CENG 484 - DATA MINING

PRACTICAL FINAL EXAM

Q1.

a. **Analysis of blood pressure to detect hypertension patients.**

I think it is a data mining task because we can analyse the given data in this case analyse the blood pressure and extract previously unseen useful information such as detecting the hypertension patients.

b. **Annual income calculation from the salaries data.**

I think it is not a data mining task because it is a straightforward calculation. We don't need to analyse or explore any data. From given salaries data, we only make mathematical calculations to get the annual income.

c. **Finding the winner in the running competition using time.**

I think it is not a data mining task because we only measure the time and we dont make predictions about who could win. We do not apply any data mining technique to find the winner.

d. **Estimation of company profit for the next year.**

I think it is a data mining task because from the given data about the company we can analyse the previous years' profit and then using data mining techniques such as classification or regression we can make predictions to estimate the company profit for the next year.

e. **Obtaining humidity, temperature data from the multiple sensors.**

I think it is not a data mining task because we can directly acquire the necessary information about humidity or temperature with the help of sensors. We do not require any technique to analyse the data and find patterns.

Q2.

a. **What is the information gain for a1 attribute ?**

   Information gain for a1 attribute : 0.25642589168200297

b. **What is the information gain for a2 attribute ?**

   Information gain for a2 attribute : 0.01997309402197489

c. **Which attribute produces best split (among a1 , a2) according to the information gain ?**

   Information gain for a1 is much higher than the information gain for a2. We should choose the split that achieves most reduction thus maximum information gain. Therefore the best split is a1 which gives the higher information gain.

Q3.

a. **How to create zero mean unit variance data, apply your solution. Why should we do this step, explain briefly.**

   The solution is applied in Final_Q_3.py
   It is a scaling feature technique. It makes sure that data is internally consistent and each data type has the same content and format. This allows us to compare different data types. For example in the diabetes dataset glucose varies from 50 to 200 DiabetesPedigreeFunction varies from 0 to 1. We cannot compare those two different data types. However when we apply the zero mean unit variance, it gives greater meaning. Thus we can acquire more meaningful results and patterns.

b. **Shuffle the data and prepare for 10-fold cross validation.**

   The solution is applied in Final_Q_3.py

c. **Apply Naive Bayes classification.**

   The solution is applied in Final_Q_3.py

d. Calculate average F1 measure after cross validation is applied.

The solution is applied in Final_Q_3.py
Average F1 score : **0.6292559380974395**

e. Which is the best fold according to F1 measure? Discuss in the report.

According to my splits on the database into folds, the best split is the 10th fold with 0.8045977011494252 F1 score. The F1 score of all the folds are :

F1 scores for each fold :
[0.6428571428571429, 0.679245283018868, 0.5714285714285715,
0.5957446808510639, 0.5714285714285714, 0.5365853658536586,
0.6551724137931035, 0.7169811320754716, 0.5185185185185185,
0.8045977011494252]

Q4.

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?

The solution is applied in Final_Q_4.py
The best attribute according to support calculation is {**whole milk}** with the support value **0.25551601423487547**

b. Calculate support values for the association rules above. Which is the best rule according to calculations?

The solution is applied in Final_Q_4.py
Support values for association rules are calculated. And the best rule is

{'other vegetables'} -> {'whole milk'}

With the support value **0.07483477376715811**

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)**

The solution is applied in Final_Q_4.py
Confidence calculation for association rules are made. And the best rule is

{'chicken', 'pork', 'beef'} -> {'other vegetables'}

With the confidence value **0.5**

These two equations are not equal. Therefore confidence is not a symmetric measure

$$c=\frac{support(a,b)}{support(a)} \neq \frac{support(a,b)}{support(b)}$$

d. **Implement your lift calculation function and calculate lift values for the association rules above. Which is the best rule according to calculations?**

The solution is applied in Final_Q_4.py
Lift values for the association rules are calculated. And the best rule is

{'chicken', 'pork'} -> {'beef'}

With lift value **3.9434643143544506**

NOTE1: In the dataset, if an item contains ‘beer’ word in it, I added it to the support count. For example beer has two types = canned beer and bottled beer and I counted both of them as beer. Also, I counted all the items that contained ‘fruit’ in it.

NOTE2: I only added the best rule’s values. Other rules’ support, confidence lift values can be found in the code. They are printed.