Data Mining

Name:

Surname:

Badge number:

Exercize I (ISLR, Chapter 6, Applied Exercize 10)

We have seen that as the number of predictors used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

a. Generate a simulated data set as follows:

```
set.seed(123)
p = 20
n = 1000
X = matrix(rnorm(p*n), ncol=p)
beta = c(rep(1,p/4),rep(0,3*p/4))
y = X%*%beta + rnorm(n)
```

b. Split your data set into a training set containing the first 100 observations and a test set containing the last 900 observations.

```
# write here
```

c. Perform best subset selection on the training set, and **plot in output** the training set MSE associated with the best model of each size.

```
# write here
```

d. **Plot in output** the test set MSE associated with the best model of each size.

```
# write here
```

e. **Print in output** the model size for which the test set MSE takes on its minimum value. Comment on your results.

```
# write here
```

Write here.

Exercize 2

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

obs.	X1	X2	X3	y
1	0	3	0	red
2	2	0	0	red
3	0	1	3	red
4	0	1	2	green
5	-1	0	1	green
6	1	1	1	red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

a. **Print in output** the Euclidean distance between each observation and the test point, $X_1 = X_2 =$

$X_3 = 0$.

```
# write here
```

b. What is your prediction with $K = 1$?

```
# write here
```

c. What is your prediction with $K = 3$?

```
# write here
```

d. If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

Write here.