Nama : Eko Putra Nugraha

NIM : 1103213212

BAB 1

Kode yang diimplementasikan mencerminkan langkah-langkah fundamental dalam membangun dan mengevaluasi model pembelajaran mesin, sebagaimana dijelaskan dalam **Bab 1 - Introduction** dari buku "Introduction to Machine Learning with Python." Dimulai dengan memuat dataset Iris, data diproses dengan membagi menjadi data latih dan data uji untuk memastikan evaluasi model yang adil dan objektif. Visualisasi scatter matrix memberikan wawasan awal tentang distribusi data dan relevansi fitur, membantu mengidentifikasi fitur yang paling informatif untuk klasifikasi.

Selanjutnya, model K-Nearest Neighbors (KNN) digunakan sebagai algoritma klasifikasi untuk memprediksi spesies bunga berdasarkan fitur yang diberikan. Model dilatih menggunakan data latih, kemudian diuji pada data baru dan data uji. Hasil evaluasi menunjukkan bahwa model memiliki akurasi tinggi (97%) pada data uji, yang menunjukkan kemampuan generalisasi model yang baik. Keseluruhan proses ini menyoroti pendekatan sistematis dalam membangun model pembelajaran mesin, mulai dari eksplorasi data hingga evaluasi, yang menjadi dasar penting untuk tugas pembelajaran mesin yang lebih kompleks.

BAB 2

Kode pada **Bab 2 - Supervised Learning** menunjukkan implementasi dasar pembelajaran terawasi, mencakup tugas klasifikasi dan regresi. Dalam klasifikasi, dataset simulasi dibuat untuk membedakan dua kelas, kemudian model **K-Nearest Neighbors (KNN)** dilatih dan diuji. Hasilnya menunjukkan akurasi yang baik, dengan visualisasi yang menggambarkan pembagian kelas berdasarkan fitur. Sementara itu, untuk tugas regresi, dataset linier dengan noise digunakan untuk melatih model **Linear Regression**. Model mampu memprediksi target dengan cukup baik, sebagaimana ditunjukkan oleh grafik prediksi regresi dan nilai **Mean Squared Error (MSE)**.

Bab ini menekankan pentingnya langkah-langkah mendasar dalam supervised learning, termasuk pembuatan dataset, pembagian data latih dan uji, pelatihan model, evaluasi, serta visualisasi hasil. Klasifikasi menggunakan KNN menunjukkan bagaimana model non-parametrik bekerja untuk membedakan kelas, sementara regresi linier memperlihatkan hubungan linier antara fitur dan target dengan keberadaan noise. Bab ini memberikan landasan kuat untuk memahami cara kerja algoritma supervised learning dalam skenario praktis.

BAB 3

Kode pada **Bab 3 - Unsupervised Learning** memperkenalkan metode clustering menggunakan algoritma **KMeans**, serta mengevaluasi kualitas kluster dengan **Silhouette Score**. Dataset simulasi yang digunakan terdiri dari data yang tidak berlabel, dan algoritma KMeans mampu membagi data ke dalam kluster yang berbeda berdasarkan pola distribusi. Visualisasi hasil clustering menunjukkan pemisahan kluster yang baik, dengan setiap kluster memiliki pusat yang dihitung oleh algoritma.

Evaluasi menggunakan Silhouette Score membantu menentukan kualitas clustering berdasarkan jumlah kluster yang dipilih. Hasil menunjukkan bahwa jumlah kluster memengaruhi kualitas clustering, dengan nilai Silhouette Score tertinggi tercapai pada tiga kluster. Hal ini menunjukkan pentingnya pemilihan jumlah kluster yang tepat untuk mencapai pemisahan yang optimal. Bab ini menggambarkan proses utama dalam unsupervised learning, yaitu menemukan struktur tersembunyi dalam data tanpa label.

BAB 4

Kode pada **Bab 4 - Representing Data and Engineering Features** menunjukkan proses penting dalam **feature engineering** dan **representasi data** yang diperlukan untuk analisis dan pelatihan model pembelajaran mesin. Dimulai dari pembuatan **DataFrame**, data kategorikal seperti City diubah menjadi representasi numerik melalui teknik **One-Hot Encoding** untuk memastikan kompatibilitas dengan algoritma pembelajaran mesin. Selanjutnya, proses **scaling** menggunakan Min-Max Scaling dilakukan untuk menyelaraskan skala fitur numerik, memastikan bahwa semua fitur diperlakukan secara seimbang oleh model.

Fitur baru juga dibuat melalui **interaksi fitur** dan **transformasi polinomial**, seperti Temp_Squared dan Temp_Rainfall, yang menambahkan kompleksitas dan dimensi baru ke dataset. Visualisasi scatter plot antara Temperature dan Rainfall memberikan wawasan awal tentang hubungan antar fitur dan membantu memahami distribusi data. Secara keseluruhan, bab ini menekankan pentingnya representasi data yang baik, transformasi fitur, dan visualisasi sebagai langkah fundamental dalam pipeline pembelajaran mesin, yang bertujuan untuk meningkatkan kualitas dan prediktabilitas model.

BAB 5

Kode pada **Bab 5 - Model Evaluation** menyoroti pentingnya mengevaluasi kinerja model secara menyeluruh dengan menggunakan teknik seperti **Cross-Validation**, **Hyperparameter Tuning**, dan evaluasi pada data uji. Proses **Cross-Validation** memastikan model memiliki kemampuan generalisasi yang baik, dengan rata-rata skor validasi yang menunjukkan kinerja model di berbagai subset data latih. Hyperparameter Tuning dengan **Grid Search** mengoptimalkan parameter model untuk menghasilkan performa terbaik, sementara evaluasi akhir menggunakan **Confusion Matrix** dan **Classification Report** memberikan metrik detail tentang kemampuan model pada data uji.

Evaluasi menunjukkan bahwa model Random Forest dengan hyperparameter terbaik memberikan kinerja sempurna pada dataset Iris, dengan akurasi, precision, recall, dan F1-score mencapai nilai **1.00**. Hal ini menunjukkan bahwa model tidak hanya mampu mempelajari pola dalam data latih dengan baik, tetapi juga mampu menggeneralisasi dengan sangat baik pada data yang belum pernah dilihat sebelumnya. Bab ini menekankan pentingnya menggunakan teknik evaluasi yang robust dan menyeluruh untuk memastikan kinerja model yang optimal dan dapat diandalkan.

BAB 6

Kode pada **Bab 6 - Model Evaluation and Improvement** menunjukkan pentingnya **Hyperparameter Tuning** menggunakan **Grid Search** untuk meningkatkan performa model. Dalam eksperimen ini, Grid Search diterapkan pada pipeline yang terdiri dari **StandardScaler** untuk normalisasi data dan **SVM** dengan kernel linear. Proses ini memastikan bahwa model mengoptimalkan parameter seperti C, gamma, dan kernel untuk mencapai kinerja terbaik. Dengan menggunakan Grid Search, model SVM mencapai skor terbaik pada parameter C=10, gamma=0.01, dan kernel='linear', yang memberikan hasil evaluasi yang sangat baik.

Hasil evaluasi pada data uji menunjukkan bahwa model berhasil memprediksi dengan sangat baik, meskipun ada sedikit penurunan kinerja pada kelas tertentu. **Confusion Matrix** dan **Classification Report** menunjukkan bahwa model mencapai **1.00** untuk beberapa kelas, dengan sedikit penurunan pada kelas lain, yang mengindikasikan bahwa model dapat lebih baik lagi dengan dataset yang lebih beragam atau pendekatan yang lebih cermat. Bab ini menekankan bahwa untuk mencapai model

yang optimal, perlu dilakukan tuning hyperparameter yang sistematis, dan Grid Search adalah metode yang efektif untuk mencapainya.

BAB 7

Eksperimen yang dilakukan pada Bab bertujuan untuk menggunakan bigram sebagai representasi teks, di mana fitur yang diambil adalah pasangan kata berturut-turut dalam teks. Pendekatan ini dirancang untuk menangkap hubungan kontekstual antara kata-kata yang tidak dapat ditangkap oleh unigram (kata tunggal). Dalam implementasi ini, **TfidfVectorizer** diatur untuk menghasilkan bigram sebagai fitur, dan model **Multinomial Naive Bayes** digunakan untuk melatih dan menguji data yang telah direpresentasikan. Hasil evaluasi menunjukkan bahwa penggunaan bigram tidak memberikan peningkatan yang signifikan dalam hal metrik evaluasi seperti precision, recall, dan f1-score pada dataset sederhana ini.

Matriks kebingungan memperlihatkan bahwa kesalahan klasifikasi masih terjadi pada data uji, di mana satu instance dari kelas aktual salah diklasifikasikan. Hal ini menunjukkan bahwa meskipun bigram memberikan konteks tambahan, kompleksitas representasi harus disesuaikan dengan karakteristik dataset. Untuk dataset sederhana seperti ini, bigram mungkin tidak menambah informasi yang cukup signifikan untuk meningkatkan performa model. Namun, dalam dataset yang lebih kompleks atau teks dengan struktur yang lebih panjang, penggunaan bigram dapat menjadi strategi yang lebih efektif. Kesimpulannya, pemilihan representasi teks harus dipertimbangkan berdasarkan kebutuhan spesifik analisis dan sifat dari data yang digunakan.

BAB 8

Kesimpulan dari eksperimen pada Bab 8 menunjukkan pentingnya penyimpanan model dengan metadata tambahan untuk mendukung manajemen model secara lebih terstruktur. Dengan menyertakan informasi seperti deskripsi model, versi, dan tanggal pembuatan, metadata membantu dokumentasi yang lebih baik untuk pengelolaan model, terutama dalam proyek jangka panjang atau kolaboratif. Teknik ini mempermudah pelacakan model yang digunakan, termasuk konteks spesifik atau pengaturan yang relevan, tanpa perlu membuka ulang kode atau pengaturan sebelumnya.

Selain itu, hasil eksperimen menunjukkan bahwa penyimpanan model dengan metadata tidak memengaruhi performa model, dengan akurasi tetap 100% setelah model dimuat dan diuji ulang. Hal ini memastikan bahwa pengelolaan tambahan melalui metadata tidak mengorbankan fungsi utama model. Pendekatan ini sangat relevan untuk pipeline pembelajaran mesin yang kompleks, di mana pengelolaan versi model dan dokumentasi yang jelas sangat diperlukan untuk keberlanjutan dan kolaborasi proyek. Strategi ini menunjukkan praktik terbaik dalam pengembangan model yang dapat digunakan ulang secara efisien.