**Preprocessing**

**Original Dataset Overview**

* **File Name:** updated\_file\_r.csv
* **Rows:** 3000
* **Original Columns:**
  1. **Identifiers & Metadata**
     + event\_id, event\_time, location, rock\_type
  2. **Static Mechanical Test Features**
     + depth, compressive\_strength, tensile\_strength, tangential\_stress,  
       brittleness\_ratio, elastic\_strain\_energy, density
  3. **Seismic / Vibration Features**
     + signal\_energy, event\_duration, dominant\_frequency, signal\_rms,  
       seismic\_event\_rate, peak\_ground\_acceleration, gutenberg\_richter\_b,  
       cumulative\_seismic\_energy, axial\_stress, radial\_stress, strain\_rate,  
       seismic\_pulses, total\_seismic\_energy
  4. **Labels / Targets**
     + intensity\_label, rockburst\_grade

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Action** | **Columns Affected** | **Purpose** |
| **Datetime Parsing** | Converted event\_time to datetime; extracted event\_year, event\_month, event\_day, event\_hour | event\_time | Enable time-based analysis |
| **Cyclic Encoding** | Created hour\_sin, hour\_cos, month\_sin, month\_cos | Derived from event\_hour, event\_month | Preserve periodic nature of time |
| **Target Encoding** | Mapped intensity\_label I–IV → 1–4 | intensity\_label | Model compatibility |
| **Missing Value Handling** | KNN Imputation (k=5) | seismic\_pulses, total\_seismic\_energy | Fill missing seismic values |
| **Categorical Encoding** | One-hot encoding | rock\_type, location | Convert text to numeric |
| **Scaling** | StandardScaler (mean=0, std=1) | All numeric except intensity\_label\_num | Normalize feature scales |

### ****Output After Preprocessing****

* **Rows:** 3000
* **Columns:** Original numeric + new encoded columns + cyclic encodings
* **Example New Columns:**
* Time features: event\_year, event\_month, hour\_sin, hour\_cos, month\_sin, month\_cos
* Encoded features: rock\_type\_\*, location\_\*
* Clean numeric seismic & static features (imputed + scaled)
* **Target Variable:** intensity\_label\_num (numeric)

**Feature Engineering**

### ****1. Starting Dataset****

* **File Name:** preprocessed\_dataset.csv
* **Rows:** 3000
* **Contains:** All cleaned/scaled features from preprocessing, plus encoded categorical variables, numeric seismic and static mechanical features, and target label intensity\_label\_num.

### ****2. Feature Engineering Steps Applied (Project PDF-Aligned)****

### A. Static Mechanical Features

|  |  |  |
| --- | --- | --- |
| **New Feature** | **Formula** | **Purpose** |
| depth\_stress\_interaction | depth \* tangential\_stress | Capture combined geological load |
| brittleness\_strain\_interaction | brittleness\_ratio \* strain\_rate | Capture material deformation tendency |

### B. Seismic / Vibration Features

(Adapted because raw vibration waveforms were not available in CSV)

|  |  |  |
| --- | --- | --- |
| **New Feature** | **Formula** | **Purpose** |
| energy\_ratio | signal\_energy / total\_seismic\_energy | Relative event energy |
| stress\_ratio | axial\_stress / radial\_stress | Compare directional stresses |
| freq\_dur\_product | dominant\_frequency \* event\_duration | Frequency-duration coupling |
| energy\_per\_pulse | total\_seismic\_energy / seismic\_pulses | Energy distribution per impulse |
| seismic\_energy\_density | cumulative\_seismic\_energy / depth | Depth-normalized energy |
| vibration\_std | Std dev of (signal\_energy, dominant\_frequency, signal\_rms) | Measure vibration variability |

### C. Geological Features

|  |  |  |
| --- | --- | --- |
| **New Feature** | **Method** | **Purpose** |
| Fault density proxy | (events per location) / depth | Simulate fault-line weakness |
| Rock type representation | Already from preprocessing (one-hot) | Capture lithology |
| Location representation | Already from preprocessing (one-hot) | Contextual geology info |

### ****Output After Feature Engineering****

* **Rows:** 3000
* **Columns:** Preprocessed features + engineered features
* **Example New Columns:**
  + depth\_stress\_interaction
  + brittleness\_strain\_interaction
  + energy\_ratio
  + stress\_ratio
  + freq\_dur\_product
  + energy\_per\_pulse
  + seismic\_energy\_density
  + vibration\_std
  + fault\_density\_proxy
* **Target Variable:** intensity\_label\_num

### ****Notes on Differences****

* **Vibration Data:**
  + Original PDF specifies **STFT** & spectral features from raw waveforms
  + **CSV lacked raw SAC/WAV files** → used normalized/derived seismic metrics instead
* **Thermal Imagery:**
  + Not in CSV → preprocessing skipped, to be done when images available
* **Geological Maps:**
  + No GIS data in CSV → used proxy (fault\_density\_proxy) instead of true fault density extraction