Analisis Perbandingan Algoritma *KNN* dan *Naïve Bayes* dalam Mendiagnosis Penyakit Diabetes Mellitus

Aryanto Bangun Widodo dan Ericks Rachmat Swedia

Fakultas Ilmu Komputer Universitas Gunadarma
Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat
E-mail: aryantounlimited123@gmail.com , erick_rs@staff.gunadarma.ac.id

Abstrak

Penyakit diabetes mellitus adalah penyakit dimana kadar glukosa di dalam darah menjadi tinggi karena tubuh tidak dapat memproduksi atau mengeluarkan insulin secara cukup. Dibidang kesehatan, diagnosa penyakit diabetes mellitus merupakan masalah penting guna mendiagnosa sejak dini sehingga cepat mendapatkan penanganan yang terbaik. Berdasarkan permasalahan tersebut salah satu teknik yang dapat digunakan dalam melakukan diagnosa adalah data mining dengan teknik *classification*. Peneliti melakukan analisis perbandingan algoritma *KNN* dan *Naïve Bayes* untuk mendiagnosis penyakit diabetes mellitus. Penelitian ini bertujuan untuk memberikan pemahaman dasar tentang kinerja algoritma klasifikasi dalam mendiagnosis diabetes mellitus. Pengujian dilakukan menggunakan perbandingan 80% sebagai *Data Training* : 20% sebagai *Data Testing* dan menggunakan 3 *fold validation*. Berdasarkan hasil penelitian dan pengujian ditarik kesimpulan bahwa algoritma *Naïve Bayes* memiliki akurasi lebih baik dibandingkan algoritma *KNN*. Algoritma *Naïve Bayes* berdasarkan perhitungan memiliki akurasi tertinggi sebesar 74.7% sedangkan algoritma KNN memiliki akurasi 68.6%. Selisih akurasi dari kedua algoritma tersebut sebesar 6.1%.

Kata kunci : *KNN, Naïve Bayes,* Diabetes Mellitus.

Pendahuluan

Data Mining adalah penggabungan sejumlah disiplin ilmu komputer yang dapat didefenisikan sebagai penemuan pola-pola baru dari kumpulan data sangat besar, meliputi metode yang merupakan irisan dari *Artificial Intelligence (AI)* dan *Machine Learning*. *Data Mining* memiliki banyak sekali manfaat dalam kehidupan ini diantaranya dalam pengolahan kumpulan data besar sehingga data tersebut dapat dijadikan sebagai sumber pengetahuan dan informasi yang baru. Sekarang ini pemanfaatan *data mining* tidak hanya sebatas dalam bidang ilmu teknologi, melainkan juga dalam bidang kesehatan. *Data mining* dapat dijadikan sebagai acuan untuk menganalisis, memprediksi, mendiagnosa suatu jenis penyakit dengan menggunakan metode yang dapat diterapkan. Dari banyaknya jenis penyakit, salah satu penyakit yang dapat diprediksi dengan menggunakan metode *data mining* yaitu penyakit diabetes mellitus. Penyakit diabetes mellitus adalah penyakit di mana kadar glukosa di dalam darah menjadi tinggi karena tubuh tidak dapat memproduksi atau mengeluarkan insulin secara cukup disebabkan ketidakmampuan pankreas dalam tubuh untuk memproduksi hormon insulin. Kekurangan insulin dapat mengakibatkan gula darah di dalam tubuh mengalami jumlah yang berlebihan. Dalam bidang kesehatan, diagnosa penyakit diabetes mellitus merupakan masalah yang sangat penting guna mendiagnosa penderita sejak dini sehingga cepat mendapatkan penanganan yang terbaik. Berdasarkan permasalahan tersebut salah satu teknik yang dapat digunakan dalam melakukan diagnosa yaitu menggunakan teknik *data mining*. Teknik data mining yang dapat digunakan yaitu teknik klasifikasi. Pemanfaatan teknik *data mining* untuk memprediksi penyakit diabetes mellitus sudah banyak dilakukan oleh peneliti. Beberapa penelitian terdahulu terkait dengan diagnosa diabetes mellitus menggunakan algoritma klasifikasi *data mining* :

Penelitian yang dilakukan oleh [1] mengimplementasikan algoritma *Naïve Bayes* yang mampu menghasilkan akurasi yang baik. Dalam hasil penelitiannya didapatkan nilai akurasi 92%. Hasil ini lebih baik dibandingkan dengan penelitian sebelumnya yang menggunakan *KNN* dengan tingkat akurasi sebesar 91%. Penelitian dilakukan oleh [2] peneliti melakukan komparasi menggunakan algo-

DOI : http://dx.doi.org/10.32409/jikstik.23.3.3617

ritma *Naive Bayes* dan *KNN*. Penelitian tersebut bertujuan untuk mengetahui algoritma mana yang cocok untuk membangun pengetahuan penyakit diabetes. Hasil penelitian menunjukkan bahwa algoritma *KNN* sebagai algoritma yang memiliki akurasi terbaik dengan *Naïve Bayes* yaitu 85,60% dan KNN sebesar 91,61%. Penelitian dilakukan oleh [3] penelitian ini menggunakan dua algoritma yaitu *Naïve Bayes* dan *KNN*. Berdasarkan hasil penelitian nilai akurasi algoritma *Naïve Bayes* lebih tinggi dibandingkan algoritma *KNN* yaitu *Naïve Bayes* sebesar 80% dan *KNN* sebesar 75%. Penelitian dilakukan oleh [4] dalam penelitian tersebut algoritma yang digunakan yaitu algoritma C.45. Penelitian tersebut melakukan klasifikasi data penderita diabetes dengan teknik *data mining* klasifikasi yang menggunakan algoritma C.45. Penelitian dilakukan oleh [5] peneliti menggunakan algoritma klasifikasi yaitu algoritma *Naïve Bayes Classifier*. Tujuan penelitian tersebut adalah untuk mengetahui hasil klasifikasi pasien ke dalam dua kategori diagnosis diabetes mellitus yaitu 'Ya' dan 'Tidak' menggunakan algoritma *Naïve Bayes Classifier* dan mengetahui tingkat akurasi dari empat proporsi data yaitu 60:40, 70:30, 80:20 dan 90:10. Berdasarkan tingkat akurasi yang telah diketahui, nilai akurasi terbaik adalah pada proporsi *data testing* 40% dan 20% dengan nilai akurasi sebesar 92,31%.

Kemudian Penelitian dilakukan oleh [6] peneliti menggunakan dua *dataset* yaitu PIDD (*Pima Indian Diabetes Dataset*) dan 130_US. Teknik yang digunakan untuk analisis dataset adalah *Random Forest, KNN, Naïve Bayes* dan J48. Kemudian peneliti menggunakan pendekatan ansambel. Keakuratan pendekatan ansambel yang diusulkan adalah 93,62% untuk PIDD dan 88,56% untuk 130_US *dataset* rumah sakit. Penelitian dilakukan oleh [7] peneliti menggunakan algoritma *Decision Tree, KNN* dan *SVM*. Tujuan utama dari penelitian tersebut adalah untuk memprediksi penyakit diabetes mellitus sejak awal. Hasil dari penelitian diperoleh bahwa SVM mengungguli *Decision Tree* dan *KNN* dengan akurasi tertinggi 90,23%. Penelitian dilakukan oleh [8] dalam penelitian tersebut bertujuan untuk menerapkan teknik *resampling bootstrapping* untuk meningkatkan akurasi dan kemudian menerapkan *Naïve Bayes*, *Decision Trees* dan *KNN* kemudian membandingkan kinerjanya. Tujuan utamanya adalah untuk menentukan pola baru dan kemudian menginterpretasikan pola tersebut untuk memberikan informasi yang signifikan dan berguna bagi pengguna. Pada penelitian ini metode yang diusulkan memberikan akurasi yang tinggi dengan nilai akurasi 90,36% dan keputusan stump memberikan akurasi yang lebih rendah dari yang lain dengan memberikan akurasi 83,72%. Penelitian dilakukan oleh [9] dalam penelitian tersebut bertujuan untuk mengklasifikasikan dan memprediksi pasien diabetes dengan merancang model classifier berdasarkan lima algoritma klasifikasi yang berbeda antara lain *Decision Tree, KNN, Naive Bayes, Random Forest* dan SVM. Eksperimen pengujian menunjukkan bahwa akurasi yang diberikan oleh model pengklasifikasi yang dikembangkan dengan menggunakan *Decision Tree, KNN, Naïve Bayes, SVM* dan *Random Forest* masing-masing yaitu 73,82%, 71,65%, 76,30%, 65,10% dan 68,74%. Dengan demikian, penelitian tersebut menunjukkan bahwa algoritma *Naïve Bayes* memberikan akurasi yang lebih baik dalam memprediksi diabetes dibandingkan dengan algoritma lainnya. Penelitian dilakukan oleh [10] penelitian tersebut dilakukan untuk mendeteksi diabetes mellitus dengan mengembangkan model hybrid yang terdiri dari dua model *machine learning* yaitu *Light Gradient Boosting Machine (LGBM)* dan *KNN*. Penelitian ini bertujuan mengembangkan model pembelajaran mesin untuk mendeteksi terjadinya diabetes pada pasien. Sistem yang diusulkan memiliki akurasi 91% dan kurva karakteristik operasi penerima yaitu 93%. Hasil percobaan menunjukkan bahwa akurasi prediksi model *hybrid* lebih baik daripada pembelajaran mesin tradisional.

Selanjutnya Penelitian ini dilakukan oleh [11] peneliti menggunakan algoritma *Random Forest* dan *SVM*. Penelitian ini bertujuan untuk mengevaluasi efisiensi metode yang digunakan berdasarkan klasifikasi. Hasil penelitian bahwa algoritma *Random Forest* menawarkan 75.7813 presisi lebih tinggi daripada *SVM*. Hasil penelitian dapat membantu profesional medis dapat membuat keputusan untuk perawatan kepada penderita diabetes. Penelitian ini dilakukan oleh [12] peneliti menerapkan empat teknik data mining seperti *Random Forest, SVM, Logistic Regresi* dan *Naive Bayes*. Hasil penelitian bahwa regresi logistik memiliki nilai tertinggi dibandingkan dengan lainnya yaitu 82,46%. Penelitian ini dilakukan oleh [13] peneliti menggunakan algoritma *Naïve Bayes, Decesion Tree, Adaboost* dan *Random Forest* untuk memprediksi pasien diabetes. Regresi logistik (LR) digunakan untuk mengidentifikasi faktor risiko penyakit diabetes berdasarkan nilai rasio dan odds. Tujuan utama dari penelitian ini adalah untuk mengembangkan sistem berbasis mesin *learning* untuk memprediksi pasien diabetes. Kombinasi *classifer* berbasis LR dan *Random Forest* berkinerja lebih baik, dimana kombinasi fitur berbasis LR seleksi dan *classifer berbasis Random Forest* memberikan nilai akurasi 94,25% dan nilai AUC 0,95. Penelitian dilakukan oleh [14] peneliti menggunakan alagoritma *Logistic Regression Method, KNN, SVM, Naive Bayes, Decesion Tree, Random Forest.* Penelitian tersebut bertujuan untuk menilai risiko diabetes dikalangan individu berdasarkan gaya hidup dan latar belakang keluarga. Hasil penelitian bahwa performa pengklasifikasi algoritma Random Forest ditemukan paling akurat dengan nilai akurasi 94,10%. Penelitian dilakukan oleh [15] peneliti menggunakan algoritma *k-nearest Neighbour, Decision Trees, Ran-*

dom Forest, AdaBoost, Naive Bayes dan *XGBoost* serta *Multilayer Perceptron (MLP)*. Hasil penelitian tersebut bahwa algoritma *XGBoost* memiliki nilai akurasi tertinggi yaitu 94,6%.

Pada penelitian ini, digunakan *dataset public* yang bersumber dari https://www.kaggle.com/datasets/mathchi/diabetes-data-set. *Dataset* ini terdiri dari sembilan (9) atribut yaitu *Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age* dan *Outcome*. Kemudian dataset ini berjumlah 768 data. Peneliti menggunakan algoritma *KNN* dan *Naïve Bayes*, karena kedua algoritma ini merupakan algoritma yang sering digunakan dalam penelitian sebelumnya dan hasilnya akurasi kedua algoritma tersebut paling baik dalam memprediksi penyakit diabetes mellitus. Sehingga perbedaan peneliti ini ingin mengetahui diantara algoritma *KNN* dan *Naïve Bayes* mana algoritma yang terbaik dalam memprediksi penyakit diabetes mellitus. Selain itu alasan peneliti memilih algoritma *KNN* dan *Naive Bayes* karena memiliki alasan khusus diantaranya kemudahan/kesederhanaan implementasi, kecepatan komputasi dan performa yang sudah terbukti baik pada penelitian sebelumnya. Perbedaan selanjutnya dimana penelitian ini menggunakan satu sumber *dataset public* dan juga perbedaan perbandingan proporsi antara data training dan data testing yaitu dengan perbandingan 80%:20%. Selain itu juga perbedaan penggunaan tools yaitu *Orange Data Mining*. Penelitian ini bertujuan untuk memberikan pemahaman dasar tentang kinerja algoritma klasifikasi dalam mendiagnosis diabetes mellitus. Hasil penelitian ini diharapkan dapat membantu dokter, ahli medis, penderita maupun masyarakat luas dalam mengambil langkah terhadap diagnosa penyakit diabetes mellitus sejak dini.

Metode Penelitian

Metode pada penelitian ini terdiri dari beberapa tahapan proses, diantaranya :

Identifikasi Masalah

Permasalahan yaitu mendiagnosa penderita diabetes mellitus dengan gejala yang timbul pada *dataset*. Diabetes mellitus adalah penyakit yang cukup berbahaya saat ini. Penyakit ini dapat menyebabkan komplikasi yang tinggi sehingga dapat mengakibatkan kematian. Penyakit diabetes mellitus perlu dikenali sejak dini, sehingga dapat ditangani sesegera mungkin dengan perawatan yang tepat. Berdasarkan hal tersebut perlu adanya klasifikasi untuk mendiagnosa penyakit diabetes mellitus secara akurat dan tepat. Algoritma *KNN* dan *Naïve Bayes* adalah algoritma klasifikasi yang digunakan dalam penelitian ini.

Studi Literatur

Studi literatur bertujuan untuk mengetahui teori yang mendukung dalam penelitian yang akan dikerjakan. Selain itu dapat menjadi referensi terkait permasalahan yang dialami dalam proses penelitian serta dalam melakukan proses diagnosa penderita yang beresiko penyakit diabetes mellitus menggunakan *data mining* dengan metode klasifikasi. Teori pendukung ini yang dijadikan dasar maupun referensi dalam melakukan penelitian yang akan dikerjakan.

Dataset

Data yang digunakan dalam penelitian ini bersumber dari *Kaggle.com* (https://www.kaggle.com/datasets/mathchi/diabetes-data-set) dengan nama diabetes.csv dengan ukuran 23.87 kB. *Dataset* ini memiliki 9 variabel dan 768 data dengan terdiri dari 2 target yaitu 0 / No Diabtes (Tidak terdiagnosa penyakit diabetes mellitus) dan 1 / Diabetes (Terdiagnosa penyakit diabetes mellitus). Target dari *dataset* tersebut yaitu untuk membedakan antara penderita yang sakit (menderita penyakit diabetes mellitus) dengan yang sehat (tidak menderita penyakit diabetes mellitus. Hasil target pada dataset ini digunakan peneliti sebagai perbandingan untuk mendapatkan akurasi terbaik dari model yang diujikan. Deskripsi variabel dataset diabetes mellitus berdasarkan nilai rangenya dapat dilihat pada Tabel 1.

Tabel 1: Variabel *Dataset*

Variabel	Nilai
Pregnancies	0 – 17
Glucose	0 – 199
BloodPressure	0 – 122
SkinThickness	0 – 99
Insulin	0 – 864
BMI	0 – 67.1
DiabetesPedigreeFunction	0.078 – 2.42
Age	21 – 81
Outcome	0 – 1

Preprocessing

Pada langkah ini dataset akan dilakukan pengecekan dan pembersihan sehingga fitur yang dilakukan uji coba hanya data yang relevan untuk penelitian. Tahapan dalam *preprocessing* harus dilakukan semua. Berikut tahapan *preprocessing* data:

Data Cleaning

Data Cleaning adalah proses mempersiapkan data untuk analisis dengan menghapus atau memodifikasi data yang tidak benar, tidak lengkap, tidak relevan, diduplikasi, atau diformat dengan tidak benar. Dalam penelitian ini tidak ditemukan adanya beberapa data yang kosong, tidak benar, tidak lengkap, tidak relevan ataupun format yang tidak benar.

Discretization

Discretization atau diskretisasi yaitu proses pengkategorian atau pengelompokkan nilai berdasarkan masing-masing attribut.

	Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
1	1.0	≥ 4.5	≥ 129.5	65.5 - 77	≥ 29.5	< 7	28.85 - 34.75	≥ 0.533	≥ 36.5
2	0.0	< 1.5	< 104.5	65.5 - 77	13.5 - 29.5	< 7	< 28.85	0.2685 - 0.533	25.5 - 36.5
3	1.0	≥ 4.5	≥ 129.5	< 65.5	< 13.5	< 7	< 28.85	≥ 0.533	25.5 - 36.5
4	0.0	< 1.5	< 104.5	65.5 - 77	13.5 - 29.5	7 - 125.5	< 28.85	< 0.2685	< 25.5
5	1.0	< 1.5	≥ 129.5	< 65.5	≥ 29.5	≥ 125.5	≥ 34.75	≥ 0.533	25.5 - 36.5
6	0.0	≥ 4.5	104.5 - 129.5	65.5 - 77	< 13.5	< 7	< 28.85	< 0.2685	25.5 - 36.5
7	1.0	1.5 - 4.5	< 104.5	< 65.5	≥ 29.5	7 - 125.5	28.85 - 34.75	< 0.2685	25.5 - 36.5
8	0.0	≥ 4.5	104.5 - 129.5	< 65.5	< 13.5	< 7	≥ 34.75	< 0.2685	25.5 - 36.5
9	1.0	1.5 - 4.5	≥ 129.5	65.5 - 77	≥ 29.5	≥ 125.5	28.85 - 34.75	< 0.2685	≥ 36.5
10	1.0	≥ 4.5	104.5 - 129.5	≥ 77	< 13.5	< 7	< 28.85	< 0.2685	≥ 36.5
11	0.0	1.5 - 4.5	104.5 - 129.5	≥ 77	< 13.5	< 7	≥ 34.75	< 0.2685	25.5 - 36.5
12	1.0	≥ 4.5	≥ 129.5	65.5 - 77	< 13.5	< 7	≥ 34.75	≥ 0.533	25.5 - 36.5
13	0.0	≥ 4.5	≥ 129.5	≥ 77	< 13.5	< 7	< 28.85	≥ 0.533	≥ 36.5
14	1.0	< 1.5	≥ 129.5	< 65.5	13.5 - 29.5	≥ 125.5	28.85 - 34.75	0.2685 - 0.533	≥ 36.5
15	1.0	≥ 4.5	≥ 129.5	65.5 - 77	13.5 - 29.5	≥ 125.5	< 28.85	≥ 0.533	≥ 36.5
16	1.0	≥ 4.5	< 104.5	< 65.5	< 13.5	< 7	28.85 - 34.75	0.2685 - 0.533	25.5 - 36.5
17	1.0	< 1.5	104.5 - 129.5	≥ 77	≥ 29.5	≥ 125.5	≥ 34.75	≥ 0.533	25.5 - 36.5
18	1.0	≥ 4.5	104.5 - 129.5	65.5 - 77	< 13.5	< 7	28.85 - 34.75	< 0.2685	25.5 - 36.5
19	0.0	< 1.5	< 104.5	< 65.5	≥ 29.5	7 - 125.5	≥ 34.75	< 0.2685	25.5 - 36.5
20	1.0	< 1.5	104.5 - 129.5	65.5 - 77	≥ 29.5	7 - 125.5	28.85 - 34.75	0.2685 - 0.533	25.5 - 36.5
21	0.0	1.5 - 4.5	104.5 - 129.5	≥ 77	≥ 29.5	≥ 125.5	≥ 34.75	≥ 0.533	25.5 - 36.5
22	0.0	≥ 4.5	< 104.5	≥ 77	< 13.5	< 7	≥ 34.75	0.2685 - 0.533	≥ 36.5
23	1.0	≥ 4.5	≥ 129.5	≥ 77	< 13.5	< 7	≥ 34.75	0.2685 - 0.533	≥ 36.5
24	1.0	≥ 4.5	104.5 - 129.5	≥ 77	≥ 29.5	< 7	28.85 - 34.75	< 0.2685	25.5 - 36.5

Gambar 1: Sampel Dataset Setelah Proses Discretization

Pada Gambar 1 dapat dilihat hasil diskretiasasi atau pengelompokkan nilai, sebagai contoh Glucose dikelompokan nilai ≥129, <104.5 dan 104.5 – 129.5. Insulin dikelompokan nilai ≥125.5, <7 dan 7 – 125.5.

Data Training dan *Data Testing*

Dataset dibagi menjadi 2 yaitu *Data Training* dan *Data Testing*. Pengujian yang dilakukan dengan menggunakan perbandingan 80% (614 data) sebagai *Data Training* : 20% (154 data) sebagai *Data Testing*. Pemilihan perbandingan ini berdasarkan penelitian yang dilakukan oleh (Khasanah, Nasution dan Amijaya. 2022) dimana nilai akurasi terbaik yaitu pada proporsi *data testing* 20% dengan nilai akurasi sebesar 92,31%.

Klasifikasi Algoritma *KNN* dan *Naïve Bayes*

K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) adalah salah satu algoritma yang dapat digunakan untuk melakukan pengklasifikasian. Algoritma *K-Nearest Neighbor* menggunakan *Neighborhood Classification* sebagai nilai prediksi dari nilai *instance* yang baru. Prinsip kerja algoritma *K-Nearest Neighbor* dengan mencari jarak terdekat antar data yang akan dievaluasi dengan tetangga terdekat dalam data training. Algoritma *K-Nearest Neighbor (KNN)* merupakan salah satu algoritma paling sederhana untuk memecahkan masalah klasifikasi dan sering menghasilkan hasil yang kompetitif dan signifikan. *K-fold cross validation* adalah sebuah metode yang digunakan untuk mengetahui rata-rata keberhasilan pengklasifikasian dengan melakukan pembagian dataset secara acak menjadi k himpunan bagian. Mengacu pada penelitian sebelumnya, penelitian ini menggunakan *3-fold cross validation*, dimana dataset dibagi menjadi empat subset. Dua di antaranya digunakan untuk melatih setiap lipatan dan satu untuk validasi. Setelah evaluasi tersebut, ketiga subset tersebut digunakan untuk melatih model akhir dan sisanya digunakan untuk menguji kinerja sistem.

Naïve Bayes

Naive Bayes yaitu pengklasifikasi sederhana berdasarkan teorema Bayesian dengan asumsi penentuan nasib sendiri yang kuat. Pengklasifikasi Naïve Bayes menganggap bahwa ada (atau tidak adanya) fitur (atribut) tertentu dari suatu kelas tidak terkait dengan ada (atau tidak adanya) fitur lain ketika variabel kelas diberikan. *Naïve Bayes* merupakan metode pengklasifikasian berdasarkan probabilitas sederhana dan dirancang agar dapat dipergunakan dengan asumsi antar variabel penjelas saling bebas (*independen*). Hipotesis dalam Teorema Bayes yaitu label kelas yang menjadi

pemetaan dalam klasifikasi, sedangkan bukti menjadi fitur pendukung.

Analisa Hasil Akurasi

Akurasi merupakan alat ukur yang biasa digunakan untuk mengevaluasi kinerja model yang telah dibangun. Untuk menghitung akurasi ditunjukkan pada persamaan berikut :

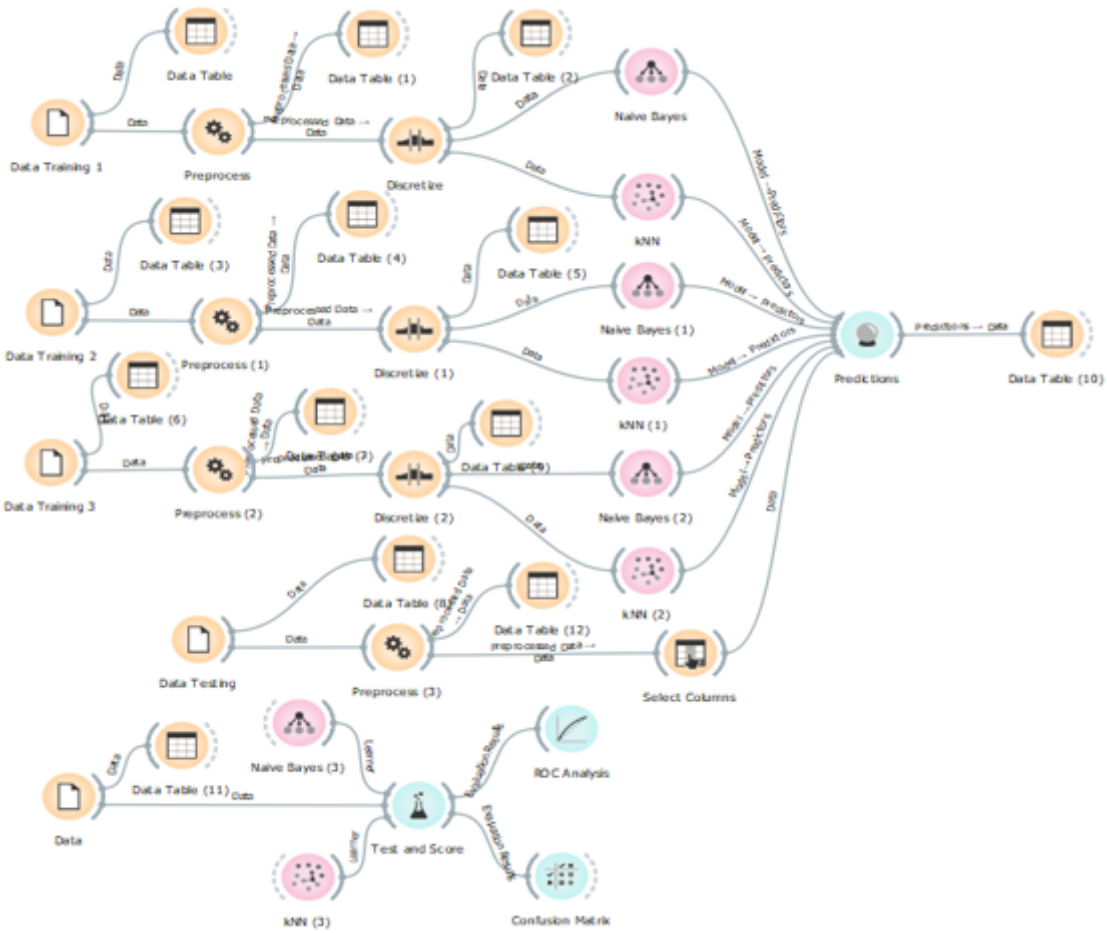
$$Akurasi = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) * 100\% \quad (1)$$

1. TP atau *True Positives* adalah jumlah tuple positif yang dilabeli dengan benar oleh *classifier*. *Tuple* positif adalah *tuple* aktual yang berlabel positif, seperti *tuple* dengan label = 'Diabetes'.
2. TN atau *True Negatives* adalah jumlah *tuple* negatif yang dilabeli dengan benar oleh *classifier*. *Tuple* negatif adalah tuple aktual yang berlabel negatif, seperti *tuple* dengan label = 'No Diabetes'.
3. FP atau *False Positives* adalah jumlah *tuple* negatif yang salah dilabeli oleh classifier. Misalnya, sebuah *tuple* pasien yang berlabel = 'No Diabetes' akan tetapi oleh *classifier* dilabeli = 'Diabetes'.
4. FN atau *False Negatives* adalah jumlah *tuple* positif yang salah dilabeli oleh *classifier*. Misalnya, sebuah *tuple* pasien yang berlabel = 'Diabetes' akan tetapi oleh classifier dilabeli = 'No Diabetes'. Empat istilah tersebut dapat digambarkan sebagai *confusion matrix* seperti yang diilustrasikan pada Tabel 2.

Tabel 2: *Confusion Matrix*

No.	Kelas Aktual	Kelas Hasil Prediksi		
		Diabetes Mellitus	No Diabetes Mellitus	Jumlah
1	Diabetes Mellitus	TP	FN	P
2	No Diabetes Mellitus	FP	TN	N
3	Jumlah	P'	N'	P + n

Hasil penelitian nantinya juga akan di uji dengan *confusion matrix* dan analisa kurva ROC untuk mengetahui tingkat akurasi dan laju kesalahan yang terdapat dalam masing-masing pengujian model.



Gambar 2: Pelatihan dan Pengujian Algoritma

Hasil dan Pembahasan

Hasil dari setiap pengujian algoritma dapat menampilkan hasil prediksi, nilai akurasi, nilai recall, nilai presisi serta kurva ROC (Receiver Operating Characteristic) pada kedua algoritma tersebut.

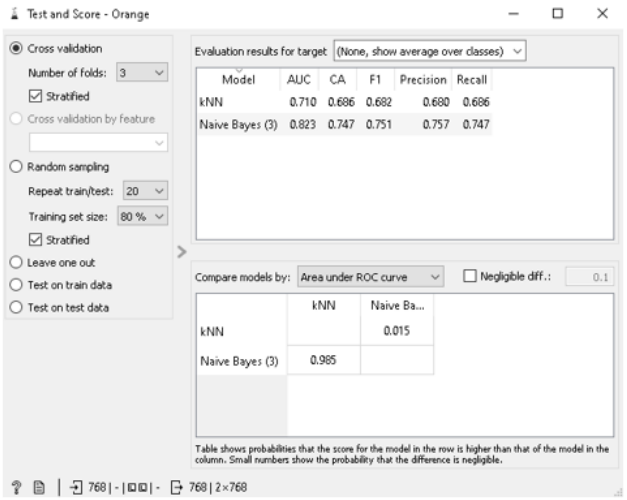
Pelatihan dan Pengujian Model Klasifikasi

Gambar 2 adalah pelatihan dan pengujian algoritma yang dilakukan untuk mengetahui hasil dari proses penelitian. *Data Training*, berfungsi untuk memasukan *dataset training*, berikut ini penjelasannya :

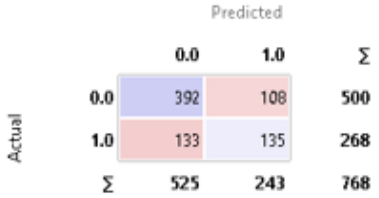
1. *Data Training*, berfungsi untuk memasukan *dataset training*
2. *Data Testing*, berfungsi untuk memasukan *dataset testing*
3. Data, berfungsi untuk memasukan dataset yang akan diuji
4. *Preprocess,* berfungsi untuk melakukan pembersihan data dari duplikasi, *noise* atau data yang kosong atau hilang
5. *Discretize,* berfungsi untuk pengelompokkan isi atribut berdasarkan numerik dan kategori
6. *Data Table*, untuk melihat isi dataset sebelum dan sesudah proses *preprocess*, sesudah proses diskretisasi, *data training, data testing*, data uji dan hasil prediksi
7. *Naïve Bayes*, Pengujian model algoritma yang dilakukan
8. *KNN*, Pengujian model algoritma yang dilakukan
9. *Test and Score*, untuk mengetahui *recall*, presisi, akurasi dari pengujian model algoritma yang dilakukan sehingga bisa diketahui akurasi
10. *Confussion Matrix*, berfungsi untuk mengetahui jumlah benar dan salah dalam pengujian yang dilakukan.
11. *ROC Analysis*, berfungsi untuk mengetahui laju kurva akurasi dalam pengujian model algoritma yang dilakukan.

Hasil Akurasi Pengujian Algoritma *K-Nearest Neighbor (KNN)*

Setelah dilakukan pengujian dan pemodelan dengan menggunakan *tools Orange Data Mining* dengan Algoritma *K-Nearest Neighbor (KNN)* menampilkan hasil sebagai berikut :



Gambar 3: Hasil Test&Score Algoritma K-Nearest Neighbor (KNN)



Gambar 4: Hasil *Confussion Matrix Algoritma K-Nearest Neighbor (KNN)*

Berdasarkan Gambar 3 dan 4, dari total 768 data yang dilakukan pengujian, terdapat 500 pasien tidak terdiagnosa penyakit diabetes mellitus. Dari data tersebut diprediksi pasien yang tidak terdiagnosa penyakit diabetes mellitus (No Diabetes) sebanyak 392 pasien dan hasilnya sesuai dengan yang ada pada *dataset*. Kemudian sebanyak 108 pasien, hasil prediksi terdiagnosa penyakit diabtes mellitus (Diabetes) namun hasil prediksi tersebut tidak sesuai dengan yang ada pada dataset. Sedangkan dari 268 pasien terdiagnosa penyakit diabtes mellitus. Dari *dataset* sebanyak 133 pasien terdiagnosa penyakit diabetes mellitus (Diabetes) namun hasil prediksi tidak terdiagnosa penyakit diabetes mellitus (No Diabetes). Kemudian sebanyak 135 pasien hasil prediksi terdiagnosa penyakit diabetes mellitus (Diabetes) dan hasil prediksi tersebut sesuai dengan yang ada pada *dataset*.

Keterangan persamaan *confussion matrix* sebagai berikut :

TP : 135

FP : 133

TN : 392

FN : 108

Presisi (Hasil = Diabetes Mellitus) = $\frac{135}{135+133}$ = $\frac{135}{268}$ = $0.504 = 50.4\%$

Nilai presisi kelas Diabetes Mellitus yaitu 50.4

Presisi (Hasil = No Diabetes Mellitus) = $\frac{392}{392+108} = \frac{392}{500} = 0.784 = 78.4$

Nilai presisi kelas No Diabetes Mellitus yaitu 78.4

Recall (Hasil = Diabetes Mellitus) = $\frac{135}{135+108} = \frac{135}{243} = 0.556 = 55.6$

Nilai *recall* kelas Diabetes Mellitus yaitu 55.6

Recall (Hasil = No Diabetes Mellitus) = $\frac{392}{392+133} = \frac{392}{525} = 0.747 = 74.7$

Nilai *recall* kelas No Diabetes Mellitus yaitu 74.7

Akurasi = $\frac{392+135}{392+135+133+108} = \frac{527}{768} = 0.686 = 68.6$

Dari hasil perhitungan yang telah dilakukan menunjukkan bahwa hasil akurasi dari algoritma *K-Nearest Neighbor (KNN)* memiliki hasil akurasi sebesar 68.6% dan sesuai dengan pengujian akurasi yang dilakukan pada *tools Orange Data Mining* sebesar 68.8%.

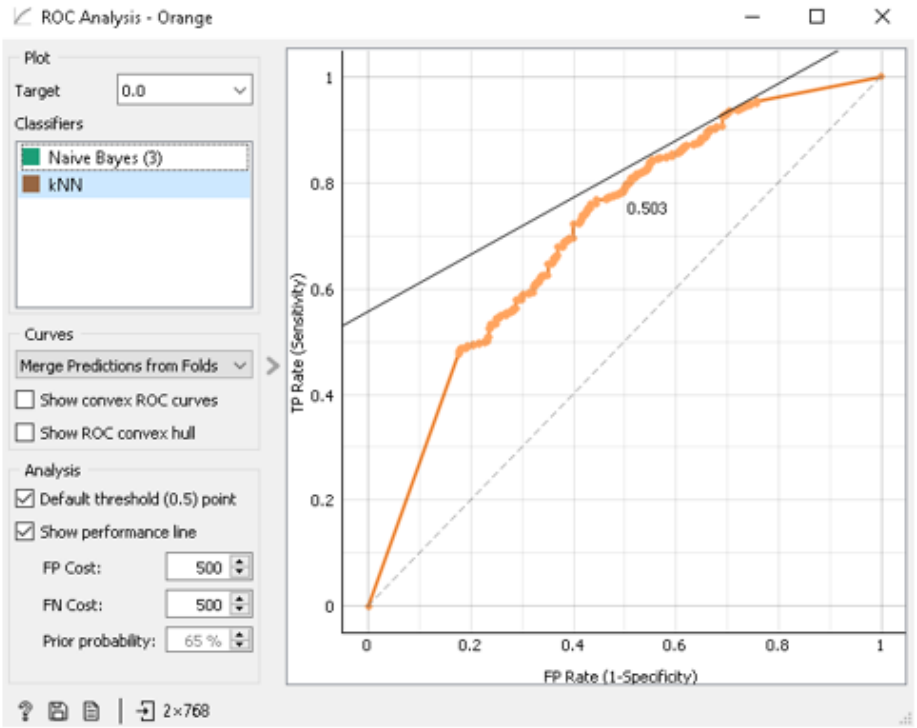
Kurva *ROC K-Nearest Neighbor (KNN)*

Kurva *ROC* dibuat berdasarkan nilai telah yang didapatkan pada perhitungan dengan *confusion matrix* yaitu antara *False Positive Rate* dengan *True Positive Rate*. Kurva *ROC*) digunakan untuk memvisualisasikan secara dua dimensi performa dari setiap klasifikasi yang diujikan, dimana garis horizontal merupakan nilai *false positive*, sedangkan garis vertical berupa *true positive*. Nilai Area *Under Curve (AUC)* merupakan area dibawah kurva *ROC*. Untuk pengkategorian hasil AUC, nilai kualitas suatu klasifikasi berdasarkan nilai *AUC* nya bisa dilihat pada Tabel 3.

Pada Gambar 5 menghasilkan Kurva *ROC Algoritma K-Nearest Neighbor (KNN)* dengan nilai *Area Under Curve (AUC)* sebesar 0.710. Termasuk kedalam kelas *Fair Classification*.

Tabel 3: Kriteria *AUC*

Nilai *AUC*	Keterangan
90% - 100%	*Excellent Classification*
80% - 90%	*Good Classification*
70% - 80%	*Fair Classification*
60% - 70%	*Poor Classification*
<60%	*Failur*

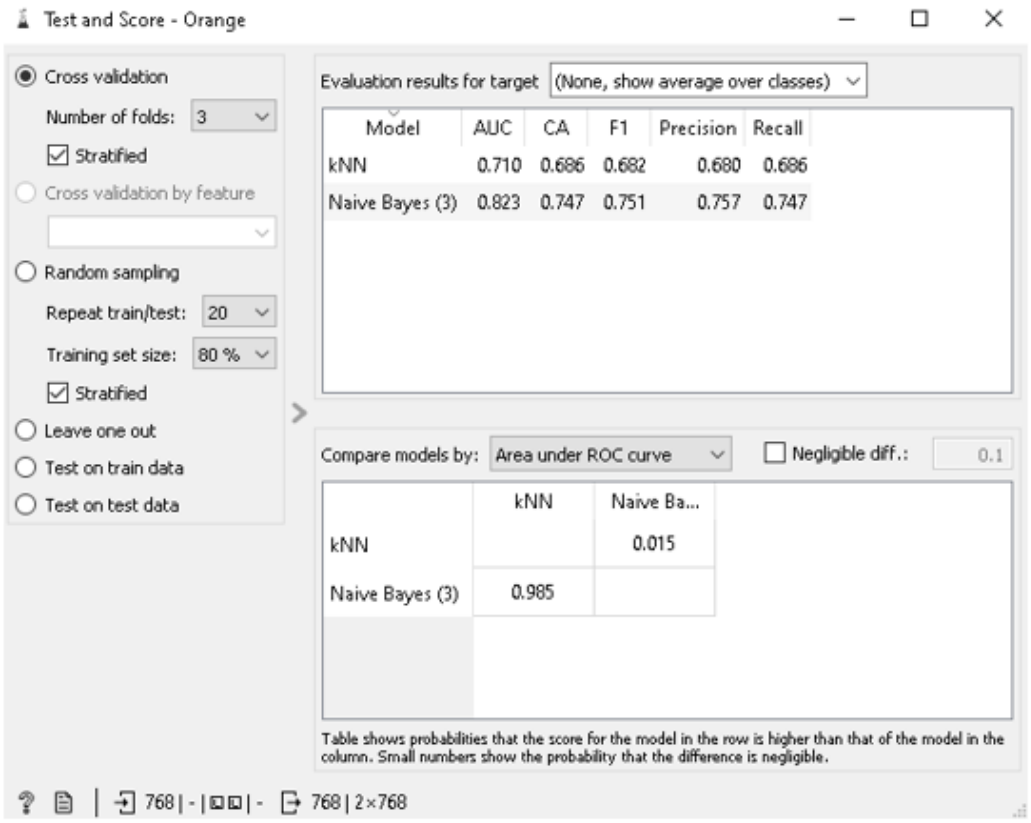


Gambar 5: Hasil Kurva *ROC K-Nearest Neighbor (KNN)*

Hasil Akurasi Pengujian Algoritma *Naïve Bayes*

Setelah dilakukan pengujian dan pemodelan dengan menggunakan tools Orange Data Mining dengan Algoritma *Naïve Bayes* menampilkan hasil seperti ditunjukkan pada Gambar 6 dan 7. Berdasarkan Gambar 6 dan 7, dari total 768 data yang dilakukan pengujian, terdapat 500 pasien tidak terdiagnosa penyakit diabetes mellitus. Dari data tersebut hasil prediksi pasien yang tidak terdiagnosa penyakit diabtes mellitus (No Diabetes) sebanyak 385 pasien dan hasil prediksi tersebut sesuai dengan yang ada pada *dataset*. Kemudian 115 pasien hasil prediksi terdiagnosa penyakit diabetes mellitus (Diabetes) namun hasil prediksi tersebut tidak sesuai dengan

yang ada pada *dataset*. Sedangkan dari 268 pasien terdiagnosa penyakit diabtes mellitus. Dari data tersebut sebanyak 79 pasien hasil prediksi tidak terdiagnosa penyakit diabetes mellitus (No Diabetes) namun hasil prediksi tersebut tidak sesuai dengan yang ada pada *dataset*. Kemudian sebanyak 189 pasien hasil prediksi terdiagnosa penyakit diabetes mellitus (Diabetes) dan hasilnya sesuai dengan yang ada pada *dataset*.



Gambar 6: Hasil *Test&Score* Algoritma *Naïve Bayes*

Keterangan persamaan confussion matrix sebagai berikut :

TP : 189

FP : 79

TN : 385

FN : 115

Presisi (Hasil = Diabetes Mellitus) = $\frac{189}{189+79} = \frac{189}{268} = 0.705 = 70.5$

Nilai presisi kelas Diabetes Mellitus yaitu 70.5

Presisi (Hasil = No Diabetes Mellitus) = $\frac{385}{385+115} = \frac{385}{500} = 0.77 = 77$

Nilai presisi kelas No Diabetes Mellitus yaitu 77

Recall (Hasil = Diabetes Mellitus) = $\frac{189}{189+115} = \frac{189}{304} = 0.622 = 62.2$

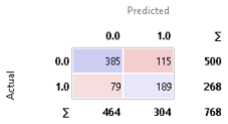
Nilai recall kelas Diabetes Mellitus yaitu 62.2

Recall (Hasil = No Diabetes Mellitus) = $\frac{385}{385+79} = \frac{385}{464} = 0.830 = 83$

Nilai recall kelas No Diabetes Mellitus yaitu 83

Akurasi = $\frac{385+189}{385+189+79+115} = \frac{574}{768} = 0.747 = 74.7$

Dari hasil perhitungan yang telah dilakukan menunjukkan bahwa hasil akurasi dari algoritma *Naïve Bayes* memiliki hasil akurasi sebesar 74.7% dan sesuai dengan pengujian akurasi yang dilakukan pada *tools Orange Data Mining* sebesar 74.7%.

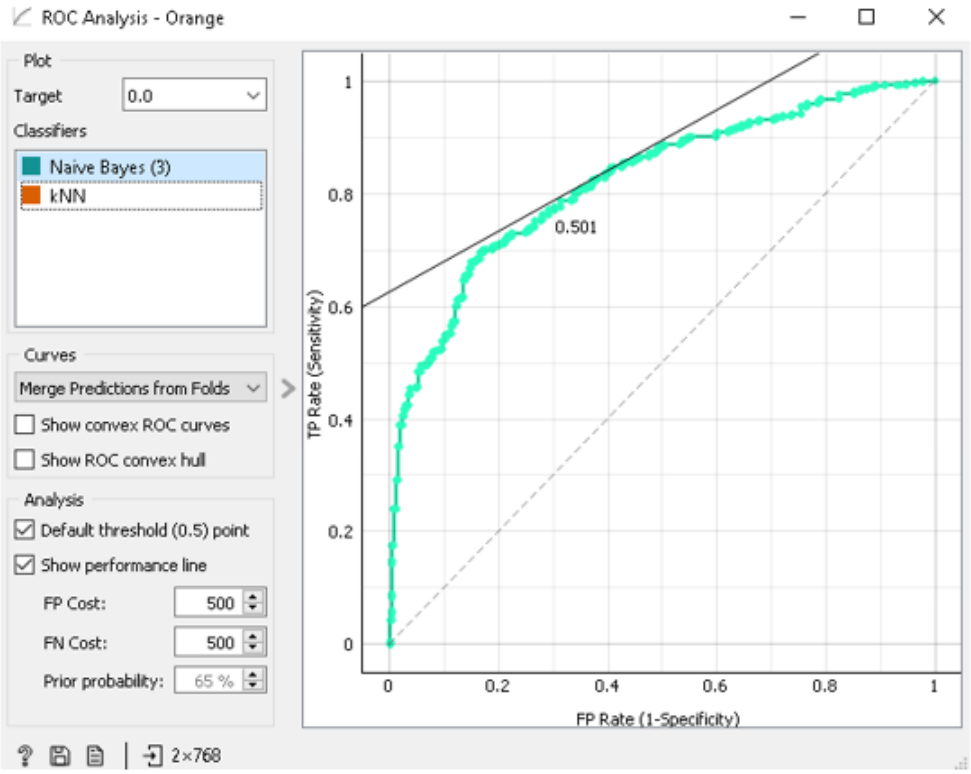


Gambar 7: *Hasil Confussion Matrix* Algoritma *Naïve Bayes*

Kurva *ROC Naïve Bayes*

Gambar 8 adalah hasil dari kurva *ROC* Algoritma *Naïve Bayes*. Kurva *ROC* dibuat berdasarkan nilai telah yang didapatkan pada perhitungan dengan *confusion matrix* yaitu antara *False Positive Rate* dengan *True Positive Rate*. Kurva *ROC*) digunakan untuk memvisualisasikan secara dua dimensi performa dari setiap klasifikasi yang diujikan, dimana garis horizontal merupakan nilai *false positive*, sedangkan garis vertical berupa *true positive*. Nilai *Area Under Curve (AUC)* merupakan area dibawah kurva *ROC*. Untuk pengkategorian hasil *AUC*, nilai kualitas suatu klasifikasi berdasarkan nilai *AUC* nya dapat dilihat pada Tabel 3. Pada Gambar 8 menghasilkan Kurva *ROC* Algoritma

Naïve Bayes Classifier dengan nilai *Area Under Curve (AUC)* sebesar 0.823. Termasuk kedalam kelas *Good Classification.*



Gambar 8: Hasil Kurva *ROC Naïve Bayess*

Analisa Hasil Komparasi

Setelah melakukan pengujian terhadap dua model algoritma yaitu *K-Nearest Neighbor (KNN)* dan *Naïve Bayes* menggunakan *Confussion Matrix* dan *AUC*, maka dapat dibuat perbandingan terhadap dua model tersebut dalam memprediksi penyakit diabtes mellitus terhadap pasien sebagai berikut :

Tabel 4: Analisa Komparasi

Algoritma	Akurasi	AUC
k-Nearest Neighbor (k-NN)	68.6%	0.710
Naïve Bayes	74.7%	0.823

Berdasarkan tabel 4 dapat dilihat dari segi akurasi bahwa algoritma *Naïve Bayes Classifier* lebih baik dari pada algoritma *K-Nearest Neighbor (KNN)*. Dengan perbandingan *dataset* 80% *data training* : 20% *data testing*, algoritma *K-Nearest Neighbor (KNN)* memiliki nilai akurasi sebesar 68.6%. Sedangkan algoritma *Naïve Bayes* memiliki nilai akurasi 74.7 % . Selisih akurasi dari kedua model tersebut sebesar 6.1%. Kemudian untuk nilai *AUC* yang ditarik dari kurva *ROC* menunjukkan bahwa dengan perbandingan *dataset* 80% *data training* : 20% *data testing*, algoritma *K-Nearest Neighbor (KNN)* memiliki nilai sebesar 0.710. Sedangkan algoritma *Naïve Bayes* memiliki nilai sebesar 0.823. Kedua model memiliki selisih nilai AUC sebesar 0.113. Dengan demikian algoritma *Naïve Bayes Classifier* lebih baik dibandingkan dengan algoritma *K-Nearest Neighbor (KNN)*.

Penutup

Berdasarkan hasil penelitian dan pengujian terhadap dataset diagnosa penyakit diabetes mellitus, maka dapat ditarik kesimpulan bahwa model algoritma *Naïve Bayes* memiliki akurasi lebih baik dibandingkan dengan model algoritma *K-Nearest Neighbor (KNN)*. Dengan penggunaan *tools Orange Data Mining* dapat membantu dalam menghitung nilai Akurasi, AUC, *Precision* dan *Recall* dari dua model yang dibandingkan tersebut. Berdasarkan perhitungan menggunakan *tools* dengan perbandingan dataset 80% *data training* : 20% *data testing*, algoritma *K-Nearest Neighbor (KNN)* memiliki nilai akurasi 68.6 %. Sedangkan algoritma *Naïve Bayes* memiliki nilai akurasi sebesar 74.7%. Selisih akurasi dari kedua model tersebut sebesar 6.1%. Kemudian perhitungan tersebut telah diuji berdasarkan nilai *recall* maupun presisi baik dari

pasien yang terdiagnosa penyakit diabetes mellitus (Diabetes Mellitus) maupun yang tidak terdiagnosa penyakit diabetes mellitus (No Diabetes Mellitus). Sedangkan untuk nilai AUC ditarik dari konversi ROC berdasarkan perhitungan dengan *tools Orange Data Mining* menunjukkan bahwa dengan perbandingan dataset 80% *data training* : 20% *data testing*, algoritma *K-Nearest Neighbor (KNN)* memiliki nilai sebesar 0.710. Sedangkan algoritma *Naïve Bayes* memiliki nilai sebesar 0.823. Kedua model memiliki selisih nilai AUC sebesar 0.113. Maka dapat ditarik kesimpulan untuk penelitian dengan dataset diagnosa penyakit diabetes mellitus, lebih baik menggunakan algoritma *Naïve Bayes* dibandingkan menggunakan algoritma *K-Nearest Neighbor (KNN)*.

Beberapa saran dari peneliti diharapkan dapat membuat penelitian selanjutnya menjadi lebih baik, diantaranya sebagai berikut:

1. Penggunaan data primer dari rumah sakit di Indonesia sehingga dapat diketahui prediksi aktual mengenai kondisi pasien terdiagnosa penyakit diabetes mellitus di Indonesia.
2. Penggunaan metode klasifikasi lain seperti *Random Forest, DecisionTree*, SVM dan lain sebagainya dapat dilakukan untuk dibandingkan dengan *Naïve Bayes Classifier* agar dapat melihat model mana yang lebih akurat dalam memprediksi pasien yang terdiagnosa penyakit diabetes mellitus.

Daftar Pustaka

[1] Devi Nurul Anisa dan Jumanto, "Klasifikasi Penyakit Diabetes Melitus Mengunakan Algoritma Naive Bayes", Jurnal Dinamika Informatika, 14(1): 33–42, 2022.

[2] Maulidya Dwi Nurmalasari, Kusrini, dan Sudarman, "Komparasi Algoritma Naïve Bayes Dan K-Nearest Neighbor Untuk Membangun Pengetahuan Diagnosa Penyakit Diabetes", Jurnal Komputasi dan Informatika, 5(1): 52–59, 2021.

[3] Naisah Marito Putri dan Betha Nurina Sari, "Komparasi Algoritma KNN Dan Naive Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Melitus", Jurnal Sains dan Manajemen, 10(1): 45–57, 2022.

[4] Aris Faizal, dan Benyamin, "Penerapan Data Mining Untuk Identifikasi Penyakit Diabetes Melitus Dengan Menggunakan Metode Klasifikasi", Jurnal Sistem Komputer dan Sistem Informasi 1(1): 1–6, 2019.

[5] Latifah Uswatun Khasanah,,Yuki Novia Nasution, dan Fidia Deny Tisna Amijaya, "Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier", Jurnal Immiah Matematika, 1(1): 41–50, 2022.

[6] Minyechil Alehegn, Rahul Raghvendra Joshi, and Preeti Mulay, "Diabetes Analysis and Prediction Using Random Forest, KNN, Naïve Bayes and J48: An Ensemble Approach", International Journal of Scientific and Technology Research, 8(9): 1346–54, 2019.

[7] Abdulhakim Salum Hassan, I. Malaserene, and A. Anny Leema, "Diabetes Mellitus Prediction Using Classification Techniques", International Journal of Innovative Technology and Exploring Engineering, 9(5): 2080–84, 2020.

[8] S. Saru and S Subashree, "Analysis and Prediction of Diabetes Using Machine Learning", International Journal of Emerging Technology and Innovative Engineering, 5(4): 167–75, 2019.

[9] Gajendra Sharma and Umesh Hengaju, "Performance Analysis of Data Mining Classification Algorithm to Predict Diabetes", Journal International Advanced Networking and Applications, 12(1): 4509–18, 2020.

[10] B. A. Omodunbi, et al, "Development of a Diabetes Melitus Detection and Prediction Model Using Light Gradient Boosting Machine and K-Nearest Neighbour", Journal of Engineering and Environmental Sciences, 3(1), 2021.

[11] Abdulazeez Abdulqadir and Zebari, "Data Mining Classification Techniques for Diabetes Prediction", Qubahan Academic Journal, Vol. 1 No. 2, DOI: 10.48161/qaj.v1n2a55, 2021.

[12] Rastogi and Bansal, "Diabetes Prediction Model Using Data Mining Techniques", Journal Measurement: Sensors, 25, 2023.

[13] Maniruzzaman et al, "Classifcation and prediction of diabetes disease using machine learning paradigm", Journal Health Information Science and Systems, 8(7): 1-14. https://doi.org/10.1007/s13755-019-0095-z, 2020.

[14] Tigga and Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods", International Conference on Computational Intelligence and Data Science,167: 706-716, 2019.

[15] Hasan et al, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers", IEEE Access, 8 (2020) 7616-76531, 2020.

[16] Mehmet Akturk, "Diabetes Dataset", diakses daring pada https://www.kaggle.com/datasets/mathchi/diabetes-data-set, 8 Febuari 2023.