1. All the experiments are done on the lab machine in CSIT building, ANU. The operating system is 64-bit Ubuntu (Linux). The memory is 15.3 GiB and the CPU is Intel Core i7-7700 CPU @ 3.60GHz. rattle
2. (a)

The purpose of data collection is to gather the answers to some questions from a group of people in order to draw some conclusions and make some predictions in the future. In this case, the collected data is the personal feedback about the experiences under the effect of the Coronavirus, the attitudes about the sexual harassment and assault [1]. The collection of experiences and views of participants is useful for the analysis of circumstances of Australians under the effect of COVID-19. The mean for data collection in this case is by online survey.

(b)

The correlation table and correlation plot for the seven attributes is generated in rattle. There are a total number of 42 pairs of attributes, and they are generally divided into 4 groups which are strongly positive correlated, slightly positive correlated, strongly negative correlated, and slightly negative correlated.

|  |  |  |  |
| --- | --- | --- | --- |
| strongly positive correlated (r >0.5) | slightly positive correlated | strongly negative correlated | slightly negative correlated |
| C3\_d & C3\_e |  |  |  |

Age 0.3-0.5 有点关系 a&b c&d c&e b&c b&f

A&B 人们相信权威正式渠道获取信息。通过a的人有一定可能也是会看b

很多人在不同渠道获取信息之后愿意分享给家人，这些渠道包括（b,d,e）.

b&f 获得专业指导后，分享社交媒体

大于0.5 d&e

D 和E都是传统媒体。 喜欢看报的人非常有可能喜欢看电视和听收音机。

age group 和所有的C3 都是0.05 以下， 获取covid 消息来源和年龄**无关**。. age 和其他C3所有都几乎是.无关。

其他所有方法无关系。 这和不同人的喜好性格有关， 人们选择自己信赖的途径来获取信息。

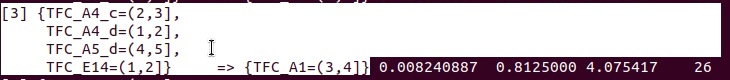
1. Association mining
2. 影响幸福指数

去掉missing

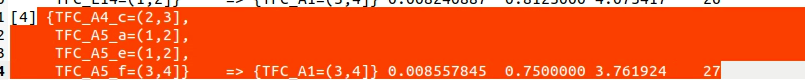
用support 有误解，比如data严重skewed， 根据support排列不太行。1137 人0-1开心，开心support太高了。要用coffidence

根据对数据的观察，数据不均匀分布，开心满意设置设置parameter 0.6 0.6

不满意的数据， support非常少， 0.009，



对GP 不自信，对hospital不自信，对政客不信任，钱不多，不开心



对政府不自信，对政治敏感关心，觉得政治是大问题，不觉得领导人重要



自己或者朋友不感染开心



在公共区域不社交开心

（2） 发现covid影响，政府反馈提高

3. 1. 为什么简单？

The opinionated is the respondent on different parts of A4 and A5. Opinioned 是根据前面的数据计算A4 A5 得到的。 Classification performance 会很高。 No ground truth label。

TREE:

Performance 不随 training 比例变化。都是 FP FN 都是0. Accuracy recall precision F not changed TN335 TP 139

Linear:

20/40/40

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

70/15/15

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

训练多了变差，overfit。 对数据很敏感，比如categorical 没有normalised。导致数据极大极小。

SVM：

20/40/40

文本

描述已自动生成

70/15/15

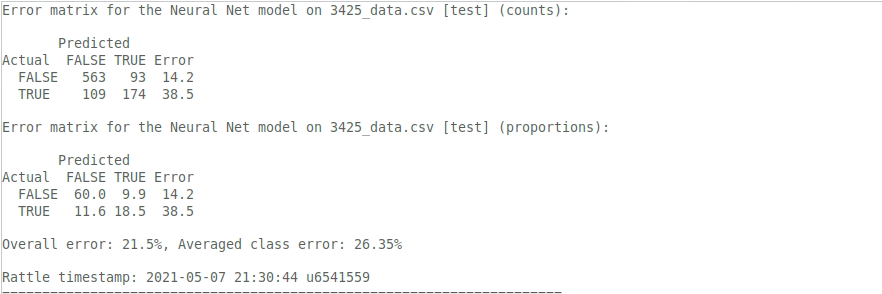
图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

提升了。Have more training data。 The data is not linear separable. More data optimise the straight line classifier.

NN

20/40/40



70/15/15

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Neural Network have better performance. Underfitting.

© Tree classifier. Performance 100% target 是由rule 生成的 tree 可以找到 rule， 忽略outlier的影响

[1] N. Biddle, “Participant Information Sheet”, ANU. [Online]. Available: https://www.srcentre.com.au/anuethics/Information\_sheet\_ANUPoll\_April\_2021.pdf