

Supervised Learning

Course Summary - Master's Degree

Carlos Alberto Botina Carpio
Universidad Internacional de la Rioja
carlos.botina621@comunidadunir.net

November 24, 2025

Abstract

This document contains a summary of the Supervised Learning course syllabus for the Master's Degree. It includes a summary of the main topics covered during the sessions, as well as additional explanations and extensions of the concepts and techniques referenced in class. The purpose of this document is to serve as study material and reference for the course contents. Notice that this document is fully customized to the author's needs, placing greater emphasis on topics that the author struggles with or has not yet mastered, while covering more briefly those topics that are already well understood by the author.

Contents

1 Descriptive Data Analysis	3
1.1 Types of Variables	3
1.1.1 Qualitative Variable	3
1.1.2 Quantitative Variable	3
1.2 Characterization of Variable Distributions	4
1.2.1 Histogram	4
2 Exploratory Data Analysis	4
2.1 Measures of Dispersion	4
2.1.1 Range	5
2.1.2 Quartiles and Interquartile Range	5
2.1.3 Variance	5
2.1.4 Standard Deviation	6
2.2 Outlier Detection	7
2.2.1 Box Plot	7
2.3 Scatter Plots	8
2.3.1 Two-dimensional Scatter Plots	8
2.3.2 Scatter Plot Matrix	10
2.4 Correlation between Variables	10
2.4.1 Linear Correlation	10
2.4.2 Pearson's Correlation Coefficient	11
2.5 Covariance Matrix	12
3 Missing Data and Normalization	12
3.1 Finding Redundant Attributes	13
3.1.1 Correlation and Covariance	13
3.1.2 Chi-square Test for Nominal Data	13
3.2 Detecting Duplicate Records	15
3.3 Mechanisms for Replacing Missing Data	15
3.3.1 Data Loss Types	16
3.3.2 Handling Mechanisms	16
3.3.3 Other Imputation Methods	17
3.4 Min-Max Normalization	17
3.5 Z-score Normalization	19
3.6 From Nominal to Binary	20

Lecture 002

1 Descriptive Data Analysis

1.1 Types of Variables

One of the first stages of descriptive analysis is to recognize the different types of data, as these observations are a representation of the world around us through what are called variables. A variable is a characteristic or attribute of an object that can be observed and measured, assuming different values. Variables can be represented in different forms: numerical (discrete and continuous) or categorical (nominal and ordinal). Figure 1 summarizes the classification of variable types.

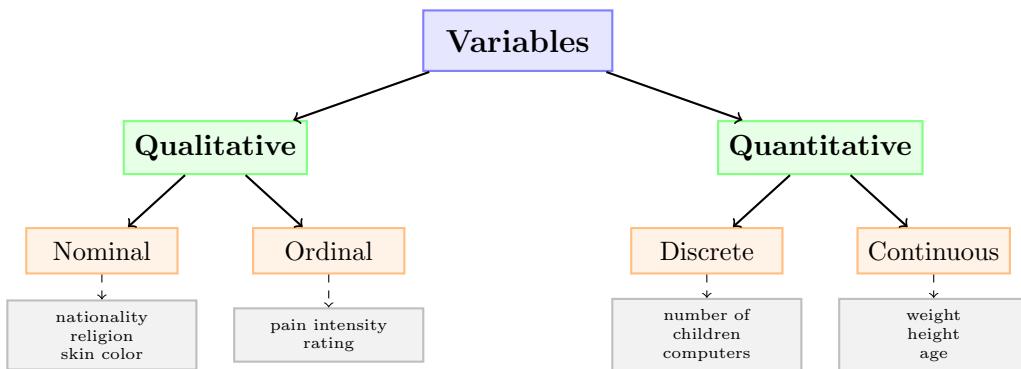


Figure 1: Classification of variable types

1.1.1 Qualitative Variable

A qualitative variable expresses a characteristic in a non-numerical form, representing a quality. Examples: color, gender, blood type, marital status. It can be of two types:

- **Nominal:** Values are codes or names that cannot be ordered. Examples: nationality, religion, skin color.
- **Ordinal:** Values imply an ordering from greater to lesser or from better to worse. Examples: pain intensity (absent, mild, moderate, severe, very severe), rating (excellent, good, fair, poor).

1.1.2 Quantitative Variable

A quantitative variable is expressed by means of a number. It can be of two types:

- **Discrete:** Only admits a finite set of numerical values, usually integers. Examples: number of children per household, number of computers in a classroom, number of electrons in an atom.
- **Continuous:** Can take infinitely many possible values within a range or interval on the real number line. These variables represent measurements whose values cannot be enumerated. Examples: weight (kg), height, age, the time it takes an athlete to run 100 meters.

1.2 Characterization of Variable Distributions

1.2.1 Histogram

A histogram is a graphical representation of the distribution of numerical data. It displays the frequency of data points within specified intervals (bins). The x-axis represents the intervals, and the y-axis represents the frequency or relative frequency. The bars are adjacent to each other, unlike bar charts for categorical data.

Histograms are used for continuous or discrete quantitative variables. The bin width affects the appearance and interpretation. The area of each bar is proportional to the frequency, with no gaps between bars (unless a bin has zero frequency). Histograms reveal the shape, center, and spread of the data distribution.

To read a histogram, examine the overall shape (symmetric, skewed, or multimodal), identify the center, assess the spread, and look for outliers or unusual patterns.

Figure 2 shows an example of a histogram. Its shape depends on the number of intervals, and the symmetry can vary; the more similar the values on both sides of the center, the more symmetric the distribution (Rodríguez Ojeda, 2007):

- If the height or y-axis of the bars are similar to each other, we would have a uniform distribution.
- If the heights are greater in the central zone, a “bell” shape is formed, which can be symmetric or asymmetric, toward the positive side (to the right) or negative (to the left).
- If there are bars far from the group, they are considered atypical data, which are probably due to measurement errors and can be discarded, as they do not belong to the group that is desired to be characterized (Rodríguez Ojeda, 2007).

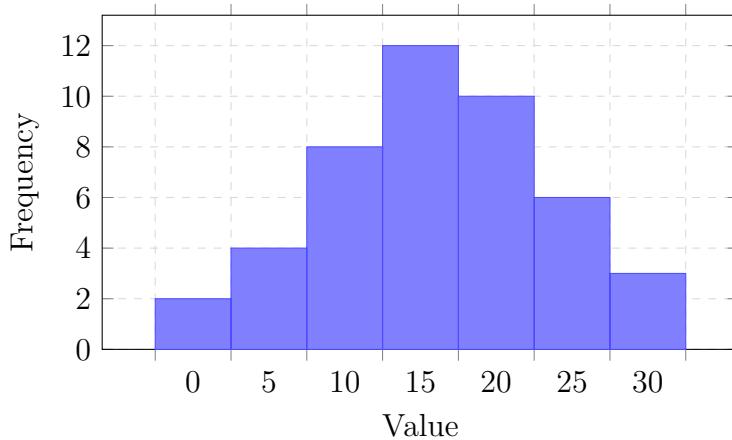


Figure 2: Example of a histogram showing a bell-shaped distribution

2 Exploratory Data Analysis

2.1 Measures of Dispersion

Data tends to cluster toward the center of the frequency distribution, but extreme values can be far from this central tendency. Measures of dispersion quantify this distance from the average.

2.1.1 Range

The difference between the upper and lower limits of a dataset. Limited utility as it is easily affected by extreme values. The interquartile range (IQR) is commonly used to suppress the influence of extremes (Witte y Witte, 2017).

2.1.2 Quartiles and Interquartile Range

Quartiles divide the distribution into four equal parts (25% each). Q1 (25th percentile), Q2 (median, 50th percentile), and Q3 (75th percentile). The IQR equals Q3 - Q1, representing the range for the middle 50% of data, effectively removing the upper and lower 25%. The IQR is resistant to extreme values and tends to be less than half the range (Witte y Witte, 2017).

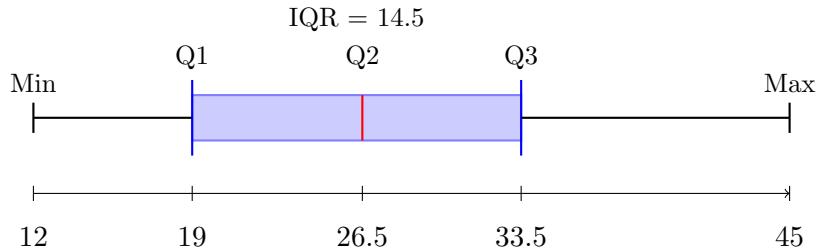


Figure 3: Box plot showing quartiles and interquartile range. The box represents the IQR (middle 50% of data), Q1 and Q3 are the box edges, and Q2 (median) is the red line.

2.1.3 Variance

The average of the squared deviations from the arithmetic mean. Deviations are squared because their sum equals zero (negative cancel positive). Notation: s^2 (sample) and σ^2 (population). Useful for comparing variability between distributions, but has squared units making interpretation complex (Spiegel y Stephen, 2009).

The formulas for variance are:

$$\text{Population variance: } \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (1)$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2)$$

where N is the population size, n is the sample size, μ is the population mean, and \bar{x} is the sample mean.

Example: Calculating Variance Consider the following dataset representing test scores: 75, 82, 68, 90, 85.

Note: This example calculates sample variance (s^2), treating the data as a sample from a larger population. If this were the entire population, we would use population variance (σ^2) and divide by N instead of $(n - 1)$.

Step 1: Calculate the mean

$$\bar{x} = \frac{75 + 82 + 68 + 90 + 85}{5} = \frac{400}{5} = 80 \quad (3)$$

Step 2: Calculate deviations from the mean

$$x_1 - \bar{x} = 75 - 80 = -5 \quad (4)$$

$$x_2 - \bar{x} = 82 - 80 = 2 \quad (5)$$

$$x_3 - \bar{x} = 68 - 80 = -12 \quad (6)$$

$$x_4 - \bar{x} = 90 - 80 = 10 \quad (7)$$

$$x_5 - \bar{x} = 85 - 80 = 5 \quad (8)$$

Step 3: Square each deviation

$$(x_1 - \bar{x})^2 = (-5)^2 = 25 \quad (9)$$

$$(x_2 - \bar{x})^2 = (2)^2 = 4 \quad (10)$$

$$(x_3 - \bar{x})^2 = (-12)^2 = 144 \quad (11)$$

$$(x_4 - \bar{x})^2 = (10)^2 = 100 \quad (12)$$

$$(x_5 - \bar{x})^2 = (5)^2 = 25 \quad (13)$$

Step 4: Sum the squared deviations

$$\sum_{i=1}^5 (x_i - \bar{x})^2 = 25 + 4 + 144 + 100 + 25 = 298 \quad (14)$$

Step 5: Calculate sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{5-1} \times 298 = \frac{298}{4} = 74.5 \quad (15)$$

Therefore, the sample variance is $s^2 = 74.5$ (units squared).

2.1.4 Standard Deviation

To resolve the problem of squared units of measurement, the square root of the variance is calculated, always taking the positive value. This produces a new measure, known as standard deviation, which describes variability in the original units of measurement. Like the previous measures, standard deviation must be distinguished: symbolized by s for the sample and σ for the population.

The formulas for standard deviation are:

$$\text{Population standard deviation: } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (16)$$

$$\text{Sample standard deviation: } s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (17)$$

Example: Calculating Standard Deviation Using the previous variance example (test scores: 75, 82, 68, 90, 85), where we calculated $s^2 = 74.5$:

$$s = \sqrt{s^2} = \sqrt{74.5} \approx 8.63 \quad (18)$$

Therefore, the sample standard deviation is $s \approx 8.63$ test score points.

Interpretation The standard deviation measures the typical distance of data points from the mean. In this example:

- The mean test score is 80 points
- The standard deviation is approximately 8.63 points
- This means that, on average, test scores deviate from the mean by about 8.63 points
- Most scores (approximately 68% in a normal distribution) fall within one standard deviation of the mean, i.e., between $80 - 8.63 = 71.37$ and $80 + 8.63 = 88.63$ points

Curious Fact

A larger standard deviation indicates greater variability (scores are more spread out), while a smaller standard deviation indicates less variability (scores are clustered closer to the mean).

2.2 Outlier Detection

An outlier is a numerical observation that is distant from the rest of the data. Outliers in a dataset can cause problems in statistical analyses by generating misleading conclusions. Usually, an outlier can indicate a measurement error or a long-tailed distribution in the population, although they can occur by chance.

2.2.1 Box Plot

A box plot is a graphical representation of numerical data through its quartiles. The boxes in this standardized method contain the following elements: the median (Q2), the interquartile range, and the quartiles (Q1 and Q3). Box plots also have lines extending from the boxes (whiskers) that indicate variability outside the upper and lower quartiles, hence the term box-and-whisker plot. Depending on the type of representation, the whiskers can extend to the minimum value at one end and the maximum value at the other, representing the full range.

For outlier identification, John Tukey proposed whiskers with a length of 1.5 times the IQR (Tukey, 1977). That is, from the hinges (Q1 and Q3), 1.5 times the IQR is added or subtracted respectively. Then, values beyond these whiskers are considered outliers and are represented as individual points.

To construct box plots, the median, first and third quartiles, and the interquartile range of each attribute are used. Then, for the whiskers, 1.5 times the IQR is added above and subtracted below the box. Finally, the last value within those limits is taken, and that is where the whiskers are graphed.

Example: Box Plot with Outliers Consider the dataset: 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 60, 70.

Calculating the quartiles:

- $Q1 = 16$, $Q2$ (median) = 23, $Q3 = 30$
- $IQR = Q3 - Q1 = 30 - 16 = 14$
- Lower whisker limit: $Q1 - 1.5 \times IQR = 16 - 21 = -5$ (minimum within limit: 10)
- Upper whisker limit: $Q3 + 1.5 \times IQR = 30 + 21 = 51$
- Outliers: 60 and 70 (both exceed the upper whisker limit)

Note: The whiskers extend to 10 and 32, not to the calculated limits (-5 and 51). This is because whiskers are drawn to the last data value within the limits, not to the limits themselves. Since -5 is below any actual data point, the lower whisker goes to the minimum value (10). Similarly, the upper whisker stops at 32, which is the last value within the upper limit (51). Values beyond these limits (60 and 70) are shown as outlier points.

Figure 4 shows the box plot for this dataset, with outliers represented as individual points.



Figure 4: Box plot showing quartiles, IQR, whiskers (extending $1.5 \times IQR$ from $Q1$ and $Q3$), and outliers (60 and 70) represented as individual red points.

2.3 Scatter Plots

The most useful graphical representations for describing the behavior of a set of variables are scatter plots, also known as dot plots or scatter diagrams. These graphs show the relationship between numerical variables, and they allow discovering and confirming anticipated relationships between two associated sets of data. Numerical variables are represented by coordinates in n-dimensional Euclidean space.

2.3.1 Two-dimensional Scatter Plots

For the case of two variables, the datasets must have the same length: one for the X-axis (horizontal) and one for the Y-axis (vertical). In Python, a scatter plot of two numerical variables X and Y can be obtained using the `plot(X, Y)` command from the matplotlib library.

The representation of each variable on each axis allows identifying the relationship of increase or decrease between the dependent and independent variables. Scatter plots are suitable when you have paired numerical data and want to see if one variable affects another. However, it is necessary to understand that correlation is not causation, and another unnoticed variable may influence the results. In the case of a database, numerical variables correspond to dataset attributes and are used to define aspects of an instance, such as size, shape, and color.

Correlation vs. Causation **Example:** Consider a scatter plot showing ice cream sales (X-axis) and drowning deaths (Y-axis) over time. The plot would likely show a positive correlation: as ice cream sales increase, drowning deaths also increase. However, this does not mean that ice cream sales *cause* drowning deaths.

Why correlation is not causation: Both variables are influenced by a third, hidden variable: the season (summer). In summer:

- More people buy ice cream (hot weather)
- More people go swimming (hot weather)
- More swimming leads to more drowning incidents

The correlation exists because both variables respond to the same underlying factor (summer weather), not because one directly causes the other. This illustrates that:

- **Correlation** means two variables tend to change together
- **Causation** means one variable directly causes changes in another
- A correlation can exist without causation when a third variable (confounding variable) influences both

To establish causation, controlled experiments or additional evidence are needed, not just observational correlation.

Example: Scatter Plot Figure 5 shows an example of a two-dimensional scatter plot displaying the relationship between two variables. Each point represents a pair of values (x, y), allowing visualization of patterns, trends, and potential correlations in the data.

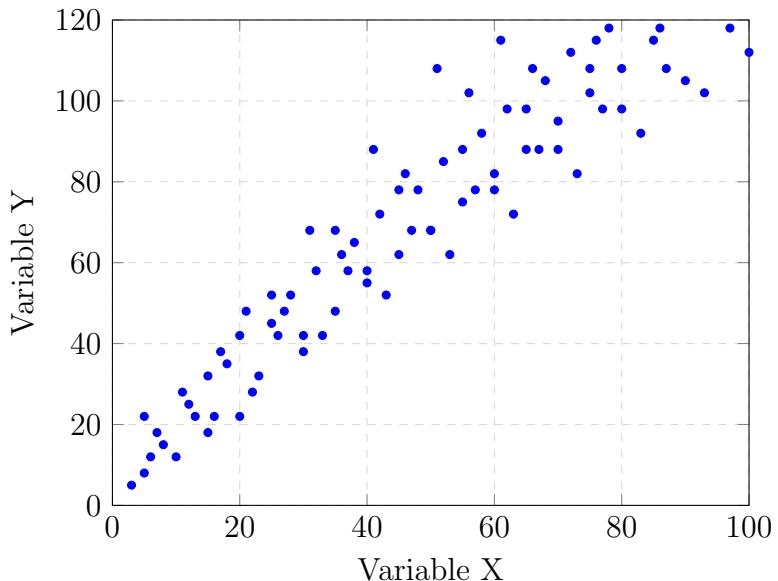


Figure 5: Example of a two-dimensional scatter plot showing a positive correlation between Variable X and Variable Y.

2.3.2 Scatter Plot Matrix

The scatter plot matrix (SPLOM—scatterplot matrix) presents multiple adjacent scatter plots for all variable comparisons in a single display. To some extent, it addresses one of the geometric problems for datasets with more than three attributes by presenting two or more pairs of variables in the same graph, which is very useful for identifying dependencies between data.

One of the disadvantages is that SPLOM requires a large amount of screen space, and the formation of multivariate associations remains a challenge. Additionally, it is necessary to incorporate additional statistical measures to organize SPLOM and guide the observer through an exploratory analysis of high-dimensional datasets.

2.4 Correlation between Variables

2.4.1 Linear Correlation

Correlation is known as a statistical measure that quantifies the degree of joint variation that exists between two variables and, specifically, evaluates the increasing or decreasing trend of the data. The types of correlation can be observed in Table 1.

Type of Correlation	Description
Positive correlation	The values of the variables increase together, given that an increase in Y (dependent variable) depends on an increase in X (independent variable).
Negative correlation	One value decreases as the other increases; therefore, an increase in the values of variable X will cause a decrease in the values of variable Y.
Null or no correlation	There is no relationship between the variables, therefore, the variables are independent. The graph does not follow any trend, causing the points to be totally dispersed.
Linear	There is a relationship between the variables, which is linear.
Non-linear	Although there is a relationship between the variables, it is not linear. It can be exponential, U-shaped, etc.

Table 1: Types of correlation

Given the subjectivity in interpreting scatter plots, it is necessary to use a coefficient that allows measuring the degree of dependence between variables. The linear correlation coefficient represents the behavior of a dependent variable Y with respect to an independent variable X.

The strength of the correlation is determined by the proximity of the points to each other in the graph. If the points are closer together, the strength will be greater (strong correlation). If they are not so concentrated, the strength will be weak. If a pattern of very dispersed points is observed, the correlation strength will be null (no correlation).

2.4.2 Pearson's Correlation Coefficient

Pearson's correlation coefficient is a statistic whose values range between -1 and 1 ($-1 \leq r \leq 1$). A correlation close to 1 indicates a positive association (both variables increase together). A correlation close to -1 indicates a negative association (one increases as the other decreases). A value near 0 indicates no linear relationship, although a non-linear relationship may exist. This coefficient can only be used for quantitative and continuous variables.

The correlation coefficient is calculated as:

$$r = \frac{\text{Cov}(X, Y)}{S_x \cdot S_y} \quad (19)$$

where:

- r is Pearson's correlation coefficient
- $\text{Cov}(X, Y)$ is the covariance between X and Y
- S_x is the standard deviation of variable X
- S_y is the standard deviation of variable Y

This coefficient allows determining the existence and intensity of linear associations between variables. The sign is interpreted the same as covariance: positive (direct correlation), negative (inverse correlation), or null (no correlation). It measures how close the points are to a straight line, reflecting the data trend.

Curious Fact: Covariance

Covariance measures how two variables vary together. It is calculated as the average of the products of deviations from the mean: $\text{Cov}(X, Y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$. A positive covariance means that when X is above its mean, Y tends to be above its mean too (they move together). A negative covariance means when X is above its mean, Y tends to be below its mean (they move in opposite directions). However, covariance is difficult to interpret because its magnitude depends on the units of measurement. This is why Pearson's correlation coefficient normalizes covariance by dividing by the product of standard deviations, making it unitless and easier to interpret.

Example: Calculating Pearson's Correlation Coefficient Consider the following paired dataset:

X	Y
2	4
4	6
6	8
8	10
10	12

Step 1: Calculate the means

$$\bar{x} = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6 \quad (20)$$

$$\bar{y} = \frac{4 + 6 + 8 + 10 + 12}{5} = \frac{40}{5} = 8 \quad (21)$$

Step 2: Calculate deviations and their products

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = (2 - 6)(4 - 8) + (4 - 6)(6 - 8) + (6 - 6)(8 - 8) + (8 - 6)(10 - 8) \dots \quad (22)$$

$$= (-4)(-4) + (-2)(-2) + (0)(0) + (2)(2) + (4)(4) \quad (23)$$

$$= 16 + 4 + 0 + 4 + 16 = 40 \quad (24)$$

Step 3: Calculate standard deviations

$$S_x = \sqrt{\frac{1}{n-1} \sum(x_i - \bar{x})^2} = \sqrt{\frac{(-4)^2 + (-2)^2 + 0^2 + 2^2 + 4^2}{4}} = \sqrt{\frac{40}{4}} = \sqrt{10} \approx 3.16 \quad (25)$$

$$S_y = \sqrt{\frac{1}{n-1} \sum(y_i - \bar{y})^2} = \sqrt{\frac{(-4)^2 + (-2)^2 + 0^2 + 2^2 + 4^2}{4}} = \sqrt{\frac{40}{4}} = \sqrt{10} \approx 3.16 \quad (26)$$

Step 4: Calculate covariance

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{40}{4} = 10 \quad (27)$$

Step 5: Calculate Pearson's correlation coefficient

$$r = \frac{\text{Cov}(X, Y)}{S_x \cdot S_y} = \frac{10}{3.16 \times 3.16} = \frac{10}{10} = 1.0 \quad (28)$$

Therefore, $r = 1.0$, indicating a perfect positive linear correlation. This makes sense as the data points lie exactly on a straight line with a positive slope.

2.5 Covariance Matrix

Statistics allow understanding relationships between multiple variables simultaneously. If covariance is positive, the linear association is positive. If covariance is negative, the linear association is negative (small X values correspond to large Y values, and vice versa).

The covariance matrix is a square matrix containing variances on the main diagonal and covariances in the off-diagonal elements. The matrix is symmetric: the covariance between X and Y equals the covariance between Y and X, so each pair appears twice.

Lecture 003

3 Missing Data and Normalization

Feature selection is one of the main stages in creating machine learning models, as in most cases, data contains more information than necessary. Feature selection not only

improves model quality but also optimizes computational resource usage, such as efficient storage and memory utilization during the training process.

In addition to feature selection, **data transformations** are performed. Data transformation is necessary to adapt data to classification problems, since data are typically not defined under the same numerical scales and in certain cases follow different distributions.

3.1 Finding Redundant Attributes

Attributes in a dataset must not be derived from other attributes; such attributes are called **redundant data**. Techniques for identifying redundant attributes include correlation, covariance, and the χ^2 test.

3.1.1 Correlation and Covariance

Pearson's correlation coefficient measures linear association between two quantitative variables. For attributes X_1 and X_2 : a high correlation coefficient indicates strong correlation and potential redundancy (one can be removed); a coefficient of 0 indicates independence (no redundancy); a negative coefficient indicates inverse relationship.

An example of redundancy:

Name	is_male	is_female
angie	0	1
juan	1	0
diego	1	0
pedro	1	0
ana	0	1
jose	1	0
carolina	0	1

Table 2: Example dataset showing redundant attributes

In Table 2, `is_male` and `is_female` are perfectly correlated (100%): if a person is not a man, they are a woman. One attribute determines the other, making one redundant and removable without information loss.

3.1.2 Chi-square Test for Nominal Data

When the feature type and target variable are categorical (i.e., a classification problem), the Chi-square test, also known as Pearson's chi-square, can be used. This test is performed on nominal data. For two attributes X_1 and X_2 in a dataset, a contingency table is created to represent data tuples. The chi-square statistic is calculated as:

$$\chi^2 = \sum \left[\frac{(Y_o - Y_e)^2}{Y_e} \right] \quad (29)$$

where Y_o is the actual count of observed values and Y_e is the expected values of joint events in the contingency table. The χ^2 test checks the hypothesis that X_1 and X_2 are independent. If this hypothesis can be rejected, we can say that X_1 and X_2 are statistically correlated and one of them can be discarded.

Step-by-Step Example Consider testing whether **Gender** and **Interest** (Sports vs Reading) are independent attributes. We have collected data from 100 people:

	Sports	Reading	Total
Male	30	20	50
Female	10	40	50
Total	40	60	100

Table 3: Observed frequencies (contingency table)

Step 1: Define the hypothesis

- H_0 : Gender and Interest are independent (not associated)
- H_1 : Gender and Interest are not independent (are associated)

Step 2: Set significance level

- $\alpha = 0.05$ (common choice: 0.01 to 0.10)

Step 3: Create contingency table

- Table 3 shows the observed frequencies (Y_o)

Step 4: Calculate expected frequencies For each cell: $E = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$

$$E_{\text{Male, Sports}} = \frac{50 \times 40}{100} = 20$$

$$E_{\text{Male, Reading}} = \frac{50 \times 60}{100} = 30$$

$$E_{\text{Female, Sports}} = \frac{50 \times 40}{100} = 20$$

$$E_{\text{Female, Reading}} = \frac{50 \times 60}{100} = 30$$

	Sports	Reading	Total
Male	20	30	50
Female	20	30	50
Total	40	60	100

Table 4: Expected frequencies

Step 5: Calculate Chi-square Using equation 29:

$$\begin{aligned} \chi^2 &= \frac{(30 - 20)^2}{20} + \frac{(20 - 30)^2}{30} + \frac{(10 - 20)^2}{20} + \frac{(40 - 30)^2}{30} \\ &= \frac{100}{20} + \frac{100}{30} + \frac{100}{20} + \frac{100}{30} \\ &= 5 + 3.33 + 5 + 3.33 \\ &= 16.67 \end{aligned}$$

Step 6: Calculate degrees of freedom

$$df = (\text{rows} - 1) \times (\text{columns} - 1) = (2 - 1) \times (2 - 1) = 1$$

Step 7: Find p-value With $\chi^2 = 16.67$ and $df = 1$, the p-value < 0.001 (very small).

Step 8: Make decision Since p-value < α (0.05), we **reject** H_0 . Gender and Interest are statistically correlated, indicating one attribute may be redundant.

3.2 Detecting Duplicate Records

When recording values in databases from non-standardized sources, data often presents missing values or **duplicate records**. Information may be recorded two or more times, which significantly affects data quality. **Duplicate data** (Record Linkage) refers to records whose values match in all study variables selected by the data analyst. This occurs due to various causes: incorrect data entry, repeated entry of the same value, or data corruption.

Duplicate data not only refers to the same entity, but also to attributes or instances that, despite having different contents, should be the same. Therefore, duplicate record detection aims to explore one or more sources for data that should be unique but are not due to different representations.

Databases containing names, addresses, or other data can easily be affected by duplicate entries as a result of manipulation by multiple people or data entry at different times and under different circumstances. An algorithm is necessary to gather all possible duplicate records and merge or remove some of them.

An example of duplicate records is shown in Table 5. The dataset contains four variables (Id, first name, last name, age, and height) for seven subjects. Records in rows 6 and 7 are identical, indicating a duplicate entry for Id 7.

Id	First Name	Last Name	Age	Height (m)
1	Angela	Castro	27	1.67
2	Adrian	Guzman	31	1.80
3	Theodoro	Rivadeneira	36	1.61
4	Angela	Castillo	27	1.77
5	Adrian	Casas	53	1.88
6	Beatriz	Perez	48	1.69
7	Olivia	Apraez	36	1.62
7	Olivia	Apraez	36	1.62

Table 5: Database with duplicate records (Id 7 appears twice)

3.3 Mechanisms for Replacing Missing Data

Missing values are values for attributes that were not entered or were lost during recording. Inadequate handling of missing values can introduce **bias** (systematic error that causes estimates to consistently deviate from true values) and lead to misleading conclusions, limiting the generalization of research results. Common reasons for missing values include manual data entry procedures, equipment errors, and incorrect measurements.

Understanding why data is missing is important for choosing the appropriate treatment method. If values are missing completely at random, the sample may still be representative. However, if values are missing systematically, the analysis may be biased. For example, in a study of IQ and income, participants with above-average IQ might skip the salary question, potentially missing a positive association.

3.3.1 Data Loss Types

- **MAR (Missing at Random):** The probability of missing an observation depends on observed variables but not on missing ones. Given particular values for observed features, the distribution of remaining features is the same between observed and missing cases.

Example: In a survey, income data is missing more often for older participants (age is observed), but the missingness doesn't depend on the actual income value itself.

- **MCAR (Missing Completely at Random):** A special case of MAR where the distribution of a sample with a missing value doesn't depend on observed or unobserved data. Missing values form another possible sample from the probability distribution. When data is MCAR, analysis has no bias; however, data is rarely MCAR.

Example: A data entry error causes random records to be lost, with no relationship to any variable values (e.g., a computer crash that randomly deletes some entries).

- **NMAR (Not Missing at Random):** Missing values depend on both observed values and the missing value itself. This is challenging because obtaining an unbiased estimate requires modeling the missingness mechanism itself, which must then be incorporated into a more complex model for estimating missing values.

Example: In an income survey, high-income individuals are more likely to skip the income question (the missingness depends on the actual income value, which is unobserved).

3.3.2 Handling Mechanisms

Missing value handling in machine learning can be addressed through two mechanisms:

- **Deletion mechanism:** Removes entire records (rows) that contain missing values. Simple but can lead to information loss, especially if many records have missing values. Most appropriate when data is MCAR and missing values are few.
- **Imputation by mean, median, and mode:** Replaces missing values with statistical measures:
 - **Mean:** Average value (for continuous numerical data)
 - **Median:** Middle value (for continuous numerical data, robust to outliers)
 - **Mode:** Most frequent value (for categorical data)

Preserves dataset size but may introduce bias if missingness is not random.

These mechanisms are important because many algorithms cannot handle missing data, making it essential to identify and handle missing values in any dataset.

3.3.3 Other Imputation Methods

Beyond simple statistical imputation, there are advanced methods that analyze relationships between attributes:

- **K-Nearest Neighbour (KNN):** Uses similar records to impute missing values based on distance metrics.
- **Expectation-Maximization (EM):** Iteratively estimates parameters of a probability distribution from incomplete data. Works best with easily maximizable distributions like Gaussian mixture models. Generates a single imputation but tends to underestimate estimation errors.
- **Multiple Imputation (MI):** Generates multiple imputed values from observed data, creating several complete datasets. Less biased than EM but computationally expensive. Uses Markov Chain Monte Carlo (MCMC) methods to introduce randomness, typically from a standard normal distribution.
- **Other methods:** SVM-based, clustering-based, logistic regression, and maximum likelihood procedures.

The choice of imputation method is independent of the learning algorithm and should be selected based on the specific situation and data characteristics.

3.4 Min-Max Normalization

Most datasets contain attributes that vary greatly in magnitude, units, and range. Some machine learning algorithms use Euclidean distance between attributes, ignoring units. Attributes with high magnitudes will weigh much more in distance calculations than features with low magnitudes, making it necessary to scale features to the same magnitude level.

Feature scaling is a critical preprocessing step that can make the difference between a weak and superior model. Scaling is necessary for correct predictions due to:

- Regression coefficients are directly influenced by attribute scale
- Features with larger scale dominate over features with smaller scale
- First-order iterative algorithms (e.g., gradient descent) run more easily with scaled values
- Some algorithms reduce execution time when features are scaled
- Distance-based algorithms are very sensitive to attribute scales

Feature scaling is essential for algorithms that calculate distances between data points, such as linear and logistic regression, artificial neural networks, support vector machines, K-Means clustering, K-Nearest Neighbors, principal component analysis, and gradient descent.

However, scaling is a monotonic transformation and does not affect rule-based algorithms, which do not require normalization. Tree-based algorithms (CART, Random

Forests, Gradient Boosting Trees) use rules (series of inequalities) and do not need feature normalization. Some algorithms like Linear Discriminant Analysis (LDA) and Naive Bayes are designed to handle different feature scales by weighting each feature, so scaling may have little effect.

Min-Max scaling transforms all numerical values $x^{(i)}$ of a numerical attribute to a specific range defined by $[x_{\min}, x_{\max}]$, where x_{\min} corresponds to the minimum and x_{\max} to the maximum of the dataset. To obtain a new transformed value $x_{\text{norm}}^{(i)}$, the following expression is used for each value $x^{(i)}$:

$$x_{\text{norm}}^{(i)} = \frac{x^{(i)} - x_{\min}}{x_{\max} - x_{\min}} \quad (30)$$

Normalization typically refers to a specific case of min-max scaling where the final interval is $[0, 1]$ (i.e., $x_{\min} = 0$ and $x_{\max} = 1$). The interval $[-1, 1]$ is also common. Normalizing all data to the same range prevents attributes with large differences ($x_{\max} - x_{\min}$) from dominating others in distance calculations and misleading the learning process. This normalization is known to accelerate learning in Artificial Neural Networks by helping weights converge faster.

Example: Normalization to [0,1] Consider an attribute `age` with values: [25, 30, 35, 40, 45]. To normalize to $[0,1]$:

- $x_{\min} = 25, x_{\max} = 45$
- For $x = 30$: $x_{\text{norm}} = \frac{30-25}{45-25} = \frac{5}{20} = 0.25$
- For $x = 40$: $x_{\text{norm}} = \frac{40-25}{45-25} = \frac{15}{20} = 0.75$

The normalized values become: [0.0, 0.25, 0.5, 0.75, 1.0].

Example: Problem without Normalization Consider a dataset with two features:

- `income`: [20000, 50000, 80000, 100000] (range: 80000)
- `age`: [25, 30, 35, 40] (range: 15)

In distance calculations, `income` differences (e.g., 30000) will dominate `age` differences (e.g., 5), even though both features may be equally important. After normalization to $[0,1]$, both features contribute equally to distance calculations.

Example: Normalization to [-1,1] To normalize to $[-1, 1]$, the formula becomes:

$$x_{\text{norm}}^{(i)} = 2 \cdot \frac{x^{(i)} - x_{\min}}{x_{\max} - x_{\min}} - 1$$

For the same age values [25, 30, 35, 40, 45]:

- For $x = 30$: $x_{\text{norm}} = 2 \cdot 0.25 - 1 = -0.5$
- For $x = 40$: $x_{\text{norm}} = 2 \cdot 0.75 - 1 = 0.5$

The normalized values become: [-1.0, -0.5, 0.0, 0.5, 1.0].

3.5 Z-score Normalization

Z-score normalization (standardization) transforms a dataset to a normal distribution with mean 0 and standard deviation 1. Values for an attribute x_i are normalized based on the mean (μ) and standard deviation (σ) of the dataset.

In some cases, min-max normalization is not feasible, especially when the minimum or maximum of an attribute $x^{(i)}$ is unknown. Additionally, the presence of outliers can bias min-max normalization by clustering values and limiting the digital precision available to represent them.

When applying this transformation, attribute values have a mean of 0 and standard deviation of 1. The result indicates how far a specific data point is from the mean in terms of standard deviations. This standardization facilitates comparison of different datasets, even when they are in different units or have very disparate values.

The Z-score formula is:

$$z = \frac{x - \mu}{\sigma} \quad (31)$$

where μ is the mean and σ is the standard deviation of the attribute.

Example: Z-score Calculation Consider an attribute `test_scores` with values: [65, 70, 75, 80, 85].

- Mean: $\mu = \frac{65+70+75+80+85}{5} = 75$
- Standard deviation: $\sigma = \sqrt{\frac{(65-75)^2+(70-75)^2+(75-75)^2+(80-75)^2+(85-75)^2}{5}} = \sqrt{50} \approx 7.07$
- For $x = 80$: $z = \frac{80-75}{7.07} \approx 0.71$ (0.71 standard deviations above the mean)
- For $x = 65$: $z = \frac{65-75}{7.07} \approx -1.41$ (-1.41 standard deviations below the mean)

The normalized values become: [-1.41, -0.71, 0.0, 0.71, 1.41].

Example: Comparing Different Datasets A professor wants to compare a student's grade in an undergraduate course (mean = 75, σ = 10) with another student's grade in a graduate course (mean = 85, σ = 5):

- Student A: score = 80 in undergraduate course
- Student B: score = 87 in graduate course

Using Z-scores:

- Student A: $z = \frac{80-75}{10} = 0.5$ (0.5 standard deviations above the mean)
- Student B: $z = \frac{87-85}{5} = 0.4$ (0.4 standard deviations above the mean)

Despite Student B having a higher raw score (87 vs 80), Student A performed better relative to their class (0.5 vs 0.4 standard deviations above the mean).

Example: Handling Outliers Consider data with an outlier: [10, 12, 14, 16, 18, 100].

With min-max normalization to [0,1], the outlier (100) compresses all other values into a narrow range [0.0, 0.09], losing precision. With Z-score normalization, the outlier is still visible but doesn't compress the other values as dramatically, preserving more information about the distribution.

3.6 From Nominal to Binary

Machine learning algorithms such as support vector machines and artificial neural networks cannot properly handle nominal attributes; therefore, the presence of these features in a dataset can be problematic. To solve this issue, nominal variables could be transformed to numerical ones, where each nominal value is encoded as an integer starting from 0 or 1. However, this option is not recommended because it assumes an order or hierarchy that does not exist in the attribute values. This transformation of nominal values to a sequence of integers establishes unequal relationships between pairs of nominal values, which is incorrect.

An alternative transformation is to map each nominal attribute to a set of newly generated attributes; this technique is known as **one-hot encoding** or **hot encoding**.

In the transformation from nominal to binary, the nominal variable is replaced by a new set of binary attributes: if the nominal attribute has N different values, it is replaced by a set of N binary attributes, each representing one of the possible values. For each instance, only one of the N newly created attributes will have a value of 1, while the rest will have the value 0. The variable with value 1 corresponds to the original value of the old nominal attribute.

Example: Problem with Integer Encoding Consider a nominal attribute `color` with values: [red, blue, green].

Integer encoding (not recommended):

Color	Encoded
red	0
blue	1
green	2

Table 6: Integer encoding implies order (red < blue < green)

This incorrectly implies that green (2) is “greater” than red (0) and that the distance between red and blue (1) is different from blue to green (1), which is meaningless for nominal data.

Example: One-Hot Encoding Using **one-hot encoding** for the same `color` attribute:

Color	color_red	color_blue	color_green
red	1	0	0
blue	0	1	0
green	0	0	1
red	1	0	0
green	0	0	1

Table 7: One-hot encoding: each color becomes a binary attribute

Each color value is now represented by three binary attributes. For each instance, exactly one attribute is 1 and the others are 0. This preserves the nominal nature without implying any order or hierarchy.

Example: Multiple Nominal Attributes Consider a dataset with two nominal attributes: `gender` (Male, Female) and `city` (New York, London, Tokyo).

Original data:

Gender	City
Male	New York
Female	London
Male	Tokyo
Female	New York

Table 8: Original dataset with nominal attributes

After one-hot encoding:

gender_Male	gender_Female	city_NY	city_London	city_Tokyo
1	0	1	0	0
0	1	0	1	0
1	0	0	0	1
0	1	1	0	0

Table 9: One-hot encoded dataset (5 binary attributes)

The 2 nominal attributes with 2 and 3 values respectively become 5 binary attributes. Each instance has exactly 2 ones (one for gender, one for city) and 3 zeros.