Machine Learning – Exercise 0: Dataset Description

Group Number - 40

Group Members

- **Pletikosic Vice**
- **Cayo Ventura Cesar**
- **Charan Naveen Kumar Ravuri**

Introduction

In this exercise, we selected two classification datasets from the UCI Machine Learning Repository: the Energy Efficiency dataset and the PhiUSIIL Phishing URL dataset. Our goal was to choose datasets with diverse characteristics to explore different preprocessing and modelling challenges.

Dataset 1: Energy Efficiency

Dataset URL - https://archive.ics.uci.edu/dataset/242/energy+efficiency

The dataset contains **768 samples** and **8 input attributes**, all of which are **numeric**. The **target attributes** are:

- **Heating Load**
- **Cooling Load**
- Both are **continuous numeric values**.

Input attributes include:

- Relative Compactness
- Surface Area
- Wall Area
- Roof Area
- Overall Height
- Orientation
- Glazing Area
- Glazing Area Distribution

All attributes are **interval-scaled numeric values**, requiring:

- **Normalisation** or **standardisation** before applying machine learning algorithms.
- Target attributes have a **wide range of values**, making it important to analyse their distribution.
- Selected input attributes like **Relative Compactness** and **Surface Area** show a **uniform distribution**.
- There are no missing values in the dataset.
- **Scaling is essential** due to the varying ranges of numeric features.

Dataset 2: PhiUSIIL Phishing URL

Dataset URL - https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset

The dataset contains **235,795 samples** and **54 attributes**, including a mix of **real, categorical, and integer** types.

The target attribute is:

- **isPhishing**
  - Binary classification:
    - 1 → Legitimate
    - 0 → Phishing

Input attributes include:

- URLLength
- Domain
- TLD (Top-Level Domain)
- CharContinuationRate
- TLDLegitimateProb
- URLSimilarityIndex
- URLTitleMatchScore
- URLCharProb

Attribute types:

- **Categorical attributes** (e.g., Domain, TLD) are **nominal** and require encoding (e.g., one-hot or label encoding).
- **Numeric attributes** (e.g., CharContinuationRate, TLDLegitimateProb) are **interval-scaled** and may require normalisation.

Target attribute characteristics:

- The distribution is **imbalanced**, with **more legitimate URLs** than phishing ones.
- This imbalance should be addressed using techniques like **resampling** or **class weighting**.
- No missing values are reported, but **encoding and scaling** are essential for model performance.

Comparison and Justification

The Energy Efficiency dataset is small and numeric, requiring scaling but no encoding. The PhiUSIIL dataset is large and categorical, requiring encoding and handling of class imbalance. This diversity in sample size, attribute types, and preprocessing needs makes them ideal for exploring different aspects of classification tasks.

Conclusion

Both datasets offer unique challenges and learning opportunities. The Energy Efficiency dataset helps understand regression-based classification with numeric features, while the PhiUSIIL dataset introduces categorical data handling and class imbalance issues. These datasets will be used in future exercises to build and evaluate classification models.