Implementación de una feed forward network para predecir si una persona es propensa a padecer de problemas del corazón

José Alejandro López Quel
21001127
Universidad Galileo
Statistical Learning II

Abstract

En este paper, se presenta un caso de predicción de la posibilidad de llegar a tener problemas del corazón. Al determinar si una persona es propensa a padecer de problemas del corazón hay factores importantes como la presión arterial, uso de sustancias nocivas como alcohol y tabaco, obesidad, entre otras, los cuales pueden ayudar a predecir la existencia de futuras enfermedades. Por lo que en este trabajo, se analiza utilizar los datos antes mencionados para predecir estos comportamientos empleando una red neuronal del tipo feed forward.

1 Introducción

El aprendizaje profundo (deep learning) se ha utilizado para resolver problemas del mundo real en muchos ámbitos. La medicina no es una excepción. Aunque son controvertidos, se han propuesto y utilizado múltiples modelos con cierto éxito. Cada año mueren en Estados Unidos unas 610.000 personas por enfermedades del corazón, lo que supone 1 de cada 4 muertes. Las enfermedades del corazón son la principal causa de muerte tanto en hombres como en mujeres.

Se considera que los factores de riesgo tradicionales de los problemas de corazón son el colesterol LDL alto, la hipertensión arterial, la diabetes mellitus, el tabaquismo, los antecedentes familiares de problemas cardiacos, la edad, la obesidad y un estilo de vida poco saludable. Sin embargo, estos problemas pueden controlarse eficazmente con un cambio de estilo de vida y la adopción de hábitos saludables, y por lo tanto esto equivale a ahorrar el alto costo de tratamientos médicos y hospitalizaciones si se detecta a tiempo. Con la detección precoz de estos problemas, los pacientes pueden recibir una serie de tratamientos aconsejados por los médicos para reducir el riesgo de futuros problemas cardíacos y aliviar o controlar los síntomas. En este contexto los historiales clínicos pueden considerarse un recurso de información útil para ayudar a los médicos en la detección o la predicción de problemas del corazón.

Los avances en el aprendizaje automático y la inteligencia artificial han motivado a muchos científicos de datos a utilizar estas tecnologías en la detección precoz de enfermedades de alto riesgo del corazón como las cardiopatías. El aprendizaje automático aplicado a los historiales clínicos puede ser una herramienta útil para predecir el evento de cardiopatía coronaria con síntomas de enfermedades cardíacas, así como para explorar las características clínicas más significativas y los factores de riesgo que pueden causar eventos como un infarto o incluso la muerte. Los médicos pueden aprovechar el aprendizaje automático para clasificar las características clínicas y y desvelar correlaciones y relaciones ocultas y no evidentes entre los datos de los pacientes.

Por lo que, en este artículo se decide estudiar un dataset de 462 historiales médicos, los cuales pertenecen a pacientes de Sudáfrica. Los objetivos de este estudio son:

- Investigar y experimentar modelos de redes neuronales del tipo feed forward para predecir si las personas son propensas o no a padecer de problemas del corazón.
- Identificar el modelo de aprendizaje automático más eficaz que logra el mejor rendimiento de predicción en el conjunto de datos dado.

2 Factores para determinar problemas del corazón

Existen múltiples factores de riesgo que pueden contribuir al desarrollo de la cardiopatía isquémica. En la Tabla 1 se muestra un resumen de los factores más importantes. Aunque la edad, el sexo y los antecedentes familiares son factores que no pueden cambiarse ni controlarse, reconocerlos como factores de riesgo puede capacitar a la persona para tomar la iniciativa iniciativa para vigilar los factores controlables. El riesgo de problemas del corazón aumenta con la edad y los antecedentes familiares de enfermedades cardíacas. Los médicos controlan de forma rutinaria muchos factores de riesgo. Las lecturas de la presión arterial La presión arterial se mide en casi todos los encuentros clínicos. La presión arterial alta es uno de los efectos efectos secundarios de la restricción del flujo sanguíneo. La obstrucción arterial que restringe el flujo sanguíneo en los vasos sanguíneos de la sangre en los vasos sanguíneos provoca hipertensión debido a la mayor resistencia del flujo sanguíneo cuando el corazón bombea. Las mediciones de la presión arterial superiores a 140/90 se asocian a un mayor riesgo de enfermedad cardíaca.

Table 1: Factores de riesgo asociados a las enfermedades cardíacas

Factor de riesgo	Controlable
Presión arterial sistólica	Si
Tabaco acumulativo	Si
Colesterol LDL	Si
Antecedentes familiares	No
Obesidad	Si
Alcohol en sangre	Si
Edad	No

3 Descripción del dataset

El conjunto de datos para este estudio se ha obtenido de Sudáfrica, el cual es un subconjunto de un conjunto de datos más amplio. Tiene un total de 462 observaciones médicas (instancias) y 10 características, 9 como características clínicas independientes, y 1 es la variable objetivo, una clase binaria etiquetada como 0 o 1, es decir, se ha detectado un evento de problema relacionado al corazón, es decir, se ha detectado para las observaciones médicas como positivo o negativo. Los datos corresponden a un grupo de hombres de una zona de alto riesgo de alto riesgo de enfermedades cardíacas en Sudáfrica. Cada paciente de alto riesgo fue supervisado en el conjunto de datos y las características recuperadas fueron las siguientes: presión arterial sistólica (Sbp), tabaco acumulado en kg (Tobacco), colesterol malo también conocido como colesterol de lipoproteínas de baja densidad (Ldl), adiposidad, antecedentes familiares de enfermedades cardíacas (Famhist), comportamiento tipo A (TypeA), obesidad, consumo actual de alcohol (Alcohol), edad de inicio (Age) y enfermedad coronaria (Chd) (sí=1 o no=0).

3.1 Preprocesamiento

El conjunto de datos original se encuentra en formato .dat, por lo que se ha convertido a .csv, y se ha editado el nombre de las columnas para que sea más expresivas. Se codifica los valores de texto categóricos existentes en el conjunto de datos original en valores numéricos para que puedan ajustarse a los modelos de aprendizaje automático.

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
0	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	1
1	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
2	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
3	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	1
4	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	1

Figure 1: Dataset antes de procesamiento

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
0	160	12.00	5.73	23.11	1	49	25.30	97.20	52	1
1	144	0.01	4.41	28.61	0	55	28.87	2.06	63	1
2	118	0.08	3.48	32.28	1	52	29.14	3.81	46	0
3	170	7.50	6.41	38.03	1	51	31.99	24.26	58	1
4	134	13.60	3.50	27.78	1	60	25.99	57.34	49	1

Figure 2: Dataset después de procesamiento

3.2 División de datos en entrenamiento y validación

Se divide el dataset en 85% de los datos para el entrenamiento y el resto (15%) para validación.

```
X_train, X_validate, y_train_cat, y_validate_cat = train_test_split(x, y, test_size=0.15)
```

```
X_train.shape
```

```
(272, 9)
```

```
y_train_cat.shape
```

```
(272,)
```

```
y_validate_cat.shape
```

```
(48,)
```

Figure 3: División del dataset

4 Metodología y experimentación

La utilidad de un buen modelo de predicción de problemas del corazón depende en gran medida de su precisión y la estabilidad. Para lograrlo, se divide la investigación en tres experimentos, en cada uno se emplea distintas capas para obtener las predicciones más adecuadas y se evalúa tanto para el dataset de entrenamiento como para el dataset de validación.

4.1 Experimento 1

Se emplea un modelo sencillo de 2 capas, cada una de ellas empleando como activación ReLU. Se emplea *binary crossentropy* como función de perdida y *ADAM* como optimizador. Se emplea la métrica de *accuracy* para validar el modelo y se emplean *checkpoint* y *earling stopping* ambos

parámetros monitoreando el valor de *accuracy* para los datos de validación con el parámetro de paciencia en 1000 iteraciones. Para el entrenamiento se utilizan 500 *epochs* y un *batch size* de 10.

```
Model: "Experimento_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_96 (Dense)             (None, 12)                120
_________________________________________________________________
dense_97 (Dense)             (None, 9)                 117
_________________________________________________________________
dense_98 (Dense)             (None, 1)                 10
=================================================================
Total params: 247
Trainable params: 247
Non-trainable params: 0
_________________________________________________________________
```

Figure 4: Sumario del modelo para el Experimento 1

4.1.1 Resultados

Se obtiene que el valor de *accuracy* para los datos de entrenamiento es de 0.6985, mientras que para el set de validación es de 0.7708.

```
28/28 [==============================] - 0s 1ms/step - loss: 0.5844 - accuracy: 0.6985
5/5 [==============================] - 0s 2ms/step - loss: 0.5467 - accuracy: 0.7708
====================

Modelo: Experimento_1.h5

Train accuracy: 69.9%

Test accuracy: 77.1%

====================
```

Figure 5: Resultados para el Experimento 1

4.2 Experimento 2

Se emplea un modelo de 2 capas, cada una de ellas empleando como activación ReLU, la primera empleando 64 unidades, seguida de una capa de regularización de batch, posteriormente la segunda capa empleando 32 unidades seguido de un droput del 25 por ciento, por ultimo se obtienen los datos empleando sigmoid. Se emplea *binary crossentropy* como función de perdida y *ADAM* como optimizador. Se emplea la métrica de *accuracy* para validar el modelo y se emplean *checkpoint* y *earling stopping* ambos parámetros monitoreando el valor de *accuracy* para los datos de validación con el parámetro de paciencia en 1000 iteraciones. Para el entrenamiento se utilizan 2000 *epochs* y un *batch size* de 10.

4.2.1 Resultados

Se obtiene que el valor de *accuracy* para los datos de entrenamiento es de 0.9301, mientras que para el set de validación es de 0.9167.

```
Model: "Experimento_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_99 (Dense)             (None, 64)                640
_________________________________________________________________
batch_normalization_32 (Batc (None, 64)                256
_________________________________________________________________
dense_100 (Dense)            (None, 32)                2080
_________________________________________________________________
dropout_30 (Dropout)         (None, 32)                0
_________________________________________________________________
dense_101 (Dense)            (None, 1)                 33
=================================================================
Total params: 3,009
Trainable params: 2,881
Non-trainable params: 128
_________________________________________________________________
```

Figure 6: Sumario del modelo para el Experimento 2

```
9/9 [==============================] - 1s 2ms/step - loss: 0.2297 - accuracy: 0.9301
2/2 [==============================] - 0s 5ms/step - loss: 0.3666 - accuracy: 0.9167
====================

Modelo: Experimento_2.h5

Train accuracy: 93.0%

Test accuracy: 91.7%

====================
```

Figure 7: Resultados para el Experimento 2

4.3 Experimento 3

Se emplea un modelo de 4 capas, cada una de ellas empleando como activación ReLU, la primera empleando 16 unidades, seguida de una capa de regularización de batch y un dropout del 45 por ciento, posteriormente la segunda capa empleando 8 unidades seguido de un droput del 45 por ciento, despúes se utiliza una capa de 2 unidades seguida de una capa de regularización de batch y por ultimo se obtienen los datos empleando sigmoid. Se emplea *binary crossentropy* como función de perdida y *ADAM* como optimizador. Se emplea la métrica de *accuracy* para validar el modelo y se emplean *checkpoint* y *earling stopping* ambos parámetros monitoreando el valor de *accuracy* para los datos de validación con el parámetro de paciencia en 1000 iteraciones. Para el entrenamiento se utilizan 2000 *epochs* y un *batch size* de 20.

4.3.1 Resultados

Se obtiene que el valor de *accuracy* para los datos de entrenamiento es de 0.6801, mientras que para el set de validación es de 0.7917.

```
Model: "Experimento_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_102 (Dense)            (None, 16)                160
_________________________________________________________________
batch_normalization_33 (Batc (None, 16)                64
_________________________________________________________________
dropout_31 (Dropout)         (None, 16)                0
_________________________________________________________________
dense_103 (Dense)            (None, 8)                 136
_________________________________________________________________
dropout_32 (Dropout)         (None, 8)                 0
_________________________________________________________________
dense_104 (Dense)            (None, 2)                 18
_________________________________________________________________
batch_normalization_34 (Batc (None, 2)                 8
_________________________________________________________________
dense_105 (Dense)            (None, 1)                 3
=================================================================
Total params: 389
Trainable params: 353
Non-trainable params: 36
_________________________________________________________________
```

Figure 8: Sumario del modelo para el Experimento 3

```
9/9 [==============================] - 0s 2ms/step - loss: 0.6118 - accuracy: 0.6801
2/2 [==============================] - 0s 5ms/step - loss: 0.5827 - accuracy: 0.7917
====================

Modelo: Experimento_3.h5

Train accuracy: 68.0%

Test accuracy: 79.2%

====================
```

Figure 9: Resultados para el Experimento 3

4.4 Discusión de resultados

Se obtiene que el modelo que mejor se desempeña es el del Experimento 2, el cual es el que obtiene un mayor valor de *accuracy* tanto para los datos de entrenamiento como con los datos de validación. Se puede observar en las gráficas siguientes el comportamiento del aprendizaje para cada uno de los modelos. Se puede observar que para el Experimento 1 el *accuracy* se va incrementando y luego empieza a disminuir a partir de la iteración 400. Se tiene que para el Experimento 2 el *accuracy* incrementa con cada iteración y posteriormente se detiene para evitar *overfitting*. Por último, el Experimento e se puede observar que se incrementa de forma lenta y luego se mantiene el mismo nivel de *accuracy*.

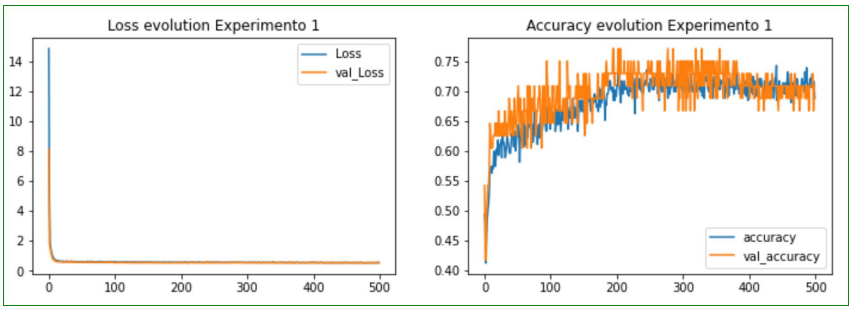

Figure 10: Gráficas de loss y accuracy para el Experimento 1

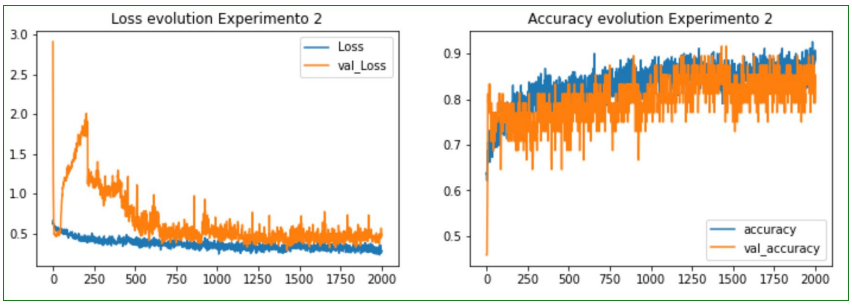

Figure 11: Gráficas de loss y accuracy para el Experimento 2

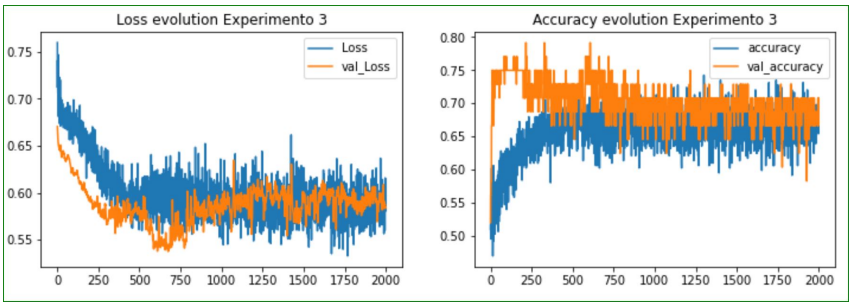

Figure 12: Gráficas de loss y accuracy para el Experimento 3

Empleando Keras se obtiene los siguientes resultados del reporte de clasificación, se puede observar que el valor de precisión para los valores 0 (Persona no propensa a padecer problemas del corazón) es de 0.87, mientras que para los valores de 1 (Persona propensa a padecer problemas del corazón) es de 0.68. Verificando el valor de recall y de f1-score se encuentra que todos son arriba de 0.71.

Los resultados demuestran que, con datos suficientes y características clínicas seleccionadas, las técnicas de aprendizaje automático son capaces de predecir la ocurrencia de eventos de problemas del corazón con un porcentaje de accuracy elevado. La selección de las características detalladas en este trabajo han demostrado y confirmado que las características clínicas y los factores de riesgo como el tabaco, el colesterol LDL, la presión arterial sistólica, la adiposidad y los antecedentes familiares se encuentran entre las características más importantes que ayudan a la detección temprana y a la predicción de la presencia de eventos de cardiopatía coronaria a partir de los registros médicos. Los médicos pueden aprovechar el análisis exploratorio de datos realizado en el conjunto de datos para mostrar las correlaciones y relaciones entre los datos de los pacientes.

El éxito del aprendizaje automático depende en gran medida de la riqueza de los datos que representan el fenómeno considerado. Aunque el conjunto de datos seleccionado tiene las características y los factores de riesgo más conocidos para predecir enfermedades relacionadas con el corazón, con un conjunto bastante rico de características, más datos y más variables pueden ayudar a

	precision	recall	f1-score	support
0.0	0.87	0.71	0.78	28
1.0	0.68	0.85	0.76	20
accuracy			0.77	48
macro avg	0.77	0.78	0.77	48
weighted avg	0.79	0.77	0.77	48

Figure 13: Resultados reporte de clasificación experimento con mejores resultados.

mejorar los resultados de la predicción. Si se dispusiera de otros conjuntos de datos externos con las mismas características y procedentes de distintas regiones, no solo limitado a un país en especifico, se podrían llegar a utilizado para validar los resultados obtenidos en este trabajo.

5 Trabajo futuro

Como trabajo futuro, se plantea aplicar el enfoque desarrollado en este trabajo de aprendizaje automático en otros conjuntos de datos de enfermedades cardiovasculares, cáncer y enfermedades infecciosas, ya que podrían llegar a obtenerse resultados que puedan llegar a ser favorables para el área de detección de enfermedades tempranas. También se plantea implementar los modelos obtenidos como servicio web e integrarlos en una aplicación para que los médicos puedan evaluar su utilidad en el mundo real.

Referencias

[1] Lakhani, P., Sundaram, B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 284, 574–582 (2017).

[2] Yasaka, K., Akai, H. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: A preliminary study. Radiology 286, 887–896 (2018).

[3] Indrakumari, R., Poongodi, T., Jena, S. R. (2020). Heart Disease Prediction using Exploratory Data Analysis. Procedia Computer Science, 173, 130–139.