Statistical Learning
Final Project

Global Websites Traffic Analysis

M102040004 柯鋅瑀
M102040016 李翊瑄

Outline

- 1. Data introduction
- 2. Data Pre-Processing
- 3. Goal setting
- 4. Methods
- 5. Result and insight
- 6 . Reference

1. Data introduction

This dataset is collected from Kaggle. It contains Top 50 ranked sites from 191 countries as on 25th May 2017 and is subject to change in future time due to the dynamic structure of ranking.

Country_Rank	Website	Trustworthiness	Avg_Daily_Visitc	Child_Safety	Avg_Daily_Page	Privacy	Facebook_likes	Twitter_mentions	Google_pluses
1	www.google.com.af	Excellent	N/A	Excellent	N/A	Excellent	9	1	37
2	www.google.com	Excellent	515 007 350	Excellent	4 192 159 833	Excellent	94.2K	11.2K	11.7M
3	www.youtube.com	Excellent	506 457 282	Excellent	2 679 159 025	Excellent	13.5K	16.5K	19.3M
4	www.facebook.com	Excellent	270 071 255	Good	1 082 985 733	Excellent	5.87M	64.4K	127K
5	www.yahoo.com	Excellent	99 572 035	Excellent	383 352 336	Excellent	17.2K	1.11K	798K
6	www.acbar.org	Unknown	100 388	Unknown	712 760	Unknown	-	12	12
7	www.bbc.com	Excellent	9 282 040	Excellent	24 690 228	Excellent	9	9.37K	7.2K
8	www.wikipedia.org	Excellent	118 921 355	Excellent	397 197 324	Excellent	476	162	126K

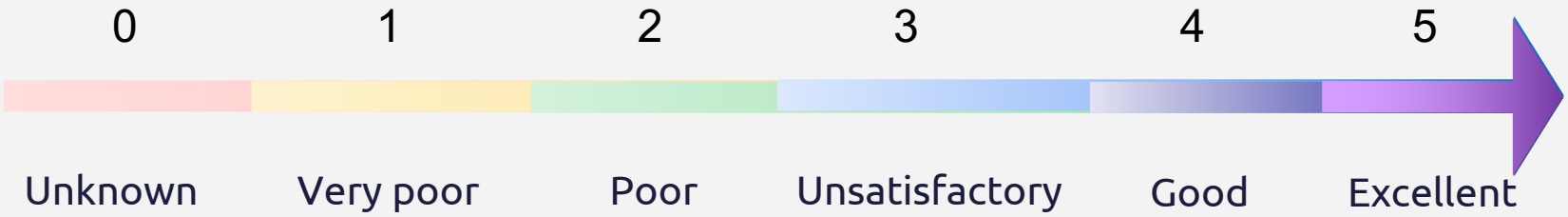
1. Data introduction

Variables:

① Websites Address
② World and Country Traffic Rank
③ Websites Traffic Information
④ Websites Safety
⑤ Social Media Traffics
⑥ Servers related Information

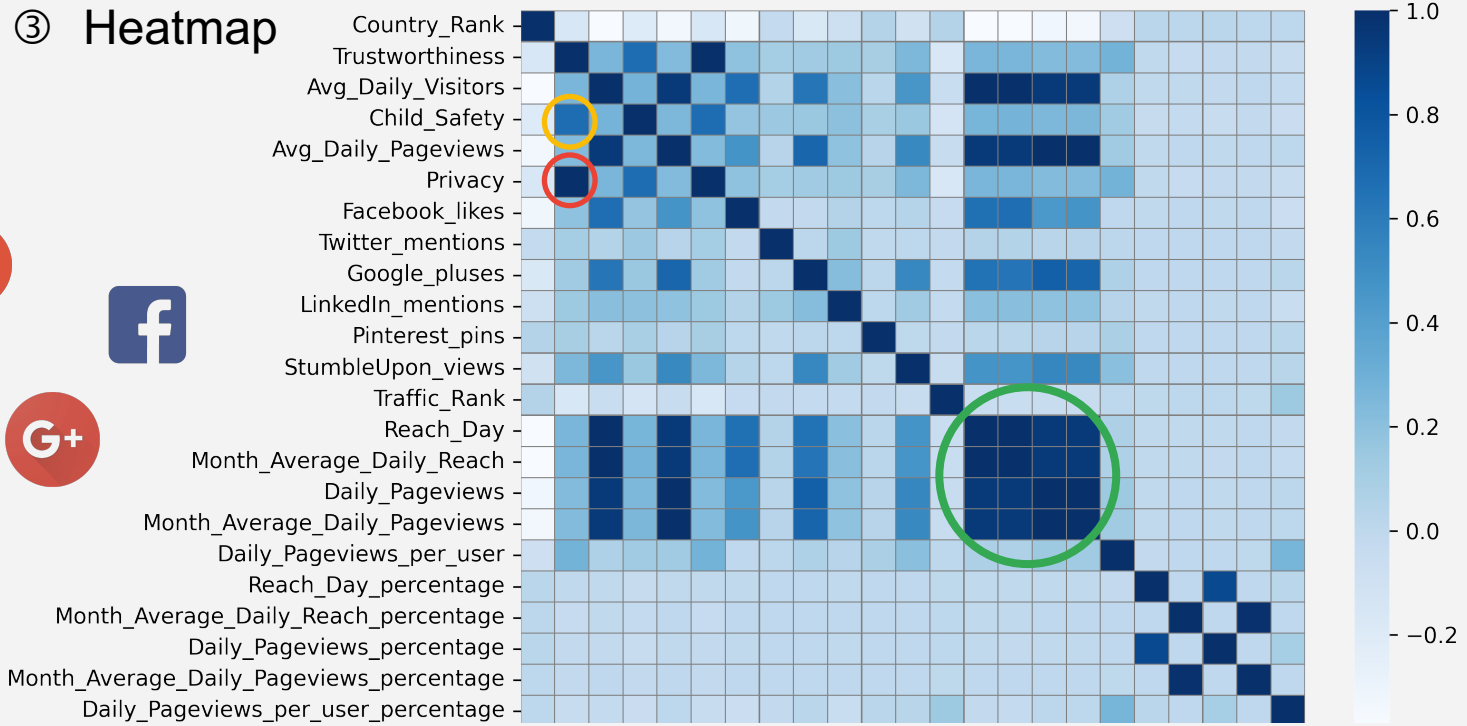
2. Data Pre-Processing

① Impute zeros in Social Media Traffics variables and as numerics
② Ordinal encoding for "Trustworthiness", "Child_Safety", "Privacy"



2. Data Pre-Processing

③ Heatmap



2. Data Pre-Processing

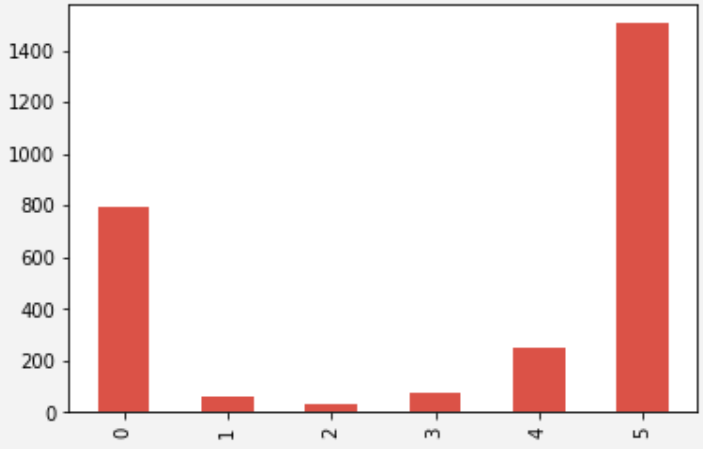
④ Remove the repetitive data(9540→3401)

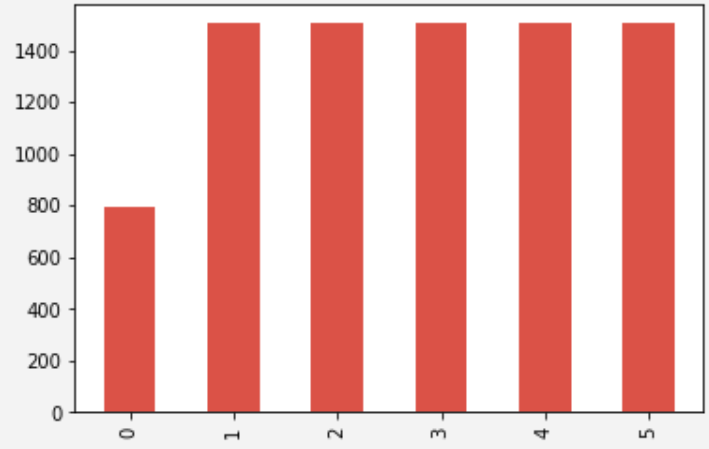
Website	Trustworthiness	Avg_Daily_ Visitors	Child_ Safety	Avg_Daily_ Pageviews	……	country
www.google.com	Excellent	515007350	Excellent	4192159833	……	Afghanistan
www.google.com	Excellent	515007350	Excellent	4192159833	……	Albania

Website	Trustworthiness	Avg_Daily_ Visitors	Child_ Safety	Avg_Daily_ Pageviews	……	Country_ Afghanistan	Country_ Albania
www.google.com	Excellent	515007350	Excellent	4192159833	……	1	1

2. Data Pre-Processing

⑤ Split data
⑥ Oversampling





3. Goal setting

Goal : Predict the level of child safety of a website

Measurement:

$$F1\ score = \frac{2\times(precision \times recall)}{precision + recall}$$

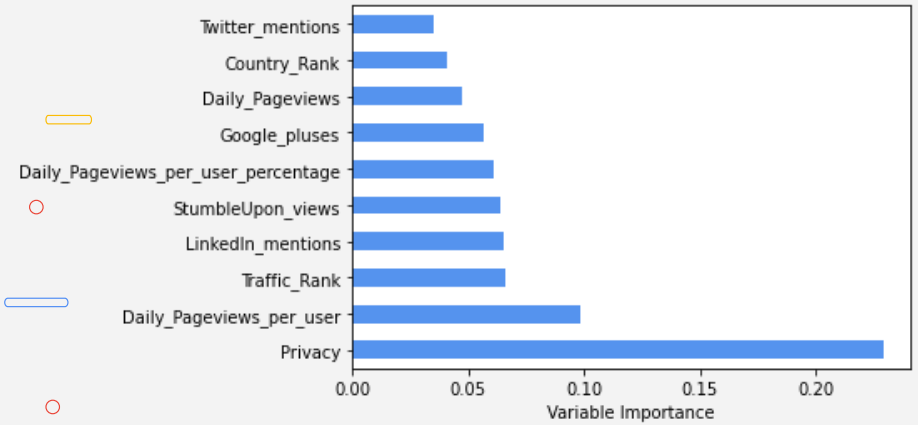
$$Weighted - Precision = P_0\times W_0 + P_1\times W_1 + \cdots + P_5\times W_5$$

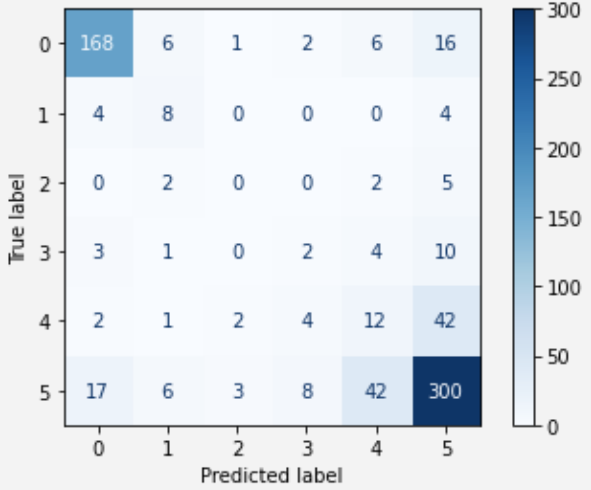
$$Weighted - Recall = R_0\times W_0 + R_1\times W_1 + \cdots + R_5\times W_5$$

4. Methods

① Decision Tree

accuracy	0.7174
f1_score(average="weighted")	0.7173

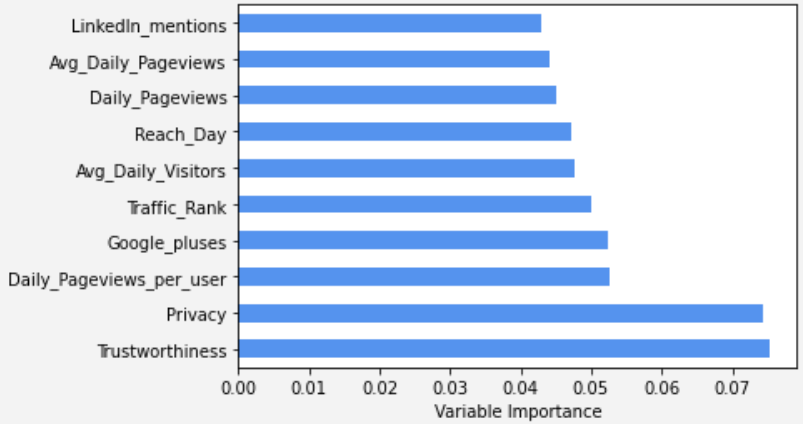


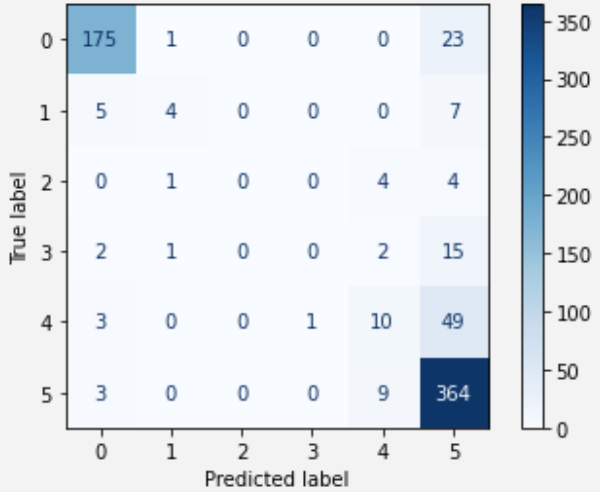


4. Methods

② Random Forest

accuracy	0.7994
f1_score(average="weighted")	0.8351

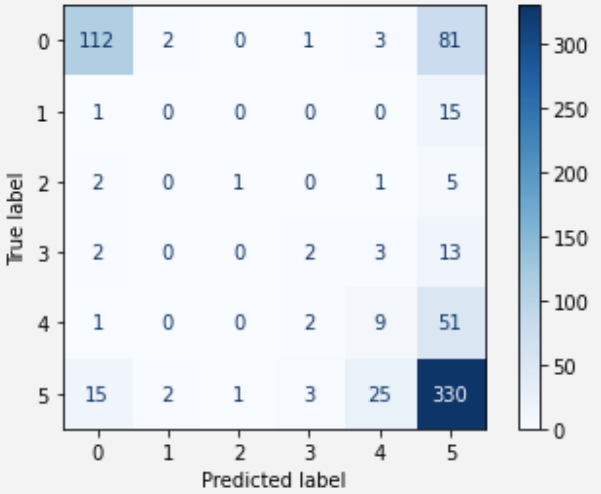




4. Methods

③ SVM

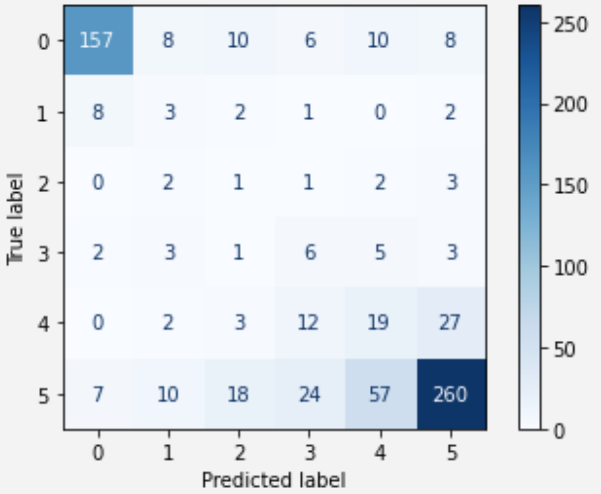
accuracy	0.6647
f1_score(average="weighted")	0.6932



4. Methods

④ Multinomial Logistic Regression

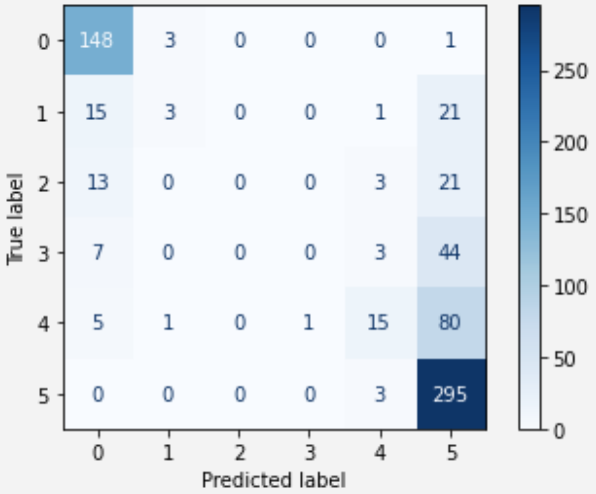
accuracy	0.6530
f1_score(average="weighted")	0.6982



4. Methods

⑤ PCA + Multinomial Logistic Regression

accuracy	0.6252
f1_score(average="weighted")	0.6839



5. Result and insight

Trade off between level 0,5 and level 1,2,3,4

	Decision Tree	Random Forest	SVM	Multinomial Logistic Regression	PCA + Multinomial Logistic Regression
Accuracy	0.7174	0.7994	0.6647	0.6530	0.6252
F1_Score	0.7173	0.8351	0.6932	0.6982	0.6839
			rbf c=100 gamma=0.5		14

6 . Reference

- https://kknews.cc/zh-tw/tech/a6bmjyx.html
- https://zhuanlan.zhihu.com/p/64315175
- https://www.kaggle.com/bpali26/popular-websites-across-the-globe

THANKS