Pengaruh Principal Component Analysis Pada Naïve Bayes dan K-Nearest Neighbor Untuk Prediksi Dini Diabetes Melitus Menggunakan Rapidminer

Siti Rokhanah[1] *, Arief Hermawan[2], Donny Avianto [3]
[123] Magister Teknologi Informasi, Universitas Teknologi Yogyakarta
Indonesia
** Corresponding Author.E-mail: siti.6210211005@student.uty.ac.id*

Abstrak

Penderita diabetesmelitus mengalami gangguan pada sistem metabolisme yang disebabkan oleh pankreas yang tidak memproduksi insulin atau penggunakan insulin dalam metabolisme belum efektif semakin banyak. Kepedulian akan hidup sehat menurun drastis, sehingga lonjakan kematian penyakit tersebut tinggi. Banyak orang belum memahami gejala dini yang muncul sehingga sulit untuk sembuh. Hal ini dikarenakan belum adanya prediksi dini penderita penyakit tersebut. Dalam kajian ini menjelaskan pengaruh analisis komponen utama (PCA) untuk menemukan fitur optimal dalam klasifikasi prediksi dini diabetes melitus pada naïve bayers dan k-nearest neighbor ditambah aplikasi rapidminer yang bersifat terbuka bisa digunakan sebagai alat uji keakuratan data. Bahan penelitian yang digunakan bersumber dari Dataset Prediksi Risiko Diabetes Tahap Awal Learning Repository dari website Kaggle yaitu diabet_data_upload.csv. Jumlah record yang digunakan adalah 520 baris data dan 17 nama tabel untuk setiap baris data yang ada. Tujuan penggunaan kedua metode pengelompokan adalah untuk menunjukkan akurasi paling akurat dari data yang diolah. Hasil penelitian memberikan kajian bahwa formula k-nearest-neighbor dengan principal component analysis dapat bekerja lebih baik dibandingkan dengan k-nearest-neighbor saja. Performansi k- nearest neighbor dengan principal component analysis (PCA) lebih baik dengan nilai akurasi sebesar 93.27%, sedangkan akurasi tanpa analisis komponen utama dalam hal ini hanya menggunakan algoritma k-nearest-neighbor hanya sebesar 90.70. Hasil ini diperoleh dengan mempertimbangkan record yang ada dan nilai k = 5, kemudian diperoleh hasil bahwa algoritma k-nearest neighbor menggunakan metode principal component analysis (PCA) untuk mengklasifikasikan diagnosis diabetes didapatkan tinggi. Hasil nilai yang tepat.

Keywords: **Kelompok; DiabetesMelitus; Naïve Bayes; k-Nearest Neighbor; PCA.**

Abstract

Patients with diabetes mellitus experience disturbances in the metabolic system caused by the pancreas not producing insulin or using insulin in metabolism that is not effective more and more. Concern for healthy living has decreased drastically, so the spike in deaths from this disease is high. Many people do not understand the early symptoms that appear, making it difficult to recover. This is because there is no early prediction of sufferers of the disease. This study explains the effect of principal component analysis (PCA) to find optimal features in the classification of early prediction of diabetes mellitus in naïve Bayers and k-nearest neighbors plus the open rapidminer application that can be used as a test tool for data accuracy. The research material used comes from the Learning Repository Early Stage Diabetes Risk Prediction Dataset from the Kaggle website, namely diabet_data_upload.csv. The number of records used is 520 rows of data and 17 table names for each existing row of data. The purpose of using the two grouping methods is to show the most accurate accuracy of the processed data. The results of the study provide a study that the k-nearest-neighbor formula with principal component analysis can work better than just k-nearest-neighbor. The performance of k-nearest neighbor with principal component analysis (PCA) is better with an accuracy value of 93.27%, while the accuracy without principal component analysis in this case only

uses the k-nearest-neighbor algorithm is only 90.70. These results are obtained by considering the existing records and the value of k = 5, then the result is that the k-nearest neighbor algorithm uses the principal component analysis (PCA) method to classify diabetes diagnoses as high. Exact value result..

***Keywords**: Group; Diabetes Mellitus; Naïve Bayes; k-Nearest Neighbor; PCA.*

1. Introduction

Prediksi dini dari diabetesmelitus adalah suatu penyakit pada metabolisme yang disebabkan ketidakmampuan tubuh dalam penggunaan hormone insulin yang belum efekti [2]. Saat ini diabetes mempengaruhi orang-orang dari segala usia. Hampir melebihi dari 1,2 juta anak dan remaja di hampir belahan bumi menderita diabetesmelitus [2]. Penyakit diabetesmelitus termasuk daftar penyakit paling menyebabkan kematian di dunia [13]. Ketika tahun 2021, orang yang terimbas akibat diabetes sedunia hampir mencapai 6.7 juta [13]. Menurut Organisasi Kesehatan Dunia, jumlah pasien diabetesmellitus akan bertambah, hingga mencapai angka empat kali lipat.

. Menurut Organisasi Kesehatan Dunia, pasien diabetes terus bertambah, hingga mencapai angka empat kali lebih cepat dari waktu sebelumnya. Penyakit diabetesmelitus merenggut hamper 999 ribu pasien setiap tahun. Kadar gula darah bisa dikendalikan asalkan pasien menjaga pola makannanya jika perlu minum obat sedini mungkin sebelum penyakitnya semakin parah. Itulah mengapa, orang awam harus mengetahui indikator penyebab penyakit diabetes [5].

Peneliti sebelumnya telah banyak melakukan penelitian untuk menganalisis identifikasi penyakit diabetes melitus, salah satunya adalah sebagai berikut:

Algoritma naive bayes pada klasifikasi diabetes mellitus dengan isi penelitian adalah menguji klasifikasi diabetes dengan algoritma naive bayes dan menunjukkan hasil dengan akurasi 90.20 pada kumpulan data yang sama dengan data dan karakteristik data. Salah satu learning machine yang digunakan adalah Rapid Miner [6].

Penelitian berikut adalah perbandingan k-nearest neighbor dan naive bayes untuk mendeteksi diabetesmelitus, dengan penelitian menggunakan k-nearest neighbor dan naive bayes dengan sekunder data, jumlah dan fitur yang sama. Hasil akurasi yang diperoleh adalah 90.70%. [7]. Mendasari hasil kajian dahulu yang telah dijelaskan di atas, digarisbawahi bahwa kNN dan Naive Bayes terindikasi mempunyai kelebihan sendiri-sendiri. Dalam sebuah kajian, keduanya ini dibandingkan dengan menggunakan metode Knowledge Discovery in Database (KDD),

Hasil kajian keduanya menunjukkan bahwa skor akurasi kNN lebih akurat. Penelitian ini membandingkan algoritma kNN dan naive bayes dalam identifikasi diabetesmelitus. Implemantasi kNN didasarkan pada kenyataan bahwa kNN salah satu cara yang efisien untuk digunakan pada

data yang luas, dapat menahan data percobaan yang acak, dan memiliki kinerja yang tepat. Pada saat yang sama, alasan untuk menggunakan algoritma naive bayes adalah cepat dan efisien di dunia cloud, dan hanya butuh data percobaan yang lebih sempit untuk beroperasi.

Terinformasi seminar penelitian yang membuat perbandiangan performa kNN dan naïve bayes dalam mendasari klasifikasi pasien diabetes dengan menggunakan metode principal component analysis (PCA) tidak ada. Misi yang digaris bawahi dengan adanya kajian ini seharusnya bisa memperkaya informasi yang bermanfaat tentang diabetes kepada banyak pihak dan kemudian kelak dapat menjadi study banding untuk penelitian pada masa setelahnya.

2. *Materials and Methods*

2.1. Materials

A. **Penyakit Diabetesmelitus**

Pengertian tentang gangguan penyakit adalah kondisi yang terjadi ketika mekanisme yang menyeimbangkan fungsi atau sistem tubuh terganggu atau gagal sehingga tubuh tidak dalam keadaan baik. Suatu gangguan tubuh begitu menyusahkan dan dapat mempengaruhi ketidakstabilan pada fungsi sistem tubuh, penyakit tidak hanya dilihat dari secara kasat mata saja, tetapi juga terdapat kelainan pada fungsi sistem metabolik [9]

Diabetesmelitus (kencing manis) merupakan penyakit yang dapat menghilangkan nyawa manusia, merenggutnya hampir 999 ribu orang setiap periodenya. Menurut Organisasi Kesehatan Dunia, pasien akan terus mengalami kenaikan, hingga menembus angka empat kali lebih cepat dari masa sebelumnya [5]

B. **Data**

Data adalah kumpulan kejadian yang berasal dari pengukuran. Hasil keputusan yang rasional adalah penarikan kesimpulan dari informasi atau fakta yang tepat dan pula digarisbawahi bahwa data adalah suatu realitas yang mengdeskripsikan peristiwa atau entitas yang masih belum disentuh dan masih akan diolah untuk menghasilkan informasi [10].

C. Klasifikasi

Klasifikasi dapat dikatakan aliran proses pembelajaran untuk menunjuk satu himpunan atribut dari suatu objek, atau dengan arti lain bahwa klasifikasi adalah suatu proses pengelompokan data. Pengelompokan adalah penataan objek ke dalam kategori-kategori yang memiliki kegunaan yang sama. Pengelompokan digunakan untuk menggambarkan kumpulan data di mana setiap tipe data adalah nominal atau biner. Saat mengklasifikasikan data yang dipantau, itu dibagi menjadi dua entitas, yaitu entitas coba dan entitas uji, dimana entitas latih dikelola menggunakan metode data minning. Diantaranya adalah beberapa algoritma untuk fungsi klasifikasi yaitu k nearest neihgbor, naive bayes dan sebagainya [11].

D. Data Minning

Data Minning adalah proses pengelolaan data besar untuk mendapatkan informasi yang akurat dari data untuk memfasilitasi penyelesaian masalah dan membuat kebijakan. Nama lain untuk data minning tersebut adalah analisis pola, ekstraksi pengetahuan, pemanenan informasi, dll. [12]

Informasi tersebut merupakan rangkuman dari pernyataan data dan statistik. Informasi ini merupakan hasil ekstraksi data. Istilah lain untuk knowledge mining adalah penemuan pengetahuan dalam data, ada juga big data yaitu. intelijen bisnis, ekstraksi pengetahuan dan analisis pola atau pengumpulan data. Ide dari data mining adalah sejumlah besar kumpulan data telah terkumpul, kemudian metode data minning diolah dengan formula yang mengambil data atau mengekstraknya sebagai informasi. Informasi ini dapat digunakan untuk satu tujuan. Proses formula diatas sangat bermanfaat untuk banyak tujuan [13].

E. Algoritme Naïve Bayes

Algoritme naive bayes adalah cara pembelajaran formula untuk masalah pengelompokan, terutama digunakan untuk pengelompokan huruf yang berisi kumpulan entitas percobaan dimensi tinggi. Diantaranya adalah analisis sentimen, pemfilteran spam, dan klasifikasi, yang selain diketahui karena kecepatannya tetapi juga karena keefektifannya. Algoritme naive bayes memungkinkan membuat model dengan cepat, menjadikannya algoritme prediksi pembelajaran tercepat. Algoritma ini menggunakan kemungkinan target. Mengapa disebut Algoritma Naive Bayes karena menganggap bahwa kemunculan suatu fitur tidak bergantung pada keberadaan fitur lainnya, meskipun fitur tersebut saling bergantung atau keberadaan fitur lain, semua fitur tersebut secara individual mempengaruhi probabilitas dan karenanya dianggap seperti algoritma naif dan terkait dengan ahli statistik dan filsuf bernama Thomas bayes. Dasar dari algoritma naive bayes adalah teorema dasar, juga dikenal sebagai aturan bayes atau hukum bayes.

Formula ini merupakan metode untuk menghitung kemungkian bersyarat, yaitu kemungkinan suatu kejadian yang diberikan informasi sebelumnya [14]. Naive Bayes memiliki kemampuan untuk membangun model dengan cepat, kemampuan prediktif, dan juga menawarkan cara baru untuk mempelajari dan bekerjasama dengan data. Formula Naive Bayes hanya berkompromi pada atribut tipe data diskrit atau diskrit, atau tidak berkompromi terhadap atribut nilai kontinyu (numerik), dan semua atribut dapat berdiri sendiri dan menjadi atribut yang mempengaruhi atribut yang diprediksi. Secara singkat, formula klasifikasi Naive Bayes adalah pengklasifikasi kumpulan data statistik yang memprediksi semua kemungkinan untuk setiap anggota kelas. Jaringan saraf dan pohon keputusan memiliki kekuatan klasifikasi yang sama dengan naive bayes, yang didasarkan pada teorema bayes.

F. Principal component analysis (PCA)

Analisa komponen utama adalah formula yang dipakai sebagai alat untuk mengurangi dimensi data menjadi bentuk dalam lingkup nilai yang hamper sama (Raysyah, Arinal, & Mulyana, 2021). Principal component analysis dimanfaatkan untuk mengekstraksi susunan dari kumpulan entitas yang cukup multidimensi, karena (Ilmaniati dan Putro, 2019) menjelaskan bahwa formul PCA (Principal Component Analysis) lebih sesuai apabila tujuan kajian adalah membandingkan data dengan meringkas yang lebih besar. nomor. variabel. Formula ini juga dapat digunakan ketika menganalisa apakah variabel yang akan dikaji saling berkesinambungan atau tidak. Formula PCA biasanya sebagai berikut (Muhtadi, 2017):

1. Tentukan rata-rata atau nilai tengah data.
2. Tentukan Matriks Kovarian, menggunakan rumus 1.

$$(\quad) = \frac{\sum(x\mathrm{i} - \bar{x})(y\mathrm{i} - \bar{y}}{n - 1} \quad \text{(I)}$$

Dimana dan adalah variant yang berbeda yang terdapat pada rumus, sedangkan $\bar{x}$ dan $\bar{y}$ adalah nilai tengah dari variant x dan y , serta n adalah total data.

3. Tentukan nilai eigen dengan rumus 2.

$(A - \lambda I) = 0$ (2)

dimana adalah matriks, λ dan I adalah nilai eigen dari .

4. Tentukan nilai vektor eigen dengan Persamaan 3.

$[A - \lambda I][\mathrm{X}] = 0$ (3)

5. Hitunglah variabel baru dengan cara mengkalikan variant asli dengan matriks vektor eigen.

G. K-Nearest Neighbor (kNN)

Algoritme k-nearest-neighbor (kNN) adalah formula dasar pembelajaran kelompok berbasis contoh (Kripsiandita, Arifianto dan A'yun, 2021). Algoritma kNN mengklasifikasikan data baru berdasarkan jarak percobaan terdekat dengan target, atau sering disebut sebagai tetangga terdekat (Yustanti, 2012). Algoritma tersebut mempunyai percobaan terawasi, dimana hasil rumus yang diuji dirangking berdasarkan tetangga terdekat dengan data yang paling banyak diuji (Novita, Harsani, & Qur'ania, 2018). Banyaknya tetangga terdekat disebut K. Nilai k dapat ditentukan dengan n+1, dimana n adalah jumlah elemen, atau dengan cara ganjil dan genap, jika nilai n ganjil maka nilai k . bersifat langsung dan sebaliknya. Langkah-langkah penentuan jarak terpendek adalah: Pertama, entitas dibagi menjadi entitas coba dan entitas uji, setelah terkumpul entitas coba dan entitas uji, dihitung jarak masing-masing entitas sampel dengan entitas coba (Setianto, Kusrini u Henderi, 2019).). Jarak Euclidean dapat digunakan untuk menghitung jarak dan rumusnya akan ditampilkan Persamaan 4.

$$d(xixj) \quad = \sqrt{\sum_{r=1}^{n}(ar(xi) - ar\ (xi))} \quad ^2$$

dimana $d(x_i x_j)$ adalah jarak *Euclidean*, $x_i x_j$ data ke i, *record* ke j, a_r adalah data ke-r, n adalah dimensi objek.

H. Rapid Miner

Aplikasi Rapidminer menggunakan metode berorientasi objek dalam hierarki Java dan dapat digunakan di hampir semua platform sistem operasi. Rapidminer fleksibel digunakan dan diimplementasikan pada level yang tidak sama. Rapidminer mengimplementasikan algoritme pembelajaran yang di implementasikan ke lembar entitas dari baris perintah. Rapidminer mencakup alat untuk preprocessing entitas, klasifikasi, pengelompokan, regresi, kerjasama, dan visual. Rapidminer dapat memproses entitas, mengintegrasikannya ke dalam rencana pembelajaran, dan menganalisis pengklasifikasi yang dihasilkan serta kinerjanya. semua ini tanpa menulis kode apa pun. Implementasi Rapidminer adalah menerapkan metode pembelajaran ke kumpulan data dan menganalisis hasilnya untuk mendapatkan wawasan dari data, atau mengimplementasikan sejumlah formula dan membandingkan kinerja sesuka hati dalam kumpulan data untuk menjaga pengguna tetap fokus pada kumpulan data yang digunakan.

RapidMiner merupakan aplikasi yang banyak memperoleh penghargaan diantaranya pada tahun 2017, RapidMiner mendapat penghargaan dari KDNuggets sebagai the most populer general platform for data mining/data science [7] RapidMiner pertama kali dikembangkan pada tahun 2001 oleh Raft Klinkenberg, Ingo Mierswa, dan Simon Fischer [7]. Perangkat lunak ini dapat bekerja pada lingkungan standalone dan jaringan. RapidMiner dapat berintegrasi dengan data mining, text mining, machine learning, analisis prediksi, dan analisis bisnis [7]. Untuk data minning sendiri dengan banyak fungsi sehingga lebih mudah untuk diimplementasikan.

2.2. Methods

Diagnosis diabetesmelitus menjadi fokus kajian yang dilakukan dalam kajian ini. Entitas yang digunakan berasal dari Kaggle.com. Entitas tersebut berasal dari National Institute of Diabetes and Digestive and Kidney Diseases. Entitas yang digunakan berjumlah 520 data. Variabel yang digunakan adalah pemasukan, kadar gula, tekanan darah, insulin, indeks massa tubuh, umur dan silsilah diabetes. Dan alat yang digunakan dalam penelitian ini adalah Rapid Minner. Entitas diolah menggunakan algoritma k-Nearest Neighbor dan Naive Bayes. Teknik data mining digunakan untuk mengklasifikasikan data diagnosis penyakit diabetes melitus.

1. Preprosesing Data

Langkah ini termasuk proses pembersihan data. Menghapus beberapa atribut (kolom) yang tidak digunakan dalam penelitian ini. Atribut tidak digunakan Penelitian ini adalah kehamilan dan ketebalan kulit. Namun entitas yang digunakan dalam penelitian ini sudah lengkap dan tidak kekurangan nilai.

2. Transformasi Entitas

Melakukan perubahan pada atribut *pemasukan* dari bentuk karakter *Y* dan *Tdk* menjadi *angka 0 dan 1*

Tabel 1 : Data Diabetes Melitus

P1	P2	P3	P4	P5	P6	P7
Tdk	89	76	37	31	0.2	23
Y	88	30	99	55	0.5	26
Tdk	118	58	94	33	0.3	23
Y	117	88	145	35	0.4	40
Tdk	136	74	204	37	0.4	24

Sumber : Dataset diabetes Kaggle

Keterangan :

1) P1 = *Pemasukan*
2) P2 = *Kadar gula*
3) P3 = *Tekanan Darah*
4) P4 = Insulin
5) P5 = BMI
6) P6 = *DPF*
7) P7 = *Umur*

Table 1 berisi data awal daridiabetesmelitus, kemudian dilakukan perubahan pada bagian atribut penghasilan sehingga dihasilkan data seperti pada Tabel 2.

Tabel 2. Data diabetes melitus setelah transformasi

H1	H2	H3	H4	H5	H6	H7
0	89	76	37	31.20	0.19	23
1	88	30	99	55	0.49	26
0	118	58	94	33.30	0.26	23
1	117	88	145	34.50	0.40	40
0	136	74	204	37.40	0.39	24

3. Data Mining

Pada tahap ini, peneliti membuat model dengan menggunakan dua algoritma klasifikasi, kNN dan Naive Bayes. Setelah itu, juga menerapkan perhitungan *Principal Component Analysis* (PCA) untuk kedua algoritma di atas.

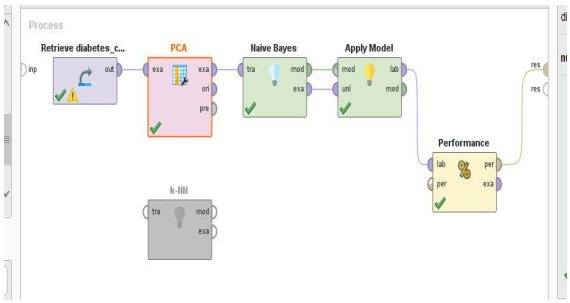
4. Evaluation

Kemudian dilakukan evaluasi model, bentuk evaluasinya terdiri dari perhitungan nilai akurasi, memori dan presisi dari kedua algoritma yang digunakan, serta analisis komponen utama (PCA) yang diterapkan

3. Results and Discussion

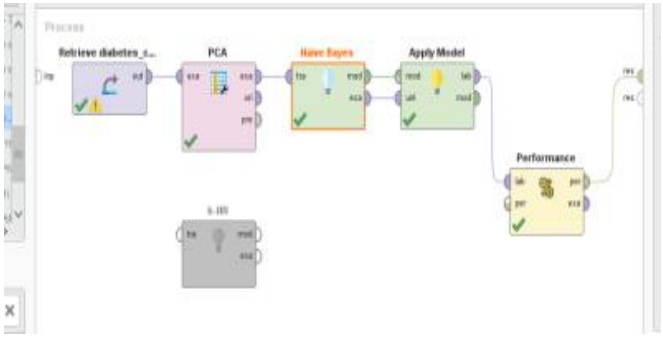
Penelitian ini melakukan pengujian dengan menerapkan prosentase untuk data latih dan data uji sebesar 80% dan 20%. Pengujian melampaui 2 tahapan yaitu pengujian pada algoritma naïve bayes dengan penerapan metode PCA dan pengujian algoritma k-nearest neighbor dengan penerapan metode PCA menggunakan mesin learning rapid miner.

Pengujian untuk algoritma naïve bayes dengan implementasi algoritme PCA didapatkan hasil akurasi, presisi dan recall seperti sesuai pada gambar 1.



Gambar 1. Visualisasi naïve bayes dengan PCA

PCA yang diterapkan adalah dengan nilai k = 5 , maka setelah melalui proses dengan rapid miner akan diperoleh nilai akurasi, recall dan precision dari algoritma naïve bayes dan PCA adalah dapat dijabarkan sesuai gambar 2 dan tabel 3 dibawah ini :



Gambar 2 . Visualisasi PCA pada Naïve Bayes dengan k = 5

accuracy: 90.19%

	true 0	true 1	class precision
pred. 0	281	12	95.90%
pred. 1	39	188	82.82%
class recall	87.81%	94.00%	

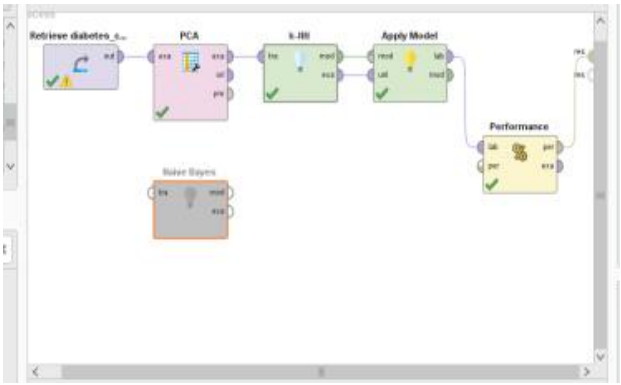
Gambar 3 . Confusion matrix Naïve Bayes dan PC

Tabel 4 . Data Hasil Akurasi Naïve Bayes + PCA

	True 1	True 0
Pred 1	39	198
Pred 0	12	281

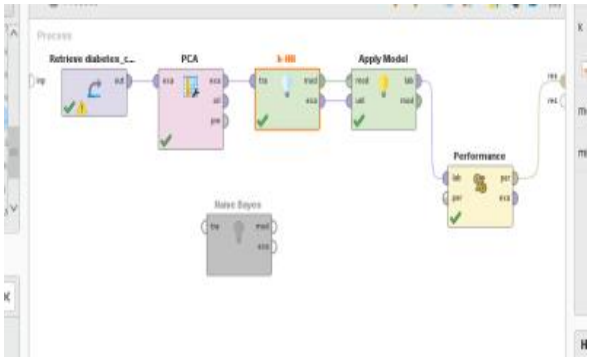
Nama Metode	Precision		Recall	
	1	0	1	0
Naïve Bayers + PCA	82.82	95.90	94.00	87.81

Pengujian untuk algoritma k-nearest neighbor dengan implementasi formula PCA didapatkan hasil akurasi, presisi dan recall seperti ditunjukkan pada gambar 4.

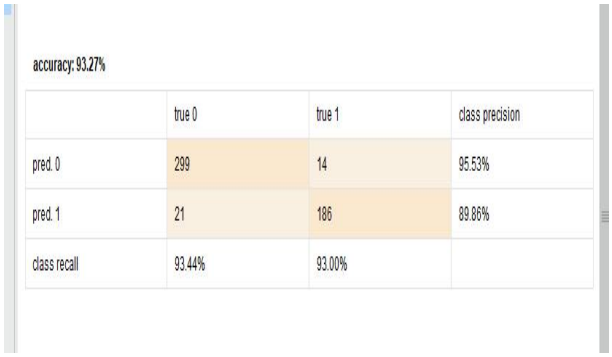


Gambar 4. Visualisasi kNN dengan PCA

PCA yang diterapkan adalah dengan nilai k = 5 , maka setelah melalui proses dengan rapid miner akan diperoleh nilai akurasi, recall dan precision dari algoritma naïve bayes dan PCA adalah dapat dijabarkan sesuai gambar 5 dan tabel 4 dibawah ini :



Gambar 5. Visualisasi kNN dengan PCA



Gambar 5. Confusion matrix kNN dengan PCA

Tabel 4. Data Hasil akurasi pada kNN + PCA

	True 1	**True 0**
Pred 1	196	21
Pred 0	14	298

Nama Metode	Precision		Recall	
	1	0	1	0
kNN + PCA	89.90	95.57	93.00	93.27

Berikut dapat dijelaskan hasil perbandingan

penelitian hanya menggunakan algoritma k-nearest neighbor tanpa adanya pengaruh PCA dan penerapan k nearest neighbor dengan adanya pengaruh principal component analisis (PCA):

Author	data	Metode	Akurasi
[11]	Dataset Prediksi Risiko Diabetes Tahap Awal	Naïve bayes	90.20
[12]	Dataset Prediksi Risiko Diabetes Tahap Awal	k-nearest neighbor	75.00
[owner]	Dataset Prediksi Risiko Diabetes Tahap Awal	k-nearest neighbor dengan PCA	93.27

4. Conclusions

Penelitian ini menganalisis pengaruh PCA terhadap dua metode data minnig yaitu kNN dan Naive Bayes dalam melakukan klasifikasi pasien diabetesmelitus. Secara keseluruhan, hasil kajian menunjukkan bahwa tingkat kinerja metode naive Bayes dan PCA menurun ketika k = 1 sampai 5 digunakan. PCA naïve Bayesian memiliki akurasi tertinggi sebesar 90.19% pada k=5, sedangkan PCA dan kNN memiliki akurasi sebesar 93.27% untuk n = 5 dan k = 5 dengan jumlah data dan fitur yang sama sebagai bahan klasifikasi prediksi dini diabetesmellitus. Pada akhirnya penelitian ini dapat memberikan manfaat untuk membantu masyarakat memprediksi dini gejala diabetesmelitus sehingga dapat mudah untuk ditangani.

References

[1] Ardiyansyah, Panny Agustia Rahayuningsih, and Reza Maulana. 2018. "Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner." Jurnal Khatulistiwa Informatika VI(1):20–28.

[2] Fernanda IS, Ratnawati ED, Adikara PP (2017) Identifikasi Penyakit Diabetes Mellitus Menggunakan Metode Modified KNearest Neighbor (MKNN), Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 1, No. 6, Juni 2017, hlm. 507-513 ← *Journal*

[3] S. E. Viswapriyaa, (2019). International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV8IS100318 (This work is licensed under a Creative Commons Attribution 4.0 International License.) Published by : www.ijert.org Vol. 8 Issue 10, October-2019 . ← *Journal*

[4] Novianto Dian, Sugihartono Tri (2020). Sistem Deteksi Kualitas Buah Jambu Air Berdasarkan Warna Kulit Menggunakan Algoritma Principal Component Analysis (Pca) dan K-Nearest Neigbor (K-NN),

Jurnal Ilmiah Informatika Global Volume 11 No. 2 Desember 2020 ISSN PRINT : 2302-500X ISSN ONLINE : 2477-3786← *Journal*

[5] Ramadhani AR, Niswatin KR (2018). Sistem Diagnosa Diabetes Menggunakan Metode K-NN Jurnal Sains dan Informatika p-ISSN: 2460-173X Volume 4, Nomor 2, November 2018 e-ISSN: 2598-58414079(200202)37:1%3C51::AID-CRAT51%3E3.0.CO;2-N ←*Journal*

[6] Li, M., Xing, S., Yang, L., Fu, J., Lv, P., Wang, Z., Yuan, Z. (2019). Nickel-loaded ZSM-5 catalysed hydrogenation of oleic acid: The game between acid sites and metal centres. *Applied Catalysis A: General.* 587, 117112. DOI: 10.1016/j.apcata.2019.117112. ←*Journal*

[7] Natasuwarna, A.P., 2019, Data Mining dengan Penerapan Aplikasi RapidMiner, Pustakaone, Jogjakarta

[8] Argina MA (2020) Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes Indonesian Journal of Data and Science ISSN: 2715-9930 ←*Journal*

[9] Singh KH, Vishnavat K, R. Srinivasan (2018) Employee Performance And Leave Management Using Data Minning Technique. International Journal of Pure and Applied Mathematics Volume 118 No. 20 2018, 2063-2069 ISSN: 1311-8080 (printed version); ISSN: 1314-3395 ← *Journal*

[10] Adebayo OA , Chaubey SM (2019). Data Minning Classification Technique On The Analysis Of Students Performance., GSJ: Volume 7, Issue 4, April 2019, Online: ISSN 2320-9186 ←*Journal*

[11] Yahya, Gunawan Indra, Harianto Bambang (2017). Penerapan PCA dan K-NN Untuk Meningkatkan Nilai Akurasi Pengenalan Wajah. JURNAL INFORMATIKA HAMZANWADI Vol. 2 No. 1, Mei 2017, hal. 81-90 ISSN: 2527 - 6069 ←*Journal*

[12] BN Azmi, A. Hermawan, D. Avianto (2020). Analisis Pengaruh PCA Pada Klasifikasi Kualitas Air Menggunakan Algoritma K-Nearest Neighbor dan Logistic Regression. ←*Journal*

[13] Ridwan Achmad (2020) Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus ←*Journal*

[14] Putry M. N, Sari NB M.Kom (2022) Komparasi Algoritma KNN dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Melitus ←*Journal*

[15] A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," Procedia Comput. Sci.,vol. 85, pp. 662–668, 2016, doi: 0.1016/j.procs.2016.05.251.

[16] Hermawan Arief, Wibowo Adityo Permana, Wijaya A Setiawan, (2022) The Improvement of Artificial Neural Network Accuracy Using Principle Component Analysis Approach←Jurnal

[17] Krismawan DA, Rachmawanto EH, (2022) Principal Component Analysis (PCA) dan K-nearest Neighbor (KNN) dalam Deteksi Masker Pada Wajah. ←*Journal*