Perbandingan Performa Algoritma KNN dan SVM dalam Klasifikasi Kelayakan Air Minum

Sopiatul Ulum[1], Rizal Fahmi Alifa[2], Putri Rizkika[3], Chaerur Rozikin[4]
[1,2,3,4]Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
E-mail: [1]*2010631170033@student.unsika.ac.id*, [2]2010631170116@student.unsika.ac.id,
[3]2010631170108@student.unsika.ac.id, [4]chaerur.rozikin@staff.unsika.ac.id
Coresponden Author: *2010631170033@student.unsika.ac.id*

Diterima Redaksi: 12 Juni 2023 Revisi Akhir: 17 Juli 2023 Diterbitkan Online: 31 Juli 2023

Abstrak – Air menjadi kebutuhan mendasar bagi kelangsungan makhluk hidup dan pembangunan. Saat ini, kesadaran masyarakat terhadap pola konsumsi air yang berkualitas dan bermutu semakin tinggi sehingga diperlukan penelitian terhadap kelayakan air. Dalam penelitian air tersebut menggunakan metode klasifikasi objek. Pada penelitian ini membahas perbandingan antara 2 metode pembelajaran mesin yaitu *K-Nearest Neighbors* (K-NN) dengan *Support Vector Machine* (SVM) berdasarkan parameter dalam kelayakan air minum yaitu: PH, *hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethane, turbidity* dan *probability*. Penelitian ini menghasilkan tingkat akurasi algoritma *K-Nearest Neighbors* (K-NN) sebesar 65,341% dan algoritma *Support Vector Machine* (SVM) menghasilkan akurasi sebesar 69,764%. Dari hasil tersebut, bisa disimpulkan bahwa algoritma *Support Vector Machine* (SVM) memiliki akurasi lebih tinggi *daripada* algoritma *K-Nearest Neighbors* (K-NN).

Kata Kunci — Kelayakan Air, Pembelajaran Mesin, *K-Nearest Neighbors, Support Vector Machine*

Abstract – *Water is a basic need for the survival of living things and development. At present, public awareness of quality and quality water consumption patterns is getting higher, so research is needed on the feasibility of water. In the water research using object classification method. This study discusses a comparison between 2 Machine Learning methods, namely K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM) based on predetermined parameters. This research produces an accuracy rate of the K-Nearest Neighbors (K-NN) algorithm of 65.341% and the Support Vector Machine (SVM) algorithm produces an accuracy of 69.764%. From these results, it can be concluded that the Support Vector Machine (SVM) algorithm has higher accuracy than the K-Nearest Neighbors (K-NN) algorithm.*

Keywords — *Water Potability, Machine Learning, K-Nearest Neighbors, Support Vector Machine.*

1. PENDAHULUAN

Air menjadi kebutuhan mendasar bagi kelangsungan makhluk hidup dan pembangunan. *World Health Organization* atau WHO yang bergerak dibidang kesehatan dunia mengungkapkan bahwasannya air bersih ialah air yang bisa dimanfaatkan oleh manusia untuk memenuhi keperluan domestic, yaitu: mulai dari konsumsi, air minum, serta persiapan makanan [1]. Kebutuhan air menjadi semakin meningkat dengan berjalannya waktu begitu cepat dan bertumbuhnya zaman. Pada tahun 2002, badan dunia yang bergerak pada bidang pendidikan, ilmu pengetahuan, kebudayaan atau UNESCO (*The United Nations Educational, Scientific and Cultural Organization*) telah menetapkan bahwa sebesar 60 liter/orang/hari menjadi hak dasar manusia atas air yang digunakan dan dibagi dalam standar kebutuhan air berdasarkan dengan wilayah oleh Direktorat Jenderal Cipta Karya Departemen Pekerjaan Umum [2]. Kehidupan yang layak dan bermartabat dapat diwujudkan dengan pemenuhan hak asasi manusia atas air & sanitasi [3]. Selain itu, pemenuhan hak dasar atas air & sanitasi juga diatur dalam target *Sustainable Development Goals* (SDGs) yang menanggung ketersediaan dan pengelolaan air bersih dan sanitasi yang berkesinambungan untuk semua [4]. Oleh karena itu, salah satu dari kategori air yang amat penting ialah air bersih.

Salah satu sumber daya air bersih yang sering dipergunakan oleh manusia ialah mata air. Menurut peneliti Hendrayana (1994) mengungkapkan bahwasannya mata air ialah sebuah tempat atau sumber adanya air tanah dapat merembes atau mengalir sampai mencapai bagian atas dari permukaan tanah dengan alami, kemudian akan melalui atau mengikuti alur sungai, sehingga seringkali menjadi sumber aliran air pada sungai [5]. Air sungai yang tercemar oleh limbah, baik itu limbah rumah tangga atau limbah industri dapat mempengaruhi kualitas air bersih.

Penurunan kualitas air bersih ialah hal yang perlu diperhatikan karena bisa berdampak pada kesehatan, pertumbuhan penduduk, ataupun dalam konteks pembangunan perkotaan [6]. Kustanto berpendapat bahwa semakin tingginya aktivitas industri dan kegiatan manusia dalam sehari-harinya, maka semakin meningkat pencemaran yang dapat mempengaruhi ekosistem perairan & sanitasi yang setimpal dalam akses ke air minum yang aman untuk dikonsumsi dan digunakan pada manusia[7].

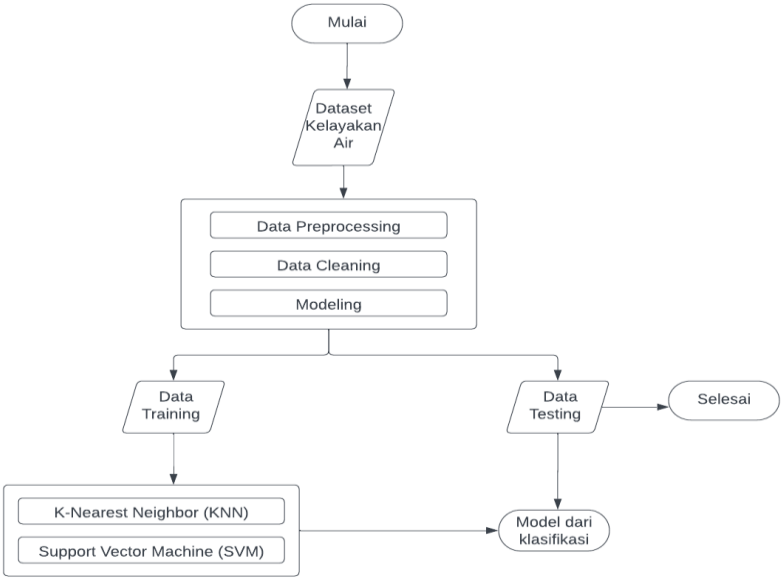
Saat ini, kesadaran masyarakat terhadap pola konsumsi air yang berkualitas dan bermutu semakin tinggi sehingga diperlukan pengujian atau pengukuran terhadap kualitas air dimana kita dapat mengetahui kondisi air yang akan dikonsumsi atau gunakan. Selain itu, Menilai kualitas air untuk tujuan penggunaan air yang berbeda, seperti penggunaan rumah tangga, irigasi, konservasi dan penggunaan industri, merupakan strategi penting untuk keamanan pangan serta kesehatan manusia.[8]. Kualitas air dapat diketahui melalui kandungan air yang ada di dalamnya. Terdapat parameter pemdukung kualitas air yang harus diuji guna menentukan kelayakan air minum, antara lain kandungan pH, *hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes,* dan *turbidity*. Beberapa metode IKA (Indeks Kualitas Air) digunakan dalam menentukan status dari kualitas air adalah metode PI (*Pollution Index*), metode CCME (*Canadian Council of Ministers of the Environment*), dan metode STORET [9]. Oleh karena itu, diperlukan adanya sebuah penelitian perbandingan dengan 2 metode dari cabang ilmu *Machine Learning* yaitu K-NN atau *K-Nearest Neighbors* dengan metode SVM atau *Support Vector Machines* untuk dapat menentukan kelayakan air.

2. METODE PENELITIAN

Pada penelitian yang dilakukan kami menggunakan *dataset* dari kaggle. Jumlah dari *dataset* ialah 3276 serta mempunyai 10 kolom yaitu : PH, *Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity, Potability* [10]. Berikut Tabel 1 merupakan parameter dan deskripsi dari *dataset* kelayakan air.

Tabel 1. Parameter Deskripsi Dari *Dataset* Kelayakan Air

Parameter	Deskripsi
PH	pH ialah indikator kondisi status air asam atau basa. pH maksimum yang diizinkan dari 6,5 hingga 8,5.
Hardness	*Hardness* atau kesadahan air merupakan debit air dalam mengendapkan sabun yang ditimbulkan oleh kalsium serta magnesium.
Solids	*Solids* ukuran untuk mengetahui jumlah total padatan yang terlarut dalam air.
Chloramines	*Chloramines* ialah kandungan kimia yang terwujud dari reaksi antara klorin dan amonia atau senyawa nitrogen lainnya dalam air. Senyawa ini sering digunakan sebagai desinfektan alternatif untuk air minum karena lebih stabil dan kurang berbau serta berpotensi lebih sedikit membentuk senyawa berbahaya dibandingkan klorin.
Sulfate	*Sulfate* adalah zat alami yang ditemukan dalam mineral, tanah, dan batuan.
Conductivity	*Conductivity* atau konduktivitas ialah keunggulan suatu zat atau bahan dalam menghantarkan arus listrik.
Organic carbon	*Organic carbon* atau karbon organik adalah unsur karbon yang terkandung dalam senyawa organik seperti bahan-bahan organik yang terdapat dalam tanah, air, dan udara.
Trihalomethane	*Trihalomethanes* ialah bahan kimia yang bisa didapat dalam air yang diolah dengan klorin.
Turbidity	*Turbidity* merupakan tingkat kekeruhan pada di air.
Probability	Indikator air yang layak digunakan untuk konsumsi manusia.



Gambar 1. Metode Pengerjaan

Penelitian ini menggunakan Spyder. Spyder adalah sektor pengembangan yang terintegrasi atau *Integrated Development Environment* (IDE) untuk pemrograman *Python* yang dirancang khusus untuk data science [12]. Pada metode pengerjaan penelitian ini ditunjukkan pada gambar 1 Pertama kita ke mencari *dataset* yang digunakan, disini kami mengambil *dataset* dari kaggle. kaggle adalah situs atau *platform* yang mengadakan kompetisi pada bidang *data science* [13]. Kami menggunakan *dataset water potability* atau *dataset* kualitas air.

Selanjutnya tahap *data cleaning*, *data cleaning* merupakan pembersihan data termasuk kedalam tahap awal dari *data preprocessing*, yaitu tugasnya untuk menyeleksi data dan membuang data yang berpotensi mengurangi akurasi dan kualitas dari hasil proses [14]. Pada tahap *data cleaning* kita melakukan beberapa langkah yaitu : mengecek apakah terdapat *missing value*, jika terdapat maka kita hapus si *missing value* tersebut atau diganti menggunakan nilai rata-rata atau mean [15]. disini kami menggunakan fillna().mean() untuk mengganti *missing value* pada *dataset water potability*.

Tahap *modeling*, yaitu penerapan dari knn dan svm terhadap *data training* dimana untuk dicarinya model atau pola pengetahuannya. pada tahap *modeling* kami melakukan pengecekan pada variabel target nya yaitu "potability", lalu kami melakukan distribusi pada fitur numerik ke kelas target setelah itu kami melakukan distribusi numerik tersebut ke dalam bentuk histogram, kemudian kita lakukan normalisasi untuk menghindari bias dalam analisis model *machine learning* ini [16].

Tahap berikutnya adalah melakukan *data training* serta *data testing*. Disini kita akan melakukan *data training* dan *data testing* untuk mengetahui model dari d*ata training* dan *data testing* itu sendiri [17]. Setelah itu kami menerapkan algoritma KNN dan SVM untuk mengetahui dari *data training* dan *data testing*. Pada pengujian data *test size* dilakukan *loop* dengan ukuran 20% sampai 80%. Misalnya, pada iterasi pertama, ukuran data test adalah 20% dari data.

2.1. Algoritma *K-Nearest Neighbor* (KNN)

Algoritma *K-Nearest Neighbor* (KNN) ialah termasuk algoritma populer di pembelajaran mesin yang sering digunakan dalam berbagai bidang, termasuk pengenalan pola, klasifikasi, regresi, dan rekomendasi. Algoritma ini menggunakan pendekatan *supervised*, di mana *instance query* terbaru akan diklasifikasikan dialndaskan dengan mayoritas dari kategori yang terdapat pada KNN [11]. Dalam kasus prediksi kelayakan air minum, algoritma KNN digunakan untuk mengklasifikasikan apakah sumber air tersebut layak atau tidak untuk dikonsumsi berdasarkan data-data pengujian kualitas air tersebut. Dalam hal ini, data yang diperlukan seperti hasil pengujian kualitas air dari sumber air tersebut dapat digunakan sebagai data latih untuk algoritma KNN.

2.2. Algoritma *Support Vector Machine* (SVM)

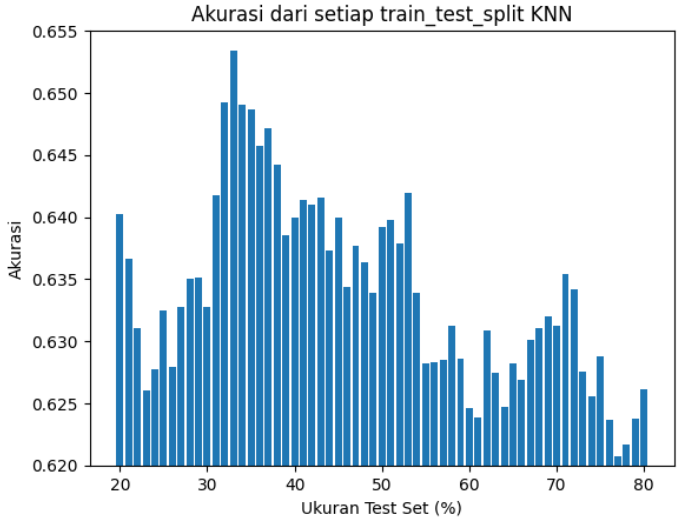
Support Vector Machine (SVM) ialah suatu sistem pembelajaran yang dilakukannya dugaan dalam bentuk fungsi linear yang terdapat fitur dengan dimensi tinggi. SVM dilatih menggunakan algoritma pembelajaran yang dilandaskan dengan teori optimasi[18]. SVM mempunyai dasar linear *classifier* pada kasus klasifikasi dimana dengan linier bisa dipisahkan, tetapi SVM sudah dibesarkan sehingga bisa bekerja pada masalah *non-linier* dengan menginput konsep dari kernel dalam ruang kerja yang berdimensi tinggi[19]. Tujuan dari algoritma SVM ini untuk mengetahui mengolah data menjadi data latih dan data uji.

2.3. *Confusion Matrix*

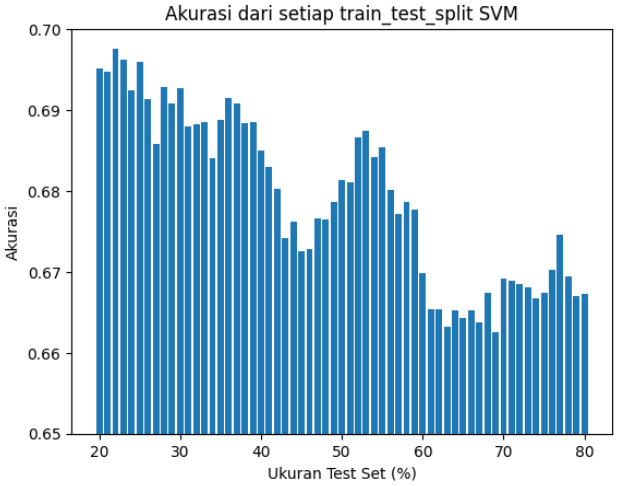
Confusion matrix ialah ukuran performa untuk menyelesaikan masalah dalam klasifikasi pembelajaran mesin dimana hasilnya berupa berisi dua kelas ataupun lebih. Matriks kebingungan juga terdiri dari empat campuran yang berbeda yaitu dari nilai prediksi serta nilai aktual. Terdapat empat yang merepresentasikan ketentuan dari proses klasifikasi pada matriks konfusi yaitu *true positive (tp), true negative (tn), false positive (fp) dan false negative (fn)* [20]. TP *(True Positive)* ialah jumlah sampel dari air minum. TN *(True Negative)* ialah jumlah sampel dari air yang benar-benar tidak bisa diminum dan diperkirakan tidak bisa diminum. FP *(False Positives)* ialah jumlah sampel dari air yang sebenarnya tidak bisa diminum tetapi diharapkan bisa diminum. FN *(False Negative)* ialah jumlah sampel dari air yang sebenarnya bisa diminum tetapi diperkirakan tidak bisa diminum.

3. HASIL DAN PEMBAHASAN

Peneliti melakukan pengujian dalam proporsi sampel uji untuk *data training* dan *data testing* yang berbeda-beda. Program melakukan *loop* pada variabel "*test_size*" dari 20% hingga 80%. Setiap iterasi *loop* akan mengubah ukuran *data testing* menjadi persentase yang berbeda-beda. Misalnya, pada iterasi pertama, ukuran *data test* adalah 20% dari data. Proporsi dengan hasil akurasi yang paling besar akan dipilih.



Gambar 2. Akurasi *Train Test Split* Algoritma KNN



Gambar 3. Akurasi *Train Test Split* Algoritma SVM

Gambar 2 dan 3 menampilkan hasil proporsi sampel uji yang berbeda untuk algoritma KNN dan SVM. Hasilnya menunjukkan bahwa algoritma KNN menghasilkan akurasi paling tinggi sebesar 65% dengan *data testing* 33% dan *data training* 65,341%. Sedangkan algoritma SVM menghasilkan akurasi paling tinggi sebesar 69,764% dengan *data testing* 22% dan *data training* 78%. Dari hasil tersebut dapat disimpulkan bahwa performa algoritma SVM lebih baik daripada algoritma KNN. Meskipun begitu, perlu diingat bahwa hasil pengujian ini hanya berdasarkan pada data sampel yang digunakan. Oleh karena itu, untuk menggeneralisasi hasil ini ke data yang lebih luas, diperlukan pengujian yang lebih komprehensif dan lebih banyak. Selain itu, perlu juga diingat bahwa pemilihan proporsi sampel uji tidak boleh dilakukan secara sembarangan, karena dapat mempengaruhi hasil pengujian dan interpretasi dari hasil pengujian. Penting dalam melakukan pertimbangan yang matang dalam pemilihan proporsi sampel uji yang tepat.

4. SIMPULAN

Air merupakan sumber utama penghidupan yang digunakan untuk menunjang berbagai kegiatan manusia. Air berasal dari sebuah mata air yang mengalir mengikuti air sungai. Air sungai yang tercemar dapat mempengaruhi kualitas dan kelayakan air bersih. Maka dari itu, diperlukan sebuah pengujian kelayakan air tersebut yang berdasarkan beberapa parameter diantaranya kandungan pH, *hardness*, *solids*, *chloramines*, *sulfate*, *conductivity*, *organic carbon*, *trihalomethanes*, *dan turbidity*. Dari parameter tersebut akan digunakan dalam memprediksi kondisi air dengan metode klasifikasi data dengan memanfaatkan 2 metode *machine learning* yang nantinya akan dibandingkan 2 metode tersebut, antara *K-Nearest Neighbors* (KNN) dan *Support Vector Machine* (SVM).

Pada penelitian ini menggunakan *dataset* dari kaggle yaitu *dataset water potability*, untuk metode pengerjaan kami melakukan *data profiling*, *data cleaning*, *modeling*, *train* dan *test* pada algoritma serta model dari algoritma yang kami gunakan yaitu KNN dan SVM. Berlandasan dari penelitian yang telah dilakukan, algoritma SVM menghasilkan performa lebih baik daripada algoritma KNN dalam klasifikasi kelayakan air minum. Hal ini dilihat dari hasil akurasi yang dihasilkan oleh masing-masing algoritma, dimana algoritma KNN memiliki tingkat akurasi sebesar sebesar 65% dengan *data testing* 33% dan *data training* 65,341%. Sedangkan pada algoritma SVM memiliki tingkat akurasi sebesar 69,764% dengan *data testing* 22% dan *data training* 78%. Oleh karena itu dalam performa perbandingan antara algoritma KNN dan algoritma SVM yang paling terbaik dalam klasifikasi kelayakan air minum adalah algoritma SVM.

5. SARAN

Berikut ialah beberapa saran dari peneliti untuk penelitian selanjutnya guna menutup kekurangan serta pengembangan dari jurnal ini:

1. Perbandingan dengan algoritma lain: Meskipun penelitian ini membandingkan performa KNN dan SVM, peneliti selanjutnya dapat melibatkan algoritma lain dalam perbandingan untuk mendapatkan hasil yang lebih komprehensif. Beberapa algoritma yang populer dalam klasifikasi data termasuk *Decision Trees*, *Random Forest*, dan *Naive Bayes*. Melibatkan algoritma-algoritma ini dapat memberikan perspektif yang lebih luas tentang performa model.

2. Perbaikan pada tahap *data preprocessing*: Untuk perbaikan pada tahap *data preprocessing*, ada beberapa saran yang dapat dicoba seperti penanganan terhadap *missing value*, normalisasi atau standarisasi, seleksi fitur, dll. Dengan melakukan *data preprocessing* yang baik, peneliti dapat meningkatkan kualitas data yang digunakan dalam penelitian dan memastikan bahwa hasil analisis dan model yang dikembangkan lebih akurat.

DAFTAR PUSTAKA

[1] Gramedia. (2021). Mengenal Ciri-cifi Air Bersih Menurut WHO yang Aman Digunakan [online]. Available: https://www.gramedia.com/literasi/ciri-ciri-air-bersih/

[2] SPARTA. (2023). Standar Kebutuhan Air Bersih Setiap Orang [online]. Available: https://www.atbbatam.com/?md=view&id=1-17070500012

[3] WHO, "The Human Right to Water and Sanitation Media brief," UN-Water Decad. Program. Advocacy Commun. Water Supply Sanit. Collab. Counc., no. April 2011, hal. 1–8, 2011, [Daring]. Tersedia pada: http://www.un.org/waterforlifedecade/pdf/human_right_to_water_and_sanitation_media_brief.pdf

[4] U. Nations, "Sustainable Development Goal (SDG)." https://sdgs.un.org/

[5] RimbaKita. (2023). Mata Air - Pengertian, Proses, Jenis, Manfaat & Pengelolaan [online]. Available: https://rimbakita.com/mata-air/

[6] G. L. Pritalia, "Analisis Komparatif Algoritma Machine Learning pada Klasifikasi Kualitas Air Layak Minum," vol. 2, 2022.

[7] A. Kustanto, "Water quality in Indonesia: The role of socioeconomic indicators," J. Ekon. Pembang., vol. 18, no. 1, pp. 47–62, Jul. 2020, doi: 10.29259/jep.v18i1.11509.

[8] C. T. Son, N. T. H. Giang, T. P. Thao, N. H. Nui, N. T. Lam, and V. H. Cong, "Assessment of Cau River water quality assessment using a combination of water quality and pollution indices," J. Water Supply Res. Technol.-Aqua, vol. 69, no. 2, pp. 160–172, Mar. 2020, doi: 10.2166/aqua.2020.122.

[9] I. G. Vidiastanta, N. Hidayat, and R. K. Dewi, "Komparasi Metode K-Nearest Neighbors (K-NN) Dengan Support Vector Machine (SVM) Untuk Klasifikasi Status Kualitas Air". [10] Kadiwal Aditya. 2021. Water Quality [online]. available: https://www.kaggle.com/*datasets*/adityakadiwal/water-potability

[11] Jamari Untung. (2022, Maret 20). PENJELASAN CARA KERJA ALGORITMA K-NEAREST NEIGHBOR (KNN) [online]. available: http://labdas.si.fti.unand.ac.id/2022/03/20/ penjelasan-cara-kerja-algoritma-k-nearest-neighbor-knn/

[12] Spyder. (2023). Ringkasan [online]. available: https://www.spyder-ide.org/

[13] Kaggle. (2018). Oprec Ristek 2018 - Data Science [online]. available: https://www.kaggle.com/competitions/oprecristekds/overview/description

[14] Lawton George (2022, Januari). DEFINISI prapemrosesan data [online]. available: https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing

[15] Science Data in Master's. (2023). Cara Mengatasi Data Hilang [online]. available: https://www.mastersindatascience.org/learning/how-to-deal-with-missing-data/

[16] Afifah Lutfia (2023). Apa Itu Bias dan Variance di Machine Learning? [online]. available: https://ilmudatapy.com/apa-itu-bias-dan-variance-di-machine-learning/

[17] Mufadhol. (2022, Mei 13). Perbedaan Data Training Dan Data Testing [online]. available: https://teknik-informatika-s1.stekom.ac.id/informasi/baca/Perbedaan-Data-Training-dan-Data-Testing/d475bd43bdae3488afe8a0f648ee5671fb6cdc40

[18] A. R. Isnain, A. I. Sakti, D. Alita, and N. S. Marga, "SENTIMEN ANALISIS PUBLIK TERHADAP KEBIJAKAN LOCKDOWN PEMERINTAH JAKARTA MENGGUNAKAN ALGORITMA SVM," J. Data Min. Dan Sist. Inf., vol. 2, no. 1, p. 31, Feb. 2021, doi: 10.33365/jdmsi.v2i1.1021

[19] G. Gunawan and Y. Reswan, "DESAIN APLIKASI PENGENALAN POLA TANDA TANGAN MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM)," J. MEDIA INFOTAMA, vol. 17, no. 1, Feb. 2021, doi: 10.37676/jmi.v17i1.1311.

[20] Anggreany Susan Maria. (2020). Confusion Matrix [online]. available: https://socs.binus.ac.id/2020/11/01/confusion-matrix/