Machine learning
Email Classification

Group 1:

Members:

1. Đàm Tuấn Anh
2. Phạm Thị Quỳnh
3. Đỗ Thị Diệu Thúy
4. Nguyễn Thị Tươi

Group 1: Email Classification

Kind of methods using:

1. Naive Bayes Model
2. Logistic Regression Model
3. The Perceptron

Data Processing:

- Replacing link with “URL_change”, date with “Date_change”,
- Deleting Numbers, words with 1 size
- Combining words into meaningfull Phrases

1. Naive Bayes Model

- **Data processing:**
  - Training data: 80 %
  - Testing data: 20 %

The number of words in dictionary				
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1888	1967	1875	1867	1929

- **Accuracy:**

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Traning	Group	51.76	51.13	54.34	48.55	52.65	51.69
	Library	98.39	91.32	89.07	91.32	90.34	90.29
Testing	Group	51.25	55.00	40.00	65.00	46.49	51.55
	Library	67.50	67.50	62.50	56.25	66.20	63.99

2. Logistic Regression Model

- **Data processing:**
  - Training data: 80 %
  - Testing data: 20 %

The number of words in dictionary				
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
2617	2737	2592	2580	2716

- **Accuracy:**

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Traning	Group	100.00	100.00	100.00	100.00	100.00	100.00
	Library	100.00	100.00	100.00	100.00	100.00	100.00
Testing	Group	80.00	88.75	86.25	85.00	90.14	86.03
	Library	85.00	95.00	87.50	90.00	90.14	89.53

3. The perceptron

- **Data processing:**
  - Training data: 80 %
  - Testing data: 20 %

The number of words in dictionary				
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
2617	2737	2592	2580	2716

- **Accuracy:**

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Traning	Group						
	Library	98.07	99.36	99.04	98.39	100.00	98.97
Testing	Group						
	Library	77.46	77.46	77.46	77.46	77.46	77.46

Conclusion

Thank for working hard of members:

1. Data processing: Tuoi, Tuan Anh, Thuy, Quynh.
2. Model coding :
    - Naive Bayes: Quynh
    - Logistic Regrestion: Tuoi, Quynh.
    - The Perceptron: Quynh, Tuoi, Tuan Anh
3. Model tesing: Thuy, Tuoi
4. Slice making: Quynh
5. Presentation: Tuan Anh
6. Following our github project: https://github.com/linhadam/machine_learning

Thanks for your attention!

Score

Work	100%
Đàm Tuấn Anh	25%
Phạm Thị Quỳnh	25%
Đỗ Thị Diệu Thúy	25%
Nguyễn Thị Tươi	25%