LECTURER: TAI LE QUY

MACHINE LEARNING

UNSUPERVISED LEARNING AND FEATURE ENGINEERING

TOPIC OUTLINE

UNIT 5

FEATURE SELECTION

STUDY GOALS

- Describe the **different techniques** used to **select relevant features**.
- Explain how to **rank features** according to certain relevant **evaluation criteria**.
- Select the best features in order to **maximize the performances** and **avoid overfitting** of the learned model.

EXPLAIN SIMPLY

1. Briefly describe the **three main influences** of **machine learning models' performances**.

2. Explain how **feature variance** can be used for **feature selection**.

3. Describe the difference between **bottom-up** and **top-down recursive feature selection**.

INTRODUCTION

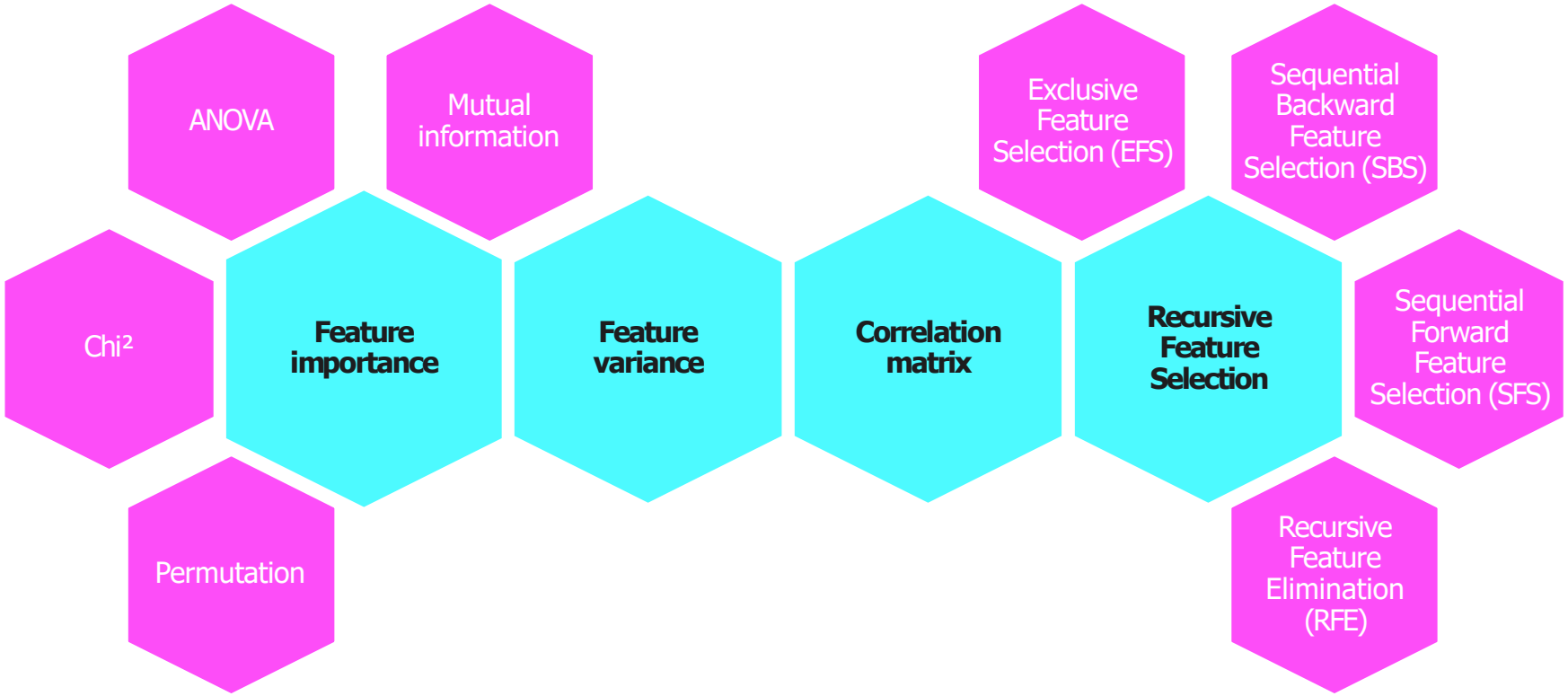
- Feature selection is the process of choosing relevant features from the original features
- Feature selection can be performed manually or with algorithms
- Feature selection algorithms:
  - Filter models: features are selected by studying their characteristics using some statistical evaluation criteria
  - Wrapper models: using a certain learning approach
- Selected features are identified either by:
  - Index: the rank of a feature (e.g., a feature with rank 1 means that it is the best feature with respect to its relevance or importance)
  - Weight: the relevance of features (the higher the feature weight, the better its relevance or rank)
- Feature selection can be either univariate or multivariate
  - Univariate feature selection: each feature is independently evaluated
  - Multivariate feature selection: each feature is evaluated with respect to other features

- Feature selection algorithms:
  - Supervised Feature Selection (SFS): labels (output) of data points are available, selects discriminant features that allow separate data points to belong to different classes.
  - Unsupervised Feature Selection (UFS): labels are not available. UFS is much harder to perform than SFS because defining a feature's relevancy in the absence of output (labels) becomes challenging.

- Feature selection can be either univariate or multivariate
  - Univariate feature selection: each feature is independently evaluated
  - Multivariate feature selection: each feature is evaluated with respect to other features

UNIT CONTENT

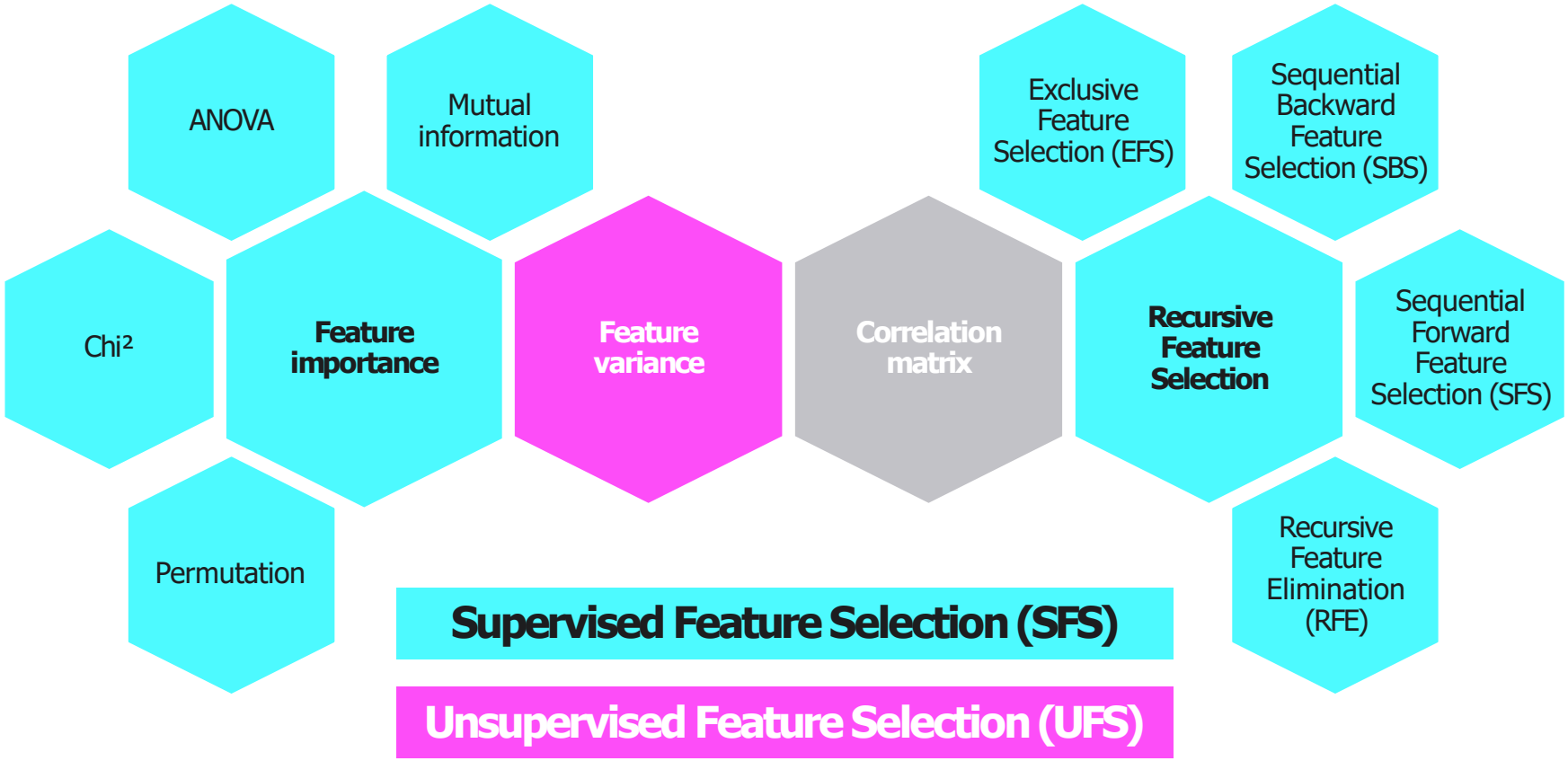
Image 1: Unit content - Feature selection



Source of the image : Müller-Kett, 2021.

UNIT CONTENT

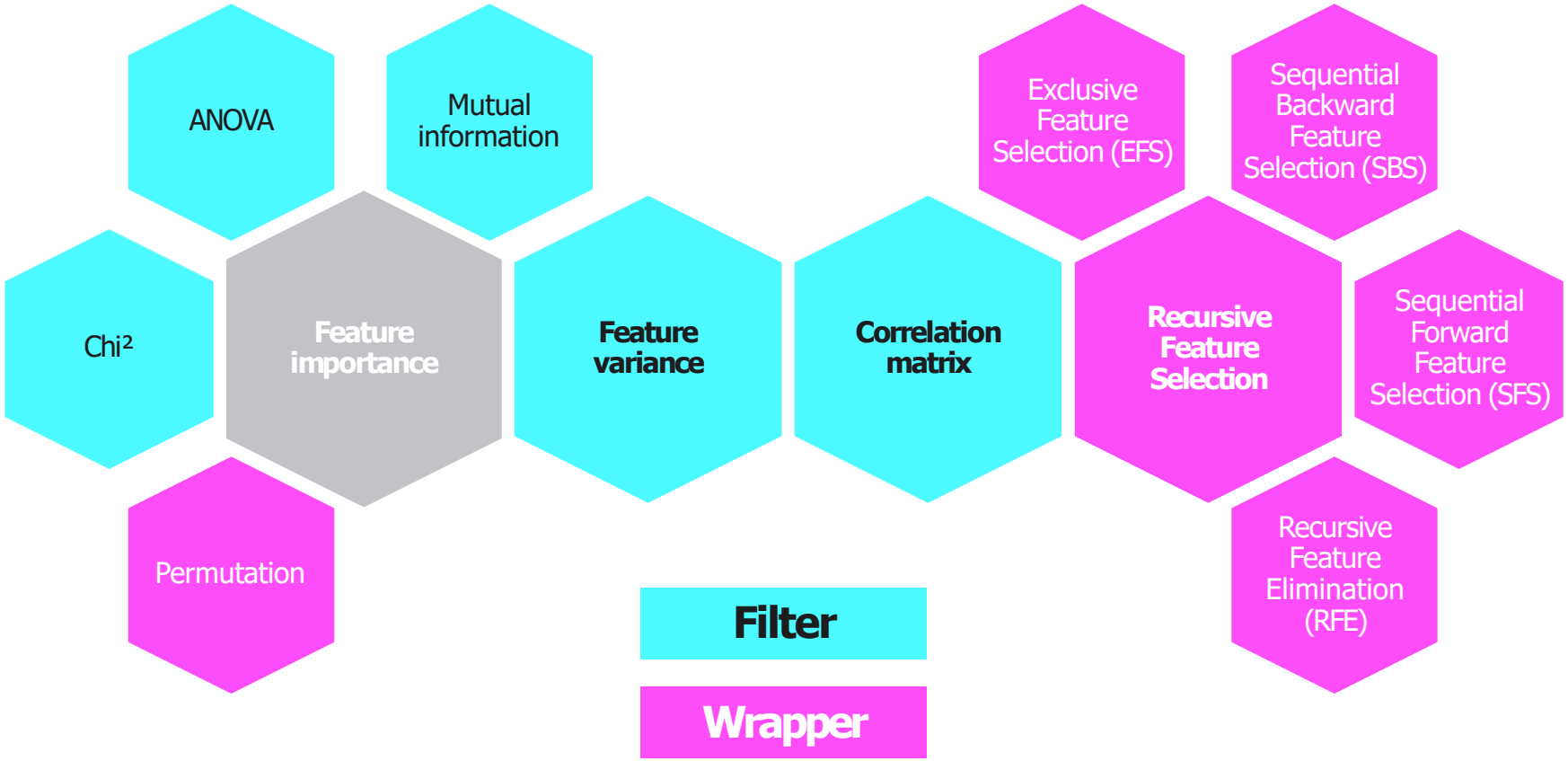
Image 1: Unit content - Feature selection



Source of the image : Müller-Kett, 2021.

UNIT CONTENT

Image 1: Unit content - Feature selection



Source of the image : Müller-Kett, 2021.

FEATURE IMPORTANCE

- Features can be ranked according to their relevance to the response variable (output and labels)
  - Assigning a score to features according to their contribution to the prediction of the response
  - Evaluate correlation between each feature and the response variable
    - The higher the correlation between an input feature and the response variable, the better the score or the importance of this feature in predicting the response
  - Feature importance is primarily a supervised feature selection technique.
- Methods
  - ANOVA Test
  - Chi-Square Test
  - Mutual Information

ANALYSIS OF VARIANCE (ANOVA) TEST

- Compares the variance of one independent feature (input feature) with one dependent feature (output feature) to assess whether they are relevant
- Formulation
  - A dataset X containing k input features (columns) and n data points (rows).
  - Let X_{jm}, j = 1, .., k, be the mean value of each feature j.
  - Let X_m be the overall mean of the dataset.
  - The sum of squares between input features (SSB)

$$SSB = \sum_{j=1}^{k} n \cdot \left(X_{jm} - X_m\right)^2$$

  - The sum of the squared differences between each data point and its corresponding mean, X_{jm}

$$\mathrm{SSE} = \sum_{j=1}^{k} \sum_{i=1}^{n} \left(X_{ji} - X_{jm}\right)^2$$

ANALYSIS OF VARIANCE (ANOVA) TEST

- Formulation
  - Mean square between input features (MSB) and the mean square of errors (MSE)

$$MSB = \frac{SSB}{k-1}, MSE = \frac{SSE}{n-k}$$

k – 1: the degree of freedom 1, n – k: and degree of freedom 2

  - F-value used to accept or reject the null hypothesis

$$F = \frac{MSB}{MSE}$$

- Hypothesis
  - Null hypothesis: feature has no relevance to the response (target) variable
  - Alternative hypothesis: feature has some relevance to the response variable
  - p-value probabilities: probability or chance that the data points occurred under the null hypothesis

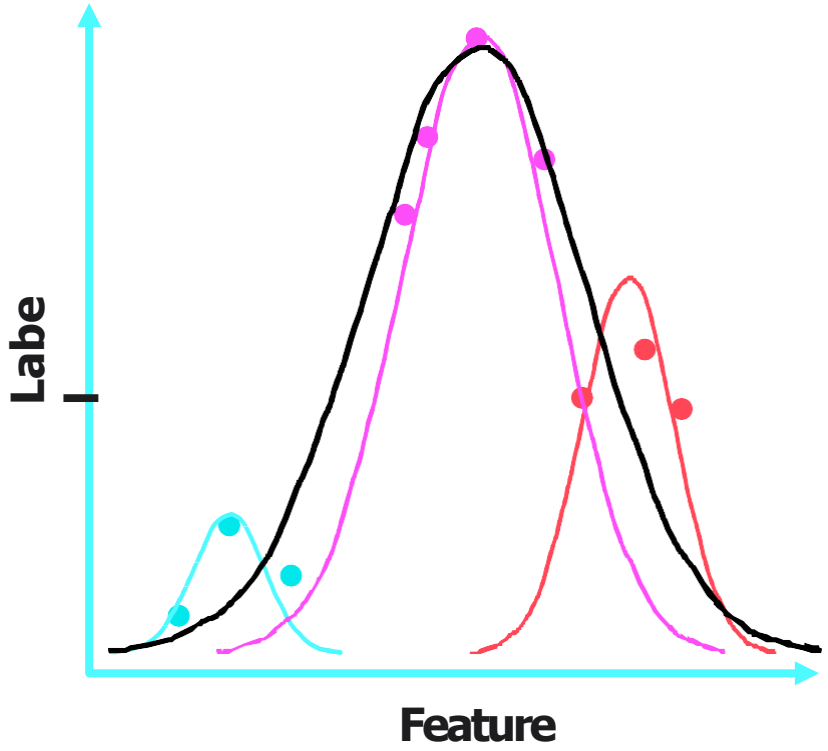
- Hypothesis
  - The lower the p-value of a certain feature, the better its chance to have some relevance (i.e., rejecting the null hypothesis for the alternative hypothesis)
  - Choose the confidence level indicating our confidence in rejecting or accepting the null hypothesis
    - E.g., if we want to be within the confidence interval bigger than or equal to 95 percent and less than 100 percent, an **Alpha level** of 5%.
    - The null hypothesis will be rejected if p <= 0.05

- $p > 5\%$ means "not significant," shown as "n.s." in graphics.
- $0.01 < p <= 5\%$ means "significant," shown with an asterisk, *, in graphics.
- $0.001 < p <= 0.01$ means "highly significant," shown with two asterisks, **, in graphics.
- $p <= 0.001$ means "very highly significant," shown with three asterisks, ***, in graphics.

ANALYSIS OF VARIANCE (ANOVA)

- Compares **differences** in **group means.**
- **Statistical hypothesis test**
- **Input feature**: Numeric
- **Labels**: Categorical
- Choose the features with the **lowest probability** that there is **no difference between groups** (low p-value).

Image 2: ANOVA



Source of text: Bejani et al., 2014.
Source of the image: Müller-Kett, 2023.

CHI² TEST

- Chi-square test allows the evaluation of the independence between 2 variables
- Calculated based on the difference between the observed O and expected E values for input feature with respect to each category of the response variable:

$$Chi - square = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

- Example:

Contingency table

Gen-der	Responded-Yes	Responded-No	Total
Male	O(Male/Yes)	O(Male/No)	n_{Male} = O(Male/Yes) + O(Male/No)
Female	O(Female/Yes)	O(Female/No)	n_{Female} = O(Female/Yes) + O(Female/No)
Total	n_{Yes} = O(Male/Yes) + O(Female/Yes)	n_{No} = O(Male/No) + O(Female/No)	n = n_{Male} + n_{Female}

CHI² TEST

- We consider the patient gender (input feature) independent from the treatment response (response variable)

$$E(Male|Yes) = n \cdot p(Male) \cdot p(Yes) = n \cdot \frac{n_{Male}}{n} \cdot \frac{n_{Yes}}{n}$$

$$E(Male|No) = n \cdot p(Male) \cdot p(No) = n \cdot \frac{n_{Male}}{n} \cdot \frac{n_{No}}{n}$$

$$E(Female|Yes) = n \cdot p(Female) \cdot p(Yes) = n \cdot \frac{n_{Female}}{n} \cdot \frac{n_{Yes}}{n}$$

$$E(Female|No) = n \cdot p(Female) \cdot p(No) = n \cdot \frac{n_{Female}}{n} \cdot \frac{n_{No}}{n}$$

- The Chi-square score

$$\begin{aligned}\text{Chi} - \text{square} \; = & \frac{(O(Male|Yes) - E(Male|Yes))^2}{E(Male|Yes)} \\ & + \frac{(O(Male|No) - E(Male|No))^2}{E(Male|No)} \\ & + \frac{(O(Female|Yes) - E(Female|Yes))^2}{E(Female|Yes)} \\ & + \frac{(O(Female|No) - E(Female|No))^2}{E(Female|No)}\end{aligned}$$

CHI² TEST

- Greater the difference between the observed and the expected values, more the Chi-square value → the input feature (patient gender) is dependent on the response variable (treatment response)
- High p-values of Chi-square tests indicate that the null hypothesis is not correct and therefore it is rejected
- Higher values of Chi-square tests indicate that the feature is more dependent on the response variable, i.e., it has more importance.

CHI^2 TEST

- Compares **expected and observed values** in a **contingency table**
- This table lists each **unique feature value** and **label**, as well as their **occurrences** and **all combinations**
- **Statistical hypothesis test**
- **Input feature**: Categorical
- **Labels**: Categorical
- Choose the features with the lowest probability of **not being related with the labels (low p-value)**

Table 1: Contingency table

	Cat	Dog	Total
Purrs	(cat|purr) 678	(dog|purr) 3	(purr) 681
Barks	(cat|bark) 16	(dog|bark) 129	(bark) 145
Total	(cat) 694	(dog) 132	826

Source of the table: Müller-Kett, 2023.

MUTUAL INFORMATION

- Measures the amount of information gained (the reduction in the uncertainty)
- One variable or feature given a known value of another feature
  - X and Y share mutual information
- The mutual information between two variables X and Y:

$$MI(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \cdot \log_2 \left(\frac{p(x,y)}{p(x) \cdot p(y)} \right)$$

where p(x) and p(y) are the marginal probabilities, and p(x, y) is the joint probability

MUTUAL INFORMATION

$$MI(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \cdot \log_2\left(\frac{p(x,y)}{p(x) \cdot p(y)}\right)$$

– Example

X = x1	X = x1	X = x1	X = x1	X = x2	X = x2
Y = 1	Y = 2	Y = 2	Y = 2	Y = 1	Y = 2

p(x1) = 4/6, p(x2) = 2/6, p(Y=1) = 2/6, p(Y = 2) = 4/6
p(x1,Y = 1)=1/6, p(x1,Y = 2) = 3/6
p(x2,Y = 1) = 1/6, p(x2,Y = 2) = 1/6

$$MI(X,Y) = p(x1, Y = 1) * \log_2\left(\frac{p(x1, Y = 1)}{p(x1) * p(Y = 1)}\right) + p(x1, Y = 2) * \log_2\left(\frac{p(x1, Y = 2)}{p(x1) * p(Y = 2)}\right) + p(x2, Y = 1) * \log_2\left(\frac{p(x2, Y = 1)}{p(x2) * p(Y = 1)}\right) + p(x2, Y = 2) * \log_2\left(\frac{p(x2, Y = 2)}{p(x2) * p(Y = 2)}\right) = -1.87$$

MUTUAL INFORMATION

- Evaluates the **information in one variable** that is **also present in another variable**
- **Marginal probabilities**
- **Joint probabilities**
- **Ranges** from **0** (independent) to $+\infty$
- **Symmetric**

Table 2: Mutual information

	Domain known (x_1)	**Domain unknown (x_2)**	**marginal p**
Spam (y_1)	joint $p(x_1,y_1)$	joint $p(x_2,y_1)$	$p(y_1)$
No spam (y_2)	joint $p(x_1,y_2)$	joint $p(x_2,y_2)$	$p(y_2)$
marginal p	$p(x_1)$	$p(x_2)$	

- $MI(X, Y) = MI(Y, X) = MI(-X, Y) = MI(X, -Y)$

- Better suited to capturing nonlinear relationships between X and Y because it is based on the use of the logarithmic function

Source of the table: Müller-Kett, 2023.

FEATURE PERMUTATION

Evaluate a feature's impact on the model prediction power

1. **Select** a feature candidate
2. **Shuffle** the values in this column
3. **Train** a machine learning model
4. **Assess** the model's performance
5. Select the **next feature** candidate and repeat steps 2 to 4
6. The feature with the **largest model performance drop** is the **most informative**

Table 3: Feature permutation

X1	X2	X3	Y
…	…	…	…
…	…	…	…
…	…	…	…

Source of the table: Müller-Kett, 2023.

FEATURE VARIANCE

1. **Calculate the variances** of each feature

2. **Drop features** with **less** variance than a pre-defined **threshold**

— Very **simple** and **fast** technique

— Useful to **quickly clean** a dataset for **informative columns**, e.g., remove singularity columns

Image 3: Singularity

Selecting High Variance Features (Example)				
	F1	F2	F3	F4
	0	2	0	3
	0	3	4	3
	0	5	1	2
Variance	0	1.55	0.88	0.22

Source of the image: Microsoft archive.

Covariance

- Simultaneous deviations of two variables from their mean

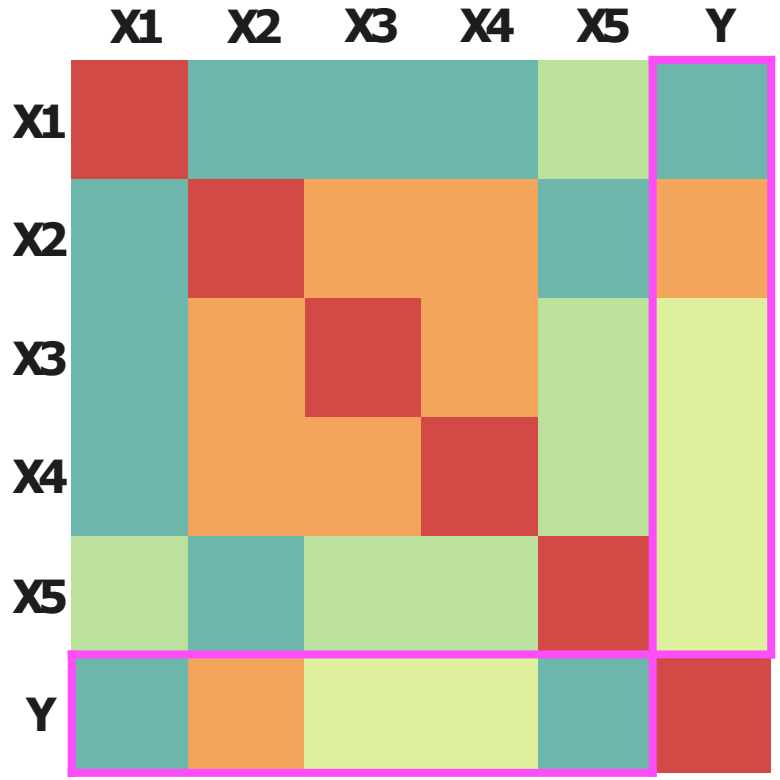
Pearson's correlation

- Normalization by the product of standard deviations
- Range between -1 and 1
- Assumes normal variable distributions

Spearman's correlation

- Based on value ranks
- No normal distribution assumption

Image 4: Correlation matrix



Source of the table: Müller-Kett, 2023.

CORRELATION MATRIX

- Covariance measures a linear relationship between two variables

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \text{mean}_x)\cdot(y_i - \text{mean}_y)$$

where n is the number of rows in X and Y.

- Pearson's correlation
  - Dividing the covariance by the product of their standard deviations

$$\text{Pearson's correlation score}\left(X,Y\right) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

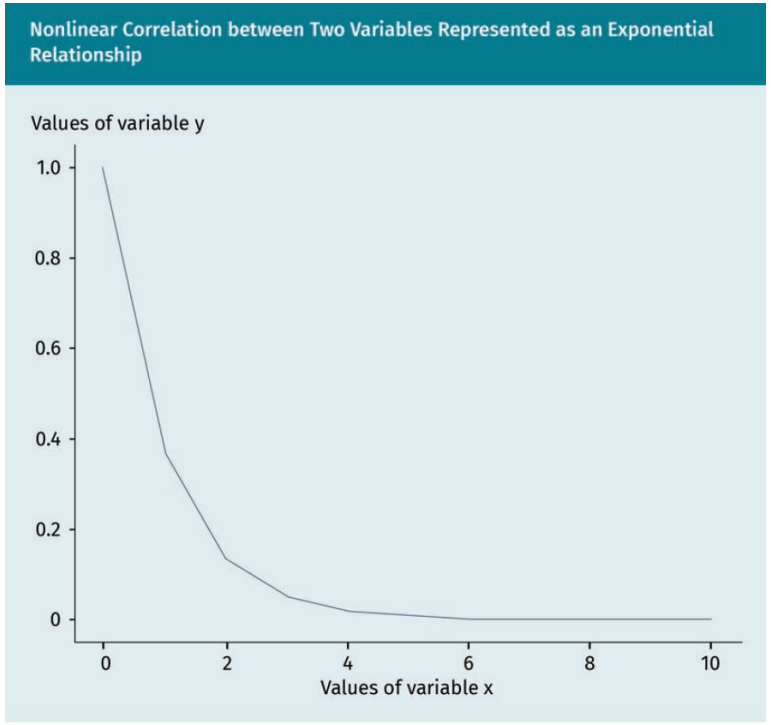
  - Assumption: the data points in each column follow a Gaussian or normal probability distribution and the relationships are linear between the different columns

CORRELATION MATRIX

- Spearman's correlation

$$Spearman's\ correlation\ score\left(X,Y\right) = \frac{Cov(rank(X), rank(Y))}{\sigma_{rank(X)} \cdot \sigma_{rank(Y)}}$$

  - The rank of the values of a variable is obtained by sorting these values in decreasing order → requires a monotonic relationship between the two variables in the sense that when x1 > x2 then y1 < y2.



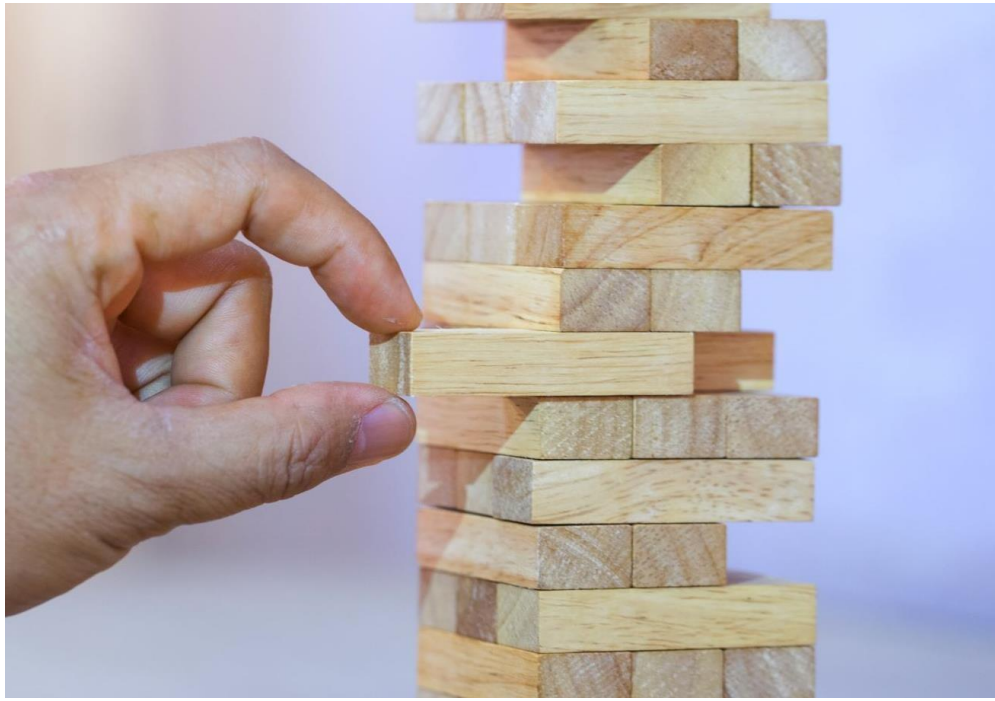
Pearson's and Spearman's Correlation for a Nonlinear Correlation Represented as an Exponential Relationship between Two Variables

x	y	rank(x)	rank(y)
0	1	0	10
1	0.36	1	9
2	0.13	2	8
3	0.05	3	7
4	0.018	4	6
5	0.006	5	5
6	0.002	6	4
7	0.0009	7	3
8	0.0003	8	2
9	0.0001	9	1
10	0.0000	10	0
Pearson's correlation coefficient		Spearman's correlation coefficient	
−0.69		−1	

RECURSIVE FEATURE SELECTION

1. Start with **all features (top-down)** or only **one feature (bottom-up)**
2. **Train** a machine learning model
3. **Assess** the model's performance
4. **Remove** one feature **(top-down)** or **add** one feature **(bottom-up)**
5. **Repeat** steps 2 and 3
6. Continue **until specified number of features** or **defined model performance** is reached

Image 5: Top-down recursive feature selection



Parameters: the number k of features to select and the model or learning algorithm

Source of the image: Microsoft archive.

Exclusive Feature Selection (EFS)

- **Bottom-up** approach
- Best feature combination out of **all possible combinations of k features**
- Computationally **expensive**

Image 5: Top-down recursive feature selection

Recursive Feature Elimination (RFE)

- **Top-down** approach
- Starts by considering all the original features
- Ranks the features according to their importance or weight
- Iterative **the least important feature**
- If the decrease in the model prediction accuracy trained with the remaining features is higher than a predefined threshold, then that feature is important and should be conserved; otherwise, it can be removed.

Source of the image: Microsoft archive.

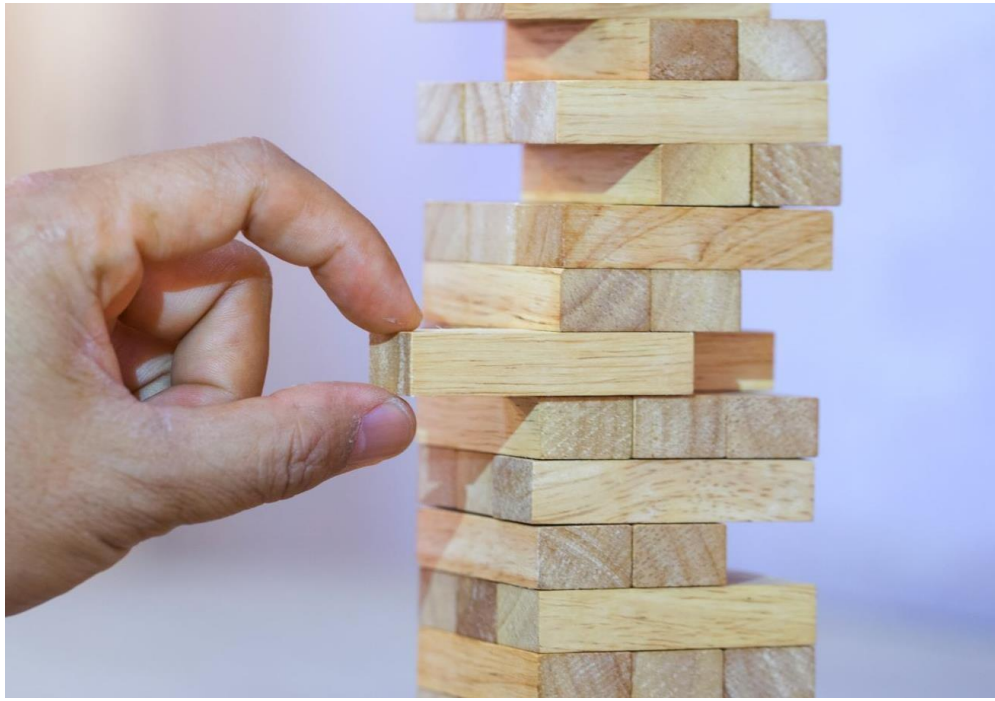
Sequential Forward Feature Selection (SFS)

- **Bottom-up** approach
- Calculates **singular feature importance** first
- Only **combines most important** features, until reaching the predefined number k of features to be selected

Sequential Backward Feature Selection (SBS)

- **Top-down** approach
- Considering all the features together and proceeds to rank all the combinations by removing one feature
- Sequentially **removes the feature not occurring** in the most important feature combination

Image 5: Top-down recursive feature selection



Source of the image: Microsoft archive.

COMPARISON

Methods	Advantages	Disadvantages
Exclusive Feature Selection (EFS) (bottom-up)	Studying all the potential combinations of the original features	Worst computation complexity
Recursive Feature Elimination (RFE) (top-down)	- Computation is less complex than SFS and SBS - Use feature weight coefficients or feature importance to remove the least important feature	RFE exacerbates overfitting because of keeping important features instead of useful ones
Sequential Forward Feature Selection (SFS) (bottom-up)	- Eliminates the least important feature by using the feature importance coefficient or weight - Better computation complexity than EFS	Unable to investigate the usefulness of a feature after being added from the feature set.
Sequential Backward Feature Selection (SBS) (top-down)	- Removes each one of these features in order to determine its importance on the model accuracy rate - Better computation complexity than EFS	Unable to investigate the usefulness of a feature after being removed from the feature set.

Describe the **different techniques** used to **select relevant features**.

Explain how to **rank features** according to certain relevant **evaluation criteria**.

Select the best features in order to **maximize the performances** and **avoid overfitting** of the learned model.

SESSION 5

TRANSFER TASK

TRANSFER TASKS

A start-up that **sells sustainable products in smaller stores worldwide** has been very successful in recent years.

To better understand **what makes satisfied customers**, a **survey** was conducted.

As a Data Scientist, you and your team are tasked to **determine the most relevant influences** that contribute to customer satisfaction.

TRANSFER TASKS

Here, you can see the **first couple of rows** (there are **1000 in total**) and **columns** (there are **500 in total**) of the dataset provided to you:

Table. 4: Transfer task data sample

Has placed an order	Has received discount	Number of orders	Satisfied
Yes	No	2	Yes
Yes	Yes	1	No
Yes	Yes	7	Yes

- Discuss **potential problems** with this dataset that will be used to **train a machine learning model** and **solution strategies**. By doing so, also evaluate potential problems of **individual columns**.
- Marketing likes to provide a metric for the **effectiveness of giving discounts**, and they also want you to provide a **test statistic** for this metric. Describe which methods are suitable in this context.
- In addition to providing feature importance for each column, management asks you to provide **metrics for feature combinations**. This is a high-priority project, and you are equipped with the **most giant machines** available. Discuss which methods are appropriate for this use case.

Source of the table: Müller-Kett, 2023.

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

1. Which one of the following characteristics applies to feature variance?
   a) wrapper feature selection technique
   b) multi-variate feature selection technique
   c) supervised feature selection technique
   d) unsupervised feature selection technique

2. Which one of the following characteristics applies to a correlation matrix?
   a) unsupervised and supervised feature selection technique
   b) supervised feature selection technique
   c) unsupervised feature selection technique
   d) wrapper feature selection technique

3. Which one of the following characteristics applies to recursive feature elimination?

a) filter feature selection technique
b) wrapper feature selection technique
c) uni-variate feature selection technique
d) unsupervised feature selection technique

LIST OF SOURCES

Text

Bejani, M., Gharavian, D., & Charkari, N. M. (2014). Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. *Neural Computing and Applications*, *24*(2), 399—412.

Images

Müller-Kett, 2021.
Müller-Kett, 2023.
Microsoft Archive.

Table

Müller-Kett, 2023.