Assignment #02: Data Preprocessing for Capstone

Due Date: July 24, 11:59 PM ET
Grade: 10 points

🎯 Objective

Apply the data preprocessing techniques learned in this lab to your capstone project dataset.

📋 Requirements

1️⃣ Data Loading and Exploration:

- Load your malware detection dataset
- Display basic dataset information (shape, data types, sample rows)
- Identify the target variable and features
- Check for class imbalance

2️⃣ Missing Value Analysis:

- Check for missing values in each column
- Analyze the pattern of missing data
- Choose and implement appropriate handling strategy
- Document your decision rationale

3️⃣ Categorical Feature Encoding:

- Identify categorical features in your dataset
- Choose appropriate encoding method (Label vs One-Hot)
- Implement the encoding
- Explain your encoding choices

4️⃣ Feature Scaling:

- Analyze the scale of your numerical features
- Determine if scaling is needed for your chosen algorithm
- Implement appropriate scaling if necessary
- Use proper train/test split workflow

5️⃣ Documentation:

- Add markdown cells explaining each preprocessing step
- Justify your preprocessing decisions
- Include before/after comparisons where relevant

- Document any data quality issues discovered

📊 Deliverables

- Jupyter notebook with all preprocessing steps
- Clean, commented code
- Comprehensive markdown documentation
- Ready-to-use dataset for model training

💡 Tips for Success

- ✅ Use the LLM prompts provided in the course materials
- ✅ Apply the techniques learned in today's exercises
- ✅ Remember the proper scaling workflow (fit on train, transform test)
- ✅ Document your thought process and decisions
- ✅ Test your preprocessing pipeline thoroughly

🚨 Common Mistakes to Avoid

- ❌ Scaling before train/test split (data leakage)
- ❌ Using test data to determine preprocessing parameters
- ❌ Not handling categorical variables properly
- ❌ Ignoring class imbalance
- ❌ Poor documentation of preprocessing decisions