Project Report

PROJECT ID: 62
ASSIGNED DATASET : 8
STUDENTS (NAME SURNAME ID) [Alessia Marino 10933660, Letizia Grassi 10577745, Susana Paoletti 10717422]

1. Setup Choices

Libraries

The following libraries were used in the project:

- **`pandas, numpy, math`**: for data management and numerical operations.
- **`matplotlib, seaborn`**: for data visualization.
- **`ydata_profiling`**: for generating an exploratory report of the dataset.
- **`recordlinkage`**: for record matching and deduplication.
- **`sklearn`**: for building machine learning models (regression, classification and clustering) and performing data preprocessing tasks like missing data imputation, feature scaling, and encoding categorical variables.

Data Preparation Techniques

Imputation techniques

Depending on the characteristics of each missing value, we applied various strategies:

- **Simple deterministic imputation** applied by using a logical relation between columns based on their interdependencies.
- **Simple imputation with a standard value** for both categorical and numerical missing data
- **ML-Based imputation** specifically the LogisticRegression and RandomForestClassifier algorithms from the *scikit-learn* library for categorical variables.

Outlier Detection Techniques

To address anomalies in the dataset, we utilized and compared the following outlier detection techniques:

- **Statistic-based** techniques including Z-score, Standard Deviation (STD), Percentile, and Interquartile Range (IQR)
- **Model-based** techniques using k-Nearest Neighbors (k-NN)

Duplicate Detection techniques

To improve the efficiency of identifying duplicate records, we employed two approaches that reduce the number of comparisons needed:

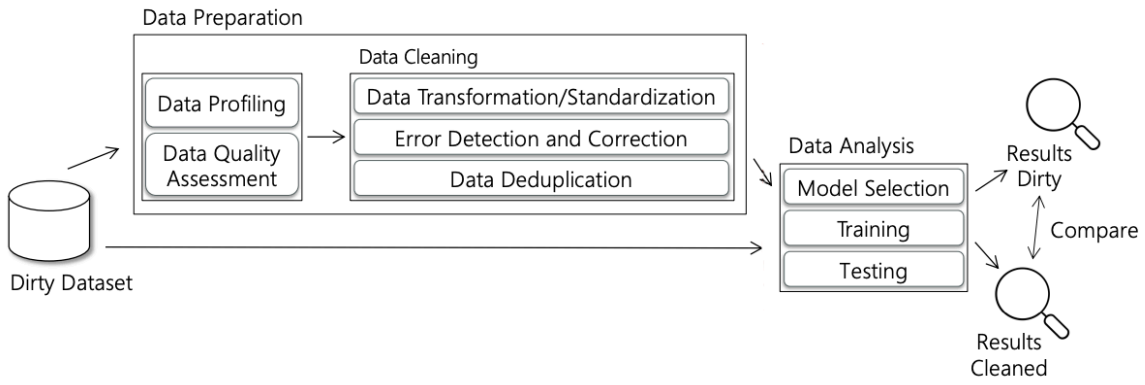
- **Blocking** approach which compares only tuples within the same block based on a certain key
- **Sorted Neighborhood** which compares only tuples that are close to each other in the sorted order based on a certain key

Data Analysis details

Given the nature of our dataset, we opted for a classification analysis using the following approach:

- **k-Nearest Neighbors Classifier (kNeighborsClassifier)** from the *scikit-learn* library was utilized to perform the classification task on both the dirty and cleaned datasets.
- **F1-Score** was selected as the primary performance metric to evaluate the effectiveness of our analysis.

2. Pipeline Implementation



Data Exploration & Profiling

During the data exploration stage, we investigated the key characteristics of the dataset. Specifically, we observed that:

- Data in 'Ubicazione' column overlaps with that in the 'Tipo Via', 'Via', 'Civico', and 'ZD' columns.
- The minimum value in the "Superficie totale" column is 0.
- Column 'Superficie altri usi' has a float data type, while the other two surface columns have an integer data type.
- There is an exact duplicate.
- The columns “ZD”, “Settore Merceologico” and “Tipo Via” have predefined classes.
- The dataset has some missing values in the following columns: 'Insegna' (350), 'Settore Merceologico' (12), 'Civico' (72) and 'Superficie altri usi' (756).
- There’s a significant correlation between "Superficie vendita", "Superficie totale" and "Superficie altri usi" and between “ZD” and “Codice Via”.
- Value distributions of columns ‘Superficie vendita’, ‘Superficie altri usi’ and ‘Superficie totale’ are skewed.

To get a comprehensive view, we also examined the report generated by Pandas Profiling, which provided a detailed summary, highlighted distributions, correlations and missing values of the dataset that validated our previous findings.

Data Quality Assessment

During the second step, we defined and evaluated these Data Quality dimensions:

- **Completeness** as the extent to which all expected data is present in a dataset and we evaluated it as the ratio of non-missing values to the total number of expected values.
- **Accuracy** as the degree to which data correctly represents the real-world values or events it is supposed to model. From the dataset documentation[1], it was clear that the dataset 'Attività commerciali di media e grande distribuzione' corresponded to commercial activities with a sales area above 250 square meters. Using this information, we assessed the accuracy of the 'superficie vendita' column.
- **Consistency** as the degree to which data is uniform and does not contain conflicting or contradictory information. In our case, we assessed consistency by checking if the sum of ‘Superficie vendita' and 'Superficie altri usi' equaled 'Superficie totale'. Although this relationship was not directly mentioned in the documentation, it appeared a reasonable logical assumption considering the structure of the dataset and the significant correlation between those three columns.

Timeliness was not considered because we didn’t have the temporal context of the data.

Data Cleaning (Data Transformation)

In this phase, we focused on normalizing and standardizing our dataset to make it more suitable for the subsequent analysis.

We converted the data type of the "Superficie altri usi" column to Integer to standardize the type of all the surface-related columns and removed the exact duplicates.

Upon analyzing the dataset, we noticed that the values in the 'Ubicazione' column, which contained address information, often differed significantly from those in the corresponding columns: Tipo Via, Via, Civico, and ZD. We assumed that the system allowed the user to manually input this information, and the corresponding columns were then automatically filled. As a result, we decided to consider 'Ubicazione' as the correct information, extracting the data from it, creating new columns with the extracted values, and removing the old columns. Moreover, we chose not to extract the remaining information from the 'Ubicazione' column, as it was not present in all rows and seemed not relevant for our analysis, such as details about the ‘isolato’ or type of ‘accesso’.

Data Cleaning (Error Detection & Correction Missing Values)

In this phase, we imputed the missing values in the 'Superficie altri usi' column with zero and corrected the values in the 'Superficie Totale', 'Superficie altri usi', and 'Superficie vendita' columns, ensuring that the value of 'Superficie Totale' matched the sum of the other two.

We deleted the rows where the values in the 'Settore Merceologico' and 'Insegna' columns were both null, as they contained too much missing information and would have not been useful for our data analysis. However, for the rows with missing values only in the 'Insegna' column, we decided to keep them since the number of cases was high. To handle the missing values in this column, we imputed them with the standard value 'Non Specificato'.

---

[1] “Comune di Milano - Attività commerciali di media e grande distribuzione (informazioni aggiuntive)”, http://www.datiopen.it/it/opendata/Comune_di_Milano_Attivit_commerciali_di_media_e_grande_distribuzione?metadati=showall#dettagli-sul-dato

For the rows with only the value missing in the 'Settore Merceologico' column, we tried two Machine Learning techniques for predicting categorical variables: *Logistic Regression* and *Random Forest Classifier*. We chose these techniques because we wanted to obtain more accurate and realistic imputations compared to Simple Imputation, especially since this column was crucial for our data analysis. The results obtained with both techniques were identical, so we used these values to replace all six missing values.

At the end of this phase, we recalculated the total number of NaN values in the dataset to check the progress and, starting from 1190, we were able to reduce the number to zero.

Data Cleaning (Error Detection & Correction Outliers)

In this phase, we considered the numerical columns for the identification of possible outliers.
As a first step, we removed the rows with a 'Civico' value equal to zero, as we considered them invalid data.

Afterward, we focused on analyzing the values in the columns related to the surfaces. We made sure that the surface values were not negative and compared various techniques for identifying outliers. We observed that STD and Z-Score were not suitable for non-Normal distributions. Therefore, we proceeded by considering other fitting distribution methods such as the Robust Z-Score (MAD), Percentage, and Interquartile Range (IQR), which are more appropriate for skewed distributions. We then decided to implement the model-based method k-Nearest Neighbors (KNN) to also account for the correlation between the surface columns, using the 95th percentile of the mean distance as the cut-off threshold. Using this method, 45 potential outliers were identified, which we chose to maintain in the dataset, as we believe that such numerical values may indeed represent actual real-world data.

Data Cleaning (Data Deduplication)

During the Data Deduplication phase, we first removed the exact duplicates and then identified candidate duplicates using the Record Linkage technique. To do so, we used two different indexing methods: Blocking and Sorted Neighbourhood. For both techniques, we chose an index containing the columns 'Tipo Via', 'Via' and 'Civico' because we thought this combination, which represents the address, would be the most effective at distinguishing between records. We believed the address was a good choice for grouping records that are likely to represent the same entity.
Regarding the Record Linkage comparing rules, we applied the same criteria for both methods. In particular we considered: 1) An exact match for the address columns ‘Tipo Via’, ‘Via’, ‘Civico’, ‘ZD’, ‘Codice Via’ and also for 'Settore Merceologico'. 2) For surface numerical columns we allowed for minor discrepancies. 3) For 'Insegna' names we decided to use a Jaro-Winkler similarity measure because it emphasizes prefix matching and is suitable for typographical errors.
When comparing the results, we found that both methods identified the same candidate pairs, with the Sorted Neighborhood method detecting one additional pair. Finally, for each of these pair of rows we decided which strategy to adopt which resulted in dropping 6 rows over the 7 candidate matches.

Data Analysis

In the final step of the pipeline, we conducted a data analysis, selecting 'Settore Merceologico' as the target column. Given that 'Settore Merceologico' is a categorical variable with well-defined classes, we chose a classification model, specifically the *KNeighborsClassifier*.
We performed the analysis on our cleaned dataset, removing the columns 'Tipo Via,' 'Via,' and 'Civico' because their combination represents a key and therefore does not provide meaningful information for the analysis itself. We used an 80:20 split for training and validation and considering the relatively small size of the dataset, we limited the number of splits to two.
We then conducted a similar analysis on the dirty dataset, this time also removing the 'Ubicazione' column, as it represents another key. Furthermore, since 'Settore Merceologico' contained missing values, we decided to retain them and simply impute them with the mode.

3. Results

In this section, we will discuss the main outcomes of our project pipeline and reflect on how our data preparation process impacted on the data quality and the data analysis.

Data Quality results

Below is a table comparing the initial and final scores for the three data quality dimensions — **accuracy**, **consistency**, and **completeness** — we defined for the pipeline. The initial values reflect the state of the dataset before cleaning, while the final values represent the results after completing the data preparation steps.

Dimension	Initial Score (%)	Final Score (%)
Completeness	88.1	100
Accuracy	99.1	100
Consistency	58	100

As we can see from the table, the changes made during the data preparation process contributed to the overall improvement in data quality across all dimensions. However, for the columns 'Insegna' and 'Superficie altri usi', a significant number of missing values were present. These were imputed with standard values (such as 'Non Specificato' for 'Insegna' and zero for 'Superficie altri usi'). Therefore, while these imputations helped us to complete the dataset and preserve valuable information, their accuracy may be questionable, as they might not accurately correspond to the real-world data.

Data Analysis results

The table below compares the classification analysis results on the column 'Settore Merceologico' using the dataset in its original (dirty) and cleaned forms. The initial results reflect the model's performance with the raw dataset, while the final results are based on the dataset after applying our data preparation techniques.

Metric	Dirty Dataset	Cleaned Dataset
Accuracy	0.8415	0.8595
F1-Score	**0.4628**	**0.6904**

Given the results of the classification analysis on both the cleaned and dirty datasets, it’s clear that the **F1-Score** provides a more meaningful assessment of model performance than accuracy. While the accuracy metric remained stable, the **F1-Score** demonstrated a much more significant improvement, rising from 0.4628 to 0.6864. These results suggest to us that the data cleaning process had a substantial positive impact, especially in terms of balancing the model's precision and recall.