# Research Question

1. **Dalam konteks bisnis dan ilmiah, apakah memprediksi nilai kesadahan (Hardness) air tanah merupakan hal yang mendesak? Jelaskan urgensinya!**

   Kesadahan air (*water hardness*) disebabkan oleh tingginya konsentrasi ion alkali tanah di dalam air, terutama Ca2+ dan Mg2+; maka, definisi parameter ini didasarkan pada konsentrasi gabungan kedua ion tersebut (Orellana, Darder, & Quílez-Alburquerque, 2023). Konsentrasi tinggi ion-ion ini dalam air dapat menjadi sumber masalah kesehatan dan juga bertanggung jawab atas pembentukan kerak kapur pada pipa industri atau rumah tangga, tangki air, dan berbagai mesin. Untuk mencegah dampaknya, kesadahan air ditentukan dengan pewarna indikator kolorimetri atau protokol laboratorium terstandar berdasarkan AAS atau ICP-MS (Dalmieda & Kruse, 2019).

   Dalam sebagian besar penelitian berskala besar, misalnya Anderson, dkk. (1975), Masironi, dkk. (1979), dan Leoni, dkk. (1985), telah dilaporkan adanya hubungan terbalik antara kesadahan air minum dan penyakit kardiovaskular. Menurut Nurullita, Astuti dan Arifin (2010), kadar kesadahan di atas 500 mg/L merupakan air sangat keras. Kadar kesadahan di atas 500 mg/L apabila dikonsumsi secara terus menerus akan menyebabkan kerusakan pada ginjal dalam kurun waktu jangka panjang. Yang (1998) menunjukkan hubungan statistik negatif dari berbagai jenis morbiditas/mortalitas kanker dengan kesadahan air dan kalsium. Hasil penelitian oleh Yang, dkk. (1997) menunjukkan bahwa terdapat pengaruh yang signifikan dari asupan magnesium dari air minum terhadap risiko penyakit serebrovaskular, tetapi hal ini masih menjadi perdebatan. Kesadahan adalah hal penting untuk air minum dari sudut pandang penerimaan dan pertimbangan operasional. Sengupta P. (2013). Oleh karena itu, konsumen perlu diberitahu tentang komposisi mineral dan kesadahan air ketika terjadi perbaikan atau hal lainnya oleh pemasok pipa atau manufaktur pemeliharaan pipa (Sengupta P. (2013).

   Menurut Nining Setyaningsih (2014), kesadahan merupakan salah satu parameter kimia yang menyebabkan berkurangnya kualitas air tanah. Kadar kesadahan yang tinggi pada air dapat menimbulkan masalah bagi rumah tangga. Kesadahan air merupakan faktor kunci keberhasilan proses mencuci. Konstadinos Abeliotis (2014) mempelajari dampak kesadahan air terhadap persepsi konsumen terhadap *laundry* di lima negara Eropa. Studi tersebut mendapatkan bahwa air sadah yang tinggi apabila digunakan untuk mencuci akan sulit berbusa sehingga akan menyebabkan pemborosan detergen. Air sadah menggumpalkan sabun membentuk *scum*, sehingga sabun sulit untuk berbusa. *Scum* dapat menjadikan noda pada pakaian sehingga pakaian menjadi kusam. Sabun bereaksi dengan kalsium dalam air sadah untuk membentuk dadih (*curds)* yang lengket: maka, jumlah sabun harus lebih banyak daripada yang dibutuhkan. Dadih sabun yang terbentuk dari reaksi yang disebutkan

sebelumnya menimbulkan masalah dalam pencucian karena kotoran menempel pada pakaian dan memerangkap tanah pada kain.
Air dengan kesadahan yang tinggi menjadikan kerak yang mengakibatkan penghantar panas ke air menjadi kurang, sehingga akan terjadi pemborosan penggunaan bahan bakar (Novitasari, 2022). Lerato Lethea (2017) telah mempelajari dampak kesadahan air terhadap konsumsi energi elemen pemanas geyser. Studi tersebut membuktikan bahwa kesadahan air yang tinggi meningkatkan konsumsi energi sekitar 4% hingga 12%. Selain itu kerak dapat menyumbat pipa saluran air panas, misalkan pada radiator. Oleh karena itu, disarankan menurunkan tingkat kesadahan air yang tinggi untuk memperlambat pembentukan kerak. Studi oleh Konstadinos Abeliotis (2014) juga menunjukan bahwa penggunaan air minum dengan kesadahan rendah di rumah tangga mempunyai beberapa dampak positif, seperti pengurangan konsumsi energi. Abdelghani, dkk menyimpulkan dalam studinya bahwa lebih baik menerapkan sistem penyaringan air (*softener*), terutama di daerah dengan kesadahan air yang tinggi. Hal tersebut diterapkan pada pasokan air gedung untuk mengurangi tagihan penggunaan energi, memperpanjang umur instalasi hidrolik, mengurangi frekuensi perawatan, membuat sabun dan deterjen lebih efisien, dan juga meningkatkan kualitas air minum.

Sheibani (2023) telah mempelajari pengaruh air sadah terhadap kualitas makanan dan efisiensi operasional. *Hard water*, yang mengandung mineral tinggi seperti kalsium dan magnesium, dapat merusak rasa dan tekstur makanan, mengurangi umur simpan produk, serta menyebabkan penumpukan kerak pada peralatan industri yang mengakibatkan peningkatan biaya perawatan dan konsumsi energi. Dengan memahami dan memprediksi kesadahan air, produsen makanan dapat mengambil langkah seperti penggunaan sistem penyaringan air (*softener*) dan perawatan peralatan yang tepat, sehingga memastikan kualitas produk tetap tinggi dan operasi berjalan lebih efisien.

Berdasarkan studi mengenai *sustainable development* dalam konteks penanganan air sadah oleh Klosok-Bazan & Witsanko-Sniezek (2017), air keras (*hard water*) memerlukan penggunaan deterjen dan bahan kimia pembersih lebih banyak untuk mencapai hasil yang sama, yang pada gilirannya meningkatkan emisi zat pencemar ke lingkungan. Penumpukan mineral akibat air keras (*hard water*) dapat merusak infrastruktur air dan sistem distribusi, meningkatkan biaya pemeliharaan dan perbaikan. Selain itu, penggunaan air keras dalam industri dapat mengurangi efisiensi proses produksi dan meningkatkan konsumsi energi. Penelitian menunjukkan bahwa pengurangan kesadahan air dapat secara signifikan menurunkan emisi karbon dioksida dan polutan lainnya, serta mengurangi penggunaan surfaktan dalam produk kebersihan. Oleh karena itu, prediksi yang akurat dan manajemen kesadahan air penting untuk mengurangi dampak negatif terhadap lingkungan, memastikan efisiensi energi, dan mendukung tujuan keberlanjutan dengan meningkatkan efisiensi penggunaan sumber daya alam.

Memprediksi nilai kesadahan air tanah merupakan hal yang mendesak dalam konteks bisnis dan ilmiah. Air sadah (*hard water*) memiliki dampak negatif yang luas. Dalam perspektif bisnis, air keras dapat mengurangi efisiensi operasional, meningkatkan biaya pemeliharaan dan perbaikan, serta merusak kualitas produk makanan dan minuman. Dari perspektif ilmiah dan lingkungan, penggunaan *hard water* meningkatkan konsumsi energi dan emisi polutan, merusak infrastruktur, serta memiliki implikasi terhadap kesehatan manusia. Oleh karena itu, prediksi kesadahan air yang akurat penting untuk mengantisipasi dampak negatif ini, meningkatkan efisiensi dan keberlanjutan sumber data, serta memastikan kualitas air yang lebih baik bagi konsumen dan lingkungan.

2. **Apakah submisi Kaggle Anda mengalami *overfit* atau *underfit*? Jika iya, jelaskan mengapa submisi tersebut Anda anggap sebagai *overfit* atau *underfit* serta jelaskan bagaimana cara mengatasi hal tersebut! Jika tidak, jelaskan mengapa!**

   Submisi Kaggle kami mengalami *overfit* karena pada saat model digunakan untuk data train kami mendapatkan R2 sebesar 0.95 sedangkan pada saat model digunakan untuk memprediksi data test dan dilakukan submisi ke kaggle hanya didapatkan R2 sebesar 0.902. Penurunan R2 yang signifikan untuk model prediksi memperlihatkan bahwa model tersebut tidak dapat tergeneralisasi dengan baik pada data test atau model terlalu sesuai dengan data train. *Overfit* dapat diatasi dengan beberapa cara berikut

   1. Menggunakan lebih banyak data train
      Dengan menggunakan lebih banyak data, membuat model lebih sulit untuk mengidentifikasi pola yang tepat dan mengharuskannya menemukan solusi yang lebih fleksibel untuk mengakomodasi lebih banyak kondisi.

   2. Menyederhanakan model
      Mengurangi kompleksitas model dengan mengurangi jumlah fitur, menghapus fitur yang tidak relevan. Hal ini akan mencegah model menggunakan terlalu banyak bidang untuk menyimpan pola tertentu, sehingga membuat model lebih fleksibel

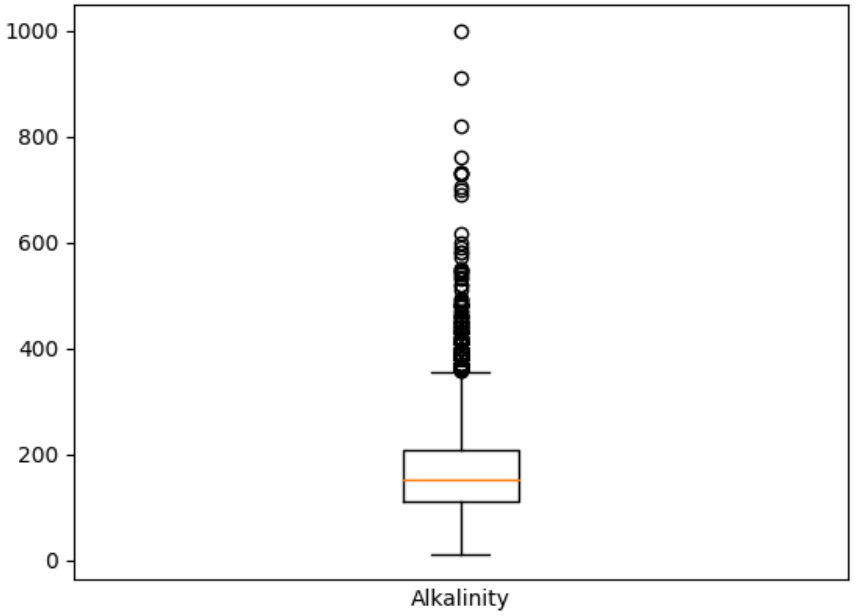
   3. Validasi silang (Cross-Validation)
      Manggunakan teknik validasi silang untuk mengevaluasi performa model pada beberapa subset data pelatihan. Hal ini membantu mendeteksi *overfitting* dan memberikan perkiraan performa model yang lebih kuat

   4. Augmentasi data
      Teknik ini memperluas kumpulan data train dengan membuat versi modifikasi dari data yang ada, menjadikan model lebih kuat dan mampu menangani berbagai kemungkinan data baru

3. **Jawablah pertanyaan-pertanyaan statistik berikut:**
   a. **Berapa Hardness rata-rata dari sumber air yang memiliki kadar sodium di atas persentil 75 dan memiliki tingkat kebasaan (Alkalinity) di atas rata-rata?**

      Persentil 75 dari sodium adalah 60 dan rata-rata alkalinity adalah 167.151046875. Sehingga ditemukan Hardness rata-rata dari sumber air yang memiliki kadar sodium di atas 60 dan tingkat alkalinity di atas 167.151046875 adalah **383.0833134765625**

   b. **Apakah ada sumber air yang memiliki tingkat kebasaan (Alkalinity) yang dapat dianggap outlier? Jelaskan!**




Terdapat beberapa metode yang dapat digunakan untuk mendeteksi outlier. Salah satu metodenya adalah metode IQR (*Interquartile Range*). Pendeteksian dilakukan dengan mencari kuartil pertama dan ketiga (Q1, Q3) lalu menghitung nilai IQR dengan mengurangi Q3 dan Q1. Nilai IQR yang telah didapatkan kemudian dikali dengan 1.5 untuk menemukan batas bawah (Q1 - 1.5*IQR) dan batas atas (Q3 + 1.5*IQR). Sehingga, tiap titik data yang berada di luar kedua batas tersebut dapat dianggap sebagai outlier.

Grafik Boxplot di atas menggunakan perhitungan yang sama dengan metode IQR. Berdasarkan grafik, terlihat bahwa terdapat beberapa titik data yang terletak di atas batas atas. Titik-titik data tersebut dapat dianggap sebagai outlier berdasarkan metode IQR. Jumlah outlier tersebut tepatnya adalah 174.

Jadi, berdasarkan metode IQR terdapat sumber air yang memiliki tingkat kebasaan (Alkalinity) yang dapat dianggap sebagai outlier.

4. **Apakah ada hubungan antara Specific Conductivity dan Hardness? Jelaskan!**

   Ya, terdapat hubungan antara Specific Conductivity dan Hardness. Berdasarkan perhitungan, didapatkan korelasi pearson adalah 0.64 dan korelasi spearman adalah 0.84.

   Nilai korelasi pearson menunjukkan bahwa Specific Conductivity dan Hardness memiliki hubungan linier positif. Hal ini berarti apabila nilai Specific Conductivity meningkat, maka nilai Hardness juga akan meningkat dan sebaliknya. Namun, hubungan linier ini tidak terlalu kuat sehingga ada kemungkinan terdapat variabilitas dalam data yang tidak dijelaskan oleh hubungan linier.

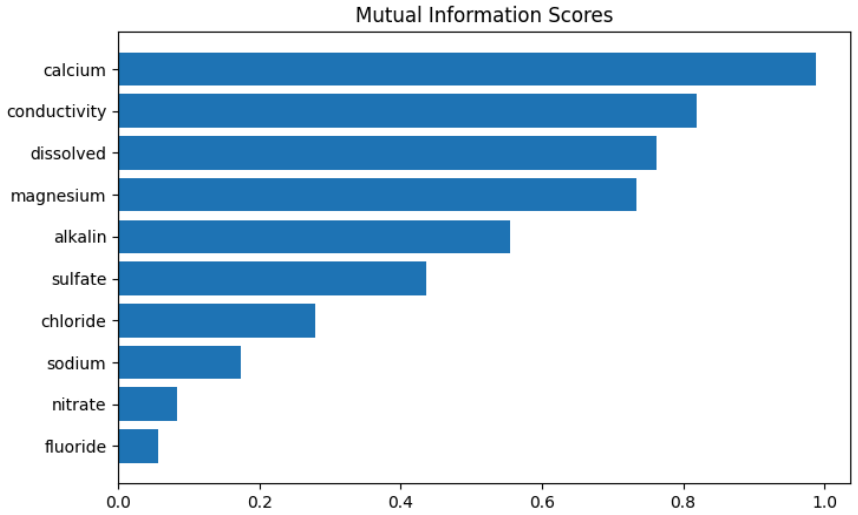
   Nilai korelasi spearman menunjukkan hubungan monoton positif yang kuat. Hal ini berarti peningkatan/penurunan salah satu variabel beriringan dengan peningkatan/penurunan variabel lainnya, meskipun hubungan tidak sepenuhnya linier.

   Berdasarkan kedua nilai korelasi, dapat disimpulkan bahwa terdapat hubungan positif yang cukup kuat antara Specific Conductivity dan Hardness.

5. **Dari zat-zat kimia yang diberikan, zat-zat apa saja yang paling mempengaruhi dan paling tidak mempengaruhi Hardness? Jelaskan!**

   <u>Sebelum penambahan fitur baru</u>

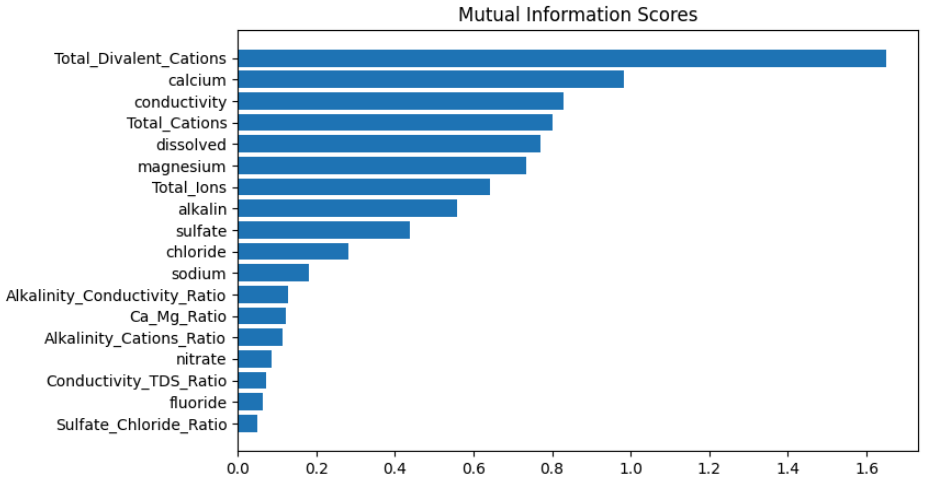
   Untuk melihat zat kimia apa saja yang mempengaruhi dan tidak mempengaruhi Hardness, kami menggunakan fungsi mutual informasi. Mutual Informasi menggambarkan hubungan dalam kaitannya dengan ketidakpastian. Mutual Informasi (MI) antara dua kuantitas adalah ukuran sejauh mana pengetahuan tentang satu kuantitas mengurangi ketidakpastian mengenai kuantitas lainnya. Semakin tinggi skor mutual informasi maka zat tersebut semakin berpengaruh signifikan terhadap Hardness. Didapatkan skor mutual informasi sebagai berikut



Dapat dilihat bahwa calcium memiliki mutual information score paling tinggi dan fluoride memiliki mutual information score paling rendah. Dengan demikian dapat disimpulkan bahwa berdasarkan mutual information scorenya calcium merupakan zat kimia yang paling mempengaruhi Hardness dan fluoride adalah zat kimia yang paling tidak mempengaruhi Hardness

Setelah penambahan fitur baru

Kelompok kami melakukan penambahan fitur baru yang disapatkan dari penggabungan dan pembagian beberapa fitur. Fitur yang kami tambahkan merupakan fitur yang menurut kami dapat membantu memprediksi Hardness antara lain Total_Cations (Calcium + Magnesium + Sodium), Ca_Mg_Ratio (Calcium / Magnesium), Total_Ions (Calcium + Magnesium + Sodium + Chloride + Sulfate + Nitrate + Fluoride), Sulfate_Chloride_Ratio (Sulfate / Chloride), Conductivity_TDS_Ratio (Conductivity / Total_Dissolved_Solids), Alkalinity_Conductivity_Ratio (Alkalinity / Conductivity), Total_Divalent_Cations (Calcium + Magnesium), Alkalinity_Cations_Ratio (Alkalinity / Total_Cations). Didapatkan skor mutual informasi sebagai berikut

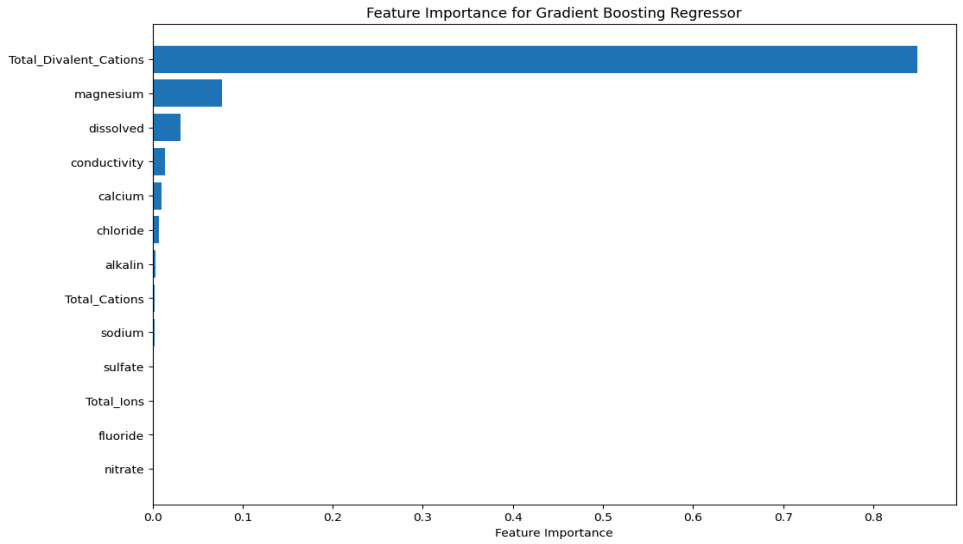



Dapat dilihat bahwa Total_Divalent_Cations (Calcium + Magnesium) memiliki mutual information score paling tinggi dan Sulfate_Chloride_Ratio (Sulfate / Chloride) memiliki mutual information score paling rendah. Dengan demikian dapat disimpulkan bahwa berdasarkan mutual information scorenya Total_Divalent_Cations (Calcium + Magnesium) merupakan zat kimia yang paling mempengaruhi Hardness dan Sulfate_Chloride_Ratio (Sulfate / Chloride) paling tidak mempengaruhi Hardness. Karena ternyata beberapa fitur baru memiliki mutual information score yang rendah, maka kami akan menghapuskan fitur tersebut.

Fitur yang mempengaruhi prediksi Hardness

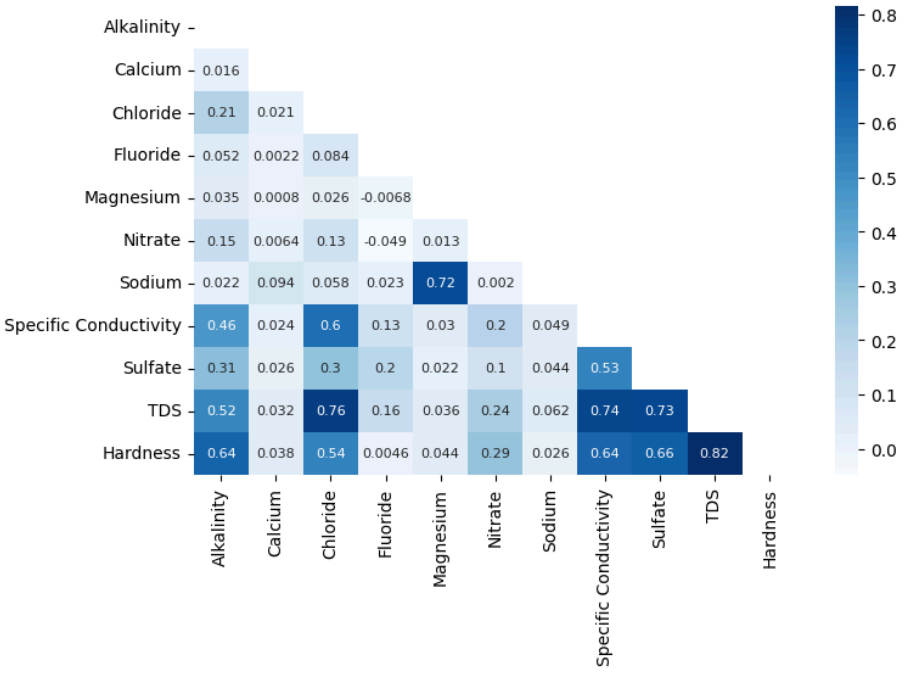
Karena model prediksi yang kami gunakan adalah gradient boosting, maka untuk menetukan fitur yang mempengaruhi prediksi Hardness akan digunakan fungsi feature importance. Feature importance dalam gradient boosting digunakan untuk mengukur

pengaruh setiap fitur dalam model prediktif yang dapat diukur berdasarkan seberapa sering dan seberapa signifikan fitur tersebut digunakan untuk membagi data dalam *decision tree*. Didapatkan feature importancenya sebagai berikut




Dari gambar di atas dapat dilihat bahwa Total_Divalent_Cations merupakan fitur yang paling mempengaruhi prediksi Hardness, hal ini sesuai dengan informasi yang didapatkan dengan mutual informasi bahwa Total_Divalent_Cations merupakan zat kimia yang paling mempengaruhi Hardness. Sedangkan untuk fitur yang paling tidak mempengaruhi prediksi Hardness adalah sulfate, Total_Ions, fluoride, dan nitrate yang keempatnya memiliki nilai 0, fluoride juga pada informasi sebelumnya dinyatakan sebagai zat kimia yang paling tidak mempengaruhi Hardness.

6. **Apakah ada pasangan zat kimia yang berkorelasi tinggi? Apakah ada efek dari korelasi tersebut?**



Berdasarkan grafik *correlation heatmap* di atas, terlihat bahwa TDS (Total Dissolved Solids) dan Chloride memiliki korelasi paling tinggi di antara pasangan zat kimia lainnya dengan nilai korelasi 0.76. Selain itu, pasangan zat kimia lain yang berkorelasi cukup tinggi adalah Sodium dan Magnesium, TDS dan Specific Conductivity, serta TDS dan Sulfate.

Korelasi yang tinggi antara beberapa pasangan zat kimia itu cukup masuk akal. Contohnya, TDS adalah jumlah total zat terlarut dalam air. Karena zat terlarut adalah konduktor, maka semakin tinggi TDS, semakin tinggi pula Specific Conductivity atau kemampuan air menghantarkan listrik. Lalu,

Untuk membuat model prediksi, pasangan zat kimia yang berkorelasi tinggi memiliki beberapa efek.

1. Multikolinearitas
   Korelasi tinggi antara variabel prediktor dapat menyebabkan multikolinearitas. Adanya multikolinearitas akan mempengaruhi akurasi model linier karena estimasi parameter regresi yang dihasilkan menjadi tidak efisien karena mempunyai bias dan variansi yang besar.
2. Overfitting
   Selain menyebabkan multikolinearitas, korelasi tinggi antar variabel prediktor juga dapat menyebabkan terjadinya overfitting. Model akan menjadi terlalu kompleks dan hanya mempelajari noise dalam data alih-alih pola yang sebenarnya. Hal ini akan mempengaruhi keakuratan dari hasil prediksi.

**7. Jelaskan fitur baru apa saja yang anda peroleh untuk membantu kemampuan model dalam memprediksi ketika anda melalui proses feature engineering.**

Kami mempertimbangkan beberapa fitur yang mungkin dapat membantu kemampuan model untuk memprediksi. Fitur-fitur tersebut adalah :

1. Total_Cations (Calcium + Magnesium + Sodium)
Jumlah konsentrasi kalsium, magnesium, dan sodium. Fitur ini dapat memberikan informasi tambahan mengenai total kation dalam air.

2. Ca_Mg_Ratio (Calcium / Magnesium)
Rasio antara kalsium dan magnesium untuk memberikan informasi tambahan mengenai dua kation divalen utama yang mempengaruhi kesadahan (*hardness*) air.

3. Total_Ions (Calcium + Magnesium + Sodium + Chloride + Sulfate + Nitrate + Fluoride)
Jumlah konsentrasi ion dari beberapa komponen untuk memberikan gambaran tentang kekayaan mineral dalam air.

4. Sulfate_Chloride_Ratio (Sulfate / Chloride)

Rasio antara sulfat dan klorida untuk memberikan informasi tambahan mengenai jenis garam yang ada dalam air, yang mungkin dapat mempengaruhi kesadahan secara tidak langsung.

5. Conductivity_TDS_Ratio (Conductivity / Total_Dissolved_Solids)
Rasio antara konduktivitas dan jumlah zat padat terlarut dalam air. Fitur ini dapat memberikan informasi mengenai karakteristik ionik air.

6. Alkalinity_Conductivity_Ratio (Alkalinity / Conductivity)
Rasio antara tingkat alkalinitas dan konduktivitas. Fitur ini dapat memberikan informasi mengenai hubungan antara alkalinitas air dan kemampuannya untuk menghantarkan listrik.

7. Total_Divalent_Cations (Calcium + Magnesium)
Jumlah kalsium dan magnesium, sebagai dua kation divalen utama yang mempengaruhi kesadahan (*hardness*) air.

8. Alkalinity_Cations_Ratio (Alkalinity / Total_Cations)
Rasio antara tingkat alkalinitas dan total kation. Fitur ini dapat memberikan informasi tambahan mengenai keseimbangan antara alkalinitas air dan total kation.

Setelah menggunakan fungsi mutual information, kami mendapati bahwa hanya 3 dari 8 fitur baru yang skor mutual informationnya cukup tinggi. Fitur-fitur itu adalah Total_Divalent_Cations, Total_Cations, dan Total_Ions. Ketiga fitur ini selanjutnya kami gunakan untuk membantu kemampuan model dalam memprediksi.

**8. Jelaskan model yang Anda gunakan dalam memprediksi label Hardness! Mengapa Anda menggunakan model tersebut?**

Kami memilih untuk menggunakan model Gradient Boosting Regressor untuk memprediksi tingkat kesadahan (hardness) air tanah. Model Gradient Boosting merupakan salah satu jenis algoritma *ensemble* yang membangun model prediksi dengan menggabungkan kekuatan dari banyak model prediksi yang lebih sederhana, yang disebut dengan "weak learners", biasanya berupa *decision trees* (Guillen, dkk. 2023).

Gradient Boosting memiliki kemampuan yang sangat baik dalam menangani *dataset* yang kompleks dan memiliki banyak fitur (Nadkari, dkk., 2023). Model ini mampu menangkap hubungan non-linear antara variabel-variabel prediktor dan variabel target, yang sangat penting mengingat data air tanah ini kemungkinan memiliki banyak interaksi kompleks antara komponen kimianya. Meskipun Gradient Boosting juga dapat dipengaruhi oleh *outlier*, pendekatan *ensemble*-nya dapat membantu dalam mengurangi dampak dari *outlier* dibandingkan dengan model linear seperti regresi linear biasa (Natekin & Knoll, 2013). Hal ini sangat penting mengingat adanya outlier dalam data *train* dan data *test*.

Gradient Boosting cenderung lebih tahan terhadap multikolinearitas dibandingkan dengan model linear (Vaulet, dkk., 2022). Dengan adanya multikolinearitas dalam data variabel prediktor, model linear seringkali memberikan nilai koefisien yang tidak stabil dan interpretasi yang sulit, sementara Gradient Boosting fokus pada pengurangan kesalahan prediksi dan cenderung tidak terganggu oleh kolinearitas antar fitur. Gradient Boosting memiliki banyak *hyperparameter* yang bisa di-*tuning* untuk meningkatkan performa model (Yuhana, 2022), seperti *learning rate*, dsb. Dengan melakukan *tuning* yang tepat, model ini bisa disesuaikan dengan karakteristik spesifik dari *dataset* yang digunakan, sehingga menghasilkan prediksi yang lebih akurat. Gradient Boosting biasanya memberikan hasil yang konsisten dan memiliki kemampuan generalisasi yang baik pada data yang belum pernah dilihat sebelumnya (Singh, dkk., 2021). Hal ini sangat penting di mana data *test* merupakan data yang baru dan model harus mampu memberikan prediksi yang akurat pada data tersebut.

Dengan mempertimbangkan faktor-faktor tersebut, Gradient Boosting Regressor adalah pilihan yang tepat untuk memprediksi tingkat kesadahan air tanah dalam konteks *dataset* yang kompleks, memiliki *outlier*, dan multikolinearitas. Model ini tidak hanya memberikan kinerja yang kuat dan konsisten tetapi juga memungkinkan fleksibilitas dalam penyesuaian hyperparameter untuk mencapai hasil yang optimal.

**9. Menurut Anda, apakah metrik penilaian R2 tepat? Jika tidak, metrik penilaian apa yang menurut anda lebih tepat digunakan? Elaborasikan jawaban Anda!**

$R^2$ (Koefisien Determinasi)

R-squared, atau koefisien determinasi, adalah metrik yang umum digunakan untuk mengevaluasi seberapa baik model regresi linear dapat menjelaskan variabilitas data. Nilai R-squared berkisar antara 0 hingga 1, dengan nilai yang lebih tinggi menunjukkan model yang lebih baik dalam menjelaskan variabilitas data target (Yasmin, dkk 2022). Selain itu, R-squared mudah dipahami dan diinterpretasikan, menjadikannya metrik yang populer dalam banyak analisis regresi.

Namun, R-squared juga memiliki sejumlah kekurangan yang perlu dipertimbangkan. Salah satu kelemahan utama adalah ketidakmampuannya untuk menangani overfitting dengan baik. Nilai R-squared dapat meningkat dengan menambahkan lebih banyak variabel prediktor, bahkan jika variabel tersebut tidak relevan, yang pada akhirnya dapat menyebabkan model menjadi overfit (Sapra, 2014). Terakhir, R-squared tidak selalu informatif untuk hubungan non-linear. Dalam model non-linear, R-squared mungkin tidak mencerminkan kualitas model dengan baik, sehingga dapat menyesatkan dalam mengevaluasi performa model (Spiess & Neumeyer, 2010).

Metrik Penilaian Lain yang Dapat Digunakan

1. Mean Absolute Error (MAE)

MAE mengukur rata-rata kesalahan absolut antara nilai prediksi dan nilai sebenarnya. MAE lebih intuitif dan kurang sensitif terhadap outlier dibandingkan dengan MSE (Mean Squared Error) (Hodson, 2022 dan Anisha, dkk., 2022). Kelebihan utama dari MAE adalah kemudahannya dalam interpretasi dan ketidaksensitifannya terhadap outlier. Karena MAE memberikan nilai kesalahan dalam satuan yang sama dengan data aslinya (Jierula, 2021), ini membuatnya intuitif dan mudah dipahami. Namun, kelemahannya adalah MAE tidak mempertimbangkan arah kesalahan.

2. Mean Squared Error (MSE)
   MSE mengukur rata-rata kuadrat kesalahan antara nilai prediksi dan nilai sebenarnya. Semakin kecil nilainya, semakin performa model tersebut (Hodson, 2022). Kelebihan utama dari MSE adalah bahwa MSE memberikan pengaruh yang lebih besar untuk error yang lebih besar, sehingga lebih sensitif terhadap outlier dan dapat membantu mengidentifikasi model yang menghasilkan kesalahan besar. Namun, kelemahannya adalah MSE sangat sensitif terhadap outlier, sehingga beberapa kesalahan besar dapat mendominasi metrik ini dan memberikan gambaran yang tidak akurat tentang performa model secara keseluruhan.

3. Root Mean Squared Error (RMSE)
   RMSE adalah akar kuadrat dari MSE, memberikan metrik dalam satuan yang sama dengan variabel target (Hodson, 2022 dan Ding, dkk., 2023). Kelebihan RMSE termasuk kemudahan interpretasinya dalam konteks yang sama dengan target, serta kemampuannya memberikan penalti besar untuk kesalahan besar, mirip dengan MSE. Namun, seperti MSE, RMSE juga sangat sensitif terhadap outlier, yang berarti beberapa kesalahan besar dapat memberikan pengaruh yang tidak proporsional terhadap nilai RMSE.

4. Mean Absolute Percentage Error (MAPE)
   MAPE mengukur rata-rata kesalahan absolut sebagai persentase dari nilai sebenarnya. MAPE membantu memahami seberapa akurat prediksi dengan memperhatikan ukuran variabel target (Chavez, dkk. 2020). Kelebihan utama MAPE adalah kemampuannya memberikan ukuran kesalahan relatif, yang membantu memahami seberapa akurat prediksi dengan memperhatikan ukuran variabel target. Namun, kelemahan MAPE adalah bahwa metrik ini tidak didefinisikan untuk nilai sebenarnya yang mendekati nol, yang dapat menjadi masalah dalam beberapa konteks data.

Untuk memprediksi tingkat kesadahan air tanah (*water hardness*) dengan dataset yang memiliki outlier dan multikolinearitas, penting memilih metrik penilaian yang tepat. $R^2$, atau koefisien determinasi, adalah metrik umum dalam regresi yang mengukur seberapa baik variabel prediktor menjelaskan varians variabel target. Namun, $R^2$ memiliki kelemahan dalam konteks ini. Cenderung meningkat dengan penambahan

variabel prediktor, $R^2$ dapat menyebabkan overfitting terutama dalam kasus multikolinearitas.

Oleh karena itu, disarankan untuk menggunakan juga metrik alternatif lain. RMSE (Root Mean Squared Error) memberikan pengaruh hasil lebih besar untuk error besar dan hasil dalam satuan yang sama dengan variabel target, cocok untuk gambaran kualitas prediksi keseluruhan. MAE (Mean Absolute Error) lebih stabil dan kurang sensitif terhadap outlier dibandingkan $R^2$, memberikan ukuran kesalahan rata-rata yang mudah diinterpretasikan. MAPE (Mean Absolute Percentage Error) memberikan ukuran kesalahan relatif dalam bentuk persentase, berguna untuk memahami kesalahan dalam konteks proporsional terhadap nilai sebenarnya.

Dalam dataset ini, $R^2$ mungkin bukan metrik penilaian yang paling tepat. Metrik seperti RMSE, MAE, dan MAPE dapat memberikan gambaran lebih akurat tentang kinerja model. Memilih metrik penilaian yang tepat memastikan evaluasi model yang akurat dan pengambilan keputusan yang lebih baik berdasarkan data, sehingga model yang dihasilkan tidak hanya *fit* dengan data *train* tetapi juga memiliki kinerja baik pada data *test* dan data dunia nyata.

10. **Jika Anda boleh mengambil data dari sumber eksternal, data tentang apa yang Anda akan ambil? Jelaskan mengapa data tersebut dapat membantu Anda memprediksi Hardness!**

    Kesadahan air adalah kandungan mineral-mineral tertentu yang terdapat di dalam air, biasanya berupa ion kalsium (Ca) dan magnesium (Mg) dalam bentuk garam karbonat. Selain itu dapat juga disebabkan oleh keberadaan ion-ion lain dari logam bervalensi 2 seperti Fe (Besi), Mn (Mangan), dan Sr (Srotium) dengan bentuk garam sulfat, klorida dan bikarbonat dalam jumlah kecil. Air yang memiliki sifat sadah umumnya ditemukan pada wilayah dengan sumber air tanah/sumur yang pada daerah tersebut lapisan tanah yang mengandung deposit garam mineral, kapur, dan kalsium (Candra, 2007) Dengan mempertimbangkan hal-hal tersebut, jika diperbolehkan mengambil data dari sumber eksternal, maka kelompok kami akan menambahkan data berikut untuk membantu memprediksi *hardness*

    1. Kandungan karbondioksida ($CO_2$) dalam air
       Karbondioksida merupakan gas yang mudah terlalut ke dalam air. Ketika karbondioksida bereaksi dengan air, maka akan terbentuk asam karbonat ($H_2CO_3$) (Effendi, 2003). Asam karbonat bereaksi dengan batu kapur kalsium karbonat ($CaCO_3$) dan magnesium karbonat ($MgCO_3$) membentuk kalsium bikarbonat [$Ca(HCO_3)_2$] dan magnesium bikarbonat [$Mg(HCO_3)_2$] yang larut dalam air. Kedua senyawa inilah yang menyebabkan air menjadi sadah. Oleh karena itu, dengan mengetahui kandungan karbondioksida dalam air kita dapat memperkirakan kesadahan air terutama air tanah dari lapisan tanah yang mengandung banyak kapur. Air yang memiliki tingkat kesadahan tinggi biasanya memiliki kandungan karbondioksida terlarut atau bebas dalam air

lebih sedikit, karena pada air sadah, $CO_2$ bereaksi dengan kalsium karbonat dan magnesium karbonat sehingga menghasilkan ion $Ca^{2+}$ dan $Mg^{2+}$ serta ion bikarbonat (kalsium dan magnesium bikarbonat).

2. pH air
   Kesadahan air sangat dipengaruhi oleh keberadaan kalsium bikarbonat [$Ca(HCO_3)_2$] dan magnesium bikarbonat [$Mg(HCO_3)_2$] dalam air. Ion bikarbonat dan karbonat termasuk golongan basa, basa menyebabkan pH meningkat. Dengan demikian kita dapat memprediksi kesadahan air dengan mengetahui pH air tersebut. Semakin tinggi pH air, maka tingkat kesadahan air semakin tinggi, hal ini disebabkan air sadah mengandung banyak ion bikarbonat yang merupakan basa.

3. Kandungan besi, mangan, dan srotium
   Seperti yang telah disebutkan sebelumnya, kesadahan air juga dapat disebabkan keberadaan ion-ion lain dari logam bervalensi 2 seperti Fe (Besi), Mn (Mangan), dan Sr (Srotium). Oleh karena itu dengan mengetahui kandungan besi, mangan, dan srotium dalam air dapat membantu dalam memprediksi kesadahan air (*water hardness*).

**Kode Colab :** Seleksi Compfest 24.ipynb

## Daftar Pustaka


Abeliotis, K., Candan, C., Amberg, C., Ferri, A., Osset, M., Owens, J., & Stamminger, R. (2014). Impact of water hardness on consumers' perception of laundry washing result in five European countries. *International Journal of Consumer Studies, 39*(1), 60-66. https://doi.org/10.1111/ijcs.12149

Anderson, T. W., Neri, L. C., Schreiber, G. B., Talbot, F. D., & Zdrojewski, A. (1975). Letter: Ischemic heart disease, water hardness and myocardial magnesium. Canadian Medical Association Journal, 113, 199-203.

Anisha, C., & Arulanand, N. (2022). Tuned Homogenous Ensemble Regressor Model for Early Diagnosis of Parkinson Disorder Based on Voice Features Modality. September 2022. https://doi.org/10.36548/jaicn.2022.3.005

Aribiyanto, M. A. A. (2016). Pemetaan Tingkat Kesadahan Air Sumur di Wilayah Surabaya Barat Berbasis Aplikasi Sistem Informasi Geografis. [Skripsi]. Universitas Airlangga.

Chavez, G., Liu, Y., Ghysels, P., Li, X., & Rebrova, E. (2020). Scalable and Memory-Efficient Kernel Ridge Regression. 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 956-965. https://doi.org/10.1109/IPDPS47924.2020.00102

Dalmieda, J., & Kruse, P. (2019). Metal Cation Detection in Drinking Water. Sensors (Basel, Switzerland), 19(23), 5134. https://doi.org/10.3390/s19235134

Ding, R., Wang, Z., Jiang, L., & Zheng, S. (2023). Radar Target Localization with Multipath Exploitation in Dense Clutter Environments. Applied Sciences. https://doi.org/10.3390/app13042032

Guillen, M. D., Aparicio, J., & Esteve, M. (2023). Gradient tree boosting and the estimation of production frontiers. Expert Systems with Applications, 214, 119134. https://doi.org/10.1016/j.eswa.2022.119134

Hodson, T. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development. https://doi.org/10.5194/gmd-15-5481-2022

Jierula, A., Wang, S., OH, T.-M., & Wang, P. (2021). Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. Applied Sciences, 11(5), 2314. https://doi.org/10.3390/app11052314

Klosok-Bazan, Iwona; Witsanko-Sniezek, Aneta (2017) : Sustainable development in the context of hard water treatment, Economic and Environmental Studies (E&ES), ISSN 2081-8319, Opole University, Faculty of Economics, Opole, Vol. 17, Iss. 4, pp. 1135-1145, https://doi.org/10.25167/ees.2017.44.31

Leoni, V., Fabiani, L., & Ticchiarelli, L. (1985). Water hardness and cardiovascular mortality rate in Abruzzo, Italy. *Archives of Environmental Health, 40*, 274-278.

Lethea, L. (2017). Impact of water hardness on energy consumption of geyser heating elements. *Water SA, 43*(4). http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S1816-79502017000400009

Nadkarni, S. B., Vijay, G. S., & Kamath, R. C. (2023). Comparative Study of Random Forest and Gradient Boosting Algorithms to Predict Airfoil Self-Noise. Engineering Proceedings, 59, 24. https://doi.org/10.3390/engproc2023059024

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21. https://doi.org/10.3389/fnbot.2013.00021

Nurullita, U., Astuti, R., & Arifin, M. Z. (2017). Pengaruh lama kontak karbon aktif sebagai media filter terhadap persentase penurunan kesadahan CaCO3 air sumur artetis. Jurnal Kesehatan Masyarakat Indonesia, 6(1). https://doi.org/10.26714/jkmi.6.1.2010.%25p

Novitasari, G. I. (2022) GAMBARAN KESADAHAN AIR SUMUR GALI SEBELUM DAN SETELAH DIREBUS DI DUSUN PLEMBON LOR DESA LOGANDENG KECAMATAN PLAYEN KABUPATEN GUNUNGKIDUL TAHUN 2021. [Skripsi]. Poltekkes Kemenkes Yogyakarta.

Masironi, R., Pisa, Z., & Clayton, D. (1979). Myocardial infarction and water hardness in the WHO myocardial infarction registry network. Bulletin of the World Health Organization, 57, 291-299.

Orellana, G., Darder, M. M., & Quílez-Alburquerque, J. (2023). Luminescence-based sensors for water quality analysis. In R. Narayan (Ed.), Encyclopedia of sensors and biosensors (1st ed., pp. 599-613). Elsevier. https://doi.org/10.1016/B978-0-12-822548-6.00116-3

Sengupta P. (2013). Potential health impacts of hard water. International journal of preventive medicine, 4(8), 866–875.

Setyowati, D. (2018). *Pengaruh Waktu Perendaman Resin Saset Terhadap Penurunan Kesadahan Air Sumur Gali* (Doctoral dissertation, Poltekkes Kemenkes Yogyakarta).

Sheibani, E. (2023). The impact of hard water on food quality: A comprehensive analysis. *Journal of Food Technology and Preservation, 7*(4), 189.

Singh, U., Rizwan, M., Alaraj, M., & Alsaidan, I. (2021). A Machine Learning-Based Gradient Boosting Regression Approach for Wind Power Production Forecasting: A Step towards Smart Grid Environments. Energies, 14, 5196. https://doi.org/10.3390/en14165196

Spiess, A. N., & Neumeyer, N. (2010). An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC pharmacology, 10, 6. https://doi.org/10.1186/1471-2210-10-6

Sungkono, J., & Nugrahaningsih, T. K. (2017). Simulasi dampak multikolinearitas pada kondisi penyimpangan asumsi normalitas. *Magistra*, *29*(102), 46-50.

Vaulet, T., Al-Memar, M., Fourie, H., Bobdiwala, S., Saso, S., Pipi, M., Stalder, C., Bennett, P., Timmerman, D., Bourne, T., & De Moor, B. (2022). Gradient boosted trees with individual explanations: An alternative to logistic regression for viability prediction in the first trimester of pregnancy. Computer methods and programs in biomedicine, 213, 106520. https://doi.org/10.1016/j.cmpb.2021.106520

Widayat, W. (2002). Teknologi pengolahan air sadah. *Jurnal Teknologi Lingkungan*, *3*(3), 256-266.

Yang, C. Y. (1998). Calcium and magnesium in drinking water and risk of death from cerebrovascular disease. *Stroke, 29*(2), 411-414.

Yasmin Makki Mohialden, Nadia Mahmood Hussien, Shatha J. Mohammed, & Itidal Saad Mohammed. (2022). Recent hybrid machine learning algorithm applications: a review. Journal of Information Technology and Informatics, 2(1), 1–4. https://www.qabasjournals.com/index.php/jiti/article/view/128

Yuhana, U. L., Purwarianti, A., & Imamah. (2022). Tuning Hyperparameter pada Gradient Boosting untuk Klasifikasi Soal Cerita Otomatis. Jurnal Edukasi dan Penelitian Informatika, 8(1), 134. http://dx.doi.org/10.26418/jp.v8i1.50506