 ISSN:2302-8149
e-ISSN:2540-9719

Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi *Random Forest*

[1]Widya Apriliah*, [2] Ilham Kurniawan, [3] Muhamad Baydhowi, [4]Tri Haryati
[1,2,4]Sistem Informasi Kampus Kabupaten Karawang, Fakultas Teknik dan Informatika,
Universitas Bina Sarana Informatika, Jl. Banten No.1 Karangpawitan Karawang Barat
[3]Sistem Informasi, Fakultas Informatika, Universitas Bina Insani,
Jl. Siliwangi No. 6 Rawa Panjang, Bekasi Barat.
*e-mail: widyaaprilia.wyr@gmail.com

(***received*:** 16 November 2020, ***revised*:** 8 Desember 2020, ***accepted*:** 14 Desember 2020)

Abstrak

Diabetes adalah salah satu penyakit kronis yang mengancam jiwa dengan pertumbuhan tercepat yang telah mempengaruhi 422 juta orang di seluruh dunia menurut laporan Organisasi Kesehatan Dunia (WHO), pada tahun 2018. Diabetes dianggap sebagai salah satu penyakit paling mematikan dan kronis yang menyebabkan peningkatan gula darah. Banyak komplikasi terjadi jika diabetes tetap tidak diobati dan tidak teridentifikasi. Namun, peningkatan pendekatan machine learning memecahkan masalah kritis ini. Tujuan dari penelitian ini adalah merancang model yang dapat memprakirakan kemungkinan terjadinya diabetes pada pasien dengan ketelitian yang maksimal. Klasifikasi adalah teknik data mining yang menetapkan kategori pada kumpulan data untuk membantu dalam memprediksi dan analisis yang lebih akurat. Oleh karena itu tiga algoritma klasifikasi machine learning yaitu Suport Vector Machine, Naive Bayes dan Random Forest digunakan dalam percobaan ini untuk mendeteksi diabetes secara dini. Eksperimen dilakukan menggunakan dataset Diabetes Hospital in Sylhet, Bangladesh yang bersumber dari UCI repository. Performa ketiga algoritma dievaluasi pada berbagai ukuran seperti Precision, Accuracy, F-Measure, dan Recall. Akurasi diukur melalui instance yang diklasifikasikan dengan benar dan salah. Hasil yang diperoleh menunjukkan Random Forest mengungguli dengan nilai akurasi tertinggi 97,88% dibandingkan algoritma lain. Hasil ini diverifikasi menggunakan kurva Receiver Operating Characteristic (ROC) secara tepat dan sistematis.

Kata Kunci: diabetes, naive bayes, random forest, akurasi, support vector machine, machine learning

Abstract

Diabetes is one of the fastest growing, life-threatening chronic diseases affecting 422 million people worldwide, according to a report by the World Health Organization (WHO) in 2018. Diabetes is considered to be one of the most deadly and chronic diseases that cause elevated blood sugar. Many complications occur if diabetes remains untreated and unidentified. However, an improved machine learning approach solves this critical problem. The aim of this study is to design a model that can predict the likelihood of diabetes occurr in patients with maximum accuracy. Therefore, three machine learning classification algorithms, namely Suport Vector Machine, Naive Bayes and Random Forest, were used in this experiment to detect diabetes early. Experiments were conducted using the Diabetes Hospital in Sylhet, Bangladesh dataset sourced from the UCI repository. The performance of the three algorithms is evaluated on various measures such as Precision, Accuracy, F-Measure, and Recall. Accuracy is measured through correctly and incorrectly classified instances. The results obtained showed that Random Forest outperformed with the highest accuracy value of 97.88% compared to other algorithms. These results are verified using the Receiver Operating Characteristic (ROC) curve accurately and systematically.

Keywords: *diabetes, naive bayes, random forest, accuracy, , machine learning, support vector machine*

 ISSN:2302-8149
e-ISSN:2540-9719

1 Pendahuluan

Strategi klasifikasi digunakan secara luas di bidang medis untuk mengklasifikasikan data ke dalam kelas yang berbeda menurut beberapa kendala yang secara komparatif merupakan pengklasifikasi individu [1]. Diabetes mellitus (DM), menurut definisi World Health Organization (WHO), adalah penyakit degeneratif kronis yang disebabkan oleh produksi insulin yang tidak mencukupi di pankreas atau oleh ketidakmampuan tubuh untuk secara efektif menggunakan insulin yang diproduksi, mengambil hyperglycemia (peningkatan glukosa darah) sebagai indikator utama [2]. Karena gejalanya yang mirip dengan kondisi sakit biasa, banyak orang yang tidak menyadari bahwa mereka mengidap penyakit diabetes dan bahkan sudah mengarah pada komplikasi. Untuk memastikan bahwa seseorang apakah mengidap diabetes atau tidak maka perlu diagnosis dokter melalui cek darah. Bagi orang awam, setidaknya harus mengenal beberapa gejala yang biasanya mengiringi penyakit diabetes ini seperti, sering buang air kecil, mudah merasa haus, mudah merasa lapar, turunnya berat badan secara drastis, kulit kering, penyembuhan luka relatif lama, dan adanya gangguan penglihatan [3]. Hampir setengah dari semua penderita diabetes memiliki faktor keturunan, yang merupakan salah satu ciri terpenting DM [4].

Penyakit diabetes tidak bisa disembuhkan sepenuhnya tapi bisa dikontrol. Diabetes tipe 1 dapat dikontrol dengan mengonsumsi insulin. Ada berbagai jenis insulin seperti insulin kerja cepat, insulin kerja pendek, insulin kerja menengah, dan insulin kerja panjang tergantung pada seberapa cepat mereka merespons kerja dan berapa lama efeknya bertahan. Diabetes tipe 2 dapat dikontrol dengan diet seimbang, pengobatan oral, dan olahraga teratur [5] Diabetes mellitus ada dalam tiga bentuk [6]: (1) Diabetes Mellitus Tipe-1 ditandai dengan produksi insulin pankreas kurang dari yang dibutuhkan oleh tubuh, suatu kondisi yang juga disebut "insulin-subordinate diabetes mellitus" (IDDM). Orang yang menderita DM tipe-1 memerlukan dosis insulin eksternal untuk mengganti lebih sedikit insulin yang diproduksi oleh pankreas. (2) Diabetes Mellitus Tipe-2 ditandai dengan tubuh melawan insulin karena sel-sel tubuh bereaksi berbeda terhadap insulin dari biasanya. Hal ini pada akhirnya dapat menyebabkan tidak adanya insulin dalam tubuh. Ini juga disebut "diabetes mellitus non-insulin subordinat" (NIDDM)". Jenis diabetes ini umumnya ditemukan pada mereka yang menjalani gaya hidup tidak aktif. (3) Diabetes gestasional adalah struktur prinsip ketiga yang diamati selama kehamilan.

Dalam kegiatan prediksi diagnostik, data mining dan text mining telah dibuktikan sebagai metode yang menjanjikan. Metode ini disusun oleh serangkaian alat dan teknik yang mampu mengeksplorasi kumpulan data, dan membantu dalam penemuan pengetahuan [7]. Machine learning dianggap sebagai salah satu fitur kecerdasan buatan terpenting yang mendukung pengembangan sistem komputer yang memiliki kemampuan untuk memperoleh pengetahuan dari pengalaman masa lalu tanpa perlu pemrograman untuk setiap kasus. Machine learning dianggap sebagai kebutuhan yang mendesak dari situasi saat ini untuk menghilangkan upaya manusia dengan mendukung otomatisasi dengan kekurangan minimum. Metode yang ada untuk deteksi diabetes adalah dengan menggunakan tes laboratorium seperti glukosa darah dan toleransi glukosa oral. Namun, metode ini memakan waktu lama [8]. Penelitian ini berfokus pada membangun model prediksi menggunakan algoritma machine learning dan teknik data mining untuk prediksi kemungkinan diabetes.

2 Tinjauan Literatur

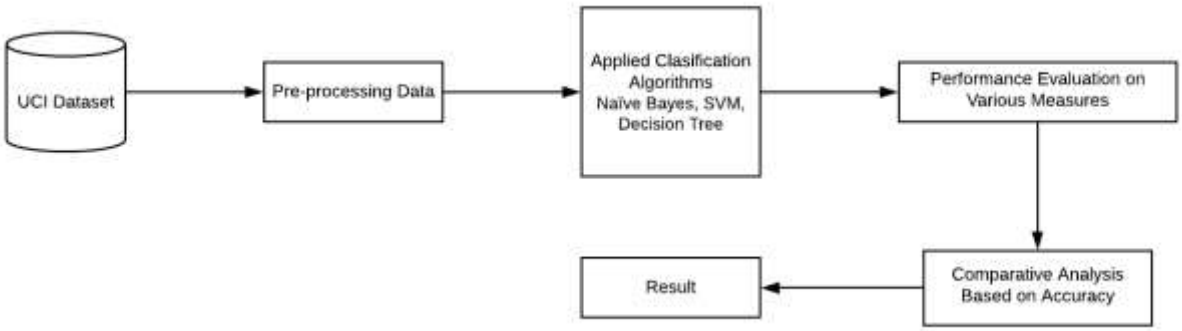
Penelitian Lukmanto dan Irwansyah [9] mengenai deteksi dini diabetes mellitus (dm) menggunakan model fuzzy hierarchical pada tahun 2013 jumlah penderita Diabetes Mellitus (DM) di dunia mencapai 382 juta. Diperkirakan prevalensinya akan meningkat 55% pada tahun 2035. Sebagai bentuk upaya kami untuk berkontribusi dalam pencegahan fenomena tersebut kami mengusulkan suatu aplikasi kecerdasan komputasi dengan menggunakan model hirarki fuzzy yang memiliki kemampuan untuk melakukan deteksi dini terhadap DM. Untuk mensukseskan metode pengusulan kami, kami bekerjasama dengan salah satu laboratorium RS Jaktim Indonesia sebagai fasilitator data-data yang kami butuhkan selama penelitian dan melakukan wawancara dengan dua dokter medis di rumah sakit yang sama. Arsitektur metode yang kami usulkan ini dirancang berdasarkan bagaimana

dokter menyimpulkan terkait indikasi seseorang berpotensi terkena DM, yang modelnya telah disesuaikan dengan data yang kami peroleh dari pihak berwenang di laboratorium. Sebagai bentuk evaluasi yang kami lakukan, kami melakukan perbandingan data yang telah kami peroleh dari metode kami dengan keputusan dokter yang dilengkapi dengan data dari laboratorium, dan hasilnya 87,46% dari 311 data yang relevan adalah setara dengan pernyataan dokter medis. Dalam menginterpretasikan kesimpulan yang kami dapatkan dengan rumah sakit yang bekerjasama dengan kami, hasil penelitian menunjukkan bahwa metode yang kami usulkan telah memenuhi kebutuhan akan efektivitas dan efisiensi dalam melakukan deteksi dini terhadap DM dan dapat membantu masyarakat dalam mengetahui potensi DM sejak dini. Penelitian Fiarni, Sipayung, dan Maemunah [10] mengenai analisis kinerja teknik klasifikasi data mining untuk memprediksi diabetes. Dalam penelitian ini faktor risiko diabetes dipersempit menjadi tujuh fitur, yaitu Usia, Jenis Kelamin, IMT, Riwayat Diabetes Keluarga, Tekanan Darah, Lama Menderita Diabetes dan Kadar Glukosa Darah. Secara keseluruhan akurasi model yang diusulkan adalah 68% sehingga dapat digunakan sebagai metode alternatif untuk membantu memprediksi penyakit komplikasi diabetes secara dini. Secara keseluruhan akurasi model yang diusulkan adalah 68% sehingga dapat digunakan sebagai metode alternatif untuk membantu memprediksi penyakit komplikasi diabetes secara dini. Penelitian Kavakiotis, et al [11] mengenai machine learning dan metode data mining dalam penelitian diabetes. Penelitian ekstensif dalam semua aspek diabetes (diagnosis, etiopatofisiologi, terapi, dll.) Telah menghasilkan sejumlah besar data. Tujuan dari penelitian ini adalah untuk melakukan tinjauan sistematis terhadap aplikasi machine learning, teknik dan data mining di bidang penelitian diabetes berkenaan dengan prediksi dan diagnosis, komplikasi diabetik, latar belakang genetik dan lingkungan , dan perawatan dan manajemen kesehatan dengan kategori pertama yang tampaknya paling populer. Berbagai macam algoritma pembelajaran mesin digunakan. Secara umum, 85% dari yang digunakan ditandai dengan pendekatan supervised learning dan 15% oleh unsupervised learning, dan lebih khusus lagi, aturan asosiasi. Support vector machine (SVM) muncul sebagai algoritma yang paling sukses dan banyak digunakan. Mengenai jenis data, kumpulan data klinis yang paling banyak digunakan. Penelitian Perveen, et al [12] mengenai analisis kinerja teknik klasifikasi data mining untuk memprediksi diabetes. Baru-baru ini upaya ekstensif sedang dilakukan untuk meningkatkan akurasi sistem tersebut menggunakan ensemble classification. Penelitian ini mengikuti teknik ensemble adaboost dan bagging menggunakan pohon keputusan J48 (c4.5) sebagai learning base beserta teknik data mining J48 mandiri untuk mengelompokkan pasien diabetes mellitus dengan menggunakan faktor risiko diabetes. Klasifikasi ini dilakukan di tiga kelompok dewasa ordinal yang berbeda di Canadian Primary Care Sentinel Surveillance network. Hasil percobaan menunjukkan bahwa secara keseluruhan kinerja metode ansambel adaboost lebih baik daripada bagging serta pohon keputusan J48 yang berdiri sendiri. Penelitian Islam, et al [13] mengenai prediksi kemungkinan diabetes pada tahap awal menggunakan teknik data mining. Dalam penelitian ini, kami telah menggunakan kumpulan data 520 kasus, yang dikumpulkan menggunakan kuesioner langsung dari pasien Rumah Sakit Diabetes Sylhet di Sylhet, Bangladesh. Kami telah menganalisis dataset dengan algoritma naive bayes, algoritma regresi logistik, dan algoritma random forest dan setelah menerapkan 10-fold cross validation dan percentage split, ditemukan bahwa algoritma random forest memiliki akurasi terbaik pada dataset ini. Sedangkan pada penelitian kali ini akan mengkaji performa dari ketiga algoritma dievaluasi pada berbagai ukuran seperti precision, accuracy, f-measure, dan recall. Akurasi diukur melalui instance yang diklasifikasikan dengan benar dan salah.

3 Metode Penelitian

Prosedur yang diusulkan disajikan pada gambar-1 di bawah ini dalam bentuk diagram model. Gambar tersebut menunjukkan alur penelitian yang dilakukan dalam membangun model.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 1. Diagram Model yang Diusulkan

Dari kerangka model penelitian yang telah diusulkan pada Gambar 1. Diagram model yang diusulkan, dapat dijelaskan sebagai berikut:

a) UCI Dataset adalah dataset yang digunakan dalam penelitian ini, merupakan dataset dari UCI repository yaitu dataset Diabetes Hospital in Sylhet, Bangladesh dengan jumlah data sebanyak 520 data, 17 atribut dan 1 kelas.
b) Pre-processing data, teknik pre-processing data yang digunakan pada penlitian ini yaitu Resample, gunanya untuk menghasilkan subsampel acak dari kumpulan data menggunakan pengambilan sampel dengan penggantian atau tanpa penggantian.
c) Applied Clasification Algorithms, algoritma klasifikasi yang digunakan dalam penelitian ini menggunakan algoritma klasifikasi naive bayes, support vector machine dan random forest.
d) Performance evaluation on various measures, Eksperimen dilakukan menggunakan 10 fold cross validation. Pengukuran Accuracy, F-Measure, Recall, Precision dan ROC (Receiver Operating Curve), digunakan untuk klasifikasi penelitian ini
e) Comparative analysis based on accuracy, pada penitian ini dilakukan percobaan menggunakan 3algoritma klasifikasi yang berbeda untuk mengetahui algoritma klasifikasi mana yang mempunyai nilai akurasi tertinggi.
f) Result, hasil penelitian menunjukan algoritma klasifikasi random forest mempunyai nilai akurasi tertinggi yaitu 97,88%.

3.1. Deskripsi Singkat Algoritma yang Digunakan:

Support Vector Machine (SVM)

Algoritma Support Vector Machine (SVM) adalah algoritma berbasis diskriminasi yang bertujuan untuk menemukan batas pemisahan optimal yang disebut hyperplane untuk membedakan kelas dari satu sama lain. Sampel yang paling dekat dengan hyperplanes ini disebut vektor dukungan, dan perbedaan tersebut dinyatakan sebagai jumlah bobot dari subset sampel yang membatasi kerumitan masalah [14]. Untuk memisahkan dua klasifikasi melalui hyperplane yang optimal menggunakan persamaan (1) dan (2) berikut:

$$f(x) = w.x + b \quad (1)$$

atau

$$f(x) = \sum_{i=1}^{n} a_i y_i K(x, x_i) + b \quad (2)$$

Harus berada diantara dua kelas sampel hyperplane pemisah yang optimal

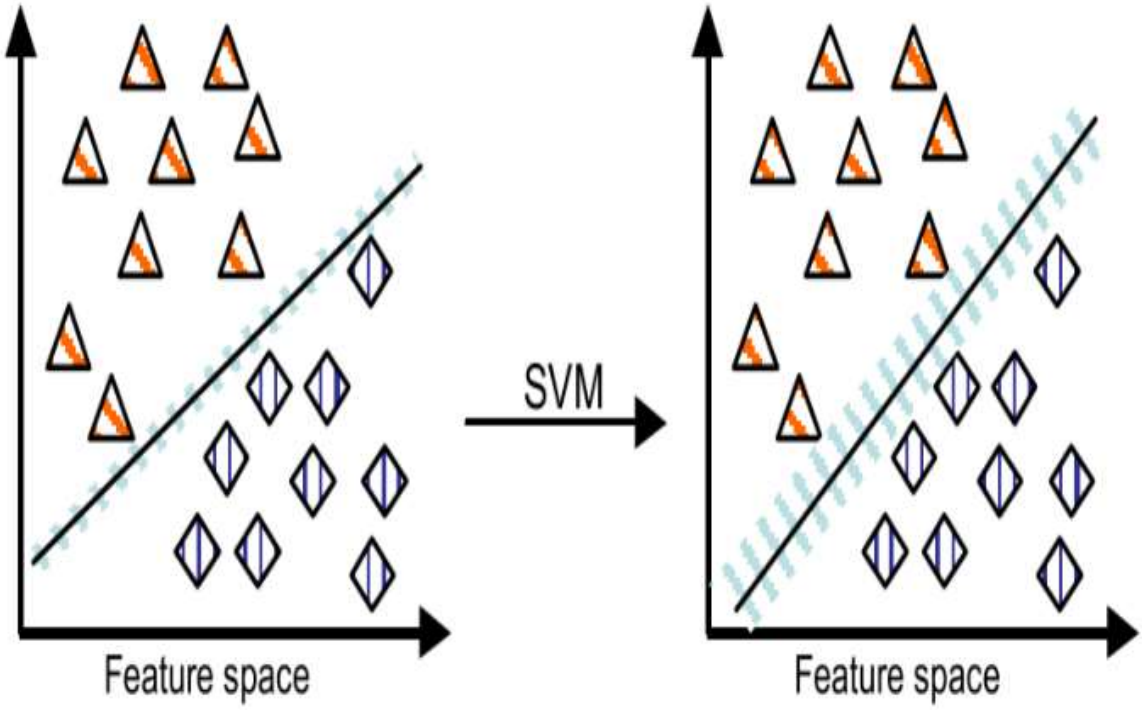
$$y_i(wx_i + b) \geq 1 \ (i = 1, 2, \ldots, n)$$

dimana :

w : beban vektor (garis tegak lurus)
x : titik data masukan SVM
b : bias
a_i : nilai bobot setiap titik data
$K(x, x_i)$: fungsi karnel

ISSN:2302-8149
e-ISSN:2540-9719

SVM bekerja berdasarkan prinsip dasar, yaitu menyisipkan hyper-plane antara kelas-kelas dan mengarahkannya sedemikian rupa sehingga menjaganya pada jarak maksimum dari titik data terdekat seperti yang terlihat pada Gambar 1. Data ini menunjukan, data yang muncul paling dekat dengan hyper-plane, dikenal sebagai Support Vectors [15].



Gambar 1. Klasifikasi Linear Margin Maksimum [15]

Kinerja yang dievaluasi dari algoritma SVM untuk prediksi diabetes [16], [17] menggunakan Confusion Matrix adalah sebagaimana Tabel 1 berikut:

Tabel 1. Confusion Matrix SVM

True Positive	False Positive	False Negative	True Negative
304	18	9	189

Naive Bayes

Metode pengklasifikasi naïve bayes adalah pengklasifikasi probabilistik sederhana yang menetapkan setiap objek ke kelas dengan asumsi independensi yang kuat di antara variabel. Kemudian teorema Bayes dijelaskan sebagaimana persamaan (3) berikut:

$$P(c\,|F) = \frac{P(F\,|c)}{P(F)} \qquad (3)$$

di mana P (c) mewakili probabilitas sebelumnya, P (F) adalah probabilitas marginal, P (c | F) adalah probabilitas posterior kelas gabungan, P (F | c) menunjukkan probabilitas bersyarat, masing-masing [18]. Klasifikasi naïve bayes perlu untuk menyelesaikan satu set estimasi kepadatan satu dimensi. Asumsi umum adalah bahwa dalam setiap kelas, nilai dari setiap atribut berasal dari

 ISSN:2302-8149
e-ISSN:2540-9719

distribusi normal. Seseorang dapat merepresentasikan distribusi tersebut dalam hal mean dan standard deviation dan dapat memperkirakan secara efisien mean dan standard deviation menggunakan metode Maximum Likelihood Estimation (MLE). Parameterisasi model ini efisien secara komputasi dan juga dapat dengan mudah digunakan dalam kasus terdistribusi. Namun demikian, asumsi bahwa distribusi atribut obeya Gaussian mungkin tidak berlaku untuk beberapa domain [19]. Kinerja algoritma Naive Bayes yang dievaluasi menggunakan Confusion Matrix adalah sebagaimana Tabel 2 berikut:

Tabel 2. Confusion Matrix Naive Bayes

True Positive	False Positive	False Negative	True Negative
288	8	34	190

Random Forest

Pendekatan radom forest yang diusulkan oleh Breiman [20] adalah algoritma pembelajaran mesin dengan banyak pohon keputusan. Random forest adalah kombinasi dari metode Bagging [21] dan Random Sub spaces [22]. Metode ini telah membuktikan keberhasilannya dalam masalah regresi dan klasifikasi dalam beberapa tahun terakhir dan merupakan salah satu algoritma machine learning terbaik yang digunakan di berbagai bidang [23][24][25][26][27][28]. Performa yang dievaluasi dari teknik Random Forest menggunakan Confusion Matrix adalah sebagaimana Tabel 3 berikut:

Tabel 3. Confusion Matrix Random Forest

True Positive	False Positive	False Negative	True Negative
315	4	7	194

4 Hasil Dan Pembahasan

Penelitian ini menggunakan aplikasi Weka versi 3.8.4 dengan menggunakan hardware Laptop processor Intel Core i5 dengan RAM 8 GB dan hardisk 500GB.

4.1. Data Set

Tools yang digunakan dalam penelitian ini adalah WEKA [29] digunakan untuk melakukan percobaan. WEKA adalah perangkat lunak yang dirancang di negara Selandia Baru oleh University of Waikato, yang mencakup kumpulan berbagai metode pembelajaran mesin untuk klasifikasi data, pengelompokan, regresi, visualisasi, dll. Salah satu keuntungan terbesar menggunakan WEKA adalah dapat dipersonalisasi sesuai dengan kebutuhan. Tujuan utama dari penelitian ini adalah untuk mencari tingkat akurasi tertinggi pada algoritma klasifikasi yang diusulkan untuk memprediksi diabetes menggunakan tools WEKA dengan menggunakan UCI dataset. Tabel 4 menunjukkan deskripsi singkat dari dataset.

Tabel 4. Deskripsi Dataset

Dataset	Jumlah Atribut	Jumlah Data
Diabetes Hospital in Sylhet, Bangladesh)	17	520

Metodologi yang diusulkan dievaluasi pada dataset diabetes yaitu Diabetes Hospital in Sylhet, Bangladesh [13], yang diambil dari UCI Repository. Dataset ini terdiri dari rincian medis 520 kasus, dengan deskrpsi seperti pada Tabel 5.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 5. Deskripsi Dataset

Attributes	Values
Age	1.20–35, 2.36–45, 3.46–55,4.56–65, 6.above 65
Sex	1.Male, 2.Female 1.Yes,
Polyuria	1.Yes, 2.No. 1.Yes,
Polydipsia	1.Yes, 2.No. 1.Yes,
Sudden Weigh Loss	1.Yes, 2.No. 1.Yes,
Weakness	1.Yes, 2.No. 1.Yes,
polyphagia	1.Yes, 2.No. 1.Yes,
Genital thrush	1.Yes, 2.No. 1.Yes,
Visual blurring	1.Yes, 2.No. 1.Yes,
Itching	1.Yes, 2.No. 1.Yes,
Irritability	1.Yes, 2.No. 1.Yes,
Delayed healing	1.Yes, 2.No. 1.Yes,
Partial paresis	1.Yes, 2.No. 1.Yes,
Muscle stiffness	1.Yes, 2.No. 1.Yes,
Alopecia	1.Yes, 2.No. 1.Yes,
Obesity	1.Yes, 2.No. 1.Yes,
Class	1.Positive, 2.Negative.

4.2. Pengukuran Akurasi

Algoritma naive bayes, SVM dan random forest digunakan dalam penelitian ini. Eksperimen dilakukan menggunakan teknik 10-fold cross-validation. Accuracy, f-measure, recall, precision and ROC (Receiver Operating Curve) measures digunakan untuk klasifikasi penelitian ini. Tabel-5 menjelaskan deskripsi dataset, Tabel 6 menjelaskan ukuran akurasi di bawah ini:

Tabel 6. Pengukuran Akurasi

Pengukuran	Definisi	Formula
Accuracy (A)	Akurasi menentukan keakuratan algoritme dalam memprediksi instance	A=(TP+TN) / (Jumlah total sampel)
Precision (P)	Classifier, correctness/accuracy diukur dengan Precision	P = TP / (TP+ FP**)**
Recall (R)	Untuk mengukur pengklasifikasi completeness atau sensitivity, menggunakan Recall	R =TP / (TP+FN)
F-Measure (F)	F-Measure adalah rata-rata dari precision dan recall.	F=2*(P*R) / (P+R)
ROC	ROC (Receiver Operating Curve) digunakan untuk membandingkan kegunaan pengujian	

Tabel 7. Kinerja Perbandingan Algoritma Klasifikasi pada Berbagai Ukuran

Classification Algorithm	Precision	Recall	F-Measure	Accuracy %	ROC
SVM	0,949	0,948	0,948	94,80	0,949
Naive Bayes	0,925	0,919	0,920	91,92	0,964
Random Forest	**0,979**	**0,979**	**0,979**	**97,88**	**0,998**

Tabel 7 menunjukkan nilai kinerja yang berbeda dari semua algoritma klasifikasi yang dihitung pada berbagai ukuran. Dari Tabel 7 dianalisis bahwa algoritma klasifikasi random forest menunjukan nilai akurasi dan nilai ROC yang paling tinggi yaitu sebesar 97,88% dan nilai ROC sebesar 0,998 mengungguli algoritma klasifikasi lainnya. ROC merupakan representasi dari algoritma klasifikasi yang dibangun untuk memprediksi diabates, semakin mendekati angka 1 maka semakin baik pula algoritma klasifikasi yang dibangun, Jadi algoritma klasifikasi random forest dapat memprediksi kemungkinan diabetes dengan lebih akurat dibandingkan dengan algoritma klasifikasi lainnya.

5 Kesimpulan

Salah satu masalah medis yang penting adalah deteksi diabetes pada tahap awal. Studi saat ini mengatakan bahwa deteksi diabetes pada tahap awal dapat memainkan peran penting dalam pengobatan. Langkah-langkah kesadaran sederhana seperti diet rendah gula, aktivitas fisik teratur, dan gaya hidup sehat dapat menghindari obesitas. Karena metode, teknik, dan alat data mining menjadi lebih menjanjikan untuk memprediksi diabetes dan pada akhirnya mengurangi jumlah pasien dan mengurangi biaya perawatan. Kontribusi utaman penelitian ini adalah untuk mengetahui algoritma klasifikasi terbaik untuk prediksi resiko diabetes. Percobaan dilakukan pada dataset Diabetes Hospital in Sylhet, Bangladesh yang diambil dari UCI repository. Hasil percobaan menentukan kecukupan sistem yang dirancang dengan akurasi yang dicapai sebesar 97,88%. Kami menemukan bahwa algoritma Random Forest telah bekerja dengan akurasi terbaik. Kedepannya, sistem yang dirancang dengan algoritma klasifikasi machine learning dapat digunakan untuk memprediksi atau mendiagnosis penyakit lain. Penelitian dapat diperpanjang dan ditingkatkan untuk otomatisasi analisis diabetes termasuk beberapa algoritma machine learning lainnya.

Referensi

[1] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[2] A. Viloria, Y. Herazo-Beltran, D. Cabrera, and O. B. Pineda, "Diabetes Diagnostic Prediction Using Vector Support Machines," *Procedia Comput. Sci.*, vol. 170, pp. 376–381, 2020, doi: 10.1016/j.procs.2020.03.065.

[3] S. Hadijah, "Gejala Diabetes, Ciri-Ciri Diabetes, Penyebab Diabetes, Serta Penanganan Penyakit Diabetes yang Perlu Kamu Tahu," *10 November*, 2017. https://www.cermati.com/artikel/gejala-diabetes-ciri-ciri-diabetes-penyebab-diabetes-serta-penanganan-penyakit-diabetes-yang-perlu-kamu-tahu (accessed Dec. 10, 2020).

[4] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics Med. Unlocked*, vol. 10, pp. 100–107, 2018, doi: 10.1016/j.imu.2017.12.006.

[5] D. J. Reddy *et al.*, "Materials Today : Proceedings Predictive machine learning model for early detection and analysis of diabetes," *Mater. Today Proc.*, 2020, doi: 10.1016/j.matpr.2020.09.522.

[6] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[7] L. B. Moreira and A. A. Namen, "A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia," *Comput. Methods Programs Biomed.*, vol. 165, pp. 139–149, 2018, doi: 10.1016/j.cmpb.2018.08.016.

[8] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[9] R. B. Lukmanto and E. Irwansyah, "The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model," *Procedia Comput. Sci.*, vol. 59, no. Iccsci, pp. 312–319, 2015, doi: 10.1016/j.procs.2015.07.571.

[10] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes

complication disease using data mining algorithm," *Procedia Comput. Sci.*, vol. 161, pp. 449–457, 2019, doi: 10.1016/j.procs.2019.11.144.

[11] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[12] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," *Procedia Comput. Sci.*, vol. 82, no. March, pp. 115–121, 2016, doi: 10.1016/j.procs.2016.04.016.

[13] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," *Comput. Vis. Mach. Intell. Med. Image Anal.*, pp. 113–125, 2020, doi: doi.org/10.1007/978-981-13-8798-2_12.

[14] S. Salcedo-Sanz, J. L. Rojo-Álvarez, M. Martínez-Ramón, and G. Camps-Valls, "Support vector machines in engineering: An overview," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 4, no. 3, pp. 234–267, 2014, doi: 10.1002/widm.1125.

[15] M. Sewak, P. Vaidya, C.-C. Chan, and Zhong-Hui Duan, "SVM Approach to Breast Cancer Classification," *Second Int. Multi-Symposiums Comput. Comput. Sci. (IMSCCS 2007)*, pp. 32–37, 2007, doi: 10.1109/IMSCCS.2007.46.

[16] H. Kucuk and I. Eminoglu, "Classification of ALS disease using support vector machines," 2015 23nd Signal Processing and Communications Application Conference (SIU), Malatya, vol. 3, no. 2, pp. 1664–1667, 2015, doi: 10.1109/siu.2015.7130171.

[17] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *Med. Informatics Decis. Mak.*, pp. 1–7, 2010.

[18] H. Zhang, C. T. Liu, J. Mao, C. Shen, R. L. Xie, and B. Mu, "Development of novel in silico prediction model for drug-induced ototoxicity by using naïve Bayes classifier approach," *Toxicol. Vitr.*, vol. 65, no. September 2019, 2020, doi: 10.1016/j.tiv.2020.104812.

[19] A. Khajenezhad, M. A. Bashiri, and H. Beigy, "A distributed density estimation algorithm and its application to naive Bayes classification," *Appl. Soft Comput.*, p. 106837, 2020, doi: 10.1016/j.asoc.2020.106837.

[20] L. Breiman, "Random forests," *Machine Learning*, vol 45 no. 1 pp. 5–32, 2001.

[21] L. Breiman, "Bagging predictors," *Machine Learning.*, vol. 24, no. 2, pp. 123–140, 1996

[22] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," vol. 20, no. 8, pp. 832–844, 1998.

[23] H. R. Pourghasemi *et al.*, *Spatial modeling, risk mapping, change detection, and outbreak trend analysis of coronavirus (COVID-19) in Iran (days between February 19 and June 14, 2020)*, vol. 98, June. International Society for Infectious Diseases, 2020.

[24] M. Jeung, S. Baek, J. Beom, K. H. Cho, Y. Her, and K. Yoon, "Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments," *J. Hydrol.*, vol. 575, May, pp. 1099–1110, 2019, doi: 10.1016/j.jhydrol.2019.05.079.

[25] E. Izquierdo-Verdiguier and R. Zurita-Milla, "An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 88, no. October 2019, p. 102051, 2020, doi: 10.1016/j.jag.2020.102051.

[26] T. Hengl, M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," *PeerJ*, vol. 2018, no. 8, 2018, doi: 10.7717/peerj.5518.

[27] S. Oliveira, F. Oehler, J. San-Miguel-Ayanz, A. Camia, and J. M. C. Pereira, "Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest," *For. Ecol. Manage.*, vol. 275, pp. 117–129, 2012, doi: 10.1016/j.foreco.2012.03.003.

[28] P. Zahedi, S. Parvandeh, A. Asgharpour, B. S. McLaury, S. A. Shirazi, and B. A. McKinney, "Random forest regression prediction of solid particle Erosion in elbows," *Powder Technol.*, vol. 338, pp. 983–992, 2018, doi: 10.1016/j.powtec.2018.07.055.

[29] R. Arora and S. Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA," *Int. J. Comput. Appl.*, vol. 54, no. 13, pp. 21–25, 2012.