PERBANDINGAN KINERJA ALGORITMA K-NEAREST NEIGHBOR MENGGUNAKAN SMOTE DAN ALGORITMA K-NEAREST NEIGHBOR TANPA SMOTE DALAM DIAGNOSIS PENYAKIT DIABETES PADA DATA TIDAK SEIMBANG

SKRIPSI

Disusun Sebagai Salah Satu Syarat
Untuk Memperoleh Gelar Sarjana Komputer
Pada Departemen Ilmu Komputer/Informatika

Disusun Oleh :
Amelia Gita Pertiwi
24010315130100

DEPARTEMEN ILMU KOMPUTER/ INFORMATIKA
FAKULTAS SAINS DAN MATEMATIKA
UNIVERSITAS DIPONEGORO
2019

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini

Nama : Amelia Gita Pertiwi

NIM : 24010315130100

Judul : Perbandingan Kinerja Algoritma *K-Nearest Neighbor* Menggunakan SMOTE dan Algoritma *K-Nearest Neighbor* Tanpa SMOTE dalam Diagnosis Penyakit Diabetes pada Data Tidak Seimbang

Dengan ini saya menyatakan bahwa dalam tugas akhir/ skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang sepengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali secara tertulis diacu dalam naskah ini dan disebutkan di daftar pustaka.

Semarang, 27 Agustus 2019

Amelia Gita Pertiwi

24010315130100

HALAMAN PENGESAHAN

Judul : Perbandingan Kinerja Algoritma *K-Nearest Neighbor* Menggunakan SMOTE dan Algoritma *K-Nearest Neighbor* Tanpa SMOTE dalam Diagnosis Penyakit Diabetes pada Data Tidak Seimbang

Nama : Amelia Gita Pertiwi

NIM : 24010315130100

Telah diujikan pada sidang skripsi pada tanggal 5 Agustus 2019 dan dinyatakan lulus pada tanggal 5 Agustus 2019.

Semarang, 27 Agustus 2019

Mengetahui,
a.n. Ketua Departemen Ilmu Komputer/ Informatika
Sekretaris Departemen

Dr. Eng. Adi Wibowo, S.Si, M.Kom
NIP. 198203092006041002

Panitia Penguji Skripsi
Ketua,

Dr. Retno Kusumaningrum, S.Si, M.Kom
NIP. 198104202005012001

HALAMAN PENGESAHAN

Judul : Perbandingan Kinerja Algoritma *K-Nearest Neighbor* Menggunakan SMOTE dan Algoritma *K-Nearest Neighbor* Tanpa SMOTE dalam Diagnosis Penyakit Diabetes pada Data Tidak Seimbang

Nama : Amelia Gita Pertiwi

NIM : 24010315130100

Telah diujikan pada sidang skripsi dan dinyatakan lulus pada tanggal 5 Agustus 2019.

Semarang, 27 Agustus 2019

Dosen Pembimbing,

Nurdin Bahtiar, S.Si, MT.

NIP. 197907202003121002

KATA PENGANTAR

Segala puji syukur penulis panjatkan kepada Allah SWT atas karunia-Nya yang diberikan, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Perbandingan Kinerja Algoritma *K-Nearest Neighbor* Menggunakan SMOTE dan Algoritma *K-Nearest Neighbor* Tanpa SMOTE dalam Diagnosis Penyakit Diabetes pada Data Tidak Seimbang".

Penyusunan dan penulisan laporan skripsi ini tidak terlepas dari bantuan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu, dengan segala hormat penulis mengucapkan terimakasih kepada:

1. Ibu Dr. Retno Kusumaningrum, S.Si, M.Kom., selaku Ketua Departemen Ilmu Komputer/ Informatika Fakultas Sains dan Matematika Universitas Diponegoro Semarang.
2. Bapak Panji Wisnu Wirawan, S.T., M.T., selaku Koordinator Skripsi Departemen Ilmu Komputer/ Informatika Fakultas Sains dan Matematika Universitas Diponegoro Semarang.
3. Bapak Nurdin Bahtiar, S.Si, MT., selaku dosen pembimbing skripsi yang telah membantu dalam membimbing dan mengarahkan penulis hingga skripsi ini dapat terselesaikan dengan baik.
4. Kedua orang tua dan teman dekat yang telah mendukung, membantu, serta memberi semangat kepada penulis dalam proses menyelesaikan skripsi ini.
5. Semua pihak yang telah membantu kelancaran dalam penyusunan skripsi, yang tidak dapat penulis sebutkan satu persatu.

Penulis menyadari bahwa laporan skripsi ini masih jauh dari kata sempurna. Oleh karena itu, kritik dan saran sangat penulis harapkan untuk perbaikan kedepannya. Semoga laporan ini dapat bermanfaat bagi pembaca pada umumnya dan bagi penulis pada khususnya.

Semarang, 27 Agustus 2019

Penulis,

Amelia Gita Pertiwi

24010315130100

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI UNTUK KEPENTINGAN AKADEMIS

Sebagai civitas akademik Universitas Diponegoro, saya yang bertanda tangan di bawah ini:

Nama	: Amelia Gita Pertiwi
NIM	: 24010315130100
Program Studi	: Informatika
Departemen	: Ilmu Komputer/ Informatika
Fakultas	: Sains dan Matematika
Jenis Karya	: Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan **Hak Bebas Royalti Noneksklusif (*Non-exclusive RoyaltiFree Right*)** kepada Universitas Diponegoro atas karya ilmiah saya yang berjudul:

Perbandingan Kinerja Algoritma K-Nearest Neighbor Menggunakan SMOTE dan Algoritma K-Nearest Neighbor Tanpa SMOTE dalam Diagnosis Penyakit Diabetes pada Data Tidak Seimbang

beserta perangkat yang ada (jika diperlukan). Dengan Hal Bebas Royalti Non-eksklusif ini Universitas Diponegoro berhak menyimpan, mengalihmedia/ formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan skripsi saya tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis/ pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Semarang, 27 Agustus 2019

Yang menyatakan

Amelia Gita Pertiwi

24010315130100

ABSTRAK

Penyakit diabetes merupakan salah satu penyebab kematian hampir 70% di seluruh dunia menurut Profil Kesehatan Indonesia tahun 2017. Tingginya angka kematian tersebut menyebabkan perlunya dilakukan upaya untuk menurunkan jumlah penderita penyakit diabetes dengan cara melakukan penelitian-penelitian yang mengarah pada melakukan suatu diagnosis sehingga dapat mendeteksi seseorang terkena penyakit diabetes secara akurat. Penelitian ini mencoba membandingkan kinerja algoritma *K-Nearest Neighbor* menggunakan *Synthetic Minority Over-sampling Technique* dan algoritma *K-Nearest Neighbor* tanpa *Synthetic Minority Over-sampling Technique* dalam mendiagnosis penyakit diabetes pada data tidak seimbang. Parameter yang diujicobakan adalah nilai *k* pada *K-Nearest Neighbor* dan *Synthetic Minority Over-sampling Technique*. Pengujian dilakukan menggunakan strategi *K-Fold Cross Validation*. Data yang digunakan pada penelitian ini sebanyak 3876 data yang berasal dari Rumah Sakit Pusat Pertamina. Berdasarkan hasil pengujian yang dilakukan menunjukkan bahwa nilai akurasi yang dihasilkan dalam mendiagnosis penyakit diabetes dengan menggunakan *Synthetic Minority Over-sampling Technique* lebih baik daripada akurasi yang dihasilkan tanpa menggunakan *Synthetic Minority Over-sampling Technique* dengan peningkatan akurasi tertinggi sebesar 8,25%. Rata-rata akurasi tertinggi didapatkan ketika nilai $k = 3$ pada *K-Nearest Neighbor,* nilai $k = 5$ pada *Synthetic Minority Over-sampling Technique*, dan nilai *fold* = 10, yaitu mencapai 78,06%.

Kata Kunci: Diagnosis penyakit diabetes, *Imbalanced datasets*, *K-Nearest Neighbor*, *Synthetic Minority Over-sampling Technique*

ABSTRACT

According to the Indonesian Health Profile in 2017, diabetes is one of the causes of death for almost 70% in the world. The high mortality rate induces the need for making the effort to reduce the number of people with diabetes by conducting studies that lead to making a diagnosis so that can detect a person with diabetes accurately. This study tries to compare the performance of the K-Nearest Neighbor algorithm using Synthetic Minority Over-sampling Technique and the K-Nearest Neighbor algorithm without Synthetic Minority Over-sampling Technique in diagnosing diabetes on imbalanced datasets. The parameters tested are the k value of the K-Nearest Neighbor and Synthetic Minority Over-sampling Technique. The testing is carried out using the K-Fold Cross Validation strategy. The data used in this study were 3876 data from Pertamina Central Hospital. Based on the results of tests conducted, it shows that the value of accuracy produced in diagnosing diabetes by using Synthetic Minority Over-sampling Technique is better than the accuracy produced without using Synthetic Minority Over-sampling Technique with the highest accuracy increase of 8,25%. The highest average accuracy is obtained when the value of k = 3 in the K-Nearest Neighbor, k = 5 in the Synthetic Minority Over-sampling Technique, and fold = 10, which reaches 78,06%.

Keywords : *Diabetes diagnosis, Imbalanced datasets, K-Nearest Neighbor, Synthetic Minority Over-sampling Technique*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

Bab ini membahas latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup, dan sistematika penulisan skripsi mengenai diagnosis penyakit diabetes menggunakan metode *K-Nearest Neighbor* (KNN) dan *Synthetic Minority Over-Sampling Technique* (SMOTE).

1.1 Latar Belakang

Diabetes merupakan penyakit kronis yang terjadi karena kelainan sekresi insulin pada kenaikan glukosa yang tidak teratur. Diabetes akan meningkatkan gula darah dalam tubuh sehingga dapat terjadi penyakit komplikasi yang menyebabkan beberapa resiko seperti stroke, penyakit jantung, kebutaan, gagal ginjal, dan kematian (Ravikumar & Veena, 2014). Berdasarkan Profil Kesehatan Indonesia, pada tahun 2013 diabetes merupakan penyakit dengan jumlah penderita terbanyak ke enam di Indonesia dengan proporsi 4,8% di bawah penyakit hipertensi, artritis, stroke, masalah gigi dan mulut, dan penyakit paru obstruktif menahun. Selain itu, penyakit diabetes juga merupakan salah satu penyebab kematian hampir 70% di seluruh dunia (Kemenkes RI, 2018).

Penyakit diabetes dapat dibagi menjadi dua tipe, yaitu diabetes tipe 1 atau *insulin dependent* diabetes yang biasanya terjadi pada kalangan anak-anak dan remaja atau bahkan sejak lahir karena tubuh mereka tidak dapat memproduksi insulin dengan baik, serta diabetes tipe 2 atau *non-insulin dependent* diabetes yang terjadi pada orang dewasa akibat dari pola hidup yang salah. Diabetes tipe 2 merupakan jenis penyakit diabetes yang paling sering ditemukan. Jika tidak dilakukan penanganan yang efektif, maka jumlah penderita penyakit diabetes akan semakin meningkat (Herwanto, 2006).

Berdasarkan penjelasan di atas, maka perlu dilakukan upaya untuk menurunkan jumlah penderita penyakit diabetes dengan cara melakukan penelitian-penelitian yang mengarah pada melakukan suatu diagnosis sehingga dapat terdeteksi apabila seseorang memiliki penyakit diabetes, dengan begitu dapat dilakukan pencegahan dan penanganan bagi penderita penyakit diabetes dengan pendekatan yang menyeluruh. Salah satu caranya yaitu menggunakan *machine learning* sebagai salah satu teknik pembelajaran.

Machine learning merupakan salah satu bidang dari *artificial intelligence* dimana kita dapat membuat sebuah komputer atau mesin untuk mempelajari sesuatu secara otomatis (Kourou, et al., 2014). Pemanfaatan *machine learning* dalam bidang kesehatan sudah banyak dilakukan, di antaranya diagnosis penyakit diabetes dengan menggunakan algoritma *Backward Elimination* dan *K-Nearest Neighbor* (Hermawanti & Ghadati, 2014), diagnosis penyakit hipertensi kehamilan dengan menggunakan algoritma *Decision Tree* (Muzakir & Wulandari, 2016), dan diagnosis penyakit diabetes dengan menggunakan algoritma C4.5 dan *K-Nearest Neighbor* (Karyono, 2016).

Terdapat beberapa metode dalam mendiagnosis penyakit diabetes, seperti *K-Nearest Neighbor* yang menghasilkan akurasi sebesar 95,29% (Hermawanti & Safriandono, 2016), *Decision Tree* yang menghasilkan akurasi sebesar 92,5% (Mirza, et al., 2018), *Naïve Bayes* yang menghasilkan akurasi sebesar 78,8% (Alghamdi, et al., 2017), dan lain sebagainya. Namun, algoritma-algoritma tersebut memiliki beberapa kekurangan, misalnya pada *Decision Tree* akan terjadi pengakumulasian jumlah error dari setiap tingkat dalam sebuah pohon keputusan yang besar, serta sulit untuk mendesain pohon keputusan yang optimal jika jumlah datanya besar. Begitu juga pada *Naïve Bayes,* sebuah probabilitas tidak bisa mengukur seberapa besar tingkat keakuratan sebuah prediksi. Dengan kata lain, jika sebuah probabilitas tidak dapat merepresentasikan sebuah data maka prediksi yang dihasilkan kurang akurat. Selain itu, terdapat permasalahan data yang sering muncul pada kelas lain dan muncul juga pada kelas yang diuji mengakibatkan kesalahan prediksi. Hal ini yang menyebabkan metode *Naïve Bayes* masih belum optimal (Socrates, 2016).

Algoritma *K-Nearest Neighbor* adalah salah satu algoritma yang biasa digunakan untuk klasifikasi, selain itu dapat pula digunakan untuk estimasi dan prediksi (Larose, 2005). Algoritma *K-Nearest Neighbor* mengelompokkan data baru yang belum diketahui kelasnya dengan memilih data sejumlah *k* yang letaknya paling dekat dari data baru. Kelas dengan frekuensi terbanyak dari data terdekat sejumlah *k* akan dipilih sebagai kelas yang diprediksi untuk data baru. Pada umumnya nilai *k* menggunakan jumlah ganjil agar tidak terdapat jarak yang sama dalam proses klasifikasi. Jauh atau dekatnya tetangga dihitung menggunakan jarak *Euclidean* (Shofia, Putri, & Arwan, 2017). Algoritma *K-Nearest Neighbor* memiliki beberapa kelebihan, yaitu sederhana dan mudah untuk diterapkan, serta efektif apabila data latihnya besar, sehingga pada penelitian skripsi ini dipilih algoritma *K-Nearest Neighbor* untuk melakukan klasifikasi. Akan tetapi, algoritma *K-Nearest Neighbor* tidak dilengkapi dengan kemampuan untuk bekerja pada dataset tidak seimbang (Siringoringo, 2018).

Salah satu permasalahan yang dapat muncul dalam melakukan klasifikasi adalah adanya data yang tidak seimbang (*imbalanced datasets*) dalam hal jumlah dari masing-masing kelas atau condong ke salah satu kutub, misalnya lebih condong ke kelas negatif atau sebaliknya. Secara umum, algoritma untuk klasifikasi akan menghasilkan suatu model dengan tingkat kepekaan yang minim terhadap kelas minoritas ketika menerima *imbalanced datasets* (He & Ma, 2013). Maka dari itu penting untuk menangani masalah pada *imbalanced datasets*.

Terdapat beberapa pendekatan yang digunakan untuk menangani masalah ketidakseimbangan pada dataset, yaitu *re-sampling* (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), *cost-sensitive learning* (Zhou & Liu, 2006), dan *one-class classification* (Juszczak & Duin, 2003). Pendekatan pertama yaitu melakukan penanganan terhadap *imbalanced dataset* dengan cara membentuk data sintetis untuk menyeimbangkan jumlah data dari masing-masing kelas, sedangkan pendekatan kedua dan ketiga melakukan penanganan dengan cara memodifikasi algoritma klasifikasinya. Pendekatan yang dilakukan pada penelitian skripsi ini yaitu teknik *re-sampling*. Teknik *re-sampling* secara umum dibagi menjadi dua, yaitu *over-sampling* kelas minor dan *under-sampling* kelas mayor. Karena tidak ingin kehilangan data-data penting yang mungkin ditimbulkan oleh teknik *under-sampling* maka pada penelitian skripsi ini menggunakan teknik *over-sampling* untuk menangani ketidakseimbangan dataset.

SMOTE (*Synthetic Minority Over-sampling Technique*) merupakan salah satu teknik *over-sampling* yang sering digunakan untuk menangani masalah *imbalanced datasets* dengan membuat data sintetis pada kelas data minor sehingga data menjadi seimbang (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) dan diharapkan berimbas pada kinerja klasifikasi yang lebih baik. Pada beberapa kasus penelitian, terjadi peningkatan kinerja (dalam hal ini akurasi) dari sekitar 65% untuk distribusi data awal menjadi sekitar 80% karena telah dilakukan SMOTE sehingga data yang awalnya tidak seimbang menjadi benar-benar seimbang (Pears, Finlay, & Connor, 2014). Algoritma SMOTE juga dapat digunakan untuk membantuk algoritma *K-Nearest Neighbor* yang tidak bisa menangani kasus dataset tidak seimbang.

Berdasarkan penjelasan di atas, maka pada penelitian ini dikaji mengenai perbandingan kinerja algoritma *K-Nearest Neighbor* (KNN) menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) dan algoritma *K-Nearest Neighbor* (KNN)

tanpa *Synthetic Minority Over-sampling Technique* (SMOTE) dalam mendiagnosis penyakit diabetes pada data tidak seimbang.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan di atas, maka dapat dirumuskan permasalahan dalam penelitian ini yaitu bagaimana perbandingan kinerja algoritma *K-Nearest Neighbor* menggunakan *Synthetic Minority Over-sampling Technique* dan algoritma *K-Nearest Neighbor* tanpa *Synthetic Minority Over-sampling Technique* dalam mendiagnosis penyakit diabetes pada data tidak seimbang?

1.3 Tujuan dan Manfaat

Tujuan yang ingin dicapai dari penelitian skripsi ini yaitu membandingkan kinerja algoritma *K-Nearest Neighbor* (KNN) menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) dan algoritma *K-Nearest Neighbor* (KNN) tanpa *Synthetic Minority Over-sampling Technique* (SMOTE) dalam mendiagnosis penyakit diabetes pada data tidak seimbang. Sedangkan manfaat dari penelitian skripsi ini adalah menghasilkan model klasifikasi yang dapat dimanfaatkan dalam mendiagnosis penyakit diabetes untuk data yang tidak seimbang.

1.4 Ruang Lingkup

Pada penyusunan skripsi ini, diberikan ruang lingkup yang jelas agar pembahasan lebih terarah dan tidak menyimpang dari tujuan penulisan, yaitu:

1. *Dataset* penelitian yang digunakan untuk membangun model klasifikasi pada penelitian skripsi ini yaitu *dataset* dari Rumah Sakit Pusat Pertamina.
2. Implementasi dilakukan dengan bahasa pemrograman *Phyton* dan berbasis *web*.
3. Penulisan bilangan desimal pada dokumentasi dibatasi maksimal tiga angka di belakang koma.

1.5 Sistematika Penulisan

Untuk memberikan gambaran yang urut dan jelas mengenai pembahasan penyusunan skripsi ini, berikut merupakan sistematika penulisan yang digunakan yaitu:

BAB I PENDAHULUAN

Bab ini berisi tentang latar belakang masalah, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan dalam penyusunan skripsi

yang berjudul "Perbandingan Kinerja Algoritma *K-Nearest Neighbor* Menggunakan SMOTE dan Algoritma *K-Nearest Neighbor* Tanpa SMOTE dalam Diagnosis Penyakit Diabetes pada Data Tidak Seimbang"

BAB II LANDASAN TEORI

Bab ini membahas mengenai penelitian-penelitian terkait yang sudah dilakukan sebelumnya dan teori-teori dalam penyusunan skripsi yang berjudul "Perbandingan Kinerja Algoritma *K-Nearest Neighbor* Menggunakan SMOTE dan Algoritma *K-Nearest Neighbor* Tanpa SMOTE dalam Diagnosis Penyakit Diabetes pada Data Tidak Seimbang"

BAB III METODOLOGI PENELITIAN

Bab ini membahas tentang tahapan-tahapan yang dilakukan dalam penelitian skripsi yang dilakukan. Tahapan tersebut meliputi pengumpulan data, *preprocessing data*, SMOTE, pembagian data latih dan data uji, pengujian, serta evaluasi.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas mengenai hasil skenario dan analisa eksperimen yang telah dilakukan sebelumnya, dimulai dari pengumpulan data hingga hasil dan analisis dari setiap eksperimen.

BAB V PENUTUP

Bab ini berisi tentang kesimpulan dari uraian yang telah dijabarkan pada bab-bab sebelumnya dan saran untuk pengembangan peneliti lebih lanjut.