Jurnal Ilmiah Dan Karya Mahasiswa
Vol. 1 No. 6 Desember 2023
e-ISSN :2985-7732, p-ISSN :2985-6329, Hal 382-396
DOI: https://doi.org/10.54066/jikma.v1i6.1162

Implementasi Model Machine Learning dalam Mengklasifikasi Kualitas Air

Stacyana Jesika
Universitas Negeri Medan
Email: stacyanajs@mhs.unimed.ac.id

Suci Ramadhani
Universitas Negeri Medan
Email: suciramadhani122002@gmail.com

Yohanna Permata Putri
Universitas Negeri Medan
Email: yohannapasaribu6@gmail.com

Jalan Willem Iskandar, Pasar V Medan Estate, Percut Sei Tuan, Deli Serdang

Abstract. *Water quality is an important factor in maintaining human health and environmental sustainability. Water pollution is a major problem in Indonesia, so it is important to monitor and classify water quality effectively. Implementation of machine learning models in classifying water quality can provide important benefits in the environmental and health fields. This research uses two machine learning algorithms, namely KNN and SVM, to classify water quality. The water quality data used comes from the website www.kaggle.com, which was uploaded by MsSmartyPants in 2021 with the title "Water Quality (Dataset for water quality classification)". Implementation of this machine learning model involves the steps of data collection, data pre-processing, selection of relevant attributes, algorithm selection, model training, evaluation, and model implementation for real-time water quality classification.*

Keywords: *Classification, K-NN, SVM, and Water Quality*.

Abstrak. Kualitas air merupakan faktor penting dalam menjaga kesehatan manusia dan keberlanjutan lingkungan. Pencemaran air menjadi masalah utama di Indonesia, sehingga penting untuk memantau dan mengklasifikasikan kualitas air dengan efektif. Implementasi model machine learning dalam mengklasifikasikan kualitas air dapat memberikan manfaat penting dalam bidang lingkungan dan kesehatan. Penelitian ini menggunakan dua algoritma machine learning, yaitu KNN dan SVM, untuk mengklasifikasikan kualitas air. Data kualitas air yang digunakan bersumber dari website *www.kaggle.com*, yang telah diunggah oleh MsSmartyPants pada tahun 2021 dengan judul "Water Quality (Dataset for water quality classification)". Implementasi model machine learning ini melibatkan langkah-langkah pengumpulan data, pra-pemrosesan data, pemilihan atribut yang relevan, pemilihan algoritma, pelatihan model, evaluasi, dan implementasi model untuk klasifikasi kualitas air secara real-time.

Kata kunci: Klasifikasi, K-NN, SVM, dan Kualitas Air.

Revised Oktober 30, 2023, Revised November 03, 2023; Accepted November 24, 2023
*Stacyana Jesika, stacyanajs@mhs.unimed.ac.id

LATAR BELAKANG

Kualitas air merupakan faktor penting dalam menjaga kesehatan manusia dan keberlanjutan lingkungan. Air memiliki standarisasi tersendiri dalam layak atau tidaknya bagi tubuh manusia. Pencemaran air merupakan isu utama di Indonesia, dimana sumber utamanya berasal dari limbah domestik dan rumah tangga. Dampaknya adalah menurunnya kualitas dan ketersediaan air bersih di berbagai wilayah di negara ini. Oleh karena itu, penting untuk memantau dan mengklasifikasikan kualitas air dengan efektif (Said et al., 2022).

Implementasi model machine learning dalam mengklasifikasikan kualitas air adalah penerapan algoritma dan teknik machine learning untuk menganalisis data terkait kualitas air dan menghasilkan model yang dapat memprediksi kualitas air berdasarkan atribut-atribut yang relevan (Savitri & Nursalim, 2023). Algoritma machine learning yang digunakan pada penelitian ini yaitu KNN dan SVM untuk mengklasifikasi serta memprediksi tingkat akurasi pada klasifikasi kualitas air.

Penelitian ini menggunakan algoritma KNN dan SVM karena beberapa alasan yaitu diantaranya adalah, SVM merupakan algoritma pembelajaran yang digunakan untuk masalah klasifikasi dan regresi. Prinsip dasar SVM adalah mencari hyperplane (bidang pemisah) yang optimal untuk memisahkan dua kelas data dengan jarak maksimum di antara kelas-kelas tersebut. SVM memiliki keunggulan dalam kemampuan generalisasi yang tinggi, kemampuan menghasilkan model klasifikasi yang baik dengan data pelatihan terbatas, dan kemudahan implementasi melalui formulasi support vector dalam masalah Quadratic Programming (QP) (Suyanto, 2017). Sementara itu untuk metode KNN merupakan algoritma pembelajaran yang berdasarkan prinsip "tetangga terdekat". KNN melakukan klasifikasi dengan cara mencari k-tetangga terdekat dari data baru yang akan diprediksi, kemudian menggunakan mayoritas kelas dari tetangga-tetangga tersebut untuk memprediksi kelas data baru. KNN memiliki beberapa keunggulan, seperti pelatihan yang cepat, sederhana dan mudah dipahami, ketahanan terhadap data pelatihan yang bising, serta efektivitasnya dalam menangani data pelatihan yang besar (Mutrofin et al., 2014).

Mengklasifikasikan kualitas air dengan menggunakan model machine learning dapat memberikan manfaat penting dalam bidang lingkungan dan kesehatan. Dengan memanfaatkan teknologi machine learning, kita dapat mengidentifikasi pola-pola kompleks dalam data air, mengklasifikasikan air berdasarkan standar kualitas yang

ditetapkan, dan mengambil tindakan yang tepat untuk menjaga dan meningkatkan kualitas air.

Data kualitas air yang didapatkan sendiri bersumber dari website www.kaggle.com yang diupload oleh MsSmartyPants pada tahun 2021 dengan judul Water Quality (Dataset for water quality classification). Implementasi model machine learning melibatkan langkah-langkah seperti pengumpulan data kualitas air, pra-pemrosesan data, pemilihan atribut yang relevan, pemilihan algoritma machine learning yang sesuai, pelatihan model menggunakan data latih, evaluasi dan pengujian model, serta implementasi dan penggunaan model untuk mengklasifikasikan kualitas air secara real-time. Oleh karena itu penelitian ini dibuat untuk membangun model machine learning yang mampu mengklasifikasikan kualitas air berdasarkan atribut-atribut yang tersedia, memprediksi tingkat akurasi dalam klasifikasi kualitas air dengan menggunakan algoritma K-NN dan SVM, serta meningkatkan pemahaman tentang kualitas air dan faktor-faktor yang mempengaruhinya melalui analisis data menggunakan model machine learning.

KAJIAN TEORITIS

1. Klasifikasi

Klasifikasi merupakan tugas krusial dalam machine learning dan data mining yang bertujuan untuk mengategorikan data ke dalam kelas atau kategori yang telah ditentukan berdasarkan atribut-atribut atau fitur-fitur tertentu. Dalam konteks klasifikasi, kita memiliki kumpulan data yang sudah diberi label sebelumnya, yang berarti setiap data di dalamnya sudah memiliki label yang sesuai dengan kelas atau kategori tertentu. Tujuan utama dari klasifikasi adalah untuk mengembangkan model yang mampu memahami pola-pola atau relasi-relasi antara atribut-atribut tersebut, lalu menggunakannya untuk memprediksi kelas atau kategori yang tepat untuk data-data baru yang belum pernah terlihat sebelumnya (Han et al., 2012).

2. Super Vector Mechine (SVM)

a. Pengertian SVM

SVM, atau Mesin Vektor Dukungan, adalah metode pembelajaran mesin yang terawasi yang digunakan untuk tugas seperti klasifikasi dan regresi. SVM pada dasarnya adalah sebuah pendekatan klasifikasi biner yang membagi titik-

titik data ke dalam dua kelas. Saat melatih SVM, tujuannya adalah menemukan hiperplane margin maksimum yang dapat memisahkan kedua kelas dengan jarak terbesar dari titik data pelatihan terdekat (McCartney, 2015).

SVM, atau Mesin Vektor Dukungan, adalah metode pembelajaran mesin yang terawasi yang digunakan untuk tugas seperti klasifikasi dan regresi. SVM pada dasarnya adalah sebuah pendekatan klasifikasi biner yang membagi titik-titik data ke dalam dua kelas. Saat melatih SVM, tujuannya adalah menemukan hiperplane margin maksimum yang dapat memisahkan kedua kelas dengan jarak terbesar dari titik data pelatihan terdekat (Han et al., 2012).

3. K-Nearest Neighbor (KNN)

K-Nearest Neighbor atau KNN, merupakan algoritma yang digunakan untuk mengklasifikasikan data dengan memanfaatkan data latih yang berasal dari K tetangga terdekat dalam sekitarnya (Maulida, 2020). k-NN dilakukan dengan mengidentifikasi kelompok k objek pada data pelatihan yang memiliki karakteristik yang serupa dengan objek pada data baru atau data pemeriksaan (Bachtiar et al., 2019). Salah satu teknik lalai belajar, K-Nearest Neighbor (K-NN) mencari sekelompok k objek dalam data pelatihan yang memiliki kemiripan paling besar dengan objek dalam data baru atau data pengujian (Maulida, 2020) .

Algoritma ini beroperasi dengan menggunakan perhitungan jarak terpendek dari sampel uji ke sampel latihan untuk menentukan K-Nearest Neighbors (KNN). Setelah mengidentifikasi KNN, langkah berikutnya adalah mengambil mayoritas dari KNN tersebut sebagai prediksi untuk sampel uji. Biasanya, tingkat kedekatan atau jarak diukur menggunakan metrik jarak Euclidean (Maulida, 2020).

METODE PENELITIAN

Penelitian ini dimulai dengan mengumpulkan dan memahami data yang diperoleh. Selanjutnya, dilakukan persiapan data melalui proses preprocessing, yang meliputi langkah-langkah seperti membersihkan data, memilih data yang relevan, dan menyeimbangkan data. Setelah itu, data yang telah diolah akan dibagi menjadi bagian-bagian yang sesuai untuk memulai proses klasifikasi menggunakan algoritma Machine Learning. Akhirnya, dilakukan evaluasi model dan perbandingan tingkat akurasi dari setiap model yang diuji untuk memprediksi kualitas air. Berikut adalah tahapan-tahapan

yang dilakukan pada penelitian ini yaitu:

1. Pengumpulan dan pemahaman data

Pengumpulan dan pemahaman data merupakan tahap awal sangat penting dalam mengelolah data. Data mining ialah metode untuk mengekstrak atau mengumpulkan informasi bernilai dari kumpulan data besar. Prosesnya sering melibatkan pemanfaatan teknik statistik dan matematika dengan menggunakan teknologi kecerdasan buatan (Said et al., 2022). Melalui analisis data, kita dapat memperoleh informasi yang beragam, seperti jumlah rekaman data, jenis data yang ada, dan jumlah kolom data yang tersedia. Data yang digunakan pada penelitian ini besifat universal (didapatkan dari berbagai tempat) yang bersumber dari website www.kaggle.com diunggah oleh MsSmartyPants pada tahun 2021 dengan judul Water Quality (Dataset for water quality classification), adapun rincian data yang digunakan sebagai berikut:

no	variabel	rentang nilai
1	Aluminium	0 – 5
2	Ammonia	-0.08 – 29.84
3	Arsenic	0 – 1.05
4	Barium	0 – 4.49
5	Cadmium	0 – 0.13
6	Chloramine	0 – 8.68
7	Chromium	0 – 0.90
8	Copper	0 – 2
9	Flouride	0 – 1.50
10	Lead	0 – 0.20
11	Nitrat	0 – 19.83
12	Nitrit	0 – 2.93
13	Mercury	0 – 0.01
14	Perchlorate	0 – 60.01
15	Radium	0 – 7.99
16	Selenium	0 – 0.100
17	Silver	0 – 0.5
18	Uranium	0 – 0.09
19	Bacteria	0 – 1
20	Virus	0 – 1
21	is_safe	0 – 1

Gambar 1: Tabel atribut dataset

2. Preprocessing Data

Data preprocessing merupakan teknik awal dalam proses penambangan data yang bertujuan untuk mengubah data mentah menjadi format yang lebih efisien dan bermanfaat. Format data mentah yang berasal dari berbagai sumber seringkali mengandung kesalahan, nilai yang hilang, dan inkonsistensi. Oleh karena itu, perlu dilakukan pembenahan format data agar hasil dari proses penambangan data menjadi

tepat dan akurat (Said et al., 2022). Dengan melakukan preprocessing data yang baik, dapat mempengaruhi hasil dari penambangan data dengan meningkatkan akurasi dan menghasilkan informasi yang lebih jelas. Tahapan dalam data preprocessing terdiri dari data cleansing, data selection, dan data balancing.

3. Split Data

Pembagian data merupakan salah satu metode yang digunakan untuk mengevaluasi kinerja model, dengan menyesuaikan persentase data dalam set pelatihan dan pengujian. Hal ini bertujuan untuk memperoleh performa maksimal pada model machine learning yang akan digunakan. Split data ini melibatkan pemisahan dataset menjadi dua bagian, yang akan digunakan sebagai data pelatihan dan data pengujian dengan proporsi tertentu. Dalam penelitian ini, variasi persentase dan jumlah pembagian data akan dilakukan dalam beberapa percobaan untuk mengamati hasil akurasi tertinggi.(Said et al., 2022)

4. Algortima Machine Learning

Adapun beberapa algortima Machine leraning yang digunakan dalam memprediksi klasifikasi kualitas air diantaranya:

a. K-Nearest Neighbors (KNN)

Adalah suatu metode machine learning yang memanfaatkan k-nearest neighbors (k tetangga terdekat) untuk melakukan klasifikasi. Dalam konteks prediksi kualitas air, KNN dapat digunakan untuk mengklasifikasikan air berdasarkan atribut-atribut yang relevan. Cara kerja algoritma ini adalah dengan menghitung jarak antara sampel data yang baru dengan tetangga terdekatnya. Kualitas air baru kemudian diklasifikasikan berdasarkan mayoritas label dari tetangga terdekatnya. Dengan demikian, KNN memanfaatkan informasi dari tetangga terdekat untuk membuat prediksi klasifikasi.

b. SVM

SVM adalah algoritma yang beroperasi dengan cara berikut: ia mengaplikasikan pemetaan non-linear untuk mengganti data train asal ke dalam dimensi yang lebih tinggi. Di dalam dimensi baru tersebut, SVM mencari hiperplane pemisah linear yang optimal (dikenal sebagai "batas keputusan") yang memisahkan tuple dari satu kelas dengan yang lain. Melalui pemetaan non-linear yang sesuai ke dimensi yang cukup tinggi, data dari kedua kelas selalu bisa

dibedakan oleh hiperplane. SVM mengidentifikasi hiperplane ini dengan memanfaatkan vektor dukungan (yang merupakan tuple pelatihan yang sangat penting) dan margin (yang ditentukan oleh vektor dukungan).(Han et al., 2012)

5. Evaluasi Model

Tahapan ini dilakukan dalam menilai Kemampuan atau hasil kerja dari suatu algoritma dalam tingkatan akurasi pada model, diperlukan penghitungan matriks kebingungan (confusion matrix). Confusion matrix adalah teknik yang dipakai dalam menilai hasil kerja dari model yang diuji terutama dalam kasus klasifikasi pada machine learning (Savitri & Nursalim, 2023). Confusion matrix memiliki empat kombinasi yang berbeda berdasarkan nilai prediksi dan nilai aktual. Tabel evaluasi model dapat ditemukan pada tabel berikut:

Confusion Matrix		Nilai Aktual	
		Positif	**Negatif**
Nilai Prediksi	**Positif**	*True Positives*	*False Positives*
	Negatif	*False Negatives*	*True Negatives*

Gambar 2: Tabel confusion matrix

Confusion matrix bertujuan untuk memvisualisasikan prediksi dan keadaan aktual dari data yang dihasilkan oleh algoritma machine learning. Hal ini dilakukan dengan menghitung akurasi, presisi, recall, dan F1-score. Metrik-metrik ini sangat penting dalam mengevaluasi kinerja klasifikasi atau algoritma machine learning yang digunakan untuk melakukan prediksi. Rumus untuk keempat metrik ini ditunjukkan dalam tabel berikut:

No.	Pengukuran	Rumus
1.	Akurasi	$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$
2.	Presisi	$Precision = \frac{TP}{TP + FP}$
3.	Recall	$Recall = \frac{TP}{TP + FN}$
4.	F1-Score	$F1 - Score = \frac{2\ x\ Recall\ x\ Precision}{Recall + Precision}$

Gambar 3: Tabel rumus confusion matrix

HASIL DAN PEMBAHASAN

Berikut ini merupakan tahapan-tahapan dalam mengklasifikasikan dataset mengenai kualitas air pada penelitian ini dengan menggunakan serta membandingkan antara 2 algoritma model mechine learning yaitu algoritma KNN dan algoritma SVM:

1. Exploratory Data Analysis (EDA)

Pada tahapan ini bertujuan untuk melakukan analisis pada data yang akan diuji, seperti memahami struktur, karakteristik dan informasi yang ada pada dataset yang dianalisis. Selain itu tahapan ini juga digunakan dalam melakukan identifikasi pola dan korelasi antar variable sebelum melakukan analisis statistik yang lebih mendalam lagi pada Pembangunan model yang akan diuji.

A. Jumlah kolom dan baris

```
[ ]  # Dataset Dimension
     df.shape

     (7999, 21)
```

Gambar 4: Jumlah baris dan kolom

Dataset pada penelitian ini terdiri dari 7999 baris dan 21 kolom

B. Mencari tipe data

```
[ ]  # Information of Data
     df.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 7999 entries, 0 to 7998
     Data columns (total 21 columns):
      #   Column       Non-Null Count  Dtype
     ---  ------       --------------  -----
      0   aluminium    7999 non-null   float64
      1   ammonia      7999 non-null   object
      2   arsenic      7999 non-null   float64
      3   barium       7999 non-null   float64
      4   cadmium      7999 non-null   float64
      5   chloramine   7999 non-null   float64
      6   chromium     7999 non-null   float64
      7   copper       7999 non-null   float64
      8   flouride     7999 non-null   float64
      9   bacteria     7999 non-null   float64
      10  viruses      7999 non-null   float64
      11  lead         7999 non-null   float64
      12  nitrates     7999 non-null   float64
      13  nitrites     7999 non-null   float64
      14  mercury      7999 non-null   float64
      15  perchlorate  7999 non-null   float64
      16  radium       7999 non-null   float64
      17  selenium     7999 non-null   float64
      18  silver       7999 non-null   float64
      19  uranium      7999 non-null   float64
      20  is_safe      7999 non-null   object
     dtypes: float64(19), object(2)
```

Gambar 5: Tipe data

Dataset pada penelitian ini memiliki 2 tipe data yaitu 19 atribut bertipe data float dan atribut lainya bertipe data object.

C. Ringkasan statistic

```
[ ] # Descriptive statistics
    df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
aluminium	7996.0	0.666396	1.265323	0.00	0.0400	0.070	0.2800	5.05
ammonia	7996.0	14.278212	8.878930	-0.08	6.5775	14.130	22.1325	29.84
arsenic	7996.0	0.161477	0.252632	0.00	0.0300	0.050	0.1000	1.05
barium	7996.0	1.567928	1.216227	0.00	0.5600	1.190	2.4825	4.94
cadmium	7996.0	0.042803	0.036049	0.00	0.0080	0.040	0.0700	0.13
chloramine	7996.0	2.177589	2.567210	0.00	0.1000	0.530	4.2400	8.68
chromium	7996.0	0.247300	0.270663	0.00	0.0500	0.090	0.4400	0.90
copper	7996.0	0.805940	0.653595	0.00	0.0900	0.750	1.3900	2.00
flouride	7996.0	0.771646	0.435423	0.00	0.4075	0.770	1.1600	1.50
bacteria	7996.0	0.319714	0.329497	0.00	0.0000	0.220	0.6100	1.00
viruses	7996.0	0.328706	0.378113	0.00	0.0020	0.008	0.7000	1.00
lead	7996.0	0.099431	0.058169	0.00	0.0480	0.102	0.1510	0.20
nitrates	7996.0	9.819250	5.541977	0.00	5.0000	9.930	14.6100	19.83
nitrites	7996.0	1.329846	0.573271	0.00	1.0000	1.420	1.7600	2.93
mercury	7996.0	0.005193	0.002967	0.00	0.0030	0.005	0.0080	0.01
perchlorate	7996.0	16.465266	17.688827	0.00	2.1700	7.745	29.4875	60.01
radium	7996.0	2.920106	2.322805	0.00	0.8200	2.410	4.6700	7.99
selenium	7996.0	0.049684	0.028773	0.00	0.0200	0.050	0.0700	0.10
silver	7996.0	0.147811	0.143569	0.00	0.0400	0.080	0.2400	0.50
uranium	7996.0	0.044672	0.026906	0.00	0.0200	0.050	0.0700	0.09
is_safe	7996.0	0.114057	0.317900	0.00	0.0000	0.000	0.0000	1.00

Gambar 6: Ringkasan statistik

Ringkasan statistik ini berguna dalam membantu analisis statistik secara keseluruhan.

Keterangan :

- Count = jumlah data
- Mean = rata-rata
- Std = standar deviasi
- Min = nilai minimum
- 25 % = kuartil pertama
- 50 % = kuartil kedua atau median
- 75 % = kuartil ketiga
- Max = nilai maksimum

2. Preproccessing Data

Sebelum melakukan pengujian model maka sangatlah diperlukan tahapan pra pemrosesan data yaitu dengan melakukan serangkaian langkah untuk merapikan data sebelum memulai analisis. Sebelum menuju ke tahap pemprosesan. Data mentah akan

diolah terlebih dahulu dengan tujuan untuk memastikan data dalam keadaan siap untuk diuji.

A. Data cleaning (Check *Missing Value*)

Tahap ini sangatlah penting dilakukan dalam melakukan prepoccessing data, karena memastikan bahwa data tidak memiliki missing value atau tidak berdampak pada kualitas data itu sendiri. Data dengan kualitas yang baik adalah data yang terbebas dari adanya missing value. Missing value ini juga memiliki dampak terhadap pengujian model karena missing value dapat berdampak kepada nilai-nilai yang diperlukan seperti mean, modus, media dan standar deviasi yang berdampak kepada tingkat keakuratan model tersebut.

```
[ ]  # CHECK MISSING VALUE
     df.isnull().sum()

     aluminium      0
     ammonia        0
     arsenic        0
     barium         0
     cadmium        0
     chloramine     0
     chromium       0
     copper         0
     flouride       0
     bacteria       0
     viruses        0
     lead           0
     nitrates       0
     nitrites       0
     mercury        0
     perchlorate    0
     radium         0
     selenium       0
     silver         0
     uranium        0
     is_safe        0
```

Gambar 7: Check missing value

Dari gambar diatas dapat disimpulkan bahwa data yang diperoleh bebas dari adanya missing value.

B. Feature scaling

```
[ ]  # Calling the Standardization library
     from sklearn.preprocessing import StandardScaler
     scaler = StandardScaler()
     scaler.fit(X_train)
     scaler.fit(X_test)
```

```
▾ StandardScaler
StandardScaler()
```

```
[ ]  X_train_scaled = scaler.transform(X_train)
     X_test_scaled = scaler.transform(X_test)

     #instead of doing fit and transform in two different steps

     X_train_scaled = scaler.fit_transform(X_train)
     X_test_scaled = scaler.fit_transform(X_test)
```

Gambar 8: Feature scaling

Tahap ini bertujuan untuk memastikan bahwasanya skala pada data ada di tingkatan yang sama, tahapan ini juga penting dilakukan pada preprocessing data sebelum melakukan pengujian model

C. Split data

Proses spilt data dilakukan sebelum memasuki pengujian model yaitu dengan cara membagi data menjadi 2 yaitu data testing dan data training. Split data ini sangat berguna dalam melakukan pengujian model, seperti membantu mengevaluasi kinerja model, mencegah overfitting dan lain-lain.

```
[ ]  #SPLIT DATA
     X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, stratify = Y, random_state = 42)
```

Gambar 9: Split data

Berdasarkan gambar diatas data data dibagi menjadi 2 yaitu 20 % data testing dan 80 % data training.

D. Uji KNN

```
[ ]  # KNeighborsClassifier library
     from sklearn.neighbors import KNeighborsClassifier
     knn = KNeighborsClassifier(n_neighbors = 4)
     knn.fit(X_train,Y_train)
```

```
KNeighborsClassifier
KNeighborsClassifier(n_neighbors=4)
```

Gambar 10: Library K-NN

Berdasarkan gambar diatas dilakukan pemanggilan library KNN terlebih dahulu sebelum melakukan ngeujian model yang ingin diuji, dan kemudian pada pengujian model KNN ini parameter k yang mewakili jumlah tetangga terdekat yang akan dipertimbangkan adalah 4 tertangga terdekat dengan menggunakan data train (data latih)

```
[ ]  # Prediction k-NN Algorithm
     y_pred = knn.predict(X_test)

     Pred = pd.DataFrame(np.array(y_pred),
                         columns=['Prediction'])
     verbal=Pred['Prediction'].map({0:'Non-safe', 1:'safe'})
     Pred['Conclusion'] = pd.DataFrame(np.array(verbal),
                          columns=['Conclusion'])
     Pred.head(250)
```

Gambar 11: Uji K-NN

Kode pada gambar diatas digunakan untuk menampilkan hasil prediksi dengan model KNN, dengan menggunakan model K-NN yang telah dilatih sebelumnya untuk memprediksi label data pengujian (‘X_test’). Hasil prediksi disimpan dalam DataFrame ‘Pred’, yang mencakup kolom 'Prediction' (hasil prediksi numerik).

```
[ ]  # Report of Model Evaluation
     print(f"Report Model Evaluation: \n {classification_report(Y_test, y_pred)}")
     print(f"Confusion Matrix: \n {confusion_matrix(Y_test, y_pred)}")
```

```
Report Model Evaluation:
                precision    recall  f1-score   support

           0.0       0.90      0.99      0.94      1418
           1.0       0.67      0.13      0.22       182

    accuracy                           0.89      1600
   macro avg       0.78      0.56      0.58      1600
weighted avg       0.87      0.89      0.86      1600

Confusion Matrix:
 [[1406   12]
 [ 158   24]]
```

Gambar 12: Evaluasi model K-NN

Pada gambar diatas dapat dilihat hasil dari pengujian knn didaptkan nilai akurasinya adalah 89%. Yang artinya tingkatan akurasi pada model KNN ini sebesar 89%.

E. Uji SVM

```
[ ]  # Calling the Support Vector Machine library
     from sklearn.svm import SVC
     model = SVC(kernel='rbf')
     model.fit(X_train, Y_train)
```

Gambar 13: Library SVM

Gambar diatas yaitu pemanggilan library SVM dalam melatih model SVM dengan kernel RBF.

```
[ ]  # Report of Model Evaluation
     print(f"Report Model Evaluation: \n {classification_report(Y_test, y_pred_SVM1)}")
     print(f"Confusion Matrix: \n {confusion_matrix(Y_test, y_pred_SVM1)}")
```

```
Report Model Evaluation:
                precision    recall  f1-score   support

           0.0       0.91      0.99      0.95      1418
           1.0       0.75      0.28      0.41       182

    accuracy                           0.91      1600
   macro avg       0.83      0.63      0.68      1600
weighted avg       0.90      0.91      0.89      1600

Confusion Matrix:
 [[1401   17]
 [ 131   51]]
```

Gambar 14: Evaluasi model SVM

Pada gambar diatas dapat dilihat hasil dari pengujian SVM didaptkan nilai akurasinya adalah 91%. Yang artinya tingkatan akurasi pada model KNN ini sebesar 91%.

KESIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan diatas dalam perbandingan antara model K-Nearest Neighbors (KNN) dan model Support Vector Machine (SVM), hasil menunjukkan bahwa model SVM memiliki tingkat akurasi yang sedikit lebih tinggi, yakni mencapai 91%, dibandingkan dengan model KNN yang mencapai 89%. Hal ini berarti bahwa model SVM lebih unggul sedikit dalam memodelkan hubungan dan pola dalam data yang digunakan dalam pengujian. Tingkat akurasi yang tinggi pada kedua model yaitu melebihi 85%, menunjukkan bahwa keduanya memiliki kemampuan yang baik dalam mengklasifikasikan data dengan benar.

DAFTAR REFERENSI

Bachtiar, F., Syahputra, I., & Wicaksono, S. (2019). PERBANDINGAN ALGORITME MACHINE LEARNING UNTUK MEMPREDIKSI PENGAMBIL MATA KULIAH. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, *6*(5), 543–548. https://doi.org/10.25126/jtiik.2019611755

Han, J., Kamber, M., & Pei, J. (2012). Techniques to Improve Classification Accuracy. In *Data Mining, Concepts and Techniques*.

Maulida, A. (2020). *Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes*. *1*(2), 29–33.

McCartney, P. R. (2015). Big Data Science. In *MCN The American Journal of Maternal/Child Nursing* (Vol. 40, Nomor 2). https://doi.org/10.1097/NMC.0000000000000118

Mutrofin, S., Izzah, A., Kurniawardhani, A., & Masrur, M. (2014). OPTIMASI TEKNIK KLASIFIKASI MODIFIED K NEAREST NEIGHBOR MENGGUNAKAN ALGORITMA GENETIKA. *GAMMA*, *10*, *Nomor*(September), 130–134. https://doi.org/10.1017/9781316534946.021

Said, H., Matondang, N. H., & Irmanda, H. N. (2022). Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi. *Techno.Com*, *21*(2), 256–267. https://doi.org/10.33633/tc.v21i2.5901

Savitri, L., & Nursalim, R. (2023). Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma Machine Learning dengan Pendekatan Supervised Learning. *Diophantine Journal of Mathematics and Its Applications*, *2*(01), 30–36. https://doi.org/10.33369/diophantine.v2i01.28260

Suyanto, D. (2017). *Data Mining Untuk Klasifikasi dan Klastering Data*. Informatika Bandung.