Quiz 3 - DATA MINING

Name:	

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.

   (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.

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

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

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?

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

(b) Model 2

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

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.

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.

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

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

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

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

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

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

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

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