Practical Machine Learning: Exploratory Data Analysis

Iran Izadyariaghmirani

Department of Computer Science, University of Wyoming

March 25, 2024

I. Introduction

In this study, we aim to conduct an initial analysis of two raw datasets before implementing machine learning techniques. The datasets under consideration are the “Primary Tumor” and "Wine Quality". The objective is to familiarize ourselves with the datasets, identify potential challenges, and determine the necessity of preprocessing for effective machine learning tasks. To this end, we utilize various pre-processing methods, such as handling missing values and normalization. In the following sections, we describe the issues identified for each dataset and explain the techniques used to mitigate these challenges. This will be then followed by an evaluation of the datasets before and after the preprocessing stages. The discussion will be concluded with a set of remarks.

II. “Primary Tumor” dataset

***Description of the dataset*:** The “Primary Tumor” dataset comprises 339 instances and 18 patient-related features. Each instance represents a patient, and the features include patient-related information such as age and medical characteristics such as histologic type. The target variable 'binaryClass' indicates whether the tumor is malignant (P) or benign (N).

***Issues identified (m*issing values):** The dataset contains 207 instances with missing values, totaling 225 missing values. The features with missing values are 'skin', 'axillar', 'sex', 'histologic-type', and 'degree-of-diffe'.

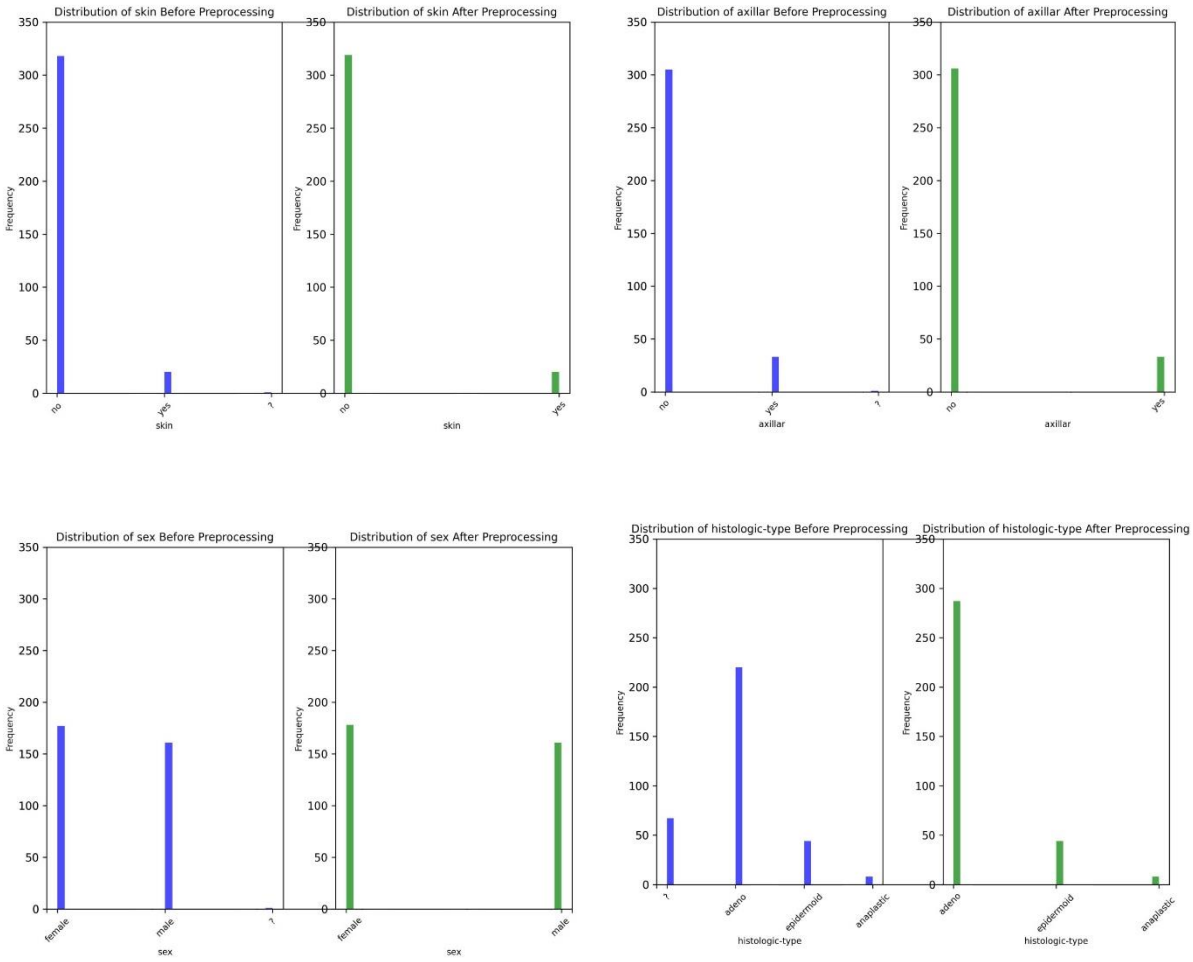
***Data preprocessing*:** To address the identified issues, we first examined the distribution of missing values and attribute frequencies for each feature and then implemented the appropriate data handling strategy as described below:

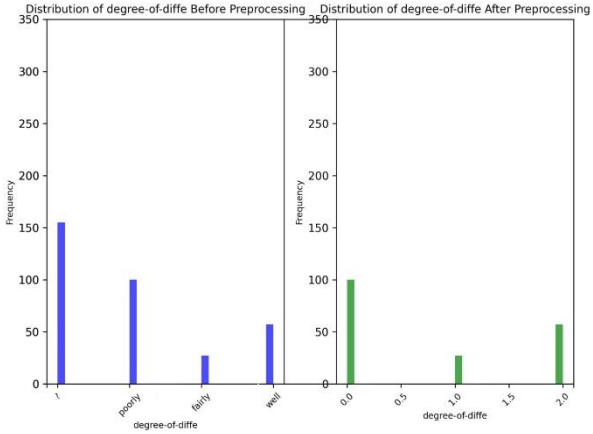
- The 'skin' and 'axillar' features each have one missing value. Furthermore, in both cases, the majority of the instances are labeled as 'no'. Similarly, the 'sex' has one missing value, with slightly more instances labeled as 'female' compared to 'male'. Due to the limited number (one) of missing values for each attribute, the missing data can be easily replaced with the mode of each feature.
- 'Histologic-type' includes three categories (i.e., 'epidermoid,' 'adeno,' and 'anaplastic'), among which 'adeno' shows a significantly higher frequency (220 instances) compared to 'epidermoid' (45 instances) and 'anaplastic' (8 instances). This feature has 67 missing values. In this case, all the

missing values were replaced with the dominant category (i.e., 'adeno'), which is also the mode of the feature.

- Finally, the 'degree-of-diff' has three categories ('well,' 'fairly,' and 'poorly') and 155 missing values. As seen, the missing values constitute a substantial portion of the dataset. In this case, it was best to utilize the ordinal encoding approach. This technique is suitable when the categories have an inherent order ('well' < 'fairly' < 'poorly'), and none of them are dominant (i.e., a relatively balanced distribution). In the ordinal encoding method, integer values are assigned to the categories in a way that reflects their ordinal relationship, while the missing values are replaced with "*NaN*" (not a number).

Figure 1 demonstrates the frequencies of various categories in each dataset before and after the preprocessing step.





The missing values of 'the 'degree-of-diff' were replaced with "NaN". It should be noted that in this scenario, the missing values will not be eliminated. Instead, they will be handled according to the strategy defined by the classifier or any preprocessing steps applied before the classification stage

Figure 1. The frequencies of various categories in each dataset before and after the preprocessing step.

Evaluation of preprocessed data: A Random Forest classifier was trained and evaluated using cross-validation. The mean accuracy obtained after the preprocessing of the datasets was approximately 0.8524, whereas the mean accuracy before the preprocessing was approximately 0.8493.

III. "Wine Quality" dataset

Description of the dataset: The dataset utilized in this exercise is the "winequality-red" dataset, which includes data on various features of wines, such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The dataset comprises a total of 1,599 rows and eleven features. The target variable, "quality" rates the wine on a scale from zero to ten (0-10). There are no missing values in the dataset.

Issues identified (skewed distributions): Some features exhibit skewed distributions, which could potentially impact the performance of machine learning models. We employ a "logarithmic transformation" to address this issue and improve model interpretability.

Preprocessing: To prepare the data for machine learning, we implemented a preprocessing pipeline consisting of the following steps:

- Logarithmic transformation: a custom logarithmic transformation was applied to address potential issues with skewed feature distributions.
- Standard scaling: standard scaling was employed to normalize the features and bring them to a comparable scale, thus enhancing the performance of machine learning models.

The results of the normalization preprocessing on the distributions of the features are presented in Figure 2.

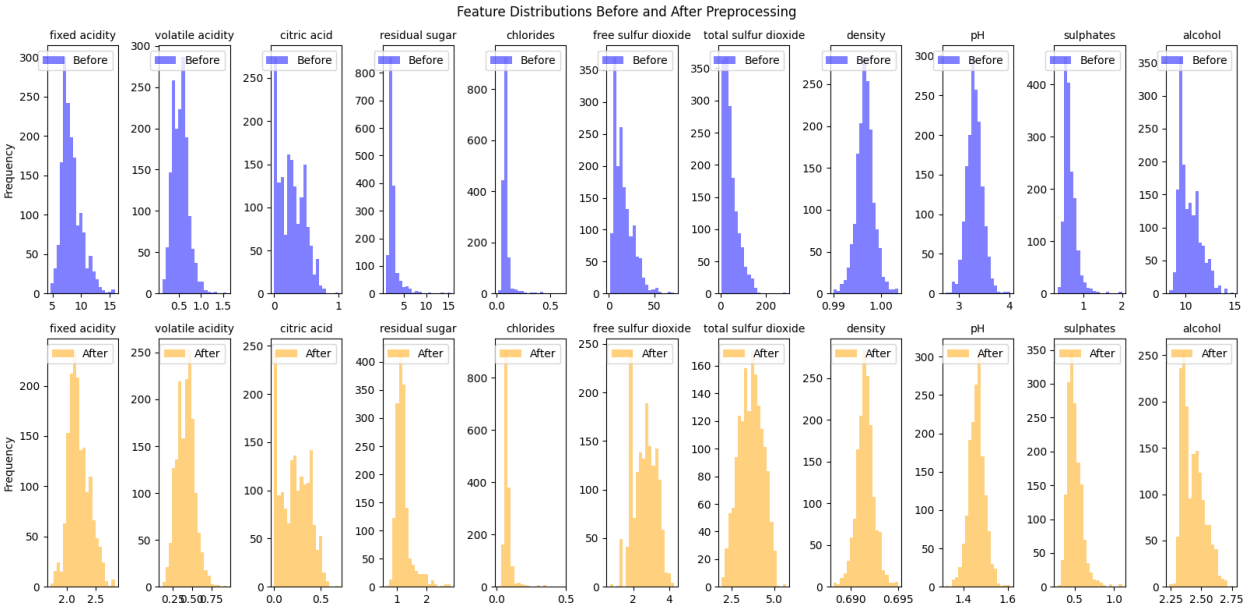


Figure 2. Results of the normalization for features of the “winequality-red” dataset.

Cross-validation scores: To further examine the data enhancement resulting from the applied pre-processing steps, a Random Forest model was trained and evaluated using 10-fold cross-validation with the preprocessing pipeline. The mean cross-validation score for the Random Forest model with the logarithmic transformation and standard scaler preprocessing steps was approximately 0.5829, compared to a mean score of 0.5710 before the pre-processing.

IV. Conclusions

In this study, we conducted an initial analysis of the “Primary Tumor” and the “Wine Quality” datasets before implementing machine learning techniques.

For the “Primary Tumor” dataset, missing values were addressed using appropriate preprocessing techniques, including replacement with mode for categorical features and ordinal encoding for ordinal categories. After preprocessing, a Random Forest classifier achieved a mean accuracy of approximately 0.8524 through cross-validation.

For the “Wine Quality” dataset, skewed distributions in some features were mitigated through a custom logarithmic transformation. Additionally, standard scaling was applied. A Random Forest model trained with these preprocessing steps achieved a mean cross-validation score of approximately 0.5829.

Overall, our analysis and preprocessing steps have prepared both datasets for further machine-learning tasks.

V. References

1. https://www.openml.org/search?type=data&sort=runs&id=1003&status=active
2. “Winequality-red” dataset - UCI Machine Learning Repository
3. Scikit-learn: Pedregosa et al., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011). [Link to the Scikit-learn paper or documentation]