Seminario Profesional

Christian Sosa - 22000102 | Pablo Siquinajay - 22006477 | Luis Pivaral - 21010442

1. QR

Link: https://l1nk.dev/XKfZs

2. Introduccion

El problema a resolver corresponde a una tarea de clasificación multiclase utilizando un modelo de inteligencia artificial supervisado. La información sobre las enfermedades dermatológicas a predecir se encuentra definida dentro del propio conjunto de datos. Este trabajo se enmarca en el ámbito de la salud, específicamente en el área de la dermatología, y busca apoyar el diagnóstico diferencial de enfermedades cutáneas mediante el análisis de características clínicas e histopatológicas de los pacientes.

3. Descripcion del dataset

El conjunto de datos contiene un total de 366 instancias, cada una representando un paciente. De los 34 atributos, 12 corresponden a características clínicas (como prurito, eritema o escamas), mientras que los 22 restantes son resultados histopatológicos obtenidos mediante análisis microscópicos (como la presencia de acantosis, paraqueratosis o infiltrado inflamatorio). Cada atributo está codificado como un valor entero que refleja la severidad o presencia del rasgo observado. La variable "age" es numérica, aunque presenta valores ausentes codificados con el carácter "?", lo que requiere tratamiento previo antes del entrenamiento del modelo. La variable objetivo o clase está codificada del 1 al 6, cada una correspondiente a un diagnóstico diferente como se muestra en la tabla. Este dataset es útil para tareas de clasificación multiclase y presenta un ligero desbalance en la distribución de clases, lo cual puede influir en el rendimiento del modelo si no se corrige adecuadamente.

Nombre de enfermedad	Valor
psoriasis	1
dermatitis seborreica	2
liquen plano	3
pitiriasis rosada	4
dermatitis crónica	5
pitiriasis rubra pilaris	6

Cuadro 1: Variables explicadas

4. Metodologia

1. Preprocesamiento
1.1 Manejo de valores faltantes: los valores ausentes en la variable "edad" (representados por "?") fueron tratados como nulos y las filas correspondientes eliminadas por ser pocas.
1.2 Normalización: se aplicó Min-Max scaling sobre las variables numéricas, excluyendo la clase, para reescalar los atributos en el rango [0, 1].
1.3 Transformación de la variable objetivo: las clases codificadas del 1 al 6 se transformaron al rango 0–5 para adaptarse a la función `SparseCategoricalCrossentropy`.
1.4 División del dataset: se usó un 80 % para entrenamiento y 20 % para prueba, reservando una fracción del entrenamiento para validación.
2. Arquitectura del Modelo
Se implementó una red neuronal secuencial con una capa de entrada de 34 nodos, dos capas ocultas densas (64 y 32 neuronas, activación ReLU, L2 y *dropout* del 30 % en la primera), y una capa de salida con 6 neuronas y activación *softmax* para clasificación multiclase.
3. Configuración del Entrenamiento
Se utilizó `SparseCategoricalCrossentropy` como función de pérdida, `Adam` como optimizador y precisión como métrica. Se aplicó balanceo de clases con `compute_class_weight`. Se incorporaron `EarlyStopping` (paciencia de 5 epochs) y `TensorBoard` para visualización.
4. Entrenamiento del Modelo
El modelo se entrenó hasta 100 epochs con *batch size* de 64, validación sobre el 20 % del entrenamiento y detención automática con `EarlyStopping` al no mejorar la pérdida de validación.
5. Evaluación del Modelo
El desempeño se evaluó sobre el conjunto de prueba usando `evaluate`, generando métricas de pérdida, precisión, matriz de confusión y reporte de clasificación (precisión, *recall* y F1-score por clase).

5. Resultados del modelo

No. NN	Descripción	Accuracy	Loss Function	Recall
NN 1	Conformada por 6 capas: 1 regularización 0.01, 2 Dropout (0.3), 2 de 32 nodos y la outlayer	0.9722	0.241	0.959
NN 2	Conformada por 4 capas: 128, 128, 64, 64. Con dropout de 0.3 cada dos capas y batch normalization luego de la segunda capa y al final.	0.9520	0.250	0.925
NN 3	Conformada por 4 capas, dos capas ocultas, con regularización L2 con parámetro alpha, batch size auto y optimización adam, ReLU y early stopping, tipo MLP.	0.9727	0.0058	0.9667

Cuadro 2: Tabla de resultados generales

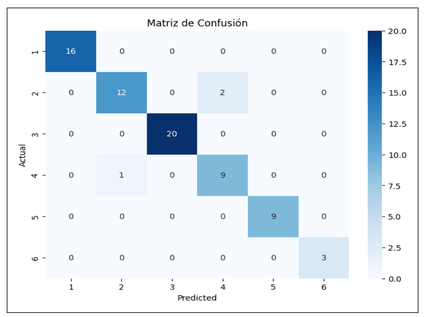

Figura 1: Matriz de confusión de **NN1**

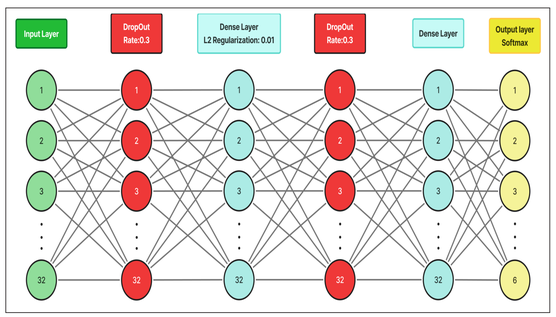

Figura 2: Diagrama de **NN1**

Figura 3: Gráfico Epoch Accuracy

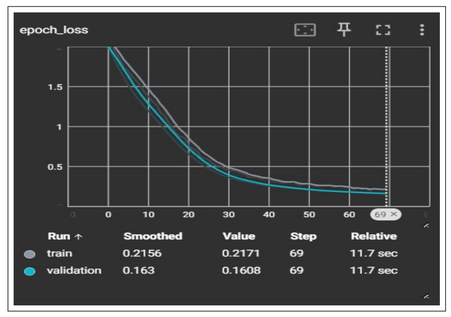

Figura 4: Curva de pérdida del entrenamiento

6. Conclusiones

El desarrollo del modelo de clasificación para enfermedades dermatológicas basado en redes neuronales artificiales demostró un desempeño altamente satisfactorio. Con una precisión de hasta 97.83 %, el modelo logró distinguir eficazmente entre seis tipos distintos de enfermedades eritemato-escamosas utilizando tanto características clínicas como histopatológicas.
Este resultado destaca la relevancia de un buen preprocesamiento de datos y de una correcta representación de las variables. A pesar de que los atributos del dataset ya estaban normalizados y eran de tipo entero, se requirieron ajustes importantes como el manejo de datos faltantes (en la columna de edad) y la adecuada codificación de la variable categórica.
Asimismo, se observó que el diseño de la arquitectura de la red neuronal influye considerablemente en el rendimiento del modelo. Diferencias en el número de capas ocultas, la función de activación y la tasa de aprendizaje afectaron directamente los resultados de cada red evaluada. Por lo tanto, se concluye que no solo los datos, sino también la estructura interna del modelo, juegan un papel fundamental en el éxito de la clasificación.
Finalmente, este proyecto reafirma el potencial de las redes neuronales como herramientas de apoyo en el diagnóstico médico, especialmente en campos como la dermatología, donde una clasificación certera puede significar un tratamiento más oportuno y adecuado para el paciente.

7. Mejoras a futuro

A pesar del éxito del modelo actual, existen múltiples líneas de mejora que podrían explorarse en futuras versiones del proyecto para aumentar su robustez, precisión y aplicabilidad en escenarios reales.
En primer lugar, una expansión del dataset sería fundamental. La incorporación de más muestras, así como de datos adicionales como imágenes dermatológicas reales, permitiría entrenar un modelo más generalizable y cercano a situaciones clínicas reales. También sería beneficioso incluir información sociodemográfica como grupo étnico, historial clínico o antecedentes familiares, lo que podría mejorar aún más la capacidad predictiva del modelo.
Desde el punto de vista de la variabilidad en los datos, sería interesante utilizar técnicas de data augmentation, incluso en datos tabulares, como el uso de algoritmos generativos (por ejemplo, GANs para datos clínicos), para mejorar la capacidad del modelo de adaptarse a nuevas muestras.
Finalmente, una mejora significativa consistiría en el desarrollo de una interfaz gráfica interactiva, dirigida a profesionales de la salud, donde puedan introducir los datos clínicos del paciente y recibir en tiempo real una predicción con su respectiva probabilidad y recomendación. Esta funcionalidad acercaría el modelo a su aplicación práctica como una herramienta de apoyo al diagnóstico médico.