# Machine Learning
# UAS


Disusun oleh:

Muhammad Ghaniyu Haq Haryanto 2041720178

Raka Bagas Fitriansyah 2041720187

Wahyu Rizky Akbari 2041720191

**PROGRAM STUDI D-IV TEKNIK INFORMATIKA**
**JURUSAN TEKNOLOGI INFORMASI**
**POLITEKNIK NEGERI MALANG**
**2022**

# Pengujian Kualitas Air Untuk Komsumsi
# Dengan Metode Gaussian Naïve Baiyes dan Random Forest

**Muhammad Ghaniyu Haq Haryanto[1], Raka Bagas Fitriansyah, Wahyu Rizky Akbari[3]**

[1,2,3] Prodi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang
[1]2041720178@student.polinema.ac.id, [2] 2041720187@student.polinema.ac.id,
[3]2041720191@student.polinema.ac.id

## ABSTRAK

Upaya membantu strategi pengelolaan sumber daya air dengan teknologi pada pemantauan kualitas air sungai dapat dilakukan dengan penerapan prediksi parameter kualitas air menggunakan metode Naïve bayes dan random fores, untuk mendapatkan hasil yang lebih cepat, efektif, akurat, murah, dan berdaya guna tinggi. Penerapan metode diatas dilakukan untuk memprediksi parameter kualitas air yaitu Ph, kandungan mineral, senyawa kimia, keasaman, metana, kekeruhan. Didapatkan sebesar 76,397 % untuk potability tidak, sedangkan untuk potability ya sebesar 23,603%. Sehingga dari hasil tersebut dapat disimpulkan potability nya cendurung pada tidak. Didapatkan dari hasil perhitungan kode program gaussian naïve bayes diatas, mengenai water potability. Didapatkan sebesar 0,640 atau 64% untuk potability tidak. Sehingga dari hasil tersebut dapat disimpulkan potability nya cendurung pada tidak. Dari salah satu tree pada random forest kami, Didapatkan hasil weighted gini index sebesar 0,712 Didapatkan dari hasil perhitungan kode program random forest diatas, mengenai water potability. Didapatkan sebesar 0,661 atau 66,1% untuk potability tidak. Sehingga dari hasil tersebut dapat disimpulkan potability nya cendurung pada tidak. Dengan menggunakan 2 metode. Gaussian Naïve Bayes serta Random Forest didapatkan hasil bahwa metode Random Forest lebih unggul dibandingkan metode Gaussian Naïve Bayes. Serta dari dataset tersebut dapat disimpulkan bahwa air tidak layak untuk dikonsumsi berdasarkan aspek aspek yang tersedia.

**Kata kunci :** Water Potability, Gaussian Naïve Bayes, Random Forest

## Pendahuluan

Air merupakan bahan alam yang diperlukan untuk kehidupan manusia, hewan dan tanaman yaitu sebagai media pengangkutan zat-zat makanan, juga merupakan sumber energi

serta berbagai keperluan lainnya. Masalah utama yang dihadapi berkaitan dengan sumber daya air adalah kuantitas air yang sudah tidak mampu memenuhi kebutuhan yang terus meningkat dan kualitas air untuk keperluan domestik yang semakin menurun dari tahun ke tahun. Kegiatan industri, domestik, dan kegiatan lain berdampak negatif terhadap sumber daya air, termasuk penurunan kualitas air. Kondisi ini dapat menimbulkan gangguan, kerusakan, dan bahaya bagi mahluk hidup yang bergantung pada sumber daya air.

Kecerdasan buatan yang menjadi penggerak revolusi industri yang saat ini masuk pada era ke 4, yang telah menghasilkan perubahan peradaban secara signifikan. Saat ini revolusi industri keempat yang sering juga disebut revolusi digitalisasi kembali menyeruak dengan jaringan sibernya. Selain itu mesin-mesin dan komputer mulai diambil alih oleh kecerdasan buatan atau artificial intelligence (AI). Jika dulunya manusia yang banyak menggunakan pikirannya, kini giliran mesin atau komputer yang banyak berpikir dengan kecerdasan buatannya untuk menggantikan atau membantu pekerjaan manusia.

Upaya membantu strategi pengelolaan sumber daya air dengan teknologi pada pemantauan kualitas air sungai dapat dilakukan dengan penerapan prediksi parameter kualitas air menggunakan metode Naïve bayes dan random fores, untuk mendapatkan hasil yang lebih cepat, efektif, akurat, murah, dan berdaya guna tinggi. Penerapan metode diatas dilakukan untuk memprediksi parameter kualitas air yaitu Ph, kandungan mineral, senyawa kimia, keasaman, metana, kekeruhan.

## METODE

1. Dataset
    1) Dataset awal

        Dataset yang kami gunakan bersumber pada kaggle.com. Dimana data tersebut membasa tentang kelayakan air untuk dikomsunsi. Berikut kami tampilkan 5 data teratas dari dataset yang digunakan:

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | NaN | 204.890456 | 20791.31898 | 7.300212 | 368.516441 | 564.308654 | 10.379783 | 86.990970 | 2.963135 | 0 |
| 1 | 3.716080 | 129.422921 | 18630.05786 | 6.635246 | NaN | 592.885359 | 15.180013 | 56.329076 | 4.500656 | 0 |
| 2 | 8.099124 | 224.236259 | 19909.54173 | 9.275884 | NaN | 418.606213 | 16.868637 | 66.420093 | 3.055934 | 0 |
| 3 | 8.316766 | 214.373394 | 22018.41744 | 8.059332 | 356.886136 | 363.266516 | 18.436525 | 100.341674 | 4.628771 | 0 |
| 4 | 9.092223 | 181.101509 | 17978.98634 | 6.546600 | 310.135738 | 398.410813 | 11.558279 | 31.997993 | 4.075075 | 0 |

    2) Deskripsi Dataset

        Terdapat 10 kolom pada dataset kami. Berikut adalah penjelasan setiap kolom pada dataset kami:

a. pH

   PH merupakan parameter penting dalam mengevaluasi keseimbangan asam-basa air. Ini juga merupakan indikator kondisi status air asam atau basa.

b. Hardeness

   Kesadahan terutama disebabkan oleh garam kalsium dan magnesium. Garam-garam ini larut dari endapan geologis yang dilalui air. Lamanya waktu air bersentuhan dengan bahan penghasil kesadahan membantu menentukan berapa banyak kesadahan yang ada dalam air baku.

c. Solids

   Air memiliki kemampuan untuk melarutkan berbagai macam mineral anorganik dan beberapa mineral atau garam organik seperti kalium, kalsium, natrium, bikarbonat, klorida, magnesium, sulfat.

d. Chloramines

   Klorin dan kloramin adalah disinfektan utama yang digunakan dalam sistem air publik. Chloramines paling sering terbentuk ketika amonia ditambahkan ke klorin untuk mengolah air minum. Tingkat klorin hingga 4 miligram per liter (mg/L atau 4 bagian per juta (ppm)) dianggap aman dalam air minum.

e. Sulfate

   Sulfat adalah zat alami yang ditemukan dalam mineral, tanah, dan batuan. Mereka hadir di udara sekitar, air tanah, tumbuhan, dan makanan.

f. Conductivity

   Air murni bukanlah penghantar arus listrik yang baik, melainkan isolator yang baik. Peningkatan konsentrasi ion meningkatkan konduktivitas listrik air.

g. Organic_Carbon

   Total Karbon Organik (TOC) di perairan sumber berasal dari bahan organik alami (NOM) yang membusuk serta sumber sintetis. TOC adalah ukuran jumlah total karbon dalam senyawa organik dalam air murni.

h. Thrihalometanes

   THM adalah bahan kimia yang dapat ditemukan dalam air yang diolah dengan klorin. Konsentrasi THM dalam air minum bervariasi sesuai dengan kadar bahan organik dalam air, jumlah klorin yang dibutuhkan untuk mengolah air, dan suhu air yang diolah.

i. Turbidity

   Kekeruhan air tergantung pada jumlah zat padat yang ada dalam keadaan tersuspensi.

j. Potability

   Menunjukkan apakah air aman untuk dikonsumsi manusia di mana 1 berarti Dapat Diminum dan 0 berarti Tidak Dapat Diminum.

2. Drop Column

   Dataset pada project kami yang tidak efektif dihapus. Dikarenakan tidak cocok pada hasil yang diharapkan. Berikut kolom yang dihapus :

```
[8]  del data['Conductivity']
     del data['Organic_carbon']
     del data['Trihalomethanes']
     del data['Turbidity']
```

   Untuk kolom yang dihapus pada kolom conductivity, organic_carbon, thrihalomethanes, turbidity. Karena kelompok kami hanyak mengambil kolom yang berkorelasi kuat dengan kolom potability.

3. Spliting Data

   Membagi dataset menjadi data train dan data test. Dengan variabel x berisi semua data kolom setelah dilakukan drop column. Dan variabel y berisikan kolom potability. Berikut disajikan gambar proses split data:

```
X = data.drop(['Potability'], axis=1)
y = data['Potability']
```

```
X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size = 0.15, random_state = 42)
```

   Proses split menggunakan test size sebesar 0.15. Dan test split ini berarti menggunakan perbandingan pembagian data sebesar 8.5 : 1.5.

4. Gaussian Naïve Bayes

   Pengklasifikasi Naive Bayes paling sederhana yang memiliki asumsi bahwa data dari masing-masing label diambil dari distribusi Gaussian sederhana. Asumsi pendistribusian nilai kontinu yang terkait dengan setiap fitur berisi nilai numerik

   a. Rumus GNB

a) Mean

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}}$$

Dimana jumlah atribut n dengan nilai X dibagi dengan Jumlah n.

b) Standar Deviasi

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

Selanjutnya adala nilai Standar Deviasi (STD) dan yang kita hitung adalah nilai varian yang mempresentasikan seluruh populasi,

Keterangan:

σ = varian satau ragam untuk populasi

xi = Titik tengah nilai dalam satu atribut

μ = rata-rata atau mean dari populasi

n = Jumlah data

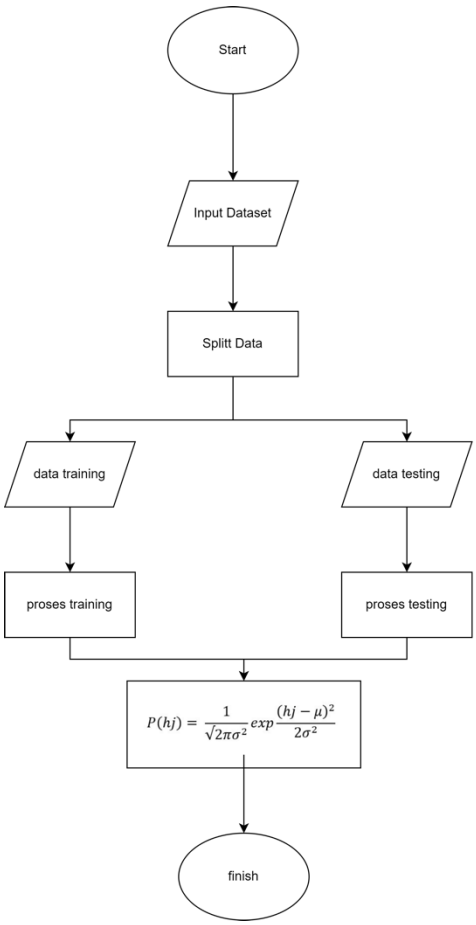
c) Gaussian Value

$$P(hj) = \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(hj - \mu)^2}{2\sigma^2}$$

Setelah kalian sudah menghitung dari masing-masing atribut dengan dua formula diatas,selanjutnya adalah menghitung dengan formula *Gaussian* di atas. Jadi ada dua **tahapan** jika ingin menggunakan algoritma *naive bayes* pada kasus atribut data bernilai numerik atau kontinyu, yaitu

1. Mendiskritkan data yang bernilai kontinu
2. Menggunakan Fungsi *Gaussian* untuk menentukan nilai probabilitas dalam distribusi normal.

b. Flowchart GNB




c. Rumus Excel GNB

a) Mean

```
=AVERAGE('FILTER BERDASARKAN POTABILITY'!B4:B1281)
```

b) Standar Deviasi

```
=STDEV.S('FILTER BERDASARKAN POTABILITY'!B3:B1280)
```

c) Probability Data

```
=COUNTIF('CONVERT DATA SET DAN SET NAN'!X2:X3277;'PROBABILITY DATA'!B4)/COUNTA('CONVERT DATA SET DAN SET NAN'!X2:X3277)
```

d) Gaussian Value

```
=1/SQRT(2*3,14*'STDEV DATA'!B4)*EXP(-((C5-'RATA RATA DATA'!B5)^2/(2*'STDEV DATA'!B4^2)))
```

```
=C6*D6*E6*F6*G6*'PROBABILITY DATA'!C4
```

5. Random Forest

Random Forest adalah algoritma dalam machine learning yang digunakan untuk pengklasifikasian data set dalam jumlah besar. Karena fungsinya bisa digunakan untuk banyak dimensi dengan berbagai skala dan performa yang tinggi. Klasifikasi ini dilakukan melalui penggabungan tree dalam decision tree dengan cara training dataset yang Anda miliki.

a. Rumus Random Forest

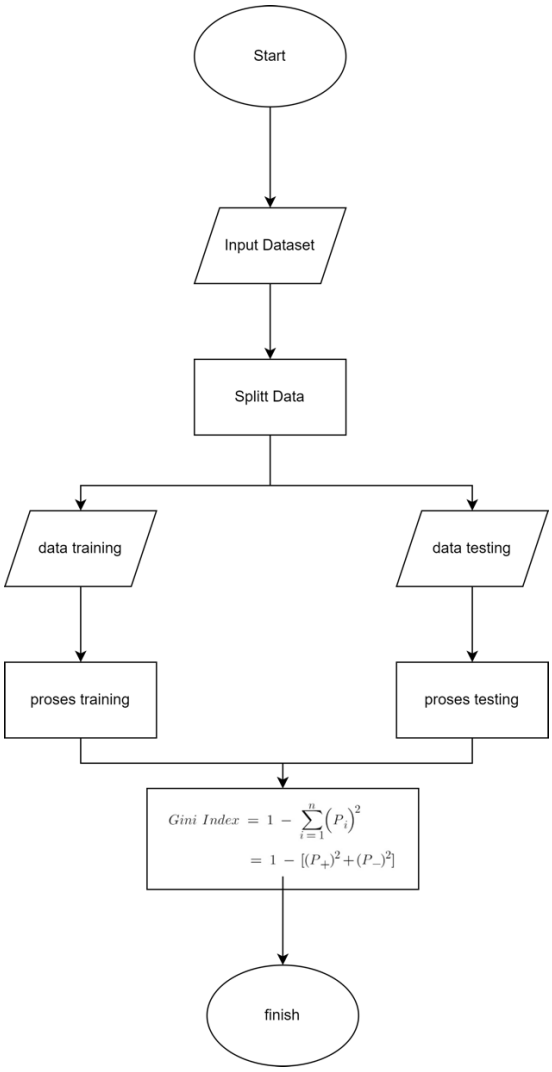
a) Gini Index

$$Gini\ Index = 1 - \sum_{i=1}^{n}\left(P_i\right)^2$$
$$= 1 - [(P_+)^2 + (P_-)^2]$$

b) Weighted Gini

$$E(S) = -p_{(+)}\log p_{(+)} - p_{(-)}\log p_{(-)}$$

b. Flowchart Random Forest



c. Rumus Excel Random Forest

a) Gini Index

```
=1-((J13 / $H$5)^2 + (J14 / $H$5)^2)
```

b) Weighted Gini

```
=((J13/H5)*J16)+((M14/H5)*M16)
```

**Hasil dan Pembahasan**

1. Gaussian Naïve Bayes

a. Hasil Perhitungan Manual

Dari perhitungan berdasarkan dataset diatas, didapatkan hasil sebagai berikut:

| KESIMPULAN | | |
|---|---|---|
| TOTAL TES | | |
| YA | 773 | 23,603% |
| TIDAK | 2502 | 76,397% |

Didapatkan dari hasil perhitungan manual gaussian naïve bayes diatas, mengenai water potability. Didapatkan sebesar 76,397 % untuk potability tidak, sedangkan untuk potability ya sebesar 23,603%. Sehingga dari hasil tersebut dapat disimpulkan potability nya cendurung pada tidak.

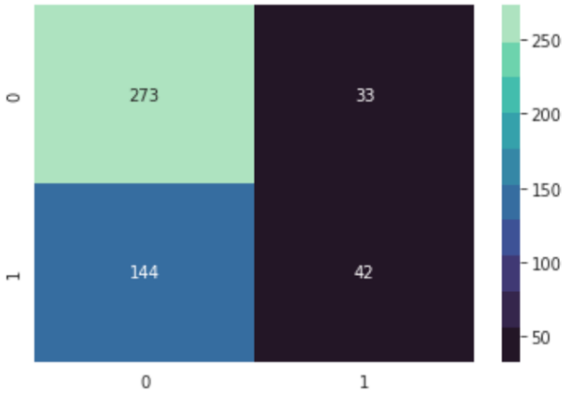
b. Hasil Perhitungan Kode Program

Dari perhitungan berdasarkan dataset diatas, didapatkan hasil sebagai berikut:

```
Naive Bayes      0.640
```

Didapatkan dari hasil perhitungan kode program gaussian naïve bayes diatas, mengenai water potability. Didapatkan sebesar 0,640 atau 64% untuk potability tidak. Sehingga dari hasil tersebut dapat disimpulkan potability nya cendurung pada tidak. Perbedaan sebesar 12% dari perhitungan manual gaussian naïve bayes.

c. Hasil Confusion Matrix




Berdasarkan model GNB didapatkan nilai confusion matrix sebesar TP = 273 FP = 33 FN = 144 TN = 42

2. Random Forest

a. Hasil Perhitungan Manual

Dari perhitungan berdasarkan dataset diatas, didapatkan hasil sebagai berikut:

| | |
|---|---|
| Weighted Gini Index | 0,71582923 |

| | |
|---|---|
| Weighted Gini Index | 0,712260855 |

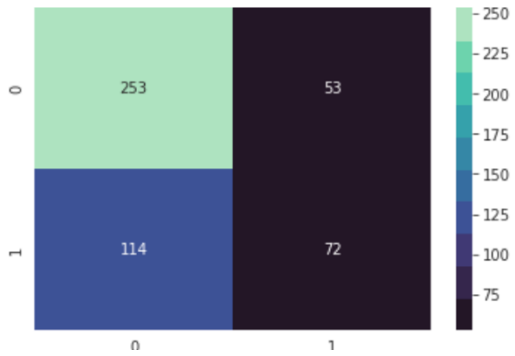
Dari salah satu tree pada random forest kami, Didapatkan hasil weighted gini index sebesar 0,712

b. Hasil Perhitungan Kode Program

```
Random Forest     0.661
```

Didapatkan dari hasil perhitungan kode program random forest diatas, mengenai water potability. Didapatkan sebesar 0,661 atau 66,1% untuk potability tidak. Sehingga dari hasil tersebut dapat disimpulkan potability nya cendurung pada tidak.

c. Hasil Confusion Matrix




Berdasarkan model RF didapatkan nilai confusion matrix sebesar TP = 253 FP = 53 FN = 114 TN = 72

**KESIMPULAN**

Dengan menggunakan 2 metode. Gaussian Naïve Bayes serta Random Forest didapatkan hasil bahwa metode Random Forest lebih unggul dibandingkan metode Gaussian Naïve Bayes. Serta dari dataset tersebut dapat disimpulkan bahwa air tidak layak untuk dikonsumsi berdasarkan aspek aspek yang tersedia.