

PERBANDINGAN METODE SUPERVISED LEARNING UNTUK PREDIKSI DIABETES GESTASIONAL

Andrew C Handoko*¹⁾, Hendry²⁾

1. Universitas Kristen Satya Wacana, Salatiga, Indonesia
2. Universitas Kristen Satya Wacana, Salatiga, Indonesia

Article Info

Kata Kunci: Diabetes; Diabetes Gestasional; Orange; Prediksi; *Supervised Learning*

Keywords: *Diabetic; Gestasional Diabetic; Orange; Prediction; Supervised Learning*

Article history:

Received 14 May 2023

Revised 28 May 2023

Accepted 11 June 2023

Available online 1 December 2023

DOI :

<https://doi.org/10.29100/jipi.v8i4.4166>

* Corresponding author.

Andrew C Handoko

E-mail address:

andrewhandoko925@gmail.com

ABSTRAK

Dalam penelitian ini dilakukan analisa terhadap metode *Supervised Learning* dengan membandingkan hasil prediksi dari tiap metode, guna mendapatkan algoritma terbaik, yang dapat dikembangkan kedepannya sebagai salah satu media untuk mempermudah deteksi Diabetes Gestasional. Prediksi dilakukan terhadap *Dataset* Diabetes Gestasional yang di dapat dari *Kaggle*, dengan judul “Diabetes Dataset” yang berasal dari *National Institute of Diabets and Digestive and Kidney Diseases*. Dimana analisis akan menggunakan bantuan *Software Orange*, sebagai tempat untuk melakukan pengolahan data dan melihat nilai hasil prediksi dari masing-masing algoritma yang ada di metode *Supervised Learning*. Algoritma yang dibandingkan ada tujuh, dengan nilai *Recall* sebagai penentu no satu algoritma yang dianggap bagus untuk melakukan prediksi, diikuti dengan nilai *Akurasi, Precision, Test Time* dan *Train Time*. Dan dengan bantuan *Orange*, maka di dapat algoritma yang paling bagus adalah *Logistic Regression*.

ABSTRACT

In this study, an analysis of the Supervised Learning method was carried out by comparing the prediction results of each method, in order to get the best algorithm, which can be developed in the future as one of the media to facilitate the detection of Gestational Diabetes. Predictions were made on the Gestational Diabetes Dataset obtained from Kaggle, entitled “Diabetes Dataset” which comes from the National Institute of Diabets and Digestive and Kidney Diseases. Where the analysis will use the help of Software Orange, as a place to do data processing and see the predicted value of each algorithm in the Supervised Learning method. There are seven algorithms compared, with the Recall value as the number one determinant of an algorithm that is considered good for making predictions, followed by the values for Accuracy, Precision, Test Time and Train Time. And with Orange's help, the best algorithm is Logistic Regression.

I. PENDAHULUAN

DIABETES merupakan salah satu penyakit yang menjadi ancaman serius terhadap kesehatan global di antara populasi manusia yang semakin menua ini. Dan menurut WHO (Organisasi Kesehatan Dunia) di tahun 2014 menunjukkan bahwa prevalensi Diabetes, terutama tipe 2 terus meningkat dalam tiga dekade terakhir, terutama pada Negara dengan penghasilan rendah dan menengah. Sedangkan untuk Indonesia sendiri di tahun 2015 menurut WHO bahwa Indonesia menempati peringkat ke tujuh dunia dimana prevalensi Diabetes tertinggi dunia dimana estimasi penderita Diabetes sebesar 10 juta orang dari total penderita Diabetes di dunia sebesar 415 juta orang. Dimana pada tahun 2017 juga menurut IDF (International Diabetes Federation) Atlas menunjukkan bahwa Diabetes di Indonesia cenderung meningkat. Indonesia menempati posisi ke enam di dunia dengan jumlah penderita diabetes di usia 20 – 79 tahun sekitar 10,3 juta orang. Dan pada diabetes gestasional, hampir 80% penderitanya berada di Negara dengan penghasilan rendah dan menengah. Di Negara Eropa diabetes gestasional ini berkembang sebesar 5,4%, di negara Afrika sebesar 14%, di negara-negara Asia sebesar 1% - 20%, sedangkan di Indonesia sendiri prevalensi diabetes gestasional sebesar 1,9% - 3,6%. Penyakit diabetes yang mana sering disebut sebagai penyakit gula ini merupakan penyakit berbahaya yang mana dapat menyebabkan kematian akibat dari komplikasi yang timbul dari penyakit ini. Diabetes sendiri merupakan penyakit metabolik yang mana tubuh penderita tidak mampu mencukupi kebutuhan insulin secara efektif sehingga gula di dalam darah menjadi berlebih. Dan Diabetes pada ibu hamil tiap tahunnya meningkat, dimana diabetes ini disebut Diabetes Gestasional. Dimana kemudahan dalam skrining atau deteksi dini terkait Diabetes Gestasional sangatlah penting untuk mengurangi risiko atau

komplikasi yang mungkin terjadi pada ibu maupun bayinya nanti. Sehingga tindakan pencegahan maupun penanganan yang harus dilakukan dapat segera diambil agar dapat melindungi perempuan atau ibu hamil serta mengurangi angka kejadian diabetes gestasional. Dan dapat menurunkan morbiditas untuk ibu dan anak. Dimana Diabetes Gestasional ini memberikan dampak buruk kepada ibu maupun janin, dimana meningkatkan risiko ibu mengalami hipertensi pada kehamilan hingga pre-eklamsia. Dan sekitar 50-70% ibu hamil penderita diabetes gestasional juga berisiko menderita diabetes tipe 2 dalam kurun waktu 5-10 tahun setelah melahirkan dan anak-anak yang lahir dari ibu penderita diabetes berisiko mengalami diabetes tipe 2 pada usia dewasa. Sehingga Diabetes Gestasional ini perlu mendapat perhatian khusus dan serius dalam pengobatan demi melindungi perempuan maupun ibu hamil serta anaknya. Maka dari itu, prediksi diperlukan guna melakukan tindakan pencegahan pada wanita yang diprediksi akan mengalami Diabetes Gestasional. Prediksi ini akan menggunakan algoritma yang ada pada metode Supervised Learning, dimana tentunya kita perlu menentukan algoritma mana yang paling baik untuk memprediksi Diabetes Gestasional. Karena pada bidang kesehatan, kesalahan prediksi kecil akan sangat fatal. Sehingga pada penelitian ini akan dilakukan analisis mengenai hasil evaluasi dari tiap algoritma yang ada, yang bertujuan untuk memilih algoritma mana yang paling baik untuk melakukan prediksi. Dimana pada penelitian ini, data yang digunakan berasal dari website Kaggle, dengan judul "Diabetes Dataset", yang didapat dari National Institute of Diabetes and Digestive and Kidney Diseases. Dan akan menggunakan aplikasi Orange dengan metode Regresi Logistik untuk melihat pengaruhnya.

Dan berikut beberapa penelitian sebagai referensi dalam penelitian ini, [1] yang pertama yaitu penelitian oleh Raras & Adhi Yoga, yang membahas tentang rancangan sistem pendeteksi untuk penyakit diabetes terutama diabetes retinopati. Dalam perancangannya digunakan pendekatan machine learning dengan metode regresi logistik. Dan dalam pelatihan data pada model sistem deteksi yang dirancang ada empat macam kondisi yaitu dengan parameter bawaan, standarisasi atribut, pemilihan atribut dan pengaturan parameter. Dan hasil yang didapatkan menunjukkan model mempunyai akurasi yang cukup bagus dengan persentase sebesar 80,17%.

[2] Lalu pada penelitian yang dilakukan oleh Yolanda dan rekan-rekannya, membahas tentang penyakit Diabetes Gestasional yang sulit didiagnosis karena dokter perlu mempertimbangkan beberapa faktor yang mempengaruhi risiko terjadinya penyakit ini. Dan diagnosis juga bergantung pada interpretasi dokter, sehingga rentan terjadinya human error. Maka salah satu solusi yang dapat diterapkan adalah dengan menggunakan algoritma klasifikasi untuk mengidentifikasi keberadaan penyakit ini. Dan hasil sensitivity terbaik yang didapatkan adalah 0,8125, specificity bernilai 0,8788 dan F1 score bernilai 0,7879 pada $K = 25$ dan $E = 2$. Sedangkan pengujian dengan algoritme KNearest Neighbor (KNN). Hasil terbaik diperoleh dari pengujian dengan nilai fold = 4, yaitu sensitivity yaitu 0,6043, specificity yaitu 0,8703, dan F1 score yaitu 0,6383.

[3] Di penelitian selanjutnya membahas tentang analisis untuk menentukan algoritma prediksi yang paling baik. Sehingga dapat meminimalisir kesalahan prediksi yang terjadi. Dalam analisa yang dilakukan menggunakan teknik Supervised Learning ini ada beberapa metode atau algoritma seperti kNN, SVM, Naïve Bayes dan lainnya. Dimana algoritma terbaik untuk melakukan prediksi pada dataset ILPD adalah algoritma Regresi Logistik.

[4] Selain itu, berikut penelitian yang membahas tentang analisa yang dilakukan untuk melihat algoritma Regresi Logistik atau Neural Network yang paling baik untuk melakukan prediksi terhadap ketepatan waktu lulus mahasiswa. Pada penelitian ini didapatkan hasil bahwa Neural Network memiliki akurasi yang lebih baik, yaitu sebesar 69%. [5] Lalu yang terakhir, yaitu penelitian yang membahas mengenai analisa algoritma prediksi. Dimana algoritma yang dianalisa yaitu Neural Network, Random Forest, Linear Regression, SVM Gaussian Process dan Polynomial Regression. Dimana hasil yang di dapatkan yaitu algoritma Linear Regression dengan akurasi prediksi RMSE sebesar 1625.280 +/- 414.224 (micro average : 1671.084 +/- 0.0000), yang merupakan paling kecil dibanding algoritma lainnya, sehingga merupakan algoritma terbaik untuk membantu prediksi total kasus COVID-19 di Indonesia.

Lalu Diabetes Melitus menurut American Diabetes Association (ADA) tahun 2010, merupakan suatu kelompok penyakit metabolik dengan karakteristik hiperglikemia yang terjadi karena kelainan sekresi insulin, kerja insulin, atau kedua-duanya. Diabetes gestasioanl atau disebut sebagai Gestasional Diabetes Mellitus (GDM) adalah diabetes yang terjadi pada saat 10 kehamilan ditandai dengan kadar gula darah yang naik biasanya pada usia 24 minggu masa kehamilan. Dan dalam pemeriksaan untuk mengetahui apakah ibu hamil mengalami GDM dilakukan dengan mengukur tingkat toleransi gula darah plasma setelah 2 jam berpuasa. GDM ini dapat terjadi pada perempuan manapun, yang risikonya akan berbeda-beda tergantung dengan gaya hidup di tiap daerah. Serta perempuan dengan usia lebih dari 35 tahun memiliki risiko lebih tinggi daripada yang hamil di usia lebih muda. Perempuan dengan usia 35 tahun lebih cenderung memiliki kadar gula darah tinggi karena kadar insulin yang diproduksi tubuh makin berkurang. Lalu ada faktor berat badan, yang mana jika Indeks Massa Tubuh masuk ke dalam obesitas, maka insulin sulit bekerja secara maksimal karena kadar lemak darah yang tinggi terutama kolestrol. Serta beberapa faktor lainnya seperti riwayat keturunan keluarga apakah mempunyai riwayat diabetes, faktor dari tekanan darah dan lainnya.

Dan komplikasi yang dapat terjadi pada ibu hamil penderita GDM ada beberapa seperti gangguan penglihatan, pre-eklamsia atau sindrom dengan tanda hipertensi, ukuran janin yang besar, keguguran, persalinan yang lama, premature serta persalinan sectio caesarea (SC). Serta komplikasi yang akan terjadi setelah bersalin pada ibu hamil yaitu adanya risiko infeksi kandung kemih, memperberat komplikasi diabetes yang sudah ada sebelumnya (gangguan pada organ lain seperti jantung, ginjal, saraf dan lainnya), lalu risiko menderita diabetes mellitus tipe 2 dalam jangka waktu 5- 10 tahun setelah melahirkan. Komplikasi pada bayi juga dapat terjadi seperti bayi kuning (Ikterus Neonatorum), sindrom gangguan pernafasan bayi, hiperglikemia akut, peningkatan risiko obesitas dan diabetes pada saat anak-anak dan remaja, risiko mengalami diabetes mellitus tipe 2 pada saat usia dewasa dan berat bayi yang baru lahir bisa lebih dari 4000 gram. Berdasarkan penjabaran diatas, serta penelitian yang sudah ada, yang berkaitan dengan prediksi dan Diabetes Gestasional, maka akan dilakukan penelitian yang akan membahas dan menganalisa algoritma mana yang paling baik pada metode Supervised Learning untuk melakukan prediksi terhadap Diabetes Gestasional.

II. METODE PENELITIAN

A. Pengumpulan Data

Data yang digunakan pada penelitian ini adalah data sekunder yang didapatkan melalui repository dataset yang bersumber dari website kaggle.com. Kaggle merupakan tempat perkumpulan data scientist dimana tersedia banyak dataset yang dapat diambil dan tersedia banyak perlombaan yang berkaitan dengan Machine Learning. Dataset yang digunakan berjudul "Diabetes Dataset" yang dimana dataset tersebut didapat dari National Institute of Diabetes and Digestive and Kidney Diseases yang dipublikasikan di website Kaggle. Dataset yang kami gunakan memuat data dari semua wanita dengan umur minimal dua puluh satu tahun yang berasal dari "Pima Indian Heritage". Dengan sembilan atribut yang digunakan serta tujuh ratus enam puluh delapan instance. Dimana ada delapan atribut yang akan digunakan sebagai variabel independen, yaitu :

1. Pregnancies : berapa kali mengalami kehamilan
2. Glucose : konsentrasi plasma glucose selama dua jam yang didapat dari tes oral terhadap toleransi glukosa
3. Blood Pressure : tekanan darah diastolic (mmHg)
4. Skin Thickness : ketebalan dari lipatan kulit pada bagian trisep (mm)
5. Insulin : Serum insulin yang diberikan selama dua jam (mu U/ml)
6. BMI : Body Mass Index atau Indeks Massa Tubuh (bobot dalam kg/(tinggi dalam m)²)
7. Diabetes Pedigree Function : riwayat keturunan apakah penderita diabetes atau tidak
8. Age : umur (tahun) Dan satu atribut sisanya sebagai variabel dependen yaitu Outcome (target/output) dengan nilai numerik 0 atau 1. Dimana jika bernilai 1 maka hasilnya positif menderita diabetes.

B. Metode Supervised Learning Dan Software Orange

Metode Supervised Learning merupakan salah satu metode yang terdapat dalam Machine Learning, yang digunakan untuk mengklasifikasikan data atau melakukan prediksi hasil secara akurat. Data yang digunakan merupakan data yang sudah diberikan label atau yang kelasnya sudah diketahui. Dimana nantinya mesin akan diajari untuk melihat pola dari data yang diinput serta label dari output yang ada, dan juga mesin akan diajari untuk melihat keterkaitan dari data yang diinput dengan output yang dihasilkan. [3] Dimana dalam metode ini, memiliki beberapa algoritma di dalamnya, dan algoritma yang akan digunakan untuk melakukan analisis perbandingan adalah Decision Tree, Random Forest, SVM (Support Vector Machine), Neural Network, Naïve Bayes, kNN dan Regresi Logistik Berikut adalah penjelasan singkat mengenai ketujuh algoritma yang akan digunakan nantinya :

1. Decision Tree
Algoritma ini menggunakan pendekatan berbasis aturan atau rule berbentuk If-Then. Untuk mendapatkan suatu keputusan atau hasil akhir, diperlukan sebuah pohon yang terdiri dari node-node yang saling terhubung. Tiap node terhubung oleh cabang ke node lain ataupun node akhir untuk mendapatkan suatu keputusan.
2. Random Forest
Algoritma ini merupakan pengembangan dari Decision Tree. Dimana pada algoritma ini terdiri dari beberapa pohon yang saling terhubung yang nantinya akan menghasilkan sebuah klasifikasi. Dimana dari hasil tersebut akan terlihat kecenderungan dari sebuah data berkelompok itu.
3. SVM
SVM atau Support Vector Machine merupakan salah satu algoritma yang dapat mengklasifikasikan data berbentuk linier dan non linier. SVM bertujuan untuk menemukan fungsi berbentuk klasifikasi atau regresi paling baik, yang terwujud secara geometris. Dimana nantinya fungsi akan dipetakan dengan pemberian label dari data training yang dilakukan.
4. Neural Network

Algoritma ini cukup bagus digunakan untuk data dengan ketidakpastian yang tinggi. Tetapi waktu yang dibutuhkan untuk menentukan suatu kecocokan yang baik, tentu akan sangat lama. Dalam algoritma ini ada lapisan input dan lapisan output, yang mana antar kedua lapisan ini, bisa jadi memiliki satu atau lebih lapisan lainnya yang tersembunyi.

5. Naïve Bayes

Algoritma ini merupakan algoritma klasifikasi secara statistik. Dimana dasarnya adalah ilmu probabilitik sederhana yang memiliki dasar dari Teorema Bayes atau peluang bersyarat.

6. kNN

K-Nearest Neighbor atau kNN merupakan algoritma untuk melakukan klasifikasi. Dimana data dikategorikan berdasar kedekatan lokasi dengan data lainnya atau dengan data/neighbor terdekat. Dalam menghitung jarak, bisa digunakan rumus dari jarak euclidian (Euclidian Distance).

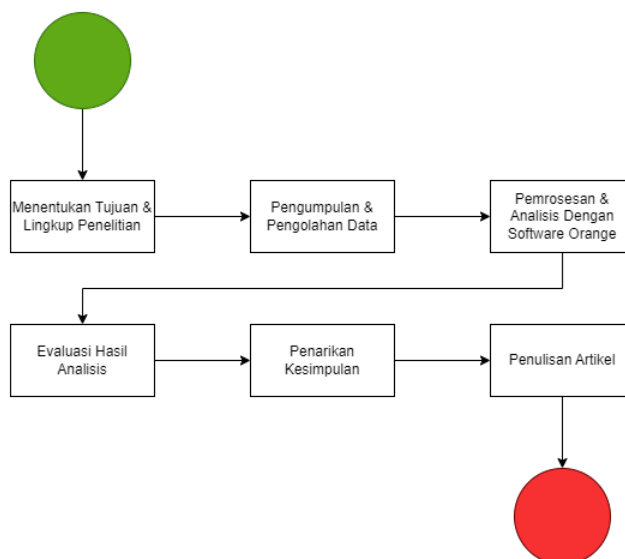
7. Regresi Logistik

Regresi termasuk dalam teknik supervised learning, dimana metode ini akan membandingkan dua atau lebih variabel untuk mengetahui bagaimana pengaruhnya satu sama lain, sehingga akan berdampak pada keputusan nantinya. Variabel yang dibandingkan adalah variabel independent dan variabel dependent. Dalam metode regresi sendiri terbagi menjadi beberapa sub-metode lagi, salah satunya adalah regresi logistik. Menurut Hosmer & Lemeshow, regresi logistik adalah metode analisis statistika yang digunakan untuk mendeskripsikan hubungan antara variabel respon dengan variabel prediktor. Dimana variabel respon (dependent) berbentuk data dikotonomi seperti 1 atau 0, ya atau tidak, hidup atau mati, sakit atau sehat dan lainnya. Dan variabel prediktor (independent) berupa data kategorik maupun numerik. Sehingga, jika variabel responnya bernilai 1 maka menyatakan bahwa variabel respon sesuai dengan kriteria dari hasil pembelajaran menggunakan variabel prediktor, dan berlaku sebaliknya, dimana jika variabel bernilai 0 maka tidak sesuai dengan kriterianya.

Orange merupakan salah satu aplikasi Data Mining yang dapat digunakan untuk melakukan analisis dan visualisasi data. Aplikasi bersifat Open Source, dan fitur-fitur yang disediakan cukup lengkap dan sangat memudahkan penggunaannya. Dimana Widget yang disediakan cukup mudah digunakan dan hasil dari analisis juga sudah tertera dengan jelas. Sehingga pengguna dapat lebih mudah memahami hasil dari data yang dianalisis. Juga baik Supervised Learning ataupun Unsupervised Learning sudah tersedia di dalam aplikasi ini.

C. Tahapan Penelitian

Dalam melakukan penelitian ini, yang perlu dipersiapkan terlebih dahulu adalah menentukan latar belakang serta tujuan apa yang diinginkan nantinya. Dan batasan serta ruang lingkup dalam penelitian juga harus didefinisikan. Tinjauan pustaka juga harus ditulis, karena dapat membantu meningkatkan pemahaman terhadap komponen-kom-



Gambar. 1. Flow Penelitian

ponen yang terlibat dalam penelitian, yaitu mengenai diabetes, machine learning, regresi serta regresi logistik.

Tahapan selanjutnya adalah langkah penelitian yaitu menyiapkan data yang akan digunakan, dimana penelitian ini menggunakan data dari website kaggle. Lalu data yang didapatkan harus dilakukan pembagian 16 variabel menjadi variabel independen dan variabel dependen. Lalu data juga harus dibagi menjadi dua, yaitu yang akan digunakan sebagai data training dan satu lagi untuk data testing.

Lalu langkah ketiga adalah menggunakan software Orange untuk melakukan analisis algoritma yang paling baik untuk melakukan prediksi Diabetes Gestasional. Dimana dengan Orange, akan dibuat sebuah Workflow yang menunjukkan alur proses prediksi dari Dataset yang ada hingga ke masing-masing algoritma. Dari Dataset yang ada, akan digunakan fungsi Select Column untuk memilih kolom yang akan digunakan. Setelah itu akan disambungkan ke fungsi yang memuat masing-masing algoritma serta fungsi untuk melakukan test dan melihat skor hasil prediksi.

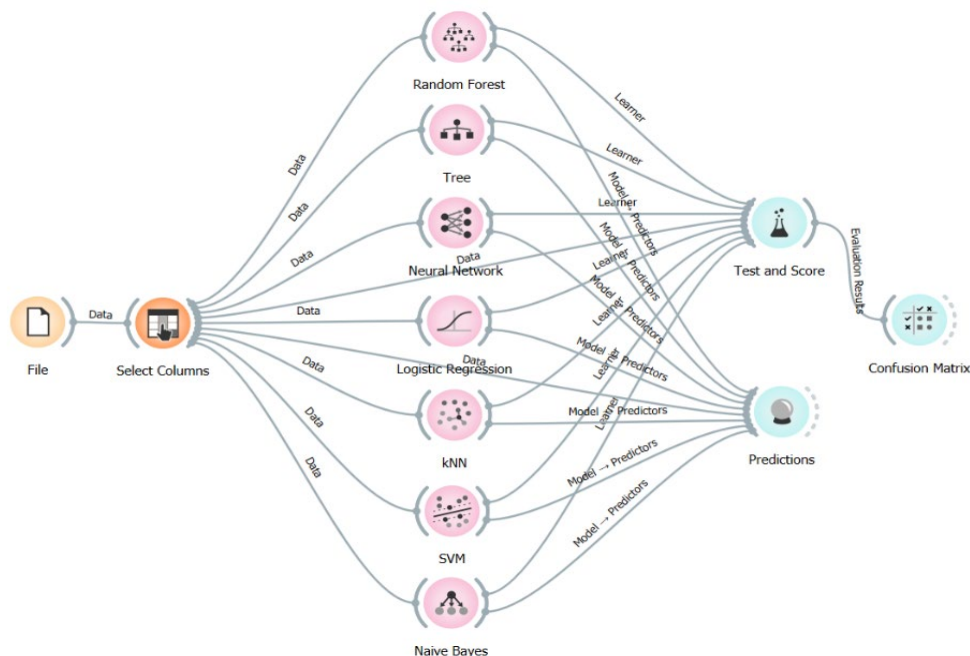
Dan langkah terakhir adalah melakukan analisa atau evaluasi terhadap hasil implementasi masing-masing algoritma terhadap dataset, agar mengetahui mana algoritma yang paling baik. Dalam mengevaluasi digunakan confusion matrix. Lalu akan didapatkan nilai Akurasi, Precision, Recall, serta Train dan Test Time. Setelah itu, baru melakukan penarikan kesimpulan dan memberikan saran dari hasil penelitian

III. HASIL DAN PEMBAHASAN

Untuk data yang digunakan berbentuk CSV dengan 9 kolom dan 768 baris. Dimana terbagi menjadi 8 kolom sebagai variabel pengaruh dan 1 kolom sebagai variabel hasil. Pada gambar 2 diatas, dapat kita lihat 20 data yang

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0

Gambar. 2. Dataset

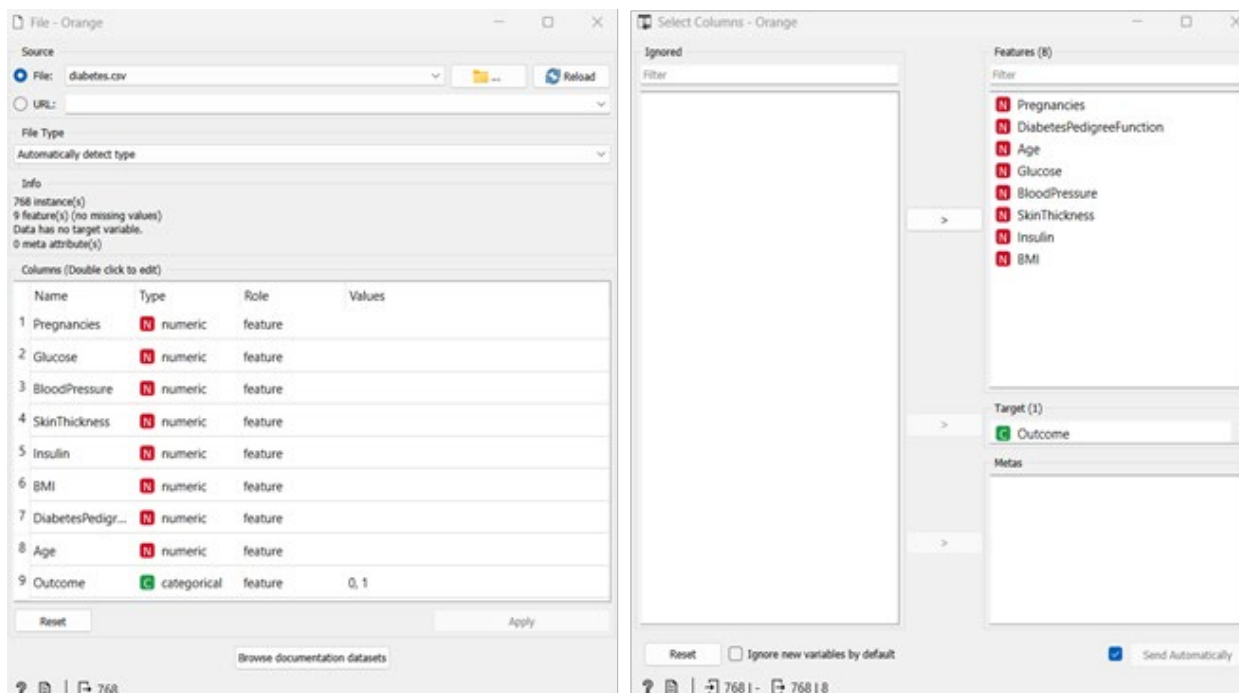


Gambar. 3. Workflow

disajikan sebagai contoh. Lalu dengan data tersebut, akan diproses di Software Orange. Dalam melakukan pemrosesan data, maka pembuatan Workflow diperlukan. Gambar 3 merupakan Workflow yang akan digunakan untuk analisis pada penelitian ini. Dari Workflow tersebut dapat dilihat ada tiga bagian yang menyusunnya, yaitu pada bagian berwarna kuning sebagai bagian untuk persiapan data, baik untuk import data, menentukan atribut bebas dan target, maupun untuk melihat data dalam bentuk spreadsheet. Lalu pada bagian berwarna merah muda adalah

algoritma yang akan 18 digunakan, dimana pada penelitian ini menggunakan tujuh algoritma yang ada pada metode Supervised Learning. Terakhir pada bagian berwarna hijau merupakan bagian untuk melakukan evaluasi dan memperoleh hasil prediksi dari masing-masing algoritma.

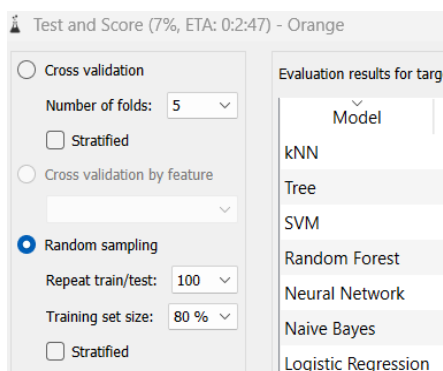
Lalu dataset yang sudah berbentuk CSV akan dimasukkan kedalam Orange dengan widget file. Dan informasi mengenai datanya, dapat kita lihat pada gambar 4. Langkah selanjutnya adalah menentukan atribut target, menggunakan widget select columns. Pada widget ini akan kita pilih kolom Outcome sebagai atribut targetnya. Dan sisa atribut lainnya dimasukkan kedalam features yang dapat dilihat pada gambar 3 sebelah kanan. Setelah itu



Gambar. 4. Informasi Dataset & Widget File

kita tarik garis agar dari widget select columns terhubung dengan widget test and score untuk melakukan training dan testing dari data yang ada dengan algoritma yang terhubung dengan widget tersebut. Pada widget test and score kita lakukan pembagian data untuk training dan testing dengan perbandingan sejumlah 80%:20%. Untuk pembagian datanya dilakukan secara random sampling dengan 100 kali pengulangan.

Untuk Hasil prediksi juga dapat dilihat pada gambar 4, dimana diambil 20 data paling atas saja. Dan untuk melihat bagaimana nilai masing-masing algoritma dalam melakukan predikis, maka langkah selanjutnya adalah melakukan evaluasi. Dimana dari hasil evaluasi menggunakan widget confusion matrix terhadap ketujuh algoritma tersebut dapat kita lihat pada gambar 5. Berdasarkan hasil dari confusion matrix ini, maka akan dilakukan perhi-



Gambar. 5. Pembagian Data

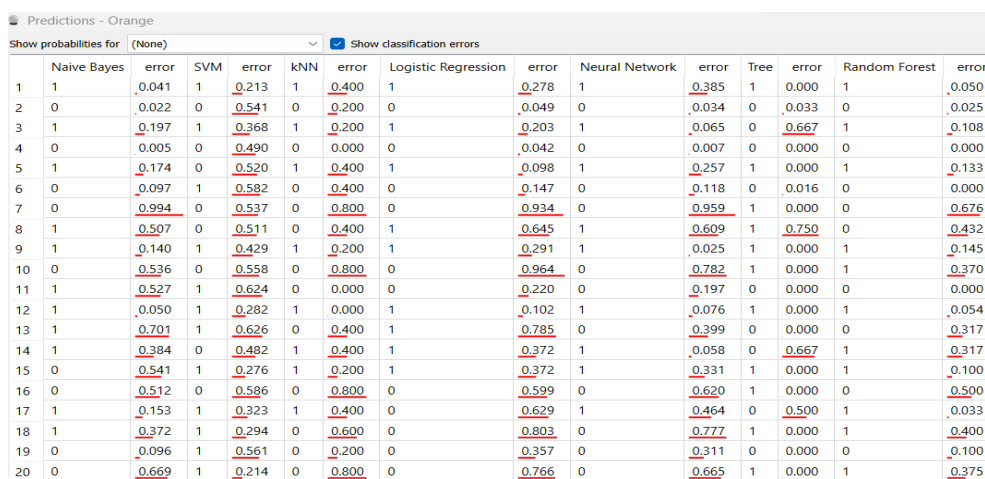
tungan untuk mendapatkan nilai Akurasi, Precision dan Recall. Nilai Akurasi kita dapatkan melalui perhitungan dengan rumus $(TP + TN) / (TP + FP + FN + TN)$. Lalu untuk nilai Precision dan Recall kita dapatkan secara otomatis dengan fungsi widget di Orange yaitu test and score.

Untuk akurasi dapat dilihat pada tabel 1. Dari perhitungan yang didapat, posisi pertama ditempati oleh Logistic Regression sebagai algoritma terbaik, yang diikuti oleh Neural Network dan selanjutnya adalah Naïve Bayes. Sedangkan algoritma dengan akurasi terburuk adalah SVM.

TABEL. 1.
Pembagian Data

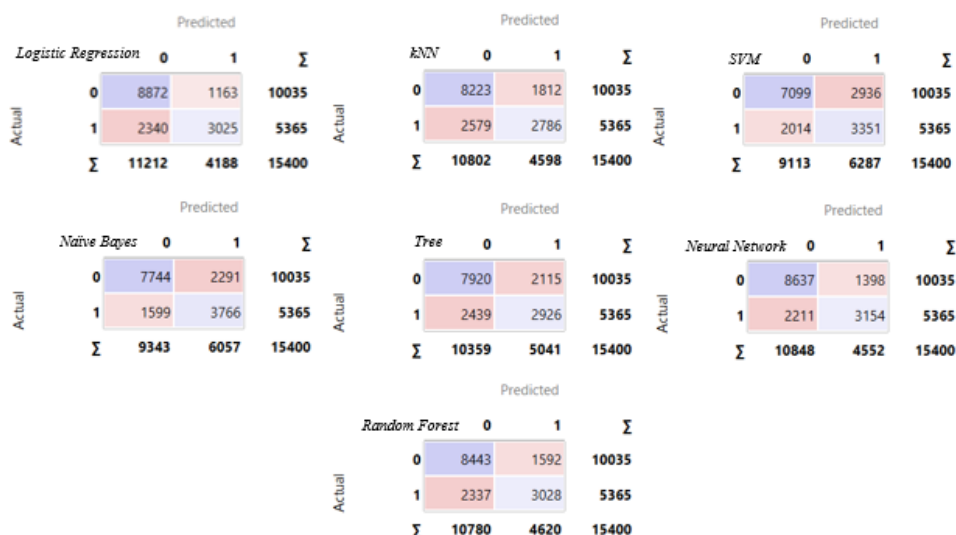
	Logistic Regression	kNN	SVM	Naïve Bayes	Tree	Neural Network	Random Forest
TP	3025	2786	3351	3766	2926	3154	3028
TN	8872	8223	7099	7744	7920	8637	8443
FP	1163	1812	2936	2291	2115	1298	1592
FN	2340	2579	2014	1599	2439	2211	2337
Akurasi	0,772532468	0,71487	0,678571	0,747403	0,704286	0,770654	0,74487

Berikut merupakan hasil evaluasi yang didapatkan untuk nilai Precision. Dimana bisa dilihat bahwa algoritma Logistic Regression memiliki nilai paling tinggi, yaitu 0,767. Lalu diikuti Neural Network diposisi kedua dan Naïve Bayes diposisi ketiga. Yang dapat kita simpulkan bahwa algoritma Logistic Regression merupakan algoritma terbaik untuk melihat hasil prediksi positif, atau perbandingan wanita yang benar-benar terkena Diabetes Gestasional



Sample	Naive Bayes	SVM	kNN	Logistic Regression	Neural Network	Tree	Random Forest
1	1	1	1	1	1	1	1
2	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1
4	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1
6	0	1	0	0	0	0	0
7	0	0	0	0	0	1	0
8	1	0	0	1	1	1	0
9	1	1	1	1	1	1	1
10	0	0	0	0	0	1	1
11	1	1	0	0	0	0	0
12	1	1	1	0	1	1	1
13	1	1	0	1	0	0	0
14	1	0	1	1	1	0	1
15	0	1	1	1	1	1	1
16	0	0	0	0	0	1	0
17	1	1	1	0	1	0	1
18	1	1	0	0	0	1	1
19	0	1	0	0	0	0	0
20	0	1	0	0	0	1	1

Gambar. 6. Hasil Prediksi



Model	Actual \ Predicted	0	1	Σ
Logistic Regression	0	8872	1163	10035
	1	2340	3025	5365
	Σ	11212	4188	15400
kNN	0	8223	1812	10035
	1	2579	2786	5365
	Σ	10802	4598	15400
SVM	0	7099	2936	10035
	1	2014	3351	5365
	Σ	9113	6287	15400
Naive Bayes	0	7744	2291	10035
	1	1599	3766	5365
	Σ	9343	6057	15400
Tree	0	7920	2115	10035
	1	2439	2926	5365
	Σ	10359	5041	15400
Neural Network	0	8637	1398	10035
	1	2211	3154	5365
	Σ	10848	4552	15400
Random Forest	0	8443	1592	10035
	1	2337	3028	5365
	Σ	10780	4620	15400

Gambar. 7. Hasil Confusion Matrix

dibandingkan dengan seluruh wanita yang diprediksi terkena Diabetes Gestasional, walaupun pada kenyataannya ada yang terkena dan tidak terkena. Lalu berikut adalah hasil evaluasi untuk nilai Recall yaitu algoritma Logistic Regression diposisi pertama dengan nilai 0,773 yang merupakan nilai tertinggi. Lalu diikuti Neural Network diposisi kedua dan diposisi ketiga ada Naïve Bayes. Dimana nilai Recall ini akan menunjukkan rasio perbandingan dari

wanita yang diprediksi terkena Diabetes Gestasional, walaupun pada kenyataannya bisa saja benar terkena atau tidak terkena, dibandingkan dengan seluruh wanita yang benar-benar terkena Diabetes Gestasional.

Lalu untuk waktu testing dan training dapat dilihat pada gambar 9, dimana waktu tersebut diukur dalam satuan detik. Untuk algoritma Naïve Bayes menjadi algoritma dengan waktu training tercepat yaitu 1,092 detik. Dimana posisi kedua dan ketiga secara berurut diisi oleh kNN dengan waktu 1,124 detik dan Logistic Regression 4,613 detik. Dan algoritma yang butuh waktu paling lama melakukan training terhadap dataset yaitu Neural Network

Model	Precision	Recall	Model	Train time [s]	Test time [s]
kNN	0.707	0.715	kNN	1.124	1.293
Tree	0.700	0.704	Tree	9.678	0.020
SVM	0.693	0.679	SVM	9.064	1.764
Random Forest	0.740	0.746	Random Forest	4.629	0.746
Neural Network	0.760	0.766	Neural Network	143.154	0.784
Naive Bayes	0.757	0.747	Naive Bayes	1.092	0.244
Logistic Regression	0.767	0.773	Logistic Regression	4.613	0.246

Gambar. 8. Nilai *Precision* dan *Recall*

Gambar. 9. Waktu *Training* dan *Testing*

dengan waktu 143,154 detik. Tetapi pada saat testing, algoritma dengan waktu tercepat berubah menjadi algoritma Tree yaitu 0,020 detik, lalu ada Naïve Bayes dengan waktu 0,244 detik dan Logistic Regression diposisi ketiga dengan waktu 0,246 detik. Dan algoritma yang paling lama adalah SVM dengan waktu 1,764 detik.

Dari hasil yang didapatkan melalui software Orange, dapat kita lihat pada masing-masing algoritma memiliki kelebihan dan kekurangannya masing-masing. Untuk algoritma Tree, dapat kita lihat waktu untuk melakukan *Training* cukup lama, karena jumlah kriteria dari dataset cukup banyak dan proses pengambilan keputusan cukup panjang. Karena dengan algoritma ini, maka keputusan akan terus menghasilkan cabang-cabang baru, hingga keputusan didapatkan, yang mana ini akan sangat memakan memori karena penyimpanan rute dari tiap cabang tidak bisa kita prediksi berapa banyak. Juga untuk waktu hingga keputusan didapatkan juga akan lama, karena banyaknya cabang yang muncul akibat pengulangan kriteria yang digunakan. Tetapi setelah dilakukan *Training*, maka algoritma ini akan sangat cepat memberikan hasil keputusan sesuai dengan input yang ada, yang dapat kita lihat pada gambar sepuluh, dimana Tree menjadi algoritma dengan waktu *Testing* tercepat dan juga untuk kepresisian prediksinya cukup bagus, serta nilai recall yang cukup bagus juga. Lalu pada algoritma *Random Forest*, yang merupakan sekumpulan Tree, sehingga output yang dikeluarkan merupakan nilai rata-rata dari tiap Tree yang ada. Dimana pada algoritma ini, tentunya akan mengeluarkan Tree yang acak, sehingga akurasi akan lebih akurat, tetapi karena acak, maka juga memerlukan waktu yang cukup lebih lama. Sehingga terkadang kurang efektif, tetapi pada penelitian kali ini, hasil yang didapatkan untuk algoritma *Random Forest* cukup bagus. Pada algoritma SVM, prediksi yang dilakukan cukup mendapat nilai yang kurang baik. Hal ini disebabkan karena SVM sendiri yang kurang cocok untuk melakukan prediksi, terutama dengan dataset yang berjumlah banyak. Tetapi dalam melakukan klasifikasi, maka algoritma ini cukup baik. Seperti pada gambar delapan dan sembilan, dapat dilihat bahwa nilai yang diberikan SVM cukup jelek terutama kepresisian prediksinya, serta waktu *Training* yang cukup lama dan waktu *Testing* yang didapat menjadi yang terlama dibanding algoritma lain. Lalu pada algoritma *Neural Network* sendiri dapat dilihat seperti pada hasil yang didapatkan, dimana waktu *Training* yang dibutuhkan sangat lama. Ini disebabkan algoritma akan didorong untuk mencari keputusan dan belajar secara mandiri, yang tentunya semakin lama waktu *Training* yang dibutuhkan, akan semakin baik hasil prediksinya, terutama pada dataset yang memiliki hubungan yang rumit pada datanya. Untuk algoritma Naïve Bayes sendiri sebenarnya cukup baik karena penerapannya yang lebih mudah. Dimana algoritma ini akan melihat probabilitas dari probabilitas lainnya yang masih berkaitan, sehingga keputusan akan didapatkan. Tetapi algoritma ini lebih cocok untuk melakukan pengklasifikasian daripada prediksi, walaupun untuk hasil yang didapatkan, baik waktu *Training* dan *Testing*, serta nilai Recall dan presisi cukup baik. Selanjutnya algoritma kNN, yang melakukan pengklasifikasian pada data yang berdimensi banyak, dimana tiap dimensi ini akan menjadi titik-titik sebagai representasi dari data *Training*. Lalu akan dicari keputusan berdasarkan nilai terdekat dari tiap titik ke titik lainnya yang saling terhubung. Dan pada penelitian ini, hasil yang didapat dengan algoritma ini cukup baik. Yang terakhir, algoritma Regresi Logistik, yang merupakan salah satu jenis Regresi. Dimana pada algoritma ini, data yang digunakan berbentuk numerik, yang mana sangat cocok dengan dataset yang digunakan. Lalu dari masing-masing fitur yang ada pada dataset, akan dicari keterkaitan antara fitur penyebab dan fitur tujuan dari data yang ada, sehingga akan menghasilkan jawaban ya atau tidak.

Karena kesederhanaan dari model matematis yang digunakan, maka algoritma ini cocok dengan dataset berjumlah besar. Karena tidak memerlukan waktu yang lama, serta memori yang relatif kecil.

Juga dapat diketahui seperti pada penelitian sebelumnya, yang sudah dibahas pada subbab pendahuluan, karena pada penelitian ini menggunakan dataset berbentuk numerik yang cocok dengan algoritma Regresi Logistik. Sehingga pada penelitian kali ini, dapat kita analisa bahwa dengan adanya kecocokan antara bentuk dataset dengan algoritma yang ingin digunakan akan sangat penting. Selain mempengaruhi waktu *Training* dan *Testing*, juga akan mempengaruhi akurasi dan Recall. Dimana algoritma Regresi Logistik yang memang digunakan untuk melakukan prediksi, sehingga memang mendapatkan nilai akurasi, Recall maupun Precision yang sangat baik dibanding algoritma lainnya. Yang mana memang ada yang tidak sesuai dengan jenis datanya dan juga tidak sesuai dengan fungsi dari algoritma itu sendiri, seperti pada algoritma Naïve Bayes.

Berdasarkan hasil analisis yang sudah dilakukan, dan dengan berdasar bahwa hasil prediksi yang kita inginkan yaitu “Wanita diprediksi Diabetes Gestasional dan pada kenyataannya benar terkena Diabetes Gestasional maupun ada yang tidak terkena, dibandingkan dengan wanita diprediksi tidak terkena Diabetes Gestasional dan pada kenyataannya wanita tersebut sebenarnya terkena Diabetes Gestasional”, dimana sesuai dengan confusion matrix kita lebih memilih terjadinya True Positif dan False Positif, dibanding dengan False Negatif. Sehingga nilai Recall akan dipakai menjadi acuan untuk pemilihan algoritma paling baik untuk melakukan prediksi. Berdasarkan hasil itu, maka nilai Recall tertinggi adalah algoritma Logistic Regression, yang tentunya juga didukung oleh hasil evaluasi lain seperti Akurasi dan Precision yang juga menempati posisi pertama, walaupun untuk waktu training dan testing hanya menempati posisi ketiga.

IV. KESIMPULAN

Berdasarkan hasil penelitian dan analisis, dapat terlihat bahwa algoritma terbaik pada metode Supervised Learning guna melakukan prediksi pada dataset Diabetes Gestasional yaitu algoritma Logistic Regression. Dimana hasil ini didapat berdasarkan nilai Recall yang paling baik dan diikuti oleh nilai lainnya seperti Akurasi dan Precision, juga waktu untuk training dan testing yang terbilang lumayan cepat. Dimana hasil ini juga berdasar dari kecocokan jenis dataset dan algoritmanya, serta fungsi dari algoritma Logistic Regression yang memang biasa digunakan sebagai algoritma untuk melakukan prediksi. Lalu dapat dilihat juga, bahwa kecepatan waktu *Training* dataset tidak terlalu mempengaruhi baik nilai akurasi, Recall dan Precision menjadi lebih baik jika waktu yang dibutuhkan lebih singkat. Tetapi tidak menutup kemungkinan untuk algoritma lain memiliki nilai Recall yang lebih baik lagi, sebagai acuan dari prediksi yang dilakukan dalam penelitian ini. Bentuk dataset juga akan mempengaruhi seberapa baik hasil prediksi yang didapatkan, maka kesesuaian bentuk dataset yang dimiliki harus disesuaikan dengan algoritma yang ada, sehingga dapat memberikan hasil prediksi yang sangat baik.

DAFTAR PUSTAKA

- [1] R. Tyasnurita and A. Y. M. Pamungkas, “Deteksi Diabetik Retinopati menggunakan Regresi Logistik,” *ILKOM Jurnal Ilmiah*, vol. 12, no. 2, pp. 130–135, Aug. 2020, doi: 10.33096/ilkom.v12i2.578.130-135.
- [2] V. Yolanda, I. Cholissodin, and P. P. Adikara, “Klasifikasi Diagnosis Penyakit Diabetes Gestasional pada Ibu Hamil menggunakan Algoritme Neighbor Weighted K-Nearest Neighbor (NWKNN),” 2021. [Online]. Available: <http://j-ptiik.ub.ac.id>
- [3] I. Rosianal *et al.*, “Perbandingan Hasil Prediksi Diagnosis pada Indian Liver Patient Dataset (ILPD) dengan Teknik Supervised Learning Menggunakan Software Orange,” *Jurnal Telematika*, vol. 16, no. 2.
- [4] N. Hamdani and A. Setyanto, “Perbandingan Algoritma Regresi Logistik Dan Neural Network Pada Prediksi Nilai Hasil Pembinaan Dan Kelulusan Tepat Waktu,” *Jurnal Teknologi Informasi*, vol. XV, no. 1, 2020.
- [5] D. Wulan Sari and M. Maharani, “ANALISIS DAN PERBANDINGAN ALGORITMA PREDIKSI DALAMMENGETAHUI PERKIRAAN PENINGKATAN JUMLAH KASUS COVID-19 DIINDONESIA DENGAN METODOLOGI CRISP-DM,” 2021. [Online]. Available: <https://www.researchgate.net/publication/348648401>
- [6] Q. Wu *et al.*, “An early prediction model for gestational diabetes mellitus based on genetic variants and clinical characteristics in China,” *Diabetol Metab Syndr*, vol. 14, no. 1, Dec. 2022, doi: 10.1186/s13098-022-00788-y.
- [7] S. Martha, W. Andani, and S. W. Rizki, “Perbandingan Metode k-Nearest Neighbor, Regresi Logistik Biner, dan Pohon Klasifikasi pada Analisis Kelayakan Pemberian Kredit,” *Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi*, vol. 10, no. 2, pp. 262–273, Nov. 2022, doi: 10.34312/euler.v10i2.16751.
- [8] P. C. Algoritma, D. Naïve Bayes Untuk Prediksi Ketepatan Waktu Studi Mahasiswa, J. Nata Permana, R. Goejantoro, and S. Prangga, “Comparison Of C4.5 Algorithm and Naïve Bayes for Prediction Of Student Study Timeliness (Case Study: Departement of Statistics Mulawarman University)”.
- [9] S. Wiyono and T. Abidin, “Perbandingan Algoritma Machine Learning SVM dan Decision Tree untuk Prediksi Keaktifan Mahasiswa,” *Jurnal & Penelitian Teknik Informatika*, vol. 3, no. 1, 2018.
- [10] S. Nanda, M. Savvidou, A. Syngelaki, R. Akolekar, and K. H. Nicolaidis, “Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks,” *Prenat Diagn*, vol. 31, no. 2, pp. 135–141, Feb. 2011, doi: 10.1002/pd.2636.
- [11] I. Gnanadass, “Prediction of Gestational Diabetes by Machine Learning Algorithms,” *IEEE Potentials*, vol. 39, no. 6, pp. 32–37, Nov. 2020, doi: 10.1109/MPOT.2020.3015190.
- [12] G. Santosa, U. Kristen, and D. Wacana, “Perbandingan Akurasi Model Regresi Logistik untuk Prediksi Kategori IPMahasiswa Jalur Prestasi dengan Non Jalur Prestasi.” [Online]. Available: <https://www.researchgate.net/publication/324990799>
- [13] X. Zhang *et al.*, “Risk prediction model of gestational diabetes mellitus based on nomogram in a Chinese population cohort study,” *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-78164-x.

- [14] S. Widaningsih, “PERBANDINGAN METODE DATA MINING UNTUK PREDIKSI NILAI DAN WAKTU KELULUSAN MAHASISWA PRODI TEKNIK INFORMATIKA DENGAN ALGORITMA C4,5, NAÏVE BAYES, KNN DAN SVM,” *Jurnal Tekno Insentif*, vol. 13, no. 1, pp. 16–25, Apr. 2019, doi: 10.36787/jti.v13i1.78.
- [15] Mufdillah *et al.*, *Mengenal dan Upaya Mengatasi Diabetes Melitus dalam Kehamilan Diabetes Melitus dalam Kehamilan*.