Implementasi Algoritma SVM dalam Memprediksi Penyakit Stroke

Alexsander[1], Ahmad Nazri[2], Rio Agus Panbudi[3], Junadhi[4]
[1234] Teknik Informatika, Sekolah Tinggi Manajemen Informatika dan Komputer AMIK Riau
[1]alexanderkomed3@gmail.com, [2]anazri@gmail.com, [3]rioaguspanbudi@gmail.com, [4]junadhi@sar.ac.id

Abstract - The aim of this research is to build a stroke prediction model using the Support Vector Machine (SVM) algorithm based on available medical data, with the aim of identifying risk factors that are potentially associated with the disease. Stroke is a medical condition that occurs when the blood supply to part of the brain is disrupted or stopped, causing damage to brain cells due to lack of oxygen and nutrients. This can be caused by blocked blood vessels (ischemic stroke) or rupture of blood vessels (hemorrhagic stroke). This study implemented SVM to produce a classification model. The SVM method is known for its ability to find the best hyperplane for separate data classes. Evaluation of the performance of the SVM algorithm is carried out using prediction accuracy, precision, recall, and F1-score. The results of this study provide a better understanding of the effectiveness of SVM in predicting stroke disease. In this research, the best results were found, namely with an accuracy level of 95 percent after carrying out 15 experiments with an 80:20 distribution of training and testing data. This research applies the Support Vector Machine (SVM) algorithm to predict stroke. The accuracy produced by the Support Vector Machine (SVM) method in making predictions is 95% and is said to be very good.

Keywords* — *Algoritma, Support Vector Machine (SVM), Stroke, Stroke Iskemik, Stroke Hemoragik.

Abstrak*—Tujuan dari penelitian ini adalah untuk membangun model prediksi stroke menggunakan algoritma Support Vector Machine (SVM) berdasarkan data medis yang tersedia, dengan tujuan untuk mengidentifikasi faktor risiko yang berpotensi terkait dengan penyakit tersebut. Stroke adalah kondisi medis yang terjadi ketika pasokan darah ke bagian otak terganggu atau berhenti, menyebabkan kerusakan pada sel-sel otak akibat kekurangan oksigen dan nutrisi. Hal ini dapat disebabkan oleh pembuluh darah yang tersumbat (stroke iskemik) atau pecahnya pembuluh darah (stroke hemoragik). Penelitian ini menerapkan SVM untuk menghasilkan model klasifikasi. Metode SVM dikenal karena kemampuannya untuk menemukan hiperplan terbaik untuk memisahkan kelas data. Evaluasi kinerja algoritma SVM dilakukan menggunakan akurasi prediksi, presisi, recall, dan F1-score. Hasil dari penelitian ini memberikan pemahaman yang lebih baik tentang efektivitas SVM dalam memprediksi penyakit stroke. Dalam penelitian ini, hasil terbaik ditemukan, yaitu dengan tingkat akurasi 95 persen setelah melakukan 15 percobaan dengan distribusi data latih dan uji 80:20. Akurasi yang dihasilkan oleh metode Support Vector Machine (SVM) dalam membuat prediksi adalah 95% dan dikatakan sangat baik.

Kata Kunci— Algoritma, Support Vector Machine (SVM), Stroke, Stroke Iskemik, Stroke Hemoragik.

I. Pendahuluan

Pertama – tama penulis menucapkan puji dan syukur kepada Tuhan Yang Maha Esa karena berkat dan rahmatnya kepada penulis sehingga laporan proposal ini dapat selesai tepat pada waktunya. Laporan ini di susun sebagai salah satu pertanggung jawaban penulis untuk memenuhi tugas dari mata kuliah Kerja Praktek (KP. Langkah praktis dalam mempersiapkan mahasiswa untuk dapat tangkas, ahli, bertanggung jawab dan trampil dalam kehidupannya pada dunia kerja. Dan diharapkan kepada mahasiswa agar mendapatkan gambaran tentang dunia kerja yang sebenarnya.

Tidak lupa penulis mengucapkan terima kasih kepada semua pihak yang telah membantu menyelesaikan laporan ini. Semoga bisa bermanfaat bagi kita dan menjadi acuan bagi mahasiswa yang nantinya mengikuti penulisan seperti ini. Dan tentunya penulis menyadari laporan ini masih sangat jauh dari sempurna. Untuk itu penulis mengharapkan saran serta kritik kepada penulis demi perbaikan pembuatan laporan penulis di masa yang akan datang.

Stroke merupakan salah satu penyakit serius yang dapat menyebabkan dampak yang signifikan terhadap kesehatan dan kualitas hidup seseorang. Menurut data dari World Health Organization (WHO), stroke merupakan penyebab utama kematian dan kecacatan di seluruh dunia. Faktor risiko untuk stroke dapat bervariasi, termasuk gaya hidup yang tidak sehat, riwayat medis, serta faktor genetik.

Pada saat yang sama, teknologi dan metode analisis data semakin berkembang, memberikan peluang untuk menerapkan pendekatan berbasis data dalam prediksi penyakit. Salah satu pendekatan yang populer adalah Support Vector Machine (SVM), yang merupakan algoritma pembelajaran mesin yang dapat digunakan untuk klasifikasi dan regresi.

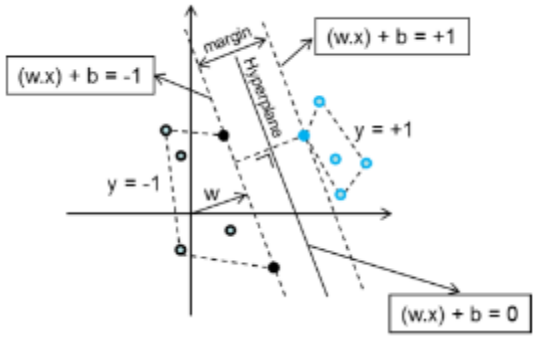
Dengan menggabungkan kemajuan dalam ilmu kedokteran dan teknologi informasi, implementasi SVM dalam prediksi penyakit stroke dapat memberikan manfaat yang signifikan. Hal ini dapat membantu dalam identifikasi faktor risiko yang mungkin terkait dengan stroke, memungkinkan deteksi dini, serta memberikan peluang untuk intervensi yang lebih efektif.

Namun, untuk menerapkan SVM dalam prediksi penyakit stroke, diperlukan ketersediaan data yang berkualitas tinggi, termasuk data medis yang lengkap dan akurat tentang pasien yang telah mengalami stroke serta pasien yang tidak mengalami stroke. Selain itu, perlu juga dilakukan validasi model untuk memastikan tingkat akurasi dan keandalan prediksi.

Dengan demikian, implementasi algoritma SVM dalam memprediksi penyakit stroke menjadi relevan dalam konteks upaya pencegahan, diagnosis dini, dan pengelolaan penyakit tersebut, dengan potensi untuk meningkatkan kualitas hidup dan mengurangi angka kematian yang disebabkan oleh stroke.

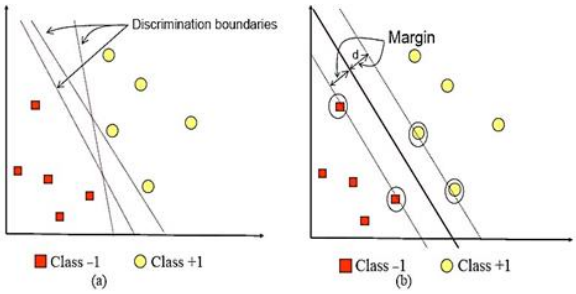
II. Metode Penelitian

Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992 bersama rekannya Bernhard Boser dan Isabelle Guyon. SVM merupakan algoritma yang bekerja menggunakan pemetaan nonlinier untuk mengubah data pelatihan asli ke dimensi yang lebih tinggi. Dalam hal ini dimensi baru, akan mencari hyperplane untuk memisahkan secara linier dan dengan pemetaan nonlinier yang tepat ke dimensi lebih tinggi, data dari dua kelas selalu dapat dipisahkan dengan hyperplane tersebut. SVM menemukan ini menggunakan support vector dan margin, Widodo (2013). Dalam teknik ini, kita berusaha untuk menemukan fungsi pemisah (klasifier) yang optimal yang bisa memisahkan dua kelas yang berbeda. Teknik ini berusaha menemukan fungsi pemisah (hyperplane) terbaik diantara fungsi yang tidak terbatas jumlahnya untuk memisahkan dua macam obyek. Hyperplane terbaik adalah hyperplane yang terletak di tengah-tengah antara dua set obyek dari dua kelas.



Gambar 1. SVM menemukan hyperlane terbaik

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah class pada input space. Gambar 2 memperlihatkan beberapa pattern yang merupakan anggota dari dua buah class : positif (dinotasikan dengan +1) dan negatif (dinotasikan dengan –1). Pattern yang tergabung pada class negatif disimbolkan dengan kotak, sedangkan pattern pada class positif, disimbolkan dengan lingkaran. Proses pembelajaran dalam problem klasifikasi diterjemahkan sebagai upaya menemukan garis (hyperplane) yang memisahkan antara kedua kelompok tersebut. Berbagai alternatif garis pemisah (discrimination boundaries) ditunjukkan pada gambar 2.2 (Nugroho, 2008).



Gambar 2. Hyperplane Terbaik Yang Memisahkan Kedua Class Negatif dan Positif

Seperti yang dituliskan di atas, konsep dari SVM yaitu sebagai usaha untuk mencari hyperplane yang terbaik dan berfungsi sebagai pemisah antara dua buah class pada input space. Pada Gambar 2.1 diperlihatkan bahwa beberapa pattern yang merupakan bagian anggota dari dua buah class : +1 dan -1. Pattern yang tergabung pada class -1 diberi simbol dengan warna merah kotak dan sedangkan pattern pada class +1 diberi simbol dengan warna kuning lingkaran. Problem yang terjadi pada klasifikasi tersebut dapat dijelaskan dengan usaha menemukan garis hyperplane yang memisahkan antara kedua kelompok tersebut. Garis solid pada Gambar 1 menunjukkan hyperplane yang terbaik, yaitu yang terdapat pada tengah – tengah kedua class, sedangkan titik merah dan kuning yang berada didalam lingkaran hitam adalah support vector.

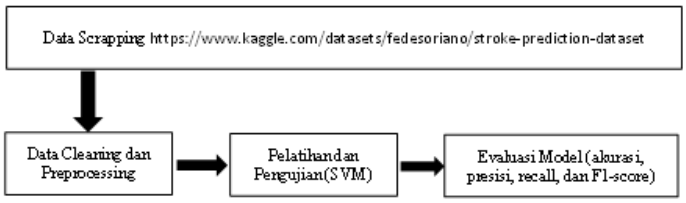
Berikut adalah beberapa macam fungsi kernel SVM yang dapat dilihat pada Tabel 1

Tabel 1. Macam-Macam Fungsi Kernel

Jenis Kernel	Definisi
Polynomial	$K(\bar{x}_i, \bar{x}_j) = \bar{x}_i, \bar{x}_j + 1)^p$
RBF	$K(\bar{x}_i, \bar{x}_j) = \exp\left(-\frac{\|\bar{x}i - \bar{x}j\|}{2\sigma}\right)$
Sigmoid	$K(\bar{x}_i, \bar{x}_j) = \tanh(\alpha\bar{x}_i, \bar{x}_j + \beta)$

Tahapan Penelitian

Berikut ini merupakan gambar yang menjelaskan langkah-langkah yang dilakukan dalam penelitian ini.



Gambar 3. Alur Metodologi Penelitian

Langkah awal dalam penelitian ini adalah melakukan pengumpulan data dari situs web *https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.* Proses scraping data dilakukan untuk mendapatkan dataset yang komprehensif, mencakup informasi

penting seperti atribut diberikan oleh pengguna. Langkah ini menjadi krusial karena dataset yang baik merupakan fondasi utama dalam melaksanakan pelatihan dan pengujian terhadap algoritma pada SVM. Setelah pengumpulan data, langkah berikutnya dalam penelitian ini adalah melakukan serangkaian proses untuk memastikan bahwa dataset yang diperoleh bersih, terstruktur, dan siap untuk diolah menggunakan algoritma SVM.

Pengambilan data dilakukan dengan mengambil data dari sumber public yaitu Kaggle repository dengan url berikut: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. Dataset yang digunakan terdiri dari 5.110 record dengan 12 atribut. Atribut dalam data tersebut seperti pada Tabel 2:

Tabel 2. Deskripsi Atribut Dataset

No.	Atribut	Deskripsi
1	id	Id pasien
2	Gender	Jenis kelamin pasien
3	Age	Umur pasien
4	Hypertension	Riwayat hipertensi / tekanan darah tinggi
5	Heart desease	Riwayat penyakit jantung
6	Ever married	Status perkawinan
7	Work type	Tipe pekerjaan pasien
8	Residence type	Jenis tempat tinggal pasien
9	Avg glucose level	Riwayat kadar gula pasien
10	Bmi	Ukuran berat badan pasien
11	Smoking status	Riwayat merokok atau tidaknya pasien
12	stroke	Prediksi apakah pasien rentan atau tidaknya terkena stroke

Langkah selanjutnya yaitu proses data cleaning, dan preprocessing. Pertama, data cleaning melibatkan identifikasi dan penanganan data yang tidak lengkap, duplikat, atau tidak konsisten. Ini mencakup penghapusan entri data yang tidak relevan atau tidak lengkap, penanganan nilai yang hilang, dan pengidentifikasian serta penanganan duplikasi data. Proses ini penting untuk memastikan integritas dan kualitas dataset sebelum dilibatkan dalam analisis lebih lanjut.

Selanjutnya, preprocessing melibatkan serangkaian langkah untuk menyiapkan data agar sesuai dengan kebutuhan algoritma SVM. Ini termasuk normalisasi data, konversi format, dan pemilihan atribut yang relevan. Proses ini bertujuan untuk meningkatkan efisiensi dan performa algoritma, serta memastikan konsistensi dan keakuratan analisis.

Kemudian kita lakukan pelatihan dan pengujian model SVM. Tahap utama dari penelitian ini adalah klasifikasi dengan menggunakan algoritma Support Vector Machine. Fitur yang sudah dipilih sebelumnya akan digunakan sebagai masukan perhitungan oleh Support Vector Machine, untuk mengklasifikasikan dokumen. Pada tahap ini digunakan dokumen training sebagai dokumen masukan. Teks dari setiap kalimat tentang penyakit stroke sebelumnya telah ditransformasikan. Algoritma klasifikasi SVM menggunakan data latih untuk membentuk model classifier, model yang terbentuk akan digunakan sebagai prediksi kelas data baru yang belum pernah ada sebelumnya. Data latih dan data uji yang digunakan adalah data yang telah memiliki label kelas, dengan perbandingan data latih dan data uji adalah 80% : 20%.

Selanjutnya, pada proses training, dilakukan perubahan komposisi split dataset dan hasil akurasi terbaik didapat pada komposisi split 80:20. Pada evaluasi dan pengujian, akan dicari nilai precision,recall, dan f1-score.

III. Hasil dan Pembahasan

Pengolahan data dilakukan setelah proses pengambilan data di kaggle berhasil dilakukan. Pada proses pengolahan data sendiri dilakukan proses pembersihan data, agar data yang diolah benar-benar data bersih bukan data mentah (data kotor). Pada tahapan ini, membutuhkan eksplorasi atau pendalaman terhadap dataset. Eksplorasi dilakukan dengan tujuan untuk menunjukan pada semua atribut dan class dalam dataset tersebut valid, sehingga bisa digunakan untuk objek penelitian yang baik. Maka dari itu, tujuan untuk mengetahui hasil predikisi terbaik.

Dataset yang digunakan memiliki 12 atribut original dari sumbernya, akan tetapi tidak semua atribut atau features tersebut akan digunakan karena terdapat features yang tidak dapat membantu dalam proses prediksi ini sehingga perlu dilakukan feature selection. Dari 12 atribut, fitur yang akan digunakan hanya berjumlah 6 atribut yaitu seperti pada Tabel 3:

Tabel 3. Atribut Dataset Yang Digunakan

No.	Atribut	Deskripsi
1	id	Id pasien
2	Age	Umur pasien
3	Hypertension	Riwayat hipertensi / tekanan darah tinggi
4	Heart desease	Riwayat penyakit jantung
5	Avg glucose level	Riwayat kadar gula pasien
6	Bmi	Ukuran berat badan pasien

Selain menghilangkan feature yang tidak penting, diperlukan juga preprocessing text untuk merubah dataset yang masih original dan masih dalam keadaan kotor dan belum siap dilakukan klasifikasi. Adapun bisa dilakukan klasifikasi, memungkinkan akurasinya rendah.

Kita tidak bisa memproses data awal begitu saja, karena dari data awal banyak terdapat data yang tidak memnuhi syarat untuk di lakukan pengujian. Disini kita akan melalui suatu tahapan yang dinamai dengan tahapan processing data. Dimana pada tahapan ini seluruh data akan di saring dan dipilih untuk di lakukan pengujian. Pada data awal, kita memiliki data sebanyak 5.110 data. Setelah kita lakukan processing data, maka data yang berjumlah 5.110 ini akan

mengalami perubahan menjadi 4.909 data. Untuk lebih jelasnya dapat kita lihat pada gambar di bawah ini:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5104	14180	Female	13.0	0	0	No	children	Rural	103.08	18.6	Unknown	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

4909 rows × 12 columns

Gambar 3. Jumlah Data Setelah Dilakukan Processing Data

Tahap utama dari penelitian ini adalah klasifikasi dengan menggunakan algoritma Support Vector Machine. Fitur yang sudah dipilih sebelumnya akan digunakan sebagai masukan perhitungan oleh Support Vector Machine, untuk mengklasifikasikan dokumen. Pada tahap ini digunakan dokumen training sebagai dokumen masukan. Teks dari setiap kalimat tentang penyakit stroke sebelumnya telah ditransformasikan. Algoritma klasifikasi SVM menggunakan data latih untuk membentuk model classifier, model yang terbentuk akan digunakan sebagai prediksi kelas data baru yang belum pernah ada sebelumnya. Data latih dan data uji yang digunakan adalah data yang telah memiliki label kelas, dengan perbandingan data latih dan data uji adalah 80% : 20%.

Setelah model yang diusulkan dibuat dan dilakukan training, maka terdapat hasil dari performa model yang dibuat sebagai berikut:

Tabel 4. Performa Model

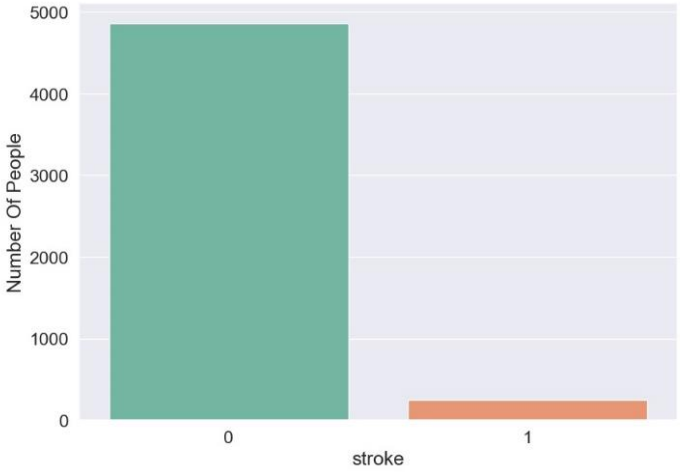
No.	Split	Akurasi
1	50:50	76%
2	60:40	82%
3	70:30	92%
4	80:20	95%
5	90:10	93%

Berdasarkan Tabel 2 pada proses training, dilakukan perubahan komposisi split dataset dan hasil akurasi terbaik didapat pada komposisi split 80:20. Pada evaluasi dan pengujian, akan dicari nilai precision dan recall. Nilai tersebut dapat kita lihat pada gambar berikut:

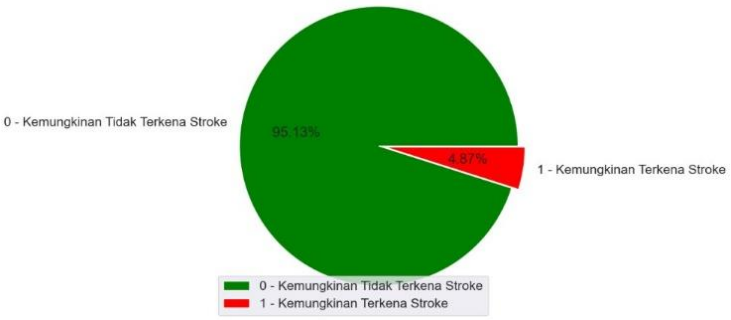
```
Akurasi: 0.95
Laporan Klasifikasi:
              precision    recall  f1-score   support

           0       0.95      1.00      0.97       929
           1       0.43      0.06      0.10        53

    accuracy                           0.95       982
   macro avg       0.69      0.53      0.54       982
weighted avg       0.92      0.95      0.92       982
```

Gambar 4. Nilai Akurasi, Precision, recall, dan f1-score

Kita juga bisa melihat hasil prediksi dalam bentuk grafik. Disini kita melihat dalam dua bentuk grafk, yaitu grafik batang dan grafik lingkaran. Untuk lebih jelasnya dapa dilihat pada gambar dibawah ini:



Gambar 5. Grafik Batang



Gambar 6. Grafik Lingkaran

IV. Kesimpulan

Setelah melakukan penelitian tentang Prediksi Penyakit Stroke, dapatditarik kesimpulan bahwa:

1. Penelitian ini menerapkan algoritma Support Vector Machine (SVM) dalam melakukan prediksi terhadap penyakit stroke.
2. Akurasi yang dihasilkan oleh metode Support Vector Machine (SVM) dalam melakukan prediksi sebesar 95% dan dikatakan sangat baik.

V. Daftar Pustaka

[2] Intan, S. F. Permana, Inggih, S. F. N. Afdal, M. Muttakin, Fitriani, “Perbandingan Algoritma KNN , NBC , dan SVM : Analisis Sentimen Masyarakat Terhadap Perparkiran di Kota Pekanbaru”2023.

[3] Shedriko, “STRING (Satuan Tulisan Riset dan Inovasi Teknologi) perbandingan algoritma svm dan knn dalam mengklasifikasi kelulusan mahasiswa pada suatu mata kuliah” 2021, volume 6, 115-122.

[4] Tasari, A. D. Tarigan, D. Nia, E. Purba, D. Br Saputra, Kana, “Perbandingan Algoritma Support Vector Machine dan KNN dalam Memprediksi Struktur Sekunder Protein”, 2022, volume 9, issue 2, 172-179.

[1] Nurmalasari, Dewi, Hermanto, T. I. Nugroho, I. Ma’ruf “Perbandingan Algoritma SVM , KNN dan NBC Terhadap Analisis Sentimen Aplikasi Loan Service”, 2023, volume 7, 1521-1530.

[5] V. Jayadi, B. Handhayani, T. D. Lauro, Manatap “Perbandingan Knn Dan Svm Untuk Klasifikasi Kualitas Udara Di Jakarta”, 2023, volume 11, issue 2.

[6] S. U. Alifa, R. F. Rizkika, P. Rozikin, Chaerur “Perbandingan Performa Algoritma KNN dan SVM dalam Klasifikasi Kelayakan Air Minum”, 2023, volume 7, issue 2, 141-146.

[7] Ishlah, A. W. Sudarno, S. Kartikasari, Puspita, “Implementasi Gridsearchcv Pada Support Vector Regression (Svr) Untuk Peramalan Harga Saham”, 2023, volume 12, issue 2, 276-286.

[8] Arifin, O. Sasongko, T. Bayu, “Analisa perbandingan tingkat performansi metode support vector machine dan naïve bayes classifier”, 2018, volume 8, issue 1, 67-72.

[9] Munawarah, R. Soesanto, O. R. Faisal, M. Yani, “Penerapan Metode Support Vector Machine Pada Diagnosa Hepatitis”, 2016, volume 4 No.01, issue 1, 1-11.

[10] S. Sripamuji, A. D. Ramadhanti, I. R. Amalia, R. Saputra, J. Prihatnowo, Bagas, “Penerapan Algoritma Support Vector Machine Dan Multi-Layer Perceptron Pada Klasifikasi Topik Berita”, 2022, volume 11, issue 2, 84-91.

[11] Arifin, N. Enri, U. Sulistiyowati, Nina, “Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification”, 2021, volume 6, issue 2.

[12] Agustina, W. Furqon, M. T. Rahayudi, Bayu, “Implementasi Metode Support Vector Machine (SVM) Untuk Klasifikasi Rumah Layak Huni (Studi Kasus: Desa Kidal Kecamatan Tumpang Kabupaten Malang)”, 2018, volume 2, issue 10, 3366-3372.

[13] Handayanto, A. Latifa, K. Saputro, N. D. Waliansyah, R. Robi, “Analisis dan Penerapan Algoritma Support Vector Machine (SVM) dalam Data Mining untuk Menunjang Strategi Promosi”, 2019, volume 7, issue 2.

[14] Lukman, “Penerapan Algoritma Support Vector Machine (SVM) dalam Pemilihan Beasiswa:Studi Kasus SMK YAPIMDA”, 2016, volume 1, issue 1, 49-57.

[15] Nafi'iyah, Nur, “Algoritma SVM untuk Memprediksi Pengunjung Wisata Musium di Jakarta”, 2020, volume 1, issue 1, 33-41.

[16] Abdusyukur, Fatwa, “Penerapan Algoritma Support Vector Machine (Svm) Untuk Klasifikasi Pencemaran Nama Baik Di Media Sosial Twitter”, 2012, volume 12, issue 1, 73-82.