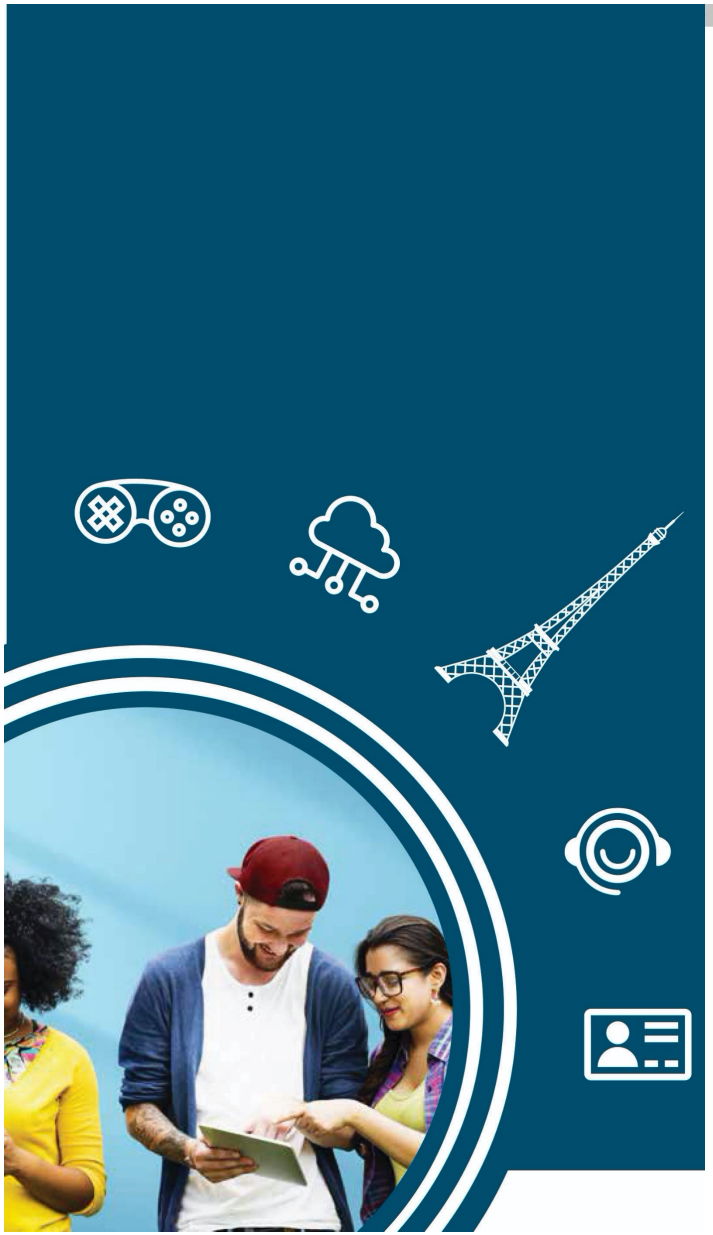


FOUNDATIONS OF STATISTICAL ANALYSIS & MACHINE LEARNING

Amric Trudel
amric.trudel@epita.fr

COURSE PROGRAM

Structure

PREPARATION	Data exploration
	Data preprocessing
REGRESSIONS	Linear regression with one variable
	Multiple and polynomial regressions
	Regression model assessment
CLASSIFICATIONS	Logistic regression
	Classification model assessment
	Discriminant Analysis
	k-NN, Decision Tree, SVM
CLUSTERING	k-means, hierarchical clustering
DIMENSIONALITY REDUCTION	Principal Components Analysis
	Extensions
ALL NOTIONS	Final assignment

PREPARE THE DATA SET

Why?

- Remove irrelevant data
- "Repair" incorrect/missing data
- Set to numeric format
- Split data: predictors/response, train/test

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
1	0	3	Braund, Mr. Owen Harris	male		1	0	A/5 21171	7.25
2	1	1	Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	2. 3101282	7.925
4	1	1	rs. Jacques Heath (Lily May Peel)						
5	0	3							
6	0	3	Moran, Mr. James						
7	0		McCarthy, Mr. Timothy J						
8	0	3	Palsson, Master. Gosta Leonard						
9	1	3	scar W (Elisabeth Vilhelmina Berg)						
10	1	2	sser, Mrs. Nicholas (Adele Achem)						
11	1		Sandstrom, Miss. Marguerite Rut						
12	1	1	Bonnell, Miss. Elizabeth						
13	0	3	Saundercock, Mr. William Henry						
14	0	3	Andersson, Mr. Anders Johan						
15		3	om, Miss. Hulda Amanda Adolfina						
16	1	2	Hewlett, Mrs. (Mary D Kingcome)						
17		3							
18	1	2	Williams, Mr. Charles Eugene						
19	0	3	lius (Emelia Maria Vandemoortele)						
20		3	Masselmani, Mrs. Fatima						
21	0	2	Fynney, Mr. Joseph J						
22	1		Beesley, Mr. Lawrence						
23		3	McGowan, Miss. Anna "Annie"						
24	1	1							
25	0	3	Palsson, Miss. Torborg Danira						
26	1	3	Selma Augusta Emilia Johansson)						
27	0	3	Emir, Mr. Farred Chehab						
28	0	1	Fortune, Mr. Charles Alexander						
29		3	O'Dwyer, Miss. Ellen "Nellie"						
30		3							
31	0	1	Uruchurtu, Don. Manuel E	male	40	0	0	PC 17601	27.7208
32	1	1	William Augustus (Marie Eugenie)	female	57	1	0	PC 17569	146.5208

	X					y
	Embarked	Pclass	Fare	Sex	Age	Survived
training	2	3	7.25	0	22.0	0
	0	1	71.2833	1	38.0	1
	2	3	7.925	1	26.0	1
	2	1	53.1	1	35.0	1
	2	3	8.05	0	35.0	0
	1	3	8.4583	0	21.0	0
	2	1	51.8625	0	54.0	0
	2	3	21.075	0	2.0	0
	2	3	11.1333	1	27.0	1
	0	2	30.0708	1	14.0	1
	2	3	16.7	1	4.0	1
	2	1	26.55	1	58.0	1
	2	3	8.05	0	20.0	0
	2	3	31.275	0	39.0	0
	2	3	7.8542	1	14.0	0
testing	1	3	7.8792	1	35.0	1
	2	3	7.8958	0	25.0	0
	0	1	27.7208	0	40.0	0
	0	1	146.5208	1	57.0	0

PREPARE THE DATA SET

Removing unrelevant data

- Removing a column

```
dataset.drop(['PassengerId'], axis=1)
```

- Removing rows

```
dataset.drop_duplicates()
dataset.drop(dataset[dataset['Fare'] <= 0].index,
   axis=0)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp
1	1	0	3	Braund, Mr. Owen Harris	male	22	1
2	2	1	1	Bradley (Florence Briggs Thayer)	female	38	1
3	3	1	3	Heikkinen, Miss. Laina	female	26	0
4	4	1	1	rs. Jacques Heath (Lily May Peel)	female	35	1
5	5	0	3	Allen, Mr. William Henry	male	35	0
6	6	0	3	Moran, Mr. James	male	21	0
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3
9	9	1	3	scar W (Elisabeth Vilhelmina Berg)	female	27	0
10	10	1	2	sser, Mrs. Nicholas (Adele Achem)	female	14	1
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1
12	12	1	1	Bonnell, Miss. Elizabeth	female	58	0
13	13	0	3	Saundercock, Mr. William Henry	male	20	0
14	14	0	3	Andersson, Mr. Anders Johan	male	39	1
15	15	0	3	om, Miss. Hulda Amanda Adolfina	female	14	0
16	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0
17	17	0	3	Rice, Master. Eugene	male	2	4
18	18	1	2	Williams, Mr. Charles Eugene	male	32	0
19	19	0	3	lius (Emelia Maria Vandemoortele)	female	31	1
20	20	1	3	Masselmani, Mrs. Fatima	female	45	0
21	21	0	2	Fynney, Mr. Joseph J	male	35	0
22	22	1	2	Beesley, Mr. Lawrence	male	34	0
23	23	1	3	McGowan, Miss. Anna "Annie"	female	15	0
24	24	1	1	Sloper, Mr. William Thompson	male	28	0
25	25	0	3	Palsson, Miss. Torborg Danira	female	8	3
26	26	1	3	Selma Augusta Emilia Johansson)	female	38	1
27	27	0	3	Emir, Mr. Farred Chehab	male	28	0
28	28	0	1	Fortune, Mr. Charles Alexander	male	19	3
29	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	35	0
30	30	0	3	Todoroff, Mr. Lalio	male	25	0
31	31	0	1	Uruchurtu, Don. Manuel E	male	40	0
32	32	1	1	William Augustus (Marie Eugenie)	female	57	1

PREPARE THE DATA SET

Filling missing data

- Replace the missing values:

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(missing_values=np.nan, strategy='mean')
imputer.fit(X[:, 1:3])
X[:, 1:3] = imputer.transform(X[:, 1:3])
```


```
imputer2 = SimpleImputer(missing_values='error', strategy='constant',
   fill_value=0)
```

PREPARE THE DATA SET

Preparing features and response sets

- Identification of the predictors

```
X = dataset.iloc[:, :-1]
```

- Identification of the response (case of supervised learning)

```
y = dataset.iloc[:, -1]
```

X

Embarked	Pclass	Fare	Sex	Age
2	3	7.25	0	22.0
0	1	71.2833	1	38.0
2	3	7.925	1	26.0
2	1	53.1	1	35.0
2	3	8.05	0	35.0
1	3	8.4583	0	21.0
2	1	51.8625	0	54.0
2	3	21.075	0	2.0
2	3	11.1333	1	27.0
0	2	30.0708	1	14.0
2	3	16.7	1	4.0
2	1	26.55	1	58.0
2	3	8.05	0	20.0
2	3	31.275	0	39.0
2	3	7.8542	1	14.0

y

Survived
0
1
1
1
0
0
0
0
1
1
1
1
0
0
0

PREPARE THE DATA SET

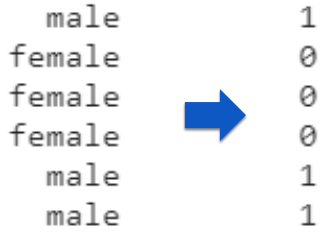
Encoding data

- Boolean data:
  - Explicit encoding:

```
encoding_dict = {'female':1, 'male': 0}
X['Sex'] = X['Sex'].replace(encoding_dict)
```

  - Implicit encoding:

```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
X['Sex'] = label_encoder.fit_transform(X['Sex'])
```



PREPARE THE DATA SET

Encoding data

- Categorical data:
  - Explicit encoding (suitable for an ordered set only):

```
encoding_dict = {'bad':0, 'middle': 1, 'good': 2}
X['Level'] = X['Level'].replace(encoding_dict)
```

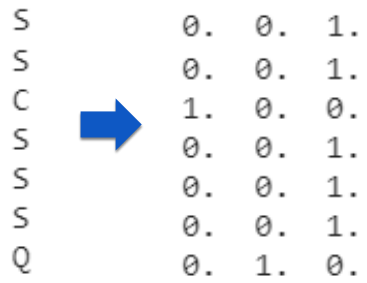
  - Implicit encoding (suitable for distinct categories):

```
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
oh_encoder = ColumnTransformer(transformers=[('encoder',
  OneHotEncoder(), ['Embarked'])], remainder='passthrough')
X = oh_encoder.fit_transform(X)
```

alternative

```
pd.get_dummies(X, columns=['Embarked'],prefix=['is'])
```

Dummy variables

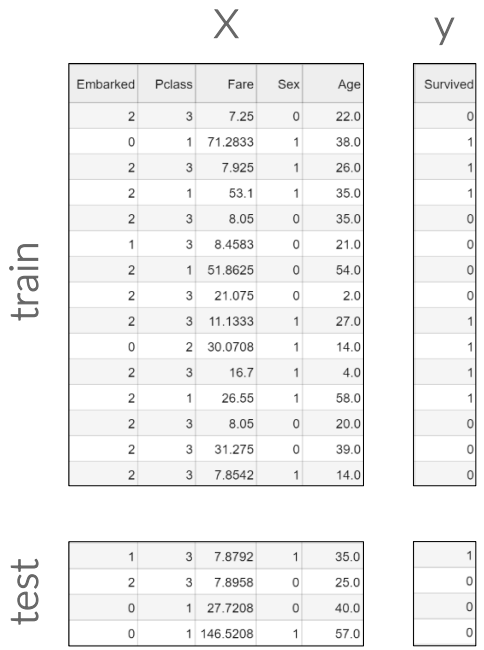


PREPARE THE DATA SET

Splitting the data set

- Split the data set so to have data dedicated to training a model and data dedicated to testing this model:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
   test_size = 0.25, random_state = 1)
```



PREPARE THE DATA SET

Scaling data

Standardisation	Normalisation
$x_{stand} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$	$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$

- Standardisation:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(X_train[:, 5:])
X_train[:, 5:] = scaler.transform(X_train[:, 5:])
```

- Normalisation:

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range=(0,1))
data_rescaled = scaler.fit_transform(data)
```

16.	-0.93
22.	-0.52
31.	0.08
50.	1.36
53.	1.56
38.	0.55
17.	-0.86
40.	0.69
20.	-0.66
47.	1.16
48.	1.23
26.	-0.25
26.	-0.25
40.	0.69
31.	0.08
20.5	-0.62
19.	-0.72
31.	0.08
24.	-0.39
32.5	0.18

PREPARE THE DATA SET

Implementation

- Titanic data set: information on passengers (age, class, survived, etc.)
- Objectives:
  - Make it ready for ML processing

PREPARE THE DATA SET

Practice

- Bank churn: list of customers of a loan bank.
- Objectives:
  - Explore the data
  - Prepare the data set so that it can be directly used for ML processing for churn prediction
- Deadline: Tuesday 29th 6 p.m.

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
5	6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
6	7	15592531	Bartlett	822	France	Male	50	7	0.00	2	1	1	10062.80	0
7	8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
8	9	15792365	He	501	France	Male	44	4	142051.07	2	0	1	74940.50	0
9	10	15592389	H?	684	France	Male	27	2	134603.88	1	1	1	71725.73	0
10	11	15767821	Bearce	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
11	12	15737173	Andrews	497	Spain	Male	24	3	0.00	2	1	0	76390.01	0
12	13	15632264	Kay	476	France	Female	34	10	0.00	2	1	0	26260.98	0
13	14	15691483	Chin	549	France	Female	25	5	0.00	2	0	0	190857.79	0
14	15	15600882	Scott	635	Spain	Female	35	7	0.00	2	1	1	65951.65	0
15	16	15643966	Goforth	616	Germany	Male	45	3	143129.41	2	0	1	64327.26	0
16	17	15737452	Romeo	653	Germany	Male	58	1	132602.88	1	1	0	5097.67	1
17	18	15788218	Henderson	549	Spain	Female	24	9	0.00	2	1	1	14406.41	0
18	19	15661507	Muldrow	587	Spain	Male	45	6	0.00	1	0	0	158684.81	0
19	20	15568982	Hao	726	France	Female	24	6	0.00	2	1	1	54724.03	0