

**EVALUASI METODE *RANDOM FOREST*, *XGBOOST* DAN C5.0 DALAM
KLASIFIKASI KUALITAS AIR BERSIH UNTUK Mendukung
PENGELOLAAN SUMBER DAYA AIR**

(Skripsi)

Oleh

**MELAN CANIADI
NPM 2017051031**



**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024**

ABSTRAK

EVALUASI METODE *RANDOM FOREST*, *XGBOOST* DAN C5.0 DALAM KLASIFIKASI KUALITAS AIR BERSIH UNTUK Mendukung PENGELOLAAN SUMBER DAYA AIR

Oleh

MELAN CANIADI

Air bersih adalah kebutuhan dasar manusia yang penting untuk kehidupan sehari-hari dan kesehatan. Namun, banyak orang di seluruh dunia masih kekurangan akses air bersih dan sanitasi yang layak. Penelitian ini bertujuan untuk mengklasifikasikan kualitas air bersih menggunakan metode *Random Forest*, *XGBoost*, dan C5.0, guna mendukung pengelolaan sumber daya air. Data yang digunakan berasal dari *Kaggle*, mencakup 971 data kualitas air dari 62 titik lokasi sungai di Amerika Serikat dari tahun 1995 hingga 2014, pembagian data menggunakan metode *hold out* dan *stratified k-fold cross-validation*. Hasil penelitian menunjukkan bahwa metode *Random Forest* dengan menggunakan *hold out* memiliki akurasi tertinggi sebesar 0.979 dengan waktu eksekusi 429.806 ms, dibandingkan dengan *stratified k-fold cross-validation* dengan akurasi 0.977 dan waktu eksekusi 8584.102 ms. *XGBoost* dan C5.0 menunjukkan akurasi tertinggi sebesar 0.966 dengan *stratified k-fold cross-validation*, meskipun waktu eksekusi lebih lama dibandingkan dengan metode *hold out*. Akurasi *XGBoost* dengan *hold out* adalah 0.964 dengan waktu eksekusi 315.998 ms, sedangkan C5.0 memiliki akurasi 0.960 dengan waktu eksekusi 62.28 ms.

Kata Kunci: *Machine Learning*, *Random Forest*, *Extreme Gradien Boosting*, C5.0, Kualitas Air, Klasifikasi.

ABSTRACT

EVALUATION OF RANDOM FOREST, XGBOOST AND C5.0 METHODS IN CLEAN WATER QUALITY CLASSIFICATION TO SUPPORT WATER RESOURCES MANAGEMENT

By

MELAN CANIADI

Clean water is a basic human need that is important for daily life and health. However, many people around the world still lack access to clean water and proper sanitation. This research aims to classify clean water quality using the Random Forest, XGBoost, and C5.0 methods, to support water resource management. The data used comes from Kaggle, includes 971 water quality data from 62 river locations in the United States from 1995 to 2014, data division uses the hold out and stratified k-fold cross-validation methods. The research results show that the Random Forest method using hold out has the highest accuracy of 0.979 with an execution time of 429,806 ms, compared to stratified k-fold cross-validation with an accuracy of 0.977 and an execution time of 8584,102 ms. XGBoost and C5.0 show the highest accuracy of 0.966 with stratified k-fold cross-validation, although the execution time is longer compared to the hold out method. The accuracy of XGBoost with hold out is 0.964 with an execution time of 315.998 ms, while C5.0 has an accuracy of 0.960 with an execution time of 62.28 ms.

Keywords: Machine Learning, Random Forest, Extreme Gradient Boosting, C5.0, Water Quality, Classification.

**EVALUASI METODE *RANDOM FOREST*, *XGBOOST* DAN C5.0 DALAM
KLASIFIKASI KUALITAS AIR BERSIH UNTUK Mendukung
PENGELOLAAN SUMBER DAYA AIR**

Oleh

MELAN CANIADI

Skripsi

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA ILMU KOMPUTER

Pada

**Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam**



**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024**

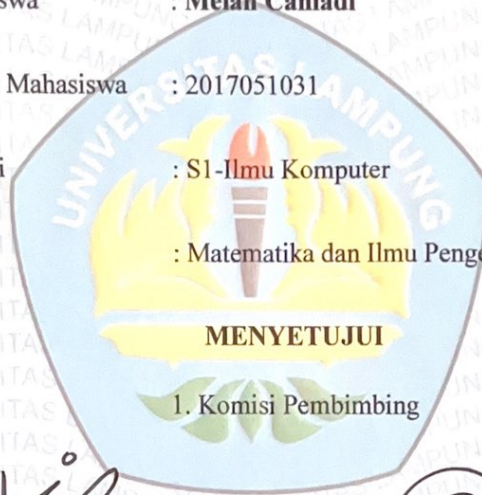
Judul Skripsi : **EVALUASI METODE *RANDOM FOREST*,
XGBOOST DAN *C5.0* DALAM KLASIFIKASI
KUALITAS AIR BERSIH UNTUK
MENDUKUNG PENGELOLAAN SUMBER
DAYA AIR**

Nama Mahasiswa : **Melan Caniadi**

Nomor Pokok Mahasiswa : 2017051031

Program Studi : **S1-Ilmu Komputer**

Fakultas : **Matematika dan Ilmu Pengetahuan Alam**



1. Komisi Pembimbing

Dewi Asiah Shofiana, S.Komp., M.Kom.
NIP. 199509292020122030

Ridho Sholehurrohman, M. Mat.
NIK. 232111970128101

2. Ketua Jurusan Ilmu Komputer

Dwi Sakethi, S.Si., M.Kom.
NIP. 196806111998021001

MENGESAHKAN

1. Tim Penguji

Ketua : Dewi Asiah Shofiana, S.Komp., M.Kom.

Sekretaris : Ridho Sholehurrohman, M. Mat.

Penguji

Bukan Pembimbing : Dr. Aristoteles, S.Si., M.Si.

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam



Dr. Eng. Heri Satria, S.Si., M.Si.

NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: 4 Juli 2024

PERNYATAAN

Saya yang bertanda tangan di bawah ini:

Nama : Melan Caniadi

NPM : 2017051031

Menyatakan bahwa skripsi saya yang berjudul "**Evaluasi Metode *Random Forest*, *Xgboost* dan C5.0 Dalam Klasifikasi Kualitas Air Bersih Untuk Mendukung Pengelolaan Sumber Daya Air**" merupakan karya saya sendiri dan bukan karya orang lain. Seluruh tulisan yang tertulis dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila di kemudian hari terbukti skripsi saya merupakan hasil penjiplakan atau dibuat orang lain, maka saya bersedia menerima sanksi berupa pencabutan gelar yang telah saya terima.

Bandar Lampung, 4 Juli 2024



Melan Caniadi
NPM. 2017051031

RIWAYAT HIDUP



Penulis bernama Melan Caniadi bertempat lahir di Way Mengaku pada tanggal 12 Maret 2002, sebagai anak ketiga dari empat bersaudara. Penulis menyelesaikan pendidikan formal di SD Negeri 3 Way Mengaku dan selesai pada tahun 2014. Kemudian melanjutkan pendidikan menengah pertama di SMP Negeri 1 Liwa yang diselesaikan pada tahun 2017, lalu melanjutkan ke pendidikan menengah atas di SMA Negeri 1 Liwa yang diselesaikan pada tahun 2020.

Pada tahun 2020 penulis terdaftar sebagai mahasiswa Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur PMPAP. Selama menjadi mahasiswa, penulis melakukan beberapa kegiatan antara lain.

1. Menjadi anggota Adapter Himpunan Mahasiswa Jurusan Ilmu Komputer pada periode 2020/2021.
2. Menjadi anggota Biro Kesekretariatan Himpunan Mahasiswa Jurusan Ilmu Komputer pada priode 2020/2021.
3. Menjadi anggota Biro Kesekretariatan Himpunan Mahasiswa Jurusan Ilmu Komputer pada priode 2021/2022.
4. Menjadi Asisten Dosen Jurusan Ilmu Komputer pada mata kuliah Sistem Operasi dan Basis Data tahun 2022, serta mata kuliah Pemrosesan Data Terdistribusi tahun 2023.
5. Menjadi anggota Divisi LCT (Lomba Cepat Tepat) pada acara Pekan Raya Jurusan Ilmu Komputer pada tahun 2021.

6. Menjadi Bendahara Pelaksana pada acara Pekan Raya Jurusan Ilmu Komputer pada tahun 2022.
7. Melaksanakan Kerja Praktik di PT Jasa Raharja Putera Cabang Bandar Lampung pada periode I tahun 2023.
8. Mengikuti *Course UI/UX Designer* Pemula pada Program Kredensial Mikro Mahasiswa Indonesia (KMMI) pada tahun 2021.
9. Melaksanakan KKN di Desa Tanjung Agung, Kecamatan Teluk Pandan, Kabupaten Pesawaran pada periode II tahun 2023.

MOTTO

“Cukuplah Allah (menjadi penolong) bagi kami dan dia sebaik-baik pelindung.”
(QS. Ali-Imran: 173)

“Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya.”
(Q.S Al-Baqarah: 286)

“Karena sesungguhnya sesudah kesulitan itu ada kemudahan.”
(Q.S Al-Insyirah: 5)

PERSEMBAHAN

Alhamdulillahirobbilalamin

Puji dan syukur tercurahkan kepada Allah Subhanahu Wa Ta'ala atas segala Rahmat dan Karunia-Nya sehingga saya dapat menyelesaikan skripsi ini. Shalawat serta salam selalu tercurahkan kepada Nabi Muhammad Shallallahu Alaihi Wasallam.

Kupersembahkan karya ini kepada:

Kedua Orang Tuaku Tercinta

Atas segala pengorbanan, perjuangan, kasih sayang, perhatian, dukungan dan do'a yang selalu menyertaiku. Kuucapkan terima kasih sebesar-besarnya karena telah mendidik dan membesarkanku dengan penuh kasih sayang yang tak akan terbalaskan. Kuucapkan juga terima kasih kepada kakak dan adikku atas dukungan dan do'a yang diberikan kepadaku.

Seluruh Keluarga Besar Ilmu Komputer 2020

Yang senantiasa memberikan semangat dan dukungan.

Almamater Tercinta, Universitas Lampung dan Jurusan Ilmu Komputer

Tempat bernaung mengemban semua ilmu untuk menjadi bekal kehidupan.

SANWACANA

Puji syukur kehadiran Allah *Subhanahu Wa Ta'ala*, karena telah memberikan limpahan nikmat, rahmat dan karunia-Nya. Shalawat serta salam semoga senantiasa tercurahkan kepada junjungan Nabi Muhammad SAW, sehingga penulis dapat menyelesaikan skripsi yang berjudul “**Evaluasi Metode *Random Forest*, *Xgboost* dan *C5.0* Dalam Klasifikasi Kualitas Air Bersih Untuk Mendukung Pengelolaan Sumber Daya Air**” dengan baik dan lancar.


Selesainya skripsi ini tidak terlepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, ucapan terima kasih ditujukan kepada:

1. Bak dan Mak yang selalu mendoakan yang terbaik, memberi dukungan, kasih sayang dan selalu memberikan semangat baik secara moral maupun material dalam menyelesaikan skripsi ini.
2. Saudara dan saudari penulis, Rinaldi, Citra Triadi dan Aditio Rahman yang selalu memberikan dukungan dan doa dalam menyelesaikan skripsi.
3. Ibu Dewi Asiah Shofiana, S.Komp., M.Kom selaku pembimbing utama dalam penelitian ini yang senantiasa memberikan arahan, ilmu dan saran serta motivasi dalam menyelesaikan penelitian ini.
4. Bapak Ridho Sholehurrohman, M. Mat selaku pembimbing kedua dalam penelitian ini yang selalu memberikan, ide, kritik dan saran sehingga skripsi ini dapat diselesaikan dengan baik.
5. Bapak Dr. Aristoteles, S.Si., M.Si. sebagai pembahas yang telah memberikan masukan serta saran yang bermanfaat dalam perbaikan skripsi ini.
6. Bapak Dr. Eng. Heri Satria, S.Si., M.Si. selaku dekan FMIPA Universitas Lampung.

7. Bapak Dwi Sakethi, S.Si., M.Kom selaku dosen pembimbing akademik dan Ketua Jurusan Ilmu Komputer FMIPA Universitas Lampung yang telah memberikan arahan dan bimbingan hingga penelitian ini selesai.
8. Ibu Anie Rose Irawati, S.T. M.Cs selaku Sekretaris Jurusan Ilmu Komputer FMIPA Universitas Lampung.
9. Seluruh Dosen, Staf, dan Karyawan Jurusan Ilmu Komputer yang telah memberikan ilmu, pelajaran, dan bantuan terbaik selama penulis menempuh pendidikan di Jurusan Ilmu Komputer Universitas Lampung.
10. Teman seperjuangan semasa kuliah Aura Husnaini P.Z, Yulia Dwi Putri, Ages Mahesa, Dita Faradila, dan Pynka Aryani Angelia Haryanto yang selalu mendukung, menemani, dan berbagi cerita indah selama masa perkuliahan.
11. Sahabat penulis Melisa Siti Febiane Ad'ha dan Dianita Oktariani yang selalu memberikan semangat kepada penulis untuk menyelesaikan studi ini.
12. Akbar Ferdian Maulana yang selalu menemani, membantu, memberikan dukungan dan semangat kepada penulis untuk menyelesaikan skripsi ini.
13. Teman-teman Himakom yang sudah mengajarkan banyak hal dalam berorganisasi dan memberikan pengalaman yang berharga.
14. Keluarga Ilmu Komputer 2020 yang telah memberikan pengalaman yang sangat berarti selama menjalankan studi di Jurusan Ilmu Komputer Universitas Lampung.
15. Seluruh pihak yang terlibat secara langsung maupun tidak langsung, atas dukungannya dalam menyelesaikan skripsi.

Penulis menyadari bahwa penyusunan skripsi ini masih jauh dari kata sempurna. Namun penulis sangat mengharapkan skripsi ini dapat bermanfaat bagi para civitas akademik Universitas Lampung pada umumnya dan mahasiswa Ilmu Komputer pada khususnya.

Bandar Lampung, 4 Juli 2024



Melan Caniadi
NPM. 2017051031

DAFTAR ISI

	Halaman
DAFTAR ISI	xiv
DAFTAR TABEL	xvii
DAFTAR GAMBAR	xix
DAFTAR KODE PROGRAM	ix
I. PENDAHULUAN	10
1.1 Latar Belakang.....	10
1.2 Rumusan Masalah	24
1.3 Batasan Masalah.....	24
1.4 Tujuan Penelitian.....	24
1.5 Manfaat Penelitian.....	25
II. TINJAUAN PUSTAKA	26
2.1 Penelitian Terdahulu.....	26
2.2 Air.....	30
2.3 Standar Kualitas Air	31
2.4 Metode CCME WQI.....	31
2.5 <i>Machine Learning</i>	35
2.5.1 <i>Supervised Learning</i>	36
2.5.2 <i>Unsupervised Learning</i>	36
2.6 Klasifikasi.....	37
2.7 <i>Hold out</i>	37
2.8 <i>Stratified K-fold Cross Validation</i>	38
2.9 Metode <i>Random Forest</i>	39

2.10	Metode <i>Extreme Gradient Boosting (XGBoost)</i>	41
2.11	Metode C5.0.....	43
2.12	<i>Confusion Matrix</i>	44
2.12.1	<i>Accuracy</i>	46
2.12.2	<i>Precision</i>	46
2.12.3	<i>Recall</i>	46
2.12.4	<i>F1 Score</i>	47
III.	METODE PENELITIAN	48
3.1	Tempat dan Waktu Penelitian	48
3.1.1	Tempat Penelitian.....	48
3.1.2	Waktu Penelitian	48
3.2	Data dan Alat.....	50
3.2.1	Data	50
3.2.2	Perangkat Penelitian.....	52
3.3	Metode.....	54
3.3.1	<i>Literature Review</i>	55
3.3.2	<i>Data Collection</i>	55
3.3.3	<i>Data Preprocessing</i>	55
3.3.4	<i>Data Split</i>	56
3.3.5	<i>Classification Modelling</i>	56
3.3.6	<i>Evaluation</i>	56
IV.	HASIL DAN PEMBAHASAN	57
4.1	<i>Import Data</i>	57
4.2	<i>Preprocessing</i>	57
4.2.1	<i>Drop Unnecessary Features</i>	57
4.2.2	<i>Rename Columns</i>	58
4.2.3	<i>New Columns Quality Classification</i>	58
4.2.4	<i>Exploratory Data Analysis (EDA)</i>	59
4.2.5	<i>SMOTE (Synthetic Minority Over-sampling Technique)</i>	69
4.3	Pembagian Data.....	70
4.4	Pemodelan	71
4.4.1	Metode <i>Random Forest</i>	71
4.4.2	Metode <i>XGBoost</i>	77

4.4.3	Metode C5.0.....	81
4.5	Pembahasan.....	86
V.	SIMPULAN DAN SARAN.....	91
5.1	Simpulan.....	91
5.2	Saran.....	92
	DAFTAR PUSTAKA.....	93

DAFTAR TABEL

Tabel	Halaman
1. Tinjauan pustaka dalam penelitian.....	26
2. Parameter standar kualitas air bersih (Patora & Morley, 2015).....	31
3. Klasifikasi indeks kualitas air CCME WQI (Lumb et al., 2011).....	34
4. Confusion matrix (Tangkelayuk & Mailoa, 2022).....	45
5. Alur waktu pengerjaan penelitian.	49
6. Penjelasan atribut pada <i>dataset</i>	50
7. Perbandingan setelah dilakukan <i>oversampling</i>	70
8. Evaluasi <i>confusion matrix Random Forest</i> dengan <i>hold out</i>	72
9. Perhitungan setiap <i>class Random Forest</i> dengan <i>hold out</i>	73
10. Metrik evaluasi <i>hold out</i> metode <i>Random Forest</i>	73
11. Iterasi SKCV pada <i>Random Forest</i>	74
12. Evaluasi <i>confusion matrix Random Forest</i> dengan SKCV.	75
13. Perhitungan setiap <i>class Random Forest</i> dengan SKCV.	76
14. Metrik evaluasi SKCV metode <i>Random Forest</i>	76
15. Evaluasi <i>confusion matrix XGBoost</i> dengan <i>hold out</i>	78
16. Perhitungan setiap <i>class XGBoost</i> dengan <i>hold out</i>	78
17. Metrik evaluasi <i>hold out</i> metode <i>XGBoost</i>	78
18. Iterasi SKCV pada <i>XGBoost</i>	79
19. Evaluasi <i>confusion matrix XGBoost</i> dengan SKCV.	80
20. Perhitungan setiap <i>class XGBoost</i> dengan SKCV.	80
21. Metrik evaluasi SKCV metode <i>XGBoost</i>	81
22. Evaluasi <i>confusion matrix C5.0</i> dengan <i>hold out</i>	82
23. Perhitungan setiap <i>class C5.0</i> dengan <i>hold out</i>	83
24. Metrik evaluasi <i>hold out</i> metode C5.0.	83

25. Iterasi SKCV pada C5.0.....	84
26. Evaluasi <i>confusion matrix</i> C5.0 dengan SKCV.....	85
27. Perhitungan setiap <i>class</i> C5.0 dengan SKCV.....	85
28. Metrik evaluasi SKCV metode C5.0.....	86
29. Perbandingan tiga metode dengan <i>holdout</i>	87
30. Perbandingan tiga metode dengan SKCV.....	88
31. <i>Runtime</i> menggunakan pembagian data metode <i>hold out</i>	89
32. <i>Runtime</i> menggunakan pembagian data metode SKCV.....	90

DAFTAR GAMBAR

Gambar	Halaman
1. Metode <i>hold out</i> (Ghazvini et al., 2014).	38
2. Metode <i>stratified k-fold cross validation</i> (Muller, 2020).....	38
3. Contoh <i>Random Forest</i> (Yang et al., 2019).	39
4. Alur kerja penelitian.....	54
5. Distribusi fitur numerik.....	59
6. <i>Heatmap dataset</i> kualitas air.....	60
7. Grafik <i>bar</i> atribut <i>quality</i>	61
8. Plot <i>violin</i> atribut <i>fecal</i>	62
9. Plot <i>violin</i> atribut <i>oxygen</i>	63
10. Plot <i>violin</i> atribut <i>pH</i>	63
11. Plot <i>violin</i> atribut <i>tot_sediment</i>	64
12. Plot <i>violin</i> atribut <i>temperature</i>	65
13. Plot <i>violin</i> atribut <i>nitrogen</i>	65
14. Plot <i>violin</i> atribut <i>phosphorus</i>	66
15. Plot <i>violin</i> atribut <i>turbidity</i>	67
16. Plot <i>scatter tot_sediment</i> dan <i>turbidity</i>	67
17. Plot <i>scatter tot_sediment</i> dan <i>phosphorus</i>	68
18. Plot <i>scatter phosphorus</i> dan <i>turbidity</i>	68
19. Plot <i>scatter pH</i> dan <i>oxygen</i>	69
20. <i>Oversampling</i> dengan SMOTE.....	70
21. <i>Confusion matrix Random Forest</i> dengan <i>hold out</i>	72
22. <i>Confusion matrix Random Forest</i> dengan SKCV.....	75
23. <i>Confusion matrix XGBoost</i> dengan <i>hold out</i>	77
24. <i>Confusion matrix XGBoost</i> dengan SKCV.	80

25. <i>Confusion matrix</i> C5.0 dengan <i>hold out</i>	82
26. <i>Confusion matrix</i> C5.0 dengan SKCV.	85

DAFTAR KODE PROGRAM

Kode Program	Halaman
1. <i>Import data</i>	57
2. <i>Drop unnecessary features</i>	58
3. <i>Rename columns</i>	58
4. <i>New columns quality classification</i>	59
5. <i>Metode hold out</i>	71
6. <i>Metode stratified k-fold cross-validation</i>	71

I. PENDAHULUAN

1.1 Latar Belakang

Air bersih merupakan kebutuhan penting dalam kehidupan manusia dan merupakan sumber daya alam yang mempunyai fungsi yang sangat penting. Air bersih dimanfaatkan manusia untuk kebutuhan sehari-hari mulai dari minum, mandi, memasak, mencuci dan keperluan lainnya (Zulhilmi et al., 2019). Air bersih dan sanitasi yang layak merupakan kebutuhan dasar setiap manusia. Dalam *Sustainable Development Goals* (SDGs) poin keenam adalah ketersediaan dan pengelolaan air bersih dan sanitasi yang aman merupakan salah satu tujuan keberlangsungan bagi seluruh manusia (Muslim et al., 2021).

Menurut laporan UNICEF (*United Nations Children's Fund*) terdapat miliar orang di seluruh dunia terus mengalami penderitaan karena kurangnya akses yang memadai terhadap air bersih, sanitasi dan kebersihan. Sekitar 2,2 miliar orang di seluruh dunia tidak memperoleh layanan air minum yang dikelola dengan aman, sementara 4,2 miliar orang tidak memiliki akses ke fasilitas sanitasi yang dikelola dengan baik. Selain itu, 3 miliar orang tidak dapat mengakses fasilitas dasar untuk mencuci tangan, tidak adanya layanan ini menggambarkan tantangan global yang signifikan dalam mencapai standar kesehatan dan kebersihan yang memadai bagi populasi dunia (UNICEF, 2019).

Penurunan kualitas air bersih hampir 70% disebabkan oleh bakteri *fecal coliform* dan *total coliform* yang berasal dari kotoran manusia dan hewan yang mengandung bakteri *panthogen* berupa *shigella sp*, *escherihia coli*,

vibrio cholerae, *campylobacter jejuni* dan *salmonella*. Fenomena ini menunjukkan bahwa pertumbuhan penduduk memberikan kontribusi negatif terhadap kualitas lingkungan hidup, khususnya kualitas air bersih (Kustanto, 2020).

Data mining adalah bagian dari ilmu kecerdasan buatan yang berkaitan dengan penggalian pola-pola dalam data untuk mengubahnya menjadi informasi yang berharga. *Data mining* melibatkan penggunaan berbagai teknik pembelajaran komputer untuk menganalisis dan mengekstraksi pengetahuan secara otomatis (Hawari et al., 2022). Tugas utama pada *data mining* meliputi klasifikasi, yaitu suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan (Imandasari et al., 2019). Metode klasifikasi *machine learning* telah menjadi perangkat yang sudah lama digunakan dalam berbagai disiplin ilmu, termasuk *data mining* serta berbagai bidang ilmu komputer lainnya (Mambang & Byna, 2017).

Penelitian ini bertujuan untuk mengklasifikasi suatu data yang disusun secara sistematis ke dalam kelompok tertentu, sehingga memungkinkan identifikasi suatu individu pada kelompok tertentu, pada penelitian ini menerapkan perbandingan metode klasifikasi *Random Forest*, *XGBoost* dan *C5.0* untuk menentukan kualitas air bersih. *Random Forest* adalah metode klasifikasi yang terdiri dari beberapa pohon keputusan, di mana setiap pohon keputusan dibangun dengan menggunakan vektor acak. Pendekatan ini umumnya digunakan untuk menyisipkan vektor acak dalam pembentukan pohon – pohon keputusan (Mambang & Byna, 2017) *Extreme Gradient Boosting (XGBoost)* merupakan metode *boosting* dengan menggabungkan beberapa kumpulan pohon keputusan yang akan digunakan dalam pembangunan pohon keputusan selanjutnya (Syukron et al., 2020). Algoritma *C5.0* adalah penyempurnaan dari algoritma sebelumnya yang dibentuk oleh Ross Quinlan pada tahun 1987 yaitu algoritma *ID3* dan *C4.5* (Purba et al., 2022).

Kualitas air yang layak konsumsi bagi masyarakat, perlu adanya identifikasi dini terhadap sumber air baku serta faktor-faktor yang mempengaruhinya. Tujuan dari penelitian ini yaitu berkontribusi untuk menentukan kualitas air bersih sebagai upaya pengelolaan sumber daya air agar lebih baik dengan menggunakan metode *Random Forest*, *XGBoost* dan *C5.0*, sehingga dapat membantu meningkatkan hasil akurasi pada proses klasifikasi serta memperoleh tingkat kualitas air bersih yang baik untuk dikonsumsi.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini berdasarkan pemaparan latar belakang adalah sebagai berikut:

1. Bagaimana mengimplementasikan metode *Random Forest*, *XGBoost* dan *C5.0* dalam mengklasifikasikan kualitas air bersih.
2. Bagaimana hasil analisis kinerja metode *Random Forest*, *XGBoost* dan *C5.0* dalam mengklasifikasikan kualitas air bersih.

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Data yang digunakan yaitu data yang bersumber dari *Kaggle* berjumlah 971 data.
2. Metode klasifikasi yang digunakan adalah metode *Random Forest*, *XGBoost* dan *C5*.

1.4 Tujuan Penelitian

Tujuan pada penelitian ini sebagai berikut:

1. Mengimplementasikan metode *Random Forest*, *XGBoost* dan *C5.0* dalam mengklasifikasikan kualitas air bersih.
2. Membandingkan kinerja metode *Random Forest*, *XGBoost* dan *C5.0* dalam mengklasifikasikan kualitas air bersih.

1.5 Manfaat Penelitian

Manfaat pada penelitian ini adalah sebagai berikut:

1. Mengetahui hasil kinerja klasifikasi kualitas air bersih sebagai upaya pengelolaan sumber daya air menggunakan perbandingan metode *Random Forest*, *XGBoost* dan *C5.0*.
2. Mendukung pencapaian *Sustainable Development Goals* (SDGs) poin keenam yang menetapkan tujuan untuk memastikan ketersediaan dan pengelolaan air bersih serta sanitasi yang aman.

II. TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Dalam melakukan suatu penelitian diperlukan penelitian yang sudah ada sebelumnya, yang berkaitan dengan penelitian tersebut. Referensi atau tinjauan pustaka dalam penelitian ini ditunjukkan pada Tabel 1.

Tabel 1. Tinjauan pustaka dalam penelitian.

No	Penelitian	Data	Metode	Hasil
1	<i>Water Quality Classification Using Machine Learning Algorithms</i> (Nasir et al., 2022).	Jumlah: 1.679 Training: 75% Testing: 25%	<i>SVM, Random Forest, Multi Layer Perceptron (MLP), Logistic Regression, XGBoost, Decision Tree dan CATBoost</i>	SVM - Akurasi: 80.7% - Precision: 81.3% - Recall: 80.7% - F1 Score: 80.6% Random Forest - Akurasi: 94% - Precision: 94% - Recall: 94% - F1 Score: 94% MLP - Akurasi: 88.6% - Precision: 88.9% - Recall: 88.6% - F1 Score: 88.6%

No	Penelitian	Data	Metode	Hasil
				<i>Logistic Regression</i> - Akurasi: 72.9% - Precision: 72.4% - Recall: 72.9% - F1 Score: 72.4%
				<i>XGBoost</i> - Akurasi: 88.1% - Precision: 88.3% - Recall: 88.1% - F1 Score: 88%
				<i>Decision Tree</i> - Akurasi: 81.6% - Precision: 81.7% - Recall: 81.6% - F1 Score: 81.5%
				<i>CATBoost</i> - Akurasi: 94.5% - Precision: 94.5% - Recall: 94.5% - F1 Score: 94.5%
2	Klasifikasi Kualitas Air Menggunakan Metode KNN, Naive Bayes Dan Decision Tree (Tangkelayuk & Mailoa, 2022).	Jumlah: 2.081 Training: 70% Testing: 30%	KNN, naive bayes dan decision tree	Akurasi KNN: 86.88% Akurasi Naive Bayes: 63.60% Akurasi Decision Tree: 80.84%

No	Penelitian	Data	Metode	Hasil
3	Klasifikasi Kualitas Air Sumur Menggunakan Algoritma <i>Random Forest</i> (Mutoffar et al., 2022).	Jumlah: 267 <i>Training</i> : 80% <i>Testing</i> : 20%	<i>Random Forest</i>	Presisi: 82% Sensitivitas (<i>Recall</i>): 83%

Penelitian terdahulu digunakan sebagai acuan dan perbandingan dalam penelitian ini. Beberapa penelitian yang digunakan adalah sebagai berikut:

1. *Water Quality Classification Using Machine Learning Algorithms*

Penelitian terdahulu yang dilakukan oleh Nasir et al., (2022) dalam penelitian *Water Quality Classification Using Machine Learning Algorithms*. Data kualitas air minum ini dikumpulkan dari berbagai negara bagian di India antara tahun 2005 dan 2014. Sebanyak 1.679 sampel dikumpulkan dan dianalisis. Atribut yang digunakan yaitu *oxygen (DO)*, *pH*, *conductivity*, *biochemical oxygen demand (BOD)*, *nitrate*, *fecal coliform*, dan *total coliform*.

Klasifikasi *machine learning* yang digunakan dalam penelitian ini adalah *Support Vector Machine*, *Random Forest*, *Logistic Regression*, *CATBoost*, *XGBoost*, *Decision Tree*, dan *Multi-Layer Perceptron*. Dalam hal ini metode *CATBoost* memiliki nilai akurasi paling tinggi yaitu 94,5%, sedangkan metode *Random Forest* 94%, *MLP* 88,6%, *XGBoost* 88,1%, *Decision Tree* 81,6%, *SVM* 80,7% dan *Logistic Regression* 72,9%.

2. Klasifikasi Kualitas Air Menggunakan Metode *KNN*, *Naive Bayes* Dan *Decision Tree*

Penelitian yang terdahulu dilakukan oleh Tangkelayuk & Malloa, (2022) penelitian kualitas air ini menggunakan *dataset* kualitas air dengan tiga algoritma yaitu *K-Nearest Neighbors*, *Naive Bayes* dan *Decision Tree*. *Dataset Water Quality* diperoleh dari situs *Kaggle* dengan jumlah data yang digunakan 2.081 baris data dengan sepuluh atribut yaitu *ph*, *hardness*, *solids*, *chloramines*, *sulfate*, *conductivity*, *organic carbon*, *trihalomethanes*, *turbidity* dan *potability*.

Analisis perbandingan akurasi *Water Quality* menggunakan data hasil klasifikasi metode *K-Nearest Neighbors*, *Naive Bayes* dan *Decision Tree* terlihat bahwa *K-Nearest Neighbors* merupakan metode yang memiliki tingkat akurasi paling tinggi yaitu 86.88% untuk klasifikasi data kualitas air yang digunakan pada penelitian ini, sedangkan metode *Naive Bayes* sebesar 63.60% dan metode *Decision Tree* sebesar 80.84%.

3. Klasifikasi Kualitas Air Sumur Menggunakan Algoritma *Random Forest*

Penelitian yang terdahulu dilakukan oleh Mutoffar et al., (2022) penelitian ini menggunakan data dalam pengambilan air sumur di Provinsi DKI Jakarta pada tahun 2017 oleh Dinas Lingkungan Hidup Provinsi DKI Jakarta. *Dataset* yang digunakan dalam penelitian ini mencakup kualitas air sumur pada periode ke-2, terdiri dari 267 data, dengan atribut-atribut seperti parameter, nilai dan indeks pencemaran. Selanjutnya, dalam tahap pengolahan data dilakukan transformasi data. Data yang awalnya berada dalam atribut "parameter" diubah menjadi atribut tersendiri dan nilai "indeks pencemaran" diubah menjadi bernilai 1 jika indeks pencemaran kurang dari atau sama dengan 1.00 menandakan bahwa kualitas air memenuhi standar. Sebaliknya jika nilai

"indeks pencemaran" diubah menjadi bernilai 0 jika indeks pencemaran lebih dari 1.00 menandakan bahwa kualitas air tidak memenuhi standar. Dengan melakukan transformasi ini, data awal yang menggambarkan kualitas air dapat lebih mudah dianalisis dan diinterpretasikan dalam konteks apakah air tersebut memenuhi standar atau tidak berdasarkan nilai indeks pencemaran yang telah ditentukan.

Kualitas air sumur di Jakarta diklasifikasikan menggunakan algoritma *Random Forest*, dengan pembagian data sebanyak 80% digunakan untuk *training* dan 20% untuk *testing*. Hasil dari penelitian menunjukkan bahwa algoritma ini menghasilkan *presisi* sebesar 0.823 dan sensitivitas sebesar 0.83, yang menunjukkan kemampuan cukup baik dalam memprediksi apakah air tersebut dapat dikonsumsi atau tidak. Lebih spesifiknya model ini dapat memprediksi dengan akurasi sekitar 82% dari data yang diuji dengan 83% dari data yang diklasifikasikan dengan benar ke dalam kategori air yang dapat dikonsumsi atau tidak.

2.2 Air

Air adalah cairan yang tidak memiliki warna, aroma atau rasa dengan rumus molekul H_2O . Karakteristik air yang paling mencolok adalah sifat polaritas yang membuatnya menjadi pelarut yang sangat efektif untuk berbagai jenis zat. Molekul air saling terikat oleh ikatan hidrogen dan pada kondisi standar, yaitu pada tekanan 100 kPa atau 1 bar, air memiliki titik beku pada 273,15 K ($0^{\circ}C$) dan titik didih pada 373,15 K ($100^{\circ}C$). Air dianggap sebagai pelarut universal karena mudah untuk mencampur dengan banyak zat kimia lainnya. Terdapat dua jenis zat yang dapat larut dalam air, yang pertama adalah zat hidrofilik dapat dengan mudah larut dalam air, seperti garam, gula, beberapa asam, beberapa gas dan berbagai molekul organik. Kemudian yang kedua adalah zat hidrofobik yang memiliki kesulitan dalam melarutkan air seperti lemak dan minyak (Ritonga, 2011).

2.3 Standar Kualitas Air

Kualitas air adalah sumber daya alam yang memegang peranan vital dalam kehidupan manusia dan perkembangan masyarakat. Standar kualitas air pada negara bagian Washington Amerika Serikat didasarkan pada peraturan WAC 173-201A (Patora & Morley, 2015). Standar kualitas air bersih dapat dilihat pada Tabel 2.

Tabel 2. Parameter standar kualitas air bersih (Patora & Morley, 2015).

No	Jenis Parameter	Satuan	Standar Kualitas Air <i>Acute/Chronic/Human Health</i>
1	<i>pH</i>	mg/L	6,5 - 8,5
2	<i>Temperature</i>	⁰ C	Max 17.5
3	<i>Dissolved Oxygen</i>	mg/L	8
4	<i>Fecal Coliform</i>	mL	200/100/0
5	<i>Total Phosphorus</i>	mg/L	10
6	<i>Total Suspended Sediment (TSS)</i>	mg/L	100
7	<i>Nitrogen</i>	mg/L	300
8	<i>Turbidity</i>	NTU	5

2.4 Metode CCME WQI

Sejumlah ilmuwan menciptakan sebuah metode untuk mengubah berbagai parameter kualitas air yang berjumlah banyak menjadi satu nilai tunggal. Salah satu metode yang dikembangkan adalah metode *CCME* (*Canadian Council of Ministers of the Environment*) yang dikembangkan oleh negara Canada (Romdania dkk., 2018). Penggunaan metode CCME WQI menghasilkan representasi yang lebih akurat tentang kondisi aktual kualitas air dibandingkan dengan nilai Indeks Pencemaran (*IP*) dan *Storet*. Metode ini juga lebih praktis dan mudah di aplikasikan (Saraswati et al., 2019). Metode CCME WQI memiliki tingkat efektivitas dan sensitivitas yang lebih tinggi dibandingkan dengan metode *IP* dan *Storet* (Romdania et al., 2018). Berikut adalah tahapan perhitungan CCME WQI (Lumb et al., 2011).

1. F1 (*Scope*), menyatakan persentase perbedaan antara variabel-variabel dengan baku mutu yang telah ditetapkan dapat ditemukan dalam rumus F1 yang tercantum dalam Persamaan 1.

$$F1 = \frac{\text{Number of failed variables}}{\text{Total number of variables}} \times 100 \quad (1)$$

2. F2 (*Frequency*), menyatakan presentase dari setiap parameter uji yang tidak memenuhi baku mutu dapat diidentifikasi melalui rumus F2 yang terdapat pada Persamaan 2.

$$F2 = \frac{\text{Number of failed test}}{\text{Total number of test}} \times 100 \quad (2)$$

3. F3 (*Amplitude*), menyatakan jumlah nilai penyimpangan. Jumlah konsentrasi yang lebih besar atau kurang dari, jika yang dicari adalah nilai minimum dari baku mutu disebut *excursion*. Untuk menghitung F3 terdapat 3 tahapan dengan rumus sebagai berikut.

- a. Jika nilai uji tidak boleh melebihi baku mutu terdapat pada Persamaan 3.

$$\text{Excursion}_i = \frac{\text{Failed Test Value}_i}{\text{Objective}_j} - 1 \quad (3)$$

- b. Jika nilai uji tidak boleh kurang dari baku mutu terdapat pada Persamaan 4.

$$\text{Excursion}_i = \frac{\text{Objective}_j}{\text{Failed Test Value}_i} - 1 \quad (4)$$

- c. Menjumlahkan nilai excursion dan membaginya dengan total tes terdapat pada persamaan 5.

$$nse = \frac{\sum_{i=1}^n excursion_i}{number\ of\ test} \quad (5)$$

Diketahui:

nse = Normalized sum of excursions

- d. F3 kemudian dihitung dengan fungsi asimtotik dengan skala jumlah dari nse dengan kisaran harga antara 0 sampai 100. Rumus dari F3 terdapat pada Persamaan 6.

$$F3 = \frac{nse}{0.01\ nse + 0.01} \quad (6)$$

4. CCME WQI jika nilai faktor telah diperoleh maka dapat dihitung menggunakan rumus pada Persamaan 7.

$$CCME\ WQI = 100 - \left[\frac{\sqrt{F1^2 + F2^2 + F3^2}}{1,732} \right] \quad (7)$$

Diketahui:

$$1,732 = \sqrt{3}$$

Hasil perhitungan nilai CCME WQI dapat dikategorikan berdasarkan Tabel 3.

Tabel 3. Klasifikasi indeks kualitas air CCME WQI (Lumb et al., 2011).

Nilai CCME WQI	Kualitas Air		Rekomendasi
	Tingkat	Kelas	
95 - 100	1	Sangat baik (Excellent)	Layak sebagai media hidup biota perairan dan mendekati kondisi alaminya, air ini dapat digunakan sebagai sumber air untuk berbagai keperluan. Kualitas air terjaga dengan baik dan tidak ada gangguan atau ancaman yang signifikan. Indeks nilai ini dapat diperoleh ketika semua pengukuran standar kualitas memiliki tujuan yang konsisten sepanjang tahun.
80 - 94	2	Baik (<i>Good</i>)	Layak sebagai media hidup biota perairan, Sumber air minum ini memerlukan tahap pengolahan awal, disertai dengan perlindungan kualitas air yang baik dan minim risiko ancaman atau gangguan. Selain itu, kondisinya jarang mengalami penyimpangan dari kondisi alamiahnya atau tujuan penggunaan yang ditentukan.

Nilai CCME WQI	Kualitas Air		Rekomendasi
	Tingkat	Kelas	
65 - 79	3	Cukup (<i>Fair</i>)	Tidak sesuai sebagai sumber air minum, namun dalam keadaan terlindungi, terkadang mengalami gangguan atau ancaman dan sesekali mengalami penyimpangan dari kondisi alamiahnya atau tujuan penggunaannya.
45 - 64	4	Kurang (<i>Marginal</i>)	Kualitas airnya sering kali terancam dan terganggu, dengan kondisi yang sering menyimpang dari tingkat alamiahnya atau tujuan penggunaannya.
0 -44	5	Buruk (<i>Poor</i>)	Kualitas airnya hampir selalu mengalami ancaman dan gangguan, serta umumnya kondisinya cenderung menyimpang dari tingkat alamiahnya atau tujuan penggunaannya.

2.5 Machine Learning

Machine learning merupakan bidang ilmu yang mengeksplorasi pola dan teori pembelajaran komputasi dalam *artificial intelligence*. *Machine learning* melibatkan proses pembelajaran dan pengembangan algoritma yang mampu belajar dan melakukan prediksi pada *dataset* (Simon et al., 2016). Kecerdasan *machine learning* dapat dilihat dari kemampuannya untuk secara efektif

menggeneralisasi informasi dari data bar yang belum pernah dipelajari sebelumnya seperti melakukan prediksi, mengklasifikasi, *ranking* dan lain-lain (Abdillah et al., 2015). Dalam klasifikasi *machine learning* terdapat dua tipe yaitu *supervised learning* yang membentuk model dengan mempelajari data latih yang sudah diberi label, sedangkan *unsupervised learning* fokus mempelajari kemiripan dalam data latih yang tidak memiliki label.

2.5.1 *Supervised Learning*

Supervised learning merupakan sebuah metode dalam *machine learning* di mana proses pembelajarannya melibatkan pengawasan. Pada *supervised learning* model dilatih menggunakan data yang telah diberi label, untuk setiap label dikategorikan di setiap titik data ke dalam satu atau beberapa kelompok. Kemudian sistem akan mempelajari bagaimana data yang telah diberi label atau biasa disebut *data training*, selanjutnya *data training* akan memprediksi hasil dari data uji. *Supervised learning* mencakup dua jenis kategori yaitu *regression* dan *classification* (Sidik & Ansawarman, 2022).

2.5.2 *Unsupervised Learning*

Unsupervised learning merupakan sebuah metode dalam *machine learning* di mana proses pembelajarannya tanpa melibatkan pengawasan yang artinya pembelajaran tanpa label. Tujuannya yaitu untuk mengidentifikasi karakteristik yang membuat titik data memiliki kesamaan, seperti pembentukan *cluster* dan pengelompokan data ke dalam *cluster* tersebut. *Unsupervised learning* memiliki beberapa kategori diantaranya adalah *clustering* dan *dimensionality reduction* (Sidik & Ansawarman, 2022).

2.6 Klasifikasi

Klasifikasi merupakan proses menemukan sebuah model atau fungsi yang menggambarkan dan membedakan kelas data. Klasifikasi adalah teknik *data mining* yang dapat digunakan saat memprediksi keanggotaan kelompok untuk data *instance* (Prima Wijaya & Muslim, 2016). Klasifikasi bertujuan untuk membuat pola atau mengelompokkan data dari *training* set ke dalam *class* tertentu berdasarkan atribut. Selanjutnya model tersebut digunakan untuk mengklasifikasikan atribut yang kelasnya belum diketahui sebelumnya (Zega, 2014). Data latih (*training*) merupakan data yang tersedia, sementara untuk data uji (*testing*) merupakan data yang sudah diberi label yang siap digunakan untuk menghitung akurasi model klasifikasi yang telah dibuat (Saifudin, 2018).

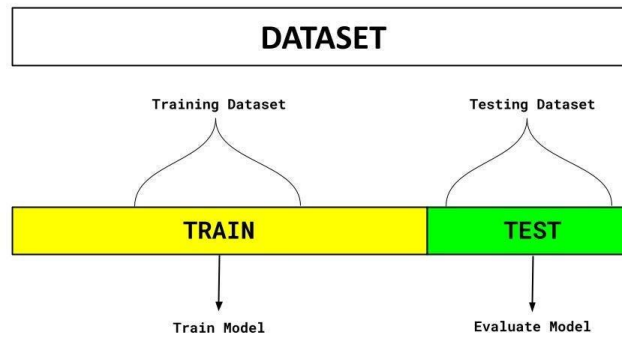
Berikut ini adalah proses klasifikasi (Suwardika et al., 2019).

1. Model dibangun untuk menggambarkan sebuah kumpulan kelas data dari populasi yang telah ditentukan sebelumnya, model tersebut dibangun dengan menganalisa data latih yang digambarkan oleh atribut. Setiap baris di diasumsikan sebagai bagian dari kelas yang telah ditentukan, ditandai oleh salah satu atribut yang disebut sebagai atribut *class label*.
2. Menguji model yang telah dibangun menggunakan data uji untuk mengevaluasi ketepatan atau kinerja model dalam mengklasifikasikan data uji. Setelah pengukuran performa selesai, pengambil keputusan dapat memutuskan apakah akan menggunakan model tersebut atau memilih untuk membuat model baru dengan data latih atau metode yang berbeda guna menghasilkan model klasifikasi yang lebih optimal.

2.7 Hold out

Hold out adalah metode pemecahan data sederhana yang membagi data menjadi dua bagian berupa *data training* dan *data testing*. Dalam teknik ini, *dataset* dengan label kelas telah dibagi menjadi dua bagian, di mana satu

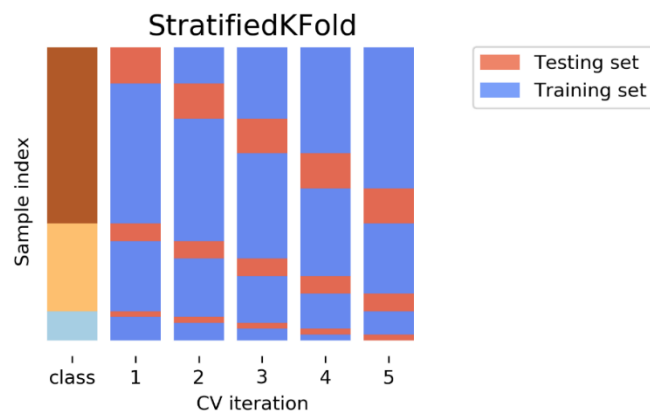
bagian digunakan untuk melatih klasifikasi dan bagian lainnya untuk menguji klasifikasi tersebut (Ghazvini et al., 2014). Berikut ilustrasi metode *hold out* dapat dilihat pada Gambar 1.



Gambar 1. Metode *hold out* (Ghazvini et al., 2014).

2.8 Stratified K-fold Cross Validation

Stratified k-fold cross validation merupakan metode yang digunakan untuk memvalidasi keakuratan suatu model dengan membagi data menjadi *data training* untuk membentuk model dan *data testing* untuk memvalidasi model. Metode *stratified k-fold cross validation* mirip dengan *k fold cross validation* hanya saja pada saat pengacakan data dibagi ke dalam *k fold* dengan komposisi jumlah yang sama untuk setiap *fold*-nya dan begitu juga untuk kategori yang lainnya (Muller, 2020). Berikut ilustrasi metode *stratified k-fold cross validation* dapat dilihat pada Gambar 2.

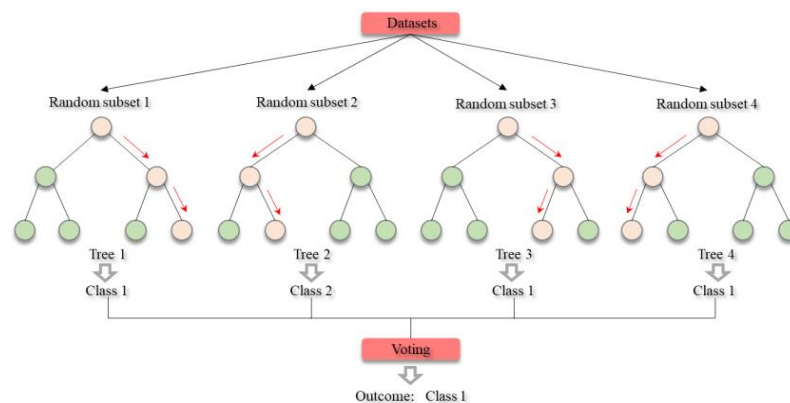


Gambar 2. Metode *stratified k-fold cross validation* (Muller, 2020).

2.9 Metode *Random Forest*

Random Forest merupakan metode klasifikasi yang terdiri dari sejumlah pohon keputusan. Setiap pohon keputusan dibangun dengan memanfaatkan vektor acak. Pendekatan umum yang digunakan dalam menyisipkan vektor acak untuk pembentukan pohon yaitu dengan memilih nilai acak, misalnya jumlah atribut F yang digunakan untuk membagi setiap *node* di pohon keputusan yang sedang dibentuk (Mambang & Byna, 2017). Metode *Random Forest* seringkali digunakan karena menghasilkan tingkat kesalahan yang lebih rendah, memberikan tingkat akurasi yang lebih baik dalam proses klasifikasi, mampu menangani *data training* yang jumlah sangat besar dan efektif dalam mengatasi data yang tidak lengkap (Primajaya & Sari, 2018). Langkah-langkah rinci dari *Random Forest* (Yang et al., 2019) adalah sebagai berikut.

1. Pilih beberapa data di *training* set sebanyak k .
2. Membuat *Decision Tree* menggunakan k data yang telah dipilih sebelumnya.
3. Tentukan jumlah *n-tree* (kumpulan pohon-pohon) yang akan dibuat ulangi langkah 1 dan 2.
4. Setelah terbentuk sejumlah besar pohon, lakukan prediksi pada data baru dengan menggabungkan hasil semua pohon, menggunakan strategi voting mayoritas.



Gambar 3. Contoh *Random Forest* (Yang et al., 2019).

Gambar 3 mengilustrasikan proses kerja metode *Random Forest*. Ketika membentuk pohon klasifikasi, pendekatan yang dilakukan yaitu dengan memisahkan sebuah masalah menjadi beberapa atau sub masalah. Pendekatan ini melibatkan pemisahan keputusan pada simpul teratas menjadi dua simpul, di mana kedua simpul tersebut mencakup pernyataan benar dan salah.

Saat menerapkan algoritma *Random Forest* untuk mengklasifikasikan data, formula indeks *gini* digunakan untuk memutuskan bagaimana *node* pada sebuah cabang pohon keputusan. Persamaan ini menggunakan *class* dan probabilitasnya untuk menghitung nilai *gini* dari setiap cabang pada sebuah simpul dan menentukan cabang mana yang memiliki kemungkinan terjadinya lebih tinggi (Erlin et al., 2022). Indeks *gini* didefinisikan pada Persamaan 8.

$$Gini = 1 - \sum_{i=1}^c (p_i)^2 \quad (8)$$

Diketahui:

P_i = Probabilitas frekuensi kelas ke- i dalam *dataset*

C = Jumlah *class*

Selain menggunakan indeks *gini*, dalam pembentukan pohon keputusan juga dapat dilakukan dengan mengukur nilai *entropy* sebagai indikator tingkat ketidakhurnian atribut (Erlin et al., 2022). Menghitung nilai *entropy* dapat dilakukan menggunakan rumus pada Persamaan 9.

$$Entropy(S) = \sum_{i=1}^c - p_i * \log_2 (p_i) \quad (9)$$

Diketahui:

S = Himpunan *dataset*

C = Jumlah kelas

P_i = Probabilitas frekuensi kelas ke- i dalam *dataset*

2.10 Metode *Extreme Gradient Boosting (XGBoost)*

Metode *XGBoost* adalah sebuah algoritma pengembangan dari *gradient tree boosting* yang menggunakan pendekatan *ensemble*, algoritma ini sangat efektif dalam menangani kasus *machine learning* yang berskala besar. Keunggulan metode *XGBoost* terletak pada fitur tambahan yang membantu mempercepat perhitungan dan mencegah *overfitting*. *XGBoost* dapat digunakan untuk menyelesaikan berbagai jenis masalah termasuk klasifikasi. Algoritma ini merupakan kumpulan pohon keputusan yang terdiri dari berbagai pohon sebelumnya (Yulianti et al., 2022). Dalam penggunaannya *XGBoost* digunakan untuk permasalahan *supervised learning*, di mana modelnya menggunakan *data training* dengan x_i sebagai variabel untuk memprediksi variabel target y_i . Secara matematis dalam menentukan nilai prediksi pada langkah ke (t) pada $\hat{y}_i^{(t)}$ (XGBoost Developers, 2023) dapat didefinisikan pada Persamaan 10.

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (10)$$

Diketahui:

$\hat{y}_i^{(t)}$ = Prediksi pada iterasi ke- t untuk data poin ke- i

t = Jumlah total iterasi

$f_k(x_i)$ = Fungsi prediktor yang dihasilkan oleh model ke- k pada iterasi ke- k , dengan x_i sebagai *input*

$\hat{y}_i^{(t-1)}$ = Prediksi kumulatif hingga iterasi ke- $t-1$

$f_t(x_i)$ = Fungsi prediktor yang dihasilkan oleh model pada iterasi ke- t

Pada *XGBoost* terdapat sebuah fungsi obyektif *training loss* dan *regularization*. Fungsi obyektif didefinisikan pada Persamaan 11.

$$obj(\theta) = L(\theta) + \Omega(\theta) \quad (11)$$

Diketahui:

L = *Training loss function*

Ω = *Regularization*

Training loss yaitu mengukur sejauh mana model dapat memprediksi *data training*. Salah satu opsi umum yang sering digunakan adalah *mean squared error* yang dapat didefinisikan pada Persamaan 12.

$$L(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (12)$$

Diketahui:

$L(\theta)$ = Fungsi *loss*

n = Jumlah total sampel dalam dataset pelatihan

y_i = Nilai target sebenarnya untuk sampel ke- i

\hat{y}_i = Nilai prediksi model untuk sampel ke- i

Selanjutnya dalam mendefinisikan kompleksitas pada *regularization* yaitu mengendalikan kompleksitas model dan mencegah *overfitting*, dapat didefinisikan pada Persamaan 13.

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (13)$$

Diketahui:

Ω = *Omega*

γ = *Gamma*

λ = *Lambda*

T = Jumlah total daun (*leaves*) dalam pohon

w_i = Berat (*weight*) dari daun ke- j

2.11 Metode C5.0

Algoritma C5.0 adalah suatu metode klasifikasi *data mining* yang khususnya digunakan dalam teknik *decision tree*. C5.0 merupakan pengembangan dari algoritma sebelumnya yang dikembangkan oleh Ross Quinlan pada tahun 1987, yaitu ID3 dan C4.5 (Pardede et al., 2019). Secara umum, proses pembuatan pohon pada kedua algoritma tersebut serupa, di mana kedua algoritma melakukan perhitungan *entropi* dan *gain*. Algoritma C4.5 berhenti setelah menghitung *gain*, sementara algoritma C5.0 akan melanjutkan dengan menghitung *gain ratio* berdasarkan nilai *gain* dan nilai *entropy*. Penggunaan nilai *gain ratio* bertujuan untuk menentukan atribut uji yang akan menjadi induk untuk setiap simpul dalam pohon. Atribut yang memiliki nilai *gain ratio* tertinggi akan dipilih sebagai induk untuk simpul berikutnya (Pratama & Andraini, 2022). Menghitung *entropy* dari setiap atribut dapat dilihat pada Persamaan 9. *Gain ratio* digunakan sebagai pembentukan *node* atau akar dan cabang pohon keputusan (Joloudari et al., 2020), *gain ratio* dapat dilihat pada Persamaan 14.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (14)$$

Diketahui:

$$\begin{aligned} A &= \text{Jumlah fitur} \\ Gain(A) &= \text{Information gain} \\ SplitInfo(A) &= \text{Information split} \end{aligned}$$

Berdasarkan Persamaan 14 $SplitInfo(A)$ dapat dihitung menggunakan Persamaan 15.

$$SplitInfo(A) = - \sum_{i=1}^n \left(\frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \right) \quad (15)$$

Diketahui:

n = Jumlah kelas dalam *dataset*

S = Jumlah *dataset* asli

S_i = *Subset* dari *dataset* S yang dibentuk berdasarkan nilai dari atribut

Berdasarkan Persamaan 14 $Gain(A)$ dapat dihitung menggunakan Persamaan 16.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} Entropy(S_i) \quad (16)$$

Diketahui:

$Entropy(S)$ = $Entropy$ dari *dataset* asli

$Entropy(S_i)$ = $Entropy$ subset S_i

2.12 Confusion Matrix

Confusion matrix adalah suatu metode klasifikasi yang mengevaluasi hasil klasifikasi yang telah dilakukan, di mana akurasi klasifikasi menjadi faktor penentu dalam kinerja klasifikasi. *Confusion matrix* memberikan perbandingan antara hasil klasifikasi yang dihasilkan oleh sistem atau model dengan hasil klasifikasi sebenarnya. Pentingnya *confusion matrix* memberikan informasi tentang sejauh mana model yang telah dibuat sebelumnya melalui pengukuran akurasi untuk menilai tingkat keakuratan model yang telah dibuat. *Confusion matrix* menggambarkan performa model klasifikasi berdasarkan serangkaian data uji di mana nilai sebenarnya diketahui. *Confusion Matrix* digunakan untuk menghitung akurasi dan biasanya ditampilkan dalam bentuk tabel (Tangkelayuk & Mailoa, 2022). *Confusion matrix* ditampilkan pada contoh Tabel 4.

Tabel 4. Confusion matrix (Tangkelayuk & Mailoa, 2022).

		<i>Predicted</i>				
		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>Actual</i>	<i>A</i>	TP _A	FP _B FN _A	FP _C FN _A	FP _D FN _A	FP _E FN _A
	<i>B</i>	FP _A FN _B	TP _B	FP _C FN _B	FP _D FN _B	FP _E FN _B
	<i>C</i>	FP _A FN _C	FP _B FN _C	TP _C	FP _D FN _C	FP _E FN _C
	<i>D</i>	FP _A FN _D	FP _B FN _D	FP _C FN _D	TP _D	FP _E FN _D
	<i>E</i>	FP _A FN _E	FP _B FN _E	FP _C FN _E	FP _D FN _E	TP _E

Istilah - istilah yang digunakan pada *confusion matrix* yang tertera pada Tabel 4, diantaranya:

1. *True Negative* merupakan jumlah prediksi negatif yang diklasifikasikan secara akurat.
2. *True Positive* merupakan jumlah prediksi positif yang diklasifikasikan secara akurat.
3. *False Positive* merupakan jumlah prediksi salah yang diklasifikasikan sebagai positif.
4. *False Negative* merupakan jumlah prediksi salah dan diklasifikasikan sebagai negatif.

Confusion matrix memberikan sejumlah statistik untuk mengevaluasi kinerja model klasifikasi (Kulkarni et al., 2020). Hasil tersebut diantaranya sebagai berikut:

2.12.1 Accuracy

Accuracy yaitu persentase sejauh mana akuratnya suatu model dalam melakukan klasifikasi atau prediksi secara benar oleh algoritma. Fungsi dari *accuracy* terdapat pada Persamaan 17.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (17)$$

2.12.2 Precision

Precision yaitu mengukur seberapa banyak prediksi positif yang benar dari semua prediksi positif yang dibuat oleh model. Fungsi dari *Precision* terdapat pada Persamaan 18.

$$Precision = \frac{TP}{(TP + FP)} \quad (18)$$

2.12.3 Recall

Recall yaitu mengukur kemampuan model untuk menemukan seberapa banyak kasus positif sebenarnya yang berhasil diidentifikasi oleh model. Fungsi dari *Recall* terdapat pada Persamaan 19.

$$Recall = \frac{TP}{(TP + FN)} \quad (19)$$

2.12.4 *F1 Score*

F1 Score yaitu perbandingan nilai rata-rata *Recall* dan *Precision* yang diberi bobot. *F1 Score* berguna ketika ingin membuat model klasifikasi dengan keseimbangan *Recall* dan *Precision* secara optimal. Fungsi dari *F1 Score* terdapat pada Persamaan 20.

$$F1\ Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (20)$$

III. METODE PENELITIAN

3.1 Tempat dan Waktu Penelitian

Tempat dan waktu penelitian adalah sebagai berikut:

3.1.1 Tempat Penelitian

Kegiatan penelitian dilakukan di Laboratorium Komputasi Dasar Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung yang beralamatkan di Jalan Prof. Dr. Ir. Sumantri Brojonegoro No. 1, Gedong Meneng, Kecamatan Rajabasa, Kota Bandar Lampung, Lampung 35141.

3.1.2 Waktu Penelitian

Penelitian dilakukan pada awal bulan Oktober 2023 hingga pada bulan Mei 2024. Alur waktu pengerjaan dapat dilihat pada Tabel 5.

Tabel 5. Alur waktu pengerjaan penelitian.

Kegiatan	2023																2024																			
	Oktober				November				Desember				Januari				Februari				Maret				April				Mei							
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4				
Studi literatur	■	■																																		
Pengumpulan <i>dataset</i>		■	■	■	■	■	■																													
Penyusunan BAB I-III	■	■	■	■	■	■	■	■	■	■	■	■																								
<i>Data Preprocessing</i>									■	■	■	■	■	■	■	■																				
Pembagian <i>dataset</i>													■	■	■	■	■	■	■	■																
Pemodelan algoritma																	■	■	■	■	■	■	■	■												
Evaluasi <i>confusion matrix</i>																					■	■	■	■	■	■	■	■								
Penyusunan BAB IV-V																									■	■	■	■	■	■	■	■				

3.2 Data dan Alat

Berikut adalah penjelasan mengenai data dan alat yang digunakan dalam penelitian ini:

3.2.1 Data

Pada penelitian ini, sumber data berasal dari *Kaggle* dan terdiri dari 13 atribut, dengan total 971 entri data. Data ini dikumpulkan dari 62 lokasi Sungai di Amerika Serikat, yang mencakup periode tahun 1995 hingga 2014. Sumber data dapat diakses melalui <https://www.kaggle.com/datasets/hailla/wqi-parameter-scores-1994-2013>. Penjelasan mengenai atribut dapat dilihat pada Tabel 6.

Tabel 6. Penjelasan atribut pada *dataset*.

No	Nama Atribut	Keterangan
1	<i>Station</i>	<i>Station</i> merupakan kode dari stasiun lokasi sampel air yang diambil.
2	<i>Station Name</i>	<i>Station name</i> merupakan nama atau label dari stasiun lokasi sampel air yang diambil.
3	<i>Year</i>	Waktu pengumpulan data sampel.
4	<i>Overall WQI</i>	<i>Overall WQI</i> adalah atribut hasil secara keseluruhan dari perhitungan setiap sampel.
5	<i>WQI FC</i>	<i>Fecal</i> merupakan organisme yang berasal dari saluran pencernaan dan kotoran manusia dan hewan.
6	<i>WQI Oxy</i>	Jumlah <i>oxygen</i> yang terlarut dalam air yang berasal dari fotosintesis dan difusi oksigen dari udara.

No	Nama Atribut	Keterangan
7	<i>WQI pH</i>	<i>pH</i> merupakan derajat keasaman suatu larutan, <i>pH</i> berkisar dari 0-14 di mana nilai 7 dianggap sebagai netral. Nilai <i>pH</i> dibawah 7 menunjukkan larutan bersifat asam, sedangkan nilai <i>pH</i> diatas 7 menunjukkan larutan bersifat basa.
8	<i>WQI TSS</i>	Total <i>suspended sediment</i> (TSS) merupakan total masa partikel padat yang terapung dalam tanah, tanpa memperhitungkan partikel yang larut atau tenggelam. TSS mencakup debu, tanah, lumpur, serpihan organik dan partikel yang tersuspensi dalam air.
9	<i>WQI Temp</i>	<i>Temperature</i> atau suhu adalah ukuran intensitas panas atau dingin dari air, jadi suhu mempengaruhi kualitas air dan keberlanjutan ekosistem akuatik.
10	<i>WQI TPN</i>	<i>Nitrogen</i> merupakan senyawa yang berasal dari limbah pertanian, limbah industri dan limbah domestik.
11	<i>WQI TP</i>	<i>Phosphorus</i> merupakan senyawa yang berasal dari pupuk pertanian, limbah domestik atau aliran air dari permukaan.
12	<i>WQI Turb</i>	<i>Turbidity</i> adalah ukuran sejauh mana partikel padat tersebar dan terlarut di dalam air.

No	Nama Atribut	Keterangan
13	<i>Location 1</i>	<i>Location 1</i> berisi tentang: a. <i>Latitude</i> (Lintang) merupakan garis horizontal yang mengukur jarak suatu titik dari garis khatulistiwa (<i>equator</i>) yang terletak di tengah-tengah bumi. b. <i>Longitude</i> (Bujur) merupakan garis vertikal yang menghubungkan antara sisi utara dan sisi selatan bumi (kutub).
14	<i>Address</i>	<i>Address</i> merupakan atribut tambahan yang menjelaskan alamat dari pengambilan sampel air.
15	<i>Plus Code</i>	<i>Plus code</i> merupakan atribut tambahan yang berfungsi sebagai penanda lokasi dari pengambilan sampel air.
16	<i>Quality</i>	<i>Quality</i> adalah atribut tambahan yang berfungsi dalam menentukan 5 kategori standar air dari <i>dataset</i> .

3.2.2 Perangkat Penelitian

Alat yang digunakan dalam penelitian ini terdiri dari *hardware* dan *software*.

3.2.2.1 Hardware

Penelitian ini menggunakan perangkat keras laptop dengan spesifikasi sebagai berikut.

- a. *Processor* : 11th Gen Intel ® Core™ i7-11800H
- b. *RAM* : 16 GB
- c. *Storage* : SSD 512 GB

3.2.2.2 Software

Penelitian ini menggunakan beberapa perangkat lunak, yaitu sebagai berikut.

- a. Sistem Operasi : *Windows 11 64-bit*
- b. Bahasa Pemrograman : *Python*
- c. *Text Editor* : *Google Colaboratory*
- d. *Web Browser* : *Google Chrome*
- e. Penyimpanan Data : *Google Drive*

3.2.2.3 Library

Penelitian ini menggunakan *library* atau *package*, yaitu sebagai berikut:

a. *Pandas*

Pandas adalah *tools* yang digunakan untuk bekerja dengan data terstruktur, seperti data statistika, keuangan, ilmu sosial dan bidang lainnya. *Library* ini menawarkan berbagai fungsi terintegrasi untuk memanipulasi data dan analisis pada data tersebut (Mckinney, 2011).

b. *Numpy*

Numpy (numerical python) adalah perpustakaan yang dapat digunakan untuk melakukan komputasi numerik secara efisien. *NumPy* menyediakan berbagai fungsi matematika yang sering digunakan oleh akademisi maupun industri (Van der Walt et al., 2011).

c. *Matplotlib*

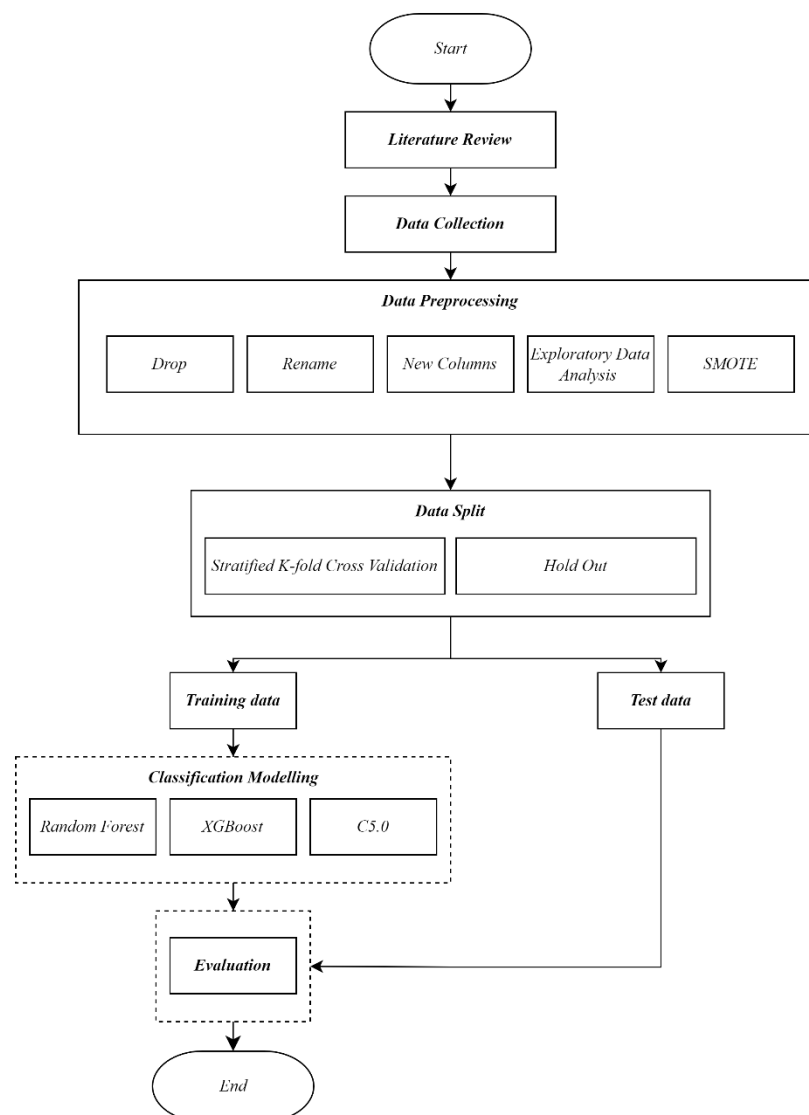
Matplotlib adalah *package* yang digunakan untuk memvisualisasikan data seperti grafik. *Matplotlib* dirancang agar dapat menghasilkan visualisasi dengan sedikit perintah (Ari & Ustazhanov, 2014)

d. *Scikit-Learn*

Scikit-Learn adalah modul *python* yang mengintegrasikan berbagai algoritma *machine learning*. Paket ini berfokus untuk mengenalkan *machine learning* untuk non spesialis (Pedregosa et al., 2011).

3.3 Metode

Alur kerja penelitian ini melalui beberapa tahapan dapat dilihat pada Gambar 4.



Gambar 4. Alur kerja penelitian.

Berdasarkan alur penelitian pada Gambar 4, berikut penjelasan setiap tahap.

3.3.1 *Literature Review*

Tahap pertama melibatkan analisis literatur sebagai upaya pengumpulan informasi yang relevan untuk menyusun penelitian. Proses ini mencakup pengumpulan informasi dari berbagai sumber seperti jurnal, buku dan sumber-sumber terpercaya lainnya yang mendukung penelitian ini.

3.3.2 *Data Collection*

Data *water quality index* diperoleh dari *Kaggle* dan terdiri dari 971 data dengan 13 atribut yang diklasifikasikan ke dalam lima kategori yang berbeda, yaitu sangat baik (*Excellent*), baik (*Good*), cukup (*Fair*), kurang (*Marginal*) dan buruk (*Poor*). Seluruh data ini disimpan dalam format *.csv*.

3.3.3 *Data Preprocessing*

Pada tahap ini, dilakukan *cleaning* untuk memastikan bahwa data siap untuk digunakan. Tahap data *cleaning* mencakup *drop* atribut yang tidak diperlukan selama proses lanjutan, seperti atribut *Station*, *Station Name*, *Year*, *Location 1*, *Address* dan *Plus Code*. Selanjutnya *rename* atribut agar tidak terjadi ambiguitas, *new columns* bertujuan membuat atribut baru yaitu atribut *Quality*, di mana pada atribut *Quality* dilakukan proses *labeling* menggunakan metode CCME dengan lima kategori standar kualitas air yaitu sangat baik (*Excellent*), baik (*Good*), cukup (*Fair*), kurang (*Marginal*) dan buruk (*Poor*), *exploratory data analysis* (EDA) bertujuan untuk memahami karakteristik dan informasi yang ada di dalam dataset, *synthetic minority over-sampling technique* (SMOTE) bertujuan untuk mengatasi ketidakseimbangan data dan Jumlah atribut yang digunakan pada penelitian sebanyak 8 atribut yaitu *WQI FC*, *WQI Oxy*, *WQI pH*, *WQI TSS*, *WQI Temp*, *WQI TPN*, *WQI TP* dan *WQI Turb*.

3.3.4 *Data Split*

Pada tahap ini, data dibagi menggunakan dua metode yaitu *hold out* dan *stratified k-fold cross-validation*. *Hold out* membagi data menjadi dua bagian yaitu data latih (*training*) dan data uji (*testing*), dengan *data training* 90% dan *data testing* 10%. *Stratified k-fold cross validation* data diacak dan dibagi menjadi k lipatan, di mana $k=10$. Setiap lipatan digunakan secara bergantian sebagai *testing* dalam 9 iterasi, sementara iterasi yang tersisa digunakan sebagai *training*. Proses ini diulang sebanyak 10 kali untuk setiap iterasi k dalam pengujian.

3.3.5 *Classification Modelling*

Tahapan selanjutnya setelah pembagian data yaitu melakukan pemodelan atau klasifikasi. Pemodelan klasifikasi pada penelitian ini menggunakan metode *Random Forest*, *XGBoost* dan *C5.0*.

3.3.6 *Evaluation*

Tahapan terakhir yang dilakukan adalah melakukan evaluasi atau penilaian klasifikasi menggunakan *confusion matrix*, dengan tujuan untuk membandingkan dan menentukan model mana yang lebih unggul diantara metode *Random Forest*, *XGBoost* dan *C5.0*.

V. SIMPULAN DAN SARAN

5.1 Simpulan

Penelitian evaluasi metode *Random Forest*, *XGBoost* dan C5.0 dalam klasifikasi kualitas air bersih untuk mendukung pengelolaan sumber daya air, dapat disimpulkan sebagai berikut:

1. Penelitian ini berhasil mengimplementasikan metode *Random Forest*, *XGBoost* dan C5.0 untuk mengklasifikasikan kualitas air bersih. Data yang digunakan adalah data kualitas air dari Kaggle, terdiri dari 971 data dengan 13 atribut yang dikumpulkan dari 62 titik lokasi sungai di Amerika Serikat mencakup periode tahun 1995 hingga 2014. Dalam analisisnya, data dibagi menggunakan dua metode pembagian data, yaitu *hold out* dengan proporsi 90% *data training* dan 10% *data testing*, serta pembagian data *stratified k-fold cross-validation* dengan proporsi $k = 10$.
2. Hasil kinerja metode *Random Forest*, *XGBoost* dan C5.0 dalam mengklasifikasikan kualitas air bersih menggunakan metode pembagian data *hold out* yaitu metode *Random Forest* memperoleh nilai *accuracy* sebesar 0.979, *precision* sebesar 0.950, *recall* sebesar 0.946 dan *f1-score* sebesar 0.947. Metode *XGBoost* memperoleh nilai *accuracy* sebesar 0.964, *precision* sebesar 0.913, *recall* sebesar 0.907 dan *f1-score* sebesar 0.907. Metode C5.0 memperoleh nilai *accuracy* sebesar 0.960, *precision* sebesar 0.906, *recall* sebesar 0.898 dan *f1-score* sebesar 0.899.

Selanjutnya, menggunakan metode pembagian data *stratified k-fold cross-validation* yaitu metode *Random Forest* memperoleh nilai *accuracy* sebesar 0.977, *precision* sebesar 0.943, *recall* sebesar 0.943 dan *f1-score* sebesar 0.943. Metode *XGBoost* memperoleh nilai *accuracy*

sebesar 0.966, *precision* sebesar 0.916, *recall* sebesar 0.916 dan *f1-score* sebesar 0.915. Metode C5.0 memperoleh nilai *accuracy* sebesar 0.966, *precision* sebesar 0.916, *recall* sebesar 0.916 dan *f1-score* sebesar 0.916.

Berdasarkan hasil kinerja ketiga metode menggunakan pembagian data *hold out*, metode *Random Forest* menunjukkan akurasi tertinggi sebesar 0.979, dibandingkan dengan metode *XGBoost* yang memiliki akurasi 0.964 dan C5.0 dengan akurasi 0.960. Namun, dalam hal waktu eksekusi, metode C5.0 lebih cepat dengan waktu 62.28 ms, dibandingkan dengan *Random Forest* yang membutuhkan 429.806 ms dan *XGBoost* yang membutuhkan 315.998 ms.

Di sisi lain, dengan pembagian data menggunakan SKCV, metode *Random Forest* tetap menunjukkan akurasi tertinggi sebesar 0.977, sementara metode *XGBoost* dan C5.0 sama-sama memiliki akurasi 0.966. Dalam hal waktu eksekusi, metode C5.0 tetap yang tercepat dengan waktu 1771.632 ms, dibandingkan dengan *Random Forest* yang membutuhkan 8584.102 ms dan *XGBoost* yang membutuhkan 4181.01 ms.

5.2 Saran

Saran yang diberikan pada penelitian ini sebagai berikut:

1. Melakukan eksplorasi terhadap penggunaan metode klasifikasi *Gradient Boosting* lainnya seperti *CatBoost*, atau metode klasifikasi berbasis *deep learning* untuk mencapai hasil yang lebih optimal.
2. Memperbanyak jumlah data yang digunakan dalam penelitian guna meningkatkan keakuratan model.

DAFTAR PUSTAKA

- Abdillah, A. A., Murfi, H., & Yudi, S. (2015). *Uji Kinerja Learning To Rank Dengan Metode Support Vector Regression*.
- Ari, N., & Ustazhanov, M. (2014). *Matplotlib In Python*. 11th International Conference on Electronics, Computer and Computation (ICECCO). <https://doi.org/10.1109/ICECCO.2014.6997585>
- Erlin, E., Desnelita, Y., Nasution, N., Suryati, L., & Zoromi, F. (2022). Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang. *Matrik: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, 21(3), 677–690. <https://doi.org/10.30812/matrik.v21i3.1726>
- Ghazvini, A., Awwalu, J., & Abu Bakar, A. (2014). Comparative Analysis of Algorithms in Supervised Classification: A Case study of Bank Notes Dataset. *International Journal of Computer Trends and Technology*, 17(1), 39–43. <https://doi.org/10.14445/22312803/IJCTT-V17P109>
- Hawari, Y., Tata Hardinata, J., & Nasution, R. A. (2022). *Buletin Big Data, Data Science and Artificial Intelligence Implementasi K-Means Clustering Dalam Menentukan Kualitas Biji Kelapa Sawit (Kasus PPKS Marihat)* (Vol. 1, Issue 1). <https://ejurnal.pdsi.or.id/index.php/zahra/index>
- Imandasari, T., Irawan, E., Perdana Windarto, A., Wanto, A., & Tunas Bangsa Pematangsiantar Jln Jendral Sudirman Blok No, S. A. (2019). *Prosiding Seminar Nasional Riset Information Science (SENARIS) Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air*.
- Joloudari, J. H., Haderbadi, M., Mashmool, A., Ghasemigol, M., Band, S. S., & Mosavi, A. (2020). Early detection of the advanced persistent threat attack using performance analysis of deep learning. *IEEE Access*, 8, 186125–186137. <https://doi.org/10.1109/ACCESS.2020.3029202>
- Kulkarni, A., Batarseh, F. A., & Chong, D. (2020). *Chapter 5: Foundations of Data Imbalance and Solutions for a Data Democracy*. Cambridge: Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-818366-3.00005-8>
- Kustanto, A. (2020). Dinamika Pertumbuhan Penduduk Dan Kualitas Air di Indonesia. *JIEP*, 20(1).

- Lumb, A., Sharma, T. C., & Bibeault, J.-F. (2011). A Review of Genesis and Evolution of Water Quality Index (WQI) and Some Future Directions. *Water Quality, Exposure and Health*, 3(1), 11–24. <https://doi.org/10.1007/s12403-011-0040-0>
- Mambang, & Byna, A. (2017). *Analisis Perbandingan Algoritma C.45, Random Forest Dengan Chaid Decision Tree Untuk Klasifikasi Tingkat Kecemasan Ibu Hamil*.
- Mckinney, W. (2011). *pandas: a Foundational Python Library for Data Analysis and Statistics*. <http://pandas.sf.net>
- Muller, A. C. (2020). *Applied Machine Learning in Python*. 2020. <https://amueller.github.io/aml/04-model-evaluation/1-data-splitting-strategies.html>
- Muslim, Abd. Q., Suci, I. G. S., & Pratama, M. R. (2021). Analisis Kebijakan Pendidikan Di Jepang, Finlandia, China Dan Indonesia Dalam Mendukung Sustainable Development Goals. *ADI WIDYA: Jurnal Pendidikan Dasar*, 6(2). <http://ejournal.ihtdn.ac.id/index.php/AW>
- Mutoffar, M. M., Naseer, M., Fadillah, A., Studi, P., Informatika, T., Tinggi, S., & Bandung, T. (2022). *Klasifikasi Kualitas Air Sumur Menggunakan Algoritma Random Forest*. 04.
- Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma'a, A. (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, 48. <https://doi.org/10.1016/j.jwpe.2022.102920>
- Pardede, M., Buulolo, E., & Ndruru, E. (2019). Implementasi Algoritma C5.0 Pada Kelulusan Peserta Ujian Kemahiran Berbahasa Indonesia (Ukbi) Pada Balai Bahasa Sumatera Utara. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, 3(1). <https://doi.org/10.30865/komik.v3i1.1569>
- Patora, K., & Morley, K. (2015). *Draft Preliminary Cost-Benefit Analysis and Least-Burdensome Alternative Analysis: Chapter 173-201A WAC Water Quality Standards for Surface Waters of the State of Washington*. Department Of Ecology State Of Washington. <https://ecology.wa.gov/getattachment/64e60bdb-4b8e-43ed-ae4-a88525d8aee4/DraftPrelimWQS-CBAformatted09282014.pdf>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* (Vol. 12). The Journal of Machine Learning Research

- Pratama, N., & Andraini, L. (2022). Model Prediksi Kesesuaian Lahan Kedelai Menggunakan C5.0 Algoritma. In *Portaldata.org* (Vol. 2, Issue 10).
- Prima Wijaya, K., & Muslim, A. (2016). Peningkatan Akurasi pada Algoritma Support Vector Machine dengan Penerapan Information Gain untuk Mendiagnosa Chronic Kidney Disease. In *Seminar Nasional Ilmu Komputer*.
- Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIMD)*, 1(1), 27–31.
- Purba, D. J., Lubis, M. R., & Siregar, Z. A. (2022). Analisis Kepuasan Pelanggan Menggunakan Algoritma C5.0. In *Journal of Informatics Management and Information Technology* (Vol. 2, Issue 2). <https://hostjournals.com/>
- Ritonga, P. S. (2011). “Air” Sebagai Sarana Peningkatan Imtaq (Integrasi Kimia Dan Agama). In *Air" Sebagai Sarana Peningkatan.....* (Vol. 8, Issue 02). Pangoloan Soleman Ritonga.
- Romdania, Y., Herison, A., Susilo, G. E., & Novilyansa, E. (2018). *Kajian Penggunaan Metode Ip, Storet, Dan Ccme Wqi Dalam Menentukan Status Kualitas Air*.
- Saifudin, A. (2018). *Metode Data Mining Untuk Seleksi Calon Mahasiswa Pada Penerimaan Mahasiswa Baru Di Universitas Pamulang*. <https://doi.org/10.24853/jurtek.10.1.25-36>
- Saraswati, S. P., Ardion, M. V., Widodo, Y. H., & Hadisusanto, S. (2019). Water Quality Index Performance for River Pollution Control Based on Better Ecological Point of View (A Case Study in Code, Winongo, Gajah Wong Streams). *Journal of the Civil Engineering Forum*, 5(1), 47. <https://doi.org/10.22146/jcef.41165>
- Sidik, A. D., & Ansawarman, A. (2022). Prediksi Jumlah Kendaraan Bermotor Menggunakan Machine Learning. *Formosa Journal of Multidisciplinary Research (FJMR)*, 1(3), 559–568. <https://doi.org/10.55927>
- Simon, A., Deo, M. S., Venkatesan, S., & Babu, D. R. R. (2016). *An Overview of Machine Learning and its Applications*. <https://www.researchgate.net/publication/289980169>
- Somasundaram, A., & Reddy, U. S. (2016). *Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data*. <https://www.researchgate.net/publication/320895020>
- Suwardika, I. G. I., Suariana, I. G. N., Bhiantara, I. P., & Arso N.Y. (2019). Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes: Studi Kasus Fakultas Ekonomi Dan Bisnis Universitas Pendidikan Nasional. *Jurnal Ilmu Komputer Indonesia (JIKI)*.

- Syukron, M., Santoso, R., & Widiharih, T. (2020). *Perbandingan Metode Smote Random Forest Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data*. <https://ejournal3.undip.ac.id/index.php/gaussian/>
- Tangkelayuk, A., & Mailoa, E. (2022). *Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes Dan Decision Tree*. 9(2), 1109–1119. <http://jurnal.mdp.ac.id>
- UNICEF. (2019, November 30). *Progress on household drinking water, sanitation and hygiene, 2000-2017*. <https://www.unicef.org/reports/progress-on-drinking-water-sanitation-and-hygiene-2019>
- Van der Walt, S., Colbert, S. C., Varoquaux, G., & Inria. (2011). *The NumPy Array: A Structure for Efficient Numerical Computation*. <http://docs.python.org>.
- Wahyuni, E. D., Arifiyanti, A. A., & Kustyani, M. (2019). *Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining*. 263–269. <http://journal.itny.ac.id/index.php/ReTII>
- XGBoost Developers. (2023). *XGBoost Release 1.5.0-dev*. <https://buildmedia.readthedocs.org/media/pdf/xgboost/latest/xgboost.pdf>
- Yang, J., Gong, J., Tang, W., Shen, Y., Liu, C., & Gao, J. (2019). Delineation of urban growth boundaries using a patch-based cellular automata model under multiple spatial and socio-economic scenarios. *Sustainability (Switzerland)*, 11(21). <https://doi.org/10.3390/su11216159>
- Yulianti, E. H., Soesanto, O., & Sukmawaty, Y. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *JOMTA Journal of Mathematics: Theory and Applications*, 4(1).
- Zega, S. A. (2014). Penggunaan Pohon Keputusan untuk Klasifikasi Tingkat Kualitas Mahasiswa Berdasarkan Jalur Masuk Kuliah. In *Seminar Nasional Aplikasi Teknologi Informasi (SNATI) Yogyakarta*.
- Zulhilmi, Efendy, I., Syamsul, D., & Idawati. (2019). Faktor Yang Berhubungan Tingkat Konsumsi Air Bersih Pada Rumah Tangga di Kecamatan Peudada Kabupaten Bireun. *Jurnal Biology Education*, 7(2). <https://doi.org/https://doi.org/10.32672/jbe.v7i2.1592>