## 1.1

1. Please list the three main stages composing the KPI design process.

*Creation of the in-house KPI team.*

*Determination of the critical success factors of the organization.*

*Selecting the correct indicators to measure.*

1. Please mark the correct statements.
* *Net profit is a possible key result indicator for an organization.*
* Key performance indicators can be financial in nature.
* It is best to measure a KPI on a yearly basis.
1. Please define the sentence below as true or false.

To better manage the KPI design process, there is a need to create a KPI team, which should be composed of individuals who are not currently working within the organization. True/*False*

## 1.2

1. Please list the two primary KPI categories.

*understanding customers*

*understanding the internal processes of the organization*

1. Please mark the correct statement.
* The customer turnover rate specifies how many customers like the company within a specific timeframe.
* *The cost of calculating the machine downtime level is not high, since the information can be directly collected from the machine data*.
1. Please complete the following sentence:

The cost of collecting the data necessary to calculate the CUR (*capacity utilization rate*) is *high*.

## 1.3

1. Please list at least three reasons to switch to process mining.

*the perfect end-to-end view of the organization’s process model*

*elimination of data differences*

*data-driven decisions*

1. Please mark the correct statements.
* Process mining cannot determine the impact each activity has on different areas of the organization.
* *The techniques associated with process mining use event logs to find bottlenecks.*
* *Process mining can automatically generate the process model with lower costs.*
1. Please mark the following statement as true or false.

Process mining techniques provide a better understanding of the overall OTC process. *True*/False

2.1

1. Please list three problems that may arise while working with the bag of words model.

*Sparse vectors requiring more computational resources.*

*The structure of a sentence is lost.*

*The meaning of the text is not considered.*

*It can be difficult to choose the right words to include in the vocabulary.*

1. Please mark the correct statement.
* TF determines the total number of times a word is found in a corpus of documents.
* *The higher the value of the IDF, the higher the significance of a word.*
* The fit\_transform method returns a matrix, which links documents with topics.
1. Please fill in the empty spaces in the following sentence.

The nltk.tokenize package, includes the *word\_tokenize* method to tokenize a string and remove *stopwords/punctuation*.

2.2

1. Please list the steps followed for the Latent Dirichlet Allocation.

*Choose a number K of topics to discover.*

*Control each document and randomly assign each word to a topic.*

*Review each word in each document.*

*Calculate the words in the document assigned to each topic and the proportion of topic assignments in all documents linked with this word.*

*Reassign a new topic to the word.*

*Repeat the above steps until the desired assignment is achieved.*

1. Please mark the correct statements.
* *LSA is an unsupervised technique in natural language processing*.
* LSA supposes that the topics that make up a document each have a normal distribution over words.
* The two parameters α and β, used in the formula to calculate the significance of a word in the LDA algorithm, should be considered with a value larger than one.
1. Please calculate the proportion of words in a document that are assigned to T1. Let: $α=0.2$, no\_topics=2, no\_words\_document=100 and the number of words linked to T1 in this document is 20.

$$p\left(document d\right)= \frac{no\_{dt}+α}{no\_{d }-1+no\_{t}\*α}=\frac{20+0.2}{100-1+2\*0.2}=0.2$$

## 3.1

1. Please list 6 of the main metrics in Web Analytics.

*number of visits*

*time spent on the website*

*bounce rate*

*exit rate*

*conversion rate*

*engagement*

1. Please mark the correct statement.
* A session ID is linked with each unique visitor entering the website.
* *A cookie is used to identify unique visitors in a website.*
* A good average session duration is about eight minutes.
1. Please define the sentence below as True/False.

Typically, a low bounce rate indicates a possible issue with the website content or user interface True/*False*

## 3.2

1. Please list some of the main types of traffic on a website:

*direct traffic*

*paid search traffic*

*organic traffic*

*social traffic*

*email traffic*

1. Please mark the incorrect statements.
* *SEO related data is found only in web analytics tools and not on external tools offered by search engines.*
* UTM parameters create a trackable URL for further analysis in Google Analytics.
* *The click rate in an email campaign can be calculated as the number of opened emails divided by the total number of emails sent.*
1. Given an example URL as below, please find the value of the parameters: frequency and name of the campaign.

<https://www.audio.com?utm_source=google&utm_medium=daily&utm_campaign=spring_sales>

*frequency: daily*

*name of campaign: spring\_sales*

## 3.3

1. Please list the main approaches used to build recommender systems:

*content-based*

*collaborative-filtering*

*knowledge-based*

*hybrid*

*graph-based*

*context-aware*

1. Please mark the correct statements.
* *Content-based recommender systems do not need to access data related to other users.*
* *A hybrid approach uses the combination of collaborative filtering and content-based approaches.*
* Collaborative filtering approaches do not need to have access to information about the user’s past activity.
1. Please define the sentence below as true or false.

The main problem in the collaborative filtering approaches is known as the cold start. *True*/False

## 4.1

1. Please list the 5 main steps included in Social Media Analytics:

*identification*

*data collection*

*data cleaning and pre-processing*

*data analysis*

*presenting the results and interpretation*

1. Please mark the correct statement.
* *Business analytic consists of analyzing the private data of a company.*
* The largest proportion of the data generated on social media is not publicly accessible.
* The data saved in an organization’s management systems is not well-structured.
1. Please define the sentence below as true or false.

The data generated on social media platforms comes mostly in an unstructured format. *True*/False

## 4.2

1. Please list the three main types of Twitter Streaming APIs.

*public streams: streams that contain public tweets*

*user streams: streams related to the tweets of a specific user*

*site streams: streams related to more than one user*

1. Please mark the correct statement.
* The Pareto principle defines that 20% of the term occurrences come from 80% of the unique terms in a document corpus.
* *The REST API is beneficial when we want to access previously published tweets.*
* The attribute in\_reply\_to\_status\_id contains the ID of the user who originally create the tweet.
1. Please define the sentence below as true or false.

Facebook and Twitter can be considered as both social and interest graphs. *True*/False

5.1

1. Please list the four main steps that should be followed to conduct an A/B test.

*understand the purpose of the website*

*collect data and identify possible bottlenecks*

*define a hypothesis*

*start testing*

1. Please mark the correct statements.
* *A/B testing is one of the cheapest experimentation techniques.*
* *A/B testing is a good way to find weaknesses in the marketing strategy of an institution.*
* A failed test does not provide any important information to the tester.
1. Please define the sentence below as true or false.

The results of A/B testing can identify the specific component on a page that provides a higher conversion rate. True/*False*

5.2

1. Please name the two main multivariate testing techniques.

*Full factorial testing*

*Fractional or factorial testing (also known as partial factorial testing)*

1. Please mark the correct statements.
* Multivariate testing does not require a page to have a high volume of traffic.
* *Multivariate testing is more complex than A/B testing.*
* *Fractional factorial testing provides faster results than full factorial testing.*
* Multivariate testing is faster than A/B testing.
1. Please fill in the blank space.

The home page of a website has three important elements. If we decide to test two versions of these elements, the number of possible variations in a multivariate test would be *eight* combinations.

## 5.3

1. Please define exploration and exploitation in the context of multi-armed bandit algorithms.

*Exploration means continuously experimenting to find the best version of a page in a website.*

*Exploitation means choosing the version which maximizes the profits of an organization.*

1. Please mark the correct statement.
* *A multi-armed bandit algorithm dynamically allocates traffic.*
* A/B testing is the best option to use when a metric needs to be optimized in a short space of time.
* A good multi-armed bandit algorithm never converges.
1. Please indicate whether the following statement is true or false.

Bandit algorithms slowly reduce exploration time and focus the resources on one version of the website.

*True*/False