Analisis dan Prediksi Tingkat Kerusakan Bangunan Akibat Gempa Bumi menggunakan Metode eXtreme Gradient Boosting (XGBoost)

Samatha Marhaendra Putra, Daffa Bil Nadzary

1. Pendahuluan

1.1. Latar Belakang

Gempa merupakan fenomena bencana alam di mana terdapat pergeseran lempeng bumi yang mengakibatkan terjadinya pelepasan energi secara spontan dan menghasilkan gelombang seismik yang bersifat destruktif [1]. Berdasarkan data dari BPS, di sepanjang tahun 2021 terdapat 10.519 gempa yang terjadi di Indonesia, mulai dari gempa berskala kecil hingga besar. Gempa yang berskala kecil umumnya tidak bersifat destruktif, namun sudah cukup untuk dapat dirasakan secara langsung oleh masyarakat sekitar. Sedangkan untuk gempa yang berskala 7.0 SR dapat merubuhkan hingga bangunan-bangunan sekitar. Kerusakan ini tidak hanya mengakibatkan kerugian secara materil, namun juga secara korban jiwa.

Salah satu cara untuk menanggulangi jumlah korban serta dampak dari gempa adalah dengan membuat sistem yang dapat memberikan *insight* berupa karakteristik bangunan yang memiliki tingkat kerusakan yang rendah serta dapat memprediksi tingkat kerusakan bangunan tersebut berdasarkan data karakteristik yang telah diberikan.

1.2. Tujuan

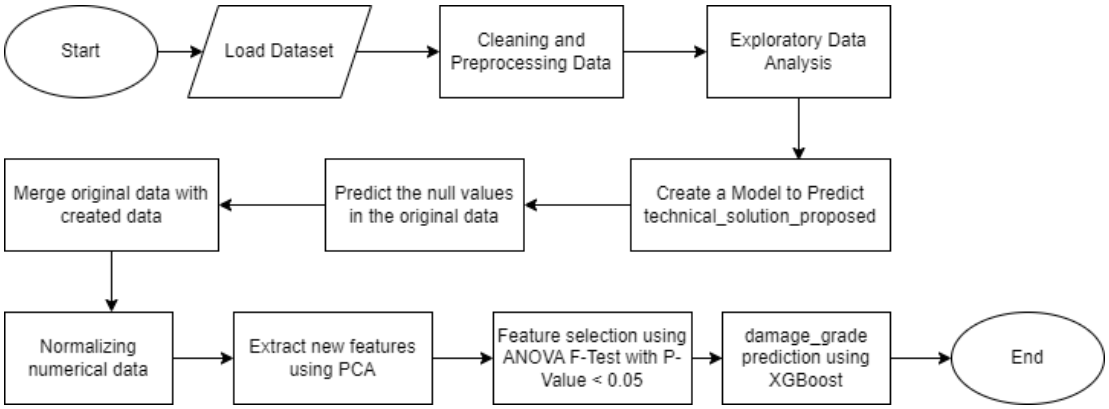
Tujuan dilakukannya penelitian ini adalah:

1.2.1. Menganalisis karakteristik bangunan yang memiliki tingkat kerusakan yang rendah terhadap gempa berskala tinggi

1.2.2. Membuat suatu model yang mampu memprediksi tingkat kerusakan dari sebuah bangunan terhadap gempa berdasarkan karakteristik-karakteristik serta perlakuan terhadap bangunan tersebut.

2. Metodologi

2.1 Alur Pengerjaan



Gambar 2. 1 *Flowchart* Pengerjaan Submisi.

Alur pengerjaan yang dilakukan pada pengerjaan ini diawali dengan pengumpulan dan pemrosesan data seperti pembersihan dan praproses awal. Data hasil praproses akan dianalisis terlebih dahulu melalui tahap *Exploratory Data Analysis* untuk melihat fitur-fitur yang memiliki tingkat korelasi yang tinggi terhadap target. Berdasarkan hasil EDA, akan dibuat model pertama yang berfungsi untuk memprediksi fitur *technical_solution_proposed,* dengan tujuan untuk memperbanyak jumlah baris data dalam membuat model untuk memprediksi *damage_grade*. Dataset baru yang berisikan nilai-nilai non-null hasil proses *impute* akan kemudian diikuti dengan proses *feature engineering*. Hasil *feature engineering* akan dimasukkan ke dalam model dan divalidasi menggunakan teknik *cross-validation* untuk melihat performa dari model tersebut. Tahap terakhir yakni pembuatan *file* submisi yang akan dikumpulkan pada *platform* Kaggle InClass Competition.

2.2 Pengumpulan Data

Dataset yang dipakai dalam pengerjaan ini adalah *dataset* yang telah disediakan oleh penyelenggara kompetisi. *Dataset* berisikan 720.000 baris data dengan 24 kolom dengan tipe numerik serta kategorikal dan 1 kolom label, yakni *damage_grade*, yang merupakan fitur kategorikal berisikan indeks tingkat kerusakan dari sebuah gempa dari 1.0 hingga 5.0.

2.3 *Preprocessing*

Proses pembersihan awal dimulai dari membuang kolom yang tidak berhubungan sama sekali terhadap label. Kemudian dilakukan proses pembuangan nilai *null* yang terdapat pada *dataset* training. Subset dari hasil *dropping* kemudian diolah berdasarkan fitur. Beberapa fitur numerik dengan nilai yang salah di-*mapping* terhadap bentuk tipe data yang sebenarnya, sedangkan beberapa fitur kategorikal yang memiliki

nilai serupa dengan nama yang berbeda akan digabung menjadi satu kategori yang sama, menyisihkan jauh lebih sedikit nilai unik.

2.4 *Exploratory Data Analysis*

EDA atau *Exploratory Data Analysis* adalah kegiatan untuk menganalisis sekumpulan data untuk melihat karakteristik utamanya. EDA bertujuan untuk melihat intisari dan pola-pola dari data yang bermanfaat dalam proses *modeling* atau eksperimen pada data [2]. Pada penelitian ini, peneliti menggunakan hasil visualisasi data dalam proses EDA untuk menentukan fitur-fitur yang penting serta membuat model yang berhubungan terhadap fitur tersebut.

2.5 *Feature Engineering*

Tahap *feature engineering* dilakukan untuk memilih, mengevaluasi, serta mengubah data agar lebih sesuai sebelum dimasukkan ke dalam proses pemodelan dengan memilih hanya beberapa fitur dengan tingkat *importance* yang sesuai. Pada tahap ini, dilakukan normalisasi nilai-nilai pada fitur numerik dengan menggunakan berbagai *scaler* untuk menyetarakan data yang sebarannya tidak seimbang (*skewed*). Setelah itu, dilakukan juga ekstraksi fitur baru dengan menggunakan teknik PCA atau *Principal Component Analysis*. Hasil fitur campuran baik dari PCA maupun fitur original akan diuji multikolinearitasnya dengan menggunakan uji ANOVA F-Test untuk memilih fitur-fitur yang penting untuk dimasukkan ke dalam model.

2.6 *Modeling*

Proses *modeling* yang dilakukan pada penelitian ini adalah menggunakan model eXtreme Gradient Boost (XGBoost). XGBoost merupakan model *machine learning* berbasis *gradient-boosted decision tree* (GBDT) yang bersifat *scalable.* Model XGBoost dipilih karena kemampuannya yang tidak hanya bagus dalam kasus regresi, namun juga kasus klasifikasi sebagaimana fokus dari penelitian ini.

Model XGBoost di-*fit* menggunakan data yang sudah diubah dan dipilih dari proses *feature engineering* sebelumnya, dan di-*tuning* lebih lanjut untuk memperoleh hasil yang maksimal. Proses *tuning* dilakukan dengan membuat sebuah *search space* yang berisikan parameter-parameter yang akan di-*tuning.* Adapun proses *hyperparameter tuning* yang dilakukan adalah dengan menggunakan pendekatan Bayesian Optimization dengan mengatur *delta stopper* untuk menghasilkan hasil kombinasi parameter yang terbaik.

2.7 *Validation*

Validasi yang dilakukan pada penelitian kali ini adalah menggunakan 5-*fold cross-validation* dengan ukuran evaluasinya adalah F1-Score Macro. *Cross-validation*

atau validasi silang adalah teknik yang digunakan untuk mengevaluasi model *machine learning* dengan data yang dibagi menjadi dua bagian; bagian pertama untuk pelatihan model, dan bagian kedua untuk validasi performa model. Dalam validasi silang, *set* pelatihan dan validasi harus saling bersilangan dalam putaran yang berurutan sehingga setiap titik data memiliki peluang untuk divalidasi [3]. Validasi silang diperlukan untuk melihat apakah model yang dibuat bersifat *overfitting* atau tidak. Gambar 2.2 menunjukkan contoh skema dari validasi silang.

Gambar 2.2 Skema 5-*fold Cross-validation*

3. Pembahasan

3.1 *Preprocessing*

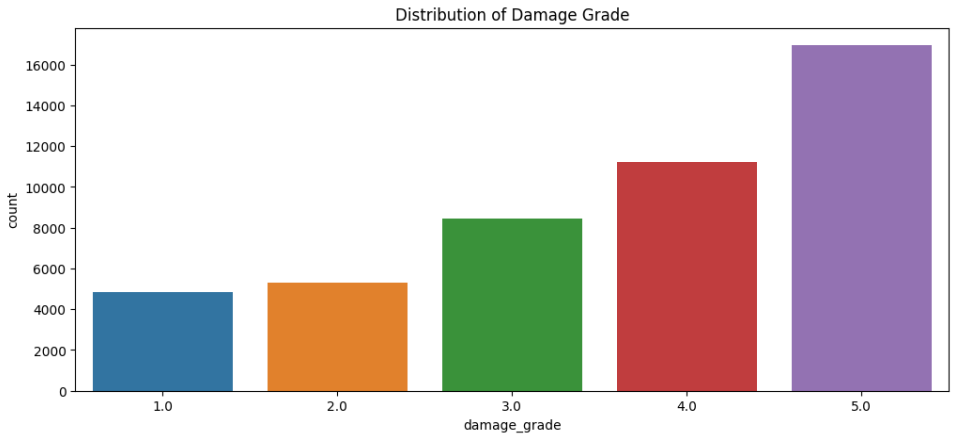
Terdapat beberapa hasil dari proses *preprocessing* ini. Pertama, beberapa fitur numerikal seperti ‘floors_before_eq (total)’ dan ‘plinth_area (ft^2)’ dimanipulasi lebih lanjut agar lebih mudah diolah. Untuk fitur ‘floors_before_eq (total)’, bentuk awal yang disediakan adalah string dengan beberapa kata kunci seperti “floor” dan “story”, menandakan jumlah lantai pada bangunan. Untuk itu, dibangun sebuah *mapping dictionary* untuk memetakan nilai-nilai tersebut ke bentuk nominalnya. Sedangkan pada fitur 'plinth_area (ft^2)’, nilai pada kolom juga mula-mula merupakan string, dengan format ‘(area) ft^2’. Namun, untuk luas yang berada di atas 1000 ft^2, ditulis sebagai ‘More than 1000 ft^2’. Untuk itu, pemrosesan dilakukan dengan dua kondisi, untuk luas yang berada di bawah 1000, maka cukup menghapus string ‘ft^2’, sedangkan untuk luas di atas 1000 akan dibuat rata menjadi 1000.

Sedangkan pada kolom kategorikal, secara umum pemrosesan yang dilakukan serupa, di mana terdapat beberapa kolom dengan nilai yang serupa, namun tidak dalam satu grup. Seperti contoh pada fitur ‘type_of_foundation’, terdapat baris data dengan nilai ‘Bamboo or Timber’ dan ‘Bamboo/Timber’, yang tentu saja merupakan satu hal yang sama. Oleh karena itu, dibuat juga beberapa *mapping dictionary* yang diterapkan pada hampir keseluruhan fitur dengan nilai unik yang tidak lazim.

Terakhir, untuk memastikan data dapat digunakan sesuai dengan jenisnya, pengubahan jenis tipe data dilakukan. Untuk kolom numerikal yang sebelumnya memiliki tipe data *object* diubah menjadi *float*, sedangkan beberapa kolom kategorikal diubah menjadi *object*.

3.2 *Exploratory Data Analysis*

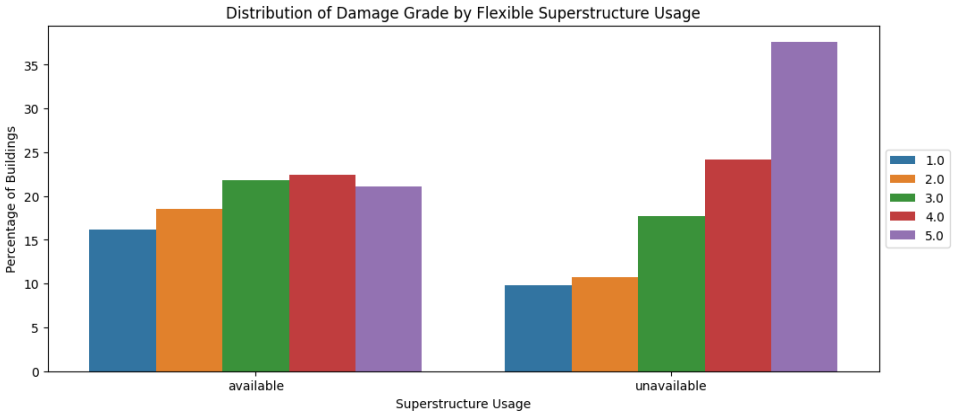
- Distribusi target 'damage_grade'

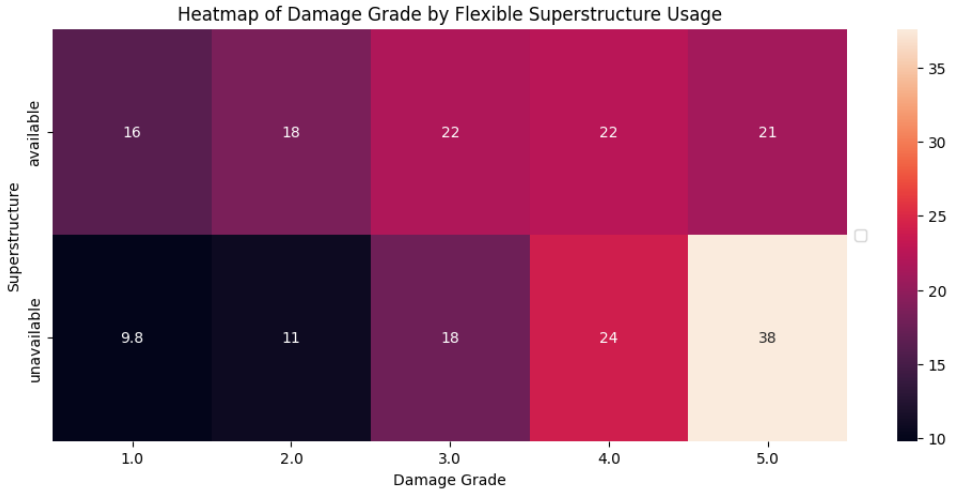


Gambar 3.1 Persebaran fitur target 'damage_grade'

Dapat dilihat pada Gambar 3.1 bahwa distribusi dari target pada dataset cenderung naik seiring dengan meningkatnya tingkat kerusakan. Jika dibandingkan dengan tingkat 1.0 yang hanya berjumlah sekitar 5.000 data, data dengan kerusakan tingkat 5.0 memiliki jumlah sebanyak 16.000 data. **Hal ini mengindikasikan bahwa diperlukan proses *undersampling* terlebih dahulu untuk menghilangkan bias pada model saat proses *training*.**

- Analisis Penggunaan Superstruktur pada Tingkat Kerusakan

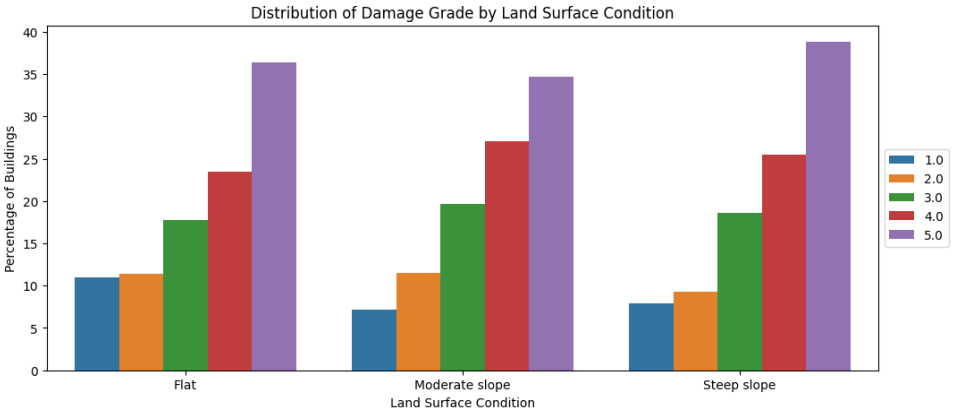




Gambar 3.2 Pengaruh penggunaan superstruktur pada tingkat kerusakan

Gambar 3.2 menunjukkan persebaran dari fitur ’flexible_superstructure’ terhadap target. Dari persebaran data sendiri menunjukkan bahwa 90% dari bangunan tidak menggunakan superstruktur pada kerangka bangunannya, sedangkan sisanya menggunakan. Dari gambar di atas, terlihat bahwa **penggunaan superstruktur dapat mengakibatkan kerusakan yang lebih minimal dibandingkan dengan bangunan yang tidak menggunakan superstruktur.**

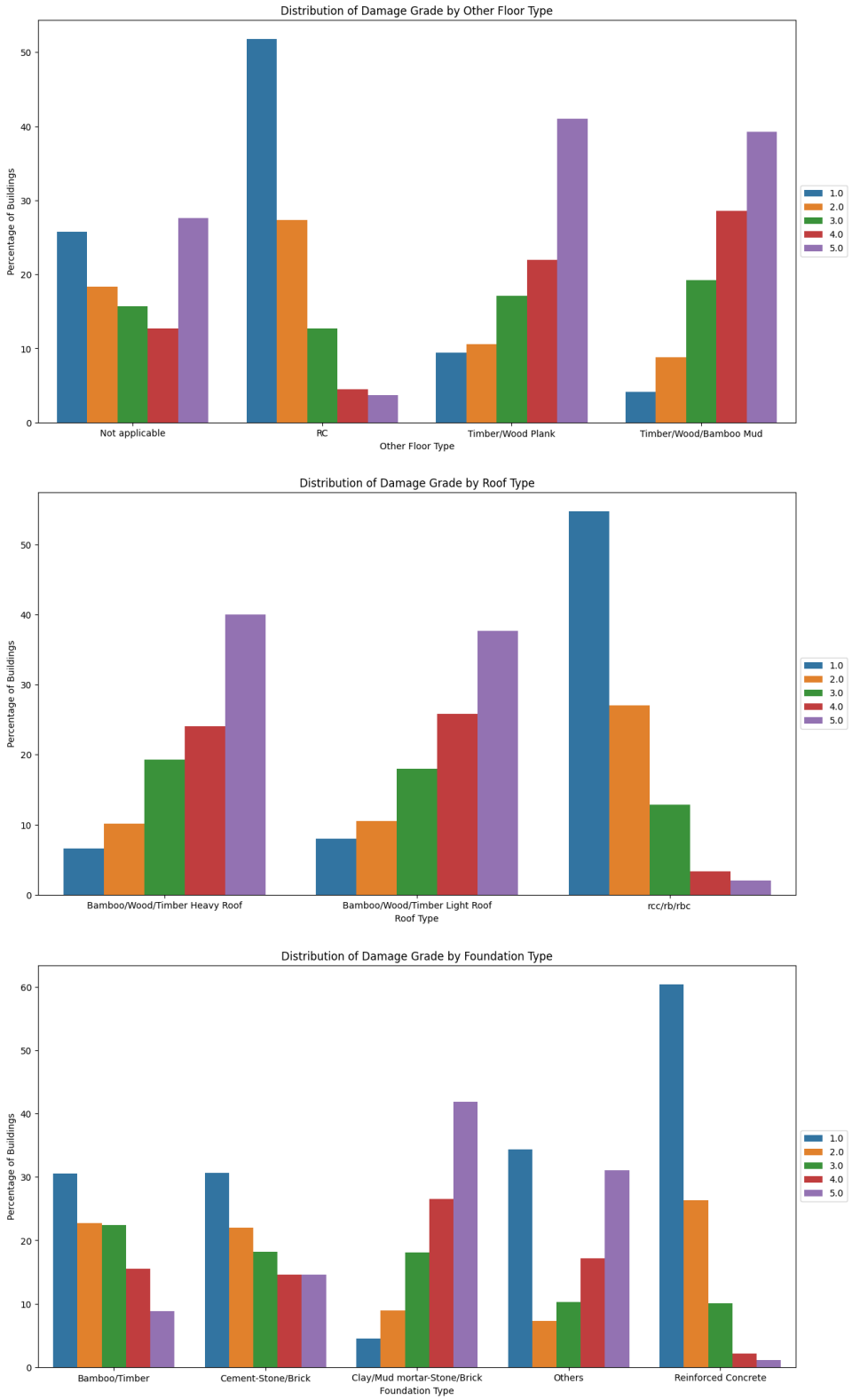
- Analisis Kondisi Lahan Tanah pada Tingkat Kerusakan



Gambar 3.3 Pengaruh kondisi struktur lahan tanah pada tingkat kerusakan

Gambar 3.3 menunjukkan persebaran dari fitur ’land_surface_condition’ terhadap target. Berbeda dengan fitur sebelumnya, pada fitur ini tidak menunjukkan adanya perbedaan yang signifikan terhadap distribusi dari kerusakan yang terjadi. **Hal ini mengindikasikan bahwa kondisi tanah tidak mempengaruhi kerusakan yang terjadi**.

- Analisis Bahan Bangunan pada Tingkat Kerusakan

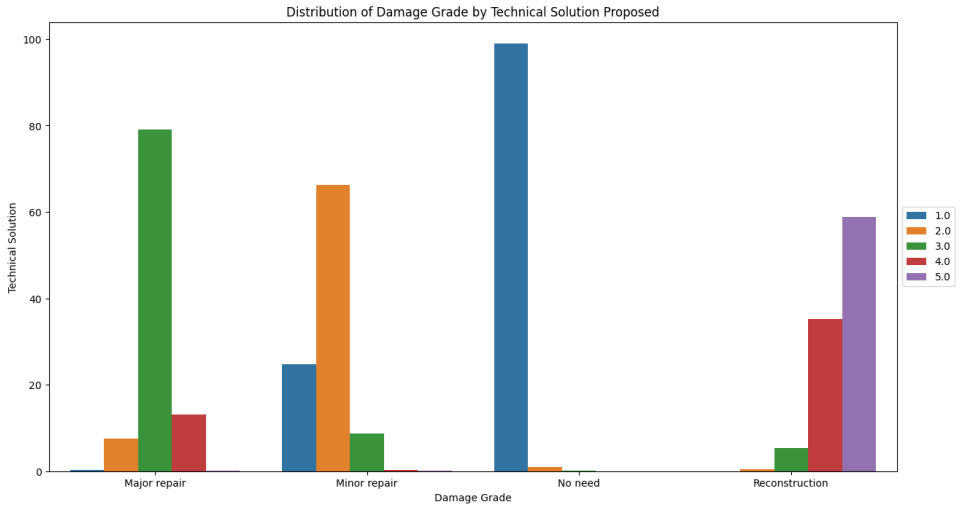


Gambar 3.4 Pengaruh bahan bangunan yang digunakan pada tingkat kerusakan

Gambar 3.4 menunjukkan persebaran dari fitur beberapa fitur seperti 'type_of_foundation', 'type_of_roof', 'type_of_ground_floor', serta

’type_of_other_floor’ yang menjelaskan bahan-bahan yang digunakan pada tiap bagian dari bangunan. Terdapat beberapa jenis bahan yang umum digunakan pada bangunan yang terdapat di dataset, yaitu *reinforced concrete*/beton bertulang, batu bata, hingga papan kayu dan bambu. **Gambar di atas mengindikasikan bahwa penggunaan beton sebagai bahan bangunan dapat menghasilkan bangunan yang lebih kokoh terhadap gempa.**

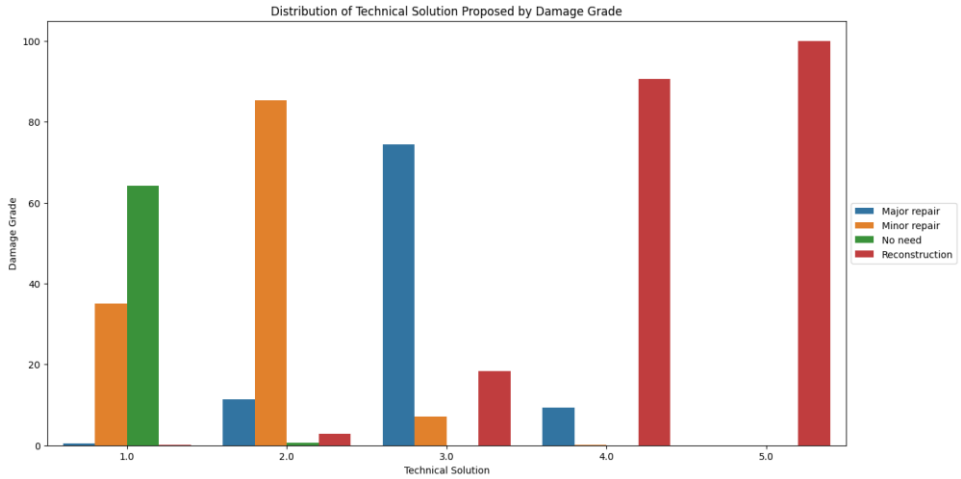
- Analisis Solusi Rekonstruksi pada Tingkat Kerusakan



Gambar 3.5 Hubungan solusi rekonstruksi terhadap tingkat kerusakan

Gambar 3.5 menunjukkan persebaran dari fitur ’technical_solution_proposed’ yang menjelaskan solusi-solusi rekonstruksi yang disarankan pada bangunan yang terdampak gempa. Terdapat beberapa kategori dari solusi ini, seperti rekonstruksi ulang, reparasi secara besar, reparasi secara kecil, hingga tidak ada reparasi sama sekali. **Gambar di atas mengindikasikan bahwa tingkat kerusakan yang dialami sangat mempengaruhi solusi yang disarankan.**

Berdasarkan analisis sementara dari beberapa fitur, terdapat satu kolom yang unik, yakni ’technical_solution_proposed’. Kolom ini memiliki nilai-nilai yang masing-masing terkait satu sama lain dengan tingkatan kerusakan yang terjadi. Gambar 3.6 menunjukkan relasi yang lebih jelas antara kedua fitur ini.



Gambar 3.5 Hubungan tingkat kerusakan terhadap solusi rekonstruksi

Berdasarkan kausalitas, kerusakan pada bangunan yang diakibatkan oleh gempa terlebih dahulu akan dinilai terlebih dahulu sebelum kemudian diberi solusi yang sesuai terhadap rekonstruksi dari bangunan tersebut. Jika dilihat pada gambar di atas, terlihat bahwa untuk tingkatan kerusakan 1.0, mayoritas tindakan yang diberikan adalah 'tidak ada tindakan' dan 'reparasi secara kecil'. Kemudian pada tingkatan 2.0, mayoritas tindakan yang diberikan adalah 'reparasi secara kecil', pada tingkatan 3.0 yaitu 'reparasi secara besar', dan tingkatan 4.0 dan 5.0 memiliki mayoritas tindakan yang sama yaitu 'rekonstruksi'.

Adanya korelasi yang tinggi ini mengakibatkan fitur ini menjadi sangat penting dalam proses pemodelan. Namun, berdasarkan eksplorasi lebih jauh, didapat bahwa fitur ini hanya memiliki sekitar 48.000 nilai non-*null,* mengakibatkannya sebagai fitur dengan nilai kosong paling banyak di antara keseluruhan fitur, sekitar 90% dari keseluruhan data. Berdasarkan informasi ini, penulis mengambil langkah selanjutnya yaitu **membuat model yang dapat memprediksi solusi rekonstruksi untuk memperbanyak jumlah baris pada data**.

3.3 *Feature Engineering*

Berdasarkan hasil dari EDA yang dilakukan sebelumnya, dibuat model yang digunakan untuk memprediksi fitur 'technical_solution_proposed' terlebih dahulu, yang di-*train* menggunakan 15.000 data hasil *undersampling* dan proses *feature engineering* lainnya secara terpisah. Dengan skor F1 macro sebesar 0.82, model digunakan untuk mengisi nilai *null* yang terdapat pada mayoritas data. Hasil praproses

yang dilakukan pada dataset yang baru ini berjumlah sekitar 300.000 data yang kemudian dipangkas menjadi 180.000 baris data melalui teknik *undersampling*.

Hasil dari proses *feature engineering* adalah *dataset* yang sudah dilakukan manipulasi sedemikian rupa yang siap dimasukkan ke dalam model untuk proses *training* serta *fitting*. Kolom-kolom seperti 'old_building' serta 'plinth_area (ft^2)' yang memiliki nilai yang cenderung ekstrem dan distribusi yang *skewed* dinormalisasi dengan menggunakan tiga macam *scaler*, yakni RobustScaler(), StandardScaler(), serta MinMaxScaler(). Hasilnya kedua fitur serta fitur numerik lainnya ternormalisasi.

Selanjutnya pada kolom kategorikal dilakukan proses *feature encoding* dengan teknik *one hot encoding* yang bertujuan untuk memanipulasi fitur kategorikal ke dalam bentuk yang lebih mudah diinterpretasikan oleh model *machine learning*. Selanjutnya, diekstrak juga fitur baru dengan menggunakan PCA atau *Principal Component Analysis*. Dengan mempertimbangkan *trade off* antara waktu serta akurasi, fitur yang didapat dari hasil PCA digabung dengan fitur original untuk kemudian diuji menggunakan uji ANOVA F-Test. Proses *feature selection* dengan menggunakan uji ANOVA F-Test ini dilakukan dengan menggunakan P-Value < 0.05. Dari hasil *feature selection*, jumlah akhir fitur yang akan digunakan untuk memprediksi fitur 'damage_grade' adalah sebanyak 88 fitur.

3.4 *Modeling*

Model XGBoost di-*train* menggunakan 180.000 baris data dan 88 fitur prediktor dengan 1 fitur target dengan objektif yaitu F1 Score Macro. Pada proses *hyperparameter tuning* dengan Bayesian Optimization, dilakukan pembuatan *search space* terlebih dahulu yang berisikan rentang nilai dari setiap parameter yang ingin dicari kombinasi yang mampu menghasilkan skor F1 yang paling baik. Potongan kode di bawah menunjukkan *search space* dari model.

```
[ ]  search_spaces = {'learning_rate': Real(0.01, 1.0, 'uniform'),
                      'max_depth': Integer(2, 12),
                      'subsample': Real(0.1, 1.0, 'uniform'),
                      'colsample_bytree': Real(0.1, 1.0, 'uniform'), # subsample ratio of columns by tree
                      'reg_lambda': Real(1e-9, 100., 'uniform'), # L2 regularization
                      'reg_alpha': Real(1e-9, 100., 'uniform'), # L1 regularization
                      'n_estimators': Integer(50, 5000)
        }
```

3.5 *Validation*

Hasil cross-validation yang dilakukan sekaligus melakukan tuning pada model memperlihatkan bahwa model XGBoost memiliki skor CV sebesar 0.78 pada *validation set* yang diambil dari *training set*, dengan skor F1 sebesar 0.68 pada *test set* yang di-*submit* melalui *platform* kompetisi Kaggle.

4. Kesimpulan dan Saran

a. Terdapat fitur dengan korelasi yang cukup besar terhadap tingkat kerusakan yang terjadi pada bangunan. Fitur dengan **korelasi yang paling besar** ditunjukkan oleh fitur 'technical_solution_proposed'.

b. Dari hasil prediksi tingkat kerusakan dengan menggunakan model XGBoost yang di-*tuning,* diperoleh bahwa model dapat menghasilkan skor **F1 Score Macro** sebesar 0.68 pada *test set* di *public leaderboard* submisi Kaggle (35% dari data tes).

c. Hasil EDA lebih lanjut menunjukkan **bahwa penggunaan bahan beton** sebagai bahan utama dalam konstruksi bangunan dapat meminimalisir tingkat kerusakan yang dialami yang diakibatkan oleh bencana alam seperti gempa, terutama gempa dengan ukuran besar.

5.
6.
7.
8.
9.
10.
11.
12.

Referensi

[1] W. H. K. Lee, International Association of Seismology and Physics of the Earth's Interior, and International Association for Earthquake Engineering, Eds., International handbook of earthquake and engineering seismology. Amsterdam ; Boston: Academic Press, 2002.

[2] Chatfield, C. Problem Solving: A Statistician's Guide (2nd ed.). Chapman and Hall. 1995. ISBN 978-0412606304.

[3] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 1995. 2 (12): 1137–1143. CiteSeerX 10.1.1.48.529.

13.