# Hands-on Project

**Question 1:** List the dataset(s) you chose for this project from the UCI Machine Learning respository (https://archive.ics.uci.edu/ml/datasets.php).

I chose Breast cancer data set from UCI Machine Learning repository.

**Question 2:** Describe the dataset in your own words. How many data points, how many attributes, how many types of attributes, how many classes (if any)? Who collected it? How was it collected?

There are 286 data points in data set.Data is multivariate ie:There are 10 attributes in data out of which 9 are categorical. deg-malig is the only numerical attribute.Default task in repo was classification. Since it was meant for classification, there is one target variable with name class. It can take either no-recurrence-events or recurrence-events.The data was provided by Oncology institute.

**Question 3:** What is your goal? Specifically, what insights do you want to learn from this data. Please be aware that clustering, classification, or itemset mining are not 'insights'. These are data mining tasks. Insights are relevant to the domain from which the data is generated.

```
My major goal was to select some categorical data instead of numerical
data and perform some data mining task since all the data mining tasks
discussed in the class took numerical data and performed mining(with few
exceptions like naive bayes which can deal with categorical data by
computing 2^d possibilities).When it comes to data, it is pretty simple
data with very few dimensions. Since it's categorical data, we cannot
visualize it. But as per data set, it is binary classification problem.
Age attribute takes range of 10 -99,inv-nodes and node-caps are also
important attributes in defining whether breast cancer can re-occur or
not.The data also contains tumor-size which shows that people who
already have tumor  with certain parameters, cancer can re-occur even
after successful surgery.Therefore goal of my data mining task is to
mine this data and provide a summary of it(What combination of
attributes assure cancer won't re-occur.Frequent sets containing class
is only considered in order to summarize).
```

**Question 4:** List the data mining task(s) and the specific algorithms you want to perform on this data. Do not pick the tasks listed in the 'Default Task' column on the UCI page.

Data mining task I'm choosing for this data set is Frequent Pattern Mining(while default task on the repo is classification). Algorithm used is Apriori.

**Question 5:** Before selecting the methods you listed in response to Question 3, what are all methods you originally considered to use for the selected data mining task? What was your rationale for selecting the methods you listed in response to Question 3? What was your rationale for not selecting other methods?

```
Before selecting frequent pattern mining, I considered converting
categorical data into numeric data using pandas replace/encoding. But
attributes like age, tumor-size and inv-nodes are interval data and we
cannot replaced with numbers which can best represent them. Also
attributes like breast, node-caps and irradiat can be converted into
binary data which cannot yield good result in clustering. Clustering
methods stumble on categorical data. Therefore I didnt select
clustering.
```

**Question 6:** What limitations does your 'selected' method(s) has(have) that may limit your ability to accomplish the goal you have set for yourself?

As |I| increases in data set,and number of datapoints increase,frequent sets increase. It becomes difficult to shorlist frequent sets containing classes among thousands of frequent sets generated by the algoritm.If the size of I is less and there are more datapoints,we can get good summary of the data.Also, this method won't give any predictive answer. Instead it just summarizes the data(descriptive data mining instead of predictive datamining). Therefore, human intervention is required to analyze the summary and take decision/conclude the case accordingly whether cancer re-occurs or not after the surgery.Also, setting confidence % is challenging.

**Question 7:** Do you have any alternative plan/strategy to overcome the above limitation(s)?

Inorder to limit frequent sets, we can increase minsup. Also, we can use itemset summarization methodologies(such as NDI) to compress the frequent sets obtained.

**Question 8:** For each of the methods you want to use, what parameter choices do you want to use and why? It does not have to be one parameter choice, it could be a collection or a range of choices you may want to consider.

```
For the algorithm i'm using, I need to give parameter minmum support and
file containing dataset/transactions. Minimum support helps in pruning
non frequent data. To get concise and meaningful frequent items, i'm
considering minsup to be 50 in my case.
```

**Question 9:** How will you evaluate that you are successful in your pursuing your goal at the end of the project? In other words, what is your evaluation criteria?

```
We can calculate confidence for dervied frequent sets. For example: if
our frequent set contains:
left no-recurrence-events 0-2 (92)
left no-recurrence-events 0-2 no (90)
left no-recurrence-events no (97)

we observe that when the breast is left and inv-node is 0-2 and node-
caps is no, we can say that chances of re-ocurrence of the cancer is
less. out of 200 no-recurrence events(approx),these co-occur 90 times.
Therefore confidence is 90/200. Therefore we can have a threshold value
for confidence and choose only those sets which cross that threshold. As
confidence gets higher, we can assert our conclusions to be accurate.
```

**Question 10:** How will you evaluate that you are successful in your pursuing your goal at the end of the project? In other words, what is your evaluation criteria?

Type *Markdown* and LaTeX: $\alpha^2$

**Question 11:** Show any visualizations you may have generated to understand your data. Please include the code you used and the plots below. If you borrowed code (entirely or partially) from the hands-on projects or anywhere else, clearly provide a link to your source.

You may use this package to load UCI data in python: https://github.com/SkafteNicki/py_uci (https://github.com/SkafteNicki/py_uci)

No visualizations were done for this data.

**Question 12:** **Perform data mining, evaluate your work and report your findings.** This should include code, plots and results you may have generated. If you borrowed code (entirely or partially) from the hands-on projects or anywhere else, clearly provide a link to your source.

In [7]:
```
!head breast-cancer.data
!wc -l breast-cancer.data
```

```
no-recurrence-events,30-39,premeno,30-34,0-2,no,3,left,left_low,no
no-recurrence-events,40-49,premeno,20-24,0-2,no,2,right,right_up,no
no-recurrence-events,40-49,premeno,20-24,0-2,no,2,left,left_low,no
no-recurrence-events,60-69,ge40,15-19,0-2,no,2,right,left_up,no
no-recurrence-events,40-49,premeno,0-4,0-2,no,2,right,right_low,no
no-recurrence-events,60-69,ge40,15-19,0-2,no,2,left,left_low,no
no-recurrence-events,50-59,premeno,25-29,0-2,no,2,left,left_low,no
no-recurrence-events,60-69,ge40,20-24,0-2,no,1,left,left_low,no
no-recurrence-events,40-49,premeno,50-54,0-2,no,2,left,left_low,no
no-recurrence-events,40-49,premeno,20-24,0-2,no,2,right,left_up,no
286 breast-cancer.data
```

In [6]:
```
import pandas as pd
import numpy as np
cancer_df= pd.read_csv("breast-cancer.csv")
cancer_df.head()
```

Out[6]:

| | class | age | menopaus | tumor-size | inv-nodes | node-caps | deg-malign | breast | breast-quad | irradiat |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | no-recurrence-events | 30-39 | premeno | 30-34 | 0-2 | no | 3 | left | left_low | no |
| **1** | no-recurrence-events | 40-49 | premeno | 20-24 | 0-2 | no | 2 | right | right_up | no |
| **2** | no-recurrence-events | 40-49 | premeno | 20-24 | 0-2 | no | 2 | left | left_low | no |
| **3** | no-recurrence-events | 60-69 | ge40 | 15-19 | 0-2 | no | 2 | right | left_up | no |
| **4** | no-recurrence-events | 40-49 | premeno | 0-4 | 0-2 | no | 2 | right | right_low | no |

In [8]:
```
!chmod u+x apriori
```

In [9]:
```
!./apriori -ts -s-50 breast-cancer.data HandsOn_AP_Freq_S50.txt
```

```
./apriori - find frequent item sets with the apriori algorithm
version 6.27 (2017.08.01)        (c) 1996-2017   Christian Borgelt
reading breast-cancer.data ... [42 item(s), 286 transaction(s)] done [
0.00s].
filtering, sorting and recoding items ... [20 item(s)] done [0.00s].
sorting and reducing transactions ... [247/286 transaction(s)] done [0
.00s].
building transaction tree ... [437 node(s)] done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing HandsOn_AP_Freq_S50.txt ... [150 set(s)] done [0.00s].
```

Some notable results in HandsOn_AP_Freq_S50.txt are: recurrence-events no (68) ge40 no-recurrence-events (94) ge40 no-recurrence-events 0-2 (79) ge40 no-recurrence-events 0-2 no (78) ge40 no-recurrence-events no (88) premeno no-recurrence-events (102) premeno no-recurrence-events 0-2 (83) premeno no-recurrence-events 0-2 no (79) premeno no-recurrence-events no (91) left no-recurrence-events (103) left no-recurrence-events 0-2 (92) left no-recurrence-events 0-2 no (90) left no-recurrence-events no (97) no-recurrence-events 0-2 (167) no-recurrence-events 0-2 no (162) no-recurrence-events no (184)

In [11]:
```
!chmod u+x fpgrowth
```

In [12]:
```
!./fpgrowth -ts -s-50 breast-cancer.data HandsOn_FP_Freq_S50.txt
```

```
./fpgrowth - find frequent item sets with the fpgrowth algorithm
version 6.17 (2017.05.30)        (c) 2004-2017   Christian Borgelt
reading breast-cancer.data ... [42 item(s), 286 transaction(s)] done [
0.00s].
filtering, sorting and recoding items ... [20 item(s)] done [0.00s].
sorting and reducing transactions ... [247/286 transaction(s)] done [0
.00s].
writing HandsOn_FP_Freq_S50.txt ... [150 set(s)] done [0.00s].
```

Some notable results: no-recurrence-events no (184) no-recurrence-events 0-2 no (162) no-recurrence-events 0-2 (167) left no-recurrence-events no (97) left no-recurrence-events 0-2 no (90) left no-recurrence-events 0-2 (92) left no-recurrence-events (103)

In [14]:
```
import datetime
start = datetime.datetime.now()
!./fpgrowth -ts -s-50 breast-cancer.data
end = datetime.datetime.now()
elapsed = end - start
print(elapsed.seconds,"secs ",elapsed.microseconds,"microsecs");
print()
import datetime
start = datetime.datetime.now()
!./apriori -ts -s-50 breast-cancer.data
end = datetime.datetime.now()
elapsed = end - start
print(elapsed.seconds,"secs ",elapsed.microseconds,"microsecs");
```

```
./fpgrowth - find frequent item sets with the fpgrowth algorithm
version 6.17 (2017.05.30)        (c) 2004-2017   Christian Borgelt
reading breast-cancer.data ... [42 item(s), 286 transaction(s)] done [
0.00s].
filtering, sorting and recoding items ... [20 item(s)] done [0.00s].
sorting and reducing transactions ... [247/286 transaction(s)] done [0
.00s].
writing <null> ... [150 set(s)] done [0.00s].
0 secs  239080 microsecs

./apriori - find frequent item sets with the apriori algorithm
version 6.27 (2017.08.01)        (c) 1996-2017   Christian Borgelt
reading breast-cancer.data ... [42 item(s), 286 transaction(s)] done [
0.00s].
filtering, sorting and recoding items ... [20 item(s)] done [0.00s].
sorting and reducing transactions ... [247/286 transaction(s)] done [0
.00s].
building transaction tree ... [437 node(s)] done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing <null> ... [150 set(s)] done [0.00s].
0 secs  246847 microsecs
```

```
The code was entirely taken from FPM Hands on module.
```

since this is a small data set with relatively less I, both apriori and fp-growth don't show any substantial difference in their computational time.

**Question 13:** Putting your findings in the context of your goal and evaluation plan, do you consider yourself successful? Provide reasons for your success or lack thereof.

In [16]:
```
cancer_df[(cancer_df['class'] == 'no-recurrence-events')]
```

Out[16]:

| | class | age | menopaus | tumor-size | inv-nodes | node-caps | deg-malign | breast | breast-quad | irradiat |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | no-recurrence-events | 30-39 | premeno | 30-34 | 0-2 | no | 3 | left | left_low | no |
| **1** | no-recurrence-events | 40-49 | premeno | 20-24 | 0-2 | no | 2 | right | right_up | no |
| **2** | no-recurrence-events | 40-49 | premeno | 20-24 | 0-2 | no | 2 | left | left_low | no |
| **3** | no-recurrence-events | 60-69 | ge40 | 15-19 | 0-2 | no | 2 | right | left_up | no |
| **4** | no-recurrence-events | 40-49 | premeno | 0-4 | 0-2 | no | 2 | right | right_low | no |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **196** | no-recurrence-events | 50-59 | premeno | 14-Oct | 5-Mar | no | 1 | right | left_up | no |
| **197** | no-recurrence-events | 40-49 | premeno | 14-Oct | 0-2 | no | 2 | left | left_low | yes |
| **198** | no-recurrence-events | 50-59 | ge40 | 15-19 | 0-2 | yes | 2 | left | central | yes |
| **199** | no-recurrence-events | 50-59 | premeno | 25-29 | 0-2 | no | 1 | left | left_low | no |
| **200** | no-recurrence-events | 60-69 | ge40 | 25-29 | 0-2 | no | 3 | right | left_low | no |

201 rows × 10 columns

In [18]:
```
(184/201)*100
```

Out[18]:
```
91.54228855721394
```

When minsup is 50, we got few frequent sets containing classes which were interesting(placed them above). Highest support count was 184 for no-recurrence-events no. This shows that with 91% confidence,we can say that whenever there are no node-caps, chances of cancer recurring after operation is very less.In the same way, we can choose some threshold for confidence and consider all the frequent occurring terms above confidence % and get a good summary of the data to analyze.

```
This is a good and quick method to get the gist of data without
performing complex operations on data or
without much pre-processing or data cleaning. If we consider our
confidence threshold as 50%, we get quite a few frequent sets which
summarizes data pretty well. Therefore I consider this as a successful
method  and good shot to get quick summary of categorical data.
```

**Question 14:** If you have an extra month to work on this project, what else would you do? Provide reasons.

```
I wanted to perform chi-squared statistic measure on this categorical
data and find out dependency between features and class. This way, we
can reduce the dimensionality in the categorical data.In short, I would
perform feature selection using chi-squared statistic measures. I would
take more complicated dataset and perform similar task and produce a
text file to domain related people as final product to see the summary
and produce conclusions from it.
```

**Question 15:** Do you consider this project to be in the 'innovative category' or a 'good application' category? Provide your reason.

```
I feel this is good application category since almost the same method is
used for rule based classification in classification module. I donot
know if it's innovative but I wanted to do something differently other
than doing mundane tasks like applying PCA/SVD and cluster/classify
data.
```