Universidad Autónoma de Querétaro

Facultad de Ingeniería

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

October 3, 2023

Author:*Lic. Ijtsi Dzaya Ramos Morales*

Machine Learning , Dr. Marco Antonio Aceves Fernández

Report 2: Data imputation and normalization

Abstract

Machine learning (ML) plays a crucial role in healthcare research, particularly in disease prediction. This report focuses on enhancing a hepatitis detection dataset through data imputation and attribute normalization techniques. Missing data patterns are analyzed, and two imputation methods, mean imputation by category and MICE, are applied. Categorical attributes are encoded using one-hot encoding. Numerical attributes undergo min-max scaling and log transformation. Comparative assessments confirm the effectiveness of the processes. The resulting dataset becomes a valuable asset for hepatitis detection research, contributing to more precise analyses and medical diagnostics.

Introduction

In medical data analysis, the quality of the dataset can significantly impact the reliability of diagnostic tools and research outcomes. The dataset that will be the focus of this report pertains to hepatitis detection, a critical health concern. However, this dataset encounters common data challenges, including missing values, categorical attributes, and varied numerical scales. The primary objectives of this project are data imputation and normalization, both aimed at enhancing the dataset's utility for medical research and diagnosis.

Theoretical Foundation

Ensuring data quality is fundamental for robust and reliable research outcomes. This section provides a comprehensive overview of the key concepts of data imputation, categorical attribute encoding, and data normalization, which are vital for enhancing the hepatitis detection dataset.

Data Imputation

Missing data can introduce biases and hinder analysis. Three common patterns of missing data include:

- **MCAR (Missing Completely at Random)**: Data is missing in a random manner, unrelated to any variables.
- **MAR (Missing at Random)**: Data is missing in a way that depends on other observed variables but not on the missing data itself.
- **MNAR (Missing Not at Random)**: Data is missing in a way that depends on the missing data itself, posing challenges for imputation.

To address these challenges, a range of data imputation methods are employed, such as:

- **Mean and Median Imputation**: Filling missing values with the mean or median of the observed data.
- **Mean Imputation by Categories**: Imputing missing values based on the mean within specific categories.
- **Nearest Neighbor Imputation**: Imputing missing values based on similarity to other observations.
- **Random Imputation**: Assigning random values from the observed data.
- **Hot Deck Imputation**: Matching missing values with similar observed cases.
- **MCMC (Markov Chain Monte Carlo) Imputation**: Utilizing probabilistic modeling to impute missing data.
- **MICE (Multivariate Imputation by Chained Equations)**: Iteratively imputing missing data using conditional distributions.

The primary objective of data imputation is to systematically fill in missing values, aiming to restore dataset completeness and enable more accurate insights into hepatitis detection.

Categorical Attribute Encoding

Many attributes within the dataset are categorical, rendering them unsuitable for mathematical analysis. These attributes must be transformed into numerical

formats using encoding techniques. This conversion simplifies mathematical analysis without compromising data integrity.

- One-Hot Encoding: It creates binary columns for each category or label present in a categorical attribute. If an observation belongs to a particular category, the corresponding binary column is set to 1; otherwise, it is set to 0. This approach is ideal when there is no inherent ordinal relationship among the categories.
- Numeric Encoding Based on Ordinal Order: Each category is assigned a numerical value based on its position in the order. This allows the transformation of categorical data into a format that preserves the ordinal relationship, making it suitable for mathematical analysis.

The choice between one-hot encoding and numeric encoding depends on the nature of the categorical attribute.

Data Normalization

Numerical attributes in the dataset may exhibit varying scales, potentially impacting analysis and modeling. Data normalization is applied to ensure that all numerical attributes share a consistent scale. This process enhances the stability and convergence of analytical methods while preserving data integrity.

- Min-Max Scaling

  This method scales numerical attributes to a specific range, typically between 0 and 1. It is especially useful when the data's original scale varies significantly.

- Logarithmic Transformation

  Logarithmic transformation is another valuable tool in data normalization. It is applied to data that exhibits exponential growth or has a skewed distribution. Taking the logarithm of such data can help make it more symmetric and suitable for analysis.

Materials and methods

1. Data set: A database was obtained from the Kaggle platform, sourced from the UCI Machine Learning Repository. This database encompasses attributes pertinent to the diagnosis of hepatitis C, contributing to a comprehensive understanding of the disease's diagnostic factors.

   The dataset comprises a total of 615 observations, encompassing 14 attributes. These attributes encapsulate laboratory values and demographic information derived from both blood donors and patients diagnosed with hepatitis C. The data is characterized by its numerical nature, except for the "Category" and "Sex" attributes.

   The dataset's context underscores its focus on laboratory values, demographics, and diagnostic categorization. The data set originates from the UCI Machine Learning Repository.

2. Software: Python, a versatile programming language, was employed to fulfill the objectives of this project. The analysis involved the utilization of the `numpy`, `pandas`, `matplotlib`,`sklearn`, `seaborn` and `math` libraries, renowned for their capabilities in numerical computations, data manipulation, and graphical analysis.

3. Methods:

   - Mean Imputation by category: This method involves grouping the dataset by the categories within the categorical attribute that has missing values. Then, for each category group, the mean (average) value of the attribute is calculated. The missing values within each category are then replaced with their respective category's mean value.
   - MICE imputation: It involves creating a predictive model for each attribute with missing data, using other attributes as predictors. These models are built iteratively, imputing missing values for each attribute based on its own model and the imputed values from other attributes. The process continues for several iterations, refining imputations with each step.
   - Min-max normalization: For each selected attribute, the minimum and maximum values within that attribute are calculated. Then, each data point in the attribute is transformed by subtracting the minimum value and dividing by the range (maximum - minimum).
   - Log Transform normalization: Log transformation is applied to numerical attributes to alter their distribution. It involves taking the natural logarithm (or another specified base) of each data point in the attribute.

Results

The results depicted provide insights into the enhanced dataset's completeness, numerical representation of categorical attributes, and standardized numerical values.

Figure 1 shows the null values distribution within the instances of the database.

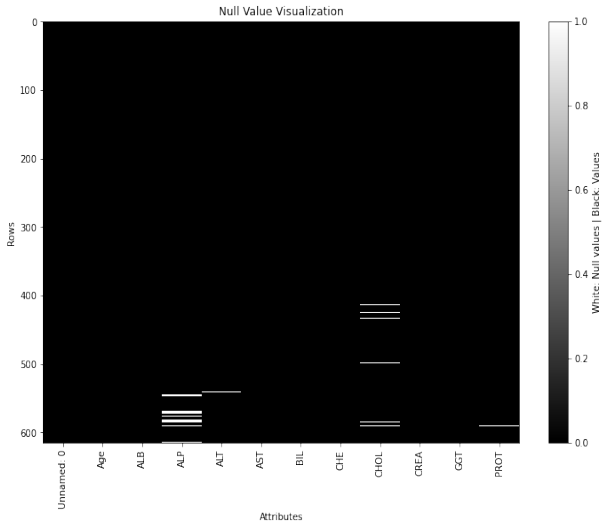

Figure 1: Null Values Visualization

Figure 2 presents the null values arranged by category and attribute. These two processes help to visualize the type of missing values that are being dealt with.

Category	Unnamed: 0	Category	Age	Sex	ALB	ALP	ALT	AST	\
0=Blood Donor	0	0	0	0	0	0	0	0	
0s=suspect Blood Donor	0	0	0	0	0	0	0	0	
1=Hepatitis	0	0	0	0	0	3	1	0	
2=Fibrosis	0	0	0	0	0	9	0	0	
3=Cirrhosis	0	0	0	0	1	6	0	0	

Category	BIL	CHE	CHOL	CREA	GGT	PROT
0=Blood Donor	0	0	7	0	0	0
0s=suspect Blood Donor	0	0	0	0	0	0
1=Hepatitis	0	0	0	0	0	0
2=Fibrosis	0	0	1	0	0	0
3=Cirrhosis	0	0	2	0	0	1

Figure 2: Null Values per attribute

Figure 3 displays the Pearson correlation between the numerical attributes, as to obtain more information regarding the relationship between the variables and to choose the best imputation method.

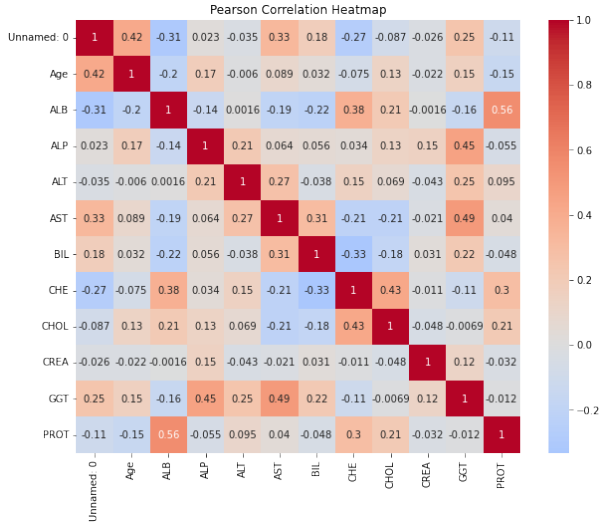

Figure 3: Pearson Correlation matrix

The distribution of before and after imputation by mean per category and MICE can be observed in figure 4 and 5 respectively.

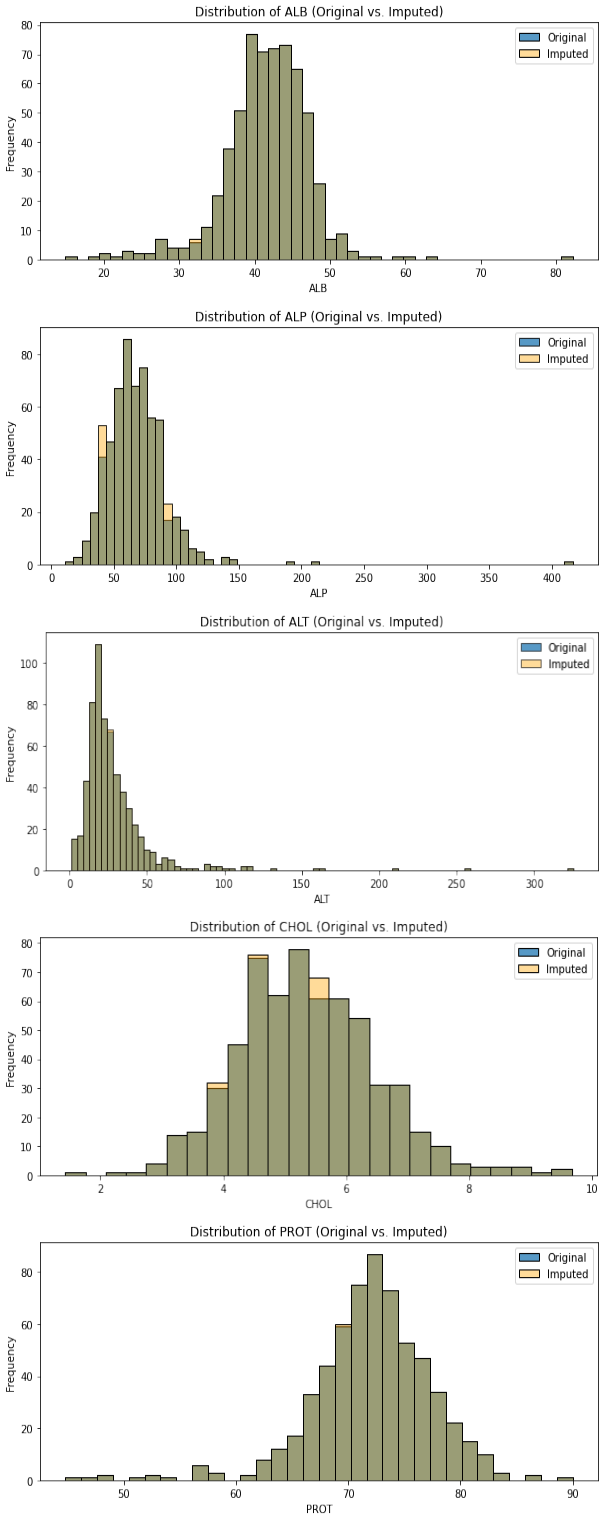

Figure 4: Comparison for Mean Imputation

The yellow colour depicts the imputed distribution, while the blue colour depicts the original distribution. The green colour is where the imputed and original distribution share the same values.

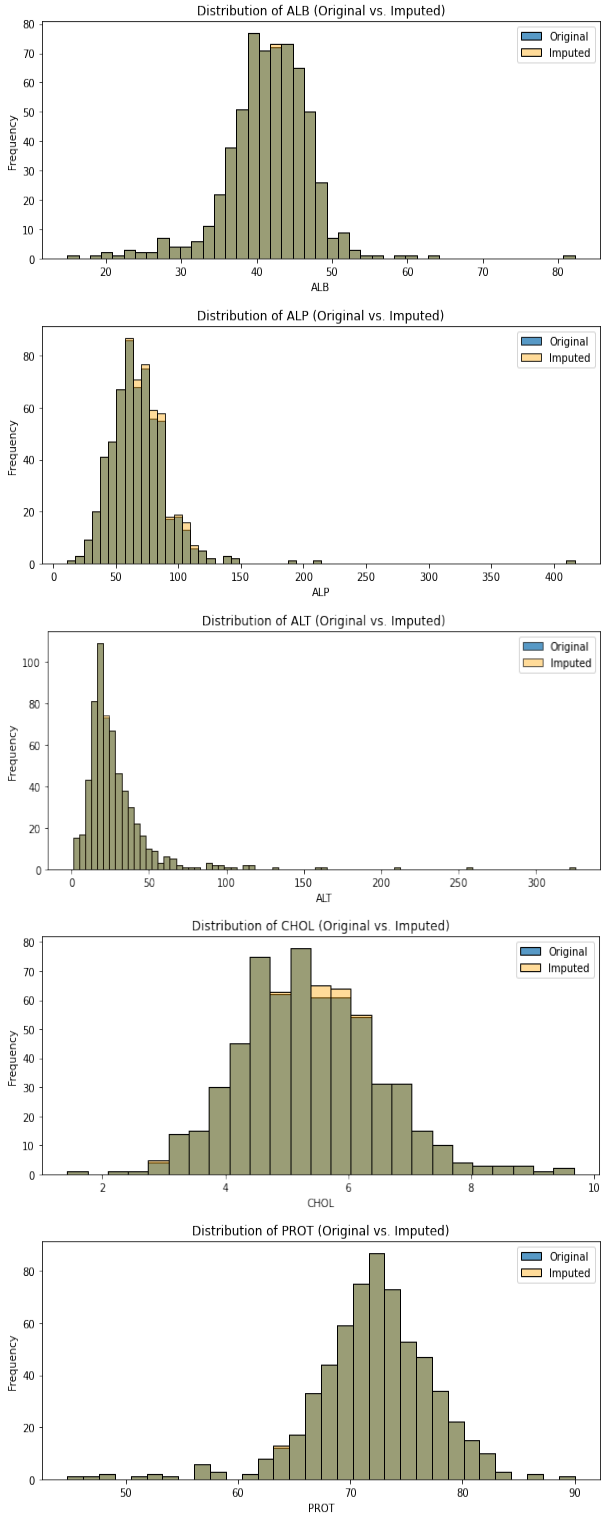

Figure 5: Comparison for MICE Imputation

Figure 6 shows the min-max normalization of the first five instances of the database and the encoding of the categorical values. Figure 7 shows the normalization via log transform for the first five instances of the database and the encoding of the categorical values.

	Unnamed: 0	Age	ALB	ALP	ALT	AST	BIL
0	-1.000000	-0.551724	-0.294118	-0.797531	-0.956923	-0.923567	-0.944882
1	-0.996743	-0.551724	-0.294118	-0.708642	-0.889231	-0.910828	-0.976378
2	-0.993485	-0.551724	-0.058824	-0.688889	-0.778462	-0.732484	-0.952756
3	-0.990228	-0.551724	-0.147059	-0.797531	-0.815385	-0.923567	-0.858268
4	-0.986971	-0.551724	-0.264706	-0.688889	-0.803077	-0.910828	-0.929134

	CHE	CHOL	CREA	GGT	PROT	Sex_f	Sex_m
0	-0.333333	-0.50	-0.816993	-0.975232	0.086957	-1.0	1.0
1	0.333333	-0.25	-0.876751	-0.965944	0.391304	-1.0	1.0
2	-0.066667	0.00	-0.854342	-0.910217	0.521739	-1.0	1.0
3	-0.200000	-0.25	-0.865546	-0.910217	0.347826	-1.0	1.0
4	0.066667	-0.25	-0.873016	-0.922601	0.043478	-1.0	1.0

Figure 6: Min-Max Normalization

	Unnamed: 0	Age	ALB	ALP	ALT	AST	BIL
0	0.301030	1.518514	1.591065	1.724276	0.903090	1.361728	0.903090
1	0.477121	1.518514	1.591065	1.851258	1.278754	1.397940	0.602060
2	0.602060	1.518514	1.672098	1.875061	1.568202	1.724276	0.845098
3	0.698970	1.518514	1.643453	1.724276	1.491362	1.361728	1.278754
4	0.778151	1.518514	1.602060	1.875061	1.518514	1.397940	1.000000

	CHE	CHOL	CREA	GGT	PROT	Sex_f	Sex_m
0	0.845098	0.602060	2.029384	1.113943	1.845098	0.0	0.30103
1	1.079181	0.698970	1.875061	1.204120	1.886491	0.0	0.30103
2	0.954243	0.778151	1.939519	1.531479	1.903090	0.0	0.30103
3	0.903090	0.698970	1.908485	1.531479	1.880814	0.0	0.30103
4	1.000000	0.698970	1.886491	1.477121	1.838849	0.0	0.30103

Figure 7: Log Transform

Discussion

In this section, the results of our missing data analysis and the methods we applied to handle missing values in the dataset are discussed.

0.1 Analysis of Missing Data Patterns

It was observed in Figure 1 that missing data appear to be distributed randomly across attributes, which means that the missing values do not depend on other attributes. Figure 2 shows was that a substantial portion of missing data occurs towards the end of the dataset, which could mean that the missing data could be due to errors during compilation of data of the last patients or during the last months.

0.2 Imputation Methods Selection

- **Unbalanced Database**: The choice of mean imputation by category and Multiple Imputation by Chained Equations (MICE) is justified by the unbalanced nature of the database. Given the database's imbalance, conventional imputation methods are preferred to minimize the risk of introducing biases.
- **Pearson Correlation Matrix (Figure 3)**: Low correlations between variables indicate that the MICE method is well-suited for this dataset, as there are no strong linear dependencies that could be leveraged by simpler imputation methods.

0.3 Evaluation of Imputation Quality

The imputation quality was evaluated by comparing the distributions between the original and imputed data, as shown in figure 4 and figure 5. The assessment indicated minimal change in the distributions.

0.4 Categorical Attribute Encoding

The dataset contains a single categorical attribute, "sex." Given that "sex" is non-ordinal and has only two distinct values (e.g., male and female), the chosen encoding method was one-hot encoding. This transformation effectively converts the categorical attribute into binary columns, ensuring that it can be incorporated into our analytical models without introducing an artificial ordinal relationship.

0.5 Normalization Techniques

Normalization is crucial to ensure that numerical attributes are on a consistent scale, facilitating robust analysis and modeling. The two normalization methods that were employed due to the presence of skewed attribute distributions were log transform and min-max.

- **Log Transform**: This transformation helps mitigate the impact of extreme values and approximates the data to a more normal distribution. It's worth noting that some original attribute values were negative. To accommodate these values, they were shifted to ensure that the log transformation could be applied without issue.
- **Min-Max Normalization**: This technique scales attribute values to a predefined range ([-1, 1]). Min-max normalization is effective when the original data spans a wide range and the purpose is to bring all values within a common interval.

1 Conclusion

The primary objectives of this study were to optimize a hepatitis detection dataset by addressing missing values and ensuring the appropriate encoding and normalization of attributes. The analysis and methodologies applied have yielded valuable insights into dataset enhancement, which can significantly benefit healthcare research and diagnostics.

The issue of missing data was successfully addressed through advanced imputation methods. The investigation revealed that missing data occurred at random and was not dependent on other attributes. However, a notable observation was the concentration of missing data in the latter instances of the dataset, possibly stemming from data compilation issues during the last months or for the final patients. To restore dataset completeness, mean imputation by category was employed as an effective method given the dataset's substantial class imbalance.

Furthermore, categorical attributes were encoded to prepare them for mathematical analysis. The dataset contained a single non-ordinal categorical attribute, "sex," which was encoded using one-hot encoding. This transformation ensured that the attribute could be incorporated into analytical models without introducing artificial ordinal relationships.

Normalization played a pivotal role in standardizing numerical attributes. Given the presence of skewed attribute distributions, two normalization techniques were employed. The log transform, applied to highly skewed attributes, mitigated the impact of extreme values and approximated data to a more normal distribution. For attributes unsuitable for log transformation, min-max normalization was chosen, scaling values to a common interval. Both methods enhanced the stability and convergence of analytical algorithms while maintaining data integrity.

Comparisons of attribute distributions before and after imputation and normalization revealed minimal changes, indicating the effectiveness of the data enhancement processes. The resulting dataset now stands as a valuable resource for healthcare research, medical diagnostics, and scientific analysis.

References

[1] D. M. Lane, "Introduction to Statistics," Rice University, 2003.

[2] M. A. Aceves Fernández, "Inteligencia Artificial para Programadores con Prisa," 2023.

[3] R. Lichtinghagen, F. Klawonn, and G. Hoffmann, "HCV data," 2020. DOI: https://doi.org/10.24432/C5D612.