10/18/2021 Name:

DATA SCIENCE

MIDTERM TEST

1. To solve a classification problem, we build three models on the data. The accuracy on the training and validation sets of the three models is summarized in the table below. (12%)
    a. Which model would you choose and why?
    b. What phenomena occur regarding the models you did not choose?
    c. Let us assume that all models are kNN models with different *k* values. How would you modify the *k* values of the models you did not choose?

ACCURACY	Training	Validation
Model 1	0.85	0.84
Model 2	0.89	0.79
Model 3	0.79	0.77

2. Are the following statements TRUE or FALSE? Support your answer with justification and correct the false statements. You only get a point if you justify your answer correctly. (18%)
    a. For interval variables, calculating the variance is a meaningful operation.
    b. For two binary vectors, the Jaccard index is always less than or equal to the SMC (simple matching coefficient).
    c. In the kNN algorithm, the variance of the model will be high for low k values.
    d. The *maximum a posteriori* and *maximum likelihood* estimates agree if we assume that the attributes and independent from each other.
    e. Laplace estimation helps to eliminate the bias due to the naive Bayes assumption.
    f. In a diagnostic test, a high recall value is more important than a high precision value.

3. We are given two 5-dimensional feature vectors (16%):

$$a = (3, 5, 0, 4, 6)$$
$$b = (6, 9, 1, 9, 13)$$

   a. Calculate the **Minkowski distance** (L_r distance) of the two vectors with exponents $r = 1, 2, \infty$.
   b. Name the **special cases** of Minkowski distance with exponents $r = 1, 2, \infty$.
   c. Calculate the **cosine** similarity and dissimilarity of the two vectors.
   d. What is the **main difference** between Minkowski distance and cosine (dis)similarity? Name a situation when you think using cosine (dis)similarity is more reasonable!

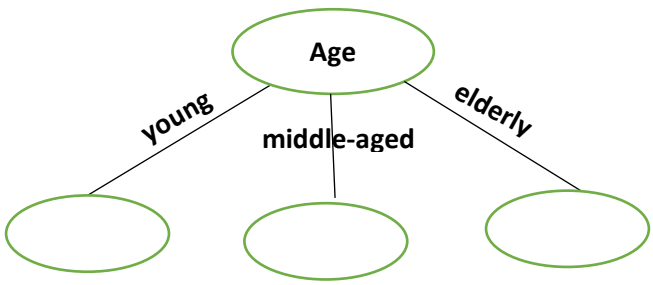
4. A company plans to launch a promotion and asked some people if they were interested. In addition, they recorded three features about each person. We use a **decision tree** to predict who would be interested in the promotion. We are given a dataset with 26 labeled records. We split the data into training and test sets. We train the model on the first 15 instances, and we test it on 10 instances. (32%)
    a. We build a decision tree with depth two on the training data. First, we choose age as the first splitting attribute (see below). Proceed, find **the best splitting attribute** in each child node to build a decision tree with depth two. Use **misclassification error** as the inhomogeneity measure.
    b. Using the decision tree built in part a., test its performance **using the test data**. Determine the **confusion matrix** and calculate the **Accuracy**, **Precision** and **Recall** metrics!
    c. Determine the **confidence scores** (ratio of positive observations) of the leaves based on the training data. Sort the confidence scores of the test instances in ascending order.
    d. Construct the **ROC curve** of the model (using the test instances). If more instances have the same confidence scores ROC curve may change diagonally!
    e. Calculate the **AUC score** of the model!
    f. What is the **probabilistic interpretation** of the AUC score?

Training set

#	*Age*	*Owns a car?*	*Owns a house?*	*Interested in promotion?*
1	young	Yes	Yes	Yes (+)
2	young	Yes	No	Yes (+)
3	young	No	No	Yes (+)
4	young	No	No	Yes (+)
5	middle-aged	Yes	Yes	No (-)
6	middle-aged	Yes	No	Yes (+)
7	middle-aged	Yes	No	Yes (+)
8	middle-aged	Yes	No	No (-)
9	middle-aged	No	Yes	No (-)
10	middle-aged	No	No	No (-)
11	middle-aged	No	No	Yes (+)
12	elderly	Yes	Yes	No (-)
13	elderly	Yes	No	No (-)
14	elderly	No	Yes	Yes (+)
15	elderly	No	No	Yes (+)
16	elderly	No	No	No (-)

Test set

#	*Age*	*Owns a car?*	*Owns a house?*	*Interested in promotion?*
17	young	Yes	Yes	Yes (+)
18	young	Yes	No	Yes (+)
19	young	No	Yes	No (-)
20	middle-aged	Yes	Yes	No (-)
21	middle-aged	Yes	No	Yes (+)
22	middle-aged	No	No	No (-)
23	elderly	Yes	Yes	No (-)
24	elderly	Yes	No	No (-)
25	elderly	Yes	Yes	Yes (+)
26	elderly	No	Yes	Yes (+)



5. Given the following data table with some distinguishing **binary attributes** and some noisy binary attributes that randomly take on the value 1. How would the **k-NN**, **decision tree** and **naive Bayes** classifier perform on the following data? Support your answers with reasoning and sketch calculations. (22%)
    a. Answer the question if the problem is treated as a binary classification problem where the **two class labels** are A and B!
    b. The two classes are further divided into two parts and now the objective is to learn the **four classes** A1, A2, B1 and B2. Evaluate the performance of the three algorithms in this case, too!