# Data Preprocessing Toolkit : An Approach to Automate Data Preprocessing

**Deepak Varma, Alwala Nehansh**

**Mrs P Swathy (Guide)**

Department of Computer Science and Engineering
Hyderabad Institute of Technology and Management

---***---

**Abstract -**Data Pre-processing transforms the data into a format that is more easily and effectively processed in data mining, machine learning, and other data science tasks. The Data Pre-processing techniques are generally used at the earlier stages of the machine learning and AI development pipeline to ensure accurate results. Data must first go through several pre-processing steps before the machine learning model can use it. We intend to build a GUI that allows users to input the data with inconsistencies and various options are provided to pre-process. After the data is processed, we provide the users an option to analyze the data which gives them clarity on the further aspects of preprocessing. Pre-processing consists of three main phases i.e. Data Cleaning, Data Transformation, and Data Reduction. Data cleaning involves correcting or removing inaccurate, corrupted, improperly formatted, duplicate, or incomplete data from a dataset. Data transformation is a technique used to convert data into a suitable format that eases data mining and retrieves strategic information. Data reduction is the process of shrinking the size of the original data so that it can be represented in a much smaller volume. Preprocessing requires a lot of work, what if we could automate it? With just one click, this web application is capable of transforming inconsistent data into consistent data that could be used further. The GUI is built using the Streamlit library in python and in the backend, we intend to use the AutoClean library along with many other libraries based on different formats of data..

***Key Words*****:** Data Cleaning, AutoClean, Data Transformation, Data Reduction, Streamlit.

## 1.INTRODUCTION

Data can be defined as a systematic record of a particular quantity. It is the different values of that quantity represented together in a set. It is a collection of facts and figures to be used for a specific purpose such as a survey or analysis. When arranged in an organized form, it can be called information. The source of data is also an important factor.

Data preprocessing changes the data into a format that can be processed in data mining, machine learning, and other data science tasks more quickly and efficiently. To ensure accurate results, the techniques are typically applied at the very beginning of the machine learning and AI development pipeline.

Proper data preprocessing determines whether a machine learning model succeeds or fails. Since better data outperforms fancier algorithms, professional data scientists typically devote a significant amount of their time to this step. If the dataset is thoroughly cleaned, there is a chance that we can get good results using straightforward algorithms as well. This can be very helpful at times, especially when it comes to computation when the dataset size is large.

Data may be qualitative or quantitative. Once you know the difference between them, you can know how to use them.

- Qualitative Data: They represent some characteristics or attributes. They depict descriptions that may be observed but cannot be computed or calculated. For example, data on attributes such as intelligence, honesty, wisdom, cleanliness, and creativity collected using the students of your class a sample would be classified as qualitative. They are more exploratory than conclusive in nature.
- Quantitative Data: These can be measured and not simply observed. They can be numerically represented

and calculations can be performed on them. For example, data on the number of students playing different sports from your class gives an estimate of how many of the total students play which sport. This information is numerical and can be classified as quantitative.

Knowledge Discovery in Databases (KDD) is a method for obtaining important data from massive data sources. Data mining is a KDD step that uses classification, clustering, association rules, and many other techniques to analyze and model large datasets. Because of their size, the number of resources used to collect them, and the methods used, the raw data are extremely susceptible to missing, noise, outliers, and inconsistent data[6]. The results of data mining will be impacted by poor quality data. Preprocessing techniques must be used on data as a result to increase their effectiveness.

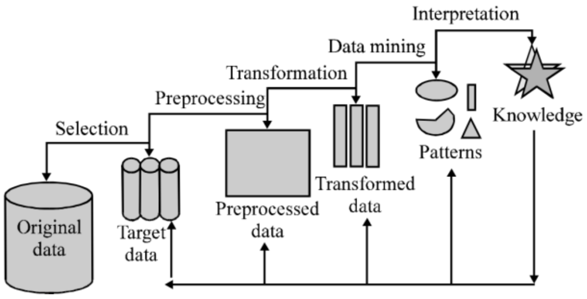


Fig. 1. KDD

The data must be chosen to determine the target data, as shown in Fig. 1, and then the chosen data must be preprocessed to increase its reliability. The data must be transformed into a format that is appropriate for the data mining process after being preprocessed. Then, the mining process will be used to extract patterns that are interrupted and evaluated in the final step, such as clustering, classification, regression, etc.

The Phases of Knowledge discovery process are:

- Data Selection: The process of choosing the appropriate data source, type, and tools to collect the data is known as data selection.
- Data Preprocessing : Data Pre-processing transforms the data into a format that is more easily and effectively processed in data mining, machine learning, and other data science tasks
- Data Transformation : Data transformation is a technique used to convert data into a suitable format that eases data mining and retrieves strategic information
- Data Mining : When large data sets are sorted through to find patterns and relationships that can be used to solve business problems through data analysis, the process is known as data mining.
- Data Interpretation : Data interpretation is the process of reviewing data using a variety of analytical techniques and drawing pertinent conclusions.

This illustrates how data preprocessing is an important step in knowledge discovery and how it can help with data mining tasks.

## 2. IMPLEMENTATION

We Intend to implement a GUI that enables users to preprocess the data with a single click. The user only needs to enter the dataset that contains some discrepancies. Once that data is entered, the user can preview the dataset and choose from several options, including handling duplicates, filling in missing values, and handling outliers. The GUI is created using Streamlit, a Python-based framework that enables the creation of interactive web applications. All necessary operations are carried out at the backend using a modified version of the AutoClean library. After the data is cleaned, the user can analyze the data using EDA(Exploratory Data Analysis) which is built using Pandas.Regarding the activity of data preprocessing, the following aspects have been implemented.

- Tabular Data Cleaning
- Data Transformation
- Image Augmentation
- Exploratory Data Analysis

## 3. DATA CLEANING

Data cleaning involves identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.A machine learning model's success or failure is determined by how well the data is cleaned. Professional data scientists typically devote a significant amount of their time to this step because better data outperforms fancier algorithms. There is a chance that if the dataset is thoroughly cleaned, we can also achieve good results with simple algorithms Especially when it comes to computation when the dataset size is large, this can be very useful at times.

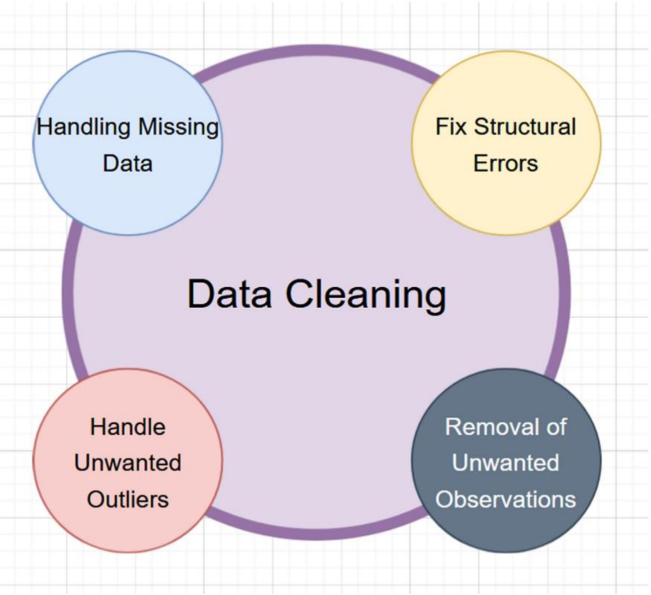


Fig. 2. Delta Debugging workflow

Data Cleaning is the first step in the process of data preprocessing,

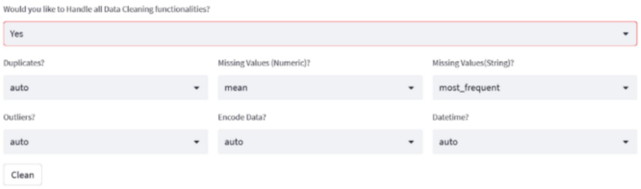
3.1 Data Cleaning Operations

The Following operations are supported by the data preprocessing toolkit :

3.1.1 Handling Missing Values

Missing values directly affect the ML Model, so handling them is necessary. Among the techniques for handling missing values are:

- Ignore The tuples
- Fill the missing value using statistical measures like mean , median.
- Replace missing value with most frequently occurring value
- Using Supervised learning algorithms like KNN to predict the missing value



Would you like to Handle all Data Cleaning functionalities?
Yes

Duplicates? auto
Missing Values (Numeric)? mean
Missing Values(String)? most_frequent
Outliers? auto
Encode Data? auto
Datetime? auto

Fig. 3. Delta Cleaning Tasks

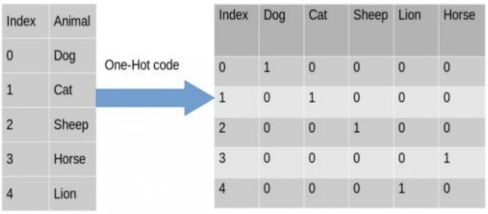
Fig. 3. shows all the data cleaning tasks supported by the Data Preprocessing toolkit.

3.1.2 Encoding

Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the models to give and improve the prediction Techniques.

There are 2 techniques used here :

- One-Hot Encoding : When the features are nominal, we employ this method of categorical data encoding. We produce a new variable in a single hot encoding for each level of a categorical feature. A binary variable with the values 0 or 1 is mapped to each category. Here, 0 denotes the absence of that category and 1 denotes its presence.



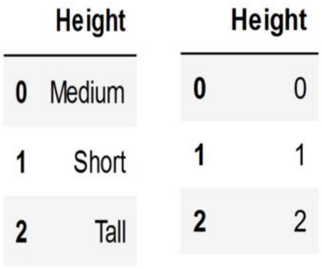
| Index | Animal |
|---|---|
| 0 | Dog |
| 1 | Cat |
| 2 | Sheep |
| 3 | Horse |
| 4 | Lion |

One-Hot code

| Index | Dog | Cat | Sheep | Lion | Horse |
|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 1 |
| 4 | 0 | 0 | 0 | 1 | 0 |

Fig. 4. One-Hot Encoding

- Label Encoding :Label encoding is the process of transforming labels into a numeric form so that they can be read by machines. computer learning After that, algorithms can decide more effectively how those labels should be used. It is a crucial step in the supervised learning pre-processing of the structured dataset.



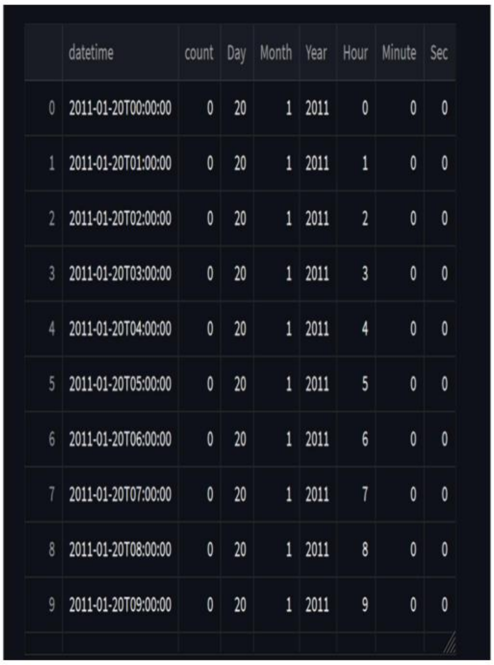
| | Height |
|---|---|
| 0 | Medium |
| 1 | Short |
| 2 | Tall |

| | Height |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 2 | 2 |

Fig. 5. Label Encoding

3.1.3 Extract Date-time

AutoClean can search the data for date time features, and extract the values to separate columns. When set to 's', it extracts the date time values up to the seconds i. e. day, month, year, hour, minutes, seconds. You can set the granularity of the extraction manually by setting extract_datetime to 'D' for day, 'M' for month, 'Y' for year, 'h' for hour, 'm' for minutes or to False if you want to skip this step.



| | datetime | count | Day | Month | Year | Hour | Minute | Sec |
|---|---|---|---|---|---|---|---|---|
| 0 | 2011-01-20T00:00:00 | 0 | 20 | 1 | 2011 | 0 | 0 | 0 |
| 1 | 2011-01-20T01:00:00 | 0 | 20 | 1 | 2011 | 1 | 0 | 0 |
| 2 | 2011-01-20T02:00:00 | 0 | 20 | 1 | 2011 | 2 | 0 | 0 |
| 3 | 2011-01-20T03:00:00 | 0 | 20 | 1 | 2011 | 3 | 0 | 0 |
| 4 | 2011-01-20T04:00:00 | 0 | 20 | 1 | 2011 | 4 | 0 | 0 |
| 5 | 2011-01-20T05:00:00 | 0 | 20 | 1 | 2011 | 5 | 0 | 0 |
| 6 | 2011-01-20T06:00:00 | 0 | 20 | 1 | 2011 | 6 | 0 | 0 |
| 7 | 2011-01-20T07:00:00 | 0 | 20 | 1 | 2011 | 7 | 0 | 0 |
| 8 | 2011-01-20T08:00:00 | 0 | 20 | 1 | 2011 | 8 | 0 | 0 |
| 9 | 2011-01-20T09:00:00 | 0 | 20 | 1 | 2011 | 9 | 0 | 0 |

Fig. 6. Date-Time Extraction

## 4. DATA TRANSFORMATION

Data transformation is a method for transforming data into a format that makes it easier to conduct data mining and retrieve strategic information.

This GUI supports data transformation using normalization.

4.1 Min-Max Normalisation

The original data is transformed linearly in this method of data normalization. The minimum and maximum value from the data are retrieved, and each value is replaced using the formula below.

$$v' = \frac{v - \text{minF}}{\text{maxF} - \text{minF}}(new\_max_F - new\_min_F) + new\_min_F \,,$$

Fig. 7. Min-Max Normalization

Where,

- v = old value of entry , V'= new value of entry
- minF and maxF maximum and minimum value of that column
- new_maxF and new_minF are maximum and minimum values from the required range.

4.2 Z-Score Normalization

In this technique, values are normalized based on mean and standard deviation of the data A. The formula used is :

$$V' = V - \bar{x} / \sigma x$$

Where,

- V' , V are new and old entry of data
- $\bar{X}$ = Mean of X and
- σx = Standard Deviation of X

## 5. IMAGE AUGMENTATION

By modifying the existing data, image augmentation creates new data that can be used for model training. In other words, it is the process of enhancing the dataset that is made available for deep learning model training.

Image Augmentation techniques include :

- Image Rotation : Rotate the image to generate new angles.
- Image Flipping : Flip the image left, right, or upside down.
- Image ReSize : change the size of image
- Image Scaling : Increase or decrease the picture size.

## 6. CONCLUSION

With this paper, we have attempted to demonstrate data preprocessing toolkit: a GUI that reduces the human effort of manual tasks. There are a number of benefits of preprocessing the data including De-Duplication of data, reduced noise in data, and handling missing or erroneous entries. The Goal here is to provide the user with a one-stop platform that allows the user to input the data and perform various preprocessing tasks like data cleaning, data transformation, data reduction, image augmentation, audio augmentation, and at the end EDA. The intuition here is to reduce as much time invested in preprocessing and put that time into further tasks after preprocessing like building the Machine Learning model or Extracting patterns using Data Mining.

## REFERENCES

1. CFamili, A. et al. 'Data Preprocessing and Intelligent Data Analysis'. 1 Jan. 1997 : 3 – 23.

2. Kotsiantis, Sotiris & Kanellopoulos, Dimitris & Pintelas, P.. (2006). Data Preprocessing for Supervised Learning. International Journal of Computer Science. 1. 111-117.

3. Puneet Mishra, Alessandra Biancolillo, Jean Michel Roger, Federico Marini, Douglas N. Rutledge,New data preprocessing trends based on ensemble of multiple preprocessing techniques.

4. Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Woźniak, Francisco Herrera, A survey on data preprocessing for data stream mining: Current status and future directions,Neurocomputing.

6. Suad A. Alasadi and Wesam S. Bhaya, 2017. Review of Data Preprocessing Techniques in Data Mining. Journal of Engineering and Applied Sciences, 12: 4102-4107.