MAKALAH

"Implementasi Metode Klasifikasi K-Nearest Neighbors (KNN) pada Dataset Penderita Diabetes"

Makalah ini disusun untuk Memenuhi Tugas Mata Kuliah Kecerdasan Buatan

Dosen Pengampu: Herfandi, M.Kom.

UNIVERSITAS
TEKNOLOGI
SUMBAWA

Disusun oleh:

Azrul Rochmad Rifa'i	(20.01.013.036)
Hifzi Rahmatullah	(20.01.013.039)
Muhammad Fiqar Ramadhan	(20.01.013.034)
Shakira Azzahra Hadi Putri	(20.01.013.041)
Jzidan Muhammad Rusdwian N.	(20.01.013.043)

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS REKAYASA SISTEM

UNIVERSITAS TEKNOLOGI SUMBAWA

2022

ABSTRAK

Diabetes adalah penyakit jangka panjang atau kronis serta ditandai dengan kadar gula (glukosa) darah yang tinggi atau di atas nilai normal. Perbedaan metode dalam dataset merupakan salah satu cara untuk menentukan metode klasifikasi yang benar. Masalah yang diangkat dalam penelitian ini adalah bagaimana mengukur kinerja metode klasifikasi dalam mengelola dataset penderita diabetes. Metode yang digunakan adalah algoritma K-Nearest Neighbor (KNN). Ini adalah metode untuk mengklasifikasikan objek berdasarkan data pelatihan yang paling dekat dengannya. Pada hasil akhir penelitian ini, telah dihitung akurasi tertinggi yaitu 90% pada K=3, presisi tertinggi yaitu 95,88% pada K=3, dan *recall* tertinggi yaitu 88,59% pada K=3.

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa yang masih memberi kesehatan, sehingga penulis dapat menyelesaikan tugas ini, yaitu membuat Makalah tentang "Implementasi Metode Klasifikasi K-Nearest Neighbors (KNN) pada Dataset Penderita Diabetes". Makalah ini dibuat untuk memenuhi salah satu tugas Mata Kuliah Kecerdasan Buatan.

Dalam penulisan makalah ini penulis merasa masih banyak kekurangan-kekurangan baik pada teknis penulisan maupun materi, mengingat akan kemampuan yang dimiliki penulis. Untuk itu kritik dan saran dari semua pihak sangat penulis harapkan demi penyempurnaan pembuatan makalah ini. Penulis mengucapkan terimakasih yang sebesar-besarnya kepada semua pihak yang telah membantu dalam menyusun laporan ini. Penulis juga berharap semoga makalah ini dapat bermanfaat bagi para pembaca.

Dengan segala kerendahan hati, kritik dan saran yang konstruktif sangat penulis harapkan dari para pembaca guna untuk meningkatkan dan memperbaiki pembuatan makalah pada tugas yang lain dan pada waktu mendatang.

Sumbawa, 05 Januari 2021

Penulis

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Diabetes adalah suatu penyakit metabolik yang diakibatkan oleh meningkatnya kadar glukosa atau gula darah. Gula darah merupakan sumber energi yang sangat penting untuk sel dan jaringan, sehingga sangat penting untuk kesehatan Anda. Jaringan komunikasi. Jika tidak diobati dengan benar, diabetes dapat menyebabkan berbagai komplikasi, termasuk penyakit arteri koroner, stroke, obesitas, mata, ginjal, dan gangguan saraf.

Terdapat banyak metode klasifikasi dalam supervised learning pada machine learning, termasuk K-Nearest Neighbor (KNN), Naive Bayes Classifier (NBC), Support Vector Machine (SVM), Neural Netowork (NN), Random Forest Classifier (RFC), Ada Boost Classifier (ABC), serta Quadratic Discriminant Analysis (QDA). Masing-masing metode ini memiliki kelebihan dan kekurangannya. Salah satu kelebihan metode klasifikasi berasal dari pengolahan objek dataset. K-Nearest Neighbor atau KNN adalah algoritma yang mengklasifikasikan data berdasarkan data pelatihan (train dataset) yang diperoleh dari tetangga terdekat K tetangga terdekatnya (Nearest Neighbor).

Dalam penelitian ini, kami menggunakan metode KNN untuk menghitung Akurasi, Presisi, Recall, dan FMeasure berdasarkan nilai K. Prosedur yang dilakukan dalam penelitian ini adalah dengan splitting data training dan data testing, menerapkan metode klasifikasi KNN, serta menghitung performa metode yang akan diuji.

1.2 Rumusan Masalah

Berdasarkan latar belakang identifikasi masalah diatas, maka kami merumuskan permasalahan sebagai berikut.

1. Bagaimana menganalisa keakurasian menggunakan Algoritma K-Nearest Neighbor (KNN) dengan kondisi nilai dari K = 3.
2. Bagaimana melihat hasil akurasi, presisi, dan recall dalam memprediksi penderita diabetes.

1.3 Batasan Masalah

Agar tidak terlepas dari maksud dan tujuan disusunnya laporan ini, maka kami membatasi blabla dengan membagi dataset sebagai data training dan data testing. Dimana perbandigan datanya adalah 80% untuk data training dan 20% untuk data testingnya.

1.4 Tujuan Penelitian

Berdasarkan dari latar belakang permasalahan, maka tujuan yang akan dicapai dalam laporan ini adalah untuk memperoleh keakurasian menggunakan Algoritma K-Nearest Neighbor (KNN) dengan kondisi nilai dari K = 3 serta melihat hasil akurasi, presisi, dan recall dalam memprediksi penderita diabetes jika menggunakan Algoritma KNN.

1.5 Manfaat Penelitian

Manfaat yang diperoleh dengan dicapainya tujuan dari laporan ini, yaitu:

1. Dapat mempermudah tenaga medis dalam memprediksi penderita diabetes.
2. Dapat digunakan sebagai informasi dan tambahan pengetahuan dengan klasifikasi memprediksi penderita diabetes.

BAB II
LANDASAN TEORI

2.1. Data Mining

Data mining adalah teknik yang merupakan gabungan metode-metode analisis data secara berkesinambungan dengan algoritma-algoritma untuk memproses data berukuran besar. Data mining merupakan proses menemukan informasi atau pola yang penting dalam basis data berukuran besar dan merupakan kegiatan untuk menemukan informasi atau pengetahuan yang berguna secara otomatis dari data yang jumlahnya besar (Kusrini, 2009). Data mining, sering juga disebut knowledge discovery in database (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan pola keteraturan, pola hubungan dalam set data berukuran besar. Keluaran dari data mining ini dapat dijadikan untuk memperbaiki pengambilan keputusan di masa depan. Dalam data mining data disimpan secara elektronik dan diolah secara otomatis, atau setidaknya disimpan dalam komputer. Data mining adalah tentang menyelesaikan masalah dengan menganalisa data yang telah ada dalam database (Kusrini, 2009).

Siklus hidup proyek data mining menurut Cross-Industry Standart Proses for Data Mining (CRISP-DM) yang dikembangkan tahun 1996 terbagi dalam 6 fase (Kusrini, 2009). Berikut gambar dari Siklus hidup proyek data mining :

1) Fase Pemahaman Bisnis (Business Understanding Phase)
    a. Penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan.
    b. Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan data mining.
    c. Menyiapkan strategi awal untuk mencapai tujuan.
2) Fase pemahaman data (Data Understanding Phase)
    a. Mengumpulkan data.
    b. Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal.
    c. Mengevaluasi kualitas data.

d. Jika diinginkan, pilih sebagian kecil group data yang mungkin mengundang pola dari permasalahan.

3) Fase pengolahan data (Data Preparation Phase)

Siapkan data awal, kumpulkan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu dilaksanakan secara intensif.

a. Pilih kasus dan variabel yang ingin dianalisis dan yang sesuai analisis yang akan dilakukan.
b. Lakukan perubahan pada beberapa variabel jika dibutuhkan.
c. Siapkan data awal sehingga siap untuk perangkat pemodelan.

4) Fase Pemodelan (Modelling Phase)

a. Pilih dan aplikasikan teknik pemodelan yang sesuai.
b. Perlu diperhatikan bahwa beberapa teknik mungkin untuk digunakan pada permasalahan data mining yang sama.
c. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data ke dalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik data mining tertentu.

5) Fase Evaluasi

a. Pengevaluasi satu atau lebih model yang digunakan dalam fase permodelan untuk mendapatkan kualitas dan efektivitas sebelum disebarkan untuk digunakan.
b. Menetapkan apakah terdapat model yang memenihi tujuan pada fase awal.
c. Menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani degan baik.
d. Mengambil keputusan yang berkaitan dengan penggunaan hasil dari data mining.

6) Fase Penyebaran

a. Menggunakan model yang dihasilkan. Terbentuknya model tidak menandakan telah terselesaikannya proyek.
b. Contoh sederhana penyebaran: pembuatan laporan.
c. Contoh kompleks penyebaran: penerapan proses data mining secara paralel pada department lain.

Menurut Larose dalam bukunya yang berjudul ”Discovering Knowledge in Data: An Introduction to Data Mining”, datamining dibagi menjadi beberapa kelompok berdasarkan tugas/pekerjaan yang dapat dilakukan (Larose, 2005), yaitu :

1. Deskripsi

   Terkadang peneliti dan analisis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Deskripsi dari pola kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderunga

2. Estimasi

   Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model dibangun menggunakan baris data (record) lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

3. Prediksi

   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

4. Klasifikasi

   Dalam klasifikasi, terdapat target variabel kategori.Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

5. Pengklasteran (Clusterring)

   Pengklasteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas obyek-obyek yang memiliki kemiripan. Klaster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan record dalam klaster yang lain. Berbeda dengan klasifikasi, pada pengklasteran tidak ada variabel target. Pengklasteran tidak melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target, akan tetapi, algoritma pengklasteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok- kelompok yang memiliki kemiripan (homogen), yang mana kemiripan record dalam satu

kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.

6. Asosiasi

Tugas asosiasi dalam data mining adalah untuk menemukan atribut yangmuncul dalam satu waktu. Salah satuimplementasi dari asosiasi adalah market basket analysis atau analisiskeranjang belanja, sebagaimana yang akan dibahas dalam penelitian ini.

2.2 Algoritma KNN (K-Nearest Neighbors)

KNN adalah sebuah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumya. Termasuk dalam supervised learning, dimana hasil query instance yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam KNN. Deskripsi kNN Diberikan titik query, akan ditemukan sejumlah k obyek atau (titik training) yang paling dekat dengan titik query. Klasifikasi menggunakan voting terbanyak diantara klasifikasi dari k obyek Algoritma k-nearest neighbor (KNN) menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari query instance yang baru.

NN adalah Number Neighbors (Jumlah Neighbors), CV adalah Cross Validation, AVG adalah rata-rata performa dari tiap CV, dan AVG Perf adalah penjumlahan dari accuracy, precision, recall, dan f1 score dibagi dengan 4. AVG Perf untuk melihat rata-rata metric performa yang dihasilkan oleh tiap algoritma.

Tabel 6. Performa KNN

Dist	NN	Metric	CV1	CV2	CV3	CV4	CV5	AVG	AVG PERF
EUCLIDEAN	5	acc	0.746	0.754	0.723	0.777	0.768	0.754	0.766
		prec	0.794	0.783	0.769	0.806	0.802	0.791	
		rec	0.746	0.754	0.723	0.777	0.768	0.754	
		f1score	0.755	0.761	0.731	0.783	0.774	0.761	
	7	acc	0.723	0.746	0.719	0.763	0.772	0.745	0.757
		prec	0.775	0.769	0.768	0.802	0.801	0.783	
		rec	0.723	0.746	0.719	0.763	0.772	0.745	
		f1score	0.732	0.751	0.728	0.770	0.774	0.751	
	9	acc	0.723	0.746	0.719	0.763	0.772	0.745	0.760
		prec	0.794	0.783	0.769	0.806	0.802	0.791	
		rec	0.723	0.746	0.719	0.763	0.772	0.745	
		f1score	0.732	0.751	0.728	0.770	0.774	0.751	
MINKOWSKI	5	acc	0.746	0.754	0.723	0.777	0.768	0.754	0.765
		prec	0.794	0.783	0.769	0.806	0.802	0.791	
		rec	0.746	0.754	0.723	0.754	0.768	0.749	
		f1score	0.755	0.761	0.731	0.761	0.774	0.756	
	7	acc	0.723	0.746	0.719	0.763	0.772	0.745	0.756
		prec	0.775	0.769	0.768	0.802	0.801	0.783	
		rec	0.723	0.746	0.719	0.746	0.772	0.741	
		f1score	0.732	0.751	0.728	0.751	0.774	0.747	
	9	acc	0.723	0.746	0.719	0.763	0.772	0.745	0.759
		prec	0.794	0.783	0.769	0.806	0.802	0.791	
		rec	0.723	0.746	0.719	0.746	0.772	0.741	
		f1score	0.732	0.751	0.728	0.751	0.774	0.747	

Berdasarkan perhitungan performa yang dilakukan, algoritma KNN dengan distance Euclidean dan jumlah neighbors 5 memiliki performa terbaik. Perbedaan performa memang tidak signifikan, namun KNN dengan distance Euclidean dan jumlah neighbors 5 akan dibandingkan performanya dengan SVM dan CNN.

2.3 Rapid Miner

RapidMiner merupakan perangkat lunak yang bersifat terbuka (open source). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik.

2.4 Crisp-Dm

CRISP-DM adalah singkatan dari Cross Industry Standard Process for Data Mining, sebuah proses penambangan data yang dikembangkan bersama oleh DaimlerChrysler, SPSS, dan NCR. Namanya adalah proses netral yang dapat digunakan di semua industri dan berbagai alat.

Sebagai sebuah metodologi, CRISP-DM menggambarkan fase dari tahapan - tahapan dalam sebuah proyek, pekerjaan yang terkait dalam tiap fase dan penjabaran terkait hubungan antar pekerjaan tersebut serta memberikan sebuah gambaran siklus hidup (life-cycle) dari Data Mining bila dilihat sebagai Model Proses.

Dari penggambaran tersebut metode ini memberikan sebuah proses standar yang bersifat umum atau tidak eksklusif dalam strategi pemecahan masalahdalam sebuah unit bisnis atau penelitian dengan menggunakan Data Mining yang sesuai atau tepat.

Pada metode CRISP-DM ini memiliki 6 model tahapan yaiu :

1. Business/Research Understanding : Melakukan pengumpulan data perihal Business objective, peniliaian terkait kondisi terkini, menetapkan tujuan dari proses data mining, dan mengembangkan rencana proyek.
2. Data Understanding : Mengumpulkan data awal, menjelaskan data, menyelidiki data, Penilaian kualitas data merupakan salah satu langkah dalam fase ini. Dalam fase ini Eksplorasi data juga dilakukan pada ringkasan statistik yang mungkin ditampilkan pada akhir fase ini. Cluster data untuk melihat pola data yang terbentuk.

3. Data Preparation : Setelah data didapatkan perlu dilakukan proses sebuah proses seleksi, cleansing, dibuat dalam bentuk tertentu, dan di format sesuai kebutuhan.
4. Modelling : Setelah data dibersihkan dan dibentuk sesuai kebutuhan kemudian dibutuhkan sebuah modeling yang sesuai dan dikalibrasi perihal pengaturan agar didapatkan hasil optimal. Bila dibutuhkan kembali dapat dilakukan data preparation agar data dapat sesuai dengan teknik data mining yang dibutuhkan.
5. Evaluation : Setelah didapatkan sebuah atau beberapa model sehingga dilakukan penilaian terkait kualitas dan efektifitas-nya. Kemudia ditentukan model seperti apa yang digunakan agar sesuai dengan objective pada fase 1 hingga diambil sebuah keputusan penggunaan dari hasil data mining.
6. Deployment : Pada fase ini secara umum ada 2 aktifitas yang dilakukan yaitu Perencanaan dan monitoring hasil dari proses deployment serta melengkapi keseluruhan aktifitas sehingga menghasilkan laporan terakhir dan melakukan review dari proyek yang dilakukan.

2.5 Dataset

Definisi dataset adalah kumpulan data yang berasal dari informasi. Di masa lalu, Anda siap mengelola dengan informasi baru. Menurut Dayat Suryana Dalam Controls Visual Basic Volume 1, istilah record adalah Penyimpanan tidak terhubung (terputus).

BAB III
METODE PENELITIAN

3.1 Objek Penelitian

Objek penelitian adalah suatu tempat yang akan diselidiki dalam kegiatan penelitian untuk menelusuri masalah dan menerapkan hasil dari penelitian tersebut. Penelitian ini dilakukan di rumah masing-masing dimulai dari tanggal 30 Desember 2021.

3.2 Jenis dan Sumber Data

3.2.1 Jenis Data

Dalam penelitian ini kami menggunakan jenis data kuantitatif yang dijadikan sebagai pendukung dalam penyelesaian tugas ini. Definisi dan Jenis dari data yang di ambil oleh penulis dari objek penelitian yaitu menggunakan Data Kuantitatif. Data kuantitatif adalah data dari hasil penelitian yang bersifat terstruktur atau berpola sehingga ragam data yang diperoleh dari sumber riset lebih mudah dibaca oleh peneliti.

3.2.2. Sumber Data

Sumber data yang digunakan penulis dalam mendukung penelitian untuk menyelesaikan tugas akhir ini yaitu data primer dan data sekunder. Adapun pengertian dan contoh dari data yang diambil penulis pada objek penelitian adalah:

1) Data Primer

   Data primer adalah jenis data yang dikumpulkan secara langsung dari sumber utamanya seperti melalui wawancara, survei, dataset statistik, dan sebagainya. Dalam pengumpulan data primer dalam penelitian ini menggunakan metode dataset statistik yang dimana penggunaan dataset statistik ini merupakan penggunaan data yang sudah tersedia.

2) Data Sekunder

Data sekunder adalah data pendukung yang sumbernya didapat dari sumber yang telah ada atau peneliti sebagai tangan kedua. Data sekunder dapat diperoleh dari berbagai sumber seperti laporan, jurnal, dan lainya. Data sekunder yang digunakan dalam penelitian ini adalah data yang berhubungan dengan data sebelumnya.

3.3 Tahapan Alur Penelitian

Tahapan penelitian yang digunakan pada laporan ini sesuai menurut Crisp-dm, yaitu sebagai berikut :

1. Business Understanding

   Pada tahap ini peneliti memahami masalah pada object penelitian kemudian mencari solusi dan tujuan untuk menyelesaikan masalah tersebut.

2. Data Understanding

   Di tahap ini peneliti menetapkan dan menungmpulkan data yang di butuhkan dari data penderita Diabetes kemudian didefinisikan berdasarkan solusi dan tujuan penelitian

3. Data Preparation

   Pada data preparation melakukan pengolahan data untuk mempersiapkan data sebelum masuk proses pemodelan. Dalam melakukan pengolahan data, akan dilakukan beberapa tahapan agar pada akhirnya akan didapatkan data yang bisa digunakan pada tahap berikutnya.

4. Modeling

   Pada tahap modeling kita menggunakan tools dari software rapidminer untuk menganalisis dataset dengan pendekatan klasifikasi dan menggunakan algoritma K-Nearest Neighbour (KNN)

5. Hasil dan Validasi

   Pada tahap ini kita menampilkan hasil yang didapatkan dari analisis berdasarkan pendekatan KNN

BAB IV
HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Kami mengambil dataset dari website Kaggle.com, dimana pada website tersebut terdapat banyak dataset yang tersaji secara lengkap. Pada laporan ini kami mengambil dataset terkait penderita diabetes dengan total data sebanyak 2000 data. Setiap baris berisi tentang informasi terkait kondisi tubuh yang dialami oleh pasien.

4.2 Business Understanding

Masalah yang dihadapi yaitu bagaimana memeproleh keakuratan predikisi penderita Diabetes dari data penderita Diabetes. Kami menawarkan solusi dengan menggunakan algoritman K-Nearest Neighbor pada proses data mining untuk menganalisa pola dari data penderita diabetes sehingga membantu untuk memperoleh hasil prediksi penderita diabetes secara akurat

4.3 Data understanding

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	2	138	62	35	0	33,6	0,127	47	Yes
3	0	84	82	31	125	38,2	0,233	23	No
4	0	145	0	0	0	44,2	0,63	31	Yes
5	0	135	68	42	250	42,3	0,365	24	Yes
6	1	139	62	41	480	40,7	0,536	21	No
7	0	173	78	32	265	46,5	1,159	58	No
8	4	99	72	17	0	25,6	0,294	28	No
9	8	194	80	0	0	26,1	0,551	67	No
10	2	83	65	28	66	36,8	0,629	24	No
11	2	89	90	30	0	33,5	0,292	42	No
12	4	99	68	38	0	32,8	0,145	33	No
13	4	125	70	18	122	28,9	1,144	45	Yes
14	3	80	0	0	0	0	0,174	22	No
15	6	166	74	0	0	26,6	0,304	66	No
16	5	110	68	0	0	26	0,292	30	No
17	2	81	72	15	76	30,1	0,547	25	No
18	7	195	70	33	145	25,1	0,163	55	Yes
19	6	154	74	32	193	29,3	0,839	39	No
20	2	117	90	19	71	25,2	0,313	21	No

Tabel Atribut Dataset dan Deskripsinya

Atribut	Singkatan	Deskripsi	Satuan	Tipe Data
Pregnant	Pregnant	Banyaknya kehamilan	-	Numerik
Plasma-Glucose	Glucose	Kadar glukosa dua jam setelah makan	Mg/dL	Numerik
Diastolic Blood- Pressure	DBP	Tekanan darah	Mm Hg	Numerik
Tricepts Skin Fold Thickness	TSFT	Ketebalan kulit	mm	Numerik
Insulin	INS	Insulin	mu U/ml	Numerik
Body Mass Index	BMI	Berat Tubuh	Kg/m2	Numerik
Diabetes pedigree Function	DPF	Riwayat Keturunan yang terkena diabetes	-	Numerik
Age	Age	Umur	Years	Numerik
Outcome	Outcome	Positif Diabetes (Yes) dan Negatif Diabetes (No)	-	Nominal

Adapun penjelasan dari masing-masing atribut tersebut ialah :

- **Pregnancies** merupakan atribut yang menjelaskan tentang kehamilan.
- **Glucose** merupakan atribut yang menjelaskan tentang kadar gula darah pasien.
- **BloodPressure** merupakan atribut yang menjelaskan tentang tekanan darah pasien.
- **SkinThickness** merupakan atribut yang menjelaskan ketebalan kulit pasien.
- **Insulin** merupakan atribut yang menjelaskan hormone penyerapan glukosa kedalam sel tubuh pasien untuk mengendalikan gula darah.

- **BMI (Body Mass Indeks)** merupakan atribut yang menjelaskan tentang berat badan pasien.
- **DiabetesPedigreeFunction** merupakan atribut yang menjelaskan riwayat keturunan yang terkena diabetes.
- **Outcome** merupakan label bagi data set yang menentukan klasifikasi diabetes.

4.4 Data Preparation

Pada laporan ini data penderita diabetes merupakan dataset yang kami teliti sudah merupakan data yang telah dibersihkan sehingga data yang disajikan sudah lengkap dan tidak memiliki data *y*ang *missing value* dan selanjutnya tinggal diaplikasikan di tahap modeling

4.5 Modeling

Pada tahap ini metode data mining untuk menemukan pengetahuan tersembunyi dan berharga dari data. Metode yang digunakan adalah Klasifikasi dengan algoritma KNN

K-Nearest Neighbor (K-NN) termasuk kelompok instance-based learning. Algoritma ini juga merupakan salah satu teknik lazy learning. kNN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing. diperlukan suatu sistem klasifikasi sebagai sebuah sistem yang mampu mencari informasi. Contoh kasus, missal diinginkan untuk mencari solusi terhadap masalah seorang pasien baru dengan menggunakan solusi dari pasien lama. Perhitungan jarak ketetanggan menggunakan algoritma euclidien seperti yang ditunjukkan pada persamaan 1.

$$euc = \sqrt{((a\_1 - b\_1)^2 + \cdots + (a\_n - b\_n)^2)}$$

Dimana a = a1,a2, …, an, dan b = b1, b2, …, bn mewakili n nilai atribut dari dua record. Untuk atribut dengan nilai kategori. Algoritma K-Nearest Neighbor (K-NN) adalah sebuah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumya. Termasuk dalam *supervised learning*, dimana hasil *query instance* yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam K-NN. Algoritma ini bekerja dengan berdasarkan pada jarak terpendek dari sample uji ke sample latih untuk menentukan KNNnya. Setelah mengumpulkan KNN, kemudian diambil mayoritas dari KNN untuk

dijadikan prediksi dari sample uji. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak Eucledian. Langkahlangkah untuk menghitung metode K-Nearest Neighbor antara lain:

1. Menentukan parameter K
2. Menghitung jarak antara *data training* dan *data testing* Perhitungan jarak yang paling umum dipakai pada perhitungan pada algoritma KNN adalah menggunakan perhitungan jarak Euclidean. Rumusannya adalah sebagai berikut:

$$euc = \sqrt{(\sum_{i=1}^{n}(pi - qi)^2)}$$

dimana :

pi = sample data / *data training*

qi = data uji / *data testing*

i = variabel data

n = dimensi data

3. Mengurutkan jarak yang terbentuk
4. Menentukan jarak terdekat sampai urutan K
5. Memasangkan kelas yang bersesuaian
6. Mencari jumlah kelas dari tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan dievaluasi

- **Akurasi**

  Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai actual. rumus akurasi dipaparkan pada persamaan 2.

- **Presisi**

  Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. Presisi dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. rumus presisi ditunjukkan pada persamaan 3.

- **Recall**

  Recall didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. Rumus Recall diuraikan pada persamaan 4.

AKURASI=((TP+TN))/((TP+TN+FP+FN))

PRESISI=TP/(TP+FP)

RECALL=TP/(TP+FN)

F-Measure=2 (Presisi x Recall)/(Presisi+Recall)

Keterangan Variabel:

TP : True Positif

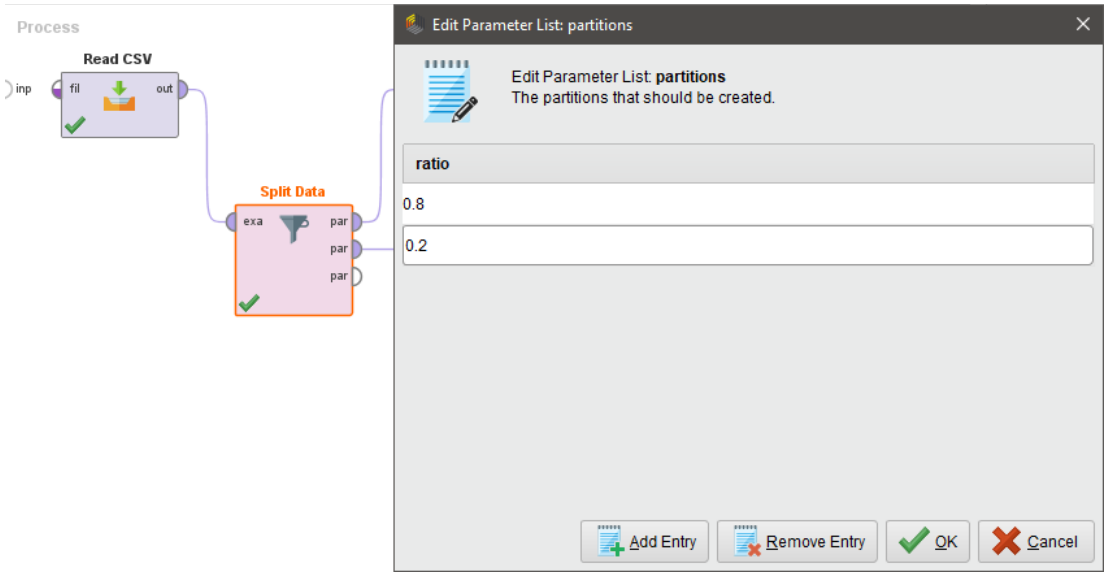
TN : True Negatif

FP : False Positif

FN : False Negatif

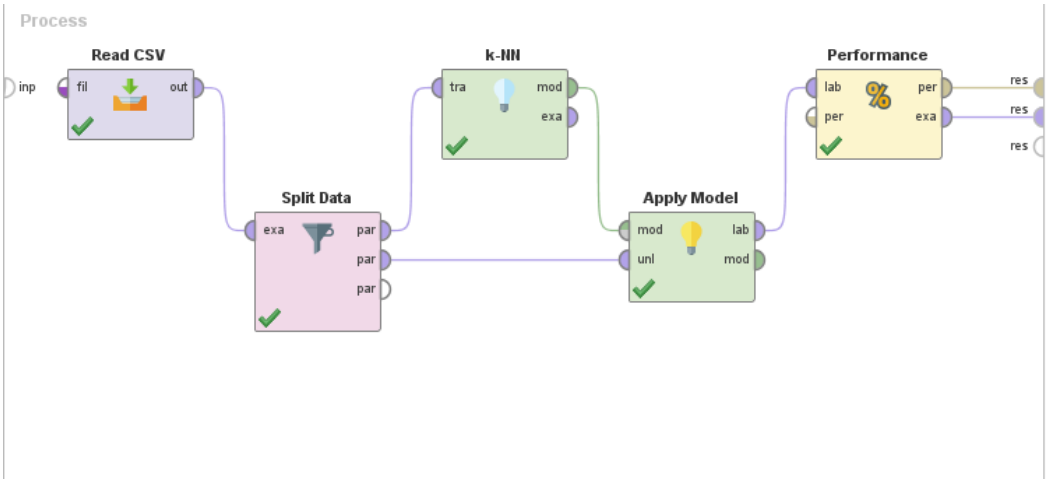
4.5.1. Implementasi pada Rapid Miner

Seperti yang telah dipaparkan sebelumnya bahwa tahapan yang dilakukan pada penelitian ini adalah dengan melakukan pembagian data training dan data testing, data yang digunakan sebanyak 2000 data, dengan pembagian sebesar 80% sebagai data training dan 20% sebagai data tesing. Tahapan selanjutnya adalah menerapkan metode KNN, pemilihan nilai K pada makalah ini yaitu nilai K = 3.



Gambar Pembagian Data Training dan Data Testing

Pada tahap ini, kami memberikan pembagian/perbandingan antara data training dan data testingnya, yaitu 80% atau 0,8 untuk data yang akan di training dan 20% atau 0,2 untuk data yang akan di testing.



Gambar Pemodelan Data Menggunakan KNN

Pada tahap ini, kami membuat model data yang akan dijalankan nantinya.

4.6 Hasil dan Validasi

accuracy: 90.00%

	true Yes	true No	class precision
pred. Yes	127	30	80.89%
pred. No	10	233	95.88%
class recall	92.70%	88.59%	

Gambar Tingkat Akurasi

precision: 95.88% (positive class: No)

	true Yes	true No	class precision
pred. Yes	127	30	80.89%
pred. No	10	233	95.88%
class recall	92.70%	88.59%	

Gambar Tingkat Presisi

recall: 88.59% (positive class: No)

	true Yes	true No	class precision
pred. Yes	127	30	80.89%
pred. No	10	233	95.88%
class recall	92.70%	88.59%	

Gambar Tingkat Recall

Berdasarkan gambar di atas, hasil yang didapat adalah dengan perbandingan antara data training sebesar 80% dan 20% untuk data testing menghasilkan tingkat akurasi cukup tinggi, yaitu sebesar 90%. Kemudian tingkat presisi sebesar 95,88% dan tingkat *recall* sebesar 88,59%.

BAB V
PENUTUP

5.1 Kesimpulan

Dari hasil perhitungan algoritma K-Nearest Neighbor (KNN) di atas, yang dimana untuk perbandingan antara data training sebesar 80% dan untuk data testing sebesar 20% telah mendapatkan pola atau *sample* sebanyak 400 data. Ini hasil pembagian dari jumlah keseluruhan data pada dataset yang berjumlah 2000 data. Sebanyak 1600 data di jadikan sebagai data training dan 400 sisanya dijadikan sebagai data testing. Dari sample tersebut, maka telah didapatkan hasil akurasi tertinggi yaitu 90% pada K=3, presisi tertinggi yaitu 95,88% pada K=3, dan *recall* tertinggi yaitu 88,59% pada K=3. Nilai yang diperoleh sangat baik dikarenakan tingkat keakurasian prediksi dari data yang disajikan memperoleh hasil yang tinggi.

5.2. Saran

Pada penelitian selanjutnya diharapkan dapat mengidentifikasi penyakit diabetes dengan pengembangan metode yang lain dan terbaru dengan tingkat akurasi yang lebih tinggi, salah satunya adalah metode C 4.5 dengan menggunakan cangkupan variable yang lebih luas lagi seperti gejala gejala yang dialami penderita penyakit diabetes, beberapa diantaranya adalah gejala sering haus dan sering kencing dimalam hari. Aplikasi ini juga diharapkan dapat dikembangkan kedalam versi mobile atau android yang dapat turut serta menampilkan hasil perhitungan jarak euclidean.

DAFTAR PUSTAKA

Naufal, Mohammad Farid. 2021. "Analisis Perbandingan Algoritma SVM, KNN, dan CNN untuk Klasifikasi Citra Cuaca" Vol. 8, No. 2. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK).

Admin Daovers. 2020. "RapidMiner : Mengenal Aplikasi Data Mining Terkemuka di Dunia", https://www.doavers.com/blog/rapidminer-mengenal-aplikasi-data-mining-terkemuka-di-dunia, diakses pada 05 Januari 2022 pukul 21.23.

Firman, Muhammad. 2019. "Pengertian Data Mining dan Penerapannya", https://www.kompasiana.com/mfirman34/5c8fb0557a6d88244e001272/pengertian-data-mining-dan-penerapannya?page=all, diakses pada 05 Januari 2022, pukul 22.48.

Journal of Applied Informatics and Computing (JAIC) Vol.5, No.2, Desember 2021, pp. 103~108 e-ISSN: 2548-6861, diakses pada 12 januai 2022 pukul 10.42