Signature Project (DA5030)

Neural network:

<https://www.analyticsvidhya.com/blog/2017/09/creating-visualizing-neural-network-in-r/>

<https://www.r-bloggers.com/2021/04/deep-neural-network-in-r/>

single-cell analysis

<https://journals.biologists.com/dev/article/148/14/dev197962/270926/Single-cell-transcriptomics-of-the-early>

<https://www.nature.com/articles/d41586-018-05882-8>

<https://www.nature.com/articles/s41586-018-0414-6>

**Dataset research:**

**Mouse Nuclear cortex (sc-RNASeq)**

[**https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression**](https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression)

*Context:*

Expression levels of 77 proteins were measured in the cerebral cortex of 8 classes of control and Down syndrome mice exposed to context fear to condition, a task used to assess associative learning.

*Content:*

The data set consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex. There are 38 control mice and 34 trisomic mice (Down syndrome), for a total of 72 mice. In the experiments, 15 measurements were registered of each protein per sample/mouse. Therefore, for control mice, there are 38x15, or 570 measurements, and for trisomic mice, there are 34x15, or 510 measurements. The dataset contains a total of 1080 measurements per protein. Each measurement can be considered as an independent sample/mouse.

The eight classes of mice are described based on features such as genotype, behavior and treatment. According to genotype, mice can be control or trisomic. 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 in recovering the ability to learn in trisomic mice, some mice have been injected with the drug and others have not.

*Classes:*

1. c-CS-s: control mice, stimulated to learn, injected with saline (9 mice)
2. c-CS-m: control mice, stimulated to learn, injected with memantine (10 mice)
3. c-SC-s: control mice, not stimulated to learn, injected with saline (9 mice)
4. c-SC-m: control mice, not stimulated to learn, injected with memantine (10 mice)
5. t-CS-s: trisomy mice, stimulated to learn, injected with saline (7 mice)
6. t-CS-m: trisomy mice, stimulated to learn, injected with memantine (9 mice)
7. t-SC-s: trisomy mice, not stimulated to learn, injected with saline (9 mice)
8. t-SC-m: trisomy mice, not stimulated to learn, injected with memantine (9 mice)

*Attribute Information:*

[1] Mouse ID

[2:78] Values of expression levels of 77 proteins; the names of proteins are followed by N indicating that they were measured in the nuclear fraction. *For example: DYRK1A\_n*

[79] Genotype: control (c) or trisomy (t)

[80] Treatment type: memantine (m) or saline (s)

[81] Behavior: context-shock (CS) or shock-context (SC)

[82] Class: c-CS-s, c-CS-m, c-SC-s, c-SC-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m

The aim is to identify subsets of proteins that are discriminant between the classes.

the Ts65Dn has been used in preclinical evaluations of drugs and small molecules proposed as potential pharmacotherapies for ID in DS.

Ts65Dn mice have shorter life expectancies and show morphological, neurological, and structural abnormalities that parallel those found in patients with DS5,6,7,8,9,10,11,12.

 Ts65Dn mice also display behavioral abnormalities similar to those seen in DS.

Segmentally trisomic Ts65Dn mice provide a postnatal model for Down syndrome

The precise locations of the Chr 16 and Chr 17 breakpoints are 84,351,351 bp and 9,426,822 bp, respectively. The Chr 16 segment contains about two-thirds of the human Chr 21 homologs in the mouse, from mitochondrial ribosomal protein L39 (*Mrpl39*) gene to the distal telomere. These data were used to generate a PCR genotyping assay for Ts65Dn (Reinholdt et al., 2011), replacing the previous methods of chromosome analysis or qPCR. Northern and Western blotting, enzyme activity assays and reverse phase protein arrays (RPPA) demonstrate that some but not all genes in the translocation product are expressed at elevated levels in segmentally trisomic animals. RPPA shows a loss of correlation among some brain proteins (Ahmed et al., 2012)

<https://rstudio-pubs-static.s3.amazonaws.com/245066_f7b5962e8ab84594829b84f06ced39b6.html>

<https://www.kaggle.com/datasets/ruslankl/mice-protein-expression/code>

**Project proposal:**

1. What data did you choose? Briefly describe it. Identify the source and where it can be found (*e.g.*, URL)

**Ans:** Data chosen: Mice Protein Expression

The data comprises the expression values of **77** proteins and protein changes that generated audible signals in the nuclear fraction of the cortex. There are a total of 72 mice, including 38 control mice and 34 trisomics (Down syndrome) animals. 15 measurements of each protein per sample and mouse were made during the tests. As a result, there are 34x15, or 510 measurements for trisomic mice, and 38x15, or 570 measurements, for control mice. Each protein has a total of **1080** measurements in the dataset.

Dataset found: <https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression>

Dataset from the study: 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. [[Web Link]](doi:10.1371/journal.pone.0119491)

1. What are you planning to predict, *i.e.*, what is the target variable? Is it categorical (classification) or continuous (regression)? Is it a data mining task?

**Ans:** The target variable is “class”. The target variable is characterized based on categories like genotype, behavior, and treatment; in combination, the eight groups of mice are outlined. It is a categorical classification. Also, a data mining task.

1. How many rows (examples) and columns (variables) are in your dataset?

**Ans:** There are 1080 rows and 82 columns in the dataset

1. Which algorithms are you planning to use and why? Explain why they are appropriate given the data set, target variable, and features.

**Ans:** For machine learning analysis, we will be performing classification (multi-class classification) as well as clustering (based on behavior and treatment). We will use algorithms such as Naïve Bayes, Rpart, Random Forest, SVM, and logistic Regression for classification because they are all non-parametric and perform well on high-dimensional data. And for clustering, we will perform k-means clustering.

1. Do you expect to do any feature engineering, feature transformations, or other data set shaping? Why and for what purpose?

**Ans:** There is no need for feature engineering or transformation as the variables are independent and numeric. Although the dataset does contain some missing values which can be imputed systematically.

1. How will you evaluate the fit of the algorithms? Which specific evaluation methods will you use and why?

**Ans:** For evaluation, we will perform predictions, estimate accuracy and get confidence intervals. Target distribution using a confusion matrix. We will also draw out the ROC curve to check the model fits.

[**https://towardsdatascience.com/precision-recall-and-f1-score-of-multiclass-classification-learn-in-depth-6c194b217629**](https://towardsdatascience.com/precision-recall-and-f1-score-of-multiclass-classification-learn-in-depth-6c194b217629)

1. Has this analysis been done before? Links? Citations? What will you do differently?

**Ans:** Several analyses have been done on the data set, which you can find here [[Analyses],](https://www.kaggle.com/datasets/ruslankl/mice-protein-expression/code) but most ofthem have used python. Only a handful of them have used R, and out of that majority have conducted binary classification for either of the three classes (genotype, behavior, treatment). So, I plan to perform a multi-class classification with a series of algorithms and use an ensemble model to rate the best algorithm.

[**https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/**](https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/)