SMARTICS Journal, Vol.9 No.1 2023. p1-6
ISSN online: 2476-9754, ISSN print: 2623-0429
DOI : https://doi.org/10.21067/smartics.v9i1.8113

Komparasi Algoritma Machine Learning dalam Identifikasi Kualitas Air

Dwi Hartanti[a,*], Afu Ichsan Pradana[b]
[ab] Universitas Duta Bangsa Surakarta
[*]correspondence email : dwihartanti@udb.ac.id

Abstract–Water is a very important part of life because it is the source of human life and living things and about 71% of the earth's area is water. Every human being has the human right to clean water which is a basis for the realization of a decent and dignified life for humans. Machine learning is a branch of artificial intelligence that focuses on using data and algorithms to mimic the way humans learn, gradually increasing accuracy and intelligence. This study uses water quality data using four algorithmic methods, namely Decission Tree, Logistic Regression, SVM, and ANN. The objective of this research is to compare which method has the maximum accuracy for water quality identification. The accuracy results obtained are the Decission Tree method of 60.19%, the Logistic Regression method of 62.80%, the SVM method of 68.59%, and the ANN method of 69.54%.

***Index Terms*-Water Quality; Decission Tree; Logistics Regression; SVM; ANN**

***Abstrak*–Air menjadi bagian sangat penting untuk kehidupan karena merupakan sumber kehidupan manusia serta mahluk hidup dan sekitar 71% wilayah bumi merupakan air. Setiap manusia memiliki Hak asasi manusia atas air yang bersih yang merupakan sebuah dasar untuk terwujudnya kehidupan yang layak dan bermatrabat bagi manusia. *Machine learning* merupakan cabang dari kecerdasan buatan yag berfokus pengguunaan data dan algoritma untuk meniru cara manusia belajar secara bertahap dapat meningkatkan akurasi dan kecerdasan. Penelitian ini menggunakan data kualitas air menggunakan empat metode algoritma yaitu Decission Tree, Logistic Regression, SVM, dan ANN. Tujuan penelitian adalah untuk membandingkan metode mana yang memiliki nilai akursi paling maksimum untuk identifikasi kualitas air. Hasil akurasi yang diperoleh adalah dengan metode *Decission Tree* sebesar 60,19%, metode *Logistic Regression* sebesar 62,80%, metode *SVM* sebesar 68,59%, dan metode *ANN* sebesar 69,54%.**

***Kata Kunci*—Kualitas Air; Decission Tree; Logistic Regression; SVM; ANN**

I. PENDAHULUAN

Air menjadi bagian sangat penting untuk kehidupan karena merupakan sumber dari kehidupan manusia serta makhluk hidup dan sekitar 71% wilayah bumi merupakan air dan memiliki senyawa komples[1][2][3]. Setiap manusia memiliki Hak asasi manusia atas air yang bersih yang merupakan sebuah dasar untuk terwujudnya kehidupan yang layak dan bermatrabat bagi manusia[4]. Maka dari itu pentingnya untuk menjaga demi keberlagsungan kehidupan manusia. Suatu ukuran kondisi air dilihat dari berbagai karakteristik fisik, kimiawi, dan biologis[2]. Banyaknya pembagunanan yang semakin pesat mengakibatkan banyak wilayah memiliki kualitas air yang kurang baik. Hal tersebut dibuktikan dengan banyaknya sungai-sungai yang berubah fungsi perubahan ini dikhawatirkan akan menurunnya kualitas air yng berada di sepanjang sungai-sungai di berbagai lingkungan sekitar.

Banyak upaya untuk menjaga kualitas air yaitu dengan melakukan kegiatan pengecekan apakah air tersebut terdapat penyakit atau bakteri sehingga akan dilakukan sebuah tindakan untuk melakukan pencegahan jika terjadi penurunan sebuah kualitas air[5]. Kualitas air memiliki banyak paramater meliputi mikrobiologi, kimia anorganik, parameter fisik, serta parameter kimia. Paramater kualitas air diketahui karena air mengangung zat mineral yang terlarut dalam air[2].

Kecerdasan buatan memiliki cabang keilmuan yaitu *Machine learning. Machine Learning* memiliki fokus dalam penggunaan data serta algoritma yang digunkan untuk meniru cara manusia belajar secara bertahap dapat meningkatkan akurasi dan kecerdasannya[6][7]. *Machine Learning* digunakan untuk memecahkan berbagai masalah dan mempelajari data yang ada dan bisa melakukan tugas tertentu [8][9]. Penelitian ini menggunakan algoritma Decission Tree, Logistic Regression, SVM, dan ANN. Metode Algortima tersebut banyak di implementasikan untuk sebuah permasalahan dengan banyak kasus dan cocok untuk kasus kualitas air[10].

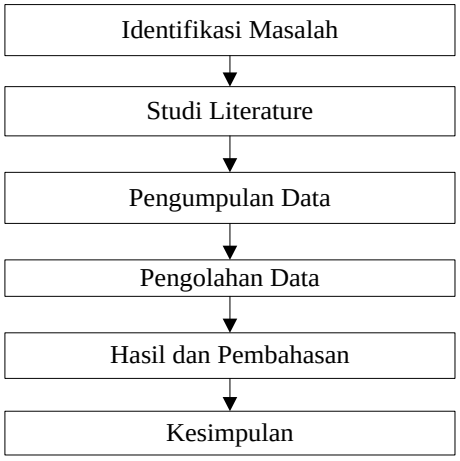
Penelitian yang dilakukan Weiskhy dkk menyimpulkan bahwa nilai Accuracy 84.81% dan nilai AUCnya 0.898 dengan menggunakan algoritma SVM-PSO dan nilai Accuracy 80.00% dan nilai AUCnya

0.787 dengan menggunakan Algoritma C4.5-PSO[11]. Penelitian yang dilakukan oleh Icha Gusti Vidiastanta dkk menyimpulkan bahwa hasil yang didapat yaitu Akurasi rata – rata yang didapatkan dengan metode K-Nearest Neighbors dan Support Vector Machine adalah sebesar 88.94% dan 87.71%. Hasilnya K-Nearest Neighbors (KNN) dengan nilai rata-rata lebih tinggi menjadi metode yang lebih baik untuk klasifikasi status air dibandingkan dengan metode Support Vector Machine (SVM)[12]. Penelitian yang dilakukan Priscolius Evrolino Jennes dkk menyimpulkan bahwa sebuah akurasi kelayakan sumber air dimana hasil akurasinya adalah 71% untuk SVM, 61% untuk decision tree, 67% untuk random forest berdasarkan akurasi diatas dapat membantu analisis dan klasifikasi kelayakan air di Indonesia[13].

Tujuan dari penelitian yang dilakukan adalah untuk membandingkan Decission Tree, Logistic Regression, SVM, dan ANN dalam kasus identifikasi kualitas air. Maka dengan empat metode algoritma tersebut dapat di ketahui metode algoritma yang paling baik digunakan dalam identifikasi kualitas air dengan hasil akurasi yang paling maksimal.

II. METODE PENELITIAN

Penelitian yang dilakukan memiliki alur penelitian yang digunakan untuk melakukan komparasi metode Decission Tree, Logistic Regression, SVM, dan ANN untuk kualitas air terlihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

1. Tahap pertama dalam penelitian dimulai dengan mengidentifikasi masalah terkait pentingnya air bagi kehidupan manusia dengan masalah kualitas air untuk di konsumsi.
2. Tahap kedua merupakan studi literatur sebagai pengumpulan informasi yang terkait kualitas air. Cara yang dilakukan dengan melakukan pengumpulan berupa artikel ilmiah, buku dan sumber-sumber untuk menjadi informasi terkait penelitian.
3. Tahap pengumpulan data dilakukan dengan teknik pengumpulan data yang menggunakan data dari *kaggle* tentang *Water Quality*.
4. Proses pengolahan data mencakup beberapa langkah, dimulai dari dataset, *Preparation*, *preprocessing*, *Visualisai Data*, *Training* dan *result*. Penogolahan data menggunkan menggunakan bahasa pemrograman Python dengan Alat Jupyter Notebook. Proses pengolahan data yang dilakukan akan menghasilkan sebuah *result* atau hasil yang akan dibahas dan menghasilkan sebuah kesimpulan dalam proses penelitian yang dilakukan.

III. HASIL

1. Dataset

Dataset yang digunakan untuk data latih dalam program ini berjumlah 3276 baris, yang memiliki 9 Fitur atau Atribut *(PH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity)* dan 1 Atribut Labelling *(Potability)*. Dataset yang digunakan dalam penelitian ini dataset yang digunakan adalah kualitas air dengan tipe data *csv* untuk proses identifikasi dalam membandingkan hasil akurasi dari ketiga metode yang digunakaan yaitu Decission Tree, Logistic Regression, SVM, dan ANN. Hasil data dapat dilihat pada Gambar 2 Data Kualitas Air

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Pot
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	

Gambar 2. Data Kualitas Air

2. Data Preparation

Data Preparation adalah suatu proses/langkah yang dilakukan untuk membuat data mentah menjadi data yang berkualitas. Dalam penelitian ini melakukan pesiapan data yang akan digunakan. Hasil dari data preparation terlihat pada Gambar 3.

```
                ph     Hardness        Solids  Chloramines      Sulfate  \
count  2785.000000  3276.000000   3276.000000  3276.000000  2495.000000
mean      7.080795   196.369496  22014.092526     7.122277   333.775777
std       1.594320    32.879761   8768.570828     1.583085    41.416840
min       0.000000    47.432000    320.942611     0.352000   129.000000
25%       6.093092   176.850538  15666.690297     6.127421   307.699498
50%       7.036752   196.967627  20927.833607     7.130299   333.073546
75%       8.062066   216.667456  27332.762127     8.114887   359.950170
max      14.000000   323.124000  61227.196008    13.127000   481.030642

       Conductivity  Organic_carbon  Trihalomethanes    Turbidity   Potability
count   3276.000000     3276.000000      3114.000000  3276.000000  3276.000000
mean     426.205111       14.284970        66.396293     3.966786     0.390110
std       80.824064        3.308162        16.175008     0.780382     0.487849
min      181.483754        2.200000         0.738000     1.450000     0.000000
25%      365.734414       12.065801        55.844536     3.439711     0.000000
50%      421.884968       14.218338        66.622485     3.955028     0.000000
75%      481.792304       16.557652        77.337473     4.500320     1.000000
max      753.342620       28.300000       124.000000     6.739000     1.000000
```

Gambar 3. Hasil Data Preparation

3. Data Preprocessing

Pada tahap data prepocesing dilakukan proses analisa data dengan value kosong atau null dan melakukan pengisian data yang kosong atu null tersebut dengan rata-rata disetiap atribut. Gambar 4 untuk menampilkan data kosong atau null dan Gambar 5 Mengisi data yang kosong dengan nilai rata-rata disetiap atribut

```
ph                491
Hardness            0
Solids              0
Chloramines         0
Sulfate           781
Conductivity        0
Organic_carbon      0
Trihalomethanes   162
Turbidity           0
Potability          0
dtype: int64
```

Gambar 4. Menampilkan data null

```
dataset["ph"].mean()
```

```
7.080794504276819
```

```
dataset["ph"]=dataset["ph"].fillna(7.080794504276819)
```

```
dataset["Sulfate"].mean()
```

```
333.7757766108134
```

```
dataset["Sulfate"]=dataset["Sulfate"].fillna(333.7757766108134)
```

```
dataset["Trihalomethanes"].mean()
```

```
66.39629294676803
```

```
dataset["Trihalomethanes"]=dataset["Trihalomethanes"].fillna(66.39629294676803)
```

Gambar 5. Mengisi data yang kosong dengan nilai rata-rata

4. Data Training dan Data Testing

Dalam tahap ini melakukan pembagian data training dan data testing. Penelitian membagi data menjadi data training sebanyak 80% dan data testing 20%. Data training akan digunakan untuk melatih model yang di bangun dan data testing unut menguji dan mengetahui performance model yang didapatkan pada tahapan testing.

5. Implementasi Metode Algoritma

a. Menerapkan Metode Decission Tree

Dalam penerapan menggunakan Metode Decission Tree. Pada implementasi menggunakan Decission Tree diperoeh nilai akurasi modelnya sebesar 60,19%. Hasil akurasi pada Gambar 6.

```
# DecisionTreeClassifier

from sklearn.tree import DecisionTreeClassifier

tree_model = DecisionTreeClassifier()
tree_model = tree_model.fit(x_train, y_train)
```

```
from sklearn.metrics import accuracy_score

y_pred = tree_model.predict(x_test)
acc_score= round(accuracy_score(y_pred, y_test), 3)
print('The accuracy for Decision Tree is : ', acc_score*100,'%')
```

```
The accuracy for Decision Tree is :  60.199999999999996 %
```

Gambar 6. Hasil akurasi metode Decission Tree

b. Menerapkan Metode Logistic Regression

Dalam penerapan menggunkan Metode Logistic Regression. Pada implementasi menggunakan Metode Logistic Regression diperoeh nilai akurasi modelnya sebesar 62,80%. Hasil akurasi pada Gambar 7.

```
#LogisticRegression
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(x_train, y_train)
y_predlr = lr.predict(x_test)
acc_lr = lr.score(x_test, y_test)
print('The accuracy for Logistic Regression is : ', acc_lr*100, '%')
```

```
The accuracy for Logistic Regression is :  62.80487804878049 %
```

Gambar 7. Hasil akurasi metode Logistic Regression

c. Menerapkan Metode Support Vector Machine (SVM)

Dalam penerapan menggunkan Metode Logistic Regression. Pada implementasi menggunakan Metode Support Vector Machine (SVM) diperoeh nilai akurasi modelnya sebesar 68,59%. Hasil akurasi pada Gambar 8.

```
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score

svm = SVC()
svm.fit(x_train, y_train)
y_predsvm = svm.predict(x_test)
```

```
acc_svm= accuracy_score(y_predsvm, y_test)
print('The accuracy for SVM is : ', acc_svm*100, '%')
```

```
The accuracy for SVM is :  68.59756097560977 %
```

Gambar 8. Hasil akurasi metode SVM

d. Menerapkan Metode Artificial Neural Network (ANN)

Dalam penerapan menggunkan Metode Artificial Neural Network (ANN). Pada implementasi menggunakan Metode Artificial Neural Network (ANN) diperoeh nilai akurasi modelnya sebesar 69,54%. Hasil akurasi pada Gambar 9.

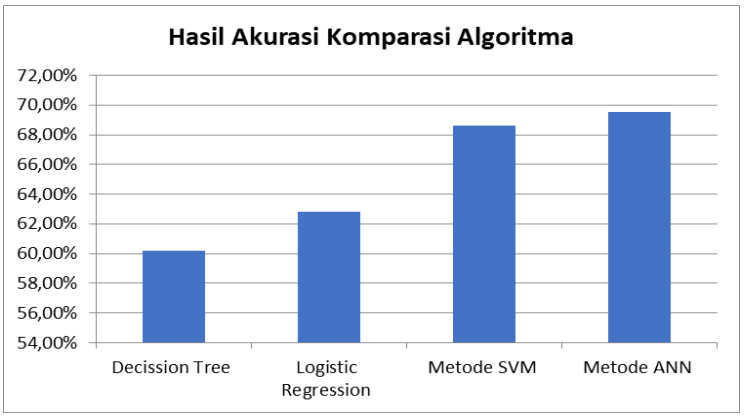
```
|: from sklearn.neural_network import MLPClassifier
   clf = MLPClassifier(activation='relu', solver='lbfgs', max_iter=500, alpha=1e-5,␣
    ↪hidden_layer_sizes=(5,3))
   clf.fit(x_train, y_train)

   from sklearn.metrics import accuracy_score
   y_model = clf.predict(x_train)
   print('The accuracy Neural Network is : ', accuracy_score(y_train, y_model)*100,␣
    ↪'%')

The accuracy Neural Network is :  69.54198473282443 %
```

Gambar 9. Hasil akurasi metode ANN

Dari hasil komprasi metode didapatkan hasi seperti Gambar 10.



Gambar 10. Hasil Akurasi Komparasi Algoritma

Hasil dari pembandingan yang telah dilakukan didapatkan nilai akurasi metode *Decission Tree* sebesar 60,19%, metode *Logistic Regression* sebesar 62,80%, metode Support Vector Machine (*SVM)* sebesar 68,59%, dan metode Artificial Neural Network *(ANN)* sebesar 69,54%. Nilai tertinggi nilai kaurasi untuk kualitas air pada metode Artificial Neural Network (ANN).

IV. KESIMPULAN

Tujuan dilakukan penelitian yang dilakuakn untuk perbandingan tingkat keakuratan dari metode penelitian Decission Tree, Logistic Regression, SVM, dan ANN. Hasil dari program ini berupa Akurasi dalam identifikasi kualitas air layak di konsumsi atau tidak. Masing-masing akurasi yang diperoleh adalah dengan metode *Decission Tree* sebesar 60,19%, metode *Logistic Regression* sebesar 62,80%, metode *SVM* sebesar 68,59%, dan metode *ANN* sebesar 69,54%.

V. REFRENCE AND CITATION

[1] E. De Buck, V. Borra, E. De Weerdt, A. Vande Veegaete, and P. Vandekerckhove, “A systematic review of the amount of water per person per day needed to prevent morbidity and mortality in (post-)disaster settings,” *PLoS One*, vol. 10, no. 5, pp. 1–14, 2015, doi: 10.1371/journal.pone.0126395.

[2] F. Y. Rahman, I. I. Purnomo, and N. Hijriana, “Penerapan Algoritma Data Mining Untuk Klasifikasi Kualitas Air,” *Technol. J. Ilm.*, vol. 13, no. 3, p. 228, 2022, doi: 10.31602/tji.v13i3.7070.

[3] M. A. Rahman, N. Hidayat, and A. Afif Supianto, “Komparasi Metode Data Mining K-Nearest

Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Vol. 2, No. 12, Desember 2018, hlm. 6346-6353 e-ISSN*, vol. 2, no. 12, pp. 925–928, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[4] Generosa Lukhayu Pritalia, "Analisis Komparatif Algoritme Machine Learning dan Penanganan Imbalanced Data pada Klasifikasi Kualitas Air Layak Minum," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 2, no. 1, pp. 43–55, 2022, doi: 10.24002/konstelasi.v2i1.5630.

[5] Y. T. K. Yunior and K. Kusrini, "Sistem Monitoring Kualitas Air Pada Budidaya Perikanan Berbasis IoT dan Manajemen Data," *Creat. Inf. Technol. J.*, vol. 6, no. 2, p. 153, 2021, doi: 10.24076/citec.2019v6i2.251.

[6] N. Muniroh and E. Agus Priatno, "PENERAPAN ALGORITMA K-NN PADA MACHINE LEARNING UNTUK KLASIFIKASI KUALITAS AIR BUDIDAYA AKUAPONIK BERBASIS IoT," *J. Teknol. dan Bisnis*, vol. 4, no. 2, pp. 73–86, 2022, doi: 10.37087/jtb.v4i2.87.

[7] I. M. Faiza and W. Andriani, "Tinjauan Pustaka Sistematis : Penerapan Metode Machine Learning untuk Deteksi Bencana Banjir," vol. 11, no. September, pp. 59–63, 2022.

[8] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.

[9] Y. Herdiana, "Penerapan Machine Learning Dengan Model Linear Regression Terhadap Analisis Kualitas Hasil Petik the Di Pt. Perkebunan …," *Comput. J. Inform.*, vol. 09, pp. 1–9, 2022, [Online]. Available: https://unibba.ac.id/ejournal/index.php/computing/article/view/855%0Ahttps://unibba.ac.id/ejournal/index.php/computing/article/download/855/710

[10] M. M. Mutoffar and A. Fadillah, "KLASIFIKASI KUALITAS AIR SUMUR MENGGUNAKAN ALGORITMA RANDOM FOREST," vol. 04, no. 02, pp. 138–146, 2022.

[11] W. S. Dharmawan, "Komparasi Algoritma Klasifikasi Svm-Pso Dan C4.5-Pso Dalam Prediksi Penyakit Jantung," *I N F O R M a T I K a*, vol. 13, no. 2, p. 31, 2022, doi: 10.36723/juri.v13i2.301.

[12] N. Anggraeni, G. Arifiana, and A. M. Abadi, "Klasifikasi Kualitas Air Sungai Winongo Menggunakan Fuzzy Inference System (FIS) Metode Mamdani," *Progr. Stud. Mat. FMIPA UNY*, no. 2014, pp. 161–170, 2017.

[13] P. E. J. Y. W. A. D. N. Fadlilah, "Prediksi Kelayakan Sumber Air Minum Menggunakan Algoritma Support Vector Machine (SVM)," vol. 3, p. 4865, 2022.

Dwi Hartanti, Meraih gelar Sarjana Teknik Informatika (S.Kom) dari Sekolah Tinggi Ilmu Manajamen Informatika Komputer Sinar Nusantara Surakarta pada tahun 2017. Kemudian gelar Master (M.Kom) dari Unoversitas Amikom Yogyakarta pada tahun 2019. Saat ini penulis menjadi Tenaga Pengajar di Universitas Duta Bangsa Surakarta.