Quiz 3 - DATA MINING - 10-17 December 2021.

Answer the following questions in the spaces reserved for this use.

1. *(1.75pt)* Write whether the following problems can be solved using supervised data mining algorithms for classification or not. In case it is not possible, explain very briefly why not.

   (a) Given a dataset describing houses sold in a given city with the sell price, predict the sell price for a new house.

   Yes, it can be solved

   (b) Given a dataset with information about the outcomes of football matches in the Spanish league, predict the outcome of a match in the English league.

   No, the dataset does not correspond with the column that is wanted to be predicted

   (c) Given a dataset with information about the outcomes of football matches in the Spanish league to predict the winner of the league.

   Yes, it can be solved

   (d) Given a dataset with pictures of hand gestures and their meaning, recognize moving hand gestures in real time.

   Yes, it can be solved

2. *(1pt)* You are given a data set on cancer detection. You've build a classification model and achieved an accuracy of 96%. Why shouldn't you be happy with your model performance? What can you do about it?

   Because measuring the accuracy only is not a good evaluation option when working with class-imbalanced sets. In this case this dataset might be unbalanced and that is not considered when calculating the accuracy.
   One of the things we can do solve that is balancing the sample by applying some of the resampling techniques we have seen in class: undersampling, oversampling, generate new examples for the smaller class. This will depend on the dataset we have.

3. *(1.5pt)* Given the following confusion matrices generated on the same testing data, show accuracy for both models. Explain also which model you think is better and why.

(a) Model 1

	Predicted positive	Predicted negative
True positive	51	101
True Negative	40	428

Ac=(51+428)/(51+428+101+40)=0,772

(b) Model 2

	Predicted positive	Predicted negative
True positive	61	91
True Negative	80	388

Ac=(61+388)/(61+388+91+80)=0,724

Although both models have a similar accuracy, I would say model 2 is better because it has less unbalanced. If the dataset is very unbalanced, accuracy is not a good measure to consider.

4. *(1.25pt)* When building a classifier using any supervised methods, should we find the best k value for the *k-fold cross-validation* method in order to obtain the best accuracy? Explain why.

No, finding a better k for the k-fold cross-validation does not increase the accuracy because this method only measures the estimating accuracy, but does not change the accuracy. Incrementing the k may say more exactly what is the accuracy but it will not actually improve it.

5. *(1.75pt)* Mark the true sentences and briefly explain your answer.

(a) In general, when training a classifier using the *k-nn algorithm* on an unbalanced training dataset, the best choice for k is to use high values.

False, using a high value of k it is more probable to have less neighbours of the same class if you have an unbalanced dataset. That would lead to errors because we would be doing the prediction with neighbours from another classes

(b) In order to use the *k-nn* method is enough to have a clean dataset without missing values and containing only numerical attributes.

True, even with some missing values this method would work too. If there are categorical variables that want to be added to the k-nn they would need to be converted.

(c) In the *k-nn* algorithm, the distance-weighted parameter is more relevant when k is large than when k is low.

False, when having a low k value that will mean we will select less neighbours and, therefore, the distance-weighted parameter will be more relevant (vice versa when the k is large)

(d) In general, the larger the value of k, the better the accuracy because we have more a more robust estimator.

False, taking the highest possible value of k, the k-nn algorithm would no longer make sense, since the rest of the samples of the dataset would be used to make the prediction and the similarity of the samples would not be taken into account. That is, similar and opposite samples would be used to obtain the result we want to predict and the distance between the data would not be taken into account.

6. *(0.5pt)*Why *Naive Bayes* algorithm is called 'naive' ?

   It's called naive because it makes the assumption that all attributes are independent of each other.

7. *(0.75pt)* Answer if each of the following sentences about the *Naïve Bayes* algorithm *is true or not.*

   (a) In general, when using *Naïve Bayes* algorithm, the larger the number of features on the dataset, the better the performance true

   (b) The smoothing technique is used to reduce the impact of the assumption of independence of features in the dataset. false

   (c) When computing the conditional probability of a numerical feature with respect to the class, we always use the normal distribution. true

8. *(0.75pt)* To reduce overfitting of a Decision Tree, mark which of the following method can be used:

   (a) Increase minimum number of examples allowed in leafs
   (b) Increase depth of trees
   (c) Set a threshold on the minimum information gain to split a node

9. *(0.75pt)* Which of the following are disadvantages of Decision Trees?

   (a) A Decision tree is not easy to interpret
   (b) Decision trees is not a very stable algorithm
   (c) Decision Trees will overfit the data easily if it perfectly describes the training dataset