Jurnal Elektronik Ilmu Komputer Udayana
Volume 12, No 1. August 2023
p-ISSN: 2301-5373
e-ISSN: 2654-5101

Evaluasi *Performance* dengan *Grid Search* Terhadap K *Nearest Neighbor* (KNN) untuk Klasifikasi Penderita Diabetes Melitus

I Gede Teguh Permana[1], Ida Bagus Gede Dwidasmara[2]

Informatika, Universitas Udayana
Bali, Indonesia
[1]teguhpermana096@gmail.com
[2]dwidasmara@unud.ac.id

Abstract

Diabetes melitus merupakan sebuah penyakit yang disebabkan oleh tingginya gula darah dalam tubuh sebagai respon ketidakmampuan sebuah pankreas untuk memproduksi insulin bahkan Indonesia mencapai posisi 6 sebagai penderita diabetes melitus terbesar di dunia.Deteksi awal mengenai diabetes melitus sangatlah penting untuk mendapatkan perawatan awal sebelum diabetes melitus berimplikasi semakin parah.Deteksi penderita penyakit diabetes melitus dapat dilakukan dengan metode klasifikasi sehingga dari hasil deteksi tersebut, penderita diabetes melitus dapat menjaga pola hidupnya untuk mengontrol variabel independent penyebab diabetes melitus (kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, Insulin, BMI, Keturunan, umur).K-Nearest Neighbour (KNN) adalah sebuah algoritma klasifikasi dapat membantu mendeteksi seseorang menderita diabetes melitus sesuai data dari variabel independen, untuk memperoleh sebuah kualitas KNN yang optimal diperlukan sebuah teknik parameter tuning hyperparameter-K dengan metode *grid search* dan membandingkannya dengan *baseline model*, evaluasi Confusion matrix terhadap *classification report*, ROC *curve.* Hasil evaluasi *performance* dengan *grid search* didapatkan hasil akurasi adalah sebesar 79% pada *training data* 76% pada *testing data*.hasil evaluasi *performance* dengan *grid search* model KNN terhadap *confusion matrix* mengenai klasifikasi penderita diabetes mellitus dapat dinyatakan bahwa model KNN lebih memiliki *ability* untuk *sensitivitas.*

Keywords : *Diabetes Melitus, KNN, grid search,classification,classification report, ROC curve*.

1. *Introduction*

Diabetes Mellitus adalah penyakit yang disebabkan oleh tingginya level gula darah dalam tubuh sebagai respon ketidakmampuan sebuah pankreas untuk memproduksi insulin [2], ketika sebuah pankreas tidak bisa memproduksi sebuah insulin dapat menyebabkan badan kesulitan dalam menjaga level kadar gula darah dalam tubuh sehingga badan dalam keadaan *hyperglycemia* yang terjadi karena level gula darah dalam tubuh yang berlebihan.Pada kasus diabetes melitus diklasifikasikan ke dalam sub kategori diabetes melitus tipe 1 dan diabetes melitus tipe 2 [1]. Diabetes tipe terjadi karena pankreas sudah tidak mampu lagi memproduksi sebuah insulin sedangkan diabetes melitus tipe 2 terjadi karena ketidakmampuan pankreas dalam menyimpan dan memproses insulin secara efektif yang cenderung disebabkan oleh kurangnya aktivitas fisik dan obesitas [3].

Menurut *international diabetes federation* (IDF) pada tahun 2019 menyatakan bahwa masyarakat berumur lebih dari 20 tahun dilaporkan sebagai penderita diabetes melitus sebesar 463 juta jiwa atau dapat dinyatakan bahwa 9.3 % dari populasi dunia menderita diabetes melitus, adapun *death rate* yang disebabkan diabetes melitus mencapai 4.6 juta jiwa. Adapun pada daerah Asia Tenggara, penderita diabetes melitus mencapai 162.6 juta jiwa atau dapat dinyatakan bahwa 9.6% dari penduduk Asia Tenggara mengalami diabetes melitus dengan *death rate* mencapai 1.2 juta jiwa, khususnya Indonesia mencapai posisi 6 sebagai penderita diabetes melitus terbesar di dunia dengan daerah Indonesia sebagai kontribusi penderita diabetes terbesar adalah DKI Jakarta, Kalimantan Timur, DI Yogyakarta,Sulawesi Utara, Jawa Timur

sehingga dapat dinyatakan bahwa sebagian besar penderita diabetes melitus tidak hanya pada kota besar [3].

Faktor yang menyebabkan tingginya gula darah dalam tubuh seringkali berhubungan dengan pola hidup yang tidak sehat, Adapun penderita diabetes melitus secara *genarly* seringkali diderita oleh perempuan dikarenakan adanya *premenstrual syndrome* [2].Faktor umur juga mempengaruhi dimana seringkali semakin bertambah umur fungsi organ vital dalam tubuh seringkali terganggu khususnya fungsi dari pankreas dalam menghasilkan insulin semakin melemah. Kondisi yang tidak memadai pada sektor kesehatan di setiap negara dalam sumber daya manusia dan fasilitas yang disediakan merupakan sebuah faktor penyebab seseorang berimplikasi diabetes melitus [8], deteksi lebih awal mengenai diabetes melitus sangatlah penting untuk mendapatkan perawatan lebih awal sebelum diabetes melitus berimplikasi semakin parah [2].

Deteksi penderita penyakit diabetes melitus dapat dilakukan dengan metode klasifikasi sehingga dari hasil deteksi tersebut, penderita diabetes melitus dapat menjaga pola hidupnya untuk mengontrol variabel independent penyebab diabetes melitus (kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, Insulin, BMI, Keturunan, umur).K-Nearest Neighbour (KNN) adalah sebuah algoritma klasifikasi dapat membantu mendeteksi seseorang menderita diabetes melitus sesuai data dari variabel independen. [5] Adapun pada penelitian yang serupa dilakukan oleh Sulastri 2019 dalam membandingkan algoritma klasifikasi untuk melakukan prediksi penderita penyakit hepatitis dengan menggunakan K-Nearest Neighbour (KNN), naive bayes, dan neural network, pada penelitiannya akurasi terbaik terjadi pada penggunaan K-Nearest Neighbour (KNN) dengan hasil 97 %, adapun K-Nearest Neighbour (KNN) merupakan sebuah algoritma yang secara umum digunakan dalam bidang kesehatan di dunia dalam mencari sebuah solusi untuk pasien baru dari kedekatannya terhadap pasien lama [5].

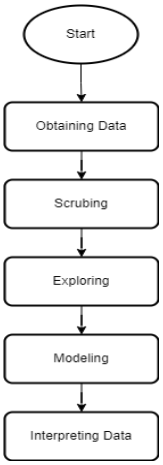
Adapun pada penelitian sebelumnya belum mengoptimasi sebuah K parameter dengan tuning *method* sebagai variabel untuk menentukan *Neighbour* pada model K-Nearest Neighbour (KNN) , jika semakin besar K dapat menghasilkan *bias classification* ataupun sebaliknya dapat menghasilkan *rigid classification* sehingga diperlukan tuning penggunaan K sebagai hyperparameter pada K-Nearest Neighbour (KNN) [4], pada penelitian "Evaluasi *Performance* dengan *Grid Search* Terhadap K *Nearest Neighbor* (KNN) untuk Klasifikasi Penderita Diabetes Melitus" dilakukan sebuah *build up* model *machine learning* KNN dengan merujuk pada metode OSEMN *Pipline* [6] yang diawali dengan *Obtaining data*, *scrubbing, exploring, modeling, interpreting data*.Pada proses *interpreting data* dilakukan sebuah evaluasi *performance* dengan *grid search*, adapun hasil yang diharapkan sebagai evaluasi *performance model machine learning* dengan *grid search* yakni; optimalisasi K *hyperparameter*, *pearson correlation coefficien*, *confusion matrix, classification report,* ROC curve. Evaluasi model dilakukan dengan tuning K parameter dengan *grid search*, *grid search* dilakukan dengan *running* setiap epoch iterasi sebanyak K parameter (1-15) dengan setiap K iterasi akan menghasilkan nilai optimal untuk melakukan proses *classification* penderita diabetes mellitus.

2. *Research Methods*

Metode penelitian yang digunakan pada proses pembuatan model *K-Nearest Neighbour* (KNN) berlandasan terhadap OSEMN *Pipeline methode* [6];

- O - *Obtaining data*
- S - *Scrubbing* (*Cleaning data)*
- E - *Exploring*
- M - *Modeling*
- N - *Interpreting data*

p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gambar 1. Alur Riset

2.1. *Obtaining Data*

Data yang digunakan pada penelitian merupakan sebuah Dataset yang diperoleh dari "*Pima Indians Diabetes Database*" di halaman www.kaggle.com, dataset tersebut merupakan sebuah dataset *original* dari *National Institute of Diabetes and Digestive and Kidney Diseases* yang bertujuan untuk mendiagnosa penderita diabetes. Dataset tersebut berisikan sebuah 9 *attribute* yang digunakan dalam mendeteksi penderita penyakit diabetes melitus, adapun *attribute* tersebut terdiri atas : kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, Insulin, BMI, Keturunan, umur, serta label *attribute* outcome.Dimana dapat di *split data* menjadi outcome sebagai dependent dan lainnya menjadi data independent dengan jumlah dataset sebesar 768 *rows* dan 9 features dalam sebuah dataset [6].

Attribute	Deskripsi	Kriteria
Kehamilan	Jumlah kehamilan yang dialami	0 - 17
Kadar glukosa	Glukosa plasma berkonsentrasi pada 2 jam dalam tes toleransi oral glukosa	0 - 197
Tekanan Darah	Tekanan darah diastolik(mm Hg)	0 - 122
Ketebalan Kulit	Ketebalan lipatan kulit trisep (mm)	0 - 60
Insulin	Insulin serum 2 jam (mu U/ml)	0 - 846
BMI	Indeks massa tubuh (berat (kg) / (tinggi (m))^2)	0 - 67,1
Keturunan	Fungsi silsilah diabetes	0,085 - 2,288
Umur	Umur (tahun)	21 - 69
Outcome	Variabel (0 atau 1) 268 dari 768 adalah 1, yang lainnya adalah 0	0 1

Table 1. Data frame

2.2. *Scrubbing*

Pada proses *scrubbing* dilakukan pada data dalam *atribut* yang bersifat *independen* ,*scrubbing* merupakan bagian penting sebagai persiapan bahan bakar yang berkualitas untuk proses training model *machine learning* KNN.Untuk persiapan bahan bakar yang baik diperlukan *shape frame* yang baik.*Shape frame* data dapat diketahui dengan identifikasi setiap *attribute* pada dataset; didapatkan bahwa seluruh *attribute* memiliki *shape* sebagai numeric sehingga semua *value* pada *attribute* pada dataset dapat dilakukan identifikasi statistik untuk mengetahui keadaan dari data di setiap *attribute*.Pada identifikasi statistik ditemukan sebuah *null value* pada setiap *attribute* (glukosa, tekanan darah, ketebalan kulit, insulin, bmi), *null value* dapat mempengaruhi proses kalkulasi model maka dari itu seluruh *null value* dapat di *convert* ke dalam *mean value* di setiap *attribute* yang terletak *null value*.Pada proses *convert mean value* ditemukan sebuah penyebaran data yang lebih kecil, adapun penyebaran data secara *left-skewed distribution*, *right-skewed distribution,* ataupun *normal distribution*.Pada penyebaran data masih terdapat penyebaran distribusi data secara *left-skewed distribution*, *right-skewed distribution* maka dari itu diperlukan penyamaan bobot untuk menjadikan data mendekati *normal distribution* dengan menggunakan persamaan :

$$Z = \frac{\chi i - \mu}{\sigma} \quad (1)$$

Penyamaan bobot dilakukan dengan proses *scaling*, hal ini disebabkan setiap *attribute* pada dataset memiliki *bobot scale* data yang berbeda satu sama lain yang dapat membuat besarnya penyebaran data pada suatu dataset [6].

2.3. *Exploring*

Adapun sebelum melakukan pemodelan diperlukan sebuah evaluasi dan mentoring data dari hasil *scrubbing* proses.Data di visualisasikan sehingga *pattern* data dapat terlihat sangat jelas.Exploring dapat dilakukan pada proses penghilangan *null value* untuk mengetahui sebuah *frame data* sudah berdistribusi normal.Pada pemodelan dilakukan proses *binary* klasifikasi untuk menunjukan prediksi penderita penyakit diabetes dari suatu pasien, dibutuhkan sebuah korelasi antara setiap fitur (*independent - independent,* ataupun *independent - dependent*) dengan metode *scatter* matriks, jika hasil *scatter* matriks menunjukan bentuk histogram maka dapat dinyatakan pada diagonal variabel terdistribusi tunggal sedangkan hasil menunjukan bentuk *scatter* pada segitiga atas atau bawah menunjukan hubungan antara dua variabel, penggunaan *scatter* matrix tidak dapat mendapatkan hubungan antara *attribute* secara numerical, untuk menunjukan hubungan antara *attribute* dapat dilakukan dengan proses *pearson correlation coefficient* yang ditunjukan jika hasil mendekati 0 maka tidak memiliki relasi ataupun jika hasil adalah mendekati -1 maka dapat memperoleh korelasi negatif ataupun mendekati +1 dapat memperoleh korelasi positif [6].

2.4. *Modeling*

K-nearest neighbors merupakan sebuah algoritma *machine learning* yang bersifat labeling (*supervised learning*), secara umum digunakan pada *problem regression,* ataupun *classification*.KNN diimplementasikan dengan "mengelompokkan" data menurut kesamaan fitur-fiturnya.KNN memiliki sebuah hyperparameter yang dapat dilakukan proses tuning untuk mencari sebuah value dari parameter yang terbaik dalam kalkulasi didalam sebuah model, “k” merepresentasikan sebuah jumlah tetangga yang digunakan untuk membandingkan data.Setiap proses *similarity* fitur diperlukan sebuah parameter *distance* untuk seberapa dekat antara fitur data. Distance parameter berdampak pada sebuah ukuran dan karakteristik dari setiap tetangga setiap fitur data, adapun metode yang dapat dilakukan untuk perhitungan antara *distance* [8] :

a. Euclidean

   Menghitung *similarity* terhadap jarak terpendek antara dua tetangga, metode euclidean sangat sensitif terhadap *scale* data maka diperlukan proses fitur normalisasi.

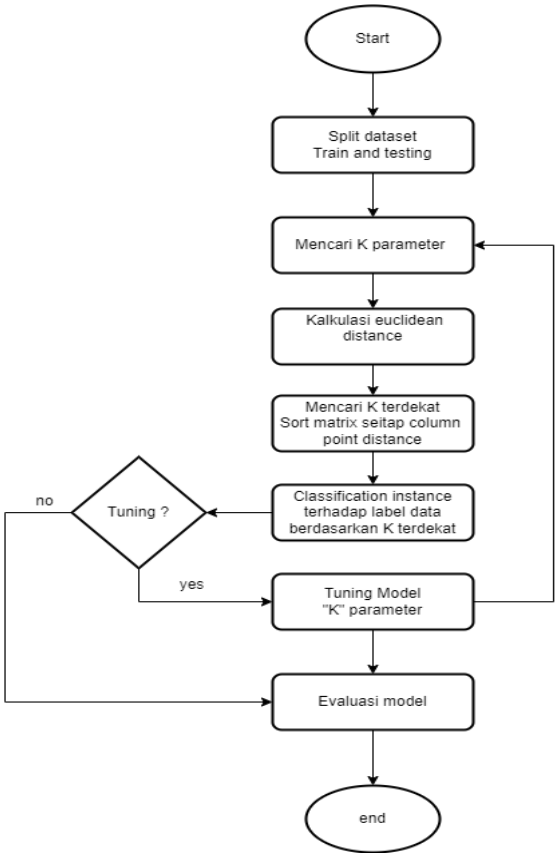
p-ISSN: 2301-5373
e-ISSN: 2654-5101

b. Taxicab or Manhattan

Menghitung *similarity* terhadap jumlah perbedaan absolut antara dua tetangga dari titik koordinat kartesius.

c. Minkowski : Gabungan dari metode Euclidean dan Manhattan

Jumlah dari fitur sangat mempengaruhi proses KNN secara signifikan dikarenakan banyaknya nilai yang dimiliki dan yang semakin "*unique*" setiap tetangga yang dimiliki yang dapat berdampak pada perhitungan setiap *distance* untuk menentukan fitur yang mana lebih dekat dengan tetangga "k", pada kalkulasi KNN penulis menggunakan metode euclidean dengan menghitung *similarity* terhadap jarak terpendek antara tetangga.Adapun tahapan proses dari KNN dapat diimplementasi sebagai berikut :



Gambar 2. Skema algoritma KNN

Skema urutan proses klasifikasi algoritma KNN ditunjukan pada :

1. Pisahkan dataset dalam data training dan testing
2. Spesifikasi K parameter
3. Kalkulasi jarak antara training dan testing data dengan menggunakan metode euclidean *distance*. Euc = $\sqrt{\sum_{i=1}^{n}(pi - qi)^2}$ (2)

   Pi = data *training*

   Qi = data *testing*

   I = data variabel

N = dimensi data

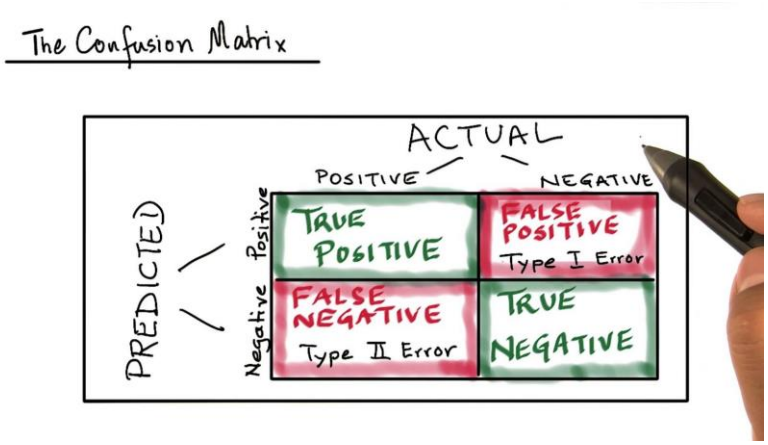
4. Mencari K terdekat dari *sorting* setiap *distance formed*
5. Klasifikasi *instance data* terhadap label data, dengan mencari *distance* classes yang terdekat dari tetangga dan tambahkan kedalam sebuah class data label “k”
6. Penentuan Tuning hyperparameter K menggunakan *grid search method*, jika iya maka proses tuning akan dilaksanakan
7. Evaluasi model dari *result training* model [6].

2.5. *Interpreting data*

Hasil dari model KNN diinterpretasikan ke dalam bentuk metode evaluasi ataupun metode optimasi parameter model;

1. *Confusion Matrix*

Sebuah *confusion matrix* merupakan teknik yang digunakan dalam merangkum sebuah *performance* dari sebuah *classification algorithm*, pada case klasifikasi penderita diabetes melitus merupakan *binary classification*.



Gambar 3.*Confusion matrix*

Pada representasi tersebut dapat dinyatakan bahwa *confusion matrix* berelasi dari *actual case* terhadap *predicted case*.Pada *true positive* (TP) merupakan sebuah *value* yang diklasifikasi benar dan kenyataannya adalah benar; pada case klasifikasi penderita diabetes melitus dapat diartikan bahwa pasien diprediksi sebagai klasifikasi penderita diabetes dengan kenyataan sebuah pasien tersebut sesungguhnya merupakan penderita diabetes.Pada *true negative* (TN) yakni pasien yang diklasifikasi tidak menderita diabetes melitus dan kenyataannya adalah pasien tersebut tidak mengalami diabetes melitus, dimana di pasien diprediksi klasifikasi tidak termasuk dalam klasifikasi label (penderita diabetes melitus).Adapun pada *false positive* merupakan sebuah *type error* I dari sebuah klasifikasi yang mana pasien dinyatakan menderita diabetes melitus namun pada kenyataannya tidak menderita diabetes melitus.Pada *false negative* merupakan sebuah *type error* II dari sebuah klasifikasi yang mana pasien dinyatakan tidak menderita diabetes melitus namun pada kenyataannya pasien tersebut menderita sebuah diabetes melitus.Pada *case true positive* dengan *case false negative* dapat dinyatakan sebuah *recal* (*type error* II), adapun pada *case true positive* dengan *case false positive* dapat dinyatakan sebagai *precision* (*type error* I) [9].

2. *Classification Report*

Classification report merupakan sebuah laporan dari hasil kalkulasi sebuah data *testing* terhadap model yang sudah *training*, adapun *report* yang dideskripsikan merupakan *precision, recal, F1-Score,* akurasi [8].

a. *Precision*

Akurasi dari *positive prediction*, *precision* = $\frac{TP}{(TP + FP)}$ (3)

TP = *True positive*

FP = *False positive*

b. *Recall*

Sensitivitas dari prediksi klasifikasi *class* atau *true positive rate*, bagian dari *positive* yang diidentifikasi benar, *recall* = $\frac{TP}{(TP + FN)}$ (4)

FN = *False negative*

c. *F1 - Score*

Digunakan untuk mewujudkan *harmonic threshold* dari *value precision* dengan *recall*, *F1 Score* = $2\frac{(precision\ x\ recall)}{(precision + recall)}$ (5)

d. Akurasi

Digunakan untuk mengetahui kemampuan model dalam mengakurasi klasifikasi model disetiap *confusion matrix*, Akurasi = $\frac{(TP + TN)}{(TP+TN +FP + FN)}$ (6)

3. ROC Curve

ROC (*Receiver Operating Characteristic*) yang membahas perihal *curve* tentang seberapa bagus sebuah model bisa membedakan perihal dua hal (pasien mengalami diabetes melitus atau tidak).Seringkali menyatakan sebuah *true positive rate* terhadap *false positive rate* sebagai kemampuan model dalam *precision class* klasifikasi antara yang terjadi pada *type error* I pada *case false positive* dengan *true positive* [8].

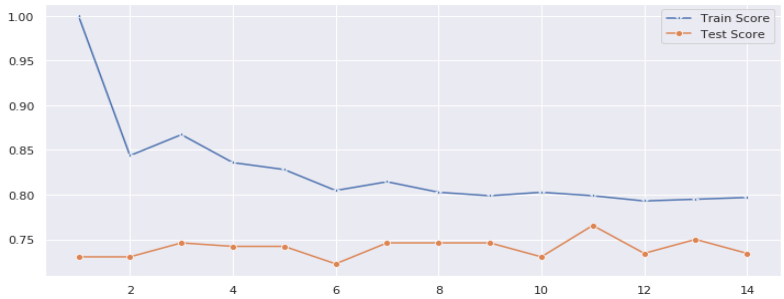
4. Grid Search Method Hyperparameter tuning

Grid search merupakan sebuah metode untuk tuning sebuah hyperparameter pada suatu mode, khususnya pada KNN model yakni parameter “K”.Tuning hyperparameter dilakukan dalam evaluasi model untuk setiap kombinasi dari algoritma parameter “K” yang secara spesifik dalam sebuah grid.*Grid search* akan memulai proses konstruksi setiap versi dari parameter K dengan sebuah *grid* yang didefinisikan ke dalam sebuah model *tuning* sehingga memperoleh parameter K terbaik dari hasil proses model *tuning* yang digunakan untuk *train model* selanjutnya, sehingga memperoleh *performance* model yang lebih baik dari *base modeling* [9]. Adapun pada penelitian sebelumnya dengan *grid search* di implementasikan untuk prediksi hasil *test* HIV/AIDS terhadap model *classification machine learning* melalui pendekatan optimisasi yang digunakan untuk evaluasi *performance* terhadap hyperparemeter tuning setiap parameter model *machine learning* untuk menghasilkan prediksi hasil test HIV/AIDS yang optimial, pada penelitian tersebut menggunakan model *machine learning*; logistic regression (LR), random forest (RF), *support vector machine* (SVM), KNN, *decision tree* (DT), gradient bosting (GB), ada boost (AB), and extra tree (ET) yang menghasilkan hasil prediksi terbaik dari model GB, ET sebesar 87.7 %, dan hasil prediksi kurang baik pada model AB sebesar 80.9 % dengan setiap hyperparameter di setiap

alogirtma dilakukan proses evaluasi *performance* dengan *grid search* [11], jadi penggunaan *grid search* dalam proses evaluasi *performance* dengan tuning setiap *hyperparameter* model *machine learning* khususnya pada model KNN sangatlah baik dan umum di implementasikan.

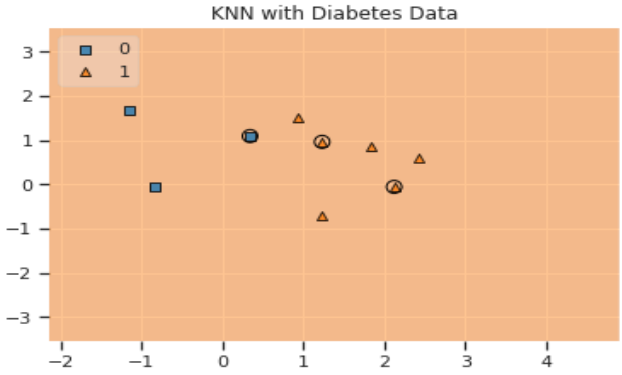
3. *Result and Discussion*

3.1. *Grid search* hyperparameter-k



Gambar 4. *Grid search* hyperparameter-k

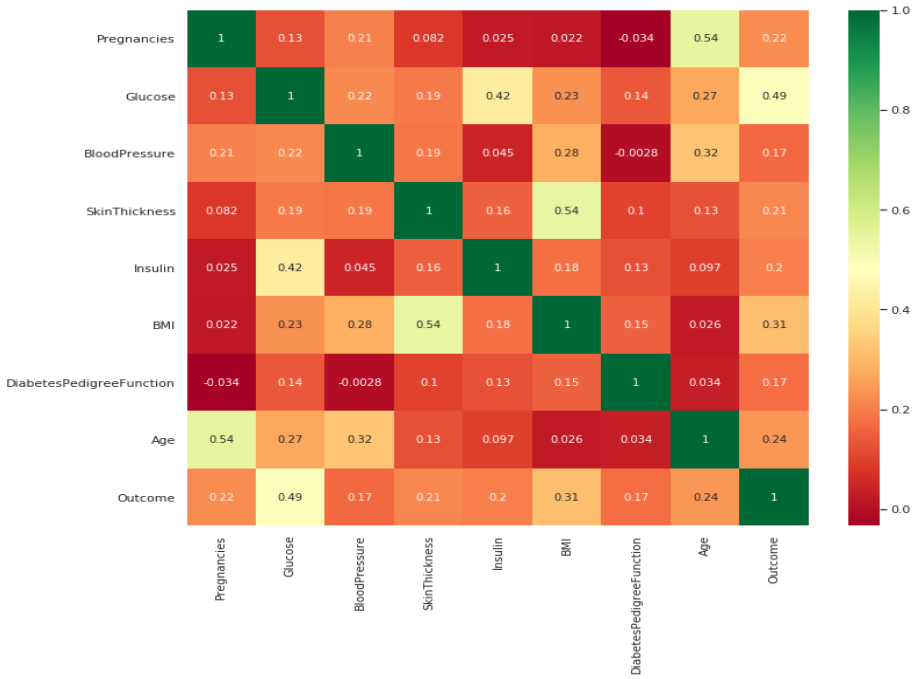
Dilihat pada grafik visualisasi hasil dari proses tuning dengan *grid search* menunjukan bahwa dengan *baseline* model yakni k=0 maka memperoleh model dalam keadaan *overfitting* yang mana model memiliki akurasi sangat baik saat melakukan proses training namun pada proses testing tidak menghasilkan akurasi yang baik.Implementasi *grid search* membantu model mencari sebuah parameter k terbaik sehingga menghasilkan model yang optimal dengan k adalah 11 dengan hasil akurasi adalah sebesar 79% pada *training data* 76% pada *testing data*, adapun jika dibandingkan dengan model yang dihasilkan pada *baseline model* diperoleh akurasi sebesar 100% pada *training data* adapun 73% pada *testing data*, jika dibandingkan *baseline model* dengan *grid search model* maka model dari hasil tuning *grid search* lebih optimal.Adapun hasil dari klasifikasi penderita diabetes melitus dapat divisualisasikan terhadap *testing data*.



Gambar 5. Klasifikasi penderita Diabetes melitus

3.2. *pearson correlation coefficient*

p-ISSN: 2301-5373
e-ISSN: 2654-5101

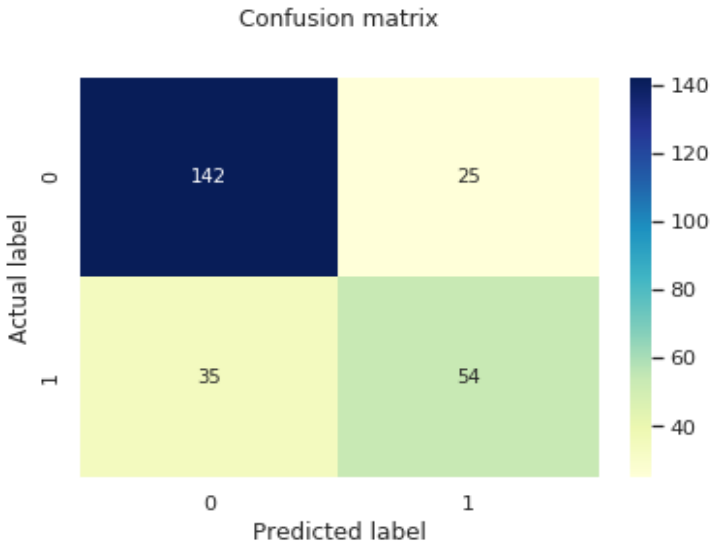


Gambar 6.*Heatmap correlations*

pearson correlation coefficient dapat dilakukan dengan metode *heatmap* yang mana dapat diketahui setiap korelasi antara *atribut* dengan *atribut* lainnya, adapun untuk mengetahui penyebaran data di setiap *atribut* dapat dipetakan dengan *pair plot*.

Gambar 7.*Pair plot*

3.3. *Confusion Matrix*



Gambar 8. *Confusion matrix*

Confusion matrix dipetakan dari y label *data testing* terhadap hasil y label klasifikasi, pada pemetaannya dapat ditunjukan bahwa tingkat kualitas model dalam membedakan pasien yang mengalami diabetes ataupun tidak dapat dilakukan dengan baik yang ditandai dengan nilai pada *case true positive* dan *true negative* cukup besar dibandingkan dengan *case error* I *and* II (*false positive, false negative*) dengan *case true positive* sebesar 54 *instance* dan *case true negative* sebesar 142, jadi terdapat 54 istance yang dapat dipetakan sebagai penderita dan 142 dapat dinyatakan sebagai pasien yang tidak menderita diabetes melitus namun tingkat klasifikasi 54 orang mengalami bias sebesar 25 orang terhadap *case false positive* yang dapat dinyatakan bahwa 25 orang yang seharusnya tidak mengalami diabetes namun di klasifikasi sebagai penderita diabetes.Sensitivitas model dapat dinyatakan cukup baik dengan *case false negative* dalam *range* yang tidak cukup besar jika dibandingkan dengan *case true positive* terhadap klasifikasi *case true positive* cukup baik dengan 35 *instance* pada case *false negative* diklasifikasi tidak mengalami diabetes namun *actual* pasien tersebut menderita diabetes melitus.

3.4. *Classification Report*

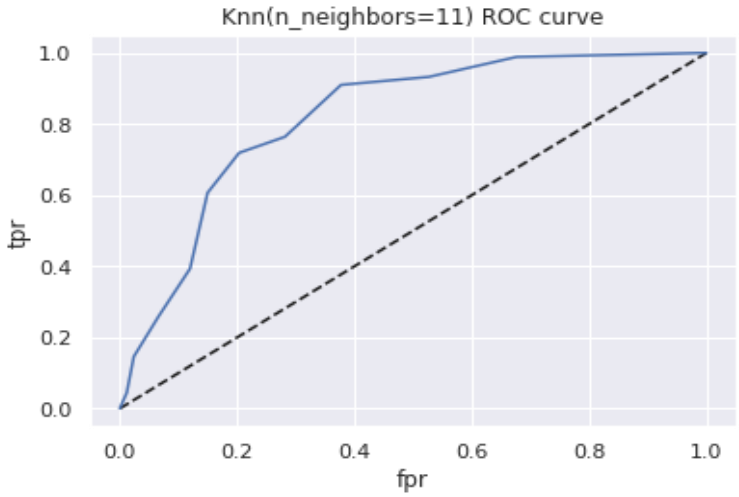
	precision	recall	f1-score	support
0	0.80	0.85	0.83	167
1	0.68	0.61	0.64	89
micro avg	0.77	0.77	0.77	256
macro avg	0.74	0.73	0.73	256
weighted avg	0.76	0.77	0.76	256

Gambar 9.*Classification report*

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Classification report mendeskripsikan nilai dari *precision*, *recall,* f1-*score* yang mana model ditunjukan dengan 76% nilai *recall* lebih besar dengan 77% nilai *precision,* dapat dinyatakan bahwa model KNN lebih memiliki kekuatan untuk *sensitivitas* dalam pengenalan sebuah *instance* sebagai *true positive* terhadap *false negative* dibandingkan *ability* untuk membedakan model termasuk ke dalam *true positive* dan *true negative* terhadap *false positive* sebagai *false alarm case* , dengan *threshold* normal antara nilai *precision* dan *recall* sebesar 76%.

3.5. ROC Curve



Gambar 10. ROC *curve*

ROC *curve* menyatakan sebuah *ability* model dalam memprediksi klasifikasi *positive* pada *case true positive* terhadap *false positive* yang mana semakin besar nilai ROC maka semakin kecil terjadinya *false alarm* (*actual negative* tapi *prediction true*), hal ini sangat penting dalam *medical* untuk menghindari salah diagnosa penderita diabetes melitus dengan hasil ROC sebesar 81%.

4. ***Conclusion***

Pada hasil yang sudah dideskripsikan maka hasil evaluasi *performance* model KNN dengan *grid search* terindikasi sebagai model yang optimal atau *good fit* dengan hyperparameter k terbaik adalah 11 dari hasil evaluasi parameter K (1-15) didapatkan hasil akurasi dari evaluasi *performance* adalah sebesar 79% pada *training data* 76% pada *testing data*, adapun jika dibandingkan dengan model yang dihasilkan pada *baseline model* (tidak dilakukan evaluasi *performance* dengan *grid search*) dapat dinyatakan sebagai model *overfitting dengan* diperoleh sebuah akurasi sebesar 100% pada *training data* adapun 73% pada *testing data*.Pada proses *pearson correlation coefficient* sebagai hasil evaluasi *performance* ditunjukan bahwa *glukosa*, bmi*, age* merupakan *atribut* yang sangat berpengaruh terhadap proses klasifikasi dengan penyebaran data yang cukup kecil terhadap *outcome attribute.*Pada hasil evaluasi *performance* model KNN terhadap *confusion matrix* mengenai klasifikasi penderita diabetes mellitus dapat dinyatakan bahwa model KNN lebih memiliki *ability* untuk *sensitivitas* dalam pengenalan sebuah *instance* sebagai *true positive* terhadap *false negative* dibandingkan *ability* untuk membedakan model termasuk ke dalam *true positive* dan *true negative* terhadap *false positive* sebagai *false alarm case* , dengan *threshold* normal antara nilai *precision* dan *recall* sebesar 76%, adapun *rate* dalam prediksi klasifikasi *case positive* (penderita diabetes) cukup baik dengan hasil ROC sebesar 81%.Adapun penulis pada pengemabangan selanjutnya dapat dilakukan *ensemble method* dengan algoritma *voting classification* yang mana sebuah algoritma dikombinasikan untuk memperoleh hasil prediksi klasifikasi yang lebih baik.

Daftar Pustaka

[1] Hartati I, Pranata AD, Rahmatullah MR. Hubungan Self Care Dengan Kualitas Hidup Pasien Diabetes Melitus di Poli Penyakit Dalam RSUD Langsa. JP2K. 2019;2(2):94– 104. Available from: http://stikescnd.ac.id/jurnal/index.php/smart/article/view/30

[2] International Diabetes Federation. Diabetes Atlas. 9th ed. 2019. 4. Kemenkes. Hari Diabetes Sedunia Tahun 2018. Jakarta: Kementerian Kesehatan RI; 2019.

[3] Kementerian Kesehatan RI. Infodatin Diabetes Mellitus. Pusat Data dan Informasi Kementerian Kesehatan RI. 2020;4.

[4] Sulastri, K. Hadiono, M. T. Anwar, "Analisis Perbandingan Prediksi Penyakit Hepatitis Dengan Menggunakan Algoritma K-Nearest Neighbor, Naive Bayes Dan Neural Network" DINAMIK, vol. 24, no. 2, p. 82 – 91, 2019.

[5] A. Fitria and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf., vol. 3, no. 2, pp. 102– 106, 2018

[6] Hasran, "Klasifikasi Penyakit Jantung Menggunakan K-Nearest Neighbor" Indonesian Journal Of Data And Science, vol. 1, no. 1, pp. 06 – 10, 2020.

[7] A. Wanto, M. N. H. Siregar, A. P. Windarto, D. Hartama, N. L. W. S. R. Ginantra, D. Napitupulu, E. S. Negara, M. R. Lubis, S. V. Dewi and C. Prianto, Data Mining: Algoritma Dan Implementasi, Medan: Yayasan Kita Menulis, 2020.

[8] D. Cahyanti, A. Rahmayani and S. A. Husniar, "Analisis Performa Metode KNN Pada Dataset Pasien Pengidap Kanker Payudara" Indonesia Journal Of Data And Science, vol. 1, no. 2, 2020.

[9] M. A. Imron and B. Prasetiyo, "Improving Algorithm Accuracy K-Nearest Neighbor Using Z-Score Normalization and Particle Swarm Optimization to Predict Customer Churn," J. Soft Comput. Explor., vol. 1, no. 1, pp. 56–62, 2020, [Online]. Available: https://shmpublisher.com/index.php/joscex/article/view/7%0Ahttps://shmpublisher.com/index.php/joscex/index.

[10] I. W. Santiyasa, G. P. A. Brahmantha, I. W. Supriana, I. G. G. A. Kadyanan, I. K. G. Suhartana, and I. B. M. Mahendra, "Identification of Hoax Based on Text Mining Using KNearest Neighbor Method," JELIKU (Jurnal Elektron. Ilmu Komput. Udayana), vol. 10, no. 2, pp. 217–226, 2021, doi: 10.24843/jlk.2021.v10.i02.p04.

[11] Mesafint.Daniel,D.H.Manjaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results" International Journal of Computers and Applications, vol.44, no.9, page.4, 2021