Nhóm 3:

Nguyễn Thùy Linh - 17520689

Lý Hồng Thiên Ân - 17520210

Phạm Viết Lưu - 17520730

Bài tập 1:

HIỂU DỮ LIỆU

1. Abalone Data Set

- Tập dữ liệu thu thập được để dự đoán tuổi của bào ngư (*abalone*) từ các phép đo vật lý.
- Có tất cả 4177 mẫu trong tập dữ liệu.
- Mỗi mẫu có 8 thuộc tính input và 1 thuộc tính output.

STT	Tên thuộc tính	Kiểu dữ liệu	Trung bình/ Số giá trị phân biệt	Phương sai/ Số giá trị duy nhất	Số mẫu bị thiếu
1	sex	nominal	3	0	0
2	length	numeric	0.524	0.0144	0
3	diameter	numeric	0.408	0.009801	0
4	height	numeric	0.14	0.001764	0
5	whole weight	numeric	0.829	0.2401	0
6	shucked weight	numeric	0.359	0.049284	0
7	viscera weight	numeric	0.181	0.0121	0
8	shell weight	numeric	0.239	0.019321	0
9	rings	numeric	9.934	10.394176	0

2. Bank Marketing Data Set

- Tập dữ liệu được thu thập dùng để xây dựng mô hình dự đoán một khách hàng có tham gia gửi tiền có kỳ hạn vào ngân hàng hay không.
- Có tất cả 41188 mẫu trong tập dữ liệu.
- Mỗi mẫu có 20 thuộc tính input và 1 thuộc tính output.

STT	Tên thuộc tính	Kiểu dữ liệu	Trung bình/ Số giá trị phân biệt	Phương sai/ Số giá trị duy nhất	Số mẫu bị thiếu
1	age	numeric	40.0241	108.601	0
2	job	nominal	12	0	0
3	marital	nominal	4	0	0
4	education	nominal	8	0	0
5	default	nominal	3	0	0
6	housing	nominal	3	0	0
7	loan	nominal	3	0	0
8	contact	nominal	2	0	0

9	month	nominal	12	0	0
10	day_of_week	nominal	7	0	0
11	duration	numeric	258.285	67225.6	0
12	campaign	numeric	2.56759	7.67296	0
13	pdays	numeric	962.475	34935.72192	0
14	previous	numeric	0.172963	0.24493	0
15	poutcome	nominal	3	0	0
16	emp.var.rate	numeric	0.0818855	2.46792	0
17	cons.price.idx	numeric	93.5757	0.33506	0
18	cons.conf.idx	numeric	-40.5026	21.42024	0
19	euribor3m	numeric	3.62129	3.0083	0
20	nr.employed	numeric	5167.04	5220.2793	0
21	y	nominal	2	0	0

3. Metro Traffic Volume Data Set

- Tập dữ liệu thu thập để xác định lưu lượng giao thông của tiểu bang Metro.
- Có tất cả 48204 mẫu trong tập dữ liệu.
- Mỗi mẫu có 8 thuộc tính input và 1 thuộc tính output.

STT	**Tên thuộc tính**	**Kiểu dữ liệu**	**Trung bình/ Số giá trị phân biệt**	**Phương sai/ Số giá trị duy nhất**	**Số mẫu bị thiếu**
1	holiday	nominal	12	0	0
2	temp	numeric	281.206	177.9	0
3	rain_1h	numeric	0.334	2006.05	0
4	snow_1h	numeric	0	0.000064	0
5	clouds_all	numeric	49.362	1522.248	0
6	weather_main	nominal	11	0	0
7	weather_description	nominal	38	1	0
8	date_time	nominal	40575	35130	0
9	traffic_volume	numeric	3259.818	3947616.63	0