Nama: M. Revaldi Sapta Wiguna

NIM: 231011400499

Kelas: 05 TPLE 004

Mata Kuliah: Machine Learning

Laporan Ujian Tengah Semester

Laporan Model Klasifikasi – Dataset Titanic

1. Deskripsi Dataset

Dataset yang digunakan adalah **Titanic Dataset** yang disediakan oleh pustaka seaborn. Dataset ini berisi data penumpang kapal Titanic yang tenggelam pada tahun 1912. Tujuan analisis ini adalah **memprediksi apakah seorang penumpang selamat atau tidak** berdasarkan karakteristik tertentu.

Beberapa variabel penting yang digunakan antara lain:

- **pclass** – kelas tiket penumpang (1 = kelas atas, 3 = kelas bawah)
- **sex** – jenis kelamin penumpang
- **age** – umur penumpang
- **sibsp** – jumlah saudara atau pasangan yang ikut di kapal
- **parch** – jumlah orang tua atau anak yang ikut di kapal
- **fare** – harga tiket yang dibayar
- **embarked** – pelabuhan keberangkatan (C = Cherbourg, Q = Queenstown, S = Southampton)
- **survived** – status keselamatan (1 = selamat, 0 = tidak selamat)

Dataset awal memiliki 891 baris data. Pada tahap preprocessing, kolom yang memiliki banyak nilai kosong seperti deck dan embark_town dihapus. Nilai kosong pada kolom age diisi dengan median, sedangkan kolom embarked diisi dengan modus.

---

2. Tahapan Analisis dan Preprocessing

Sebelum melakukan pemodelan, dilakukan beberapa langkah persiapan data (EDA dan preprocessing), yaitu:

1. **Pemeriksaan data kosong** untuk mengetahui variabel yang perlu diisi atau dihapus.
2. **Pembersihan data** dengan mengganti nilai kosong pada kolom age dan embarked.
3. **Encoding variabel kategorikal** menggunakan LabelEncoder agar bisa diolah oleh model machine learning.
4. **Pemilihan fitur** yang relevan dengan target, sehingga hanya fitur-fitur penting yang digunakan.
5. **Split data** menjadi data latih (80%) dan data uji (20%) agar model dapat dievaluasi secara objektif.

Setelah preprocessing, dataset sudah bersih, seluruh kolom memiliki tipe data numerik, dan siap digunakan untuk modeling.

---

3. Model Klasifikasi yang Digunakan

Dalam tugas ini digunakan dua algoritma klasifikasi yang populer, yaitu:

1. **Logistic Regression**
   Merupakan model linear yang digunakan untuk memprediksi probabilitas suatu data termasuk ke dalam kelas tertentu. Model ini cocok untuk klasifikasi biner seperti kasus Titanic. Logistic Regression juga menghasilkan model yang mudah diinterpretasikan dan stabil terhadap noise data.
2. **Decision Tree Classifier**
   Merupakan model non-linear berbasis struktur pohon keputusan. Model ini membagi data berdasarkan fitur yang paling informatif untuk meminimalkan kesalahan klasifikasi. Kelebihan Decision Tree adalah mudah dipahami secara visual, tetapi cenderung overfit jika tidak dikontrol dengan parameter tertentu.

Kedua model dilatih menggunakan dataset yang sama agar hasil evaluasi bisa dibandingkan secara adil.

---

4. Hasil Evaluasi Model

Evaluasi dilakukan menggunakan beberapa metrik utama: **accuracy, precision, recall, dan F1-score**, serta divisualisasikan melalui **confusion matrix** dan **ROC curve**.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.79	0.74	0.70	0.72
Decision Tree	0.76	0.73	0.67	0.70

Dari hasil tersebut, **Logistic Regression** memiliki performa sedikit lebih baik dibanding **Decision Tree**. ROC curve juga menunjukkan area yang lebih luas pada Logistic Regression, yang menandakan kemampuan model ini lebih baik dalam membedakan kelas positif dan negatif.

Pada confusion matrix, Logistic Regression menunjukkan prediksi yang lebih seimbang antara kelas “selamat” dan “tidak selamat”, sedangkan Decision Tree sedikit lebih condong ke salah satu kelas.

---

5. Pembahasan

Perbedaan hasil antara kedua model dapat dijelaskan sebagai berikut:

- Logistic Regression lebih stabil karena mengasumsikan hubungan linear antara variabel prediktor dan target. Hal ini membuat model tidak terlalu terpengaruh oleh variasi kecil dalam data.
- Decision Tree mampu menangkap hubungan non-linear, namun tanpa tuning atau pruning, model ini dapat mengalami **overfitting**, sehingga akurasi pada data uji sedikit lebih rendah.
- Dari sisi interpretasi, Decision Tree lebih mudah dipahami secara visual, karena struktur pohon dapat menunjukkan urutan pengambilan keputusan berdasarkan fitur (misalnya, jenis kelamin dan kelas penumpang menjadi faktor penting dalam keselamatan).

---

6. Kesimpulan

Berdasarkan hasil analisis dan perbandingan kedua model, dapat disimpulkan bahwa:

1. Kedua model sama-sama dapat digunakan untuk kasus klasifikasi biner seperti prediksi keselamatan penumpang Titanic.
2. **Logistic Regression** memberikan hasil akurasi dan kestabilan yang sedikit lebih baik dibandingkan Decision Tree.

3. Untuk dataset dengan karakteristik sederhana seperti Titanic, **Logistic Regression lebih direkomendasikan** karena efisien, stabil, dan tidak mudah overfit.
4. Namun, Decision Tree tetap berguna untuk eksplorasi awal dan analisis faktor-faktor yang paling berpengaruh terhadap hasil prediksi.