

**Data**

**Pre-processing in Artificial Intelligence**

**Released By: Imène ALLOUCHE**

Table of Content

1. Introduction ……………………………………. 01
2. What is Data pre-processing …………………… 02
3. Why Data Pre-processing ………………………. 03
4. Crucial Steps of Data Preprocessing ……………. 04
5. Data Cleaning ……………………………… 05
6. Data Transformation ………………………. 06
7. Data Standardization ………………………
8. Data Reduction ……………………………. 07
9. Data splitting ………………………………. 08
10. Data sampling ……………………………… 09
11. Data Augmentation ………………………… 10
12. Conclusion

Introduction

Artificial intelligence models rely heavily on data. The quality and the quantity of data available is often a major determining factor in the performance of Artificial intelligence models.

The process of building an AI model typically involves training it on a large dataset which allows it to learn and recognize patterns and relationships between labels in data.

That says, the quality of the data is so important, the data should be clean, accurate and representative of the real-world problem.

Here comes the duty of Data preprocessing which takes care of improving the quality of data by cleaning it, transforming it in an adequate format, and augmenting it in order to make it suitable for analysis and modeling.

In short terms: Data preprocessing prepares the dataset to be used in a model, by sculpturing the raw data to an organized and a structured one in order to improve the performances of any AI model.

What is Data Preprocessing

Data preprocessing also known as data preparation is the process of preparing raw data for analysis and modeling.

It is a crucial step in the data science pipeline, as it ensures that the data is in the format that can be effectively used for analysis and modeling.

Data Preprocessing involves several techniques and steps to go through in order to ensure the improvement of the accuracy of the model which we can resume to:

1. Data cleaning: identify and removing irrelevant or inaccurate data by deleting duplicate records, correcting errors and handling missing values.
2. Data transformation: convert data into a more suitable format for analysis by encoding categorical variables, creating new features, applying mathematical functions to data.
3. Data reduction: reduce dimensionality f data by removing features that doesn’t add value to the model.
4. Data splitting: divide data into training and testing sets for model evaluation.
5. Data sampling: selecting a subset of data from a larger dataset to use for analysis
6. Data augmentation: generating new data points from the existing dataset to increase its size and diversity.

Why Data Preprocessing

The importance of data preprocessing lies on the fact that it is a crucial step in data analysis and machine learning process. It helps to ensure that the data is accurate, reliable, and suitable for the task at hand.

By performing data preprocessing, you can:

1. Improve the quality of the data by removing errors, inconsistencies, or missing values.
2. Prepare the data for analysis by transforming it into a format that is more suitable for the task at hand.
3. Extract the relevant features that are important for the task at hand and reduce the dimensionality of the data.
4. Improve model performance by increasing the number of data points through data augmentation and reducing overfitting.
5. Reduce the risk of errors or biases in the results by handling missing values and sampling the data appropriately.

In addition, data preprocessing can also save time and resources in the long run by reducing the need for additional cleaning and preparation once the analysis or modeling has begun. Overall, data preprocessing is an essential step that should not be overlooked in any data analysis or machine learning project as it can greatly improve the quality and accuracy of the results.

Crucial Steps of Data preprocessing

1. Data Cleaning:

Data cleaning is the process of identifying and correcting inaccuracies, inconsistencies, and missing data in a dataset. It involves cleaning and validating the data to ensure that it is accurate, consistent and complete. It can be classified into 3 categories:

|  |  |  |
| --- | --- | --- |
| Duplication | Handling Missing Values | Inconsistent data entry |
| * Comparing between the records * Removing the duplicates | * Drop them if they were missing because they don’t exist * Replace them (estimate them) if the values were missing because they haven’t been recorded | * Efficiently fixing the typos in the data * Reunite the names that have the same meanings (example: USA and United States of America under the same name) |

1. Data Transformation:
2. Data Standardization:

|  |  |
| --- | --- |
|  |  |
|  |  |

Sub Steps vocabulary:

|  |  |
| --- | --- |
| Step | Explanation |
| Feature selection | Identifying the most relevant features for the analysis and removing those that are not informative |
| Data exploration | Visualizing and summarizing the data to understand its distribution and relationships |
| Data Imputation | Replacing missing values with estimated values |
| Data integration | Merging multiple datasets together to form a more complete dataset |
| Data validation | Checking the quality and integrity of the dataset |
| Data anonymization | Masking or removing any personal or sensitive information from the dataset |
| Data Discretization | Converting numerical values into categorical values |
| Data balancing | Resampling the data to balance the ratio of different classes in the dataset |
| Data Harmonization | Making different datasets consistent and comparable, for example by converting units or resolving naming inconsistencies |
| Data enrichment | adding additional information to the dataset from external sources |
| Data standardization | Converting the data into a standard format |
| Data verification | Cross-checking the data with external sources or ground truth to confirm its accuracy |