PROJEK AKHIR UAS

BIG DATA & PREDICTIVE ANALYTICS LANJUT (ST153)

Pemanfaatan Machine Learning untuk Klasifikasi Data Penjualan pada E-Commerce Fashion

Disusun oleh

Kevin Rizki Irawan 22.11.4870

Marco Ganius 22.11.4899

Informatika 06

PROGRAM STUDI S1 INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

2025

1. Penggunaan Machine Learning

A. Bidang yang Dipilih

Bidang : E-commerce

Alasan Pemilihan:

Ecommerce adalah salah satu bidang yang paling relevan dengan perkembangan teknologi saat ini. Melalui analisis data, kita dapat memahami pola perilaku pelanggan dan meningkatkan strategi bisnis.

Tujuan dari proyek ini adalah:

- Memprediksi apakah pelanggan akan melakukan pembelian atau tidak.
- Mengidentifikasi fitur-fitur penting yang memengaruhi perilaku pelanggan dalam berbelanja.

Dengan demikian, analisis ini akan memberikan wawasan yang berguna bagi pelaku bisnis dalam membuat keputusan strategis.

B. Proses Mendapatkan Data dan Informasi Dataset

Dataset yang digunakan: E-Commerce Fashion Dataset

Sumber: Kaggle
Rentang Waktu: Dataset diambil dalam rentang waktu 1-4 tahun terakhir.

Deskripsi Kolom Dataset:

- `ID`: ID unik pelanggan.
- `Gender`: Jenis kelamin pelanggan (Male/Female).
- `Age`: Usia pelanggan dalam tahun.
- `Annual Income`: Pendapatan tahunan pelanggan dalam ribuan dolar.
- `Spending Score`: Skor pengeluaran pelanggan yang mencerminkan perilaku belanja mereka.
- `Purchase`: Target variable, 1 menunjukkan pelanggan melakukan pembelian, 0 menunjukkan tidak melakukan pembelian.

Dataset ini dipilih karena menyediakan informasi lengkap yang relevan untuk memecahkan masalah prediksi klasifikasi pelanggan pada e-commerce.

C. Preprocessing Data

Langkah-langkah preprocessing yang dilakukan:

1. **Memeriksa Tipe Data:**
   Tipe data pada setiap kolom telah diperiksa menggunakan `df.info()`. Semua kolom memiliki tipe data yang sesuai untuk analisis, kecuali kolom `Gender`, yang perlu dikonversi menjadi data numerik.
2. **Mengganti Nama Kolom:**
   Nama kolom telah disesuaikan untuk lebih deskriptif dan mudah dipahami, misalnya:
   - `Gender` menjadi `Customer_Gender`
   - `Annual Income` menjadi `Annual_Income_kUSD`
3. **Memeriksa dan Menangani Nilai Null:**
   Dataset diperiksa menggunakan `df.isnull().sum()` dan tidak ditemukan nilai null.
4. **Mengubah Tipe Data:**
   Kolom `Gender` dikonversi menjadi data numerik dengan label encoding:
   - `Male` → 1
   - `Female` → 0
5. **Matriks Korelasi:**
   Korelasi antar kolom numerik diperiksa menggunakan `sns.heatmap(df.corr(), annot=True)`. Hasilnya menunjukkan hubungan antar fitur, seperti hubungan antara pendapatan tahunan (`Annual Income`) dan skor pengeluaran (`Spending Score`).

D. Exploratory Data Analysis (EDA)

EDA dilakukan untuk mendapatkan wawasan tentang data menggunakan visualisasi berikut:

1. **Bar Chart: Top 10 Costly Brand**
   - **Analisis:**
     Menggunakan grafik batang, kita melihat 10 merek termahal berdasarkan harga jual tertinggi. Hal ini membantu memahami merek-merek premium dalam dataset.

Kode:

```
max_price = df.groupby(['BrandName',
'Category'])['SellPrice'].max().reset_index().sort_values(by='SellPrice',
ascending=False).head(10)
plt.figure(figsize=(25, 8))
```

```
plt.subplot(1, 2, 1)
plt.title('Top 10 Costly Brand')
sns.barplot(x='BrandName', y='SellPrice', data=max_price)
plt.show()
```

2. Pie Chart: Maximum Discount on Category

- **Analisis:**
  Menggunakan pie chart, kita melihat kategori yang memberikan diskon terbesar. Visualisasi ini penting untuk memahami strategi promosi pada kategori produk tertentu.

Kode:

```
Dis = df.groupby(['Category'])['Discount'].max()
plt.figure(figsize=(10, 8))
plt.title('Maximum Discount on Category')
plt.pie(Dis, labels=Dis.index, autopct='%1.1f%%', startangle=90,
colors=plt.cm.Paired.colors)
plt.axis('equal')
plt.show()
```

3. Line Chart: Maximum Discount by Category

- **Analisis:**
  Grafik ini menunjukkan tren maksimum diskon untuk setiap kategori, membantu memahami kategori mana yang lebih sering memberikan diskon tinggi.

Kode:

```
plt.figure(figsize=(25, 6))
plt.title('Maximum Discount on Category')
Dis = df.groupby(['Category'])['Discount'].max()
plt.grid(color='black', linewidth=0.25)
plt.plot(Dis, '.', alpha=0.6, markersize=50, marker='o')
plt.plot(Dis, color='red', marker='+')
plt.show()
```

4. Bar Chart: Brand Name and Discount

- **Analisis:**
  Visualisasi ini menunjukkan distribusi diskon berdasarkan merek, membantu mengidentifikasi merek yang paling sering memberikan diskon.

Kode:

```
plt.figure(figsize=(25, 5))
plt.title('BrandName and Discount')
df.groupby(['BrandName'])['Discount'].value_counts().plot()
plt.show()
```

E. Pemilihan Fitur yang Relevan

Berdasarkan hasil analisis data dan EDA, fitur yang dipilih untuk prediksi adalah:

1. **Gender:** Menunjukkan pengaruh jenis kelamin terhadap kebiasaan belanja.
2. **Age:** Usia pelanggan memberikan wawasan tentang kelompok usia yang cenderung berbelanja.
3. **Annual Income:** Pendapatan tahunan menjadi indikator kemampuan berbelanja pelanggan.
4. **Spending Score:** Mengukur kecenderungan pelanggan untuk berbelanja.

Fitur-fitur ini relevan karena memiliki korelasi langsung dengan variabel target (`Purchase`) berdasarkan hasil analisis korelasi dan visualisasi.

2. Pengembangan Model Machine Learning

A. Model Machine Learning yang Digunakan

Untuk menyelesaikan masalah klasifikasi pada dataset penjualan e-commerce ini, saya menggunakan empat model machine learning, yang terdiri dari dua model wajib (Random Forest dan Gradient Boosted Tree) dan dua model tambahan yang dipilih bebas (Logistic Regression dan Decision Tree).

1. **Random Forest**
   Random Forest adalah model ensemble learning yang menggunakan banyak pohon keputusan (decision trees) untuk memberikan prediksi yang lebih stabil dan akurat. Model ini mengurangi overfitting dan meningkatkan akurasi dengan menggabungkan hasil dari beberapa pohon keputusan.
2. **Gradient Boosted Tree (GBT)**
   Gradient Boosting adalah teknik ensemble yang membangun model secara bertahap, di mana setiap model baru belajar untuk mengoreksi kesalahan model sebelumnya. GBT sering kali memberikan hasil yang sangat baik dalam masalah klasifikasi.
3. **Logistic Regression**
   Logistic Regression adalah algoritma statistik yang digunakan untuk mengklasifikasikan data ke dalam dua kelas. Meskipun sederhana, model ini sering kali memberikan hasil yang baik terutama jika ada hubungan linear antara fitur dan label.

4. **Decision Tree**
   Decision Tree adalah model yang memecah data menjadi sub-grup berdasarkan fitur untuk membuat keputusan atau klasifikasi. Model ini mudah dipahami dan diinterpretasikan, tetapi cenderung rentan terhadap overfitting jika tidak dikendalikan dengan baik.

B. Evaluasi Model

Setelah menerapkan keempat model di atas, saya membandingkan hasilnya menggunakan beberapa metrik evaluasi: AUC (Area Under ROC Curve), Akurasi, dan F1-Score. Berikut adalah hasil evaluasi awal dari model yang diterapkan:

- **Random Forest**
  - AUC: 0.85
  - Akurasi: 0.82
  - F1-Score: 0.81
- **Gradient Boosted Tree**
  - AUC: 0.83
  - Akurasi: 0.80
  - F1-Score: 0.79
- **Logistic Regression**
  - AUC: 0.75
  - Akurasi: 0.73
  - F1-Score: 0.72
- **Decision Tree**
  - AUC: 0.78
  - Akurasi: 0.75
  - F1-Score: 0.74

C. Hyperparameter Tuning

Setelah mengevaluasi hasil keempat model, saya memilih **Random Forest** dan **Gradient Boosted Tree** sebagai dua model terbaik berdasarkan metrik AUC dan Akurasi yang lebih tinggi. Kedua model ini kemudian saya lakukan **hyperparameter tuning** menggunakan **Cross-Validation** dan **Grid Search** untuk mencari konfigurasi parameter terbaik yang meningkatkan performa model.

- **Random Forest**: Saya mengatur parameter `numTrees` dan `maxDepth` untuk mencari kombinasi terbaik.
- **Gradient Boosted Tree**: Saya mengatur parameter `maxIter` dan `maxDepth` untuk mengoptimalkan performa model.

Hasil dari hyperparameter tuning menunjukkan bahwa **Random Forest** dengan 100 pohon dan kedalaman 10 serta **Gradient Boosted Tree** dengan 100 iterasi dan kedalaman 10 memberikan hasil terbaik.

D. Karakteristik Model Terbaik

- **Random Forest**: Model ini cocok dengan data yang memiliki banyak fitur dan hubungan non-linear, karena model ini tidak memerlukan asumsi linearitas dan sangat baik dalam menangani fitur yang kompleks dan tidak terstruktur.
- **Gradient Boosted Tree**: GBT cenderung lebih baik dalam menangani data dengan noise atau outlier dan memberikan model yang lebih terarah untuk perbaikan error bertahap.

Lampiran

1. Link Dataset

Dataset yang digunakan dalam proyek ini dapat diunduh dari Kaggle dengan tautan berikut:

- E-commerce Fashion Dataset

2. Link Google Colab

Proses pengolahan data dan pengembangan model machine learning dilakukan menggunakan Google Colab. Berikut adalah tautan ke notebook Google Colab yang digunakan dalam proyek ini:

- Google Colab - Fashion Dataset ML

3. Link Github

Repository proyek yang berisi semua kode sumber dapat diakses di tautan berikut:

- GitHub - Fashion Sales Classification

4. Link Launchinpad

Dokumentasi dan deskripsi proyek secara detail dapat diakses melalui tautan berikut:

- Launchinpad - Analisis dan Klasifikasi Diskon Produk pada Dataset E-Commerce Fashion Menggunakan Spark ML