PERBANDINGAN AKURASI ALGORITMA *K-NEAREST NEIGHBOR* DAN *LOGISTIC REGRESSION* UNTUK KLASIFIKASI PENYAKIT DIABETES

Raharjo Putra Kurniadi[1], Rd. Rohmat Saedudin[2], Vandha Pradwiyasma Widartha[3]

[1,2,3] Universitas Telkom, Bandung
putrakurniadi@student.telkomuniversity.ac.id[1], rdrohmat@telkomuniversity.ac.id[2], vandhapw@telkomuniversity.ac.id[3]

Abstrak

Diabetes atau sering disebut sebagai penyakit kencing manis merupakan suatu penyakit akibat kelainan metabolik yang diakibatkan oleh tingginya kadar glukosa darah di tubuh dalam waktu yang lama. *International Diabetes Federation* (IDF) memperkirakan sedikitnya terdapat 463 juta jiwa di seluruh dunia menderita penyakit diabetes pada tahun 2019. Negara Indonesia berada di urutan ke-7 dari 10 negara dengan jumlah penderita diabetes terbanyak, yaitu sebesar 10,7 juta dan diprediksi akan berjumlah 16,6 juta jiwa pada tahun 2045. Banyak orang terdiagnosis penyakit diabetes setelah mengalami komplikasi. Pendeteksian penyakit dapat dilakukan dengan menggunakan *data mining* dalam menggali informasi dari kumpulan data penyakit diabetes. D*ataset* yang digunakan pada penelitian ini adalah *dataset Pima Indians Diabetes Database*. *Dataset* ini berisikan 768 pasien wanita dengan 8 atribut diagnosa kondisi medis yang berbeda dan 1 atribut tujuan atau atribut label. Penelitian ini membandingkan algoritma *K-Nearest Neighbor* dan *Logistic Regression* untuk klasifikasi data *Pima Indians Diabetes Database*. Pada penelitian ini, penulis melakukan penanganan *missing value* terhadap data dan menggunakan metode *Grid Search* untuk menemukan model dengan hasil akurasi yang optimal. Hasil akurasi dievaluasi dengan menggunakan *confusion matrix* dan menghitung nilai AUC. Diperoleh hasil algoritma *K-Nearest Neighbor* dengan nilai akurasi sebesar 85,06% dan algoritma *Logistic Regression* dengan akurasi sebesar 77,92%.

Kata Kunci : ***diabetes, data mining, klasifikasi, k-nearest neighbor, logistic regression***

Abstract

Diabetes or often referred to as diabetes is a disease due to metabolic disorders caused by high blood glucose levels in the body for a long time. The International Diabetes Federation (IDF) estimates that at least 463 million people worldwide suffer from diabetes in 2019. Indonesia is ranked 7th out of 10 countries with the highest number of people with diabetes, which is 10.7 million, and is predicted to number 16. ,7 million people in 2045. Many people are diagnosed with diabetes after experiencing complications. Disease detection can be done by using data mining in extracting information from the diabetes data set. The dataset used in this study is the Pima Indians Diabetes Database dataset. This dataset contains 768 female patients with 8 different medical condition diagnostic attributes and 1 goal attribute or label attribute. This study compares the K-Nearest Neighbor and Logistic Regression algorithms for data classification of the Pima Indians Diabetes Database. In this study, the authors handle missing values on the data and use the Grid Search method to find models with optimal accuracy results. Accuracy results were evaluated by using a confusion matrix and calculating the AUC value. The results of the K-Nearest Neighbor algorithm with an accuracy value of 85.06% and the Logistic Regression algorithm with an accuracy of 77.92% are obtained.

Keywords: diabetes, data mining, classification, k-nearest neighbor, logistic regression

1. Pendahuluan

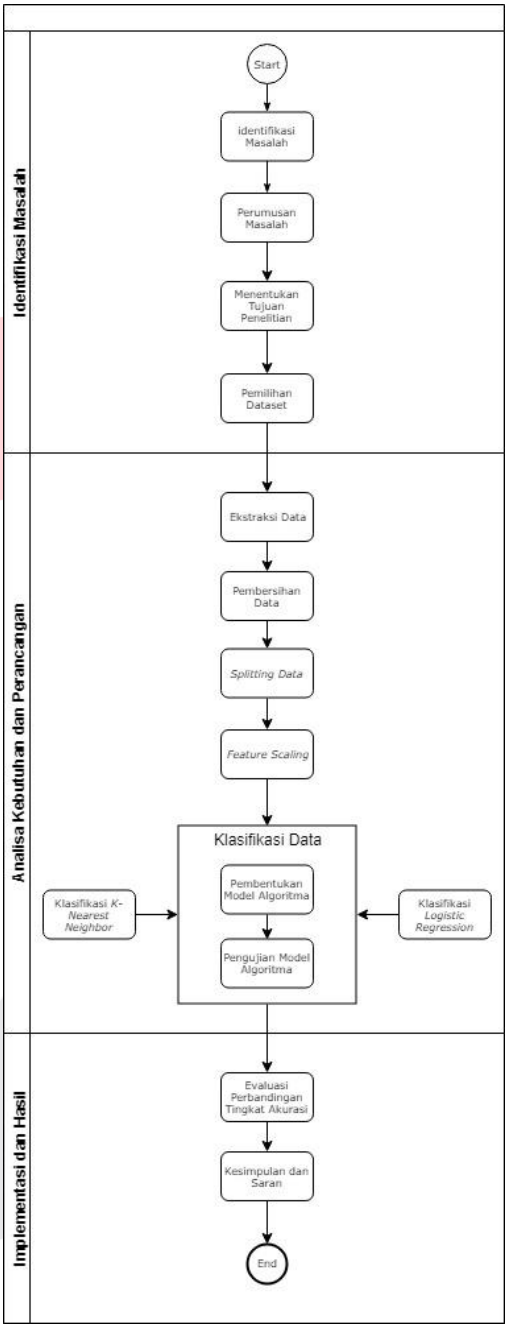
Diabetes atau sering disebut sebagai penyakit kencing manis merupakan suatu penyakit akibat kelainan metabolik yang diakibatkan oleh tingginya kadar glukosa darah di tubuh dalam waktu yang lama. Jika tidak ditangani lebih awal, diabetes bisa menyebabkan terjadinya komplikasi terhadap penyakit lain [1]. Berdasarkan rilisan Infodatin Diabetes Melitus 2020 Pusat Data dan Informasi Kementrian Kesehatan RI, *International Diabetes Federation* (IDF) memperkirakan sedikitnya terdapat 463 juta jiwa di seluruh dunia pada usia rentang 20-79 tahun menderita penyakit diabetes pada tahun 2019. Angka prevalensi tersebut diprediksi akan terus

mengalami peningkatan hingga mencapai 578 juta jiwa pada tahun 2030 dan 700 juta jiwa pada tahun 2045 [2]. Negara Indonesia sendiri berada pada posisi ke-7 dari 10 negara di dunia dengan jumlah penderita diabetes terbanyak, dengan jumlah 10,7 juta jiwa. Jumlah tersebut diprediksi akan terus mengalami kenaikan dengan jumlah 13,7 juta jiwa pada tahun 2030 dan 16,6 juta jiwa pada tahun 2045 [3]. Banyak orang terdiagnosis penyakit diabetes setelah mengalami komplikasi. Padahal, jika diagnosis dilakukan secara dini, penanganan diabetes dapat dilakukan lebih cepat dan dapat menghindari komplikasi penyakit lain yang berbahaya. Hal tersebut dapat didukung dengan penggunaan teknik *data mining* dalam menggali informasi berharga dari kumpulan data penyakit diabetes [4]. Salah satu teknik *data mining* adalah tipe pembelajaran *supervised learning* [5]. Salah satu jenis dari *supervised learning* adalah klasifikasi. Klasifikasi digunakan untuk menentukan keputusan sesuai dengan pola baru yang didapat dari pola data lama dengan menggunakan perhitungan algoritma [6]. Pada penelitian ini, penulis menggunakan algoritma *K-Nearest Neighbor* dan *Logistic Regression*.

KNN adalah suatu algoritma dengan pendekatan untuk menghitung kedekatan jarak antara kasus baru dengan kasus yang lama berdasarkan pada bobot dari yang sudah ada [7]. *Logistic Regression* adalah memperkirakan peluang log dari suatu peristiwa [8]. Tujuan dari penelitian ini adalah mengukur performa terbaik untuk klasifikasi *dataset* yang dihasilkan dengan membandingkan akurasi dari algoritma *K-Nearest Neighbor* dan *Logistic Regression*. *Dataset* yang digunakan pada penelitian ini adalah *dataset* yang berasal *dari National Institute of Diabetes and Digestive and Kidney Diseases* dan dapat diakses di UCI *Machine Learning Repository: Pima Indians Diabetes Database*. *Dataset* ini berisikan 768 pasien wanita dengan 8 atribut diagnosa kondisi medis yang berbeda dan 1 atribut tujuan atau atribut label [9].

2. Metode Penelitian

Pada penelitian ini, penulis menetapkan tiga tahapan umum sebagai prosedur penelitian. Tahapan-tahapan tersebut adalah tahap identifikasi masalah, tahap analisa perancangan dan kebutuhan, dan tahap implementasi dan hasil. Berikut adalah tahapan yang ditempuh yang dapat dilihat pada Gambar 1.



Gambar 1 (Sistematika Penyelesaian Masalah)

A. Identifikasi Masalah

Pada tahap awal ini dilakukan identifikasi masalah sesuai dengan studi kasus yang ada. Selanjutnya penulis menentukan rumusan masalah yang tepat berdasarkan masalah yang ada. Proses selanjutnya ialah menentukan tujuan dari penelitian

berdasarkan rumusan masalah yang sudah ditentukan. Proses terakhir adalah pemilihan *dataset* yang akan digunakan pada penelitian ini.

B. Analisa Kebutuhan dan Perancangan

Pada tahap ini, dilakukan proses ekstraksi data yang digunakan. Data tersebut akan melalui tahapan *data preprocessing*. *Data preprocessing* adalah proses yang dilakukan untuk membuat data mentah menjadi data yang berkualitas [10]. Diawali dengan melakukan pembersihan data untuk mengganti nilai *null (missing value)* dengan nilai *median* dari tiap kolom pada kelasnya masing-masing [11], kemudian melakukan *splitting data* dengan tujuan membagi data menjadi data *training* dan data *testing* dan melakukan proses *feature scaling* untuk penyamarataan skala dari data.

C. Implementasi dan hasil

Dalam implementasi algoritma *K-Nearest Neighbor* dan *Logistic Regression*, akan dilakukan proses pencarian nilai akurasi terbaik dari masing-masing algoritma, dengan melakukan beberapa pengujian seperti penggunaan 3 rasio berbeda, pengujian dengan menggunakan K-*Fold Cross Validation* dengan *fold* yang digunakan sebesar 10 *fold*. 10 *fold cross validation* merupakan jumlah *fold* dari *K-fold cross validation* yang direkomendasikan untuk pemilihan model yang terbaik karena memberikan estimasi hasil akurasi yang kurang bias [12]. Selanjutnya melakukan *Tunning Hyperparameter* dengan penentuan parameter yang akan digunakan untuk mendapatkan hasil akurasi terbaik dengan menggunakan metode *Grid Search* [13]. Setelah mendapatkan akurasi terbaik dari masing-masing algoritma, dilakukan tahapan perbandingan hasil akurasi dan evaluasi model menggunakan *confusion matrix*. *Confusion matrix* berisi informasi tentang kelas aktual *(actual*) dan prediksi (*predicted*) pada sistem klasifikasi. Perhitungan ini ditabulasikan ke dalam tabel yang disebut sebagai *confusion matrix* [4]. Selanjutnya dilakukan perhitungan nilai *precision, recall* dan *F1-Score* serta nilai *Area Under the Curve* (AUC). AUC mengukur dengan cara memperkirakan probabilitas nilai *outpu*t dari sampel yang dipilih acak dari populasi positif ataupun negatif, semakin besar nilai dari AUC, maka semakin kuat klasifikasinya [14].

3. Hasil dan Pembahasan

Dataset yang digunakan pada penelitian ini adalah *dataset Pima Indians Diabetes Database* yang berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases* dan diakses melalui situs atapdata.ai. Isi data yang diperoleh dapat dilihat pada Gambar 2.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
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

768 rows × 9 columns

Gambar 2 Dataset

Dataset Pima Indians Diabetes Database memiliki jumlah data sebanyak 768 baris dan 9 kolom di antaranya 8 kolom *feature* yang terdiri dari kolom *Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction*, dan *Age* serta 1 kolom target yang berisi kolom *Outcome*. Untuk deskripsi tiap kolom dari *dataset* dapat dilihat pada Tabel 1.

Tabel 1 Deskripsi Kolom Dataset

Atribut	Keterangan / Deskripsi	Satuan	Tipe Data
Pregnancies	Banyaknya kehamilan	-	Numerik
Glucose	Konsentrasi glukosa plasma	Mg/dL	Numerik
BloodPressure	Tekanan darah Diastolik	Mm Hg	Numerik
SkinThickness	Ketebalan lipatan kulit	Mm	Numerik
Insulin	Insulin	Mm U/ml	Numerik
BMI	Indeks massa tubuh	Kg/m2	Numerik
DiabetesPedigreeFunction	Riwayat diabetes dalam keluarga	-	Numerik
Age	Umur	Tahun	Numerik
Outcome	Positif Diabetes (1) dan negative Diabetes (0)	-	Numerik

Pada tahap selanjutnya, penulis melakukan tahapan *data preprocessing* agar data tersebut dapat digunakan secara tepat dan optimal dalam proses pembelajaran *machine learning*. Tahapan yang dilakukan dimulai dari *data cleansing* dengan mencari apakah ada data yang bernilai *null (missing value)* dari data yang digunakan. Setelah itu mencari nilai *median* dari setiap kolom untuk diisi ke dalam data yang bernilai *null*. Hasil dari perhitungan nilai median dapat dilihat pada Tabel 2.

Tabel 2 (Nilai *Median* Tiap Kolom)

Outcome	*Insulin*	*Glucose*	*SkinThickness*	*BloodPressure*	*BMI*
0	102,5	107	27	70	30,1
1	169,5	140	32	74,5	34,3

Selanjutnya memisahkan data variabel X dan y serta melakukan *splitting data*. Tahapan terakhir yang dilakukan adalah melakukan *feature scaling* untuk menyetarakan skala atau rentang nilai dari data.

Pengujian awal hasil akurasi yang dilakukan oleh penulis dari algoritma KNN dan *Logistic Regression* yang diperoleh berdasarkan 3 perbandingan rasio pembagian data yang berbeda dengan rasio pembagian dan hasil akurasi yang dapat dilihat pada Tabel 3 dan Tabel 4.

Tabel 3 (Perbandingan Hasil Akurasi KNN)

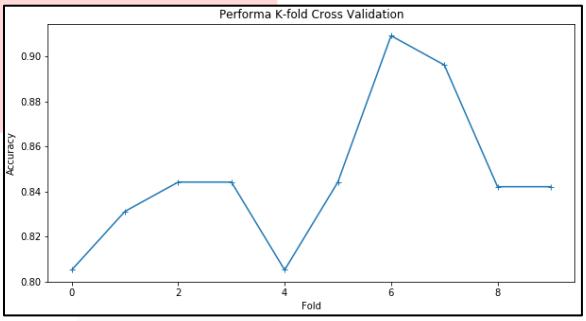
Rasio	Akurasi KNN
70% (*training*) dan 30% (*testing*)	80.51%
75% (*training*) dan 25% (*testing*)	81.25%
80% (*training*) dan 20% (*testing*)	80.51%

Tabel 4 (Perbandingan Hasil Akurasi *Logistic Regression*)

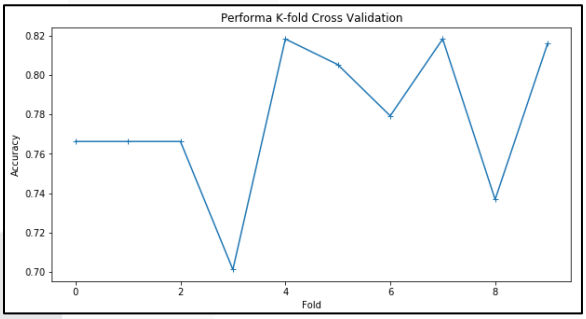
Rasio	Akurasi *Logistic Regression*
70% (*training*) dan 30% (*testing*)	76.62%
75% (*training*) dan 25% (*testing*)	75%
80% (*training*) dan 20% (*testing*)	77.92%

Pada Tabel 3 dan Tabel 4 dapat dilihat bahwa hasil akurasi terbaik dari algoritma KNN diperoleh dari rasio 75:25 yang menghasilkan nilai akurasi sebesar 81.25%. Untuk rasio 70:30 dan rasio 80:20 sebesar 80,51%. Hasil akurasi terbaik dari pengujian awal algoritma *Logistic Regression* diperoleh dari rasio 80:20 yang menghasilkan nilai akurasi sebesar 77.92%. Nilai akurasi yang diperoleh dari rasio 70:30 adalah sebesar 76.62% dan nilai akurasi dari rasio 75:25 menghasilkan akurasi sebesar 75%.

Dalam memvalidasi hasil pengujian klasifikasi terhadap *dataset*, penulis juga menggunakan *K-Fold Cross Validation* dengan jumlah *fold* sebanyak 10. Hasil dari masing-masing *fold* akan dihitung nilai rata-ratanya sebagai hasil akurasi dari pengujian. Grafik performa *K-Fold Cross Validation* untuk pengujian ini dapat dilihat pada Gambar 3 dan Gambar 4.



Gambar 3 (Grafik *K-Fold Cross Validation* dari KNN)



Gambar 4 (Grafik *K-Fold Cross Validation* dari *Logistic Regression*)

Dari pengujian hasil *K-Fold Cross Validation* dapat dilihat hasil akurasi yang didapatkan dari 10 *fold* yang digunakan serta grafik nilai pengujian tiap *foldnya*. Berdasarkan hasil akurasi 10 *fold* dalam *K-Fold Cross Validation* diperoleh nilai rata-rata skor dari algoritma KNN sebesar 84,6% dan untuk algoritma *Logistic Regression* sebesar 77,73%.

Untuk mendapatkan nilai akurasi yang lebih tinggi dari pengujian awal, penulis melakukan *tunning hyperparameter* dari model yang sudah dibuat dengan menggunakan *Grid Search Cross Validation*. Pengujian juga dilakukan terhadap 3 rasio berbeda seperti yang dilakukan pada pengujian awal. Dalam

model ini juga dilakukan tahap validasi pengujian hanya pada data *X_train* menggunakan *K-Fold Cross Validation* sebanyak 10 *fold*. Pada algoritma KNN, parameter yang digunakan pada model *Grid Search CV* adalah nilai K dari rentang nilai 1 hingga 49, penggunaan "*weights*" yaitu *uniform* dan *distance*, serta penggunaan rumus p=1 dan p=2.
Pada algoritma *Logistic Regression*, parameter yang digunakan pada model *Grid Search CV* adalah penggunaan *penalty* l1 dan l2, nilai C dengan np.logspace(-4, 4, 20), dan penggunaan "*solver*" yaitu *lbfgs* dan *liblinear*.
Hasil dari pengujian kedua algoritma dengan menggunakan *Grid Search CV* dapat dilihat pada Tabel 5 dan Tabel 6.

Tabel 5 (Perbandingan Akurasi setalah *Tunning* dari KNN)

Rasio	Akurasi KNN
70% (*training*) dan 30% (*testing*)	84.84%
75% (*training*) dan 25% (*testing*)	82.81%
80% (*training*) dan 20% (*testing*)	85.06%

Tabel 6 (Perbandingan Akurasi setelah *Tunning* dari *Logistic Regression*)

Rasio	Akurasi *Logistic Regression*
70% (*training*) dan 30% (*testing*)	76.62%
75% (*training*) dan 25% (*testing*)	75%
80% (*training*) dan 20% (*testing*)	77.92%

Dari hasil pengujian didapatkan bahwa model dengan nilai akurasi terbaik untuk algoritma KNN didapat dari rasio pembagian data 80:20 dengan menggunakan parameter nilai k=23, menggunakan rumus p=1, dan *weights* yang digunakan adalah *uniform* dengan nilai akurasi sebesar 85,06%.
Model dengan nilai akurasi terbaik untuk Algoritma *Logistic Regression* didapat dari rasio pembagian data 80:20 dengan menggunakan parameter nilai C=0.03, *penalty*=l2, dan *solver* yang digunakan adalah *lbfgs* dengan nilai akurasi sebesar 77,92%.

Berdasarkan hasil akurasi dari pengujian model yang dibuat, penulis mengukur performa model menggunakan *confusion matrix* dengan hasil yang dapat dilihat pada Tabel 7 dan Tabel 8.

Tabel 7 (Confusion matrix KNN)

	***Predicted Healthy* (0)**	***Predicted Diabetic* (1)**
***Actual Healthy* (0)**	90 TN	10 FP
***Actual Diabetic* (1)**	13 FN	41 TP

Tabel 8 (Confusion matrix *Logistic Regression*)

	***Predicted Healthy* (0)**	***Predicted Diabetic* (1)**
***Actual Healthy* (0)**	92 TN	8 FP
***Actual Diabetic* (1)**	26 FN	28 TP

a) Hasil pengujian algoritma KNN menghasilkan nilai *True Negative* (TN) dengan jumlah 90, nilai *False Negative* (FN) dengan jumlah 13, nilai *False Positive* (FP) dengan jumlah 10, dan nilai *True Positive* (TP) dengan jumlah 41.
b) Hasil pengujian algoritma *Logistic Regression* menghasilkan nilai *True Negative* (TN) dengan jumlah 92, nilai *False Negative* (FN) dengan jumlah 26, nilai *False Positive* (FP) dengan jumlah 8, dan nilai *True Positive* (TP) dengan jumlah 28.

Dari hasil *confusion matrix* yang didapat, penulis menghitung nilai akurasi dengan persamaan berikut [12]:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

Ukuran yang dapat digunakan dalam mengevaluasi pengklasifikasian adalah p*recision* dan r*ecall. Precision* adalah sebuah ukuran ketepatan dari proses pengklasifikasian atau proporsi klasifikasi positif hasil dari prediksi yang benar terhadap seluruh hasil prediksi yang bernilai positif. Persamaan dari *precision* adalah sebagai berikut:

$$Precission = \frac{TP}{TP+FP} \quad (2)$$

Recall adalah ukuran *completeness* dari proses klasifikasi atau proporsi klasifikasi positif hasil dari prediksi yang benar terhadap seluruh kelas aktual bernilai positif. Persamaan dari *recall* adalah sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Untuk menghitung kombinasi antara *precision* dan *recall* digunakan *F1-Score* [15]. Persamaan dari *F1-Score* adalah sebagai berikut [12]:

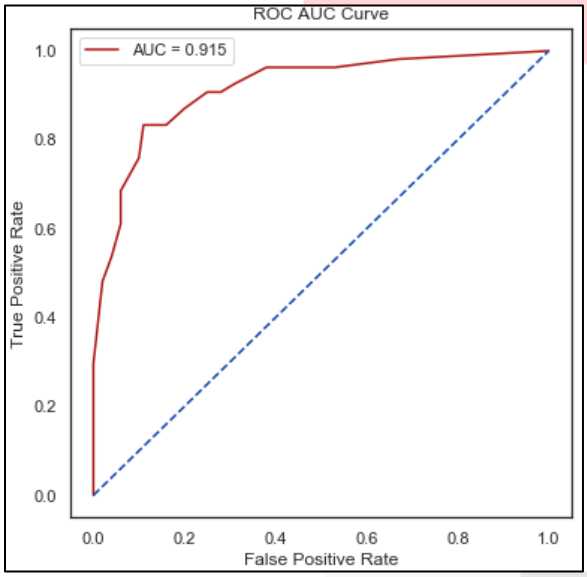
$$F1 - Score = \frac{2(Recall*Precission)}{Recall+Precission} \quad (4)$$

Hasil dari perhitungan *confusion matrix* dari algoritma KNN dan *Logistic Regression* dapat dilihat pada Tabel 9.

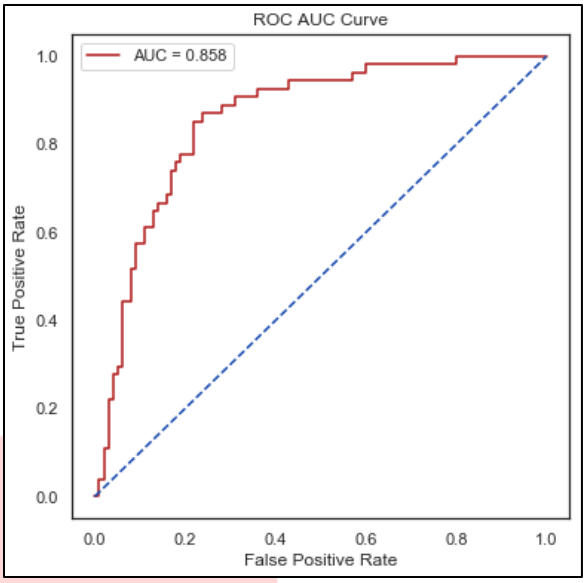
Tabel 9 (Perbandingan Hasil confusion mtarix)

	Algoritma	
	KNN	***Logistic Regression***
Accuracy	85,06%	77.92%
Precision	0,8	0,78
Recall	0,76	0,52
F1-Score	0,78	0,62

Setelah mengukur kinerja suatu model dengan *confusion matrix*, pengukuran dapat juga dilakukan dengan menampilkan informasi kinerja algoritma klasifikasi dalam bentuk grafik dengan menggunakan kurva *Receiver Operating Characteristic* (ROC) dan perhitungan skor *Area Under the Curve* (AUC). Grafik dari kurva ROC dan hasil skor AUC pada algoritma KNN dapat dilihat pada Gambar 5 dan Gambar 6.



Gambar 5 (Kurva ROC dari KNN)



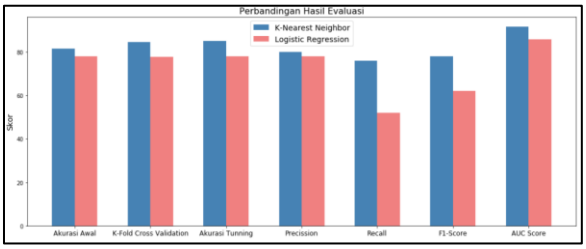
Gambar 6 (Kurva ROC dari *Logistic Regression*)

Kurva dari ROC yang dihasilkan dari algoritma KNN menampilkan grafik yang dapat digunakan untuk menghitung skor dari AUC dengan hasil skor 0,915 sedangkan algoritma *Logistic Regression* mendapat skor 0,858.

Dari hasil pengujian hasil akurasi dan evaluasi model pada algoritma *K-Nearest Neighbor* dan *Logistic Regression*, penulis membandingkan performa dari masing-masing algoritma dan didapatkan hasil seperti pada Tabel 10 serta grafik perbandingan hasil evaluasi dalam bentuk *Bar Chart* dapat dilihat pada Gambar 7.

Tabel 10 (Perbandingan Hasil Evaluasi)

	Algoritma	
	K-Nearest Neighbor	***Logistic Regression***
Akurasi Awal	81,52%	77.92%
K-Fold Cross Validation	84,6%	77,73%
Akurasi *Tunning*	85,06%	77.92%
Precision	0,8	0,78
Recall	0,76	0,52
F1-Score	0,78	0,62
AUC *Score*	0,915	0,858



Gambar 7 (Grafik Perbandingan Hasil Evaluasi)

Berdasarkan hasil perbandingan yang sudah dibuat, dapat dijelaskan bahwa hasil algoritma KNN menghasilkan nilai akurasi awal sebesar 81,52% dengan rasio *splitting data* 75:25, *K-Fold Cross Validation* dengan rata-rata 84,6%, nilai akurasi setelah *tunning* dengan rasio *splitting data* 80:20 sebesar 85,06% dengan hasil *Precision* bernilai 0,8, *Recall* bernilai 0,76, *F1-Score* bernilai 0,78, dan skor AUC sebesar 0,915. Algoritma *Logistic Regression* menghasilkan nilai akurasi awal sebesar 77,92% dengan rasio *splitting data* 80:20, *K-Fold Cross Validation* dengan rata-rata 77,73%, nilai akurasi setelah *tunning* dengan rasio *splitting data* 80:20 sebesar 77,92% dengan hasil *Precision* bernilai 0,78, *Recall* bernilai 0,52, *F1-Score* bernilai 0,62, dan skor AUC sebesar 0,858. Dengan hasil tersebut, dalam melakukan klasifikasi terhadap *dataset* penyakit diabetes dapat disimpulkan bahwa algoritma *K-Nearest Neighbor* menghasilkan nilai akurasi dan performa yang lebih baik dibandingkan algoritma *Logistic Regression*. Hal tersebut dapat dilihat dari hasil akurasi terbaik dari algoritma KNN sebesar 85,06% sedangkan algoritma *Logistic Regression* memiliki nilai akurasi terbaik sebesar 77,92%.

4. Kesimpulan

Dari hasil penelitian ini dapat disimpulkan bahwa Algoritma KNN menghasilkan nilai akurasi sebesar 85,06% dibandingkan dengan algoritma *Logistic Regression* dengan nilai sebesar 77,92%. Pada saat perhitungan nilai AUC didapatkan nilai dari algoritma KNN sebesar 0,915 dan algoritma *Logistic Regression* sebesar 0,858. Hasil tersebut menunjukkan bahwa algoritma *K-Nearest Neighbor* menghasilkan nilai akurasi dan performa yang lebih baik dibandingkan algoritma *Logistic Regression* dalam melakukan klasifikasi terhadap penyakit diabetes.

Referensi

[1] A. A. Abdillah and Suwarno, "Sistem Deteksi Penyakit Diabetes Menggunakan Metode Support Vector Machine," vol. 2, no. 2, pp. 27–36, 2014.

[2] S. Pangribowo, "Infodatin 2020 Diabetes Melitus," *Pusdatin Kemkes*, 2020.

[3] "IDF Diabetes Atlas Ninth edition 2019," 2019. doi: 10.1016/S0140-6736(55)92135-8.

[4] Isbandiyo, "Penerapan Sequential Methods untuk Handling Missing Value pada Algoritma C4.5 dan Naïve Bayes untuk Memprediksi Penyakit Diabetes Mellitus," Universitas Dian Nuswantoro Semarang, 2016.

[5] J. A. Putra and A. L. Akbar, "Klasifikasi Pengidap Diabetes Pada Perempuan Menggunakan Penggabungan Metode Support Vector Machine dan K-Nearest Neighbour," *Informatics J.*, vol. 1, no. 2, p. 47, 2016, [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.

[6] Indrayanti, D. Sugianti, and M. A. Al Karomi, "Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur Gain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus," *IC-Tech*, vol. 7, no. 2, pp. 1–6, 2017, [Online]. Available: https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/3.

[7] M. S. Mustafa and I. W. Simpen, "Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba," in *Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, 2019, vol. VIII, no. 1, pp. 1–10, [Online]. Available: https://ejurnal.dipanegara.ac.id/index.php/sisiti/article/view/1 -10/68.

[8] M. I. Gunawan, "Sistem Prediksi Penyakit Diabetes Melitus dengan Metode Logistic Regression pada Cloud Computing," 2020.

[9] V. K. Putri and F. I. Kurniadi, "Klasifikasi Diabetes Menggunakan Model Pembelajaran Ensemble Blending," *J. Ultim.*, vol. 10, no. 1, pp. 11–15, 2018, doi: 10.31937/ti.v10i1.709.

[10] R. Asmara, J. Setiawan, and M. N. Tentua, "Komparasi Algoritma C45, Naïve Bayes dan K-Nearest Neighbor Pada Pasien yang Terkena Penyakit Diabetes," *Semin. Nas. Din. Inform.*, 2020, doi: 10.33633/tc.v16i2.1322.

[11] S. Nahzat and M. Yağanoğlu, "Diabetes Prediction Using Machine Learning Classification Algorithms," *Eur. J. Sci. Technol.*, no. 24, pp. 53–59, 2021, doi: 10.31590/ejosat.899716.

[12] P. R. Sihombing and O. P. Hendarsin, "Perbandingan Metode Artificial Neural Network (ANN) dan Support Vector Machine (SVM) untuk Klasifikasi Kinerja Perusahaan Daerah Air Minum (PDAM) di Indonesia," *J. Ilmu Komput.*, vol. 13, no. 1, p. 9, 2020, doi: 10.24843/jik.2020.v13.i01.p02.

[13] M. I. Gunawan, D. Sugiarto, and I. Mardianto, "Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 6, no. 3, pp. 280–284, 2020.

[14] Ardiyansyah, P. A. Rahayuningsih, and R. Maulana, “Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner,” *J. Khatulistiwa Inform.*, vol. VI, no. 1, pp. 20–28, 2018.

[15] T. T. Hanifa, Adiwijaya, and S. Al-faraby, “Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging,” *Univ. Telkom*, vol. 4, no. 2, p. 78, 2017.