Exploratory Analysis and Machine Learning of Gene Expression Data

Alexander T. Straub | Computer Science – Machine Learning Track
College Park Scholars – Science & Global Change Program
alex1@umd.edu | Academic Showcase, April 30, 2021

Introduction

Machine learning can be applied to almost every type of data. Models created from machine learning algorithms can 1) provide insight into what independent variables are important in predicting categorical or quantitative outcome and 2) predict of the categories or values of observations in an independent dataset.

In this project, we used machine learning to find genes that predict whether a given person in the study is a schizophrenic case (SCZ) or a normal control (CON).

Software

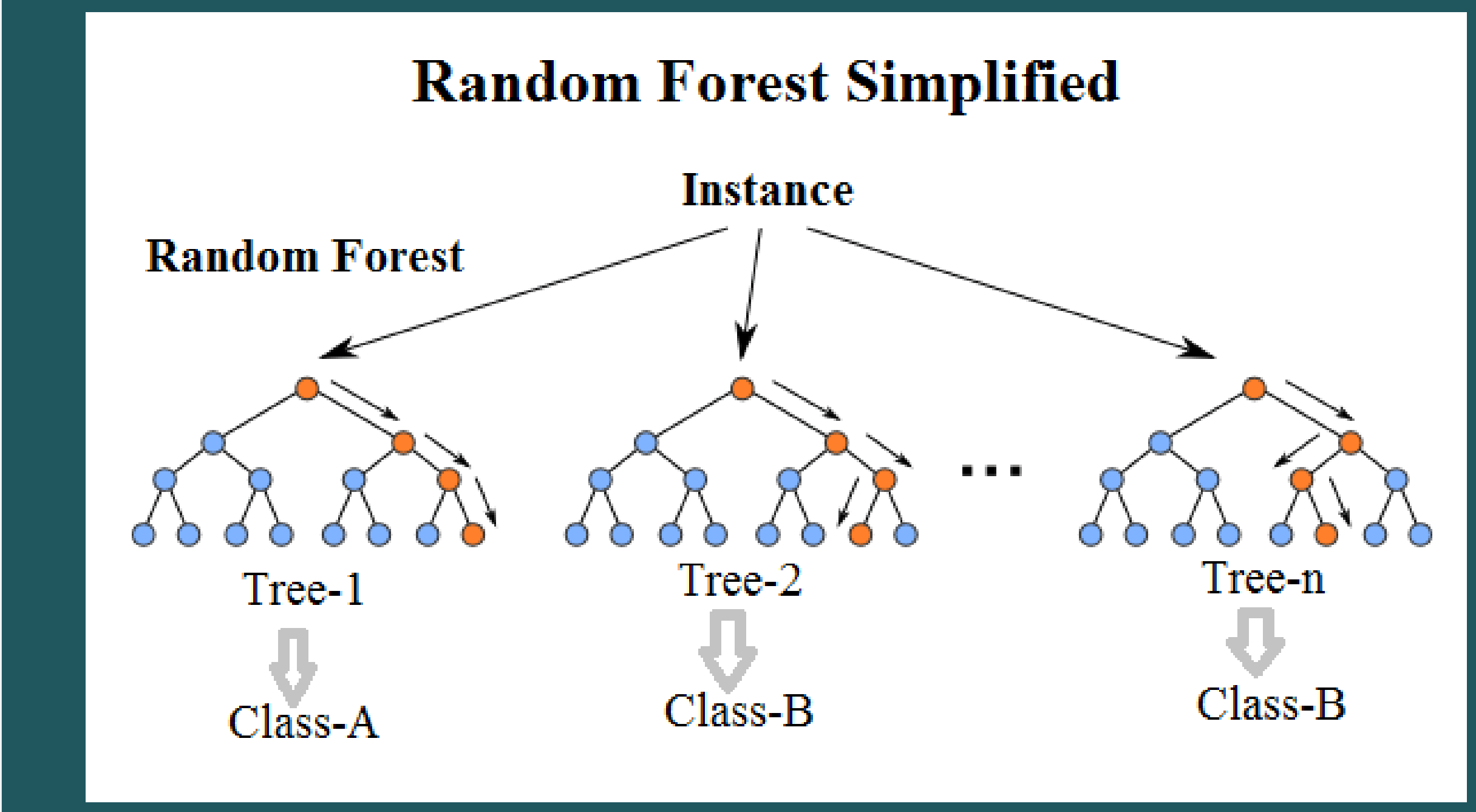
(Documentation for the software and packages used can be found in the bibliography section below)

- Preparing/cleaning data and analyzing output: Excel
- Exploratory data analysis, machine learning: R/R Studio and Python
  - Python packages: TensorFlow and Scikit (ML), Pandas (Data)

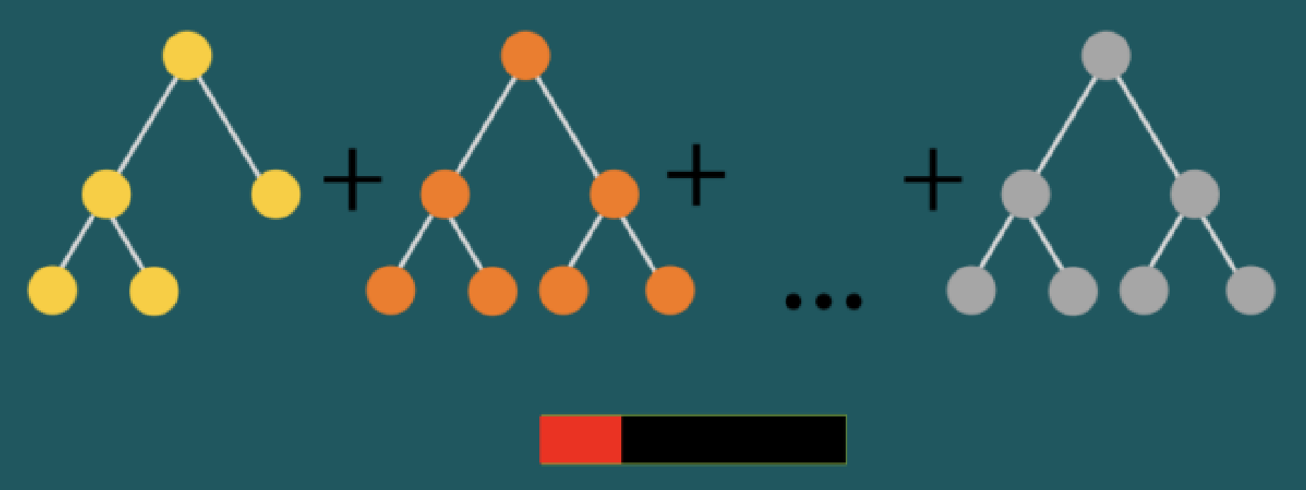
Personal Impact:

Ever since I heard about machine learning I wanted to explore and learn more about the field. This project was a great opportunity to not only gain experience working with machine learning techniques, but to use these technologies in a real world application.

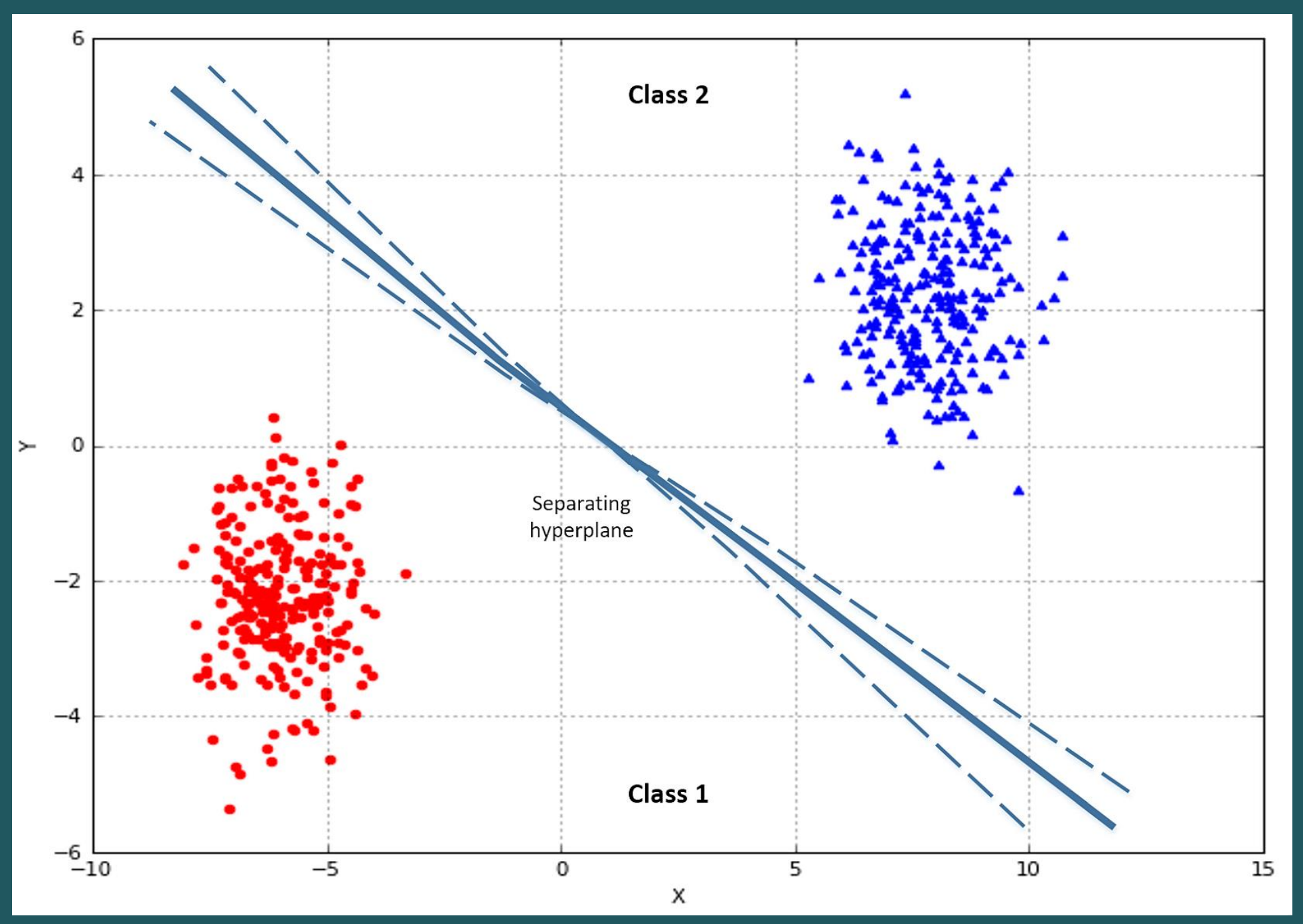
Algorithms Used: Random Forest (RF), Boosted Trees, Linear Classifier



A visualization of Random Forest classification
Image from: https://sefiks.com/2017/11/19/how-random-forests-can-keep-you-from-decision-tree/



A visualization of gradient Boosted Trees
Image from: https://developer.nvidia.com/blog/catboost-fast-gradient-boosting-decision-trees/



A visualization of linear classification (Generic data used) Image from: https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781785889622/5/ch05lvl1sec39/linear-classification

Procedure

1. Input data: molecular Gene Expression dataset of ~30k genes, measured in cells from 12 people with schizophrenia (SCZ) and 14 controls (CONs)
2. Characterization and exploratory data analysis to determine inter-individual variation
3. Produce multiple subsets of the data, based on prior differential expression (SCZ vs. CON) results combined with variation statistics
4. Run all data subsets through two versions of Random Forest (R, Scikit), a Boosted Tree classifier (Tensorflow) and a Linear Classifier (TensorFlow)
5. Compare the Step 4 variable importance (Vi) list outputs from the different models to find genes that are the most important in predicting SCZ vs. CON status

Site Information

Lieber Institute for Brain Development (Remote) https://www.libd.org/

855 N Wolfe St Suite 300, Baltimore, MD 21205

Site Supervisor: Richard E. Straub Ph.D. Senior Research Scientist

The mission of the Lieber Institute for Brain Development Maltz Research Laboratories is to translate the understanding of basic genetic and molecular mechanisms of schizophrenia and related developmental brain disorders into clinical advances that change the lives of affected individuals.

Results / Discussion

After looking at the variable importance lists from the various machine learning algorithms, it was apparent that the models vary considerably in which genes are implicated.

Even the same algorithm implemented in two different programs (e.g. RF in R or Python) had different importance values for a given gene.

Nonetheless, there were genes that had a high predictive value in multiple different Vi lists

Future Work

These biological functions of these genes are now being investigated by my site supervisor. This procedure can be applied to other clinical datasets, which will help in predicting case vs. control for future projects.

Proceed with only those genes important in the current models and retrain the models to increase prediction accuracy.

In addition, we will optimize the procedure and include different types of clinical test results to see how that influences case vs. control predictions.

Bibliography

- Anonymous. March 2021. “tf.estimator.BoostedTreesClassifier” [https://www.tensorflow.org/api_docs/python/tf/estimator/BoostedTreesClassifier/]”. TensorFlow. Accessed 10 March 2021
- Macklin, A. July 2019 “Random Forest In R” [https://towardsdatascience.com/random-forest-in-r-f66adf80ec9/]”. Towards Data Science. Accessed 5 February 2021
- Anonymous. January 2020 “1.11. Ensemble methods” [https://scikit-learn.org/stable/modules/ensemble.html#forest]”. Scikit Learn. Accessed 12 February 2021
- Anonymous. April 2021 “User Guide” [https://pandas.pydata.org/docs/user_guide/index.html#user-guide]”. Pandas. Accessed 21 February 2021

Acknowledgments

My supervisor Richard Straub was an excellent mentor and teacher, providing me with all the support and resources I needed along the way. He explained numerous biological factors that were beyond the scope of my knowledge. I would also like to acknowledge the College Park Scholars Science and Global Change program, along with Drs. Holtz & Merck who made this opportunity possible