Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

PREDIKSI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA REGRESI LOGISTIK

Sugianto[1]
Andi Maslan[2]

[1]Mahasiswa Program Studi Teknik Informatika, Universitas Putera Batam
[2]Dosen Program Studi Teknik Informatika, Universitas Putera Batam
email: pb200210039@upbatam.ac.id

ABSTRACT

The annual count of individuals afflicted with diabetes is rising. As to a survey carried out by the International Diabetes Federation (IDF), the global diabetes population is approximated to be 537 million by 2021, and this number is projected to increase further to exceed 780 million by 2045. The study's primary goal is to diagnose and forecast whether or not a patient has diabetes. The approach makes use of logistic regression, a statistical tool for modelling individual classifications of diabetes presence or absence. According to the diabetes risk prediction results, 43% of respondents gave consideration to the condition. Consequently, it has been demonstrated that normalisation enhances the accuracy of diabetes risk prediction using logistic regression methods. Based on the variables included, it is anticipated that the predictions made by this model will serve as a guide for the general public in understanding healthy living and diabetes prevention.

Keywords: *Diabetes Prediction; Logistical Regression; Risk Prediction; Diabetes Variable; Healthy Lifestyle Guidance*

PENDAHULUAN

Menurut survei yang dilakukan IDF di tahun 2021, diperkirakan ada sekitar 537 juta orang yang menderita diabetes di belahan dunia ini, dan akan terus meningkat menjadi lebih dari 780 juta pada tahun 2045. Peningkatan jumlah penderita penyakit diabetes ini menimbulkan tantangan yang significan bagi sistem Kesehatan secara keseluruhan dari individu yang terkena, Oleh karena itu, upaya pencegahan dan pengelolaan diabetes menjadi sangat penting untuk mengurangi dampak buruk penyakit ini.Salah satu tantangan utama dalam penanganan diabetes adalah kemampuan untuk melakukan prediksi dini mengenai risiko seseorang mengidap penyakit tersebut. Prediksi yang akurat dapat membantu dalam pengambilan keputusan medis yang tepat waktu dan strategi pencegahan yang lebih efektif. Namun, prediksi tidaklah sederhana karena melibatkan beberapa faktor risiko yang kompleks dan variabel klinis yang beragam. Untuk mengatasi permasalahan ini, penelitian ini akan menggunakan algoritma regresi logistik sebagai alat statistik untuk memodelkan klasifikasi adanya atau tidaknya diabetes pada individu (Cahyani et al., 2022).

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

Regresi logistik dipilih karena kemampuannya dalam mengolah data kategori dan memberikan probabilitas terkait risiko penyakit berdasarkan variabel yang relevan. Melalui pendekatan ini, diharapkan dapat diperoleh model prediksi yang akurat dan dapat diandalkan (Putra et al., 2024). Tujuan dari penelitian ini adalah untuk mengidentifikasi variabel yang berpengaruh signifikan terhadap risiko diabetes, membangun model prediksi diabetes yang akurat dan dapat diandalkan, serta mengevaluasi kinerja model prediksi dengan menggunakan metrik-metrik yang relevan seperti akurasi, presisi, *recall* dan *f1-score.*

KAJIAN TEORI

2.1 *Machine Learning*

Menurut IBM, machine learning adalah bagian dari cabang kecerdasan buatan. Bidang ini merupakan cabang ilmu komputer yang memusatkan perhatian tentang penerapan data dan algoritma untuk memecahkan masalah dan mencapai penilaian, dengan tujuan mereplikasi pembelajaran manusia dan meningkatkan kapasitas akurasi (Pratama et al., 2023).

2.2 *Classification*

Klasifikasi adalah metide yang digunakan dalam pembelajaran terawasi dari sebuah Teknik ML yang memanfaatkan *dataset*yang sudah ditraining (Purwono et al., 2021).

2.3 KNN (*K-Nearest Neighbors*)

Algoritma ini adalah algoritma yang dikenal sebagai non-numerik dalam data *mining* dan dapat digunakan untuk klasifikasi maupun regresi. Prinsip kerja dari K-NN sendiri adalah mencari jarak terdekat atau jarak *Euclidean* berdasarkan nilai k (A'yuniyah & Reza, 2023).

2.4 *Decision Tree*

Struktur data yang terdiri dari satu *node* dan satu sisi disebut pohon. Pohon keputusan dengan *node* nama *internal* dan atributnya, sisi yang ditandai dengan nilai atribut potensial , dan *node* lembar yang menampilkan banyak kela, adalah ilustrasi sederhana dari metode klasifikasi untuk beberapa kelas (Nasrullah, 2021)

2.5 SVM (*Support Vector Machine*)

Sistem pembelajaran menggunakan hipotesis fungsi linear dalam fitur dimensi tinggi. Algoritma pembelajaran yang didasarkan pada teori optimalisasi digunakan untuk melatih SVM (Isnain et al., 2021).

2.6 *Random Forest*

Berdsarkan klasifikasi pohon dan regresi, algoritma RF secara berulang menerapkan pendekatan pemisahan biner untu mencapai *node* terakhir dalam struktur pohon. Keuntungan dari algoritma ini meliputi kemampuan untuk menghasilkan sangat sedikit kesalahan, kinerja klasifikasi yang kuat, kemudahan penggunaan saat menangani sejumlah besar data pelatihan, dan teknik yang berguna untuk memperkirakan data yang hilang (Pamuji & Ramadhan, 2021)

2.7 Regresi Logistik

Teknik klasifikasi yang digunakan untuk menentukan seberapa banyak faktor prediktor berkelanjutan dan variabel respons kategori terkait. Memprediksi probabilitas atau probabilitas suatu peristiwa berdasarkan predictor yang

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

relevan adalah tujuan utama metode ini (Fitriyani et al., 2023).

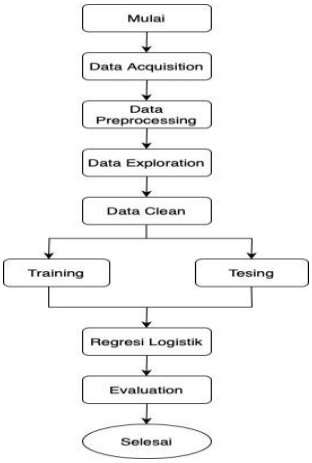
2.8 *K-Means Clustering*
Langkah pertama dari pendekatan ini adalah untuk membuat partisi *cluster* awal, yang kemudian diperbaiki secara iterative sampai partisi kluster tidak berubah secara signifikan (Sulistiyawati & Supriyanto, 2021)

2.9 *Hierarchical Clustering*
Proses mengumpulkan dua atau lebih data yang sebanding dengan satu titik pertama, kemudian bergerak ke objek lain yang sama dekat. Teknik ini dikenal sebagai *hierarki*. Proses ini terus berlanjut sampai *cluster* mulai mengambil bentuk pohon (Sadewo et al., 2019).

2.10 DBSCAN (*Density-Based Spatial Clustering of Applications Noise*)
DBSCAN adalah algoritma yang dapt mengidentifikasi kebisingan atau *outliers* dalam sejumlah besar data dan menghasilkan *cluster* yang lebih baik dan lebih akurat tanpa memerlukan (Indini et al., 2022).

METODE PENELITIAN

3.1 Regresi Logistik
Berikut ini merupakan tahapan yang ada pada algoritma regresi logistik:



Gambar 1. Proses Analisis Data

1. Data *Acquisition*
   Proses pengumpulan data yang diperlukan untuk analisis dan pemodelan data.
2. Data *Pre-processing*
   Dalam proses ini data akan dibersihkan yang akan dinormalkan untuk memastikan semua variabel berada dalam rentang yang sama.
3. Data *Exploration*
   Tahap ini merupakan salah satu langkah yang penting karena peneliti akan memeriksa dan memahami data yang sudah dikumpulkan.
4. Data *Clean*
   Tujuan dari tahap ini adalah untuk memperbaiki dan menghapus data yang kotor atau tidak konsisten sehingga data menjadi lebih akurat dan siap untuk dianalisis.
5. *Training* dan *Testing*
   Pada tahapan ini data akan belajar dari model *machine learning* yang sudah diajarkan. Agar model dapat memahami pola dan hubungan maka diperlukanlah *training.* Setelah itu data akan diuji keakurasiannya pada tahap *testing*.

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

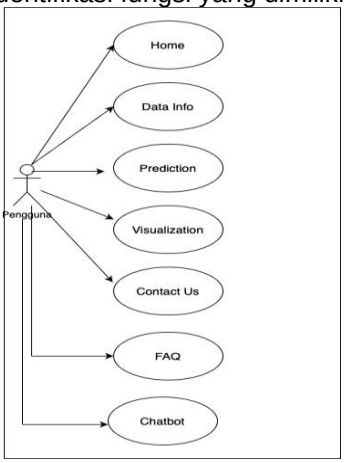
ISSN (Online) 2715-6265

6. Regresi Logistik
   Mengacu pada proses pengembangan dan penerapan model regresi logistic untuk memprediksi variabel respons biner, seperti keberadaan atau ketiadaan suatu kondisi atau peristiwa.
7. *Evaluation*
   Akan diterapkannya suatu model pada set pengujian untuk mendapatkan prediksi yang akurat untuk data yang akan digunakan.

3.2 UML (*Unified Model Language*)

1. *Use Case Diagram*

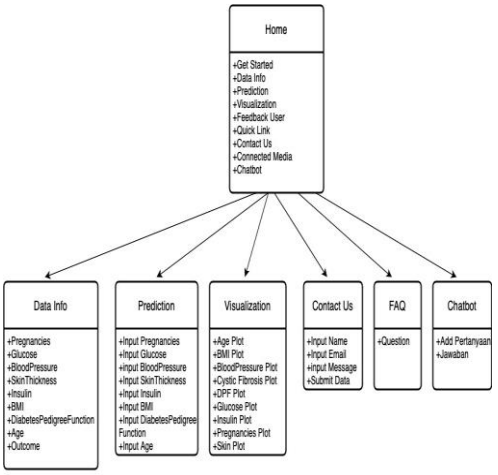
Diagram ini menggambarkan bagaimana sistem beroperasi dari perspektif internal, dengan fokus pada pengamatan internal sistem dan digunakan untuk mengidentifikasi fungsi yang dimiliki.



Gambar 2. *Use Case Diagram*

2. *Class Diagram*

Diagram kelas disusun untuk memastikan bahwa pembuat program, saat membuat kelas-kelas, mengikuti desain yang telah direncanakan, sehingga dokumentasi perancangan selaras dan konsisten menjalankan fungsi sesuai yang sistem inginkan.



Gambar 3. *Class Diagram*

HASIL DAN PEMBAHASAN

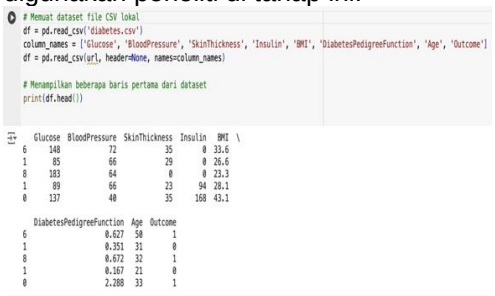
4.1 Hasil Penelitian

a. *Dataset*

Dataset yang digunakan adalah *Pima Indians Diabetes* dengan total jumlah data sebanyak 768 data yang diambil dari situs *Kaggle*.

b. *Pre-processing*

Tujuannya untuk mempersiapkan data sehingga dapat diolah dengan benar oleh algoritma ML. Ada 8 variabel yang digunakan peneliti di tahap ini.



```
# Memuat dataset file CSV lokal
df = pd.read_csv('diabetes.csv')
column_names = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome']
df = pd.read_csv(url, header=None, names=column_names)

# Menampilkan beberapa baris pertama dari dataset
print(df.head())
```

```
   Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
6      148             72             35        0  33.6
1       85             66             29        0  26.6
8      183             64              0        0  23.3
1       89             66             23       94  28.1
0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
6                     0.627   50        1
1                     0.351   31        0
8                     0.672   32        1
1                     0.167   21        0
0                     2.288   33        1
```

Gambar 4. Tampilan Parameter

c. *Prediction*

Untuk menggunakan model yang telah dikembangkan untuk menggeneralisasi

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

dan membuat estimasi tentang data yang belum diketahui hasilnya.

```
# Prediksi data testing
y_pred = model.predict(X_test)
y_pred_proba = model.predict_proba(X_test)[:, 1]  # Probabilitas prediksi

# Evaluasi model
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
print(classification_report(y_test, y_pred))

# Menampilkan beberapa prediksi dan probabilitas
for i in range(10):
    print(f'Prediksi: {y_pred[i]}, Probabilitas: {y_pred_proba[i]:.2f}')
```

Gambar 5. *Code Prediction*

d. *Training Dataset*

Pada awalnya data yang berjumlah 768 data menjadi 718 data karena adanya nilai yang hilang pada data. Kemudian data akan dilatih dengan rasio 70:30. Data *training* sebanyak 502 dan data *testing* sebanyak 216 data.

e. *Testing Dataset*

Hasilnya ada 2 yaitu yang dinormalisasi dan tanpa normalisasi.

a. Normalisasi

Tabel 4.1 *Confussion Matrix* dengan Normalisasi

		Prediksi	
		Diabetes	Tidak Diabetes
Aktual	Diabetes	38	31
	Tidak Diabetes	20	127

$$Accuray = \frac{TP+TN}{TP+TN+FN+FP} = \frac{38+127}{38+127+31+20} = \frac{165}{216} = \frac{55}{72} = 0.76$$

$$Precision = \frac{TP}{TP+FP} = \frac{38}{38+20} = \frac{38}{58} = 0.66$$

$$Recall = \frac{TP}{TP+FN} = \frac{38}{38+31} = \frac{38}{69} = 0.55$$

$$F1 - Score = 2 \text{ x } \frac{precision \text{ x } recall}{precicion + recall} = 2 \text{ x } \frac{0.66 \, x \, 0.55}{0.66 + 0.55} = 2 \text{ x } \frac{0.363}{1.21} = \frac{0.726}{1.21} = 0.60$$

b. Tanpa Normalisasi

Tabel 4.2 *Confussion Matrix* Tanpa Normalisasi

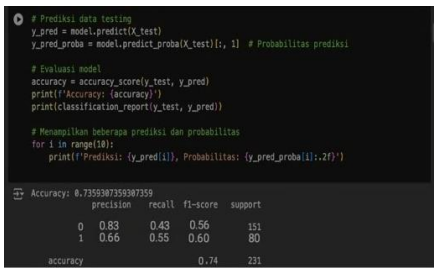
		Prediksi	
		Diabetes	Tidak Diabetes
Aktual	Diabetes	35	47
	Tidak Diabetes	7	127

$$Accuray = \frac{TP+TN}{TP+TN+FN+FP} = \frac{35+127}{35+127+47+7} = \frac{162}{216} = 0.75$$

$$Precision = \frac{TP}{TP+FP} = \frac{35}{35+7} = \frac{35}{42} = 0.83$$

$$Recall = \frac{TP}{TP+FN} = \frac{35}{35+47} = \frac{35}{82} = 0.43$$

$$F1 - Score = 2 \text{ x } \frac{precision \text{ x } recall}{precicion + recall} = 2 \text{ x } \frac{0.83 \, x \, 0.43}{0.83 + 0.43} = 2 \text{ x } \frac{0.3569}{1.26} = \frac{0.71}{1.26} = 0.56$$



Gambar 6. *Code Testing*

4.2 Hasil Rancangan

a. Tampilan Halaman *Home*

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

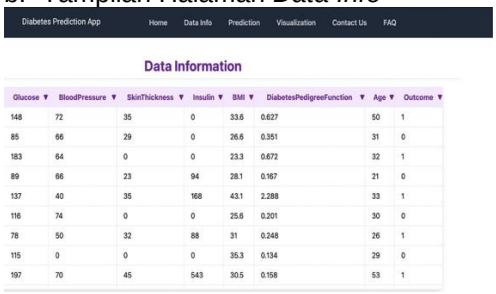
Jurnal Comasie

ISSN (Online) 2715-6265

Gambar 7. Halaman Utama

Pengguna pertama kali memasuki halaman utama untuk mengakses menu selanjutnya dapat menekan tombol *get started*.

b. Tampilan Halaman *Data Info*

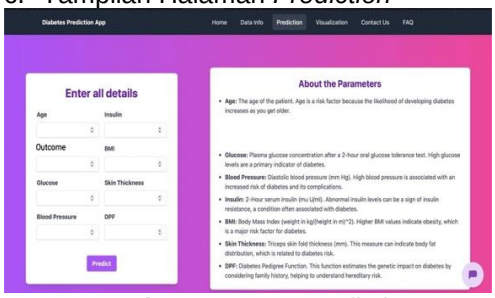


Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
148	72	35	0	33.6	0.627	50	1
85	66	29	0	26.6	0.351	31	0
183	64	0	0	23.3	0.672	32	1
89	66	23	94	28.1	0.167	21	0
137	40	35	168	43.1	2.288	33	1
116	74	0	0	25.6	0.201	30	0
78	50	32	88	31	0.248	26	1
115	0	0	0	35.3	0.134	29	0
197	70	45	543	30.5	0.158	53	1

Gambar 8. Menu *Data Info*

Halaman ini menyajikan seluruh parameter yang digunakan pada *dataset*.

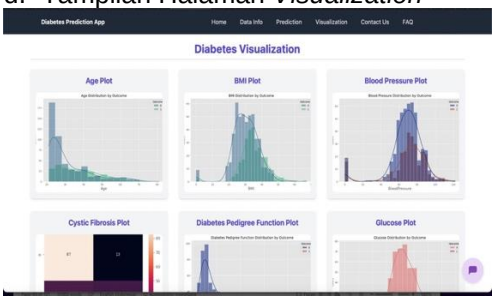
c. Tampilan Halaman *Prediction*



Gambar 9. Menu *Prediction*

Pengguna akan mengisi data dari setiap parameter, sehingga ML akan memprediksinya.

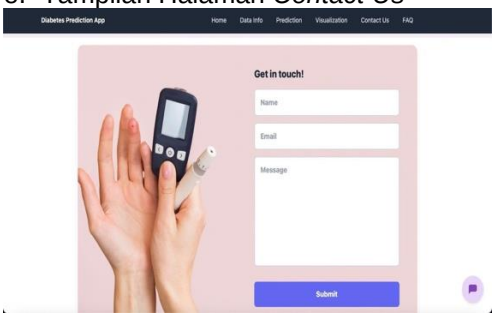
d. Tampilan Halaman *Visualization*



Gambar 10. Menu *Visualization*

Di halaman ini akan ditampilkan beberapa macam grafik dari setiap parameter yang digunakan pada data.

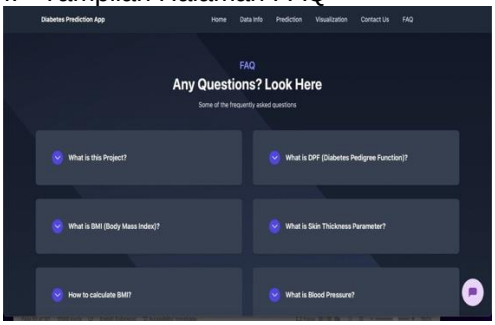
e. Tampilan Halaman *Contact Us*



Gambar 11. Menu *Contact Us*

Menu ini digunakan oleh pengguna untuk memberikan kesan dan saran bagi *website* ini.

f. Tampilan Halaman FAQ



Gambar 12. Menu FAQ

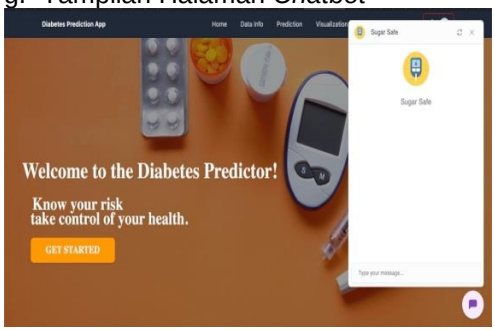
Di halaman ini berisi pertanyaan yang sering ditanyakn oleh pengguna.

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

g. Tampilan Halaman *Chatbot*



Gambar 13. Menu *Chatbot*

Menu ini digunakan oleh pengguna untuk bertanya seputar penyakit diabetes mulai dari tips dan saran serta pertanyaan lainnya.

4.3 Pengujian

a. Pengujian *Alpha*

Pada fase pengetesan ini, peneliti menggunakan metode *blackbox* untuk menguji secara internal. Untuk hasilnya dapat dilihat pada **Tabel 4.**

Tabel 4. Pengujian *Alpha*

No.	Input	Output	Hasil
1.	Membuka menu *Home*	Menampilkan tampilan yang ada pada *Home*	Berhasil
2.	Menekan tombol *Get Started*	Menampilkan halaman di menu selanjutnya	Berhasil
3.	Menekan tombol *reset page*	Menampilkan halaman kosong	Berhasil
4.	Menekan tombol *stop custom*	Menampilkan halaman semula sebelum di *reset*	Berhasil
5.	Membuka menu *Data Info*	Menampilkan tampilan yang ada pada *menu Data Info*	Berhasil

(Sumber: Data Penelitian, 2024)

SIMPULAN

Penelitian dengan judul "PREDIKSI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA REGRESI LOGISTIK" menghasilkan Kesimpulan sebagai berikut:

1. *Website* prediksi penyakit diabetes ini berhasil menerapkan teknologi MLmenggunakan algoritma regresi logistik.
2. *Website* yang dirancang ini juga memberikan akses fleksibel kepada masyarakat dalam bidang kesehatan.
3. Hasil dari pengujian *alpha* menunjukan aplikasi ini sangat bermanfaat bagi pengguna untuk mengetahui prediksi penyakit diabetes.

DAFTAR PUSTAKA

A'yuniyah, Q., & Reza, M. (2023). Penerapan Algoritma K-Nearest

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

Neighbor Untuk Klasifikasi Jurusan Siswa Di Sma Negeri 15 Pekanbaru. *IJIRSE: Indonesian Journal of Informatic Research and Software Engineering*, *3*(1), 1–7.

Cahyani, Q. R., Finandi, M. J., Rianti, J., Arianti, D. L., & Putra, A. D. P. (2022). Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik. *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, *1*(2), 1–8.

Fitriyani, N., Amalia, D. R., Handayani, H. H., & Masruriyah, A. F. N. (2023). Aplikasi Berbasis Web Berdasarkan Model Klasifikasi Algoritma SVM dan Logistic RegressionTerhadap Data Diabetes. *Riset Dan E-Jurnal Manajemen Informatika Komputer*, *7*(4), 1–10.

Indini, D. P., Siburian, S. R., Nurhasanah, Utomo, D. P., & Mesran. (2022). *IMPLEMENTASI ALGORITMA DBSCAN UNTUK CLUSTERING SELEKSI PENENTUAN MAHASISWA YANG BERHAK MENERIMA BEASISWA YAYASAN*. 1–8.

Isnain, A. R., Sakti, A. I., Alita, D., & Marga, N. S. (2021). SENTIMEN ANALISIS PUBLIK TERHADAP KEBIJAKAN LOCKDOWN PEMERINTAH JAKARTA MENGGUNAKAN ALGORITMA SVM. *Jurnal Data Mining Dan Sistem Informasi*, *2*(1), 1–7.

Nasrullah, A. H. (2021). IMPLEMENTASI ALGORITMA DECISION TREEUNTUK KLASIFIKASI PRODUK LARIS. *Jurnal Ilmiah Ilmu Komputer*, *7*(2), 1–7.

Pamuji, F. Y., & Ramadhan, V. P. (2021). Komparasi Algoritma Random Forest Dan Decision Tree Untuk Memprediksi Keberhasilan Immunotheraphy. *Jurnal Teknologi Dan Manajemen Informatika*, *7*(1), 1–5.

Pratama, R., Herdiansyah, M. I., Syamsuari, D., & Syazili, A. (2023). Prediksi Customer Retention Perusahaan Asuransi Menggunakan Machine Learning. *Jurnal SISFOKOM (Sistem Informasi Dan Komputer*, *12*(1), 1–10.

Purwono, Wirasto, A., & NIsa, K. (2021). Komparasi Algoritma Machine Learning Untuk Klasifikasi Kelompok Obat. *JURNAL SISFOTENIKA*, *11*(2), 1–12.

Putra, R. P., Zebua, R. S. Y., Budiman, Rahayu, P. W., Bangsa, M. T. A., Zulfadhilah, M., Choirina, P., Wahyudi, F., & Andiyan, A. (2024). *Data Mining Algoritma dan Penerapannya* (Vol. 1).

Sadewo, M. G., Eriza, A., Windarto, A. P., & Hartama, D. (2019). Algoritma K-Means Dalam Mengelompokkan Desa/Kelurahan Menurut Keberadaan Keluarga Pengguna Listrik dan Sumber Penerangan Jalan Utama Berdasarkan Provinsi. *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, 1–8.

Sulistiyawati, A., & Supriyanto, E. (2021). Implementasi Algoritma K-means Clustering dalam Penetuan Siswa Kelas Unggulan. *Jurnal TEKNO KOMPAK*, *15*(2), 1–12.

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

Sugianto
Mahasiswa Program Studi Teknik Informatika dari Universitas Putera Batam

Andi Maslan
Seorang Dosen di Universitas Putera Batam. Ia mengajar program studi Teknik Informatika di Fakultas Teknik dan Komputer