Outline

Outline

1. Introduction

1.1. Objectives

Objectives

- Understanding data to get the most out of it.
- Use descriptive statistical techniques to summarise data.
- Analyse the relationships present in the data numerically and graphically.
- Know the principles and know how to apply the techniques of data preprocessing.

1.2. Data

Understanding the data

- The first step to getting the best results in any Machine Learning project is to know and understand the data you are working with.
- A key starting point:
  - Visualising the raw data.

- Knowing the dimensions and type of attributes of the dataset.
- Analysing the balancing of classes.
- Discovering the relationships between its variables.
- Seeing the skewness of the distributions of each feature.

- For this, we are aided by descriptive statistics, a discipline that provides the techniques to collect, present and characterise the variables of a dataset through its basic parameters, tables or graphs.

Common data format

- The terms: instances, patterns, points, observations, rows... refer to the same thing, each of the data available for analysis.
- Characteristics, factors, dimensions, variables, attributes, property, field, columns... are the attributes that describe each instance of the dataset.

	Var 1	Var 2	...	Var k
Pattern 1				
Pattern 2				
...				
Pattern n				

2. First Analysis

2.1. Descriptive Statistics

Counting

- Knowing the number of patterns n and features k in our database is important.
- Depending on them, it will be more appropriate to apply one or another Machine Learning technique.

	Var 1	Var 2	...	Var k
Pattern 1				
Pattern 2				
...				
Pattern n				

Mean

- It is a measure of central tendency.
- It is calculated as the sum of the values divided by the number of them.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

where x_i is the i-th value of a given characteristic.

Standard deviation

- It is a measure of the amount of variation or dispersion of a set of values.
- A low standard deviation indicates that the values tend to be close to the mean.
- A high one indicates that the values are spread out over a wider range.

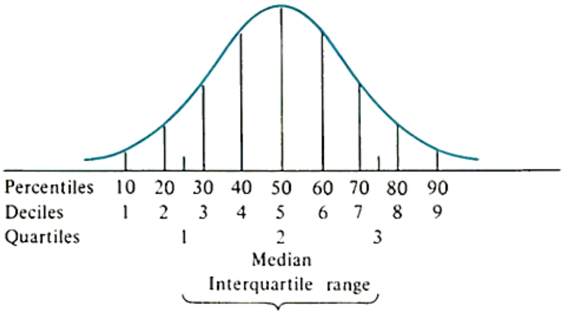
$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n}}.$$

Maximum and minimum

- Maximum value in a set of values.
- Minimum value in a set of values.
- The range is the difference between the maximum and minimum values.

Quantiles

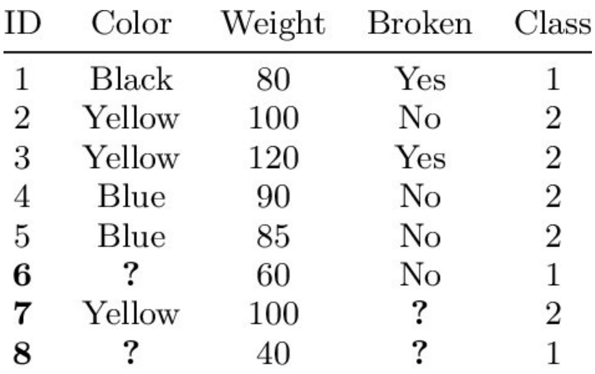
- Quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way.
- Common quantiles have special names:
  - Quartiles: four groups (Q_1, Q_2, Q_3).
  - Deciles: ten groups $(d_1, d_2, \ldots, d_9)$.
  - Percentiles: 100 groups $(p_1, p_2, \ldots, p_{99})$.
- Median $= Q_2 = d_5 = p_{50}$.


Percentiles 10 20 30 40 50 60 70 80 90
Deciles 1 2 3 4 5 6 7 8 9
Quartiles 1 2 3
Median
Interquartile range

2.2. Missing Values

Missing values

- In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation.
- Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

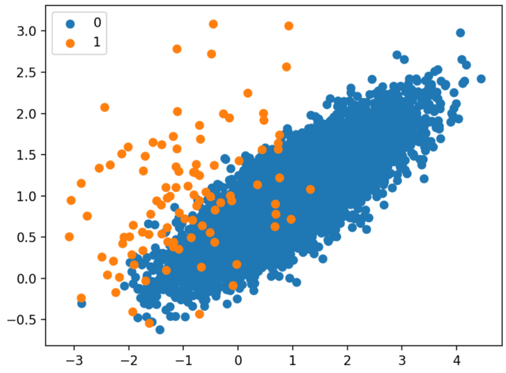


ID	Color	Weight	Broken	Class
1	Black	80	Yes	1
2	Yellow	100	No	2
3	Yellow	120	Yes	2
4	Blue	90	No	2
5	Blue	85	No	2
6	?	60	No	1
7	Yellow	100	?	2
8	?	40	?	1

2.3. Class Distribution

Class distribution

- For classification problems, it is vital to know how balanced the number of patterns per class is.
- Highly imbalanced problems, i.e. many more observations for one class than for another, are pretty common and need specific treatment in the data preparation phase of the project.



2.4. Correlation between Variables

Correlation between variables

- It is the relationship between two variables and how they may or may not change simultaneously.
- Pearson's correlation coefficient, which assumes a normal distribution, is the most commonly used method.

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

  - A value of -1 or 1 shows a completely negative or positive correlation between two variables.
  - A value of 0 shows no correlation between the two.
- Some ML methods do not work well if there are correlated variables in the data set.

2.5. Skewness

Skewness

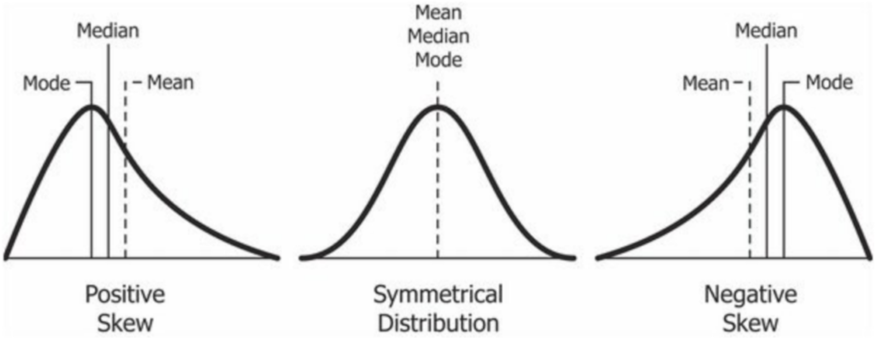
- Many ML techniques assume a normal distribution of data.
- Knowing whether the data distribution is skewed is essential to correct for skewness and improve the performance of the models.

$$skewness = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^3}{\sigma^3}.$$

  - A value clearly greater than zero is evidence of right-skewed data distribution.
  - A value equal to or close to zero shows a variable with a normal distribution.
  - Negative values will indicate a variable skewed to the left.

Skewness

- A value clearly greater than zero is evidence of right-skewed data distribution.
- A value equal to or close to zero shows a variable with a normal distribution.
- Negative values will indicate a variable skewed to the left.



3. Data Visualisation

Data visualisation

The best way to gain an initial understanding of the data is to use data visualisation techniques:

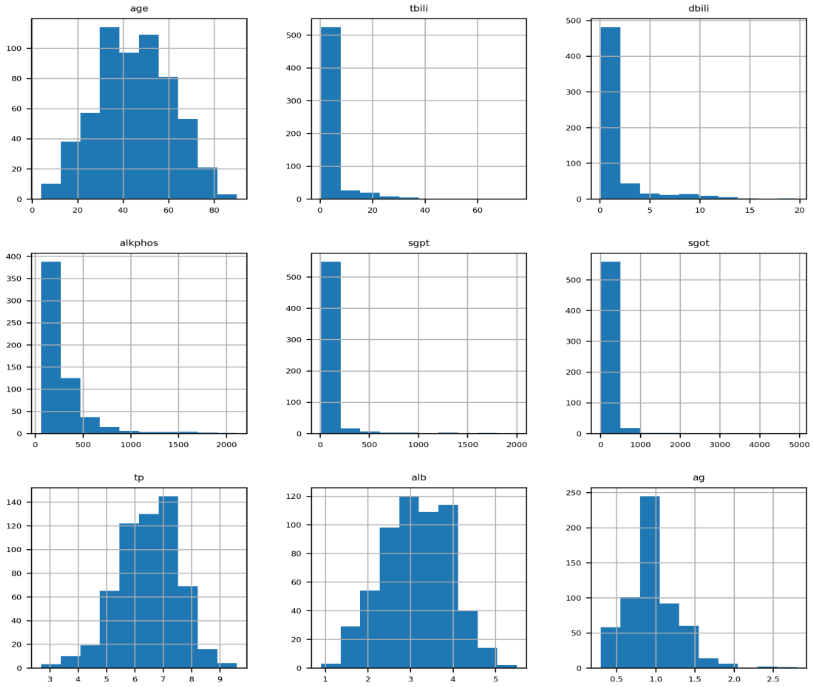
- Univariate graphs: these can be used to understand each attribute independently.
  - Histograms.
  - Density plots.
  - Boxplots.
- Multivariate plots: used to visualise the interactions between multiple variables in a data set.
  - Correlation matrix graph.
  - Scatter matrix graph.

3.1. Univariate Graphs

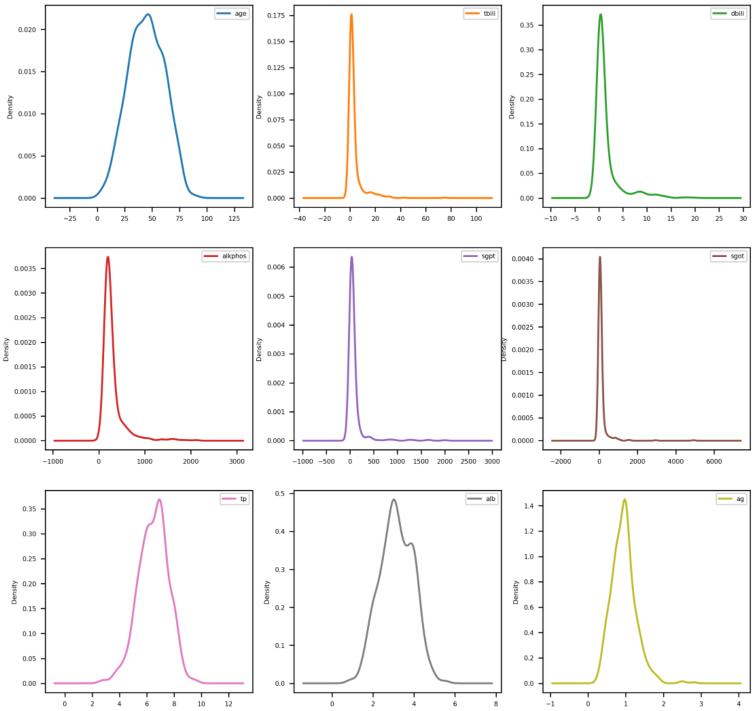
Histograms

- As discussed above, knowing the distribution of variables is important.
- A quick way is to use histograms.
- These are bar charts, which group the data into intervals and provide a count of the number of observations in each interval.
- They allow outliers to be detected.

Histograms



Density plots

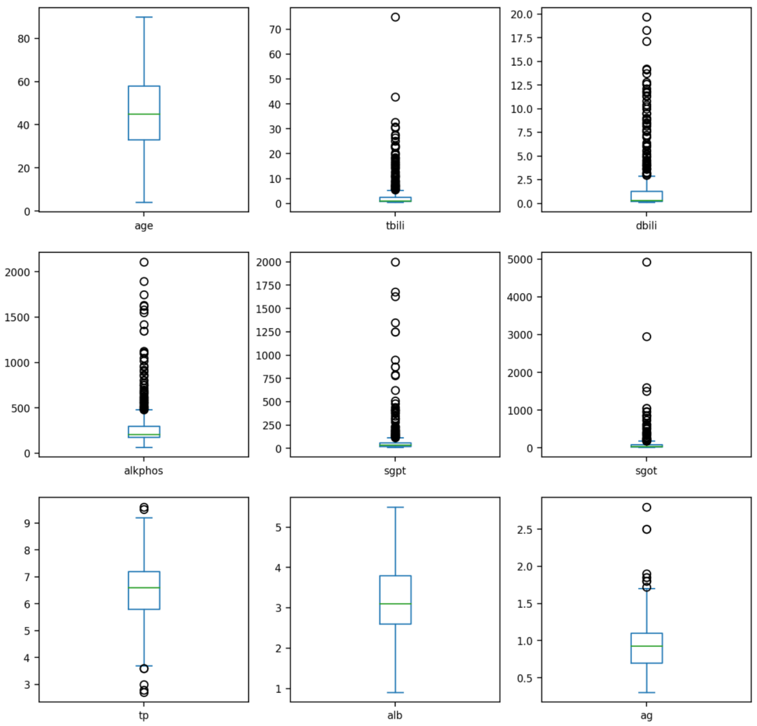


Boxplots

- Boxplots summarise the distribution of each attribute, drawing a line for a median line and a box around the Q1 and Q3 quartiles (25th and 75th percentiles).

- The legs give an idea of the dispersion of the data.
- Points outside the legs show outliers (greater than 1.5 in the interquartile range).
- The ends of the legs identify the most extreme values that are not considered outliers.

Boxplots

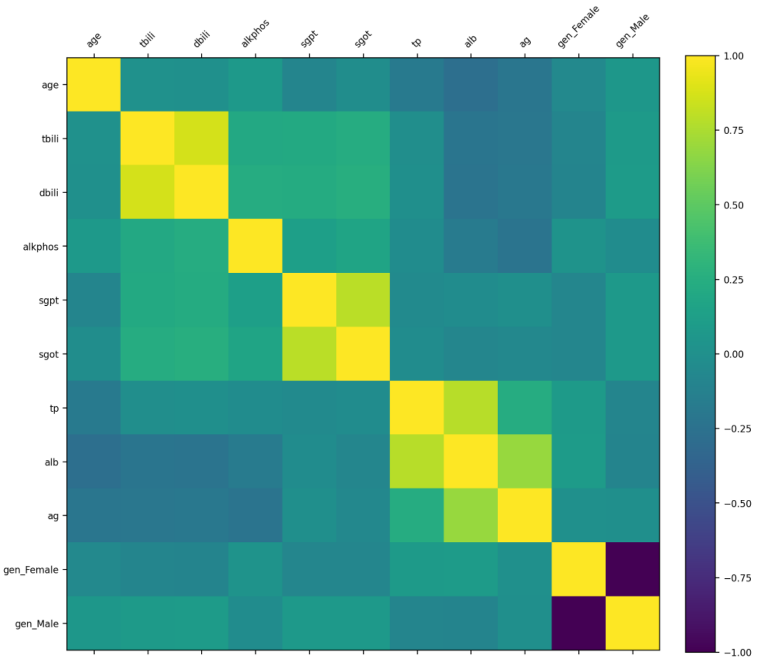


3.2. Multivariate Graphs

Correlation matrix graph: heat map

- As discussed above, correlation indicates how closely two variables are related.
- Some models perform poorly if the input variables are highly correlated.
- A visual way to see this is to use a heat map of the correlation matrix.

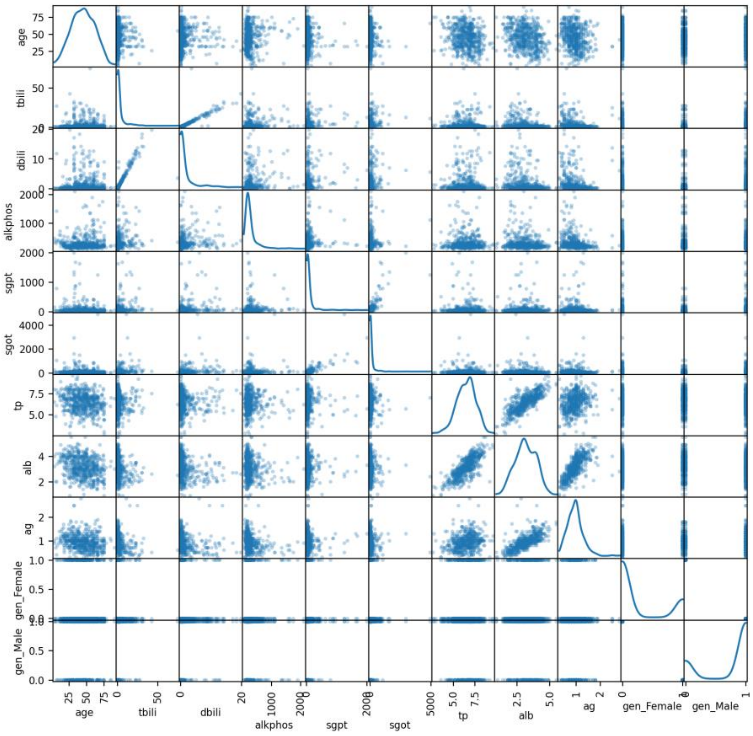
Correlation matrix graph: heat map



Scatter matrix graph

- A scatter plot shows the relationship between two variables as points in two dimensions, one axis for each attribute.
- The figure of the scatter plots of each pair of attributes together is called the scatter plot matrix.
- Scatter plots help detect structured relationships → attributes may be correlated.
- The main diagonal shows the histograms or density plots for each variable.

Scatter matrix graph



4. Preprocessing

Preprocessing

- Data preprocessing is a regular and necessary part of any Machine Learning process.
- Often, the results will depend more on the quality of the data in relation to the problem than on the Machine Learning part of the model.
- In the following slides, we will talk about noise in the data relevance of variables, ... always with respect to an objective.
- Thus, a variable that we consider noise may be useful information for a different problem.

Preprocessing

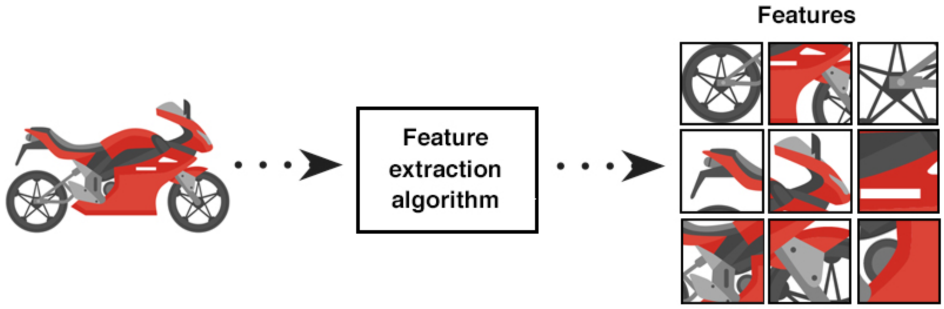
The most important tasks at this stage are:

- Data selection.
- Data cleaning.
- Data transformation.
- Dimensionality reduction.
- Imbalance treatment.

4.1. Data Selection

Feature extraction

- This process consists of determining which variables we need to characterise a problem.
- For example, when processing multimedia data, features are extracted that allow the construction of fixed-size vectors needed for models.



Feature selection

- Typically, in a high-dimensional dataset (high number of features), some features are likely to be irrelevant, insignificant and unimportant.
- The contribution of such features in the modelling can be minor, null and even counterproductive.
- This indicates that they introduce noise that harms the performance of the model.
- Moreover, the more features there are, the more resources are required to utilise them.

Feature selection

Feature selection is the process of selecting the most significant or relevant features from a given data set. The main advantages of feature selection are:

- It reduces the complexity of a model and facilitates its interpretation.
- It allows the training process of the Machine Learning model to be faster.
- It improves the performance of the models.
- It reduces overfitting.

FS: Filter methods

- They try to choose a subset of features without including any Machine Learning algorithms.
- The features are classified using statistics that determine the features' correlation with the target variable.
- Correlation is a highly context-bound term and varies from paper to paper depending on whether the data type is continuous or categorical.

FS: Filter methods

- Numerical input variables:
  - **Numerical output (regression)**:
    - Pearson's correlation coefficient. Assumes variables that follow a normal distribution. Linear correlation.
    - Spearman's ranking coefficient. Non-linear correlation.
  - **Categorical output (classification)**:
    - ANOVA correlation coefficient. Linear.
    - Kendall's ranking coefficient. Non-linear.
- Categorical input variables:
  - **Numerical output (regression)**:
    - ANOVA correlation coefficient. Linear.
    - Kendall's ranking coefficient. Non-linear.
  - **Categorical output (classification)**:
    - Chi-square test.
    - Mutual information.

FS: Wrapper methods

- They require a Machine Learning algorithm, using their performance as an evaluation criterion for the selected subset of attributes.

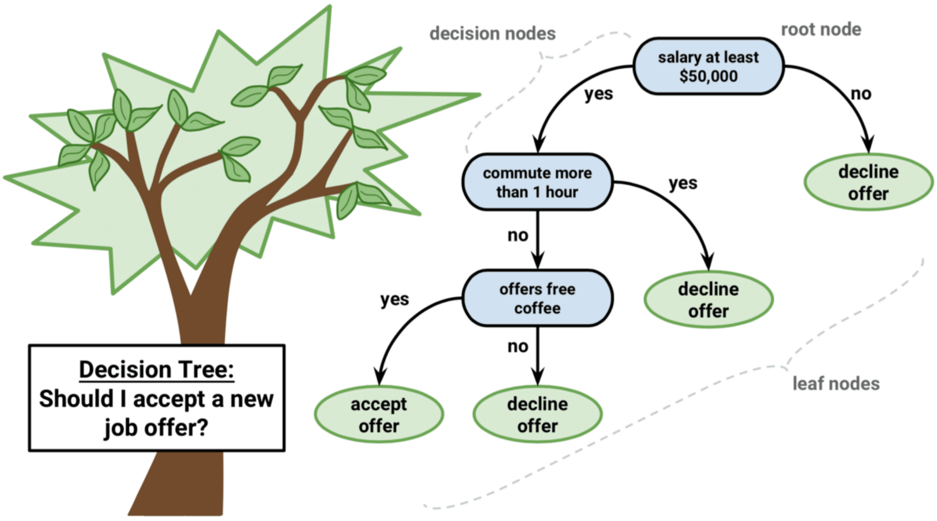
- Ideal coupling between the new subset and the Machine Learning method.

- Iteratively adds or removes variables based on the performance of successive trained models.

- A popular method is Recursive Feature Elimination, which recursively eliminates attributes by training a new model with the remaining ones.

FS: Embedded methods

- The model itself, and specifically its training algorithm, is the one that makes a selection of variables.



4.2. Data Cleaning

Data cleaning

- Eliminate inconsistencies: are points where a combination of feature values appear that violate the patterns generally observed. Example: the country “California” and the time zone “Central European Time”.
- Noise smoothing: irrelevant information that makes the predictor less accurate. Example: the weather during the summer months to determine the risk of fog in airports.
- Anomalous data (outliers): data that is outside of the norm of deviates from what is expected. An age of 200 years.
- Imputation of missing values.

Data cleaning: missing values imputation

- Many Machine Learning models do not support the use of missing data.
- Unfortunately, data from the real world is not perfect and often this type of data will be present in our dataset.
- There are two strategies for dealing with missing data:
  - Remove instances or features with missing data.
  - Impute missing values.

Remove instances or features

Different alternatives can be used:

- Delete rows with at least one missing value.
- Delete rows where all values are missing.

- Delete columns with at least one missing value.
- Delete columns with all missing values.
- Delete rows/columns when at least n values are missing.

Impute missing values

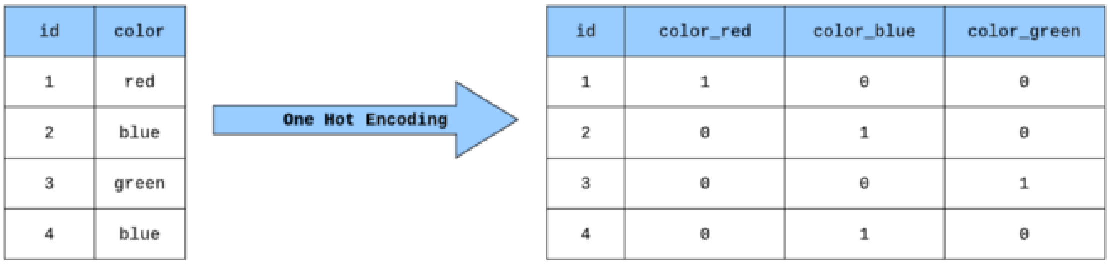
Two strategies:

- Univariate imputation: uses the values of that dimension to retrieve missing data. For example, replace by a constant value, by the mean, by an interpolated value, etc.
- Multivariate imputation: models each attribute with missing values as a function of the other characteristics. For example, using k-nearest neighbors (KNN).

4.3. Data Transformation

Binarisation of categorical variables

In general, most Machine Learning techniques do not support categorical data.



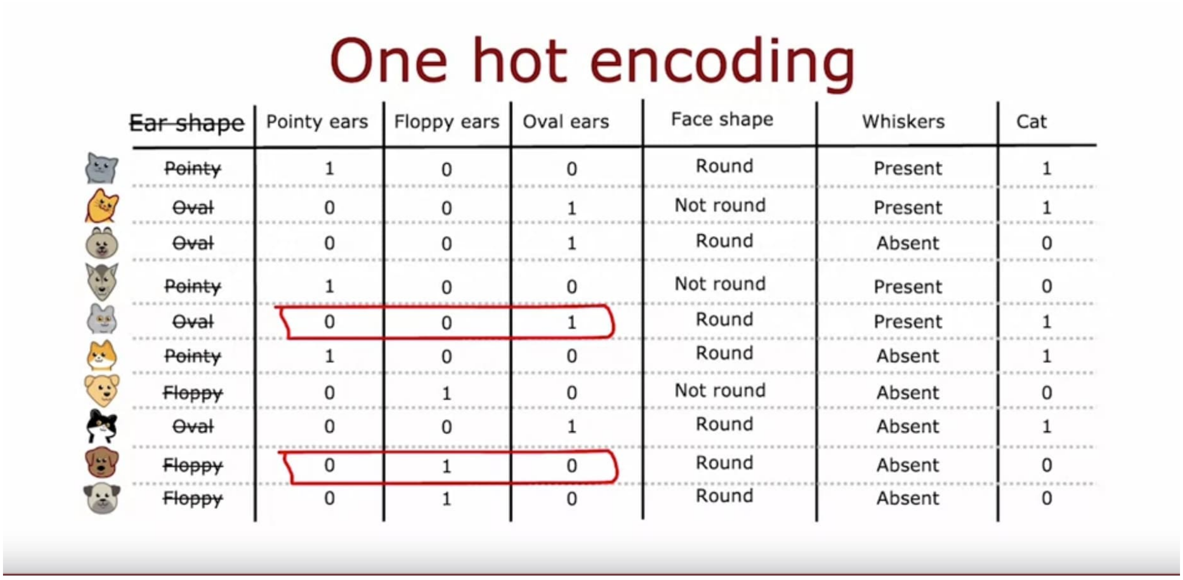
id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Binarisation of categorical variables

In general, most Machine Learning techniques do not support categorical data.



One hot encoding

~~Ear shape~~	Pointy ears	Floppy ears	Oval ears	Face shape	Whiskers	Cat
~~Pointy~~	1	0	0	Round	Present	1
~~Oval~~	0	0	1	Not round	Present	1
~~Oval~~	0	0	1	Round	Absent	0
~~Pointy~~	1	0	0	Not round	Present	0
~~Oval~~	0	0	1	Round	Present	1
~~Pointy~~	1	0	0	Round	Absent	1
~~Floppy~~	0	1	0	Not round	Absent	0
~~Oval~~	0	0	1	Round	Absent	1
~~Floppy~~	0	1	0	Round	Absent	0
~~Floppy~~	0	1	0	Round	Absent	0

Data scaling

- Many techniques work better if all attributes are in the same range of values.
- For example: gradient descent algorithms, regression, neural networks and methods using distance measures such as KNN and Clustering.

$$x'_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)},$$

where x_{ij} is the value of pattern i in the j-th characteristic.

Data normalisation - standardisation

- It is a helpful technique for transforming data distribution to a normal distribution of mean 0 and standard deviation 1.
- This technique is well suited for methods that assume a Gaussian distribution on the input variables, e.g. linear regression, logistic regression and linear discriminant analysis.

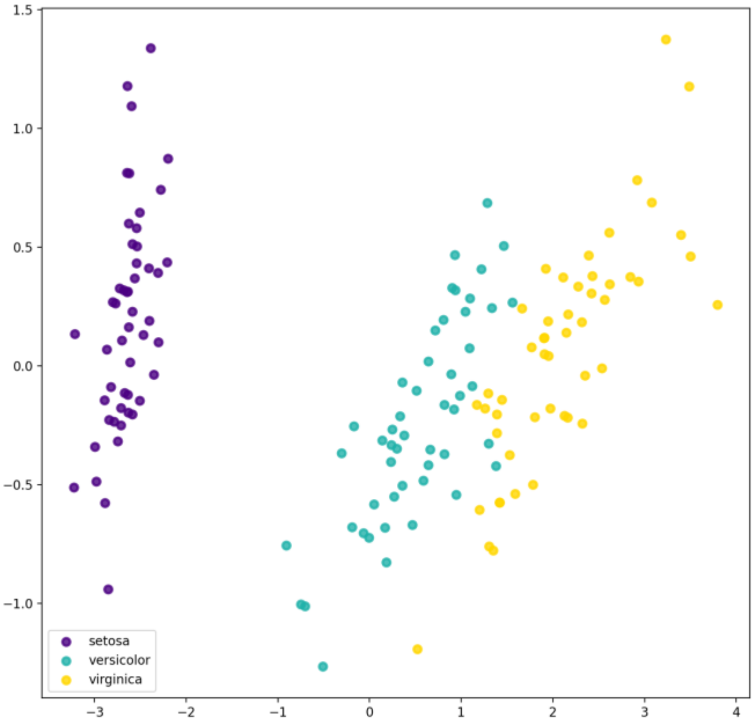
$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}.$$

4.4. Dimensionality Reduction

Principal component analysis

- This topic will be discussed in more detail in Unit 4.
- Different from feature selection.
- Principal Component Analysis (PCA) is used to reduce dimensionality.
- Data are described in terms of uncorrelated and orthogonal variables (components).
- They are further ordered by the amount of original variance they can describe.
- In summary, PCA transforms a set of possibly correlated variables into a set of non-linearly correlated variables called “principal components”.

Principal component analysis



4.5. Imbalanced Treatment

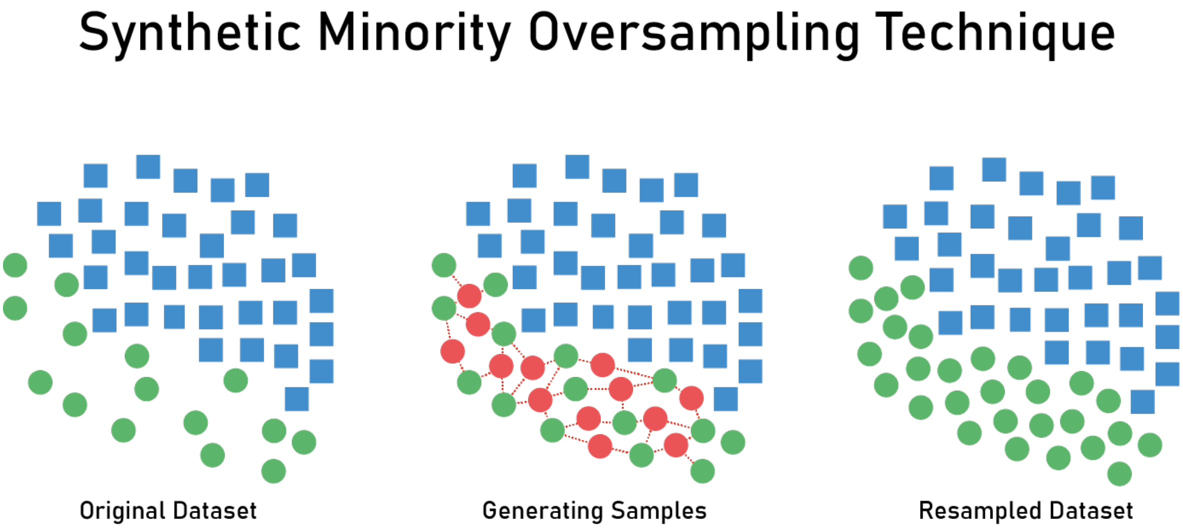
Imbalanced treatment

It may happen that, in supervised classification problems, there is a different number of patterns for each class. This leads to imbalance, i.e. the number of examples in one class is very large relative to another. In this sense, Machine Learning algorithms tend to "learn" to classify the majority classes well. To avoid this problem, we have two alternatives:

- Oversampling: synthetic patterns are generated in those classes with fewer patterns.

- Undersampling: the number of majority class patterns is reduced.

SMOTE



https://www.youtube.com/watch?v=FheTDyCwRdE

References

- Benítez-Iglésias, R., Escudero-Bakx, G., Kanaan-Izquierdo, S., Masip-Rodó, D. (2014). Inteligencia artificial avanzada. Editorial UOC.
- Bishop, CM. (2006). Pattern recognition and machine learning. Springer.
- Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow (Second Edition). O'Reilly.
- James, G.; Witten, D.; R Hastie, T.; Tibshirani, R.; Taylor, J. (2023). An Introduction to Statistical Learning with Applications in Python. Springer.
- Mohri, M., Talwalkar, A., Rostamizadeh, A. (2019). Foundations of machine learning. Cambridge: The MIT Press.

Questions?