Laporan

FINAL PROJETC DATA MINING

"Penerapan Teknik Data Mining (Klasifikasi Non Regresi dan Regresi)"

Asisten:

1. Saparuddin
2. Isma Fauziah

Oleh:

Kelompok : 5

1) Ahmad Raenaldy (60900122060)
2) Andi Isratul Aulia (60900122057)
3) Muhammad Farid Irsyadillah (60900122055)

Kelas : C

LABORATORIUM KOMPUTER TERPADU

JURUSAN SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI ALAUDDIN MAKASSAR

2024/2025

I. Pendahuluan

1. Latar Belakang

   Data mining merupakan teknik yang penting untuk menggali informasi berharga dari data mentah. Dalam laporan ini, kami menerapkan dua teknik utama : klasifikas non regresi, dan Regresi yang banyak digunakan dalam berbagai bidang, seperti pemasaran, dan cuaca.

2. Tujuan Penelitian / Pengolahan Data

   - Klasifikasi Non Regresi : Membangun model yang dapat memprediksi apakah seorang nasabah akan menerima atau menolak penawaran produk tertentu, berdasarkan berbagai fitur yang ada dalam dataset, seperti usia, pekerjaan, pendidikan, status perkawinan, saldo rekening, dan faktor lainnya..
   - Regresi : Membangun model prediktif yang mampu memperkirakan suhu dalam skala Celsius (Temperature (C)) berdasarkan berbagai parameter cuaca lainnya seperti kelembapan, kecepatan angin, arah angin, jarak pandang, tekanan atmosfer, dan informasi temporal (jam, hari, bulan, dan tahun)

3. Deskripsi Dataset

- Dataset Klasifikasi:
  - Sumber : https://archive.ics.uci.edu/dataset/222/bank+marketing
  - Ukuran Dataset : 45211 record, 16 atribut, 1 label target.
  - Atribut :
    1) age – Usia nasabah
    2) job – Pekerjaan nasabah
    3) marital – Status perkawinan nasabah
    4) education – Tingkat pendidikan nasabah

5) default – Apakah nasabah memiliki tunggakan utang (1 jika ya, 0 jika tidak)
6) balance – Saldo rekening nasabah
7) housing – Apakah nasabah memiliki pinjaman perumahan (1 jika ya, 0 jika tidak)
8) loan – Apakah nasabah memiliki pinjaman pribadi (1 jika ya, 0 jika tidak)
9) contact – Metode kontak yang digunakan untuk menghubungi nasabah
10) day – Hari dalam bulan saat nasabah terakhir dihubungi
11) month – Bulan saat nasabah terakhir dihubungi
12) duration – Durasi percakapan terakhir dengan nasabah
13) campaign – Jumlah kontak yang dilakukan selama kampanye ini
14) pdays – Jumlah hari sejak nasabah terakhir dihubungi dalam kampanye sebelumnya
15) previous – Jumlah kontak yang dilakukan dengan nasabah dalam kampanye sebelumnya
16) poutcome – Hasil dari kampanye pemasaran sebelumnya
17) y – Target klasifikasi, di mana 1 berarti nasabah menerima penawaran, dan 0 berarti menolak penawaran

- Dataset Regresi :
  - Sumber : https://www.kaggle.com/datasets/zubairmustafa/shopping-mall-customer-segmentation-data
  - Ukuran Dataset : 96453 record, 13 atribut.
  - Atribut :

1) Summary – Ringkasan cuaca
2) Precip Type – Jenis presipitasi (hujan, salju, dll.)
3) Temperature (C) – Suhu (dalam Celcius), yang merupakan target regresi
4) Humidity – Kelembaban udara
5) Wind Speed (km/h) – Kecepatan angin dalam kilometer per jam
6) Wind Bearing (degrees) – Arah angin dalam derajat
7) Visibility (km) – Jarak pandang dalam kilometer
8) Loud Cover – Tutupan awan
9) Pressure (millibars) – Tekanan udara dalam millibar
10) Hour – Jam pada hari tersebut
11) Day – Hari dalam bulan tersebut
12) Month – Bulan dalam tahun tersebut
13) Year – Tahun tercatat

4. Pembagian Dataset dalam Tugas
    - Dataset Untuk Klasifikasi : Bank Marketing Dataset.
    - Dataset Untuk Regresi : Weather History Dataset.

II. Metodologi

1. Preprocessing data

- Klasifikasi :
    - Memuat Dataset - Mengimpor dataset dari file CSV.
    - Memeriksa Informasi Dataset - Menampilkan beberapa baris pertama, tipe data, dan ukuran dataset.
    - Memilih Kolom yang Relevan - Memilih kolom yang diperlukan untuk analisis.

- Penanganan Missing Value - Memeriksa dan menangani nilai yang hilang dalam dataset.
- Label Encoding - Mengubah variabel kategorikal menjadi numerik dengan menggunakan LabelEncoder.
- Normalisasi Data - Menggunakan MinMaxScaler untuk menormalisasi fitur numerik.
- Visualisasi Data - Menampilkan grafik distribusi target, korelasi antar fitur, dan hubungan antara fitur numerik.
- Menyimpan Dataset yang Sudah Dibersihkan - Menyimpan dataset yang telah diproses ke file CSV baru.

- Regresi :
  - Memuat Dataset - Mengimpor dataset dari file CSV.
  - Memeriksa Tipe Data - Memeriksa tipe data untuk setiap kolom agar sesuai dengan analisis yang diinginkan.
  - Memeriksa Ukuran Dataset - Memeriksa jumlah baris dan kolom dalam dataset untuk memahami skala data.
  - Memilih Kolom yang Diperlukan - Memilih kolom relevan yang akan digunakan dalam analisis atau model.
  - Mengonversi Kolom Waktu - Mengubah kolom waktu menjadi format datetime dan mengekstrak fitur waktu seperti jam, hari, bulan, dan tahun.
  - Memeriksa Missing Value - Memeriksa apakah ada nilai yang hilang dalam dataset.
  - Menangani Missing Value - Mengganti nilai kosong dengan nilai yang paling sering muncul (mode) agar analisis tidak terpengaruh.

- Label Encoding - Mengubah variabel kategorikal menjadi numerik menggunakan LabelEncoder agar dapat digunakan dalam model.
- Normalisasi Data - Menggunakan MinMaxScaler untuk menormalkan fitur numerik agar berada dalam rentang yang konsisten.
- Menampilkan Statistik Deskriptif - Menggunakan fungsi describe() untuk menampilkan statistik deskriptif fitur numerik.
- Visualisasi Data - Menampilkan grafik untuk memahami distribusi data, korelasi antar fitur, dan hubungan antara fitur numerik.
- Menyimpan Dataset yang Sudah Dibersihkan - Menyimpan dataset yang telah diproses ke dalam file CSV baru.

2. Pemilihan Model dan Tahapannya

- Klasifikasi :
  - Memuat Dataset - Mengimpor dataset dari file CSV untuk digunakan dalam analisis lebih lanjut.
  - Memisahkan Kolom Fitur dan Target - Memisahkan kolom yang berisi fitur (X) dan target (y) dalam dataset. Target dalam hal ini adalah kolom 'y', yang digunakan untuk klasifikasi.
  - Membagi Data - Membagi dataset menjadi dua bagian: data pelatihan (70%) dan data pengujian (30%) menggunakan train_test_split.
  - Inisialisasi Model - Menyiapkan model klasifikasi, yaitu Decision Tree dan Random Forest, yang digunakan untuk pelatihan dan prediksi.
  - Menyimpan Hasil Evaluasi - Melatih setiap model dengan data pelatihan, melakukan prediksi pada data pengujian, lalu menghitung metrik evaluasi seperti akurasi, presisi, recall, F1 score, dan cross-validation.

- Menampilkan Hasil Evaluasi - Menampilkan hasil evaluasi model, termasuk akurasi, precision, recall, F1 score, dan confusion matrix, untuk membandingkan performa model.
- Visualisasi Hasil Evaluasi - Menampilkan grafik perbandingan metrik evaluasi (akurasi, presisi, recall, F1 score) dan hasil cross-validation untuk visualisasi yang lebih mudah dipahami.
- Analisis dan Interpretasi Hasil - Menyimpulkan model mana yang terbaik berdasarkan nilai akurasi dan memberikan interpretasi mengenai metrik yang digunakan.
- Contoh Prediksi - Menampilkan contoh perbandingan antara nilai aktual dan prediksi untuk beberapa data pengujian sebagai contoh hasil prediksi model terbaik.

- Regresi :
  - Membaca Dataset - Mengimpor dataset dari file CSV untuk digunakan dalam analisis.
  - Memisahkan Fitur dan Target - Memisahkan kolom fitur (X) dan target (y), di mana target adalah 'Temperature (C)'.
  - Membagi Data - Membagi dataset menjadi data pelatihan (70%) dan data pengujian (30%).
  - Melakukan Inisialisasi Model - Menyiapkan dan menginisialisasi model regresi (Linear Regression, Random Forest, dan Gradient Boosting).
  - Menyimpan Hasil Evaluasi untuk Masing-Masing Model - Melatih model dan menghitung metrik evaluasi seperti MAE, MSE, R2, dan cross-validation R2.
  - Menampilkan Hasil Evaluasi - Menampilkan hasil evaluasi untuk setiap model yang meliputi metrik kinerja seperti MAE, MSE, dan R2.

- Visualisasi Hasil Evaluasi untuk Perbandingan - Menampilkan grafik bar untuk membandingkan metrik evaluasi dan hasil cross-validation.
- Analisis dan Interpretasi Hasil - Menganalisis dan menginterpretasikan hasil model terbaik berdasarkan nilai R2, serta menjelaskan metrik evaluasi.
- Contoh Hasil Prediksi Berdasarkan Model Terbaik - Menampilkan contoh prediksi untuk beberapa data uji dan membandingkannya dengan nilai aktual.

III. Hasil dan Analisis

1. EDA Analysis

- Klasifikasi :

1) Ukuran dan Kebersihan Data: Dataset terdiri dari 45,211 baris dan 17 kolom, tanpa nilai yang hilang, yang berarti data ini bersih dan siap untuk analisis lebih lanjut.
2) Tipe Data: Sebagian besar fitur adalah numerik setelah normalisasi, sementara beberapa fitur seperti job, marital, dan education merupakan kategori yang sudah di-encode menjadi numerik.
3) Normalisasi: Fitur numerik seperti age, balance, dan duration sudah dinormalisasi ke dalam rentang seragam, yang membantu dalam pemodelan.
4) Statistik Deskriptif: Rata-rata usia pelanggan sekitar 30 tahun, dengan saldo rata-rata rendah. Durasi percakapan relatif panjang, menunjukkan ketertarikan terhadap produk.
5) Ketidakseimbangan Kelas: Kolom target y menunjukkan ketidakseimbangan kelas yang signifikan, dengan hanya 12%

pelanggan yang tertarik pada produk (y=1), sementara sisanya tidak tertarik.

6) Distribusi Data: Mayoritas pelanggan menikah dan memiliki tingkat pendidikan secondary atau lebih tinggi.

- Regresi :

1) Ukuran dan Kebersihan Data: Dataset terdiri dari 96,453 baris dan 12 kolom. Setelah penanganan nilai yang hilang, dataset ini tidak memiliki nilai kosong, dengan kolom 'Precip Type' yang telah diimputasi menggunakan nilai modus.
2) Tipe Data: Sebagian besar fitur dalam dataset ini adalah numerik setelah normalisasi, sementara beberapa fitur seperti 'Summary', 'Precip Type', dan 'Visibility (km)' yang sebelumnya berupa kategori telah di-encode menjadi numerik.
3) Normalisasi: Fitur numerik seperti 'Temperature (C)', 'Humidity', 'Wind Speed (km/h)', 'Wind Bearing (degrees)', 'Visibility (km)', dan 'Pressure (millibars)' telah dinormalisasi ke dalam rentang antara 0 dan 1, yang membantu proses analisis lebih lanjut.
4) Statistik Deskriptif: Rata-rata kelembapan (Humidity) adalah 65.43%, sementara suhu ('Temperature (C)') menunjukkan kisaran nilai yang lebih tinggi setelah normalisasi. Tekanan udara ('Pressure (millibars)') memiliki rentang yang cukup besar, menunjukkan variasi yang signifikan dalam data cuaca.
5) Distribusi Data: Data menunjukkan bahwa sebagian besar cuaca yang tercatat memiliki kondisi yang cukup bervariasi dalam suhu, kelembapan, dan kecepatan angin. Namun, sebagian besar suhu berada dalam kisaran yang lebih tinggi setelah normalisasi.

2. Model Performance

- Klasifikasi :
  - Decision Tree:
    1) Accuracy: 0.8739 – Model ini berhasil melakukan prediksi dengan akurasi sekitar 87.39%.
    2) Precision: 0.4651 – Proporsi prediksi positif yang benar, menunjukkan model ini masih banyak menghasilkan prediksi positif yang salah.
    3) Recall: 0.4675 – Kemampuan model untuk mendeteksi kasus positif, model cukup baik dalam menemukan kasus positif tetapi masih ada yang terlewat.
    4) F1 Score: 0.4663 – Rata-rata harmonis antara precision dan recall, menunjukkan keseimbangan antara keduanya.
    5) CV Accuracy Mean: 0.8757 – Akurasi rata-rata model dalam cross-validation, cukup konsisten di berbagai fold.
    6) CV Accuracy Std: 0.0042 – Variasi kecil antara fold, menunjukkan bahwa model stabil.
    7) Confusion Matrix:

       True Positive (TP): 747

       True Negative (TN): 11107

       False Positive (FP): 859

       False Negative (FN): 851
    8) Interpretasi: Model cenderung lebih banyak mengklasifikasikan negatif dengan benar (TN), namun ada beberapa prediksi positif yang salah (FP).
  - Random Forest:

1) Accuracy: 0.9026 – Model ini berhasil melakukan prediksi dengan akurasi sekitar 90.26%, lebih tinggi dibandingkan Decision Tree.
2) Precision: 0.6346 – Model ini lebih baik dalam memprediksi positif yang benar dibandingkan Decision Tree.
3) Recall: 0.4086 – Kemampuan model untuk mendeteksi kasus positif lebih rendah dibandingkan Decision Tree, lebih banyak positif yang terlewat.
4) F1 Score: 0.4971 – Nilai F1 lebih tinggi dibandingkan Decision Tree, menunjukkan keseimbangan antara precision dan recall yang lebih baik.
5) CV Accuracy Mean: 0.9044 – Akurasi rata-rata model dalam cross-validation lebih tinggi dari Decision Tree, menunjukkan performa yang stabil dan baik.
6) CV Accuracy Std: 0.0030 – Variasi kecil antar fold, menunjukkan model cukup stabil.
7) Confusion Matrix:

   True Positive (TP): 653

   True Negative (TN): 11590

   False Positive (FP): 376

   False Negative (FN): 945
8) Interpretasi: Model lebih baik dalam memprediksi kasus negatif (TN) dan memiliki lebih sedikit prediksi positif yang salah (FP) dibandingkan Decision Tree, namun ada banyak kasus positif yang terlewat (FN).

- Regresi :
  - Linear Regression:

1) MAE (Mean Absolute Error): 0.1963 – Rata-rata selisih absolut antara prediksi dan nilai sebenarnya. Model ini menghasilkan kesalahan yang cukup besar.
2) MSE (Mean Squared Error): 0.0532 – Kesalahan kuadrat rata-rata. Semakin tinggi MSE, semakin besar penalti untuk kesalahan besar.
3) R2 (R-squared): 0.3118 – Hanya 31.18% dari variabilitas target yang dapat dijelaskan oleh model, menunjukkan penjelasan yang rendah.
4) CV R2 Mean: 0.3105 – Hasil rata-rata R2 dari cross-validation, konsisten dengan nilai R2.
5) CV R2 Std: 0.0055 – Variasi kecil antara fold, model cukup stabil.

- Random Forest:
  1) MAE: 0.1099 – Prediksi lebih akurat dibandingkan Linear Regression, dengan kesalahan yang lebih kecil.
  2) MSE: 0.0265 – MSE lebih kecil, menunjukkan kesalahan yang lebih rendah dibandingkan Linear Regression.
  3) R2: 0.6565 – Menjelaskan 65.65% dari variabilitas target, jauh lebih baik dibandingkan Linear Regression.
  4) CV R2 Mean: 0.6737 – Rata-rata R2 dari cross-validation menunjukkan performa yang lebih baik dan konsisten.
  5) CV R2 Std: 0.0049 – Variasi kecil antar fold, menunjukkan stabilitas model.
- Gradient Boosting:
  1) MAE: 0.1604 – Prediksi sedikit lebih buruk dibandingkan Random Forest, dengan MAE lebih tinggi.
  2) MSE: 0.0418 – MSE lebih kecil dari Linear Regression, tetapi lebih tinggi dibandingkan Random Forest.

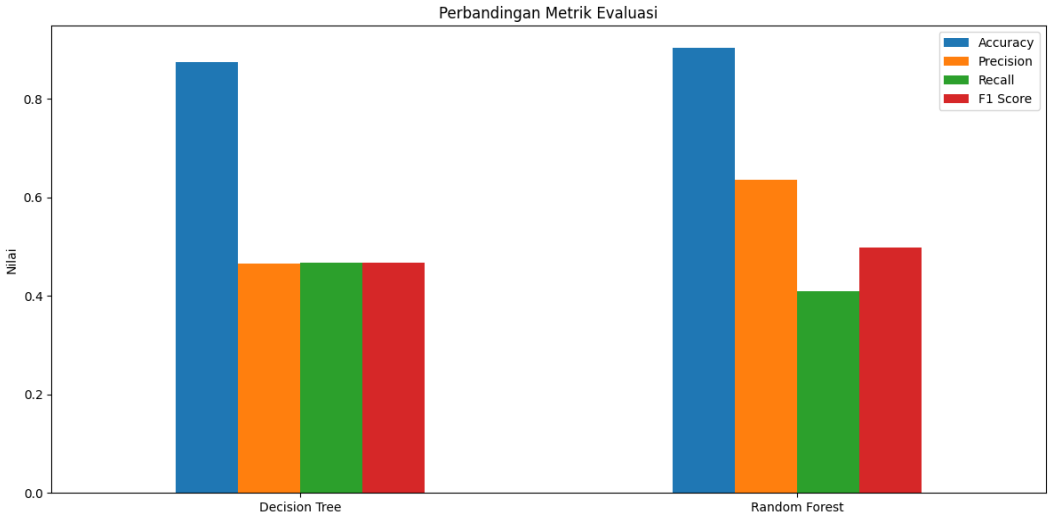
3) R2: 0.4592 – Menjelaskan 45.92% dari variabilitas target, lebih baik dari Linear Regression namun tidak sebaik Random Forest.
4) CV R2 Mean: 0.4597 – Rata-rata R2 dari cross-validation mendekati nilai R2 pada data uji.
5) CV R2 Std: 0.0058 – Variasi kecil antara fold, menunjukkan stabilitas model.

3. Comparative Analysis

- Klasifikasi :

1) Akurasi: Random Forest memiliki akurasi yang lebih tinggi (90,26%) dibandingkan Decision Tree (87,39%), menunjukkan performa yang lebih baik dalam memprediksi data secara keseluruhan.
2) Precision: Random Forest lebih unggul dalam precision (63,46%) yang berarti model ini lebih tepat dalam memprediksi kelas positif dibandingkan Decision Tree (46,51%).
3) Recall: Decision Tree sedikit lebih baik dalam recall (46,75%) yang berarti lebih baik dalam menangkap kasus positif dibandingkan Random Forest (40,86%).
4) Stabilitas Model: Random Forest lebih stabil dengan CV accuracy mean yang lebih tinggi (90,45%) dan standar deviasi lebih rendah, menunjukkan model ini lebih konsisten dalam berbagai percobaan.

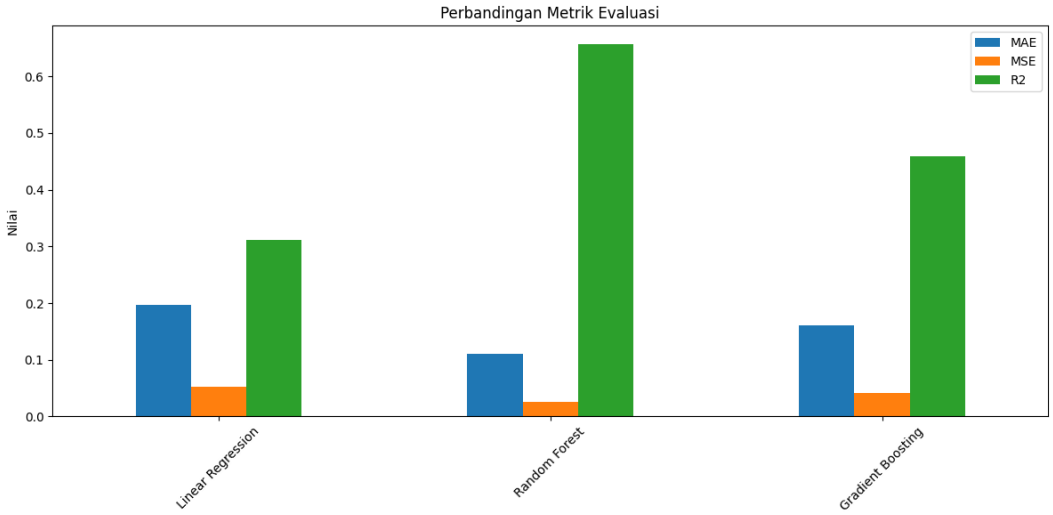
Kesimpulan: Random Forest lebih unggul dalam akurasi, precision, dan stabilitas, sementara Decision Tree lebih baik dalam recall, lebih sensitif terhadap deteksi kasus positif.



- Regresi :
  1) MAE (Mean Absolute Error): Random Forest memiliki MAE terendah (0.11), menunjukkan bahwa model ini menghasilkan prediksi yang lebih akurat dengan kesalahan prediksi yang lebih kecil dibandingkan dengan model lainnya. Linear Regression memiliki MAE tertinggi (0.20), yang menunjukkan bahwa model ini lebih banyak melakukan kesalahan dalam prediksi.
  2) MSE (Mean Squared Error): Random Forest juga unggul dalam MSE dengan nilai terendah (0.03), yang mengindikasikan bahwa model ini memiliki kesalahan prediksi yang lebih kecil secara keseluruhan dibandingkan dengan Gradient Boosting (0.04) dan Linear Regression (0.05).
  3) R^2 (Coefficient of Determination): Random Forest memiliki R^2 tertinggi (0.66), yang menunjukkan bahwa model ini dapat menjelaskan 66% variabilitas data dengan sangat baik. Gradient Boosting berada di posisi kedua (0.46), sedangkan Linear Regression memiliki nilai R^2 terendah (0.31), yang berarti model ini kurang mampu menjelaskan variabilitas data.

4) CV R^2 Mean dan CV R^2 Std (Cross-Validation): Random Forest menunjukkan hasil terbaik dengan nilai CV R^2 Mean 0.67 dan standar deviasi rendah (0.005), menunjukkan stabilitas dan konsistensi model di berbagai subset data. Gradient Boosting memiliki nilai CV R^2 Mean 0.46, sedangkan Linear Regression memiliki nilai CV R^2 Mean terendah (0.31) dan lebih fluktuatif.

Kesimpulan: Random Forest unggul dalam semua metrik evaluasi, menunjukkan akurasi dan stabilitas terbaik. Gradient Boosting berada di posisi kedua, sementara Linear Regression memiliki performa terburuk di hampir semua metrik.



IV. Kesimpulan

1. Summary Findings

Dalam teknik klasifikasi, dataset yang digunakan terdiri dari 45.211 baris dan 17 kolom, dengan beberapa fitur kategori yang telah di-encode menjadi numerik dan fitur numerik yang telah dinormalisasi. Setelah data dibersihkan dan missing values ditangani, data siap untuk analisis lebih lanjut. Berdasarkan hasil klasifikasi menggunakan Decision Tree dan Random Forest, Random Forest menunjukkan performa yang lebih baik

dibandingkan Decision Tree. Random Forest memiliki akurasi lebih tinggi, mencapai 90,26%, dibandingkan dengan Decision Tree yang hanya 87,39%. Dalam hal precision, Random Forest juga unggul dengan nilai 63,46%, yang menunjukkan model ini lebih tepat dalam memprediksi kelas positif. Sementara itu, Decision Tree lebih baik dalam recall (46,75%), yang berarti lebih sensitif terhadap deteksi kasus positif, meskipun akurasinya lebih rendah. Dalam hal stabilitas model, Random Forest lebih konsisten dengan CV accuracy mean yang lebih tinggi dan standar deviasi lebih rendah, menunjukkan bahwa model ini lebih stabil dalam berbagai percobaan. Secara keseluruhan, meskipun Decision Tree lebih sensitif terhadap deteksi kasus positif, Random Forest lebih unggul dalam akurasi, precision, dan stabilitas, menjadikannya pilihan yang lebih baik untuk masalah klasifikasi ini.

Dalam teknik regresi, dataset yang digunakan terdiri dari 96,453 baris dan 12 kolom, dengan beberapa fitur yang awalnya memiliki missing values. Setelah dilakukan penanganan terhadap missing values dan normalisasi fitur numerik, data siap untuk dianalisis lebih lanjut. Model regresi yang diuji, yaitu Linear Regression, Random Forest, dan Gradient Boosting, menunjukkan hasil yang bervariasi. Random Forest menonjol sebagai model terbaik, dengan R^2 sebesar 0.656, menunjukkan kemampuan model untuk menjelaskan lebih dari 65% varians dalam data. Selain itu, Random Forest juga memiliki MAE dan MSE yang lebih rendah, yang menunjukkan akurasi prediksi yang lebih baik. Sementara itu, Gradient Boosting memberikan hasil yang lebih baik dibandingkan Linear Regression, tetapi masih tidak sebanding dengan Random Forest dalam hal akurasi dan kesalahan prediksi. Linear Regression, meskipun sering

digunakan untuk masalah regresi, menunjukkan performa yang lebih rendah, dengan R^2 hanya 0.312. Secara keseluruhan, Random Forest adalah model yang paling unggul untuk masalah regresi ini, menawarkan prediksi yang lebih akurat dan stabil dibandingkan dengan model lainnya.

- Recommendations :

  Untuk masalah klasifikasi, Random Forest merupakan model yang paling unggul dibandingkan dengan Decision Tree. Dengan akurasi, precision, dan stabilitas yang lebih tinggi, Random Forest sebaiknya diprioritaskan sebagai model utama. Namun, meskipun Random Forest unggul dalam hal akurasi dan precision, recall-nya masih lebih rendah dibandingkan Decision Tree. Oleh karena itu, jika fokusnya adalah menangkap lebih banyak kasus positif (kelas yang jarang terjadi), disarankan untuk mengoptimalkan recall pada Random Forest, yang bisa dilakukan dengan teknik hyperparameter tuning atau menggunakan metode balancing data seperti SMOTE. Selain itu, penting untuk terus mengevaluasi model dan melakukan eksperimen lebih lanjut, seperti mencoba teknik balancing data, untuk mencapai performa yang lebih baik.

  Dalam regresi, Random Forest juga terbukti menjadi model yang paling unggul dengan R^2 yang lebih tinggi (0,656), menunjukkan bahwa model ini dapat menjelaskan lebih banyak varians dalam data dan memberikan prediksi yang lebih akurat. Oleh karena itu, Random Forest sebaiknya digunakan sebagai model utama untuk regresi. Gradient Boosting menunjukkan performa yang lebih baik dibandingkan dengan Linear Regression dan bisa dipertimbangkan sebagai alternatif kedua. Sementara itu, Linear Regression, meskipun

sering digunakan untuk masalah regresi, menunjukkan performa yang lebih rendah dengan R^2 hanya 0,312. Untuk meningkatkan performanya, disarankan untuk memeriksa pemilihan fitur atau mempertimbangkan teknik regularisasi seperti Lasso atau Ridge Regression.

- Future Improvements :

  Untuk perbaikan di masa depan dalam tugas klasifikasi dan regresi, beberapa langkah dapat diambil untuk meningkatkan performa model. Pertama, feature engineering yang lebih baik sangat penting, seperti mencoba teknik encoding yang lebih kompleks atau melakukan transformasi fitur non-linier untuk meningkatkan kualitas data yang dimasukkan ke dalam model. Selain itu, pemilihan dan eliminasi fitur yang kurang relevan juga dapat membantu meningkatkan akurasi model. Kedua, hyperparameter tuning perlu dilakukan untuk menemukan kombinasi parameter yang optimal menggunakan teknik seperti grid search atau random search. Hal ini akan memberikan model yang lebih presisi dan sesuai dengan data yang ada. Untuk meningkatkan akurasi lebih lanjut, eksperimen dengan model ensembling seperti Voting Classifier atau Stacking bisa dilakukan pada tugas klasifikasi, sementara teknik seperti Bagging atau Boosting dapat diterapkan pada tugas regresi.