Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975

Prediksi Diabetes Menggunakan Algoritma K-Nearest (KNN) Teknik SMOTE-ENN

Zaenul Amri[1*], Muhammad Rodi[2], M. Nurul Wathani[3], Amir Bagja[4], Zulkipli[5]
[1,3,5]Program Studi Informatika, Universitas Hamzanwadi
[2]Program Studi Sistem Informasi, STMIK Lombok
[4]Program Studi Sistem Informasi, Universitas Hamzanwadi
*zaenulamri@hamzanwadi.ac.id

Abstrak

Dewasa ini, diabetes merupakan penyakit umum yang memengaruhi jutaan orang di seluruh dunia, dan umumnya wanita lebih banyak terkena penyakit ini. Penelitian kesehatan terbaru telah menerapkan berbagai teknologi inovatif dan canggih untuk mendiagnosis orang dan memprediksi penyakit mereka berdasarkan data klinis. Salah satu teknologi tersebut adalah pembelajaran mesin (ML) di mana diagnosis dan prediksi dapat dilakukan dengan lebih akurat. data yang digunakan pada penelitian ini yaitu dataset diabetes wanita Pima Indian dari Kaggle dan repositori data UCI yang dimana pada penelitian ini akan dilakukan prediksi menggunakan model algoritma KNN dengan menerapkan optimasi pada dataset menggunakan Teknik SMOTE-ENN untuk dapat meningkatkan hasil akurasi prediksi pada penyakit diabetes wanita Pima Indian. Pelatihan dan pengujian pada dataset dilakukan sebanyak lima kali pembagian dengan bantuan tools jupyter notebook untuk melihat hasil akurasi terbaik pada kinerja model algoritma KNN dengan parameter akurasi klasifikasi, error klasifikasi dan kurva ROC serta melihat variable apa saja yang dapat mempengaruhi resiko terkena penyakit diabetes. Hasil penelitian menunjukkan bahwa dengan menerapkan optimasi SMOTE-ENN pada dataset penelitian dapat meningkatkan hasil akurasi prediksi menggunakan model algoritma KNN dengan pembagian data training 70% dan testing 30% mencapai akurasi klasifikasi 0.96, error klasifikasi 0.04, dan AUC sebesar 0.95. Hasil prediksi tersebut menunjukkan bahwa cenderung pada Wanita pima Indian mengalami diabetes yang dipengaruhi oleh Umur di atas 33 tahun, jumlah kehamilan, konsumsi gula yang berlebihan, tekanan darah, ketebalan kulit, insulin, IMT (Indeks Massa Tubuh), dan fungsi pewarisan diabetes.

Kata kunci : KNN, Prediksi, SMOTE-ENN

Abstract

Nowadays, diabetes is a common disease affecting millions of people worldwide, and it is generally more prevalent among women. Recent health research has adopted various innovative and advanced technologies to diagnose individuals and predict diseases based on clinical data. One such technology is Machine Learning (ML), which enables more accurate diagnosis and prediction. The data used in this study is the Pima Indian women diabetes dataset from Kaggle and the UCI data repository. This study focuses on predicting diabetes using the KNN algorithm model by applying optimization to the dataset using the SMOTE-ENN technique to enhance prediction accuracy for Pima Indian women. The dataset was trained and tested with five different splits using Jupyter Notebook to determine the best accuracy for the KNN algorithm model. Parameters such as classification accuracy, classification error, and the ROC curve were evaluated, along with identifying the variables influencing the risk of diabetes. The results showed that applying SMOTE-ENN optimization to the research dataset significantly improved the prediction accuracy using the KNN algorithm model. With a 70% training and 30% testing data split, the model achieved a classification accuracy of 0.96, a classification error of 0.04, and an AUC of 0.95. These predictions indicated that Pima Indian women are more likely to develop diabetes due to factors such as age above 33 years, the number of pregnancies, excessive sugar consumption, blood pressure, skin thickness, insulin levels, BMI (Body Mass Index), and genetic predisposition to diabetes.

Keywords : *KNN, Prediction, SMOTE-ENN.*

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975

1. Pendahuluan

Diabetes merupakan penyakit tidak menular yang secara serius memengaruhi sistem kesehatan dengan mengurangi efisiensi seseorang [1] [2] [3] [4] [5] [6] Pada penyakit ini, kadar glukosa darah naik melebihi kadar normal dalam tubuh [7] [8] [9] [10] [11]. Penting untuk dicatat bahwa glukosa adalah bentuk gula yang dibutuhkan oleh tubuh untuk metabolisme yang lebih baik. Semua sel membutuhkan glukosa sebagai sumber energi. Namun, jika kadar glukosa darah meningkat karena kekurangan hormon insulin dalam tubuh, maka glukosa darah menjadi tidak seimbang dan menyebabkan kerusakan serius pada bagian lain tubuh, seperti mata, jantung, ginjal, dan banyak lagi [12][13][14][7]. Pengendalian dapat dilakukan dengan mengubah gaya hidup, seperti kebiasaan makan, pengobatan, dan olahraga, sebagai contoh. Jika penyakit ini didiagnosis tepat waktu melalui prediksi, maka kesehatan seseorang dapat diperbaiki. Oleh karena itu, prediksi perlu dilakukan untuk mendapatkan *knowledge* secara berurutan berdasarkan bukti-bukti dengan menggunakan teknik statistik, matematik, kecerdasan buatan (*Artificial Intelegent*) dan *machine learning* untuk mengestrak pengetahuan atau menemukan pola dari suatu data yang besar [1][2][15]. salah satunya adalah teknik penambangan data (data mining). Data mining yaitu proses menemukan korelasi baru yang bermakna, pola dan tren dengan memilah-milah sejumlah data besar yang tersimpan dalam repositori, menggunakan teknologi penalaran pola serta teknik-teknik statistik dan matematika [6] [8]. Data mining adalah proses untuk mendapatkan ilmu pengetahuan dari sebuah informasi yang berasal dari gudang basis data. Teknik data mining merupakan cara yang mudah dan relatif cepat untuk memperoleh pengetahuan secara otomatis [16] [17].

Adapun model *data mining* yang digunakan untuk klasifikasi data yaitu, *algoritma naïve bayes, decision tree, artificial neural network*, K- *nearest neighbor* (K-NN), *support vector machine* (SVM) dan lain-lain. Namun pada penelitian ini akan menggunakan model algoritma KNN yang yang dikenal luas dan memiliki kinerja yang baik dalam melakukan klasifikasi dengan menerapkan optimasi pada dataset dengan Teknik SMOTE-ENN [18]. Algoritma KNN ini bekerja dengan cara menemukan tetangka terdekat dari suatu titik data berdsarkan Euclidean atau metrix jarak lainnya. Oleh karena itu, penelitian ini bertujuan untuk meningkatkan kinerja model algoritma KNN dalam melakukan prediksi dengan menerapkan teknik SMOTE-ENN pada dataset Pima Indian. Teknik SMOTE (*oversampling*) dan *Edited Nearest Neighbors* (ENN) dapat meningkatkan representasi kelas minoritas dengan menciptakan sampel sintetis menggunakan SMOTE, lalu menghilangkan sampel-sampel sintetis yang dihasilkan yang dianggap tidak valid dengan

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975

menggunakan ENN. Teknik ini terbukti dapat meningkatkan hasil akurasi klasifikasi pada dataset yang tidak seimbang [18]

2. Tinjauan Pustaka

2.1. Penelitian Terkait

Beberapa penelitian yang dilakukan sebelumnya membahas prediksi penyakit diabetes dengan menggunakan machine learning sebagai berikut.

- Penelitian yang dilakukan oleh Sourav Kumar Bhoi, et. Al memprediksi penyakit diabetes pada Wanita keturunan suku Indian pima dengan menggunakan supervised learning. Pada penelitian tersebut didapatkan hasil akurasi terbaik dengan menggunakan algoritma Linier Regresion dengan AUC sebesar 0.83 [19].
- Selanjutnya oleh Aziz Perdana prediksi penyakit diabetes menggunakan model algoritma KNN dan menganalisis fitur apa saja yang berpengaruh pada penyakit diabetes. Hasil penelitian tersebut menunjukkan bahwa akurasi model algoritma KNN sebesar 83,12% yang dipengaruhi oleh fitur glukosa, usia, insulin, tekanan darah, Massa Tubuh Indeks, kehamilan, ketebalan kulit, dan fungsi silsilah diabetes [20].
- Selanjutnya oleh Mariem Benarbia memprediksi penyakit diabetes menggunakan supervised learning dengan membandingkan dataset tidak seimbang dengan yang seimbang. Hasil terbaik didapatkan oleh dataset yang tidak seimbang atau dataset awal dengan model logistic regression akurasi didapat sebesar 82,00% [21]. Kemudian oleh Mani Abedini memprediksi penyakit diabetes menggunakan model algoritma supervised learning dan didapatkan akurasi model mencapai 83,08% [22].
- Penelitian yang dilakukan oleh Victor Chang memprediksi penyakit diabetes menggunakan supervised learning didapatkan hasil akurasi terbaik pada model algoritma random forest sebesar 79,57% [23]. Beberapa penelitian diatas menunjukkan akurasi model yang terbaik menggunakan model algoritma supervised learning, namun pada penelitian tersebut belum menerapakan metode optimasi SMOTE-ENN seperti yang dilakukan pada penelitian sebelumnya oleh Zaenul Amri dalam melakukan prediksi dengan menerapkan teknik optimasi Smotenn pada algoritma KNN yang terbukti dapat meningkatkan hasil akurasi model algoritma KNN yang mencapai 96,95% [18]. Adapun model *algoritma* KNN ini dalam mengklasifikasikan data dengan menerapkan optimasi data dengan Teknik SMOTE-ENN dapat dengan baik menangani kompleksitas keputusan nonlinier dan intraksi antar fitur dan mampu mengurangi overtfitting atau mengurangi jumlah sampel yang tidak

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975

konsisten pada dataset yang relative kecil karena dapat menyesuaikan diri dengan data yang terbatas sehingga akurasi model yang dihasilkan sangat tinggi dalam melakukan prediksi [18]. Pada penelian selanjutnya akan mencoba menerapkan model algoritma KNN dengan melakukan optimasi pada dataset pima Indian untuk meningkatkan hasil akurasi dalam melakukan prediksi penyakit diabetes.

2.2. Landasan Teori

1. Diabetes

Diabetes mellitus adalah penyakit kronis yang ditandai dengan kadar gula darah yang tinggi akibat gangguan produksi atau penggunaan insulin. Diabetes dapat diklasifikasikan menjadi tiga tipe utama: diabetes tipe 1, diabetes tipe 2, dan diabetes gestasional. Faktor-faktor yang memengaruhi risiko diabetes meliputi usia, indeks massa tubuh (BMI), tekanan darah, kadar glukosa, ketebalan kulit, dan faktor genetik. Identifikasi dini terhadap faktor-faktor tersebut sangat penting untuk mencegah komplikasi yang lebih serius.

2. Machine Learning

Machine learning adalah cabang dari kecerdasan buatan yang menggunakan data untuk melatih model agar dapat memprediksi atau mengklasifikasi suatu fenomena. Dalam konteks medis, algoritma machine learning sering digunakan untuk mendeteksi pola pada data pasien dan memberikan prediksi diagnosis atau prognosis penyakit. Algoritma berbasis pembelajaran terawasi (*supervised learning*) seperti *K-Nearest Neighbor (KNN)*, *Logistic Regression*, dan *Random Forest* telah banyak digunakan untuk memprediksi penyakit seperti diabetes.

3. Algoritma *K-Nearest Neighbor (KNN)*

Algoritma *K-Nearest Neighbor (KNN)* adalah salah satu metode *supervised learning* yang digunakan untuk klasifikasi dan regresi. KNN bekerja dengan mencari sejumlah k tetangga terdekat dari data baru berdasarkan metrik jarak, seperti *Euclidean distance*. Data baru kemudian diklasifikasikan berdasarkan mayoritas kelas tetangga terdekat.

4. Teknik SMOTE-ENN

Teknik SMOTE-ENN adalah kombinasi dari dua pendekatan untuk menangani dataset yang tidak seimbang. SMOTE (Synthetic Minority Oversampling Technique) Teknik ini menghasilkan sampel sintetis untuk kelas minoritas dengan melakukan interpolasi antara sampel asli. Hal ini membantu meningkatkan representasi kelas minoritas dalam dataset. ENN (Edited Nearest Neighbor) Teknik ini membersihkan dataset dengan menghapus sampel yang salah diklasifikasikan oleh algoritma KNN. Proses ini membantu mengurangi noise dan memperbaiki kualitas data. Kombinasi SMOTE dan ENN bertujuan untuk mengatasi masalah

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975

dataset yang tidak seimbang sekaligus mengurangi noise, sehingga meningkatkan performa model prediksi.

5. Evaluasi Model Machine Learning

Kinerja algoritma prediksi biasanya dievaluasi menggunakan metrik seperti:

- Akurasi: Persentase prediksi yang benar dari total prediksi.
- Precision, Recall, dan F1-Score: Metrik yang mempertimbangkan keseimbangan antara kelas positif dan negatif, terutama pada dataset yang tidak seimbang.

Area Under the Curve (AUC): Metrik untuk mengukur kemampuan model dalam membedakan antara kelas positif dan negatif.

3. Metode Penelitian

3.1. Jenis metode Penelitian

Jenis penelitian yang digunakan oleh peneliti adalah eksperimen dimana peneliti melakukan sebuah eksperimen penelitian dengan menentukan variable dependen dan variable independen pada dataset penelitian. yang akan digunakan. Dalam hal ini dataset tersebut dilakukan klasifikasi dengan menggunakan model algoritma *K- Nearest Neighbor* (K-NN) dalam memprediksi penyakit diabetes atau tidak diabetes. Dataset ini [24] diambil dari National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) berjumlah 768 data dengan 9 variabel diantaranya; Age, Pregnancies, Glukose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, dan Outcome.

3.2. Metode Analisis Data

Metode analisis data mengikuti tahapan yang ada pada proses *knowledge discovery in database* (KDD). Proses KDD dimulai dengan menetapkan tujuan dan diakhiri dengan evaluasi [25]. Tahapan (KDD) yang digunakan pada penelitian ini yaitu, seleksi data (*selection*) yang berjumlah 768 Data dengan 9 variabel. Pemilihan data (*preprocessing/cleaning*) untuk melakukan pembersihan data ganda atau bersifat outlier pada dataset dan mengidentifikasi *missing value* pada dataset. Transformasi (*transformation*) mengubah data awal ke dalam bentuk numerik. Data mining melakukan perhitungan pengujian model algoritma. Interpretasi/evaluasi melakukan perhitungan kinerja model algoritma. hasil perhitungan atau prediksi dengan bantuan *tools software exel* dan *jupyter notebook*.

4. Hasil dan Pembahasan

4.1. Hasil Penelitian

Pada tahap ini dilakukan seleksi dataset yang berjumlah 768 Data dengan 9 variabel seperti pada Tabel 1 di bawah ini.

Tabel 1. Dataset

Variable	Keterangan
Age	Usia 21 - 81 tahun, dengan usia rata-rata 33 tahun.
Pregnancies	Jumlah kehamilan. Dari 0 - 17, dengan rata-rata 4 kehamilan.

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975

Variable	Keterangan
Glucose	Tingkat konsentrasi glukosa plasma (2 jam setelah makan). Rentangnya adalah 0 hingga 199, dengan rata-rata 121.
Blood Pressure	Tekanan darah diastolik dalam mm Hg. Rentangnya adalah 0 hingga 122, dengan rata-rata 69.
Skin Thickness	Ketebalan kulit triceps dalam mm. Rentangnya adalah 0 hingga 99, dengan rata-rata 21.
Insulin	Rentangnya dari 0 hingga 846, dengan rata-rata 80.
BMI	Indeks massa tubuh dalam Kg/m2. Rentangnya adalah 0 hingga 67.1, dengan rata-rata 32.
Diabetes Pedigree Function	Fungsi ini memberi skor pada kemungkinan diabetes. Rentangnya adalah dari 0.078 hingga 2.42, dengan rata-rata 0.47.
Outcome	Nilainya dapat berupa 0 atau 1. Di sini, 0 berarti bahwa seorang wanita tidak menderita diabetes, dan 1 berarti bahwa seorang wanita menderita diabetes.

Pada tahap selanjutnya dilakukan pemilihan data atau *preprocessing/cleaning* data, pertama melakukan pembersihan data ganda atau bersifat *outlier* pada dataset, kedua mengidentifikasi *missing value* pada dataset, ketiga melihat korelasi fitur atau variable dalam bentuk visualisasi gambar, keempat melihat korelasi atau hubungan antara variable *dependent* terhadap *independent*, kelima menentukan variable *independent* dan variable *dependent*. keenam visualisasi data *histogram variable dependent* untuk melihat jumlah kelas diabetes dan tidak diabetes sebelum dilakukan teknik optimasi dan sesudah dilakukan Teknik optimasi SMOTE dan SMOTE-ENN. Hasil preprocessing disajikan dalam bentuk Tabel dan gambar, dapat dilihat pada Tabel 2, 3, 4, 5 dan Gambar 2, 3, 4, 5 di bawah ini.

Tabel 2. Hasil Penghapusan Data Ganda

	P	G	BP	ST	I	BMI	DPF	A	O
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Pada Table 2 diatas dapat dilihat bahwa dari hasil penghapusan data ganda pada dataset diketahui bahwa dari dataset tersebut tidak memiliki nilai yang ganda.

Tabel 3. Hasil Identifikasi Missing Value

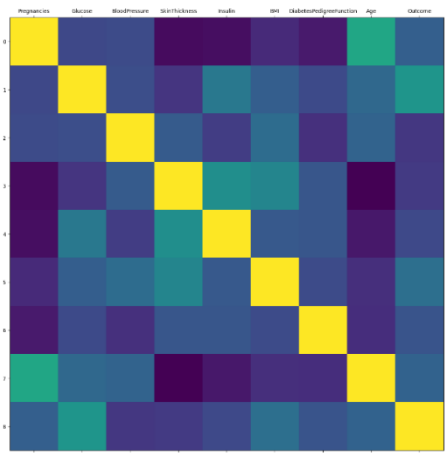
Variable	Nilai
Age	0
Pregnancies	0
Glucose	0
Blood Pressure	0
Skin Thickness	0
Insulin	0
BMI	0
Diabetes Pedigree Function	0
Outcome	0

Pada table 3 di atas dapat dilihat bahwa pada dataset yang digunakan tidak memiliki nilai yang kosong pada setiap baris atau kolom.

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975



Gambar 2. Hasil Korelasi Fitur/Variabel pada Dataset

Pada Gambar 2 dapat dilihat bahwa fitur/variable yang digunakan pada dataset ini cenderung memiliki data yang saling berhubungan antar fitur/variable pada dataset.

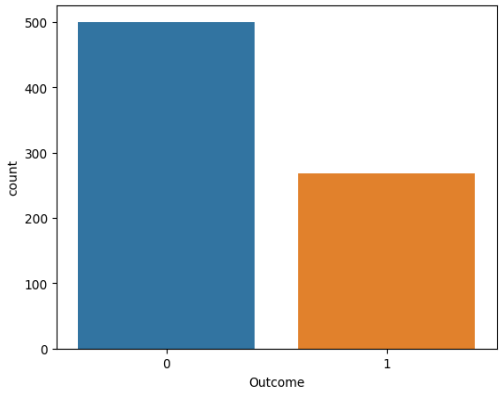
Table 4. Korelasi Variable Dependen Terhadap Variable Independent

Variabel Independen	Variabel Dependen	Korelasi
Age	Outcome	0.22189815303398686
Pregnancies	Outcome	0.4665813983068735
Glucose	Outcome	0.06506835955033277
Blood Pressure	Outcome	0.07475223191831941
Skin Thickness	Outcome	0.13054795488404797
Insulin	Outcome	0.2926946626444454
BMI	Outcome	0.1738440656529599
Diabetes Pedigree Function	Outcome	0.23835598302719757
Outcome	Outcome	0.07475223191831941

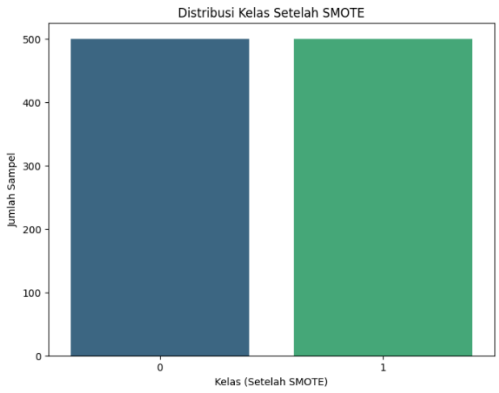
Pada table 4 dapat dilihat bahwa semua variable independen memiliki korelasi atau hubungan yang kuat dengan variable dependen karena memiliki nilai positif, sehingga dapat dipastikan dataset tersebut memiliki pengaruh yang signifikan terhadap variable dependen.

Tabel 5. Hasil Variabel Dependen dan Independen

No	Variabel	Keterangan
1.	Age	Independen
2.	Pregnancies	
3.	Glucose	
4.	Blood Pressure	
5.	Skin Thickness	
6.	Insulin	
7.	BMI	
8.	Diabetes Pedigree Function	
9.	Outcome	Dependen



Gambar 3. Sebelum Optimasi

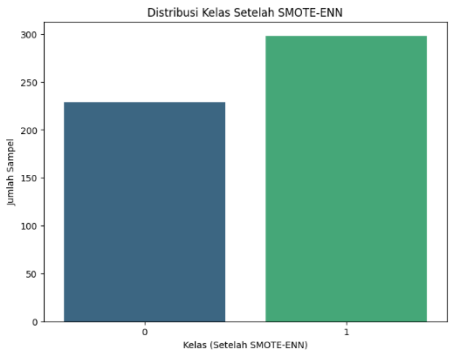


Gambar 4. Setelah Optimasi SMOTE

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975



Gambar 5. Setelah Optimasi SMOTE-ENN

Setelah itu dilakukan transformasi pada dataset karena variable pada dataset sudah bisa dikatakan sudah valid dan dataset tersebut sudah berbentuk angka (numerik) sehingga sudah bisa dilakukan pengolahan langsung ke tahap mining. Selanjutnya pada tahap ini dilakukan teknik data mining menggunakan algoritma KNN dengan menggunakan tiga kali skenario untuk melihat perbandingan antara data asli dengan data yang dilakukan optimasi dengan melakukan lima kali pembagian data *training* dan *testing* pada masing-masing skenario untuk melihat performa terbaik dari tiga skenario yang digunakan. Hasil data mining dapat dilihat pada Tabel 6 di bawah ini.

Tabel 6. Hasil Pengujian Model KNN dengan 3 Skenario

Algoritma	Split Data	Keterangan	Precision	Recall	F1-Score	Support	Akurasi
KNN	90/10	Tidak Diabetes	0.74	0.70	0.72	50	64.94%
		Diabetes	0.50	0.56	0.53	27	
	80/20	Tidak Diabetes	0.75	0.71	0.73	99	66.23%
		Diabetes	0.52	0.58	0.55	55	
	70/30	Tidak Diabetes	0.77	0.75	0.76	151	68.83%
		Diabetes	0.55	0.56	0.56	80	
	60/40	Tidak Diabetes	0.78	0.79	0.78	206	70.78%
		Diabetes	0.56	0.55	0.55	102	
	50/50	Tidak Diabetes	0.78	0.78	0.78	254	70.31%
		Diabetes	0.56	0.56	0.56	130	
KNN SMOTE	90/10	Tidak Diabetes	0.80	0.63	0.71	52	73.00%
		Diabetes	0.68	0.83	0.75	48	
	80/20	Tidak Diabetes	0.77	0.65	0.70	99	73.00%
		Diabetes	0.70	0.81	0.75	101	
	70/30	Tidak Diabetes	0.77	0.64	0.70	149	72.33%
		Diabetes	0.69	0.81	0.75	151	
	60/40	Tidak Diabetes	0.77	0.64	0.70	203	71.75%
		Diabetes	0.68	0.80	0.74	197	
	50/50	Tidak Diabetes	0.81	0.58	0.67	265	70.20%
		Diabetes	0.64	0.84	0.73	235	
KNN SMOTE-ENN	90/10	Tidak Diabetes	1.00	0.89	0.94	27	94.34%
		Diabetes	0.90	1.00	0.95	26	
	80/20	Tidak Diabetes	0.98	0.91	0.94	54	94.34%
		Diabetes	0.91	0.98	0.94	52	
	70/30	Tidak Diabetes	0.97	0.93	0.95	73	95.60%
		Diabetes	0.94	0.98	0.96	86	
	60/40	Tidak Diabetes	0.98	0.91	0.94	100	94.79%
		Diabetes	0.92	0.98	0.95	111	
	50/50	Tidak Diabetes	0.99	0.89	0.94	122	94.32%
		Diabetes	0.91	0.99	0.95	142	

Berdasarkan hasil yang telah disajikan pada Table 6 dapat dilihat hasil pengujian model algoritma KNN dengan menggunakan 3 skenario perhitungan Dimana pada skenario pertama dengan melakukan lima kali pembagian data training dan data testing hasil akurasi yang didapatkan dengan menggunakan dataset asli yang berjumlah 768 data, didapatkan akurasi terbaik pada pembagian data training 60% dan testing 40% mencapai akurasi sebesar 70.78%. Pada skenario kedua dengan menggunakan data yang telah dilakukan optimasi menggunakan Teknik SMOTE atau menyeimbangkan kelas antara kelas diabetes dan tidak diabetes dengan hasil terbaik didapatkan pada pembagian data training 90% testing 10% dan training 80% testing 20% mencapai akurasi sebesar 73.00%.

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975

Pada skenario ketiga dilakukan optimasi pada dataset menggunakan Teknik SMOTE-ENN Dimana Teknik ini memiliki kemampuan untuk menyeimbangkan kelas yang tidak seimbang dan mengurasi overfitting pada kelas yang sudah dilakukan penyeimbangan dan didapatkan hasil terbaik pada pembagian data training 70% dan testing 30% mencapai akurasi sebesar 95.60%.

Pada tahap selanjutnya dilakukan evaluasi dengan menggunakan metode *Confusion Matrix* dan kurva ROC (*Receiver Operating Characteristic*). Nilai performansi yang digunakan yaitu *accuracy klasifikasi, error klasifikasi dan Area Under Curve (AUC)*. Hasil evaluasi dapat dilihat pada Tabel 7 di bawah ini.

Tabel 7. Evaluasi Confusion Matrix dan Kurva ROC

Algoritma	Split Data	Akurasi klasifikasi	Error klasifikasi	AUC
KNN	90/10	0.65	0.35	0.63
	80/20	0.66	0.34	0.64
	70/30	0.69	0.31	0.66
	60/40	0.71	0.29	0.67
	50/50	0.70	0.30	0.67
KNN SMOTE	90/10	0.73	0.27	0.73
	80/20	0.73	0.27	0.73
	70/30	0.72	0.28	0.72
	60/40	0.72	0.28	0.72
	50/50	0.70	0.30	0.71
KNN SMOTE-ENN	90/10	0.94	0.06	0.94
	80/20	0.94	0.06	0.94
	70/30	0.96	0.04	0.95
	60/40	0.95	0.05	0.95
	50/50	0.94	0.06	0.94

Berdasarkan hasil yang telah disajikan pada Table 7 dapat dilihat bahwa hasil evaluasi kinerja model algoritma KNN menggunakan *confusion matrik* dan Kurva ROC dengan menggunakan 3 skenario, Dimana pada skenario pertama hasil evaluasi terbaik didapatkan pada pembagian data training 60% dan testing 40% mencapai akurasi klasifikasi 0.71, error klasifikasi 0.29 dan AUC sebesar 0.67. Pada skenario kedua hasil terbaik didapatkan pada pembagian data training 90%, testing 10% dan Training 80%, testing 20% mencapai akurasi klasifikasi 0.73, error klasifikasi 0.27 dan AUC sebesar 0.73. Pada skenario ketiga hasil terbaik didapatkan pada pembagian data training 70% dan testing 30% mencapai akurasi klasifikasi 0.96, error klasifikasi 0.04, dan AUC sebesar 0.95.

4.2. Pembahasan

Dataset yang digunakan pada penelitian ini berjumlah 768 Data dengan 9 variabel diantaranya, age, pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, dan outcome. dari hasil preprocessing/cleaning data tersebut tidak ditemukan data ganda dan nilai yang kosong atau *missing value* pada dataset. Selanjutnya hasil korelasi fitur/variable cenderung memiliki data yang erat atau saling berhubungan satu sama lain antara variable independent dan variable dependen sehingga memiliki pengaruh yang signifikan dalam melakukan prediksi. hasil pengujian model dan evaluasi kinerja model algoritma KNN dengan menggunakan 3 skenario

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975

dimana pada scenario pertama perhitungan tanpa melakukan optimasi pada dataset mencapai akurasi sebesar 70.78% akurasi klasifikasi 0.71, error klasifikasi 0.29 dan AUC sebesar 0.67. hal ini menunjukkan bahwa tingkat resiko penyakit diabetes pada dataset yang tidak seimbang adalah tidak diabetes karena support data yang digunakan lebih dominan atau lebih banyak data yang mengarah ke kelas tidak diabetes dibandingkan dengan kelas yang diabetes. hasil ini memprediksi secara keseluruhan pada data kelas yang tidak seimbang yaitu kelas diabetes dan tidak diabetes dengan prediksi yang cenderung tidak terkena diabetes dan termasuk dalam katagori cukup baik. Pada skenario kedua perhitingan dengan melakukan optimasi menggunakan teknik smote mencapai akurasi sebesar 73.00% akurasi klasifikasi 0.73, error klasifikasi 0.27 dan AUC sebesar 0.73. Hal ini menunjukkan bahwa Tingkat resiko penyakit diabetes pada data ini adalah diabetes karena support data yang digunakan lebih dominan atau lebih banyak data yang mengarah ke kelas diabetes dibandingkan dengan data kelas tidak diabetes. hasil ini memprediksi secara keseluruhan pada data kelas yang seimbang setelah dilakukan optimasi SMOTE pada dataset yaitu kelas diabetes dan tidak diabetes dengan prediksi yang cenderung terkena diabetes dan termasuk dalam katagori baik. Pada skenario ketiga mencapai akurasi sebesar 95.60% akurasi klasifikasi 0.96, error klasifikasi 0.04, dan AUC sebesar 0.95. hal ini menunjukkan bahwa Tingkat resiko penyakit diabetes pada data ini adalah diabetes karena support data yang digunakan lebih dominan atau lebih banyak data yang mengarah ke kelas diabetes dibandingkan dengan data kelas tidak diabetes. hasil ini memprediksi secara keseluruhan pada data kelas yang tidak seimbang setelah dilakukan optimasi SMOTE-ENN pada dataset yaitu kelas diabetes dan tidak diabetes dengan prediksi prediksi yang cenderung terkena diabetes dan termasuk dalam katagori sangat baik. Berdasarkan hasil tiga skenario di atas menunjukkan bahwa dengan melakukan optimasi pada dataset menggunakan teknik SMOTE-ENN dapat meningkatkan hasil akurasi klasifikasi yang sangat baik dalam melakukan prediksi.

5. Kesimpulan

Dalam penelitian ini, peneliti telah memprediksi diabetes pada wanita indians dengan menggunakan dataset diabetes Pima Indians. Dalam dataset ini, peneliti mengambil variabel kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, IMT (Indeks Massa Tubuh), diabetes. Fungsi pedigree, usia, dan atribut hasil digunakan untuk melakukan prediksi. Berbagai algoritma pembelajaran terawasi telah digunakan seperti CT, SVM, K-NN, NB, RF, NN, AB, dan LR, dan menghasilkan akurasi yang masih rendah dengan

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975

menerapkan optimasi data dengan Teknik SMOTE, dibandingkan dengan hasil penelitian ini dengan parameter akurasi klasifikasi, error klasifikasi dan kurva ROC dengan menerapkan optimasi pada dataset penelitian dengan Teknik SMOTE-ENN dapat meningkatkan hasil akurasi prediksi pada model algoritma KNN dengan menunjukkan hasil prediksi pada dataset yaitu terkena diabetes, yang dipengaruhi oleh Umur, kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, IMT (Indeks Massa Tubuh), dan fungsi pewarisan diabetes. Sehingga untuk mengurangi resiko penyakit diabetes perlu mengurangi konsumsi gula yang berlebihan untuk mencegah tekanan darah naik apalagi bagi para Wanita yang sudah di atas usia 33 tahun

6. Daftar Pustaka

[1] R. Zolfaghari, “Diagnosis Of Diabetes In Female Population Of Pima Indian Heritage With Ensemble Of Bp Neural Network And Svm,” Int. J. Comput. Eng. Manag, Vol. 15, No. 4, Pp. 2230–7893, 2012.

[2] H. N. A. Pham And E. Triantaphyllou, “Prediction Of Diabetes By Employing A New Data Mining Approach Which Balances Fitting And Generalization,” In Computer And Information Science, Springer, 2008, Pp. 11–26.

[3] J. Wu, Y.-B. Diao, M.-L. Li, Y.-P. Fang, And D.-C. Ma, “A Semi-Supervised Learning Based Method: Laplacian Support Vector Machine Used In Diabetes Disease Diagnosis,” Interdiscip. Sci. Comput. Life Sci., Vol. 1, Pp. 151–155, 2009.

[4] S. K. Dey, A. Hossain, And M. M. Rahman, “Implementation Of A Web Application To Predict Diabetes Disease: An Approach Using Machine Learning Algorithm,” In 2018 21st International Conference Of Computer And Information Technology (Iccit), 2018, Pp. 1–5.

[5] S. Nithya, M. Sangeetha, K. N. A. Prethi, K. S. Sahoo, S. K. Panda, And A. H. Gandomi, “Sdcf: A Software-Defined Cyber Foraging Framework For Cloudlet Environment,” Ieee Trans. Netw. Serv. Manag., Vol. 17, No. 4, Pp. 2423–2435, 2020.

[6] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, And H. Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques,” Front. Genet., Vol. 9, P. 515, 2018.

[7] S. Srivastava, L. Sharma, V. Sharma, A. Kumar, And H. Darbari, “Prediction Of Diabetes Using Artificial Neural Network Approach,” In Engineering Vibration, Communication And Information Processing: Icoevci 2018, India, 2019, Pp. 679–687.

[8] V. Karthikeyani And I. P. Begum, “Comparison A Performance Of Data Mining Algorithms (Cpdma) In Prediction Of Diabetes Disease,” Int. J. Comput. Sci. Eng., Vol. 5, No. 3, P. 205, 2013.

[9] D. Martens, J. Huysmans, R. Setiono, J. Vanthienen, And B. Baesens, “Rule Extraction From Support Vector Machines: An Overview Of Issues And Application In Credit Scoring,” Rule Extr. From Support Vector Mach., Pp. 33–63, 2008.

[10] Y. Guo, G. Bai, And Y. Hu, “Using Bayes Network For Prediction Of Type-2 Diabetes,” In 2012 International Conference For Internet Technology And Secured Transactions, 2012, Pp. 471–472.

[11] L. O. Schulz Et Al., “Effects Of Traditional And Western Environments On Prevalence Of Type 2 Diabetes In Pima Indians In Mexico And The Us,” Diabetes Care, Vol. 29, No. 8, Pp. 1866–1871, 2006.

[12] M. Maniruzzaman, M. J. Rahman, B. Ahammed, And M. M. Abedin, “Classification And Prediction Of Diabetes

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal.193-204

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.27975 Link : https://dx.doi.org/10.29408/jit.v8i1.27975

Disease Using Machine Learning Paradigm," Heal. Inf. Sci. Syst., Vol. 8, Pp. 1–14, 2020.

[13] J. Han, J. C. Rodriguez, And M. Beheshti, "Diabetes Data Analysis And Prediction Model Discovery Using Rapidminer," In 2008 Second International Conference On Future Generation Communication And Networking, 2008, Vol. 3, Pp. 96–99.

[14] S. A. Saji And K. Balachandran, "Performance Analysis Of Training Algorithms Of Multilayer Perceptrons In Diabetes Prediction," In 2015 International Conference On Advances In Computer Engineering And Applications, 2015, Pp. 201–206.

[15] M. Qusyairi, "Analisi Prediksi Tingkat Kesejahteraan Masyarakat Nelayan Lombok Timur Dengan Algoritma Naïve Bayes," Infotek J. Inform. Dan Teknol., Vol. 7, No. 2, Pp. 563–574, 2024.

[16] Suyanto, Data Mining Untuk Klasifikasi Dan Klasterisasi Data. Bandung: Bandung: Informatika, 2017.

[17] M. Saiful, H. Bahtiar, And M. T. Hidayat, "Penerapan Algoritma K-Means Clustering Dalam Mengelompokkan Smartphone Yang Rekomendasi Berdasarkan Spesifikasi," Infotek J. Inform. Dan Teknol., Vol. 7, No. 2, Pp. 478–488, 2024.

[18] Z. Amri, K. Kusrini, And K. Kusnawi, "Prediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Naïve Bayes, Decision Tree, Ann, Knn, Dan Svm," Edumatic J. Pendidik. Inform., Vol. 7, No. 2, Pp. 187–196, 2023.

[19] S. K. Bhoi, "Prediction Of Diabetes In Females Of Pima Indian Heritage: A Complete Supervised Learning Approach," Turkish J. Comput. Math. Educ., Vol. 12, No. 10, Pp. 3074–3084, 2021.

[20] A. Perdana, A. Hermawan, And D. Avianto, "Analyze Important Features Of Pima Indian Database For Diabetes Prediction Using Knn," J. Sisfokom (Sistem Inf. Dan Komputer), Vol. 12, No. 1, Pp. 70–75, 2023.

[21] M. Benarbia, "A Machine Learning Approach To Predicting The Onset Of Type Ii Diabetes In A Sample Of Pima Indian Women," 2022.

[22] M. Abedini, A. Bijari, And T. Banirostam, "Classification Of Pima Indian Diabetes Dataset Using Ensemble Of Decision Tree, Logistic Regression And Neural Network," Int. J. Adv. Res. Comput. Commun. Eng, Vol. 9, No. 7, Pp. 7–10, 2020.

[23] V. Chang, J. Bailey, Q. A. Xu, And Z. Sun, "Pima Indians Diabetes Mellitus Classification Based On Machine Learning (Ml) Algorithms," Neural Comput. Appl., Vol. 35, No. 22, Pp. 16157–16173, 2023.

[24] U. C. I. M. Learning, "Pima Indians Diabetes Database," Kaggle. Com/Uciml/Pima-Indians-Diabetes-*Database*, 2016.

[25] D. Nofriansyah And G. W. Nurcahyo, *Algoritma Data Mining Dan Pengujian*. Deepublish, 2015