

1 Conceptualización del problema

La enfermedad de Parkinson (EP) es un trastorno neurodegenerativo que afecta a las neuronas productoras de dopamina. Los síntomas parten desde temblor en extremidades hasta dificultad del habla y desplazamiento. De aquí surge la necesidad de diagnosticar lo antes posible la enfermedad para su tratamiento. El problema planteado en este estudio consiste en crear un modelo de red Neuronal (ML) capaz de clasificar y predecir con precisión pacientes que padezcan de la enfermedad del Parkinsons, utilizando datos relacionados a sus tonos y frecuencias vocales.

2 Materiales y Metodología

2.1 Preprocesamiento de datos

El conjunto de datos utilizados corresponden a un data set compuesto por mediciones de señales vocales de 31 personas, obtenidas por National Centre for Voice and Speech, Denver, Colorado. Hay 195 registros de señales de voz en total.

Para un correcto análisis, se revisa y limpia el dataset de datos que puedan entorpecer el aprendizaje. Se balancea el dataset de acuerdo a la variable de interés. Posteriormente se analiza la correlación entre variables de interés para eliminar aquellas características innecesarias (Anexo 1 y 2).

Luego se realiza un Principal Component Analysis (PCA) para reducir la dimensionalidad del dataset expresando las características más importantes (Anexo 3). Después se crean el grupo de entrenamiento y el de prueba (80% y 20% respectivamente). Finalmente los datos se estandarizan para nivelar orden de magnitud entre los datos y estos sean procesados correctamente por el modelo neuronal(Anexo 3).

2.2 Diseño de red Neuronal Secuencial

Se propone el diseño de red neuronal secuencial parametrizado, el cual los hiperparámetros se experimentan bajo ciertos rangos y opciones; funciones de pérdida, funciones de activación y regularizadores. El modelo posteriormente será evaluado con indicadores de rendimiento tales como el Loss, Accuracy, matriz de confusión y F1-Score.

3 Resultados

Luego de efectuar 72 iteraciones con las múltiples opciones de hiperparámetros acotados, se realizó un ranking con los modelos que presentaron mejores resultados. De estos, el mejor modelo se aprecia en la figura 1:

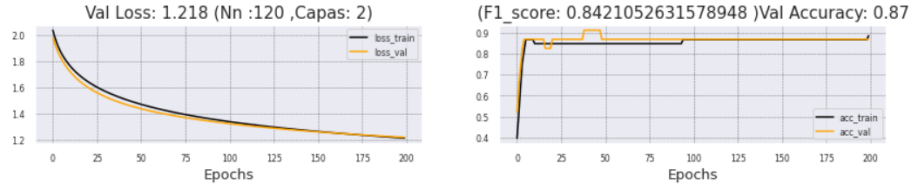


Figure 1: Resultados modelos neuronales secuenciales
Fuente: Elaboración propia

Con el modelo obtenido, se consiguen 0.1 falsos positivos y 0.0 falsos negativos, con un F1-Score de 0.84:

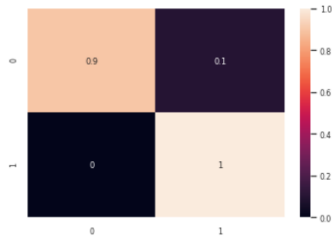


Figure 2: Matriz de confusión
Fuente: Elaboración propia

De esta manera, la arquitectura del modelo secuencial seleccionado es de 2 capas con 120 neuronas cada una, funcion de pérdida Binary Crossentropy, función de activación relu por cada neurona y sigmoide en capa final, junto con regularizador $l2 = 0.001$ con 200 epochs de entrenamiento. Este modelo clasifica de mejor manera los pacientes diagnosticados con Parkinsons (Anexo 4).

4 Conclusión

La red neuronal secuencial obtenida de 2 capas y 120 neuronas, con funciones de activacion relu, acompañada de regularización $l2$, permite efectivamente clasificar con gran precisión aquellos grupos de individuos que padecen la enfermedad de Parkinsons con solo sus registros vocales. A pesar de la poca cantidad de datos en el dataset, se espera que pueda generalizar bien con otro conjunto de datos, al poseer buenos indicadores como el accuracy de 0.87 y un F1-Score de 0.84

5 Anexos

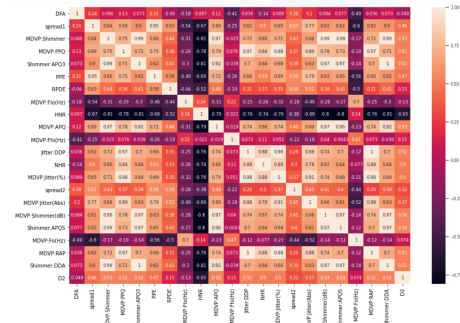


Figure 3: Matriz de correlaciones entre variables independientes
Fuente: Elaboración propia

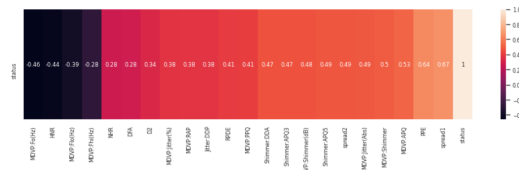


Figure 4: Correlación de variables v/s status
Fuente: Elaboración propia

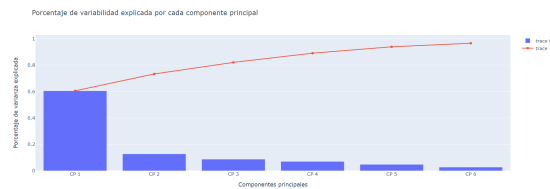


Figure 5: Principal Component Analysis (PCA)
Fuente: Elaboración propia

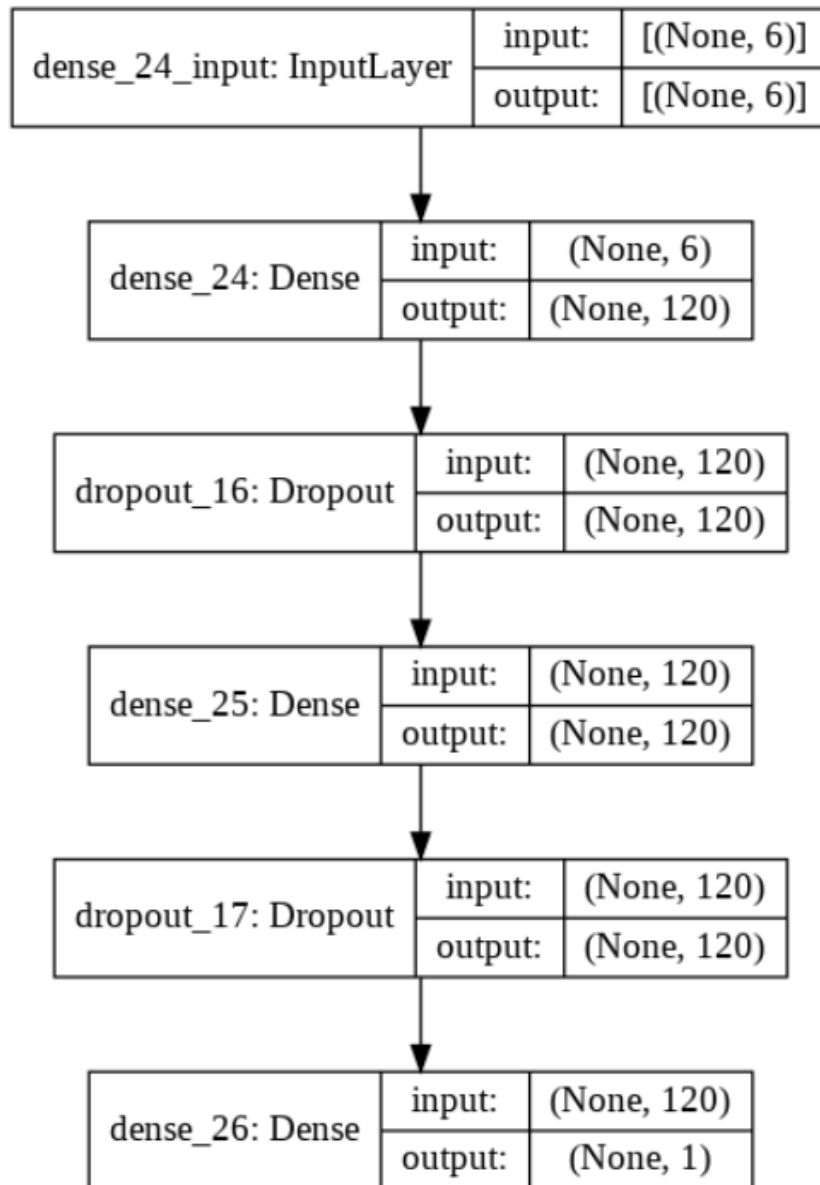


Figure 6: Resumen modelo neuronal
Fuente: Elaboración propia