Conclusion:

* The competition is focusing on data cleaning/feature selection. The learning model selection is less important comparing to the previous tasks.
* I think some of the submissions that have highest scores might’ve used neural networks.
* There is a large room of improvement for us since we only used a small portion of data. After the proper data filling, I am confident that we can obtain much more accurate results.
* The nature of the data might be a problem, e.g. some data are in the range(0,100), and some are binary, some are string. I’m currently working on this, but I’m not very certain about the best option. I will start from normalization/standardization like what I did last week.
* Since the final submission asks for the probability number, I was thinking how to generate it. As I know the SVM will naturally generate this number, and Decision Tree (/Random Forest) has the attribute/option to generate this. I’m not so sure but other algorithms but I think it should be fine because most

Draft Plan for this week & timeline

* Dropping the rows missing too many values ( > 90 out of 180 is the threshold I selected)

[by tonight/tomorrow morning]

* Filling missing values according to the distribution

[by tonight/tomorrow morning]

* Train all the features using decision tree since it can do feature selection automatically, then we decide if we need to do feature selection again using other methods. [by Wednesday morning]
* Check if the problem is actually a multi classification problem using cluster. [by Wednesday night]
* Submit a result.[Thursday morning]
* Adjust the threshold/parameters. Finetune the algorithm. [rest]