https://bufnets.tech
https://doi.org/10.59688/bufnets

BULLETIN OF NETWORK ENGINEER AND INFORMATICS
E-ISSN 2986-8017 | P-ISSN 2987-4858
Vol. 1 No 2 – October 2023

PREDIKSI PENYAKIT STROKE MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM)

STROKE PREDICTION USING A SUPPORT VECTOR MACHINE (SVM)

Patmawati

PJJ Magister Teknik Informatika, Universitas Amikom Yogyakarta
Jl. Padjajaran, Ring Road Utara, Sleman, Daerah Istimewa Yogyakarta 55283, Indonesia

e-mail: patmawati@students.amikom.ac.id

Received : 27 February 2023	Accepted : 1 March 2023	Published : 20 April 2023

Abstract

Based on data from the Indonesian Ministry of Health, there has been an increase in the number of stroke cases by 3.9% from 2013 to 2018. Nationally, the number of stroke cases often occurs in groups that have an age range between 55-64 years and the least occur in the 15-24 age range. A stroke (Cerebrovascular Accident) is a condition where blood flow to the brain is suddenly interrupted or reduced. This can be caused by a blockage or rupture of blood vessels so that cells in the brain area do not get a blood supply that is full of nutrients and oxygen. Early detection is needed which aims to reduce the number of potential deaths from stroke. Stroke prediction is still a challenge in the field of medicine, one of the reasons is the volume of data on medical data which has high heterogeneity and complexity. Machine learning techniques are data analysis models that can be used to predict stroke. Various machine learning models have been proposed by previous researchers, one of which is the Support Vector Machine. This study tries to re-implement the SVM algorithm to get better performance results than previous studies. In this study, the accuracy value was 100% and ROC-AUC values were 100%. Further studies need to be carried out regarding the results obtained to reach 100%.

Keywords: *Support Vector Machine(SVM), Machine Learning, Stroke*

Abstrak

Berdasarkan data dari Kementerian Kesehatan Indonesia, telah terjadi peningkatan jumlah pada kasus penyakit stroke sebesar 3.9% mulai dari tahun 2013 sampai dengan tahun 2018. Secara nasional, jumlah kasus stroke sering terjadi pada kelompok yang memiliki rentang umur antara 55-64 tahun dan paling sedikit terjadi pada rentang umur 15-24. Stroke atau (Cerebrovascular Accidents) merupakan sebuah keadaan dimana aliran darah ke otak mengalami gangguan mendadak atau berkurang. Hal tersebut dapat disebabkan oleh penyumbatan atau pecah pembuluh darah, sehingga sel-sel pada area otak tidak mendapatkan pasokan darah yang nutrisi dan oksigen. Diperlukan deteksi dini yang bertujuan untuk mengurangi jumlah potensi kematian akibat stroke. Prediksi stroke masih menjadi tantang dalam bidang kedokteran, salah satu penyebabnya adalah volume data pada data medis yang memiliki heterogenitas dan kompleksitas yang tinggi. Teknik machine learning merupakan model analisis data yang dapat digunakan untuk memprediksi penyakit stroke. Berbagai model pembelajaran machine learning telah diusulkan oleh peneliti-peneliti sebelumnya, salah satunya Support Vector Machine. Penelitian ini mencoba menerapkan kembali algoritma SVM dengan mendapatkan hasil kinerja lebih baik dari penelitian sebelumnya. Dalam penelitian ini didapatkan nilai accuracy sebesar 100% dan nilai ROC-AUC sebesar 100%. Perlu dilakukan pengkajian lagi terkait hasil yang didapatkan hingga mencapai 100%.

Kata Kunci: *Support Vector Machine(SVM), Machine Learning, Stroke*

1. PENDAHULUAN

Menurut WHO, stroke adalah penyakit penyebab kematian peringkat kedua setelah penyakit iskemik. Terdapat lima belas juta penderita stroke di seluruh dunia setiap tahunnya. Dan setiap 4-5 menit, terdapat penderita sroke yang meninggal di seluruh dunia [1]. Berdasarkan data dari Kementerian Kesehatan Indonesia, adanya peningkatan jumlah terhadap kasus stroke dengan presentasi sebesar 3,9% dari tahun 2013 menuju ke 2018. Secara nasional, jumlah kasus stroke sering terjadi pada pada kelompok yang memiliki rentang umur antara 55-64 tahun dan kemudian paling sedikit terjadi pada kelompok yang berumur 15-24. [2]

Stroke (Cerebrovascular Accidents) merupakan sebuah keadaan dimana aliran darah ke otak mengalami gangguan mendadak atau berkurang. Hal tersebut dapat disebabkan oleh penyumbatan atau pecah pembuluh darah, sehingga sel-sel pada area otak tidak mendapatkan pasokan darah yang nutrisi dan oksigen [3].

Diperlukan deteksi dini dan penanganan yang tepat guna meminimalkan kerusakan lebih lanjut pada bagian otak serta komplikasi yang terjadi pada bagian tubuh lainnya [4]. Deteksi dini dapat dilakukan melalui perancangan sebuah model pendekatan yang dapat digunakan untuk mengidentifikasi serta melakukan prediksi terhadap risiko stroke [5] dengan mempertimbangkan beberapa faktor yang merupakan faktor-faktor yang berisiko umum dan memiliki jangka panjang antara lain hyperglikemia, hypertens, hyperlipimedia, tekanan tinggi serta stress karena emosi [6]. Prediksi stroke memiliki tujuan untuk mengurangi jumlah kematian yang diakibatkan oleh penyakit stroke.

Prediksi stroke masih menjadi tantang dalam bidang kedokteran [7]. Salah satu penyebabnya adalah volume data pada data medis yang memiliki heterogenitas dan kompleksitas yang tinggi [8]. Teknik machine learning merupakan model analisis data yang dapat digunakan untuk memprediksi penyakit stroke [7].

Berbagai model pembelajaran machine learning telah diusulkan oleh peneliti-peneliti sebelumnya, antara lain Decision Tree [9], Support Vector Machine, Naïve Bayes [10] ,[11], Random Forest, Logistic Regression [12], [13].

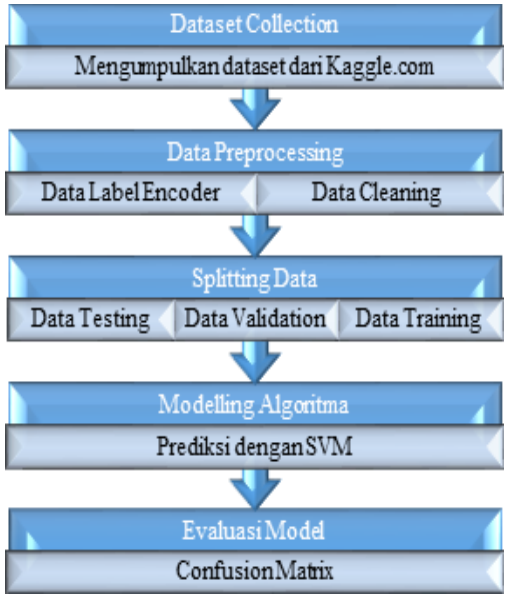
Penelitian sebelumnya [11] melakukan analisis prediksi stroke dengan menggunakan metode machine learning. Dihasilkan nilai akurasi untuk Logistic Regresion 78%, Decisition Tree Classification 66%, Random Forest Classification 73%, KNN 80%, Support Vector Machine 80% dan Naïve Bayes 82%.

Selanjutnya penelitian [14] yang juga melakukan analisis prediksi stroke dengan menggunakan model machine learning yang sama. Didapatkan akurasi algoritma Decistion Tree 74.31%. Random Forest sebesar 74.53% dan multilayer parceptron sebesar 75.02%. Berikutnya penelitian [15] melakukan perbandingan kinerja algoritma machine learning KNN, Naïve bayes, SVM dan Decision Tree. Diperoleh nilai akurasi Naïve bayes sebesar 93.93%, KNN 91.19%, Support Vector Machine 93.15% dan Decision Tree 90.90%.

Penelitian ini mencoba menerapkan kembali algoritma SVM dengan mendapatkan hasil kinerja lebih baik dari penelitian sebelumnya. Sehingga penelitian ini, dapat memprediksi penyakit stroke dengan performa yang lebih baik lagi menggunakan algoritma machine learning.

2. METODE PENELITIAN

Tahapan dalam penelitian ini dilakukan sesuai dengan Gambar 1. Tahapan penelitian.



Gambar 1. Tahapan Penelitian

Tahapan penelitian dimulai dari proses pengumpulan dataset. Selanjutnya adalah proses preprocessing data yang dilakukan dengan cara data cleaning serta dan kemudian label encoding. Setelah itu, data yang telah diprocessing kemudian dibagi menjadi data latih (data train) dan data uji (data test). Berikutnya adalah membuat model machine learning untuk melakukan prediksi stroke, yaitu Support Vector Machine,Terakhir adalah evaluate dari model

yang telah dibuat dengan menggunakan confusion matrix.

3. HASIL DAN PEMBAHASAN

Dataset Description

Pada penelitian ini, dataset yang digunakan berasal dari kaggle.com [16], dimana dataset tersebut terdiri atas 12 field atau attribute serta 5110 baris atau entri data pasien. Dataset tersebut telah digunakan dalam penelitian-penelitian sebelumnya, diantaranya [17]. Dalam dataset tersebut terdapat 11 feature yang dijadikan sebagai parameter utama dalam memprediksi kemungkinan pasien mengindap stroke. Kemudian 11 feature tersebut dibagi menjadi tiga faktor, yaitu gaya hidup, resiko medis dan faktor-faktor yang0tidak bisa dikendalikan [18]. Faktor-faktor gaya0hidup merupakan faktor yang terdiri atas kebiasaan pada setiap individu berdasarkan keinginan dan kemampuan ekonomi. Contohnya seperti aktivitas makan, aktivitas minum, aktivitas-aktivitas fisik dan juga merokok. Faktor untuk resiko medis merupakan variable yang memiliki hubungan dengan peningkatan resiko terhadap penyakit stroke, seperti level glukosa, Riwayat pasien terhadap penyakit jantung dan tekanan darah. Dan terakhir faktor tidak dapat dikendalikan yaitu faktor secara telah melekat atau tidak dapat diubah pada pasien, seperti usia dan jenis kelamin.[8][12].

Penjelasan attribute dan dekripsi dataset dapat dilihat pada Tabel 1. Deskripsi Dataset dibawah ini

Table 1. Deskripsi Dataset

Attribute	Description	Tipe Data
ID	ID atau nomor data pasien	Data numerik
Gender	Jenis Kelamin pada pasien	Data kategorik
Age	Usia si pasien	Data numerik
Hypertension	Berkategori, yaitu 0 artinya tidak mengalami hipertensi 1 artinya mengalami hipertensi	Data numerik
Heart disease	0 artinya tidak memiliki riwayat penyakit jantung 1 artinya memiliki riwayat penyakit jantung	Data numerik
Marital status	Status perkawinan si pasien	Data kategorik
Work type	Jenis pekerjaan si pasien	Data kategorik
Residence area	Wilayah tempat yang ditinggali si pasien	Data kategorik
Avg-glukose	Nilai rata-rata tingkat level glukosa dalam darah si pasien yang diukur	Data numerik
BMI	Body Mass Index si pasien	Data numerik
Smoking Status	Status merokok pasien	Data kategorik
Stroke status	Kesimpulan, apakah 0 tidak mengalami stroke 1 mengalami stroke	Data numerik

Data Processing

Pada tahap ini dilakukan preprocessing agar performa algoritma dalam memprediksi dapat bekerja dengan baik dan menghasilkan accuracy yang tinggi, Yang dilakukan pada tahap ini diantaranya yaitu data cleaning terhadap dataset yang mengalami missing values. Selain itu, dilakukan juga pengkodean label menggunakan fungsi label encoding untuk mengkodekan data kategori menjadi data numerik. Selain itu juga mengubah data bertipe string menjadi integer/angka.

Splitting Data

Sebelum data dimodelkan menggunakan algoritma machine learning, terlebih dahulu data dibagi atau displitting menjadi data training dan data testing dengan rasio perbandingan 80:20.

Modelling Algorithm

Modelling algoritma machine learning yang digunakan untuk penelitian ini yaitu Support Vector Machine dan ANN (Artifical Neural Network). SVM Merupakan algoritma untuk membuat sebuah garis pemisah (hyperlane) ideal pada ruang komponen dimensi yang lebih tinggi agar dapat memetakan informasi dengan resiko yang minim [19].

Hyperlane atau pemisah dibangun menggunakan support vector, data yang lebih dekat terhadap hyperlane. Data tersebut terletak pada batas irisan kelas pertama yang disebut dengan support vector + (positif), dan kemudia kelas kedua support vector – (negative). Jarak antara support vector disebut margin, dimana maximum margin merupakan hyperlane yang baik. Tujuan dari SVM yaitu mencari hyperlance optimal dalam membagi data agar benar-benar terpisah menjadi dua bagian.

Evaluasi Model

Modelling Matriks evaluasi model dilakukan untuk mengukur kinerja suatu metode klasifikasi sehingga dapat diketahui seberapa baik sistem dalam melakukan klasifikasi data. Pengujian dilakukan menggunakan matrics performance dan ROC-AUC.

Metrics performance terdiri atas parameter *accuracy* , *precision* , *recall* , *f1-score* . Accuracy merupakan perbandingan antara data sampel yang diprediksi benar dengan jumlah total data sampel [20]. Berikut adalah rumus untuk mencari nilai accuracy

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Precision merupakan perbandingan antara sampel berkategori positif benar yang dibandingkan dengan total data sampel yang diprediksikan positif

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall adalah nilai perbandingan data sampel yang diprediksi bernilai positif dan memiliki kategori positif benar

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-score adalah nilai rata-rata antara nilai yang didapatkan dari precision dan nilai dari recall

$$\text{F1-score} = \frac{2x(Precison\ X\ Recall\ )}{Precision + Recall} \quad (4)$$

ROC-AUC adalah grafik yang menggambarkan hubungan antara dua dimensi yaitu antara parameter true positive rate terhadap parameter false positive rate [8].

3.5 Evaluasi Model

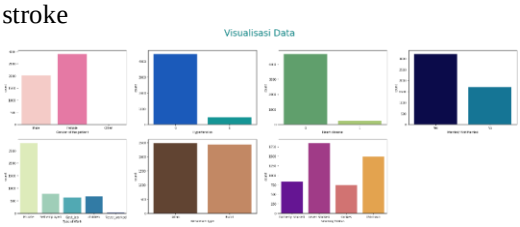
Berikut adalah tampilan 5 data dari dataset stroke yang didapatkan pada situs kaggle.com yang di tunjukan pada Gambar 2.

	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
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never 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

Gambar 2. *Dataset Stroke*

Selanjutnya dilakaukan dilakukan preprocing data. Ditemukan terdapat data yang bernilai null sebanyak pada field bmi. Data bernilai null tersbut kemudian dihapus. Selanjutnya dilakanan visualisasi terhadap field atau attribute yang digunakan untuk klasifikasi.

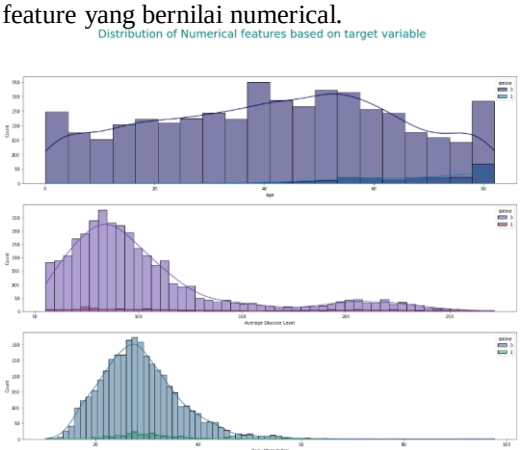
Berikut adalah Gambar 3. Visualisasi parameter stroke



Gambar 3. Visualisasi Parameter *Stroke*

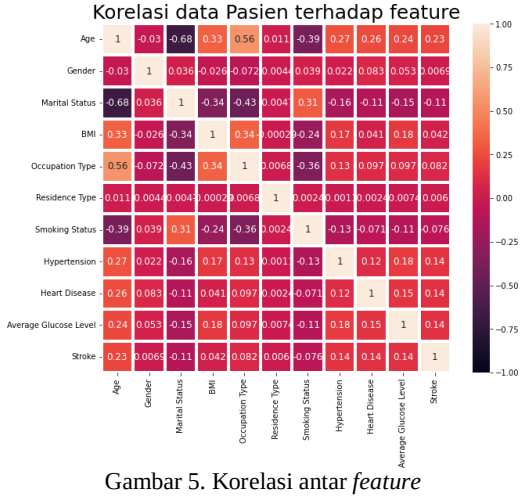
Dari hasil visualisasi diatas, ditemukan data bias yang terjadi pada parameter gender. Terdapat data gender yang bernilai other. Data bias tersebut kemudian dihapus, sehingga total data yang akan di modelling sebanyak 4908.

Berikut adalah Gambar 4, distribution feature yang bernilai numerical.



Gambar 4. *Distribution Numerical Feature*

Dari distribution angka diatas, dapat disimpulkan bahwa penyakit strok rentang terjadi pada umur 40 tahun keatas dengan rata-rata glukosa level sebesar 100 dan Index max body berada sekitar 30. Selanjutnya, Gambar 5 adalah korelasi antar feature pada dataset

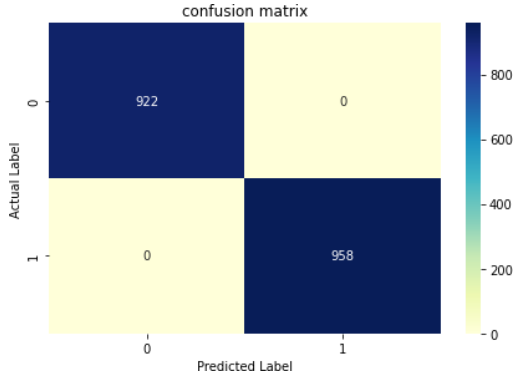


Gambar 5. Korelasi antar *feature*

Dari gambar diatas, dapat disimpulkan bahwa terdapat

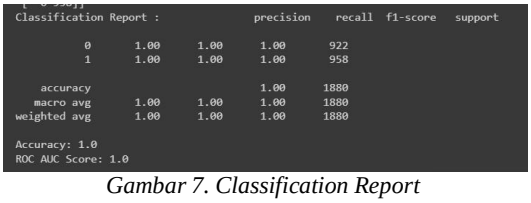
- korelasi positif kecil antara attribute *Average Glucose Level ,Age, , Heart Disease, Hypertension.*
- Selanjutnya, terdapat korelasi positif yang kecil juga antara Age and Stroke, Heart Disease ,Hypertension, BMI , Average Glucose Level.
- Terdapat juga korelasi positif kecil antara Smoking Status and Marital Status, Occupation Type and BMI.
- Terakhir, terdapat korelasi positif medium antar Age and Occupation Type.
- Serta korelasi negative medium antara Age and Marital Status.

Hasil evaluate dari model yang dibuat dengan menggunakan confusion matrix seperti pada Gambar 6 berikut ini:



Gambar 6. *Confusion Matrix Predicted*

Dari Gambar 6 diatas, diketahui bahwa sebanyak 922 data yang diprediksi benar pada label 0. Dan sebanyak 958 data yang diprediksi benar pada label 1. Berikut Gambar 7, Classification Report yang menampilkan data precision, recall dan f1-score. Nilai akurasi yang didapatkan dalam penelitian mencapai hingga 100%.



Classification Report :	precision	recall	f1-score	support
0	1.00	1.00	1.00	922
1	1.00	1.00	1.00	958
accuracy			1.00	1880
macro avg	1.00	1.00	1.00	1880
weighted avg	1.00	1.00	1.00	1880

Accuracy: 1.0
ROC AUC Score: 1.0

Gambar 7. Classification Report

4. KESIMPULAN

Penelitan yang dilakukan diatas berjalan dengan baik. Dengan menggunakan dataset yang sama serta metode algoritma yang sama yaitu Suppot vector machine, Nilai akurasi pada penelitian ini lebih baik dari yang didapatkan pada penelitian sebelumnya. Nilai akurasi yang didapatkan sebesar 100%. Bukan tidak mungkin suatu accuracy mendapatkan nilai 100%. Bisa jadi dikarenakan beberap hal terkait imbalanced data yang menyebabkan terjadi accuracy yang sempurna. Saran penulis untuk penelitian kedepannya, agar ditambahkan lagi process imbalanced, agar didapatkan nilai akurasi yang tertinggi.

PERNYATAAN PENGHARGAAN

Terimakasih kepada para penulis yang di kutip pada artikel ini, secara tidak langsung tulisan yang di kutip sangat membantu di dalam menyelesaikan artikel ini, semoga artikel ini dapat bermanfaat bagi yang membacanya.

DAFTAR PUSTAKA

[1] Campbell, B. C., &; Khatri, P. (2020). Stroke. The Lancet, 396(10244), 129-142. doi:10.1016/s0140-6736(20)31179-x.

[2] Kemenkes, R. I. (2019). Infodatin: Stroke Don't Be The One. Kementerian Kesehatan Republik Indonesia.

[3] Qiao, M., Jiang, C., Zhu, Y., & Li, G. (2016). Research on design method and electromagnetic vibration of six-phase fractional-slot concentrated-winding PM motor suitable for ship propulsion. IEEE Access, 4, 8535-8543.

[4] Kissane, J., Neutze, J. A., & Singh, H. (Eds.). (2020). Radiology fundamentals: Introduction to imaging & technology. Springer Nature.

[5] Janowski, T., & Mohanty, H. (2010). Distributed Computing and Internet Technology. Springer Berlin Heidelberg.

[6] Guzik, A., & Bushnell, C. (2017). Stroke epidemiology and risk factor management. CONTINUUM: Lifelong Learning in Neurology, 23(1), 15-39.

[7] Bhuyan, M. K. (2019). Computer vision and image processing: Fundamentals and applications. CRC Press.

[8] Faisal, A., & Subekti, A. (2021). Deep Neural Network untuk Prediksi Stroke. JEPIN (Jurnal Edukasi dan Penelitian Informatika), 7(3), 443-449.

[9] Chamchong, R., & Wong, K. W. (Eds.). (2019). Multi-disciplinary Trends in Artificial Intelligence: 13th International Conference, MIWAI 2019, Kuala Lumpur, Malaysia, November 17–19, 2019, Proceedings (Vol. 11909). Springer Nature.

[10] Chantamit-O-Pas, P., & Goyal, M. (2017). Prediction of stroke using deep learning model. In Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part V 24 (pp. 774-781). Springer International Publishing.

[11] Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6).
[12] Zuama, R. A., Rahmatullah, S., & Yuliani, Y. (2022). Analisis Performa Algoritma Machine Learning pada Prediksi Penyakit Cerebrovascular Accidents. JURNAL MEDIA INFORMATIKA BUDIDARMA, 6(1), 531-534.
[13] Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Al Mamun, M. S., & Kaiser, M. S. (2020, November). Performance analysis of machine learning approaches in stroke prediction. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1464-1469). IEEE.
[14] Sakib, S., Yasmin, N., Tasawar, I. K., Aziz, A., Siddique, M. A. B., & Khan, M. M. R. (2021, September). Performance Analysis of Machine Learning Approaches in Diabetes Prediction. In 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 1-6). IEEE.
[15] Cahyani, D. E. (2022). PENERAPAN MACHINE LEARNING UNTUK PREDIKSI PENYAKIT STROKE. Jurnal Kajian Matematika dan Aplikasinya (JKMA), 3(1), 15-22.
[16] Fedesoriano. (2021, January 26). Stroke prediction dataset. Retrieved March 1, 2023, from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.
[17] Abd Mizwar, A. R., Sunyoto, A., & Arief, M. R. (2022). Stroke Prediction using Machine Learning Method with Extreme Gradient Boosting Algorithm. MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 21(3), 595-606.
[18] Wu, Y., Chan, E., Melton, J. R., & Verseghy, D. L. (2017). A map of global peatland distribution created using machine learning for use in terrestrial ecosystem and earth system models. Geoscientific Model Development Discussions, 1-21.
[19] Kumar, I., Virmani, J., Bhadauria, H. S., & Panda, M. K. (2018). Classification of breast density patterns using PNN, NFC, and SVM classifiers. In Soft Computing Based Medical Image Analysis (pp. 223-243). Academic Press.
[20] Tharwat, A. (2021). Classification assessment methods. Applied computing and informatics, 17(1), 168-192.