**Data Preprocessing**

1. Dealing with missing data

It is crucial that we take care of those missing values before we proceed with further analyses.

1. Eliminating samples or features with missing values

df.dropna()

**Pro:** convenient

**Con:** end up removing too many samples, which will make a reliable analysis impossible;losing valuable information that our classifier needs to discriminate between classes.

1. Imputing missing values

**Mean imputation:** simply replace the missing value by the mean value of the entire feature column.

Code: Imputer

Github

1. dealing categorical data
2. mapping ordinal features

We need to convert the categorical string values into integers.

Code: map,use size\_mapping and inv\_size\_mapping

1. Encoding class labels

Many machine learning libraries require that class labels are encoded as integer values.

1. Performing one-hot encoding on nominal features

**Onehot encoding**: create a new dummy feature for each unique in the nominal feature column,Here, we would covert the color feature into three new features:blue,green and red.

**OneHotEncoder**

Reference :

Python machine learning.books