1. **Implement a feedforward neural network with a hidden layer and use a training dataset to learn how to classify Y in a given dataset. Compute the confusion matrix in the replication dataset.**

2. **Why a random forest is a "random" and a "forest"?**

3. **Which type of pre-processing would you apply to the data and why?**

4. **Which types of layers are typically used in a convolutional neural network?**
   Convolutional, activation (ReLU), pooling, dropout layers…

5. **What is a support vector? In which statistical learning technique is it applied?**
   A support vector is used in **SVM** and is meant to be the points in the **categories** (not clusters, because we are using a **supervised** analysis) that are the closest to the **hyperplanes** that divide the categories.

6. **What is an activation function? In which statistical learning technique is it applied?**
   Activation function is a function that is used in ANN and is meant to transform the combination of inputs into an output inside a neuron. Normally, they are nonlinear and derivable.

7. **Which are the minimum number of hyperparameters that must be defined in a random forest? Why?**
   The number of decision trees, the number of features and samples you want to extract for each tree to generate the boost.

8. **When are you going to use autoencoders?**
   When you want to reduce the noise (image restoration), feature extraction…

9. **In which technique the gain of information statistic is used?**
   Decision trees to decide which feature will be each node.

10. **How can you conduct non-linearly classifications using SVM?**
    Add a new dimension and then you can use a hyperplane to differentiate the categories.

11. **Define the common components that any artificial neural network must have**
    Input, bias unit, weights, activation function, output, neuron, hidden layers…

12. **In which situation will you apply a natural solution algorithm (natural computing)?**
    In NP hard problems such as feature selection

13. Explain differences between a clustering using k-means vs fuzzy k-means

Instead of saying that a sample belongs to one cluster and another sample belongs to another cluster, it says “you belong to one cluster in probability”. This probability is computed based on the distance.
The weight of a point to a centroid is estimated inversely proportional to its distance. Closer points to the prototype of the cluster “j” have more weight. If you are far away, you are not going to contribute much.

So, the new centroid of the cluster is a weighted mean based on the distance of each of the points to that cluster.

14. What is a pattern? How does it relate to overfitting?

A pattern is a set of features characteristic of an individual/sample that are repeated, therefore they are not random.
Overfitting means that the model has learned the noise of the data and not the features. The model learns the specific peculiarities of the data and not on the specific peculiarities that any dataset could have. It is saying that each individual is a pattern.

15. Which are the steps of the algorithm modeling a Gaussian mixture of distributions?

We want to estimate to which cluster, in probability, each point belongs and the mean and sd of each of the clusters.The algorithm typically follows the Expectation-Maximization (EM) framework:

- Initialization: Initialize the parameters of the model. This includes the mean, covariance, and mixing coefficients for each Gaussian component. Common initialization methods include using k-means clustering to get initial cluster centers and estimating covariance matrices.
- Expectation Step (E-step): For each data point, calculate the probability that it belongs to each Gaussian component.
- Maximization Step (M-step): Update the parameters of the Gaussian components to maximize the log-likelihood of the data.
- Iterate between the E-step and M-step until convergence is reached.

16. What is the difference between feature selection and feature extraction? When would you use one or another?

FS: Interpretability, ignores potential interactions between features
FE: Uncover underlying structures, not interpretable

17. Explain the steps to be used in a canonical GA?

A canonical Genetic Algorithm (GA) is an optimization algorithm inspired by the process of natural selection:

- Initialization (initial population), evaluation of fitness, selection (tournament), recombination, mutation, create a new population for the next generation (replace old by new individuals), repeat…

18. Describe another method for doing unsupervised clustering other than k-means family, hierarchical clustering or modeling the data as a mixture of Gaussian distributions.

K-medoids, k-medians, k-mode, fuzzy k-means…

19. Define the hyperparameters in a classical k-Means algorithm

Number of clusters, number of iterations or convergence criteria (when to stop), initialization method and distance metric.

20. Explain the different roles of the internal validation set and the external validation set.

The **internal** validation set has the purpose of controlling the complexity of the algorithm (choosing the optimal parameters). So, it is used to optimize the model. Once we have the final model, we need to **externally** validate (cross-validation) it with more data that has not been used during the calibration dataset. Then we make a performance estimation.

21. Explain in your own words what is a feature vector and a feature table

A combination of d-features is represented as a d-dimensional column vector called feature vector. The feature table is a table that contains the feature vector as columns and the different samples/individuals as rows. There may be an additional column that represents the class or label of each sample. The feature table contains the values of each feature and label for each sample.

22. Explain what a decision boundary is in the context of classifiers.

The feature space, which is the d-dimensional space defined by the feature vector, will be divided in different regions that correspond to each class. The decision boundaries are the lines that delimit the different regions. Meaning that they are the boundaries of each region.

23. Explain the roles of internal and external validation sets in predictive model development

The role of internal validation is to control the complexity of the algorithm. It is used to optimize the different hyperparameters of the model to obtain the best result. For example, the "k" value of the KNN classifier.
The role of external validation is to assess the generalizing power of the model. Meaning that it is used at the final step to make a performance assessment of the final model. In external validation we use new data.

24. Name two figures of merit for regression problems. One of them should be robust to outliers. Explain the reason why

For regression models we can use the L2 loss function. Another figure of merit that is robust to outliers is L1.

25. **A data analyst is doing supervised feature selection with all the dataset. Then he does a data partition to develop the classifier and test it. Is this methodologically correct? Yes/No. Motivate your answer**
    No. I would first make the partition of the data, then I would perform the supervised feature selection and find in internal validation which is the optimum feature selection to make the reduction of dimensionality. So, the complexity control also involves finding the optimum dimensionality.

26. **Explain the basic differences between a clustering using k-means and a clustering based on a mixture of Gaussian distributions**
    K-means is the most popular partitional clustering algorithm that consists in obtaining the clusters that minimize an objective function for a given "k" number of clusters. So, we partition the dataset into clusters by finding the minimum squared error between the various data points in the data set and the centroid of a cluster, then assigning each data point to the nearest cluster.

27. **What is the “curse of dimensionality” phenomenon? How does this relate to the overfitting of a model?**
    For a given number of samples, if we increase the number of features (increase dimensionality) the performance of the model will decrease. As we introduce more dimensions, the space will become more empty. Therefore, each sample will have its own distinct characteristics and all samples will become equally different. As a consequence, the model will not be capable of learning a common pattern between the different samples.
    In light of the above, the model will be overfitted. Meaning that it will learn the noise and not the features, since each sample can be considered as a different pattern

28. **According to the session of unsupervised clustering, and as it is summarized in the article “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data”, which are the limitations of the Kmeans clustering?**
    The limitations are that you need to provide the number of "k" clusters, you don't know if your random inicialization will converge or not (the space is not flat), you may also want to use different definitions of distance to find clusters with different shapes, for big dimensionalities you may need a lot of iterations, etc. For this reason there are different "flavours" of this algorithm that try to solve these specific problems (but they will have other disadvantages).

29. **Which are the minimal steps that any algorithm based on the evolutionary paradigm should implement?**
    Initial data, tournament selection that simulates natural selection, mating of the individuals selected, recombination/mutations to add variability.

30. **What is the purpose of the EM algorithm in the mixture of Gaussian distributions approach)**

31. **What is the difference between a partitional algorithm and a hierarchical algorithm in unsupervised clustering? (0.25 Pnt)**
    Partitional algorithm: You have a set of points and you partition them in 'k' clusters according to an objective function. So, in each iteration you try to maximize your objective function.
    Hierarchical algorithm: At first you are just looking at the relationships between the samples, not looking for clusters. By looking at the relationships you build a dendrogram and then you define a threshold of the deepness of the tree. This threshold will define the different clusters.
    For this reason, in hierarchical algorithms if you use k=4 and then k=5, a cluster will be divided into 2 clusters and all the other points will remain in their respective clusters. This may not happen in partitional algorithms.

1. **Identify all the errors in this paragraph. Write the sentence of each error and explain why it is an error. (0.8)**

   "Unsupervised learning comprises algorithms that require training and a replication dataset. In contrast, supervised learning techniques generate a statistical model based on an expected output. Furthermore, unsupervised training suffers from overfitting. In the case of supervised learning, overfitting is a rare event related to exploding gradients"

   Unsupervised learning does not require a training dataset nor replication dataset, we only have unlabeled input data.
   Furthermore, supervised training suffers from overfitting because we know a priori the output and the training dataset is constructed precisely to that model, so the model is just specific for this data. In the case of unsupervised learning, overfitting is a rare event related to exploding gradients because we predict the model with unlabeled data (explicit model).

2. **Name the steps you must conduct in a dataset when generating a supervised statistical model with many different techniques. (0.4)**
   First we have to pre-process the data (noise filtering, feature extraction, normalization), dimensionality reduction (feature selection and projection), then split the data into the training and test subset and do the prediction (classification). Train the model and evaluate the results using metrics. Finally, decide if the model is good enough or try another model and later do an external validation.

3. **Name a statistical learning technique found by reverse engineering (0.2)**
   Reverse engineering lets us know how a system works. One technique can be perceptron, an algorithm that enables neurons to learn and process information.

4. **Name a way to minimize the overfitting in the case of a decision tree. (0.2)**
   For decision trees we can minimize the overfitting by pruning the branches that are less significant, with a small number of values.

5. **What is a hyperparameter? Name a hyperparameter in a random forest (0.4)**
   Hyperparameters are a set of variables that have to be specified before training a model to be executable. In decision random forest one hyperparameter is the number of features (m, the number of variables used for each tree) and number of items (k)

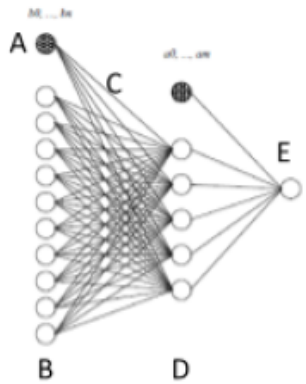
6. **How do you detect the presence of overfitting in a trained statistical model? (0.4)**
   To detect the presence of overfitting there are two ways: by looking in the plot at the creating an hyperplane between the two subgroups or by fitting the model with the test dataset, if the model performed better in the training dataset it is overfitted.

7. **What is a "support vector" in a support vector machine? (0.4)**
   When plotting a model we can see a set of points in the space. To divide the subgroups we have to define a hyperplane between the two groups in a way that the distance between them is maximized. So the support vector is the margin at each side that shows the maximum distance between the hyperplane and the points.

8. **In which statistical learning technique is used the bias unit? What is the purpose of this unit? (0.4)**
   Bias unit is used in the artificial neural network. It is used as an extra neuron in the input and hidden layers that act as the intercept term and has a value of 1 to never go through the origin point.

9. **Explain how to generate "new" datasets in a random forest (0.4)**
   We can generate new datasets using bootstrap. By taking random features from the dataset and generating new subsets.

10. **Name each part of the next feedforward neural network (0.4):**



A: bias unit B: Input layer C: Weight D: Hidden layer E: Output layer

1) Overfitting affects:
   a) supervised methods
   b) unsupervised methods
   c) both

2) A hyperparameter:
   a) Is required in supervised methods but not unsupervised
   b) Is the output parameter estimated in unsupervised methods
   c) Must be specified before running the algorithm

3) A step you always do with the data is:
   a) generate a training and a replication dataset
   b) preprocessing it
   c) scaling it

4) PCA
   a) Is a natural computing algorithm
   b) Uses uncorrelated features
   c) Reduces the dimensionality of the data

5) Bootstrapping is applied to
   a) Decision trees
   b) Random forest
   c) a and b

6) A linear model
   a) Is a special case of artificial neural network
   b) Is a type of unsupervised learning
   c) Does not suffer of overfitting

8) A Confusion matrix
   a) Can be computed with categorical and continuous output variables
   b) Is used to summarize the structure of the data
   c) None of the above —> determines the performance of the classification model

9) ROC curves
   a) Refer to the maximum value that the error can reach in the replication dataset
   b) Are a way to quantify the performance of the generated model
   as double-cross and permutations test
   c) Minimize the area under the curve in the replication dataset

10) Choose one of the following methods to do feature selection in the context of Multi Linear Regression
   a) Ridge Regression → used to analyse any data that suffers from multicollinearity

   b) Partial Least Squares → for baseline correction
   c) LASSO→regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model
   d) Principal Component Regression → for feature extraction

In the context of mass spectrometry data, please explain the differences between binning and mean filtering. Which one can be used for dimensionality reduction?

Mean filtering uses a sliding window to calculate the moving average. It does not perform dimensionality reduction.

Binning partitions the axis into non overlapping windows, and the mean of each window is calculated. It performs dimensionality reduction.

- In contrast the mean filtering is a smoothing technique that reduces noise in the data by averaging the data for a given window length

What is a support vector?

1. The vector of weights of a neural network
2. The closest point to a linear regression plane that separates two categories
3. A closest point from a category to the hyperplane that separates the two categories
4. The vector that defines the centroid (supporting vector) of the K-means algorithm

What is an activation function?

1. It is the weight between two perceptrons
2. It is the non-linear transformation of the output of a perceptron
3. It is the transformation of the output of a perceptron
4. It is the function applied to the transformation of the data prior to applying ML techniques

Which are the minimum number of hyperparameters that must be defined in a random forest?

1. Three.
2. One.
3. Two.
4. Five.

Reinforcement learning

1. Implies defining the policy of the problem
2. Requires prior knowledge of the problem
3. Requires an expected output
4. Does not use Deep Learning

In which technique is used the gain of information statistic?

1. In support vector machines
2. In PCA
3. In decision tree
4. K-means