SKRIPSI

PREDIKSI RISIKO TERDIAGNOSIS PENYAKIT DIABETES MENGGUNAKAN METODE *RANDOM FOREST*

BILLY SUTAWIJAYA

NPM: 2017710035

PROGRAM STUDI MATEMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2023

FINAL PROJECT

RISK PREDICTION OF DIAGNOSED DIABETES USING RANDOM FOREST METHOD

BILLY SUTAWIJAYA

NPM: 2017710035

DEPARTMENT OF MATHEMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2023

LEMBAR PENGESAHAN

PREDIKSI RISIKO TERDIAGNOSIS PENYAKIT DIABETES MENGGUNAKAN METODE *RANDOM FOREST*

Billy Sutawijaya

NPM: 2017710035

Bandung, 9 Agustus 2023

Menyetujui,

Pembimbing 1	Pembimbing 2
Dr. Erwinna Chendra	Liem Chin, M.Si.

Ketua Penguji	Anggota Penguji
Benny Yong, Ph.D.	Rizky Reza Fauzi, D.Phil.Math.

Mengetahui,

Ketua Program Studi

Dr. Livia Owen

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

PREDIKSI RISIKO TERDIAGNOSIS PENYAKIT DIABETES MENGGUNAKAN METODE *RANDOM FOREST*

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
9 Agustus 2023

Billy Sutawijaya
NPM: 2017710035

ABSTRAK

Diabetes adalah salah satu penyakit kronis yang diderita oleh hampir 537 juta orang dewasa di dunia. Penggunaan metode pembelajaran mesin untuk kebutuhan kesehatan telah banyak dilakukan, termasuk diantaranya untuk prediksi diagnosis penyakit diabetes. Salah satu metode yang dapat digunakan untuk memprediksi diagnosis diabetes adalah metode *Random Forest*. Metode *Random Forest* telah terbukti efektif dalam berbagai aplikasi prediksi dan klasifikasi. *Random Forest* adalah metode *ensemble learning* yang memanfaatkan sejumlah besar pohon keputusan untuk memperoleh prediksi akurat. Model yang dibangun kemudian dievaluasi menggunakan *Matriks Konfusi* untuk mengevaluasi model. Penelitian ini menggunakan *Pima Indian Diabetes Dataset* dan *130 US Hospital Dataset* dengan skala data yang berbeda. Proses pengolahan data dilakukan dengan dan tanpa *resampling*. Hasil evaluasi model menunjukkan metode *Random Forest* dapat diandalkan pada data skala kecil maupun besar. Selain itu, data dengan *resampling* menunjukkan performa yang lebih baik dan meminimumkan risiko kesalahan prediksi.

Kata-kata kunci: Diabetes; Metode *Random Forest*; *Resampling*; Matriks Konfusi.

ABSTRACT

Diabetes is a chronic disease that affects nearly 537 million adults worldwide. The use of machine learning methods for health needs has been carried out a lot, including to predict the diagnosis of diabetes. One of the methods that can be used to predict the diagnosis of diabetes is the Random Forest method. The Random Forest method has proven effective in various prediction and classification applications. Random Forest is a ensemble learning method that utilizes a large number of decision trees to obtain accurate predictions. The developed model is then evaluated using Confusion Matrix to evaluate the model. This study used Pima Indian Diabetes Dataset and 130 US Hospital Dataset with different data scales. Data processing is done with and without resampling. The model evaluation results show that the Random Forestmethod can be relied upon on both small and large scale data. In addition, data with resampling shows better performance and minimizes prediction errors.

Keywords: Diabetes; Random Forest Method; Resampling; Confusion Matrix.

KATA PENGANTAR

Puji syukur kepada Allah Tuhan semesta alam yang Maha Pengasih lagi Maha Penyayang sehingga penulis dapat menyelesaikan skripsi ini dengan judul "Prediksi Risiko Terdiagnosis Penyakit Diabetes menggunakan Metode Random Forest". Skripsi ini disusun sebagai salah satu syarat unutuk menyelesaikan studi Strata-1 Program Studi Matematika, Fakultas Teknologi Informasi dan Sains (FTIS), Universitas Katolik Parahyangan, Bandung. Penulis sangat berharap skripsi ini dapat menjadi manfaat bagi orang yang membacanya, khususnya dalam dunia Kesehatan.

Penulis ingin memberikan ucapan terima kasih kepada pihak-pihak yang telah berkontribusi dan memberikan bantuan baik berupa tindakan, saran, kritikan, ataupun bantuan moril sehingga skripsi ini dapat diselesaikan. Terima kasih yang sebesar-besarnya kepada:

1. Papi, Mami, dan Cece yang selalu memberikan bantuan, mendoakan, mendukung, dan memberikan semangat dalam proses pengerjaan skripsi ini hingga dapat diselesaikan dengan baik.
2. Ibu Dr. Erwinna Chendra sebagai dosen pembimbing-1 yang telah sabar dalam mengarahkan, memberikan ide dan masukan serta berbagai saran selama proses pengerjaan skripsi ini.
3. Bapak Liem Chin, M.Si. selaku dosen pembimbing-2 yang telah sabar dalam memberikan perbaikan tulisan, memberi masukan berupa saran dan kritikan, serta menambahkan ide-ide untuk menyelesaikan skripsi ini.
4. Bapak Dr. Daniel Salim selaku koordinator skripsi yang telah memberikan ilmu, saran serta bantuan pada masa perkuliahan dan penyusunan skripsi.
5. Bapak Benny Yong, Ph.D. dan Bapak Rizky Reza Fauzi, D.Phil.Math. yang telah menguji penulis pada sidang skripsi, memberikan saran, kritik, dan masukan sehingga skripsi ini bisa diselesaikan.
6. Bapak Ibu seluruh dosen yang telah sabar dalam mengajarkan penulis pada setiap mata kuliahnya, memberikan ilmunya, serta membantu penulis dalam berbagai hal.
7. Enrico dan Anthony sebagai sahabat seangkatan penulis yang selalu berupaya memberikan bantuan serta motivasi kepada penulis dalam menyelesaikan skripsi ini.
8. Teman-teman Matematika Angkatan 2017 yang tidak dapat disebutkan satu persatu.
9. Dan kepada segala pihak yang telah membantu penulis dalam proses pengerjaan skripsi ini.

Penelitian ini tentu tidak dapat terhindarkan dari kekurangan dan kesalahan yang disebabkan karena kelalaian penulis. Oleh karena itu, penulis sangat berharap pembaca dapat memberikan masukan berupa kritik dan saran yang dapat memperbaiki skripsi ini menjadi lebih baik.

Bandung, 9 Agustus 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Diabetes adalah salah satu penyakit kronis yang menyebabkan terjadinya gangguan metabolisme pada tubuh manusia dan ditandai dengan peningkatan kadar glukosa dalam darah. Sekitar 537 juta orang dewasa di dunia yang berusia 20 hingga 79 tahun telah menderita diabetes selama tahun 2021 dan dikhawatirkan akan mencapai 783 juta pada tahun 2045. Sementara itu, sekitar 19,4 juta orang di Indonesia terdiagnosis diabetes. Dalam waktu yang lama, diabetes dapat menyebabkan kerusakan yang lebih serius pada jantung, pembuluh darah, ginjal, mata, dan saraf [1].

Diagnosis penyakit diabetes biasanya dilakukan oleh dokter berdasarkan hasil pemeriksaan gejala medis tertentu. Beberapa faktor yang berpengaruh secara signifikan adalah faktor usia, genetik, dan pola hidup. Faktor usia dan genetik menjadi faktor yang tak dapat diubah, tetapi untuk faktor pola hidup masih dapat diubah untuk mencegah terjadinya diabetes.

Dalam beberapa kasus, dokter maupun tenaga medis memiliki keterbatasan dalam mendiagnosis penyakit yang dapat diakibatkan oleh kelelahan dan ketidaktelitian. Dalam hal ini, peran teknologi dapat membantu mengurangi kesalahan dan ketidaktelitian tenaga medis. Salah satu teknologi yang dapat membantu peran medis dalam mendiagnosis suatu penyakit adalah pembelajaran mesin.

Pembelajaran mesin saat ini berkembang pesat dan banyak digunakan pada berbagai bidang, salah satunya dalam bidang kesehatan. Pembelajaran mesin digunakan untuk menemukan pola tertentu dalam data sehingga dapat terbentuk model yang memungkinkan untuk melakukan prediksi terhadap masalah-masalah kesehatan. Salah satu pemanfaatan pembelajaran mesin dalam bidang kesehatan adalah untuk mengklasifikasikan seseorang terdiagnosis diabetes atau tidak berdasarkan gejala-gejala yang diperoleh dari hasil pemeriksaan medis.

Beberapa penelitian sebelumnya yang menggunakan pembelajaran mesin, berhasil memprediksi diabetes berdasarkan berbagai data dan metode. Di antaranya menggunakan *Pima Indian Diabetes Dataset* [2, 3, 4, 5] di mana penelitian yang dilakukan oleh Alehegn memberikan hasil akurasi sebesar 82,80% menggunakan metode *Random Forest* [2]. Dengan menggunakan metode yang sama, Kumari dapat memberikan hasil akurasi sebesar 99,7% [4]. Namun, penelitian ini tidak melalui proses penanganan *imbalanced data* dan tidak melakukan normalisasi pada data. Sementara itu, Daniah dengan metode *Fuzzy Expert System* memberikan hasil akurasi 92,5% [3] dan Iyer yang menggunakan metode *Naive Bayes* berhasil memperoleh akurasi sebesar 79,5% [5]. *Shanghai Health Record Dataset* yang digunakan Zhenga memberikan akurasi 98% [6]. Arun yang menggunakan data dari pusat kesehatan lokal di Thailand membandingkan tiga belas model di mana metode *Random Forest* dinyatakan sebagai model terbaik dengan akurasi 85,5% [7].

Qing Liu dan peneliti lainnya menggunakan empat algoritma pembelajaran mesin terhadap 388.420 responden di Kota Wuhan yang berusia lebih dari 65 tahun di mana metode Regresi Logistik memiliki akurasi yang lebih baik daripada tiga algoritma lainnya yaitu Pohon Keputusan, *Random Forest*, dan *XGBoost*, tetapi perbedaan akurasi yang dihasilkan tidak terlalu jauh [8]. Pada penelitian yang menggunakan *dataset* dari Riyadh, metode *Random Forest* menghasilkan akurasi 88%, yang mana lebih unggul dari pada metode Regresi Logistik dengan akurasi 70,8 % [9]. Alehegn pada konklusi penelitiannya, tidak memberikan kesimpulan terkait penggunaan metode *Random Forest* pada data skala kecil ataupun besar [2].

Penelitian ini melakukan pengembangan terhadap penggunaan model *Random Forest* dengan menerapkan metode *Resampling* untuk menyeimbangkan banyaknya kelas pada data. Selain itu, penelitian ini menggunakan dua skala data, skala kecil dan skala besar. Penggunaan dua skala data bertujuan untuk membandingkan dan menguji kestabilan performa *Random Forest* pada data dengan skala yang berbeda.

1.2 Rumusan Masalah

Permasalahan yang akan dibahas dalam penelitian ini adalah:

1. Bagaimana memprediksi seseorang terdiagnosis penyakit diabetes dengan menerapkan metode *Random Forest* ?
2. Bagaimana pengaruh fitur dalam memprediksi penyakit diabetes dengan menggunakan metode *Random Forest* ?
3. Bagaimana konsistensi akurasi metode *Random Forest* dalam memprediksi pasien terdiagnosis diabetes pada data berskala kecil maupun besar ?
4. Bagaimana perbandingan hasil yang diperoleh dari data yang seimbang dengan data yang tidak seimbang ?

1.3 Tujuan

Tujuan penulisan penelitian ini adalah untuk:

1. Membangun model prediksi berbasis metode *Random Forest* untuk mendeteksi diagnosis penyakit diabetes pada pasien berdasarkan fitur-fitur yang relevan dari data klinis dan riwayat kesehatan.
2. Mengevaluasi pengaruh fitur-fitur atau variabel dalam memprediksi penyakit diabetes dengan menerapkan metode *Random Forest*, serta mengidentifikasi fitur-fitur yang paling berpengaruh dalam model prediksi.
3. Memahami sejauh mana performa model *Random Forest* dapat dipertahankan pada ukuran data yang berbeda dan memberikan wawasan tentang skalabilitas metode tersebut dalam konteks prediksi penyakit diabetes.

4. Membandingkan konsistensi akurasi metode *Random Forest* dalam memprediksi pasien yang terdiagnosis diabetes antara data yang seimbang dan data yang tidak seimbang.

1.4 *State of the Art*

Pada penelitian ini digunakan dua referensi utama yang berhubungan dengan prediksi penyakit diabetes menggunakan metode pembelajaran mesin. Penelitian pertama dilakukan oleh Minyecil Alehegn, Rahul Raghvendra Joshi, dan Preeti Mulay dengan judul *Diabetes Analysis and Prediction Using Random Forest, KNN, Naive Bayes, and J48: An Ensemble Approach*, yang mana penelitian ini bertujuan untuk membuat model prediksi terhadap penyakit diabetes sebagai langkah awal pencegahan. Penelitian ini menggunakan dua buah dataset dengan ukuran data yang berbeda, *Pima Indian Diabetes Dataset* (PIDD) sebagai *dataset* kecil dengan jumlah observasi 768 dan *130 US Hospital Dataset* sebagai *dataset* besar dengan 93.743 observasi. Alehegn membandingkan beberapa metode pembelajaran mesin dalam penelitiannya, dan melakukan penggabungan metode (*Ensemble Approach*) guna mendapat hasil yang lebih baik. Metode gabungan yang dikembangkan memberikan hasil yang lebih baik daripada metode pembelajaran mesin tunggal dengan akurasi sebesar 93,62%. Selain itu, penelitian ini memberikan kesimpulan bahwa metode *Naive Bayes* dan Pohon Keputusan J48 hanya cocok digunakan untuk data berskala besar [2]. Penelitian ini tidak menjelaskan mengenai metode yang digunakan dalam mengatasi *missing value*. Selain itu, penelitian ini juga tidak mengatasi masalah data yang tidak seimbang.

Oleh karena itu, dengan menggunakan data yang sama, skripsi ini melakukan pengembangan berupa menambahkan langkah penyeimbangan banyaknya kelas pada data dengan menerapkan metode *Resampling* dan melakukan standarisasi pada data numerik menggunakan *Z-Score*. Metode pengolahan *missing value* yang berbeda juga dilakukan pada skripsi ini dengan menggunakan nilai median dan modus sebagai nilai pengganti *missing value*.

Penelitian kedua yang berjudul *Predicting the Risk of Incident Type 2 Diabetes Mellitus in Chinese Elderly Using Machine Learning Techniques* yang dilakukan oleh Qing Liu, Miao Zhang, dan beberapa peneliti lainnya pada tahun 2022 di kota Wuhan, China, melakukan identifikasi individu yang memiliki risiko tinggi terkena diabetes. Penelitian ini membangun model pembelajaran mesin untuk memprediksi diabetes mellitus tipe 2 seefektif mungkin terhadap penduduk kota Wuhan yang berusia lebih dari 65 tahun. Qing Liu menggunakan empat metode pembelajaran mesin dalam penelitiannya dan mengevaluasi model tersebut menggunakan nilai *Area Under Curve* (AUC), sensitivitas, spesifisitas, dan akurasi. Menurut penelitian ini, *Extreme Gradient Boosting* (XGBoost) menghasilkan nilai AUC tertinggi yaitu 0.7805 dibandingkan metode Regresi Logistik, Pohon Keputusan, dan *Random Forest*. Lima variabel penting yang paling berpengaruh terhadap model adalah *Fasting Plasma Glucose* (FPG), pendidikan, olahraga, jenis kelamin, dan ukuran lingkar pinggang. Penelitian ini memberikan kesimpulan bahwa XGBoost dapat digunakan untuk melakukan penyaringan individu yang beresiko tinggi sebagai langkah awal pencegahan penyakit diabetes agar dapat ditangani lebih awal [8]. Penggunaan metrik evaluasi, tingkat pengaruh suatu variabel, serta pemilihan variabel bebas atau fitur diterapkan pada skripsi ini. Model dengan pemilihan variabel dibandingkan terhadap model dengan data yang menggunakan seluruh variabel.

1.5 Batasan Masalah

Batasan masalah yang ditemukan dalam penelitian ini antara lain:

1. Data yang digunakan dalam penelitian ini yaitu *Pima Indian Diabetes Dataset* terbatas hanya pada jenis kelamin perempuan yang berusia lebih dari 21 tahun. *Pima Indian Diabetes Dataset* yang digunakan terbatas pada tahun 1965 dan berlokasi di Arizona, Amerika Serikat.

2. Data *130 US-Hospital* yang digunakan terbatas hanya pada tahun 1999-2008 dan berlokasi di Amerika Serikat.

1.6 Sistematika Pembahasan

Sistematika pembahasan pada penelitian ini terdiri dari 5 bab, yaitu:

BAB 1 : Pendahuluan

Pada bab ini akan dibahas latar belakang, rumusan masalah, tujuan penulisan, dan sistematika pembahasan.

BAB 2 : Landasan Teori

Pada bab ini akan membahas teori pendukung seperti Diabetes, *Bootstrap*, *Bootstrap Aggregating*, Pohon Keputusan, *Random Forest*, Matriks Konfusi, Standarisasi Data, dan *Imbalanced Data*.

BAB 3 : Metode Penelitian

Bab ini berisi mengenai metode penelitian yang dilakukan berupa pengumpulan data, eksplorasi data, pemrosesan data, pemodelan data dengan *Random Forest*, hingga evaluasi data.

BAB 4 : Hasil dan Analisis

Bab ini menjelaskan hasil pemodelan yang dilakukan menggunakan *Random Forest* dalam bentuk Matriks Konfusi dan beberapa pengukuran seperti akurasi, sensitivitas, presisi, spesifisitas, dan *F1-Score*.

BAB 5 : Kesimpulan

Bab ini memberikan kesimpulam akhir penelitian ini serta rencana pengembangan penelitian yang mungkin dapat ditambahkan pada penelitian selanjutnya.