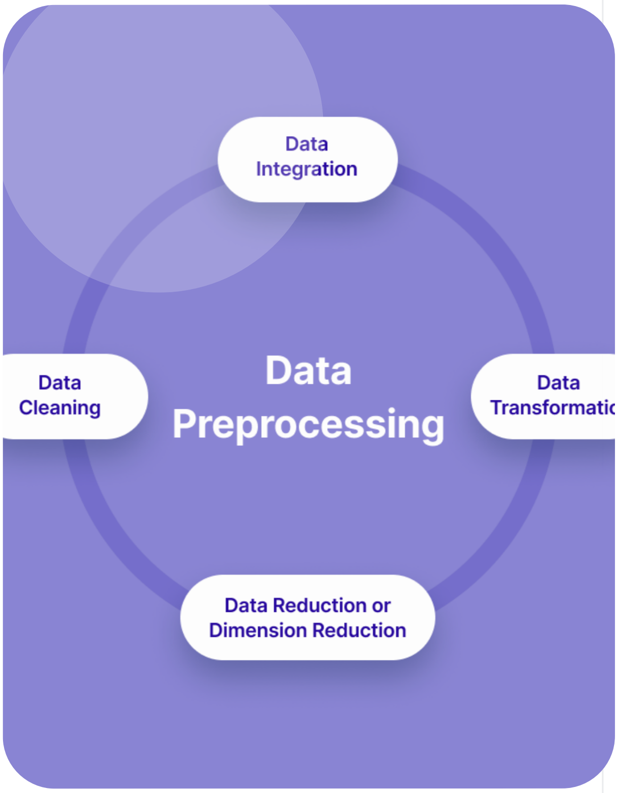
A Simple Guide to Data Preprocessing in Machine Learning

How to improve your data quality to build more accurate AI models? Learn about data preprocessing steps you need to take in order to convert raw data into the processed form.



5 min read · July 19, 2022

Pragati Baheti
Microsoft

Contents

Data is no less than an asset in today's world. But—

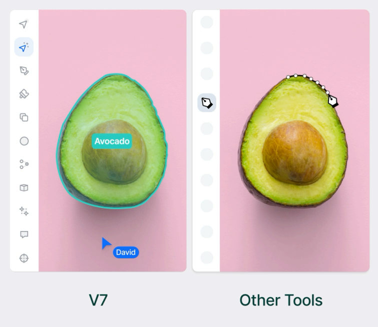
Well, not exactly.

Data in the real world is quite dirty and corrupted with inconsistencies, noise, incomplete information, and missing values. It is aggregated from diversified sources using data mining and warehousing techniques.

It is a common thumb rule in machine learning that the greater the amount of data we have, the better models we can train.

In this article, we will discuss all Data Preprocessing steps one needs to follow to convert raw data into the processed form.

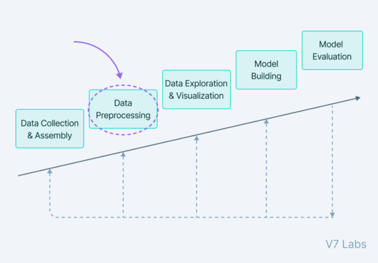
Here's what we'll cover:



What is Data Preprocessing?

Data Preprocessing includes the steps we need to follow to transform or encode data so that it may be easily parsed by the machine.

The main agenda for a model to be accurate and precise in predictions is that the algorithm should be able to easily interpret the data's features.



Why is Data Preprocessing important?

The majority of the real-world datasets for machine learning are highly susceptible to be missing, inconsistent, and noisy due to their heterogeneous origin.

Applying data mining algorithms on this noisy data would not give quality results as they would fail to identify patterns effectively. Data Processing is, therefore, important to improve the overall data quality.

- Duplicate or missing values may give an incorrect view of the overall statistics of data.
- Outliers and inconsistent data points often tend to disturb the model's overall learning, leading to false predictions.

Quality decisions must be based on quality data. Data Preprocessing is important to get this quality data, without which it would just be a *Garbage In, Garbage Out* scenario.

💡 **Pro tip:** Check out An Introductory Guide to Quality Training Data for Machine Learning to learn more.

Features in machine learning

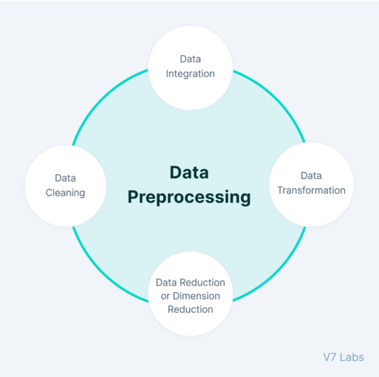
Individual independent variables that operate as an input in our machine learning model are referred to as features. They can be thought of as representations or attributes that describe the data and help the models to predict the classes/labels.

For example, features in a structured dataset like in a CSV format refer to each column representing a measurable piece of data that can be used for analysis: Name, Age, Sex, Fare, and so on.

4 Steps in Data

Preprocessing

Now, let's discuss more in-depth four main stages of data preprocessing.



Data Cleaning

Data Cleaning is particularly done as part of data preprocessing to clean the data by filling missing values, smoothing the noisy data, resolving the inconsistency, and removing outliers.

1. Missing values

Here are a few ways to solve this issue:

- Ignore those tuples

This method should be considered when the dataset is huge and numerous missing values are present within a tuple.

- Fill in the missing values

There are many methods to achieve this, such as filling in the values manually, predicting the missing values using regression method, or numerical methods like attribute mean.

2. Noisy Data

It involves removing a random error or variance in a measured variable. It can be done with the help of the following techniques:

- Binning

It is the technique that works on sorted data values to smoothen any noise present in it. The data is divided into equal-sized bins, and each bin/bucket is dealt with independently. All data in a segment can be replaced by its mean, median or boundary values.

- Regression

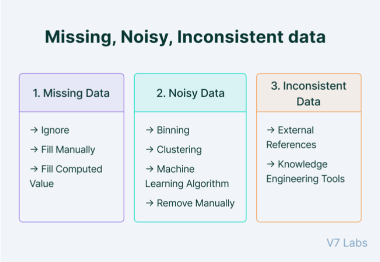
This data mining technique is generally used for prediction. It helps to smoothen noise by fitting all the data points in a regression function. The linear regression equation is used if there is only one independent attribute; else Polynomial equations are used.

- Clustering

Creation of groups/clusters from data having similar values. The values that don't lie in the cluster can be treated as noisy data and can be removed.

3. Removing outliers

Clustering techniques group together similar data points. The tuples that lie outside the cluster are outliers/inconsistent data.



Data Integration

Data Integration is one of the data preprocessing steps that are used to merge the data present in multiple sources into a single larger data store like a data warehouse.

Data Integration is needed especially when we are aiming to solve a real-world scenario like detecting the presence of nodules from CT Scan images. The only option is to integrate the images from multiple medical nodes to form a larger database.

💡 **Pro tip:**
Looking for quality medical datasets? Check out 21+ Best Healthcare Datasets for Computer Vision.

We might run into some issues while adopting Data Integration as one of the Data Preprocessing steps:

- Schema integration and object matching: The data can be present in different formats, and attributes that might cause difficulty in data integration.
- Removing redundant attributes from all data sources.
- Detection and resolution of data value conflicts.

Data Transformation

Once data clearing has been done, we need to consolidate the quality data into alternate forms by changing the value, structure, or format of data using the below-mentioned Data Transformation strategies.

Generalization

The low-level or granular data that we have converted to high-level information by using concept hierarchies. We can transform the primitive data in the address like the city to higher-level information like the country.

Normalization

It is the most important Data Transformation technique widely used. The numerical attributes are scaled up or down to fit within a specified range. In this approach, we are constraining our data attribute to a particular container to develop a correlation among different data points. Normalization can be done in multiple ways, which are highlighted here:

- Min-max

normalization

- Z-Score normalization
- Decimal scaling normalization

Attribute Selection

New properties of data are created from existing attributes to help in the data mining process. For example, date of birth, data attribute can be transformed to another property like is_senior_citizen for each tuple, which will directly influence predicting diseases or chances of survival, etc.

Aggregation

It is a method of storing and presenting data in a summary format. For example sales, data can be aggregated and transformed to show as per month and year format.

Data Reduction

The size of the dataset in a data warehouse can be too large to be handled by data analysis and data mining algorithms.

One possible solution is to

obtain a reduced representation of the dataset that is much smaller in volume but produces the same quality of analytical results.

Here is a walkthrough of various Data Reduction strategies.

Data cube aggregation

It is a way of data reduction, in which the gathered data is expressed in a summary form.

Dimensionality reduction

Dimensionality reduction techniques are used to perform feature extraction. The dimensionality of a dataset refers to the attributes or individual features of the data. This technique aims to reduce the number of redundant features we consider in machine learning algorithms. Dimensionality reduction can be done using techniques like Principal Component Analysis etc.

Data compression

By using encoding technologies, the size of the data can significantly reduce. But compressing data can be either lossy or non-lossy. If original data can be obtained after reconstruction from compressed data, this is referred to as lossless reduction; otherwise, it is referred to as lossy reduction.

Discretization

Data discretization is used to divide the attributes of the continuous nature into data with intervals. This is done because continuous features tend to have a smaller chance of correlation with the target variable. Thus, it may be harder to interpret the results. After discretizing a variable, groups corresponding to the target can be interpreted. For example, attribute age can be discretized into bins like below 18, 18-44, 44-60, above 60.

Numerosity reduction

The data can be represented as a model or equation like a regression model. This would save the burden of storing huge datasets instead of a model.

Attribute subset selection

It is very important to be specific in the selection of attributes. Otherwise, it might lead to high dimensional data, which are difficult to train due to underfitting/overfitting problems. Only attributes that add more value towards model training should be considered, and the rest all can be discarded.

Data Quality Assessment

Data Quality Assessment includes the statistical approaches one needs to follow to ensure that the data has no issues. Data is to be used for operations, customer management, marketing analysis, and decision making—hence it needs to be of high quality.

The main components of Data Quality Assessment include:

1. The completeness with no missing attribute values

2. Accuracy and reliability in terms of information
3. Consistency in all features
4. Maintain data validity
5. It does not contain any redundancy

Data Quality Assurance process has involves three main activities.

1. **Data profiling**: It involves exploring the data to identify the data quality issues. Once the analysis of the issues is done, the data needs to be summarized according to no duplicates, blank values etc identified.
2. **Data cleaning**: It involves fixing data issues.
3. **Data monitoring**: It involves maintaining data in a clean state and having a continuous check on business needs being satisfied by the data.

💡 **Pro tip:** Have a look at our Data Annotation Tutorial to learn more about data labeling.

Data Preprocessing: Best practices

Here's a short recap of everything we've learnt about data preprocessing:

- The first step in Data Preprocessing is to understand your data. Just looking at your dataset can give you an intuition of what things you need to focus on.
- Use statistical methods or pre-built libraries that help you visualize the dataset and give a clear image of how your data looks in terms of class distribution.
- Summarize your data in terms of the number of duplicates, missing values, and outliers present in the data.
- Drop the fields you think have no use for the modeling or are closely related to other attributes. Dimensionality reduction is one of the very important aspects of Data Preprocessing.
- Do some feature

engineering and figure out which attributes contribute most towards model training.

💡 **Read more:**

The Ultimate Guide to Object Detection

Optical Character Recognition: What is It and How Does it Work [Guide]

A Gentle Introduction to Image Segmentation for Machine Learning and AI

Image Classification Explained: An Introduction

The Ultimate Guide to Semi-Supervised Learning

The Beginner’s Guide to Contrastive Learning

9 Reinforcement Learning Real-Life Applications

Mean Average Precision (mAP) Explained: Everything You Need to Know

A Step-by-Step Guide to Text Annotation [+Free OCR Tool]

The Essential Guide to Data Augmentation in Deep Learning

Related articles

MACHINE LEARNING

What is Overfitting in Deep Learning and How to Avoid It

Pragati Baheti 7 min read

MACHINE LEARNING

Data Annotation Tutorial: Definition, Tools, Datasets

Nilesh Barla 13 min read

MACHINE LEARNING

Autoencoders in Deep Learning: Tutorial & Use Cases [2022]

Hmrishav Bandyopadhyay 7 min read

Subscribe to our blog

1 personalized email from V7's CEO per month

Your email →

COMPANY

About
Pricing
Contact Us
Jobs
News
Data Security

PLATFORM

Image Annotation
Video Annotation
Dataset Management
Document Processing
Model Training
Labeling Services

SOLUTIONS

Agriculture
Automotive
Construction
Energy
Food & Beverage
Government
Healthcare
Insurance & Finance
Life Sciences & Biotech
Logistics
Manufacturing
Retail
Software & Internet
Sports

COMMUNITY

Blog
Documentation
Community
Academy
Open Datasets
ML Glossary
Ethics & CoC

RESOURCES

V7 vs Scale AI
V7 vs Labelbox
V7 vs Dataloop
V7 vs Superannotate
V7 vs CVAT
V7 vs Clarifai
V7 vs Hive
V7 vs Hasty
V7 vs Appen
V7 vs Roboflow
V7 vs Supervisely
V7 vs Playment
V7 vs Kili Technology
V7 vs Innotescus