**Rapid Miner Project Description**

Jack Bauer family is going to move to Pittsburgh. The family is recruiting a butler to help them make decisions. The tasks is:

1. A house. Jack Bauer family wants to buy a house. The requirements are:
2. The price is less than 500,000 USD.
3. It has investment potential.
4. Close to medical centers/hospitals, universities and supermarkets/malls (Target, Walmart, Whole Food, Costco, etc.).
5. Excellent traffic in surrounding areas.

You are required to give a presentation to Jack Bauer family to help them make the above decision.

* For Task 1, you are suggested to use (but not limited to) Decision Tree and search for some more data from web to prepare a rich and exciting presentation.
* Data: Pittsburgh property price data.

 \* You are encouraged to additionally collect your own data to conduct more solid analysis.

* A visualization system to help:

<https://vietexob.shinyapps.io/real_estate_app>

<https://vietexob.shinyapps.io/traffic_real_estate>

This is an open project, please feel free to use your resources and power!

Jack Bauer family is looking forward to your presentation!

**Task 1 Guide:**

1. Whether a house should be recommended is a multi-factor decision, including its price, investment potential, traffic, proximity to public services, crime rate, neighborhood etc. Try to manually rate several (say, 20~) housing options by discussing among group members. Determine an overall rating of each housing option based on the multiple factors your group chose to judge upon.

1.1 This rating can be somewhat subjective, but the more options you rate, the more objective your analysis becomes in the later training step.

1.2 Suggested scale is 1~10, but the actual rating scale is up to you.

1.3 Theoretically, if you manually rate all the 3318 options in csv data, your job is done because you can recommend the houses with highest ratings to Jack family. However, do you have time for that?

2. The subset of data you manually rated (i.e., labeled) is your training data + test data. Again, the more data you label, the more time you need, but the more useful your trained model will be. You need to find a balance by yourselves.

Select 90% of labeled data as training data, the rest 10% as test data to optimize your decision tree parameters. You want to train a decision tree model with good performance on your test data.

Things to think about:

2.1 How do you sample the training data and test data?

2.2 Do you need to use all the attributes provided in csv? Any preprocessing of the raw data?

3. When you trained a satisfactory decision tree model, apply the model to unlabeled data to automatically label them (make rating prediction).

Finally, provide suggestion to the family **with good reasons**.

**Note: You need to submit PowerPoint and 2 pages’ project report and rapid miner file.**