Berdasarkan penjelasan di atas, penelitian ini akan berfokus pada peningkatan performa klasifikasi *Support Vector Machine* dengan menerapkan *borderline*-SMOTE untuk mengatasi *dataset* yang tidak seimbang, *grid search* untuk mencari *hyperparameter* SVM yang optimal serta menghindari *overffiting*, dan *bagging* untuk menghindari *overfitting* serta mengurangi variansi.

1.2 Perumusan Masalah

Sesuai dengan latar belakang yang telah dijabarkan sebelumnya, maka yang menjadi permasalahan pada penelitian ini:

1. Bagaimana pengaruh algoritma *oversampling borderline*-SMOTE dan *bagging* terhadap klasifikasi *support vector machine* pada *dataset* diabetes *Pima Indians*?
2. Apakah optimasi *hyperparameter grid search* dapat meningkatkan performa model klasifikasi *support vector machine* yang telah dilakukan *oversampling borderline*-SMOTE dan *bagging*?
3. Bagaimana performa model klasifikasi yang dihasilkan oleh *support vector machine* dengan memanfaatkan *bagging*, *borderline*-SMOTE dan *grid search*.

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengetahui performa dari model klasifikasi *support vector machine* dengan *bagging, borderline*-SMOTE dan *grid search* terhadap *dataset* diabetes.

1.4 Manfaat Penelitian

Adapun manfaat yang dapat diberikan oleh penelitian ini adalah:

1. Membuat model klasifikasi penyakit diabetes dengan menerapkan algoritma *support vector machine* dengan *bagging*, *borderline*-SMOTE dan dioptimasi menggunakan algoritma *grid search*.
2. Diharapkan dapat membantu instansi kesehatan dalam memprediksi penyakit diabetes.

1.5 Batasan Penelitian

Adapun yang menjadi batasan dalam penelitian ini adalah:

1. Data yang digunakan untuk melakukan penelitian ini berasal dari *dataset Pima Indians Diabetes Database* dan *dataset diabetes Frankfurt.*
2. *Dataset Pima Indians* akan digunakan untuk membentuk model klasifikasi serta melihat performanya.
3. *Dataset Frankfurt* akan digunakan untuk melihat performa model yang terbentuk dari *dataset Pima Indians*.
4. Algoritma *borderline*-SMOTE digunakan untuk melakukan *oversampling* pada *dataset Pima Indians*.
5. Algoritma yang digunakan untuk membuat model klasifikasi penyakit diabetes adalah s*upport vector machine*.
6. Algoritma *bagging* digunakan untuk menghindari *overfitting* serta mengurangi variansi pada model klasifikasi *support vector machine*.
7. Algoritma *grid search* digunakan untuk mengoptimasi *hyperparameter* yang ada pada algoritma *support vector machine* dan *bagging*.

1.6 Luaran yang diharapkan

Adapun luaran yang diharapkan dari penelitian ini yaitu berupa model klasifikasi yang dapat mengklasifikasi penyakit diabetes dengan baik.

1.7 Sistematika Penulisan

Berikut adalah sistematika penulisan dalam menyusun proposal seminar teknologi ini:

BAB 1 PENDAHULUAN

Pada bab ini terdapat Latar Belakang, Rumusan Masalah, Tujuan Penelitian, Manfaat Penelitian, Ruang Lingkup, Luaran yang Diharapkan, dan Sistematika Penulisan.

BAB 2 LANDASAN TEORI

Pada bab ini berisi teori mendasar yang digunakan sebagai acuan dalam penyusuan laporan penilitan ini.

BAB 3 METODOLOGI PENELITIAN

Pada bab ini berisi langkah-langkah yang dilakukan dalam penelitian, serta metode yang ada dalam penelitian ini.

BAB 4 HASIL DAN PEMBAHASAN

Pada bab ini berisi tentang penjelasan segala proses yang dilakukan serta hasil yang didapat pada penelitian ini.

BAB 5 PENUTUP

Pada bab ini berisi kesimpulan yang dibuat berdasarkan uraian masalah dan hasil yang disajikan, serta saran-saran yang disampaikan oleh penulis. Dimana saran tersebut dapat dijadikan sebagai acuan untuk pengembangan penelitian selanjutnya.

DAFTAR PUSTAKA

LAMPIRAN