



IMPLEMENTASI DATA MINING MENGGUNAKAN METODE RANDOM FOREST DAN SUPPORT VECTOR MACHINE DALAM KLASIFIKASI PENYAKIT DIABETES

¹⁾Anggita Ghozali ²⁾Hasih Pratiwi ³⁾Sri Sulistijowati Handajani
Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Sebelas Maret

¹⁾Ghozalianggit@student.uns.ac.id

Abstract

Received :
29/12/2022

Accepted :
27/2/2023

Published :
31/07/2023

Diabetes is a deadly disease. This disease is included in the top ten causes of death globally. In 2045, diabetes patients will increase by 12.2%, with the number of people with diabetes amounting to 642.7 million. In addition, people with diabetes in Indonesia will rank fifth in the world. Therefore, people need to be aware of this disease. From existing diabetes data, it can be classified using data mining. The method used is Random Forest and Support Vector Machine. The data used in this study is secondary data from Kaggle, which comes from Sylhet Hospital, Bangladesh, with a total of 520 data and 17 variables, of which 1 variable is a label. The data is divided into three conditions, namely 80%:20%, 75%:25%, and 70%:30%. Data was balanced using Synthetic Minority Oversampling (SMOTE). From the results of the study, it was found that the Random Forest algorithm with a data split of 80%:20% got the best results with an accuracy of 0.98, a precision of 0.96, a recall of 1, a specificity of 0.95, and an F1-score of 0.98. The three most influential variables in the classification of diabetes are polyuria, polydipsia, and gender.

Keywords: diabetes, data mining, Random Forest, SMOTE, Support Vector Machine

Abstrak

Diabetes menjadi salah satu penyakit yang mematikan. Penyakit ini masuk dalam sepuluh besar penyebab kematian secara global. Pada tahun 2045, pasien diabetes akan meningkat sebanyak 12,2%, dengan jumlah penderita penyakit diabetes sebesar 642,7 juta. Selain itu, penderita penyakit diabetes di Indonesia akan menempati ranking kelima di dunia. Oleh karena itu, masyarakat perlu waspada terhadap penyakit ini. Dari data diabetes yang ada, dapat diklasifikasikan menggunakan *data mining*. Metode yang digunakan yaitu *Random Forest* dan *Support Vector Machine*. Data yang digunakan pada penelitian ini merupakan data sekunder dari Kaggle, yang berasal dari Rumah Sakit Sylhet Bangladesh, dengan jumlah data sebesar 520 dan 17 variabel, dimana 1 variabel merupakan label. Data dibagi dengan tiga kondisi, yaitu 80%:20%, 75%:25%, dan 70%:30%. Data diseimbangkan menggunakan *Synthetic Minority Oversampling* (SMOTE). Dari hasil penelitian didapatkan bahwa algoritma *Random Forest* dengan *split* data 80%:20% mendapatkan hasil terbaik dengan akurasi yang didapatkan sebesar 0,98, presisi sebesar 0,96, *recall* sebesar 1, *specificity* sebesar 0,95, dan *F1-score* sebesar 0,98. Tiga variabel yang paling berpengaruh dalam klasifikasi penyakit diabetes ini secara berturut turut adalah *polyuria*, *polydipsia*, dan jenis kelamin.

Kata Kunci: diabetes, data mining, *Random Forest*, SMOTE, *Support Vector Machine*

1. Pendahuluan

Diabetes adalah salah satu penyakit yang sangat serius dan menjadi masalah kesehatan global. Menurut WHO (2020), penyakit diabetes ini masuk dalam sepuluh besar penyakit yang menyebabkan kematian di dunia. Penyakit diabetes merupakan sebuah penyakit yang disebabkan karena insulin yang dihasilkan pankreas tidak mencukupi (WHO, 2022). Hal tersebut menyebabkan terjadinya hiperglikemia pada

tubuh manusia. Selain itu, jika tidak ada penanganan secara lebih cepat, maka diabetes akan menyebabkan komplikasi (Utomo dkk., 2020). Penyakit ini akan merusak jantung, saraf, ginjal, dan yang lainnya.

Ada beberapa keluhan klasik yang mencirikan pasien terkena penyakit diabetes. Contohnya yaitu *polyuria*, yaitu kondisi dimana tubuh manusia mengeluarkan urine secara berlebihan. Selain itu, karena urine yang dikeluarkan terlalu banyak, maka akan menyebabkan *polyphagia*, atau kondisi dimana manusia akan merasa lapar secara berlebihan. Selanjutnya yaitu ada *polydipsia*, yaitu kondisi dimana terjadi rasa haus yang berlebihan. Keluhan klasik lainnya yaitu seperti penurunan berat badan secara tiba-tiba, terdapat luka yang susah sembuh, dan lain lain. Pasien diabetes akan meningkat seiring berjalannya waktu. Menurut IDF (2021), diperkirakan akan terjadi peningkatan sebesar 12,2% pada jumlah pasien diabetes pada tahun 2045, dengan jumlah total mencapai 642,7 juta orang. Selain itu, pada tahun tersebut Indonesia akan diprediksi menempati ranking kelima. Oleh karena itu, penyakit ini harus diwaspadai semenjak dini.

Data mining adalah ekstraksi pola yang menarik dari sebuah data. Suatu pola dikatakan menarik apabila pola tersebut implisit, tidak diketahui sebelumnya, dan dapat dimanfaatkan. Dalam data mining, terdapat pembelajaran *supervised learning* dan *unsupervised learning* (Nengsih, 2019). *Supervised learning* merupakan sebuah pembelajaran dimana model dilatih untuk mengenali pola dari input data dan label output. Sedangkan untuk *unsupervised learning* merupakan pembelajaran dimana modelnya akan menemukan pola dari suatu data tanpa bantuan, dengan mengidentifikasi pola yang tersembunyi.

Dari data diabetes yang ada dapat dimanfaatkan menggunakan *supervised learning*, yaitu klasifikasi. Klasifikasi merupakan analisis data menggunakan model yang menggambarkan kelas data. Klasifikasi ini digunakan untuk mengambil data dan ditempatkan kedalam suatu kelompok (Aryani & Wijayanto, 2021). Menurut Mutmainah (2021), klasifikasi dapat didefinisikan sebagai teknik yang mempelajari tentang sekumpulan data sehingga mendapatkan aturan yang dapat mengenali data baru yang belum dipelajari sebelumnya. Metode yang populer dalam klasifikasi adalah *Random Forest*. Metode ini akan membentuk sejumlah pohon, dimana satu pohon akan memberikan satu unit suara. Hasil akhir pada klasifikasi metode *Random Forest* dilakukan dengan *majority voting*, atau akan diambil suara yang terbanyak. *Random Forest* memiliki keunggulan dapat memberikan akurasi yang tinggi (Hamami & Dahlan, 2022). Selain *Random Forest*, terdapat juga metode *Support Vector Machine* (SVM). SVM memiliki kelebihan diantaranya adalah dalam menentukan jarak menggunakan *support vector* sehingga proses komputasi menjadi cepat.

Sebuah penelitian tentang klasifikasi penyakit diabetes melitus telah dilakukan oleh Putri dkk. (2021). Penelitian tersebut menggunakan metode C.45. Dari hasil penelitian tersebut didapatkan akurasi sebesar 90%. Selain itu, penelitian menggunakan data medis telah dilakukan oleh Samosir dkk (2021). Penelitian tersebut menggunakan data penyakit jantung dengan metode yang digunakan yaitu *Random Forest*, *Naïve Bayes*, dan *K-Nearest Neighbor*. Dari hasil penelitian didapatkan bahwa metode *Naïve Bayes* mendapatkan performa yang paling baik, dimana akurasinya sebesar 0,91, AUC

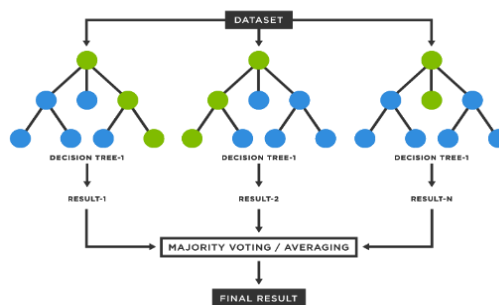
(*area under the curve*) 0,84, dan *F1-Score* sebesar 0,84. Penelitian dengan data penyakit diabetes juga telah dilakukan oleh Nugraha & Sabaruddin (2021). Dataset yang digunakan yaitu Pima Indians Diabetes, dengan total variabel sebanyak 9, dimana 1 variabelnya merupakan output/label. Dari penelitian tersebut didapatkan hasil bahwa dengan metode *Random Forest* akurasi sebesar 75,82% (*oversampling*), 71,24% (*undersampling*), dan 73,86 (*original*).

Berdasarkan uraian yang telah dijelaskan sebelumnya, maka dilakukan penelitian klasifikasi menggunakan data medis penyakit diabetes. Penelitian ini dilakukan untuk mengetahui metode yang cocok terhadap data penyakit diabetes menggunakan *data mining*. Metode yang digunakan dalam penelitian ini yaitu *Random Forest*. Selain itu, digunakan juga metode *Support Vector Machine* untuk membandingkan hasil dari metode *Random Forest*, sehingga didapatkan metode yang lebih baik pada klasifikasi penyakit diabetes ini.

2. Metode Penelitian

2.1. *Random Forest*

Random Forest merupakan sebuah metode yang terdiri dari beberapa pohon keputusan. Metode ini akan menghasilkan kelas pada kasus klasifikasi, dan prediksi rata rata pada kasus regresi. Cara kerja Random Forest dapat dilihat melalui Gambar 1.

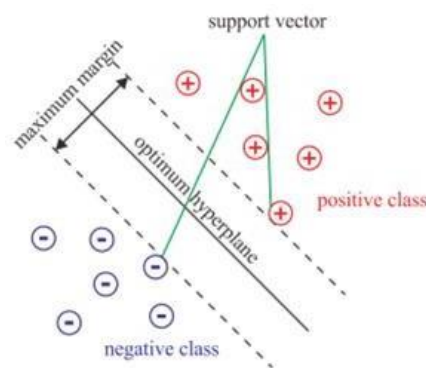


Gambar 1. Cara kerja *Random Forest*

Terdapat tiga poin penting dalam tahapan Random Forest. Pertama, melakukan bootstrap sampling untuk membangun pohon keputusan. Kedua, dengan prediktor yang acak, masing masing pohon akan memprediksi suatu hasil. Terakhir, Random Forest akan mengombinasikan hasil dari setiap pohon keputusan dengan cara mengambil suara terbanyak atau majority voting pada klasifikasi, atau rata rata untuk regresi. Menurut Pamuji dan Ramadhan (2021), algoritma Random Forest ini memiliki kelebihan yaitu dapat menghasilkan eror yang relatif rendah, performa yang baik dalam klasifikasi, dan cocok untuk data yang berjumlah besar.

2.2. Support Vector Machine

Support Vector Machine merupakan metode yang termasuk dalam supervised learning yang dapat digunakan dalam klasifikasi maupun regresi. Support Vector Machine merupakan teknik dengan tingkat presisi/akurasi dan kualitas yang tinggi (Nurhafida & Sembiring, 2022). Support Vector Machine menggunakan hyperplane untuk memisahkan dua kelas yang berbeda. Kelebihan dari SVM yaitu kemampuannya dalam mengidentifikasi hyperplane yang terpisah sehingga dapat memaksimalkan margin dari kelas yang berbeda (Bei & Saepudin, 2021). Ilustrasi cara kerja SVM dapat dilihat melalui Gambar 2.



Gambar 2. Cara kerja Support Vector Machine

Support Vector Machine akan memisahkan dua kelas yang berbeda menggunakan *hyperplane*. Untuk mencari *hyperplane* dapat dicari menggunakan rumus:

$$\min \frac{1}{2} \|w\|^2 \quad (1)$$

Dengan subjek, $y_i(w \cdot x_i + b) \geq 1$, dimana x_i merupakan data input, y_i merupakan nilai keluaran dari data x_i , sedangkan w dan b merupakan parameter yang akan dicari (Abdulloh & Pambudi, 2021).

2.3. Confusion Matrix

Pengukuran kinerja suatu metode merupakan suatu hal yang sangat penting. Pengukuran tersebut dapat dilihat melalui *confusion matrix*. *Confusion matrix* merupakan metode evaluasi yang digunakan dalam menghitung performa dalam sebuah klasifikasi. Menurut Amaliah dkk (2022), *confusion matrix* adalah tabel yang digunakan untuk melihat akurasi dan seberapa baik algoritma yang digunakan dari klasifikasi untuk memprediksi data *testing*. Tabel *confusion matrix* dapat dilihat melalui Tabel 1.

Tabel 1. Confusion Matrix

Prediksi	Aktual		Total
	+	-	
+	TP	FP	TP+FP
-	FN	TN	FN+TN
Total	TP+FN	FP+TN	TP+FP+FN+TN

Dalam *confusion matrix* terdapat empat istilah sebagai hasil klasifikasi. *True Positive* (TP) merupakan data positif yang terdeteksi secara benar. *True Negative* (TN) merupakan jumlah data negatif yang diklasifikasikan secara benar. Selain itu, terdapat juga *False Positive* (FP) dan *False Negative* (FN). *False Positive* (FP) merupakan data negatif namun diklasifikasikan sebagai data positif, sedangkan *False Negative* (FN) merupakan data positif namun dideteksi sebagai data negatif. Dari nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) dapat dihitung performanya melalui beberapa indikator. Indikator tersebut diantaranya yaitu akurasi, presisi, *recall*, *specificity*, dan *F1-Score*. *F1-Score* merupakan operasi gabungan dari presisi dan *recall*, sehingga kedua evaluasi tersebut dapat diwakilkan oleh *F1-Score* (Nuraini dkk., 2022). Perhitungan dari setiap indikator dapat diperoleh melalui rumus:

$$\text{akurasi} = \frac{TP+TN}{(TP+FP+TN+FN)} \quad (2)$$

$$\text{presisi} = \frac{TP}{(TP+FP)} \quad (3)$$

$$\text{recall} = \frac{TP}{(TP+FN)} \quad (4)$$

$$\text{specificity} = \frac{TN}{(TN+FP)} \quad (5)$$

$$\text{F1-score} = \frac{2 \times \text{presisi} \times \text{recall}}{(\text{presisi} + \text{recall})} \quad (6)$$

2.4. Deskripsi Data dan Langkah-Langkah Penelitian

Data yang digunakan dalam penelitian ini yaitu data penyakit diabetes. Data tersebut diperoleh dari Kaggle, yang bersumber dari Rumah Sakit Shyllet Bangladesh. Data ini terdiri dari 520 data dengan 17 variabel. Terdapat 200 data tidak menderita diabetes, dan 320 data menderita penyakit diabetes. Rincian variabel data dapat dilihat pada Tabel 2.

Tabel 2. Variabel Penelitian

Variabel Penelitian	Type Data	Keterangan
Usia dalam tahun	Integer	20 - 90
Jenis Kelamin	String	Laki- Laki/Perempuan
<i>Polyuria</i>	String	Ya/Tidak
<i>Polydipsia</i>	String	Ya/Tidak
Penurunan Berat Badan Secara Tiba-Tiba	String	Ya/Tidak
<i>Fatigue and Weakness</i>	String	Ya/Tidak
<i>Polyphagia</i>	String	Ya/Tidak
Infeksi Candidiasis	String	Ya/Tidak
Daya Penglihatan yang Berkurang	String	Ya/Tidak
Gatal	String	Ya/Tidak
<i>Irritability</i>	String	Ya/Tidak
Luka yang Sulit Sembuh	String	Ya/Tidak
<i>Partial Paresis</i>	String	Ya/Tidak
<i>Muscle Stiffness</i>	String	Ya/Tidak
<i>Alopecia</i>	String	Ya/Tidak
Obesitas	String	Ya/Tidak
Kelas	String	Positif/Negatif

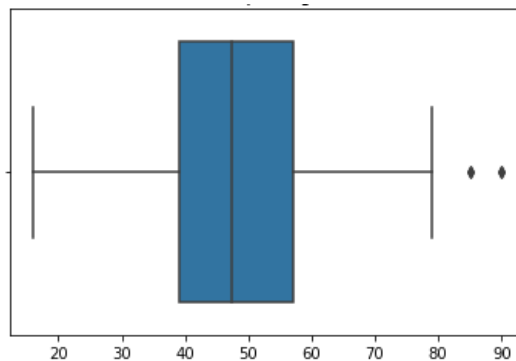
Langkah – Langkah yang dilakukan dalam penelitian ini yaitu:

1. Melakukan *pre-processing* data.
 - a. Melakukan *label encoding* terhadap data kategorik menggunakan *label encoder* pada Python.
 - b. Melakukan pengecekan *missing value*, *outliers*, dan keseimbangan data.
 - c. Membagi data menjadi data *training* dan *testing* menggunakan *stratified random sampling* dengan perbandingan 80%:20%, 75%:25%, dan 70%:30%.
2. Memproses data menggunakan *Random Forest*
 - a. Melatih data *training* yang sudah diseimbangkan dengan *Synthetic Minority Oversampling* (SMOTE) dan mencari parameter yang terbaik.
 - b. Melakukan prediksi pada data *testing*.
 - c. Melakukan perhitungan performa model menggunakan persamaan (2), (3), (4), (5), dan (6).
3. Memproses data menggunakan *Support Vector Machine*
 - a. Melatih data *training* yang sudah distandarisasi dan diseimbangkan dengan *Synthetic Minority Oversampling* (SMOTE). dan mencari parameter yang terbaik.
 - b. Melakukan prediksi pada data *testing*.
 - c. Melakukan perhitungan performa model menggunakan persamaan (2), (3), (4), (5), dan (6).
4. Mengevaluasi model yang lebih baik antara *Random Forest* dan *Support Vector Machine* berdasarkan akurasi, presisi, *recall*, *specificity*, dan *F1-score* yang diperoleh.

3. Hasil dan Pembahasan

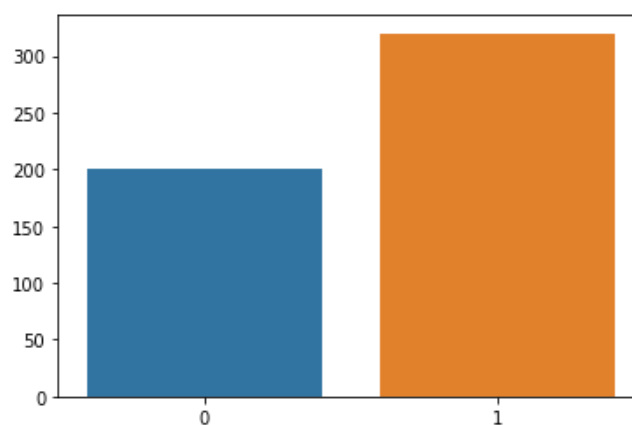
3.1. *Pre-processing Data*

Pada *pre-processing* data, langkah pertama yang dilakukan yaitu melakukan *label encoding* pada data kategorik. Data dengan label 0 bernilai “Tidak” dan “Negatif”. Label 1 diberikan pada data yang mengandung nilai “Ya” dan “Positif”. Setelah dilakukan *label encoding*, dilakukan pengecekan *missing value*. Dari data data yang ada, tidak ditemukan adanya *missing value*. Langkah selanjutnya yaitu pengecekan terhadap *outliers* pada data. Dari variabel yang ada, hanya variabel umur yang memiliki *outliers*. *Outliers* tersebut dapat dilihat melalui Gambar 3.



Gambar 3. *Boxplot* usia

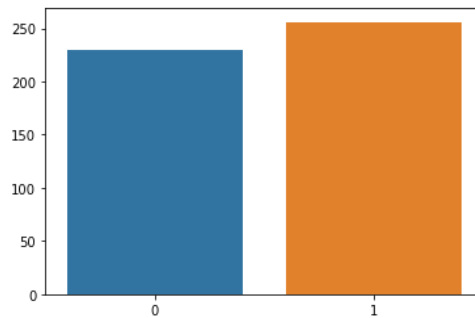
Gambar 3 memperlihatkan bahwa terdapat *outliers* pada variabel usia. Akan tetapi, *outliers* pada variabel umur tidak dibuang. Hal itu dikarenakan *outliers* tersebut merupakan fenomena dari subjek penelitian, sehingga tidak dilakukan pembuangan data. Langkah selanjutnya yaitu dilakukan pengecekan keseimbangan data. Keseimbangan data dapat dilihat melalui Gambar 4.



Gambar 4. Keseimbangan data

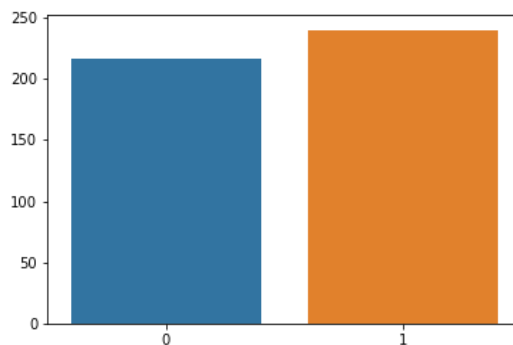
Gambar 4 memperlihatkan bahwa data tidak seimbang. Oleh karena itu, harus dilakukan penanganan keseimbangan data. Data dapat diseimbangkan menggunakan *Synthetic Minority Oversampling Technique* (SMOTE). Data tersebut diseimbangkan setelah data di *split* menjadi data *training* dan data *testing*. Data di *split* dengan

perbandingan 80%:20%, 75%:25%, dan 70%:30% menggunakan *stratified random sampling*.



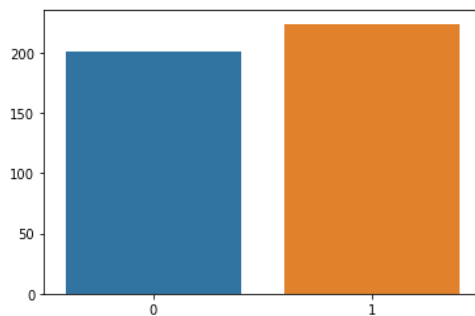
Gambar 5. Data *training* 80%:20% setelah dilakukan *balancing* data

Pada kondisi *split* dengan rasio 80%:20%, jumlah data *training* dengan label 0 sebanyak 160 dan label 1 sebanyak 256. Karena data tidak seimbang, maka data diseimbangkan dengan SMOTE. Hasil *resampling* data dapat dilihat pada Gambar 5. Dari Gambar 5 terlihat bahwa data sudah seimbang. Data dengan label 0 setelah diseimbangkan menjadi sebanyak 230 data.



Gambar 6. Data *training* 75%:25% setelah dilakukan *balancing* data

Pada rasio 75%:25%, jumlah data *training* dengan label 0 berjumlah sebanyak 150, sedangkan data dengan label 1 berjumlah sebanyak 240 data. Data tersebut dikatakan belum seimbang. Oleh karena itu dilakukan penanganan keseimbangan data menggunakan SMOTE. Gambar 6 merupakan kondisi data setelah dilakukan *resampling* menggunakan SMOTE. Data dengan label 0 bertambah, dimana data ini bertambah menjadi 216 data. Setelah dilakukan *resampling*, data sudah menjadi seimbang/*balance*.



Gambar 7. Data *training* 70%:30% setelah dilakukan *balancing* data

Sama seperti sebelumnya, data dengan kondisi *split* 70%:30% juga belum seimbang. Dari data tersebut terdapat data dengan label 0 sebanyak 140 dan label 1 sebanyak 224. Gambar 7 memperlihatkan data setelah dilakukan *resampling*, dimana data menjadi seimbang/*balance*. Data dengan label 0 bertambah menjadi 201 data.

3.2. Klasifikasi Menggunakan Algoritma *Random Forest*

Random Forest merupakan salah satu dari algoritma klasifikasi dalam *data mining*. Dalam penelitian ini dilakukan klasifikasi dengan tiga *split* data yang berbeda. Hasil dari klasifikasi menggunakan algoritma *Random Forest* yaitu:

a. *Split* data 80%:20%

Setelah dilakukan *pre-processing* data, selanjutnya dilakukan diklasifikasikan menggunakan algoritma *Random Forest*. Algoritma akan mempelajari data melalui data *training* yang sudah seimbang/*balance*. Untuk mendapatkan parameter yang terbaik, maka dilakukan *hyperparameter tuning*. Dari *hyperparameter tuning* didapatkan bahwa parameter yang paling optimal yaitu jumlah pohon sebesar 200. *Confusion matrix* dapat dilihat melalui Tabel 3.

Tabel 3. *Confusion Matrix Random Forest Split Data 80%:20%*

Prediksi	Aktual		Total
	Diabetes	Tidak Diabetes	
Diabetes	62	2	64
Tidak Diabetes	0	40	40
Total	62	42	104

Tabel 3 menjelaskan bahwa algoritma *Random Forest* dengan *split* data 80%:20% dapat mengklasifikasikan pasien yang tidak terkena diabetes dengan benar sebanyak 40, dan pasien yang terkena diabetes secara benar sebanyak 62. Selain itu, terdapat 2 data yang seharusnya tidak diabetes namun diklasifikasikan sebagai diabetes. Dari Tabel 2 dapat dihitung evaluasi modelnya, yaitu akurasi, presisi, *recall*, *spesifity*, dan *F1-score* menggunakan persamaan (2), (3), (4), (5), dan (6). Dari perhitungan, didapatkan akurasi sebesar 0,98, presisi sebesar 0,96, *recall* sebesar 1, *specificity* sebesar 0,95, dan *F1-score* sebesar 0,98. Algoritma ini mampu mengklasifikasikan sebanyak 102 data sampel dari total 104 data *testing* secara benar. Dengan kata lain, algoritma ini mampu mengklasifikasikan 98% secara benar.

b. *Split* data 75%:25%

Kondisi yang kedua yaitu *split* data dengan rasio data *training* sebesar 75% dan data *testing* sebesar 25%. Setelah data diseimbangkan, dilakukan *hyperparameter tuning* pada data *training*. Didapatkan bahwa jumlah pohon terbaik sebesar 100. Hasil *confusion matrix* dapat dilihat melalui Tabel 4.

Tabel 4. *Confusion Matrix Random Forest Split Data 75%:25%*

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	75	5	80
Tidak Diabetes	1	49	50
Total	76	54	130

Tabel 4 memvisualisasikan kinerja dari metode *Random Forest*. Tabel 4 memperlihatkan bahwa metode ini dapat memprediksi secara benar penyakit diabetes sebesar 75 data dan memprediksi 49 data tidak diabetes secara benar. Namun, ada 1 data yang salah klasifikasi, yang seharusnya diabetes namun diklasifikasikan sebagai tidak diabetes, dan 5 data yang seharusnya tidak diabetes namun diklasifikasikan sebagai diabetes. Dari perhitungan evaluasi model diketahui bahwa akurasi sebesar 0,95, presisi bernilai 0,3, *recall* sebesar 0,98, *specificity* 0,9, dan *F1-score* sebesar 0,95.

c. *Split data 70%:30%*

Sama seperti sebelumnya, setelah data di *split* dengan rasio 70% dan 30% dan diseimbangkan, dilakukan *hyperparameter tuning*. Hal tersebut dilakukan untuk mendapatkan parameter yang baik, sehingga menghasilkan klasifikasi yang maksimal. Pada kondisi ini didapatkan jumlah pohon yang maksimal yaitu sebesar 200. Hasil klasifikasi dapat dilihat melalui Tabel 5.

Tabel 5. *Confusion Matrix Random Forest Split Data 70%:30%*

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	91	5	96
Tidak Diabetes	1	59	60
Total	92	64	156

Dari hasil klasifikasi dapat diketahui bahwa metode dengan *split data* 70% dan 30% mampu mengklasifikasikan 91 data terkena diabetes secara benar. Selain itu, algoritma ini juga mampu mengklasifikasikan 59 data tidak terkena diabetes dengan benar. Akan tetapi, pada metode ini mengklasifikasikan 5 data yang sebenarnya tidak diabetes diklasifikasikan ke diabetes, dan 1 data yang seharusnya diabetes namun diklasifikasikan sebagai tidak diabetes. Dari hasil perhitungan evaluasi model, dapat diketahui bahwa algoritma ini memprediksi 150 data secara benar dari 156 data *testing* yang ada. Akurasi yang didapatkan dari model ini yaitu sebesar 96%, presisi sebesar 0,94, *recall* sebesar 0,98, *specificity* sebesar 0,92, dan *F1-score* sebesar 0,96.

3.3. Klasifikasi Menggunakan Algoritma *Support Vector Machine*

Pada klasifikasi ini, dilakukan juga *splitting* data dengan tiga kondisi yang sama seperti sebelumnya, yaitu 80%:20%, 75%:25%, dan 70%:30%. Setelah di *split*, data *training* yang sudah diseimbangkan dilatih dengan metode ini. Hasil klasifikasi menggunakan *Support Vector Machine* yaitu sebagai berikut.

a. *Split data 80%:20%*

Seperti halnya pada *Random Forest*, data yang sudah seimbang dilatih untuk menemukan pola yang berguna. Data ini dilatih menggunakan algoritma *Support Vector Machine* serta dilakukan *hyperparameter tuning*. Dari hasil *hyperparameter tuning*

didapatkan kernel terbaiknya yaitu linear dan C 0,1. Hasil klasifikasi pada data *testing* dapat dilihat pada Tabel 6.

Tabel 6. *Confusion Matrix Support Vector Machine Split Data 80%:20%*

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	57	7	64
Tidak Diabetes	1	39	40
Total	58	46	104

Tabel 6 memperlihatkan hasil klasifikasi menggunakan *Support Vector Machine*. metode ini secara benar mengklasifikasikan diabetes secara benar sebanyak 57 data, dan tidak diabetes sebanyak 39 data. Selain itu, algoritma ini juga mengklasifikasikan yang seharusnya diabetes diklasifikasikan ke tidak diabetes sebanyak 1 data, dan juga data yang seharusnya tidak diabetes diklasifikasikan ke diabetes sebesar 7 data. Oleh karena itu, akurasi dari model ini bernilai 0,92 atau 92%. Nilai presisi sebesar 0,89, *recall* sebesar 0,98, *specificity* sebesar 0,84, dan *F1-score* pada model ini bernilai sebesar 0,93.

b. *Split data 75%:25%*

Pada *split data* dengan rasio data *training* dan data *testing* sebesar 75%:25%, sama seperti sebelumnya, dilakukan pemilihan parameter yang terbaik. Dari hasil *hyperparameter tuning* didapatkan bahwa kernel terbaiknya yaitu linear dan C sebesar 0,1. Hasil klasifikasi dapat dilihat melalui Tabel 7.

Tabel 7. *Confusion Matrix Support Vector Machine Split Data 75%:25%*

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	69	11	80
Tidak Diabetes	2	48	50
Total	71	59	130

Tabel 7 dapat menjelaskan kinerja dari metode *Support Vector Machine*. Metode ini mampu mengklasifikasikan secara benar penyakit diabetes sebanyak 69 data, dan tidak diabetes sebanyak 48 data. Selain itu, algoritma ini mengklasifikasikan 11 data yang seharusnya tidak diabetes namun diklasifikasikan sebagai diabetes, dan 2 data yang seharusnya diabetes diklasifikasikan sebagai tidak diabetes. Maka dari itu, didapatkan akurasi dari model sebesar 0,90 atau sebesar 90%, serta presisi sebesar 0,86, *recall* sebesar 0,97, *specificity* sebesar 0,81, dan *F1-score* sebesar 0,90.

c. *Split data 70%:30%*

Pada kondisi *split data* 70%:30%, didapatkan kernel yang terbaik dalam klasifikasi ini yaitu kernel linear dan C 0,1. Dengan kernel linear akurasi yang didapatkan yaitu sebesar 96%. Hasil klasifikasi secara lebih detail dapat dilihat melalui Tabel 8.

Tabel 8. *Confusion Matrix Support Vector Machine Split Data 70%:30%*

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	85	11	96
Tidak Diabetes	4	56	60
Total	89	67	156

Akurasi model yang didapatkan yaitu sebesar 0,9 atau 90%. Hal tersebut berarti bahwa algoritma ini mampu mengklasifikasikan secara benar sebanyak 85 data diabetes, dan 56 data tidak diabetes dari 156 data yang ada. Selain itu, SVM mengklasifikasikan 11 data yang seharusnya tidak diabetes namun diklasifikasikan sebagai diabetes, dan 4 data yang seharusnya diabetes namun diklasifikasikan sebagai tidak diabetes. Nilai dari presisi sebesar 0,88, *recall* 0,95, *specificity* 0,83, dan *F1-Score* sebesar 0,91.

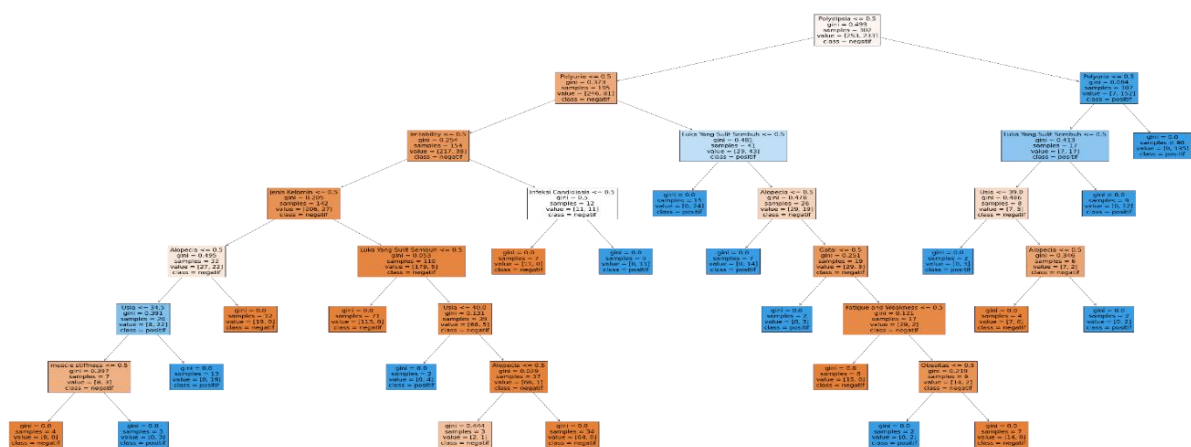
3.4. Perbandingan Hasil Klasifikasi Menggunakan *Random Forest* dan *Support Vector Machine*

Setelah dilakukan klasifikasi, tahap selanjutnya yaitu pemilihan model. Model yang sudah dibentuk sebelumnya dibandingkan untuk memilih model yang terbaik dalam klasifikasi penyakit diabetes ini. Perbandingan model dari *Random Forest* dan *Support Vector Machine* dapat dilihat melalui Tabel 9.

Tabel 9. Perbandingan Hasil Klasifikasi Menggunakan *Random Forest* & *Support Vector Machine*

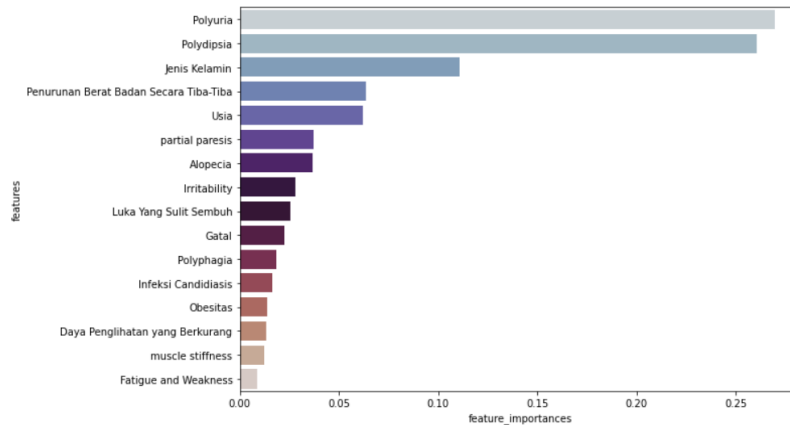
	<i>Random Forest</i>			<i>Support Vector Machine</i>		
	80%:20%	75%:25	70%:30	80%:20	75%:25	70%:30
		%	%	%	%	%
Akurasi	0,98	0,95	0,96	0,92	0,90	0,90
Presisi	0,96	0,93	0,94	0,89	0,86	0,88
Recall	1	0,98	0,98	0,98	0,97	0,95
Specificity	0,95	0,9	0,92	0,84	0,81	0,83
F1-score	0,98	0,95	0,96	0,93	0,90	0,91

Tabel 9 memperlihatkan bahwa metode *Random Forest* dan *Support Vector Machine* memiliki performa lebih dari 0,9 atau 90%. Menurut Aulianita dalam Azies dan Anuraga (2021), apabila akurasi lebih besar dari 90% maka dapat dikategorikan sebagai sangat baik. Model yang terbaik didapatkan oleh algoritma *Random Forest* dengan *split* data sebesar 80%:20%. Hal tersebut dikarenakan pemilihan model terbaik dilihat melalui *F1-score*. *F1-score* merupakan perbandingan terbobot dari presisi dan *recall*. Selain itu, *Random Forest* dengan *split* data 80%:20% dipilih sebagai model yang terbaik karena metode ini juga memiliki akurasi yang paling tinggi.



Gambar 7. Visualisasi salah satu *tree* di *Random Forest*

Gambar 7 merupakan visualisasi dari salah satu pohon di *Random Forest*. Gambar tersebut merupakan pohon dengan *split* data 80%:20%. Selain itu, pohon tersebut merupakan pohon ke 190 dari total 200 pohon yang ada.



Gambar 8. *Feature importance* Random Forest

Melalui *Random Forest* dapat diketahui juga variabel yang paling berpengaruh terhadap penyakit diabetes. Gambar 8 merupakan *feature importance* dari *Random Forest* dengan *split* data 80%:20%. Urutan variabel yang paling berpengaruh terhadap diabetes dalam penelitian ini yaitu *polyuria*, *polydipsia*, dan jenis kelamin. Dengan data yang lebih banyak, informasi yang didapatkan tentunya juga lebih banyak dan lebih mudah dipelajari oleh model. Dengan 17 variabel ciri dan faktor risiko diabetes didapatkan akurasi sebesar 98%. Hal ini lebih baik dibandingkan penelitian yang dilakukan oleh Nugraha & Sabaruddin (2021) dengan data diabetes dimana variabel yang digunakan sebanyak 7 variabel, dengan akurasi paling baik yang dihasilkan yaitu sebesar 75,82%.

4. Kesimpulan dan Saran

Berdasarkan hasil analisis yang telah dilakukan, didapatkan kesimpulan bahwa metode *Random Forest* dengan *split* data 80%:20% merupakan metode yang terbaik dalam klasifikasi penyakit diabetes ini. Akurasi yang didapatkan sebesar 0,98, presisi 0,96, *recall* 1, *specificity* 0,95, dan *F1-score* sebesar 0,98. Selain itu, variabel yang paling berpengaruh dalam klasifikasi ini berturut turut yaitu *polyuria*, *polydipsia*, dan jenis kelamin.

Pustaka

- Abdulloh, F. F, & Pambudi, I. R. (2021). Analisis Sentimen Pengguna Youtube Terhadap Program Vaksin Covid-19. *CSRID (Computer Science Research And Its Development Journal)*, 13(3), 141-148.
- Al Azies, H., & Anuraga, G. (2021). Classification of Underdeveloped Areas In Indonesia Using The SVM And K-NN Algorithms. *Jurnal ILMU DASAR*, 22(1), 31. DOI: 10.19184/Jid.V22i1.16928
- Amaliah, S., Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi Di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, 4(3), 121-127.

- Aryani, Y., & Wijayanto, A. W. (2021). Klasifikasi Pengembalian Radar dari Ionosfer Menggunakan SVM, Naive Bayes dan Random Forest. *Komputika: Jurnal Sistem Komputer*, 10(2), 111-117.
- Bei, F., & Saepudin, S., “Analisis Sentimen Aplikasi Tiket Online di Play Store Menggunakan Metode Support Vector Machine (SVM),” *Sismatik*, Vol. 01, No. 01, pp. 91–97, 2021
- Hamami, F. & Dahlan, A. (2022). Klasifikasi Cuaca Provinsi DKI Jakarta Menggunakan Algoritma Random Forest Dengan Teknik Oversampling. *Jurnal TEKNOINFO* (Vol. 16, Issue 1).
- IDF. (2021). IDF Diabetes Atlas 10th Edition. Diakses dari www.diabetesatlas.org
- Iman, Q., & Wijayanto, A. W. (2021). Klasifikasi Rumah Tangga Penerima Beras Miskin (Raskin)/Beras Sejahtera (Rastra) di Provinsi Jawa Barat Tahun 2017 dengan Metode Random Forest dan Support Vector Machine. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, 9(2), 178-184.
- Kaggle. (2020). Diakses dari www.kaggle.com/datasets/alakaaay/diabetes-uci-dataset
- Mutmainah, S. (2021). Penanganan Imbalanced Data Pada Klasifikasi Kemungkinan Penyakit Stroke. *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*, 1(1)
- Nengsih, W. (2019). Analisa Akurasi Permodelan Supervised dan Supervised Learning Menggunakan Data Mining. *Sebatik*, 23(2), 285-291.
- Nugraha, W., & Sabaruddin, R. (2021). Teknik *Resampling* untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Diabetes Menggunakan C4. 5, *Random Forest*, dan SVM. *Techno. Com*, 20(3), 352-361.
- Nuraini, A. F., Pertiwi, R. D., Subarkah, M. Z., & Ferawati, K. (2022, November). Analisis Sentimen Pelaksanaan Vaksinasi Covid-19 secara Massal pada Media Sosial Twitter. In *Seminar Nasional Official Statistics* (Vol. 2022, No. 1, pp. 1385-1394).
- Nurhafida, S. I., & Sembiring, F. (2022). Analisis Sentimen Aplikasi Novel Online Di Google Play Store Menggunakan Algoritma Support Vector Machine (SVM). *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, 6(1), 317-327.
- Pamuji, F. Y. & Ramadhan, V. P. (2021). Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotherapy. *Jurnal Teknologi dan Manajemen Informatika*, 7(1), 46-50.
- Putri, S. U., Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4. 5. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, 2(1), 39-46.
- Samosir, A., Hasibuan, M. S., Justino, W. E., & Hariyono, T. (2021, September). Komparasi Algoritma Random Forest, Naïve Bayes dan K-Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung. In *Prosiding Seminar Nasional Darmajaya* (Vol. 1, pp. 214-222).
- Utomo, A. A., Rahmah, S., & Amalia, R. (2020). Faktor Risiko Diabetes Mellitus Tipe 2: A Systematic Review. *AN-NUR: Jurnal Kajian dan Pengembangan Kesehatan Masyarakat*, 1(1), 44-53.

WHO. (2020). The Top 10 Causes of Death. Diakses dari <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death> pada 8 Agustus 2022.

WHO. (2022). Diabetes. Diakses dari <https://www.who.int/news-room/fact-sheets/detail/diabetes> pada 21 September 2022

