Analisis Algoritma *Gradient Boosting*, *AdaBoost* dan *CatBoost* dalam Klasifikasi Kualitas Air

http://dx.doi.org/10.28932/jutisi.v8i2.4906

Riwayat Artikel
Received: 8 Juni 2022 | *Final Revision*: 2 Agustus 2022 | *Accepted*: 7 Agustus 2022

Taufik Zulhaq Jasman ✉ #1, Muhammad Alief Fadhlullah #2, Arnold Listanto Pratama #3, Rismayani*4

Teknik Informatika, Universitas Dipa Makassar
Jalan Perintis Kemerdekaan Km. 9, Makassar, 90245, Indonesia
[1]taufikzulhak42@gmail.com
[2]alieffadhlullah9@gmail.com
[3]kalebuciha@gmail.com

* *Rekayasa Perangkat Lunak Universitas* Dipa *Makassar*
Jalan Perintis Kemerdekaan Km. 9, Makassar, 90245, Indonesia
[4]rismayani@undipa.ac.id

✉Corresponding author: [1]taufikzulhak42@gmail.com

Abstrak — **Penelitian ini bertujuan untuk mencari akurasi tertinggi dari ketiga algoritma klasifikasi yang digunakan. Akurasi yang tinggi sangat diperlukan untuk membantu masyarakat dan instansi terkait seperti PDAM (Perusahaan Daerah Air Minum) untuk mendistribusikan air kepada masyarakat. Oleh karena itu, klasifikasi kualitas air ini akan memilih algoritma dengan akurasi tertinggi. dan menguji kinerja ketiga model. Metode yang digunakan dalam analisis ini, yaitu untuk mengatasi data yang hilang adalah metode median. Kemudian untuk menangani data yang tidak seimbang digunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*). Dalam studi ini, dilakukan perbandingan akurasi dan kinerja *Gradient Boosting*, *AdaBoost*, dan *CatBoost*. dengan hasil bahwa algoritma *CatBoost* memiliki akurasi dan kinerja tertinggi sebesar 68%, diikuti oleh *Gradient Boosting* sebesar 60% dan *AdaBoost* sebesar 58%. Kemudian performa nilai AUC (*Area Under the Curve*) *CatBoost* sebesar 0,678, *Gradient Boosting* sebesar 0,595, dan *AdaBoost* sebesar 0,584. Namun hasil akurasi dan performanya masih kurang.**

Kata kunci— **Akurasi; Algoritma; Kualitas Air; Model; Penelitian.**

Analysis of Gradient Boosting, Adaboost, Catboost Algorithms in Water Quality Classification

Abstract — ***This study aims to find the highest accuracy of the three classification algorithms. High accuracy is needed to help the community and related agencies such as PDAM (Perusahaan Daerah Air Minum) to distribute water to the community. Therefore, this water quality classification will choose the algorithm with the highest accuracy. and test the performance of the three models. The method used in this analysis, namely to overcome missing data, is the median method. Then to handle unbalanced data, SMOTE (Synthetic Minority Over-sampling Technique) method is used. In this study, we compare the accuracy and performance of Gradient Boosting, AdaBoost, and CatBoost. with the result that the CatBoost algorithm has the highest accuracy and performance of 68%, followed by Gradient Boosting of 60% and AdaBoost of 58%. Then the performance of the AUC (Area Under the Curve) CatBoost value is 0.678, Gradient Boosting is 0.595, and AdaBoost is 0.584. However, the results of accuracy and performance are still lacking.***

Keywords— ***Accuracy; Algorithm; Model; Research; Water Quality.***

I. PENDAHULUAN

Kebutuhan utama sehari-hari makhluk di dunia ini salah satunya yang tak bisa dipisahkan ialah air. Bukan hanya penting bagi kita manusia, salah satu komponen utama dari flora dan fauna adalah air. Jika air tidak ada, kemungkinan tidak ada kehidupan di dunia ini karena semua makhluk hidup yang ada di dunia ini sangat memerlukan air agar dapat menjalani hidupnya [1].

Sustainable Development Goals (SDGs), dikenal juga sebagai *Global Goals* (Cita – cita Global), diadopsi oleh Perserikatan Bangsa-Bangsa pada tahun 2015 sebagai seruan universal untuk bertindak demi mengakhiri kemiskinan, memelihara planet ini, dan memastikan bahwa pada tahun 2030 semua orang menikmati perdamaian dan kemakmuran [2]. Saat ini ada 17 SDGs dari program ini. SDGs 6 bertujuan untuk memastikan ketersediaan dan pengelolaan air dan sanitasi yang berkelanjutan untuk semua orang, dan menegaskan pentingnya air dan sanitasi dalam agenda politik global. Dibandingkan dengan SDGs lain, lingkungan indikator di SDG 6 masih baru, artinya bahwa hal ini merupakan pertama kalinya banyak negara anggota PBB harus mengumpulkan, menyerahkan dan menganalisis jenis data yang diperlukan untuk mengukur kemajuan pada *goals* ini [3].

Menurut publikasi dari BPS, produksi air bersih pada tahun 2020 sebesar 5,262.1 juta m^3. Di mana jumlah pelanggan perusahaan air bersih menurut provinsi dari tahun 2015, 2017, 2018, 2019 dan 2020 berturut – turut sebesar 11,746,504, 13,160,707, 14,128,479, 14,985,944 dan 15,345,992 [4]. Dari pemaparan data di atas ditarik benang merah bahwa jumlah permintaan air bersih meningkat dari tahun ke tahun.

Dengan bertambahnya jumlah penduduk kebutuhan air yang bersih oleh orang – orang akan semakin meningkat fakta yang ada lapangan, kualitas dan kuantitas air bersih semakin merosot. Tingkatan sisi kualitas air yang diperlukan untuk setiap aktivitas tertentu mempunyai standar mutu yang berbeda – beda maka dari itu dibutuhkan percobaan demi mengetahui kesesuaian kualitas air dengan peruntukannya [5].

Hal inti yang seringkali dihadapi berkaitan dengan sumber daya air bersih ialah kuantitas air yang tidak mampu memenuhi kebutuhan yang terus menanjak dan kualitas air untuk keperluan lokal yang semakin hari kian merosot [6]. Diperparah lagi di negara Indonesia, akses untuk bersih masih menjadi problematika. Hampir sebagian besar air tawar yang dipakai berasal dari air sungai, danau, waduk dan air sumur [7].

Dikutip dari Keputusan Menkes RI No. 1405/menkes/sk/xi/2002 tentang Persyaratan Kesehatan Lingkungan Kerja Perkantoran dan industri adanya pengertian berkaitan dengan Air Bersih yaitu air yang dipakai untuk kebutuhan sehari-hari dan kualitasnya memenuhi seluruh persyaratan kesehatan air bersih sesuai dengan peraturan perundang-undangan yang berlaku dan dapat dikonsumsi apabila dimasak [1].

Pihak PDAM merupakan pelaku penting dalam mengalirkan air bersih yang mempunyai kualitas mumpuni ke tiap - tiap rumah tangga yang menggunakan jasa dari PDAM. Demi membantu pihak dari PDAM dalam menentukan tingkatan kualitas air tadi memenuhi standar atau tidak memenuhi agar lebih cepat dan tepat untuk membantu warga yang ingin tahu status kualitas air apakah layak diminum atau tidak, dibuatlah sebuah teknik dengan memakai metode klasifikasi [8].

Klasifikasi merupakan aktivitas yang membutuhkan penggunaan algoritma *machine learning* yang mempelajari cara mengelompokkan kelas - kelas ke dalam suatu kelas tertentu. Contoh sederhana yang mudah diketahui ialah mengklasifikasikan pasien ke dalam kelas "sembuh" atau "belum sembuh". Klasifikasi berdasar pada permodelan prediksi masalah di mana kelas kelas diprediksi untuk input data [9]. Fitur atau atribut dari klasifikasi bisa dalam bentuk bilangan biner, juga bisa dalam bentuk kategorikal, dan sebagainya [10].

Penelitian – penelitian sebelumnya pernah dilakukan dengan menggunakan dataset yang sama yaitu dataset *water potability*.

Penelitian pertama dilakukan oleh Rouqi Yang. Di dalam penelitian ini, dilakukan komparasi *imputation of missing values* dengan metode menghapus data yang hilang, menggunakan nilai tengah (median) atau nilai rata – rata (*mean*), *arbitrary value imputation* dan menggunakan KNN. Hasil dari komparasi tersebut yaitu metode KNN dan Median memiliki akurasi tertinggi. Jika dibandingkan dari kedua metode tersebut, KNN memiliki akurasi tertinggi. Namun KNN memiliki beberapa kekurangan yaitu pertama, KNN lemah dalam dataset yang besar. Kedua, KNN tidak bekerja maksimal dalam dimensi data yang tinggi [11].

Penelitian kedua dilakukan komparasi algoritma dengan dataset yang sama yaitu algoritma *Logistic Regression*, KNN, *Random Forest* dan ANN. dengan kesimpulan Random Forest memiliki akurasi tertinggi yaitu sebesar 70.42% dan akurasi terendah berada pada *Logistic Regression* dengan akurasi 60.51% [12].

Dari penelitian kedua, hasil akurasi yang di dapat masih kurang, bahkan belum mencapai 80%. Maka dari itu dilakukan langkah inisiatif yaitu mencari algoritma yang lain. Untuk mendapatkan akurasi atau performa yang tinggi. dari kedua penelitian di atas belum ada dari satupun menggunakan algoritma *Gradient Boosting Classifier*, *CatBoost Classifier* dan

AdaBoost Classifier. Ketiga algoritma tersebut termasuk ke dalam algoritma *boosting* [13] [14]. Pada Tabel 1 memuat kelebihan dan kekurangan dari masing – masing algoritma [15] [16] [17].

TABEL 1

KELEBIHAN dan KEKURANGAN MASING – MASING ALGORITMA

Gradient Boosting		***CatBoost***		***AdaBoost***	
Kelebihan	Kekurangan	Kelebihan	Kekurangan	Kelebihan	Kekurangan
Seringkali memberikan akurasi prediksi yang tidak dapat direkayasa**.**	Model ini akan terus dioptimasi untuk meminimalkan semua kesalahan. Hal ini dapat memperbanyak *outlier* dan menyebabkan *overfitting*.	Memberikan hasil yang bagus untuk data kategorikal	Performanya hanya lebih baik pada data kategorikal	Lebih mudah digunakan karena hanya sedikit membutuhkan parameter, tidak seperti seperti SVM.	*AdaBoost* sensitif terhadap data *noisy* dan *outlier*, jika berencana untuk menggunakan *AdaBoost* maka sangat disarankan untuk menghilangkannya.
Banyak fleksibilitas, menyediakan beberapa opsi *hyper parameter tuning* yang membuat fungsi tersebut sangat fleksibel.	Seringkali membutuhkan banyak pohon (>1000 pohon) yang dapat menghabiskan banyak waktu dan memori.	Dapat melatih model pada GPU yang secara signifikan meningkatkan kecepatan learning	Kinerjanya sangat buruk jika variabel tidak diatur dengan benar	*AdaBoost* dapat digunakan untuk meningkatkan akurasi pengklasifikasi yang lemah, sehingga membuatnya fleksibel.	*AdaBoost* juga lebih lambat dari *XGBoost*.
Tidak memerlukan proses *preprocessing*	Fleksibilitas yang tinggi di mana menghasilkan banyak parameter yang berinteraksi dan sangat memengaruhi banyak parameter (*number of iterations*, *tree depth*, *regularizations parameters*, dll.).			Sekarang telah diperluas di luar klasifikasi biner dan telah digunakan dalam klasifikasi teks dan gambar juga.	
Dapat menangani data yang hilang	Sifatnya kurang interpretatif, meskipun hal ini mudah diatasi dengan menggunakan berbagai *tools*.				

Tujuan akhir dari penelitian ini adalah untuk menemukan akurasi tertinggi dari ketiga algoritma klasifikasi tersebut. di mana algoritma yang memiliki akurasi tertinggi akan dijadikan acuan dalam klasifikasi kualitas air ini. Serta menguji performa dari ketiga model tersebut. dengan harapan dapat menemukan akurasi yang tinggi dan performa yang baik.

II. METODE PENELITIAN

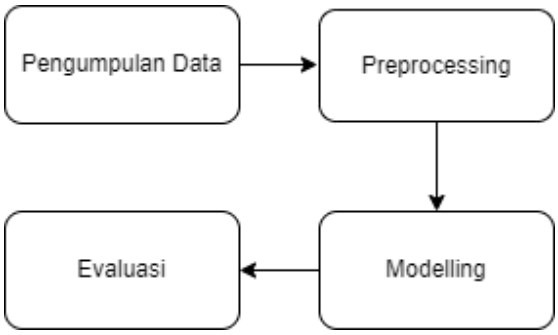
Tahapan yang pertama kali dilakukan adalah pengumpulan data. Dataset sangat penting karena dataset itulah yang akan kita olah. Dataset yang akan diolah adalah dataset *Water Quality* yang dapat diunduh gratis dari *Kaggle*. Di dalam dataset ini terdapat 9 fitur dan 1 kelas. Diantaranya:

1) ***Ph Value***: pH menunjukkan jumlah kadar ion hidrogen dalam larutan air. Selain itu, pH adalah indikator yang baik untuk air lunak atau keras. Misalnya, sampel air keras memiliki tingkat pH tinggi yang melebihi 8,5, sedangkan sampel air lunak memiliki tingkat pH rendah yaitu kurang dari 6,5 [18].
2) ***Hardness***: Didefinisikan sebagai kandungan terukur dari logam kation divalen. Kalsium terlarut (Ca^{++}) dan magnesium (Mg^{++}) adalah dua kation divalen yang ditemukan pada tingkat yang cukup tinggi di sebagian besar perairan. Pada kandungan air yang alami, kalsium dan magnesium terikat pada bikarbonat, sulfat atau klorida [19].
3) ***Solids***: *Total* Di*ssolve Solids* (TDS) adalah jumlah bahan organik dan anorganik, seperti logam, mineral, garam, dan ion, yang terlarut dalam volume air tertentu, TDS pada dasarnya merupakan takaran semua yang terlarut dalam air yang bukan merupakan molekul H2O [20].
4) ***Chloramines***: Kloramin dibuat dari sekelompok bahan kimia yang mengandung amonia dan klorin. Kloramin yang paling umum digunakan dalam pengolahan air kota adalah monokloramin. Monokloramin ditambahkan ke air dalam

jumlah yang telah diukur, memastikan bahwa mikroorganisme telah dibunuh, tetapi air masih aman untuk diminum [21].

5) ***Sulfate***: Sulfat adalah kombinasi belerang dan oksigen dan merupakan bagian dari mineral alami di beberapa bagian tanah dan batuan yang mengandung air tanah. Mineral larut dari waktu ke waktu dan dilepaskan ke dalam air tanah [22].
6) ***Conductivity***: Ukuran kemampuan air untuk mengalirkan arus listrik terutama dipengaruhi dalamnya air oleh adanya padatan terlarut anorganik. Seperti klorida, nitrat, sulfat, natrium, magnesium, dan sebagainya. Hal ini membantu mengetahui kualitas air berdasarkan: garam terlarut dan juga membantu untuk menentukan jumlah reaksi kimia atau teknik perawatan yang diperlukan untuk memurnikan air [23].
7) ***Organic Carbon***: *Total Organic Carbon* adalah jumlah karbon organik yang ada dalam batuan utama yang dinyatakan sebagai persen berat. TOC adalah mewakili untuk jumlah total bahan organik yang ada dalam sedimen dan digunakan sebagai indikator kekayaan sumber sehubungan dengan seberapa banyak hidrokarbon yang dapat dihasilkan oleh sedimen [24].
8) ***Trihalomethanes***: THMs adalah cairan yang mudah menguap pada suhu kamar dan berbagai efek racun telah dihubungkan dengan paparan jangka pendek dan jangka panjang dari hewan percobaan pada dosis tinggi. THM yang diberikan oleh saluran cairan pada jagung menyebabkan keracunan yang lebih signifikan daripada dosis setara yang diberikan dalam emulsi cair [25].
9) ***Turbidity***. *Turbidity* atau kekeruhan mengacu pada sifat hamburan cahaya pada sampel. Kekeruhan dapat digambarkan sebagai "kabur" atau "putih", dan disebabkan oleh partikel halus yang menghamburkan cahaya pada kurang lebih 90 derajat ke arah dari cahaya memasuki sampel. Kekeruhan tidak sama dengan warna, atau warna dengan kekeruhan [26].
10) ***Potability***. Merupakan kelas dari data ini di mana kelas 0 adalah kelas yang di mana air tersebut tidak dapat diminum. Sedangkan kelas 1 merupakan kelas air yang dapat diminum.

Dataset ini memiliki total 3277 *record*. Setelah pengumpulan data, proses berikutnya yaitu *Preprocessing* kemudian dilanjutkan dengan *Modelling* atau klasifikasi dan tahapan terakhir adalah tahapan Evaluasi. Untuk gambarannya divisualisasikan pada Gambar 1.



Gambar 1. Proses Penelitian

1. *Preprocessing*: terdapat beberapa *Missing Values* yang mana mempengaruhi akurasi data nantinya. Dilakukan teknik *Replacing Missing Values* yaitu mengganti atau mengisi data yang hilang. Data tersebut akan diganti dengan menggunakan Median. Mengapa tidak menggunakan KNN atau *Mean*? Berdasarkan penelitian dari Rouqi Yang, bahwa KNN dan Median memiliki akurasi tertinggi dalam *Replacing Missing Values*. Namun KNN lemah dalam memproses data yang sangat banyak [11].
   Mengingat *record* dataset sangat banyak maka Median cocok dalam proses ini. Setelah melakukan teknik *Replacing Missing Values*, juga melakukan Teknik *Over Sampling* yaitu SMOTE. Karena terdapat ketidakseimbangan antara antara jumlah kelas 0 dan 1.
   Persentase jumlah dari kelas 0 sebesar 60.99% dan kelas 1 sebesar 39.01%. Di mana selisihnya sebesar 21.98%. Karena jumlah selisihnya lumayan banyak maka akan digunakan Teknik *Oversampling* menggunakan SMOTE. SMOTE adalah teknik dalam *machine learning* untuk menangani masalah yang muncul saat berhadapan dengan sekumpulan kelas yang tidak seimbang [27].
   Setelah Teknik *Replacing Missing Values* dan *Oversampling* dengan menggunakan SMOTE. Selanjutnya tahapannya yaitu membagi data latih dan data uji. Di sini di terapkanlah teknik *Split Validation* dengan membagi data

latih sebesar 80% dan data uji sebesar 20%. Terakhir, sebelum *modelling* dilakukan teknik standarisasi data dengan menggunakan *Standard Scaler* tujuannya agar tidak terjadi ketimpangan tiap data.

2. *Modelling*: setelah melakukan tahap *Preprocessing* selanjutnya *modelling*. Pada tahap ini data data latih dan data uji diolah atau dihitung berdasarkan algoritma yang digunakan, yaitu *Gradient Boosting Classifier*, *CatBoost Classifier* dan *AdaBoost Classifier*.

   Gradient Boosting Classifier merupakan kumpulan algoritma *machine learning* yang menggabungkan banyak model yang lemah bersama - sama untuk membuat model prediksi yang kuat. *Decision tree* biasanya digunakan saat melakukan *gradient boost*. Model ini menjadi populer karena keefektifannya dalam mengklasifikasikan kumpulan data yang kompleks [28].

   CatBoost atau *Categorical Boosting* merupakan salah satu implementasi dari *Gradient Boosting* yang mana menggunakan bilangan biner dari algoritma *decision tree* sebagai dasar dari prediksi data [29].

   AdaBoost atau *Advanced Boosting* adalah algoritma yang paling baik digunakan untuk meningkatkan kinerja *decision tree* pada dataset dengan kelas label. *AdaBoost* juga dapat digunakan untuk meningkatkan kinerja algoritma *machine learning* apa pun. Algoritma ini paling efektif digunakan algoritma yang lemah. Model ini adalah model yang mencapai akurasi paling tepat di atas peluang acak pada kasus klasifikasi [30].

3. Evaluasi: di tahapan ini akan di diperoleh hasil akurasi, presisi, *recall*, dan *f1-score*. Sebelum itu dilakukan pengujian performa model dengan menggunakan *confusion matrix*. *Confusion Matrix* adalah metode populer yang digunakan pada tahapan klasifikasi. *Confusion matrix* adalah matriks N x N yang digunakan untuk menilai kinerja model klasifikasi, di mana N adalah jumlah kelas target. Matriks ini mengkomparasikan nilai target aktual dengan yang diprediksi oleh model *machine learning* [31]. Setelah itu ditampilkan kurva ROC dan menampilkan nilai AUC untuk menilai performa dari model.

   Receiver Operator Characteristic (ROC) adalah kurva grafik yang dipakai untuk menunjukkan kemampuan analisa pengklasifikasi kelas biner. Kurva ini pertama kali digunakan dalam teori deteksi sinyal tetapi sekarang digunakan di banyak bidang lain seperti kedokteran, radiologi, alarm bahaya, dan *machine learning* [32]. Kurva ROC sangat berguna untuk mengevaluasi performa dari suatu model dan akurasi terutama pada kelas berbentuk biner [33]. Selain ROC juga ada nilai AUC. *Area Under the Curve* (AUC) adalah ukuran kemampuan model klasifikasi untuk membedakan antara kelas dan digunakan sebagai ringkasan dari kurva ROC [34].

TABEL 2

CONFUSION MATRIX

		PREDICTION	
		NEGATIVE	**POSITIVE**
TRUE	**NEGATIVE**	TN	FP
	POSITIVE	FN	TP

Pada Tabel 2 terdapat istilah TN, FP, FN dan TP di mana:

- TN (*True Negative*): model memprediksi negatif dan faktanya negatif .
- FP (*False Positive*): model memprediksi positif dan faktanya negatif.
- FN (*False Negative*): model memprediksi negatif namun faktanya positif.
- TP (*True Positive*): model memprediksi positif dan faktanya positif.

Rumus untuk menghitung akurasi, presisi, recall, dan nilai f1-score adalah sebagai berikut:

a. $\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}}$ (1)

b. $\boldsymbol{Precision} = \frac{TP}{TP+FP}$ (2)

c. $\boldsymbol{Recall} = \frac{TP}{TP+FN}$ (3)

d. $\boldsymbol{F1\ Score} = 2 * \frac{Precision*Recall}{Precision+Recall}$ (4)

III. HASIL DAN PEMBAHASAN

A. *Gradient Boosting Classifier*.

Pada algoritma *Gradient Boosting* di dapati *Confusion Matrix* sebagai berikut:

TABEL 3

CONFUSION MATRIX dari *GRADIENT BOOSTING*.

		PREDICTION	
		0	**1**
ACTUAL	**0**	212	174
	1	149	265

Dari Tabel 3 di diperoleh bahwa:

1. Model memprediksi ada 212 data air yang tidak layak diminum dan faktanya benar 212 data tersebut merupakan air yang tidak layak diminum (TN).
2. Model memprediksi ada 149 data air yang tidak layak diminum namun faktanya 149 data tersebut layak diminum (FN).
3. Model memprediksi ada 174 data air yang layak diminum namun faktanya 174 data tersebut tidak layak diminum (FP).
4. Model memprediksi ada 265 data air yang layak diminum dan faktanya benar 265 data tersebut memang merupakan air yang layak diminum (TP).

Berdasarkan dari tabel confusion matrix bahwa model memprediksi banyak FP dan FN yang mana memengaruhi akurasi dari model tersebut.

Hasil akurasi, *precision*, *recall*, *f1-score*, dan *support*:

TABEL 4

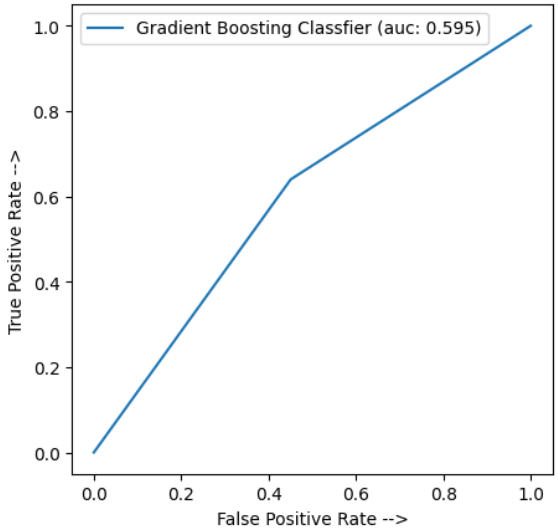
CLASSIFICATION REPORT dari *GRADIENT BOOSTING*.

ACCURACY	**PRECISION**	**RECALL**	**F1-SCORE**	**SUPPORT**
60%	60%	60%	60%	800

Dari Tabel 4 didapati hasil akurasi sebesar 60%. Akurasi tersebut masih tergolong rendah, standar akurasi dapat dikatakan baik apabila setidaknya akurasi tersebut lebih atau sama dengan 70%.

Kurva ROC dan nilai AUC dari *Gradient Boosting* adalah sebagai berikut:

p-ISSN : 2443-2210
e-ISSN : 2443-2229



Gambar 2. Kurva ROC dari *Gradient Boosting*

Pada Gambar 2 didapati bahwa titik kurva sedikit menjauhi dari garis *True Positive Rate*. Serta didapati juga nilai luas *Area Under Curve* sebesar 0.595.

B. *AdaBoost Classifier*

Tabel 5 merupakan *Confusion Matrix* dari *AdaBoost classifier*.

TABEL 5

CONFUSION MATRIX dari *ADABOOST*

		PREDICTION	
		0	**1**
ACTUAL	**0**	209	177
	1	155	259

Dari Tabel 5 di diperoleh bahwa:

1. Model memprediksi ada 209 data air yang tidak layak diminum dan faktanya benar 209 data tersebut merupakan air yang tidak layak diminum (TN).
2. Model memprediksi ada 155 data air yang tidak layak diminum namun faktanya 155 data tersebut layak diminum (FN).
3. Model memprediksi ada 177 data air yang layak diminum namun faktanya 177 data tersebut tidak layak diminum (FP).
4. Model memprediksi ada 259 data air yang layak diminum dan faktanya benar 259 data tersebut memang merupakan air yang layak diminum (TP).

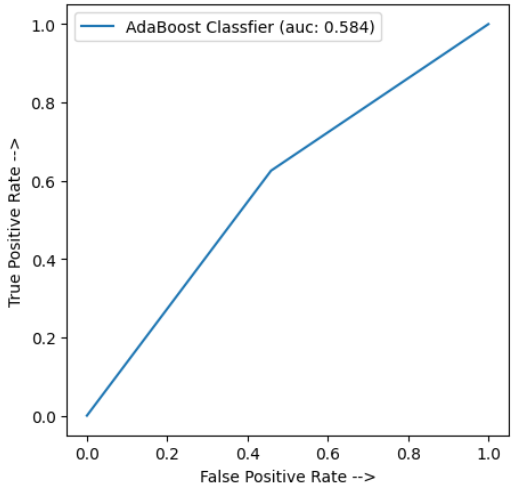
TABEL 6

CLASSIFICATION REPORT dari *ADABOOST*

ACCURACY	**PRECISION**	**RECALL**	**F1-SCORE**	**SUPPORT**
58%	58%	58%	58%	800

Dari Tabel 6 didapati hasil akurasi sebesar 58%. Sedikit lebih rendah dari model sebelumnya.

Kurva ROC:



Gambar 3. Grafik ROC dan Nilai AUC dari *AdaBoost*

Dari kurva yang terlihat pada Gambar 3, didapati bahwa nilai *Area Under Curve* sebesar 0.567.Titik kurva juga sedikit lebih jauh dari garis TPR. Menandakan bahwa performa dari kinerja model ini masih kurang.

C. *CatBoost Classifier*

Tabel 7 merupakan tabel *Confusion Matrix* dari model *CatBoost* adalah sebagai berikut:

TABEL 7

CONFUSION MATRIX dari *CATBOOST*

		PREDICTION	
ACTUAL		**0**	**1**
	0	243	143
	1	113	301

Dari Tabel 7 di diperoleh bahwa:

1. Model memprediksi ada 243 data air yang tidak layak diminum dan faktanya benar 243 data tersebut merupakan air yang tidak layak diminum (TN).
2. Model memprediksi ada 113 data air yang tidak layak diminum namun faktanya 113 data tersebut layak diminum (FN).
3. Model memprediksi ada 143 data air yang layak diminum namun faktanya 143 data tersebut tidak layak diminum (FP).
4. Model memprediksi ada 301 data air yang layak diminum dan faktanya benar 301 data tersebut memang merupakan air yang layak diminum (TP).

Hasil akurasi, *precision*, *recall*, *f1-score*, dan *support*:

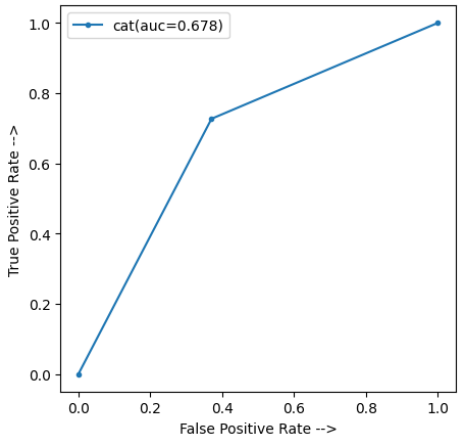
TABEL 8

CLASSIFICATION REPORT dari *CATBOOST*

ACCURACY	**PRECISION**	**RECALL**	**F1-SCORE**	**SUPPORT**
68%	68%	68%	68%	800

Dari Tabel 8 didapati hasil akurasi sebesar 68%. Paling tinggi diantara dua model yang lain.

p-ISSN : 2443-2210
e-ISSN : 2443-2229

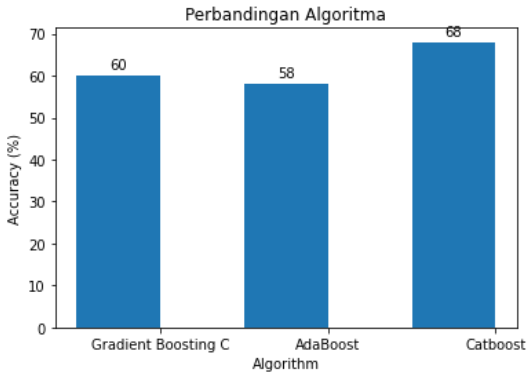


Gambar 4. Grafik ROC dan Nilai AUC dari *CatBoost*

Dari Gambar 4, diperoleh bahwa kurva hampir mendekati garis TPR. Sedikit lebih baik dari kedua model sebelumnya. Dilihat juga nilai *Area Under Curve* sebesar 0.678. juga lebih tinggi dari nilai AUC yang lain.

D. *Komparasi Akurasi*

Dari data – data di atas kemudian dilakukan komparasi untuk melihat algoritma yang manakah yang paling baik dalam menganalisis data ini.

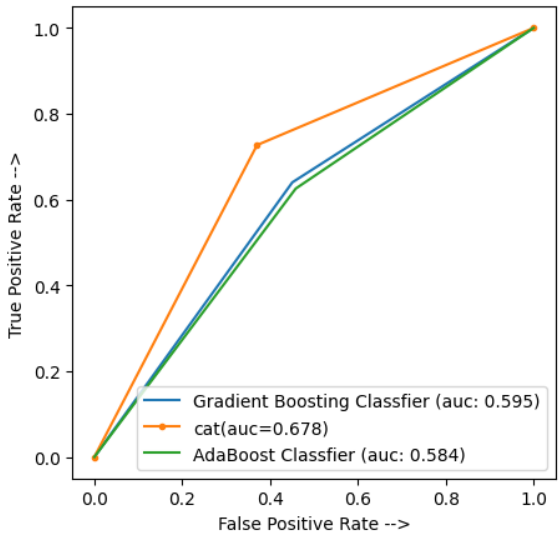


Gambar 5. Hasil Akurasi dari Ketiga Model

Gambar 5 merupakan hasil akurasi dari ketiga model tersebut. Didapati bahwa algoritma *CatBoost* memiliki akurasi paling tinggi yaitu sebesar 68%, disusul *Gradient Boosting* sebesar 60% dan algoritma *AdaBoost* yang paling rendah yaitu sebesar 58%. Hal ini membuat algoritma *CatBoost* lebih dapat dipercaya dibandingkan kedua algoritma lain. Meskipun ketiga kinerja algoritma di atas masih kurang. Dikarenakan akurasi dari ketiga model ini tergolong rendah, ada beberapa teknik atau metode yang dapat dilakukan demi meningkatkan akurasi dan kinerja yaitu 1. Mengganti algoritma yang digunakan, masih banyak algoritma klasifikasi lain yang dapat digunakan untuk mencari akurasi yang terbaik. 2. Menggunakan teknik *feature selection*, yaitu mereduksi fitur dan memilih fitur yang terpenting. 3. *Hyperparameter Tuning*, yaitu mencari *hyperparameter* terbaik dari algoritma ini, teknik ini menggunakan *GridSearchCV* (*Cross Validation*). 4. Selain metode SMOTE, juga ada metode yang lain untuk mengatasi *imbalance data* yang dapat digunakan seperti *ADASYN* (*Adaptive Synthetic*), *Borderline SMOTE*, *SVMSMOTE*, dll. 5. Menghapus *outlier*, mengingat di penelitian ini *outlier* diabaikan yang kemungkinan berpengaruh dalam hasil akurasi dan kinerja model.

E. Komparasi Kurva ROC dan Nilai AUC

Komparasi kurva ROC dan nilai AUC dari ketiga algoritma tersebut yaitu:



Gambar 6. Komparasi Kurva ROC dan Nilai AUC

Dari Gambar 6 merupakan kurva evaluasi kinerja dari ketiga algoritma yang diteliti. Algoritma *CatBoost* paling mendekati *True Positive Rate* dengan nilai sebesar 0.678. sedangkan di bawah *CatBoost* ada *Gradient Boosting* dengan nilai sebesar 0.595. serta *AdaBoost* yang paling di bawah dengan nilai sebesar 0.584. semakin dekat garis kurva dengan garis TPR maka semakin terbaik juga model tersebut. Dari pembahasan akurasi dan kurva ROC dan nilai AUC didapatkan bahwa kinerja ketiga model ini masih kurang. Sama seperti pada pembahasan di komparasi akurasi, beberapa teknik yang perlu dilakukan untuk mendapatkan akurasi tertinggi yaitu mengganti algoritma, menggunakan *feature selection*, *hyperparameter Tuning*, *imbalance data* selain SMOTE, dan menghapus *outlier*.

IV. SIMPULAN

CatBoost merupakan algoritma yang memiliki akurasi dan performa yang tertinggi kemudian disusul *Gradient Boosting* dan yang paling rendah yaitu algoritma *AdaBoost*. Namun akurasi dan performanya masih kurang dan perlu ditingkatkan lagi dengan penggunaan berbagai macam metode. Sehingga untuk pengklasifikasian pada dataset kualitas air ini, performa dari ketiga algoritma ini masih kurang. Di penelitian selanjutnya dapat digunakan algoritma yang berbeda atau dengan menggunakan *feature selection* dan *hyperparameter tuning*, dll. Sehingga mendapatkan akurasi yang lebih tinggi lagi di mana dapat membantu masyarakat maupun pihak – pihak terkait dalam mengklasifikasikan kualitas air.

UCAPAN TERIMA KASIH

Terima kasih kepada Universitas Dipa Makassar yang telah mensupport penelitian sehingga penelitian ini dapat terselesaikan dengan baik, kemudian tak lupa juga mengucapkan pada semua pihak yang terlibat secara langsung atau tidak langsung yang tidak dapat diucapkan satu persatu.

DAFTAR PUSTAKA

[1] R. D. Ambarwati, "Air bagi Kehidupan Manusia," 2016. [Online]. Available: http://dsdap.bantenprov.go.id/upload/Advetorial/1.%202%20ARTIKEL%20AIR%20BERSIH%20(RDA)_EDITOR.pdf.

[2] "Sustainable Development Goals," [Online]. Available: https://www.undp.org/sustainable-development-goals. [Accessed 15 Juli 2022].

[3] "GOAL 6: Clean Water and Sanitation," [Online]. Available: https://www.unep.org/explore-topics/sustainable-development-goals/why-do-sustainable-development-goals-matter/goal-6. [Accessed 15 Juli 2022].

[4] I. W. Pradipta and S. Harsanto, Statistik Air Bersih, 2021.

[5] N. A. Firdaus, "Analisis Kualitas Air (Studi Kasus Mata Air Citrosono Di Kecamatan Grabag Kabupaten Magelang)," *Jurnal Geo Rafflesia: Artikel Ilmiah Pendidikan Geografi*, vol. 4, no. 2, 2020.

[6] E. B. Sasongko, E. Widyastuti and R. E. Priyono, "Kajian Kualitas Air dan Penggunaan Sumur Gali Oleh Masyarakat Di Sekitar Sungai Kaliyasa Kabupaten Cilacap," *Jurnal Ilmu Lingkungan*, vol. 12, p. 72–82, 2014.

[7] D. E. Puspitasari, "Dampak Pencemaran Air Terhadap Kesehatan Lingkungan Dalam Perspektif Hukum Lingkungan (Studi Kasus Sungai Code Di Kelurahan Wirogunan Kecamatan Mergangsan dan Kelurahan Prawirodirjan Kecamatan Gondomanan Yogyakarta)," *Mimbar Hukum*, vol. 21, no. 1, 2009.

[8] R. M. S. Tumangger, N. Hidayat and Marji, "Komparasi Metode Data Mining Support Vector Machine dengan Naive Bayes untuk Klasifikasi Status Kualitas Air," 2019. [Online]. Available: http://j-ptiik.ub.ac.id.

[9] J. Brownlee, "4 Types of Classification Tasks in Machine Learning," 8 April 2020. [Online]. Available: https://machinelearningmastery.com/types-of-classification-in-machine-learning/ . [Accessed 10 May 2022].

[10] H. Tan, "Machine Learning Algorithm for Classification," *Journal of Physics: Conference Series*, vol. 1994, no. 1, 2021.

[11] R. Yang, "Analyses of Approaches to Deal with Missing Data in Water Quality Data Set," in *Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development*, 2022.

[12] D. Poudel, D. Shrestha, S. Bhattarai and A. Ghimire, "Comparison of Machine Learning Algorithms in Statistically Imputed Water Potability Dataset," 2008. [Online]. Available: https://www.researchgate.net/publication/358783243_Comparison_of_Machine_Learning_Algorithms_in_Statistically_Imputed_Water_Potability_Dataset.

[13] Z. Zhang, "Boosting Algorithms Explained. Theory, Implementation, and Visualization," 16 June 2019. [Online]. Available: https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30. [Accessed 15 July 2022].

[14] "CatBoost - Open-source Gradient Boosting Library.," [Online]. Available: https://CatBoost.ai/. [Accessed 15 July 2022].

[15] V. Kurama, "Gradient Boosting for Classification.," [Online]. Available: https://blog.paperspace.com/gradient-boosting-for-classification/. [Accessed 15 July 2022].

[16] M. Yawar, "CatBoost-ML.," [Online]. Available: https://www.codingninjas.com/codestudio/library/CatBoost-ml . [Accessed 15 July 2022].

[17] V. Kurama, "A Guide To Understanding AdaBoost.," [Online]. Available: https://blog.paperspace.com/AdaBoost-optimizer/. [Accessed 15 July 2022].

[18] S. M, "What Is the Best pH Level for Drinking Water?.," [Online]. Available: https://www.medicinenet.com/what_is_the_best_ph_level_for_drinking_water/article.htm. [Accessed 28 May 2022].

[19] H. E. Diggs and J. M. Parker, "Aquatic Facilities," in *Planning and Designing Research Animal Facilities*, Elsevier, 2009, p. 323–331.

[20] J. Woodard, "What is TDS in Water & Why Should You Measure It?," 24 March 2021. [Online]. Available: https://www.freshwatersystems.com/blogs/blog/what-is-tds-in-water-why-should-you-measure-it . [Accessed 28 May 2022].

[21] B. Campbell, "What is Chloramine in Water Treatment?," 20 December 2021. [Online]. Available: https://www.wqpmag.com/water-disinfection/what-chloramine-water-treatment. [Accessed 28 May 2022].

[22] B. Oram, "Water Research Center - SulfateSulfate, Hydrogen Sulfide, Sulfate Reducing Bacteria - How to Identify and Manage.," [Online]. Available: https://www.water-research.net/index.php/sulfates. [Accessed 28 May 2022].

[23] R. Bhatta, S. Baral and R. Khanal, "Water Quality of Wetlands in Nepal: a Case Study of Jagadishpur Reservoir Ramsar Site," September 2015. [Online]. Available: https://www.researchgate.net/publication/286138506.

[24] J. H. Dembicki, "Source Rock Evaluation," in *Practical Petroleum Geochemistry for Exploration and Production*, Elsevier, 2017, p. 61–133.

[25] S. S. Anand, B. K. Philip and H. M. Mehendale, "Chlorination Byproducts," in *Encyclopedia of Toxicology: Third Edition*, Elsevier, 2014, p. 855–859.

[26] I. Woodard & Curran, "Waste Characterization," in *Industrial Waste Treatment Handbook*, Elsevier, 2006, p. 83–126.

[27] V. Lendave, "How can SMOTE technique improve the performance of weak learners?," 22 January 22. [Online]. Available: https://analyticsindiamag.com/how-can-smote-technique-improve-the-performance-of-weak-learners/. [Accessed 28 May 2022].

[28] D. Nelson, "Gradient Boosting Classifiers in Python with Scikit-Learn.," [Online]. Available: https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/. [Accessed 28 May 2022].

[29] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush and A. Gulin, "CatBoost: unbiased boosting with categorical features," 2017. [Online]. Available: https://github.com/CatBoost/CatBoost.

[30] J. Brownlee, "Boosting and AdaBoost for Machine Learning," 25 April 2016. [Online]. Available: https://machinelearningmastery.com/boosting-and-AdaBoost-for-machine-learning/. [Accessed 28 May 2022].

[31] A. Suresh, "What is a confusion matrix?," 17 November 2020. [Online]. Available: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5. [Accessed 28 May 2022].

[32] K. Chan, "What is a ROC Curve and How to Interpret It," [Online]. Available: https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/. [Accessed 28 May 2022].

[33] K. H. Zou, A. J. O'Malley and L. Mauri, "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models," *Circulation*, vol. 115, no. 5, p. 654–657, 2007.

[34] A. Bhandari, "AUC-ROC Curve in Machine Learning Clearly Explained," 16 June 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/. [Accessed 28 May 2022].