YILDIZ TECHNICAL UNIVERSITY
FACULTY OF MECHANICAL ENGINEERING
DEP OF INDUSTRIAL ENGINEERING
2021/2022 FALL SEMESTER
END3971 – Artificial Intelligence and Expert Systems
EXAM: Final

MUDEK Criteria								Total
Question(s)	1	2	3	4				
Grade								

Name-Surname :
Student ID :
Signature :
Res. Asst.:

Duration: 60 minutes
Date: 12.01.2022-10:00
Exam Place: B201 B202 B402

PLEASE SHOW ALL YOUR CALCULATIONS!!!
PLEASE WRITE YOUR RESULTS TO THE BOXES!

Q1. **(15p)** Recall the example that we have solved in the lecture to find the price of a house. We will do the same thing here, in the exam. I have a dataset of 10 houses given in the following

No	Size	Price		No	Size	Price
1	45	96		**6**	83	243
2	55	129		**7**	92	210
3	57	170		**8**	99	199
4	64	129		**9**	110	220
5	80	160		**10**	112	280

Each observation has size as its feature and the output is the price. For example the first house's size is 45m2 and its price is 96TL.
I will use two models to make predictions. KNN with k=1 and KNN with k=2. We will use ordinary KNN, i.e., we will take arithmetic mean and use Euclidian distances.
a. (5p) What is the price of a house with size 60 m2 if we use a KNN model with k=2.

b. (5p) What is the price of a house with size 60 m2 if we use a KNN model with k=3.

c. (5p) We want to see what happens to the model with the choice of k. Please plot the train error and the test error versus the value of k.

Q2. **(20p)** Assume that I am trying to fit a two different models to a given dataset using a 3 fold cross validation. The models has the following results:

			Model 1 R2 Values		Model 2 R2 Values	
Fold1	**Fold2**	**Fold3**	**Train**	**Test**	**Train**	**Test**
Train	*Train*	*Test*	0.85	0.81	0.86	0.78
Train	*Test*	*Train*	0.8	0.75	0.88	0.8
Test	*Train*	*Train*	0.81	0.81	0.75	0.7

a) (10p) Which model would you choose and why? Show your calculations.

b)(10p) What is a n-fold cross validation? Why do we need such a technique?

Q3. (40p) We have a dataset of patients that has a specific tumor type. Each data point has the size of the tumor as its single feature and an output that shows whether a tumor is really a cancer **(y=1)** or not **(y=0).** The logistic regression model for this data is: $h_\theta(x) = g(x/4 - 7)$ Here ***x*** is the size of the tumor and *g(z)* is the sigmoid function. where $g(z) = {}^{1}/_{1+e^{-z}}$

We have the following **test set** that has five patients.

Patient	1	2	3	4	5	6	7	8
Size	12	18	22	42	36	72	14	25
Cancer?	0	0	1	1	1	0	0	0
Prediction								

Finally recall the error metric that is given in the lecture:

$$\frac{1}{m}\sum_{i=1}^{m} err\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right)$$

Where $err\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) = 0$ if your prediction is correct and 1 otherwise. Use natural logarithm (LN) whenever necessary.

a) (10) Find the confusion matrix of this algorithm. Please use the table below. Also use the last line of the above table to create the confusion matrix. Show your calculations.

TN	**FP**
FN	**TP**

b) (5p) Please plot the decision boundary.

c) (25p) Assume you are give a confusion matrix as below. Calculate the required values. Please show your calculations.

TN=300	**FP=5**
FN=15	**TP=25**

Accuracy	
Precision	
Recall	
FPR	
F1 score	

Q4(20) The K-means algorithm is given as follows:

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in R^N$

Repeat {

For i=1 to m

c(i) := index (from 1 to K) of cluster centroid closest to x(i)

μ_k := average of points assigned to cluster k

}

I have two features, say X1 and X2 and 5 sample points and I want to divide them into two clusters using K-means algorithm. The points are (0,0), (0,2) , (1,4), (5,4) , (5,0). As the algorithm proposes, we will start the algorithm by choosing two random data points as the centroids of the clusters. Let the center of the first cluster be (0,2) and let the center of the second cluster be (5,0). Now please iterate the algorithm for one cycle, i.e., first find which cluster does each data point belong to, **and** then update the centroid of the two clusters.

Q5. (10p) Briefly explain why do we use test and cross validation sets in finding appropriate models for predicting in machine learning.

Bonus Question: **You run each model for three times in question 2 (since you are making a three fold CV) hence you have three different parameter set for each model. After you decided to pick a model, which model would you use to make predictions? Or how do you calculate the parameters? You can write at the back of the paper if you need.**