CENG 484 - Data Mining
Practical Final Exam

(10p) 1- Find out if the scenarios below are data mining tasks according to your experiences you gained in the lessons, **discuss them briefly**.

Example answer: I think it is a data mining task because .....

a) Analysis of blood pressure to detect hypertension patients.
b) Annual income calculation from the salaries data.
c) Finding the winner in the running competition using time.
d) Estimation of company profit for the next year.
e) Obtaining humidity, temperature data from the multiple sensors.

(18p) 2- Calculate the **information gain** for some attributes from the table below. You should implement information gain function by hand, if you have from the assignment before, you can utilize. You will create information gain function which takes **numerical values (such as number of positive, negative samples or probabilities)** as parameters. It is not required to count positive and negative samples by writing code, you can choose Python or R. (Final_Q_2)

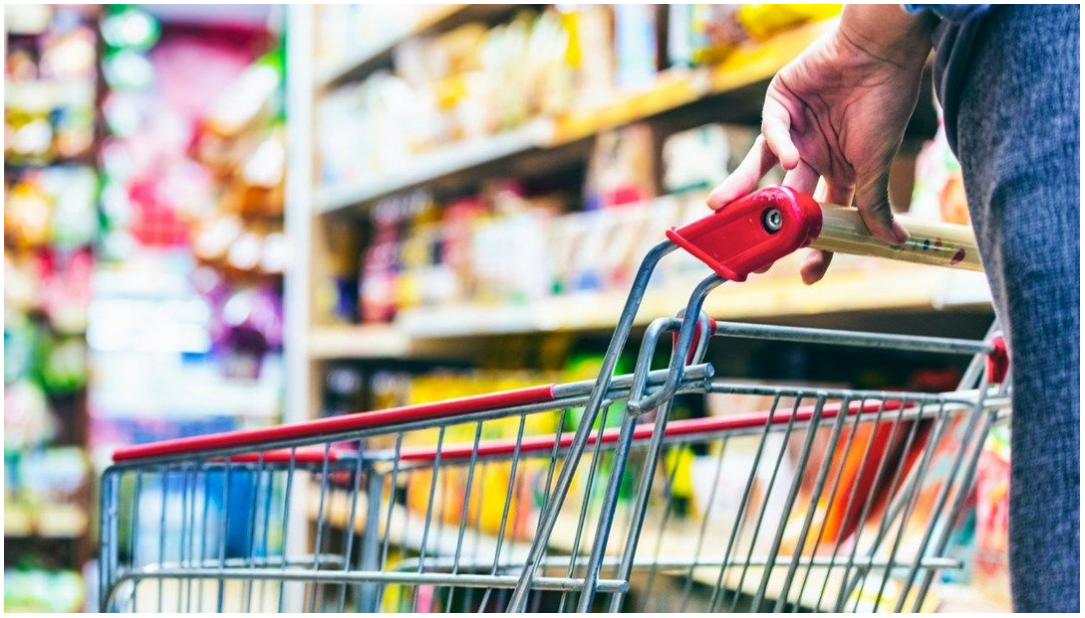
a1	a2	a3	Target
T	T	6.0	+
F	T	4.0	-
F	T	3.0	-
T	T	7.0	+
T	F	1.0	+
F	F	8.0	-
T	F	5.0	+
F	F	4.0	+
T	F	2.0	+
T	F	9.0	-

a) What is the information gain for a1 attribute ?
b) What is the information gain for a2 attribute ?
c) Which attribute produces best split (among a1 , a2) according to the information gain ?

(32p) 3- Try to implement **Naive Bayes** algorithm on **diabetes** data. It will be detect whether a person has diabetes according to some features such as age, glucose, insulin. You will use diabetes.csv and choose Python or R for coding. (Final_Q_3)

a) How to create **zero mean unit variance** data, apply your **solution**. Why should we do this step, **explain briefly**. (You have to write your own function, do not use a library.)
b) Shuffle the data and prepare for **10-fold cross validation**. (You have to write your own function, do not use a library.)
c) Apply **Naive Bayes** classification. (You can use a library to only implement Naive Bayes algorithm.)
d) Calculate **average F1 measure** after cross validation is applied. (You have to write your own calculation function, do not use a library.)
e) Which is the **best fold** according to F1 measure? Discuss in the report.

(40p) 4- Try to analyze **market sales records** which contains purchased goods for each transaction. You will use market_sales.csv and choose Python or R for coding. Complete the steps below. (Final_Q_4)



Association Rules

{whole milk} -> {yogurt}
{other vegetables} -> {whole milk}
{coffee} -> {fruit}
{coffee} -> {sugar}
{soda} -> {coffee}
{hamburger meat} -> {ketchup}
{whole milk,yogurt} -> {coffee}
{coffee,soda} -> {beer}
{chicken,pork} -> {beef}
{chicken,pork,beef} -> {other vegetables}

a) Implement your **support** calculation function and calculate support values for the **attributes** {whole milk}, {yogurt}, {coffee}, {fruit}, {sugar}, {hamburger meat}, {ketchup}, {soda}, {chicken}, {pork}. Which is the best attribute according to calculations?
b) Calculate **support** values for the **association rules** above. Which is the best rule according to calculations?

c) Implement your **confidence** calculation function and calculate confidence values for the **association rules** above. Which is the best rule according to calculations? Is confidence a **symmetric measure**? (Is a->b or b->a equal)
d) Implement your **lift** calculation function and calculate lift values for the **association rules** above. Which is the best rule according to calculations?

Note: This final exam is based on practical data mining applications that you experienced in your assignments. It is like a written final exam. Plagiarism checks will be made, students determined to have **cheated will get zero**. Please complete the entire section **personally**.

Please submit your solutions until **25 June 2020** 09:30. You should upload a zip file "Student_Number_Name.zip" as shown below.