TESIS

KOMPARASI ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK MEMBANGUN PENGETAHUAN DIAGNOSA PENYAKIT DIABETES

Disusun oleh:

Nama : **Maulidya Dwi Nurmalasari**
NIM : **20.52.1302**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2022

TESIS

KOMPARASI ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK MEMBANGUN PENGETAHUAN DIAGNOSA PENYAKIT DIABETES

COMPARISON OF NAÏVE BAYES AND MODIFIED K-NEAREST NEIGHBOR (KNN) ALGORITHM TO BUILD DIAGNOSIS KNOWLEDGE OF DIABETES

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama	**: Maulidya Dwi Nurmalasari**
NIM	**: 20.52.102**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2022

HALAMAN PENGESAHAN

KOMPARASI ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK MEMBANGUN PENGETAHUAN DIAGNOSA PENYAKIT DIABETES

COMPARISON OF NAÏVE BAYES AND MODIFIED K-NEAREST NEIGHBOR (KNN) ALGORITHM TO BUILD DIAGNOSIS KNOWLEDGE OF DIABETES

Dipersiapkan dan Disusun oleh

Maulidya Dwi Nurmalasari

20.52.1302

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Rabu, 02 Februari 2022

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 02 Februari 2022
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

KOMPARASI ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK MEMBANGUN PENGETAHUAN DIAGNOSA PENYAKIT DIABETES

COMPARISON OF NAÏVE BAYES AND MODIFIED K-NEAREST NEIGHBOR (KNN) ALGORITHM TO BUILD DIAGNOSIS KNOWLEDGE OF DIABETES

Dipersiapkan dan Disusun oleh

Maulidya Dwi Nurmalasari

20.52.1302

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Rabu, 02 Februari 2022

Pembimbing Utama

Prof.Dr. Kusrini, M.Kom.
NIK. 190302106

Anggota Tim Penguji

Prof. Dr. Ema Utami, S.Si., M.Kom.
NIK. 190302037

Dr. Arief Setyanto, S.Si., M.T.
NIK. 190302036

Pembimbing Pendamping

Sudarmawan, M.T.
NIK. 190302035

Prof.Dr. Kusrini, M.Kom.
NIK. 190302106

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta,02 Februari 2022
Direktur Program Pascasarjana

Prof.Dr. Kusrini, M.Kom.

NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : **Maulidya Dwi Nurmalasari**
NIM : **20.52.1302**
Konsentrasi : **Business Intelligence**

Menyatakan bahwa Tesis dengan judul berikut:
KOMPARASI ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK MEMBANGUN PENGETAHUAN DIAGNOSA PENYAKIT DIABETES
Dosen Pembimbing Utama : Prof.Dr. Kusrini, M.Kom.
Dosen Pembimbing Pendamping : Sudarmawan, M.T.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 02 Februari 2022
Yang Menyatakan,

Maulidya Dwi Nurmalasari

v

HALAMAN PERSEMBAHAN

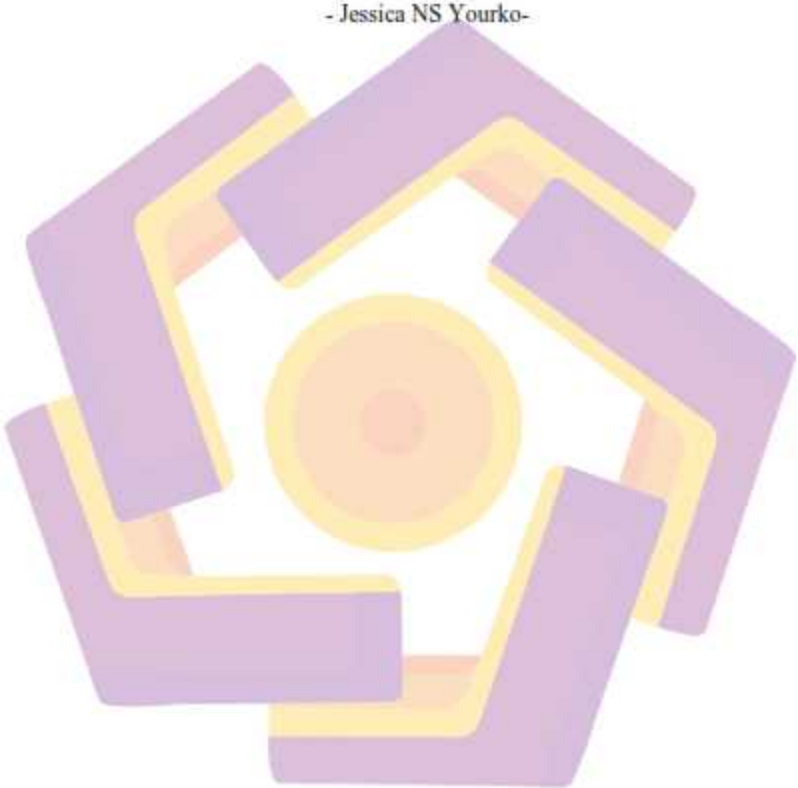
Puji dan syukur penulis ucapkan kepada Allah SWT atas anugerah dan nikmat yang tak terkira sehingga penulis dapat menyelesaikan karya tulis ini. Pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Eyang Pungki dan Eyang Susanti yang sudah memberikan dukungan , doa , dan motivasi untuk Saya dalam pengerjaan thesis ini.
2. Kedua orang tua Saya terutama Ibu Saya , Almarhum Ayah Saya dan Ayah sambung Saya Pak Agus, Kakak Saya Abi, serta Adik Saya Ira yang selalu memberikan motivasi, doa dan dukungan dalam pengerjaan thesis ini.
3. Bapak M. Suyanto, Prof., Dr., MM. selaku Rektor Universitas Amikom Yogyakarta.
4. Prof.Dr. Kusrini, M.Kom. dan Pak Sudarmawan, M.T. selaku pembimbing 1 dan pembimbing 2, yang telah banyak memberikan masukan, arahan dan motivasi kepada Saya.
5. Kucing Saya Milea, Molly, Gembul, Bagong, Gogon, Paijo, Coco, Tembong,Gembil, Bonbon, dan Gemoy yang menjadi motivasi terbesar Saya dalam pengerjaan thesis ini.
6. Teman-teman kelas MTI Angkatan 24 yang selama 3 semester telah berjuang dan belajar bersama. Terimakasih atas segala doa dan dukungannya, semoga kita semua menjadi orang-orang yang berguna bagi nusa, bangsa dan agama.
7. Sahabat-sahabat seperjuangan saya Aljinor, Frizka, Hendrik, Mbak Dhana,Mas Jati serta seluruh pihak yang telah membantu kelancaran thesis ini yang tidak dapat disebutkan satu-persatu.
8. Sahabat Saya Lola, Egga, Maryani dan Anggi yang selalu mengingatkan dan memotivasi untuk mengerjakan thesis Saya.

HALAMAN MOTTO

"Miliki cukup keberanian untuk memulai dan cukup hati untuk menyelesaikan."

- Jessica NS Yourko-



KATA PENGANTAR

Puji dan syukur penulis persembahkan untuk Allah SWT yang telah memberikan rahmat, hidayah dan kekuatan sehingga penulis dapat menyelesaikan thesis dengan judul Komparasi Algoritma Naïve Bayes Dan K-Nearest Neighbor (Knn) Untuk Membangun Pengetahuan Diagnosa Penyakit Diabetes ini sesuai dengan waktu yang diharapkan. Thesis ini disusun sebagai salah satu syarat kelulusan bagi setiap mahasiswa Universitas Amikom Yogyakarta. Selain itu juga merupakan suatu bukti bahwa mahasiswa telah menyelesaikan kuliah jenjang program Magister dan untuk memperoleh gelar Magister Komputer. Penulis sangat menyadari bahwa dalam penulisan thesis ini sangat jauh dari kesempurnaan. Walaupun sangat sederhana, pastinya penulis akan mengalami berbagai macam kesulitan. Oleh karena itu dalam kesempatan ini, penulis mengucapkan terima kasih kepada:

1.Bapak M.Suyanto, Prof., Dr., MM. selaku Rektor Universitas Amikom Yogyakarta.

2.Prof.Dr. Kusrini, M.Kom selaku Direktur Program Pascasarjana Universitas Amikom Yogyakarta.

3. Prof.Dr. Kusrini, M.Kom. dan Pak Sudarmawan, M.T. selaku pembimbing 1 dan pembimbing 2

4.Bapak dan Ibu Dosen Universitas Amikom Yogyakarta yang telah banyak memberikan ilmunya selama penulis kuliah.

5.Kedua orang tua, Ayah sambung dan saudara-saudara yang selalu mendukung penulis dalam segala hal.

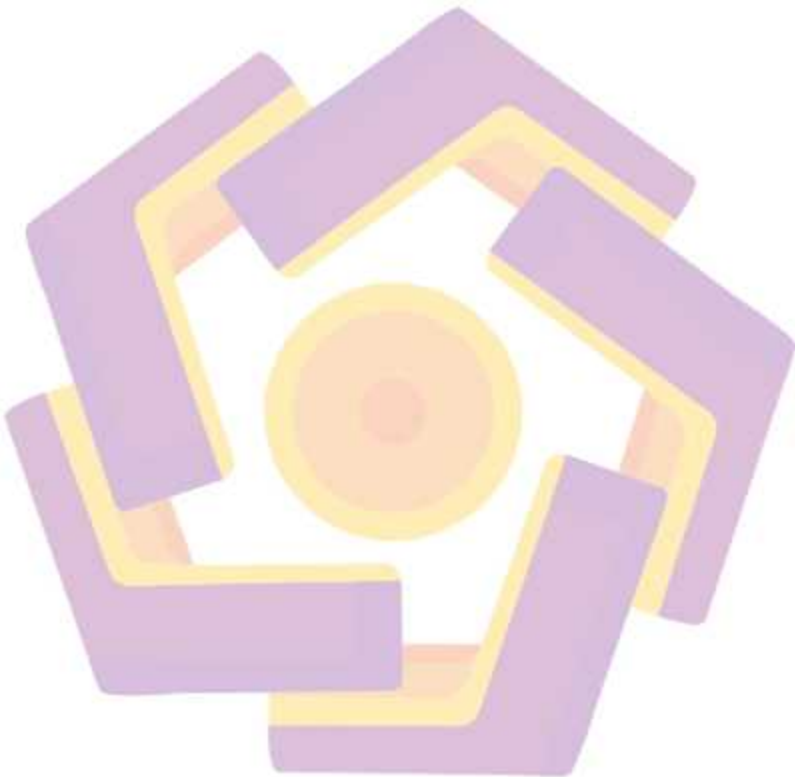
Penulis menyadari bahwa dalam pembuatan thesis ini masih banyak kekurangan dan kelemahannya. Oleh karena itu penulis berharap kepada semua pihak agar dapat menyampaikan kritik dan saran yang membangun untuk menambah kesempurnaan thesis ini. Namun penulis tetap berharap thesis ini akan bermanfaat bagi semua pihak yang membacanya.

Yogyakarta, 12 Februari 2022

Penulis

DAFTAR ISI

DAFTAR TABEL



DAFTAR GAMBAR

INTISARI

Diabetes disebabkan oleh kekurangan hormon insulin yang dikeluarkan oleh pankreas untuk menurunkan kadar gula darah. Faktor-Faktor yang memicu terjadinya penyakit diabetes berasal dari berbagai faktor seperti kombinasi faktor genetik dan lingkungan.Munculnya fenomena munculnya berbagai outlet brand minuman bisa menjadi salah satu pemicu kadar gula darah pada manusia. Kadar gula darah normal pada tubuh berkisar antara 70-130 mg/dL pada saat sebelum makan, kurang dari 180 mg/dL pada saat dua jam setelah makan, kurang dari 100 mg/dL pada saat setelah tidak makan atau berpuasa selama delapan jam, dan 100-140 mg/dL pada saat menjelang tidur.

Masalah kompleksitas pengetahuan dan data dalam sistem diagnosa ini diatasi dengan menggunakan metode *Naïve Bayes* dan *K Nearest Neighbor(KNN)*. Proses penentuan keputusan dalam sistem diagnosa penyakit diabetes ini diawali dengan menggunakan data dari UCI Machine Learning untuk selanjutnya dilakukan perhitungan dengan menggunakan 2 metode tersebut dan dilakukan evaluasi menggunakan Confusion Matrix

Hasil akhir dari penelitian ini adalah sebuah perbandingan sistem prediksi dengan *Naïve Bayes* dan *K Nearest Neighbor(KNN)* untuk melakukan keputusan atau diagnosa penyakit diabetes dengan performa metode terbaik, yang dapat dijadikan diagnosa penyakit diabetes dan pengaruh normalisasi minimax pada perfoma metode.

Kata kunci: *Naïve Bayes*, *K Nearest Neighbor(KNN)*, *Minimax*, Diabetes

ABSTRACT

Diabetes is caused by a deficiency of the hormone insulin, which is secreted by the pancreas to lower blood sugar levels. The factors that trigger the occurrence of diabetes are derived from various factors such as a combination of genetic and environmental factors. The phenomenon of the emergence of various beverage brand outlets can be one of the triggers for blood sugar levels in humans. Normal blood sugar levels in the body range from 70-130 mg/dL before eating, less than 180 mg/dL two hours after eating, less than 100 mg/dL after not eating or surviving for eight hours, and 100-140 mg/dL at bedtime.

Knowledge and data problems in this diagnostic system are overcome by using the Naïve Bayes and K Nearest Neighbor (KNN) methods. The decision-making process in the diabetic diagnosis system begins by using data from UCI Machine Learning for further calculations using those 2 methods and the evaluation is carried out using the Confusion Matrix.

The final result of this study is a comparison of the prediction system with Naïve Bayes and K Nearest Neighbor (KNN) to make decisions or diagnose diabetes by using the best method, which can be used as a diagnosis of diabetes and the effect of minimax normalization on the performance method.

Keyword: Naïve Bayes, K Nearest Neighbor(KNN), Minimax, Diabetes

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Diabetes merupakan salah satu penyakit yang sangat menakutkan bagi sebagian besar orang didunia. Menurut Kementrian Kesehatan Republik Indonesia (Kemenkes), Diabetes disebabkan oleh kekurangan hormon insulin yang dikeluarkan oleh pankreas untuk menurunkan kadar gula darah. Kombinasi faktor genetik dan lingkungan juga memicu terjadinya diabetes mellitus type 2. Kadar gula darah normal pada tubuh berkisar antara 70-130 mg/dL pada saat sebelum makan, kurang dari 180 mg/dL pada saat dua jam setelah makan, kurang dari 100 mg/dL pada saat setelah tidak makan atau berpuasa selama delapan jam, dan 100-140 mg/dL pada saat menjelang tidur (Putra, 2019).

Kemenkes juga menyatakan ada beberapa faktor resiko penyakit diabetes yang bisa diubah diantaranya adalah kegemukan (Berat badan lebih /IMT > 23 kg/m2) dan lingkar perut (Pria > 90 cm dan Perempuan > 80cm), kurang olahraga/aktivitas fisik, Dislipidemia (Kolesterol HDL ≤ 35 mg/dl, trigliserida ≥250 mg/dl, riwayat penyakit jantung, Hipertensi/ Tekanan darah Tinggi (> 140/90 mmHg) dan diet tidak seimbang. Diabetes tidak hanya disebabkan oleh kadar gula yang tinggi pada darah, ternyata juga ada faktor lain yang meningkatkan resiko penyakit diabetes. Faktor lain inilah yang perlu diperhatikan. Jika menggunakan sistem pakar, pakar juga tidak akan sepenuhnya yakin berapa besar nilai faktor dari setiap faktor yang mempengaruhi. Ketidakyakinan pakar ini tentu akan

menurunkan akurasi jika menggunakan sistem pakar sebagai pembangun pengetahuan mengenai diabetes.

Penelitian yang dilakukan oleh (Pria et all,2020) mengimplementasikan sistem pakar deteksi penyakit diabetes mellitus (DM) dengan menggunakan metode forward chaining dan certainty factor yang dimana tidak tercantum asal data gejala yang didapat dari pakar atau dengan observasi dan gejala yang masih sedikit untuk mendeteksi penyakit diabetes. Peneliti sebelumnya yang dilakukan oleh (Sri Handayani,2020) melakukan penelitian tentang diagnosa penyakit diabetes dengan metode forward chaining dengan terdapat banyak gejala yang tidak bisa dipastikan memang gejala dari penyakit diabetes.

Dalam penelitiannya (Fauzia et all, 2019) melakukan komparasi terhadap kejang pada penyakit epilepsi dengan tes EGG menggunakan 3 metode yaitu Naïve Baues, Random Tree Forest dan K-Nearest Neighbor(KNN) dengan akurasi terbaik adalah KNN 92,7%, Random Tree Forest 86,6%) dan Naïve Bayes sebesar 55,6%

Penelitian yang dilakukan oleh (Dayanand & Neethi , 2020) melakukan penelitian pada penyakit Thyroid dengan mengunakan beberapa algoritma yaitu Naïve Bayes dengan accuracy 0,85 , SVM dengan accuracy 0,82, dan KNN dengan accuracy sebesar 0.85. Selanjutnya penelitian yang dilakukan oleh (Bindiya et all, 2020) melalukan penelitian pada penyait diabete mellitus dengan menggunakan beberapa algoritma machine learning yaitu KNN dengan accuracy 100% , Naïve Bayes dengan accuracy 86%, Logistic Regression dengan accuracy 76%,Random Forest Classifier dengan accuracy 82% ,Decision Tree Classifier dengan accuracy 80%, dan Support vector Machine dengan accuracy 80%.

Dalam penelitiannya (Lia Dwi et all, 2021) mengimplementasikan beberapa algoritma untuk menentukan tingkat keberhasilan Immunotherapy untuk pengobatan penyakit kulit dengan pengujjian data denga aplikasi Weka..Hasil nilai accuracy K-Nearest Neighbor 91,1111%, dan Naïve Bayes dengan accuracy 82,2222% selanjutnya pengujian F1-Score dengan hasil KNN yaitu 0,909 % dan Naïve Bayes 0, 809%, selanjutnya pengujian untuk Kappa Statistic dengan hasil KNN yaitu 0,722 dan Naïve Bayes dengan hasil 0,39 , dan pengujian MAE dengan hasil KNN yaitu 0,097 dan Naïve Bayes yaitu 0,262.Data yang digunakan pada penelitian ini berasal dari UCI machine learning.

Penelitian yang dilakukan oleh (Annida et all, 2020) melakukan penelitian untuk mendeteksi penyakit daun pada tanaman padi dengan hasil penelitian yaitu 3 macam model yaitu Model Overfit (Random Forest, Decission Tree dan Naive Bayes). Model Underfit (SVM) dan Good Models (KNN). Jadi metode terbaik diantara kelima algoritma yaitu metode KNN dengan nilai akurasi 87%, karena model ini konsisten baik pada kedua evaluasi. KNN tidak terbukti memiliki masalah overfiting karena secra konsisten berkinerja baik pada data train dan data test. Pada penelitian ini dilakukan pengujian yaitu dengan accuracy, f1-score,MCC,training time, dan predict time. Data set yang digunakan berasal dari PIMA Indian Diabetes Dataset

Penelitian sebelumnya dilakukan oleh (Harianto & Didi , 2020) dengan melakukan penelitian sistem pendukung keputusan dengan menaikkan jumlah peserta didik dengan menggunakan Algoritma C45 , Naïve Bayes, dan K-Nearest Neighbor. Hasil dari penelitian ini menunjukkan accuracy dari algoritma Naïve

Bayes memiliki accuracy paling tinggi yaitu 86,50%, lalu KNN dengan accuracy 80%, dan algoritma c4.5 dengan accuracy sebesar 79,50%. Penelitian ini juga melakukan oengujian precission dengan Naïve bayes dengan nilai precission paling tiggi yaitu 86,14% , algoritma C.45 dengan 78,64%, dan KNN dengan 73,44 %. Peneliti juga melakukan pengujian recall dengan hasil pada algoritma KNN sebesar 94%, Naïve Bayes 87%, dan algoritma c45 dengan 81%. Data set yang digunakan berasal dari repositori UCI (Universitas California, Invene). Data set yang akan digunakan adalah UCI Machine Learning Repository.

Peneliti yang dilakukan pleh (Sumit & Mahesh, 2020) dengan melakukan prediksi penyakit jantung dengan menggunakan beberapa model Deep Learning Neural Network. Hasil dari penelitian menunjukan bahwa akurasi pada Logistic Regression sebesar 82,25%, KNN sebesar 90,16%, SVM sebesar 81,97%,Naïve Bayes sebesar 82,25%,Hyper-paramerer optimization(Talos) sebesar 90,78%, dan Random forest sebsar 85,15%.

Dalam penelitian yang dilakukan oleh (Shehsaib et all, 2020) melakukan penelitian untuk memprediksi kejang pada penyakit epilepsy. Hasil dari penelitian menunjukan bahwa akurasi pada model algoritma KNN sebesar 95,165%, Naïve Bayes sebear 95,739%,Linear Classification Model sebsar 58,339, Discriminant 82,086 %, SVM sebesar 81,573%, dan Decision Tree sebsar 94,26%. Data set yang digunakan pada penelitian ini yaitu Kaggle.

Penelitian yang dilakukan oleh (Theshay,2020) melakukan penelitian untuk mendeteksi kanker payudara dengan menggunakan improvasi dari KNN untuk melakukan komparasi terhadap perfomance. Hasil dari penggunakan parameter

default yaitu 90,10% dan KNN dengan menggunakan hyper -parameter sebesar 94,35%. Data set pada penelitian ini berasal dari Wisconsin's breast cancer data repository sejumlah 569 data set.

Penelitian yang dilakukan oleh Fida Maisa Hana (2020) melakukan penelitian klasifikasi menggunakan metode algoritma Decision Tree C4.5 pada penderita penyakit diabetes. Hasil dari pengujian menghasilkan akurasi yang cukup besar yaitu 97,12 % Precision sebesar 93,02% %, dan Recall sebesar 100,00%. Pada penelitian ini menggunakan dataset yang sama dengan penelitian yang dilakukan namun memiliki perbedaan penelitian ini tidak menggunakan normalisasi

Dari beberapa uraian diatas, penulis tertarik ingin melakukan penelitian menggunakan algortitma *Naïve Bayes* dan *K Nearest Neighbor(KNN)* dengan penyakit diabetes sebagai objek penelitian. Pemilihan algoritma *Naïve Bayes* dan K *Nearest Neighbor* karena memiliki kinerja yang terbaik dibandingkan dengan algoritma yang lain berdasarkan dari hasil penelitian – penelitian sebelumnya.Data set yang akan digunakan pada penelitian ini dari UCI Machine Learning Repository dengan mendownload pada website dengan link https://archive.ics.uci.edu/ml/machine-learning-databases/00529/ berdasarkan dengan penelitian sebelumnya yang dilakukan oleh Zhi-Hua Zhou and Yuan Jiang(2004) agar mendapat data yang dapat dipercaya Selain itu, penelitian yang dilakukan dengan sistem pakar yang terdapat pada gejala penyakit diabetes masih belum lengkap dan ada yang menggunakan gejala yang banyak tetapi tidak terbukti sumbernya dari mana sehingga diharapkan dengan adanya penelitian untuk pembentukan pengetahuan diagnosis penyakit diabetes dapat membantu para pakar

dalam menentukan pengetahuan tentang penyakit diabetes. Pada penelitin – penelitian sebelumnya melakukan pengujian dengan menggunakan coffusion matrix salah satu perhitungan yang digunakan yaitu F1-Score yang tenyata memiliki fungsi untuk memahami algoritma mana yang lebih cocok dengan data berdasarkan nilai Precision dan Recall Penelitian yang akan dilakukan penelitian adalah melakukan komparasi algoritma Naïve Bayes dan K-Nearest Neighbor untuk membangun pengetahuan tentang diagnosa penyakit diabetes dengan menghitung nilai perfoma kedua model algoritma tersebut pada cofunsion matrix yaitu akurasi, precision, recall, F1 Score, dan AUC dengan menggunakan scenario yaitu menggunakan normalisasi Minimax dan tanpa normalisasi Minimax untuk melakukan eksperimen apakah akan berpengaruh pada hasil klasifikasi.

1.2. Rumusan Masalah

Berikut beberapa rumusan masalah pada penelitian ini yaitu :

a. Metode algoritma manakah dari *Naïve Bayes* dan *K- Nearest Neighbor (KNN)* yang memiliki nilai perfoma yang baik untuk membangun pengetahuan diagnosa penyakit diabetes?

b. Apakah optimasi algoritma Minimax berperngaruh terhadap algoritma Naïve Bayes dan K-Nearest Neighbor (KNN)?

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini adalah :

a. Algoritma yang digunakan dalam penelitian ini yaitu *Naïve Bayes* dan *K-Nearest Neighbor*(K-NN).

b. Data set yang akan digunakan dari UCI *Machine Learning Repository* dengan mendownload pada website dengan link https://archive.ics.uci.edu/ml/machine-learning-databases/00529/
c. Model pengukuran validasi performa model algoritma prediksi menggunakan metode atau teknik k fold 5.
d. Nilai Perfoma akan diukur atau dihitung menggunakan perfoma Confusion Matrix yaitu Accuracy, Precision, Recall, dan F1- Score.
e. Pre-Processing menggunakan tools Rapid Miner.
f. Komparasi yang dilakukan pada algoritma Naïve Bayes dan K Nearest Neighbor(KNN) dilakukan secara manual

1.4. Tujuan Penelitian

Bagian ini memuat penjelasan secara spesifik:

a. Mendapatkan nilai perfoma yang terbaik berdasarkan hasil penelitian dari algoritma Naïve Bayes dan K- Nearest Neighbor(KNN) untuk membangun pengetahuan diagnosa penyakit diabetes.
b. Mengetahui apakah optimasi dengan algoritma Minimax berpengaruh terhadap Algoritma Naïve Bayes dan K-Nearest Neighbor(KNN).

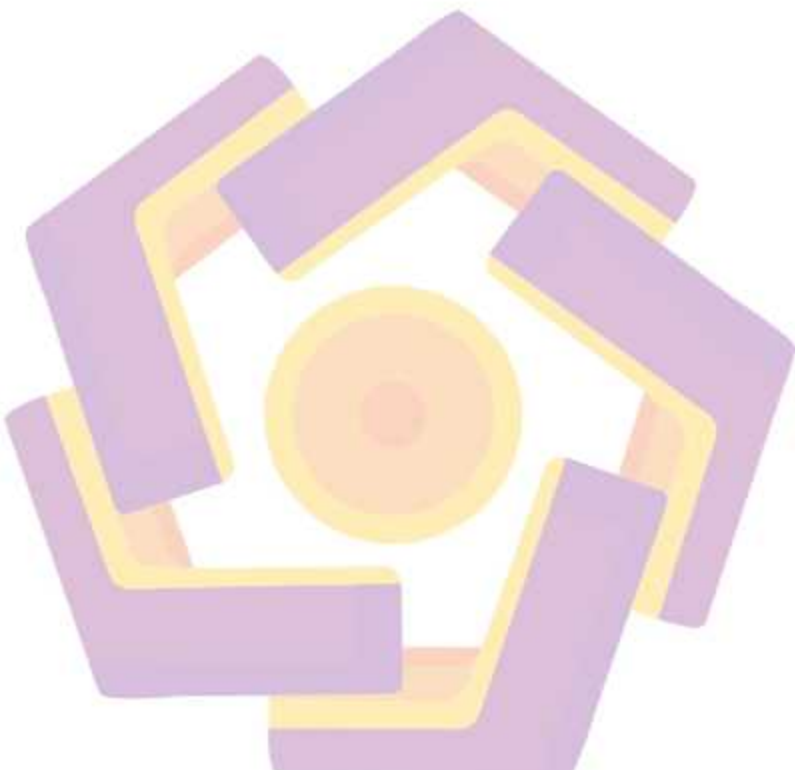
1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah :

a. Dapat menjadi pendoman pengembangan aplikasi untuk membangun pengetahuan diagnosa penyakit diabetes.

b. Dapat membantu orang untuk mengetahui pengetahuan tentang penyakit diabetes lebih dalam

c. Dapat memberikan rekomendasi algoritma yang dapat digunakan dalam membangun pengetahuan diagnosa penyakit diabetes.



BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Dalam melakukan penelitian diperlukan rujukan atau tinjauan pustaka sebagai acuan dalam melakukan penelitian. Penelitian yang dilakukan oleh Fauzia et all (2020) membahas membandingkan algoritma Naïve Bayes, KNN, dan Random Tree Forest pada penyakit kejang epilepsy dengan hasil Klasifikasi terbaik adalah metode KNN (akurasi: 92,7%, presisi: 82,5%, sensitivitas 73,2% dan spesifisitas: 96,7%), lalu Random Tree Forest (akurasi: 86,6%, presisi 68,2%, sensitivitas: 42,2%, dan spesifisitas: 96,7%) dan pengklasifikasi naïve bayes (akurasi: 55,6%, presisi: 25,3%, sensitivitas: 80,3%, dan spesifisitas: 50,4%). Waktu training data naïve bayes 0,166030 detik, sedangkan waktu pelatihan random tree forest 2,4094 detik dan KNN paling lambat dalam pelatihan yaitu 4,789 detik.

Pada penelitian yang dilakukan oleh Dayanand & Neethi (2020) membahas tentang melakukan klasifikasi dengan beberapa algoritma machine learning yaitu SVM , Naïve Bayes dan KNN untuk memprediksi penyakit thyroid. Hasil klasifikasi dari beberapa metode didasarkan pada keakuratan dan perfoma model algoritma. Hasil dari akurasi menggunakan SVM 0.82, Naïve Bayes 0.83 dan KNN 0.85.

Pada penelitian dilakukan oleh Bindiya et All(2020) membahas tentang menerapkan beberapa metode yaitu KNN,Naïve Bayes, Logistic Regression,

Random Forest Classifier , Decision Tree Classifierm dan SVM untuk prediksi penyakit Diabetes Mellitus. Hasil dari penerapan algoritma machine learning yaitu KNN dengan accuracy 100% , Naïve Bayes dengan accuracy 86%, Logistic Regression dengan accuracy 76%,Random Forest Classifier dengan accuracy 82% ,Decision Tree Classifier dengan accuracy 80%, dan Support vector Machine dengan accuracy 80%.

Pada penelitian yang dilakukan oleh Lia Dwi et all(2020) membahas tentang mengimplementasikan algoritma Naïve Bayes dan K-Nearest Neighbor dalam menentukan tingkat keberhasilan Immunotherapy untuk pengobatan penyakit kanker kulit. Klasifikasi yang didapat dari kolaborasi pengujian menggunakan metode machine learning yaitu naïve bayes dan k-nearest neighbor menunjukkan bahwa akurasi terbaik untuk keberhasilanimmunotherapy ketika menggunakan metode K-Nearest Neighbor sebesar 91,1111%. Sedangkan menggunakan metode naïve bayes rata-rata akurasi sebesar 82,2222% .

Pada penelitian yang dilakukan oleh Annida et all(2020) membahas tentang deteksi penyakit daun pada tanaman padi dengan menggunakan algoritma Decision Tree, Random Forest, Baive Bayes, SVM, dan KNN. Hasil penelitian yang dibahas mengenai machine learning clasification penyakit tanaman padi menyimpulkan dari hasil penelitian yaitu 3 jenis hasil; Overfit Models yaitu algoritma Random Forest, Decission Tree dan Naive Bayes, Underfit Models yaitu algoritma SVM dan Good Models yaitu algoritma KNN. Metode terbaik diantara kelima tersebut yaitu metode KNN dengan nilai akurasi 87%.

Pada penelitian dilakukan oleh Harianto & Didi (2020) membahas tentang membandingkan algoritma Komparasi Algoritma Decision Tree, Naive Bayes Dan K-Nearest Neighbor Untuk Memprediksi Mahasiswa Lulus Tepat Waktu. Hasil eksperimen terhadap accuracy dengan algoritma C4.5 adalah sebesar 79,50% dan kemudian dengan algoritma Naïve Bayes di dapat nilai accuracy sebesar 86,50% serta juga dengan perhitungan accuracy terhadap algoritma k-nearest neighbor di dapat nilai sebesar 80,00%. Dari perhitungan dengan metode algoritma Naïve Bayes, diketahui atribut yang paling tinggi untuk faktor penentu orang tua siswa mendaftarkan anaknya ke sekolah SMP Cenderawasih adalah atribut umur, yang berumur 31 sampai 50 tahun, kemudian atribut faktor, karena tidak masuk negeri, atribut transportasi, menggunakan motor, kemudian atribut jarak, dimana rumah mereka diatas 1 km, serta atribut informasi, dimana informasi tentang SMP Cenderawasih di dapat dari teman atau saudara.

Pada penelitian yang dilakukan oleh Sumit & Mahesh (2020) membahas tentang prediksi penyakit jantung dengan menggunakan Deep Learning Neural Network Model. Hasil dari penelitian menunjukan bahwa akurasi pada Logistic Regression sebesar 82,25%, KNN sebesar 90,16%, SVM sebesar 81,97%,Naïve Bayes sebesar 82,25%,Hyper-paramerer optimization(Talos) sebesar 90,78%, dan Random forest sebesar 85,15%.

Pada penelitian yang dilakukan oleh Shehzaib et all(2020) membahas tentang komparasi KNN, Naïve Bayes, Linear Classification Mode, Discriminan Analysis Model, SVM, Decision Tree untuk prediksi kejang epilepsy. Hasil dari enam pengklasifikasi, Naive Bayes memiliki akurasi 95.739% dan K-Nearest

Neighbor (KNN) menunjukkan akurasi terbaik kedua dengan nilai 95.165%, Decision Tree menunjukkan akurasi terbaik ketiga sebesar 94,260%. Discriminan Analysis Model dan SVM menunjukkan akurasi masing-masing 82.026% dan 81.573%. Linear Classification Model dengan tingkat akurasi terendah sebesar 58,33%

Pada penelitian yang dilakukan oleh Annida et all(2020) membahas optimasi KNN untuk deteksi kanker payudara. Model KNN yang dioptimalkan untuk prediksi kanker payudara menggunakan pendekatan pencarian grid untuk mencari hyper-parameter terbaik. Hasil dari perbandingan yang dilakukan KNN dengan parameter default adalah 90,10% dan KNN dengan hyper-parameter adalah 94,35%.

Pada penelitian yang dilakukan oleh Fida Maisa Hana (2020) melakukan klasifikasi menggunakan metode algoritma Decision Tree C4.5 pada penderita penyakit diabetes. Hasil dari pengujian menghasilkan akurasi yang cukup besar yaitu 97,12 % Precision sebesar 93,02% %, dan Recall sebesar 100,00%.

2.2. Keaslian Penelitian

Tabel 2.1. Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Epileptic Seizure Detection in EEGs by Using Random Tree Forest, Naïve Bayes and KNN Classification	Fauzia P. Lestari, Mohammad Hackal, Rizki Edmi Edison, Fikry Ravi Fauzy, Siti Nurul Khotimah and Freddy Haryanto, Journal of Physics: Conference Series,2019	Penelitian ini bertujuan untuk membandingkan algoritma Naïve Bayes Dan KNN pada penyakit kejang epilepsi	Machine Learning memberikan performa yang baik untuk mengklasifikasikan data EEG kejang dan non-kejang. Antara Klasifikasi KNN, klasifikasi Naïve bayes dan Random Tree Forest. Klasifikasi terbaik adalah metode KNN (akurasi: 92,7%, presisi: 82,5%, sensitivitas 73,2% dan spesifisitas: 96,7%), bukan dari Random Tree Forest (akurasi: 86,6%, presisi 68,2%, sensitivitas: 42,2%, dan spesifisitas: 96,7%) dan pengklasifikasi naïve bayes (akurasi: 55,6%, presisi: 25,3%, sensitivitas: 80,3%, dan spesifisitas: 50,4%). Waktu training data naïve bayes 0,166030 detik, sedangkan waktu pelatihan pohon acak Hutan 2,4094 detik dan KNN paling lambat dalam pelatihan yaitu 4,789 detik.	Penambahan 2 perhitungan untuk pengujian perfoma	Penambahan perhitungan perfoma yaitu dengan 2 perhitungan Recall, dan F1-Score, pada penelitian ini juga akan dilakukan proses preparasi data untuk melakukan analisis missing value, inkonsistensi dan transformasi data
2	*Thyroid Disease Prediction Using Feature Selection And Machine Learning Classifiers*	Dr. Dayanand Jamkhandikar , Neethi Priya, Johan, The International journal of analytical and	Penelitian ini bertujuan untuk melakukan klasifikasi dengan beberapa algoritma machine	Kesimpulannya adalah hasil klasifikasi dari beberapa metode didasarkan pada keakuratan dan perfoma model algoritma. Hasil dari akurasi menggunakan SVM 0.82, Naïve Bayes 0.83 dan KNN 0.85.	Penambahan perhitungan untuk pengujian hasil dari klasifikasi	Pada penelitian ini proses preprocessing data tidak hanya menggunakan feature selection tetapi juga manambahkan analsisis

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
		experimental modal analysis, 2020	learning untuk memprediksi penyakit thyroid.			Penambahan perhitungan perfoma yaitu dengan 4 perhitungan Accuracy, Precision, Recall, dan F1-Score
3	Diabetes Mellitus Prediction usingMachine Learning Algorithms	A. R. Bindiya , K. Nikhil , M. S. Sindhu Rashmi , Shafinaz Banu , International Journal for Research in Applied Science & Engineering Technology (IJRASET),2020	Penelitian ini bertujuan untuk menerapkan beberapa metode yaitu KNN,Naïve Bayes, Logistic Regression, Random Forest Classifier , Decision Tree Classifierm dan SVM untuk prediksi penyakit Diabetes Mellitus.	Kesimpulannya adalah hasil dari penerapan algoritma machine learning yaitu KNN dengan accuracy 100% , Naïve Bayes dengan accuracy 86%, Logistic Regression dengan accuracy 76%,Random Forest Classifier dengan accuracy 82% ,Decision Tree Classifier dengan accuracy 80%, dan Support vector Machine dengan accuracy 80%.	Penambahan perhitungan untuk pengujian hasil daari klasifikasi	Pada penelitian ini akan melakukan validasi pengujian dengan menggunakan k-fold, dimana algoritma akan dilakukan proses validasi pengujian sebanyak k perulangan. Penambahan perhitungan perfoma yaitu dengan 4 perhitungan Accuracy, Precision, Recall, dan F1- Score dan penambahan algoritma sebelum masuk keklasifikasi yaitu algoritma Minimax dan penambahan perhitungan pengujian perfoma yaitu Precission dan Recall
4	Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor	F. Lia Dwi Cahyanti , Windu Gata , Fajar Sarasuti, Jurnal Ilmiah Universitas	Penelitian ini bertujuan untuk mengimplementasikan algoritma Naïve Bayes	Klasifikasi yang didapat dari kolaborasi pengujian menggunakan metode machine learning yaitu naïve bayes dan k-nearest	Data set yang digunakan 90 records dapat ditambahkan beberapa datas set lagi	Pada penelitian ini proses validasi sebanyak n perulangan untuk validasi hasil yang lebih valid.

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	Dalam Menentukan Tingkat Keberhasilan Immunotherapy Untuk Pengobatan Penyakit Kanker Kulit	Batanghari Jambi,.2021	dan K-Nearest Neighbor dalam menentukan tingkat keberhasilan Immunotherapy untuk pengobatan penyakit kanker kulit.	neighbor menunjukkan bahwa akurasi terbaik untuk keberhasilan immunotherapy ketika menggunakan metode K-Nearest Neighbor sebesar 91,1111%. Sedangkan menggunakan metode naïve bayes rata-rata akurasi sebesar 82,2222% .	untuk mengetahui tingkat keberhasilan algoritma	Penambahan perhitungan perfoma yaitu dengan perhitungan Precision dan Recall. Penambahan algoritma Minimax sebelum klasifikasi data.
5	Deteksi Penyakit Daun pada Tanaman Padi Menggunakan Algoritma Decision Tree, Random Forest, Naïve Bayes, SVM dan KNN	Annida Purnamawati , Wawan Nugroho , Destiana Putri, Wahyutama Fitri Hidayat: Jurnal Nasional Informatika dan Teknologi Jaringan,2020	Penelitian ini bertujuan untuk deteksi penyakit daun pada tanaman padi dengan menggunakan algoritma Decision Tree, Random Forest, Baive Bayes, SVM, dan KNN	Kesimpulan dari hasil penelitian yang dibahas mengenai machine learning clasification penyakit tanaman menyimpulkan dari hasil penelitian yaitu 3 jenis hasil; Overfit Models yaitu algoritma Random Forest, Decission Tree dan Naive Bayes, Underfit Models yaitu algoritma SVM dan Good Models yaitu algoritma KNN. Metode terbaik diantara kelima tersebut yaitu metode KNN dengan nilai akurasi 87%,	Tidak disebutkan hasil dari 4 algoritma yang lain.	Pada penelitian ini akan dilakukan proses pre-processing data untuk mendapatkan data yang lebih baik serta melakukan validasi pengujian sebanyak n perulangan untuk mendapatkan hasil yang lebih valid Penambahan algoritma sebelum masuk keklasifikasi yaitu algoritma Minimax dan penambahan perhitungan pengujian perfoma yaitu Precission dan Recall
6	Komparasi Algortima C4.5, Naïve Bayes, dan k-Nearest	Harianto , Didi Rosiyadi, Jurnal Informatika,2020	Penelitian ini bertujuan untuk membandingkan algoritma Komparasi	Kesimpulannya adalah hasil eksperimen terhadap accuracy dengan algoritma C4.5 adalah sebesar 79,50% dan kemudian dengan algoritma Naïve Bayes di dapat	Pada penelitian ini data yang digunakan testing 200 data uji.	Pada penelitian ini akan melakukan pre-processing data melalui feature selection,

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	Neighbor Sebagai Sistem Pendukung Keputusan Menaikkan Jumlah Peserta Didik		Algoritma Decision Tree, Naive Bayes Dan K-Nearest Neighbor Untuk Memprediksi Mahasiswa Lulus Tepat Waktu	nilai accuracy sebesar 86,50% serta juga dengan perhitungan accuracy terhadap algoritma k-nearest neighbor di dapat nilai sebesar 80,00%. Dari perhitungan dengan metode algoritma Naïve Bayes, diketahui atribut yang paling tinggi untuk faktor penentu orang tua siswa mendaftarkan anaknya ke sekolah SMP Cenderawasih adalah atribut umur, yang berumur 31 sampai 50 tahun, kemudian atribut faktor, karena tidak masuk negeri, atribut transportasi, menggunakan motor, kemudian atribut jarak, dimana rumah mereka diatas 1 km, serta atribut informasi, dimana informasi tentang SMP Cenderawasih di dapat dari teman atau saudara		missing value, inkonsistensi dan transportasi data serta Penambahan algoritma sebelum masuk keklasifikasi yaitu algoritma Minimax dan penambahan perhitungan pengujian F1- Score untuk mengetahui perhitungan pengujian Precission atau Recall yang lebih baik.
7	Heart Diseases Prediction using Deep Learning Neural Network Model	Sumit Sharma, Mahesh Parmar, International Journal of Innovative Technology and Exploring Engineering (IJITEE),2020	Penelitian ini bertujuan untuk prediksi penyakit jantung dengan menggunakan Deep Learning Neural Network Model	Kesimpulannya adalah hasil dari penelitian menunjukan bahwa akurasi pada Logistic Regression sebesar 82,25%, KNN sebesar 90,16%, SVM sebesar 81,97%,Naïve Bayes sebesar 82,25%,Hyper-paramerer optimization(Talos) sebesar 90,78%, dan Random forest sebsar 85,15%.	Perhitungan perfoma hanyak pada akurasinya.	Pada penelitian yang akan dibuat membandingkan peforma melalui perbandingan 2 algoritma untuk valdiasi hasil, juga malkukan penambahan validasi pengujian menggunakan n perulangan untuk hasil yang lebih valid

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
						Penambahan algoritma sebelum masuk keklasifikasi yaitu algoritma Minimax dan penambahan perhitungan pengujian Precission, Recall,dan F1- Score
8	Comparative Analysis of Classifiers for Prediction of Epileptic Seizures	Shehzaib Shafique, Saba Sarfraz,Usman Qamar Shaikh, Aamash Nadeem and Zia Ur Rehman, Pakistan Journal of Engineering and Technology (PakJET),2020	Penelitian ini bertujuan untuk komparasi KNN, Naïve Bayes, Linear Classification Mode, Discriminan Analysis Model, SVM, Decision Tree untuk prediksi kejang epilepsi	Kesimpulannya adalah hasil dari enam pengklasifikasi, Naive Bayes memiliki akurasi 95.739% dan K-Nearest Neighbor (KNN) menunjukkan akurasi terbaik kedua dengan nilai 95.165%, Decision Tree menunjukkan akurasi terbaik ketiga sebesar 94,260%. Discriminan Analysis Model dan SVM menunjukkan akurasi masing-masing 82.026% dan 81.573%. Linear Classification Model dengan tingkat akurasi terendah sebesar 58,33%	Perhitungan perfoma hanya pada akurasinya.	Pada penelitian ini akan melakukan pre-processing data melalui feature selection, missing value, inkonsistensi dan transportasi data serta Penambahan algoritma sebelum masuk keklasifikasi yaitu algoritma Minimax dan penambahan perhitungan pengujian F1- Score untuk mengetahui perhitungan pengujian Precission atau Recall yang lebih baik.
9	An optimized K-Nearest Neighbor based breast cancer detection	Tsehay Admassu Assegie, Journal of Robotics and Control (JRC),2020	Penelitian ini bertujuan untuk optimasi deteksi kanker payudara menggunakan algoritma K-Nearest Neighbor	Kesimpulannya adalah model KNN yang dioptimalkan untuk prediksi kanker payudara menggunakan pendekatan pencarian grid untuk mencari hyper-parameter terbaik. Hasil dari perbandingan yang dilakukan KNN dengan parameter	Perhitungan perfoma hanya pada akurasinya.	Pada penelitian ini akan melakukan pre-processing data melalui feature selection, missing value, inkonsistensi dan transportasi data serta Penambahan algoritma sebelum masuk keklasifikasi

Tabel 2.1. Matriks literatur review dan posisi penelitian (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
				default adalah 90,10% dan KNN dengan hyper-parameter adalah 94,35%.		yaitu algoritma Minimax dan penambahan perhitungan pengujian F1- Score untuk mengetahui perhitungan pengujian Precission atau Recall yang lebih baik.
10	Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5	Fida Maisa Hana,Jurnal Sistem Komputer&Kecerdasan Buatan(SISKOM-B),2020	Penelitian ini bertujuan untuk klasifikasi penderita penyakit diabetes dengan algoritma Decision Tree C4.5	Kesimpulannya adalah algoritma Decision Tree C4.5 untuk klasifikasi penderita Diabetes hasil Pengujian menghasilkan akurasi yang cukup besar yaitu 97,12 % Precision sebesar 93,02% %, dan Recall sebesar 100,00%.	Perhitungan performa pada akurasi, precision dan recall dan tanpa menggunakan optimasi	Pada penelitian ini akan melakukan pre-processing data melalui feature selection, missing value, inkonsistensi dan transportasi data serta Penambahan algoritma sebelum masuk keklasifikasi yaitu algoritma Minimax dan penambahan perhitungan pengujian F1- Score untuk mengetahui perhitungan pengujian Precission atau Recall yang lebih baik.

2.3. Landasan Teori

2.3.1.Klasifikasi

Dalam bukunya (Santoso, 2007) yang dikutip dari Agus Mulyanto, 2009 klasifikasi merupakan langkah atau cara dalam upaya membentuk suatu model atau fungsi yang diguankan dalam menjelaskan atau membedakan konsep kelas data. Dengan melakukan proses ini suatu objek dapat dikenali dan dikelompokkan berdasarkan kelasnya dengan cara memperkirakan berdasarkan hasil dari model yang telah dibentuk. Klasifikasi dokumen adalah bagian proses yang penting dalam bidang sistem informasi, khusunya untuk pengetahuan bisnis.

Penelitian oleh (Han & Kamber, 2001) menjelaskan bahwa dalam proses klasifikasi terbagi menjadi dua tahapan yaitu pelatihan (learning) dan pengujian (testing). Kedua tahapan ini saling berurutan dalam proses klasifikasi, pada tahap learning (pelatihan) ini merupakan tahap di mana dilakukan proses pembelajaran terhadap suatu data yang telah diketahui kelasnya atau sering disebut data latih. Tahapan ini dimaksudkan agar komputer dapat belajar mengenal beberapa objek (data latih) berdasarkan kelasnya sehingga dihasilkan satu model klasifikasi. Kemudian tahapan kedua adalah proses testing (pengujian), tahap ini berfungsi untuk melakukan evaluasi tingkat kinerja dari model hasil dari tahap learning dengan data baru yang disebut data uji. Keluaran dari tahap ini berupa tingkat keakuratan suatu model dalam memprediksi data yang belum diketahui kelasnya (data uji). Jika nilai akurasi dari tahap testing ini tinggi, maka dapat model hasil pembelajaran pada tahap learning layak untuk digunakan dalam memprediksi data-data baru yang belum diketahui kelasnya.

Proses klasifikasi sangat erat hubungannya dengan teknik atau algoritma yang dapat belajar dan mengelompokkan data ke dalam kelas-kelasnya. Beberapa algoritma yang dapat digunakan dalam proses klasifikasi diantaranya adalah Naive Bayes Classifier, Decision Tree, Rule Based Classifier dan Jaringan Saraf Tiruan atau lebih dikenal Neural Network. Masing-masing algoritma memiliki cara kerja yang berbeda-beda dalam proses klasifikasi data, selain itu setiap algoritma juga memiliki kelebihan dan kekurangan masing-masing dalam melakukan klasifikasi pada banyak kasus (Han & Kamber 2001).

2.3.2. Naïve Bayes

Naïve Bayes merupakan sebuah metode penggolongan berdasarkan probabilitas sederhana dan dirancang untuk dipergunakan dengan asumsi bahwa antar satu kelas dengan kelas yang lain tidak saling tergantung (independen). Pada klasifikasi Naïve Bayes, proses pembelajaran lebih ditekankan pada mengestimasi probabilitas. Keuntungan dari pendekatan ini yaitu pengklasifikasian akan mendapatkan nilai error yang lebih kecil ketika data set berjumlah besar (Berry, 2006). Selain itu menurut Han and Kamber (2006) klasifikasi Naïve Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam basis data dengan jumlah yang besar.

Formulasi Naïve Bayes untuk klasifikasi menurut Prasetyo (2012) adalah sebagai berikut:

$$P(X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \quad (1)$$

Dimana:

$P(X)$ = probabilitas data dengan vector X pada kelas Y

$P(Y)$ = probabilitas awal kelas Y (*prior probability*)

$\prod_{i=1}^{q} P(X_i|Y)$ = probabilitas independen kelas Y dari semua fitur dala vector X

Nilai $P(X)$ = probabilitas dari X

Probabilitas P(X) selalu tetap sehingga dalam perhitungan prediksi nantinya dapat diabaikan dan hanya menghitung bagian $P(Y) \prod_{i=1}^{q} P(X_i|Y)$ saja dengan memilih nilai yang terbesar sebagai kelas hasil prediksi atau yang biasa dikenal dengan sebutan *Maximum A Posteriori* (MAP) dimana MAP ini dapat dinotasikan dengan:

$hMAP = arg\ arg\ (P(Y)\ \prod_{i=1}^{q} P(X_i|Y))$

Sementara probabilitas independensi $\prod_{i=1}^{q} P(Y)$ merupakan pengaruh semua fitur dari data terhadap setiap kelas Y, yang dinotasikan dengan:

$P(Y = y) = \prod_{i=1}^{q} P(X|Y) = y)$

Setiap set fitur $X = [X_1, X_2, X_3, \ldots X_q]$ terdiri atas q atribut.

2.3.3. K-Nearest Neighbor(K-NN).

Klasifikasi K-Nearest Neighbor Menurut Prasetyo (2012), algoritma Nearest Neighbor (kadang disebut K- Nearest Neighbor atau K-NN) merupakan algoritma yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain. Dekat atau jauhnya lokasi (jarak) biasanya dihitung berdasarkan jarak Euclidean dengan rumus sebagai berikut (Han and Kamber, 2006)

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{N} (diff(x_{il}, x_{jl}))^2} \quad (2)$$

Dengan :

x_{it} = data testing ke-i pada variabel ke-l

x_{ij} = data training ke-i pada variabel ke-l

$d(x_i, x_j)$ = jarak

N = dimensi data variabel bebas

$diff(x_{it}, x_{jt})$ = *difference* atau ketidaksamaan

Penghitungan nilai *difference* atau ketidaksamaan pada persamaan (1) tergantung pada tipe data yang digunakan. Menurut Prasetyo (2012), penghitungan nilai ketidaksamaan berdasarkan tipe data untuk tiap variabel dapat diringkas seperti pada Tabel 2.2.

Tabel 2.2 Ketidakmiripan Dua Data dengan Satu Atribut

Tipe Atribut Formula Jarak	Formula Jarak
Nominal	$diff_{(x_{it}x_{jt})} = \{0 \quad Jika\ x_{it} = x_{jt}\ 1 \quad Jika\ x_{it} \neq x_{jt}$
Ordinal	$diff_{(x_{it}x_{jt})} = \lvert x_{it}, x_{jt}\rvert/(n-1)$ n adalah banyaknya penkategorian dalam x
Interval atau Rasio	$diff_{(x_{it}x_{jt})} = \lvert x_{it}, x_{jt}\rvert$

2.3.4. Confusion Matrix

Dalam bukunya (Sokolova & Lapalme, 2009) *Confusion matrix* adalah salah satu metode yang digunakan untuk mengevaluasi kinerja algoritma klasifikasi. Tabel 2.3 merupakan gambaran sedehana untuk

mempermudah pemahaman tentang istilah *confusion matrix* dalam keluaran klasifikasi.

Tabel 2.3 Confusion Matrix

		Kelas Prediksi	
		Positif	Negatif
Kelas Sesungguhnya	Positif	TP	FN
	Negatif	FP	TN

Nilai *True Negative* (TN) adalah data yang di klasifikasi dengan tepat sebagai keluaran negatif atau salah. *True Positive* (TP) adalah data yang diklasifikasi dengan tepat sebagai keluaran positif atau benar. *False Positive* (FP) adalah data yang diklasifikasi dengan kurang tepat apabila keluaran berupa positif atau benar. *False Negative* (FN) adalah data yang diklasifikasi dengan kurang tepat.

$$\text{Precision} = \frac{\sum_i^n \frac{TPi}{TPi+FPi}}{n} \quad (3)$$

Persamaan (3) merupakan perhitungan rata-rata nilai *precision* yaitu dari data hasil klasifikasi seberapa banyak data yang benar antara nilai sebenarnya dengan prediksi yang diberikan oleh sistem.

$$\text{Recall} = \frac{\sum_i^n \frac{TPi}{TPi+FNi}}{n} \quad (4)$$

Persamaan (4) merupakan perhitungan rata-rata nilai *recall* yaitu dari seluruh data benar seberapa banyak data yang keluar dalam hasil klasifikasi. Evaluasi *recall* digunakan apabila lebih memilih nilai *False Positive* daripada *False Negative* (Ghoneim, 2019).

$$\text{Akurasi} = \frac{\sum_i^n \frac{TPi+TNi}{TPi+FPi+TNi+FNi}}{n} \quad (5)$$

$$\text{F-score} = \frac{\sum_i^n \frac{recall\ i+presisi\ i}{\beta(recall\ i)+(1-\beta)(presisi\ i)}}{n} \quad (6)$$

Persamaan (5) merupakan perhitungan rata-rata nilai akurasi untuk menunjukkan tingkat efektifitas per kelas dari sebuah klasifikasi (Sokolova & Lapalme, 2009). Sedangkan persamaan (6) merupakan perhitungan rata-rata nilai F-score yang merupakan nilai kombinasi dari perhitungan *recall* dan *precission*.

2.3.5. Evaluasi dan Validasi

Validasi yaitu tahap mengevaluasi akurasi prediksi dari suatu model. Bootstrap, stratified sampling, cross-validation, random sub-sampling, dan holdout adalah beberapa metode validasi yang berfungsi untuk memvalidasi sebuah model yang bersumber pada data yang diperoleh . K-fold cross validation adalah metode validasi yang memisahkan data awal secara acak kedalam k bagian yang sama-sama terbagi atau "fold" Fungsi k-fold adalah supaya tidak ada data overlapping terhadap data testing.

Menurut Fawcett (2006) Grafik Receiver Operating Characteristics (ROC) adalah teknik untuk menggambarkan, mengorganisasi dan memilih pengklasifikasi berdasarkan kinerja mereka. Kurva ROC digunakan untuk mengukur nilai Area Under Curve (AUC). Menurut Gorunescu (2011) pedoman untuk mengklasifikasikan keakuratan pengujian menggunakan nilai AUC, sebagai berikut:

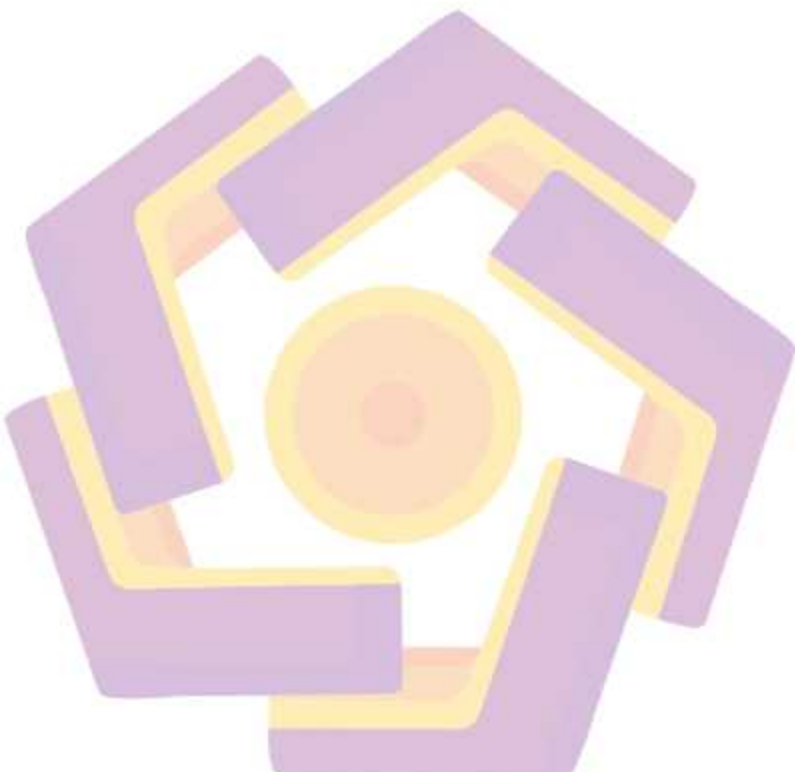
a. 0.90 - 1.00 = Excellent Classification

b. 0.80 - 0.90 = Good Classification

c. 0.70 - 0.80 = Fair Classification

d. 0.60 - 0.70 = Poor Classification

e. 0.50 - 0.60 = Failure



BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Adapun jenis, sifat dan pendekatan penelitian yang akan dilakukan pada penelitian ini sebagai berikut :

1. Jenis Penelitian Eksperimen

Penelitian ini merupakan penelitian eksperimen yaitu: meneliti perbandingan tingkat Accuracy, Precision, Recall, dan F1- Score pada metode algoritma Naïve Bayes dan K-Nearest Neighbor untuk membangun pengetahuan diagnosa penyakit diabetes.

2. Sifat Penelitian Deskriptif

Tujuan dari penelitian ini yaitu : mengetahui perbandingan tingkat Accuracy, Precision, Recall, dan F1- Score untuk membangun pengetahuan diagnosa penyakit diabetes menggunakan algoritma Naïve Bayes dan K-Nearest Neighbor dan mengetahui apakah jumlah record data set berpengaruh terhadap performa kedua algoritma tersebut.

3. Pendekatan Penelitan Kuantitatif

Pada penelitian ini menggunakan pendekatan kuantitatif yang nantinya hasil dari penelitian ini merupakan informasi-informasi berupa angka dan diagram hasil dari eksperimen perbandingan dua metode yang dilakukan. Pengumpulan data dilakukan melalui hasil eksperimen yang kemudian data tersebut dilakukan perbandingan dan analisis seperti dibuatkannya tabel dan diagram untuk melihat

perbandingan metode mana yang paling baik dalam membangun pengetahuan diagnose penyakit diabetes.

3.2. Metode Pengumpulan Data

Penelitian ini menggunakan tahap eksperimen sebagai metode pengumpulan data, namun selain itu peneliti juga mengumpulkan data awal sebagai bahan referensi melalui observasi, wawancara, dan studi pustaka kepada praktisi langsung yang terkait langsung dengan bidangnya. Tahap eksperimen dilakukan dengan melakukan perbandingan model algoritma Naïve Bayes dan K- Nearest Neighbor(KNN) sebagai metode yang akan dibandingkan, kemudian hasil perbandingan kedua model algoritma tersebut di evaluasi berdasarkan tingkat Accuracy, Precision, Recall, dan F1- Score untuk dilihat hasil perbandingannya.

Proses pengambilan data pada objek yang akan diteliti, dalam kasus ini peneliti mengambil data history penyakit diabetes dari UCI Machine Learning Repository. Data set yang diperoleh tersebut akan dijadikan model data untuk melakukan perbandingan pada kedua model algoritma yang akan diteliti.

3.3. Metode Analisis Data

Metode analisis data pada penelitian ini adalah membandingkan hasil eksperimen mulai dari awal sampai akhir. Eksperimen dimulai dari melakukan cleaning atau pembersihan pada data yang telah dikumpulkan. Tahap selanjutnya akan dilakukan transformasi data, dimana pada tahap ini akan dilakukan pengclusteran untuk dijadikan menjadi beberapa group atau kelompok data. Setelah data berhasil dikelompokan makan proses selanjutnya akan dilakukan proses pemodelan data dimana data yang sudah dikelompokan tersebut akan diubah

menjadi nilai-nilai yang dipisahkan dengan tanda koma atau Comma Sparated Value (CSV) sebagai format data yang akan dimasukan ke dalam database

Setelah proses pemodelan data atau memasukan data ke dalam database selesai dilakukan, maka proses selanjutnya melakukan proses preprocessing data melalui feature selection, missing value, inkonsistensi dan transportasi data serta penambahan optimasi menggunakan algoritma minimax dan melakukan perbandingan dua model algoritma yaitu Naïve Bayes dan K Nearest Neighbor(K-NN) menggunakan data yang telah diperoleh. Kedua model algoritma ini akan dibandingkan dengan menggunakan 5-FOLD Cross-Validation. Metode ini membagi secara acak 10 subset dan kemudian setiap subset n=1 akan menjadi data training sedangkan n=2,3,...,10 akan menjadi data testing. Kemudian dilakukan sebanyak 10 iterasi dengan data testing n+1 dan syarat data testing tidak sama dengan data training.

Penarikan kesimpulan dilakukan berdasarkan hasil evaluasi pada kedua model algoritma yang terpilih dengan menggunakan pengukuran nilai nilai performa dengan Confusion Matrix yaitu Accuracy, Precision, Recall, dan F1-Score. Hasil pengukuran tersebut akan diadikan acuan atau pedoman dalam menentukan hasil atau rekomendasi pada penelitian ini.

3.4. Alur Penelitian

Pada bagian ini berisi diagram alur langkah penelitian secara lengkap dan terinci termasuk di dalamnya tercermin algoritma, rute, pemodelan-pemodelan, desain, yang terkait dengan aspek perancangan sistem. Alur penelitian untuk membangun pengetahuan diagnosa penyakit diabetes dijelaskan pada gambar 1 .

Gambar 3.1. Alur Penelitian

Berikut penjelasan alur penelitian pada gambar 1 yaitu

a. Identifikasi masalah

Dalam proses identifikasi masalah dengan mencari tahu permasalahan yang ada pada objek penelitian dengan membaca beberapa artikel terkait

b. Studi literatur

Dalam studi literatur yang dilakukan adalah dengan mencari dan membaca jurnal serta buku yang relevan dengan permasalahan yang akan diangkat

sebagai bahan rujukan dalam memilih metode atau algoritma yang sesuai dan menentukan objek penelitian.

c. Pengumpulan data penyakit diabetes

Setelah membaca beberapa literatur yang relevan, tahap berikutnya adalah proses pengumpulan data berupa history dari diagnosis penyakit diabetes dari sumber internet dengan UCI Machine Learning Repository dengan mendownload pada website dengan link https://archive.ics.uci.edu/ml/machine-learning-databases/00529/.

d. *Preprocessing*

Dalam proses ini dilakukan pengolahan terhadap kumpulan history diagnosis penyakit diabetes dari UCI Machine Learning Repository sebelum dilakukan klasifikasi. Preprocessing data dilakukan melalui feature selection, missing value, inkonsistensi dan transportasi data sehingga nanti menjadi sebuah dataset.

e. Penentuan skenario

Pada tahap ini penulis menentukan 4 skenario percobaan dalam penelitian ini yaitu kombinasi antara arsitektur dengan menggunakan algoritma *Minimax*) dan tanpa menggunakan algoritma Minimax pada algoritma *Naïve Bayes*, dan *K Nearest Neighbor*(KNN

f. Evaluasi dan Validasi

Pada tahap ini dilakukan proses evaluasi dan validasi pada model klasifikasi yang telah dibuat untuk setiap skenario.Validasi pada penelitian ini menggunakan 5 k–fold lalu selanjutnya melakukan evaluasi yang

didapatkan dari nilai-nilai *confusion matrix* dari setiap skenario untuk menilai Accuracy, Precision, Recall, dan F1- Score dari model klasifikasi dengan hasil data dalam bentuk kurva ROC untuk mengukur nilai AUC.

g. Kesimpulan

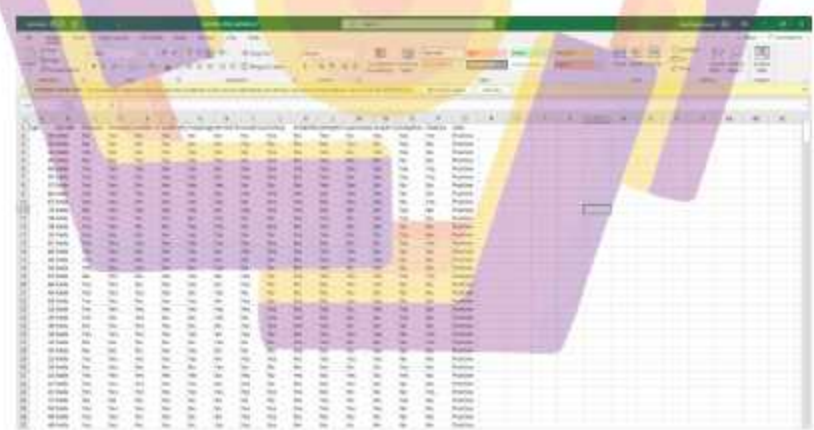
Setelah didapatkan beberapa hasil fakta dengan 4 skenario percobaan, tahap selanjutnya adalah membuat kesimpulan dengan menyajikan hasil dari percobaan yang telah dilakukan dengan beberapa fakta terkait arsitektur *Naïve Bayes dan K Nearest Neighbor* (KNN) dan pengaruh penggunaan algoritma *Minimax* terhadap tingkat akurasi dan arsitektur yang memiliki kinerja terbaik.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1.Pengumpulan data penyakit diabetes

Penelitian disini menggunakan dataset dari UCI Machine Learning Repository sebanyak 521 dataset dengan 17 featur atau atribut yaitu Age,Gender,Polyuria,Polydipsia,SuddenWeightLoss,Weakness,Polyphagia,GenitalThrush,Visual Blurring,Itching,Irritability,Delayed Healing,Partial Paresis,Muscle Stiffness,Alopecia,Obesity,Classmyang ada pada website dengan link https://archive.ics.uci.edu/ml/machine-learning-databases/00529/. Dataset diabetes dapat dilihat pada gambar 4.1 dan penjelasan tentang atribut yang digunakan dijelaskan pada tabel 1.



Gambar 4.1. Dataset Penyakit Diabetes

Tabel 4.1. Keterangan Atribut

No	Atribut	Keterangan Atribut
1	Age	Umur pada pasien
2	Gender	Jenis kelamin pada pasien
3	Polyuria	Sering buang air kecil lebih sering
4	Polydipsia	Haus berlebihan
5	Sudden Weight Loss	Penurunan berat badan secara tiba-tiba
6	Weakness	Badan merasa lemah
7	Polyphagia	Nafsu makan bertambah
8	Genital Thrush	Sariawana pada Thrush
9	Visual Blurring	Penglihata kabur
10	Itching	Kulit kering dan gatal
11	Irritability	Pasien gampang marah
12	Delayed Healing	Menghambat penyembuhan
13	Partial Paresis	Mengalami kelumpuhan sebagian
14	Muscle Stiffness	Nyeri otot atau kekakuan pada sendi
15	Alopecia	Kerontokan
16	Obesity	Mengalami kegemukan
17	Class	Hasil dari diagnosa pasien

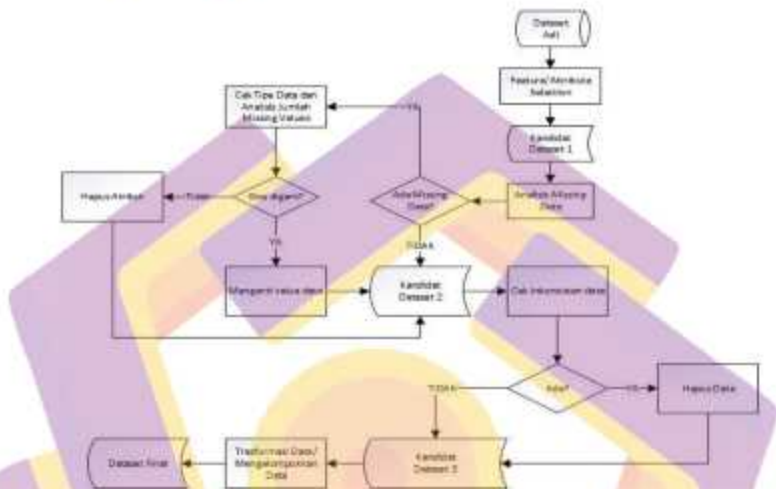
4.2.Preprocessing

Tahap *preprocessing* adalah data yang terkait pada data yang akan diuji atau dievaluasi oleh *machine learning*, tahap ini terdiri dari pemilihan atribut atau fitur yang mendukung penelitian, *data cleaning*, dan menghasilkan dataset final yang selanjutnya akan dievaluasi.

4.2.1. Preparasi Data

Pada bagian ini, analisis dataset akan dilakukan terlebih dahulu sebelum dilakukan ui performa menggunakan *machine learning*. Database original yang tersedia berisikan informasi yang *redundant* dan tidak lengkap sesuai dengan data yang didapat dari database real. Hal ini perlu dilakukan karena data yang biasanya

berasal dari dunia nyata memiliki nilai atau data yang bersifat *redundant*, *noisy*, *multiple* (ganda), dan tidak lengkap. Tahap-tahap yang akan dilakukan dalam preparasi data dapat dilihat pada Gambar 4.2.



Gambar 4.2. Alur Preparasi Data

Pada langkah pertama adalah langkah pemilihan atribut atau *feature selection*. Langkah ini adalah langkah dimana akan dilakukan tahap pemilihan atribut atau *feature* yang akan digunakan. Setelah dilakukan analisis data, terdapat beberapa atribut yang memiliki nilai *missing values* yang cenderung tinggi. Pada data yang tersedia terlihat bahwa terdapat 17 atribut *feature* pada dataset ini. Setelah dianalisa kembali terdapat semua atribut disimpulkan bahwa semua atribut berperngaruh terhadap hasil perhitungan karena atribut sangat erat hubungannya dengan kasus diagnosa penyakit diabetes akan dijadikan kandidat sebagai atribut atau *feature* yang akan digunakan. Proses *feature selection* dapat dilihat pada Tabel 4.2.

Selanjutnya langkah yang kedua adalah menganalisis jumlah *record* yang tidak ada pada setiap atribut atau langkah identifikasi *missing values*. Data yang terdapat *missing values* tidak akan dapat diolah karena tidak memiliki nilai yang dapat dijadikan acuan untuk perhitungan. Oleh karena itu untuk mengatasi data kosong tersebut dapat dilakukan dua hal yaitu memberikan nilai secara acak terhadap attribut tersebut atau dapat dihilangkan data pada atribut tersebut. data yang tidak memiliki *missing values* akan dijadikan kandidat atribut yang akan digunakan dalam perhitungan. Hasil dari proses *missing values* atau analisis jumlah *record* dapat dilihat pada Tabel 4.2

Tabel 4.2. Hasil Proses *missing values*

No	Atribut	Missing Data
1	Age	0%
2	Gender	0%
3	Polyuria	0%
4	Polydipsia	0%
5	Sudden Weight Loss	0%
6	Weakness	0%
7	Polyphagia	0%
8	Genital Thrush	0%
9	Visual Blurring	0%
10	Itching	0%
11	Irritability	0%
12	Delayed Healing	0%
13	Partial Paresis	0%
14	Muscle Stiffness	0%
15	Alopecia	0%
16	Obesity	0%
17	Class	0%

Langkah ke tiga adalah langkah identigfikasi duplikasi data. Langkah ini dilakukan untuk menghindari terjadi data yang sama untuk setiap atribut. Pada tahap ini dilakukan analisis data terhadap data yang sama atau *double* yang apabila terdapat data yang sama maka data tersebut akan dihapus atau dihilangkan. Dataset pada Tabel 4.2 Menunjukan bahwa tidak adanya duplikasi data sehingga dataset tersebut merupakan kandidat dataset yang akan di proses pada langkah selanjutnya.

Selanjutnya adalah langkah ke empat atau tranformasi data. Pada tahap ini akan dilakukan proses tranformasi data menjadi sebuah nilai atau bobot. Proses tranformasi data menggunakan teknik normalisasi dengan menggunakan metode min-max untuk mendapatkan nilai bobot yang akan dijadikan nilai input untuk setiap parameter atribut yang digunakan untuk menentukan dataset final. Contoh proses transformasi data untuk menentukan nilai bobot IPK dapat dilihat seperti pada perhitungan berikut.

$$xBaru = \frac{(xlama - minlama)}{(maxlama - minlama)}(maxbaru - minbaru) + minbaru$$

$$IPK\ \frac{(3{,}92-3)}{(4-3)}(1-0)+0 = \frac{(0{,}92)}{(1)}\ 1+0 = 0.92$$

4.2.2. Dataset Final

Pada bagian ini merupakan dataset final yang akan digunakan pada panelitian ini. Berikut ini adalah daftar kriteria-kriteria atribut yang akan digunakan dalam penelitian ini. Setelah melalui tahapan preparasi data tersebut yang telah dilakukan didapatkan data sebanyak 521 records dengan jumlah 17 fitur atau atribut.Pada penelitian ini terdapat 2 dataset final yaitu dataset tanpa menggunakan normalisasi minimax dan dengan normalisasi minimax.Contoh dataset final tanpa

minimax pada Tabel 4.3 dan contoh dataset final dengan normalisasi minimax dan contoh dataset final dengan normalisasi minimax pada Table 4.4.

Tabel 4.3. Contoh Dataset Final Tanpa Normalisasi *Minimax*

No	Atribut	M1	M2	M3	.....	M520
1	Age	40	58	41	.....	42
2	Gender	Male	Male	Male	.....	Male
3	Polyuria	No	No	Yes	.....	No
4	Polydipsia	Yes	No	No	.....	No
5	Sudden Weight Loss	No	No	No	.....	No
6	Weakness	Yes	Yes	Yes	.....	No
7	Polyphagia	No	No	Yes	.....	No
8	Genital Thrush	No	No	No	.....	No
9	Visual Blurring	No	Yes	No	.....	No
10	Itching	Yes	No	Yes	.....	No
11	Irritability	No	No	No	.....	No
12	Delayed Healing	Yes	No	Yes	.....	No
13	Partial Paresis	No	Yes	No	.....	No
14	Muscle Stiffness	Yes	No	Yes	.....	No
15	Alopecia	Yes	Yes	Yes	.....	No
16	Obesity	Yes	No	No	.....	No
17	Class	Positive	Positive	Positive	.....	Negative

Tabel 4.4. Contoh Dataset Final Dengan Normalisasi *Minimax*

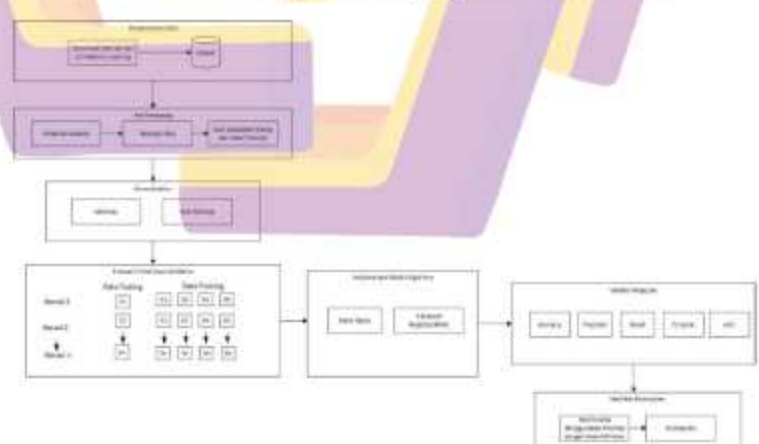
No	Atribut	M1	M2	M3	.....	M520
1	Age	0.43	0.56	0.33	.....	0.35
2	Gender	1	1	1	.....	1
3	Polyuria	0	0	1	.....	0
4	Polydipsia	1	0	0	.....	0
5	Sudden Weight Loss	0	0	0	.....	0
6	Weakness	1	1	1	.....	0
7	Polyphagia	0	0	1	.....	0
8	Genital Thrush	0	0	0	.....	0

Tabel 4.4. Contoh Dataset Final Dengan Normalisasi *Minimax* (Lanjutan)

No	Atribut	M1	M2	M3	.....	M520
9	Visual Blurring	0	1	0	.....	0
10	Itching	1	0	1	.....	0
11	Irritability	0	0	0	.....	0
12	Delayed Healing	1	0	1	.....	0
13	Partial Paresis	0	1	0	.....	0
14	Muscle Stiffness	1	0	1	.....	0
15	Alopecia	1	1	1	.....	0
16	Obesity	1	0	0	.....	0
17	Class	Positive	Positive	Positive	.....	Negative

4.3.Penentuan skenario

Pada penelitian ini dilakukan beberapa penentuan scenario yang digunakan dan penjelasan penentuan skenario dijelaskan pada gambar 4.2 yang dimulai dari pengumpulan data, preprocessing, normalization, evaluasi k fold cross validation, implementasi model algoritma, validasi pengujian, dan hasil dan kesimpulan. Penentuan implementasi skenario digambarkan pada gambar 4.3



Gambar 4.3. Alur Implementasi

Berikut penjelasan dari alur skenario pada penelitian ini

1. Pengumpulan Data

   Pada tahap ini dilakukan pengumpulan data pada UCI Machine Learning dengan mendownload pada website dengan link https://archive.ics.uci.edu/ml/machine-learning-databases/00529/ dengan Repository sebanyak 521 dataset dengan 15 featur atau atribut yaitu*Age,Gender,Polyuria,Polydipsia,SuddenWeightLoss,Weakness,Polyphagia,GenitalThrush,VisualBlurring,Itching,Irritability,Delayed Healing,Partial Paresis,Muscle Stiffness,Alopecia,Obesity,Class* dan akhirnya menjadi dataset pada penelitian ini

2. Preprocessing

   Pada tahap ini dilakukan *preprocessing* dari data yang sudah didownload sebelumnya denga beberapa tahap yang pertama yaitu pelabelan dataset pada penelitian ini pada atribut class lalu dilanjutkan dengan missing value yaitu menghilangkan data set yang duplikasi atau data yang tidak lengkap setelah itu yang terakhir dilakukan split data yaitu membuat data testing dan data training dengan perbandingan 80% untuk data testing dan 20% training

3. Normalization

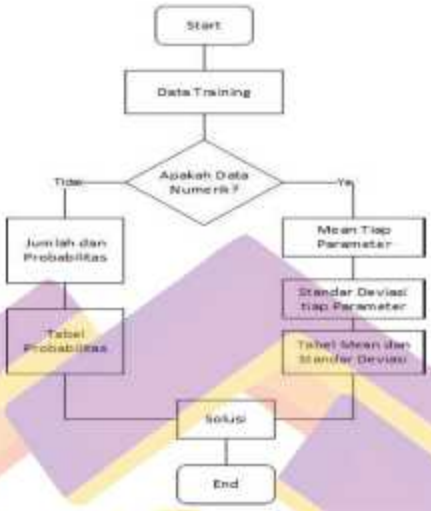
   Pada tahap ini dilakukan skenario penggunaan normalisasi dengan minimax atau tanpa normalisasi *minimax* yang selanjutnya akan digunakan dengan percobaan dengan algoritma *Naïve Bayes* dan *K-Nearest Neighbor*(KNN)

4. Evaluasi K Fold Cross Validation

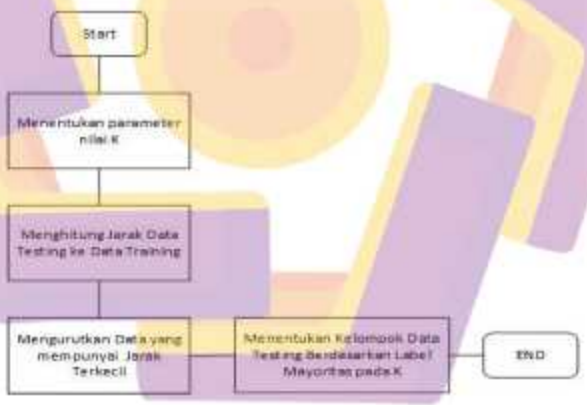
   Pada tahap dataset yang digunakan dengan jumlah 521 data, data tersebut akan dibagi kedalam 5-*fold* atau 5 bagian subset untuk menentukan nilai performa dan kemudian akan dihitung rata-rata setiap hasil performa model algoritma tersebut.

5. Implementasi Model Algoritma

   Pada tahap ini dilakukan percobaan skenario dengan penggunaan normalisasi dan tanpa normalisasi dengan menggunakan algoritma *Naïve Bayes* dan *K-Nearest Neighbor*.Pada gambar 4.4 penjabaran tentang klasifikasi algoritma Naïve Bayes yaitu dimulai dengan penginputan data training dilanjutkan dengan pengecekan apakah data bertipe numerik jika ya maka dilakukan perhitungan mean tiap parameter, perhitungan standar deviasi tiap parameter lalu terbuatnya tabel mean dan standar deviasi lalu menghasilkan solusi sedangkan jika data tidak numerik maka selanjutnya menghitung jumlah dan probabilitas dan menghasilkan table probabilitas dan menghasilkan solusi. Pada gambar 4.5 mengambarkan proses algoritma klasifikasi *K Nearest Neighbor* (KNN) yaitu proses yang dimulai dari menentukan parameter nilai K selanjutnya menghitung jarak data testing ke data training lalu mengurutkan data yang mempunyai jarak terkecil selanjutnya yang terakhir menentukan kelompok data testing berdasarkan mayoritas pada nilai k



Gambar 4.4. Proses Klasifikasi Naive Bayes



Gambar 4.5. Proses Klasifikasi KNN

6. Validasi Pengujian

Pada tahap ini dilakukan pengujian dengan menggunakan ilai-nilai *confusion matrix* dari setiap skenario untuk menilai *Accuracy,*

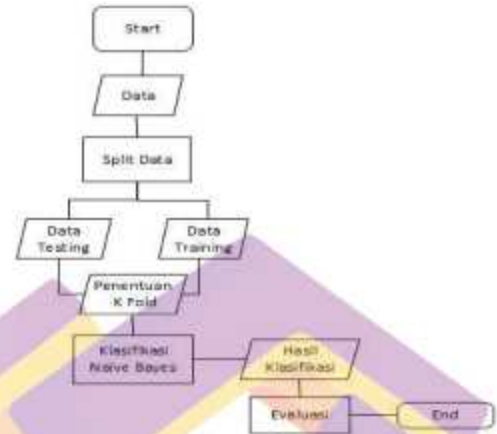
Precision, Recall, dan *F1- Score* dari model klasifikasi dengan hasil data dalam bentuk kurva ROC untuk mengukur nilai AUC

7. Hasil dan Kesimpulan

   Pada tahap ini dituliskan 4 hasil dari penelitian yaitu algoritma Naïve Bayes dengan normalisasi *minimax*, algoritma *Naïve Bayes* tanpa normalisasi *minimax*, algoritma *K-Nearet Neighbor(KNN)* dengan normalisasi *minimax*,dan algoritma *K-Nearet Neighbor(KNN)* tanpa normalisasi *minimax*

4.3.1. Algoritma *Naïve Bayes* Tanpa Normalisasi *Minimax*

Pada skenario pertama dilakukan dengan tanpa menggunakan normalisasi minimax pada algoritma *Naïve Bayes*.Pada tahap ini dimulai dengan menginputkan data yang sebelumnya sudah dilakukan *preprocessing* dengan melabeli class menjadi label untuk perhitungan perfoma dan replace value untuk menghilangkan data redundan, ke dua dengan split data menjadi data testing dan data training lalu yang ketiga menentukan k fold yang digunakan untuk membagi subset sebanyak 5, ke empat memproses data dengan menggunakan algoritma *Naïve Bayes*.yang keenam yaitu hasil dari penggunaan algoritma lalu pada tahap terakhir dilakuan dengan validasi data menggunakan confusion matrix yaitu *Accuracy, Precision, Recall*, dan *F1- Score* dari model klasifikasi dengan hasil data dalam bentuk kurva ROC untuk mengukur nilai AUC.Alur dari proses penelitian tersebut dijelaskan pada gambar 4.6



Gambar 4.6. Alur Skenario Algoritma *Naïve Bayes* Tanpa Normalisasi *Minimax*

Pada bagian ini dilakukan dengan implementasi dengan rapid miner. Berikut adalah beberapa tahapan yang dilakukan dengan menggunakan rapid miner yaitu pertama dengan memasukan data yang sudah didownload dalam UCI *Machine Learning*,yang kedua yaitu menghilangkan data yang duplikasi atau tidak lengkap,ketiga yaitu split data menjadi data testing yaitu 80% dan data training sebesar 20%,ke empat yaitu untuk evaluasi data menggunakan *Cross Validation* yaitu untuk memasukan k fold sebanyak 5 subset. Tahapan implementasi digambarkan pada gambar 4.7

Gambar 4.7. Implementasi Rapid Miner Algoritma *Naïve Bayes* Tanpa Normalisasi *Minimax*

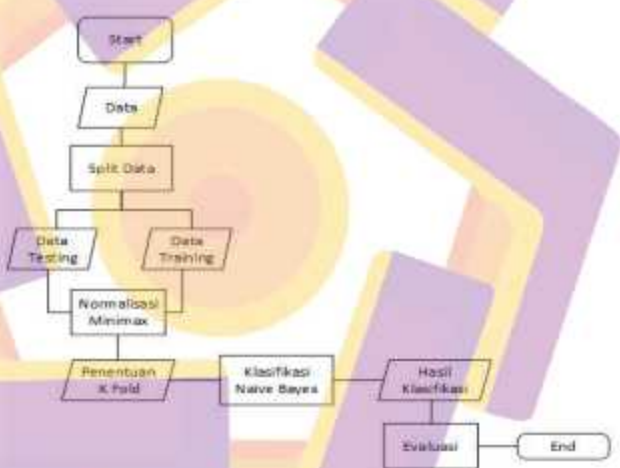
Pada bagian ini saat diklik cross validation maka terdapat implementasi dengan algoritma Naïve Bayes yang selanjutnya disambungkan dengan apply model untuk validasi data menggunakan confusion matrix yaitu Accuracy, Precision, Recall, dan F1- Score dari model klasifikasi dengan hasil data dalam bentuk kurva ROC untuk mengukur nilai AUC. Tahapan tersebut dijelaskan pada gambar 4.8

Gambar 4.8. *Cross Validation* Algoritma *Naïve Bayes* Tanpa Normalisasi *Minimax*

4.3.2. Algoritma *Naïve Bayes* Dengan Normalisasi *Minimax*

Pada skenario pertama dilakukan dengan tanpa menggunakan normalisasi minimax pada algoritma *Naïve Bayes*.Pada tahap ini dimulai dengan menginputkan data yang sebelumnya sudah dilakukan *preprocessing* dengan melabeli class menjadi label untuk perhitungan perfoma dan *replace value* untuk menghilangkan data redundan , kedua dengan split data menjadi data testing dan data training lalu yang ketiga melakukan normalisasi dengan *minimax* lalu yang keempat yaitu

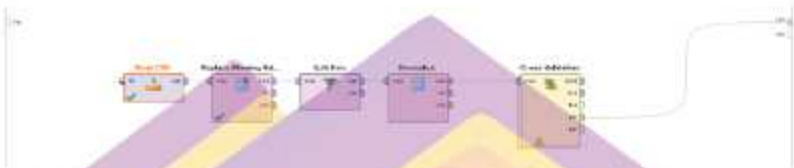
menentukan k fold yang digunakan untuk membagi subset sebanyak 5, ke lima memproses data dengan menggunakan algoritma *Naïve Bayes*,yang ke tujuh yaitu hasil dari penggunaan algoritma dan pada tahap terakhir dilakuan dengan validasi data menggunakan confusion matrix yaitu *Accuracy, Precision, Recall*, dan *F1-Score* dari model klasifikasi dengan hasil data dalam bentuk kurva ROC untuk mengukur nilai AUC.Alur dari proses penelitian tersebut dijelaskan pada gambar 4.9



Gambar 4.9. Alur Skenario Algoritma *Naïve Bayes* Dengan Normalisasi *Minimax*

Pada bagian ini dilakukan dengan implementasi dengan rapid miner. Berikut adalah beberapa tahapan yang dilakukan dengan menggunakan rapid miner yaitu pertama dengan memasukan data yang sudah didownload dalam UCI *Machine Learning*,yang kedua yaitu menghilangkan data yang duplikasi atau tidak lengkap,ketiga yaitu split data menjadi data testing yaitu 80% dan data training

sebesar 20%,ke empat yaitu normalisasi data dengan *minmax* lalu yang kelima untuk evaluasi data menggunakan *Cross Validation* yaitu untuk memasukan k fold sebanyak 5 subset. Tahapan implementasi digambarkan pada gambar 4.10



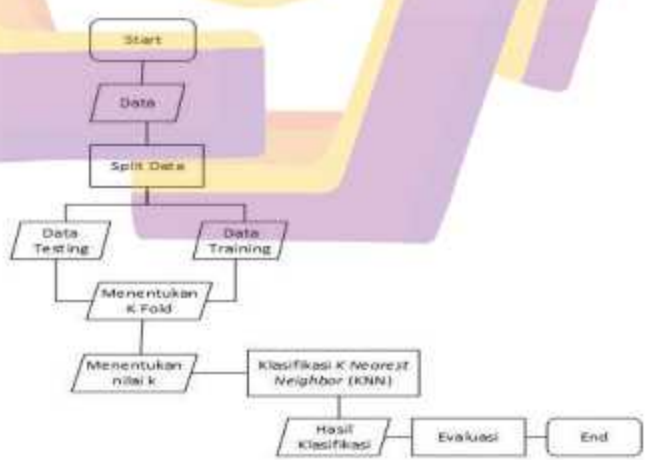
Gambar 4.10. Implementasi Rapid Miner Algoritma *Naïve Bayes* Dengan Normalisasi *Minimax*

Pada bagian ini saat diklik *cross validation* maka terdapat implementasi dengan algoritma *Naïve Bayes* yang selanjutnya disambungkan dengan apply model untuk validasi data menggunakan *confusion matrix* yaitu *Accuracy, Precision, Recall*, dan *F1- Score* dari model klasifikasi dengan hasil data dalam bentuk kurva ROC untuk mengukur nilai AUC. Tahapan tersebut dijelaskan pada gambar 4.11

Gambar 4.11 *Cross Validation* Algoritma *Naïve Bayes* Dengan Normalisasi *Minimax*

4.3.3. Algoritma KNN Tanpa Normalisasi *Minimax*

Pada skenario pertama dilakukan dengan tanpa menggunakan normalisasi minimax pada algoritma *Naïve Bayes*.Pada tahap ini dimulai dengan menginputkan data yang sebelumnya sudah dilakukan *preprocessing* dengan melabeli class menjadi label untuk perhitungan perfoma dan replace value untuk menghilangkan data redundan ,ke dua dengan split data menjadi data testing dan data training lalu yang ketiga menentukan k fold yang digunakan untuk membagi subset sebanyak ,ke empat yaitu menentukan nilai k pada *K Nearest Neighbor* yaitu 3, kelima memproses data dengan menggunakan algoritma *K Nearest Neighbor(KNN)* umtuk dilakukan klasifikasi,keenam yaitu hasil dari penggunaan algoritma lalu pada tahap terakhir dilakuan dengan validasi data menggunakan confusion matrix yaitu *Accuracy, Precision, Recall*, dan *F1- Score* dari model klasifikasi dengan hasil data dalam bentuk kurva ROC untuk mengukur nilai AUC.Alur dari proses penelitian tersebut dijelaskan pada gambar 4.12



Gambar 4.12 Alur Skenario Algoritma KNN Tanpa Normalisasi *Minimax*

Pada bagian ini dilakukan dengan implementasi dengan rapid miner. Berikut adalah beberapa tahapan yang dilakukan dengan menggunakan rapid miner yaitu pertama dengan memasukan data yang sudah didownload dalam UCI *Machine Learning* yang kedua yaitu menghilangkan data yang duplikasi atau tidak lengkap,ketiga yaitu split data menjadi data testing yaitu 80% dan data training sebesar 20%,ke empat yaitu untuk evaluasi data menggunakan *Cross Validation* yaitu untuk memasukan k fold sebanyak 5 subset. Implementasi Rapid Miner digambarkan pada gambar 4.13

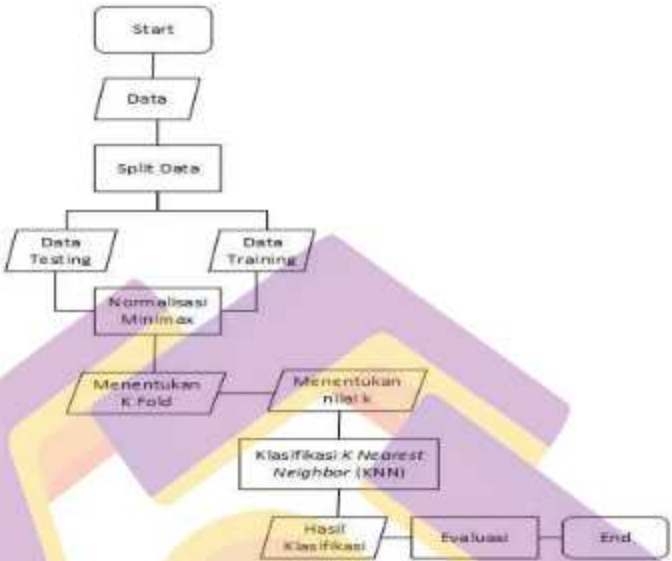
Gambar 4.13. Implementasi Rapid Miner Algoritma KNN Tanpa Normalisasi Minimax

Pada bagian ini saat diklik *cross validation* maka terdapat implementasi dengan algoritma *Naïve Bayes* yang selanjutnya disambungkan dengan apply model untuk validasi data menggunakan confusion matrix yaitu *Accuracy, Precision, Recall*, dan *F1- Score* dari model klasifikasi dengan hasil data dalam bentuk kurva ROC untuk mengukur nilai AUC. Tahapan tersebut dijelaskan pada gambar 4.14

Gambar 4.14. *Cross Validation* Algoritma KNN Tanpa Normalisasi *Minimax*

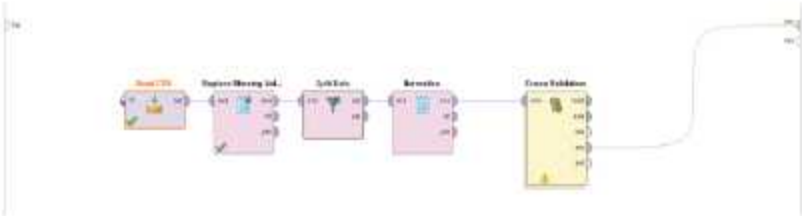
4.3.4. Algoritma KNN Dengan Normalisasi *Minimax*

Pada skenario pertama dilakukan dengan tanpa menggunakan normalisasi minimax pada algoritma Naïve Bayes.Pada tahap ini dimulai dengan menginputkan data yang sebelumnya sudah dilakukan preprocessing dengan melabeli class menjadi label untuk perhitungan perfoma dan *replace value* untuk menghilangkan data redundan,ke dua dengan split data menjadi data testing dan data training lalu yang ketiga melakukan normalisasi dengan *minimax* lalu yang keempat yaitu menentukan k fold yang digunakan untuk membagi subset sebanyak 5 subset, ke lima menentukan nilai k yaitu 3 dan yang ke enam yaitu memproses data dengan menggunakan algoritma *K Nearest Neighbor(KNN)* untuk dilakukan klasifikasi , ke tujuh yaitu hasil dari penggunaan algoritma dan pada tahap terakhir dilakuan dengan validasi data menggunakan confusion matrix. Implementasi Rapid Miner dijelaskan pada gambar 4.15



Gambar 4.15. Alur Skenario Algoritma KNN Dengan Normalisasi *Minimax*

Pada bagian ini dilakukan dengan implementasi dengan rapid miner. Berikut adalah beberapa tahapan yang dilakukan dengan menggunakan rapid miner yaitu pertama dengan memasukan data yang sudah didownload dalam UCI Machine Learning,yang kedua yaitu menghilangkan data yang duplikasi atau tidak lengkap,ketiga yaitu split data menjadi data testing yaitu 80% dan data training sebesar 20%,ke empat yaitu untuk evaluasi data menggunakan Cross Validation yaitu untuk memasukan k fold sebanyak 5 subset. Implementasi Rapid Miner dijelaskan pada gambar 4.16



Gambar 4.16. Implementasi Rapid Miner algoritma KNN Dengan Normalisasi Minimax

Pada bagian ini saat diklik *cross validation* maka terdapat implementasi dengan algoritma *Naïve Bayes* yang selanjutnya disambungkan dengan apply model untuk validasi data menggunakan *confusion matrix* yaitu *Accuracy, Precision, Recall, dan F1- Score* dari model klasifikasi dengan hasil data dalam bentuk kurva ROC untuk mengukur nilai AUC. Tahapan tersebut dijelaskan pada gambar 4.17

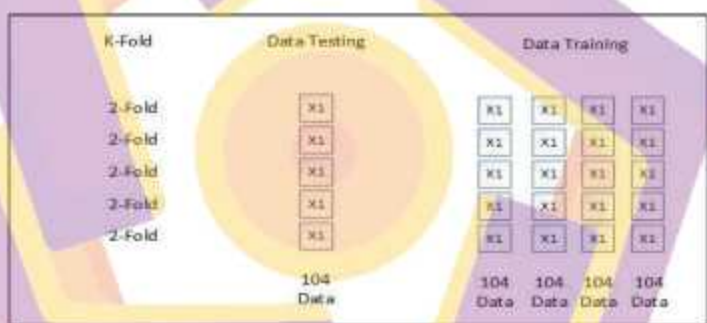
Gambar 4.17. Cross Validation KNN Dengan Normalisasi Minimax

4.4.Evaluasi dan Validasi

Pada tahap ini dilakukan evaluasi dan validasi pada pengujian yaitu evaluasi menggunakan K Fold 5 dan pengujian tingkat perfoma yaitu *Accuracy, Precision, Recall,* dan *F1- Score.*

4.4.1. Evaluasi

Evaluasi merupakan tahap dimana dataset final akan dilakukan pengujian atau dievaluasi performanya. Pada tahap ini implementasi model algoritma *Naïve Bayes* dan *K Nearest Neighbor* akan dilakukan untuk menguji atau mengevaluasi performa dari algoritma tersebut. Dataset yang akan digunakan dalam penelitian ini memiliki jumlah 521 data, data tersebut akan dibagi kedalam 5-*fold* atau 5 bagian subset untuk menentukan nilai performa dan kemudian akan dihitung rata-rata setiap hasil performa model algoritma tersebut. Mekanisme pembagian 5-*fold* atau 5 bagian subset dapat dilihat pada Gambar 4.18



Gambar 4.18. Pembagian K Fold

4.4.2. Validasi Algoritma *Naïve Bayes* Tanpa Normalisai *Minimax*

Pada pengujian ini implementasi dari model algoritma *Naïve Bayes* tanpa normalisasi Minimax akan digunakan untuk mendapatkan nilai tingkat performa. Pada pengujian ini terdapat beberapa teknik pengujian sesuai kapabilitas yang dimiliki oleh model algoritma tersebut. Adapun konfigurasi parameter dalam mekanisme pengujian ini antara lain:

1. Algoritma yang digunakan *Naïve Bayes*.
2. Split data menjadi data training dan data testing yang digunakan adalah 0,8 untuk data training dan 0,2 untuk data testing.
3. Pengujian validasi yang digunakan 5-*fold* atau 5 subset.
4. Pengujian tingkat perfoma yaitu *Accuracy, Precision, Recall,* dan *F1- Score*

Pengujian pertama dilakukan untuk nilai tingkat perfoma akurasi dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.5

Tabel 4.5. *Accuracy* Algorima *Naïve Bayes* Tanpa Normalisasi *Minimax*

Penentuan K fold	Accuracy
Fold 1	87.26%
Fold 2	85.82%
Fold 3	86.54%
Fold 4	86.31%
Fold 5	85.81%

Hasil dari pengujian nilai tingkat perfoma akurasi tertinggi diperoleh dengan menggunakan penentuan Fold 1 dengan nilai akurasi yang dihasilkan yaitu 87,26%

Berikut ini adalah grafik dari perbandingan akurasi dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.19

Gambar 4.19. Grafik Akurasi *Naïve Bayes* Tanpa Normalisasi *Minimax*

Pengujian kedua dilakukan untuk nilai tingkat perfoma *precision* dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.6

Tabel 4.6. *Precision* Algorima *Naïve Bayes* Tanpa Normalisasi *Minimax*

Penentuan K fold	Precision
Fold 1	80.01%
Fold 2	78.19%
Fold 3	79.17%
Fold 4	79.24%
Fold 5	78.51%

Hasil dari pengujian nilai tingkat perfoma *precision* tertinggi diperoleh dengan menggunakan penentuan Fold 1 dengan nilai *precision* yang dihasilkan yaitu 80.01%

Berikut ini adalah grafik dari perbandingan *precision* dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.20

Gambar 4.20. Grafik *Precision Naïve Bayes* Tanpa Normalisasi *Minimax*

Pengujian ketiga dilakukan untuk nilai tingkat perfoma *recall* dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.7

Tabel 4.7. *Recall* Algorima *Naïve Bayes* Tanpa Normalisasi *Minimax*

Penentuan K fold	*Recall*
Fold 1	89.38%
Fold 2	87.50%
Fold 3	88.75%
Fold 4	88.12%
Fold 5	86.97%

Hasil dari pengujian nilai tingkat perfoma *recall* tertinggi diperoleh dengan menggunakan penentuan Fold 1 dengan nilai *recall* yang dihasilkan yaitu 89.38%

Berikut ini adalah grafik dari perbandingan *recall* dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.21

Gambar 4.21. Grafik *Recall Naïve Bayes* Tanpa Normalisasi *Minimax*

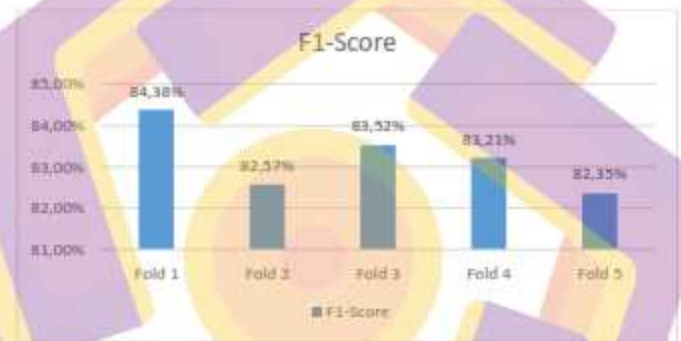
Pengujian keempat dilakukan untuk nilai tingkat perfoma F1- *Score* dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.8

Tabel 4.8. *F1 Score* Algorima *Naïve Bayes* Tanpa Normalisasi *Minimax*

Penentuan K fold	F1- Score
Fold 1	84.38%
Fold 2	82.57%
Fold 3	83.52%
Fold 4	83.21%
Fold 5	82.35%

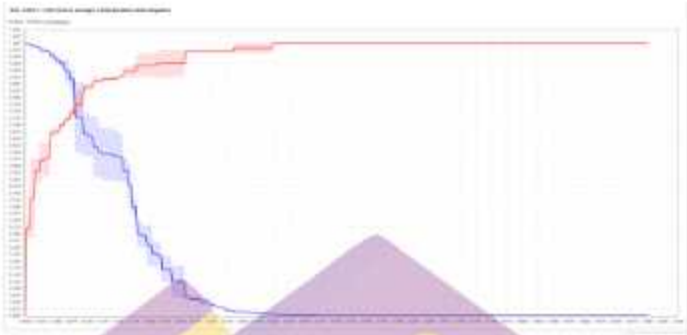
Hasil dari pengujian nilai tingkat perfoma *F1- Score* tertinggi diperoleh dengan menggunakan penentuan Fold 1 dengan nilai *F1- Score* yang dihasilkan yaitu 84.38%

Berikut ini adalah grafik dari perbandingan *F1- Score* dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.22



Gambar 4.22. Grafik *F1-Score Naïve Bayes* Tanpa Normalisasi *Minimax*

Pengujian kelima dilakukan untuk nilai AUC untuk mengetahui apakah klasifikasi yang sudah dilakukan sudah baik pada gambar 4.23 adalah hasil AUC dari nilai pengujian akurasi tertinggi yaitu 87,26% dengan nilai AUC 0.944.



Gambar 4.23. AUC *Naïve Bayes* Tanpa Normalisasi *Minimax*

4.4.3. Validasi Algoritma *Naïve Bayes* Dengan Algoritma *Minimax*

Pada pengujian ini implementasi dari model algoritma *Naïve Bayes* akan digunakan untuk mendapatkan nilai tingkat performa. Pada pengujian ini terdapat beberapa teknik pengujian sesuai kapabilitas yang dimiliki oleh model algoritma tersebut. Adapun konfigurasi parameter dalam mekanisme pengujian ini antara lain:

1. Algoritma yang digunakan *Naïve Bayes*.
2. Split data menjadi data training dan data testing yang digunakan adalah 0,8 untuk data training dan 0,2 untuk data testing.
3. Normalisasi dengan *Minimax* dengan min 0 dan max 1
4. Pengujian validasi yang digunakan 5-*fold* atau 5 subset.
5. Pengujian tingkat perfoma yaitu *Accuracy*, *Precision*, *Recall*, dan *F1-Score*

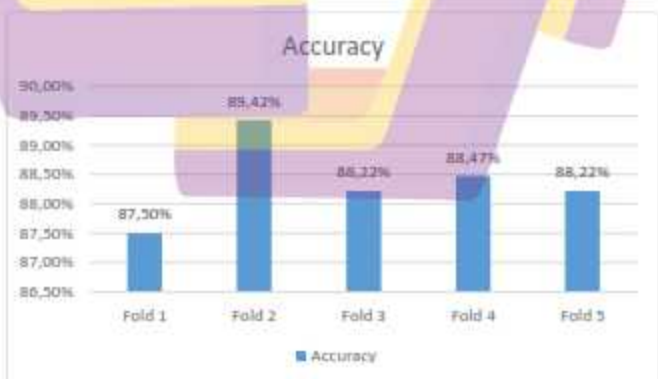
Pengujian pertama dilakukan untuk nilai tingkat perfoma akurasi dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.9

Tabel 4.9. *Accuracy* Algorima *Naïve Bayes* Dengan Normalisasi *Minimax*

Penentuan K fold	Accuracy
Fold 1	87.50%
Fold 2	89.42%
Fold 3	88.22%
Fold 4	88.47%
Fold 5	88.22%

Hasil dari pengujian nilai tingkat perfoma akurasi tertinggi diperoleh dengan menggunakan penentuan Fold 2 dengan nilai akurasi yang dihasilkan yaitu 89.42%

Berikut ini adalah grafik dari perbandingan akurasi dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.24



Gambar 4.24. Grafik Akurasi *Naïve Bayes* Dengan Normalisasi *Minimax*

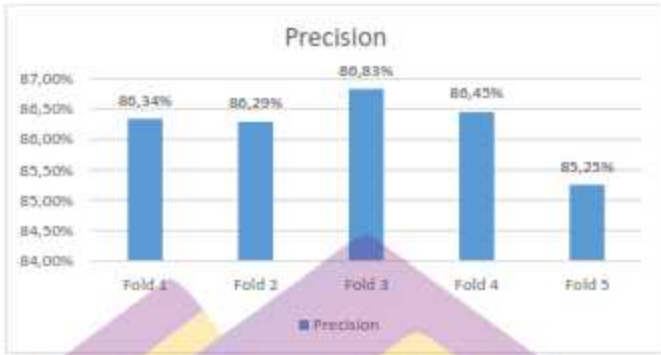
Pengujian kedua dilakukan untuk nilai tingkat perfoma *precision* dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.10

Tabel 4.10. *Precision* Algorima *Naïve Bayes* Dengan Normalisasi *Minimax*

Penentuan K fold	Precision
Fold 1	86.34%
Fold 2	86.29%
Fold 3	86.83%
Fold 4	86.45%
Fold 5	85.25%

Hasil dari pengujian nilai tingkat perfoma *precision* tertinggi diperoleh dengan menggunakan penentuan Fold 3 dengan nilai *precision* yang dihasilkan yaitu 86.83%

Berikut ini adalah grafik dari perbandingan *precision* dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.25



Gambar 4.25. Grafik *Precision Naïve Bayes* Dengan Normalisasi *Minimax*

Pengujian ketiga dilakukan untuk nilai tingkat perfoma *recall* dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.11

Tabel 4.11. *Recall* Algorima *Naïve Bayes* Dengan Normalisasi *Minimax*

Penentuan K fold	Recall
Fold 1	81.25%
Fold 2	86.25%
Fold 3	83.12%
Fold 4	83.75%
Fold 5	85.04%

Hasil dari pengujian nilai tingkat perfoma *recall* tertinggi diperoleh dengan menggunakan penentuan Fold 2 dengan nilai *recall* yang dihasilkan yaitu 86.25%

Berikut ini adalah grafik dari perbandingan *recall* dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.26

Gambar 4.26. Grafik *Recall Naïve Bayes* Dengan Normalisasi *Minimax*

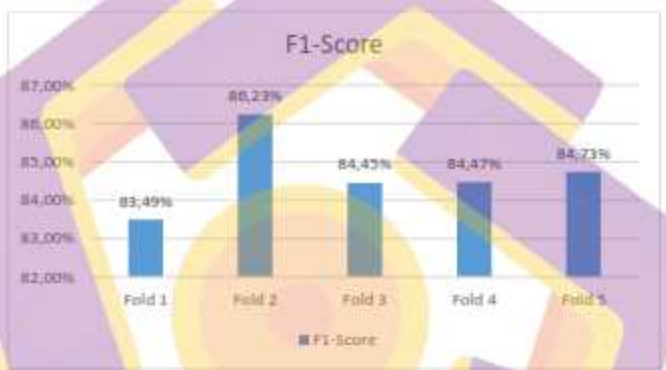
Pengujian keempat dilakukan untuk nilai tingkat perfoma *F1- Score* dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.12

Tabel 4.12. *F1 Score* Algorima *Naïve Bayes* Dengan Normalisasi *Minimax*

Penentuan K fold	F1- Score
Fold 1	83.49%
Fold 2	86.23%
Fold 3	84.45%
Fold 4	84.47%
Fold 5	84.73%

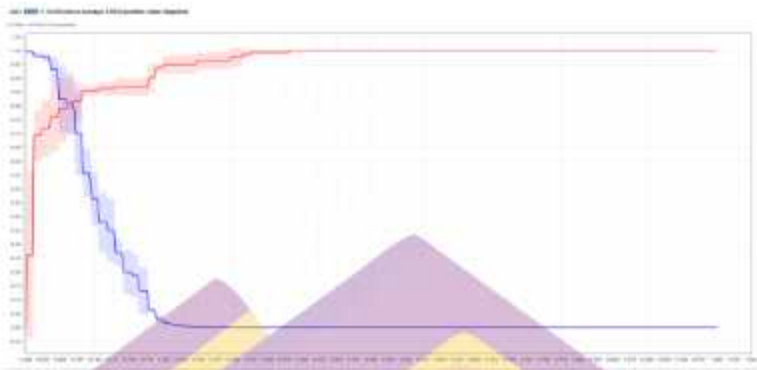
Hasil dari pengujian nilai tingkat perfoma F1- *Score* tertinggi diperoleh dengan menggunakan penentuan K Fold 2 dengan nilai F1- *Score* yang dihasilkan yaitu 86.23%

Berikut ini adalah grafik dari perbandingan F1- *Score* dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.27



Gambar 4.27. Grafik *F1-Score Naïve Bayes* Dengan Normalisasi *Minimax*

Pengujian kelima dilakukan untuk nilai AUC untuk mengetahui apakah klasifikasi yang sudah dilakukan sudah baik pada gambar 4.28 adalah hasil AUC dari nilai pengujian akurasi tertinggi yaitu 89.42%dengan nilai AUC 0.955



Gambar 4.28. AUC Algoritma *Naïve Bayes* Dengan Normalisasi *Minimax*

4.4.4. Validasi Algoritma KNN Tanpa Algoritma *Minimax*

Pada pengujian ini implementasi dari model algoritma *K-Nearest Neighbor(KNN)* akan digunakan untuk mendapatkan nilai tingkat performa. Pada pengujian ini terdapat beberapa teknik pengujian sesuai kapabilitas yang dimiliki oleh model algoritma tersebut. Adapun konfigurasi parameter dalam mekanisme pengujian ini antara lain:

1. Algoritma yang digunakan *K-Nearest Neighbor(KNN)*
2. Split data menjadi data training dan data testing yang digunakan adalah 0,8 untuk data training dan 0,2 untuk data testing.
3. Pengujian validasi yang digunakan 5-*fold* atau 5 subset.
4. Menggunakan nilai k yaitu 3
5. Pengujian tingkat perfoma yaitu *Accuracy, Precision, Recall*, dan *F1-Score*

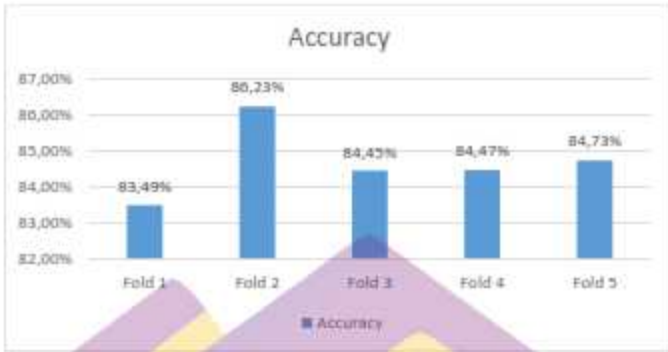
Pengujian pertama dilakukan untuk nilai tingkat perfoma akurasi dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.13

Tabel 4.13. *Accuracy* Algorima KNN Tanpa Normalisasi *Minimax*

Penentuan K fold	Accuracy
Fold 1	84.13%
Fold 2	88.95%
Fold 3	88.94%
Fold 4	91.12%
Fold 5	89.41%

Hasil dari pengujian nilai tingkat perfoma akurasi tertinggi diperoleh dengan menggunakan penentuan K Fold 4 dengan nilai akurasi yang dihasilkan yaitu 91.12%

Berikut ini adalah grafik dari perbandingan akurasi dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.29



Gambar 4.29. Grafik Akurasi Algoritma KNN Tanpa Normalisasi *Minimax*

Pengujian kedua dilakukan untuk nilai tingkat perfoma *precision* dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.14

Tabel 4.14 *Precision* Algorima KNN Tanpa Normalisasi *Minimax*

Penentuan K fold	Precision
Fold 1	78.47%
Fold 2	83.51%
Fold 3	81.15%
Fold 4	85.78%
Fold 5	83.62%

Hasil dari pengujian nilai tingkat perfoma *precision* tertinggi diperoleh dengan menggunakan penentuan Fold 4 dengan nilai *precision* yang dihasilkan yaitu 86.23%

Berikut ini adalah grafik dari perbandingan *precision* dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.30

Gambar 4.30. Grafik *Precision* Algoritma KNN Tanpa Normalisasi *Minimax*

Pengujian ketiga dilakukan untuk nilai tingkat perfoma recall dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.15

Tabel 4.15. *Recall* Algorima KNN Tanpa Normalisasi *Minimax*

Penentuan K fold	Recall
Fold 1	80.62%
Fold 2	88.78%
Fold 3	93.13%
Fold 4	92.50%
Fold 5	90.08%

Hasil dari pengujian nilai tingkat perfoma *recall* tertinggi diperoleh dengan menggunakan penentuan Fold 3 dengan nilai *recall* yang dihasilkan yaitu 93.13%

Berikut ini adalah grafik dari perbandingan recall dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.31

Gambar 4.31. Grafik *Recall* Algoritma KNN Tanpa Normalisasi *Minimax*

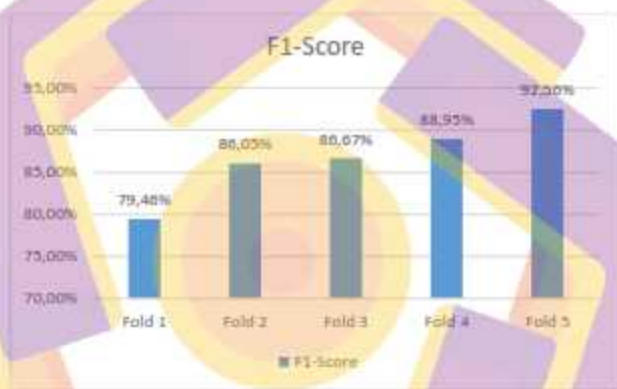
Pengujian keempat dilakukan untuk nilai tingkat perfoma F1- Score dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.16

Tabel 4.16. *F1 Score* Algorima KNN Tanpa Normalisasi *Minimax*

Penentuan K fold	F1- Score
Fold 1	79.46%
Fold 2	86.05%
Fold 3	86.67%
Fold 4	88.95%
Fold 5	86.63%

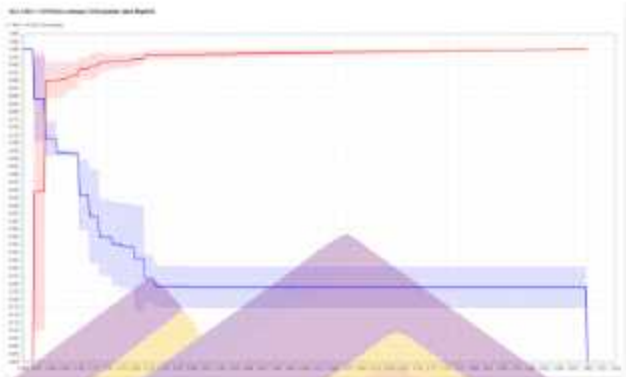
Hasil dari pengujian nilai tingkat perfoma F1- Score tertinggi diperoleh dengan menggunakan penentuan Fold 4 dengan nilai F1- Score yang dihasilkan yaitu 88.95%

Berikut ini adalah grafik dari perbandingan F1- Score dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.32



Gambar 4.32 Grafik *F1-Score* Algoritma KNN Tanpa Normalisasi *Minimax*

Pengujian kelima dilakukan untuk nilai AUC untuk mengetahui apakah klasifikasi yang sudah dilakukan sudah baik pada gambar 4.33 adalah hasil AUC dari nilai pengujian akurasi tertinggi yaitu 91.12% dengan nilai AUC 0.954.



Gambar 4.33. AUC Algoritma KNN Tanpa Normalisasi *Minimax*

4.4.5. Validasi Algoritma KNN Dengan Normalisaal *Minimax*

Pada pengujian ini implementasi dari model algoritma *K-Nearest Neighbor(KNN)* akan digunakan untuk mendapatkan nilai tingkat performa. Pada pengujian ini terdapat beberapa teknik pengujian sesuai kapabilitas yang dimiliki oleh model algoritma tersebut. Adapun konfigurasi parameter dalam mekanisme pengujian ini antara lain:

1. Algoritma yang digunakan *K-Nearest Neighbor(KNN)*
2. Split data menjadi data training dan data testing yang digunakan adalah 0,8 untuk data training dan 0,2 untuk data testing.
3. Normalisasi dengan *Minimax* dengan min 0 dan max 1
4. Pengujian validasi yang digunakan 5-*fold* atau 5 subset.
5. Menggunakan nilai k yaitu 3
6. Pengujian tingkat perfoma yaitu *Accuracy, Precision, Recall*, dan *F1-Score*

Pengujian pertama dilakukan untuk nilai tingkat perfoma akurasi dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.17

Tabel 4.17. *Accuracy* Algorima KNN Dengan Normalisasi *Minimax*

Penentuan K fold	Accuracy
Fold 1	91.11%
Fold 2	94.47%
Fold 3	94.71%
Fold 4	96.15%
Fold 5	95.43%

Hasil dari pengujian nilai tingkat perfoma akurasi tertinggi diperoleh dengan menggunakan penentuan Fold 4 dengan nilai akurasi yang dihasilkan yaitu 96.15%

Berikut ini adalah grafik dari perbandingan akurasi dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.34

Gambar 4.34. Grafik Akurasi Algoritma KNN Dengan Normalisasi *Minimax*

Pengujian kedua dilakukan untuk nilai tingkat perfoma *precision* dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.18

Tabel 4.18. *Precision* Algorima KNN Dengan Normalisasi *Minimax*

Penentuan K fold	Precision
Fold 1	85.07%
Fold 2	90.06%
Fold 3	90.21%
Fold 4	91.45%
Fold 5	91.42%

Hasil dari pengujian nilai tingkat perfoma *precision* tertinggi diperoleh dengan menggunakan penentuan Fold 4 dengan nilai *precision* yang dihasilkan yaitu 91.45%

Berikut ini adalah grafik dari perbandingan *precision* dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.35

Gambar 4.35. Grafik *Precision* Algoritma KNN Dengan Normalisasi *Minimax*

Pengujian ketiga dilakukan untuk nilai tingkat perfoma *recall* dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.19

Tabel 4.19. *Recall* Algorima KNN Dengan Normalisasi *Minimax*

Penentuan K fold	Recall
Fold 1	93.12%
Fold 2	96.25%
Fold 3	96.88%
Fold 4	99.38%
Fold 5	97.53%

Hasil dari pengujian nilai tingkat perfoma *recall* tertinggi diperoleh dengan menggunakan penentuan Fold 4 dengan nilai *recall* yang dihasilkan yaitu 99.38%

Berikut ini adalah grafik dari perbandingan *recall* dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.36

Gambar 4.36. Grafik *Recall* Algoritma KNN Dengan Normalisasi *Minimax*

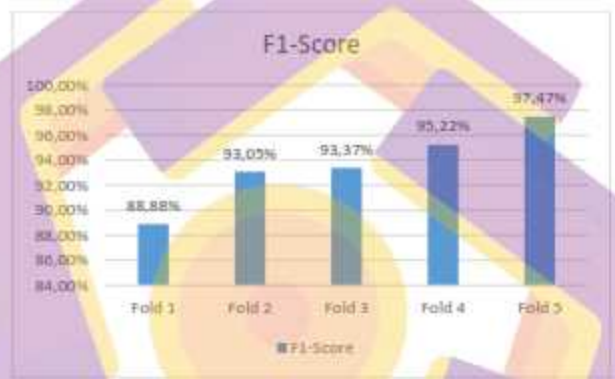
Pengujian keempat dilakukan untuk nilai tingkat perfoma F1- *Score* dengan menggunakan konfigurasi yang telah dilakukan sebelumnya dengan hasil seperti pada tabel 4.20

Tabel 4.20. *F1 Score* Algorima KNN Dengan Normalisasi *Minimax*

Penentuan K fold	F1- Score
Fold 1	88.88%
Fold 2	93.05%
Fold 3	93.37%
Fold 4	95.22%
Fold 5	94.30%

Hasil dari pengujian nilai tingkat perfoma F1- Score tertinggi diperoleh dengan menggunakan penentuan Fold 4 dengan nilai F1- Score yang dihasilkan yaitu 95.22%

Berikut ini adalah grafik dari perbandingan F1- Score dengan konfigurasi penentuan K Fold secara berturut-turut pada Fold 1 s.d Fold 5 seperti pada gambar. 4.37



Gambar 4.37. Grafik *F1 Score* Algoritma KNN Dengan Normalisasi *Minimax*

Pengujian kelima dilakukan untuk nilai AUC untuk mengetahui apakah klasifikasi yang sudah dilakukan sudah baik atau belum.Pada gambar 4.38 adalah hasil AUC dari nilai pengujian akurasi tertinggi yaitu 96.15%dengan nilai AUC 0,980.

Gambar 4.38. AUC Algoritma KNN Dengan Normalisasi *Minimax*

4.4.6. Analisis Perbandingan Hasil Pengujian

Pada bagian ini hasil percobaan dari implementasi *Naïve Bayes* dan *K Nearest Neighbor* dengan percobaan menggunakan normalisasi *Minimax* dan tanpa normalisasi *Minimax* akan dilakukan perbandingan. Hasil yang akan dibandingkan merupakan rata-rata terbaik dari keseluruhan percobaan pengujian dengan parameter Akurasi, *Precision*, *Recall*, F1 *Score* dan AUC yang telah dilakukan pada sub bab sebelumnya. Hasil nilai rata-rata performa pada percobaan terbaik dari semua subset dengan menggunakan dua algoritma dapat dilihat pada Tabel 4.21.

Tabel 4.21 . Hasil Perbandinan Tingkat Performa

No	Skenario	Akurasi	Precision	Recall	F1-Score	AUC
1	*Naïve Bayes*	87.26%	80.01%	89.38%	84.38%	0.944
2	*Naïve Bayes* + Normalisasi *Minimax*	89.42%	86.83%	86.25%	86.23%	0.955
3	KNN	91.12%	86.23%	93.13%	88.95%	0.954
4	KNN + Normalisasi *Minimax*	96.15%	91.45%	99.38%	95.22%	0.980

Berdasarkan hasil pada Tabel 4.19 dapat ditarik suatu kesimpulan bahwa nilai performa unuk akurasi model algoritma *K Nearest Neighbor(KNN)* dengan normalisasi memiliki tingkat akurasi yang lebih baik dengan nilai akurasi sebesar 96.15% jika dibandingkan dengan model algoritma *Naïve Bayes* dengan normalisasi *Minimax* dengan nilai yang lebih rendah sebesar 89.42% sedangkan untuk model algoritma *K Nearest Neighbor(KNN)* tanpa normalisasi Minimax memiliki tingkat akurasi yang lebih baik dengan nilai akurasi sebesar 91.12%jika dibandingkan dengan model algoritma *Naïve Bayes* dengan normalisasi *Minimax* dengan nilai yang lebih rendah sebesar 87,26% Adapun grafik dari perbandingan nilai akurasi pada Gambar 4.39.

Gambar 4.39. Perbandingan Performa Akurasi

Kemudian untuk nilai tingkat performa *precision* model algoritma *K Nearest Neighbor(KNN)* dengan normalisasi memiliki tingkat precision yang lebih baik dengan nilai precision sebesar 91.45% jika dibandingkan dengan model algoritma *Naïve Bayes* dengan normalisasi Minimax dengan nilai yang lebih rendah

sebesar 91.45% sedangkan untuk model algoritma *K Nearest Neighbor(KNN)* tanpa normalisasi *Minimax* memiliki tingkat precision yang lebih baik dengan nilai precision sebesar 86.23% jika dibandingkan dengan model algoritma *Naïve Bayes* dengan normalisasi *Minimax* dengan nilai yang lebih rendah sebesar 89.38%.Adapun grafik dari perbandingan nilai akurasi pada Gambar 4.40.

Gambar 4.40. Perbandingan Performa *Precision*

Kemudian untuk nilai tingkat performa *recall* model algoritma *K Nearest Neighbor(KNN)* dengan normalisasi memiliki tingkat *recall* yang lebih baik dengan nilai *recall* sebesar 99.38% jika dibandingkan dengan model algoritma *Naïve Bayes* dengan normalisasi Minimax dengan nilai yang lebih rendah sebesar 86.25% sedangkan untuk model algoritma *K Nearest Neighbor(KNN)* tanpa normalisasi *Minimax* memiliki tingkat *recall* yang lebih baik dengan nilai recall sebesar 93.13% jika dibandingkan dengan model algoritma *Naïve Bayes* dengan normalisasi *Minimax* dengan nilai yang lebih rendah sebesar 80.01% Adapun grafik dari perbandingan nilai akurasi pada Gambar 4.41.

Gambar 4.41. Perbandingan Performa *Recall*

Kemudian untuk nilai tingkat performa *F1-Score* model algoritma *K Nearest Neighbor(KNN)* dengan normalisasi memiliki tingkat precision yang lebih baik dengan nilai *F1-Score* sebesar 95.22% jika dibandingkan dengan model algoritma *Naïve Bayes* dengan normalisasi *Minimax* dengan nilai yang lebih rendah sebesar 86.23% sedangkan untuk model algoritma *K Nearest Neighbor(KNN)* tanpa normalisasi *Minimax* memiliki tingkat *F1-Score* yang lebih baik dengan nilai F1-Score sebesar 88.95% jika dibandingkan dengan model algoritma *Naïve Bayes* dengan normalisasi *Minimax* dengan nilai yang lebih rendah sebesar 88.95% Adapun grafik dari perbandingan nilai akurasi pada Gambar 4.42.

Gambar 4.42. Perbandingan Performa *Recall*

Kemudian untuk nilai AUC model algoritma *K Nearest Neighbor(KNN)* dengan normalisasi memiliki nilai yang lebih baik dengan nilai precision sebesar 0.980 jika dibandingkan dengan model algoritma *Naïve Bayes* dengan normalisasi Minimax dengan nilai yang lebih rendah sebesar 0.955 sedangkan untuk model algoritma *K Nearest Neighbor(KNN)* tanpa normalisasi *Minimax* memiliki nilai AUC yang lebih baik dengan nilai AUC sebesar 0.954 jika dibandingkan dengan model algoritma *Naïve Bayes* dengan normalisasi Minimax dengan nilai yang lebih rendah sebesar 0.944 Adapun grafik dari perbandingan nilai akurasi pada Gambar 4.43.

Gambar 4.43. Perbandingan AUC

BAB V

PENUTUP

5.1. Kesimpulan

Setelah melalui tahap pengujian tingkat peforma dengan parameter akurasi, precision,recall,F1 Score, dan AUC pada model algoritma *Naïve Bayes dan K Nearest Neighbor(KNN)* untuk diagnose penyakit diabetes, maka dapat diambil beberapa kesimpulan antara lain:

1. Nilai perfoma tertinggi yang didapatkan pada implementasi model algoritma *K Nearest Neighbor(KNN)* dan *Naïve Bayes* yaitu pada algoritma *K Nearest Neighbor(KNN)* yaitu mendapatkan *accuracy* sebesar 96.15%, precision sebesar 91.45%, recall sebesar 99.38%,F1 Score sebesar 95.22%, dan AUC sebesar 0.980 sedangkan untuk algoritma *Naive Bayes* mendapatkan *accuracy* sebesar 89.42%, *precision* sebesar 86.83%, *recall* sebesar 86.25%, F1 *Score* sebesar 86.23%, dan AUC sebesar 0.955. Dari penjelasan diatas dapat disimpulkan bahwa algoritma *K Nearest Neighbor(KNN)* memilki nilai perfoma tertinggi daripada algoritma *Naïve Bayes* yaitu dengan selisih nilai *accuracy* sebesar 6.73%, *precision* sebesar 4.62%, *recall* sebesar 13.13%, *F1 Score* sebesar 8.99%,dan AUC sebesar 0.025
2. Dengan skenario penggunaan normalisasi *Minimax* dan tanpa normalisasi *Minimax* pada implementasi model *K Nearest Neighbor(KNN)* dan *Naïve Bayes* disimpulkan yaitu pada algoritma *K Nearest Neighbor(KNN)* tanpa normalisasi yaitu mendapatkan *accuracy* sebesar 91.12% *precision* sebesar

86.23% *recall* sebesar 93.13% *F1 Score* sebesar 88.95% dan AUC sebesar 0.954 sedangkan algoritma *K Nearest Neighbor(KNN)dengan* normalisasi *Minimax* yaitu mendapatkan *accuracy* sebesar 96.15%, *precision* sebesar 91.45%, *recall* sebesar 99.38%,*F1 Score* sebesar 95.22%, dan AUC sebesar 0.980. Dari penjelasan implementasi algoritman *K Nearest Neighbor(KNN)* normalisasi Minimax dan tanpa normalisasi memiliki selisih nilai *accuracy* sebesar 5.03 %, *precision* sebesar 5.22 %, *recall* sebesar 6.25%, F1 *Score* sebesar 6.27%,dan AUC sebesar 0.026.

Pada algoritma *Naïve Bayes* tanpa normalisasi *Minimax* yaitu mendapatkan *accuracy* sebesar 87,26%, *precision* sebesar 80.01% *recall* sebesar 89.38% *F1 Score* sebesar 89.38% dan AUC sebesar 0.944 sedangkan untuk algoritma *Naive Bayes* dengan normalisasi *Minimax* mendapatkan *accuracy* sebesar 89.42%, *precision* sebesar 86.83%, *recall* sebesar 86.25%, *F1 Score* sebesar 86.23%, dan AUC sebesar 0.955. Dari penjelasan implementasi algoritman*Naïve Bayes* normalisasi *Minimax* dan tanpa normalisasi memiliki selisih nilai *accuracy* sebesar 2.16%, *precision* sebesar 6.8%, *recall* sebesar -3.13%, *F1 Score* sebesar 1.85%,dan AUC sebesar 0.011.

Dari penjelasan diatas dengan penggunaan normalisasi minimax pada model algoritma *K Nearest Neighbor(KNN)* dan *Naïve Bayes* berpengaruh pada hasil implementasi klasifikasi.

5.2. Saran

Saran yang dapat digunakan oleh peneliti-peneliti selanjutnya untuk meningkatkan hasil kesimpulan pada penelitian ini, adalah sebagai berikut :

1. Pada penelitian ini tahap *preprocessing* dilakukan secara manual dan pada penelitian selanjutnya tahap *preprocessing* dapat dilakukan dengan mengimplementasikan teknik *replace value* menggunakan teknik atau algoritma tertentu.
2. Pada penelitian selanjutnya dapat dilakukan pengujian dengan mengimplementasikan model-model algoritma *machine learning* lain serta pendekatan statistik ataupun squensial lain.
3. Pada penelitian selanjutnya dapat dilakukan dengan menggunakan dataset yang lebih baru dan lengkap

DAFTAR PUSTAKA

PUSTAKA BUKU

Santoso, B. 2007. Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu

Sulaeman, M. (1998). Ilmu Budaya Dasar Suatu Pengantar. Bandung: Rafika Aditama

Berry, I. H. and Browne, M. 2006. Lecture Notes in DATA MINING. USA: World Scientific.

Han, J and Kamber, M. 2006. Data Mining Concepts and Techniques, second edition. California: Morgan Kaufman.

Prasetyo, E. 2012. Data Mining Konsep dan Aplikasi Menggunakan MATLAB. Yogyakarta: ANDI Yogyakarta

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Ramesh D and Katheria Y S (2019) Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach Health Technol. (Berl)

Guariguata L, Whiting D R, Hambleton I, Beagley J, Linnenkamp U, and Shaw J E (2014) Global estimates of diabetes prevalence for 2013 and projections for 2035 Diabetes Res. Clin. Pract. 103(2) pp 137–149

Yulianti et all (2021) Sistem Pakar Deteksi Penyakit Diabetes Mellitus (DM) menggunakan Metode Forward chaining dan Certainty factor Berbasis Android , Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi) 5 (1) 2021, 49-55

S. Hardani (2020), "Diagnosa Penyakit Diabetes Dengan Metode Forward Chaining", jitk(Jurnal Ilmu Pengetahuan dan Teknolog), vol. 5, no. 2, pp. 231-236, Feb. 2020.

Fauzia et all (2019), "Epileptic Seizure Detection in EEGs by Using Random Tree Forest, Naïve Bayes and KNN Classification"Journal of Physics: Conference Series. 1505 012055

Dr. Dayanand Jamkhandikar , Neethi Priya, Johan(2020),"Thyroid Disease Prediction Using Feature Selection And Machine Learning Classifiers", The International journal of analytical and experimental modal analysis. ISSN NO:0886-9367

A. R. Bindiya , K. Nikhil , M. S. Sindhu Rashmi, Shafinaz Banu (2020), "Diabetes Mellitus Prediction using Machine Learning Algorithms",International Journal for Research in Applied Science & Engineering Technology (IJRASET). ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue VII July 2020

F. Lia Dwi Cahyanti, Windu Gata, Fajar Sarasati(2021), "Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Dalam Menentukan Tingkat Keberhasilan Immunotherapy Untuk Pengobatan Penyakit Kanker Kulit" Jurnal Ilmiah Universitas Batanghari Jambi. ISSN 1411-8939 (Online), ISSN 2549-4236 (Print) DOI 10.33087/jiubj.v21i1.1189

Annida et all (2020)," Deteksi Penyakit Daun pada Tanaman Padi Menggunakan Algoritma Decision Tree, Random Forest, Naïve Bayes, SVM dan KNN" Jurnal Nasional Informatika dan Teknologi Jaringan. VOL. 5 NO.1 (2020) EDISI SEPTEMBER

Harianto, Didi Rosiyadi(2020), "Komparasi Algortima C4.5, Naïve Bayes, dan k-Nearest Neighbor Sebagai Sistem Pendukung Keputusan Menaikkan Jumlah Peserta Didik" Jurnal Informatika. Vol.7 No.1 April 2020, Halaman 55-61 ISSN: 2355-6579 | E-ISSN: 2528-2247

Sumit Sharma, Mahesh Parmar(2020)," Heart Diseases Prediction using Deep Learning Neural Network Model" International Journal of Innovative Technology and Exploring Engineering (IJITEE). ISSN: 2278-3075, Volume-9 Issue-3, January 2020

Shehzaib et all(2020)," Comparative Analysis of Classifiers for Prediction of Epileptic Seizures", Pakistan Journal of Engineering and Technology (PakJET). Volume: 03, Number: 03, Pages: 84 - 88, Year: 2020

Tsehay Admassu Assegie(2020)," An optimized K-Nearest Neighbor based breast cancer detection" Journal of Robotics and Control (JRC). Volume 2, Issue 3, May 2020 ISSN: 2715-5072 DOI: 10.18196/jrc.2363

Sokolova, M., & Lapalme, G. (2009). Information Processing and Management.A systematic analysis of performance measures for classification tasks, 427-437

E. Indrayuni(2016), "Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization," Jurnal Evolusi, Vol. 4, No.2, hal. 20-27,

Valentino et all(2020) "Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization" Jurnal Nasional Teknik Elektro dan Teknologi Informasi | Vol. 9, No. 2, Mei 2020

Zhi-Hua Zhou and Yuan Jiang(2004). NeC4.5: Neural Ensemble Based C4.5. IEEE Trans. Knowl. Data Eng, 16. 2004.

PUSTAKA ELEKTRONIK

Ghoneim, S. (2019, April 2). Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on? Retrieved April 2020, from Towards Data Science: https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124

Putra, S. (2019, November 27). Kadar Gula darah Normal Berapa Sih?. Dibuka pada 6 Mei 2021. https://health.detik.com/berita-detikhealth/d-4801052/kadar-gula-darah-normal-berapa-sih

F. Gorunescu(2011), Data Mining: Concepts, Models and Techniques, Berlin: Germany: Springer-Verlag Berlin Heidelberg,.

Fawcett, Tom. 2005. An introduction to ROC Analysis. Pattern Recognition Letters, 27, 861-874. doi:10.1016/j.patrec.2005.10.010

LAMPIRAN