

Banana Quality ML Analysis

1. Pendahuluan

Proyek Banana Quality ML Analysis bertujuan untuk membangun model machine learning yang dapat mengklasifikasikan kualitas pisang berdasarkan fitur-fitur tertentu. Menggunakan dataset yang terdiri dari variabel-variabel numerik dan metode Random Forest Classifier, proyek ini berupaya membantu dalam prediksi kualitas pisang, yang dapat digunakan oleh produsen atau pengecer untuk menentukan kelayakan produk mereka.

2. Tujuan Proyek

Tujuan utama dari proyek ini adalah untuk:

1. Mengembangkan model machine learning yang mampu mengklasifikasikan kualitas pisang (baik atau buruk).
2. Memahami faktor-faktor kunci yang memengaruhi kualitas pisang dengan mengeksplorasi korelasi antar variabel.
3. Meningkatkan akurasi klasifikasi dengan menggunakan algoritma Random Forest dan evaluasi model melalui akurasi dan laporan klasifikasi.

3. Metodologi

Proses yang diikuti dalam proyek ini melibatkan beberapa tahapan penting:

- a. Pengumpulan dan Pembersihan Data
 - Data kualitas pisang dimuat dari file CSV.
 - Pembersihan data dilakukan dengan menghapus nilai yang hilang (missing values) dan memastikan data siap untuk diproses.
- b. Preprocessing Data
 - Setelah data bersih, variabel fitur dan target dipisahkan.
 - Scaling data dilakukan menggunakan StandardScaler untuk memastikan bahwa semua fitur numerik berada dalam skala yang sama, yang penting untuk algoritma seperti Random Forest.
- c. Pembagian Data
 - Data dibagi menjadi training set (80%) dan testing set (20%) menggunakan `train_test_split`. Ini membantu melatih model pada sebagian data dan menguji kinerja pada data yang belum pernah dilihat sebelumnya.
- d. Pelatihan Model

- Model Random Forest Classifier dilatih menggunakan data training. Model ini dipilih karena kemampuannya menangani data yang kompleks dan interpretasi hasil yang baik.

e. Evaluasi Model

- Setelah model dilatih, prediksi dilakukan pada test set dan hasilnya dievaluasi menggunakan metrik akurasi dan classification report. Metrik ini memberikan informasi tentang seberapa baik model dapat memprediksi kualitas pisang.

f. Visualisasi dan Analisis Data

- Analisis distribusi data dan korelasi antar variabel dilakukan untuk memahami lebih baik hubungan antar fitur dan target.
- Heatmap korelasi antar variabel digunakan untuk memvisualisasikan hubungan antar fitur.

4. Hasil dan Diskusi

- Model Random Forest yang dilatih mencapai akurasi sekitar [X%] (sesuai hasil evaluasi model).
- Evaluasi menggunakan classification report memberikan rincian tentang precision, recall, dan F1-score untuk setiap kelas kualitas pisang.
- Analisis korelasi menunjukkan fitur-fitur tertentu yang memiliki hubungan kuat dengan kualitas pisang, yang dapat digunakan untuk lebih mengoptimalkan model di masa mendatang.

5. Kesimpulan

Proyek Banana Quality ML Analysis berhasil mengembangkan model machine learning yang akurat dalam mengklasifikasikan kualitas pisang. Model Random Forest menunjukkan performa yang memadai dan analisis data memberikan wawasan berharga tentang fitur yang memengaruhi kualitas pisang. Proyek ini berpotensi digunakan dalam proses inspeksi kualitas produk secara otomatis.

6. Rekomendasi

Beberapa peningkatan yang dapat dilakukan di masa depan antara lain:

- Penambahan fitur baru yang dapat lebih memperkaya model, seperti data terkait cuaca atau metode penyimpanan pisang.
- Penggunaan model lain seperti Gradient Boosting atau Support Vector Machine untuk melihat perbandingan kinerja model.
- Pemrosesan lebih lanjut dengan teknik lain seperti dimensionality reduction untuk mengurangi kompleksitas model.

7. Kode Sumber

Seluruh kode yang digunakan dalam proyek ini dapat ditemukan di struktur direktori proyek, termasuk kode untuk preprocessing data, pelatihan model, dan evaluasi:

<https://github.com/haekalsetiawan/banana-quality-ml-analysis.git>