|  |  |  |
| --- | --- | --- |
| Nama Akun Kaggle | : | Whisnumurty GA |
| Link Akun Kaggle | : | <https://www.kaggle.com/whisnumurtyga> |
| Nama Competition | : | Steel Plate Defect Prediction |
| Deskripsi Challenge | : | Mengembangkan model prediktif yang dapat mengidentifikasi dan memprediksi keberadaan cacat pada pelat baja. Peserta diberikan kumpulan data yang berisi informasi tentang berbagai atribut fisik dan kimia dari pelat baja, serta label yang menunjukkan apakah pelat baja tersebut mengandung cacat atau tidak. |

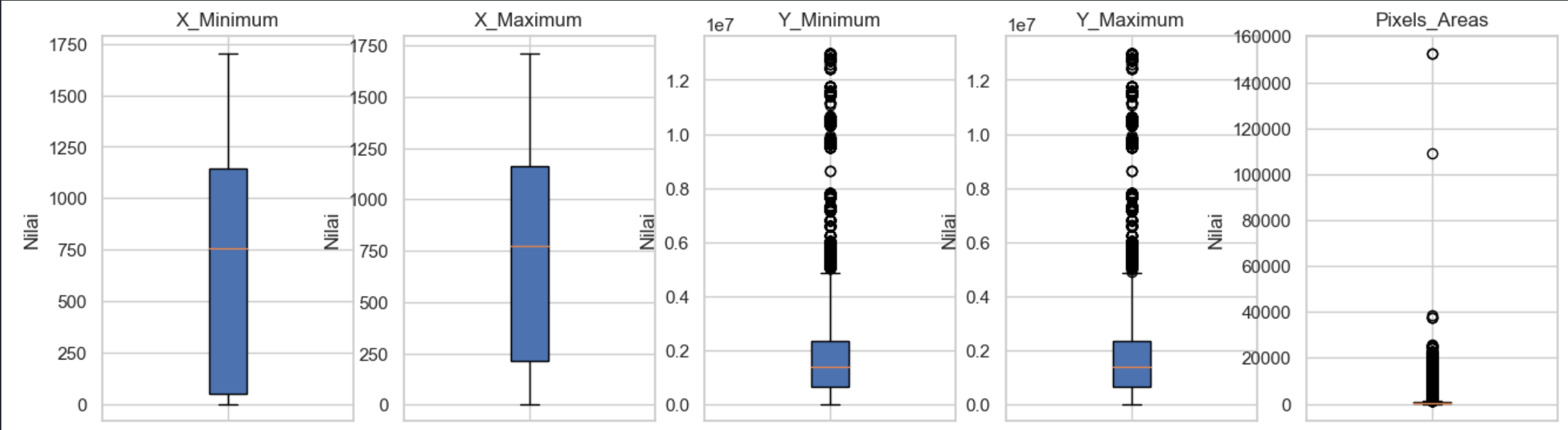
# Step 1 – Data Understanding

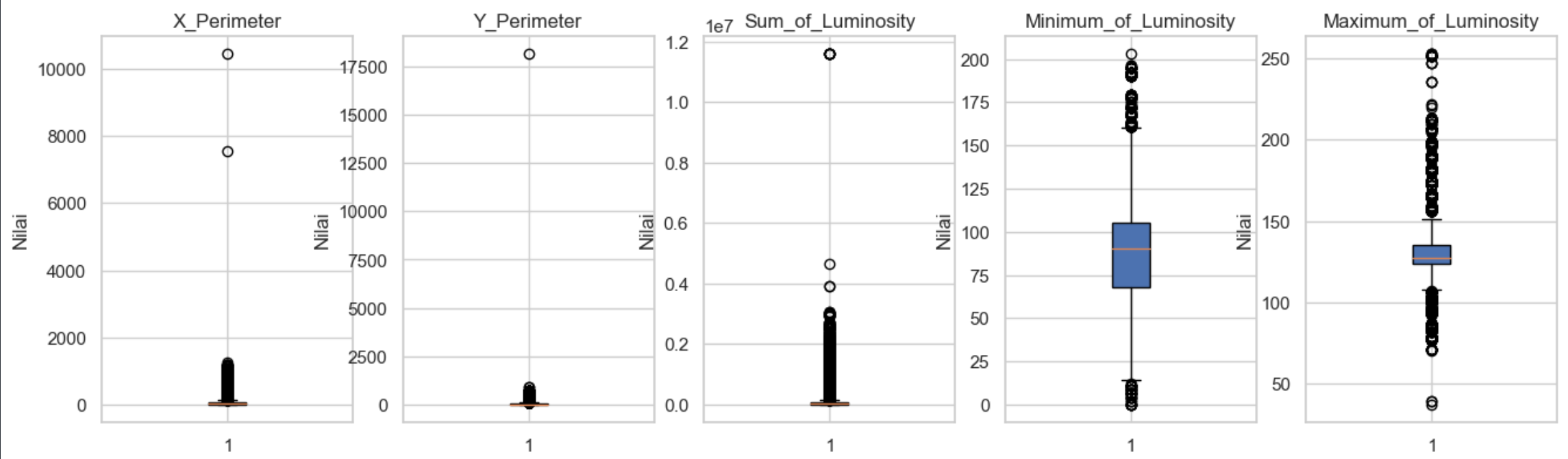
|  |  |
| --- | --- |
| Nama Feature | Deskripsi |
| Id | ID unik untuk setiap sampel data. |
| X\_Minimum, X\_Maximum, Y\_Minimum, Y\_Maximum | Koordinat batas persegi panjang yang mengelilingi cacat pada pelat baja. |
| Pixel\_Areas | Jumlah piksel yang tercakup oleh cacat tersebut. |
| X\_Perimeter, Y\_Perimeter | Panjang perimeter (keliling) dalam arah X dan Y dari cacat. |
| Sum\_of\_Luminosity, Minimum\_of\_Luminosity, Maximum\_of\_Luminosity | Informasi tentang luminositas cacat, termasuk jumlah total luminositas, luminositas minimum, dan maksimum. |
| Length\_of\_Conveyer | Panjang konveyor di mana pelat baja bergerak. |
| TypeOfSteel\_A300, TypeOfSteel\_A400 | Variabel biner yang menunjukkan jenis baja (A300 atau A400). |
| Steel\_Plate\_Thickness | Ketebalan pelat baja. |
| Edges\_Index, Empty\_Index, Square\_Index, Outside\_X\_Index, Edges\_X\_Index, Edges\_Y\_Index, Outside\_Global\_Index | Berbagai indeks yang mungkin dihitung dari citra pelat baja. |
| LogOfAreas, Log\_X\_Index, Log\_Y\_Index, Orientation\_Index, Luminosity\_Index, SigmoidOfAreas | Fitur-fitur lain yang dihitung dari citra pelat baja. |
| Pastry, Z\_Scratch, K\_Scatch, Stains, Dirtiness, Bumps, Other\_Faults: | Label yang menunjukkan jenis cacat yang ada pada pelat baja. Misalnya, "Pastry" mungkin menunjukkan adanya cacat berbentuk pastry, dan seterusnya. |

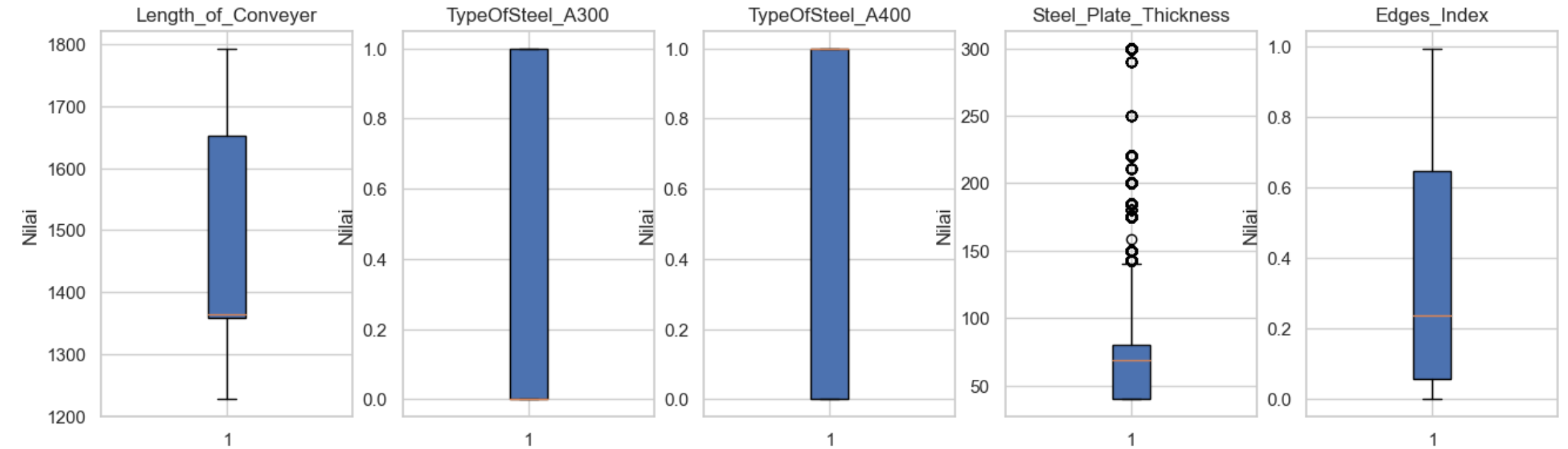
# Step 2 – Set Requirements

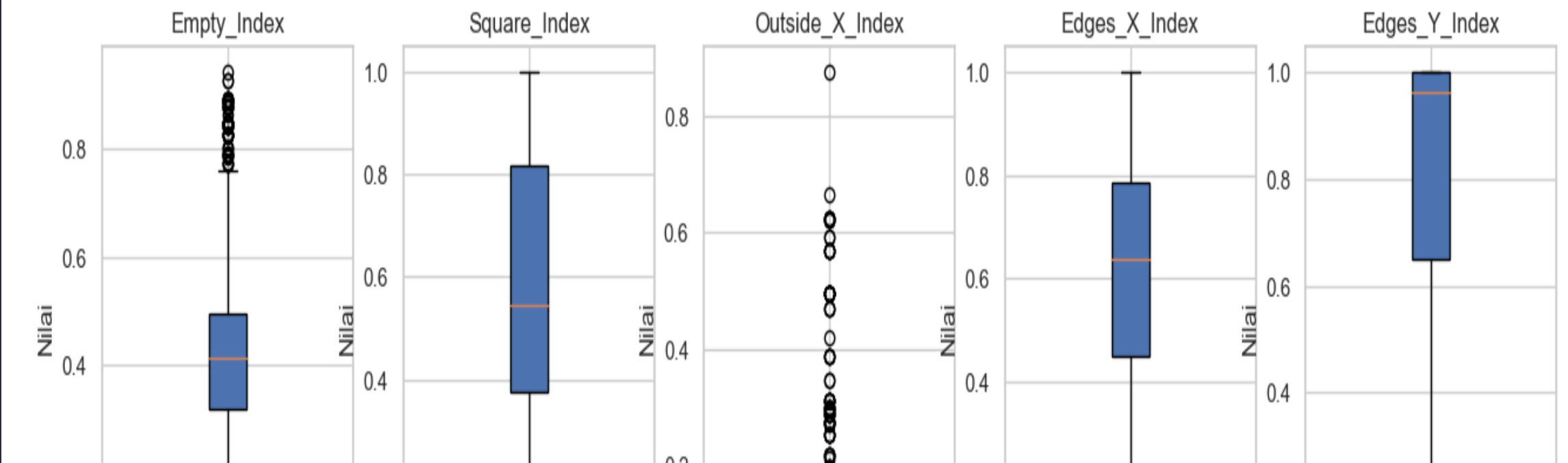
|  |  |
| --- | --- |
| Library | Fungsi |
| NumPy | melakukan operasi numerik untuk bekerja dengan array dan matriks |
| Pandas | Manipulasi dan analisis data |
| Pretty Table | Membuat table rapi dan mudah dibaca |
| Matplotlib & Seaborn | Visualisasi Data |
| DeepCopy | Mencopy model |
| Partial | Untuk membuat fungsi baru dengan beberapa argumen dari fungsi yang ada. |
| gc | Garbage Collector untuk memerikan control atas pemulihan memori |
| Scikit Learn | Untuk machine learning dan analisis data. Serta menyediakan berbagai algoritma machine learning, evaluasi model, dan alat-alat pra-pemrosesan data. |
| Optuna | Mengoptimasi parameter secara otomatis |
| XGBoost | Library untuk pemodelan prediktif decision tree based sebagai model dasar |
| CatBoost | Library untuk menangani data kategorikal |

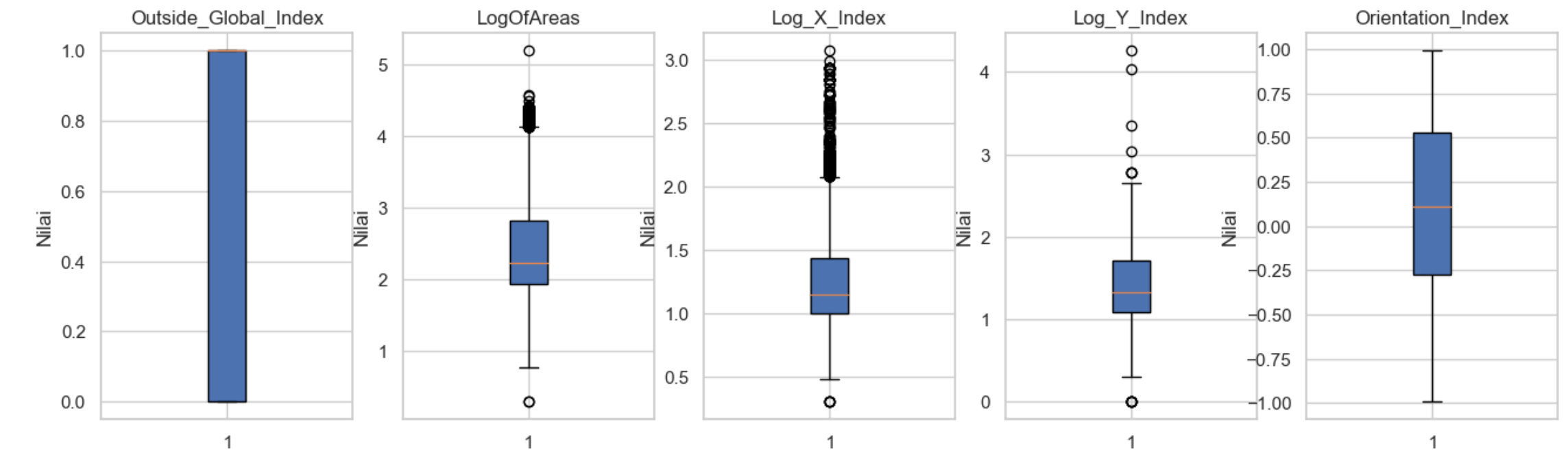
# Step 3 – Exploratory Data Analysis

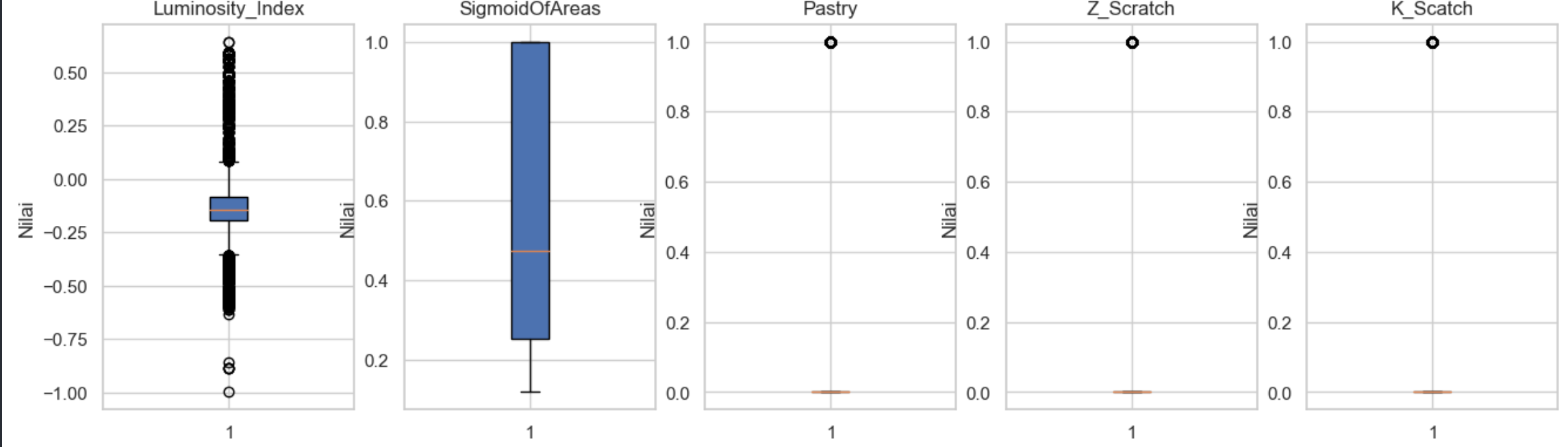


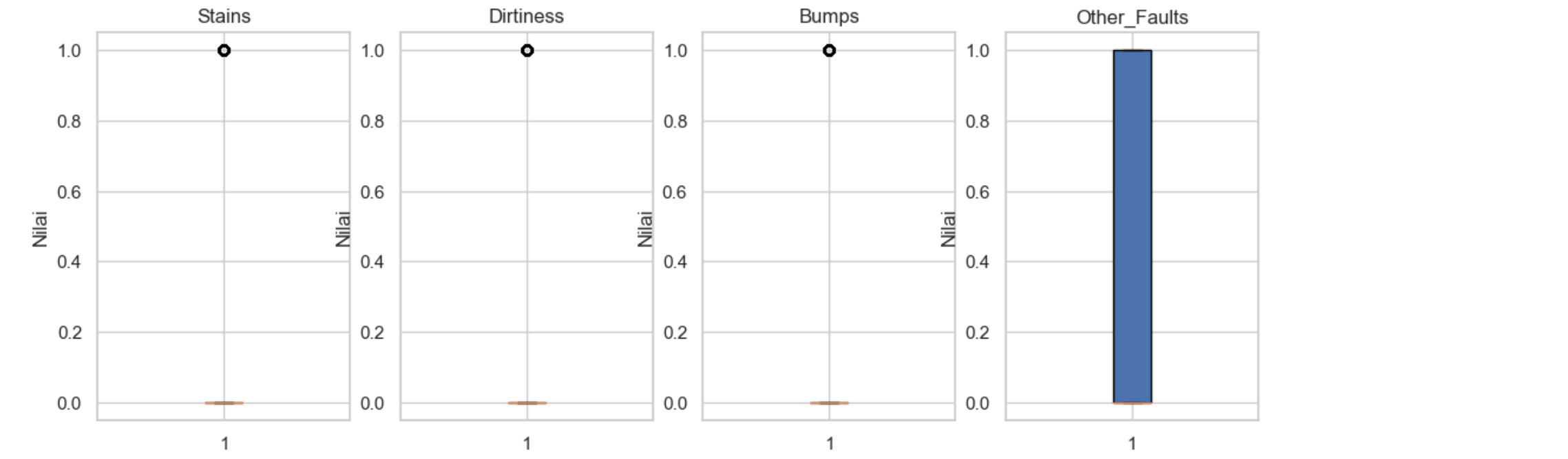




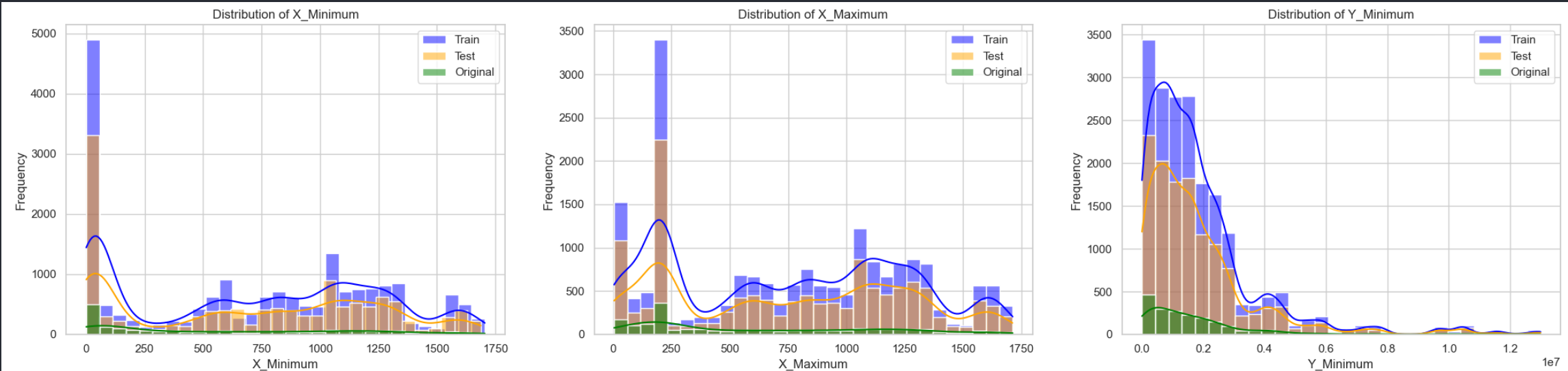


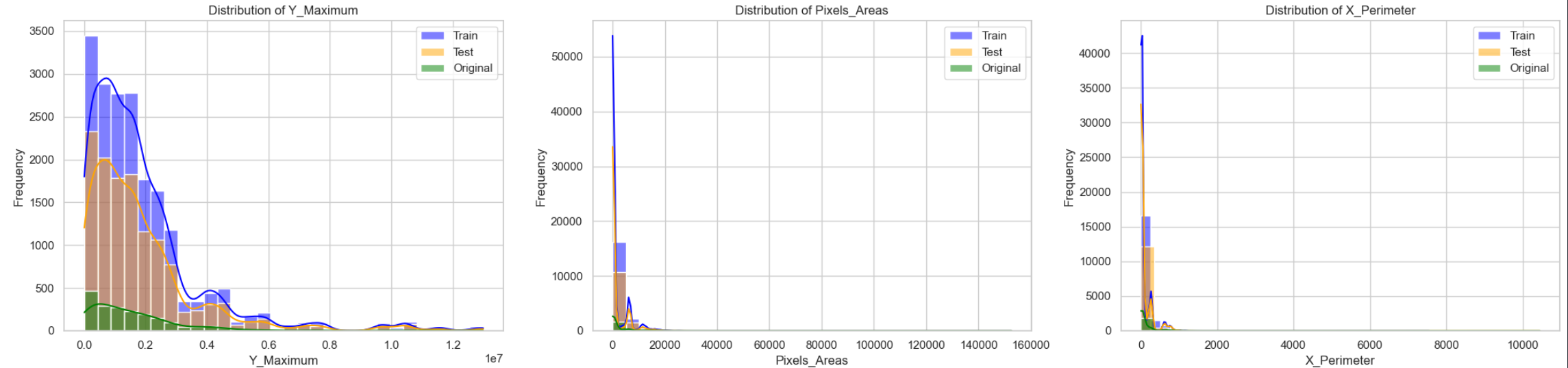






Dari hasil boxplot tersebut dapat kita ketahui bahwa terdapat kolom kategorik yang sudah dijadikan bentuk numerik seperti (Pastry, Z\_Scratch, K\_Scatch, Strains, Dirtiness, Bumps, Other Faults, Outside\_Global Index), dari box plot tersebut kita juga dapat melihat beberapa feature terutama pada y\_minimum dan y\_maximum (artinya Panjang baja secara vertical sangat bervariasi)





Dari grafik histogram tersebut bis akita lihat bahwa distribusi data numerik memiliki pola kurva distribusi yang agak berbeda dari data original. Hal tersebut menandakan bahwa data sintesis yang dihasilkan dari data original memiliki pola yang berbeda.

# Step 4 - Feature Engineering

## 4.1 One Hot Encoding

## 4.2 Stabilize Categorical Columns

## 4.3 High Freq OHE

## 4.4 Category Encoding

# Step 5 – Modelling

## 5.1 Create Model

## 5.2 Optimize Ensemble Weight

## 5.3 Fit the Model

# Step 6 – Submission

# Step 7 – Generalization Ensemble

setiap pdf harus memuat:

1. deskripsi challenge

2. link atau nama akun kalian di kaggle.

3. penjelasan preprocess kalian dan juga EDA yang kalian lakukan terhadap data

4. model yang dipilih dan kenapa memilih model tersebut

5. berapa akurasi terakhir yang kalian dapatkan

6. kalian boleh coba berkali-kali, yang kami catat adalah peringkat kalian di leader board dan akurasi yang kalian dapatkan di tanggal 29 Maret 2024 jam 09.00 WIB