PENERAPAN SMOTE UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS PADA KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN C5.0, RANDOM FOREST DAN SVM

Skripsi

Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Strata-1 Ilmu Komputer

Oleh

M. KHAIRUL REZKI

NIM 1811016310001

PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU

JUNI 2024

PENERAPAN SMOTE UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS PADA KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN C5.0, RANDOM FOREST DAN SVM

Skripsi

Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Strata-1 Ilmu Komputer

Oleh

M. KHAIRUL REZKI
NIM 1811016310001

PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
JUNI 2024

SKRIPSI

PENERAPAN SMOTE UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS PADA KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN C5.0, RANDOM FOREST DAN SVM

Oleh:

M. KHAIRUL REZKI

NIM. 1811016310001

Telah dipertahankan di depan Dosen Penguji pada tanggal 4 Juni 2024.

Susunan Dosen Penguji:

Pembimbing I

Muhammad Itqan Mazdadi, S.Kom., M.Kom.

NIP. 199006122019031013

Dosen Penguji I

Muliadi, S.Kom, M.Sc.

NIP. 197804222010121002

Pembimbing II

Fatma Indriani, S. T., M. I. T., Ph. D.

NIP. 198404202008122004

Dosen Penguji II

Triando Hamonangan Saragih, S.Kom., M.Kom.

NIP. 199308242019031012

Banjarbaru, 15 Juni 2024

Koordinator Program Studi Ilmu Komputer

Irwan Budiman, S. T., M. Kom.

NIP. 197703252008121001

SURAT PERNYATAAN

Dengan ini saya menyatakan bahwa dalam jurnal ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Pustaka.

Banjarbaru, 30 Juni 2024

M. Khairul Rezki

ABSTRAK

Penerapan SMOTE Untuk Mengatasi Ketidakseimbangan Kelas Pada Klasifikasi Penyakit Diabetes Menggunakan C5.0, Random Forest Dan SVM

(Oleh: M. Khairul Rezki; Pembimbing: Muhammad Itqan Mazdadi, S.Kom., M.Kom. dan Fatma Indriani, S.T M.I.T; 2024; halaman)

Penerapan metodologi kategorisasi dalam klasifikasi diabetes sering kali memberikan hasil yang tidak memuaskan sebagai akibat dari sifat dataset yang rumit dan distribusi kelas yang tidak seimbang di dalam data. Oleh karena itu, penerapan SMOTE untuk mengatasi ketidakseimbangan kelas dalam klasifikasi sering kali memberikan hasil yang tidak memuaskan karena kerumitan dataset dan banyaknya faktor yang terlibat. Akibatnya, serangkaian tes dilakukan untuk mengevaluasi ketepatan berbagai metode klasifikasi. Penelitian ini bertujuan untuk mengevaluasi ketepatan model klasifikasi C5.0, Random Forest, dan SVM dengan menggunakan pendekatan reguler dan berbasis SMOTE. Metodologi terdiri dari pemilihan dataset, tinjauan umum algoritma kategorisasi seperti C5.0, Random Forest, dan SVM, pemanfaatan teknik SMOTE, validasi melalui validasi terpisah, prapemrosesan yang menggabungkan normalisasi min-max, dan evaluasi eksekusi dengan menggunakan matriks kerancuan dan analisis AUC. Dataset ini diperoleh melalui Kaggle dengan tujuan untuk mengurangi distribusi kelas yang tidak seimbang dalam dataset diabetes dengan menggunakan SMOTE. Dataset ini terdiri dari total 768 contoh, dengan 268 sampel untuk individu dengan diabetes dan 500 sampel untuk individu tanpa diabetes. Sebelum menggunakan SMOTE, nilai presisi untuk kategorisasi menggunakan C5.0, Random Forest, dan SVM masing-masing adalah 0.714, 0.733, dan 0.746. Nilai AUC yang sesuai dengan data yang diberikan adalah 0,745, 0,824, dan 0,799. Teknik SMOTE menghasilkan nilai akurasi 0,603, 0,727, dan 0,727 untuk proses yang identik. Nilai AUC yang terkait adalah 0,734, 0,831, dan 0,794. Analisis menunjukkan bahwa penggunaan SMOTE memiliki efek yang terbatas pada tiga model kategorisasi. Hal ini dikarenakan adanya risiko overfitting pada dataset, yang mengakibatkan ketergantungan yang berlebihan pada data yang dihasilkan secara artifisial untuk kelas minoritas. Akibatnya, hal ini menyebabkan penurunan kinerja model, termasuk nilai presisi dan AUC.

Kata Kunci: SMOTE, C5.0, Random Forest, SVM, Diabetes.

ABSTRAK

Application of SMOTE to Address Class Imbalance in Diabetes Disease Categorization Utilizing C5.0, Random Forest, and Support Vector Machine

(By: M. Khairul Rezki; Pembimbing: Muhammad Itqan Mazdadi, S.Kom., M.Kom. dan Fatma Indriani, S.T M.I.T; 2024; page)

The application of categorization methodologies in the classification of diabetes frequently produces unsatisfactory outcomes as a result of the intricate nature of the dataset and the imbalanced distribution of classes within the data. Therefore, the application of SMOTE to tackle class imbalance in classification frequently yields unsatisfactory outcomes due to the intricacy of the dataset and the multitude of factors involved. consequently, a series of tests were conducted to evaluate the precision of various classification methods. This work aims to evaluate the precision of C5.0, Random Forest, and SVM classification models by employing regular and SMOTE-based approaches. The methodology consists of the selection of datasets, an overview of categorization algorithms such as C5.0, Random Forest, and SVM, the utilization of the SMOTE technique, validation through split validation, preprocessing incorporating min-max normalization, and evaluation of execution using confusion matrices and AUC analysis. The dataset was acquired via Kaggle with the aim of mitigating the imbalanced class distribution in a diabetes dataset by employing SMOTE. The dataset consists of a total of 768 instances, with 268 samples being to individuals with diabetes and 500 samples corresponding to individuals without diabetes. Prior to employing SMOTE, the precision values for categorization using C5.0, Random Forest, and SVM were 0.714, 0.733, and 0.746 respectively. The AUC values corresponding to the given data were 0.745, 0.824, and 0.799. The SMOTE technique resulted in accuracy values of 0.603, 0.727, and 0.727 for the identical processes. The associated AUC values were 0.734, 0.831, and 0.794. The analysis suggests that the use of SMOTE has a limited effect on the three categorization models. This is because there is a risk of overfitting on the dataset, which results in an excessive dependence on artificially generated data for minority classes. Consequently, this leads to a decrease in the performance of the models, including precision and AUC scores.

Keywords: SMOTE, C5.0, Random Forest, SVM, Diabetes.

KATA PENGANTAR

Puji syukur penulis panjatkan ke Tuhan kita Yang Maha Esa karena atas berkat rahmat dan karunia-Nya penulis dapat menyelesaikan jurnal yang berjudul "*Application of Smote to Address Class Imbalance in Diabetes Disease Categorization Utilizing C5.0, Random Forest, and Support Vector Machine"* untuk memenuhi syarat dalam menyelesaikan pendidikan program S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat.

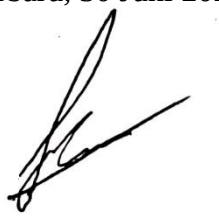
Pada lembar ini penulis ingin menyampaikan ucapan terimakasih kepada semua pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan jurnal ini, adapun yang dimaksud adalah sebagai berikut:

1. Diri saya sendiri yang tidak pernah patah semangat walaupun banyak menemui kesulitan baik disebabkan oleh diri sendiri maupun hal lain.
2. Keluarga besar yang selalu memberikan bantuan, semangat, doa dan dukungan dalam proses penyelesaian jurnal ini.
3. Bapak Muhmmad Itqan Mazdadi, S.Kom., M.Kom. selaku dosen pembimbing utama yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian jurnal ini.
4. Ibu Fatma Indriani, S.T., M. I. T., Ph. D. selaku dosen pembimbing pendamping yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian jurnal ini.
5. Bapak Irwan Budiman, S.T., M. Kom. selaku Ketua Program Studi Ilmu Komputer FMIPA ULM, atas bantuan dan izin beliau jurnal ini dapat diselesaikan.
6. Seluruh Dosen dan staff Program Studi Ilmu Komputer FMIPA ULM atas ilmu dan bantuan yang diberikan selama ini yang sangat bermanfaat.
7. Sigit, Fawwaz, Hamdani, Hevny, Akbar, dan Nadem yang selalu mendukung dalam berbagai hal.

8. Teman-teman dan sahabat-sahabat keluarga Ilmu Komputer yang memberikan dukungan dan selalu mengingatkan serta mendoakan dalam proses mengerjakan jurnal.
9. Serta semua pihak yang tidak dapat disebutkan satu persatu yang telah turut membantu dalam penyelesaian jurnal ini.

Akhir kata penulis menyadari sepenuhnya bahwa penulisan ini jauh dari sempurna. Semoga tulisan ini dapat bermanfaat bagi ilmu pengetahuan dan pembaca khususnya serta mendapat keridhaan Allah SWT.

Banjarbaru, 30 Juni 2024



M. Khairul Rezki