Data 624 Project 2

# Problem Statement

Data related to the manufacturing process of a soft drink are provided, including values of various parameters that control the process. The objective is to build a predictive model to predict the pH content based on the manufacturing process data.

# Executive Summary

Since predicting the value of a numerical variable such as PH is a regression problem, our methodology consists of building five different models for the problem, and ranking their performance on a holdout portion of the training data provided. We then selected the model with the lowest RMSE value on the holdout data set. We developed predictors using the following models:

- Support Vector Machine  
- Stochastic Gradient Boosting  
- Random Forest  
- Neural Network  
- Generalized Linear Model

Comparing the model prediction errors on the holdout data set, we found that the Random Forest model consistently produced the best results, and we therefore selected it to make predictions on the test data set that was provided.

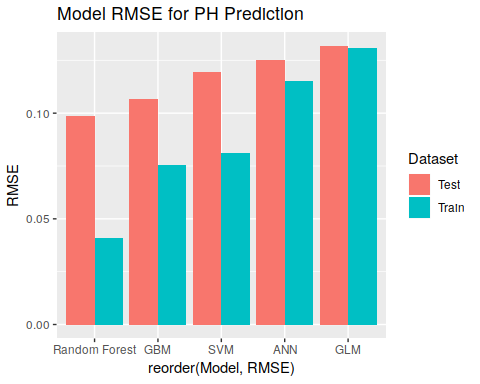
We also examined the most important predictors related to PH values. The more we understand the relationships between PH and the variable, the better we can manage PH through manipulating those variables effectively. The top 1 - 6 of most important variables are:

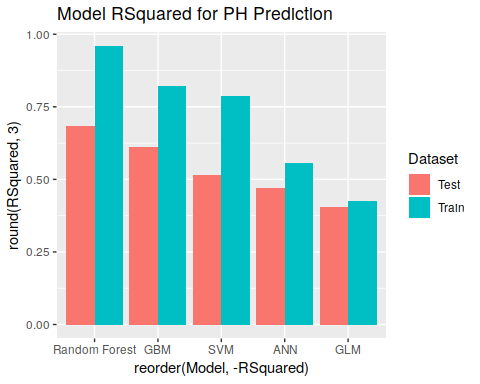
- Mnf.FLow,  
- Band.Code\_C (via one-hot-encoding of Brand.Code),  
- Pressure.Vacuum,  
- Oxygen.Filler, and  
- Temperature.

# Comparison of Model Performance

We now compare and rank the RMSE errors produced by the various models on the portion of the training data that was withheld from the training of the models. Using the model with the smallest RMSE on the withheld training data, we also make a final prediction on the test data provided. The predictions are written to an Excel file.

Since we see that the Random Forest model produced the lowest RMSE on the withheld training data, we select it as the best model to predict PH in the manufacturing process data set. A final set of predictions is made using this model.





bestFit = model5  
PH.pred = predict(bestFit, test)  
write.xlsx(PH.pred, file="DATA624\_Proj2.xlsx", sheetName="PH", append=F)

# Conclusions

All of our models performed fairly well, with a measured RMSE of less than one-tenth of a pH point. Random Forecast outperformed other models with the smallest RMSE and largest RSquared values in training (0.0410 and 0.96) and test (0.099 and 0.684), and is selected as our best model. Random Forecast is a tree-based model, it is hard to get coefficients for each variables, but it does have a way to show the importance of the variables in affecting the target variables, as showed in the important variable lists above. The top 6 variables are

- Mnf.FLow,  
- Band.Code\_C (via one-hot-encoding of Brand.Code),  
- Pressure.Vacuum,  
- Oxygen.Filler, and  
- Temperature

By changing these variables, we can maximize the control over pH of the product while minimizing the things we have to monitor and control. Further cost analysis on these processes would help determine our route forward.

# Next Steps

There are many ways we can continue improving the model performance, one method could be running more times of cross validation on more folds than 3 times 3-fold we have now for SVM and GBM models. It would take a longer time to compute, but the results would likely be better. Finally, more data would help us build a better model, in particular because the gap between the test and train sets tends to be relatively large across all of the models.

# References

1. Random Forests. <https://uc-r.github.io/random_forests>
2. Kevin Murphy (2012). Machine Learning a Probabilistic Perspective.
3. Support Vector Machine. <https://uc-r.github.io/svm>
4. Support Vector Machines. <http://web.mit.edu/6.034/wwwbob/svm.pdf>
5. Kuhn et al (2013). Applied Predictive Modeling