CERIAS

The Center for Education and Research in Information Assurance and Security

Impact of Data Quality and Data Preprocessing on ML Model Fairness

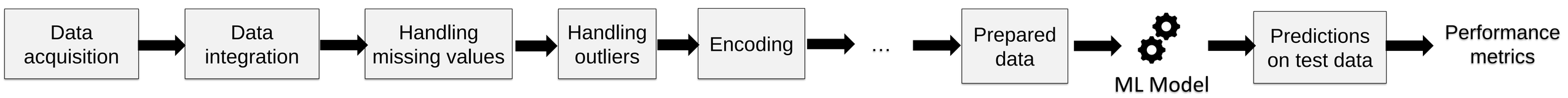
Sathvika Kotha, Mentor: Romila Pradhan

Computer and Information Technology

Motivation:

The increasing adoption of machine learning (ML) techniques has shown us that the more data a model contains, the more accurate it will be. However, the **quality of the data may be more significant than the quantity**. Before being fed into an ML model, training data undergoes a number of preprocessing steps. Prior work has considered the effect of data cleaning on ML classification tasks without any consideration to fairness of downstream model bias. In this study, we systematically examine the **effects of data quality issues and data preprocessing techniques on model fairness**.

A Typical Data Science Pipeline:



Experimental Setup:

Datasets:

Standard data in fairness literature:

- Adult Census Income dataset [1]
- German Credit dataset [1]
- COMPAS [2]

ML models:

- Logistic regression
- Random forest classifier
- Support vector machine
- k-nearest neighbors

Performance evaluation metrics:

- Accuracy
- Fairness metrics e.g., statistical parity difference, predictive parity difference, equalized odds [3]

Implementation:

Preprocessing Step	Detection Method	Repair Method
Missing Values	isNaN()	• Removal • Mode imputation • Mean imputation • KNN imputation
Outliers	• Local Outlier Factor • Inter Quartile Range • Z-score	• Removal • Mode Imputation
Rescaling	---	• Z-score normalization • MaxAbx scaler • MinMax - Normalization
Encoding	Categorical Variables	• Label Encoding • One-Hot Encoding

Results:

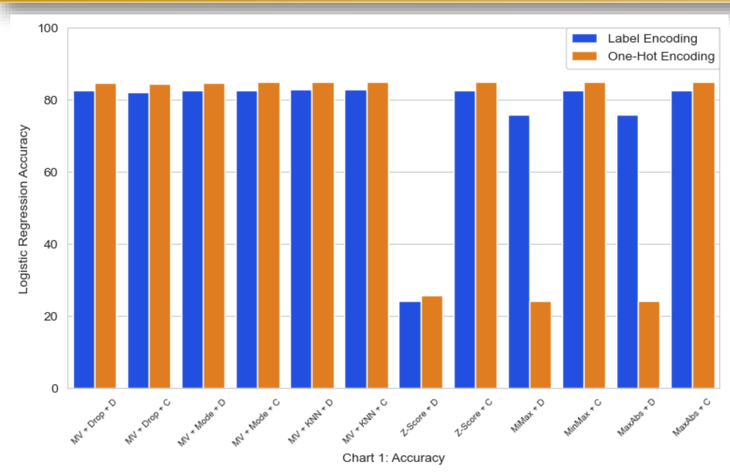


Chart 1: Accuracy

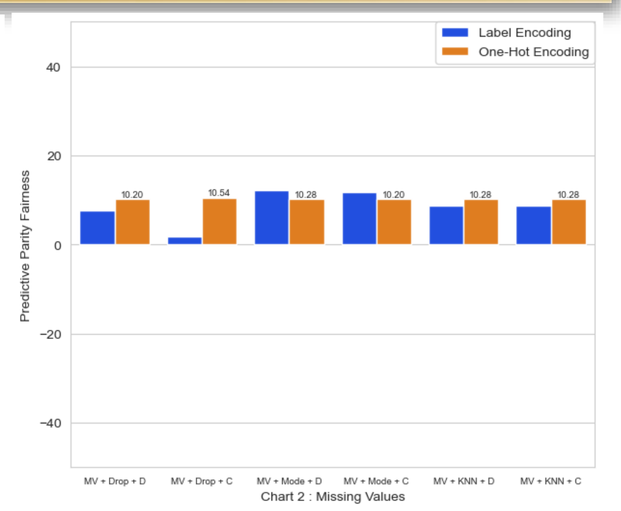


Chart 2 : Missing Values

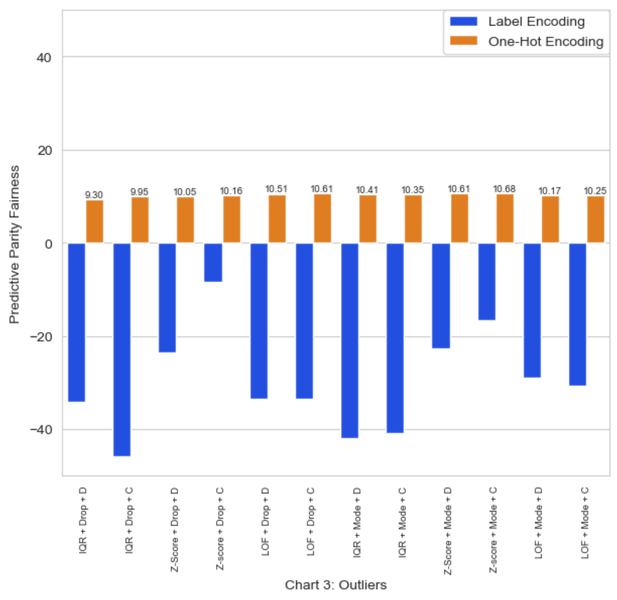


Chart 3: Outliers

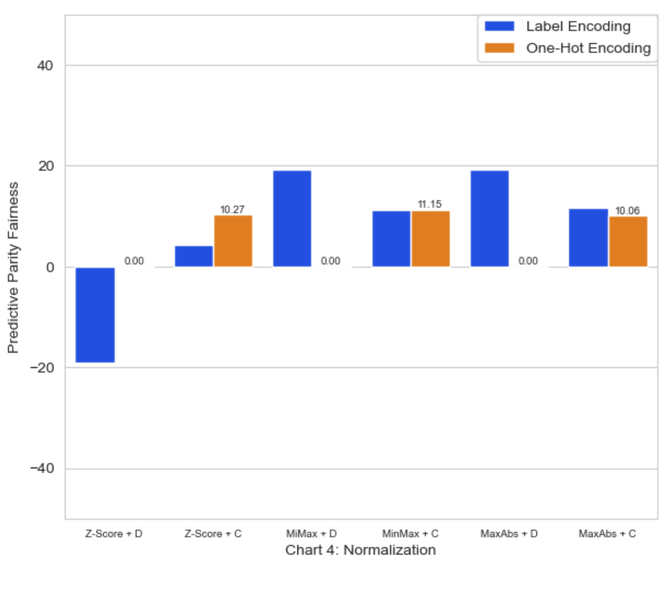


Chart 4: Normalization

- Experiments were carried out in 1. Dirty test data (with no preprocessing) 2. Clean test data (sample preprocessing applied to test data as well)
- Dirty test data is represented as D whereas clean test data is represented as C in the output charts.

Conclusions And Future Work:

- Different data preprocessing steps impact downstream machine learning model performance differently:
  1. One-hot encoding of categorical data improves fairness (reduces bias) more than label encoding; the impact on accuracy, however, is similar in both cases.
  2. Manipulating outliers have a notable effect on predictive parity where the model is more biased toward unprivileged groups.
- We plan to augment this study including ML preprocessing steps with traditional data-cleaning techniques.

References:

1. Adult. (1996). UCI Machine Learning Repository.
2. Angwin et al. Machine Bias. (2016). ProPublica.
3. Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the international workshop on software fairness.*
4. Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., & Zhang, C. (2021). CleanML: A study for evaluating the impact of data cleaning on ML classification tasks. In *IEEE 37th International Conference on Data Engineering (ICDE)*

PURDUE UNIVERSITY

CERIAS