POLITECNICO
MILANO 1863

Data & Information quality
Academic Year 2022 - 2023

Accuracy assessment

Project ID:	23
Project number:	1
Assigned datasets:	adult, soybean
Assigned task:	Classification

Stefano Civelli (10628574)
Vlad Marian Cimpeanu (10606922)

Contents

1 Introduction **2**

2 Setup Choices **2**
2.1 Chosen ML algorithms . . . 2
2.2 Chosen ML evaluation metrics . . . 2
2.3 Selected imputation techniques . . . 2

3 Pipeline implementation **3**
3.1 Missing data injection . . . 3
3.2 Correlation Analysis . . . 3
3.3 Data imputation . . . 4
3.4 Drop duplicates . . . 4
3.5 Preprocessing for ML . . . 4

4 Results **6**
4.1 Dataset accuracy assessment . . . 6
4.1.1 Define the accuracy . . . 6
4.1.2 Accuracy assessment analysis Soybean dataset . . . 6
4.1.3 Accuracy assessment analysis Adult dataset . . . 7
4.2 ML performance results . . . 7
4.2.1 Soybean dataset . . . 7
4.2.2 Adult dataset . . . 7

1 Introduction

The project consists in applying data imputation techniques to 5 versions of the same dataset with different degrees of completeness: 50% to 90%.

After the imputation, we measured the **accuracy** of the obtained results using the correct (100%) dataset as reference.

Finally, we trained a couple of ML algorithms with the goal of observing *performance changes* with the different datasets and different imputation methods.

2 Setup Choices

2.1 Chosen ML algorithms

Since we were assigned a **classification** task we selected 2 classification algorithms from the scikit learn library that we deemed fit for the task:

- **Logistic regression** (softmax regression for the soybean dataset, since it contains more than 2 classes)
- **SVM** with round basis function kernel

2.2 Chosen ML evaluation metrics

Since both datasets are quite unbalanced we opted for a metric that takes into account the imbalance of the different classes such as the weighted f1 score.
Furthermore, we decided to plot the confusion matrices to have a deeper insight on the results.

2.3 Selected imputation techniques

As for imputation, we chose 1 basic and 1 advanced technique as requested by the specification.

Basic imputation in both datasets is done with a combination of *median* for numerical features and *mode* for categorical and ordinal features.
Advanced imputation on the other hand uses the **MICE** algorithm with both a `KNeighborsClassifier` for categorical features and `KNeighborsRegressor` for numerical ones.

NOTE: for some features couples like [sex-relationship] and [education-educationNum] a simple correlation-based imputation was applied since they were very highly correlated (100% in the case of education-educationNum). See the pipeline section for more details on this aspect.

3 Pipeline implementation

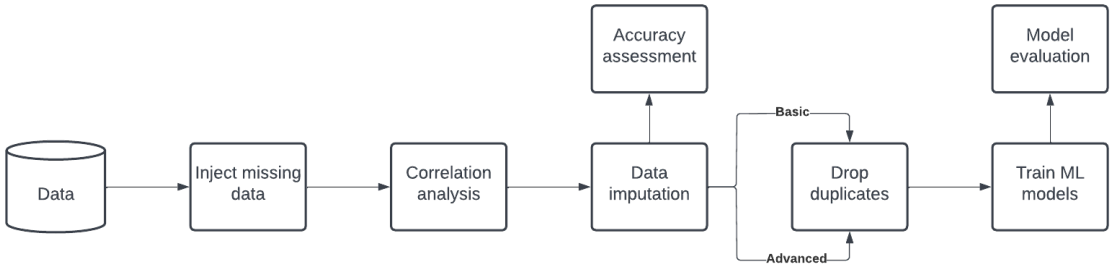

Figure 1: Pipeline

3.1 Missing data injection

In this phase, we used the provided script to inject the `NaN` values and therefore create the 5 versions of each dataset

3.2 Correlation Analysis

In this step of the pipeline, we performed correlation analysis to gain more insights about the data. This process may help to find some strong correlations, useful for the imputation process.

Correlation analysis for the adult dataset: The profiling operation suggested that there were some very simple **correlations** that could be exploited for the imputation process. These imputations are exact since some features (or some aspects of them) were 100% correlated and therefore using them before starting any more complex imputation was an obvious choice.
For example, Figure 2 shows the perfect correlation between education and education-num. They are essentially the same attribute.

Correlation analysis for the soybean dataset: Looking at Figure 3, the correlation matrix for the soybean dataset suggests us there are some covariates that are highly correlated with the target variable *class*. We want to investigate if this correlation can help during the prediction or imputation phase.
We focus on the covariates with a correlation index higher than 0.9, so *leafspots-marg*,*leafspot-size*, *leaf-mild*, *int-discolor*, *sclerotia*. From Figure 4, we can notice it is almost always possible to infer one of those covariates starting from the *class* variable.
The opposite is not true, thus, these covariates do not seem to carry important information during the prediction phase, nevertheless, one may consider applying imputation based on this weak functional dependency.

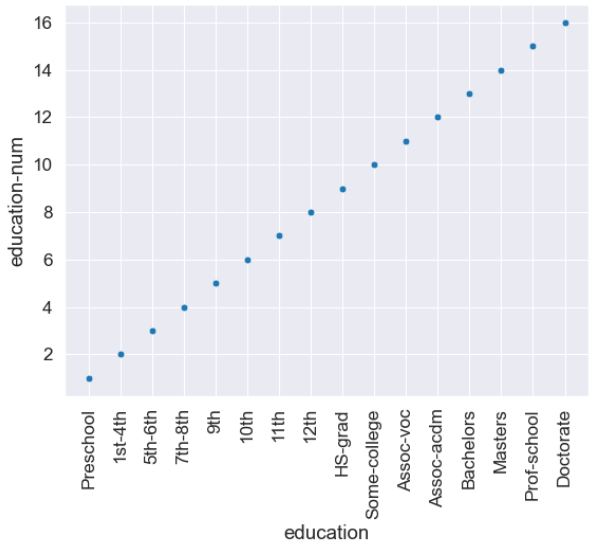

Figure 2: Correlation between education and education-num

3.3 Data imputation

As briefly mentioned in the introduction, **basic** and **advanced** imputation techniques were applied in this phase of the pipeline. Once that was done we evaluated the accuracy of both imputation methods on all of the datasets versions. Accuracy results will be discussed in the "Results" section.

3.4 Drop duplicates

Before proceeding with the training of ML algorithms we checked for duplicates in the dataset. If so, drop them because they may hinder ML performances.

On the other hand, we decided *not to drop duplicates present in the original (100%) dataset* since it would have given us an unfair advantage that we wouldn't have had in a **"real"** scenario where, of course, the original (perfect) dataset doesn't even exist. As a matter of fact, introducing `NaN` values may make some duplicates no longer detectable.

3.5 Preprocessing for ML

Other than dropping duplicates we did some pre-processing before feeding the data to the ML algorithms.

- We one-hot-encoded all categorical features in both datasets
- Since the adult dataset contained also numerical features we applied **z-score** standardization in order to bring them to the same scale, otherwise the bigger values would have influenced more the algorithm.

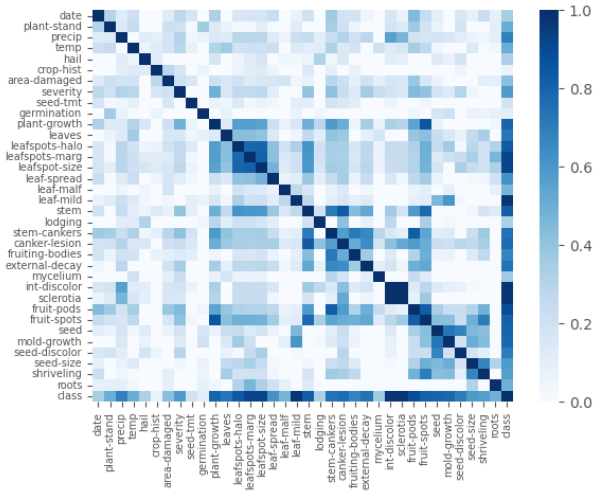

Figure 3: Correlation analysis for soybean dataset.

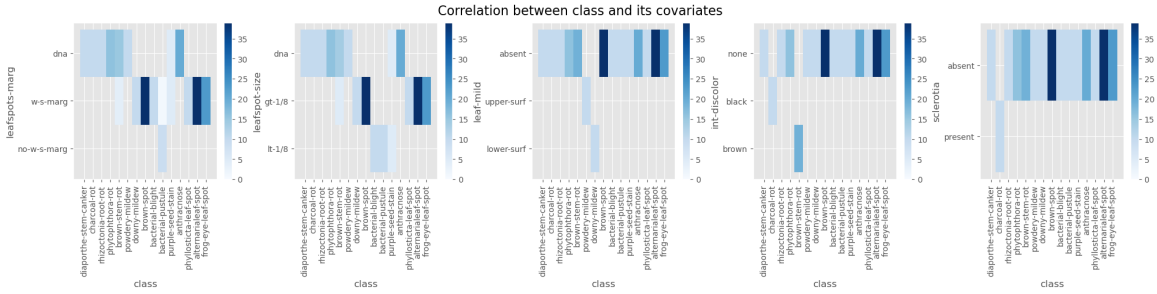

Figure 4: Distribution of pairs (*class*, x) where x is highly correlated with *class*, for the soybean dataset.

- Here we could (and should) have also dropped one feature between [education and education-Num] because they are essentially the same. Since we have shown imputation using correlation on them we decided to not drop them.

4 Results

In this section, we will present the main results obtained when assessing the dataset accuracy and the ML performances

4.1 Dataset accuracy assessment

4.1.1 Define the accuracy

For the accuracy assessment, we have decided to use both the exact matching and the similarity-based approach, since there are some numerical and some categorical features.
To have a similarity-based accuracy comparable to the exact matching accuracy, we need to define a distance function that maps to $[0, 1]$. For the case of ordinal and continuous variables, we decided to use the following distance function:

$$k(x_1, x_2) = \frac{1}{M - m}|x_1 - x_2| \tag{1}$$

where M is the maximum value of the column which v_1 and v_2 belongs to, whereas m is the minimum.
For the soybean dataset, there is also a cyclic variable ($date$) which has the domain of months. It is possible to determine if two months are close enough or not, with an appropriate transformation. First, we map the months to a discrete domain:

$$[January, February, \ldots, December] \longrightarrow [0, 1, \ldots, 11]. \tag{2}$$

This transformation is not enough, since not all the distances are correctly captured, for instance, according to this mapping, January and December appear very distant, which should not be true. To handle this problem, we map all these points to a 2D-dimensional space using a harmonic transformation as follows:

$$h(x) = \left[\sin\left(2\pi\tfrac{1}{12}x\right), \cos\left(2\pi\tfrac{1}{12}x\right)\right] \tag{3}$$

As we can see from Figure 5, now, all the distances are correctly captured by this mapping.
To compute the distance between two months we can use the following distance function:

$$k(x_1, x_2) = \frac{1}{2\|x_1\|\|x_2\|}(1 - x_1 \cdot x_2) \tag{4}$$

4.1.2 Accuracy assessment analysis Soybean dataset

By looking at Figure 6 we can notice that given the same dataset and imputation method, the exact matching accuracy is always lower than the similarity matching one, as we expected.
More interestingly, it seems that advanced and basic approaches produce comparable results, thus, one may think an advanced approach such as a MICE or ML-based is not worth it, given the speed and easiness to use a basic imputation method.
Furthermore, it is interesting to note that starting with very little data (50% missing values), it is possible to retrieve a lot of information.

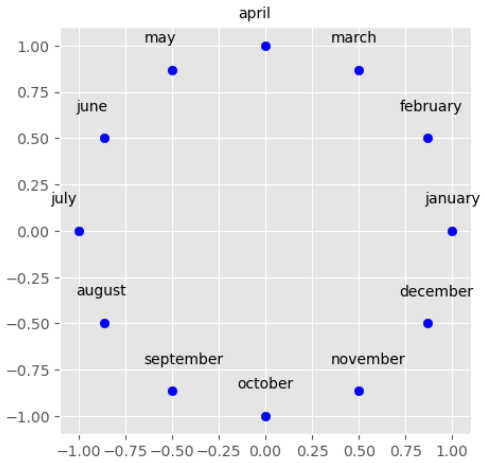

Figure 5: 2D-transformation for date variables.

4.1.3 Accuracy assessment analysis Adult dataset

Similarly to what happens with the soybean dataset, as we can see in Figure 7, the advanced imputation method works slightly better than the basic one. As we will discuss below, even if accuracy seems to be similar, ML performance using the dataset imputed with advanced techniques is quite a bit better so the more complicated approach is still worth the trouble.

NOTE: only the **distance-based accuracy** metric has been evaluated for the adult dataset since there are numerical values where the exact metric would be obviously too strict.

4.2 ML performance results

4.2.1 Soybean dataset

By looking at Figure 8, interestingly, we can notice that for the datasets with a lower degree of completeness, the f1-score after using an advanced imputation method is much higher than the one got after using a basic one. The difference becomes negligible for datasets with a higher completeness degree.
For this reason, we believe the accuracy assessment may mislead people, and encourage them to use basic imputation methods

4.2.2 Adult dataset

In Figure 9 you can see **confusion matrixes** of the **Logistic regression classifier** for both basic and advanced imputation and for all percentages of completeness. It is interesting to note that the algorithm trained on the "basic dataset" tends to predict the majority class most of the time, while the "advanced dataset" seems to have learned the data better and therefore predicts the minority class more often when it is right to do so. This phenomenon is visible in the shift of dark blue hue from left to right in the 2nd row of the matrices.

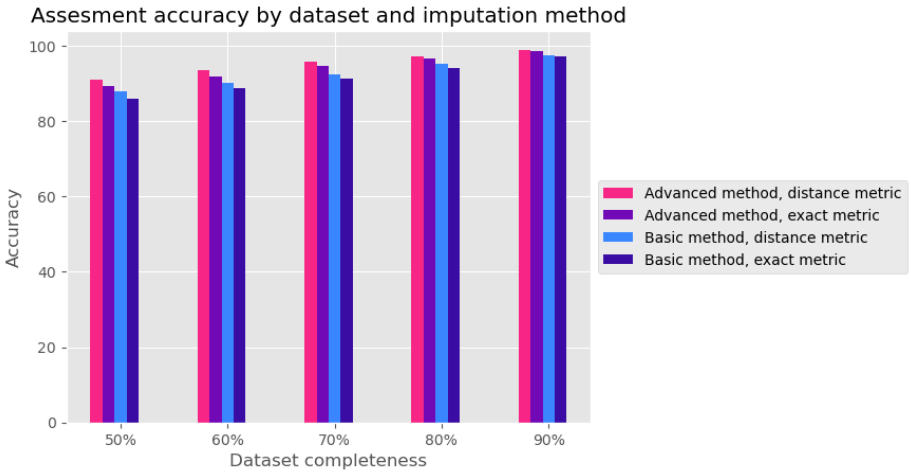

Figure 6: Accuracy assessment for soybean dataset.

A very similar pattern can be observed for the **SVM classifier** and for that reason plots are not provided in this report.

Since f1 scores results for the Adult dataset are very similar to the ones obtained for the soybean we will not put here Figures showing them so as to not overflow this report with plots. All of them are available for reference in the notebooks that come with this report.

Figure 7: Accuracy assessment for the adult dataset.

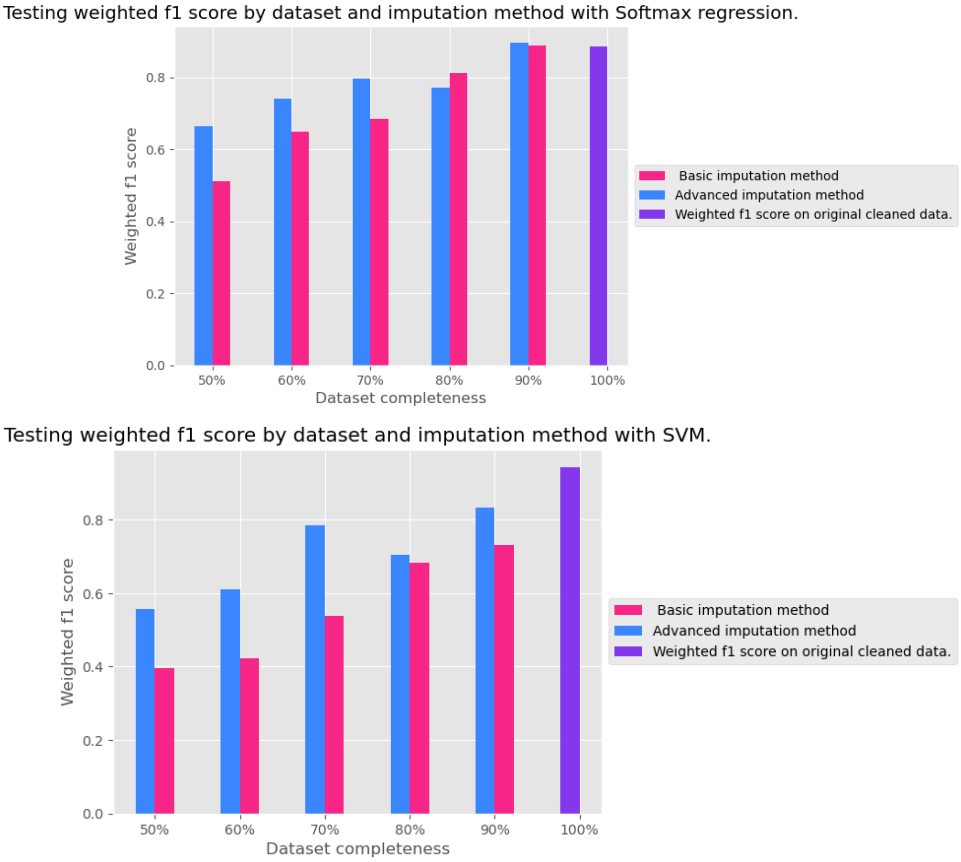

Figure 8: F1-scores for soybean dataset.

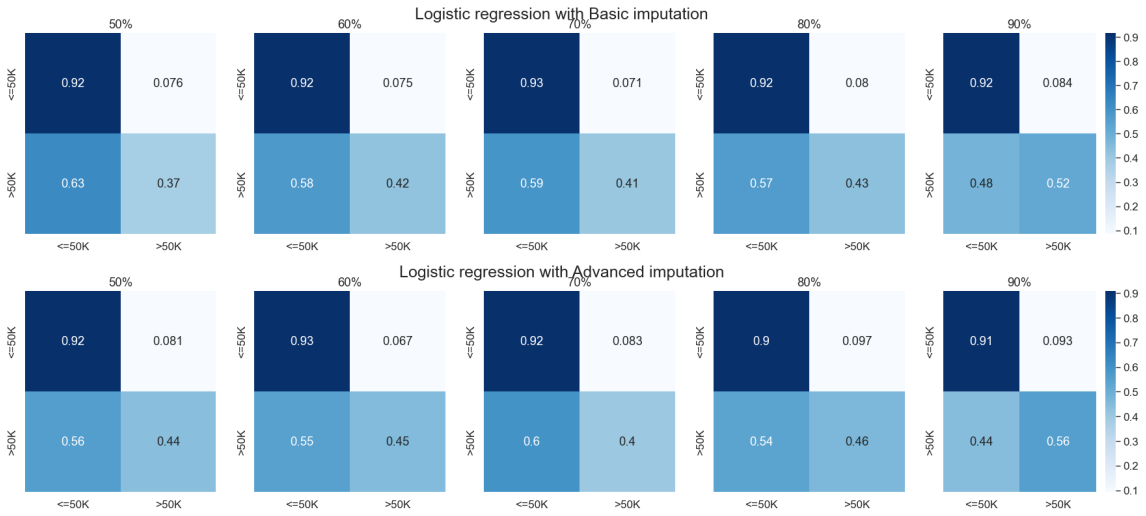

Figure 9: Logistic regression confusion matrixes for adult dataset.