Data Mining Project

Mice Protein Expression

Charles FRANCHI

April 2024

Abstract

This report is about an analysis performed on the dataset named **Mice Protein Expression**[1].
We will present the dataset, how we deal with it, and the different results we found.
This report is related to a Github page where you can find the R code used for this project : `https://github.com/FRANCHI-Charles/R-project-2024`

Contents

1 Introduction

The Down's syndrome is a main problem of today's medical research. This genetic illness is not fully understood, and it is the subject of many contemporary studies.
As often in medical research, one of our possibilities to understand it is to study mice. Mice are known to have a similar genetic behavior than us, and we often use them to test a new drug or social behavior. Also, mice can also have an equivalent of Down's syndrome.
This is the idea behind the dataset **Mice Protein Expression**[1]. The goal is to detect the impact of a treatment, the Memantine, on the capacity to learn for a specimen suffering Down's syndrome[1].

2 Data Understanding

The dataset is composed of 1080 instances, each with 82 features.

The instances correspond to 15 measurement samples of 38 control mice (for thus a total of 570 instances) and 34 trisomic mice (510 instances). According to the authors of the dataset[1], each measurement can be considered as independant. However, we should be aware that there could be some bias due to the fact that we only have data from 72 mice, so measurements could be correlated.

The first feature is the mouse ID.
Then, 77 features is some proteins expression : a protein is an organic molecule coded with the DNA. They are vitals for living being and are the same along all organisms. The amount of each protein is regulated by different factors, such as the presence of other molecules, like drugs. The *expression level* of a protein is a value without unit, representing the different quantities of each protein in a cell. There is not clear defintion of these values and they makes sense only by comparing several values together.
The 4 last features are 3 classes :

1. Control mice[2] or trisomy mice.
2. Simulated to learn or not (they used *context-shock*, a technic with electronic stimulation of the brain to learn specific contexts[3]).
3. Treated with Memanine or not : the ones which weren't treated got saline injection (basically salt water).
4. The last class is the concateantion of the 3 previous one.

Finally, it is important to notice some missing values.
For the 38 first features, we have less than 18 instances with missing values in total. It represents less than 2% of the instances, so we can get rid of it.
For the rest, we have only 6 proteins with missing values, sum up in Table 1.

[1] Memantine is known for its effect against demancia, especially against Alzheimer's disease[4].
[2] Without Down's syndrome.

Protein	BAD	BCL2	pCFOS	H3AcK18	EGR1	H3MeK4
Number of Missing Values	213	285	75	180	210	270

Table 1: Important amount of missing values.

3 Data Preparation

3.1 Missing values

To deal with the features with an important number of missing values (Table 1), we will fill all the missing values with the mean of the column. In this way, we aim to introduce as less bias as possible.

Also, we notice there is no clear unbalancy in the dataset : each class have about 500 elements, no matter the class attributes.

3.2 Redundancies

Despite the multiple measurements per mouse, we do not have explicit reduduncies in the dataset between instances.

However, between features, we can notice high correlation :

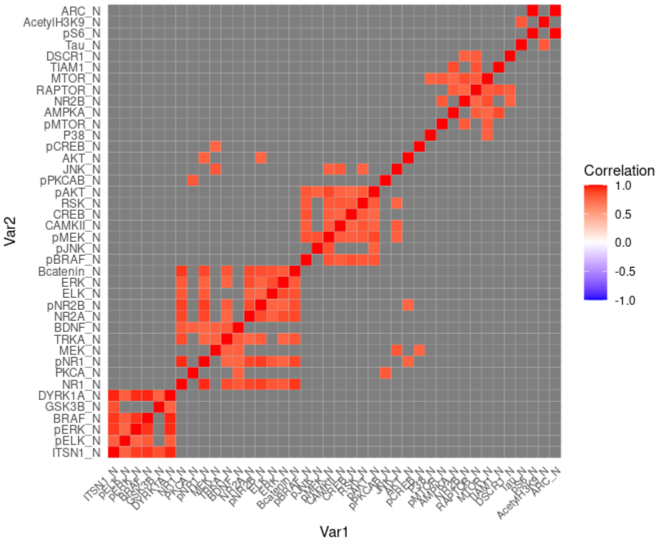

Figure 1: High correlated features. We only show features with more than 0.75 correlation in absolute value with at least one other feature in the dataset.

As shown in Figure 1, some protein expression are really dependant of each other. It can make sense, because some protein regulation system can impact several proteins at the same time. However, it means we don't need this features anymore in our dataset because it doesn't actually add new information. We will thus keep only one representative per "group" of highly correlated features, eliminating 32 features.

4 Modeling

4.1 Visualisation

To have an idea of our global distribution, we will do some PCA (Principal Component Analysis) plotting and a LDA (Linear Descriminant Analysis) to take into account the labels.

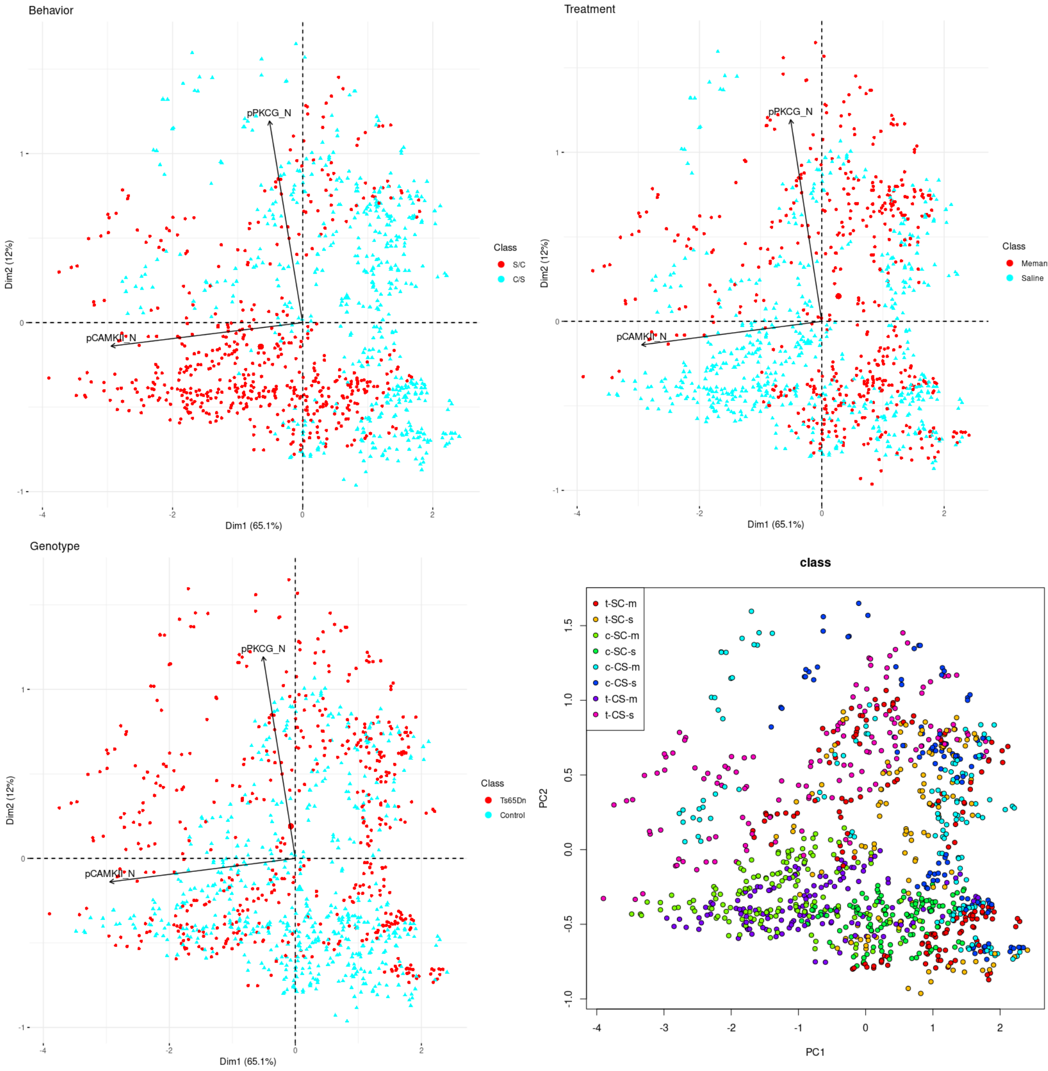

Figure 2: PCA plots, one by class attributes of the dataset

With the PCA plots, we can extract the major features of the dataset : there is two of them, the pPKCG and the pCAMKII.

The first is known to have several neuronal functions as synapse reinforcement[2] and the other to have an impact on long-term memory[2]. What is intersting here is that both are related with memory and both can be regulated with calcium. This suggests that Calcium may have a high correlation with both Behavior and Treatment, as the PCA plots in Figure 2 seem effective in discriminating these categories.

Also, the LDA plot (3) is intersting. It shows us we have some sub-classes which seems easily linearly separable (like t-CS-s and t-CS-m). This suggest than a SVM method for classification may be relevant.

4.2 SVM analysis

Support Vector Machine (SVM) is a method of supervised machine learning used for classification. Here, we will use the linear kernel, that means we do not project the data in higher dimensional space. The goal here is not to find a good classifier for the data. Of course, it could be usefull to have a method to know if a mouse have received a treatment

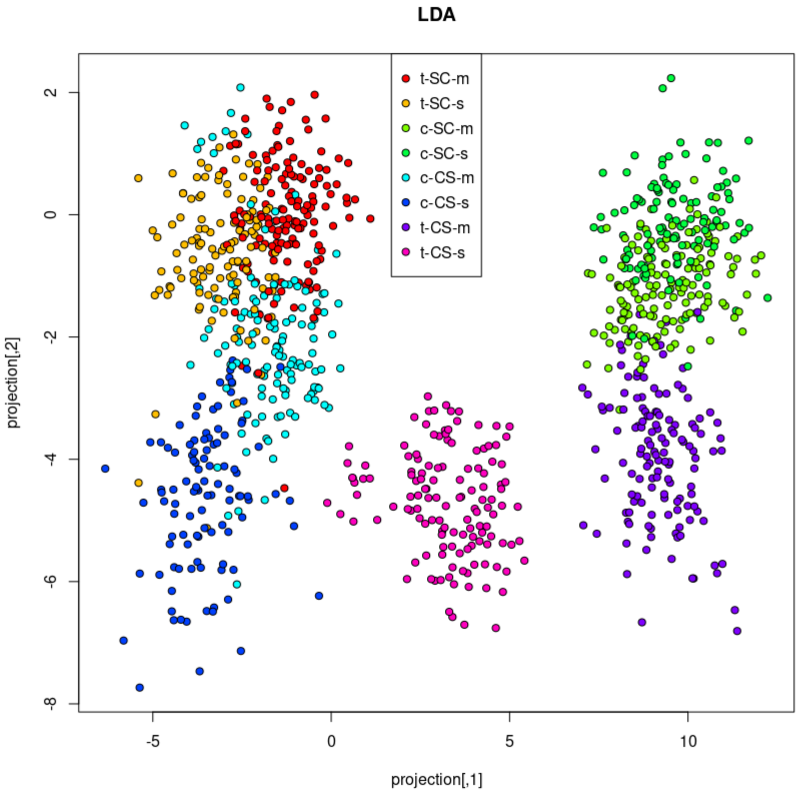

Figure 3: LDA plot on all the labels.

or not, or if it was simulated to learn, but we have in fact cheaper method than observing the DNA proteins to detect that.

Our goal is to train a robust SVM linear classifier, from which we can deduce the features that play a crucial role in the discrimination of different labels through their coefficients.

5 Evaluation

The SVM classifier performs well on the dataset.

For the class that combines Behavior, Treatment, and Genotype, we have achieved an accuracy of more than 98%.

If we look at the Treatment class, we notice we allready have more than 90% of accuracy. Thus, Memantine actually has a significant impact on protein distributions.

To get which proteins are the most impacted by the treatment, we can look at the different coefficient of the model : the higher[3] means the biggest impact on the classification, so the biggest difference between a mouse with or without the treatment. The principal proteins are sum up in Table 2.

Names	SOD1	pNR2A	pGSK3B	H3AcK18	pCAMKII	pCASP9	nNOS1
Weights	1.87	1.53	1.19	1.14	1.13	1.10	1.09

Table 2: Most impacted proteins by the treatment.

These proteins have similarity : SOD1, pGSK3B, pCAMKII, pCASP9 and nNOS1

[3] the higher in absolute value.

are known to have an impact on the brain, precisely in the memory process, and thus are related to Alzeihmer's Disease. Also, bad regulation of SOD1 is known to create complication of the Down's syndrome. pNR2A is also a protein related with the brain : it regulates some molecules between synapses[2].

All these results could have been expected, such the Memantine is a drug used because it has direct influence on synapses. This is however a good proof of the concrete impact of this drug.

Moreover, the effect zone of pNR2A and pCAMKII in the brain is influenced by calcium ion. Thus, has seen in Section 4.1, we may have an interesting correlation to study in Down's syndrome : the calcium ions may play a role on which we may have an influence.

6 Conclusion

The **Mice Protein Expression** dataset provides abundant information about the relationships between various proteins and mice. Through our study, we identify a clear impact of the Memantine drug on the brain and a potential study zone related to calcium impact. Eventually, it would be worthwhile to conduct a deeper study of the major proteins identified here in relation to Down's syndrome, to aid in the fight against this disease.

References

[1] Higuera C., Gardiner K., and Cios K. *Mice Protein Expression.* UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C50S3Z. 2015.

[2] National Library of Medicine. *National Center for Biotechnology Information.* URL: https://www.ncbi.nlm.nih.gov (visited on 04/10/2024).

[3] Stanford Medicine. *Fear Conditioning, Behavioral and Functional Neuroscience Laboratory.* URL: https://med.stanford.edu/sbfnl/services/bm/lm/bml-fear.html (visited on 04/02/2024).

[4] Cerner Multum. *Memantine Uses, Side Effects & Warnings.* 2023. URL: https://www.drugs.com/mtm/memantine.html (visited on 04/02/2024).