Clasificación de patrones de tratamiento de ratones para estímulo de aprendizaje con trisomía

Kaleb Alfaro, Email: *kaleb.23415@gmail.com,*

***Resumen*—En este proyecto se analiza la base de datos sobre la expresión de proteínas de ratones encontrada en la publicación [1]. El trabajo de los autores se oriento de asociar las dinámicas en la expresión de proteínas en ratones con síndrome de Down bajo tratamiento de memantina. En el siguiente proyecto, se elaboraron modelos de clasificación para desarrollar una herramienta capaz de distinguir los distintos factores vistos en la expresión de proteínas para asociar información del roedor. En este estudio se trató de un problema de clasificación de multi clase y se hayó que los modelos de desarrollados por SVC y una red neuronal artificial cumplen mejor los esquemas de clasificación.**

***Index Terms*—Machine Learning, Bioinformática, Redes neuronales, expresión de proteínas, Keras.**

I. Introducción

El síndrome de Down es una enfermedad genética causada por el surgimiento de una copia extra del cromosoma 21. Este problema también se le conoce como trisomía. Normalmente, se encuentran un total de 46 cromosomas dentro del genoma humano, pero en los pacientes afectados se encuentran un total de 47. Entre los afectados de este padecimiento, se reportan síntomas como desarrollo pobre del cuerpo y cerebro, por consecuencia dificultades para el aprendizaje. Esta enfermedad no ha podido ser curada en su totalidad con alguna clase de fármaco.

Por parte de los investigadores en [1], reportaron un análisis de ratones con este mismo padecimiento. Estos mismos cumplían la descripción de problemas para el aprendizaje y memoria. De acuerdo con [1], el tratamiento por dosis de memantina, ha ayudado a reducir los problemas de déficit intelectual en pacientes con trisomía. Estos tratamientos no han sido todavía acompañados por estudios relacionados con las dinámicas moleculares. En [1], se desarrolló el primer acercamiento objetivo de este tratamiento por medio de las dinámicas en las proteínas sobre el núcleo del cortéx.

La base de datos recompila 1080 muestras de 72 ratones, por cada uno se extrajeron información de 77 tipos de proteínas por medio de las señales detectables en el núcleo del cortéx cerebral. En la población de ratones, se distinguen dos tipos, 38 control (saludable) y 34 trisómicos (Síndrome de Down). En total se realizaron 15 mediciones durante periodos de tiempo. En la Figura 1, se muestra como se encuentra ordenada la información de la base de datos y la determinación de las etiquetas.

Alfaro es estudiantes del programa de maestría de ingeniería de electrónica con énfasis de sistemas empotrados de la Escuela de Ingeniería de Electrónica del Instituto Tecnológico de Costa Rica.

Este informe técnico expone los resultados del Proyecto Final del curso de Reconocimiento de Patrones del plan de ingeniería de electrónica con énfasis de sistemas empotrados.

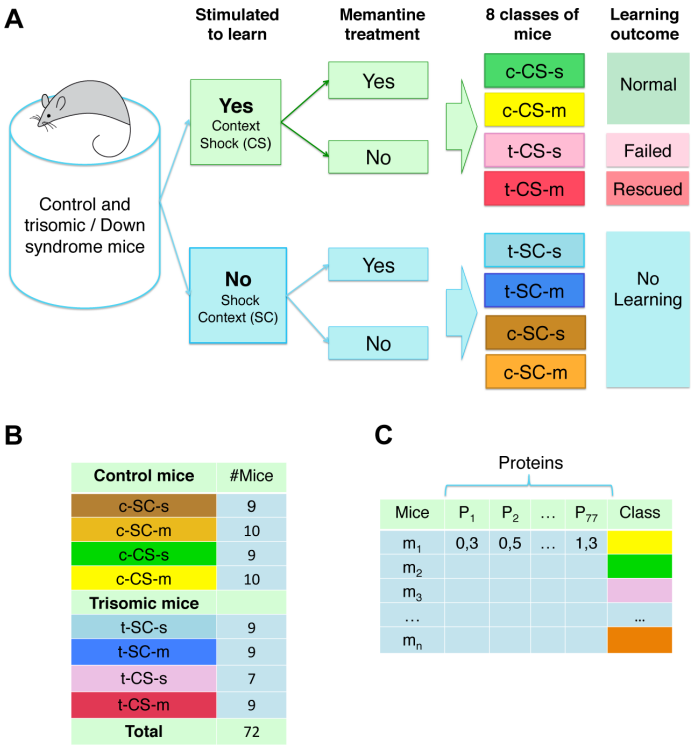

Figura 1. Descripción ilustrativa de la base de datos utilizada. En **A** se observa como se determina la clasificación de los ratones de acuerdo a su estímulo para aprender, genotipo y comportamiento. En **B**, se presenta como se identifican cada medición de acuerdo a la identificación del ratón y en **C** como se encuentra tabulada la información. Imagen tomada de [2].

De este estudio, se establecen 3 categorías principales:

1. **Genotipo:** control o trisómico.
2. **Tratamiento:** estimulación por memantina o salina (placebo).
3. **Comportamiento:** estimulados para aprender o no hubo estimulación aprendizaje.

II. Exploración de los datos

Como primera parte, se realizó un chequeo de la base de datos. De ella se encontró con valores faltantes dentro de ella. Estos datos se decidieron no descartarlos para la aproximación del modelo. Estos espacios fueron rellenados con las medias aritméticas de la base de datos. Se considera que esta técnica no debería alterar la clasificación de estas clases. En la Fig2, se muestran en los espacios oscuros, los valores faltantes de la base de datos.

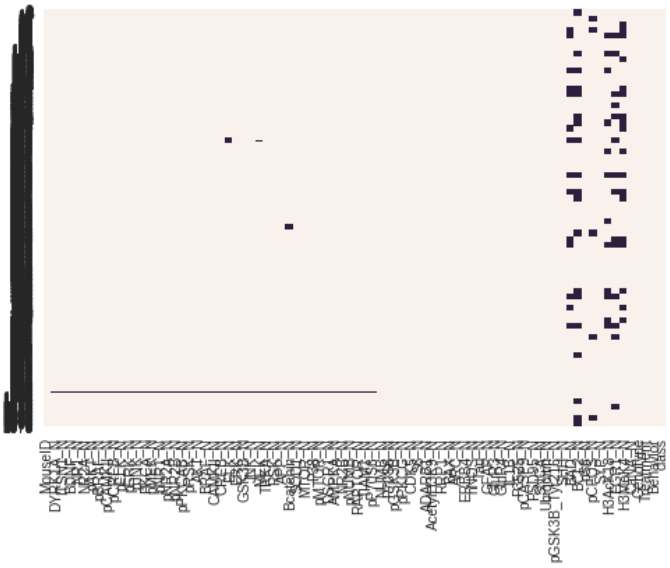

Figura 2. Visualización de los valores faltantes de la base de datos. Estos valores se encuentran relacionados con la expresión de proteínas de algunas mediciones.

III. Etapa de preprocesamiento de la base de datos

Antes de realizar los ajustes para los modelos de aprendizaje, primero se exploró reducir las dimensiones para cada medición. Esto con el fin de trabajar en vez de 77 atributos por instancia a un número menor. La técnica utiliza para este apartado, fue el análisis por componentes principales (**PCA**). En la Figura 3, se muestran los resultados obtenidos con este análisis hasta 10 componentes principales.

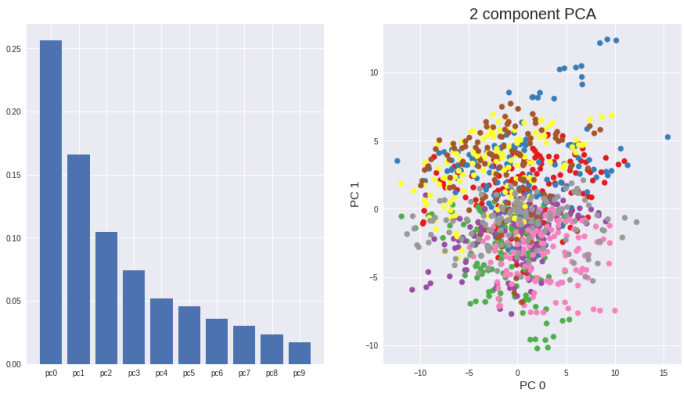

Figura 3. Visualización de los pesos de las componentes principales. Nótese como el 50 % de la representación de las covarianzas de los datos se estima con solo las primeras 3 componentes principales. Para visualizar la ubicación de los datos, se proyectaron sobre las primeras dos componentes principales como se muestra en la gráfica. En totalidad se muestran 8 clases/colores

Para tener un estimado de cuantas componentes principales se requieren como mínimo para poder tener buenas aproximaciones de clasificación, se desarrolló el experimento de clasificación con diferentes dimensiones. De esta forma, se concluyó que con 10 componentes principales es suficiente para determinar buenos modelos de clasificación.

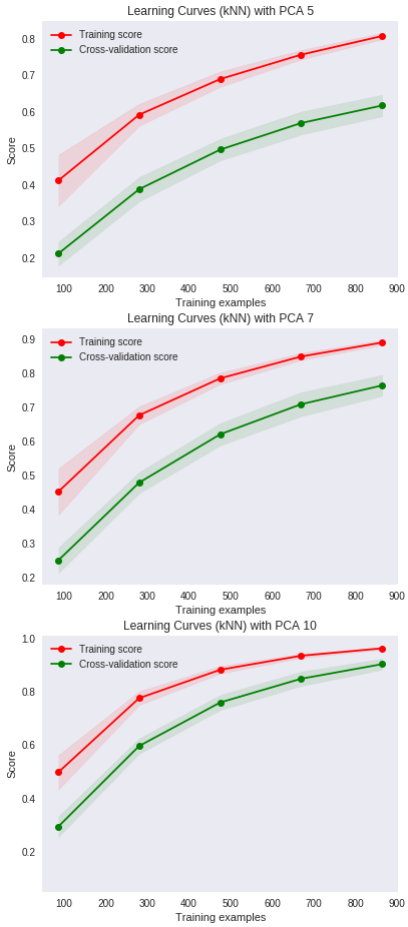

Figura 4. Comparación de aprendizaje utilizando KNN para distintas dimensiones. Utilizando métrica euclidiana.

IV. Modelos de Clasificación

Al inicio se separaron dos subgrupos de la base de datos, con una diferencia de 80/20 para el conjunto de datos orientado para entrenamiento y el resto pequeño para validación final de los algoritmos utilizados de clasificación. Los modelos de clasificación utilizados consistieron:

1. **KNN**: este fue parametrizado para 5 vecinos y métrica euclidiana.
2. **Decision Trees**: este fue asignado para una profundidad de 8 del árbol máxima.
3. **LinearSVC**: este fue elaborado con kernel RBF y $\gamma = 0{,}1$.
4. **Gaussian Naive Bayes**: no utilizó conocimiento a *priori*.
5. **Multi-Layer Neural Network**: implementación de la biblioteca de *scikit-learn*. 2 capas ocultas (9,8) neuronas, optimizador por LBFGS y $\alpha = 9 \times 10^{-5}$.
6. **Keras Multi-Layer Neural Network**: implementación de la biblioteca de *Keras*. Se asignaron 4 capas (15,12,10,8) neuronas, función de activación RELU y SOFTMAX para la última capa. La función de *loss* se

asignó `categorical_crossentropy` y optimizador *Adam* con los parámetros *learning rate*= 0,005, $\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$, $\epsilon = 0$ y *decay*= 0,0. La curva de aprendizaje por épocas se muestra en la Figura 5.

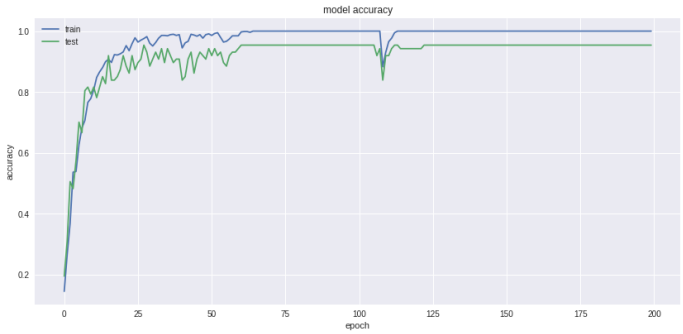

Figura 5. Curva de aprendizaje de la red neuronal desarrollada con la biblioteca Keras.

V. Comparación entre los modelos modelos

Para evaluar cada modelo elaborado, se utilizó el grupos de datos apartado desde el inicio con el fin de validarlos. Cada modelo elaboró sus predicciones y con ellos se elaboraron la matriz de confusión entre las 8 clases. Las matrices en sus valores normalizados se muestran en las Figuras 6, 7, 8, 9, 10 y 11.

De estas matrices, se puede observar como la clase con mayor confusión es la llamada `c-CS-m`. El modelo de *Gaussian Naive Bayes* y la *multi-layer neural network* se suelen confundir la clase `c-CS-m` con la `t-CS-m`. Mientras tanto los modelos de *KNN* y *SVC* apuestan por `c-CS-m` cuando se trata de `t-CS-s`. *Gaussian Naive Bayes* y *Decision Trees* suelen confundir `c-CS-m` con varias clases, aunque si predicen bien esta clase cuando se trata de esta. El mejor modelo con muy poca confusión, fue la red neuronal desarrollada con Keras y su mayor confusión se da con la clase `t-CS-s`. Esto demuestra que el clasificador utilizando 10 componentes principales es posible y suficiente para lograr un modelo suficientemente bueno.

Gaussian Naive Bayes puede tener malos resultados; pero al observar su matriz de confusión, se observa como si logra atinar bien los dos casos extremos de `c-CS-m` y `t-SC-s`. Quiere decir que la expresión de proteínas no se puede comprender como un problema de atributos independientes uno del otro, sino que están entrelazados entre ellos.

VI. Conclusiones

1. La red neuronal construida en Keras el el modelo más balanceado en precisión y confusión entre clases.
2. Por lo general, los modelos suelen confundir con la clase `c-CS-m` (control, estimulado para aprender, con tratamiento). La mayoría de modelos se confunde por `t-CS-m` (con síndrome de Down). Puede que la combinación de factores similares, causen gran similitud (medicamento, estímulo a aprender), para que se confundan.

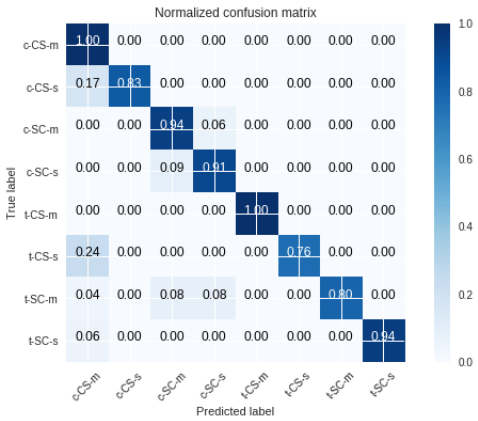

Figura 6. Matriz de confusión para el estimador de *KNN*

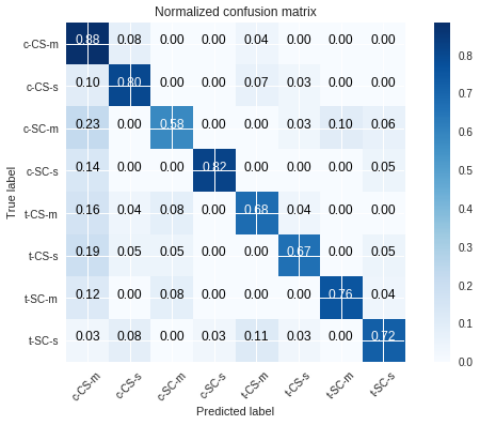

Figura 7. Matriz de confusión para el estimador de *Decision Tree*

3. La simplificación del problema utilizando componentes principales ha ayudado a desarrollar modelos de clasificación más eficientes y rápidos de procesar. Los resultados obtenidos pueden ser considerados buenos

Referencias

[1] Ahmed MM, Dhanasekaran AR, Tong S, Wiseman FK, Fisher EMC, Tybulewicz VLJ, et al. 2013. *Protein profiles in Tc1 mice implicate novel pathway perturbations in the Down syndrome brain.* Hum Mol Genet. 2013 May 1;22(9):1709–24.

[2] 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

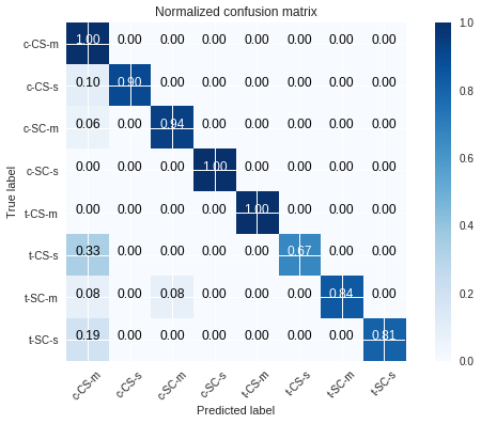

Figura 8. Matriz de confusión para el estimador de *SVC*

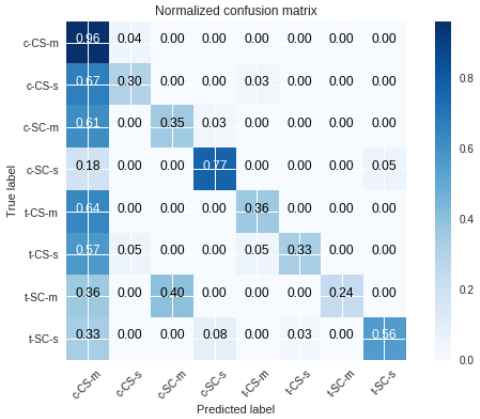

Figura 9. Matriz de confusión para el estimador de *Gaussian Naive Bayes*

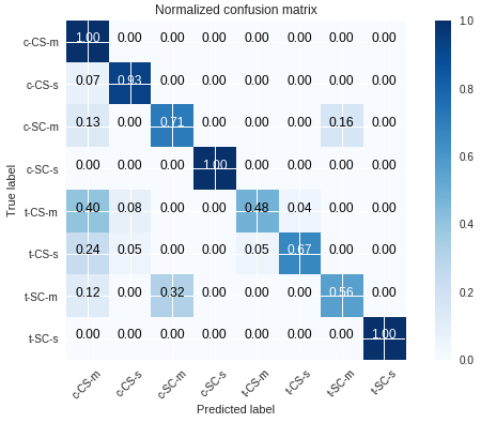

Figura 10. Matriz de confusión para el estimador de *Multi-layer neural network*

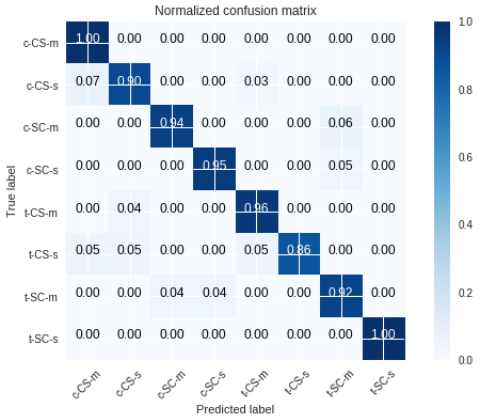

Figura 11. Matriz de confusión para el estimador de *Keras Multi-layer neural network*