



Bài giảng môn học: _____
Học Máy (Machine Learning)

Chương 2: Quy trình xây dựng một hệ thống học máy – Phần 2

Đặng Văn Nam
dangvannam@hmg.edu.vn

Nội dung chương 2 – Phần 2

1. Các bước cơ bản xây dựng một mô hình học máy
2. Một số nguồn dữ liệu cho học tập
3. Thu thập và tiền xử lý dữ liệu
 1. Ví dụ tập Data_Patient.csv
 2. Ví dụ tập Data_Titanic.csv
4. Bài tập chương 2

2. Một số nguồn dữ liệu cho học tập

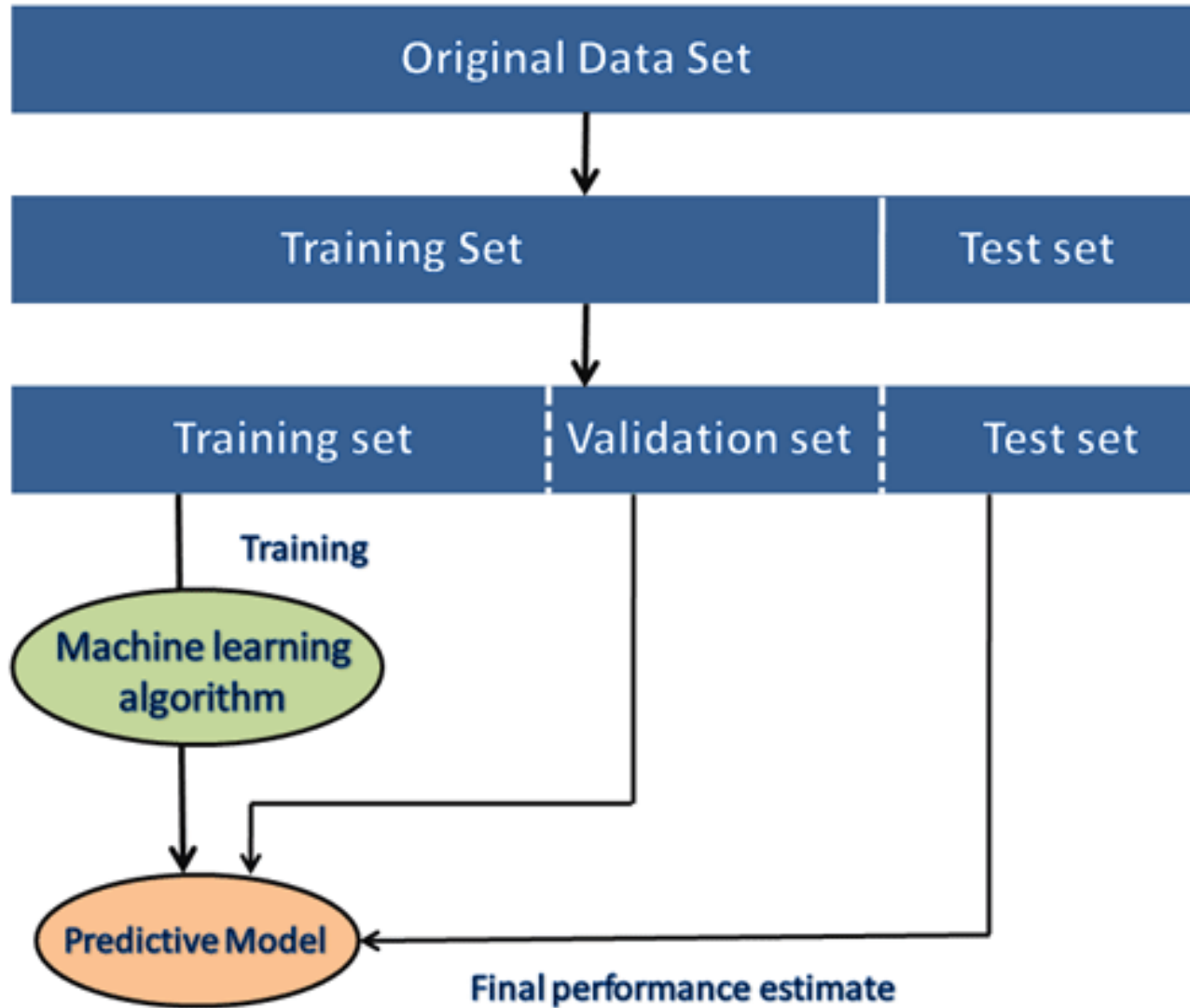
1. Datasets

Dữ liệu có vai trò quan trọng trong việc xây dựng ứng dụng học máy cho bất kỳ bài toán nào.

Chất lượng và **khối lượng** dữ liệu ảnh hưởng trực tiếp đến độ chính xác của mô hình học máy.



Datasets (t)

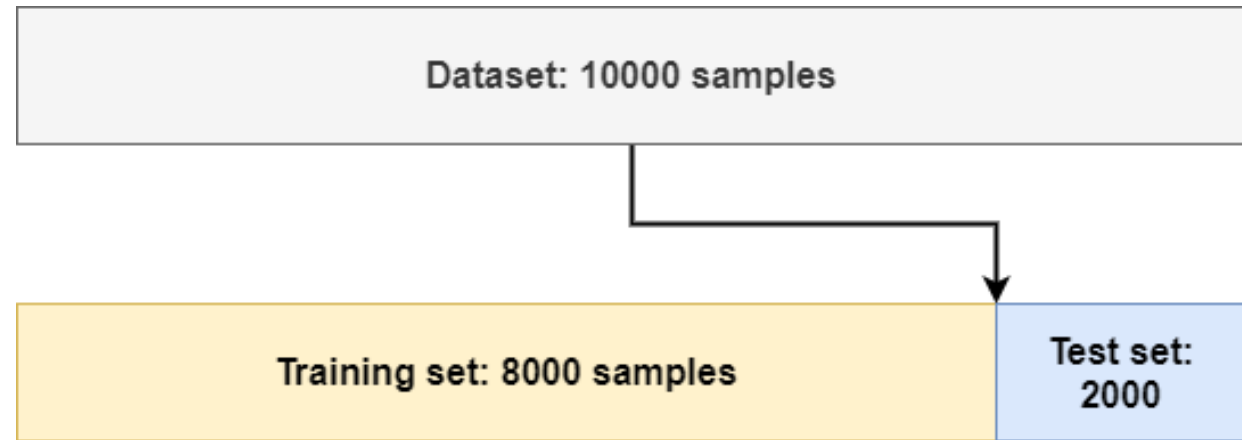
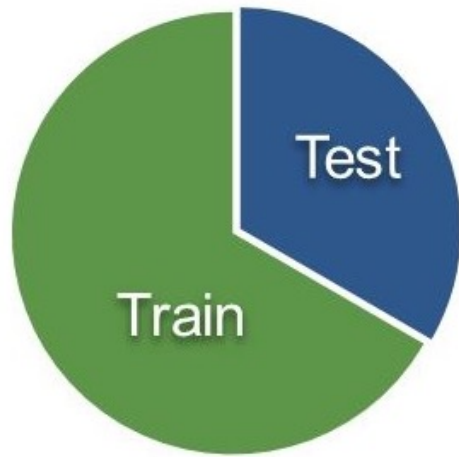


Dữ liệu từ các nguồn sau khi được tổng hợp và xử lý sẽ thu được các tập dữ liệu – **Datasets** phục vụ cho việc xây dựng model.

1. Tập huấn luyện (Training Set)
2. Tập kiểm tra (Test Set)
3. Tập kiểm chéo (Validation Set)

Datasets (t)

- **Tập huấn luyện (Training Set)** bao gồm các điểm dữ liệu sử dụng trực tiếp trong việc xây dựng mô hình.
- **Tập kiểm tra (Test set)** gồm các dữ liệu được dùng để đánh giá hiệu quả của mô hình. Tập kiểm tra đại diện cho dữ liệu mà mô hình chưa từng thấy, có thể xuất hiện trong quá trình vận hành mô hình trên thực tế.

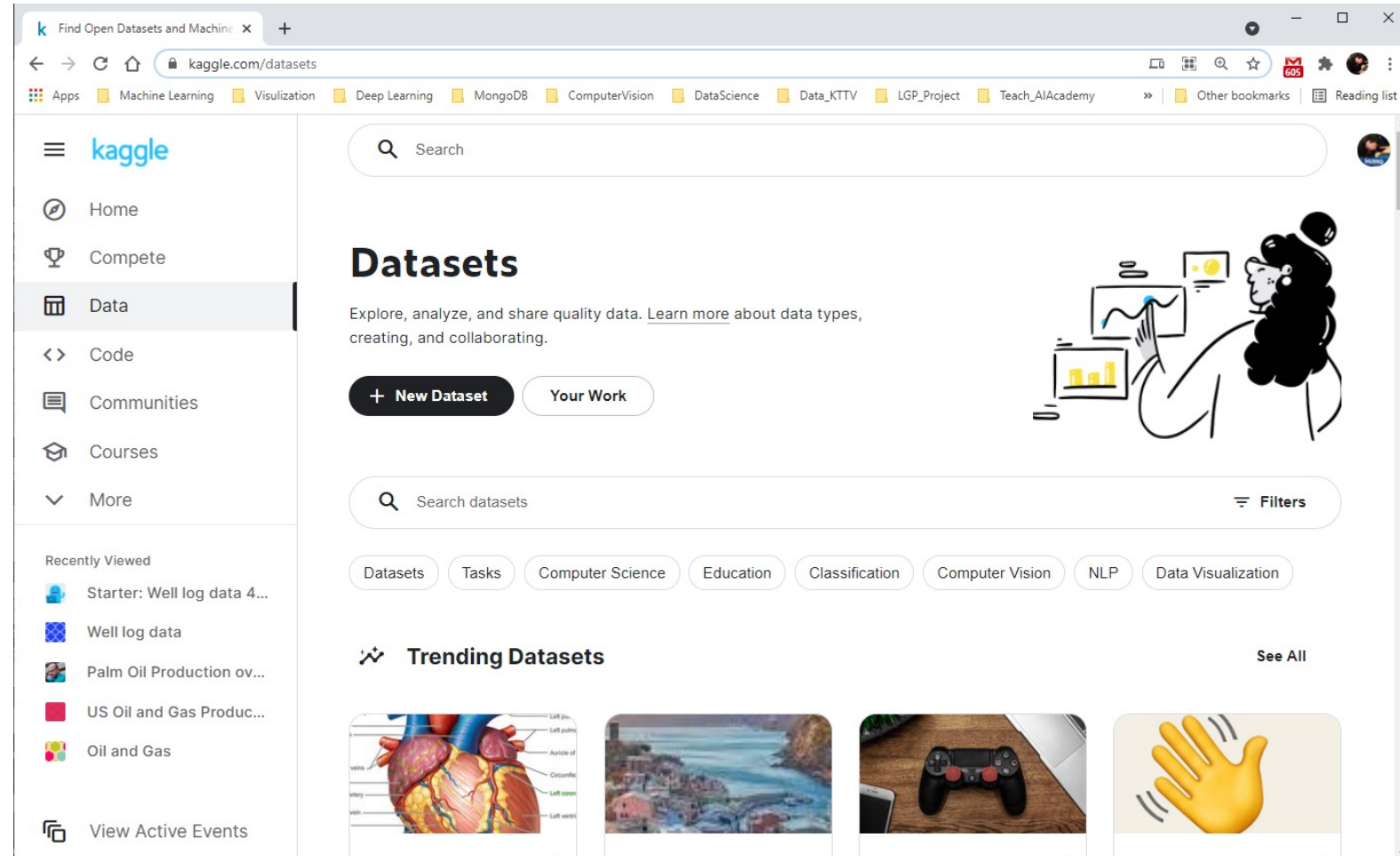


- Để đảm bảo tính phổ quát, dữ liệu kiểm tra không được sử dụng trong quá trình xây dựng mô hình.
- Điều kiện cần để một mô hình hiệu quả: **Kết quả đánh giá trên tập huấn luyện và tập kiểm tra đều cao.**

2. Một số nguồn Dataset cho ML

Kaggle

- Kaggle là một trong những nguồn cung cấp dữ liệu tốt nhất cho các nhà khoa học dữ liệu và những người học về ML.
- <https://www.kaggle.com/datasets>.



Một số nguồn Dataset cho ML (t)

UCI Machine Learning Repository

- UCI là một nguồn cung cấp Dataset tuyệt vời cho việc xây dựng các model học máy. Ra đời năm 1987, UCI được các sinh viên, giáo sư, nhà nghiên cứu sử dụng rộng rãi
- <https://archive.ics.uci.edu/ml/index.php>.

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)






Search

☐ Repository ☐ Web

[View ALL Data Sets](#)

Browse Through: 488 Data Sets

[Table View](#) [List View](#)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
 Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
 Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998

Một số nguồn Dataset cho ML (t)

Datasets via AWS

Registry of Open Data on AWS



About

This registry exists to help people discover and share datasets that are available via AWS resources. [Learn more about sharing data on AWS.](#)

See [all usage examples](#) for datasets listed in this registry.

See datasets from [Facebook Data for Good](#), [NASA Space Act Agreement](#), [NIH STRIDES](#), [NOAA Big Data Program](#), [Space Telescope Science Institute](#), and [Amazon Sustainability Data Initiative](#).

Search datasets (currently 190 matching datasets)

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the [Registry of Open Data on AWS GitHub repository](#).

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

- <https://registry.opendata.aws/>.

Một số nguồn Dataset cho ML (t)

Google's Dataset Search Engine

The screenshot shows the Google Dataset Search interface. At the top, the search bar contains the word "classification". Below the search bar, there are filters for "Updated Date", "Download Format", "Usage Rights", and "Free". The results section shows "100+ results found". The first result is a dataset titled "Classification" with a circular icon containing the letter "D". It lists sources: "catalog.data.gov", "data.nasa.gov", and "+1more", and notes it was "Updated May 2, 2019". To the right of this result, there are three blue buttons: "Explore at catalog.data.gov", "Explore at Rally - Open Data Portal", and "Explore at data.wu.ac.at". Below the first result, there are two more results from Kaggle: "Mushroom Classification" (updated Dec 1, 2016) and "Question Classification". On the right side of the interface, there is a detailed view for the "Classification" dataset, showing it was updated on May 2, 2019, provided by Dashlink, and a description of supervised learning.

Google Dataset Search

classification

Updated Date Download Format Usage Rights Free

100+ results found

Classification
catalog.data.gov
data.nasa.gov
+1more
Updated May 2, 2019

Classification

Explore at catalog.data.gov

Explore at Rally - Open Data Portal

Explore at data.wu.ac.at

Dataset updated May 2, 2019

Dataset provided by
Dashlink

Description

A supervised learning task involves constructing a mapping from an input data space (normally described by several features) to an output space. A set of training examples—examples with known output values—is used by a learning

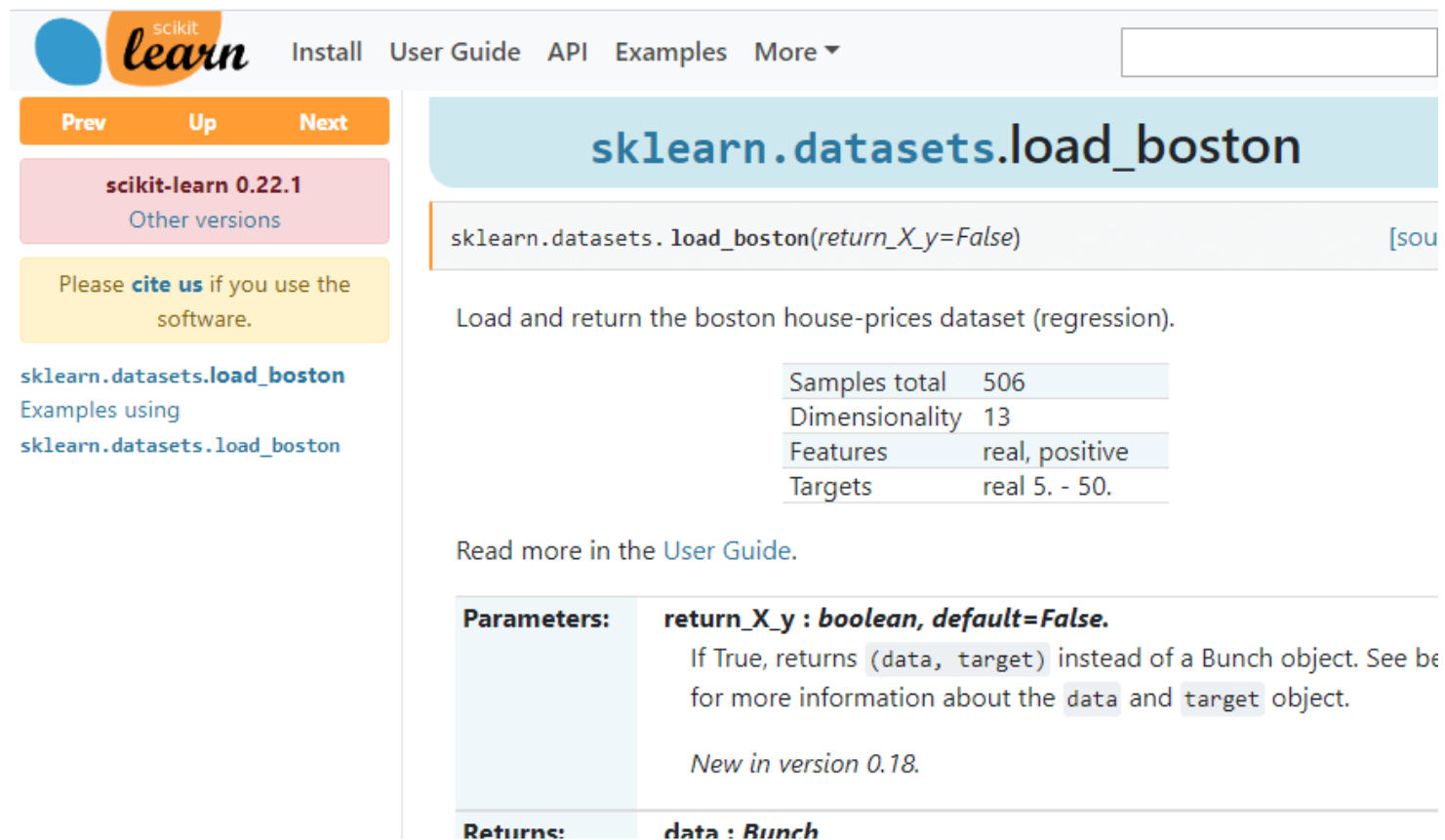
kaggle Mushroom Classification
www.kaggle.com
Updated Dec 1, 2016

kaggle Question Classification
www.kaggle.com

- <https://toolbox.google.com/datasetsearch>.

Một số nguồn Dataset cho ML (t)

Scikit-learn dataset



The screenshot shows the Scikit-learn website's documentation for the `sklearn.datasets.load_boston` function. The page includes navigation links (Prev, Up, Next), the current version (0.22.1), and a sidebar with links to the function name and examples. The main content area displays the function signature `sklearn.datasets.load_boston(return_X_y=False)`, a description of the dataset, a table of its characteristics, and details about its parameters and return value.

sklearn.datasets.load_boston

```
sklearn.datasets.load_boston(return_X_y=False)
```

Load and return the boston house-prices dataset (regression).

Samples total	506
Dimensionality	13
Features	real, positive
Targets	real 5. - 50.

Read more in the [User Guide](#).

Parameters: **return_X_y** : *boolean, default=False.*
If True, returns (data, target) instead of a Bunch object. See below for more information about the data and target object.
New in version 0.18.

Returns: **data** : *Bunch*

- <https://scikit-learn.org/stable/datasets/index.html>.

(Các ví dụ trong việc xây dựng model trong môn học sẽ chủ yếu lấy từ nguồn này)

3. Thu thập và tiền xử lý dữ liệu

1. Giới thiệu



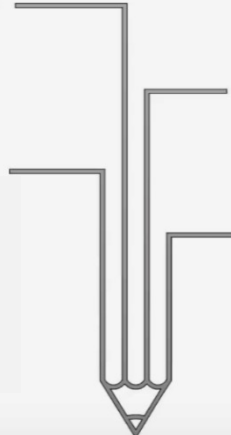
- Thu thập và chuẩn bị dữ liệu là bước đầu tiên và quan trọng trong bất kỳ một dự án học máy nào.
- Là bước **quan trọng**, chiếm **nhều thời gian** và **nguồn lực** nhất trong bất kỳ một dự án nào (80%)

Làm sạch dữ liệu

Chỉnh sửa dữ liệu bằng cách bổ sung các dữ liệu còn thiếu, thay thế và hiệu chỉnh các dữ liệu nhiễu

Giảm kích thước dữ liệu

Đảm bảo chất lượng ban đầu của dữ liệu

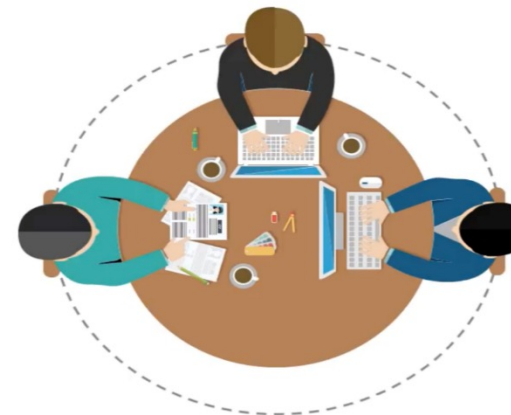


Chuyển đổi dữ liệu

Bao gồm chuẩn hóa, chuyển đổi và tổng hợp dữ liệu sử dụng các phương pháp ETL.

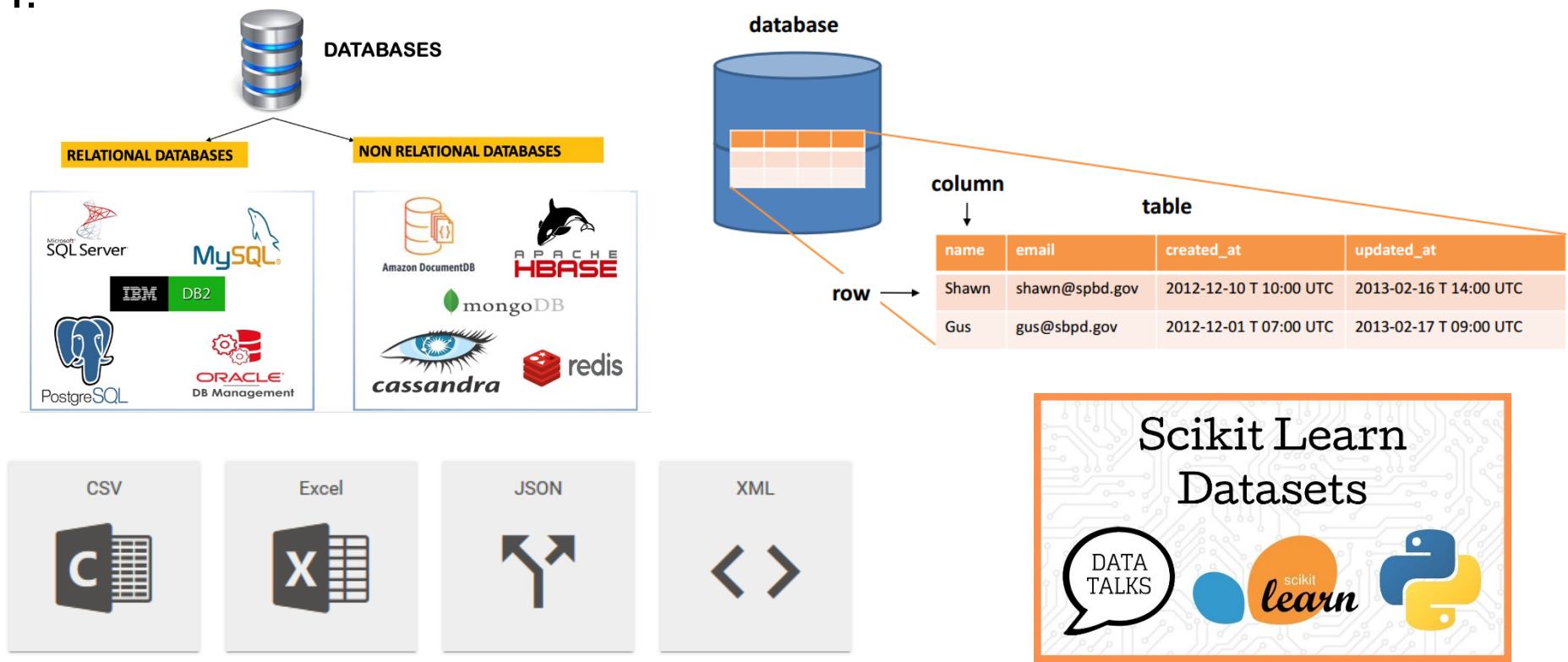
Tích hợp dữ liệu

Xử lý sự không tương thích giữa các dữ liệu



1. Giới thiệu

Dữ liệu tồn tại trong rất nhiều dạng khác nhau, từ dữ liệu trong các file cơ bản như Excel, CSV, text, Json, XML...hay lưu trữ trong các CSDL. Dữ liệu còn có thể được cung cấp thông qua các API.



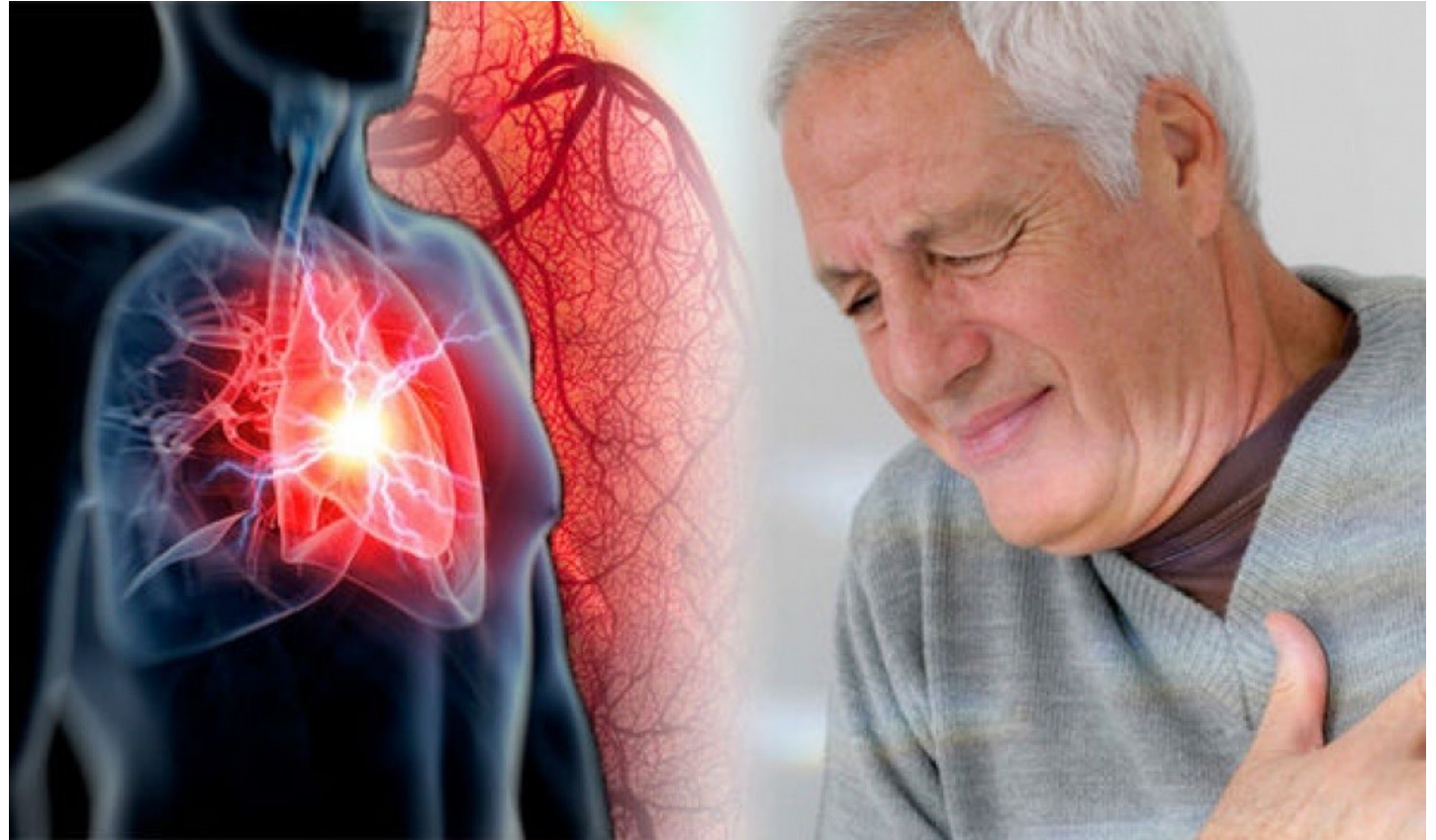
Trong bài học này chúng ta sẽ tìm hiểu cách đọc dữ liệu từ một số dạng cơ bản và phổ biến nhất.

Một số kỹ thuật xử lý dữ liệu cho ML

1. Tập dữ liệu Data_Patient

Một số kỹ thuật tiền xử lý dữ liệu: Data_Patient.csv

(Learning-Project based) Áp dụng cho tập dữ liệu Patient: Dự đoán bệnh đau tim



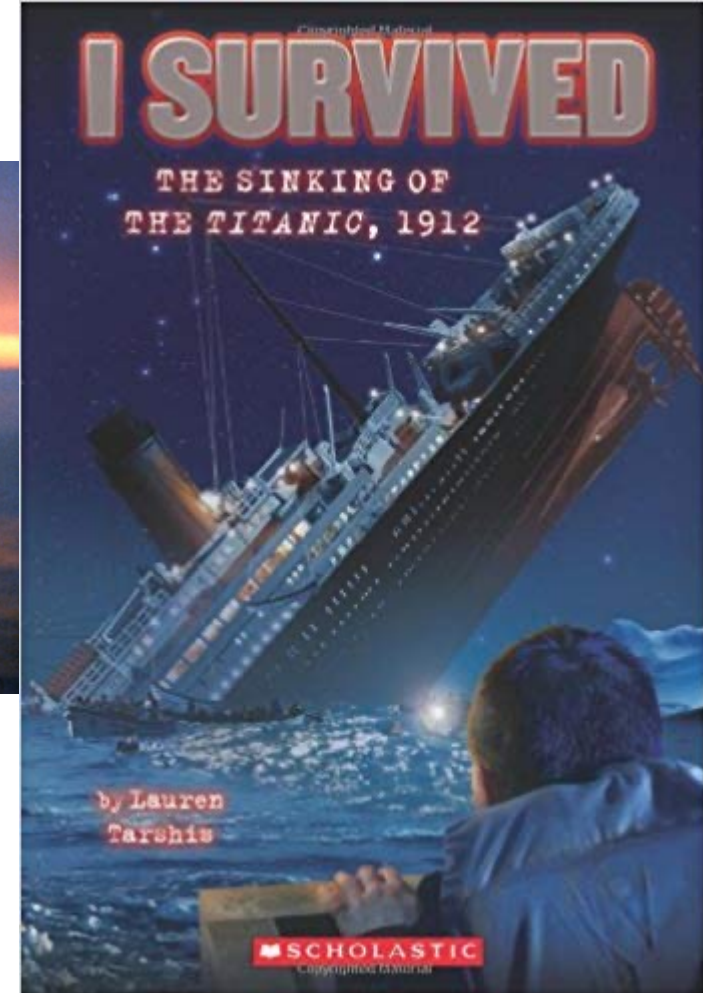
- Sinh viên xem lại phần chuẩn bị dữ liệu trong ví dụ chi tiết các bước xây dựng mô hình học máy với tập dữ liệu 300 bệnh nhân...

2. Tập dữ liệu Data_Titanic

Một số kỹ thuật tiền xử lý dữ liệu: Data_Titanic.csv

(Learning-Project based) Áp dụng cho tập dữ liệu TITANIC

- Thảm họa đắm tàu Titanic là thảm họa hàng hải lớn nhất trong lịch sử. Chuyến ra khơi đầu tiên của con tàu vào ngày 15/4/1912, Titanic đã bị chìm sau khi va vào một tảng băng trôi, làm chết 1502 trong tổng số 2224 hành khách và thủy thủ đoàn.
- Một số người may mắn thoát chết như: Phụ nữ, trẻ em và những người thuộc tầng lớp thượng lưu.

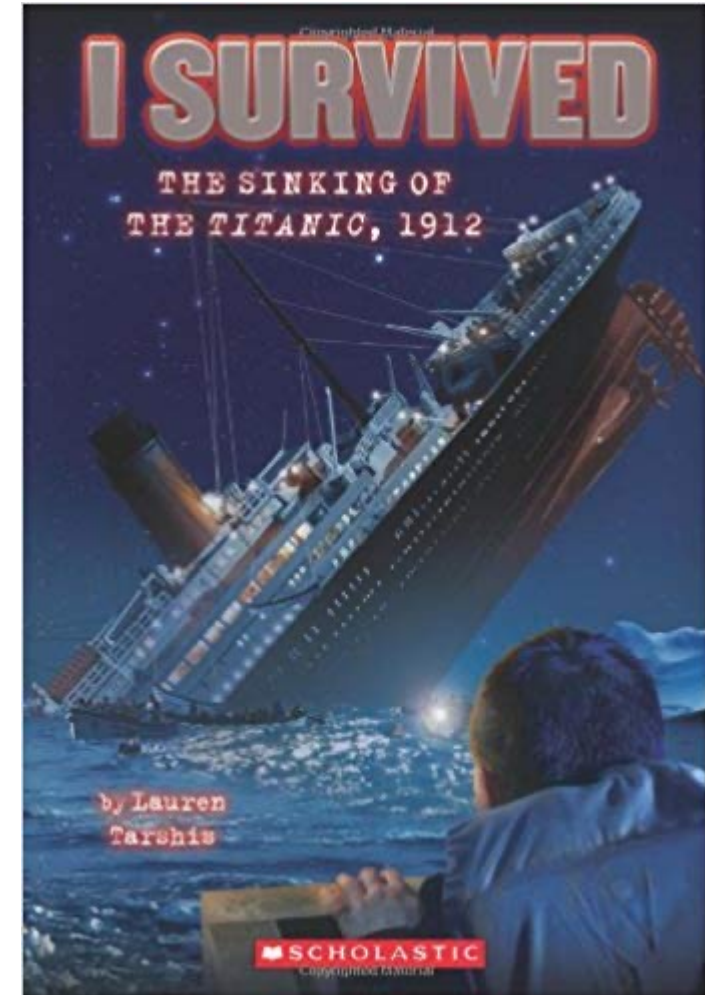


Một số kỹ thuật tiền xử lý dữ liệu: Data_Titanic.csv

- Cần xây dựng một mô hình học máy có khả năng dự đoán một hành khách với các thông số liên quan được cứu hay không được cứu nếu xảy ra tai nạn tương tự!

Titanic Survival Prediction With Python

Machine Learning Project



Một số kỹ thuật tiền xử lý dữ liệu

Thu thập dữ liệu: **Data_Titanic.csv**

Bao gồm:

- 1309 dòng (ứng với thông tin của 1309 hành khách)
- Mỗi dòng có 12 cột (tương ứng với 12 thuộc tính của hành khách)

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Eli	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

1) **PassengerId**: Thuộc tính này cho biết Mã (ID) của hành khách trên tàu (Kiểu dữ liệu: Số).

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (El	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

2) **Survived**: Thuộc tính cho hành khách ngày có được cứu hay không? 0 (No – Không được cứu) | 1 (Yes – Được cứu) (Kiểu dữ liệu: Số).

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (El	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

3) **Pclass**: Loại vé tàu của hành (cho biết địa vị xã hội) 1 – Hạng nhất | 2 – Hạng 2 | 3 – Hạng ba (Kiểu dữ liệu: Số).

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Eli	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D. Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

4) Name: Tên của hành khách (Kiểu dữ liệu: chuỗi).

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradle	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heat	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (El	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William I	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

5) **Sex:** Giới tính của hành khách: male – Nam | female – Nữ (Kiểu dữ liệu: chuỗi).

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradle	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heat	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (El	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William I	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

6) Age: Tuổi của hành khách đi tàu (Kiểu dữ liệu: số)

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cummings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (El	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

7) SibSp (Siblings/Spouses): Số lượng anh chị em| vợ chồng của hành khách cùng trên tàu(Kiểu dữ liệu: số)

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Eli	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Tập dữ liệu huấn luyện:

8) Parch(Parents/Children): Số lượng bố mẹ| con cái của hành khách cùng trên tàu(Kiểu dữ liệu: số)

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Eli	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

9) Ticket: Số vé của hành khách (Kiểu dữ liệu: chuỗi kết hợp ký tự và số)

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (El	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

10) Fare: Giá vé đi tàu của hành khách (Kiểu dữ liệu: Số)

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Eli	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

11) Cabin: Số hiệu cabin trên tàu của hành khách (Kiểu dữ liệu: chuỗi bao gồm ký tự và số)

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Eli	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett Mrs. (Mary D Kin	female	55	0	0	248706	16		S

Dữ liệu Titanic

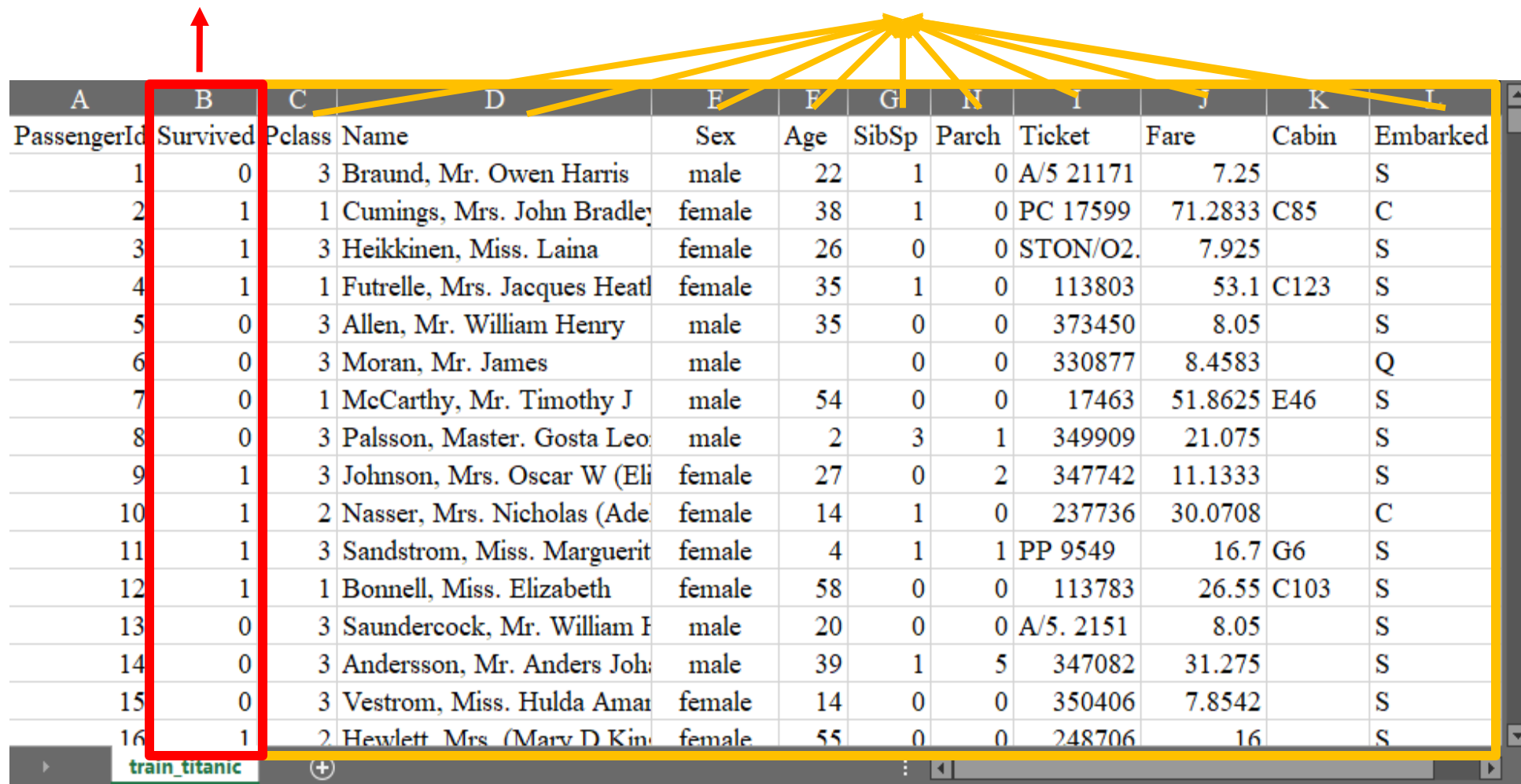
12) Embarked: Ký hiệu cho biết cảng mà hành khách lên tàu, Tàu Titanic đón khách ở 3 cảng: C = Cherbourg | Q = Queenstown | S =Southampton (Kiểu dữ liệu: ký tự)

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Eli	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett Mrs. (Mary D King	female	55	0	0	248706	16		S

Dữ liệu Titanic

Grown truth
(Target)

Feature
(input)



A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heatl	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Eli	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William F	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D. Kin	female	55	0	0	248706	16		S

Một số kỹ thuật tiền xử lý dữ liệu

Raw Data												
	A	B	C	D	E	F	G	H	I			
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
5	4	1	1	Futrelle, Mrs. Jacques Hea	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. James	male			0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Master. Gosta Le	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs. Oscar W (E	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs. Nicholas (Ac	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, Miss. Marguer	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saundercock, Mr. William	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson, Mr. Anders Jo	male	39	1	5	347082	31.275		S
16	15	0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, Mrs. (Mary D Ki	female	55	0	0	248706	16		S
18	17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
19	18	1	2	Williams, Mr. Charles Euge	male			0	244373	13		S
20	19	0	3	Vander Planke, Mrs. Julius	female	31	1	0	345763	18		S

Data Titanic

Chi tiết các bước tiền xử lý dữ liệu trong file jupyter notebook.

Data Titanic

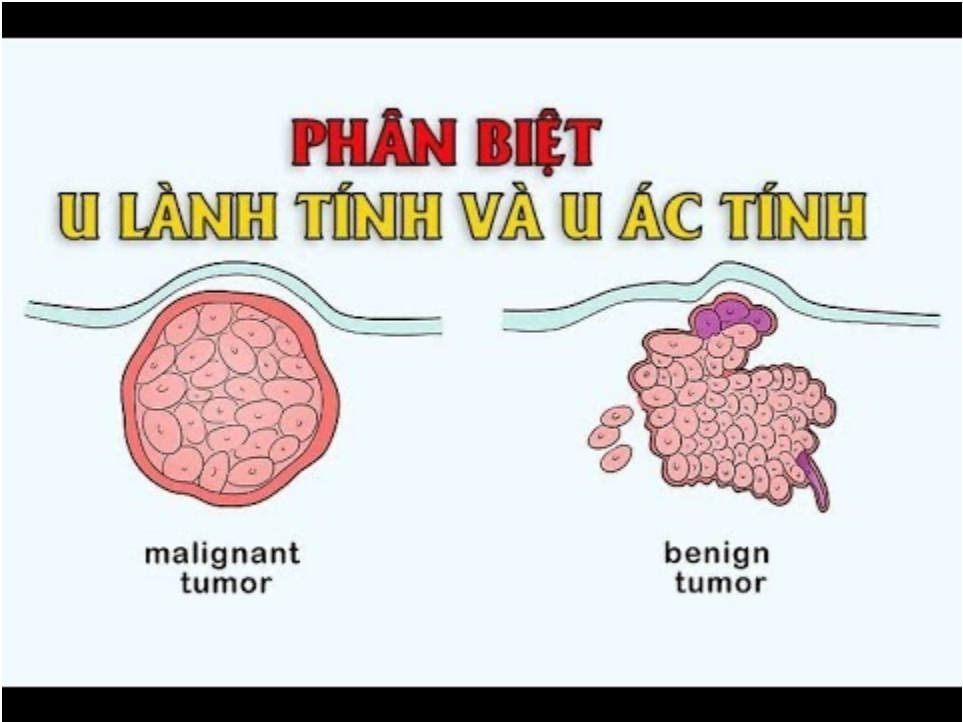


							New Data	
	A	B	C	D	E			
1	Survived	Pclass	Sex	Age	SibSp	Parch	Embarked	
2	0	3	0	1	1	0	0	
3	1	1	1	2	1	0	1	
4	1	3	1	1	0	0	0	
5	1	1	1	2	1	0	0	
6	0	3	0	2	0	0	0	
7	0	3	0	1	0	0	2	
8	0	1	0	3	0	0	0	
9	0	3	0	0	3	1	0	
10	1	3	1	1	0	2	0	
11	1	2	1	0	1	0	1	
12	1	3	1	0	1	1	0	
13	1	1	1	3	0	0	0	
14	0	3	0	1	0	0	0	
15	0	3	0	2	1	5	0	

Data_Titanic_ok

BÀI TẬP CHƯƠNG 2

Tập dữ liệu Data_Practice_ML.xlsx chứa dữ liệu của các bệnh nhân bị u vú

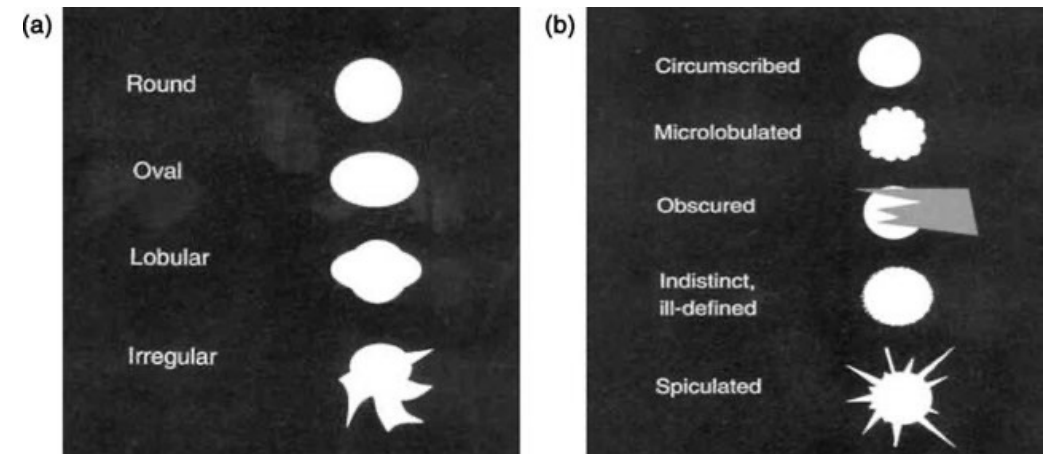


PatientID	Sex	Age	weight	Shape	Margin	Density	Target
P1	Female	67	55	Lobular	Spiculated	Low	1
P2	Female	43	66	Round	Circumscribed		1
P3	Female	58	66	Irregular	Spiculated	Low	1
P4	Female	28	46	Round	Circumscribed	Low	0
P5	Female	74	51	Round	Spiculated		1
P6	Female	65	71	Round		Low	0
P7	Female	70	75			Low	0
P8	Female	42	65	Round		Low	0
P9	Female	57	77	Round	Spiculated	Low	1
P10	Female	60	62		Spiculated	High	1
P11	Female	76	52	Round	ill-defined	Low	1
P12	Female	42	61	Oval	Circumscribed	Low	1
P13	Female	64	76	Round		Low	0
P14	Female	36	69	Lobular	Circumscribed	Iso	0
P15	Female	60	63	Oval	Circumscribed	Iso	0
P16	Female	54	55	Round	Circumscribed	Low	0
P17	Female	52	79	Lobular	ill-defined	Low	0

Thực hành

Mỗi bản ghi tương ứng với một bệnh nhân, Bao gồm các thuộc tính:

1. **PatientID:** Thuộc tính cho biết mã số của bệnh nhân
2. **Sex:** Thuộc tính cho biết giới tính bệnh nhân
3. **Age:** Thuộc tính cho biết tuổi của bệnh nhân, dữ liệu số
4. **Weight:** Thuộc tính cho biết cân nặng của bệnh nhân (Kg)
5. **Shape:** Thuộc tính cho biết hình dạng của khối u, bao gồm 4 giá trị: **Round, Oval, Lobular, Irregular**
6. **Margin:** Thuộc tính cho biết dạng đường biên của khối u, bao gồm 5 giá trị: **Circumscribed, Microlobulated, Obscured, ill-defined, Spiculated**
7. **Density:** Thuộc tính cho biết mật độ của khối u, bao gồm 4 giá trị: **High, Iso, Low, Fat-containing**
8. **Target:** Thuộc tính cho biết khối u là lành tính **(0)** - hay ác tính **(1)**



Thực hành

Yêu cầu 1:

Đọc tập dữ liệu Data_Practice_ML.xlsx vào biến DataFrame,

- Hiển thị thông tin của biến,
- Hiển thị dữ liệu 5 bản ghi đầu tiên; 5 bản ghi cuối cùng, 5 bản ghi ngẫu nhiên
- Thống kê dữ liệu các thuộc tính số, các thuộc tính Object; đưa ra các nhận xét về dữ liệu; Có thể sử dụng các biểu đồ để thể hiện trực quan

Thực hành

Yêu cầu 2:

- Kiểm tra các bản ghi trùng lặp, các thông số bất thường trong tập dữ liệu nếu có
- Thống kê dữ liệu thiếu (missing) cho từng thuộc tính, và liệt kê ra các bản ghi bị missing tương ứng với thuộc tính đó.

Yêu cầu 3:

Phân tích, thống kê - xác định mức độ ảnh hưởng của các thuộc tính độc lập [Age, Shape, Margin, Density] tới thuộc tính phụ thuộc [Target]

- Xác định thuộc tính quan trọng, không quan trọng ảnh hưởng việc u lành tính hay ác tính của bệnh nhân

Thực hành

Yêu cầu 4:

- Đề xuất và Áp dụng các kỹ chuẩn hóa, tiền xử lý dữ liệu phù hợp cho tập dữ liệu này để có thể đưa vào các mô hình học máy và
- Lưu kết quả dữ liệu sau xử lý ra file **Data_Practice_ML_OK.csv**



LƯU Ý: Sinh viên trình bày phần bài tập thực hành như trong file ví dụ mẫu với dữ liệu Data_Patient, Data_Titanic; Sử dụng các cell Markdown để mô tả, giải thích....



Thank you!