

Laporan Final Project 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 - 50242310xx

*Departemen Teknik Komputer
Institut Teknologi Sepuluh Nopember
Surabaya*

Penulis: Kelompok 6
20 Juni 2025

Daftar Isi

Dasar Teori

1 Latar Belakang dan Relevansi

Di sistem pendidikan kita sekarang, bantuan untuk siswa yang kesulitan seringkali datang terlambat. Biasanya, sekolah baru akan bertindak—misalnya dengan mengadakan kelas remedial—setelah nilai ujian akhir keluar dan hasilnya kurang memuaskan. Tentu saja, pendekatan seperti ini punya kelemahan besar: bantuan diberikan saat semuanya sudah terjadi, di mana siswa mungkin sudah kehilangan motivasi.

Di sinilah *machine learning* bisa mengubah keadaan. Dengan menganalisis data-data siswa yang sudah ada—mulai dari data pribadi, pergaulan, dan terutama nilai-nilai sebelumnya—kita bisa membuat sebuah model prediksi. Model ini bisa berfungsi seperti “sistem peringatan dini” yang bisa menandai siswa mana yang kemungkinan besar akan gagal, jauh sebelum ujian akhir. Dengan kemampuan ini, para guru bisa memberikan bantuan yang lebih personal dan tepat waktu, sehingga peluang siswa untuk berhasil jadi lebih besar.

2 Tujuan dan Ruang Lingkup Proyek

Berdasarkan masalah di atas, tujuan utama dari proyek ini adalah:

1. Membuat dan melatih beberapa model *machine learning* untuk memprediksi nilai akhir matematika siswa (G3).
2. Membandingkan model-model tersebut untuk menemukan mana yang paling akurat dan bisa diandalkan.
3. Menganalisis fitur atau faktor apa saja yang paling berpengaruh menurut model terbaik.

Untuk mencapai tujuan itu, proyek ini punya batasan (ruang lingkup) sebagai berikut:

- **Dataset:** Kami menggunakan dataset publik “Student Performance” dari UCI, yang berisi data dari 395 siswa di Portugal.
- **Metodologi:** Kami mengikuti alur kerja standar, mulai dari membuat fitur baru (*feature engineering*), memproses data (*preprocessing*), hingga melatih model.
- **Model yang Diuji:** Kami membandingkan beberapa jenis model: **Regresi Linear** (sebagai baseline), **Random Forest** dan **XGBoost** (sebagai model ensemble canggih), serta **Multi-Layer Perceptron (MLP)** (sebagai perwakilan deep learning).
- **Metrik Evaluasi:** Performa model diukur menggunakan **R-squared (R^2)** dan **Root Mean Squared Error (RMSE)**.

Pendekatan untuk membandingkan berbagai model seperti ini sejalan dengan banyak penelitian di bidang *Educational Data Mining*. Sebagai contoh, sebuah studi oleh Al-Barrak dan Al-Razgan (2016) juga membandingkan beberapa algoritma, termasuk *Decision Trees* (yang menjadi dasar dari Random Forest dan XGBoost), untuk memprediksi IPK mahasiswa. Studi mereka juga menemukan bahwa nilai-nilai sebelumnya adalah faktor prediksi terkuat, sebuah temuan yang menjadi fondasi utama dalam analisis kami.

Referensi

- Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting Students Final GPA Using Decision Trees: A Case Study. *International Journal of Information and Education Technology*, 6(7), 528-533. doi:10.7763/IJIET.2016.V6.745.
- Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5TG7T>.

DataSet

3 Deskripsi Dataset

Dataset yang kami pilih untuk proyek ini, yaitu “Student Performance Data Set”, lebih dari sekadar nilai-nilai tak bermakna. Ini adalah potret lengkap kehidupan para siswa. Keistimewaannya adalah data yang sangat beragam, mencakup semuanya mulai dari nilai-nilai akademik, data diri (seperti usia dan alamat), latar belakang keluarga (seperti pendidikan orang tua), hingga gaya hidup mereka (seperti waktu belajar, absensi, dan seberapa sering mereka main). Karena datanya begitu kaya dan bervariasi, ini menjadi dasar yang sempurna untuk membuat model prediksi yang bagus dan menggali faktor apa saja yang sebenarnya mengukur kesuksesan akademis seorang siswa.

3.1 Sumber Data

- **Judul:** Student Performance Data Set
- **Author:** P. Cortez and A. Silva
- **Tanggal Terbit:** 2008
- **Sumber:** UCI Machine Learning Repository
- **Relevansi:** Dataset ini sangat relevan karena secara spesifik mencatat berbagai atribut demografis, sosial, dan akademik siswa yang dapat digunakan untuk tujuan prediksi kinerja, sejalan dengan tujuan proyek ini.

3.2 Ukuran dan Kualitas Dataset

- **Jumlah Data:** Dataset ini terdiri dari **395 entri** (siswa).

- **Jumlah Fitur:** Terdapat total **33 fitur** (atribut) awal untuk setiap siswa.
- **Data Kosong:** Analisis awal menunjukkan bahwa dataset ini memiliki kualitas yang sangat baik, dengan **tidak ada nilai kosong** (*missing values*) sama sekali, sehingga tidak memerlukan langkah mengeluarkan data.

3.3 Fitur-fitur Relevan

Berikut adalah deskripsi dari fitur-fitur utama yang terdapat dalam dataset: —p0.15—p0.15—p0.6—

Nama Variabel	Tipe Data	Deskripsi
school	Kategorikal	Sekolah siswa ('GP' atau 'MS')
sex	Biner	Jenis kelamin siswa ('F' atau 'M')
age	Integer	Usia siswa (15 hingga 22)
address	Kategorikal	Tipe alamat rumah ('U' - urban atau 'R' - rural)
famsize	Kategorikal	Ukuran keluarga ('LE3' - ≤3 atau 'GT3' - >3)
Pstatus	Kategorikal	Status kohabitasi orang tua ('T' - bersama atau 'A' - berpisah)
Medu	Integer	Pendidikan Ibu (0 - tidak ada, 4 - pendidikan tinggi)
Fedu	Integer	Pendidikan Ayah (0 - tidak ada, 4 - pendidikan tinggi)
Mjob	Kategorikal	Pekerjaan Ibu ('teacher', 'health', 'services', dll.)
Fjob	Kategorikal	Pekerjaan Ayah ('teacher', 'health', 'services', dll.)
studytime	Integer	Waktu belajar mingguan (1 - ≤2 jam, 4 - ≤10 jam)
failures	Integer	Jumlah kegagalan kelas sebelumnya (0-4)
schoolsup	Biner	Dukungan pendidikan ekstra dari sekolah (yes/no)
famsup	Biner	Dukungan pendidikan dari keluarga (yes/no)
activities	Biner	Mengikuti kegiatan ekstrakurikuler (yes/no)
higher	Biner	Ingin melanjutkan ke pendidikan tinggi (yes/no)
internet	Biner	Akses internet di rumah (yes/no)
romantic	Biner	Dalam hubungan romantis (yes/no)
famrel	Integer	Kualitas hubungan keluarga (1 - sangat buruk, 5 - sangat baik)
goout	Integer	Frekuensi pergi keluar bersama teman (1 - sangat jarang, 5 - sangat sering)
health	Integer	Status kesehatan saat ini (1 - sangat buruk, 5 - sangat baik)
absences	Integer	Jumlah absensi sekolah (0 hingga 93)
G1	Target	Nilai periode pertama (0 hingga 20)
G2	Target	Nilai periode kedua (0 hingga 20)

G3 Target Nilai akhir (0 hingga 20) — **Output Target Utama**

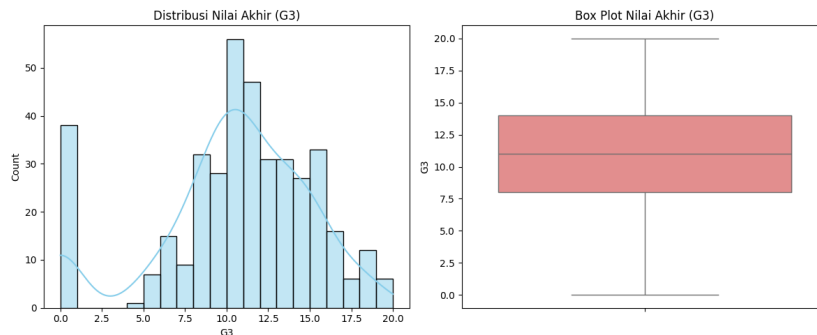
Deskripsi Fitur dalam Dataset

4 Langkah Preprocessing dan Transformasi Data

Data mentah diolah melalui beberapa tahapan kunci untuk memaksimalkan potensinya dalam pemodelan.

4.1 Eksplorasi Data (EDA)

Langkah awal ini krusial untuk memahami data. Dengan membuat visualisasi seperti histogram dan heatmap, kami dapat mengidentifikasi pola, distribusi data, serta korelasi antar fitur. Wawasan dari EDA, seperti korelasi kuat antara G1/G2 dengan G3, menjadi dasar untuk pemilihan fitur di tahap selanjutnya.



Gambar 1: Distribusi value G3

4.2 Rekayasa Fitur (Feature Engineering)

Daripada hanya menggunakan fitur asli, kami membuat fitur-fitur baru yang lebih informatif. Contohnya, `grade_avg_prev` (rata-rata G1 dan G2) dibuat untuk menangkap sinyal kinerja akademik sebelumnya dalam satu variabel kuat. Langkah ini relevan karena seringkali kombinasi fitur memberikan informasi yang lebih berpengaruh daripada fitur individual.



Model machine learning memerlukan input numerik. Fitur-fitur kategorikal seperti sex ('F'/'M') atau higher ('yes'/'no') diubah menjadi format biner (0 dan 1) melalui One-Hot Encoding. Transformasi ini wajib dilakukan agar algoritma dapat memproses semua informasi yang tersedia dalam dataset.


```
Proses One-Hot Encoding (Contoh Konversi Teks ke Numerik):  
- Fitur 'sex': 'M' -> 1, 'F' -> 0  
- Fitur 'address': 'U' -> 1, 'R' -> 0  
- Fitur 'higher': 'yes' -> 1, 'no' -> 0  
- Fitur 'activities': 'yes' -> 1, 'no' -> 0
```

Gambar 4: One-Hot Encoding untuk fitur dengan value non-numerik

Hasil Praktikum

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur 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

Kami melakukan uji coba model dengan skenario dataset yang memiliki total 33 fitur, dan fitur yang dipakai untuk memprediksi kinerja akademik siswa. Fitur-fitur tersebut meliputi nilai ujian sebelumnya (G1, G2, dan rata-ratanya), data demografi dan keluarga (usia, pendidikan orang tua, jenis kelamin, alamat), serta kebiasaan belajar dan sekolah (waktu belajar, kegagalan sebelumnya, absensi, akses internet, partisipasi kegiatan). Kami juga menyertakan aspek gaya hidup dan sosial siswa, seperti frekuensi bersosialisasi, status kesehatan, status hubungan romantis, dan konsumsi alkohol, dengan jumlah siswa total 395. Setelah melakukan uji coba didapatkan hasil sebagai berikut

5 Uji Coba dengan Data Siswa

```
=== LANGKAH 7: UJI COBA REAL-TIME ===

--- Data Siswa Baru ---
G1          15
G2          15
failures    0
higher      yes
age         16
Medu        4
Fedu        3
studytime   3
absences    2
goout       2
health      5
sex         M
address     U
internet    yes
romantic    no
freetime    4
Dalc        1
Walc        2
activities  yes
dtype: object
...
>>> Prediksi Nilai Akhir (G3) oleh MLP: 15.13
(Catatan: RMSE model ini ~2.66)
```

Gambar 5: Data Siswa Pertama


Di sini pada nilai aktualnya nilai G3 siswa adalah 15, dan saat dilakukan uji coba prediksi didapatkan setiap model seperti yang ada pada gambar berikut.

```
--- Prediksi Menggunakan Linear Regression (Simple) ---
>>> Prediksi Nilai Akhir (G3) oleh LinReg (Simple): 10.91
(Catatan: RMSE model ini ~4.67)

--- Prediksi Menggunakan Linear Regression (Lengkap) ---
>>> Prediksi Nilai Akhir (G3) oleh LinReg (Lengkap): 18.06
(Catatan: RMSE model ini ~2.24)

--- Prediksi Menggunakan RandomForest Regressor ---
>>> Prediksi Nilai Akhir (G3) oleh RandomForest: 15.23
(Catatan: RMSE model ini ~1.75)

--- Prediksi Menggunakan XGBoost Regressor ---
>>> Prediksi Nilai Akhir (G3) oleh XGBoost: 14.82
(Catatan: RMSE model ini ~1.82)

--- Prediksi Menggunakan MLP TensorFlow/Keras ---
1/1  0s 158ms/step
>>> Prediksi Nilai Akhir (G3) oleh MLP: 15.13
(Catatan: RMSE model ini ~2.66)
```

Gambar 6: Hasil Uji Coba Model Pertama

Dilanjutkan uji coba dengan data siswa sebagai berikut:

Dengan data siswa tersebut, kami mengasumsikan bahwa siswa tersebut memiliki nilai G3 sebesar 14,5 yang didapat dengan perkiraan secara intuitif menggunakan heatmap yang ada dan didapatkan hasil seperti pada gambar berikut.

6 Evaluasi Model

Berikut adalah visualisasi untuk RMSE (Root Mean Square Error) dan R-Squared, di mana nilai RMSE yang baik adalah mendekati 0 dan R-Squared mendekati 1.

Dari data tersebut RandomForest Regressor mendapatkan nilai RMSE yang lebih rendah daripada semua model yang ada, diikuti oleh XGBoost, Linear Regression dengan fitur yang lengkap, lalu MLP TensorFlow. Nilai RMSE menunjukkan berapa rentang error yang dihasilkan oleh data. Untuk R-Squared, RandomForest Regressor juga merupakan yang paling baik

```

...
=== LANGKAH 7: UJI COBA REAL-TIME ===

--- Data Siswa Baru ---
G1      15
G2      16
failures 0
higher  yes
age      17
Medu     4
Fedu     4
studytime 3
absences 2
goout    2
health   5
sex      M
address  U
internet yes
romantic no
freetime 3
Dalc     1
Walc     1
activities yes
dtype: object

--- Prediksi Menggunakan Linear Regression (Simple) ---
>>> Prediksi Nilai Akhir (G3) oleh LinReg (Simple): 10.91
(Catatan: RMSE model ini ~4.67)

```

Gambar 7: Data Siswa Kedua

```

--- Prediksi Menggunakan Linear Regression (Simple) ---
>>> Prediksi Nilai Akhir (G3) oleh LinReg (Simple): 10.91
(Catatan: RMSE model ini ~4.67)

--- Prediksi Menggunakan Linear Regression (Lengkap) ---
>>> Prediksi Nilai Akhir (G3) oleh LinReg (Lengkap): 18.76
(Catatan: RMSE model ini ~2.24)

--- Prediksi Menggunakan RandomForest Regressor ---
>>> Prediksi Nilai Akhir (G3) oleh RandomForest: 15.67
(Catatan: RMSE model ini ~1.75)

--- Prediksi Menggunakan XGBoost Regressor ---
>>> Prediksi Nilai Akhir (G3) oleh XGBoost: 15.01
(Catatan: RMSE model ini ~1.82)

--- Prediksi Menggunakan MLP TensorFlow/Keras ---
1/1 [56ms/step]
>>> Prediksi Nilai Akhir (G3) oleh MLP: 16.23
(Catatan: RMSE model ini ~2.68)

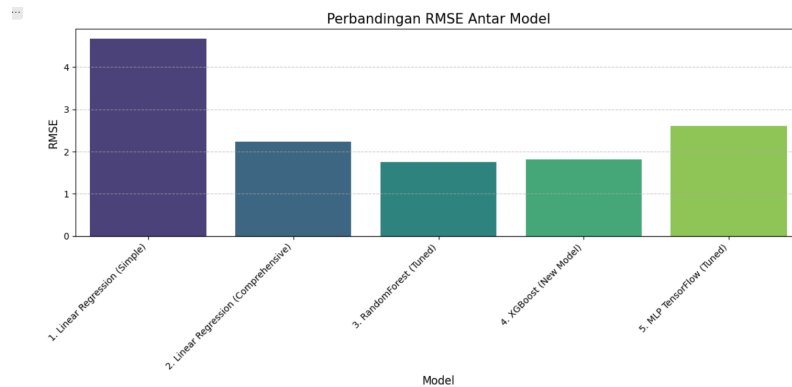
```

Gambar 8: Hasil Uji Coba Model Kedua

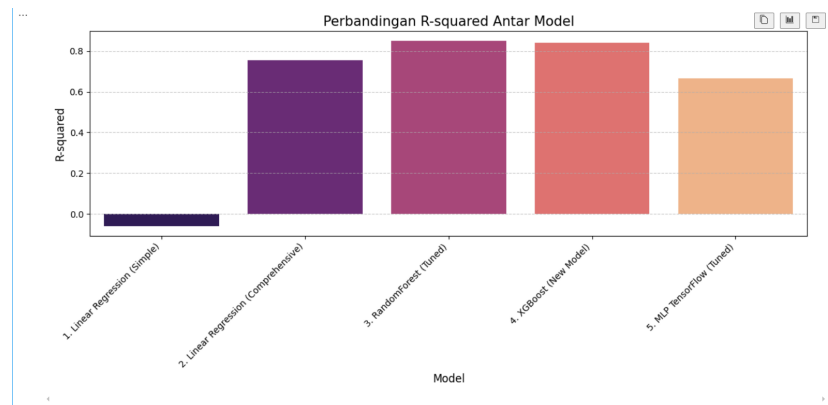
di antara model yang ada. R-Squared menunjukkan informasi mengenai bagaimana model tersebut menjelaskan variabilitas data terhadap model yang digunakan. R-Squared menggambarkan seberapa baik model sesuai dengan data yang digunakan.

Secara teknis, RandomForest Regressor mungkin lebih cocok untuk dataset yang relatif lebih kecil dibandingkan MLP TensorFlow dan XGBoost. RandomForest umumnya dianggap lebih tidak rumit dalam interpretasi dan implementasi dibandingkan dua model lainnya.

Dapat dilihat bahwa performa model RandomForest Regressor adalah model dengan prediksi yang paling akurat dibandingkan dengan model-model lain. Hal ini dapat terjadi dikarenakan RandomForest Regressor merupakan



Gambar 9: Visualisasi RMSE

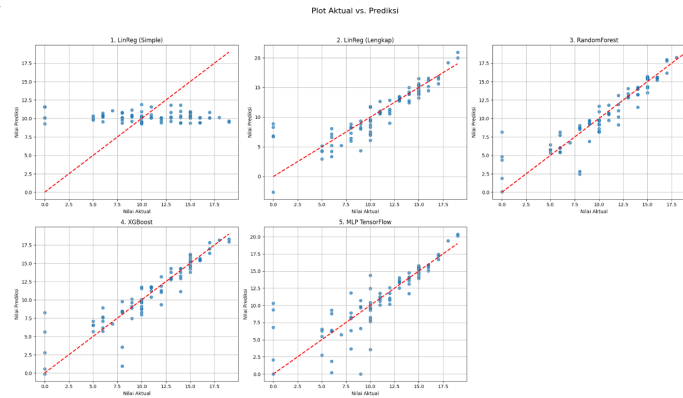


Gambar 10: Visualisasi R-Squared

model yang memang cocok untuk dataset yang lebih kecil daripada MLP TensorFlow dan XGBoost. Jika dibandingkan secara teknis, RandomForest merupakan model yang paling tidak rumit dibanding dua model tersebut. RandomForest Regressor bekerja dengan membuat decision tree secara paralel dengan memilih data secara acak, dan dari decision tree yang sudah dibuat secara paralel nilai prediksi dari setiap decision tree tersebut akan dirata-rata untuk mendapatkan hasilnya. XGBoost juga masih menggunakan decision tree, namun dengan cara memperbaiki kesalahan berulang kali dari decision tree sebelumnya. Untuk penjelasan simpelnya, MLP TensorFlow sendiri merupakan salah satu model jaringan saraf tiruan yang memiliki tiga komponen, yaitu *Input Layer*, *Hidden Layer*, dan *Output Layer* yang memiliki cara pemodelan yang lebih kompleks dibandingkan RandomForest Regressor dan XGBoost.

7 Visualisasi Scatter Plot

Kami juga melakukan visualisasi menggunakan scatter plot.



Gambar 11: Visualisasi Scatter Plot

Dari visualisasi scatter plot yang membandingkan nilai aktual dengan nilai prediksi untuk setiap model, beberapa wawasan penting dapat ditarik mengenai performa dan karakteristik setiap algoritma. Idealnya, titik-titik pada scatter plot harus terkonsentrasi rapat di sepanjang garis diagonal merah ($y=x$), menunjukkan bahwa nilai prediksi sangat mendekati nilai aktual.

Model RandomForest Regressor dan XGBoost menunjukkan performa yang paling baik dalam visualisasi ini. Titik-titik data untuk kedua model ini tampak paling padat dan terpusat di sekitar garis diagonal. Ini mengindikasikan bahwa baik RandomForest maupun XGBoost mampu menangkap hubungan kompleks dalam data dan menghasilkan prediksi yang konsisten dan akurat di berbagai rentang nilai G3. Kerapatan titik di sekitar garis menunjukkan bahwa error prediksi (perbedaan antara aktual dan prediksi) cenderung kecil dan terdistribusi merata.

Sebaliknya, model Linear Regression yang menggunakan fitur baseline menunjukkan sebaran titik yang lebih luas dan menyebar jauh dari garis diagonal. Ini terutama terlihat pada nilai-nilai G3 yang lebih ekstrem (sangat rendah atau sangat tinggi), di mana prediksi cenderung menyimpang lebih jauh dari nilai aktual. Sebaran yang lebih tersebar tidak merata mencerminkan keterbatasan model linear dalam menangkap hubungan non-linear atau interaksi kompleks antar fitur, yang mengakibatkan nilai RMSE yang lebih tinggi dibandingkan model berbasis pohon atau jaringan saraf.

Model MLP TensorFlow menunjukkan performa yang kuat, berada di antara Linear Regression dan RandomForest/XGBoost. Meskipun titik-titiknya tidak sepadat RandomForest atau XGBoost, mereka tetap menun-

jukkan konsentrasi yang baik di sekitar garis diagonal. Ini menegaskan bahwa MLP adalah model yang efektif dan mampu memberikan prediksi yang cukup akurat, meskipun dalam kasus ini RandomForest sedikit unggul.

Secara keseluruhan, scatter plot ini secara visual mengkonfirmasi metrik evaluasi yang disajikan sebelumnya. Model-model yang memiliki nilai R-Squared tinggi dan RMSE rendah (seperti RandomForest dan XGBoost) ditunjukkan dengan titik-titik yang sangat dekat dengan garis ideal, sementara model dengan metrik yang kurang baik menunjukkan sebaran yang lebih lebar. Visualisasi ini juga membantu mengidentifikasi rentang nilai di mana suatu model mungkin berkinerja kurang baik atau di mana terdapat outlier prediksi.

Diskusi

Hasil yang diperoleh dari uji model menggunakan dataset ini adalah model RandomForest Regressor yang terbaik pada skenario ini yaitu dataset yang memiliki banyak fitur namun jumlah data yang ada pada dataset hanya 395 dimana data ini relatif cukup kecil jika dibandingkan dengan lainnya. Pada dataset kami, kami memiliki keterbatasan untuk menguji model ini pada dataset yang lebih banyak sehingga dapat melihat potensial penuh dari XGboost,dan MLP TensorFlow. Rekomendasi dari kami penulis adalah untuk mencoba model ini pada dataset yang mempunyai data yang lebih banyak daripada dataset ini untuk mendapatkan potensi penuh dari XGboost,dan MLP TensorFlow.

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.