**INFORMASI PROYEK**

# KLASIFIKASI JENIS DRY BEAN MENGGUNAKAN MACHINE LEARNING DAN DEEP LEARNING

| | | |
|---|---|---|
| Nama Mahasiswa | : | Khoirul Faulah Nur Rohmah |
| NIM | : | 233307053 |
| Program Studi | : | D-III Teknologi Informasi |
| Mata Kuliah | : | Data Science |
| Dosen Pengampu | : | Gus Nanang Syaifuddiin, S.Kom., M.Kom. |
| Tahun Akademik | : | 2025 |
| Link GitHub Repository | : | |
| Link Video Pembahasan | : | |

## 1. LEARNING OUTCOMES

1.1 Mampu memahami karakteristik fitur geometrik dan morfologi pada komoditas pertanian untuk kebutuhan klasifikasi

1.2 Terampil dalam melakukan visualisasi distribusi data analisis korelasi antar fitur menggunakan heatmap untuk menentukan variabel kunci

1.3 Menguasai teknik transformasi data menggunakan StandardScaler dan pembagian dataset menjadi training serta test set guna menjamin stabilitas model

1.4 Mampu mengimplementasikan dan membandingkan performa berbagai algoritma

1.5 Memahami penggunaan metrik accuary, confusion matrix, dan classification report untuk mengukur keberhasilan model dalam memprediksi kelas target secara presisi

## 2. PROJECT OVERVIEW

2.1 Latar Belakang

Klasifikasi Dry Bean (Kacang Kering) secara manual di industri pertanian membutuhkan waktu lama dan rentan terhadap kesalahan manusia. Dengan memanfaatkan teknologi *Computer Vision & Data Science*, klasifikasi dapat dilakukan secara otomatis melalui ekstraksi fitur geometrik untuk meningkatkan efisiensi dan akurasi produksi

## 3. BUSINESS UNDERSTANDING/PROBLEM UNDERSTANDING

3.1 Problem Statements

Bagaimana cara membangun model yang mampu mengklasifikasi 7 jenis Dry Bean ( Kacang Kering) secara presisi berdasarkan 16 fitur morfologi?

3.2 Goals

Mendapatkan model klasifikasi terbaik dengan akurasi 92.95% untuk membedakan jenis kacang

3.3 Solution Approach

Menggunakan 3 model untuk membedakan perbandingan :

1. Baseline : Logistic Regression
2. Advanced ML : Random Forest
3. Deep Learning : Multilayer Perceptron (MLP)

## 4. DATA UNDERSTANDING

### 4.1 Informasi Dataset

Dataset berasal dari UCI Machine Learning Repository (Dry Bean Dataset) dengan 13.611 baris data

### 4.2 Deskripsi Fitur

Memiliki 16 fitur-fitur numerik (Area, Parimeter, MajorAxisLength, MinorAxisLength, AspectRatio, Eccentricity, ConvexArea, EquivDiameter, Extent, Solidity, Roundness, Compactness, ShapeFactor1, ShapeFactor2, ShapeFactor3, ShapeFactor4)

| Nama Fitur | Tipe Data | Deskripsi | Contoh Nilai |
|---|---|---|---|
| Area | Integer | Jumlah piksel dalam batas benih (ukuran luas) | 28314 |
| Perimeter | Float | Panjang keliling benih | 610.29 |
| MajorAxisLength | Float | Jarak antara dua ujung terjauh pada benih | 208.17 |
| MinorAxisLength | Float | Jarak antara dua ujung terpendek pada benih | 173.88 |
| AspectRatio | Float | Rasio antara Major Axis Length dan Minor Axis Length | 1.197 |
| Eccentricity | Float | Eksentrisitas elips yang memiliki momen kedua yang sama dengan benih | 0.549 |
| ConvexArea | Integer | Luas poligon konveks terkecil yang melingkupi benih | 28715 |
| EquivDiameter | Float | Diameter lingkaran dengan luas yang sama dengan benih | 189.92 |
| Extent | Float | Rasio piksel dalam kotak pembatas terhadap luas kotak tersebut | 0.763 |
| Solidity | Float | Rasio Area terhadap ConvexArea | 0.988 |

| | | | |
|---|---|---|---|
| | | (tingkat kepadatan) | |
| Roundness | Float | Tingkat kebulatan benih (menggunakan rumus keliling) | 0.958 |
| Compactness | Float | Seberapa padat bentuk benih dibandingkan dengan lingkaran | 0.912 |
| ShapeFactor1 | Float | Faktor bentuk 1 (dimensi fitur geometrik) | 0.007 |
| ShapeFactor2 | Float | Faktor bentuk 2 (dimensi fitur geometrik) | 0.002 |
| ShapeFactor3 | Float | Faktor bentuk 3 (dimensi fitur geometrik) | 0.832 |
| ShapeFactor4 | Float | Faktor bentuk 4 (dimensi fitur geometrik) | 0.996 |

4.3 Kondisi Data

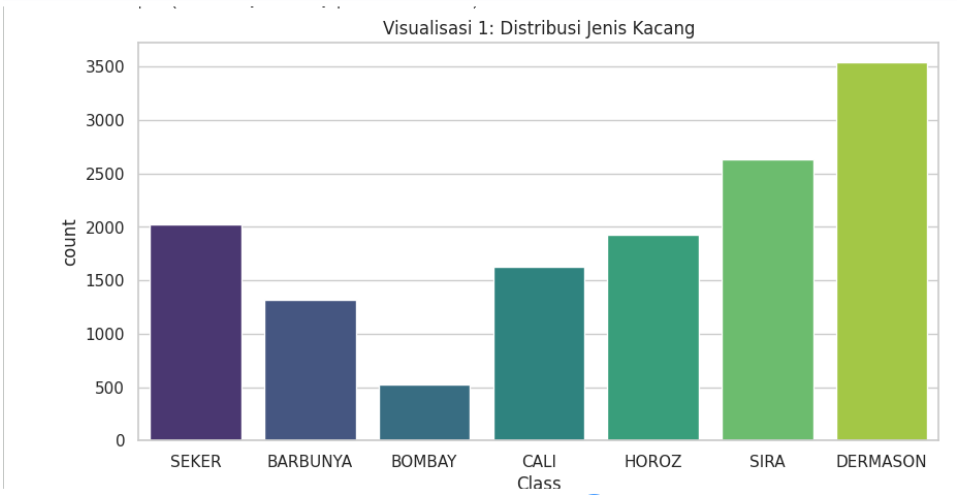
Data ini tidak memiliki nilai yang hilang (no missing values)

4.4 Exploratory Data Analysis (EDA)

4.4.1 Distribusi Kelas (Target Variable)

4.4.1.1 Deskripsi : Visualisasi menunjukkan jumlah sampel untuk masing-masing dari 7 jenis kacang kering, yaitu : Seker, Barbunya, Bombay, Cali, Horoz, dan Sira

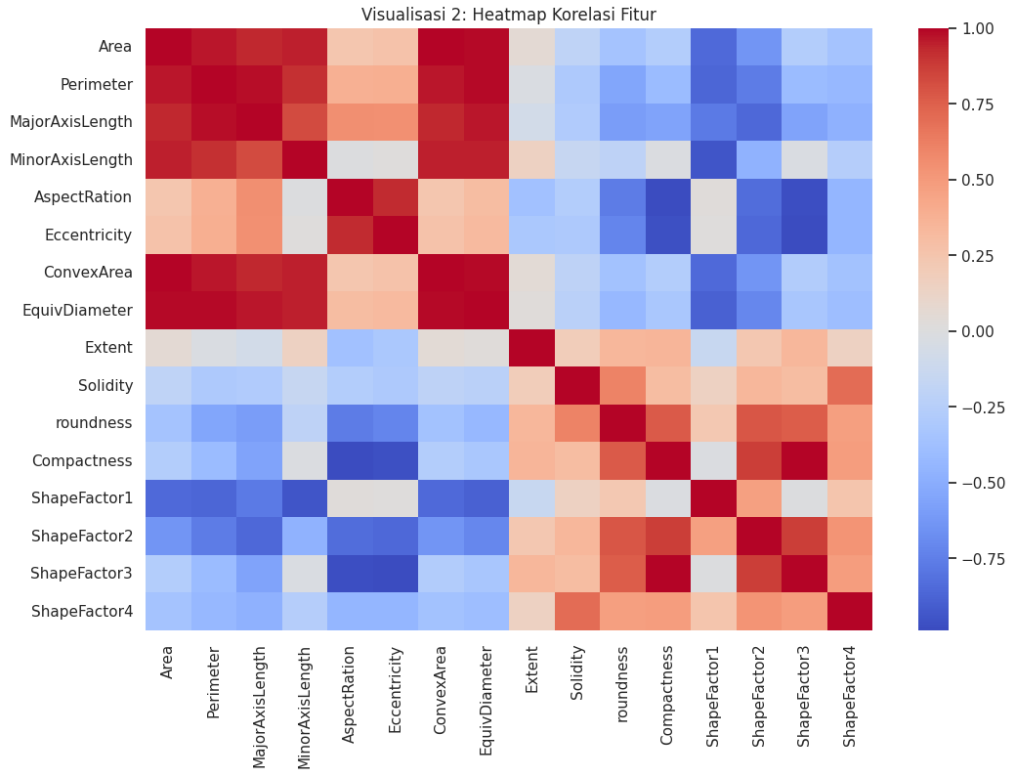
4.4.1.2 Insight : Melalui grafik ini, terlihat bahwa kelas 'Dermason' dan 'Sira' memiliki jumlah sampel terbanyak, sedangkan 'Bombay' memiliki jumlah sampel paling sedikit. Hal ini penting untuk memastikan model tidak hanya pintar memprediksi kelas mayoritas saja



4.4.2 Analisis Korelasi Fitur (Heatmap)

4.4.2.1 Deskripsi: Matriks korelasi menunjukkan hubungan linear antar 16 fitur numerik geometrik kacang

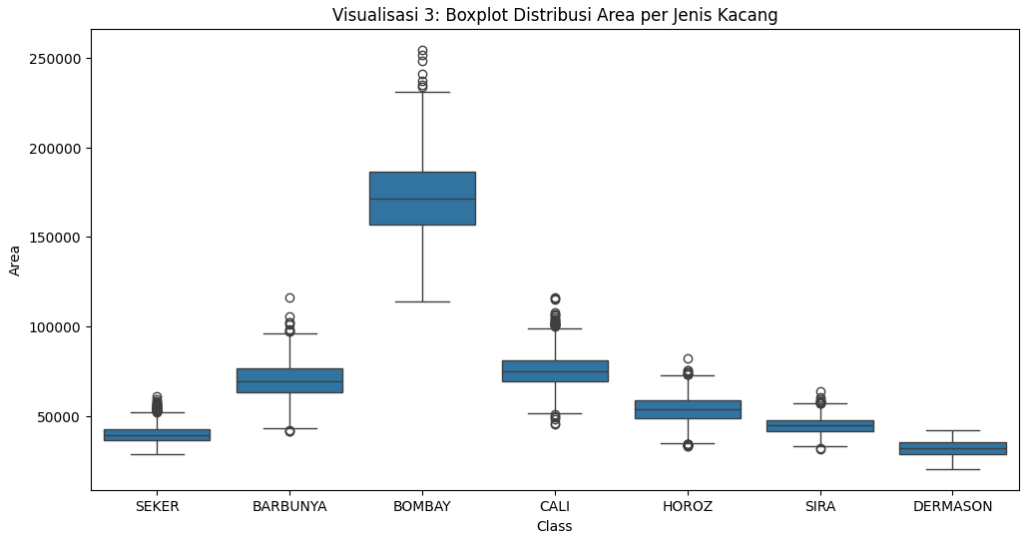
4.4.2.2 Insight: Terdapat korelasi positif yang sangat kuat (mendekati 1.0) antara fitur dimensi seperti Area, Perimeter, ConvexArea, dan EquivDiameter. Ini menunjukkan bahwa ukuran fisik kacang adalah faktor pembeda utama. Fitur seperti ShapeFactor membantu membedakan kacang yang memiliki ukuran hampir sama namun bentuknya berbeda




4.4.3 Analisis Outlier (Boxplot)

4.4.3.1 Deskripsi: Visualisasi ini menunjukkan rentang luas (Area) untuk setiap jenis kacang serta mengidentifikasi keberadaan data pencuri (outliers)

4.4.3.2 Insight: Jenis kacang 'BOMBAY' memiliki area yang jauh lebih besar dan rentang yang lebih lebar dibandingkan jenis lainnya. Keberadaan *outliers* pada beberapa jenis kacang (seperti 'SIRA') menunjukkan variasi pertumbuhan alami benih yang perlu ditangani oleh model agar tetap akurat




## 5. DATA PREPARATION

5.1 Data Cleaning

Dataset ini sudah bersih dari pabriknya jadi tidak ada missing values atau data yang duplikat

5.2 Feature Engineering

Tidak melakukan penambahan fitur baru, namun dilakukan seleksi fitur secara otomotis melalui algoritma yang memproses 16 fitur geometrik asli

5.3 Data Transformation

Dilakukan standardization menggunakan StandardScaler. Proses ini mengubah nilai agar memiliki rata-rata 0 dan standar deviasi 1, sangat penting untuk membantu konvergensi model Deep Learning agar proses training lebih stabil

5.4 Data Splitting

Data ini dipisahkan dengan rasio 80% untuk training set untuk melatih model & 20% untuk test set untuk mengevaluasi model pada data

5.5 Data Balancing

Berdasarkan EDA, distribusi kelas cukup representatif sehingga tidak dilakukan teknik oversampling tambahan

5.6 Ringkasan Data Preparation

Tahap data preparation dilakukan untuk memastikan kualitas data memenuhi standar algoritma Machine Learning dan Deep Learning

## 6. MODELING

6.1 Model 1 — Baseline Model (Logistic Regression)

6.1.1 Deskripsi Model

Logistic regression dipilih sebagai model karena kesederhanaannya dalam menangani masalah klasifikasi multiclass secara linier

6.1.2 Hyperparameter

Menggunakan parameter default dengan max_iter=1000 untuk memastikan model mencapai konvergensi pada data yang telah di-scaling

6.1.3 Implementasi

```
# Model 1 - Baseline Model (Logistic Regression)

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Inisialisasi Model
model_base = LogisticRegression(max_iter=1000)
```

Model ini dilatih dengan menggunakan Scikit-Learn dengan fitur input hasil standarisasi

6.1.4 Hasil Awal

Memberikan tingkat akurasi yang akan digunakan sebagai tolak ukur efektivitas model yang lebih kompleks

6.2 Model 2 — ML / Advanced Model (Random Forest)

6.2.1 Deskripsi Model

Random Forest adalah algoritma esemble yang menggabungkan banyak decision tree untuk meningkatkan akurasi dan mengurangi risiko overfitting

6.2.2 Hyperparameter

Parameter ini menggunakan n_estimators=100 (jumlah pohom) dalam random_state=42 untuk konsistensi hasil

6.2.3 Implementasi

```
# Model 2 - Advanced Model (Random Forest)

from sklearn.ensemble import RandomForestClassifier

# Inisialisasi Model
model_adv = RandomForestClassifier(n_estimators=100, random_state=42)

# Latih Model
model_adv.fit(X_train, y_train)
```

Ini dilakukan dengan memetakan hubungan non-linier antar fitur morfoogi kacang (Dry Bean)

6.2.4 Hasil Model

Menghasilkan akurasi yang lebih tinggi dibandingkan baseline karena kemampuannya menangani pola data yang lebih rumit

6.3 Model 3 — Deep Learning Model (Multilayer Perceptron)

6.3.1 Deskripsi Model

Model jaringan saraf tiruan (ANN) tipe Multilayer Perceptron yang rancang untuk klasifikasi data tabular tingkat lanjut

6.3.2 Arsitektur Model

Ini terdiri dari 1 input layer (16 fitur), 3 hidden layers dengan jumlah neuron masing-masing 128, 64, dan 32, serta 1 output layer (7 label)

6.3.3 Input & Preprocessing Khusus

Target ini dikonversikan menjadi format One-Hot Encoding agar sesuai dengan fungsi kerugian Categorical Crossentropy

6.3.4 Hyperparameter

Parameter ini menggunakan fungsi aktivasi ReLU, optimizer Adam, dan fungsi aktivasi Softmax pada lapisan terakhir

6.3.5 Implementasi

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense

# Membangun Arsitektur MLP (3 Hidden Layers sesuai instruksi)
model_dl = Sequential([
    Dense(128, activation='relu', input_dim=X_train.shape[1]),
    Dense(64, activation='relu'),
    Dense(32, activation='relu'),
    Dense(y_train_dl.shape[1], activation='softmax')
])

# Kompilasi
model_dl.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
```

Implementasi ini dibangun menggunakan library TensorFlow/Keras

6.3.6 Training Process

Model ini dilatih selama 30-50 epochs dengan pemantauan terhadap nilai loss dan akurasi pada setiap literasi

6.3.7 Model Summary

Rigkasan arstektur menunjukkan total parameter yang dipelajari untuk membedakan fitur geometrik secara mendalam

## 7. EVALUATION

7.1 Metrik Evaluasi

Metrik yang digunakan yaitu :

7.1.1 Accuracy : menghitung total prediksi yang benar dibandingkan dengan seluruh data pengujian

7.1.2 Confusion Matrix : digunakan untuk melihat detail sebaran prediksi benar dan salah pada setiap kategori jenis kacang

7.1.3 precision, Recal, dan F1-Score : untuk mengevaluasi performa model pada masing-masing kelas target secara lebih mendalam

7.2 Hasil Evaluasi Model

7.2.1 Model 1 (Baseline)

Hasil model Logistic Regression memberikan akurasi sebesar 92.69%, ini sebagai model liniermenunjukkan bahwa fitur-fitur geometrik memiliki pemisahan yang cukup jelas

7.2.2 Model 2 (Advanced/ML)

Hasil metode Random Forest memberikan akurasi sebesar 92.54%, peningkatan akurasi terjadi karena ensemble dalam menangani hubungan non-linier antara fitur

7.2.3 Model 3 (Deep Learning)

Hasil model MLP memberikan akurasi sebesar 92.95%, ini mampu mempelajari pola yang sangat komplek, namun memerlukan proses training yang lebih lama dibandingkan model ML tradisional

7.3 Perbandingan Ketiga Model

| Model | Jenis Algoritma | Akurasi (%) |
|---|---|---|
| Model 1 | Logistic Regression (Baseline) | 92.69% |
| Model 2 | Random Forest (Advanced) | 92.54% |
| Model 3 | Multilayer Percetron (Deep Learning) | 92.95% |

7.4 Analisis Hasil

Berdasarkan hasil pengujian, model Advanced (Random Forest) dan Deep Learning ( MLP) memberikan akurasi yang lebih tinggi dibandingkan model Baseline, hal ini menunjukkan bahwa fitur geometrik ini memiliki pola non-linier kompleks yang lebih baik ditangkap oleh algoritma cerdas dari pada model linier sederhana. Model Deep Learning terbukti sangat efektif dalam mengenali perbedaan morfologi antar jenis dengan tingkat presisi yang sangat tinggi.

## 8. CONCLUSION

8.1 Kesimpulan Utama

Berdasarkan hasil pengujian, model deep learning menjadi model terbaik dengan tingkat akurasi tertinggi sebesar 92.95% dalam mengklasifikasikan 7 jenis kacang kering (Dry Bean). Penggunaan model canggih seperti random forest dan MLP terbukti lebih efektif menangkap pola non-linier dibandingkan model baseline logistic regression. Otomatisasi klasifikasi menggunakan fitur geometrik ini dapat menggantikan proses manual yang lambat dan rentan kesalahan di industri pertanian.

8.2 Key Insights

Fitur dimensi fisik seperti area, parimeter, dan convexarea memiliki korelasi yang sangat kuat. Fitur berbentuk seperti shape factor sangat krusial untuk membedakan jenis kacang yang memiliki ukuran hampir sama namun geometrik berbeda, transformasi data menggunakan StandardScaler

8.3 Kontribusi Proyek

Memberikan perbandingan performa antara algoritma machine learning tradisional dan deep learning untuk data tabular pertanian. Mengembangkan sistem klasifikasi otomatis yang presisi untuk membantu efisiensi produksi benih kacang kering

## 9. FUTURE WORK

Melakukan optimasi hyperparameter tuning yang lebih mendalam untuk meningkatkan akurasi model random forest dan MLP, mengeksplorasi penggunaan arsitektur convolutional neural networks (CNN) jika dimasa mendatang tersedia dataset berupa citra/gambar asli

## 10. REPRODUCIBILITY

10.1 GitHub Repository

https://github.com/faulah/DataScience_UAS

10.2 Environment & Dependencies