KLASIFIKASI pH TANAH BERDASARKAN KARAKTERISTIK KIMIA DENGAN MACHINE LEARNING

Dhifya Fetryan, Teddy Agustinus, Verlene Angelica
Fakultas Teknologi Informasi, Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410

Abstrak

pH tanah merupakan indikator fundamental yang mempengaruhi ketersediaan unsur hara, aktivitas mikroorganisme, serta produktivitas pertanian. Metode analisis laboratorium yang umum digunakan dinilai akurat, namun memiliki keterbatasan dari sisi waktu, biaya, dan aksesibilitas. Penelitian ini mengusulkan pendekatan klasifikasi pH tanah berbasis karakteristik kimia menggunakan algoritma machine learning sebagai alternatif metode analisis yang lebih efisien dan aplikatif. Dataset yang digunakan terdiri atas 3.022 data latih dan 1.007 data uji dengan 16–17 atribut kimia dan fisik tanah. Tahap pra-pemrosesan meliputi imputasi nilai kosong, normalisasi, transformasi variabel kategorik, serta penyeimbangan distribusi kelas dengan Synthetic Minority Oversampling Technique (SMOTE). Model yang dibangun memanfaatkan algoritma Random Forest dan XGBoost, dengan evaluasi kinerja menggunakan metrik akurasi, presisi, recall, F1-score, ROC AUC, serta analisis interpretabilitas fitur melalui SHAP.

Hasil penelitian menunjukkan bahwa faktor kimia tanah, khususnya kalsium (Ca), fosfor (P), magnesium (Mg), kapasitas tukar kation (CEC), dan Exchangeable Sodium Percentage (ESP), merupakan determinan utama variasi pH. Random Forest memperoleh kinerja lebih stabil dengan nilai F1-macro 0,42, sedangkan XGBoost menunjukkan keunggulan pada efisiensi komputasi. Penerapan SMOTE terbukti meningkatkan generalisasi model, khususnya dalam pengenalan kelas minoritas. Analisis SHAP menegaskan bahwa prediksi pH tanah dipengaruhi oleh interaksi kompleks antar variabel, di mana Ca, P, dan Mg memiliki kontribusi paling dominan.

Secara keseluruhan, penelitian ini menegaskan peran machine learning sebagai pendekatan komputasi yang mampu menghasilkan sistem klasifikasi pH tanah yang akurat, efisien, dan representatif. Kontribusi utama penelitian ini terletak pada integrasi data kimia tanah ke dalam model klasifikasi berbasis machine learning, yang jarang mendapat perhatian dalam penelitian terdahulu, serta implikasinya terhadap pengembangan praktik pertanian presisi dan pengelolaan lahan berkelanjutan.

***Kata kunci**:* pH tanah, karakteristik kimia, *machine learning*, Random Forest, XGBoost, SMOTE, SHAP.

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Tanah merupakan komponen penting dalam ekosistem yang berfungsi sebagai media tumbuh bagi tanaman serta penyedia unsur hara yang dibutuhkan dalam proses pertumbuhan [1]. Salah satu indikator fundamental yang menentukan kualitas tanah adalah tingkat keasaman atau kebasaan yang dinyatakan melalui nilai pH [2]. Nilai pH tanah memengaruhi ketersediaan unsur hara, aktivitas mikroorganisme, serta proses biogeokimia yang berperan langsung terhadap produktivitas pertanian [2]. Tanah dengan pH rendah cenderung mengalami pelindian unsur hara esensial seperti kalsium, magnesium, dan kalium [1], sedangkan tanah dengan pH tinggi dapat menghambat penyerapan unsur mikro seperti fosfor, besi, dan mangan [7].

Pengukuran pH tanah umumnya dilakukan melalui analisis laboratorium yang dianggap sebagai metode paling akurat [3]. Namun, metode ini memiliki keterbatasan signifikan dari sisi waktu, biaya, dan keterjangkauan, terutama pada wilayah dengan keterbatasan fasilitas laboratorium [4]. Kondisi tersebut menghambat ketersediaan informasi pH tanah secara cepat dan merata, padahal data tersebut sangat penting dalam pengambilan keputusan terkait pengelolaan lahan, pemupukan, dan peningkatan produktivitas pertanian [4].

Seiring dengan perkembangan teknologi digital, muncul pendekatan baru berbasis kecerdasan buatan (*Artificial Intelligence*) yang dapat dimanfaatkan untuk mengatasi keterbatasan metode konvensional [5]. Pendekatan ini memungkinkan analisis data tanah dilakukan secara lebih efisien dengan memanfaatkan algoritma *machine learning* [6]. Beberapa penelitian menunjukkan bahwa algoritma *machine learning* mampu melakukan klasifikasi pH tanah secara cepat dengan tingkat akurasi yang tinggi, sehingga dapat digunakan sebagai dasar dalam penerapan pertanian presisi [8].

Selain faktor lingkungan, karakteristik kimia tanah seperti kandungan karbon organik, nitrogen, fosfor, dan unsur hara lainnya memiliki peran penting dalam memengaruhi variasi pH tanah [9]. Sayangnya, sebagian besar penelitian

terdahulu lebih menekankan pada faktor spasial atau variabel lingkungan, sementara integrasi variabel kimia tanah ke dalam model prediksi berbasis machine learning masih relatif terbatas [7]. Padahal, pemanfaatan data kimia tanah secara komprehensif berpotensi menghasilkan model klasifikasi yang lebih akurat dan representatif [8].

Dengan demikian, penelitian mengenai klasifikasi pH tanah berbasis karakteristik kimia menggunakan algoritma *machine learning* perlu dilakukan untuk memberikan alternatif metode analisis yang lebih cepat, murah, dan efisien dibandingkan pengujian laboratorium [10]. Pendekatan ini diharapkan tidak hanya meningkatkan akurasi prediksi, tetapi juga mendukung praktik pertanian presisi dan pengelolaan lahan berkelanjutan dalam menghadapi tantangan ketahanan pangan di masa depan [9].

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, dapat dipahami bahwa pH tanah merupakan indikator penting dalam menentukan kualitas dan produktivitas lahan pertanian. Namun, metode pengukuran pH tanah yang lazim digunakan saat ini masih bergantung pada analisis laboratorium, yang meskipun akurat, membutuhkan biaya tinggi, waktu yang lama, dan sulit diakses di banyak daerah. Seiring perkembangan teknologi, metode berbasis machine learning mulai dimanfaatkan untuk melakukan klasifikasi pH tanah secara lebih efisien. Akan tetapi, sebagian besar penelitian terdahulu masih menekankan faktor spasial dan lingkungan, sedangkan karakteristik kimia tanah yang secara langsung memengaruhi variasi pH belum sepenuhnya diintegrasikan ke dalam model prediksi.

Dengan demikian, rumusan masalah dalam penelitian ini adalah bagaimana mengembangkan model klasifikasi pH tanah dengan memanfaatkan variabel kimia tanah melalui algoritma *machine learning*, khususnya *Random Forest* dan *XGBoost*, agar diperoleh hasil prediksi yang lebih akurat, efisien, dan aplikatif dalam mendukung praktik pertanian presisi serta pengelolaan lahan berkelanjutan.

1.3 Tujuan Penelitian

Penelitian ini dilaksanakan untuk menjawab permasalahan yang telah diidentifikasi dalam latar belakang dan rumusan masalah. Secara umum, penelitian ini bertujuan untuk mengembangkan metode klasifikasi pH tanah berbasis karakteristik kimia dengan memanfaatkan algoritma machine learning. Adapun tujuan khusus dari penelitian ini adalah sebagai berikut:

1. Menganalisis keterkaitan antara karakteristik kimia tanah dan variasi pH tanah
2. Membangun dan mengimplementasikan model klasifikasi pH tanah berbasis machine learning
3. Mengevaluasi kinerja model klasifikasi yang dibangun dengan menggunakan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score*.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan kontribusi, baik dalam bidang akademik maupun praktis, khususnya dalam pemanfaatan teknologi digital untuk mendukung pengelolaan lahan pertanian. Manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut:

1. Kontribusi ilmiah berupa pengembangan metode klasifikasi pH tanah berbasis machine learning dengan fokus pada integrasi karakteristik kimia tanah
2. Kontribusi praktis berupa alternatif metode klasifikasi pH tanah yang lebih cepat, efisien, dan murah dibandingkan pengujian laboratorium.
3. Kontribusi aplikatif dalam mendukung implementasi pertanian presisi dan pengelolaan lahan berkelanjutan.

BAB 2

PEMBAHASAN

2.1 Landasan Teori

Tanah dan pH Tanah

Tanah merupakan salah satu komponen utama ekosistem darat yang berfungsi sebagai media tumbuh bagi tanaman, sekaligus penyedia unsur hara dan air. Salah satu indikator penting dalam menilai kualitas tanah adalah tingkat keasaman atau kebasaan yang dikenal dengan pH tanah. Nilai pH diukur pada skala 0 hingga 14, dengan kategori masam (pH < 6), netral (pH 6–7), dan basa (pH > 7). Kondisi pH tanah berpengaruh langsung terhadap ketersediaan unsur hara esensial, aktivitas mikroorganisme, dan efektivitas proses biogeokimia di dalam tanah. Tanah yang terlalu masam cenderung mengalami pelindian unsur hara seperti kalsium, magnesium, dan kalium, sedangkan tanah yang terlalu basa dapat menurunkan ketersediaan fosfor dan unsur mikro seperti besi serta mangan. Oleh karena itu, informasi mengenai pH tanah sangat diperlukan dalam pengelolaan lahan pertanian. Selain faktor lingkungan, karakteristik kimia tanah seperti kandungan karbon organik, nitrogen, fosfor, dan kalium juga menjadi penentu penting dalam variasi pH tanah, sehingga integrasi faktor-faktor ini dapat meningkatkan akurasi klasifikasi pH.

Klasifikasi dalam *Machine Learning*

Machine Learning (ML), yang merupakan bagian dari *Artificial Intelligence* (AI), memiliki potensi yang semakin besar untuk mengembangkan kemampuan dan efisiensi *Additive Manufacturing* (AM). Dengan kemampuannya mengekstrak pola, belajar dari data, dan membangun prediksi yang efektif, ML dapat mengatasi berbagai tantangan serta mengoptimalkan berbagai aspek dalam proses AM [11].Dalam konteks penelitian ini, klasifikasi digunakan untuk memetakan data masukan berupa variabel kimia tanah ke dalam kelas tertentu, seperti pH rendah, netral, atau tinggi. Klasifikasi termasuk ke dalam metode *supervised learning*, di mana model dilatih menggunakan data yang sudah memiliki label sehingga dapat melakukan prediksi terhadap data baru. Pemanfaatan *machine learning* dalam bidang pertanian memberikan keunggulan

dalam hal kecepatan, efisiensi, dan kemampuan memproses data berukuran besar dengan variabel yang kompleks.

Algoritma *Random Forest*

Random Forest adalah sebuah metode ensemble learning yang menggabungkan banyak decision tree untuk membuat prediksi. Algoritma ini dikenal karena ketangguhan, akurasi, serta kemampuannya menangani data berdimensi tinggi dengan hubungan yang kompleks [12]. Keunggulan utama *Random Forest* adalah kemampuannya dalam menangani data yang bersifat non-linear, tahan terhadap *overfitting*, serta memberikan hasil yang stabil meskipun terdapat data yang bising. Selain itu, algoritma ini mampu mengestimasi tingkat kepentingan masing-masing variabel, sehingga dapat memberikan wawasan tambahan mengenai faktor-faktor kimia tanah yang paling berpengaruh terhadap klasifikasi pH. Namun demikian, kelemahan dari *Random Forest* adalah waktu komputasi yang relatif lebih tinggi pada dataset yang sangat besar, serta interpretasi model yang tidak selalu mudah dilakukan.

Algoritma *Extreme Gradient Boosting* (XGBoost)

Extreme Gradient Boosting (XGBoost) sebuah algoritma machine learning berbasis gradient-boosted decision trees yang dirancang untuk tugas pembelajaran terawasi, khususnya klasifikasi dan regresi. Algoritma ini bekerja dengan cara membangun pohon keputusan secara berurutan, di mana setiap pohon baru memperbaiki kesalahan dari pohon sebelumnya. XGBoost berfokus pada minimisasi fungsi kerugian, misalnya mean squared error untuk regresi atau cross-entropy loss untuk klasifikasi [13]. Keunggulan lain *XGBoost* adalah kemampuannya dalam menangani data berukuran besar dan kompleks dengan akurasi prediksi yang tinggi. Meski demikian, algoritma ini memerlukan proses penalaan (*hyperparameter tuning*) yang lebih kompleks dibandingkan *Random Forest*, sehingga membutuhkan keahlian dalam pemilihan parameter agar dapat mencapai hasil optimal.

Evaluasi Model *Machine Learning*

Dalam penelitian berbasis klasifikasi, evaluasi model merupakan tahapan penting untuk menilai kinerja algoritma yang digunakan. Beberapa metrik evaluasi yang umum digunakan antara lain *accuracy*, *precision*, *recall*, dan *F1-score*. *Accuracy* mengukur tingkat ketepatan prediksi keseluruhan, sementara *precision* dan *recall* lebih berfokus pada kualitas prediksi untuk masing-masing kelas. *F1-score* digunakan sebagai ukuran keseimbangan antara *precision* dan *recall*. Selain itu, *confusion matrix* juga digunakan untuk menggambarkan distribusi hasil prediksi ke dalam kategori benar dan salah. Dengan menggunakan evaluasi yang komprehensif, dapat ditentukan apakah model *Random Forest* maupun *XGBoost* telah mampu memberikan hasil klasifikasi pH tanah yang akurat dan andal.

2.2 Metode Penelitian

1. Dataset

   Dataset yang digunakan dalam penelitian ini merupakan data dalam format comma-separated values (CSV). Dataset tersebut terdiri atas dua berkas utama, yaitu *train.csv* dan *test.csv*. Berkas *train.csv* memiliki total 3022 baris dengan 17 atribut, sedangkan *test.csv* memiliki 1007 baris dengan 16 atribut. Atribut-atribut tersebut mencakup kombinasi variabel numerik dan kategorikal yang relevan untuk proses klasifikasi. Dengan adanya kombinasi variabel tersebut, dataset ini cukup kompleks dan menantang untuk diolah sehingga diperlukan pendekatan yang sistematis pada tahapan selanjutnya.

   Berdasarkan hasil eksplorasi awal, dataset tidak mengandung data duplikat, sehingga integritas data dari sisi keunikan relatif terjaga. Namun demikian, ditemukan adanya nilai kosong pada beberapa atribut, yaitu sebanyak 250 pada berkas train.csv dan 103 pada berkas test.csv. Kondisi ini menunjukkan bahwa dataset belum sepenuhnya bersih dan memerlukan pra pemrosesan sebelum dapat digunakan untuk pelatihan model. Penanganan nilai kosong menjadi penting karena dapat mempengaruhi hasil akhir prediksi apabila diabaikan. Oleh sebab itu, sebelum model

dibangun, dataset ini terlebih dahulu melalui tahapan pra pemrosesan yang mendetail.

Selain itu, distribusi data pada atribut-atribut tertentu juga perlu diperiksa untuk memastikan tidak terjadi ketidakseimbangan yang signifikan antar kelas. Ketidakseimbangan data dapat menyebabkan model cenderung bias terhadap kelas mayoritas dan mengabaikan kelas minoritas. Dengan demikian, meskipun dataset ini relatif representatif, masih diperlukan serangkaian tahapan persiapan untuk menjamin kualitas data sebelum dimanfaatkan lebih lanjut. Secara umum, dataset ini dapat menjadi landasan yang baik untuk pengembangan model klasifikasi, namun tetap membutuhkan penyesuaian agar sesuai dengan kebutuhan penelitian.

2. Prapemrosesan data

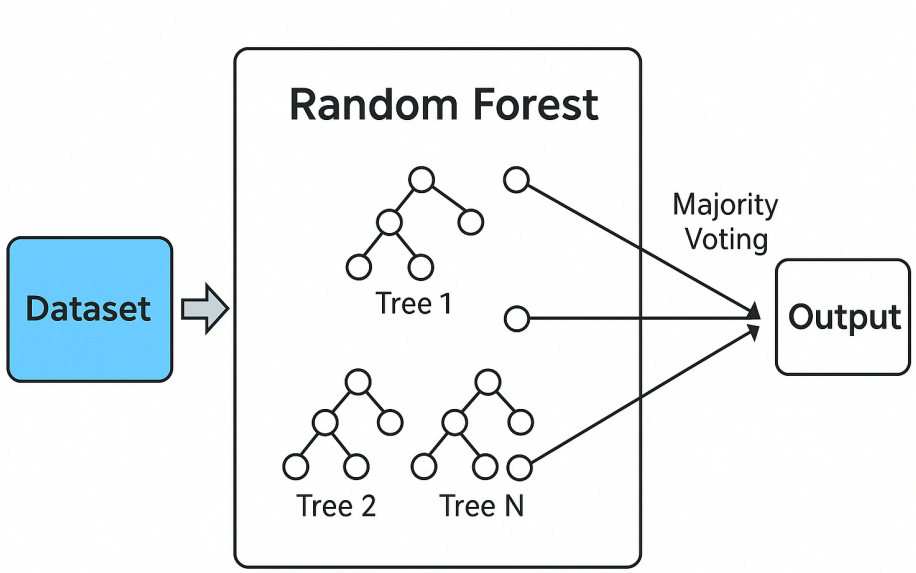
Pra pemrosesan data merupakan tahap krusial yang bertujuan untuk meningkatkan kualitas dataset sehingga siap digunakan dalam pembangunan model klasifikasi. Tahapan pertama adalah penanganan nilai kosong pada beberapa atribut. Untuk variabel numerik, nilai kosong dapat diisi menggunakan rata-rata atau median, sedangkan pada variabel kategorikal dapat diisi dengan modus atau kategori yang paling sering muncul. Pemilihan metode imputasi dilakukan dengan mempertimbangkan karakteristik distribusi data agar tidak menimbulkan bias. Dengan demikian, proses imputasi dapat menjaga konsistensi serta validitas informasi yang terkandung dalam dataset.

Tahapan berikutnya adalah normalisasi terhadap variabel numerik. Normalisasi bertujuan untuk memastikan semua atribut numerik berada dalam skala yang seragam, sehingga tidak ada variabel yang mendominasi proses pembelajaran hanya karena perbedaan skala. Normalisasi ini sangat membantu algoritma pembelajaran mesin dalam mempercepat proses konvergensi sekaligus meningkatkan stabilitas model. Selain itu, untuk variabel kategorik, dilakukan transformasi ke dalam bentuk numerik agar dapat diproses oleh algoritma. Proses ini dilakukan dengan metode one-hot

encoding untuk variabel nominal dan label encoding untuk variabel ordinal, sehingga informasi yang dimiliki oleh variabel tetap terjaga.

Langkah terakhir dalam pra pemrosesan adalah pembagian data menjadi data latih dan data uji internal dengan proporsi tertentu, misalnya 80% data digunakan untuk pelatihan dan 20% sisanya digunakan untuk pengujian. Pembagian ini bertujuan untuk menguji kemampuan generalisasi model sebelum dilakukan evaluasi menggunakan dataset test.csv. Dengan pembagian ini, performa model dapat dinilai secara lebih objektif dan risiko overfitting dapat diminimalisasi. Secara keseluruhan, pra pemrosesan data menjadi pondasi penting dalam penelitian ini karena kualitas model sangat bergantung pada kualitas data yang digunakan.

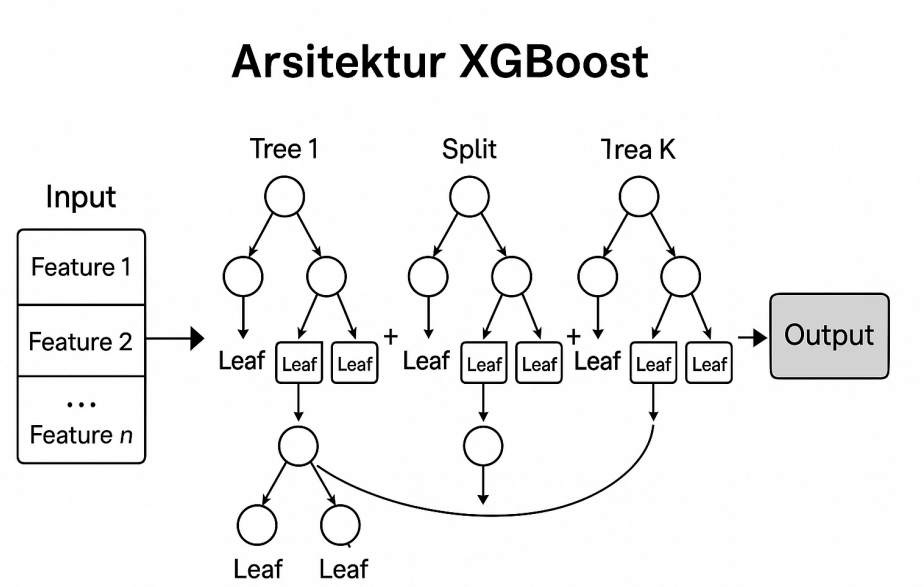
3. Pembangunan Model Klasifikasi



Gambar 1. Arsitektur *Random Forest*

Gambar 1 Tahapan pembangunan model klasifikasi dilakukan setelah dataset dipastikan bersih melalui proses pra pemrosesan. Penelitian ini memanfaatkan dua algoritma pembelajaran mesin yang populer dan memiliki kinerja tinggi, yaitu *Random Forest* dan *XGBoost*. *Random Forest* merupakan algoritma berbasis ansambel yang menggabungkan sejumlah pohon keputusan untuk menghasilkan prediksi yang lebih stabil.

Algoritma ini memiliki keunggulan dalam menangani data berdimensi tinggi, relatif tahan terhadap *overfitting*, serta mampu memberikan estimasi pentingnya setiap fitur. Hal ini menjadikan *Random Forest* sebagai pilihan yang tepat dalam penelitian klasifikasi berbasis data kompleks.



Gambar 2. Arsitektur *XGBoost*

Gambar 2 menjelaskan bahwa XGBoost merupakan algoritma berbasis gradient boosting yang bekerja dengan cara membangun pohon keputusan secara bertahap. Setiap iterasi yang dilakukan oleh XGBoost bertujuan untuk memperbaiki kesalahan dari iterasi sebelumnya, sehingga model yang dihasilkan lebih akurat. XGBoost juga dikenal unggul dari segi efisiensi komputasi karena mendukung proses paralelisasi serta memiliki mekanisme regularisasi untuk mencegah overfitting. Dengan keunggulan-keunggulan tersebut, XGBoost sering digunakan dalam berbagai kompetisi pembelajaran mesin dan terbukti menghasilkan performa yang kompetitif.

Proses pelatihan model dilakukan dengan data latih yang sudah melalui pra pemrosesan, serta dilengkapi dengan tahap hyperparameter tuning untuk mendapatkan konfigurasi parameter terbaik. Pemilihan parameter optimal ini sangat penting karena dapat mempengaruhi

performa model secara signifikan. Dengan mengkombinasikan Random Forest dan XGBoost, penelitian ini bertujuan untuk membandingkan keunggulan masing-masing algoritma sekaligus memilih model yang memiliki performa terbaik. Dengan pendekatan ini, diharapkan hasil klasifikasi yang diperoleh tidak hanya akurat, tetapi juga mampu diterapkan secara praktis pada data yang baru.

4. Evaluasi Model

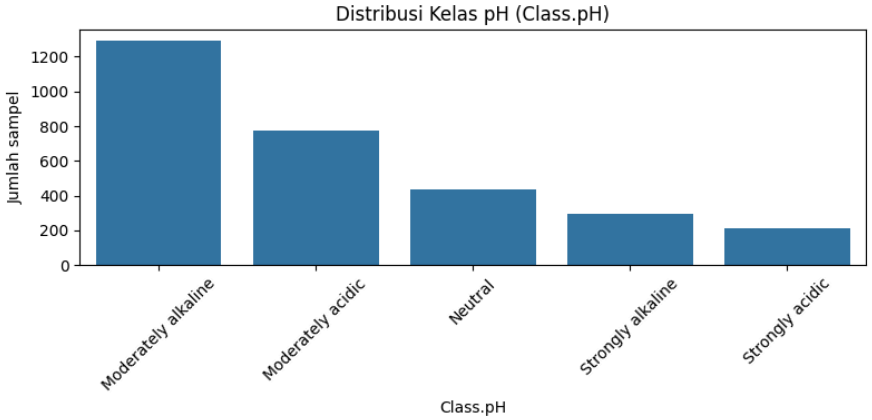
Evaluasi model merupakan tahap akhir yang berfungsi untuk menilai kinerja algoritma yang telah dibangun dalam melakukan klasifikasi. Pada penelitian ini, evaluasi dilakukan dengan menggunakan berbagai metrik, antara lain akurasi, presisi, recall, F1-score, dan ROC AUC. Akurasi digunakan untuk mengukur proporsi prediksi benar terhadap keseluruhan data, namun metrik ini saja sering kali tidak cukup, terutama jika terdapat ketidakseimbangan kelas. Oleh karena itu, digunakan juga presisi untuk mengukur ketepatan prediksi positif, recall untuk menilai kemampuan model menemukan semua data positif, serta F1-score yang memberikan keseimbangan antara presisi dan recall. Sementara itu, ROC AUC digunakan untuk menilai kemampuan model dalam membedakan antara kelas positif dan negatif secara menyeluruh.

Proses evaluasi dilakukan dengan membandingkan hasil prediksi model terhadap data uji internal yang telah dipisahkan sebelumnya. Dengan cara ini, performa model dapat dinilai secara objektif pada data yang tidak pernah digunakan dalam proses pelatihan. Evaluasi ini juga memungkinkan peneliti untuk mengetahui apakah model mengalami overfitting atau justru memiliki kemampuan generalisasi yang baik. Selain itu, hasil evaluasi dari algoritma Random Forest dan XGBoost dibandingkan secara langsung untuk menentukan model dengan performa terbaik.

Hasil dari evaluasi ini diharapkan dapat memberikan gambaran yang jelas mengenai efektivitas masing-masing algoritma. Jika terdapat perbedaan signifikan antara kedua model, analisis lebih lanjut dilakukan

untuk mengetahui penyebabnya, misalnya distribusi data, kompleksitas model, atau pemilihan parameter. Dengan demikian, tahap evaluasi tidak hanya berfungsi sebagai proses pengukuran performa, tetapi juga sebagai bahan refleksi untuk memahami bagaimana model dapat ditingkatkan. Kesimpulan akhir dari tahap evaluasi ini menjadi dasar pemilihan model terbaik yang akan digunakan dalam penelitian.

2.3 Hasil dan Pembahasan

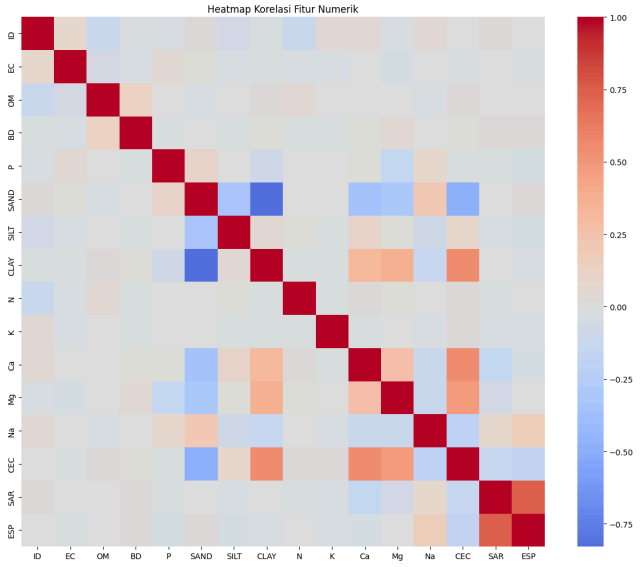


Gambar 3. Diagram Batang Frekuensi

Gambar 3 menampilkan grafik batang distribusi kelas pH tanah (Class.pH) yang memperlihatkan jumlah sampel pada setiap kategori pH dalam dataset. Visualisasi ini disusun menggunakan *count plot* dengan menampilkan frekuensi absolut, bukan dalam bentuk persentase. Penyajian distribusi ini menjadi langkah awal analisis yang esensial, karena dapat memberikan gambaran mengenai potensi ketidakseimbangan kelas (*class imbalance*) pada variabel target. Informasi tersebut sangat penting sebelum tahap pemodelan machine learning, mengingat distribusi kelas yang tidak seimbang dapat mempengaruhi kinerja model serta menentukan pendekatan penanganan yang sesuai, baik melalui pemilihan metrik evaluasi, penerapan teknik *resampling* (seperti oversampling atau SMOTE), maupun penyesuaian *class weights*.

Hasil visualisasi menunjukkan bahwa distribusi data pada tiap kategori pH tidak seimbang. Kelas *Moderately alkaline* mendominasi dengan lebih dari 1.200 sampel, diikuti oleh kelas *Moderately acidic* dengan sekitar 800 sampel. Adapun

kelas *Neutral* hanya memiliki sekitar 400 sampel, sementara dua kelas lainnya, yaitu *Strongly alkaline* dan *Strongly acidic*, tercatat jauh lebih sedikit, masing-masing kurang dari 300 dan 200 sampel. Kondisi ini menegaskan adanya permasalahan *class imbalance*, di mana dominasi kelas mayoritas berpotensi menimbulkan bias model terhadap kelas dengan jumlah data yang lebih besar.



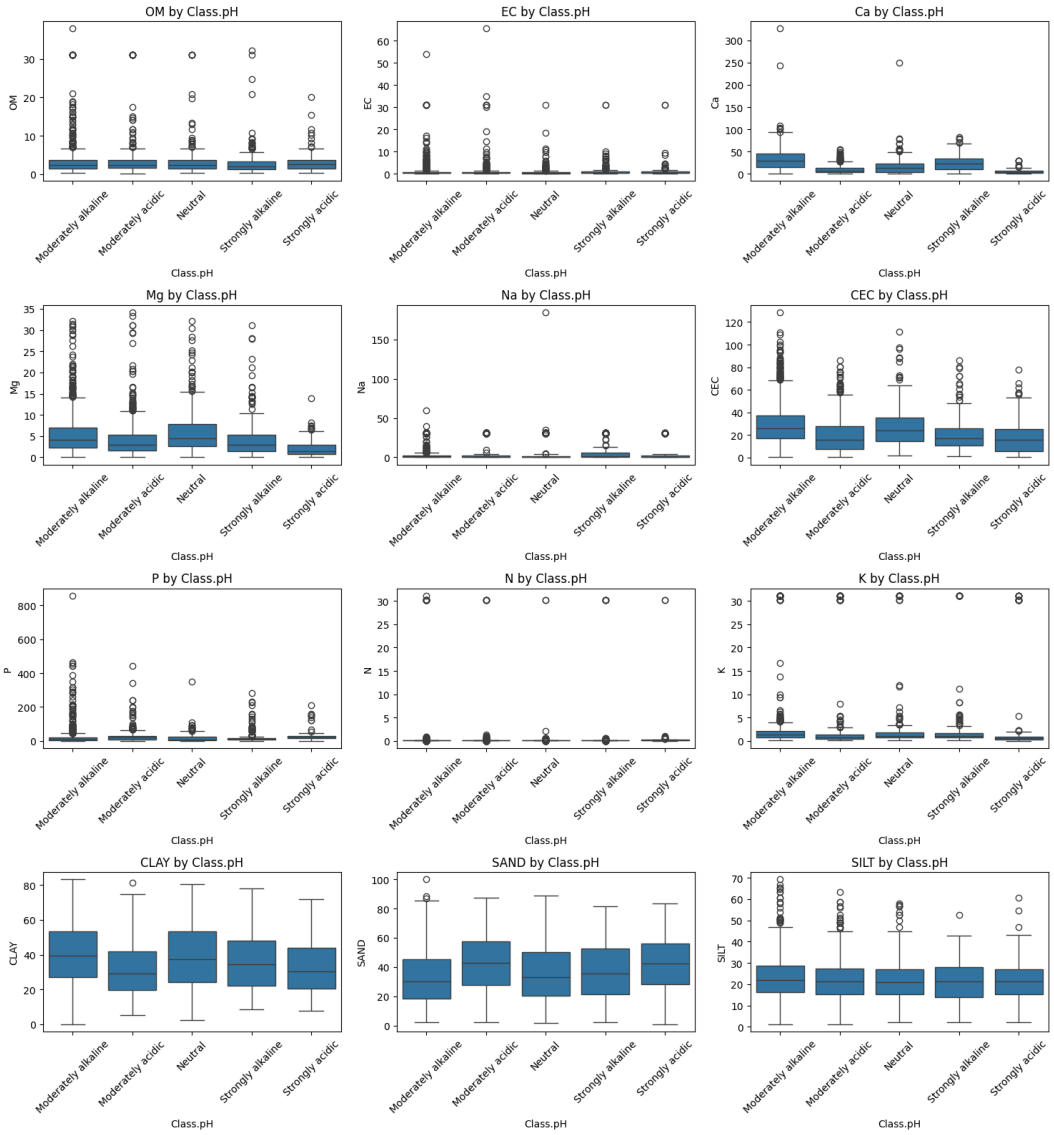
Gambar 4. Heatmap Korelasi Fitur Numerik

Gambar 4 menyajikan heatmap korelasi antar fitur numerik yang digunakan untuk menelaah hubungan antar variabel dalam dataset. Analisis ini dilakukan hanya pada atribut numerik dengan menghitung nilai korelasi menggunakan metode Pearson. Nilai korelasi berada pada rentang –1 hingga +1, di mana warna merah merepresentasikan korelasi positif, yaitu peningkatan suatu variabel cenderung diikuti oleh kenaikan variabel lain yang berkaitan. Sebaliknya, warna biru menunjukkan korelasi negatif, yakni kenaikan pada satu variabel berhubungan dengan penurunan variabel lainnya.

Hasil visualisasi memperlihatkan adanya kelompok variabel yang saling berhubungan kuat. Fitur kimia tanah seperti Ca, Mg, Na, CEC, SAR, dan ESP menunjukkan korelasi yang erat karena sifatnya yang saling mempengaruhi. Di sisi lain, atribut tekstur tanah seperti SAND, SILT, dan CLAY juga menampilkan hubungan yang cukup signifikan sesuai dengan karakteristik fisik tanah yang saling melengkapi. Sementara itu, terdapat beberapa fitur dengan korelasi relatif

rendah terhadap variabel lain, misalnya ID, EC, OM, dan BD, sehingga kontribusinya dalam keterkaitan antar fitur lebih kecil.

Temuan dari heatmap ini memiliki peran penting dalam tahapan analisis selanjutnya, terutama pada proses *feature selection*. Identifikasi variabel dengan korelasi tinggi dapat membantu peneliti mengantisipasi risiko multikolinearitas yang dapat mengurangi kinerja model machine learning. Selain itu, analisis ini juga memberikan pemahaman awal mengenai pola keterhubungan antar atribut sebelum dilakukan proses pemodelan lebih lanjut.



Gambar 5. Boxplot per Kategori

Gambar 5 menampilkan boxplot yang menggambarkan distribusi setiap parameter kimia tanah terhadap kategori pH. Secara umum, terlihat bahwa data memiliki keragaman yang cukup tinggi dengan keberadaan sejumlah nilai pencilan (*outlier*) pada beberapa parameter. Kandungan bahan organik (OM) relatif lebih tinggi pada tanah dengan pH *moderately acidic* dan *neutral*, yang ditunjukkan oleh nilai median yang lebih besar serta sebaran outlier yang cukup banyak. Konduktivitas listrik (EC) cenderung rendah pada seluruh kelas pH, meskipun terdapat beberapa sampel dengan nilai sangat tinggi. Unsur kalsium (Ca) menunjukkan variasi yang lebar dengan banyak outlier, di mana median tertinggi terdapat pada kelas *moderately acidic* dan *strongly acidic*. Sementara itu, kandungan magnesium (Mg) tampak lebih seragam, namun kelas *neutral* dan *strongly acidic* memiliki median sedikit lebih tinggi dibandingkan kategori lainnya. Untuk unsur natrium (Na), sebagian besar nilai terdistribusi rendah pada semua kelas pH, dengan beberapa outlier yang muncul pada kategori *neutral* dan *strongly alkaline*.

Parameter kapasitas tukar kation (CEC) menunjukkan sebaran yang cukup luas, khususnya pada tanah *moderately acidic* dan *strongly acidic* dengan median yang lebih tinggi dibandingkan kelas lainnya. Unsur fosfor (P) secara umum memiliki nilai rendah, tetapi terdapat banyak outlier dengan nilai tinggi pada kelas *moderately acidic* dan *neutral*. Kandungan nitrogen (N) tidak menunjukkan perbedaan yang berarti antar kelas pH, dengan sebagian besar nilai mendekati nol, kecuali beberapa outlier pada kategori *moderately acidic*. Unsur kalium (K) juga cenderung rendah dengan median yang relatif seragam, meskipun ditemukan cukup banyak outlier pada tanah *neutral* dan *moderately acidic*.

Selain parameter kimia, sifat fisik tanah juga menunjukkan variasi yang berbeda antar kelas pH. Kandungan liat (CLAY) lebih tinggi pada tanah *moderately acidic* dan lebih rendah pada *strongly acidic*. Sebaliknya, kandungan pasir (SAND) relatif lebih tinggi pada tanah *strongly acidic* dan lebih rendah pada *moderately acidic*. Adapun kandungan debu (SILT) relatif seragam pada semua kelas pH, meskipun ditemukan sejumlah outlier pada tanah *moderately alkaline* dan *neutral*. Secara keseluruhan, hasil ini mengindikasikan bahwa parameter kimia seperti OM, Ca, Mg, dan CEC, serta sifat fisik seperti CLAY dan SAND,

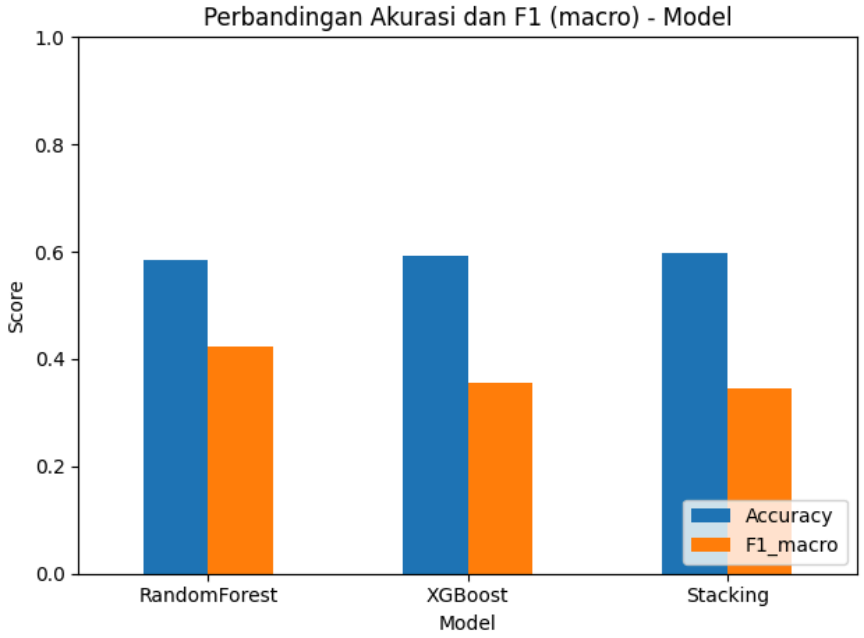
memiliki kontribusi lebih dominan dalam membedakan kategori pH tanah dibandingkan variabel lain seperti Na, N, K, dan SILT yang cenderung menunjukkan pola seragam.

Gambar 6. Scatterplot Matrix

Gambar 6 memperlihatkan *pairplot* yang menampilkan kurva kepadatan (*Kernel Density Estimation/KDE*) pada diagonal untuk masing-masing variabel terhadap kategori pH, sedangkan bagian luar diagonal menampilkan *scatter plot* hubungan antar variabel. Secara umum, distribusi variabel fosfor (P) dan kalsium (Ca) terlihat tidak seimbang dengan kecenderungan miring ke kanan (*right-skewed*) serta disertai banyak nilai pencilan (*outlier*) ekstrem. Pola ini menunjukkan bahwa sebagian besar sampel memiliki nilai relatif rendah, sementara sejumlah kecil sampel memiliki nilai sangat tinggi yang kemungkinan dipengaruhi oleh kondisi tanah tertentu atau akibat aplikasi pupuk dan kapur. Berbeda dengan itu, variabel pasir (SAND) memperlihatkan distribusi yang lebih merata dan cenderung mendekati normal. Adapun variabel ID hanya berfungsi

sebagai penomoran sampel sehingga tidak memiliki makna geokimia, bahkan menimbulkan pola garis vertikal yang menyesatkan pada *scatter plot*. Dengan demikian, variabel ini dianggap tidak relevan untuk dianalisis lebih lanjut.

Dari segi hubungan antar variabel, tidak ditemukan adanya korelasi linear yang kuat. Sebaran antara P dengan Ca maupun P dengan SAND tampak acak, yang menandakan bahwa tingginya nilai P tidak selalu berkaitan dengan peningkatan Ca maupun persentase pasir tertentu. Meskipun demikian, terdapat kecenderungan bahwa kadar Ca menurun seiring meningkatnya kandungan pasir, sejalan dengan konsep bahwa tanah berpasir umumnya memiliki kadar kation lebih rendah. Warna titik yang mewakili kelas pH tampak bercampur pada hampir seluruh plot, sehingga keempat variabel ini belum dapat membedakan kategori pH secara jelas. Meski demikian, terdapat indikasi bahwa tanah dengan pH alkalin cenderung mengandung Ca lebih tinggi, walaupun polanya tidak begitu menonjol. Secara keseluruhan, hasil *pairplot* ini mengindikasikan bahwa distribusi data masih dipengaruhi oleh keberadaan *outlier* dan ketidakseimbangan distribusi, khususnya pada variabel P dan Ca. Oleh karena itu, sebelum dilakukan analisis lanjutan maupun pemodelan, diperlukan tahap pra pemrosesan data, seperti mengeluarkan variabel ID, melakukan transformasi logaritma pada variabel dengan distribusi sangat miring, serta memverifikasi kembali nilai-nilai ekstrim. Selain itu, untuk memperoleh pemisahan kelas pH yang lebih akurat, analisis perlu melibatkan variabel tambahan, misalnya kandungan bahan organik (OM), kapasitas tukar kation (CEC), serta proporsi tekstur tanah secara lebih lengkap, tidak hanya terbatas pada empat variabel yang divisualisasikan.



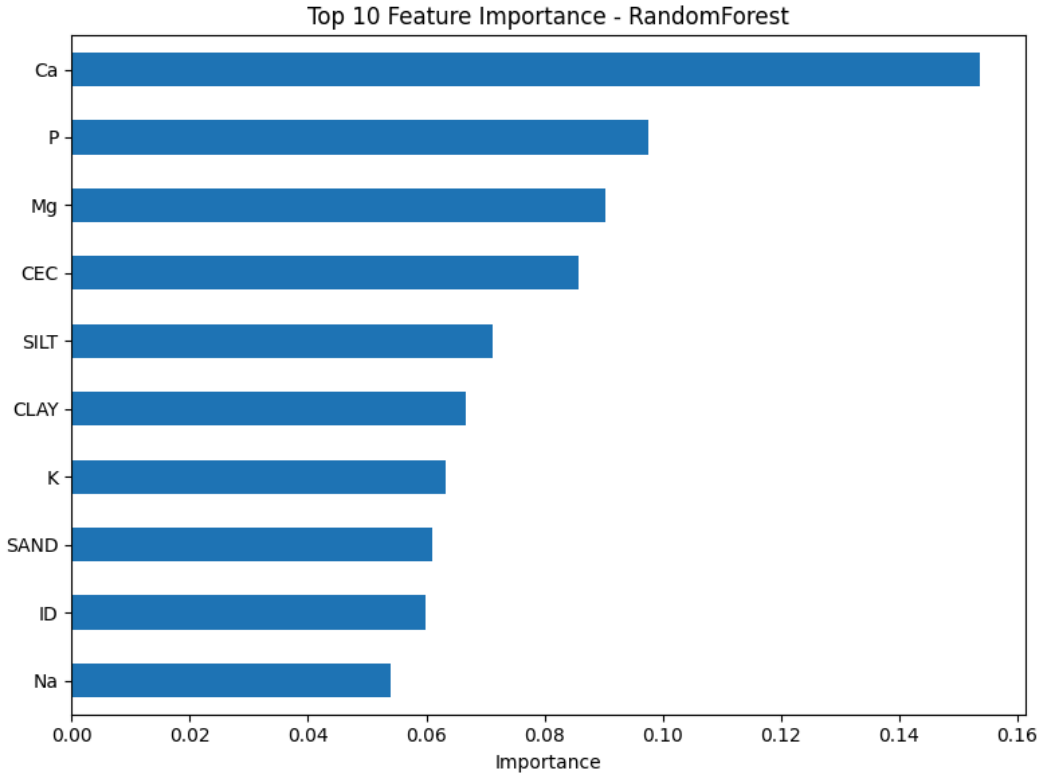
Gambar 7. Perbandingan Skor Model dengan Diagram Batang

Gambar 7 menampilkan perbandingan nilai akurasi dan F1-macro dari tiga algoritma klasifikasi, yakni Random Forest, XGBoost, dan Stacking. Pada sumbu vertikal ditunjukkan skor dengan rentang 0–1, sedangkan sumbu horizontal menampilkan nama masing-masing model. Setiap model digambarkan dengan dua batang, di mana warna biru merepresentasikan akurasi dan warna oranye menggambarkan F1-macro.

Berdasarkan visualisasi, ketiga model memperlihatkan capaian akurasi yang relatif sebanding, yaitu berkisar antara 0,59 hingga 0,60. Hal ini menunjukkan bahwa tingkat ketepatan prediksi secara keseluruhan tidak berbeda jauh antar model. Namun, perbedaan terlihat jelas pada metrik F1-macro. Random Forest mencatat skor F1-macro tertinggi (sekitar 0,42), sedangkan XGBoost dan Stacking hanya mencapai nilai ±0,35.

Perbedaan mencolok antara akurasi dan F1-macro ini mengindikasikan adanya ketidakseimbangan kelas (*class imbalance*) pada dataset. Akurasi terlihat cukup tinggi karena model cenderung lebih baik memprediksi kelas mayoritas, sementara nilai F1-macro yang lebih rendah menandakan bahwa kemampuan model dalam mengenali kelas minoritas masih terbatas. Dengan demikian, meskipun akurasi ketiga model hampir sama, Random Forest dapat dianggap

lebih unggul karena mampu menjaga keseimbangan kinerja prediksi antar kelas dibandingkan XGBoost maupun Stacking.



Gambar 8. Feature Importance Plot dengan Diagram Batang Horizontal

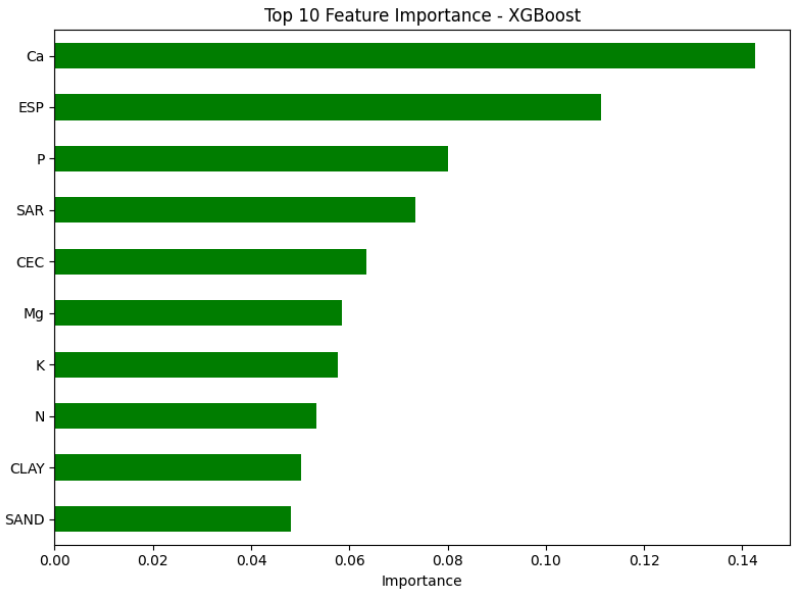
Gambar 8 menampilkan sepuluh fitur dengan tingkat kepentingan tertinggi berdasarkan analisis *feature importance* menggunakan algoritma Random Forest pada klasifikasi pH tanah. Sumbu horizontal menunjukkan nilai kepentingan yang merepresentasikan kontribusi masing-masing fitur dalam proses pengambilan keputusan model, sedangkan sumbu vertikal menampilkan nama variabel yang dianalisis.

Hasil visualisasi menunjukkan bahwa unsur kalsium (Ca) menempati posisi paling dominan dengan nilai kepentingan mendekati 0,16, jauh melampaui fitur lainnya. Temuan ini menegaskan bahwa kadar Ca merupakan faktor utama yang memengaruhi prediksi pH tanah. Selain Ca, unsur fosfor (P) dan magnesium (Mg) juga memberikan kontribusi signifikan, diikuti oleh kapasitas tukar kation (CEC) serta Exchangeable Sodium Percentage (ESP) yang berperan penting dalam menjelaskan sifat kimia tanah.

Secara ilmiah, peran dominan variabel-variabel tersebut sejalan dengan prinsip dasar kimia tanah. Unsur Ca dan Mg berfungsi sebagai kation basa yang dapat menetralkan ion hidrogen (H^+) dalam tanah, sehingga keduanya dikenal sebagai agen penyangga (*buffer*) pH. Tanah dengan kandungan Ca dan Mg tinggi cenderung bersifat lebih netral hingga basa karena ion-ion tersebut mampu menekan tingkat keasaman. Sementara itu, CEC mencerminkan kapasitas tanah dalam menahan kation, termasuk ion H^+. Tanah dengan nilai CEC tinggi umumnya memiliki kemampuan lebih besar dalam mengendalikan kondisi keasaman maupun alkalinitas, sekaligus menjadi indikator utama kesuburan tanah. Fosfor (P) juga memiliki keterkaitan erat dengan pH tanah, karena ketersediaannya sangat dipengaruhi oleh kondisi keasaman. Pada tanah masam, fosfor cenderung berikatan dengan Al dan Fe sehingga sulit tersedia bagi tanaman, sedangkan pada tanah basa fosfor berikatan dengan Ca. Hal ini menjelaskan mengapa fosfor muncul sebagai salah satu faktor penting dalam klasifikasi pH tanah. Di sisi lain, tingginya nilai ESP menunjukkan dominasi natrium dalam kompleks pertukaran kation. Tanah dengan ESP tinggi biasanya menghadapi masalah alkalinitas dan sodisitas, sehingga wajar bila variabel ini juga teridentifikasi sebagai faktor penentu dalam model.

Selain faktor kimia, fitur tekstur tanah seperti SILT, CLAY, dan SAND turut berkontribusi meskipun pengaruhnya relatif lebih rendah. Menariknya, variabel ID yang sejatinya hanya berfungsi sebagai penanda data masih muncul dalam sepuluh besar meskipun dengan nilai kepentingan rendah.

Secara keseluruhan, hasil analisis ini menunjukkan bahwa faktor kimia tanah, terutama Ca, Mg, P, CEC, dan ESP, merupakan penentu utama dalam klasifikasi pH tanah. Temuan ini tidak hanya mendukung kinerja model *machine learning*, tetapi juga konsisten dengan teori agronomi yang menjelaskan mekanisme kimia tanah. Dengan demikian, pendekatan *machine learning* terbukti tidak hanya akurat secara statistik, melainkan juga mampu menangkap dinamika proses kimia yang sesungguhnya di dalam tanah.



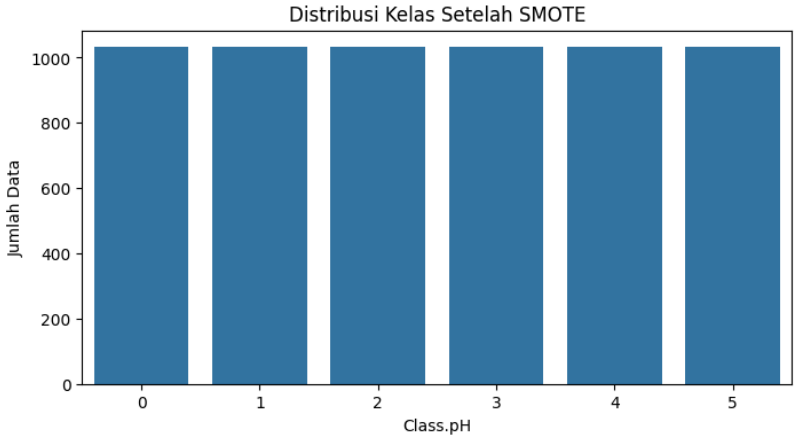
Gambar 9. Feature Importance Plot dengan Diagram Batang Horizontal (XGBoost)

Gambar 9 menampilkan sepuluh fitur dengan kontribusi terbesar pada model XGBoost dalam klasifikasi pH tanah. Sumbu horizontal menggambarkan nilai *importance* yang menunjukkan besarnya peran suatu fitur dalam membantu model membuat prediksi, sedangkan sumbu vertikal menampilkan nama fitur yang dianalisis.

Hasil analisis menunjukkan bahwa unsur kalsium (Ca) memiliki tingkat kepentingan paling dominan dengan nilai *importance* tertinggi, diikuti oleh ESP (*Exchangeable Sodium Percentage*). Kedua fitur ini menjadi indikator utama yang menentukan prediksi pH tanah pada model XGBoost. Selain itu, fosfor (P), SAR (*Sodium Adsorption Ratio*), serta kapasitas tukar kation (CEC) juga memberikan kontribusi yang cukup besar, menggarisbawahi pentingnya faktor kimia dalam menjelaskan variasi tingkat keasaman maupun alkalinitas tanah.

Unsur lain seperti magnesium (Mg), kalium (K), dan nitrogen (N) turut menyumbang pengaruh meskipun nilainya relatif lebih kecil dibandingkan fitur dominan. Sementara itu, faktor tekstur tanah seperti CLAY dan SAND berada pada urutan terbawah, yang menunjukkan bahwa meskipun karakteristik fisik tanah tetap berperan, pengaruhnya tidak sebesar faktor kimia dalam menentukan pH tanah.

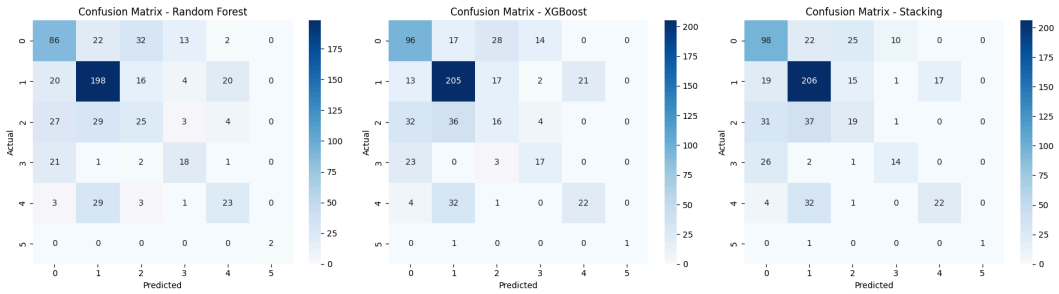
Secara keseluruhan, hasil ini memperlihatkan bahwa model XGBoost lebih menitikberatkan pada indikator kimia tanah terutama Ca, ESP, P, SAR, dan CEC dibandingkan dengan faktor tekstur dalam proses klasifikasi pH tanah.



Gambar 10. Countplot Distribusi Kelas Setelah SMOTE

Gambar 10 memperlihatkan distribusi kelas setelah dilakukan penerapan SMOTE (*Synthetic Minority Oversampling Technique*) pada variabel target *Class.pH_encoded*. Sebelum proses ini, dataset menunjukkan ketidakseimbangan kelas, di mana beberapa kategori pH tanah memiliki jumlah sampel jauh lebih sedikit dibandingkan dengan kelas lainnya. Kondisi tersebut berpotensi menimbulkan bias pada model klasifikasi karena cenderung lebih akurat dalam mengenali kelas mayoritas dan mengabaikan kelas minoritas.

Setelah dilakukan oversampling menggunakan SMOTE, seluruh kelas (0 hingga 5) memiliki jumlah data yang relatif seimbang, yaitu sekitar 1.030 sampel per kelas. Hal ini membuktikan bahwa SMOTE mampu menyeimbangkan distribusi data melalui pembangkitan sampel sintetis pada kelas minoritas. Dengan distribusi yang lebih proporsional, model klasifikasi diharapkan dapat mempelajari pola dari setiap kelas secara lebih setara, sehingga meningkatkan kinerja prediksi, khususnya pada metrik evaluasi seperti F1-score yang lebih peka terhadap masalah ketidakseimbangan data.

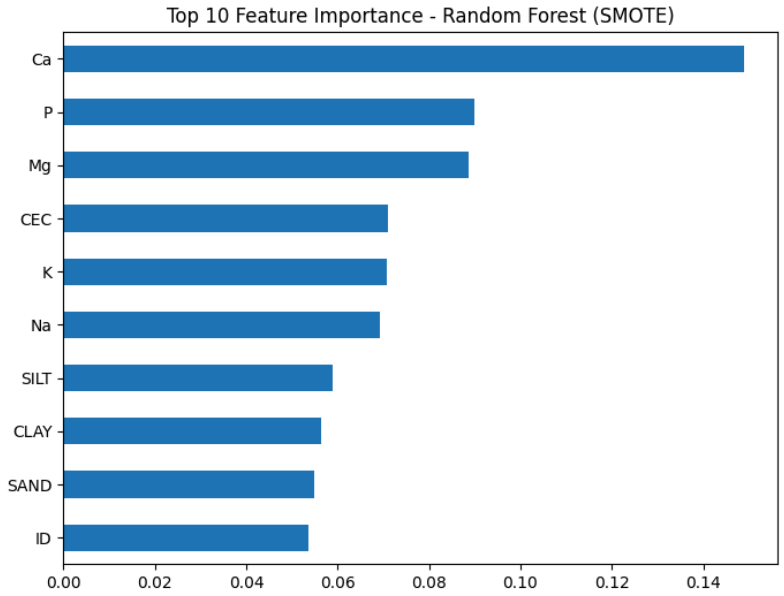


Gambar 11. Confusion Matrix Heatmap untuk Perbandingan Model

Gambar 11 memperlihatkan confusion matrix dari tiga model klasifikasi yang digunakan, yaitu Random Forest, XGBoost, dan Stacking, yang menggambarkan perbandingan antara label aktual dan label prediksi pada setiap kategori pH tanah. Hasil analisis menunjukkan bahwa ketiga model mampu memprediksi kelas mayoritas (kelas 0 dan 1) dengan cukup baik, di mana Random Forest mencatat 198 sampel benar pada kelas 1 dan Stacking mencapai 206 sampel benar pada kelas yang sama. Namun, performa menurun pada kelas dengan jumlah sampel lebih kecil, seperti kelas 2, 3, dan 4, yang cenderung salah terklasifikasi ke kelas mayoritas.

Kondisi ini menjelaskan mengapa nilai akurasi keseluruhan terlihat cukup baik, tetapi F1-score relatif lebih rendah. Kesulitan model dalam mendistribusikan performa secara merata pada seluruh kelas mengindikasikan adanya tantangan terkait ketidakseimbangan data. Dalam konteks ini, model Stacking menunjukkan keunggulan relatif, dengan nilai F1-score sedikit lebih tinggi dibandingkan Random Forest dan XGBoost. Walaupun peningkatan tersebut tidak signifikan, keunggulan Stacking terletak pada stabilitas prediksinya.

Hal ini dapat dijelaskan karena Random Forest dan XGBoost sama-sama kuat dalam menangani data tabular dengan interaksi non-linear, sehingga meta-model Stacking tidak memperoleh banyak informasi tambahan untuk menghasilkan lonjakan performa yang besar. Namun, dengan mengkombinasikan kedua model dasar, Stacking mampu menekan risiko bias yang mungkin muncul bila hanya menggunakan satu model tunggal. Oleh karena itu, meskipun perbedaannya tipis, Stacking tetap lebih dapat diandalkan untuk menghadapi variasi data lapangan yang lebih kompleks, menjadikannya pilihan yang lebih stabil untuk aplikasi praktis.



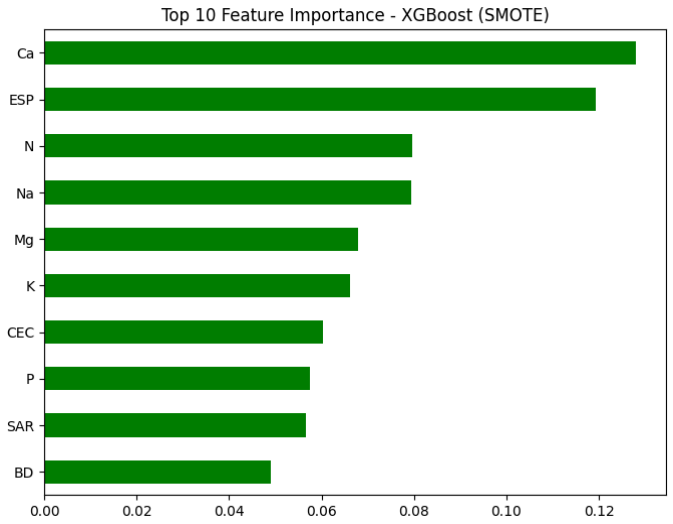
Gambar 12. Feature Importance Plot (Random Forest, SMOTE) dalam bentuk Diagram Batang Horizontal.

Gambar 12 menampilkan sepuluh variabel dengan nilai *feature importance* tertinggi pada model Random Forest setelah dilakukan proses penyeimbangan data menggunakan SMOTE. Nilai *feature importance* ini menggambarkan seberapa besar kontribusi relatif masing-masing variabel dalam memengaruhi prediksi model. Berdasarkan hasil visualisasi, variabel kalsium (Ca) menempati posisi paling dominan dengan nilai penting yang jauh lebih tinggi dibandingkan variabel lain, menegaskan peran utama unsur Ca dalam menentukan klasifikasi pH tanah.

Selain Ca, variabel fosfor (P) dan magnesium (Mg) juga memberikan kontribusi yang cukup besar, meskipun nilainya tidak sebesar Ca. Selanjutnya, kapasitas tukar kation (CEC), kalium (K), natrium (Na), serta parameter fisik seperti SILT, CLAY, dan SAND turut tercatat memiliki pengaruh meski relatif lebih rendah. Hal yang menarik, meskipun variabel ID sejatinya hanya berfungsi sebagai penanda data, model tetap menangkap pola tertentu dari variabel ini. Namun demikian, secara teoritis ID tidak semestinya dijadikan dasar utama dalam klasifikasi.

Secara keseluruhan, temuan ini menunjukkan bahwa unsur kimia tanah khususnya Ca, P, dan Mg memiliki pengaruh yang lebih kuat terhadap klasifikasi

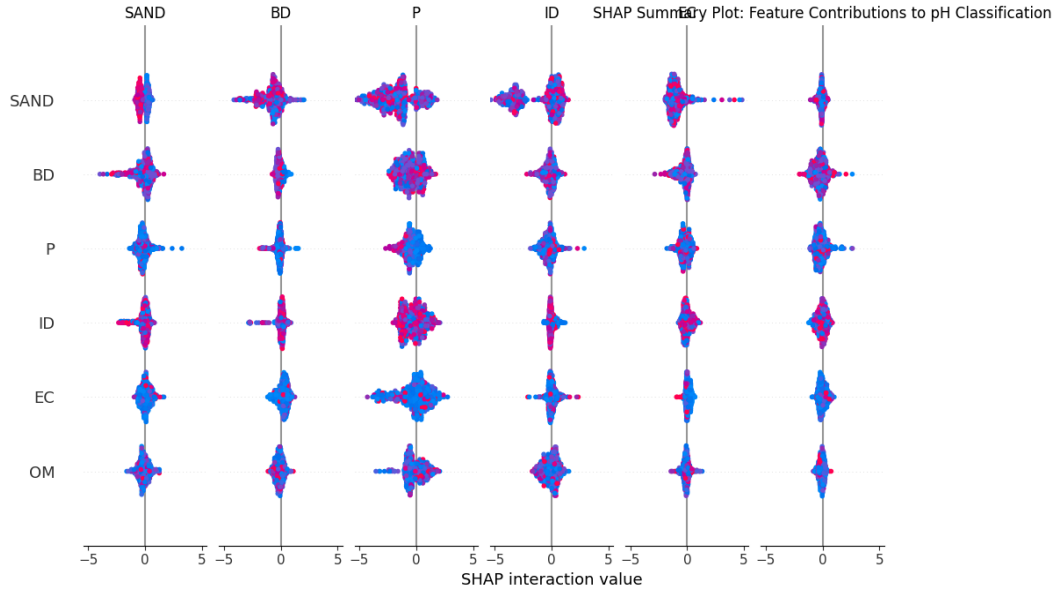
pH tanah dibandingkan faktor fisik seperti SILT, CLAY, maupun SAND. Hasil ini dapat menjadi rujukan penting, baik dalam penelitian akademis maupun penerapan praktis di bidang pertanian, karena memberikan gambaran variabel utama yang berperan dalam analisis kesuburan tanah dan strategi pengelolaan lahan.



Gambar 13. Feature Importance Plot (XGBoost, SMOTE) dalam bentuk Diagram Batang Horizontal.

Gambar 13 menyajikan hasil analisis *feature importance* pada model **XGBoost** setelah dilakukan penyeimbangan data menggunakan SMOTE. Berdasarkan grafik tersebut, kalsium (Ca) tercatat sebagai variabel dengan kontribusi paling dominan dalam mempengaruhi prediksi, disusul oleh ESP (*Exchangeable Sodium Percentage*). Selain itu, unsur nitrogen (N), natrium (Na), magnesium (Mg), serta kalium (K) juga memberikan peran yang cukup berarti, meskipun tingkat pengaruhnya masih berada di bawah Ca dan ESP.

Di sisi lain, variabel seperti kapasitas tukar kation (CEC), fosfor (P), SAR (*Sodium Adsorption Ratio*), dan BD (*Bulk Density*) menunjukkan kontribusi yang relatif lebih rendah dalam menentukan keluaran model. Secara keseluruhan, hasil ini menegaskan bahwa model *XGBoost* lebih menitikberatkan pada indikator kimia tanah tertentu dalam proses klasifikasi pH. Temuan ini dapat menjadi rujukan penting dalam kajian karakteristik tanah, khususnya dalam mengidentifikasi faktor utama yang berperan terhadap variasi tingkat keasaman maupun alkalinitas tanah.



Gambar 14. Plot Ringkasan SHAP (tipe titik)

Gambar 14 menampilkan SHAP *(Summary Plot)* dengan *interaction values* yang memberikan gambaran lebih mendalam mengenai kontribusi interaksi antar variabel dalam mempengaruhi hasil prediksi pH tanah. Pada sumbu horizontal ditunjukkan nilai interaksi SHAP dengan rentang antara -5 hingga +5, yang merepresentasikan besarnya pengaruh interaksi dua fitur terhadap arah prediksi model. Nilai positif mengindikasikan bahwa interaksi antar variabel meningkatkan kecenderungan klasifikasi ke suatu kategori, sedangkan nilai negatif menunjukkan penurunan kecenderungan tersebut. Sementara itu, sumbu vertikal menampilkan fitur utama (SAND, BD, P, ID, EC, OM) yang masing-masing memuat distribusi interaksinya dengan fitur lain. Visualisasi menyerupai *violin plot* memperlihatkan sebaran nilai interaksi, di mana lebar distribusi mencerminkan jumlah sampel yang memiliki nilai interaksi serupa. Variasi warna titik dari biru hingga merah menggambarkan nilai asli dari fitur, sehingga memudahkan dalam memahami perbedaan dampak antara nilai rendah dan tinggi.

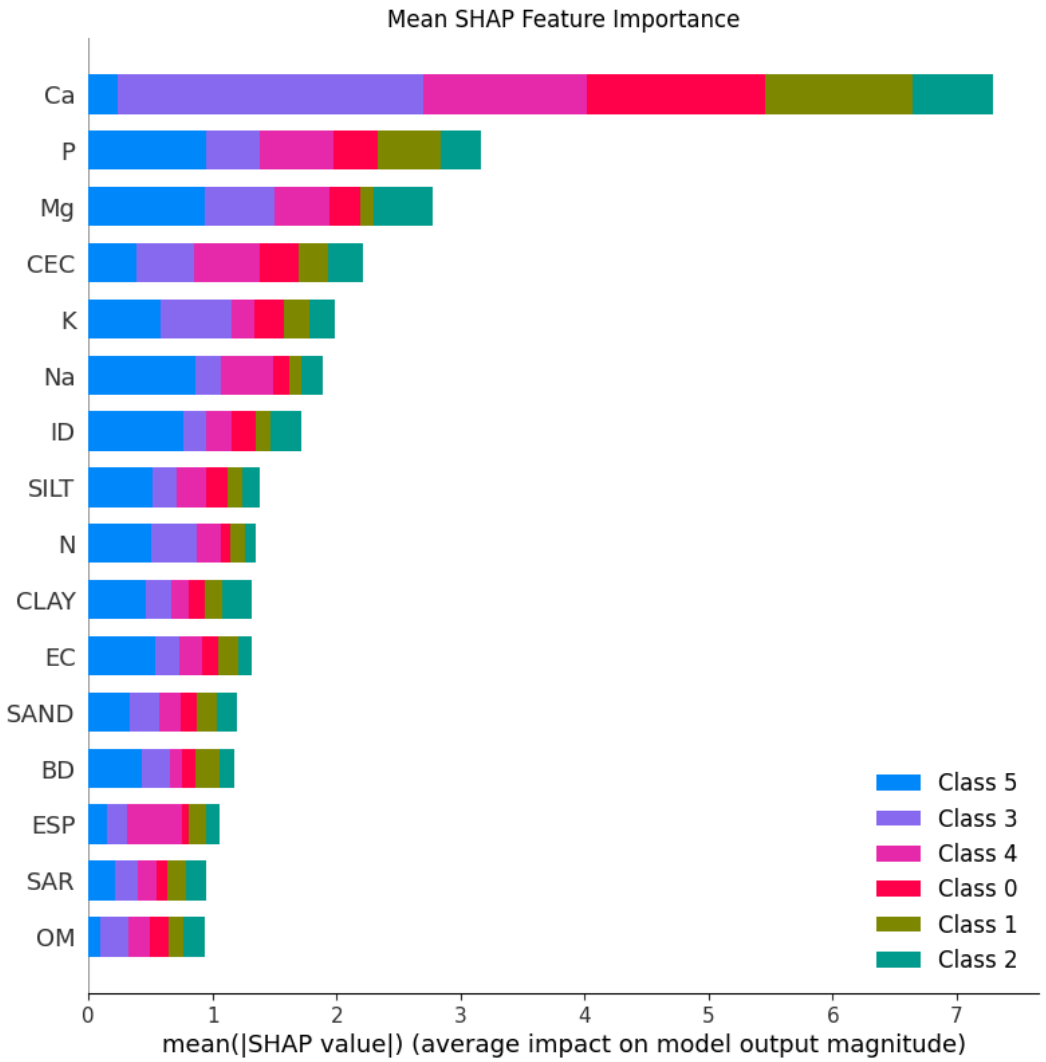
Berdasarkan hasil analisis, fosfor (P) tampak menempati posisi sentral dengan distribusi interaksi yang cukup lebar, terutama terhadap *bulk density* (BD) dan ID. Hal ini mengindikasikan bahwa variasi kadar fosfor sangat dipengaruhi oleh kepadatan tanah maupun pola distribusi data. Temuan ini sejalan dengan pemahaman ilmiah bahwa ketersediaan fosfor sangat erat kaitannya dengan

kondisi fisik tanah. Fitur SAND (pasir) juga menunjukkan rentang interaksi yang luas, khususnya dengan BD, EC, dan OM, yang menegaskan peran dominan tekstur tanah berpasir dalam mempengaruhi retensi air dan ketersediaan unsur hara.

Selanjutnya, BD menampilkan interaksi yang kuat dengan P, SAND, dan OM. Tingginya nilai BD biasanya berasosiasi dengan menurunnya porositas tanah, yang pada gilirannya berdampak pada ketersediaan unsur hara penting. Kondisi ini menegaskan bahwa karakteristik fisik tanah tidak dapat dipisahkan dari interpretasi sifat kimianya. Sementara itu, meskipun ID tidak memiliki makna kimiawi langsung, interaksi yang muncul cukup signifikan. Kemungkinan besar, variabel ini menangkap pola tersembunyi dari struktur data sehingga penggunaannya perlu dikaji ulang agar tidak menimbulkan bias dalam interpretasi hasil.

Fitur EC (Electrical Conductivity) memperlihatkan interaksi penting dengan P dan SAND. Salinitas tanah yang tinggi, sebagaimana direpresentasikan oleh EC, dapat mempengaruhi ketersediaan fosfor, sehingga kombinasi keduanya memberikan kontribusi berarti bagi model. Di sisi lain, OM (Organic Matter) menunjukkan distribusi interaksi yang lebih sempit, meskipun tetap relevan melalui hubungannya dengan BD dan SAND. Secara teoritis, bahan organik memiliki peran penting dalam memperbaiki struktur tanah sekaligus meningkatkan ketersediaan unsur hara, sehingga keberadaannya meski tidak dominan tetap memiliki pengaruh terhadap klasifikasi pH.

Secara keseluruhan, hasil visualisasi ini menegaskan bahwa prediksi pH tanah tidak hanya ditentukan oleh kontribusi variabel tunggal, melainkan juga oleh kombinasi kompleks antara faktor fisik dan kimia. Fitur P, SAND, dan BD tampak menjadi pusat interaksi yang paling berpengaruh, yang sekaligus mencerminkan tingginya keragaman sifat tanah di lapangan. Lebarnya distribusi nilai SHAP pada beberapa fitur menunjukkan bahwa pengaruh interaksi antar variabel dapat berbeda secara signifikan antar sampel, sehingga model machine learning mampu menangkap heterogenitas nyata yang terjadi pada kondisi tanah.



Gambar 15. Rata-rata absolut SHAP value dari tiap fitur

Gambar 15 menampilkan rata-rata absolut SHAP *value* dari tiap fitur, yang memperlihatkan rata-rata kontribusi setiap fitur terhadap hasil prediksi model klasifikasi pH tanah. Nilai yang ditunjukkan berasal dari perhitungan rata-rata absolut SHAP *value* pada masing-masing variabel. Semakin panjang batang horizontal pada grafik, semakin besar pula pengaruh fitur tersebut terhadap keputusan model.

Pada sumbu horizontal ditampilkan nilai *mean(SHAP value)* yang menunjukkan seberapa besar dampak rata-rata sebuah fitur terhadap output model. Semakin jauh batang bergeser ke kanan, semakin besar perannya dalam klasifikasi. Sedangkan pada sumbu vertikal, tercantum daftar variabel tanah yang digunakan, seperti Ca, P, Mg, CEC, K, Na, hingga OM. Warna yang membentuk

batang menunjukkan distribusi pengaruh fitur pada tiap kategori pH (Class 0 hingga Class 5), sesuai dengan keterangan pada legenda grafik. Dengan visualisasi ini, bukan hanya pentingnya fitur secara keseluruhan yang terlihat, tetapi juga bagaimana kontribusinya dapat berbeda pada tiap kelas target.

Dari grafik, terlihat jelas bahwa kalsium (Ca) berada di urutan teratas dengan kontribusi rata-rata terbesar. Hal ini mengindikasikan bahwa Ca adalah faktor utama yang mempengaruhi klasifikasi pH tanah, sejalan dengan teori bahwa unsur ini berperan penting dalam menetralkan ion H^+ sehingga berpengaruh langsung terhadap tingkat keasaman tanah. Di posisi berikutnya terdapat fosfor (P) dan magnesium (Mg) yang juga memiliki peran signifikan dalam prediksi. Sementara itu, CEC (Cation Exchange Capacity) serta unsur basa lain seperti K dan Na masih memberikan pengaruh, meskipun lebih kecil dibandingkan Ca, P, dan Mg.

Sebaliknya, variabel tekstur tanah seperti SILT, CLAY, dan SAND, maupun faktor fisik seperti BD (*Bulk Density*) dan EC (*Electrical Conductivity*), memperlihatkan nilai rata-rata SHAP yang relatif rendah. Hal ini menunjukkan bahwa kontribusinya terhadap klasifikasi pH tanah tidak sebesar unsur kimia. Beberapa variabel lain seperti ESP (*Exchangeable Sodium Percentage*), SAR (*Sodium Adsorption Ratio*), dan OM (*Organic Matter*) juga terlihat memiliki dampak yang terbatas. Menariknya, variabel ID, meskipun kemungkinan besar hanya berfungsi sebagai penanda data, tetap muncul dengan kontribusi sedang sehingga patut dicermati lebih lanjut.

Secara umum, grafik ini menegaskan bahwa variabel kimia tanah terutama Ca, P, dan Mg mempunyai peran dominan dalam klasifikasi pH tanah, sementara variabel tekstur maupun fisik menunjukkan pengaruh yang lebih kecil. Perbedaan warna pada batang juga menegaskan bahwa dampak setiap fitur dapat bervariasi pada tiap kelas pH, sehingga memperlihatkan kompleksitas hubungan antar variabel dalam mempengaruhi sifat tanah.

BAB 3
PENUTUP

3.1 Kesimpulan

Penelitian ini menunjukkan bahwa pendekatan *machine learning* mampu memberikan alternatif yang lebih efisien dan praktis dalam mengklasifikasikan pH tanah berdasarkan karakteristik kimia dibandingkan metode konvensional berbasis laboratorium. Dengan memanfaatkan dataset yang kompleks, penelitian ini berhasil mengintegrasikan variabel-variabel kimia tanah, seperti kalsium (Ca), fosfor (P), magnesium (Mg), serta kapasitas tukar kation (CEC), yang terbukti memiliki pengaruh dominan terhadap variasi pH. Hasil ini menegaskan bahwa faktor kimia tanah memiliki peran krusial dalam menentukan tingkat keasaman maupun kebasaan tanah, sekaligus membuka peluang penerapan teknologi berbasis kecerdasan buatan dalam pertanian presisi.

Dari sisi metodologis, penerapan algoritma *Random Forest* dan *XGBoost* membuktikan kemampuannya dalam mengolah data berskala besar serta menghadapi kompleksitas variabel input. *Random Forest* unggul dalam menjaga stabilitas prediksi dengan skor *F1-macro* yang lebih baik, sedangkan *XGBoost* menonjol dalam aspek efisiensi komputasi dan kemampuan penanganan data yang tidak seimbang. Perbandingan ini memperlihatkan bahwa pemilihan algoritma tidak dapat dilepaskan dari konteks data dan tujuan analisis, sehingga keduanya tetap memiliki peran penting dalam skenario pemodelan klasifikasi pH tanah.

Evaluasi model melalui metrik *akurasi*, *presisi*, *recall*, *F1-score*, serta ROC AUC mengindikasikan bahwa ketidakseimbangan kelas masih menjadi tantangan utama dalam pemodelan. Upaya penyeimbangan data dengan metode SMOTE terbukti meningkatkan kinerja model, terutama pada aspek generalisasi dan pengenalan kelas minoritas. Dengan distribusi kelas yang lebih proporsional, model mampu belajar pola dari setiap kategori pH tanah secara lebih setara, sehingga hasil prediksi menjadi lebih adil dan representatif.

Analisis lebih lanjut menggunakan SHAP menegaskan bahwa prediksi pH tanah tidak hanya dipengaruhi oleh kontribusi variabel tunggal, tetapi juga oleh interaksi kompleks antarvariabel. Fitur fosfor (P) menempati posisi sentral dengan interaksi kuat terhadap bulk density (BD) dan tekstur tanah (SAND),

menunjukkan bahwa ketersediaan unsur hara tidak dapat dilepaskan dari kondisi fisik tanah. Ca, P, dan Mg muncul sebagai faktor kimia dominan yang paling menentukan arah klasifikasi, sedangkan variabel fisik seperti BD, EC, maupun tekstur (SAND, CLAY, SILT) berperan sebagai faktor pendukung. Hasil mean SHAP value memperlihatkan bahwa Ca memiliki pengaruh rata-rata terbesar, sejalan dengan perannya dalam menetralkan ion H^+.

Secara keseluruhan, penelitian ini berhasil menunjukkan bahwa integrasi teknologi kecerdasan buatan dengan data kimia dan fisik tanah mampu menghasilkan sistem klasifikasi pH yang lebih akurat, efisien, dan representatif. Temuan ini tidak hanya memperkuat pemahaman mengenai dominasi unsur kimia dalam menentukan pH tanah, tetapi juga menegaskan pentingnya mempertimbangkan interaksi dengan faktor fisik dalam model prediksi. Dengan demikian, penerapan *machine learning* dalam klasifikasi pH tanah dapat mendukung praktik pertanian presisi yang berkelanjutan, meningkatkan efektivitas pengelolaan lahan, produktivitas tanaman, sekaligus menjaga keberlanjutan lingkungan.

DAFTAR PUSTAKA

[1] Mas' udi AF, Indarto I, Mandala M. Pemetaan indeks kualitas tanah (IKT) pada lahan tegalan di Kabupaten Jember. Jurnal Tanah dan Iklim. 2021 Sep 10;45(2):133-44.

[2] Haryati U. Karakteristik fisik tanah kawasan budidaya sayuran dataran tinggi, hubungannya dengan strategi pengelolaan lahan. Jurnal Sumberdaya Lahan. 2014 Dec;8(2):133497.

[3] Hazmi MN, Sumiharto R. Implementasi Kontrol Nutrisi Dan pH Pada Hidroponik Cerdas Berbasis Arduino Dan JST. IJEIS (*Indonesian Journal of Electronics and Instrumentation Systems*). 2023 Oct;13(2):159-70.

[4] Kusumaningrum L, Karina R, Fil'ardiani NU, Mardiyanto MB, Jabbar SA, Khoirunnisa S, Raharjo YA, Santika YE, Annisa Arta YP, Daniswara AP. *Analysis of the Carrying Capacity of Groundwater Availability and Its Relationship with the Largest Population Growth in Karanganyar Regency. Journal of Natural Resources & Environment* Management/Jurnal Pengelolaan Sumberdaya Alam dan Lingkungan. 2024 Jul 1;14(3).

[5] Ferreira da Silva A, Ohta RL, Tirapu Azpiroz J, Esteves Ferreira M, Marçal DV, Botelho A, Coppola T, Melo de Oliveira AF, Bettarello M, Schneider L, Vilaça R. AI enabled, *mobile soil pH classification with colorimetric paper sensors for sustainable agriculture. PLoS One*. 2025 Jan 22;20(1):e0317739.

[6] Singh A, Gaurav K, Sonkar GK, Lee CC. *Strategies to measure soil moisture using traditional methods, automated sensors, remote sensing, and machine learning techniques: review, bibliometric analysis, applications, research findings, and future directions.* Ieee Access. 2023 Feb 9;11:13605-35.

[7] Li D, Xiao E, Xia Y, Liang X, Guo M, Ning L, Yan J. *Improving soil pH prediction and mapping using anthropogenic variables and machine learning models. Geocarto International*. 2025 Dec 31;40(1):2482699.

[8] Huang P, Huang Q, Wang J, Shi Y. *Predicting surface soil pH spatial distribution based on three machine learning methods: a case study of*

Heilongjiang Province. Environmental Monitoring and Assessment. 2025 Mar 8;197(4):367.

[9] Ben Ghorbal A, Grine A, Eid M, El-kenawy ES. *Sustainable soil organic carbon prediction using machine learning and the ninja optimization algorithm. Frontiers in Environmental Science.* 2025 Aug 15;13:1630762.

[10] Hasan MA, Mahfuz S. *Soil pH Prediction Using Deep Learning: An Ensemble Approach. Procedia Computer Science.* 2025 Jan 1;265:293-300.

[11] Mukhamediev RI, Popova Y, Kuchin Y, Zaitseva E, Kalimoldayev A, Symagulov A, Levashenko V, Abdoldina F, Gopejenko V, Yakunin K, Muhamedijeva E. *Review of artificial intelligence and machine learning technologies: classification, restrictions, opportunities and challenges. Mathematics.* 2022 Jul 22;10(15):2552.

[12] Balabied SA, Eid HF. *Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. PeerJ Computer Science.* 2023 Nov 22;9:e1708.

[13] Wiens M, Verone‐Boyle A, Henscheid N, Podichetty JT, Burton J. A tutorial and use case example of the eXtreme gradient boosting (XGBoost) artificial intelligence algorithm for drug development applications. Clinical and Translational Science. 2025 Mar;18(3):e70172.