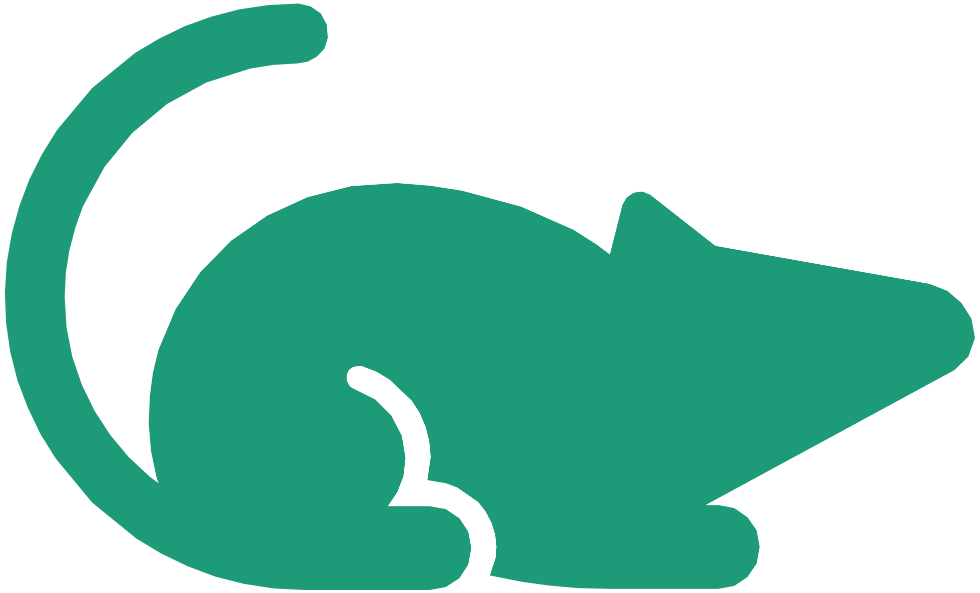
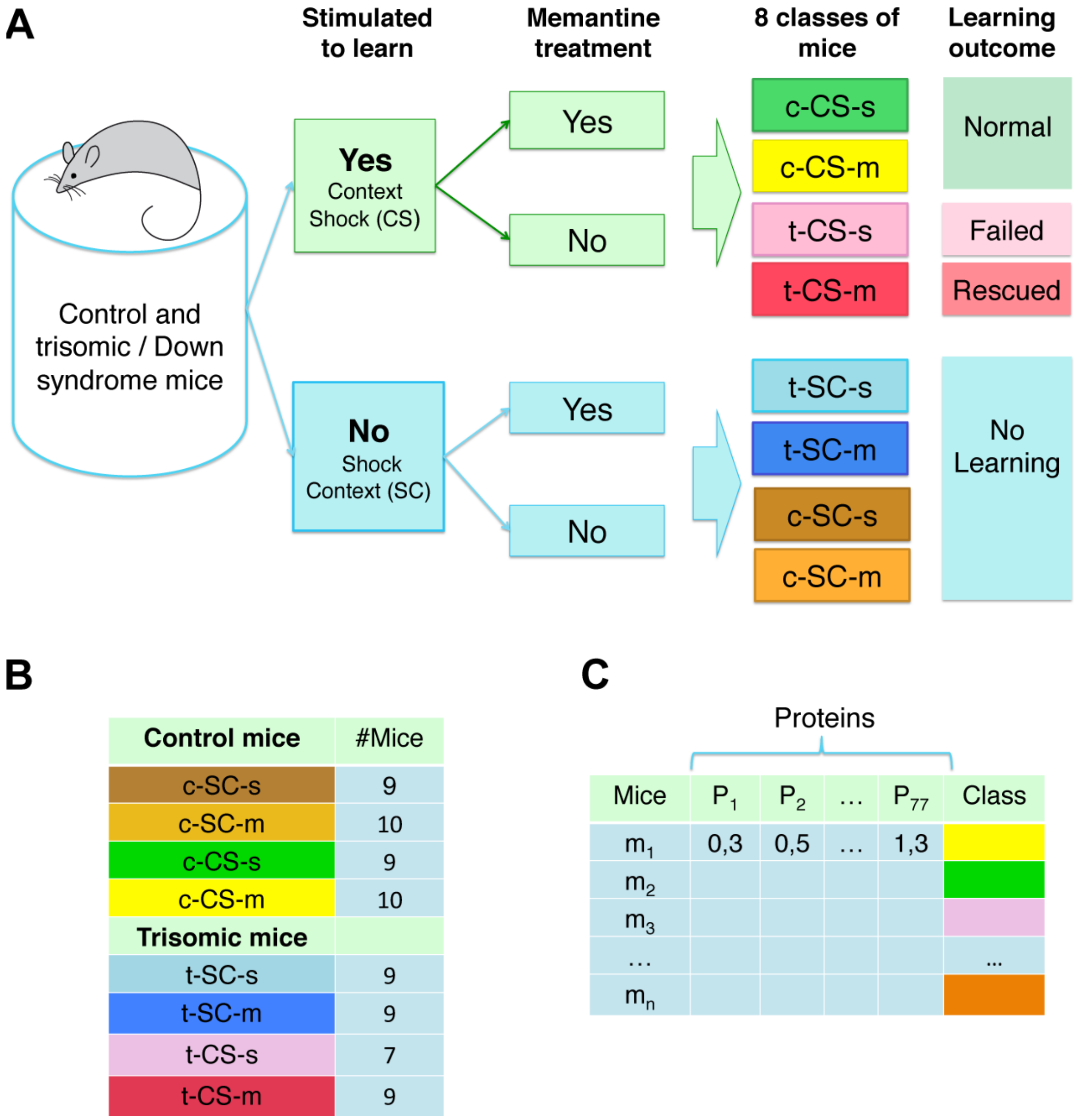


Classification of Mice by their Protein Expressions

by Alexandre Yano

Introduction

- The dataset contains a total of 1080 measurements per protein(77 proteins in total) where each measurement can be considered as an independent sample.
- There are eight classes of mice which are described based on features such as genotype, behavior and treatment.
- According to genotype, mice can be control or trisomic(down syndrome mice).
- According to behavior, some mice have been stimulated to learn (context-shock) and others have not (shock-context).
- And in order to assess the effect of the **drug memantine**, some mice have been injected with the drug and others have not.
- Control mice learn successfully while the **trisomic** mice fail, unless they are first treated with a drug, which **rescues** their learning ability.
- The data was obtained from the UCI repository .



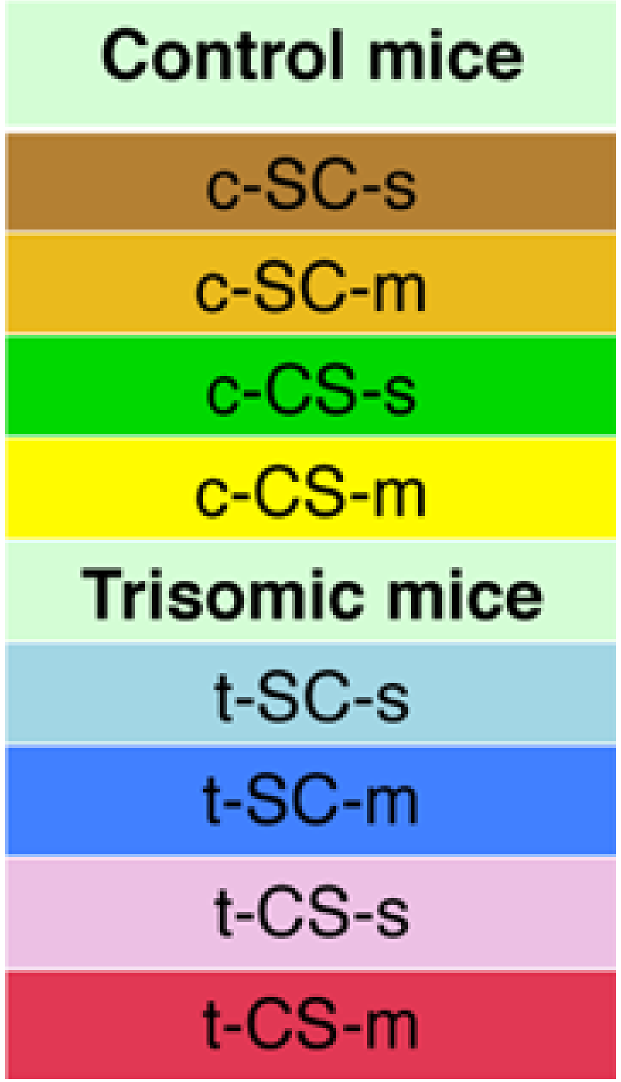
Control mice	#Mice
c-SC-s	9
c-SC-m	10
c-CS-s	9
c-CS-m	10
Trisomic mice	
t-SC-s	9
t-SC-m	9
t-CS-s	7
t-CS-m	9

	Proteins				
Mice	P_1	P_2	…	P_{77}	Class
m_1	0,3	0,5	…	1,3	🟨
m_2					🟩
m_3					🟪
…					...
m_n					🟧

The **Ts65Dn mouse model** of down syndrome display many features relevant to those seen in Down Syndrome in humans(including deficits in learning and memory) and for this reason, it can be used to learn more about Down Syndrome.

The 8 Mice Classes

- c-CS-s: control mice, stimulated to learn, injected with saline
- c-CS-m: control mice, stimulated to learn, injected with memantine
- c-SC-s: control mice, not stimulated to learn, injected with saline
- c-SC-m: control mice, not stimulated to learn, injected with memantine
- t-CS-s: trisomy mice, stimulated to learn, injected with saline
- t-CS-m: trisomy mice, stimulated to learn, injected with memantine
- t-SC-s: trisomy mice, not stimulated to learn, injected with saline
- t-SC-m: trisomy mice, not stimulated to learn, injected with memantine



Control mice

c-SC-s

c-SC-m

c-CS-s

c-CS-m

Trisomic mice

t-SC-s

t-SC-m

t-CS-s

t-CS-m

Objective

- The goal of this project is to find a model that is able to extract a subset of proteins that can help us classify mice by their protein expressions.
- The model should contain just a subset while maintaining higher accuracy.

- Again, the data set consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex.

Method

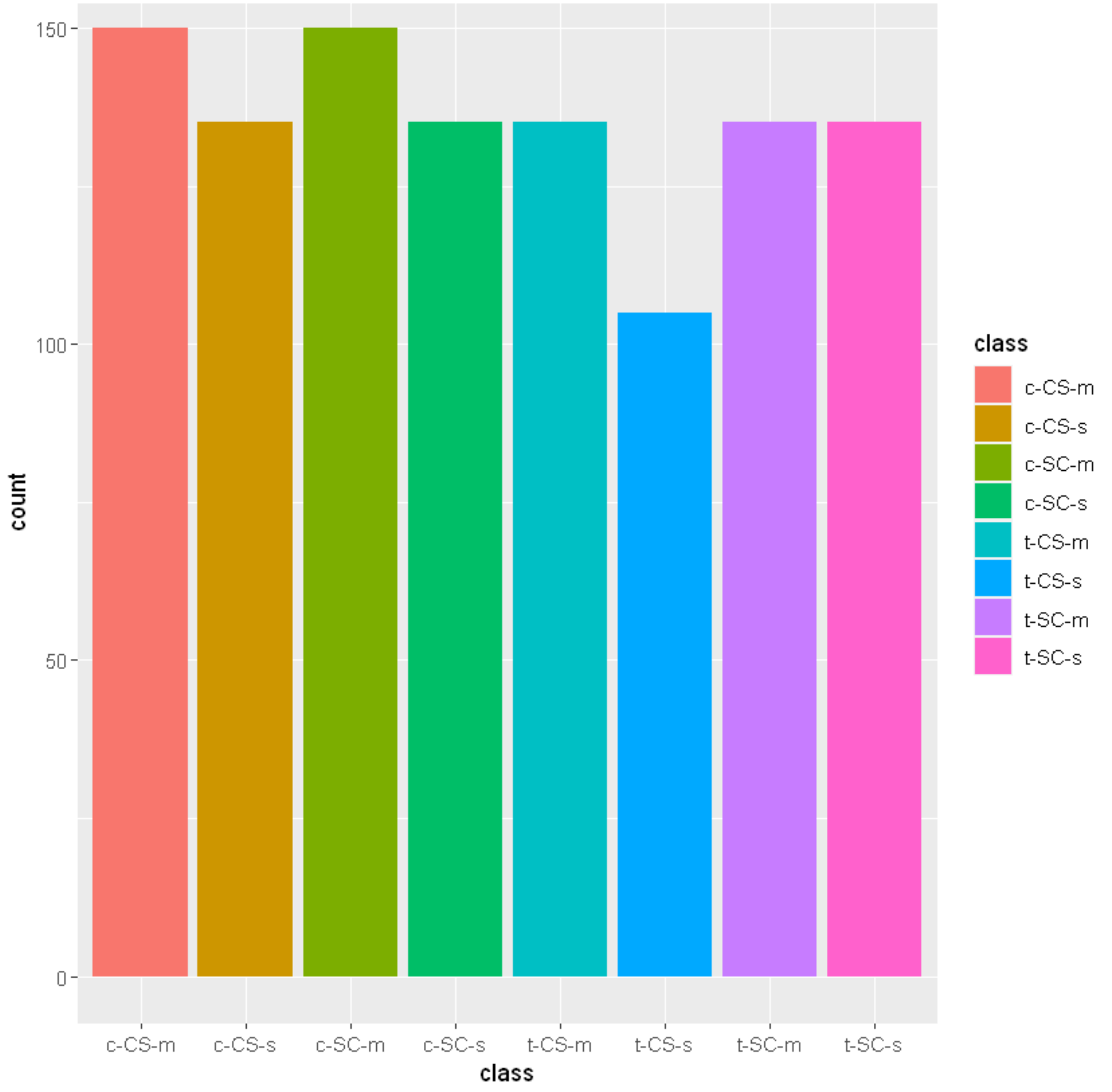
- We will use XGBOOST and Random Forest to extract a subset of proteins that help us classify each mouse based on their protein expression.

- XGOOBST was chosen as it is particularly famous for outperforming others machine learning algorithms.

- Random Forest was chosen was for practical purposes as it easily help visualize the most important features in the dataset which leads to feature reduction.

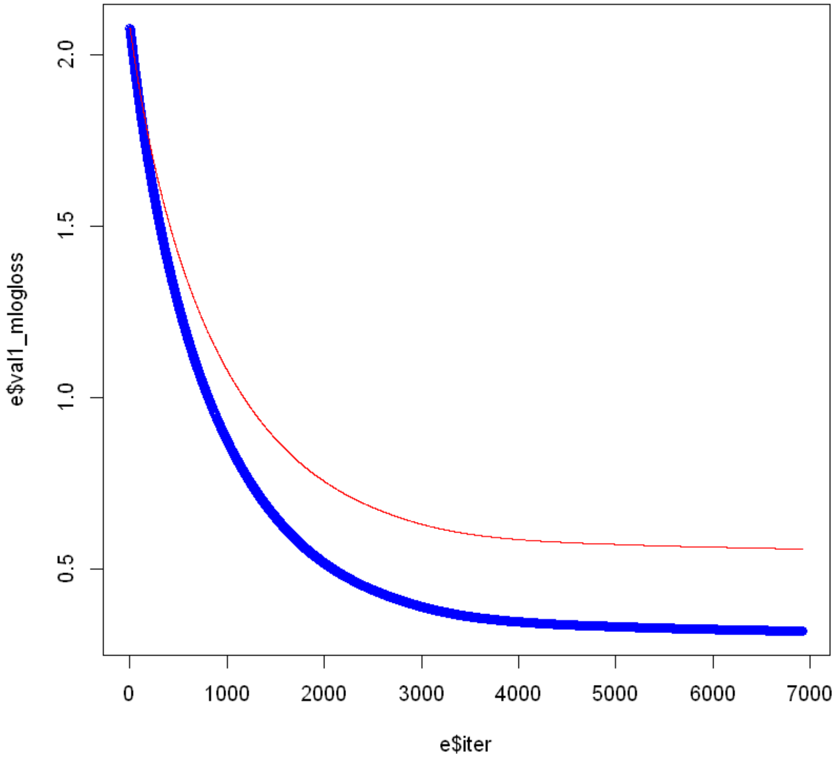
Checking for Class imbalance

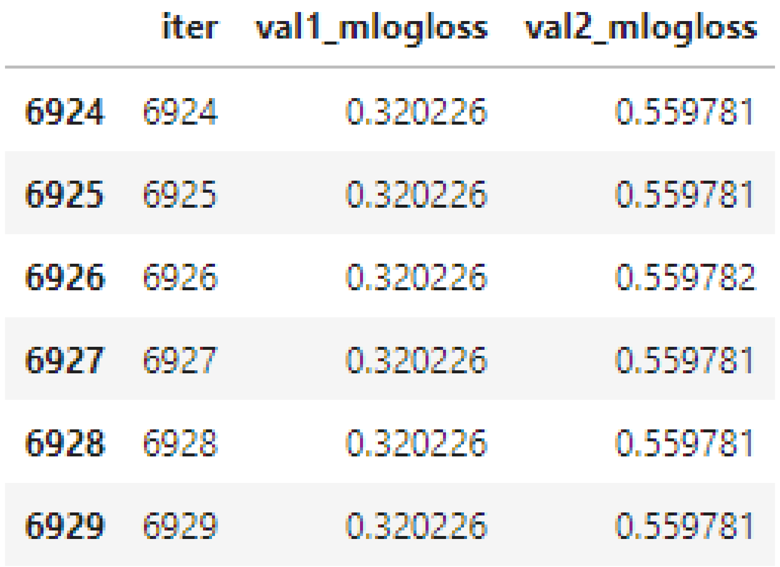
- The “class variable” is almost evenly distributed, that is, none of the mice type is totally dominating the other types.



Classification of mice by their protein expressions

- XGBOOST model

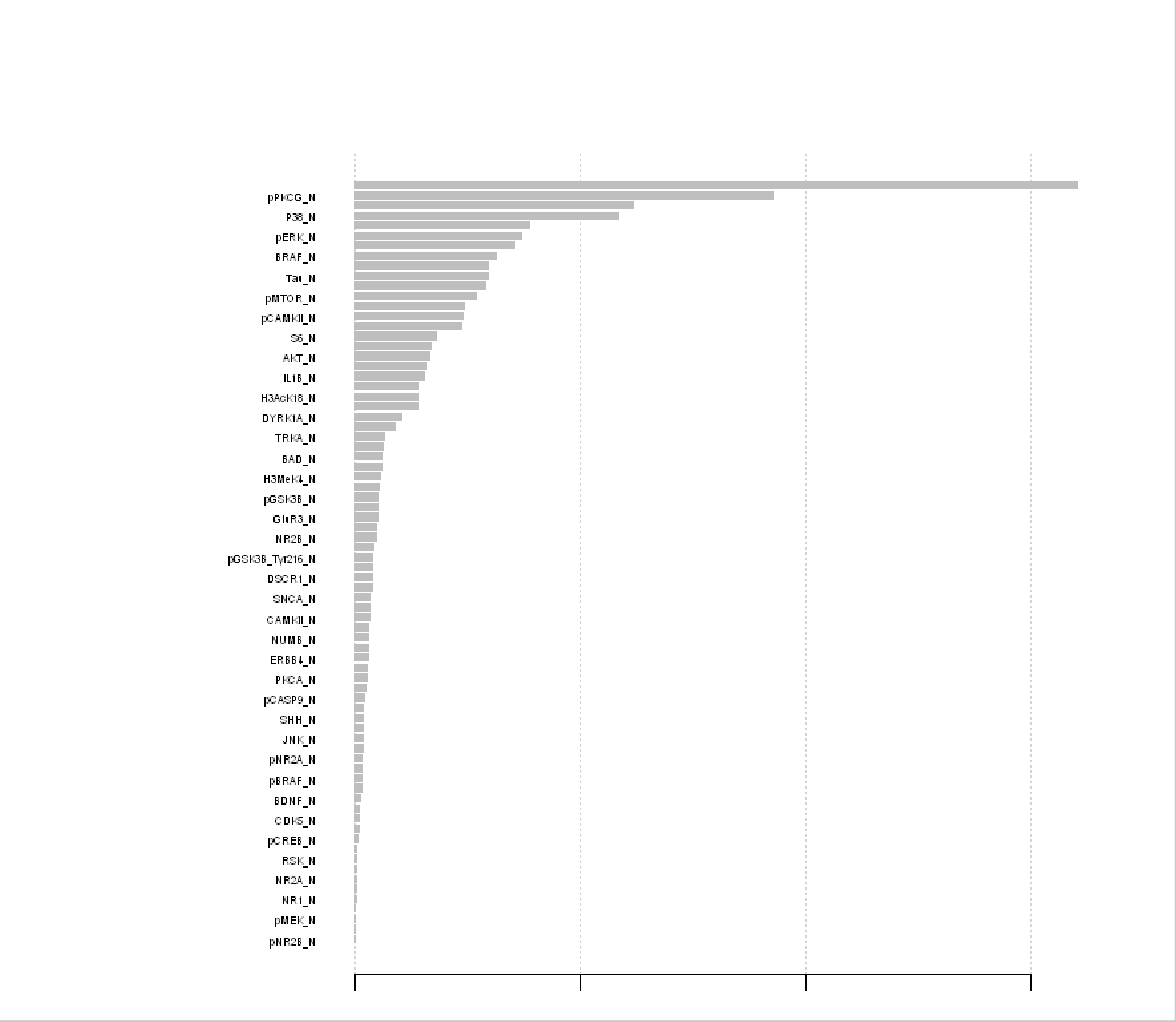




	iter	val1_mlogloss	val2_mlogloss
6924	6924	0.320226	0.559781
6925	6925	0.320226	0.559781
6926	6926	0.320226	0.559782
6927	6927	0.320226	0.559781
6928	6928	0.320226	0.559781
6929	6929	0.320226	0.559781

- XGBOOST feature importance using the 77 proteins

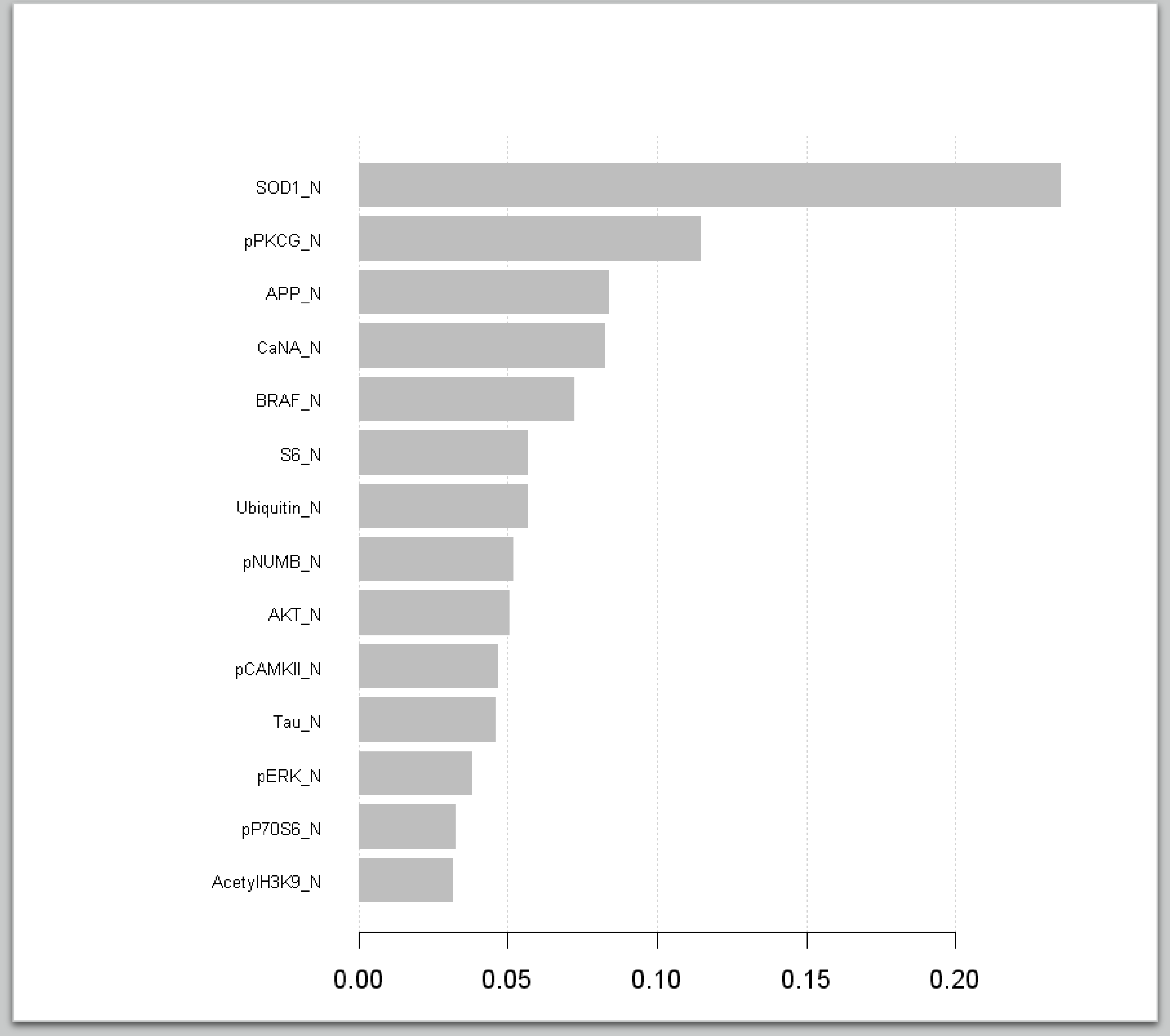
- Accuracy of 90%



XGBOOST Short Model

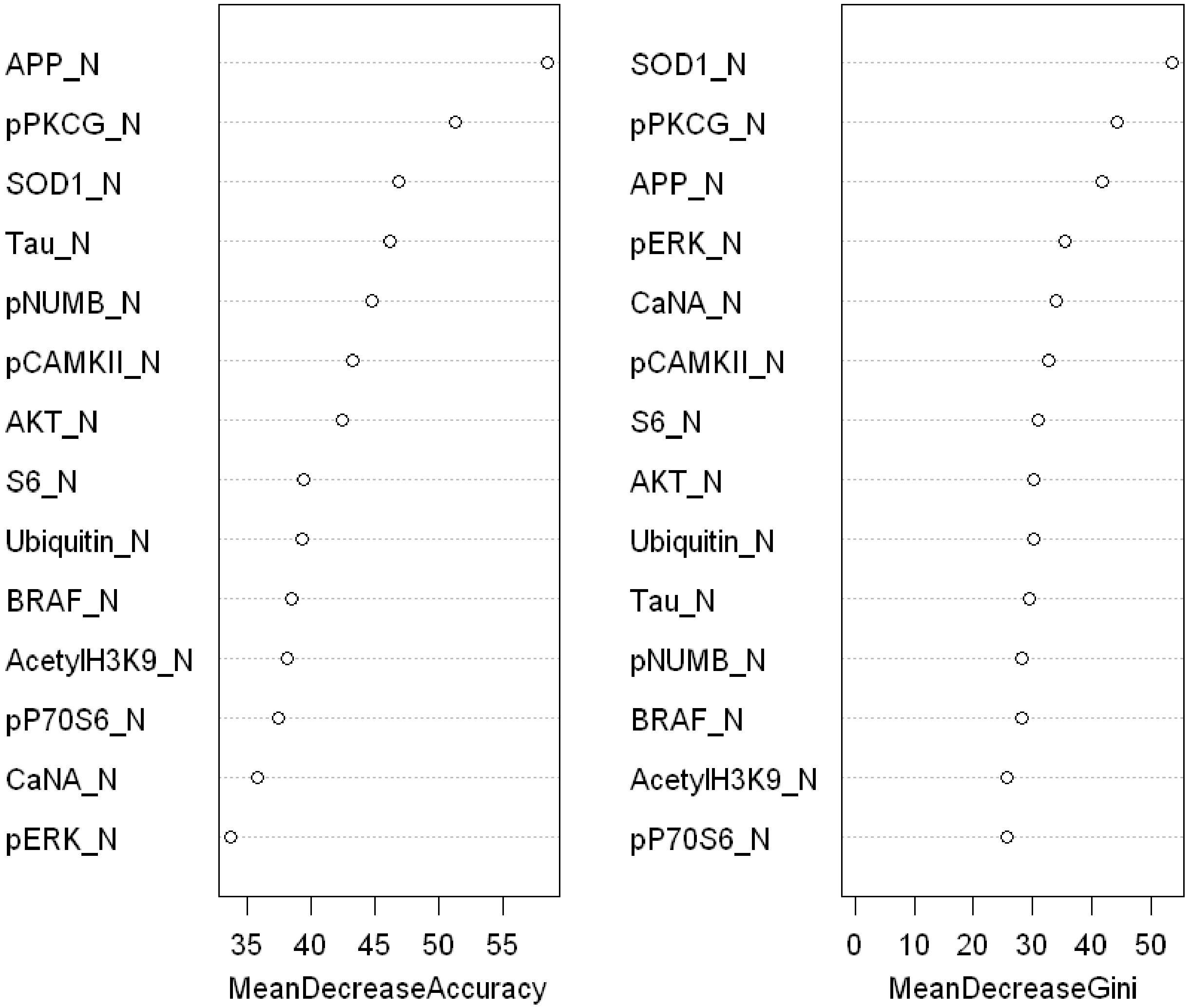
using the 14 most important proteins

accuracy of 85%



Classification of mice by their protein expressions

- Random Forest



14 Proteins

accuracy of 96.11%

[web]

Results

- Random Forest outperformed Extreme Gradient Boosting in overall, for the large dataset and for the smaller sample which not usually the case.

- XGBOOST had an accuracy of 90% and in a smaller the accuracy went down to just 85.93%.

- Random forest was able to predict with 96.11% accuracy on a smaller sample. So, in particular, the subset of proteins for our final model were:

  ('SOD1_N', 'APP_N', 'pPKCG_N', 'pERK_N', 'pCAMKII_N’, 'CaNA_N', 'Tau_N', 'pP70S6_N', 'pNUMB_N', 'BRAF_N’, 'Ubiquitin_N', 'AKT_N', 'S6_N ', 'AcetylH3K9_N', ‘c)

Limitations

- Due to the nature of the experiment, not many data point could be collected as it involved live animal which were then euthanized.

- The experiment involved context fear conditioning (CFC), which, if not done correctly, can give rise to unethical issues.

- The was no description on the type of proteins in the dataset, so I did not know the meaning or the relevance of each type of protein.

References

- Morde, V. (2019, April 08). XGBoost Algorithm: Long May She Reign! Retrieved October 18, 2020, from https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

- Ahmed MM, Dhanasekaran AR, Block A, Tong S, Costa ACS, Stasko M, et al. (2015) Protein Dynamics Associated with Failed and Rescued Learning in the Ts65Dn Mouse Model of Down Syndrome. PLoS ONE 10(3): e0119491. https://doi.org/10.1371/journal.pone.0119491

- Higuera C, Gardiner KJ, Cios KJ (2015) Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome. PLoS ONE 10(6): e0129126. https://doi.org/10.1371/journal.pone.0129126

- Chromosome 21: MedlinePlus Genetics. (2020, August 18). Retrieved October 20, 2020, from https://medlineplus.gov/genetics/chromosome/21/