Clasificación de pacientes con alta probabilidad de sufrir un ataque al corazón usando modelos simples de machine learning

Javier Pita (Universidad Carlos III, Madrid, España)

La finalidad de este trabajo es explorar técnicas sencillas de aprendizaje automático para un problema de pronosticación de que pacientes tienen alta probabilidad de sufrir un ataque al corazón. Para ello se ha enfocado el estudio en el uso de Regresores Logísticos y Árboles de Decisión para realizar dicha tarea.

1 Introducción

La medicina ha avanzado considerablemente y ha añadido modelos de aprendizaje automático e inteligencia artificial en general a su día a día. Existen muchas aplicaciones de estos modelos en el ámbito de la medicina: ayuda al diagnóstico y tratamiento, análisis genéticos, desarrollo de fármacos y vacunas, prótesis inteligentes y mucho más. Este trabajo podría estar enmarcado en la ayuda al diagnóstico o prevención. Aunque no se ha especificado en el enunciado de la práctica exactamente el uso que se le daría al modelo a optimizar podría ser útil para asignar más recursos de monitorización y control a aquellos pacientes con más probabilidad de sufrir un infarto, evitar ciertas operaciones que entrañen riesgo para el paciente o guiar al médico en la selección del tratamiento para evitar fármacos contraindicados a patologías del corazón.

El trabajo se ha dividido en una serie de secciones diferenciadas que guiarán al lector hasta la resolución y las conclusiones. Primero se expondrán los datos con los que se cuenta y un breve análisis de estos. Posteriormente se explicará el preprocesado de datos específicos para los modelos entrenados. Más adelante se verán los modelos usados y su evaluación. Finalmente se cerrará el trabajo con unas conclusiones y la explicación de un trabajo futuro.

2 Datos

Los datos para realizar la práctica han sido facilitados por el profesor de la asignatura y constan de 4984 tuplas de pacientes sobre los cuales tenemos una serie de atributos que ayudan a pronosticar la posibilidad de infarto.

2.1 Variables

Los datos constan de 12 variables muy diferentes:

ID Identificador del paciente. Rango: [1,4984]

Género Variable categórica que incluye el género del paciente. Posibles valores: Hombre y Mujer.

Tipo de trabajo Variable categórica que explica la ocupación del paciente. Posibles valores: Sector Privado, Autónomo, Funcionario y Menor.

Residencia Habitual Variable categórica que explica el tipo de entorno en el cual vive el paciente. Posibles valores: Ciudad y Rural.

Casado Variable categórica que explica el estado civil del paciente. Posibles valores: Si y No.

Edad Variable continua muestra la longevidad del paciente . Rango [0.08, 82.0].

Hipertensión Variable discreta que muestra si el paciente padece de esta condición. Posibles valores: 0 y 1.

Enfermedad del corazón Variable discreta que muestra si el paciente padece de esta condición. Posibles valores: 0 y 1.

Fumador Variable categórica que explica si un paciente fuma o ha fumado en el pasado. Posibles valores: Nunca ha fumado, Fumó en el pasado, Fuma o No Contesta.

Nivel de glucosa en sangre Variable continua mide la media de glucosa del paciente. Rango [55.12, 271.74].

IMC Variable continua mida el índice de masa corporal del paciente. Rango [14.0, 48.9].

Infarto Variable discreta que muestra si el paciente padece ha padecido un infarto. Posibles valores: 0 y 1.

2.2 Análisis univariante

Para el análisis univariante se han graficado cada una de las clases independientemente para ver de manera visual la relación de los datos. En función del tipo de variable se ha optado por diagramas de barra o histograma. Cabe destacar que en este punto uno se da cuenta que va a tener que lidiar con un dataset no balanceado en la variable de clasificación, esto supone un problema para los algoritmos de aprendizaje automático ya que pueden favorecer, si no se toman medidas de control, a la predicción de la clase más representada.

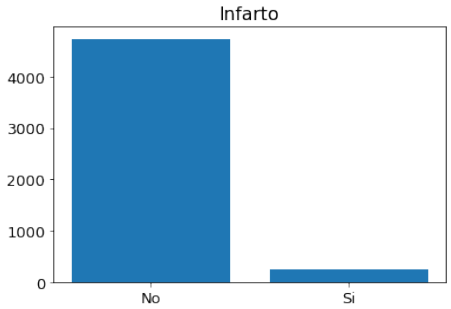

Figure 1: Clase Infartos en el dataset

Es interesante ver también los histogramas de las clases edad, nivel de glucosa medio e índice de masa corporal para confirmar que la distribución de las edades esta balanceada en cuanto a la cantidad, el índice de masa corporal puede asemejarse a una normal y en cambio el nivel medio de glucosa está completamente desplazado.

2.3 Análisis bivariante

El análisis bivariante pretende hacer un análisis visual de las variables en su relación con la clase infarto para poder pronos-

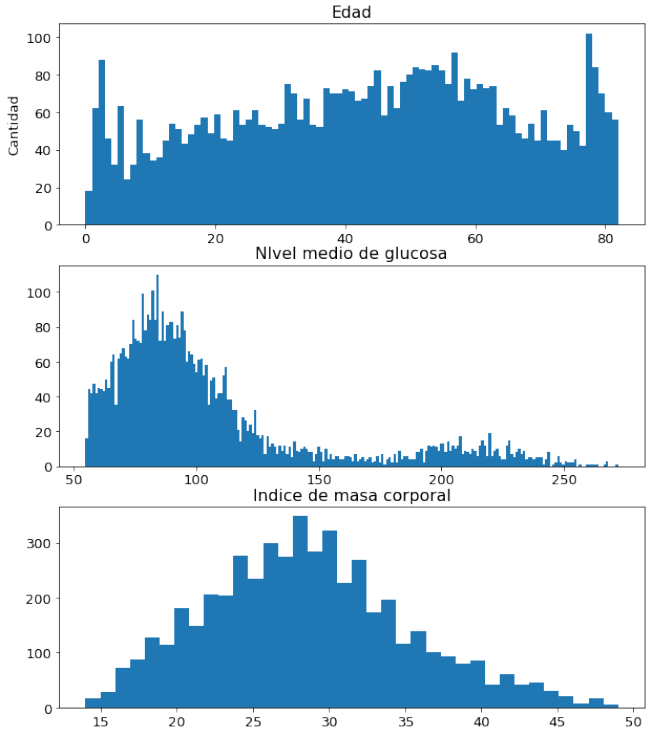

Figure 2: Histogramas de Edad, IMC y Glucosa en sangre

ticar como de bueno va a ser el modelo y tomar las decisiones necesarias para mejorar los resultados.

Lo primero realizado que se ha realizado ha sido la matriz de correlación donde hemos podido observar que la variable más correlada con el infarto es la variable edad seguida por orden de, la enfermedad de corazón, la hipertensión y el nivel medio de glucosa.

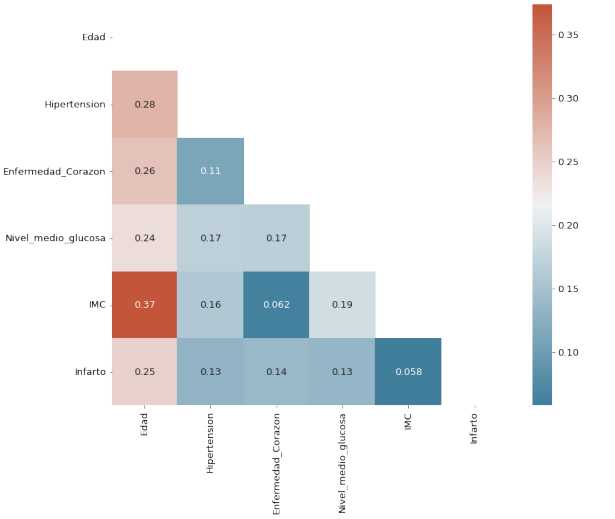

Figure 3: Mapa de correlación

Posteriormente se han realizado diagramas de barras de las variables categóricas y discretas con respecto a la variable infarto. De este análisis, que se puede ver en documento anexo de la entrega, podemos ver que ninguno de los posibles valores de las variables tiene una frecuencia similar con respecto a infarto.

Finalmente se han graficado boxplot de la edad, índice de masa corporal y nivel medio de glucosa en sangre y hemos podido apreciar en estos dos últimos que existe la posibilidad de descartar valores outliers. Finalmente se ha decidido no hacer para mantener la integridad del dataset de entrada.

En las figuras podemos ver el caso de la edad con escasos outliers para pacientes con infarto en contraposición a IMC que presenta outliers en ambos posibles valores de infarto.

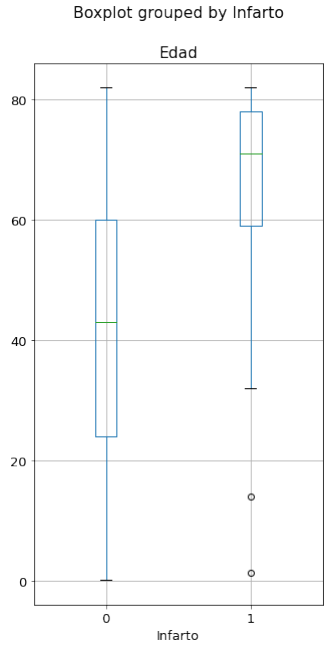

Figure 4: Boxplot de Edad

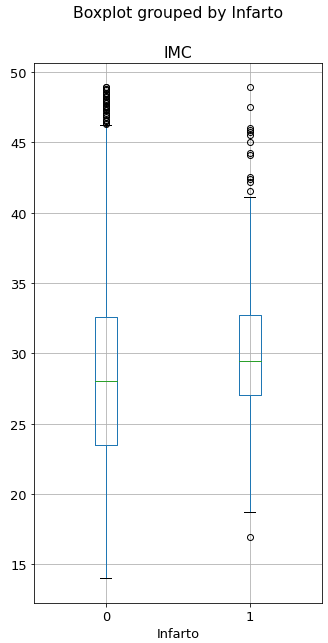

Figure 5: Boxplot de IMC

3 Pre-procesado de datos

El preprocesado de datos ha variado para cada uno de los modelos con los cuales se quería pronosticar la posibilidad de infarto. Aun así, ha existido una previa comprobación de

atributos nulos en algunas de las variables, la eliminación de la variable ID que no aporta nada al modelo y puede hacer que se identifique la relación numero de paciente infato en el entrenaiento, así como un redondeo de la variable Edad ya que se considera insignificante la diferencia de unos meses para pronosticar la probabilidad de padecer un infarto. Una persona de 34 años puede tener diferente probabilidad que una de 34,9 pero esa diferencia no radica en la edad.

3.1 Preprocesado Regresión logistica

Para presentar los datos a la regresión logística se ha realizado un One-Hot-Encoding de las variables Genero, Tipo de Trabajo, Residencia Habitual y Fumador. Para el caso de Casado se ha traducido el Si y No a 0s o 1s y las variables continuas han sufrido una normalización entre sus respectivos valores máximo y mínimos.

3.2 Preprocesado Árboles

Para presentar los datos a los árboles se ha seguido una aproximación similar que la anterior exceptuando para las variables continuas. En este caso se ha decidido no normalizar sino dejar las variables como vienen exceptuando el redondeo. Además, en este punto se ha probado a generar un nuevo dataset con estas tres variables discretizadas para comparar el comportamiento de los árboles generados con ambos datasets.

Con el fin de hacer la discretización de Edad, IMC y Glucosa en sangre la decisión se ha apoyado en unas figuras de análisis bivariante llamadas diagramas de violín que nos ayuda a separar la muestra en componentes en base a la cantidad de tuplas en la muestra y con información del dominio en cuestión.

Se va a poner el ejemplo del IMC y Edad como muestra, para más detalle de cada discretización acudir al notebook anexo a la entrega.

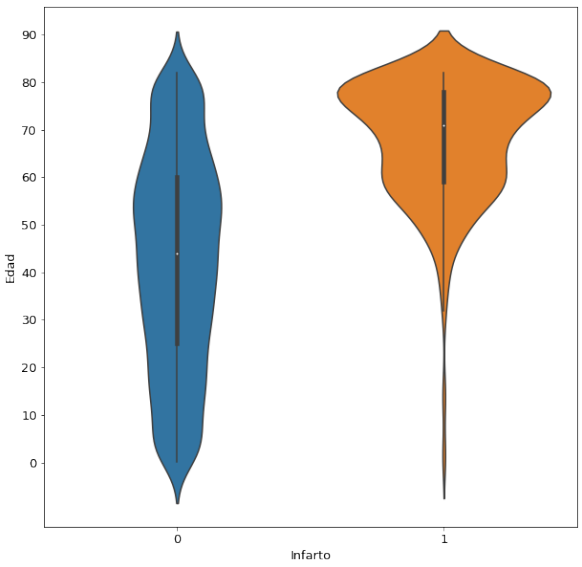

Figure 6: Violin de Edad

En el caso de la edad podemos apreciar que agrupar la muestra para el caso de infarto es sencillo ya que se generan en las partes altas dos conjuntos diferenciados. Para la parte baja, nos hemos fijado en No Infarto, ya que Infarto carece de muestras, y se ha tomado la decisión de dividir en dos subconjuntos por conocimiento del dominio, pero uno único hubiera sido valido también. Los rangos sobre los que se ha discretizado son: [0,18]; (18,45]; (45,70]; (70,adelante].

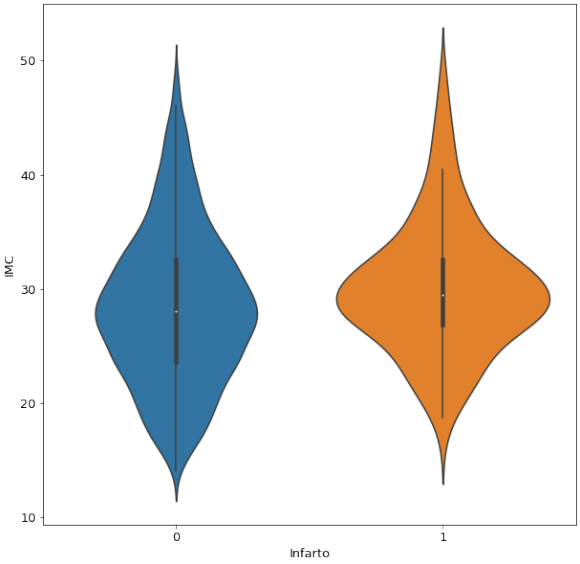

Figure 7: Violin de IMC

Para el caso del IMC nos hemos basado en la división grupos de sanidad. Esto grupos son Bajo Peso, Normal, Sobrepeso, Obesidad T1, Obesidad T2 y Obesidad T3. Al analizar la muestra se podía ver esas mismas diferencias en la parta derecha de la figura. No se ha visto necesario dividir en tantos conjuntos, pero se han usado los valores de referencia obteniendo los siguiente conjuntos: [0,24]; (24,35]; (35, en adelante].

IMC	DIAGNÓSTICO
BAJO PESO	<18.5
PESO NORMAL	18.5-24.9
SOBREPESO	25-29.9
OBESIDAD TIPO 1	30-34.9
OBESIDAD TIPO 2	35-39.9
OBESIDAD TIPO 3	≥40

Figure 8: Condicion en función del IMC

Finalmente añadir que existía la posibilidad de reducir el porcentaje de individuos catalogados como No Contesta para la variable fumador basándose en las leyes y estudios de inicio medio de estos hábitos, aunque se ha descartado. La mayoría de pacientes por debajo de 18 tienen asignado a esta variable No Contesta cuando la mayoría de ellos podría reevaluarse como Nunca ha fumado.

3.3 División del dataset

El dataset se ha dividido para todos los entrenamientos y las pruebas en dos subconjuntos manteniendo el desbalanceo de la clase a pronosticar. Se podría haber dividido en tres (entrenamiento, validación y test) para que la decisión del mejor modelo no estuviera sesgada por la busqueda de hiperparámetros, pero se dejó claro que estaba fuera del scope de esta práctica y se implementaría en la final.

3.4 Balanceo del dataset

Una de las conclusiones del análisis univariante era la necesidad de atajar el problema de un dataset desbalanceado en la clase a clasificar. Para ello se han seguido 4 técnicas diferenciadas y se han utilizado todas ellas para los árboles y el regresor logístico. Estas técnicas son:

Balanceo El balanceo se base en asignar un peso durante el entrenamiento al conocimiento extraído de las tuplas de cada tipo de manera que se aprenda más de la tuplas de la clase menos representada para optimizar el modelo. Estos pesos se basan normalmente en la frecuencia relativa de los valores de la clase a pronosticar, pero se puede variar y se han realizado pruebas con valores diferentes. Es importante puntualizar que no se generan nuevas tuplas en este approach.

Oversampling Esta técnica consiste en crear nuevas instancias de la clase menos representada para igualar a la clase más representada. Esta técnica genera nuevos datos por lo que solo se puede hacer en el dataset de entrenamiento, no en el de test ya que dañaría la realidad de los resultados. Existen diversas técnicas, en este trabajo se han probado la técnica aleatoria (RandomOverSampling) y el SMOTE. En la primera se duplican de manera aleatoria las tuplas mientras en la segunda se usa el modelo de KNN para generar tuplas de manera artificial.

Undersampling Por el contrario esta técnica consiste en sustituir o eliminar instancias de la clase más representada para igualar a la clase menos representada. Esta técnica elimina datos reales por lo que solo se puede hacer en el dataset de entrenamiento, no en el de test ya que dañaría la realidad de los resultados. Existen diversas técnicas, en este trabajo se han probado la técnica aleatoria (RandomUnderSampling) y la de centroides. En la primera se eliminan de manera aleatoria las tuplas mientras en la segunda se usa el modelo de KNN para generar tuplas de manera artificial sustituyendo las tuplas por los centroides de los grupos.

Soluciones Mixtas Estas soluciones han mezclado el balanceo con Undersampling y Oversampling para disminuir el error que producen. En las pruebas se ha dejado que la mitad de la diferencia de clases la corrija la técnica de balanceo mientras que la otra mitad generando o eliminando tuplas.

4 Modelos

Se han entrenado en total 50 modelos diferentes combinando las cuatro técnicas de balanceo del dataset mostradas en el apartado anterior con tres clasificadores: Regresión Logística, DecisionTree y Random Forest. Los dos primeros han sido vistos en clase, pero el ultimo no. El ultimo algoritmo computa una serie definida de árboles mediante una sub-selección de los datos de entrada. La salida se pronostica con el conjunto completo de árboles. La salida de cada árbol se llama voto y la salida final del modelo es la clase más votada. En esta práctica no se ha buscado el número de árboles optimo, sino que se ha elegido un tamaño base para realizar todas las pruebas.

5 Evaluación

Con el fin de evaluar los modelos se han usado las métricas y figuras vitas en clase. Para cada uno de los modelos presentados se ha computado:

Matriz de confusión La matriz de confusión es una matriz cuadrada con tantos valores como posibles valores de la clase a predecir que muestra como ha acertado o se ha equivocado y de qué forma un modelo. En nuestro caso muestra las instancias del test clasificadas correctamente en la diagonal principal y las incorrectas en la segunda diagonal. En función de estos datos se computan los siguientes valores.

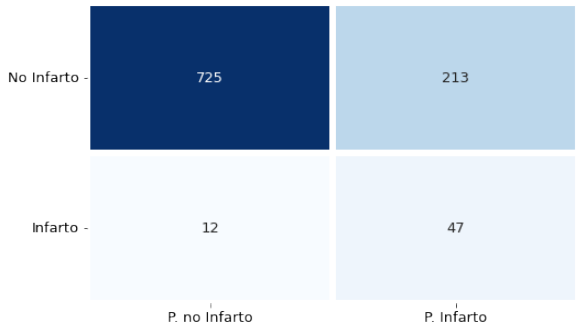

Figure 9: Matriz de confusión ejemplo

Accuracy Porcentaje de instancias correctamente clasificadas

Precision Porcentaje de verdaderos positivos con respecto a las equivocaciones de esa clase.

Recall Porcentaje de las instancias realmente positivas clasificadas de esa forma.

F1-Score Valor que une las dos medidas previas

Kappa Mismo valor que el accuracy pero corregido con la probabilidad de acertar de manera aleatoria.

Curva Roc Grafica la sensibilidad frente a la especificidad

Auroc Área bajo la curva Roc

Entre todas las que se han mostrado para cada modelo presentado se ha decidido usar F1-Score, Kappa y Auroc como las más importantes para la selección de los mejores modelos.

5.1 Mejores modelos

Es complicado determinar qué modelo es el mejor sin entender las circunstancias sobre las que se va a usar dicho modelo. No es lo mismo utilizar el modelo para asignar recursos de vigilancia sobre ciertas personas en base a la salida que un sistema de ayuda para médicos pronosticando quien puede padecer un infarto en el transcurso de una operación. Como no tenemos el contexto general del problema se han elegido

los algoritmos en base a las métricas previamente descritas. De entre los 50 modelos entrenados se han escogido los 5 mejores. Los 5 mejores modelos son todos modelos de regresión lineal entrenados bajo diferentes técnicas o datos. Los 5 primeros incluyen: modelo únicamente balanceado, solución mixta de balanceo y SMOTE, OverSampling aleatorio, únicamente SMOTE con diferentes valores del conjunto de vecinos (K)

MODELOS	F1_score	Accuracy	Kappa	AUROC
LR_balanced	0,294671	0,774323	0,219367	0,831719
LR_geldedsmote_balanced	0,294671	0,774323	0,219367	0,831900
LR_randomover	0,293750	0,773320	0,218289	0,832677
LR_smote_k_9	0,290221	0,774323	0,214562	0,832081
LR_smote_k_6	0,289308	0,773320	0,213492	0,832243
LR_smote_k_5	0,286604	0,770311	0,210320	0,831358

Figure 10: Tabla de mejores modelos

Para encontrar el mejor modelo de tipo árbol hay que irse a la posición número quince donde encontramos el árbol normal entrenado sobre el dataset no discretizado y con las instancias no balanceadas. En la siguiente figura podemos ver los seis primeros árboles entrenados y sus respectivas posiciones con respecto a mejores algoritmos.

Posicion	MODELOS	F1_score	Accuracy	Kappa	AUROC
15	SimpleTree	0,250000	0,909729	0,201989	0,602598
16	Tree_smote_k_4	0,251969	0,904714	0,251969	0,607875
19	Tree_smote_k_8	0,222222	0,901705	0,222222	0,590392
20	Tree_geldedsmote_balanced	0,217054	0,898696	0,217054	0,588920
22	Tree_randomover	0,198020	0,918756	0,198020	0,567688
23	RF_randomunder_discretized	0,237288	0,729188	0,206349	0,794767

Figure 11: Curva Roc mejores 16 modelos

La siguiente figura muestra la curva ROC para los dieciseis primeros modelos y se puede apreciar que aunque con pequeñas diferencias en algunos puntos todos los modelos presentan el mismo área.

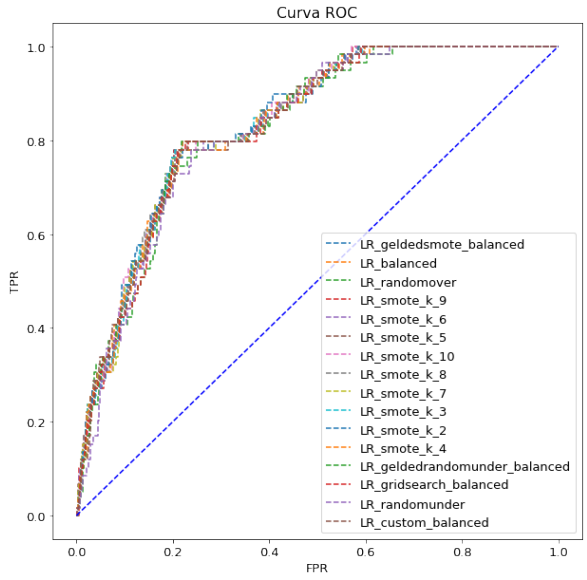

Figure 12: Curva Roc de todos los modelos

6 Conclusiones

Es fácilmente apreciable que los modelos presentados no han sido capaces de predecir de manera correcta la posibilidad de infarto implicando que, o otras técnicas más complejas son necesarias para solventarlo o un análisis más optimizado de las presentadas era necesario. Una de las vías no exploradas para optimizar los algoritmos es analiar las instancias incorrectamente clasficadas y ver patrones o similitudes con otros conjuntos con el fin de entender la motivación y ponerle remedio. Cabe puntualizar que la selección del modelo más oportuno una vez realizado el análisis depende mucho del uso y el entorno en el cual se ha planteado.

7 Experimentación

Fuera del scope de esta práctica se ha probado una solución ensemble unificando los dos mejores clasificadores (uno logístico y uno de árbol) en un único modelo que tiene como entrada la probabilidad asignada en la salida de los modelos anteriores. Este modelo unifica el conocimiento adquirido por los clasificadores anteriores y pronostica una salida "mas correcta". La idea era corregir el error de ambas y presentar un modelo más correcto pero no ha sido correcto. Más detalles de la arquitectura y el entrenamiento de este modelo se pueden encontrar en el notebook anexo en la sección de Experimentación.