Jurnal Tekno Insentif | Vol. 19 | No. 2 | Desember 2025
DOI : https://doi.org/10.36787/jti.v19i2.2072 Halaman 197-212

Implementasi *Naïve Bayes* dan *Decision Tree* Untuk Klasifikasi Jenis Tanaman

RONI, ASRUL ABDULLAH, RACHMAT WAHID SALEH INSANI

Program Studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Muhammadiyah Pontianak
Email: ronn.7ex@gmail.com

ABSTRAK

Sektor pertanian berkontribusi penting bagi perekonomian Indonesia, namun pemilihan tanaman masih mengandalkan cara tradisional yang kurang efisien. Penelitian ini mengembangkan sistem klasifikasi tanaman berbasis parameter tanah dan iklim dengan algoritma *Naïve Bayes* serta *Decision Tree*. Proses penelitian mengikuti enam tahap *CRISP-DM*. Data diambil dari *Kaggle* dengan variabel nitrogen, fosfor, kalium, suhu, kelembapan, pH, dan curah hujan. Evaluasi memakai *Confusion Matrix* dan *Cross-Validation* dengan akurasi, presisi, *recall*, dan *F1-score*. Hasilnya, *Decision Tree* akurat pada data latih (97,95%) namun turun di data uji (91,57%), sedangkan *Naïve Bayes* lebih stabil (95,25%–95,32%) sehingga direkomendasikan karena hasil yang konsisten dan lebih dapat diandalkan. Perbedaan ini terjadi karena kompleksitas struktur *Decision Tree* membuatnya lebih rentan terhadap *overfitting*, sedangkan *Naïve Bayes* yang bersifat probabilistik lebih stabil terhadap variasi data.

Kata kunci: Pertanian, Klasifikasi Tanaman, *Naïve Bayes, Decision Tree, CRISP-DM*

ABSTRACT

The agricultural sector plays an important role in Indonesia's economy, yet crop selection still relies on traditional practices that are often inefficient. This study develops a crop classification system based on soil and climate parameters using the Naïve Bayes and Decision Tree algorithms. The research process follows the six stages of CRISP-DM. The dataset, obtained from Kaggle, includes nitrogen, phosphorus, potassium, temperature, humidity, soil pH, and rainfall. Evaluation was conducted with a Confusion Matrix and Cross-Validation using accuracy, precision, recall, and F1-score. Results indicate that Decision Tree achieved 97.95% accuracy on training data but decreased to 91.57% on testing data, while Naïve Bayes remained more stable (95.25%–95.32%), thus recommended for its consistent and more reliable performance. This difference occurs because the complexity of the Decision Tree structure makes it more prone to overfitting, while the probabilistic Naïve Bayes is more stable against data variations.

***Keywords**: Agriculture, Crop Classification, Naïve Bayes, Decision Tree, CRISP-DM.*

1. PENDAHULUAN

Pertanian memainkan peran yang sangat penting dalam perekonomian dan kehidupan sehari-hari masyarakat Indonesia, yang dikenal sebagai negara agraris **(Setiadi et al., 2020)**. Berdasarkan data dari Badan Pusat Statistik (BPS) untuk tahun 2023, sektor pertanian mengalami peningkatan sebesar 12,53% dibandingkan dengan PDB, menunjukkan bahwa mayoritas penduduk masih sangat mendukung sektor ini untuk memenuhi kebutuhan dasar mereka **(Kementerian Pertanian, 2023)**. Pertanian berperan penting dalam mendukung ketahanan pangan nasional. Namun demikian, permasalahan kekurangan pangan masih menjadi tantangan berkelanjutan, sebagaimana ditunjukkan dalam *Laporan Tahunan Statistik Ketahanan Pangan Indonesia* yang mencatat 4,5% penduduk mengalami kerawanan pangan **(Badan Pangan Nasional, 2023)**. Serta dalam *Analisis Komoditas Pangan Strategis* yang menyoroti distribusi pangan yang belum merata **(Kementerian Pertanian Republik Indonesia, 2023)**.

Kendala tersebut kerap ditemui di negara-negara berkembang, termasuk Indonesia. Pada praktiknya, sebagian besar petani masih mengandalkan pengetahuan turun-temurun dan pengalaman lokal untuk menentukan jenis tanaman yang akan ditanam. Pendekatan tradisional ini memiliki keterbatasan yang dapat mengakibatkan rendahnya produktivitas. Salah satu solusi untuk mengatasi masalah ini adalah dengan menerapkan pertanian presisi **(Gupta et al., 2023)**. Pertanian presisi mengintegrasikan strategi manajemen dan teknologi untuk mengoptimalkan penggunaan sumber daya sekaligus meminimalkan dampak negatif terhadap lingkungan **(Anguraj et al., 2021)**.

Dalam pertanian presisi, pemilihan jenis tanaman berdasarkan kondisi tanah dan iklim merupakan aspek yang sangat penting. Unsur-unsur seperti nitrogen (N), fosfor (P), kalium (K), suhu, kelembapan, pH tanah, dan curah hujan berperan penting dalam menentukan kesesuaian suatu tanaman **(Gupta et al., 2023)**. Pemanfaatan data terkait kondisi tanah dan iklim melalui metode pembelajaran mesin *(Machine Learning, ML)* memungkinkan rekomendasi tanaman yang sesuai dengan karakteristik lahan, sekaligus membantu memprediksi produktivitas serta mendeteksi potensi masalah dalam proses budidaya **(Durai & Shamili, 2022)**.

Berbagai penelitian sebelumnya telah menunjukkan bahwa teknologi *data mining* dan *machine learning* dapat meningkatkan produktivitas pertanian. Teknik-teknik ini dapat dibagi menjadi dua kategori besar, yakni metode statistik dan *algoritma machine learning*. Metode statistik bekerja dengan struktur data yang sudah diketahui sebelumnya, sedangkan *algoritma machine learning* belajar langsung dari *dataset* yang tersedia, sehingga mampu beradaptasi dan berkembang seiring waktu **(Motamedi & Villányi, 2024)**. Sebagai contoh, penelitian yang dilakukan oleh Sita Rani yang mengembangkan model pemilihan tanaman optimal menggunakan *machine learning* berbasis data cuaca dan parameter tanah. Hasil penelitian tersebut menunjukkan bahwa *machine learning* mampu memberikan efisiensi dan akurasi yang tinggi dalam proses pengambilan keputusan, sehingga potensial diterapkan untuk mendukung sektor pertanian di Indonesia **(Rani et al., 2023)**.

Algoritma *Naïve Bayes* dan *Decision Tree* digunakan dalam penelitian ini untuk mengklasifikasikan jenis tanaman berdasarkan unsur tanah dan iklim. Kedua algoritma ini memiliki kelebihan masing-masing. *Naïve Bayes* dikenal sederhana, cepat, dan efisien dalam mengolah data dengan pendekatan probabilistik **(Chen et al., 2021)**. Sementara itu, *Decision Tree* unggul dalam menghasilkan hasil klasifikasi yang mudah dipahami secara visual dan memiliki kemampuan menangani data dengan baik **(Tarumingkeng, 2025)**. Dibandingkan dengan algoritma lain seperti *Random Forest, Support Vector Machine (SVM)*, atau *Neural*

Network yang membutuhkan sumber daya komputasi lebih besar dan proses pelatihan yang lebih kompleks **(Lee et al., 2019)**. *Naïve Bayes* dan *Decision Tree* cocok digunakan pada aplikasi web yang ringan dan memiliki kemudahan dalam integrasi.

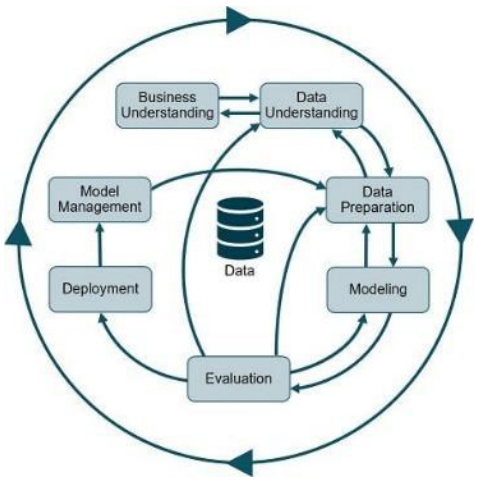
Tujuan dari penelitian ini adalah untuk menggunakan algoritma *Naïve Bayes* dan *Decision Tree* dalam klasifikasi jenis tanaman berdasarkan parameter tanah dan iklim. Penelitian ini juga mengevaluasi akurasi kedua algoritma tersebut untuk menentukan pendekatan mana yang lebih efektif. Selain itu, penelitian ini menghasilkan aplikasi klasifikasi berbasis web menggunakan *Streamlit* yang dapat mengintegrasikan model klasifikasi dan memberikan hasil prediksi interaktif, membantu pengguna atau pengguna lain dalam menyelesaikan tugas dengan lebih akurat dan efisien.

Perumusan masalah dalam penelitian ini adalah bagaimana membangun dan membandingkan model klasifikasi jenis tanaman berdasarkan parameter tanah dan iklim menggunakan algoritma *Naïve Bayes* dan *Decision Tree*, serta menentukan algoritma mana yang memberikan hasil paling akurat dan stabil untuk diimplementasikan dalam aplikasi berbasis web.

2. METODE

Penelitian ini menggunakan pendekatan kuantitatif, yang berfokus pada analisis data numerik, pengamatan objektif, dan analisis statistik sehingga temuan penelitian dapat digeneralisasikan **(Abdullah et al., 2021)**. Penelitian kuantitatif juga bertujuan menguji hipotesis melalui data yang terukur dan dianalisis menggunakan teknik statistik untuk memperoleh kesimpulan yang valid **(John W. Creswell & J. David Creswell, 2018)**.

Proses pengolahan data dalam penelitian ini mengikuti metode *CRISP-DM (Cross-Industry Standard Process for Data Mining)*. Kerangka ini terdiri atas enam tahapan penting, yakni *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation*, dan *Deployment* **(Schröer et al., 2021)**. Penerapan struktur tersebut memudahkan peneliti dalam mengorganisasi pengumpulan data dan melakukan analisis untuk klasifikasi tanaman berdasarkan kondisi tanah dan iklim.



Gambar 1. Tahapan metodologi *CRISP-DM* (Wehrstein, 2020)

Langkah-langkah *CRISP-DM* yang digunakan dalam penelitian ini dapat dijelaskan sebagai berikut:

2.1 *Business Understanding*

Tahap ini bertujuan memahami permasalahan serta menetapkan tujuan penelitian. Permasalahan utama yang diangkat adalah bagaimana membangun model klasifikasi jenis tanaman berdasarkan parameter tanah dan iklim dengan menggunakan algoritma *Naïve Bayes* dan *Decision Tree*. Tujuan penelitian difokuskan pada pengembangan model klasifikasi yang dapat memberikan rekomendasi tanaman sesuai dengan faktor lingkungan. Penelitian ini menggunakan variabel Nitrogen (N), Fosfor (P), Kalium (K), suhu, kelembapan, pH tanah, dan curah hujan, karena variabel tersebut memiliki pengaruh besar terhadap pertumbuhan tanaman **(Nugroho & Nasruddin, 2020)**.

2.2 *Data Understanding*

Data yang digunakan dalam penelitian ini berasal dari *Crop Yield Production dataset* yang tersedia di *Kaggle* **(Smith, 2020)** dengan jumlah 2.200 baris dan 8 atribut. Atribut tersebut mencakup faktor tanah (N, P, K, dan pH) serta faktor iklim (suhu, kelembapan, dan curah hujan), ditambah label jenis tanaman. *Dataset* ini dijadikan dasar untuk membangun model klasifikasi tanaman berdasarkan kondisi lingkungan. Struktur dataset ditunjukkan pada Tabel 1. Untuk membedakan parameter dalam dataset, penelitian ini merujuk pada *Petunjuk Teknis Evaluasi Lahan untuk Komoditas Pertanian* **(Djaenudin et al., 2011)**.

Tabel 1. *Sample dataset crop yield production*

N	P	K	*Temperature*	*Humidity*	pH	*Rainfall*	*Label*
98	30	47	25.34	47.06	6.23	331.79	Padi
93	59	48	27.16	67.61	5.55	248.61	Padi
109	31	31	27.43	65.41	6.70	292.21	Padi
116	38	40	24.09	52.20	5.99	235.24	Padi
118	50	43	28.00	63.12	6.96	220.73	Padi

2.3 *Data Preparation*

Dataset yang diperoleh dari *Kaggle* dibersihkan dan diproses agar siap digunakan pada pemodelan *machine learning*. Tahapan yang dilakukan meliputi pengecekan struktur data, penanganan *missing values* dan duplikasi, deteksi serta penghapusan *outlier*, penyeimbangan kelas dengan *SMOTE*, *encoding* variabel kategorikal, dan normalisasi fitur menggunakan *MinMaxScaler*. Pertama, dilakukan pengecekan struktur data untuk memastikan setiap atribut memiliki tipe data yang sesuai dengan karakteristik variabel yang digunakan. Hasil pemeriksaan menunjukkan bahwa dataset terdiri atas 2.200 baris dan 8 atribut, di mana fitur nitrogen (N), fosfor (P), dan kalium (K) bertipe *integer*, sedangkan suhu, kelembapan, pH tanah, dan curah hujan bertipe *float*, serta satu kolom label bertipe *object*.

Kedua, dilakukan pemeriksaan *missing values* untuk mendeteksi adanya nilai kosong pada setiap kolom. Berdasarkan hasil pengecekan, tidak ditemukan nilai yang hilang sehingga tidak diperlukan proses imputasi. Pemeriksaan duplikasi data juga dilakukan untuk menghindari pengaruh data ganda terhadap distribusi model, dan tidak ditemukan baris yang duplikat.

Tahap selanjutnya adalah deteksi dan penanganan *outlier* menggunakan metode *Interquartile Range (IQR)*. Nilai kuartil pertama (Q1) dan ketiga (Q3) digunakan untuk menghitung *IQR*, dengan batas bawah dan atas ditentukan oleh rumus $Q1 - 1{,}5\times IQR$ dan $Q3 + 1{,}5\times IQR$. Nilai di luar rentang ini dianggap sebagai *outlier* dan dihapus dari dataset. Pendekatan ini memastikan bahwa distribusi data tidak dipengaruhi oleh nilai ekstrem yang dapat menyebabkan bias pada proses pelatihan model.

Selanjutnya dilakukan analisis keseimbangan data (*data imbalance*) pada kolom label. Ketidakseimbangan distribusi kelas dapat menyebabkan model bias terhadap kelas mayoritas, sehingga digunakan metode *SMOTE (Synthetic Minority Oversampling Technique)* untuk menambah sampel sintetis pada kelas minoritas hingga distribusi antar kelas menjadi seimbang. Setelah data seimbang, dilakukan proses *encoding* terhadap variabel kategorikal menggunakan *LabelEncoder* dari pustaka *scikit-learn*. Proses ini mengonversi setiap kategori tanaman menjadi representasi numerik unik agar dapat dikenali oleh algoritma *machine learning*.

Tahap terakhir adalah normalisasi fitur menggunakan *MinMaxScaler* untuk mengubah skala nilai numerik setiap atribut ke dalam rentang [0,1]. Langkah ini memastikan bahwa seluruh fitur memiliki skala yang seragam sehingga model tidak terpengaruh oleh perbedaan skala antar variabel. Secara keseluruhan tahapan data preparation ini memastikan dataset berada dalam kondisi bersih, seimbang dan terstandarisasi, sehingga dapat digunakan secara optimal pada tahap pemodelan.

2.4 *Modelling*

Pemodelan dilakukan menggunakan dua algoritma, yaitu *Naïve Bayes* dan *Decision Tree*, untuk mengklasifikasikan jenis tanaman berdasarkan parameter tanah dan iklim. Data set dibagi menggunakan metode *train-test split* menjadi data latih (80%) dan data uji (20%). Model kemudian dilatih menggunakan data latih dan dievaluasi menggunakan data uji untuk menghasilkan prediksi yang kemudian dievaluasi menggunakan metrik klasifikasi.

2.4.1 Algoritma *Naïve Bayes*

Naïve Bayes adalah metode klasifikasi probabilistik yang menggunakan *teorema Bayes* dengan asumsi variabel-variabel independen. Algoritma ini menghitung kemungkinan suatu data termasuk ke dalam kelas tertentu berdasarkan nilai atribut yang dimiliki. Menurut **(Valentinus et al., 2023)** Teorema Bayes dapat dituliskan pada Persamaan (1).

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \tag{1}$$

Keterangan:

X : Data kelasnya tidak yang diketahui
H : hipotesis bahwa data X dalam kelas
$P(H|X)$: probabilitas hipotesis H diberikan data X (*posterior*)
$P(H)$: probabilitas awal hipotesis (*prior*)
$P(X|H)$: probabilitas data X muncul jika H benar (*likelihood*)
$H\ P\ (X)$: probabilitas total data X (*evidence*).

2.4.2 Algoritma *Decision Tree*

Decision Tree digunakan sebagai metode klasifikasi dengan struktur berbentuk pohon, di mana simpul mewakili atribut, sementara daun merepresentasikan kelas. Pada algoritma ID3, pemilihan atribut didasarkan pada nilai entropy yang menggambarkan tingkat ketidakpastian data. Menurut **(Sankaravadivel et al., 2023)** Nilai entropy dapat dihitung menggunakan Persamaan (2).

$$Entropy\ (S) = \sum_{i=1}^{n} p_i\ log_2(p_i) \tag{2}$$

Selanjutnya, atribut terbaik ditentukan melalui *information gain*, sebagaimana ditunjukkan pada Persamaan (3).

$$Gain(S,A) = Entropy(\mathrm{S}) - \sum_{i=1}^{n} - \frac{|Si|}{|S|} * Entropy(Si) \tag{3}$$

Keterangan:

S	:	himpunan data
n	:	jumlah partisi
p_{i}	:	proporsi data pada kelas ke-i
$\|Si\|$	:	banyaknya data pada partisi ke-i
$\|S\|$	:	total data
A	:	atribut pemisah

2.5 *Evaluation*

Tahap evaluasi dilakukan untuk menilai sejauh mana model mampu mengklasifikasikan jenis tanaman berdasarkan parameter tanah dan iklim. Dua metode digunakan, yaitu *Confusion Matrix* dan *Cross-Validation*. Melalui *Confusion Matrix*, diperoleh metrik akurasi, presisi, *recall*, dan *F1-score* yang memberikan gambaran detail terhadap performa model. Akurasi dihitung menggunakan Persamaan (4).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Presisi ditunjukkan pada Persamaan (5), *recall* pada Persamaan (6), dan *F1-score* pada Persamaan (7):

$$precision = \frac{TP}{TP + TN} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - Score = \frac{1}{\frac{1}{recall} + \frac{1}{precision}} \tag{7}$$

Keterangan:

TP	:	*True Positive*
TN	:	*True Negative*
FP	:	*False Positive*
FN	:	*False Negative*

Selain itu, *k-fold cross-validation* diterapkan untuk menguji konsistensi model dan mendeteksi potensi *overfitting*. Teknik ini menghitung rata-rata akurasi dari beberapa iterasi pelatihan dan pengujian, sehingga hasil evaluasi lebih stabil dan representatif.

Dalam penelitian ini digunakan beberapa variasi nilai *k* (3, 5, 7, 9, dan 10) untuk melihat pengaruh jumlah *fold* terhadap kestabilan model. Nilai *k = 10* dipilih sebagai acuan utama karena memberikan hasil yang paling stabil dengan standar deviasi terendah serta merupakan konfigurasi yang umum digunakan dalam evaluasi model *machine learning* **(Han et al., 2023)**. Pemilihan nilai ini juga mempertimbangkan ukuran dataset (2.200 sampel), di mana pembagian menjadi 10 *fold* masih memberikan proporsi data latih dan uji yang seimbang tanpa kehilangan representativitas data.

2.6 *Deployment*

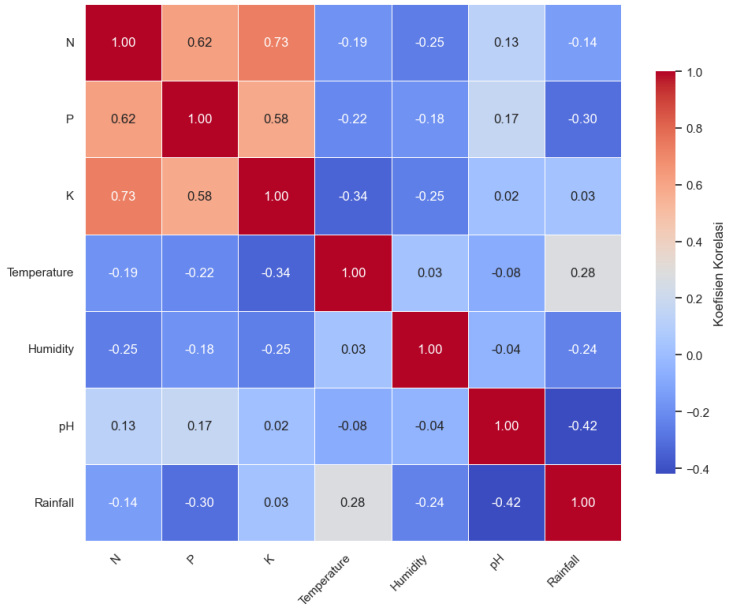
ada tahap ini, model klasifikasi yang telah dikembangkan diintegrasikan ke dalam aplikasi berbasis web sehingga pengguna dapat mengakses dan memanfaatkannya secara langsung. Model yang telah dilatih disimpan dalam format *file pickle* sehingga dapat dipanggil kembali saat diperlukan, kemudian diintegrasikan ke dalam sistem aplikasi. Aplikasi ini dirancang agar pengguna dapat memasukkan parameter tanah dan iklim melalui antarmuka interaktif, kemudian sistem secara otomatis menghasilkan prediksi jenis tanaman yang sesuai. Dengan demikian, model tidak hanya diuji pada tahap eksperimen, tetapi juga siap digunakan sebagai alat bantu pengambilan keputusan di bidang pertanian.

3. HASIL DAN PEMBAHASAN

Bagian ini menampilkan temuan penelitian yang diperoleh melalui tahapan eksplorasi data, *preprocessing*, penerapan model klasifikasi, evaluasi kinerja, serta implementasi model ke dalam aplikasi berbasis web. Hasil yang ditampilkan bertujuan untuk menunjukkan bagaimana data yang telah diolah dapat digunakan dalam membangun model klasifikasi jenis tanaman berdasarkan parameter tanah dan iklim, serta bagaimana model tersebut dapat dimanfaatkan secara langsung melalui aplikasi interaktif.

3.1 Hasil Eksplorasi Data

Eksplorasi data bertujuan untuk mendeskripsikan variabel penelitian, termasuk kandungan Nitrogen (N), Fosfor (P), Kalium (K), pH tanah, serta kondisi lingkungan berupa suhu, kelembapan, dan curah hujan. Analisis korelasi melalui *heatmap* menunjukkan adanya hubungan cukup kuat antar variabel, misalnya antara kelembaban dengan curah hujan serta suhu dengan pH tanah. Visualisasi ini memberikan gambaran relevansi variabel sebelum tahap pemodelan.



Gambar 2. *Heatmap* korelasi antar fitur numerik

Gambar 2 menampilkan *heatmap* korelasi antar fitur. Terlihat korelasi positif cukup tinggi antara Nitrogen (N) dengan Kalium (K) (0,73) dan Fosfor (P) (0,62), serta korelasi negatif antara pH dengan curah hujan (-0,42) dan Kalium dengan suhu (-0,34). Sementara itu,

beberapa fitur lain menunjukkan korelasi rendah, yang menandakan hubungan lemah. Visualisasi ini membantu memahami keterkaitan antar variabel sebelum tahap pemodelan.

3.2 Hasil *Preprocessing Data*

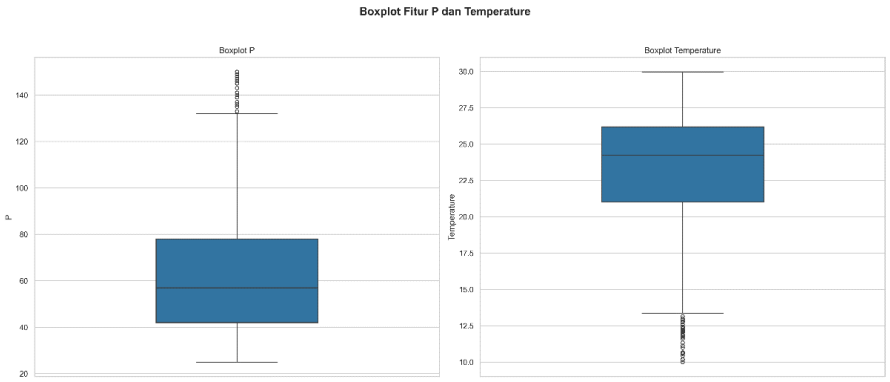
Tahap *preprocessing* bertujuan untuk menyiapkan data agar siap digunakan dalam pemodelan. Langkah-langkah yang dilakukan meliputi pemeriksaan kelengkapan data, pembersihan data, penanganan *outlier*, penyeimbangan kelas, normalisasi, serta *label encoding*.

3.2.1 Pengecekan *Missing Value* dan Duplikasi

Dataset telah diperiksa dan tidak ditemukan nilai hilang maupun duplikasi, sehingga dapat langsung digunakan untuk pemodelan.

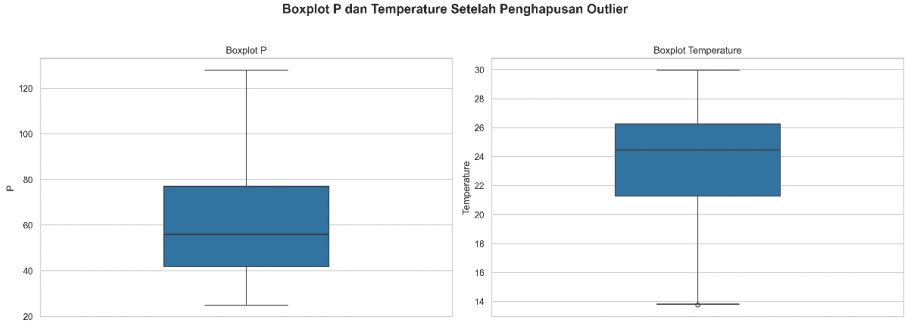
3.2.2 Pengecekan dan Penanganan *Outlier*

Outlier adalah data dengan nilai ekstrem yang menyimpang dari mayoritas data. Kehadirannya dapat mengganggu proses pelatihan model karena dapat mengubah distribusi data dan memengaruhi hasil klasifikasi Meskipun beberapa *outlier* dapat mencerminkan kondisi aktual di lapangan, dalam penelitian ini seluruh *outlier* dihapus menggunakan metode *Interquartile Range (IQR)* untuk menjaga kestabilan model.



Gambar 3. *Outlier* fitur p dan *temperature*

Gambar 3 memperlihatkan hasil pengecekan *outlier* menggunakan *boxplot* pada fitur Fosfor (P) dan *Temperature*. Tampak adanya titik data yang berada jauh di luar *whisker*, menandakan keberadaan *outlier*. Fitur lainnya (N, K, *Humidity*, pH, *Rainfall*) tidak menunjukkan *outlier* signifikan.

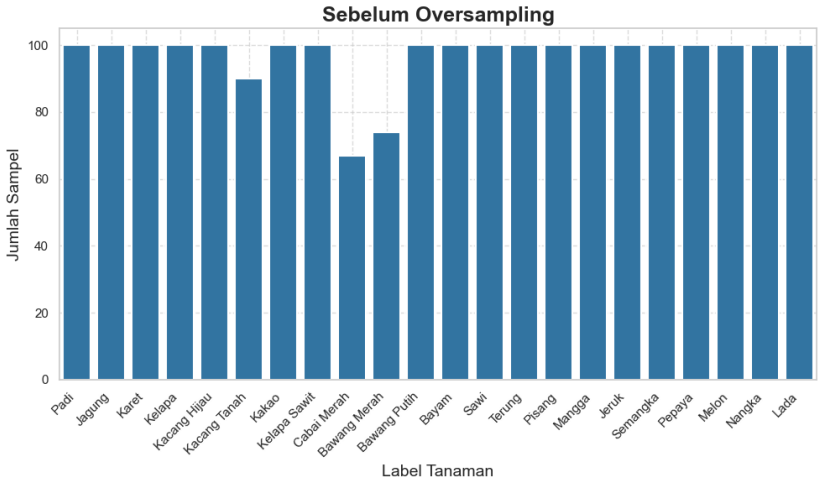


Gambar 4. *Boxplot* fitur p dan *temperature* tanpa *outlier*

Gambar 4 menampilkan *boxplot* setelah penanganan *outlier* menggunakan metode *IQR (Interquartile Range)*. Data yang berada di luar rentang (Q1 – 1.5*IQR) dan (Q3 + 1.5*IQR) dihapus. Hasilnya, titik ekstrem yang sebelumnya terlihat sudah tidak muncul lagi, sehingga data menjadi lebih rapi dan siap untuk digunakan pada tahap pemodelan.

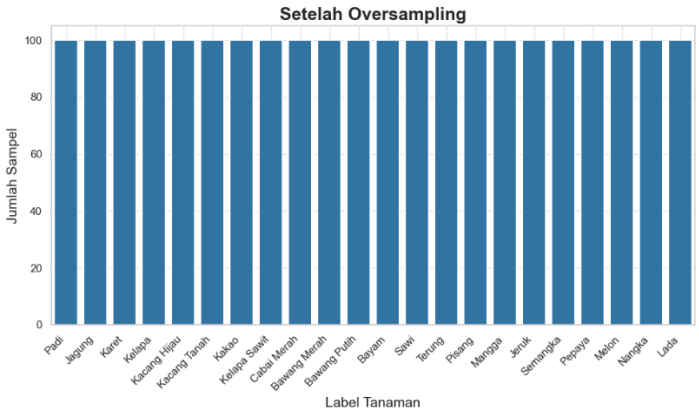
3.2.2 *Data Imbalance*

Ketidakseimbangan data muncul jika jumlah sampel antar kelas berbeda signifikan, yang dapat menyebabkan model bias terhadap kelas mayoritas dan mengabaikan kelas minoritas. Oleh karena itu, perlu dilakukan analisis distribusi label tanaman sebelum proses pemodelan.



Gambar 5. Distribusi label sebelum *oversampling*

Gambar 5 menunjukkan distribusi label tanaman sebelum dilakukan *oversampling*. Sebagian besar label berjumlah seimbang (100 sampel), namun terdapat beberapa kelas dengan jumlah lebih sedikit, seperti Cabai Merah (67 sampel), Bawang Putih (73 sampel), dan Kacang Tanah (90 sampel). Ketidakseimbangan ini berpotensi menurunkan performa klasifikasi.



Gambar 6. Distribusi label setelah *oversampling*

Untuk mengatasi masalah tersebut, digunakan metode *SMOTE (Synthetic Minority Over-sampling Technique)* yang menambahkan sampel sintetis pada kelas minoritas. Gambar 6. memperlihatkan distribusi label setelah *oversampling*, di mana seluruh kelas sudah seimbang. Dengan demikian, *dataset* menjadi lebih representatif dan siap digunakan pada tahap pemodelan.

3.2.3 *Normalisasi Data*
Data numerik dinormalisasi menggunakan *Min-Max Scaling* sehingga semua fitur berada dalam rentang [0,1]. Langkah ini bertujuan untuk menghindari dominasi fitur dengan skala besar dan meningkatkan performa algoritma yang sensitif terhadap skala, seperti *Naïve Bayes* dan *Decision Tree*.

3.2.4 Encoding Data Kategorikal
Pada tahap ini, fitur *Label* yang berisi jenis tanaman dikonversi dari bentuk kategorikal menjadi numerik menggunakan teknik *Label Encoding*. Setiap jenis tanaman diberikan representasi angka unik agar dapat diproses oleh algoritma *machine learning*. Tabel 2 menampilkan sebagian hasil *encoding*, sementara seluruh label tanaman dalam *dataset* telah berhasil diubah ke dalam format numerik.

Tabel 2. Contoh hasil *label encoding*

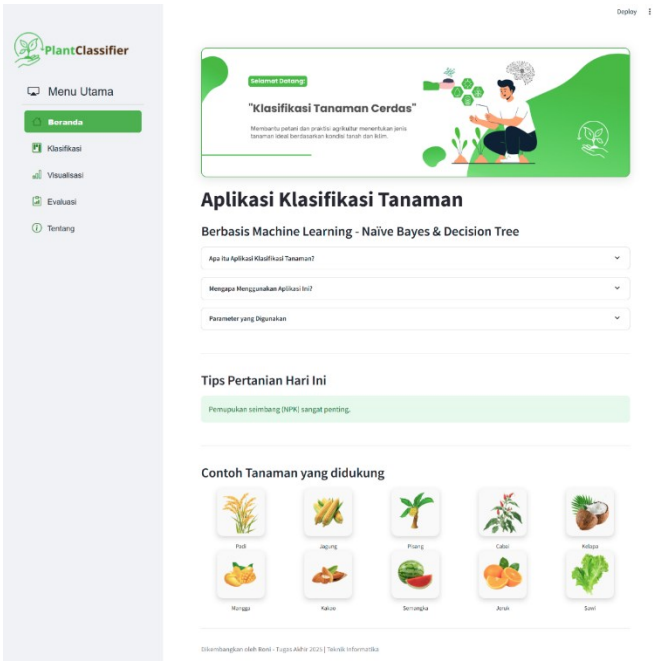
No.	Label Asli	Label *Encode*
1	Padi	16
2	Jagung	4
3	Karet	9
4	Kelapa	10
5	Kacang Hijau	6

3.3 Hasil Implementasi Model Klasifikasi
Model klasifikasi dikembangkan setelah eksplorasi dan *preprocessing*, menggunakan *Naïve Bayes* dan *Decision Tree*. Dataset dipisahkan menjadi 80% untuk pelatihan dan 20% untuk pengujian, dengan stratifikasi label agar kelas tetap seimbang.

Model *Naïve Bayes* dibangun dengan pendekatan *Gaussian Naïve Bayes* yang mengasumsikan distribusi *Gaussian* pada setiap fitur numerik. Sementara itu, model Decision Tree dikembangkan dengan proses *hyperparameter tuning* menggunakan *Grid Search* dan validasi silang (*cross-validation*) untuk memperoleh parameter terbaik. Kedua model yang telah dilatih kemudian digunakan untuk melakukan prediksi jenis tanaman berdasarkan parameter tanah dan iklim.

3.4 Hasil Integrasi Model Ke Aplikasi Web
Setelah model *Naïve Bayes* dan *Decision Tree* dilatih serta dievaluasi, langkah berikutnya adalah integrasi ke aplikasi web berbasis *Streamlit*. Model disimpan dalam format .pkl menggunakan *pickle*, kemudian dimuat kembali dengan *pickle.load()*. Cara ini memungkinkan klasifikasi dilakukan secara cepat dan interaktif melalui antarmuka web, tanpa perlu pelatihan atau komputasi ulang.

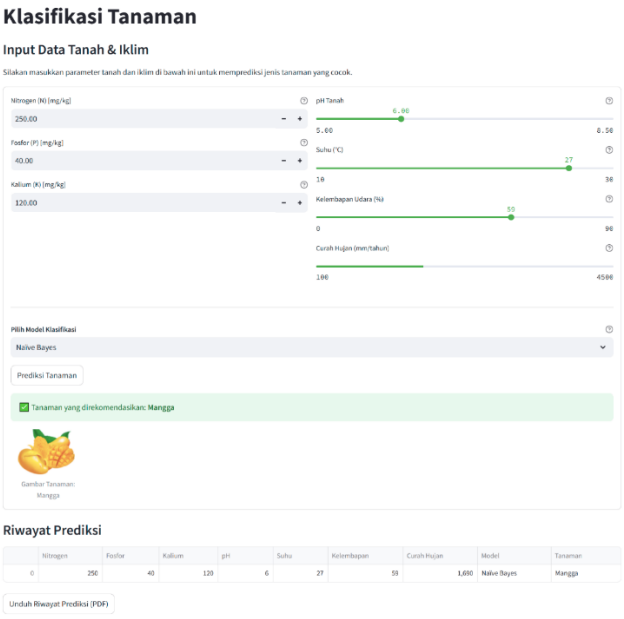
3.4.1 Halaman Beranda
Halaman beranda merupakan tampilan utama aplikasi klasifikasi tanaman. Di halaman ini disediakan informasi mengenai tujuan aplikasi, parameter tanah dan iklim yang digunakan, serta daftar jenis tanaman yang dapat diklasifikasikan.



Gambar 7. Halaman beranda *website*

3.4.2 Halaman Klasifikasi

Halaman klasifikasi digunakan pengguna untuk memasukkan data tanah dan iklim serta memilih model klasifikasi untuk mendapatkan rekomendasi jenis tanaman.



Gambar 8. Halaman klasifikasi

3.5 Hasil Evaluasi Model

Pada tahap ini dilakukan evaluasi terhadap model *Naïve Bayes* dan *Decision Tree* untuk mengetahui kinerja serta kemampuan generalisasi. Evaluasi dilakukan menggunakan *Confusion Matrix* dan *k-Fold Cross-Validation*.

3.5.1 Hasil Evaluasi *Naïve Bayes*

Model *Naïve Bayes* diuji menggunakan data latih dan uji. Tabel 3 menyajikan hasil metrik evaluasi.

Tabel 3. Hasil *confusion matrix naïve bayes*

No.	Metrik	Data Latih	Data Uji
1	Akurasi	95.25%	95.32%
2	Presisi	95.56%	95.43%
3	Recall	95.25%	95.32%
4	F1-Score	95.25%	95.30%

Model menunjukkan performa tinggi dan konsisten, dengan selisih kecil antara data latih dan uji. Selain itu, evaluasi *cross-validation* pada berbagai nilai K (3, 5, 7, 9, 10) menghasilkan akurasi stabil dengan rentang perbedaan.

Tabel 4. Hasil *cross-validation naïve bayes*

Fold	Akurasi
3	95.05%
5	95.12%
7	94.98%
9	95.01%
10	95.07%

Rata-rata akurasi sebesar 95.05% dengan standar deviasi rendah, yang menunjukkan stabilitas model.

3.5.2 Hasil Evaluasi *Decision Tree*

Model *Decision Tree* dievaluasi dengan pendekatan serupa. Hasil metrik evaluasi ditunjukkan pada Tabel 5.

Tabel 5. Hasil *confusion matrix decision tree*

No.	Metrik	Data Latih	Data Uji
1	Akurasi	97.95%	91.57%
2	Presisi	98.02%	92.14%
3	Recall	97.95%	91.57%
4	F1-Score	97.97%	91.76%

Hasil menunjukkan performa sangat tinggi pada data latih, namun mengalami penurunan pada data uji, yang mengindikasikan potensi *overfitting*.

Tabel 6. Hasil *cross-validation decision tree*

Fold	Akurasi
3	92.31%
5	91.74%
7	92.12%
9	92.48%
10	92.87%

Rata-rata akurasi sebesar 92.30%, lebih rendah dan fluktuatif dibandingkan *Naïve Bayes*

3.5.3 Perbandingan Kinerja Model

Perbandingan kinerja dilakukan untuk menilai performa *Naïve Bayes* dan *Decision Tree* pada data latih, data uji, serta melalui teknik *cross-validation*.

Tabel 7. Perbandingan metrik evaluasi *naïve bayes* dan *decision tree*

Model	Data	Akurasi	Presisi	*Recall*	*F1-Score*
Naïve Bayes	Latih	95.25%	95.56%	95.25%	95.25%
Naïve Bayes	Uji	95.32%	95.43%	95.32%	95.30%
Decision Tree	Latih	97.95%	98.02%	97.95%	97.97%
Decision Tree	Uji	91.57%	92.14%	91.57%	91.76%

Perbedaan hasil antara kedua model ini dapat dijelaskan oleh karakteristik algoritmanya. *Decision Tree* memiliki kemampuan belajar yang tinggi terhadap pola data latih, tetapi hal tersebut juga membuatnya cenderung mengalami *overfitting*, yaitu terlalu menyesuaikan diri dengan data pelatihan sehingga performanya menurun pada data uji. Sebaliknya, *Naïve Bayes* menggunakan pendekatan probabilistik dengan asumsi independensi antar fitur, sehingga menghasilkan model yang lebih sederhana dan general terhadap data baru. Hal inilah yang menyebabkan performa *Naïve Bayes* lebih stabil pada data latih maupun uji.

Untuk menguji stabilitas model, dilakukan *cross-validation*.

Tabel 8. Hasil *cross-validation*

Model	Rata-rata Akurasi CV
Naïve Bayes	95.05%
Decision Tree	92.30%

Berdasarkan hasil ini, dapat disimpulkan bahwa meskipun *Decision Tree* unggul pada data latih, model tersebut cenderung *overfitting* dan kurang generalisasi pada data baru. Sebaliknya, *Naïve Bayes* mempertahankan performa yang lebih konsisten pada data latih, data uji, maupun validasi silang, sehingga lebih andal untuk klasifikasi jenis tanaman.

4. KESIMPULAN

Algoritma *Decision Tree* menunjukkan performa lebih tinggi pada data latih, namun mengalami penurunan signifikan pada data uji sehingga mengindikasikan *overfitting*. Sebaliknya, algoritma *Naïve Bayes* tampil lebih stabil dengan akurasi konsisten di atas 95% pada data latih maupun uji, karena metode ini menggunakan pendekatan probabilistik sederhana dengan asumsi independensi antar fitur sehingga tidak mudah terpengaruh oleh variasi data. Pendekatan ini membuat Naïve Bayes mampu melakukan generalisasi lebih baik dibandingkan Decision Tree yang cenderung menyesuaikan diri secara berlebihan terhadap data pelatihan. Hasil *cross-validation* memperkuat temuan tersebut, di mana *Naïve Bayes* memberikan akurasi lebih stabil dan andal dibandingkan *Decision Tree*. Aplikasi berbasis web yang dibangun dengan *Streamlit* juga berhasil mengintegrasikan kedua model secara interaktif sebagai sistem pendukung keputusan. Dengan demikian, tujuan penelitian terkait implementasi algoritma, evaluasi performa, dan pengembangan aplikasi web telah tercapai secara menyeluruh.

UCAPAN TERIMA KASIH

Dengan penuh rasa syukur kepada Tuhan Yang Maha Esa, penulis mengucapkan terima kasih kepada Universitas Muhammadiyah Pontianak, khususnya Fakultas Teknik dan Ilmu Komputer

serta Program Studi Teknik Informatika, atas dukungan dan fasilitas yang diberikan selama penelitian ini. Ucapan terima kasih juga ditujukan kepada pembimbing penelitian, Bapak Asrul Abdullah, S.Kom., M.Cs., dan Bapak Rachmat Wahid Saleh Insani, S.Kom., M.Cs., atas bimbingan, arahan, dan saran yang sangat berarti. Penulis juga menyampaikan penghargaan kepada seluruh dosen dan staf Program Studi Teknik Informatika, serta orang tua dan keluarga yang selalu memberikan doa, dukungan, dan motivasi selama proses penelitian.

DAFTAR RUJUKAN

Abdullah, K., Jannah, M., Aiman, U., Hasda, S., Fadilla, Z., Taqwin, Masita, Ardiawan, K. N., & Sari, M. E. (2021). *Metodologi Penelitian Kuantitatif*. Yayasan Penerbit Muhammad Zaini.

Anguraj, K. A., Thiyaneswaran, B. B., Megashree, G. C., Preetha Shri, J. G., Navya, S. E., & Jayanthi, J. F. (2021). Crop Recommendation on Analyzing Soil Using Machine Learning. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(6), 1784–1791. https://turcomat.org/index.php/turkbilmat/article/view/4033

Badan Pangan Nasional. (2023). *Laporan Tahunan NFA 2023: Statistik Ketahanan Pangan Indonesia*. https://badanpangan.go.id/storage/app/media/2024/LAPORAN%20TAHUNAN%202024/LAPORAN%20TAHUNAN%20NFA%202023.pdf

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing*, *2021*(1), 30. https://doi.org/10.1186/s13634-021-00742-6

Djaenudin, D., Marwan, H., Subagjo, H., & Hidayat, A. (2011). *Petunjuk Teknis Evaluasi Lahan untuk Komoditas Pertanian*. Balai Besar Penelitian dan Pengembangan Sumberdaya Lahan Pertanian, Badan Penelitian dan Pengembangan Pertanian.

Durai, S. K. S., & Shamili, M. D. (2022). Smart farming using Machine Learning and Deep Learning techniques. *Decision Analytics Journal*, *3*, 100041. https://doi.org/10.1016/j.dajour.2022.100041

Gupta, S., Chatterjee, P., Rastogi, R., & Gonzalez, E. D. R. S. (2023). A Delphi fuzzy analytic hierarchy process framework for criteria classification and prioritization in food supply chains under uncertainty. *Decision Analytics Journal*, *7*, 1–4. https://doi.org/10.1016/J.DAJOUR.2023.100217

John W. Creswell, & J. David Creswell. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (5th edition). SAGE Publications.

Kementerian Pertanian. (2023). Analisis PDB sektor pertanian tahun 2023. In S. S. Wahyuningsih (Ed.), *Pusat Data dan Sistem Informasi Pertanian* (Mas'ud, S.E., M.Si). Kementerian Pertanian.

Kementerian Pertanian Republik Indonesia. (2023). *Analisis Komoditas Pangan Strategis 2023*. https://satudata.pertanian.go.id/assets/docs/publikasi/Analisis_Komoditas_Pangn_Strategis_2023-gab-ttd.pdf

Lee, S. H., Maenner, M. J., & Heilig, C. M. (2019). A comparison of machine learning algorithms for the surveillance of autism spectrum disorder. *PLOS ONE*, *14*(9), e0222907. https://doi.org/10.1371/journal.pone.0222907

Motamedi, B., & Villányi, B. (2024). A predictive analytics model with Bayesian-Optimized Ensemble Decision Trees for enhanced crop recommendation. *Decision Analytics Journal*, *12*, 100516. https://doi.org/10.1016/j.dajour.2024.100516

Nugroho, A. R., & Nasruddin. (2020). *Buku Ajar Geografi Tanah* (1st ed.). Program Studi Geografi Fakultas Ilmu Sosial dan Ilmu Politik niversitas Lambung Mangkurat .

Rani, S., Mishra, A. K., Kataria, A., Mallik, S., & Qin, H. (2023). Machine Learning-Based Optimal Crop Selection System in Smart Agriculture. *Scientific Reports*, *13*(1), 15997. https://doi.org/10.1038/s41598-023-42356-y

Sankaravadivel, V., Thalavaipillai, S., Rajeswar, S., & Ramlingam, P. (2023). Feature based analysis of endometriosis using machine learning. *Indonesian Journal of Electrical Engineering and Computer Science*, *29*(3), 1700. https://doi.org/10.11591/ijeecs.v29.i3.pp1700-1707

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Setiadi, T., Tarmuji, A., Fadlil, A., Noviyanto, F., Hardianto, H., & Wibowo, M. (2020). Implementation Of Naïve Bayes Method In Food Crops Planting Recommendation. *International Journal of Scientific & Technology Research*, *9*, 2. www.ijstr.org

Smith, K. (2020). *Crop Yield Production*. Kaggle. https://www.kaggle.com/datasets/kevinsmith94624/crop-yield-production

Tarumingkeng, R. C. (2025). *Decision Tree (Pohon Keputusan)*. Rudyct.com. https://rudyct.com/ab/Decision.Tree%28Pohon.Keputusan%29.pdf

Valentinus, F., Sujono, F., Ariansyah, I., & Capah, D. A. H. (2023). Implementation Of Data Mining With Classification And Forecasting Method Use Model Gaussian Naïve Bayes For Building Store (Studi Case: Tb Sinar Jaya). *Jurnal Teknik Informatika (Jutif)*, *4*(2), 413–420. https://doi.org/10.52436/1.jutif.2023.4.2.701

Wehrstein, L. (2020). *CRISP-DM ready for Machine Learning Projects*. https://towardsdatascience.com/crisp-dm-ready-for-machine-learning-projects-2aad9172056a/