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.

   No because in this case we want to predict a number (regression problem) and not a category (classification)

   (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.

   It depends. May be not because populations are different. We cannot predict labels from examples of a different population.

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

   Not directly because the thing we want to predict is not available in the description of instances and it cannot derived from a single instance.

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

   True

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?

   No because population is not well balanced. There are a lot of more people without cancer than with cancer. So a stupid model that says "nobody has cancer" could have this high accuracy. We should concentrate on Recall and Precision of the people with cancer class. Use unbalanced dataset methods.

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

   Model 1 has higher accuracy but lower F1 measure on smaller class. In general (depends on the miss-classification cost) model 2 would be better.

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. k value for the k-fold cross-validation says how many data will you use for training and validation and how many times you will repeat the learning process to obtain performance of the system. It is not a parameter of the learning process. Moreover, all classifiers tested should be compared using same k-fold cross-validation and so k.

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.

   No. small values for k are better because being dataset unbalanced, when considering several neighbours, most part of them would belong to the majority class. It is better to use only one neighbour or weighting voting with proportion of examples in training dataset.

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

   No. Data set should be also normalised.

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

(d) In general, the larger the value of k, the better the accuracy because we have more a more robust estimator.
No. See for instance in question (a) what happens with unbalanced datasets.

6. *(0.5pt)*Why *Naive Bayes* algorithm is called 'naive' ?
It's called Naive because it assume that columns in dataset are independent, which usually is not true.

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.
False.

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
True.

(b) Increase depth of trees
False.

(c) Set a threshold on the minimum information gain to split a node
True.

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

(a) A Decision tree is not easy to interpret
False.

(b) Decision trees is not a very stable algorithm
True.

(c) Decision Trees will overfit the data easily if it perfectly describes the training dataset
True.