

Analisis Perbandingan Algoritma Naïve Bayes dan Decision Tree Untuk Memprediksi Kehamilan Pada Penderita Diabetes

Aulia Ulfah

Teknik Informatika, Program Pascasarjana, Universitas Pamulang

e-mail: s2auliaulfah@gmail.com

Abstrak—Penyakit diabetes atau penyakit gula darah adalah kondisi kronis jangka panjang yang ditandai oleh peningkatan kadar gula dalam darah atau glukosa. Kondisi ini tidak hanya berdampak pada aspek fisik, tetapi juga dapat berpengaruh signifikan terhadap kesuburan dan kesehatan reproduksi, baik pada pria maupun wanita. Diabetes dapat menyebabkan gangguan hormonal yang pada gilirannya dapat menghambat proses implantasi atau bahkan pembuahan yang terjadi. Dalam upaya meningkatkan pemahaman dan pendekatan terhadap kesehatan reproduksi penderita diabetes, teknologi machine learning menjadi sebuah solusi inovatif. Algoritma machine learning, seperti Algoritma Naïve Bayes dan Algoritma Decision Tree, memiliki sifat konstruktif yang memungkinkan pembuatan prediksi yang akurat serta memberikan analisis yang mendalam. Penelitian ini bertujuan untuk membandingkan akurasi dan kinerja kedua algoritma tersebut dalam memprediksi kehamilan pada penderita diabetes. Hasil penelitian menunjukkan bahwa Algoritma Naïve Bayes mampu mencapai tingkat akurasi yang lebih tinggi dibandingkan dengan Algoritma Decision Tree, memberikan harapan untuk pengembangan pendekatan yang lebih efektif dalam manajemen kesehatan reproduksi bagi mereka yang mengidap diabetes.

Kata Kunci—Decision Tree; Diabetes; Naive Bayes; Orange; Prediksi.

I. PENDAHULUAN

Penyakit diabetes atau penyakit gula darah merupakan suatu kondisi kronis jangka panjang yang muncul karena tubuh pengidapnya tidak mampu mengambil gula (glukosa) ke dalam sel dan menggunakannya sebagai sumber energi. Hal ini mengakibatkan penumpukan gula ekstra dalam aliran darah tubuh. Diabetes yang tidak terkontrol dengan baik dapat menyebabkan konsekuensi serius, merusak berbagai organ dan jaringan tubuh. Kondisi ini juga memiliki dampak signifikan pada kesuburan dan kesehatan reproduksi, tidak hanya pada pria tetapi juga pada wanita. Diabetes dapat menyebabkan ketidakseimbangan hormonal yang pada akhirnya dapat mempengaruhi proses implantasi atau pembuahan, bahkan menyebabkan keterlambatan.

Dalam era kemajuan teknologi medis, *machine learning* menjadi sebuah terobosan yang dapat digunakan untuk memprediksi kehamilan pada penderita diabetes. Algoritma *machine learning*, seperti Algoritma Naïve Bayes dan Algoritma Decision Tree, memberikan kontribusi konstruktif dengan kemampuannya dalam membuat prediksi yang akurat dan memberikan analisis yang mendalam.

Penelitian ini memiliki tujuan yaitu membandingkan akurasi dan kinerja dua algoritma *machine learning*, Algoritma Naïve Bayes dan Algoritma Decision Tree, dalam memprediksi kehamilan pada penderita diabetes. Hasil penelitian menunjukkan bahwa Algoritma Naïve Bayes mencapai tingkat akurasi yang lebih tinggi dibandingkan Algoritma Decision Tree, memberikan pandangan optimis terhadap potensi penerapan metode ini dalam pemantauan dan manajemen kesehatan reproduksi bagi individu yang hidup dengan diabetes.

II. TINJAUAN PUSTAKA

Penelitian ini menggunakan beberapa penelitian terdahulu sebagai acuan. Penelitian ini membandingkan akurasi dua algoritma, yaitu Algoritma Naïve Bayes dan Algoritma Decision Tree untuk mengetahui nilai akurasi tertinggi dalam memprediksi tingkat kehamilan pada penderita diabetes, data sampel yang digunakan didapatkan dari Kaggle.

A. Analisis

Menurut Kamus besar bahasa Indonesia, analisis adalah penguraian suatu pokok atas berbagai bagiannya dan penelaahan bagian itu sendiri serta hubungan antar bagian untuk memperoleh pengertian yang tepat dan pemahaman arti keseluruhan. Menurut Nana Sudjana menyatakan Analisis adalah usaha memilah suatu integritas menjadi unsur- unsur atau bagian bagian sehingga jelas hierarkinya dan susunannya [1].

B. Perbandingan

Menurut Sjachran Basah perbandingan merupakan suatu metode pengkajian atau penyelidikan dengan mengadakan perbandingan di antara dua objek kajian atau lebih untuk menambah dan memperdalam pengetahuan tentang objek yang dikaji [2]. Jadi di dalam perbandingan ini terdapat objek yang hendak diperbandingkan yang sudah diketahui sebelumnya, akan tetapi pengetahuan ini belum tegas dan jelas.

C. Algoritma Naïve Bayes

Menurut Olson Delen menjelaskan Naïve Bayes untuk setiap kelas keputusan, menghitung probabilitas dengan syarat bahwa kelas keputusan adalah benar, mengingat vektor informasi obyek. Algoritma ini mengasumsikan bahwa atribut obyek adalah independen. Probabilitas yang terlibat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dari master tabel keputusan [3].

D. Algoritma Decision Tree

Algoritma yaitu sebuah algoritma yang digunakan untuk membentuk decision tree (pohon keputusan). Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Algoritma mengkonstruksi pohon keputusan dari data pelatihan, yang berupa kasus-kasus atau *record* dalam basis data. Setiap kasus berisikan nilai dari atribut-atribut untuk sebuah kelas. Setiap atribut dapat berisi data diskret atau kontinyu (numerik). Algoritma juga menangani kasus yang tidak memiliki nilai untuk sebuah atau lebih atribut.

E. Prediksi

Prediksi adalah suatu proses memperkirakan secara sistematis tentang sesuatu yang paling mungkin terjadi di masa depan berdasarkan informasi masa lalu dan sekarang yang dimiliki, agar kesalahannya (selisih antara sesuatu yang terjadi dengan hasil perkiraan) dapat diperkecil. Prediksi tidak harus memberikan jawaban secara pasti kejadian yang akan terjadi, melainkan berusaha untuk mencari jawaban sedekat mungkin yang akan terjadi [4].

F. Kehamilan

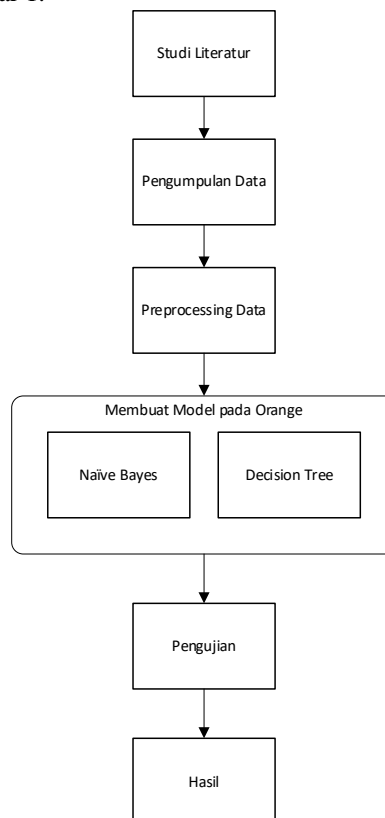
Kehamilan merupakan suatu proses fisiologis yang terjadi pada perempuan akibat adanya pembuahan antara sel kelamin laki-laki dan sel kelamin perempuan. Dengan kata lain, kehamilan adalah pembuahan ovum oleh spermatozoa, sehingga mengalami nidasi pada uterus dan berkembang sampai kelahiran janin [5].

G. Diabetes

Diabetes melitus adalah penyakit gangguan metabolik menahun akibat pankreas tidak memproduksi cukup insulin atau tubuh tidak dapat menggunakan insulin yang diproduksi secara efektif, akibatnya terjadi peningkatan konsentrasi glukosa di dalam darah atau hiperglikemia (PUSDATIN Kemenkes RI, 2019). Menurut Smeltzer & Bare, diabetes melitus merupakan suatu penyakit kronis yang menimbulkan gangguan multisistem dan mempunyai karakteristik hiperglikemia yang disebabkan defisiensi insulin atau kerja insulin yang tidak kuat [6].

III. METODE PENELITIAN

Penelitian ini dilakukan dalam beberapa tahap yaitu tinjauan Pustaka, pengumpulan data, membuat model, pengujian dan hasil. Alur tahapan ini dapat dilihat pada gambar 1.



Gambar 1.
Alur Tahapan Penelitian

A. Studi Literatur

Pada tahap ini dilakukan untuk mempelajari masalah kehamilan yang terkait dengan penyakit diabetes, informasi tentang algoritma Naïve Bayes dan algoritma Decision Tree dan informasi lain yang terkait dengan masalah yang sedang dibahas. Referensi pada penelitian ini diperoleh dari berbagai macam sumber, diantaranya yaitu ebook dan laporan penelitian sebelumnya.

B. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah dataset *Diabetes Prediction System* yang tersedia pada Kaggle yang merupakan salah satu situs terkenal di dunia *Data Science* dan *Machine Learning* yang terdiri dari ribuan dataset. Atribut yang digunakan dalam penelitian ini yaitu: Glukosa, Tekanan Darah, Ketebalan Kulit, Insulin, BMI (Body Massa Index), Genetik, Usia, dan Hasil.

C. Preprocessing Data

Preprocessing data atau pra-pemrosesan data adalah proses mengubah data mentah menjadi format data yang dapat digunakan.

D. Membuat Model pada Orange

Pada tahap ini akan dibuat dua model dengan menggunakan algoritma Naïve Bayes dan algoritma Decision Tree yang selanjutnya akan dilakukan perbandingan nilai akurasi, presisi, *recall* dan *f1-score* masing-masing algoritma dalam melakukan prediksi kehamilan pada penderita diabetes. Akurasi merupakan rasio dari hasil prediksi yang benar. Presisi adalah rasio prediksi positif yang benar dengan hasil keseluruhan dari prediksi positif. *Recall* adalah rasio antara prediksi yang benar-benar positif dengan data global yang sebenarnya positif. Kemudian *f1-score* adalah perbandingan antara presisi rata-rata dan *recall*.

E. Pengujian Model

Pada tahapan ini, dilakukan pengujian model yang dibuat menggunakan algoritma Naïve Bayes dan Decision Tree. pengujian model dilakukan menggunakan aplikasi Orange yang bertujuan untuk membandingkan akurasi dan kinerja dua algoritma untuk memprediksi kehamilan pada penderita diabetes.

Rumus untuk menghitung akurasi, presisi, dan *recall* masing-masing ditunjukkan pada rumus dibawah ini (Goutte & Gaussier, 2005). [7]

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

$$\text{Presisi} = \frac{TP}{FP + TP} * 100\%$$

$$\text{Recall} = \frac{TP}{FN + TP} * 100\%$$

Keterangan:

- TP : Banyaknya data positif yang benar
- TN : Banyaknya data negative yang benar
- FN : Banyaknya data negative yang salah
- FP : Banyaknya data positif yang salah

Berikut adalah perhitungan manual untuk mendapatkan nilai akurasi dari algoritma Naive Bayes dan Algoritma Decision Tree menggunakan rumus yang tertera diatas untuk nantinya dilakukan persamaan dengan *Test and Score* yang didapatkan menggunakan aplikasi Orange.

1. Menentukan nilai akurasi pada Algoritma Naive Bayes

$$\begin{aligned} \text{Akurasi} &= \frac{TP + TN}{TP + TN + FP + FN} * 100\% \\ &= \frac{9 + 11}{9 + 11 + 0 + 5} * 100\% \\ &= \frac{20}{25} * 100\% \\ &= 0.800 \end{aligned}$$

Dari perhitungan diatas, maka didapatkan nilai akurasi untuk algoritma Naive Bayes sebesar 0.800

2. Menentukan nilai akurasi pada Algoritma Decision Tree

$$\begin{aligned}
 \text{Akurasi} &= \frac{TP + TN}{TP + TN + FP + FN} * 100\% \\
 &= \frac{7 + 18}{7 + 18 + 8 + 7} * 100\% \\
 &= \frac{25}{40} * 100\% \\
 &= 0.625
 \end{aligned}$$

Dari perhitungan diatas, maka didapatkan nilai akurasi untuk algoritma Decision Tree sebesar 0.625.

Algoritma Naïve Bayes adalah pengklasifikasi menggunakan metode probabilistik dan statistik yang diusulkan oleh seorang ilmuwan Inggris, Thomas Bayes. Naïve Bayes adalah model yang populer dalam aplikasi pembelajaran mesin karena kesederhanaannya yang memungkinkan semua atribut berkontribusi sama pada keputusan akhir. Kesederhanaan ini setara dengan efisiensi komputasi, yang membuat teknik Naïve Bayes menarik dan cocok untuk banyak bidang. Rumus teorema Bayes dapat dilihat pada persamaan berikut:

$$P(Q|X) = \frac{P(X|Q) P(Q)}{P(X)}$$

Keterangan:

X : Data dengan kelas tidak diketahui

Q : Hipotesis X pada class spesifik

$P(Q|X)$: Probabilitas hipotesis Q berdasarkan kondisi X (posteriori probabilitas) 0

$P(Q)$: Probabilitas hipotesis Q

$P(X|Q)$: Probabilitas X terhadap hipotesis Q

$P(X)$: Probabilitas X

Algoritma Decision Tree merupakan salah satu algoritma yang dapat digunakan untuk mengklasifikasikan data berdasarkan atribut numerik dan kategorik. Hasil dari proses klasifikasi dapat digunakan dalam bentuk aturan untuk memprediksi nilai atribut diskrit dari record yang baru [8]. Langkah-langkah dalam pengolahan data menggunakan algoritma antara lain mencari nilai entropi, mencari nilai gain, membentuk *decision tree* dan aturan. Rumus untuk menghitung entropi dan gain masing-masing dapat dilihat pada persamaan berikut.

$$\text{Entropy}(S) = \sum_{i=0}^n -p_i * \log_2 p_i$$

Keterangan:

S : Himpunan Kasus

n : Jumlah Partisi S

p_i : Proporsi dari S_i terhadap S

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

Keterangan:

S : Himpunan Kasus

n : Jumlah Partisi Atribut A

$|S|$: Jumlah Kasus Dalam S

$|S_i|$: Jumlah Kasus Pada Partisi ke i

IV. HASIL DAN PEMBAHASAN

Untuk menentukan analisis perbandingan tingkat akurasi dalam memprediksi kemungkinan kehamilan pada penderita diabetes, maka algoritma yang digunakan pada penelitian ini ada dua algoritma, yaitu algoritma Naïve Bayes dan algoritma Decision Tree untuk mengaplikasikan data yang telah mengalami *preprocessing data* atau pembersihan data pada aplikasi Orange. Aplikasi Orange atau biasa dikenal dengan Orange Data Mining merupakan suatu *software open source* yang digunakan untuk melakukan

data analisis. Hasil prediksi dari kedua algoritma tersebut akan dibandingkan berdasarkan tingkat keakuratan hasil prediksinya melalui *confusion matrix* yaitu membandingkan berdasarkan nilai akurasi, presisi dan *recall*.

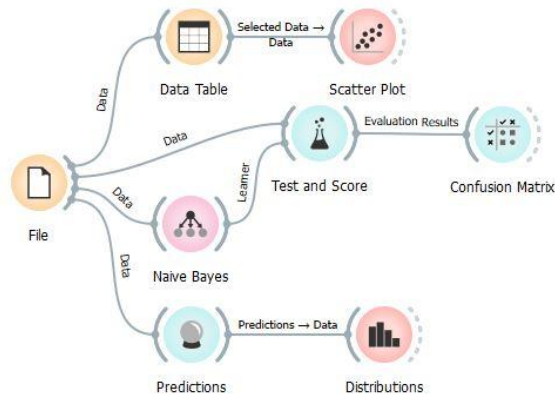
Pada saat melakukan analisis diperlukan beberapa atribut yang akan digunakan, yaitu: Glukosa, Tekanan Darah, Ketebalan Kulit, Insulin, BMI (Body Massa Index), Genetik, Usia, dan Hasil. Terdapat 25 sampel data yang digunakan sebagai berikut:

Tabel 1.

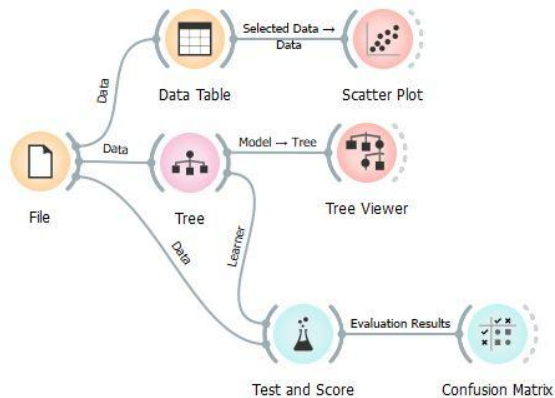
Data Sampel Prediksi Kehamilan Pada Penderita Diabetes

Glukosa	Tekanan Darah	Ketebalan Kulit	Insulin	BMI	Gen	Usia	Hasil
148	72	35	0	33.6	0.627	50	1
85	66	29	0	26.6	0.351	31	0
183	64	0	0	23.3	0.672	32	1
89	66	23	94	28.1	0.167	21	0
137	40	35	168	43.1	0.288	33	1
116	74	0	0	25.6	0.201	30	0
78	50	32	88	31	0.248	26	1
115	0	0	0	35.3	0.134	29	0
197	70	45	543	30.5	0.158	53	1
125	96	0	0	0	0.232	54	1
110	92	0	0	37.6	0.191	30	0
168	74	0	0	38	0.537	34	1
139	80	0	0	27.1	0.441	57	0
189	60	23	846	30.1	0.398	59	1
166	72	19	175	25.8	0.587	51	1
100	0	0	0	30	0.484	32	1
118	84	47	230	45.8	0.551	31	1
107	74	0	0	29.6	0.254	31	1
103	30	38	83	43.3	0.183	33	0
115	70	30	96	34.6	0.529	32	1
126	88	41	235	39.3	0.304	27	0
99	84	0	0	35.4	0.388	50	0
196	90	0	0	39.8	0.451	41	1
119	80	35	0	29	0.263	29	1
143	94	33	146	36.6	0.254	51	1

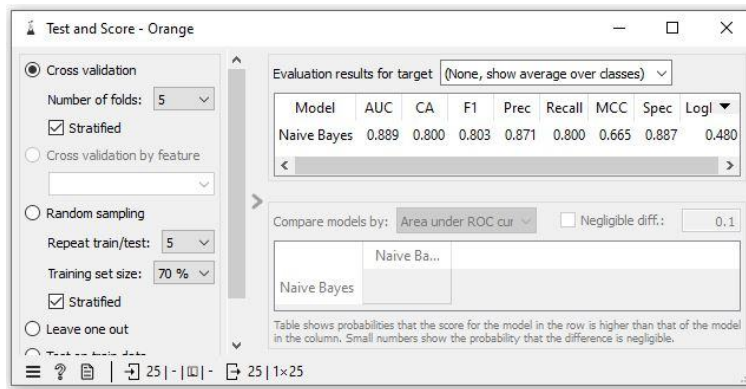
Gambar 2 adalah model algoritma Naive Bayes dan algoritma Decision Tree yang dibuat di aplikasi Orange. Berikut didalamnya sudah dimasukkan data sampel dengan atribut yang digunakan yaitu: Glukosa, Tekanan Darah, Ketebalan Kulit, Insulin, BMI (Body Massa Index), Genetik, Usia, dan Hasil.



Gambar 2.
Model Naive Bayes

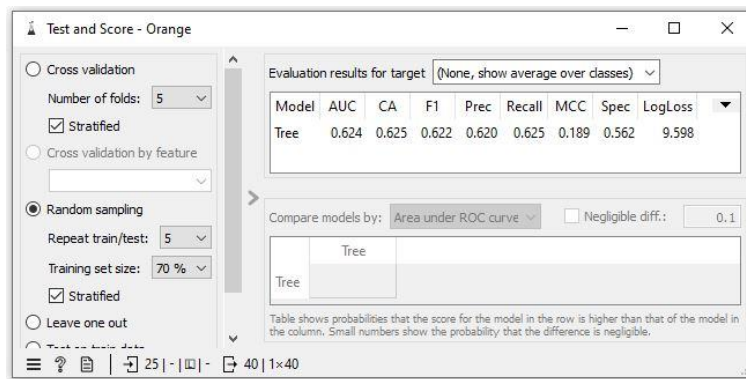


Gambar 3.
Model Decision Tree



Gambar 4.
 Test and Score Naïve Bayes

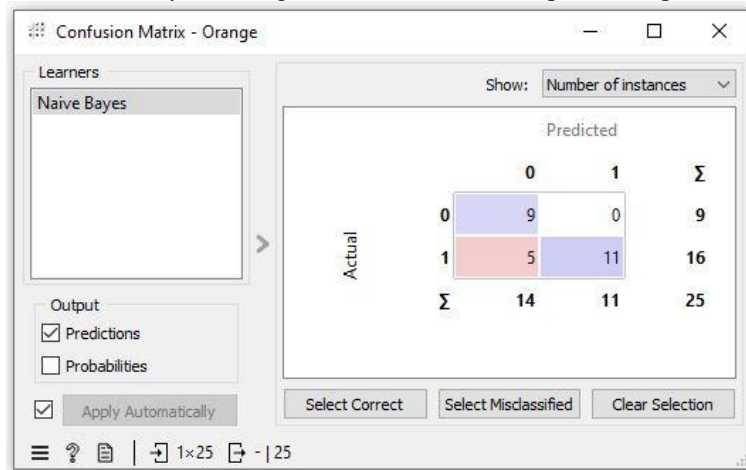
Dari gambar diatas, didapatkan *Test and Score* dari algoritma Naïve Bayes, akurasi = 0.800, presisi = 0.871, recall = 0.800, dan F1-Score 0.083.



Gambar 5.
 Test and Score Decision Tree

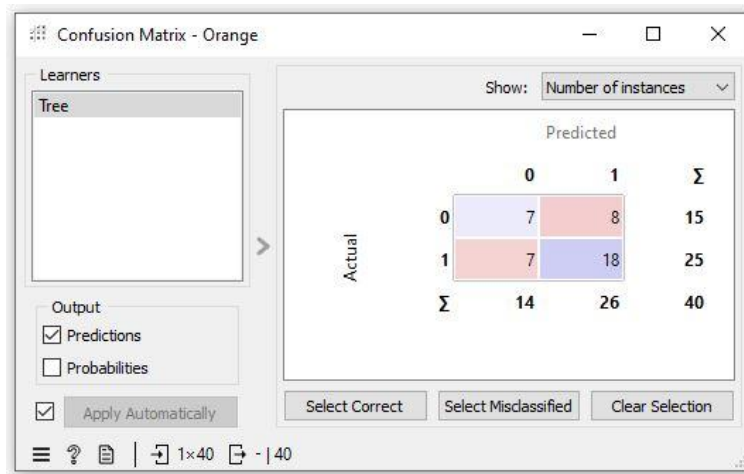
Dari gambar diatas, didapatkan *Test and Score* dari algoritma Decision Tree, akurasi = 0.625, presisi = 0.620, recall = 0.625, dan F1-Score 0.622.

Confusion matrix dari algoritma Naïve Bayes dan algoritma Decision Tree dapat dilihat pada Gambar 6 dan Gambar 7.



Gambar 6.
 Confusion Matrix Algoritma Naïve Bayes

- Banyaknya data positif yang benar : 9
- Banyaknya data negative yang benar : 11
- Banyaknya data negative yang salah : 5
- Banyaknya data positif yang salah : 0



Gambar 7.
 Confusion Matrix Algoritma Decision Tree

Banyaknya data positif yang benar : 7
 Banyaknya data negative yang benar : 18
 Banyaknya data negative yang salah : 7
 Banyaknya data positif yang salah : 8

V. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa Algoritma Naive Bayes menunjukkan performa yang lebih unggul, dengan tingkat akurasi mencapai 80%, serta nilai presisi sebesar 87%, recall, dan F1-Score masing-masing sebesar 80%. Sementara itu, Algoritma Decision Tree menghasilkan tingkat akurasi yang lebih rendah, hanya sebesar 62%, dengan nilai presisi, recall, dan F1-Score sebesar 62% dalam memprediksi kehamilan pada penderita diabetes.

Pengujian yang dilakukan menggunakan aplikasi Orange juga memberikan hasil yang konsisten, dengan akurasi Algoritma Naive Bayes sebesar 80% dan akurasi Algoritma Decision Tree sebesar 62%. Rincian hasil dapat dilihat pada tabel di bawah ini.

Tabel 2.
 Perbandingan Kinerja

Model	Accuracy	Precision	Recall Score	F1-Score
Naive Bayes	0.800	0.871	0.800	0.803
Decision Tree	0.625	0.620	0.625	0.622

DAFTAR PUSTAKA

- [1] Sudjana, Nana. 2016. Penilaian Hasil Proses Belajar Mengajar. Bandung: Pt Remaja Rosdakarya.
- [2] Basah, Sjachran, 1994. Hukum Tata Negara Perbandingan, Bina Aksara, Jakarta
- [3] Olson & Delen. Advanced Data Mining Techniques. USA: Springer-Verlag Berlin Heidelberg. 2008
- [4] Herdianto,(2013),Prediksi Kerusakan Motor Induksi Menggunakan Metode Jaringan Saraf Tiruan Backpropagation,Tesis,Universitas Sumatera Utara, Medan.
- [5] Pratiwi M. A dan Fatimah. 2019. Patologi Kehamilan Memahami Berbagai Penyakit & Komplikasi Kehamilan. Yogyakarta. Pustaka Baru Press
- [6] Smeltzer, S.C. & Bare, B.G. (2013). Buku Ajar Keperawatan Medikal Bedah
- [7] Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and Fscore, with Implication for Evaluation. European Conference on Information
- [8] Elisa, E. (2017). Analisa dan Penerapan Algoritma C4.5 DSalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT. Arupadhatu Adisesanti