In [3]:

```
# import pandas for csv data loading
import pandas as pd
```

Task 1 Logistic regression

The dataset we are gonna use is Glass identification dataset from UCI Machine Learning repository
https://archive.ics.uci.edu/ml/datasets/Glass+Identification
(https://archive.ics.uci.edu/ml/datasets/Glass+Identification)

In [2]:

```
pd.read_csv('data/glass_ident/glass.data').head()
```

Out[2]:

	id	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	class
0	1	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.0	0.0	1
1	2	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.0	0.0	1
2	3	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.0	0.0	1
3	4	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.0	0.0	1
4	5	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.0	0.0	1

This dataset has 10 attributes:

- id is a number representing the specific data point
- RI is the refractive index
- Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
- Mg: Magnesium
- Al: Aluminum
- Si: Silicon
- K: Potassium
- Ca: Calcium
- Ba: Barium
- Fe: Iron

This dataset has 6 class, which is 1,2,3,5,6,7, NOTE THAT IN THIS DATASET THERE IS NO CLASS 4

- 1 building_windows_float_processed
- 2 building_windows_non_float_processed
- 3 vehicle_windows_float_processed
- 4 vehicle_windows_non_float_processed (None in this dataset)
- 5 containers
- 6 tableware
- 7 headlamps

You have two sub-tasks on this dataset;

1. Build a logistic regression model to classify class 2 and not class 2, i.e. a binary classifier to separate class 2 from everything else. This binary classifier should be able to get an accuracy higher than 85%
2. Build a multiclass classification model by build 6 binary classifiers. This multiclass classifier should be able to get an accuracy higher than 50%

Task 2 Linear regression

We are gonna use the Computer Hardware Data Set for this task,
https://archive.ics.uci.edu/ml/datasets/Computer+Hardware
(https://archive.ics.uci.edu/ml/datasets/Computer+Hardware)

In [4]:

```
pd.read_csv('data/cpu_performance/machine.data').head()
```

Out[4]:

	vendor	name	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP	ERP
0	adviser	32/60	125	256	6000	256	16	128	198	199
1	amdahl	470v/7	29	8000	32000	32	8	32	269	253
2	amdahl	470v/7a	29	8000	32000	32	8	32	220	253
3	amdahl	470v/7b	29	8000	32000	32	8	32	172	253
4	amdahl	470v/7c	29	8000	16000	32	8	16	132	132

This dataset has 10 attribute:

- vendor: is the name of the vender
- name: many unique symbols
- MYCT: machine cycle time in nanoseconds (integer)
- MMIN: minimum main memory in kilobytes (integer)
- MMAX: maximum main memory in kilobytes (integer)
- CACH: cache memory in kilobytes (integer)
- CHMIN: minimum channels in units (integer)
- CHMAX: maximum channels in units (integer)
- PRP: published relative performance (integer)
- ERP: estimated relative performance from the original article (integer)

The task is to use the first 8 attribute to predict the 9th attribute (which is the published relative performance)

You will need to build a linear regression model for this task, by using MSE(mean square error) loss function, you are expected to get a loss lower than 6000 on you test set