

**SKRIPSI**  
**SISTEM KLASIFIKASI KELAYAKAN AIR MINUM DENGAN SELEKSI**  
**FITUR MENGGUNAKAN METODE *RECURSIVE FEATURE***  
***ELIMINATION WITH CROSS-VALIDATION (RFECV)* DAN**  
**SELECTKBEST**

**Disusun dan diajukan oleh :**

**NURHALISA**

**D121201053**



**PROGRAM STUDI SARJANA TEKNIK INFORMATIKA**  
**FAKULTAS TEKNIK**  
**UNIVERSITAS HASANUDDIN**  
**GOWA**  
**2024**

**LEMBAR PENGESAHAN SKRIPSI**

**SISTEM KLASIFIKASI KELAYAKAN AIR MINUM DENGAN SELEKSI  
FITUR MENGGUNAKAN METODE *RECURSIVE FEATURE  
ELIMINATION WITH CROSS-VALIDATION (RFECV)* DAN  
SELECTKBEST**

Disusun dan diajukan oleh

**Nurhalisa**

**D121201053**

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka  
Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika  
Fakultas Teknik Universitas Hasanuddin  
Pada tanggal 28 November 2024  
dan dinyatakan telah memenuhi syarat kelulusan

**Menyetujui,**

Dosen Pembimbing,

Ketua Program Studi,

Prof. Dr. Ir. Indrabayu, ST., MT.,  
M.Bus.Sys., IPM, ASEAN. Eng.

Prof. Dr. Ir. Indrabayu, ST., MT.,  
M.Bus.Sys., IPM, ASEAN. Eng.

NIP. 197507162002121004

NIP. 197507162002121004

## PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini :

Nama : Nurhalisa  
NIM : D121201053  
Program Studi : Teknik Informatika  
Jenjang : S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

Sistem Klasifikasi Kelayakan Air Minum dengan Seleksi Fitur menggunakan  
Metode *Recursive Feature Elimination with Cross-Validation* (RFECV) dan  
SelectKbest

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain dan bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Semua informasi yang ditulis dalam skripsi yang berasal dari penulis lain telah diberi penghargaan, yakni dengan mengutip sumber dan tahun penerbitannya. Oleh karena itu semua tulisan dalam skripsi ini sepenuhnya menjadi tanggung jawab penulis. Apabila ada pihak maupun yang merasa ada kesamaan judul dan atau hasil temuan dalam skripsi ini, maka penulis siap untuk diklarifikasi dan mempertanggungjawabkan segala resiko.

Segala data dan informasi yang diperoleh selama proses pembuatan skripsi, yang akan dipublikasikan oleh Penulis di masa depan harus mendapat persetujuan dari Dosen Pembimbing.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa Sebagian atau keseluruhan isi skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Gowa, 02 Desember 2024

Yang Menyatakan

  
Nurhalisa

## ABSTRAK

**NURHALISA.** (*Sistem Klasifikasi Kelayakan Air Minum dengan Seleksi Fitur menggunakan Metode Recursive Feature Elimination with Cross-Validation (RFECV) dan SelectKbest* dibimbing oleh Indrabayu Amirullah)

Proses pengolahan air minum membutuhkan pengukuran berkala setiap jam untuk memastikan air yang diproduksi memenuhi standar air minum. Penelitian ini bertujuan untuk mengklasifikasikan potensi air di Gowa dan melakukan seleksi fitur untuk mengidentifikasi parameter yang paling optimal. Penelitian ini menggunakan SVM dan XGBoost untuk klasifikasi dan menggunakan RFECV dan SelectKBest untuk seleksi fitur. Hasil penelitian menunjukkan bahwa sebagian besar korelasi antara parameter dan target lemah, yang mengindikasikan bahwa setiap parameter beroperasi secara independen dan memiliki nilai yang unik dalam menentukan potensi air minum. Penelitian ini mencapai akurasi model yang tinggi, dengan 95,8% untuk SVM dan 97,8% untuk XGBoost. Setelah pemilihan fitur, akurasi akhir untuk model SVM adalah 95,8% menggunakan SelectKBest dengan 3 fitur yang dipilih. Dengan menggunakan RFECV, akurasinya adalah 96% dengan 5 fitur yang dipilih. Untuk XGBoost, akurasi akhir setelah pemilihan fitur adalah 97,8% menggunakan SelectKBest dengan 5 fitur yang dipilih. Pemilihan fitur RFECV untuk XGBoost juga mempertahankan akurasi yang sama yaitu 97,8%. Berdasarkan hasil, XGBoost berkinerja sedikit lebih baik daripada SVM, tetapi RFECV meningkatkan akurasi SVM sambil mempertahankan akurasi XGBoost sedangkan metode SelectKBest juga mempertahankan akurasi untuk kedua model. Oleh karena itu, RFECV lebih unggul dibandingkan SelectKBest dengan 5 fitur terpilih yaitu kekeruhan, klorin bebas, suhu, pH, dan alkalinitas.

Kata Kunci :SVM, XGBoost, RFECV, SelectKbest, Kelayakan Air

## ABSTRACT

**NURHALISA.** (*Drinking Water Potability Classification System with Feature Selection using Recursive Feature Elimination with Cross-Validation (RFECV) and SelectKbest Method supervised by Indrabayu Amirullah*)

*Drinking water treatment process requires periodic measurements every hour to ensure the produced water meets drinkable water standards. This research aims to classify water potability in Gowa and perform feature selection to identify the most optimal parameters. The research uses the SVM and XGBoost for classification and employs RFECV and SelectKBest for feature selection. The results show that most correlations between parameters and the target are weak, indicating that each parameter operates independently and has unique value in determining drinking water potability. The study achieves high model accuracy, with 95.8% for SVM and 97.8% for XGBoost. After feature selection, the final accuracy for the SVM model is 95.8% using the SelectKBest with 3 selected features. Using the RFECV, the accuracy is 96% with 5 selected features. For XGBoost, the final accuracy after feature selection is 97.8% using the SelectKBest with 5 selected features. The RFECV feature selection for XGBoost also maintains the same accuracy of 97.8%. Based on the results, XGBoost performs slightly better than SVM, but RFECV improves SVM accuracy while maintaining XGBoost accuracy. The SelectKBest method also maintains the accuracy for both models. Therefore, RFECV is superior to SelectKBest with the 5 selected features of turbidity, free chlorine, temperature, pH and alkalinity.*

*Keywords: SVM, XGBoost, RFECV, SelectKbest, Water Potability*

## DAFTAR ISI

|                                                                              |            |
|------------------------------------------------------------------------------|------------|
| <b>ABSTRAK .....</b>                                                         | <b>iii</b> |
| <b>ABSTRACT .....</b>                                                        | <b>iv</b>  |
| <b>DAFTAR ISI.....</b>                                                       | <b>v</b>   |
| <b>DAFTAR GAMBAR.....</b>                                                    | <b>vii</b> |
| <b>DAFTAR TABEL .....</b>                                                    | <b>ix</b>  |
| <b>DAFTAR SINGKATAN DAN ARTI SIMBOL .....</b>                                | <b>x</b>   |
| <b>DAFTAR LAMPIRAN .....</b>                                                 | <b>xi</b>  |
| <b>KATA PENGANTAR.....</b>                                                   | <b>xii</b> |
| <b>BAB I PENDAHULUAN.....</b>                                                | <b>1</b>   |
| 1.1 Latar Belakang.....                                                      | 1          |
| 1.2 Rumusan Masalah .....                                                    | 3          |
| 1.3 Tujuan Penelitian.....                                                   | 3          |
| 1.4 Manfaat Penelitian.....                                                  | 4          |
| 1.5 Ruang Lingkup/Asumsi Perancangan .....                                   | 4          |
| 1.6 Sistematika Penulisan.....                                               | 4          |
| <b>BAB II TINJAUAN PUSTAKA.....</b>                                          | <b>6</b>   |
| 2.1 Air Minum .....                                                          | 6          |
| 2.2 <i>Machine Learning</i> .....                                            | 9          |
| 2.2.1 <i>Supervised Learning</i> .....                                       | 9          |
| 2.2.2 <i>Unsupervised Learning</i> .....                                     | 10         |
| 2.2.3 <i>Semi Supervised Learning</i> .....                                  | 11         |
| 2.3 Klasifikasi.....                                                         | 11         |
| 2.4 Seleksi Fitur.....                                                       | 12         |
| 2.5 <i>Recursive Feature Elimination with Cross-Validation (RFECV)</i> ..... | 13         |
| 2.6 <i>SelectKbest with f_classif</i> .....                                  | 14         |
| 2.7 <i>Support Vector Machine (SVM)</i> .....                                | 14         |
| 2.8 <i>eXtreme Gradient Boosting (XGBoost)</i> .....                         | 16         |
| 3.1 Waktu dan Lokasi Penelitian.....                                         | 18         |
| 3.2 Benda Uji dan Alat .....                                                 | 18         |
| 3.4 Tahapan Penelitian .....                                                 | 19         |
| 3.5 Rancangan Sistem .....                                                   | 21         |

|                                                    |    |
|----------------------------------------------------|----|
| 3.5.1 Pengumpulan Data.....                        | 22 |
| 3.5.1 <i>Exploratory Data Analysis</i> .....       | 23 |
| 3.5.3 <i>Preprocessing Data</i> .....              | 27 |
| 3.5.4 Pelabelan Data .....                         | 29 |
| 3.5.5 <i>Tuning Hyperparameter</i> .....           | 31 |
| 3.5.6 Model Klasifikasi SVM.....                   | 31 |
| 3.5.7 Klasifikasi Model XGBoost .....              | 34 |
| 3.5.8 Seleksi Fitur RFECV .....                    | 35 |
| 3.5.9 Seleksi Fitur SelectKbest .....              | 37 |
| 3.5.10 Evaluasi Model .....                        | 38 |
| <b>BAB IV ANALISIS DAN PEMBAHASAN</b> .....        | 40 |
| 4.1 Analisis Korelasi Fitur.....                   | 40 |
| 4.1.1 Korelasi antar Fitur .....                   | 41 |
| 4.1.2 Korelasi Fitur terhadap Target.....          | 41 |
| 4.2 Analisis Hyperparameter Tuning .....           | 41 |
| 4.3 Analisis Model Klasifikasi .....               | 42 |
| 4.3.1 SVM.....                                     | 42 |
| 4.3.2 XGBoost .....                                | 46 |
| 4.4 Analisis Seleksi Fitur RFECV .....             | 49 |
| 4.4.1 SVM.....                                     | 49 |
| 4.4.2 XGBoost .....                                | 53 |
| 4.4 Analisis Seleksi Fitur dengan SelectKbest..... | 55 |
| 4.4.1 SelectKbest dengan SVM .....                 | 55 |
| 4.4.2 SelectKbest dengan XGBoost.....              | 58 |
| 4.5 Perbandingan Kinerja Model.....                | 61 |
| <b>BAB V KESIMPULAN DAN SARAN</b> .....            | 65 |
| 5.1 Kesimpulan.....                                | 65 |
| 5.2 Saran.....                                     | 66 |
| <b>DAFTAR PUSTAKA</b> .....                        | 68 |

## DAFTAR GAMBAR

|                                                                                                                                      |    |
|--------------------------------------------------------------------------------------------------------------------------------------|----|
| Gambar 1 Workflow supervised learning (Mahesh, 2020).....                                                                            | 10 |
| Gambar 2 Unsupervised learning (Mahesh, 2020). ....                                                                                  | 11 |
| Gambar 3 Teknik feature selection RFECV (Shi dkk., 2024) .....                                                                       | 14 |
| Gambar 4 Ilustrasi kerja SVM (Sudianto, 2022) .....                                                                                  | 15 |
| Gambar 5 Architecture model XGBoost (Wang dkk., 2019).....                                                                           | 17 |
| Gambar 6 Lokasi penelitian .....                                                                                                     | 18 |
| Gambar 7 Alur penelitian.....                                                                                                        | 19 |
| Gambar 8 Alur Rancangan Sistem.....                                                                                                  | 21 |
| Gambar 9 Hasil identifikasi nilai yang hilang.....                                                                                   | 24 |
| Gambar 10 Jumlah data outliers.....                                                                                                  | 25 |
| Gambar 11 Boxplot data outliers; (a) kekeruhan; (b) pH; (c) Suhu; (d) Suhu udara;<br>(e) Klor bebas; (f) Alkalinitas; (g) TDS .....  | 25 |
| Gambar 12 Diagram distribusi data (a) kekeruhan; (b) pH; (c) Suhu; (d) Suhu<br>udara; (e) Klor bebas; (f) Alkalinitas; (g) TDS ..... | 27 |
| Gambar 13 Missing values handling.....                                                                                               | 28 |
| Gambar 14 Standar scaling .....                                                                                                      | 29 |
| Gambar 15 Contoh hasil pelabelan data.....                                                                                           | 30 |
| Gambar 16 Distribusi kelas .....                                                                                                     | 30 |
| Gambar 17 Ilustrasi proses tuning hyperparameter (Putri dkk., 2022).....                                                             | 31 |
| Gambar 18 Flowchart Model Klasifikasi SVM .....                                                                                      | 32 |
| Gambar 19 Flowchart Model klasifikasi XGBoost.....                                                                                   | 34 |
| Gambar 20 Flowchart seleksi fitur RFECV .....                                                                                        | 35 |
| Gambar 21 Alur kerja seleksi fitur SelectKbest.....                                                                                  | 37 |
| Gambar 22 Heatmap korelasi fitur setelah pelabelan .....                                                                             | 40 |
| Gambar 23 Best Parameter.....                                                                                                        | 42 |
| Gambar 24 Confusion matrix training data model SVM .....                                                                             | 43 |
| Gambar 25 Hasil Evaluasi training data model SVM .....                                                                               | 43 |
| Gambar 26 Confusion matrix test data model SVM.....                                                                                  | 44 |
| Gambar 27 Hasil Evaluasi test data model SVM.....                                                                                    | 44 |
| Gambar 28 Confusion matrix training data XGBoost .....                                                                               | 46 |
| Gambar 29 Evaluasi model training data XGBoost.....                                                                                  | 47 |
| Gambar 30 Evaluasi model testing data XGBoost.....                                                                                   | 48 |
| Gambar 31 Confusion matrix testing data XGBoost .....                                                                                | 49 |
| Gambar 32 Grafik skor cross-validation RFECV-LinearSVC.....                                                                          | 50 |
| Gambar 33 Hasil seleksi fitur RFECV-LinearSVC .....                                                                                  | 50 |
| Gambar 34 Confusion matrix model akhir dengan RFECV-RBF .....                                                                        | 51 |
| Gambar 35 Hasil Evaluasi model akhir dengan RFECV-RBF .....                                                                          | 52 |
| Gambar 36 Skor cross-validation RFECV-XGBoost .....                                                                                  | 53 |
| Gambar 37 Hasil seleksi fitur RFECV-XGBoost .....                                                                                    | 53 |
| Gambar 38 Confusion matrix model akhir dengan RFECV-XGBoost.....                                                                     | 54 |
| Gambar 39 Evaluasi model akhir dengan RFECV-XGBoost.....                                                                             | 54 |
| Gambar 40 Skor fitur terpenting SelectKbest .....                                                                                    | 55 |



|                                                                                        |    |
|----------------------------------------------------------------------------------------|----|
| Gambar 41 Confusion matrix model akhir SVM setelah seleksi fitur SelectKbest k=3 ..... | 57 |
| Gambar 42 Hasil Evaluasi model akhir SVM dengan SelectKbest k=3.....                   | 58 |
| Gambar 43 Confusion matrix model akhir XGBoost menggunakan SelectKbest k=5 .....       | 60 |
| Gambar 44 Hasil Evaluasi model akhir XGBoost dengan SelectKbest k=5 .....              | 60 |

## DAFTAR TABEL

|                                                                                                       |    |
|-------------------------------------------------------------------------------------------------------|----|
| Tabel 1 Parameter wajib kualitas air menurut Permenkes No.2 tahun 2023 .....                          | 6  |
| Tabel 2 Parameter khusus Kualitas Air Menurut Permenkes No.2 Tahun 2023 ....                          | 7  |
| Tabel 3 Penjelasan kolom dataset .....                                                                | 23 |
| Tabel 4 Standar kualitas air minum .....                                                              | 29 |
| Tabel 5 Perbandingan hasil evaluasi model melalui seleksi fitur SelectKbest-SVM .....                 | 55 |
| Tabel 6 Perbandingan confusion matrix model akhir SVM dengan SelectKbest.                             | 56 |
| Tabel 7 Perbandingan hasil evaluasi model seleksi fitur SelectKbest-XGBoost..                         | 58 |
| Tabel 8 Perbandingan hasil confusion matrix model akhir XGBoost dengan seleksi fitur SelectKbest..... | 59 |
| Tabel 9 Perbandingan Kinerja model dengan data pengukuran air harian .....                            | 61 |
| Tabel 10 Perbandingan hasil confusion matrix model .....                                              | 62 |
| Tabel 11 Perbandingan Kinerja model dengan data air sampel pelanggan .....                            | 63 |
| Tabel 12 Perbandingan hasil confusion matrix model .....                                              | 64 |

## DAFTAR SINGKATAN DAN ARTI SIMBOL

| Lambang/Singkatan | Arti dan Keterangan                                            |
|-------------------|----------------------------------------------------------------|
| SVM               | <i>Support Vector Machine</i>                                  |
| RFECV             | <i>Recursive Feature Elimination with Cross-Validation</i>     |
| ANOVA             | <i>Analysis of Variance</i>                                    |
| RBF               | <i>Radial Basis Function</i>                                   |
| SMO               | <i>Sequential Minimal Optimization</i>                         |
| ML                | <i>Machine Learning</i>                                        |
| XGBoost           | <i>eXtra Gradient Boosting</i>                                 |
| $\sigma$          | Sigma (Parameter skala)                                        |
| exp               | Fungsi Exponensial                                             |
| NTU               | <i>Nephelometric Turbidity Unit</i> (Pengukuran kekeruhan air) |
| TDS               | <i>Total Dissolve Solids</i> (Padatan terlarut total)          |
| EDA               | <i>Exploratory Data Analysis</i>                               |
| IQR               | <i>Interquartile Range</i> (Rentang Interkuartil)              |
| Q1                | Kuartil ke-1                                                   |
| Q3                | Kuartil ke-3                                                   |
| C                 | <i>Cost</i>                                                    |
| a                 | Alpha ( <i>Lagrange multipliers</i> )                          |
| b                 | Bias                                                           |
| TP                | <i>True Positive</i>                                           |
| TN                | <i>True Negative</i>                                           |
| FN                | <i>False Negative</i>                                          |
| FP                | <i>False Positive</i>                                          |

## DAFTAR LAMPIRAN

|                                                                  |    |
|------------------------------------------------------------------|----|
| Lampiran 1 Contoh Dataset: Pengukuran air harian .....           | 71 |
| Lampiran 2 Contoh dataset : Data Sampling Internal .....         | 72 |
| Lampiran 3 Source Code Program : Model SVM.....                  | 73 |
| Lampiran 4 Source Code Program : Model XGBoost .....             | 77 |
| Lampiran 5 Souce Code Program : Seleksi Fitur RFECV .....        | 80 |
| Lampiran 6 Source Code Program : Seleksi fitur SelectKbest ..... | 82 |

## KATA PENGANTAR

Puji dan syukur penulis panjatkan atas kehadiran Allah SWT karena atas berkat dan rahmat-Nya sehingga dapat menyelesaikan tugas akhir dengan judul “Sistem Klasifikasi Kelayakan Air Minum dengan Seleksi Fitur menggunakan Metode *Recursive Feature Elimination with Cross-Validation* (RFECV)” sebagai salah satu persyaratan akademik untuk menyelesaikan program strata-1 pada Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin.

Penulis menyadari banyak kesulitan dan kendala yang dihadapi saat penyusunan tugas akhir ini. Dalam prosesnya, penulis memperoleh banyak bantuan, dukungan, dan bimbingan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan terimakasih kepada :

1. Allah SWT karena telah memberikan penulis berbagai nikmat-Nya, seperti nikmat sehat, rezeki, ilmu, dan masih banyak lagi sehingga skripsi ini dapat terselesaikan.
2. Kedua orang tua penulis, Bapak Baharuddin dan Ibu Mardiani yang senantiasa mendoakan penulis dan memberikan dukungannya kepada penulis.
3. Keluarga penulis, terutama saudara-saudara penulis yaitu kak Jannah dan Ahmad karena senantiasa memberikan semangat, doa dan bantuan materi kepada penulis selama penyelesaian skripsi. Terimakasih juga kepada kak rahmi, kak vira dan putri yang memberikan semangat, doa dan bantuannya kepada penulis.
4. Prof. Dr. Ir. Indrabayu, ST., MT., M.Bus.Sys., IPM, ASEAN. Eng., selaku dosen pembimbing yang senantiasa meluangkan waktu, tenaga, pikiran dan perhatian yang luar biasa untuk membimbing penulis dalam proses penyelesaian tugas akhir mulai dari penyusunan proposal hingga penyusunan skripsi.
5. Segenap Dosen dan Staf Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin yang telah memberikan ilmu dan pengetahuan baru, serta bantuan kepada penulis selama masa perkuliahan.
6. Kak Nublan yang senantiasa memonitoring progres dan memberikan masukan serta semangat kepada penulis.
7. Nabila, Nisa dan Echa yang merupakan sahabat penulis. Terimakasih kepada Nabila yang selalu membantu dan memberi nasehat kepada penulis, Nisa yang selalu menemani penulis saat bimbingan dan senantiasa memberikan saran dalam penelitian serta Echa yang selalu memberikan dukungan dan semangat kepada penulis.
8. Kak Nunu selaku staf lab IPA Pandang-Pandang PDAM Gowa, yang telah membantu penulis untuk mendapatkan dataset dan informasi yang dibutuhkan penulis.
9. Kak Herlina dan teman-teman Lab AIMP yang telah membantu dalam memberikan masukan kepada penulis.
10. Nuni, agunawan dan teman-teman REZOLVER 20 lainnya yang telah memberikan semangat dan banyak membantu selama masa perkuliahan.

## **BAB I PENDAHULUAN**

### **1.1 Latar Belakang**

Beberapa negara semakin dihadapkan pada berbagai masalah yang terkait dengan kelangkaan air, peristiwa kontaminasi sumber air, dan kerusakan lingkungan yang bergantung pada air. Setiap tahunnya, sekitar 829.000 orang meninggal karena penyakit yang disebabkan langsung oleh air yang tidak aman, sanitasi yang tidak memadai dan praktik kebersihan yang buruk (WHO, 2023). Untuk mencapai cakupan universal pada tahun 2030, diperlukan peningkatan akses terhadap air minum yang aman dan terjangkau untuk meningkatkan kesehatan dan kesejahteraan masyarakat demi kemajuan di bidang kesehatan dan pendidikan (Hakim Zulkarnain, 2023).

Air dapat bersumber dari mana saja, seperti air tanah, air system, air danau air hujan dan lainnya. Sumber air yang paling sering digunakan sebagai sumber air minum yaitu air tanah dan air permukaan. Namun, beberapa daerah memiliki kualitas air yang tidak terjamin kualitasnya seperti di daerah perkotaan. Air dapat terkontaminasi akibat aktivitas manusia melalui pembuangan limbah industri, kebocoran sistem penyimpanan bahan bakar, atau rembesan dari tempat pembuangan sampah (Kementerian Kesehatan Republik Indonesia, 2024) seperti halnya sumber air baku yang digunakan di daerah Gowa yang berasal dari sungai Jeneberang. Hal ini dapat menyebabkan air tercemar oleh bakteri dan terkontaminasi secara kimia. Untuk mengatasi ini, diperlukan pengolahan air menjadi air yang layak minum. Berdasarkan hasil wawancara yang peneliti lakukan di IPA Pandang-Pandang PDAM Gowa bahwa air baku yang telah diolah menjadi air bersih perlu di lakukan pengecekan rutin setiap 1 jam sekali dan pengecekan parameter lengkap setiap 2 kali setahun. Hal ini dilakukan untuk memastikan bahwa air yang akan di distribusikan ke pelanggan aman untuk dikonsumsi.

Pengecekan kualitas air secara berkala tentu sulit dilakukan secara manual karena memerlukan pengecekan air setiap 1 jam sekali dengan melakukan pengukuran pada tiap parameter kemudian mengecek masing-masing parameter

apakah memenuhi standar atau tidak. Oleh karena itu, dibutuhkan sebuah teknologi yang mampu melakukan monitoring kualitas air secara otomatis menggunakan *machine learning*.

Dalam penilaian kualitas air, memerlukan pengukuran sejumlah parameter kualitas air. Menurut Permenkes no.492 tahun 2010, parameter wajib sebagai persyaratan kualitas air minum terdiri dari 26 jenis parameter yang mencakup parameter mikrobiologi, fisik, dan kimia (Menteri Kesehatan Republik Indonesia, 2010). Selain itu terdapat parameter khusus yang ditetapkan oleh pemerintah Daerah sesuai dengan kondisi geohidrologi wilayah. IPA Pandang-Pandang PDAM Gowa sendiri melakukan pengukuran parameter lengkap dengan 16 parameter hanya 2 kali setahun karena keterbatasan alat dan bahan. Oleh karena itu, dalam penelitian ini akan dilakukan seleksi fitur untuk mengetahui parameter apa saja yang perlu diutamakan dalam monitoring air secara rutin. Metode yang digunakan dalam proses seleksi fitur pada penelitian ini adalah metode RFECV. Metode ini digunakan karena menggunakan *cross-validation* untuk memperkirakan performa model pada setiap iterasi penghapusan fitur/parameter. Hal ini dapat membantu menghindari terjadinya *overfitting* dan estimasi yang akurat. Selain itu, penggunaan RFECV juga dapat meningkatkan performa SVM seperti yang dilakukan pada beberapa penelitian sebelumnya yaitu penelitian yang dilakukan oleh Irfan Pratama yang melakukan seleksi fitur menggunakan RFECV pada data terkait demografis pegawai dan rekam jejak dari pekerjaan pegawai. Hasilnya menunjukkan bahwa jumlah fitur yang paling optimal yaitu 6 dari 29 yaitu “job\_duration\_in\_current\_job\_level”, “job\_duration\_in\_current\_person\_level”, “job\_duration\_in\_current\_branch”, “Employee\_type”, “Avg\_achievement\_%” dan “Last\_achievement\_%”. Model dengan seleksi fitur ini dapat meningkatkan akurasi model sebanyak 3% dibandingkan model yang tidak menggunakan seleksi fitur (I. Pratama dkk., 2022). Selain itu, penelitian dengan metode yang sama juga dilakukan oleh Arief Riski Indra Pratama yang melakukan optimasi klasifikasi curah hujan menggunakan SVM dan RFE. Hasilnya menunjukkan bahwa penerapan RFE pada SVM dapat meningkatkan akurasi sebesar 2% dari 77% menjadi 79%. Fitur terbaik yang dihasilkan berjumlah 3 dari 10 fitur yaitu Tavg (temperatur rata-rata), ss (lamanya penyinaran matahari) dan Tn (temperatur

minimum)(A. R. I. Pratama dkk., 2022). Selain RFECV, penelitian yang kami lakukan juga akan menggunakan metode SelectKbest dalam proses seleksi fitur sebagai perbandingan kinerja pada metode RFECV.

Berdasarkan permasalahan tersebut, penulis mengajukan judul **“Sistem Klasifikasi Kelayakan Air Minum dengan Seleksi Fitur Menggunakan Metode Recursive Feature Elimination with Cross-Validation (RFECV) dan SelectKbest”**. Tujuan utama dari penelitian ini yaitu seleksi fitur menggunakan metode RFECV untuk menentukan parameter yang paling optimal dalam proses klasifikasi kualitas air minum. Penelitian ini merupakan *preliminary research* yang bertujuan untuk mengevaluasi apakah metode yang digunakan dapat menjadi landasan untuk pengembangan pada penelitian selanjutnya yang melibatkan jumlah parameter yang lebih besar dan kompleks.

## 1.2 Rumusan Masalah

1. Bagaimana korelasi antar fitur dan terhadap target?
2. Bagaimana mengembangkan sistem klasifikasi kelayakan air minum menggunakan machine learning?
3. Bagaimana menemukan parameter yang perlu diutamakan dalam proses klasifikasi kelayakan air minum?
4. Bagaimana akurasi sistem klasifikasi kelayakan air minum melalui seleksi fitur dan klasifikasi tanpa seleksi fitur?

## 1.3 Tujuan Penelitian

Berdasarkan permasalahan yang telah dirumuskan, maka tujuan dari penelitian yang akan dilakukan adalah:

1. Untuk mengetahui hubungan antar fitur dan terhadap target
2. Untuk mengembangkan sistem klasifikasi kelayakan air minum menggunakan machine learning.
3. Untuk menemukan parameter yang perlu diutamakan dalam proses klasifikasi kelayakan air minum
4. Untuk mengetahui nilai akurasi dari sistem klasifikasi kelayakan air minum menggunakan seleksi fitur dan tanpa seleksi fitur.



#### **1.4 Manfaat Penelitian**

Penelitian ini diharapkan dapat memberikan manfaat antara lain:

1. Membantu Perusahaan penyedia air dalam pemantauan kualitas air secara otomatis dan peringatan dini terjadinya kontaminasi.
2. Sebagai dasar penelitian untuk penelitian selanjutnya terkait seleksi fitur dengan dataset yang memiliki jumlah parameter yang besar.
3. Sebagai informasi terkait parameter kualitas air yang perlu diutamakan dalam prediksi kualitas air minum.
4. Sebagai referensi bagi peneliti berikutnya dalam melakukan penelitian terkait sistem klasifikasi kelayakan air.

#### **1.5 Ruang Lingkup/Asumsi Perancangan**

1. Penelitian ini hanya fokus pada pengembangan AI dalam pembuatan sistem klasifikasi kelayakan air minum.
2. Metode yang digunakan untuk pengembangan sistem klasifikasi yaitu metode SVM dan XGBoost
3. Menggunakan 2 kategori kelas yaitu layak minum(1) dan tidak layak minum(0).
4. Penelitian menggunakan 6 parameter kualitas air yaitu kekeruhan, Ph, suhu, klor bebas, alkalinitas dan TDS.

#### **1.6 Sistematika Penulisan**

Untuk menjelaskan secara keseluruhan isi dari laporan tugas akhir ini, berikut ini sistematika penulisan dilakukan.

**BAB I PENDAHULUAN** Bagian ini memuat penjelasan mengenai latar belakang, masalah yang dirumuskan, tujuan penelitian, manfaat penelitian, batasan masalah, dan juga sistematika penulisan.

**BAB II TINJAUAN PUSTAKA** Bagian ini memuat beberapa referensi literatur yang berkaitan dengan topik penelitian dan teori-teori yang mendukung penelitian yang dilakukan.

BAB III METODOLOGI PENELITIAN Bagian ini memberikan gambaran tentang desain dan alur sistem serta berisi tentang metode-metode yang akan digunakan dalam penelitian.

BAB IV HASIL DAN PEMBAHASAN Bagian ini menggambarkan hasil penelitian serta membahas dan mengevaluasi penelitian secara keseluruhan.

BAB V PENUTUP Bagian ini menyajikan simpulan dari hasil penelitian serta memberikan saran untuk pengembangan sistem di masa depan.

## BAB II TINJAUAN PUSTAKA

### 2.1 Air Minum

Air minum adalah air yang melalui pengolahan atau tanpa pengolahan yang memenuhi syarat kesehatan dan dapat langsung diminum. Air minum digunakan untuk keperluan minum, masak, mencuci peralatan makan dan minum, mandi, mencuci bahan baku pangan yang akan dikonsumsi, peturasan dan ibadah (Permenkes, 2023). Air bersih untuk keperluan air minum, harus memenuhi persyaratan yang sudah ditetapkan oleh pemerintah. Air minum aman bagi kesehatan apabila memenuhi persyaratan secara fisika, mikrobiologi, kimia, dan radioaktif (Wandrivell dkk., 2012). Menurut Efendi (2003) kualitas air yaitu: sifat air, kandungan makhluk hidup, zat energi atau komponen lain dalam air. Kualitas air dinyatakan dengan beberapa parameter yaitu parameter fisika (suhu, kekeruhan, padatan terlarut), parameter kimia (PH, oksigen terlarut, BOD, kadar logam), parameter biologi (keberadaan lankton, bakteri dan sebagainya) (Gusril, 2016). Dalam Peraturan Menteri Kesehatan No.2 Tahun 2023, parameter dibagi menjadi parameter utama dan parameter khusus. Penetapan tambahan parameter khusus menjadi tanggung jawab pemerintah daerah melalui kajian ilmiah. Berikut ini merupakan parameter wajib tercantum dalam tabel 1.

Tabel 1 Parameter wajib kualitas air menurut Permenkes No.2 tahun 2023

| No       | Jenis Parameter                             | Kadar Maksimum<br>yang<br>diperbolehkan | Satuan     | Metode<br>pengujian |
|----------|---------------------------------------------|-----------------------------------------|------------|---------------------|
| <b>A</b> | <b>Mikrobiologi</b>                         | 0                                       | CFU/100 ml |                     |
| 1        | <i>Escherichia coli</i>                     | 0                                       | CFU/100 ml | SNI/APHA            |
| 2        | <i>Total Coliform</i>                       | 0                                       |            | SNI/APHA            |
| <b>B</b> | <b>Fisik</b>                                |                                         |            |                     |
| 3        | Suhu                                        | Suhu udara $\pm$ 3                      | °C         | SNI/APHA            |
| 4        | <i>Total Dissolve Solid</i>                 | <300                                    | Mg/L       | SNI/APHA            |
| 5        | Kekeruhan                                   | <3                                      | NTU        | SNI/APHA            |
| 6        | Warna                                       | 10                                      | TCU        | SNI/APHA            |
| 7        | Bau                                         | Tidak berbau                            | -          | APHA                |
| <b>C</b> | <b>Kimia</b>                                |                                         |            |                     |
| 8        | pH                                          | 6.5-8.5                                 | -          | SNI/APHA            |
| 9        | Nitrat (sebagai NO <sup>3</sup> )(terlarut) | 20                                      | mg/L       | SNI/APHA            |
| 10       | Nitrit (sebagai NO <sup>2</sup> )(terlarut) | 3                                       | mg/L       | SNI/APHA            |

|    |                                                 |                                        |      |          |
|----|-------------------------------------------------|----------------------------------------|------|----------|
| 11 | Kromium valensi 6 (Cr <sup>6+</sup> )(terlarut) | 0.01                                   | mg/L | SNI/APHA |
| 12 | Besi (Fe) (terlarut)                            | 0.2                                    | mg/L | SNI/APHA |
| 13 | Mangan (Mn)(terlarut)                           | 0.1                                    | mg/L | SNI/APHA |
| 14 | Sisa klor (terlarut)                            | 0.2 – 0.5 dengan waktu kontak 30 menit | mg/L | SNI/APHA |
| 15 | Arsen(As)(terlarut)                             | 0.01                                   | mg/L | SNI/APHA |
| 16 | Kadmium (Cd)(terlarut)                          | 0.003                                  | mg/L | SNI/APHA |
| 17 | Timbal (Pb)(terlarut)                           | 0.01                                   | mg/L | SNI/APHA |
| 18 | Flouride(F)(terlarut)                           | 1.5                                    | mg/L | SNI/APHA |
| 19 | Aluminium (Al)(terlarut)                        | 0.2                                    | mg/L | SNI/APHA |

(Sumber : Permenkes, 2023)

Selain parameter wajib, juga ditetapkan parameter khusus oleh Pemerintah Daerah sesuai dengan kondisi geohidrologi wilayah dan jenis kegiatan lingkungan wilayahnya berdasarkan hasil penelitian dan pengkajian. Berikut ini merupakan parameter khusus tercantum dalam tabel 2.

Tabel 2 Parameter khusus Kualitas Air Menurut Permenkes No.2 Tahun 2023

| No       | Jenis Parameter                                 | Kadar Maksimum yang diperbolehkan | Satuan | Metode pengujian |
|----------|-------------------------------------------------|-----------------------------------|--------|------------------|
| <b>A</b> | <b>Wilayah Pertanian/ Perkebunan/ Kehutanan</b> |                                   |        |                  |
| 1        | Fosfat (Fosfat sebagai P)                       | 0.2                               | mg/L   | SNI/APHA/US EPA  |
| 2        | Amoniak (NH <sup>3</sup> )                      | 1.5                               | mg/L   | SNI/APHA/US EPA  |
| 3        | Benzena                                         | 0.01                              | mg/L   | SNI/APHA/US EPA  |
| 4        | Toluen                                          | 0.7                               | mg/L   | SNI/APHA/US EPA  |
| 5        | Aldin                                           | 0.00003                           | mg/L   | SNI/APHA/US EPA  |
| 6        | Dieldrin                                        | 0.00003                           | mg/L   | SNI/APHA/US EPA  |
| 7        | Karbon organik total                            | 0.0007                            | mg/L   | SNI/APHA         |
| 8        | Kalium                                          | NA                                | mg/L   | SNI/APHA/US EPA  |
| 9        | Karbonat diklorida                              | NA                                | mg/L   | SNI/APHA/US EPA  |
| 10       | Aluminium fosfida                               | NA                                | mg/L   | SNI/APHA/US EPA  |
| 10       | Magnesium fosfida                               | NA                                | mg/L   | SNI/APHA/US EPA  |
| 11       | Sulfuril fluorida                               | NA                                | mg/L   | SNI/APHA/US EPA  |
| 12       | Metil bromida                                   | NA                                | mg/L   | SNI/APHA/US EPA  |
| 13       | Seng fosfida                                    | NA                                | mg/L   | SNI/APHA/US EPA  |
| 14       | Dkuat dibromida                                 | NA                                | mg/L   | SNI/APHA/US EPA  |
| 15       | Etil format                                     | NA                                | mg/L   | SNI/APHA/US EPA  |
| 16       | Fosfin                                          | NA                                | mg/L   | SNI/APHA/US EPA  |
| 17       | Parakuat diklorida                              | NA                                | mg/L   | SNI/APHA/US EPA  |
| 18       | Asam sulfur                                     | NA                                | mg/L   | SNI/APHA/US EPA  |
| 19       | Formaldehida                                    | NA                                | mg/L   | SNI/APHA/US EPA  |

|                                                     |                                                     |            |      |                 |
|-----------------------------------------------------|-----------------------------------------------------|------------|------|-----------------|
| 20                                                  | <i>Metanol</i>                                      | NA         | mg/L | SNI/APHA/US EPA |
| 21                                                  | <i>N-metil pirolidon</i>                            | NA         | mg/L | SNI/APHA/US EPA |
| 22                                                  | <i>Pridin base</i>                                  | NA         | mg/L | SNI/APHA/US EPA |
| 23                                                  | <i>Lindan</i>                                       | NA         | mg/L | SNI/APHA/US EPA |
| 24                                                  | <i>Heptaklor</i>                                    | NA         | mg/L | SNI/APHA/US EPA |
| 25                                                  | <i>Endrin</i>                                       | NA         | mg/L | SNI/APHA/US EPA |
| 26                                                  | <i>Endosulfan</i>                                   | NA         | mg/L | SNI/APHA/US EPA |
| 27                                                  | <i>Residu karbonat</i>                              | NA         | mg/L | SNI/APHA/US EPA |
| 28                                                  | <i>Organokhlorin</i>                                | NA         | mg/L | SNI/APHA/US EPA |
| 29                                                  | $\alpha$ -BHC                                       | NA         | mg/L | SNI/APHA/US EPA |
| 30                                                  | 4.4 -DDT                                            | NA         | mg/L | SNI/APHA/US EPA |
| 31                                                  | <i>Chlordan</i>                                     | NA         | mg/L | SNI/APHA/US EPA |
| 32                                                  | <i>Toxaphen</i>                                     | NA         | mg/L | SNI/APHA/US EPA |
| 33                                                  | <i>Heptaklor</i>                                    | NA         | mg/L | SNI/APHA/US EPA |
| 34                                                  | <i>Mirex</i>                                        | NA         | mg/L | SNI/APHA/US EPA |
| 35                                                  | <i>Polychlorinated biphenyl (PCB)</i>               | NA         | mg/L | SNI/APHA/US EPA |
| 36                                                  | <i>Hexachlorobenzene (HCB)</i>                      | NA         | mg/L | SNI/APHA/US EPA |
| 37                                                  | <i>Organofosfat</i>                                 | NA         | mg/L | SNI/APHA/US EPA |
| 38                                                  | <i>Pyretroid</i>                                    | NA         | mg/L | SNI/APHA/US EPA |
| 39                                                  | <i>Profonofos</i>                                   | NA         | mg/L | SNI/APHA/US EPA |
| 40                                                  | <i>Hexachlorobenzene</i>                            | NA         | mg/L | SNI/APHA/US EPA |
| <b>B Wilayah Industri</b>                           |                                                     |            |      |                 |
| 1                                                   | <i>Total Chromium (Cr)</i>                          | 0.05       | mg/L | SNI/APHA        |
| 2                                                   | <i>Amonia (NH<sub>3</sub>) (terlarut)</i>           | 1.5        | mg/L | SNI/APHA        |
| 3                                                   | <i>Hidrogen Sulfida (H<sub>2</sub>S) (terlarut)</i> | 0.05 – 0.1 | mg/L | SNI/APHA        |
| 4                                                   | <i>Sianida (CN)</i>                                 | 0.07       | mg/L | SNI/APHA        |
| 5                                                   | <i>Tembaga (Cu)</i>                                 | 2          | mg/L | SNI/APHA        |
| 6                                                   | <i>Selenium (Se)</i>                                | 0.01       | mg/L | SNI/APHA        |
| 7                                                   | <i>Seng (Zn)</i>                                    | 3          | mg/L | SNI/APHA        |
| 8                                                   | <i>Nikel (Ni)</i>                                   | 0.07       | mg/L | SNI/APHA        |
| 9                                                   | <i>Senyawa aktif biru metilen (MBAS)</i>            | -          |      | SNI/APHA        |
| 10                                                  | <i>Fenol (C<sub>6</sub>H<sub>5</sub>OH)</i>         | -          |      | SNI/APHA        |
| 11                                                  | <i>Fosfat (PO<sub>4</sub>)</i>                      | -          |      | SNI/APHA        |
| 12                                                  | <i>Methylene Blue Active Substances (MBAS)</i>      | -          |      | SNI/APHA        |
| 13                                                  | <i>Deterjen</i>                                     | -          |      | SNI/APHA        |
| <b>C Wilayah Pertambangan</b>                       |                                                     |            |      |                 |
| <b>Minyak, Gas, Panas Bumi, Sumber Daya Mineral</b> |                                                     |            |      |                 |
| 1                                                   | <i>Hidrogen Sulfida (H<sub>2</sub>S) (terlarut)</i> | 0.05 -0.1  | mg/L | SNI/APHA        |
| 2                                                   | <i>Merkuri (Hg)</i>                                 | 0.001      | mg/L | SNI/APHA        |
| 3                                                   | <i>Tembaga (Cu)</i>                                 | 2          | mg/L | SNI/APHA        |
|                                                     | <i>Radioaktif</i>                                   |            | mg/L | SNI/APHA        |
| 4                                                   | <i>Gross alpha activity</i>                         | 0.1        | Bq/L | SNI/APHA        |

|    |                                                                                |        |      |          |
|----|--------------------------------------------------------------------------------|--------|------|----------|
| 5  | <i>Gross beta activity</i>                                                     | 1      | Bq/L | SNI/APHA |
| 6  | <i>Hidrokarbon<br/>polyaromatis</i>                                            | 0.0007 | mg/L | SNI/APHA |
| 7  | <i>Nikel (Ni)</i>                                                              | 0.07   | mg/L | SNI/APHA |
| 8  | <i>Timbal</i>                                                                  | 0.01   | mg/L | SNI/APHA |
| 9  | <i>Amonia (NH<sub>3</sub>) (terlarut)</i>                                      | 1.5    | mg/L | SNI/APHA |
| 10 | <i>Fenol (C<sub>6</sub>H<sub>6</sub>CO)<br/>(C<sub>6</sub>H<sub>5</sub>OH)</i> |        |      | SNI/APHA |

(Sumber : Permenkes, 2023)

## 2.2 Machine Learning

*Machine Learning* adalah cabang ilmu komputer yang secara luas bertujuan untuk memungkinkan komputer ‘belajar’ tanpa diprogram secara langsung. Hal ini berawal dari gerakan kecerdasan buatan pada tahun 1950-an dan menekankan pada tujuan dan aplikasi praktis, khususnya prediksi dan optimasi (Bi dkk., 2019). *Machine learning* mengacu pada deteksi otomatis pola yang bermakna dalam data (Babcock University dkk., 2017). *Machine learning* ini digunakan untuk mengajarkan mesin cara menangani data dengan lebih efisien dengan mengandalkan algoritma yang berbeda untuk menyelesaikan masalah data. Jenis algoritma yang digunakan tergantung pada jenis masalah yang ingin diselesaikan, jumlah variabel, jenis model yang paling sesuai dan sebagainya (Mahesh, 2020). Berikut ini tiga kategori *machine learning*.

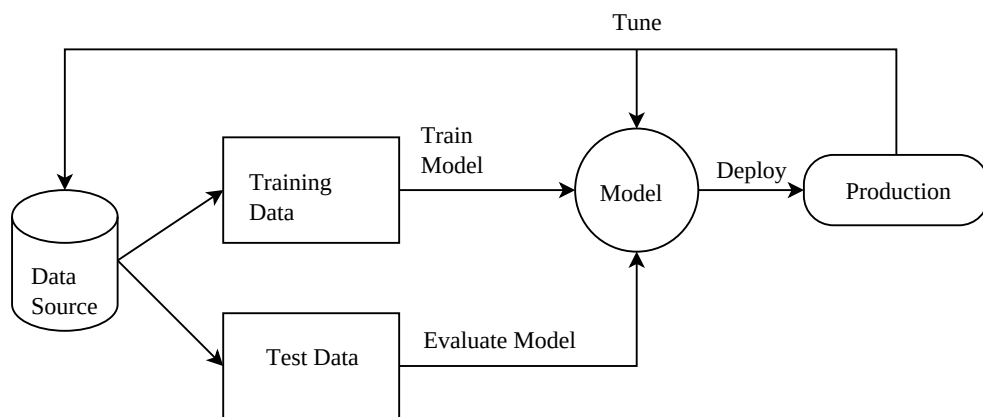
### 2.2.1 Supervised Learning

*Supervised learning* adalah tugas pembelajaran mesin untuk belajar input ke output berdasarkan contoh pasangan input-output. Ini menyimpulkan sebuah fungsi dari pelatihan berlabel yang terdiri dari sekumpulan contoh pelatihan (Mahesh, 2020). *Supervised learning* mirip dengan jenis model fitting yang menjadi standar dalam praktik epidemiologi. Nilai dari hasil (variabel dependen), sering disebut “label” dalam pembelajaran mesin, diketahui untuk setiap pengamatan (Bi dkk., 2019). Model fungsi yang terlatih dengan baik berdasarkan algoritma pembelajaran yang diawasi dapat secara akurat memprediksi label kelas untuk fenomena tersembunyi yang tertanam dalam atau contoh data yang tidak teramati. Tujuan dari *supervised learning* adalah untuk meminimalkan kesalahan untuk sekumpulan input yang diberikan (set

pelatihan (Award & Khanna, 2015). *Supervised learning* terdiri dari dua bentuk yang berbeda yaitu :

1. Regresi: Output targetnya yaitu bilangan riil atau vector bilangan riil utuh seperti harga saham dalam jangka waktu 6 bulan atau suhu pada siang hari besok.
2. Klasifikasi: Output target adalah label kelas seperti pada kasus sederhana yaitu memilih antara positif dan negatif (Dasgupta & Nath, 2016).

Beberapa algoritma *supervised learning* yang paling populer yaitu *Decision tree*, *Support Vector Machine*, *Random Forest* dan *Naive Bayes*.

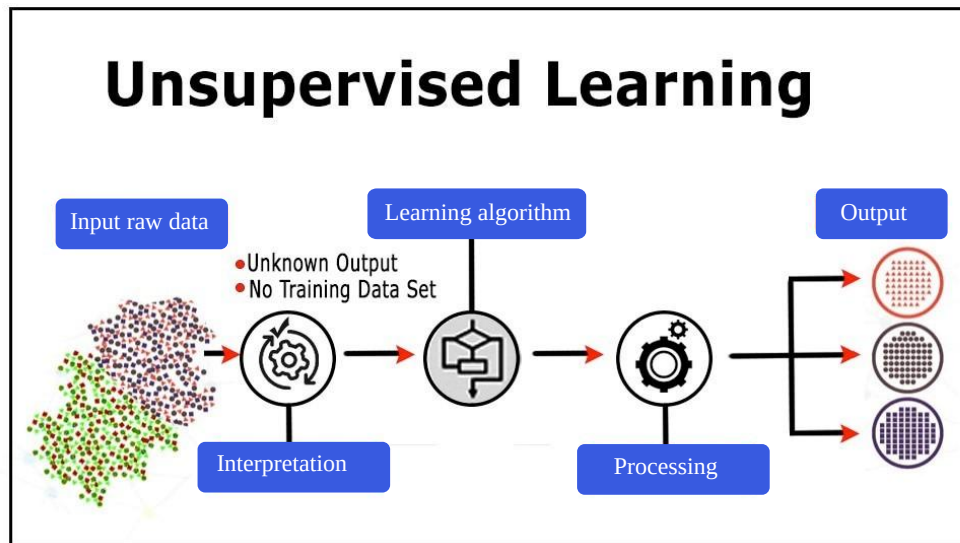


Gambar 1 Workflow supervised learning (Mahesh, 2020)

### 2.2.2 Unsupervised Learning

*Unsupervised learning* adalah algoritma yang dirancang untuk menemukan struktur tersembunyi dalam kumpulan data yang tidak berlabel, dimana keluaran yang diinginkan tidak diketahui (Award & Khanna, 2015). *Unsupervised learning* memiliki kesamaan dalam tujuan dan struktur dengan pendekatan statistik yang mencoba untuk mengidentifikasi subkelompok yang tidak ditentukan dengan karakter yang sama (misalnya “*laten*”, variabel dan kelas) (Bi dkk., 2019). *Unsupervised learning* mempelajari beberapa fitur dari data. Ketika data baru diperkenalkan, ia menggunakan fitur yang telah dipelajari sebelumnya untuk mengenali kelas data. Mekanisme ini memiliki banyak kegunaan di bidang kompresi data, deteksi *outlier*, klasifikasi, pembelajaran manusia dan sebagainya. Contoh implelementasi *unsupervised*

*learning* yaitu jeruk dan bola pantai berbentuk bulat. Beberapa algoritma yang biasa digunakan yaitu *Principal Component Analysis* dan *K-Means*.



Gambar 2 *Unsupervised learning* (Mahesh, 2020).

### 2.2.3 Semi Supervised Learning

*Semi Supervised Learning* yaitu kombinasi dari metode *supervised learning* dan *unsupervised learning*. Metode ini cocok dengan model untuk data berlabel dan data yang tidak berlabel. Pelabelan data (hasil) sering kali memakan waktu dan mahal, terutama untuk kumpulan data yang besar. *Semi supervised learning* melengkapi data berlabel yang terbatas dengan berlimpahnya data yang tidak berlabel dengan tujuan untuk meningkatkan kinerja model (Biddik, 2019). Beberapa algoritma yang umum digunakan yaitu *Generative models*, *self training* dan *transductive support vector machine*.

## 2.3 Klasifikasi

Klasifikasi merupakan proses menemukan pola yang bertujuan untuk memperkirakan kelas dari objek yang belum diketahui. Klasifikasi dapat dideskripsikan sebagai metode untuk membuktikan sebuah objek data sebagai salah satu jenis yang telah dideskripsikan sebelumnya (Yuniar, 2023). Sebuah pengklasifikasi dibuat dari sekumpulan data latih dengan kelas yang telah ditentukan. Vektor fitur pelatihan tersedia dan telah diketahui kelas-kelasnya, kemudian vektor fitur pelatihan tersebut dimanfaatkan untuk merancang pemilah.



Pengenalan pola ini disebut terbimbing, *supervised* (Wibawa dkk., 2018). Secara umum, proses klasifikasi dapat dilakukan dalam dua tahap, yaitu proses belajar dari data pelatihan dan klasifikasi kasus. Pada proses belajar, algoritma klasifikasi mengolah data *training* untuk menghasilkan sebuah model. Setelah model diuji dan dapat diterima, pada tahap klasifikasi model tersebut digunakan untuk memprediksi kelas dari kasus baru untuk membantu proses pengambilan keputusan (Lestari dkk., 2020).

## 2.4 Seleksi Fitur

Seleksi fitur merupakan proses pemilihan atribut yang sesuai dari sekumpulan atribut. Algoritma memproses data lebih cepat dengan mengurangi atribut atau fitur yang tidak relevan (Zulaikhah Hariyanti Rukmana dkk., 2022). Seleksi fitur merupakan pemilihan otomatis atribut dalam data yang paling relevan dengan masalah pemodelan yang dilakukan. Metode pemilihan fitur membantu dalam memilih fitur yang akan memberikan akurasi yang sama atau lebih baik dengan data yang lebih sedikit. Metode pemilihan fitur dapat digunakan untuk mengidentifikasi dan menghapus atribut yang tidak diperlukan, tidak relevan dan berlebihan dari data yang tidak berkontribusi pada akurasi model prediktif atau bahkan dapat menurunkan akurasi model. Selain itu, atribut yang lebih sedikit lebih baik karena dapat mengurangi kompleksitas model, dan model yang lebih sederhana akan lebih mudah dipahami dan dijelaskan. (Brownlee, 2014).

Seleksi fitur memiliki 3 kategori algoritma yaitu metode *filter*, *wrapper*, dan *embedded*.

### 1. Metode Filter

Metode pemilihan fitur filter menerapkan ukuran statistik untuk memberikan skor pada setiap fitur. Fitur-fitur diberi peringkat berdasarkan skor dan dipilih untuk disimpan atau dihapus dari kumpulan data. Metode ini seringkali bersifat univariat dan mempertimbangkan fitur secara independen atau berkaitan dengan variabel dependen.

### 2. Metode Wrapper

Metode *wrapper* merupakan metode pemilihan fitur yang mempertimbangkan pemilihan sekumpulan fitur sebagai masalah pencarian, Dimana kombinasi yang berbeda disiapkan, dievaluasi, dan dibandingkan dengan kombinasi lainnya. Sebuah model prediktif digunakan untuk mengevaluasi kombinasi fitur dan memberikan skor berdasarkan akurasi model. Contohnya yaitu algoritma *Recursive Feature Elimination*.

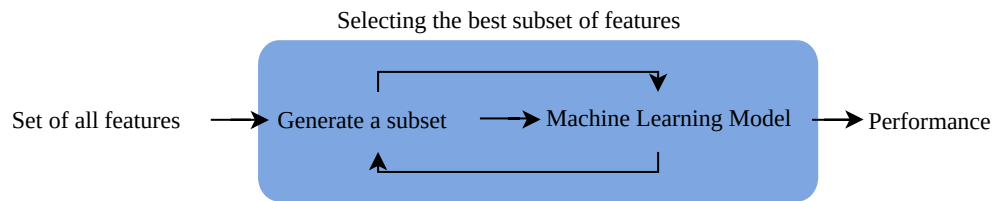
### 3. Metode *Embedded*

Metode *embedded* merupakan metode yang mempelajari fitur mana yang paling berkontribusi terhadap akurasi model saat model sedang dibuat. Jenis metode pemilihan fitur *embedded* yang paling umum adalah metode regularisasi. Contoh algoritma regularisasi yaitu LASSO, *Elastic Net*, dan *Ridge Regression*.

## 2.5 *Recursive Feature Elimination with Cross-Validation (RFECV)*

RFE adalah metode pemilihan fitur *wrapper* yang membungkus model *machine learning* dan memilih jumlah fitur yang optimal untuk setiap model secara rekursif menghilangkan fitur 0-n disetiap putaran. Selajutnya, metode ini memilih subset fitur dengan kinerja terbaik berdasarkan akurasi atau skor validasi silang. RFECV juga menghilangkan ketergantungan dan kolinearitas yang ada di dalam model (Mowri dkk., 2023). RFECV memperluas RFE dengan memasukkan proses resampling dan validasi silang. Ini menghitung kesalahan validasi untuk semua subset fitur dan memilih subset dengan tingkat kesalahan terendah sebagai fitur yang optimal (Shi dkk., 2024). RFECV mencakup empat langkah berikut :

1. Melatih model dasar menggunakan dataset fitur asli
2. Menghitung nilai kepentingan setiap fitur
3. Membentuk subset fitur baru dengan menghapus variabel fitur melalui validasi silang
4. Mengulangi langkah a hingga c hingga jumlah variabel fitur yang sesuai ditentukan.



Gambar 3 Teknik *feature selection* RFECV (Shi dkk., 2024)

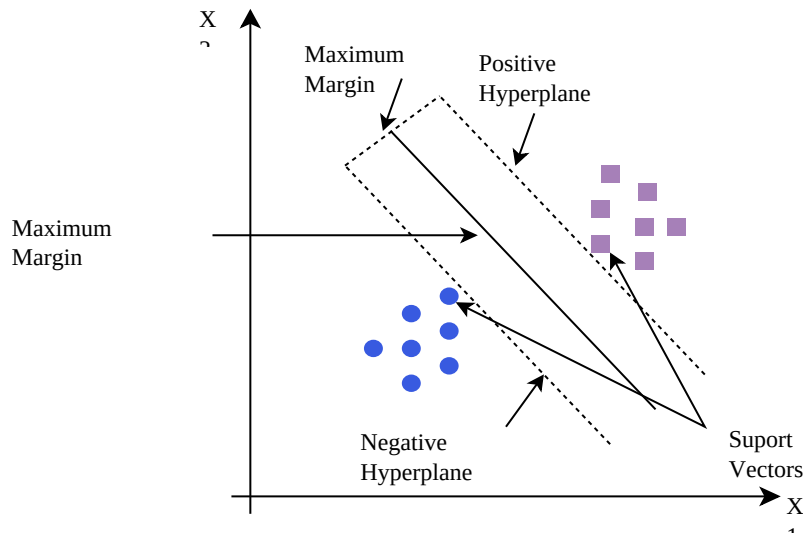
## 2.6 SelectKbest with $f\_classif$

SelectKbest adalah algoritma pemilihan fitur metode filter. SelectKbest merupakan teknik seleksi fitur yang memilih sejumlah fitur terbaik berdasarkan skor statistik tertentu. Ini dapat membantu mengurangi dimensi data, meningkatkan efisiensi, dan performa model dengan mengeliminasi fitur yang kurang relevan (Febrianti & Najah, 2024). Algoritma ini bertujuan untuk memilih fitur input dengan membandingkan hasilnya ketika disertakan dalam fungsi tertentu. SelectKbest adalah bagian dari tahapan seleksi fitur univariat yang memilih fitur terbaik berdasarkan uji statistik univariat atau uji ANOVA (Analysis of Variance) (Fitri dkk., 2023). *F-classification* menggunakan uji ANOVA untuk menentukan fitur mana yang paling signifikan dalam membedakan kelas-kelas dalam dataset. Skor F tinggi menunjukkan fitur yang informatif dan relevan untuk model (Febrianti & Najah, 2024). SelectKbest menghapus semua kecuali fitur yang memiliki skor tertinggi. SelectKbest memilih  $k$  fitur teratas dengan relevansi terbesar dengan variabel target (Fitri dkk., 2023).

## 2.7 Support Vector Machine (SVM)

SVM adalah metode yang kuat untuk pengembangan klasifikasi. Hal ini bertujuan untuk membuat batas keputusan antara dua kelas yang memungkinkan prediksi label dari satu atau lebih fitur vektor (Huang dkk., 2018). SVM biner adalah pengklasifikasi yang membedakan titik data dari dua kategori. Setiap objek data (titik data) diwakili oleh vector  $n$ -dimensi. Setiap titik data ini hanya termasuk dalam salah satu dari dua kelas. Algoritma SVM bekerja dengan cara membuat garis atau batas keputusan terbaik yang dapat memisahkan ruang dimensi  $n$  ke dalam kelas-kelas sehingga mudah dalam menempatkan titik data baru kedalam kategori yang benar. Batas keputusan ini disebut *hyperplane*. Selain itu, pada

algoritma SVM terdapat titik atau vektor ekstrim yang dapat membantu dalam menciptakan *hyperplane* yang disebut *support vector* (Sudianto, 2022). Berikut ini ilustrasi cara kerja SVM pada Gambar 4.



Gambar 4 Ilustrasi kerja SVM (Sudianto, 2022)

Untuk mencapai pemisahan maksimum antara dua kelas, SVM memilih *hyperplane* yang memiliki margin terbesar. Margin adalah penjumlahan jarak terpendek dari *hyperplane* pemisah ke titik data terdekat dari kedua kategori. *Hyperplane* seperti itu cenderung menggeneralisasi dengan lebih baik, yang berarti bahwa *hyperplane* tersebut mengklasifikasikan dengan benar titik data yang “tidak terlihat” atau titik data pengujian (Yu & Kim, 2012).

Pada dasarnya, metode SVM digunakan untuk proses klasifikasi data yang hanya memiliki dua kelas dan *linearly separable* sehingga ketika suatu masalah tidak dapat menemukan *hyperplane* pemisah yang kuat yang meminimalkan jumlah titik data yang salah klasifikasi dan dapat digeneralisasi dengan baik (Awad & Khanna, 2015). Untuk mengatasi masalah data tidak terpisah secara linear tersebut, digunakan fungsi kernel *Radial Basis Function* (RBF) (Ma'ruf dkk., 2019). Kernel RBF memiliki dua parameter yaitu Gamma dan *Cost*. Parameter *Cost* merupakan parameter yang berfungsi untuk menghindari misklasifikasi pada setiap sampel *training*. Berikut persamaan dari kernel RBF :

$$\text{Radial Basis Function (RBF): } K(\vec{x}_i, \vec{x}_j) = \exp\left(-\gamma \times \|\vec{x}_i - \vec{x}_j\|^2\right) \quad (1)$$

dimana,

|                           |                                                                                           |
|---------------------------|-------------------------------------------------------------------------------------------|
| $K(\vec{x}_i, \vec{x}_j)$ | = Fungsi kernel RBF                                                                       |
| $\vec{x}$                 | = Vektor input                                                                            |
| $\exp$                    | = Fungsi eksponensial                                                                     |
| $\gamma$                  | = Parameter gamma, menentukan pengaruh individu sampel training terhadap fungsi keputusan |

## 2.8 eXtreme Gradient Boosting (XGBoost)

XGBoost merupakan salah satu implementasi dari gradient boosting yang dikenal sebagai salah satu algoritma dengan performa terbaik yang digunakan untuk *supervised learning*. Algoritma ini dapat digunakan untuk masalah prediksi dan klasifikasi. XGBoost juga memiliki kecepatan eksekusi yang tinggi diluar komputasi inti (Ibrahim Ahmed Osman dkk., 2021). XGBoost adalah algoritma yang ditingkatkan berdasarkan *gradient boosting decision tree* dan dapat membangun *boosted trees* secara efisien dan beroperasi secara paralel. Didalam pohon regresi, *nodes* bagian dalam mewakili nilai-nilai untuk tes atribut dan *leaf nodes* dengan skor yang mewakili keputusan (Karo, 2020). Cara XGBoost membuat prediksi dapat dilihat pada persamaan (2).

$$\hat{y}_i = \sum_{k=1}^n f_k(x_i), f_k \in F \quad (2)$$

dimana,

|            |                                                          |
|------------|----------------------------------------------------------|
| $y_i$      | = Nilai prediksi model XGBoost                           |
| $x_i$      | = Vektor input                                           |
| $k$        | = Indeks pada setiap fungsi $f_k$                        |
| $f_k$      | = Fungsi ke-k, fungsi pohon keputusan ke-k               |
| $f_k(x_i)$ | = Fungsi pohon keputusan ke-k saat diberikan input $x_i$ |
| $F$        | = Himpunan dari seluruh fungsi pohon keputusan           |

Adapun fungsi objektif dari XGBoost dapat dilihat pada persamaan (3)

$$\text{Obj}(\theta) = L(\theta) + \Omega(\theta) \quad (3)$$

dimana,

|                      |                   |
|----------------------|-------------------|
| $\text{Obj}(\theta)$ | = Fungsi objektif |
|----------------------|-------------------|

$\theta$  = Parameter model

$L(\theta)$  = Fungsi loss/kerugian

$\Omega(\theta)$  = Fungsi regularisasi yang mengontrol kompleksitas model

$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i)$  merupakan fungsi loss,  $\hat{y}_i$  adalah prediksi dan  $y_i$  adalah target dan  $\Omega(\theta) = \sum_{k=1}^K \Omega(f_k)$  merupakan regularisasi yang menjumlahkan setiap kompleksitas setiap pohon keputusan. Kemudian model dilatih dengan cara aditif. Kemudian membiarkan  $\hat{y}_i^{(t)}$  menjadi prediksi dari instance  $i$  pada iterasi  $t$ , dan dapat dinyatakan pada persamaan (4).

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (4)$$

dimana,

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

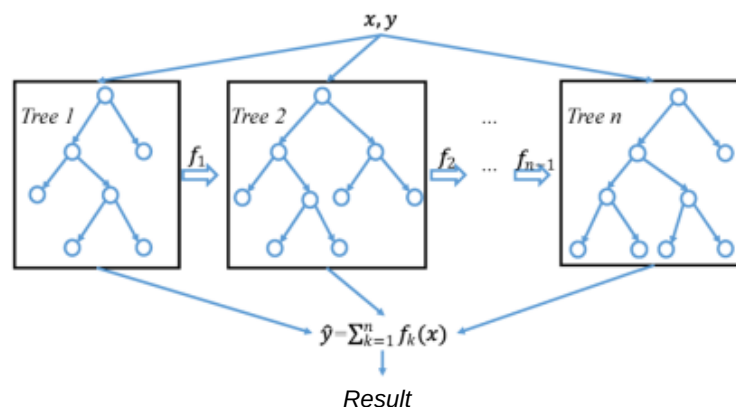
$\hat{y}_i^{(t-1)}$  = Prediksi untuk instance/ data ke- $i$  pada iterasi sebelumnya( $t-1$ )

$f_t$  = Fungsi pohon keputusan baru yang ditambahkan pada iterasi ke- $t$

$x_i$  = input untuk instance/data ke- $i$

$f_t(x_i)$  = Prediksi dari pohon baru untuk instance ke- $i$

Berikut ini merupakan *architecture* umum pada model XGBoost dapat dilihat pada Gambar 5.



Gambar 5 *Architecture* model XGBoost (Wang dkk., 2019)

Gambar 5 menunjukkan *architecture* umum dari model XGBoost. Berdasarkan gambar tersebut, prediksi akhir dari XGBoost dihasilkan dari jumlah hasil seluruh pohon keputusan yang digunakan dimana setiap pohon (Tree 1,2,...n) dilatih untuk memperbaiki kesalahan prediksi dari pohon sebelumnya.