Quiz 3 - DATA MINING

Name:	Marc Nebot Moyano

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 using supervised data mining algorithms

   (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 cannot be solved because it has nothing to do with the data they have provided us with the data we want to predict, therefore, it would be meaningless

   (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 using supervised data mining algorithms

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

   Yes it can be solved using supervised data mining algorithms

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?

We should not be happy because we should not only measure accuracy to know how well a model works when we can have class-imbalanced sets. Most likely our dataset is unbalanced and that is causing this false score because it is not taking into account these factors. To solve this we can apply undersampling, oversampling or generate more examples to the small class.

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

accuracyModel1=(51+428)/(51+428+40+101)=0.773
accuracyModel2=(61+388)/(61+388+91+80)=0.724

We can observe that both models have unbalanced datasets, as we only have 151 positive and 468 negative samples. So, whichever one we go with we would probably be overestimating the models, I would probably try to choose model 2 and try to balance the samples because model 2 is more balanced.

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.

Finding a better k will not give us a better accuracy model as this method only helps us to estimate the accuracy and therefore does not change it. The higher the k, the better estimates we will get, although obviously it will take a lot longer to estimate it and sometimes we do not need as much accuracy. In conclusion, it will not improve the accuracy of the model.

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.
It is false because if we have few values of the class, if we select a very large k we could select values that do not belong to that class.

(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, as it would work also with some missing values, therefore under better conditions it would also work.

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

False, if we have a low k we will have fewer neighbours, therefore, the distance-weighted parameter will be more relevant.

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

False, if we take the highest k-value k-nn it would not make sense as we would be taking neighbours from other classes and we could generalise too much.

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

We call it naive because it assumes that the attributes are independent, because if they were not, this type of Bayesian probability-based model would not work properly.

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
   (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