PREDIKSI SKALA KERUSAKAN BANGUNAN DENGAN MENGGUNAKAN PEMBELAJARAN MESIN: PENDEKATAN *MULTI-CLASS CLASSIFICATION*

NAUFAL FARRAS PRATAMA
HELMY LUQMANULHAKIM
ACHMAD RYVALDY

train.csv

Joints Data Competition

Universitas Gadjah Mada

2023

I. PENDAHULUAN

I.1. Latar Belakang

Gempa bumi dapat menyebabkan kerusakan yang signifikan pada bangunan dan fasilitas umum. Ada beberapa jenis gempa bumi yang dapat terjadi, seperti gempa tektonik, gempa vulkanik, gempa runtuhan, gempa jatuhan, dan gempa buatan. Faktor-faktor seperti intensitas seismik, jarak sumber, kepadatan penduduk, kondisi geologi setempat, dan infrastruktur dapat mempengaruhi dampak yang ditimbulkan oleh gempa bumi. Oleh karena itu, penting untuk memahami dampak gempa bumi terhadap bangunan dan cara merancang bangunan yang tahan terhadap guncangan gempa.

Kompetisi prediksi tingkat kerusakan gempa bumi menggunakan data historis dengan berbagai fitur seperti *floors_before_eq (total), old_building*, *plinth_area (ft^2), height_before_eq (ft),* dan fitur lainnya. Analisis data ini dapat membantu dalam prediksi tingkat kerusakan bangunan akibat gempa bumi. Sebagai penunjang kompetisi, Microsoft Azure dan Visual Studio Code merupakan perangkat lunak yang dapat membantu dalam proses eksplorasi data, pengolahan data, dan pengembangan model pembelajaran mesin. Dengan teknologi ini, diharapkan dapat membantu memprediksi kerusakan bangunan secara lebih akurat dan meminimalkan dampak yang ditimbulkan oleh gempa bumi.

I.2. Rumusan Masalah

1. Bagaimana membangun model pembelajaran mesin untuk memprediksi tingkat kerusakan yang akurat?

I.3. Tujuan Penelitian

1. Menemukan model pembelajaran mesin yang sesuai untuk memprediksi skala kerusakan dengan akurasi yang tinggi.
2. Diharapkan penelitian ini dapat memberikan kontribusi dalam meningkatkan pemahaman tentang hal-hal apa sajakah yang mempengaruhi kerusakan bangunan akibat gempa bumi, serta dapat membantu dalam mitigasi dan penanggulangan bencana gempa bumi.

II. PENGOLAHAN DATA

II.1. Interpolasi Data

Data *train* dan *test* memiliki beberapa fitur yang mengandung nilai kosong. Adapun persentase besar nilai kosong yang terdapat di data *train* dan *test* adalah sebagai berikut,

Tabel II-1 Persentase missing values pada data train dan test

Kolom	Persentase (%)		Kolom	Persentase(%)	
	train	*test*		*train*	*test*
floors_before_eq(total)	46	0	legal_ownership_status	17	0
old_building	33	0	has_secondary_use	27	0
plinth_area (ft^2)	58	0	type_of_reinforcement_concrete	40	0
height_before_eq (ft)	46	0	residential_type	37	0
land_surface_condition	42	0	no_family_residing	20	0
type_of_foundation	33	0	public_place_type	0	0
type_of_roof	58	0	industrial_use_type	16	0
type_of_ground_floor	46	0	govermental_use_type	35	0
type_of_other_floor	42	0	flexible_superstructure	9	0
position	43	0	wall_binding	9	0
building_plan_configuration	42	0	wall_material	32	0
technical_solution_proposed	94	0	damage_grade (variabel target)	0	0

Dengan total 723000 baris dan 25 kolom pada data *train* yang memiliki data yang kosong. Data pada fitur yang memiliki data hilang lebih dari 40% akan di buang.

Pada langkah selanjutnya akan dilakukan 2 macam imputasi, imputasi pertama mengisi nilai yang hilang berdasarkan median untuk *non-categorical* dan imputasi kedua menggunakan metode *Feature Based Imputation* yaitu imputasi yang dilakukan dengan mengisi value pada fitur yang kosong pada *categorical feature* berdasarkan fitur tertentu dan diambil modusnya.

II.2. Menggabungkan *unique value* pada tiap kolom

Pada proses pra-pengolahan data dilakukan perhitungan korelasi *train_data* terhadap kolom *type_of_foundation* untuk mengisi nilai yang hilang dengan melihat korelasi antar data hampir sama dalam karakteristiknya. Melakukan

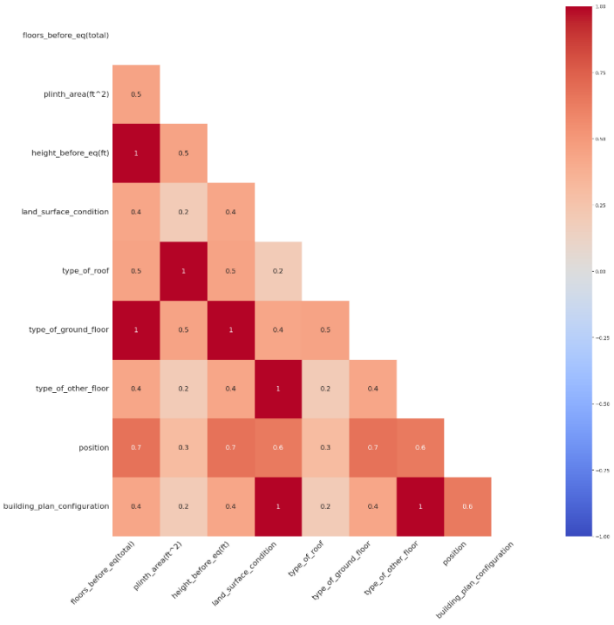
pengelompokan fitur di kolom *legal_ownership_status* yang memiliki distribusi tingkat kerusakan yang relatif sama namun dengan kata berbeda, seperti *private use*, *private*, *prvt*, *privst*. Langkah selanjutnya adalah memanipulasi nilai fitur-fitur yang telah dikelompokkan menjadi 4 (*private*, *public*, *institutional* dan *other*), Adapun beberapa kolom lain yang nilainya memiliki karakteristik mirip berdasarkan analisis data yang telah dilakukan akan digabungkan menjadi 1 kategori nilai.

II.3. Mengganti tipe data

Pada data *train* dan *test* diberlakukan proses merubah tipe data pada masing-masing fitur sesuai dengan nilai kolom yang dimuat.

III. EKSPLORASI ANALISIS DATA

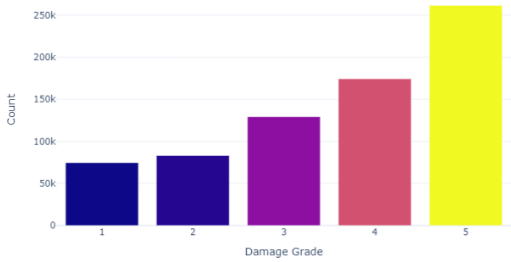
III.1. Mengukur korelasi *Nullity value* tiap kolom



Gambar III-1 Grafik korelasi tiap kolom yang memiliki nilai kosong

Diatas merupakan grafik korelasi *nullity value* antar kolom. Terlihat cukup banyak nilai korelasi yang menyentuh angka maksimal, sehingga dapat disimpulkan bahwa beberapa baris, mungkin memiliki nilai kosong yang memang tidak dapat diimputasi.

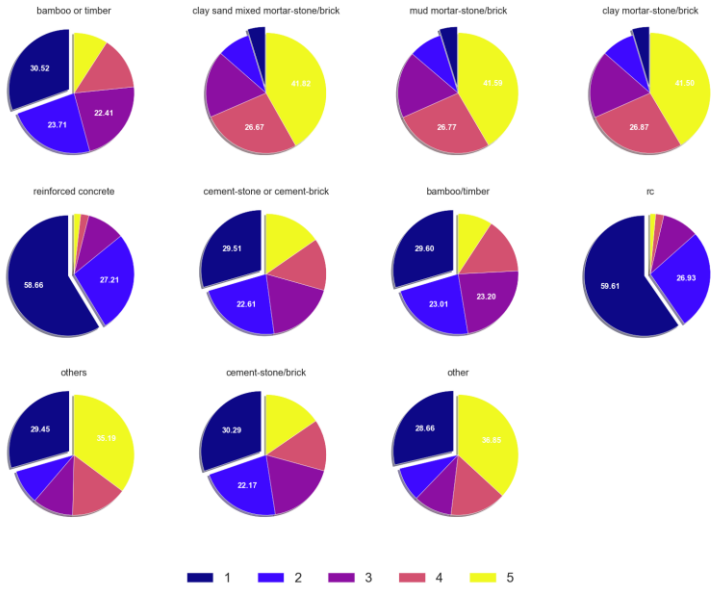
III.2. Nilai distribusi skala kerusakan pada data



Gambar III-2 Grafik distribusi damage_grade

Diatas merupakan grafik distribusi dari label *damage_grade* pada data *train*. Dari grafik tersebut dapat disimpulkan bahwa kerusakan dengan skala 5 paling banyak terjadi pada data dibanding kerusakan yang berada dibawahnya. Hal ini menandakan data yang terdistribusi pada data *train* didominasi oleh data bangunan yang memiliki kerusakan berat.

III.3. Analisis tipe pondasi dengan skala kerusakan yang diterima

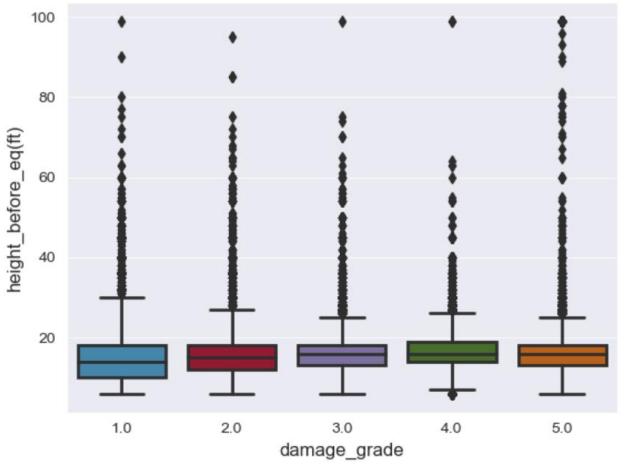


Gambar III-3 Grafik Korelasi type_of_foundation dengan damage_grade

Diatas merupakan grafik korelasi antara *type_of_foundation* dengan *damage_grade*. Dapat disimpulkan, bahwa *rc* memiliki tingkat ketahanan terhadap gempa cukup tinggi, hal ini dikarenakan pada kerusakan skala 1 didominasi oleh jenis pondasi *rc* dan pada skala kerusakan 5 jenis pondasi *rc* jauh lebih sedikit muncul. Pada data tersebut juga terdapat beberapa jenis

fondasi yang memiliki pola yang sama, data seperti ini dapat diagregrasi atau digabungkan menjadi 1 jenis pondasi.

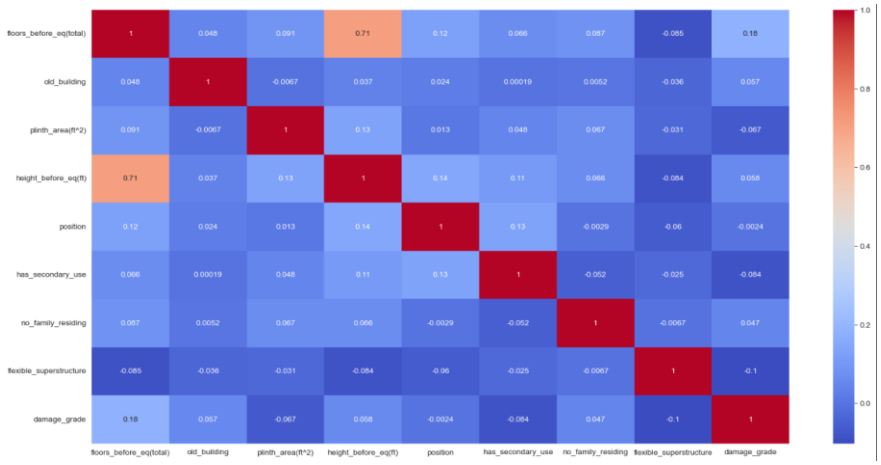
III.4.Distribusi height_before_eq terhadap skala kerusakan



Gambar III-4 Grafik box-plot antara height_before_eq dengan damage_grade

Dapat diamati pada grafik diatas memiliki *outlier* yang cukup tinggi pada data ketinggian bangunan sebelum gempa, hal ini dapat menurunkan korelasi antara variabel bebas dengan variabel independen dan juga mengacaukan prediksi pada model yang akan kita gunakan.

III.5. Analisis Multivariat



Gambar III-5 Heatmap korelasi antar fitur

Di atas merupakan plot *heatmap* yang menyatakan korelasi *pearson* antar variabel. Jika dipilih variabel terikat adalah *damage_grade* dan yang lain merupakan variabel independen maka variabel yang memiliki korelasi paling positif dengan variabel terikat adalah *floor_before_eq(total)* dan variabel yang memiliki korelasi paling negatif dengan variabel terikat adalah

flexible_superstructure. Pada variabel *floors_before_eq(total)* dengan variabel *height_before_eq(ft)* memiliki korelasi yang cukup tinggi sehingga ada kemungkinan kedua variabel ini saling mempengaruhi satu sama lain

IV. REKAYASA FITUR

IV.1. Membagi kehadiran kolom *technical_solution_proposed*

Melihat banyaknya nilai kosong pada kolom *technical_solution_proposed*, yang menyentuh persentase di angka 93% nilai yang kosong, diberlakukan distribusi nilai kehadiran berdasarkan *unique value* menggunakan nilai *binary*.

IV.2. Mengubah *categorical* menjadi *numerical*

Untuk mengubah *categorical feature menjadi numerical feature* perlu dilakukan *encoding*. *Encoding* yang perlu dilakukan pada data train dan test ada 2 jenis, yaitu ordinal *encoding* dan *one-hot encoding*. Ordinal *encoding* dilakukan pada data yang valuenya memiliki tingkatan seperti *type_of_foundation,* dan untuk *one-hot encoding* dilakukan pada data yang memiliki nilai yang setara. Hal ini dilakukan untuk mempermudah model pembelajaran mesin untuk mengenali pola data.

IV.3. Mengurutkan ordinal *value*

Dikarenakan ordinal *value* merepresentasikan tingkatan nilai, diperlukan pengurutan berdasarkan *damage_grade* dengan menghitung persentase jumlah kehadiran damage grade pada *unique value* kolom, lalu dikalikan dengan skala kerusakan.

IV.4. Menambahkan fitur baru

Setelah melakukan EDA didapatkan beberapa informasi yang dapat digunakan untuk membuat fitur baru untuk meningkatkan akurasi prediksi pada model yang akan digunakan. Beberapa fitur yang ditambahkan antara lain :

- *Height_per_floor*

  Dikarenakan fitur *floor_before_eq(total)* memiliki korelasi yang cukup baik dengan *hight_before_eq(ft)* maka dibuatlah fitur baru sebagai perbandingan antara kedua variabel tersebut yang didefinisikan sebagai berikut :

$$f(x, y) = \frac{x}{y} \times 0.1$$

x = *floor_before_eq(total)*

y = *height before_eq(ft)*

0.1 = untuk memperbaiki permasalahan *floating point*

- *Pressure*

  Menggunakan fitur *plinth_area(ft2)*, *height_before_eq(ft)* dan konstanta gravitasi standar (bumi) maka didapat fitur baru yaitu *pressure* sebagai fitur tekanan yang dialami oleh bangunan. Fitur ini didefinisikan sebagai berikut :

$$f(x,y) = g \times (0.92093 \times x \times y \times 0.3048)$$

g = konstanta gravitasi standar (9.80665 m/s2)

x = *plinth_area(ft2)*

y = *height_before_eq(ft)*

0.92093 = *conversion factor feet square to meter square*

0.3048 = *conversion factor feet to meter*

IV.5. Melakukan scaling

Untuk memudahkan proses komputasi model pembelajaran mesin, diperlukan proses scaling. Melihat hasil distribusi *box-plot* yang cenderung *right skewed* dengan beberapa kehadiran *outlier*. Diberlakukan RobustScaler pada beberapa data numerik. Dengan fungsi yang didefinisikan sebagai berikut :

$$x'_i = \frac{x_i - Q1(x)}{Q3(x) - Q1(x)}$$

V. PEMODELAN

V.1. Pembagian Data

Membagi data menjadi train set dan test set adalah salah satu teknik penting dalam pengolahan data. Tujuannya adalah untuk menguji kinerja model pembelajaran mesin yang dibuat sebelum diimplementasikan pada data yang belum dikenal. Dalam pembuatan model pembelajaran mesin, data dibagi menjadi dua bagian*: train set* dan *test set*. Pada proses pengolahan kali ini, akan digunakan rasio 80:20, di mana 80% data akan digunakan sebagai *train set* dan 20% sebagai *test set*.

V.2. Menentukan model pembelajaran mesin

Tabel V-1 Perbandingan akurasi model pembelajaran mesin

Model Pembelajaran Mesin	Akurasi
Random Forest	0.34727
Extreme Random Trees	0.42257
LightGBM	0.42151
Gradient Boosting	0.41383
Logistic Regression	0.45242
XGBoost Classifier	0.46896

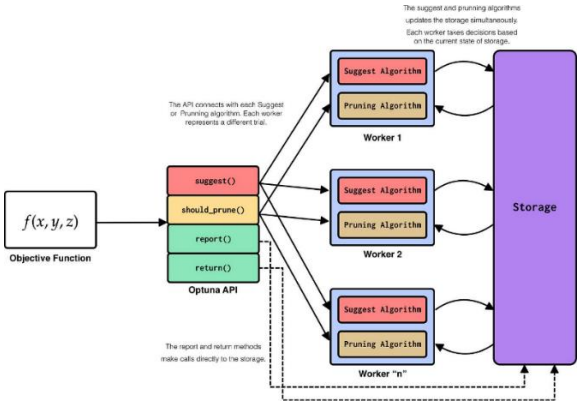
Didapati bahwa, XGBoost memiliki nilai akurasi yang cukup tinggi terhadap data yang digunakan.

VI. EVALUASI DAN VALIDASI

VI.1. *Hyperparameter Tuning*

Hyperparameter tuning merupakan perhitungan parameter untuk mendapatkan nilai metrik terbaik. Pencarian paremeter menggunakan optuna untuk model XGBoost. Metode optuna menggunakan distribusi probabilitas untuk mempertimbangkan parameter dan mengkombinasi secara acak lalu memperbarui distribusi probabilitas sampai menemukan hasil kinerja terbaik.

Gambar VI-1 Cara kerja Optuna

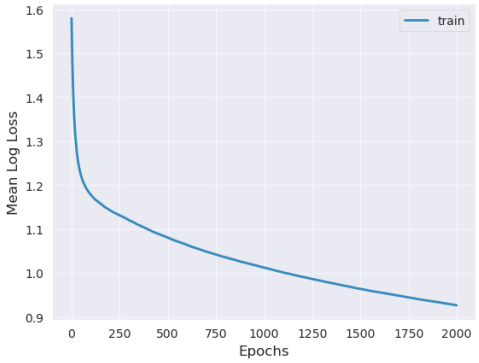


Tabel VI-1 Hyperparameter terbaik yang didapat optuna

Parameter	Nilai parameter
Max depth	9
Learning_rate	0.092

N_estimators	2000
Min_child_weight	1
gamma	0.89
subsample	0.92
Colsample_bytree	0.393
Reg_alpha	0.56
Reg_lambda	0.21
Max_delta_step	1

Gambar VI-2 Grafik kinerja pembelajaran mesin XGBoost



Dari hasil *hyperparameter tuning* menggunakan optuna untuk model XGboost, diperoleh akurasi sebesar 0.48 yang sebelum diberlakukan hyperparameter tuning hanya memperoleh akurasi sebesar 0.46. Dengan *mlogloss* terkecil 0.92595.

VI.2. Laporan akurasi tiap *class*

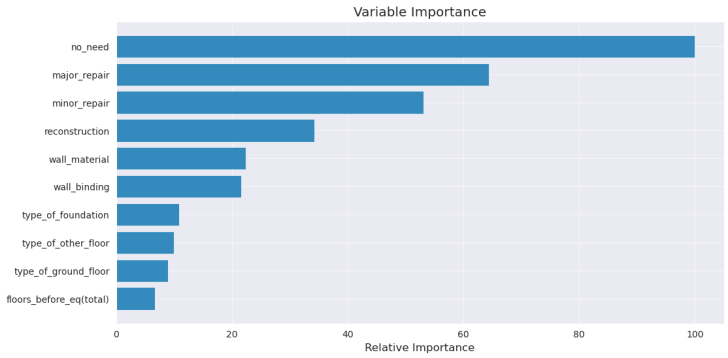
Tabel VI-2 Nilai akurasi tiap class damage_grade

damage_grade	precision	recall	F1-score	support
1	0.57	0.60	0.59	7390
2	0.43	0.60	0.30	8083
3	0.44	0.23	0.31	12453
4	0.38	0.27	0.32	15979
5	0.50	0.80	0.62	23186

Dari laporan akurasi tiap *class* tersebut didapati bahwa, model cenderung memprediksi dengan akurat skala kerusakan tingkat 5. Hal ini berkaitan dengan eksplorasi data yang sebelumnya telah dilakukan.

VI.3. Fitur penting pada model

Gambar VI-3 Variabel paling berpengaruh pada model



Dapat disimpulkan bahwa model yang digunakan merupakan model yang benar benar cocok dengan data yang diberikan.

VII. PENUTUP

VII.1. Kesimpulan

Berdasarkan validasi yang telah dilakukan, data yang digunakan pada kompetisi ini memiliki akurasi yang cukup baik dengan model pembelajaran mesin berbasis *Extreme Gradient Boosting* (XGBoost). Dengan nilai akurasi berdasarkan *cross-validation* sebesar 0.48. Pada data, diperoleh bahwa variabel yang mempengaruhi besar tidaknya kerusakan pada bangunan adalah :

1. *Technical Solution* = Solusi yang ditawarkan untuk bangunan yang terdampak gempa.
2. *Wall Material* = Material dasar sebagai pembangun dinding.
3. *Wall Binding* = Material yang digunakan sebagai perekat bahan pembentuk dinding.
4. *Type of Foundation* = Jenis pondasi yang dipakai untuk bangunan.
5. *Type of Other Floor* = Jenis lantai yang dipakai untuk selain ground-floor (lantai dasar).

DAFTAR PUSTAKA

Roiz-Pagador, J., Chacon-Maldonado, A., Ruiz, R., & Asencio-Cortes, G. (2022). *Earthquake Prediction in California Using Feature Selection Techniques* (pp. 728–738). https://doi.org/10.1007/978-3-030-87869-6_69

Petrozziello, A., & Jordanov, I. (2019). *Feature Based Multivariate Data Imputation* (pp. 26–37). https://doi.org/10.1007/978-3-030-13709-0_3

Han, W., Gan, Y., Chen, S., & Wang, X. (2020). Study on Earthquake Prediction Model Based on Traffic Disaster Data. *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)*, 331–334. https://doi.org/10.1109/ICSESS49938.2020.9237667

Zulfiar, M. H., & Zai, M. I. I. (2021). Penilaian Kerentanan Bangunan Terhadap Gempa Bumi pada Gedung Perkuliahan Berlantai Tinggi di Yogyakarta. *Bulletin of Civil Engineering*, *1*(2), 73–80. https://doi.org/10.18196/bce.v1i2.11075