What Is Data Preprocessing?

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

Machines like to process nice and tidy information – they read data as 1s and 0s. So calculating structured data, like whole numbers and percentages is easy. However, unstructured data, in the form of text and images must first be cleaned and formatted before analysis.

Data Preprocessing Importance

When using data sets to train machine learning models, you'll often hear the phrase "garbage in, garbage out" This means that if you use bad or "dirty" data to train your model, you'll end up with a bad, improperly trained model that won't actually be relevant to your analysis.

Good, preprocessed data is even more important than the most powerful algorithms, to the point that machine learning models trained with bad data could actually be harmful to the analysis you're trying to do – giving you "garbage" results.

Depending on your data gathering techniques and sources, you may end up with data that's out of range or includes an incorrect feature, like household income below zero or an image from a set of "zoo animals" that is actually a tree. Your set could have missing values or fields. Or text data, for example, will often have misspelled words and irrelevant symbols, URLs, etc.

When you properly preprocess and clean your data, you'll set yourself up for much more accurate downstream processes. We often hear about the importance of "data- driven decision making," but if these decisions are driven by bad data, they're simply bad decisions.

Understanding Machine Learning Data Features

Data sets can be explained with or communicated as the "features" that make them up. This can be by size, location, age, time, color, etc. Features appear as columns in datasets and are also known as attributes, variables, fields, and characteristics.

A machine learning data feature as "an individual measurable property or characteristic of a phenomenon being observed". It's important to understand what "features" are when preprocessing your data because you'll need to choose which ones to focus on depending on what your business goals are. Later, we'll explain how you can improve the quality of your dataset's features and the insights you gain with processes like feature selection

First, let's go over the two different types of features that are used to describe data:

categorical and numerical:

Categorical features: Features whose explanations or values are taken from a defined set of possible explanations or values. Categorical values can be colors of a house; types of animals;

months of the year; True/False; positive, negative, neutral, etc. The set of possible categories that the features can fit into is predetermined.

Numerical features: Features with values that are continuous on a scale, statistical, or integer-related. Numerical values are represented by whole numbers, fractions, or percentages. Numerical features can be house prices, word counts in a document, time it takes to travel somewhere, etc.

Data Preprocessing Steps

Let’s take a look at the established steps you’ll need to go through to make sure your

data is successfully preprocessed.

1. Data quality assessment

2. Data cleaning

3. Data transformation

4. Data reduction

1. Data quality assessment

Take a good look at your data and get an idea of its overall quality, relevance to your project, and consistency. There are a number of data anomalies and inherent problems to look out for in almost any data set, for example:

- Mismatched data types: When you collect data from many different sources, it may come to you in different formats. While the ultimate goal of this entire process is to reformat your data for machines, you still need to begin with similarly formatted data. For example, if part of your analysis involves family income from multiple countries, you’ll have to convert each income amount into a single currency.
- Mixed data values: Perhaps different sources use different descriptors for features – for example, man or male. These value descriptors should all be made uniform.
- Data outliers: Outliers can have a huge impact on data analysis results. For example if you're averaging test scores for a class, and one student didn’t respond to any of the questions, their 0% could greatly skew the results.
- Missing data: Take a look for missing data fields, blank spaces in text, or unanswered survey questions. This could be due to human error or incomplete data. To take care of missing data, you’ll have to perform data cleaning.

2. Data cleaning

Data cleaning is the process of adding missing data and correcting, repairing, or removing incorrect or irrelevant data from a data set. Dating cleaning is the most important step of preprocessing because it will ensure that your data is ready to go for your downstream needs.

Data cleaning will correct all of the inconsistent data you uncovered in your data quality assessment. Depending on the kind of data you're working with, there are a number of possible cleaners you'll need to run your data through.

Missing data

There are a number of ways to correct for missing data, but the two most common are:

 Ignore the tuples: A tuple is an ordered list or sequence of numbers or entities. If multiple values are missing within tuples, you may simply discard the tuples with that missing information. This is only recommended for large data sets, when a few ignored tuples won't harm further analysis.

 Manually fill in missing data: This can be tedious, but is definitely necessary when working with smaller data sets.

Noisy data

Data cleaning also includes fixing "noisy" data. This is data that includes unnecessary data points, irrelevant data, and data that's more difficult to group together.

 Binning: Binning sorts data of a wide data set into smaller groups of more similar data. It's often used when analyzing demographics. Income, for example, could be grouped: $35,000-$50,000, $50,000-$75,000, etc.

 Regression: Regression is used to decide which variables will actually apply to your analysis. Regression analysis is used to smooth large amounts of data. This will help you get a handle on your data, so you're not overburdened with unnecessary data.

 Clustering: Clustering algorithms are used to properly group data, so that it can be analyzed with like data. They're generally used in unsupervised learning, when not a lot is known about the relationships within your data.

If you're working with text data, for example, some things you should consider when cleaning your data are:

 Remove URLs, symbols, emojis, etc., that aren't relevant to your analysis

 Translate all text into the language you'll be working in

 Remove HTML tags

 Remove boilerplate email text

 Remove unnecessary blank text between words

 Remove duplicate data

After data cleaning, you may realize you have insufficient data for the task at hand. At this point you can also perform data wrangling or data enrichment to add new data sets and run them through quality assessment and cleaning again before adding them to your original data.

3. Data transformation

With data cleaning, we've already begun to modify our data, but data transformation will begin the process of turning the data into the proper format(s) you'll need for analysis and other downstream processes.

This generally happens in one or more of the below:

1. Aggregation

2. Normalization

3. Feature selection

4. Discreditization

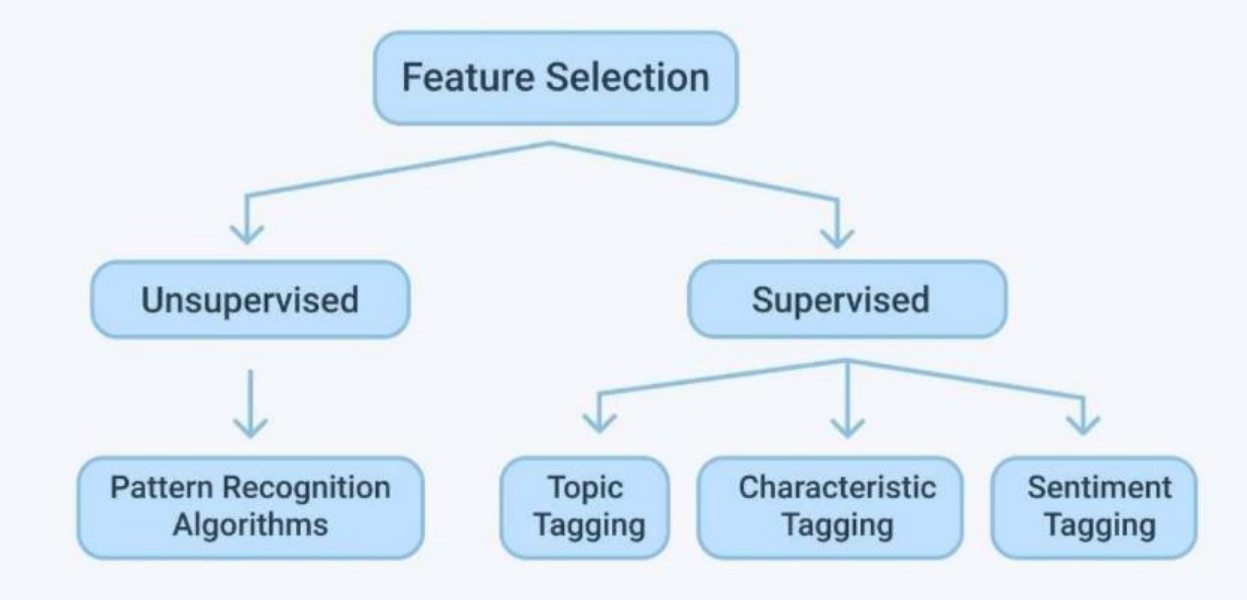
5. Concept hierarchy generation

 Aggregation: Data aggregation combines all of your data together in a uniform format.

 Normalization: Normalization scales your data into a regularized range so that you can compare it more accurately. For example, if you're comparing employee loss or gain within a number of companies (some with just a dozenemployees and some with 200+), you'll have to scale them within a specified range, like -1.0 to 1.0 or 0.0 to 1.0.

 Feature selection: Feature selection is the process of deciding which variables (features, characteristics, categories, etc.) are most important to your analysis. These features will be used to train ML models. It's important to remember,

that the more features you choose to use, the longer the training process and, sometimes, the less accurate your results, because some feature characteristics may overlap or be less present in the data.



 Discreditization: Discreditiization pools data into smaller intervals. It's somewhat similar to binning, but usually happens after data has been cleaned. For example, when calculating average daily exercise, rather than using the exact minutes and seconds, you could join together data to fall into 0-15 minutes, 15-30, etc.

 Concept hierarchy generation: Concept hierarchy generation can add a hierarchy within and between your features that wasn't present in the original data. If your analysis contains wolves and coyotes, for example, you could add the hierarchy for their genus: canis.

4. Data reduction

The more data you're working with, the harder it will be to analyze, even after cleaning and transforming it. Depending on your task at hand, you may actually have more data than you need. Especially when working with text analysis, much of regular human speech is superfluous or irrelevant to the needs of the researcher. Data reduction not only makes the analysis easier and more accurate, but cuts down on data storage.

It will also help identify the most important features to the process at hand.

 Attribute selection: Similar to discreditization, attribute selection can fit your data into smaller pools. It, essentially, combines tags or features, so that tagslike male/female and professor could be combined into male professor/female professor.

 Numerosity reduction: This will help with data storage and transmission. You can use a regression model, for example, to use only the data and variables that are relevant to your analysis.

 Dimensionality reduction: This, again, reduces the amount of data used to help facilitate analysis and downstream processes. Algorithms like K-nearest neighbors use pattern recognition to combine similar data and make it more manageable.

Data Preprocessing Examples

Take a look at the table below to see how preprocessing works. In this example, we have three variables: name, age, and company. In the first example we can tell that #2 and #3 have been assigned the incorrect companies.

Name	Age	Company
Karen Lynch	57	CVS Health
Elon Musk	49	Amazon
Jeff Bezos	57	Tesla
Tim Cook	60	Apple

We can use data cleaning to simply remove these rows, as we know the data was improperly entered or is otherwise corrupted.

Name	Age	Company

Karen Lynch	57	CVS Health
Tim Cook	60	Apple