Report

The unsupervised learning for clustering.

Clustering might be the first type of task coming to mind when the unsupervised learning is spoken about. The first dataset is given on https://www.kaggle.com/. The description of dataset is given below and originally goes to baseline data.

Clustering the Countries by using Unsupervised Learning for HELP International

Objective:

To categorize the countries using socio-economic and health factors that determine the overall development of the country.

About organization:

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

Problem Statement:

HELP International have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. So, CEO has to make decision to choose the countries that are in the direst need of aid. Hence, your Job as a Data scientist is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

At first we apply data preprocessing via dealing with missed values. Then take a look on basic statistics of dataset with describe function. The following result is obtained on fig.1.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

Fig.1 Statistics for dataset

After denoting the description for all features is applied. To accomplish this pandas function info was used and an appropriate result is given on figure 2.

```
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   country     167 non-null    object
 1   child_mort  167 non-null    float64
 2   exports     167 non-null    float64
 3   health      167 non-null    float64
 4   imports     167 non-null    float64
 5   income      167 non-null    int64
 6   inflation   167 non-null    float64
 7   life_expec  167 non-null    float64
 8   total_fer   167 non-null    float64
 9   gdpp        167 non-null    int64
```

Fig.2 Description for datatypes

Let’s plot the correlation graphs to show the feature importance of particular values. For that we use seaborn library. The results are given at figures 3 and 4 correspondingly.

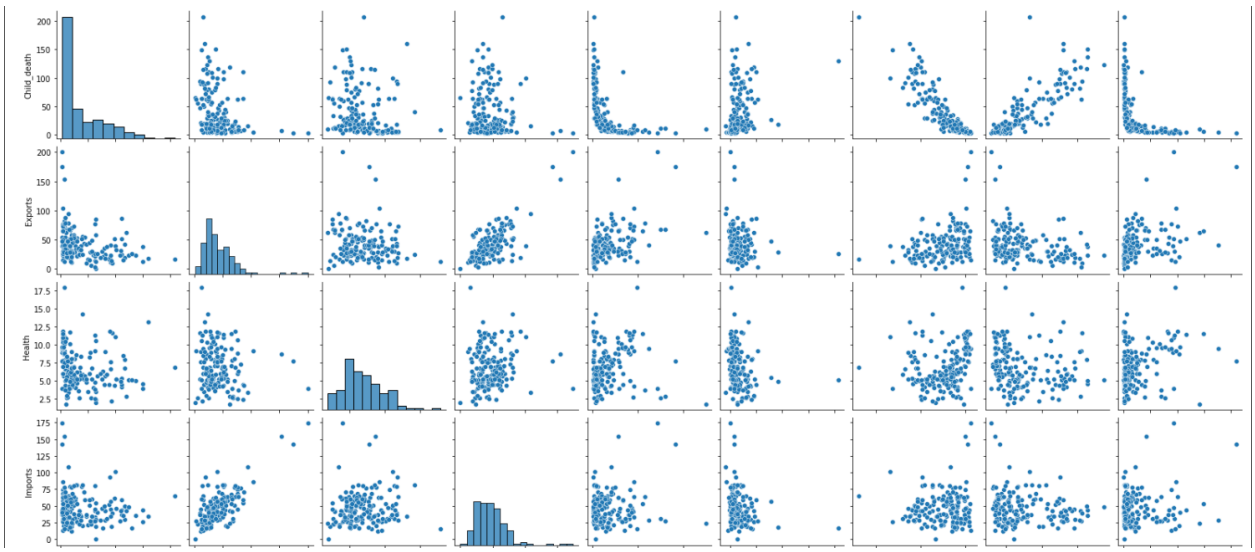

Fig.3 Plot for correlation of high values

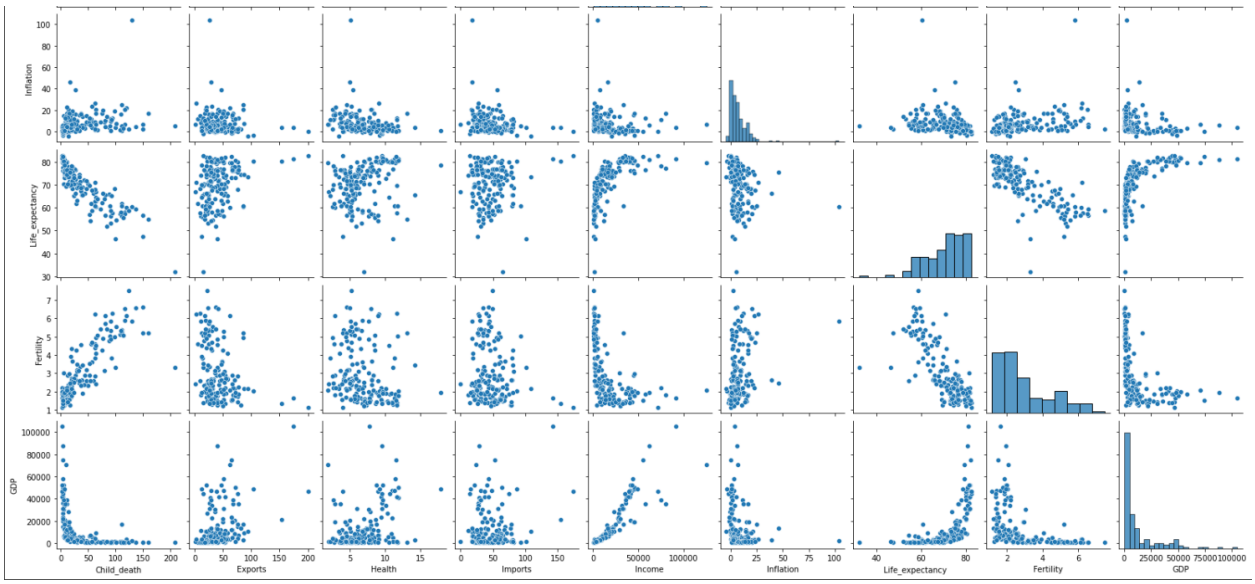

Fig.4 Plot for correlation of low values

Normalization with standard scaler has also been performed and result is given on figure 5.

	Child_death	Exports	Health	Imports	Income	Inflation	Life_expectancy	Fertility	GDP
0	1.291532	-1.138280	0.279088	-0.082455	-0.808245	0.157336	-1.619092	1.902882	-0.679180
1	-0.538949	-0.479658	-0.097016	0.070837	-0.375369	-0.312347	0.647866	-0.859973	-0.485623
2	-0.272833	-0.099122	-0.966073	-0.641762	-0.220844	0.789274	0.670423	-0.038404	-0.465376
3	2.007808	0.775381	-1.448071	-0.165315	-0.585043	1.387054	-1.179234	2.128151	-0.516268
4	-0.695634	0.160668	-0.286894	0.497568	0.101732	-0.601749	0.704258	-0.541946	-0.041817

Fig.5 Normalized results for features

As a training model simple K-means algorithm has been used with parameter of n_clusters equal to 2. The final result for clustering is given on figure 6.

	Child_death	Exports	Health	Imports	Income	Inflation	Life_expectancy	Fertility	GDP	Clusters
0	90.2	10.0	7.58	44.9	1610	9.440	56.2	5.82	553	0
1	16.6	28.0	6.55	48.6	9930	4.490	76.3	1.65	4090	1
2	27.3	38.4	4.17	31.4	12900	16.100	76.5	2.89	4460	1
3	119.0	62.3	2.85	42.9	5900	22.400	60.1	6.16	3530	0
4	10.3	45.5	6.03	58.9	19100	1.440	76.8	2.13	12200	1
5	14.5	18.9	8.10	16.0	18700	20.900	75.8	2.37	10300	1
6	18.1	20.8	4.40	45.3	6700	7.770	73.3	1.69	3220	1
7	4.8	19.8	8.73	20.9	41400	1.160	82.0	1.93	51900	1
8	4.3	51.3	11.00	47.8	43200	0.873	80.5	1.44	46900	1
9	39.2	54.3	5.88	20.7	16000	13.800	69.1	1.92	5840	1

Fig.6 Final results for clustering

The linear regression for regression task on predicting value of salary.

The dataset is available on https://www.kaggle.com/. The first thing we have to do is getting a crystal clear idea about what exactly we are trying to find out. In this particular case, the problem statement is actually super straight-forward. We have to build a model that takes years of experience as input and predict the salary based on that. YearsExperience - It shows number of years of experience an employee has.

Salary - It shows corresponding salaries (in thousand) based on years of experience.

The describe function gives us the following result – figure 7.

	YearsExperience	Salary
count	30.000000	30.000000
mean	5.313333	76003.000000
std	2.837888	27414.429785
min	1.100000	37731.000000
25%	3.200000	56720.750000
50%	4.700000	65237.000000
75%	7.700000	100544.750000
max	10.500000	122391.000000

Fig.7 Statistics for dataset

To see an actual regression trend we perform scatter plot on features

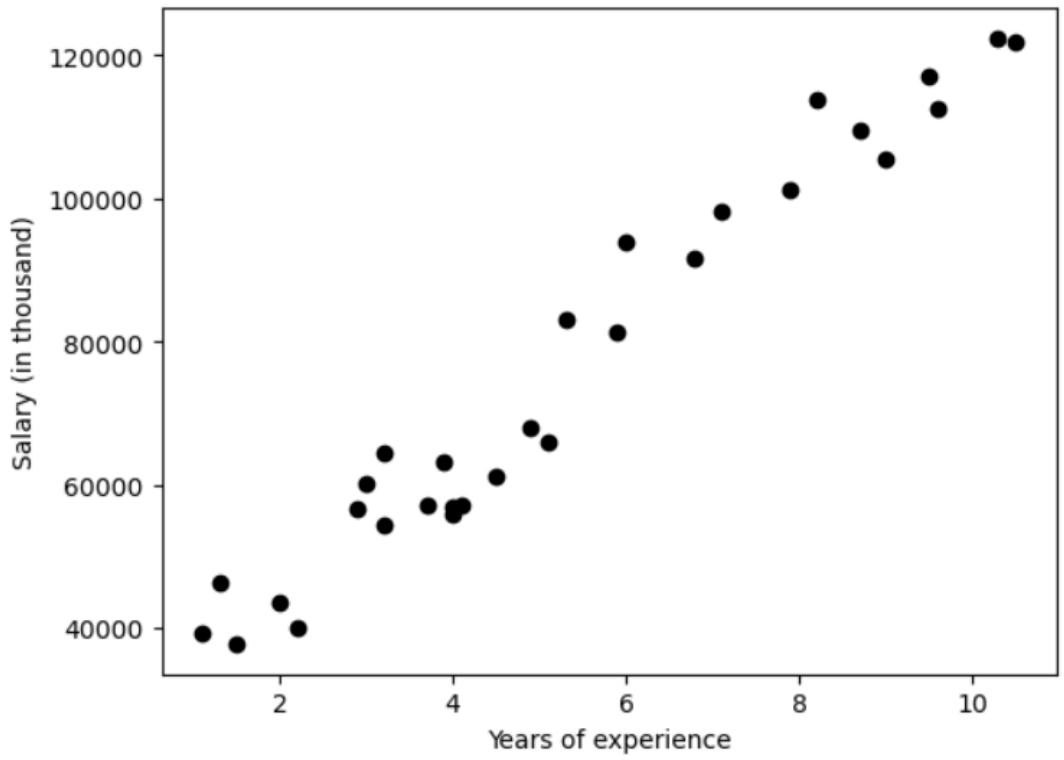

Fig.8 Statistics for dataset

Now we have to split the dataset into two parts, independent variable (Predictor) and dependent variable(target). We know, how much salary you will get is (usually) dependent of how many experience do you have in that respective field. So in our case the dependent variable is Salary and the independent variable is YearsExperience. Let's split the data into X (predictor) and Y (target). The training data is the data set used to train the machine learning model.The testing data is used to evaluate the performance of the trained model. It is used to estimate how well the model will perform on new, unseen data.

The main purpose of using linear regression in machine learning is to predict a continuous(numerical) output variable based on one or more input variables. As the output variable(Salary) is continuous so we will use Linear regression model here Also Linear regression is a simple and straightforward model to implement and interpret, making it a good choice for beginners in machine learning. The coefficients of the model provide insight into how each input variable affects the output variable, making it easy to understand and explain the results.

The result on linear regression is given – figure 9.

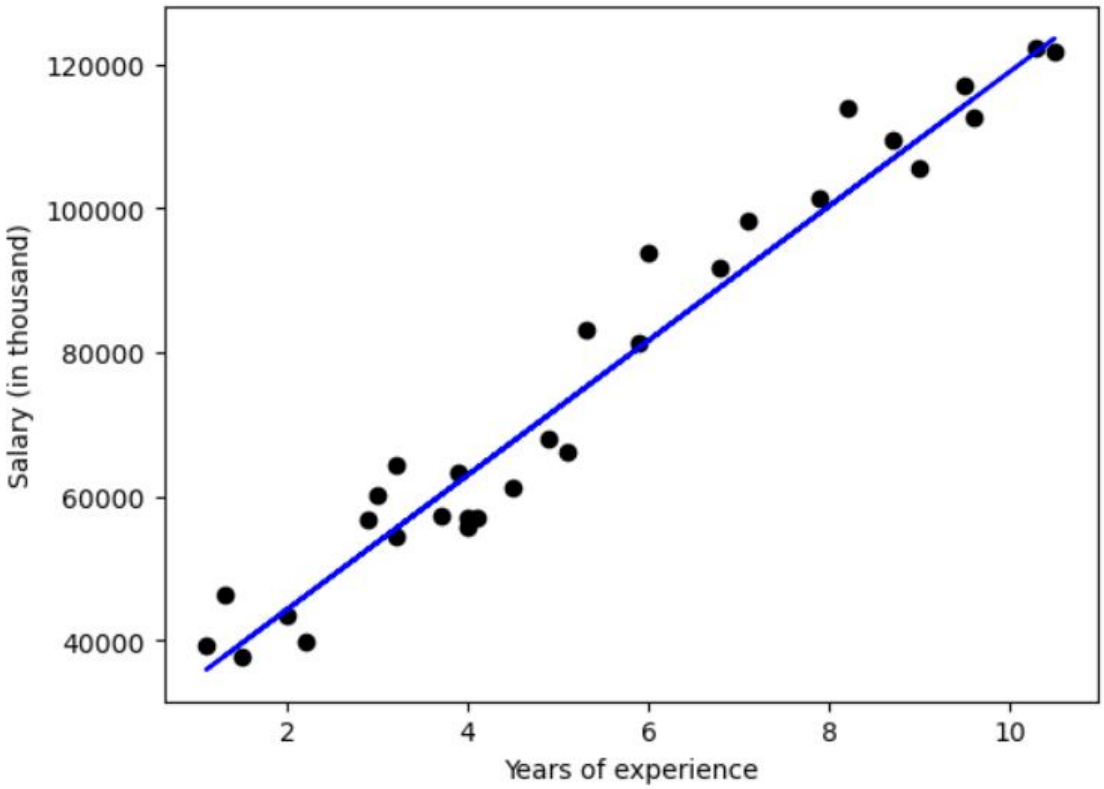

Fig.9 Result for regression.

Now we will check the accuracy/efficiency of our model that how well did our model perform. Actually there are several method for doing that.

RMSE (Root Mean Squared Error) is a commonly used metric in machine learning for evaluating the performance of regression models. RMSE is similar to MSE (Mean Squared Error), but is in the same units as the target variable, making it more interpretable.

RMSE is calculated as the square root of the average of the squared differences between the predicted values and the true values. Mathematically, it can be represented as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - y_i)^2}$$

The result is 19.899.

The classification task for types of mushrooms.

The dataset is available on https://www.kaggle.com/. The following description is used:

Context

Although this dataset was originally contributed to the UCI Machine Learning repository nearly 30 years ago, mushroom hunting (otherwise known as "shrooming") is enjoying new peaks in popularity. Learn which features spell certain death and which are most palatable in this dataset of mushroom characteristics. And how certain can your model be?

Content

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be" for Poisonous Oak and Ivy.

The head function gives us the following interpretation of top features of this dataset.

	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	population
0	p	x	s	n	t	p	f	c	n	k	...	s	w	w	p	w	o	p	k	s
1	e	x	s	y	t	a	f	c	b	k	...	s	w	w	p	w	o	p	n	n
2	e	b	s	w	t	l	f	c	b	n	...	s	w	w	p	w	o	p	n	n
3	p	x	y	w	t	p	f	c	n	n	...	s	w	w	p	w	o	p	k	s
4	e	x	s	g	f	n	f	w	b	k	...	s	w	w	p	w	o	e	n	a

Fig.10 Mushrooms dataset.

```
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   class                     8124 non-null   object
 1   cap-shape                 8124 non-null   object
 2   cap-surface               8124 non-null   object
 3   cap-color                 8124 non-null   object
 4   bruises                   8124 non-null   object
 5   odor                      8124 non-null   object
 6   gill-attachment           8124 non-null   object
 7   gill-spacing              8124 non-null   object
 8   gill-size                 8124 non-null   object
 9   gill-color                8124 non-null   object
 10  stalk-shape               8124 non-null   object
 11  stalk-root                8124 non-null   object
 12  stalk-surface-above-ring  8124 non-null   object
 13  stalk-surface-below-ring  8124 non-null   object
 14  stalk-color-above-ring    8124 non-null   object
 15  stalk-color-below-ring    8124 non-null   object
 16  veil-type                 8124 non-null   object
 17  veil-color                8124 non-null   object
 18  ring-number               8124 non-null   object
 19  ring-type                 8124 non-null   object
 20  spore-print-color         8124 non-null   object
 21  population                8124 non-null   object
 22  habitat                   8124 non-null   object
dtypes: object(23)
```

Fig.11 Mushrooms dataset datatypes.

For these types of values we do perform split into train and test datasets. To achieve the classification goal we use decision tree algorithm. For criterion we have chosen gini.

On figure 12 the classification report is given to perform the results.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	799
1	1.00	1.00	1.00	826
accuracy			1.00	1625
macro avg	1.00	1.00	1.00	1625
weighted avg	1.00	1.00	1.00	1625

Fig.12 Classification report