The Labeling types in ML

```
In the machine learning when we have features in string format we need to
implement different methods to convert the string to numeric calues
```

Build dataframe

In [70]:
```
import pandas as pd

raw_data = {'first_name': ['Jason', 'Molly', 'Tina', 'Jake', 'Amy'],
        'last_name': ['Miller', 'Jacobson', ".", 'Milner', 'Cooze'],
        'age': [42, 52, 36, 24, 73],
        'preTestScore': [4, 24, 31, ".", "."],
        'postTestScore': ["25,000", "94,000", 57, 62, 70]}
df = pd.DataFrame(raw_data, columns = ['first_name', 'last_name', 'age', 'preTes
df
```

Out[70]:

	first_name	last_name	age	preTestScore	postTestScore
0	Jason	Miller	42	4	25,000
1	Molly	Jacobson	52	24	94,000
2	Tina	.	36	31	57
3	Jake	Milner	24	.	62
4	Amy	Cooze	73	.	70

In [63]:
```
import pandas as pd

url = 'http://bit.ly/kaggletrain'
train = pd.read_csv(url)
train.head(5)
```

Out[63]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na

In [5]:
```
import pandas as pd

mushroom = pd.read_table('mushrooms.csv',sep=',',header=1)
mushroom.columns
```

Out[5]:
```
Index(['p', 'x', 's', 'n', 't', 'p.1', 'f', 'c', 'n.1', 'k', 'e', 'e.1', 's.1',
       's.2', 'w', 'w.1', 'p.2', 'w.2', 'o', 'p.3', 'k.1', 's.3', 'u'],
      dtype='object')
```

Type *Markdown* and LaTeX: α^2

In [61]:
```
from sklearn.preprocessing import LabelEncoder
nmushroom=mushroom
n2mushroom=mushroom
nmushroom[nmushroom.columns].head(5)
```

Out[61]:

	p	x	s	n	t	p.1	f	c	n.1	k	...	w	w.1	p.2	w.2	o	p.3	k.1	s.3	u	new_p
0	e	x	s	y	t	a	f	c	b	k	...	w	w	p	w	o	p	n	n	g	0
1	e	b	s	w	t	l	f	c	b	n	...	w	w	p	w	o	p	n	n	m	0
2	p	x	y	w	t	p	f	c	n	n	...	w	w	p	w	o	p	k	s	u	1
3	e	x	s	g	f	n	f	w	b	k	...	w	w	p	w	o	e	n	a	g	0
4	e	x	y	y	t	a	f	c	b	n	...	w	w	p	w	o	p	k	n	g	0

5 rows × 24 columns

Label Types

1- LabelEncoder

LabelEncoder: change the data to the sequence and us it. for example in ['A','A','C','E','C'] => [1,1,2,3,2]

In [62]:
```
#mushroom[mushroom.columns].apply(lambda col: le.fit_transform(col))

#fit = mushroom.apply(lambda x: d[x.name].fit_transform(x))

nmushroom['new_p'] = LabelEncoder().fit_transform(nmushroom['p'])
nmushroom[['p', 'new_p']].head(5)
```

Out[62]:

	p	new_p
0	e	0
1	e	0
2	p	1
3	e	0
4	e	0

Apply the LabelEncoder to the DataFrame

In [65]:
```
n2mushroom.apply(LabelEncoder().fit_transform).head(5)
```

Out[65]:

	p	x	s	n	t	p.1	f	c	n.1	k	...	w	w.1	p.2	w.2	o	p.3	k.1	s.3	u	new_p
0	0	5	2	9	1	0	1	0	0	4	...	7	7	0	2	1	4	3	2	1	0
1	0	0	2	8	1	3	1	0	0	5	...	7	7	0	2	1	4	3	2	3	0
2	1	5	3	8	1	6	1	0	1	5	...	7	7	0	2	1	4	2	3	5	1
3	0	5	2	3	0	5	1	1	0	4	...	7	7	0	2	1	0	3	0	1	0
4	0	5	3	9	1	0	1	0	0	5	...	7	7	0	2	1	4	2	2	1	0

5 rows × 24 columns

4- MultiLabelBinarizer

Change the specift rows and get the features and use it to change to columns

In [56]:
```
# Creating an MultiLabel Array
multilabel_feature = [("New Delhi", "New York"),
                      ("New York", "Sydney", "Hyderabad", "Bangalore"),
                      ("Hyderabad", "Sydney", "Chennai"),
                       ("Chennai", "New Delhi", "Bangalore"),
                      ("Bangalore", "Chennai","Iraq")]
# Printing the MultiLabel Array
print(multilabel_feature)
```

```
[('New Delhi', 'New York'), ('New York', 'Sydney', 'Hyderabad', 'Bangalore'),
('Hyderabad', 'Sydney', 'Chennai'), ('Chennai', 'New Delhi', 'Bangalore'), ('Ba
ngalore', 'Chennai', 'Iraq')]
```

In [59]:
```
# Encoding MultiLabel data using MultiLabel Binarizer
from sklearn.preprocessing import MultiLabelBinarizer
multilabelbinarizer = MultiLabelBinarizer()
multilabel_encoded_results = multilabelbinarizer.fit_transform(multilabel_feature
# Classes created in MultiLabel data after Encoding
multilabelbinarizer.classes_
```

Out[59]:
```
array(['Bangalore', 'Chennai', 'Hyderabad', 'Iraq', 'New Delhi',
       'New York', 'Sydney'], dtype=object)
```

In [60]:
```
df_multilabel_data = pd.DataFrame(multilabel_encoded_results, columns=multilabel
# Viewing few rows of data
df_multilabel_data.head()
```

Out[60]:

	Bangalore	Chennai	Hyderabad	Iraq	New Delhi	New York	Sydney
0	0	0	0	0	1	1	0
1	1	0	1	0	0	1	1
2	0	1	1	0	0	0	1
3	1	1	0	0	1	0	0
4	1	1	0	1	0	0	0

3- get_dummies

We have to spacify the columns to compare them all columns

Compare columns (p,x) with all another columns

In [55]:
```
import pandas as pd

print(nmushroom[['p','x','e']].head(5))
print(pd.get_dummies(nmushroom[['p','x']], columns=["p", "x"], prefix=["p", "x"]
```

```
   p  x  e
0  e  x  e
1  e  b  e
2  p  x  e
3  e  x  t
4  e  x  e
   p_e  p_p  x_b  x_c  x_f  x_k  x_s  x_x
0    1    0    0    0    0    0    0    1
1    1    0    1    0    0    0    0    0
2    0    1    0    0    0    0    0    1
3    1    0    0    0    0    0    0    1
4    1    0    0    0    0    0    0    1
```

Change the specify column (rows) to columns

In [53]:
```
# making data frame from csv at url
data = pd.read_csv("https://media.geeksforgeeks.org/wp-content/uploads/employees

# making dataframe using get_dummies()
dummies = data["Team"].str.get_dummies()
dummies.head(5)
```

Out[53]:

	Business Development	Client Services	Distribution	Engineering	Finance	Human Resources	Legal	Marketing	Produ
0	0	0	0	0	0	0	0	1	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	1	0	0	0	
3	0	0	0	0	1	0	0	0	
4	0	1	0	0	0	0	0	0	

In [ ]:
```
import pandas as pd
```

2- LabelBinarizer

change the value to binary for example ['A','A','C','D'] => [001,001,010,100]

In [34]:
```
from sklearn.preprocessing import LabelBinarizer
import pandas as pd
lb = LabelBinarizer()

train["new_sex"]=lb.fit_transform(train["Sex"])
train.head(5)
```

Out[34]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na