Project Report

Data Preprocessing Report – Startup Dataset

Tasks Performed

🔷 Feature Engineering

- Extracted founded_year from the founded_at column.
- Created a new column company_age using the formula: 2025 - founded_year.

🔷 Data Normalization

- Applied MinMaxScaler to scale numerical columns between 0 and 1:

  funding_total_usd, funding_rounds, investment_rounds, company_age.
- Replaced missing values with 0 before applying scaling.

🔷 Encoding

- Used pd.get_dummies() for category_code and country_code (One-Hot Encoding).
- Dropped the first dummy variable to avoid multicollinearity.

🔷 Labeling (Target Column)

- Created a new binary column Active_Status:

  1 → Active (Operating/IPO)

  0 → Not Active (Acquired/Closed)
- Normalized status column using .str.lower().

🔷 Output

- Saved the processed dataset as:

  Python :- df.to_csv("processed_companies_dataset.csv", index=False)

⚠️ Challenges Handled

- Missing or invalid `founded_at` values → handled with `errors='coerce'`

- NaN’s filled with 0 before scaling

- Geolocation data (`lat/lng`) incomplete → can be enhanced later