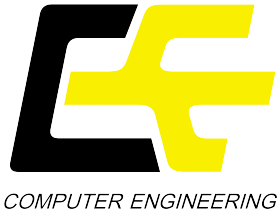

Laporan Final Project Machine Learning
Machine Learning

Prediksi Nilai Ujian Pada Sekolah Menengah ke Atas Menggunakan Model RandomForest Regressor,XGboost,MLP Tensorflow,dan Linear Regresi

Muhammad Risang - 50242310xx

Akhmad Rizqullah Ridlohi - 5024231037

Abraham Napitupulu - 5024231037

Departemen Teknik Komputer
Institut Teknologi Sepuluh Nopember
Surabaya

Penulis: Kelompok 6
20 Juni 2025

Daftar Isi

Dasar Teori

Pada bagian ini, akan dibahas landasan teoretis yang menjadi dasar pelaksanaan proyek. Penjelasan mencakup konsep fundamental machine learning, alur kerja yang diterapkan, metodologi pra-pemrosesan data, serta arsitektur dan cara kerja dari setiap model yang digunakan untuk memprediksi kinerja akademik siswa.

1 Machine Learning dan Supervised Learning

Machine learning adalah cabang dari kecerdasan buatan (AI) yang berfokus pada pengembangan algoritma yang memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit. Dalam proyek ini, pendekatan yang digunakan adalah Supervised Learning (Pembelajaran Terarah).

Dalam Supervised Learning, model dapat belajar dari dataset yang telah memiliki label atau target yang diketahui. Model menganalisis serangkaian fitur input (variabel independen, X) dan hubungannya dengan variabel output (variabel dependen, y). Tujuan utamanya adalah untuk mempelajari fungsi pemetaan (y=f(X)) sehingga ketika data baru tanpa label diberikan, model dapat memprediksi outputnya dengan akurasi tinggi.

Solusi dari permasalahan yang ingin diselesaikan dalam proyek ini meliputi model regresi, yaitu salah satu bentuk Supervised Learning di mana variabel output (y) yang diprediksi bersifat kontinu atau numerik. Dalam konteks ini, model bertujuan untuk memprediksi nilai akhir siswa (G3) yang merupakan sebuah angka.

2 Alur Kerja Proyek Machine Learning

Untuk memastikan hasil yang sistematis dan dapat dipertanggungjawabkan, proyek ini mengikuti alur kerja standar dalam pengembangan model machine learning, yang terdiri dari beberapa tahapan utama:

1. **Eksplorasi Data (EDA):** Memahami karakteristik, distribusi, dan hubungan antar variabel dalam dataset melalui statistik deskriptif dan visualisasi data.

2. **Rekayasa Fitur (Feature Engineering):** Membuat fitur-fitur baru yang lebih informatif dari fitur yang sudah ada (misalnya, grade$_avg_prev)untukmeningkatka$

3. **Pemodelan (Modeling):** Memilih, melatih, dan mengevaluasi beberapa algoritma yang berbeda untuk menemukan model dengan performa terbaik.

4. **Evaluasi Model:** Mengukur kinerja model terbaik menggunakan metrik kuantitatif seperti R-squared dan RMSE untuk menilai akurasi dan keandalannya.

5. **Prediksi (Inference):** Menggunakan model yang telah dilatih untuk membuat prediksi pada data baru yang belum pernah dilihat sebelumnya, seperti pada simulasi uji coba real-time.

3 Pra-pemrosesan Data

Data mentah jarang sekali bisa langsung digunakan oleh model. Oleh karena itu, diperlukan beberapa teknik pra-pemrosesan untuk membersihkan dan menstrukturkan data.

3.1 One-Hot Encoding

Sebagian besar model machine learning hanya dapat bekerja dengan data numerik. Fitur-fitur kategorikal dalam dataset ini (seperti sex, address, higher) yang berbentuk teks perlu diubah menjadi representasi numerik. One-Hot Encoding adalah teknik yang digunakan untuk tujuan ini, di mana setiap kategori unik dalam sebuah fitur diubah menjadi kolom biner baru (bernilai 0 atau 1).

3.2 Penskalaan Fitur (Feature Scaling)

Fitur-fitur numerik seringkali memiliki rentang nilai yang sangat berbeda (misalnya, age (15-22) vs absences (0-93)). Jika tidak diskalakan, fitur dengan rentang nilai yang lebih besar dapat mendominasi proses pembelajaran model secara tidak adil. StandardScaler digunakan untuk mentransformasi setiap fitur sehingga memiliki rata-rata 0 dan standar deviasi 1. Langkah ini sangat krusial untuk model yang sensitif terhadap skala seperti Regresi Linear dan MLP.

4 Model yang Digunakan

Pemilihan model dilakukan secara strategis untuk membandingkan pendekatan dari berbagai paradigma dengan tingkat kompleksitas yang berbeda.

4.1 Regresi Linear

Model ini dipilih sebagai baseline atau titik awal perbandingan. Regresi Linear bekerja dengan mengasumsikan adanya hubungan linear antara fitur-fitur input dan variabel target. Ia mencoba menemukan satu formula matematis terbaik untuk memetakan input ke output. Meskipun sederhana, model ini sangat berguna untuk mengukur seberapa baik masalah dapat diselesaikan dengan pendekatan linear dan untuk menilai relevansi fitur yang dipilih.

4.2 Random Forest

Random Forest adalah perwakilan dari model ensemble berbasis pohon keputusan. Cara kerjanya adalah dengan membangun ratusan pohon keputusan secara independen, di mana setiap pohon dilatih pada sampel data yang sedikit berbeda. Untuk membuat prediksi, model ini mengambil rata-rata dari semua prediksi pohon individual. Pendekatan "wisdom of the crowd" ini membuatnya sangat kuat dalam menangkap interaksi fitur yang kompleks, robust terhadap noise, dan cenderung tidak overfitting, sehingga sangat cocok untuk data tabular seperti pada proyek ini.

4.3 XGBoost

XGBoost (eXtreme Gradient Boosting) dipilih sebagai perwakilan dari metode gradient boosting. Berbeda dari Random Forest, XGBoost membangun pohon keputusan secara sekuensial. Setiap pohon baru yang dibuat secara spesifik bertugas untuk memperbaiki kesalahan prediksi dari pohon-pohon sebelumnya. Proses belajar iteratif yang fokus pada kesalahan ini membuat XGBoost menjadi salah satu model dengan performa tertinggi di industri untuk masalah regresi dan klasifikasi pada data tabular.

4.4 Multi-Layer Perceptron (MLP)

MLP dipilih untuk menjelajahi pendekatan deep learning. Model ini bekerja seperti jaringan otak buatan, terdiri dari lapisan-lapisan "neuron" yang saling terhubung. MLP mampu belajar dan mengenali pola-pola yang sangat

kompleks dan abstrak dari kombinasi semua fitur secara bersamaan. Model ini dipilih untuk menguji apakah arsitektur non-linear yang sangat fleksibel ini dapat menemukan pola yang terlewatkan oleh model berbasis pohon.

5 Metrik Evaluasi Model

Untuk mengukur dan membandingkan kinerja model secara objektif, dua metrik utama digunakan:

- **R-squared (R^2):** Metrik ini mengukur seberapa besar persentase variasi (perbedaan) dalam variabel target (G3) yang dapat dijelaskan oleh model. Nilai R^2 berkisar dari 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan model yang lebih baik.
- **RMSE (Root Mean Squared Error):** Metrik ini mengukur rata-rata besaran kesalahan prediksi model dalam satuan yang sama dengan variabel target. Dalam proyek ini, RMSE menunjukkan rata-rata berapa poin nilai prediksi meleset dari nilai sebenarnya. Nilai RMSE yang lebih rendah menunjukkan model yang lebih akurat.

Hasil Praktikum

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Hasil

Setelah

Kesimpulan

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.