Modelo predictivo para validación de datos sintéticos clínicos en diabetes

Javier I. Ramírez[1,6], Francisco M. Ortuño[1,2], Carlos Loucera[2,3], Joaquin Dopazo[2,3,4,5]

[1] Departamento de Ingeniería de Computadores, Automática y Robótica, Universidad de Granada (UGR), Granada, España.

[2] Plataforma de Medicina Computacional, Fundación Progreso y Salud (FPS), Sevilla, España.

[3] Medicina Computacional de Sistemas, Instituto de Biomedicina de Sevilla (IBiS), Sevilla, España.

[4] Bioinformática en Enfermedades Raras (BiER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Seville, Spain.

[5] FPS/ELIXIR-ES, Virgen del Rocío Hospital, Seville, Spain.

[6] FIDESOL- Fundación en Investigación y Desarrollo del Software Libre, Granada, España.

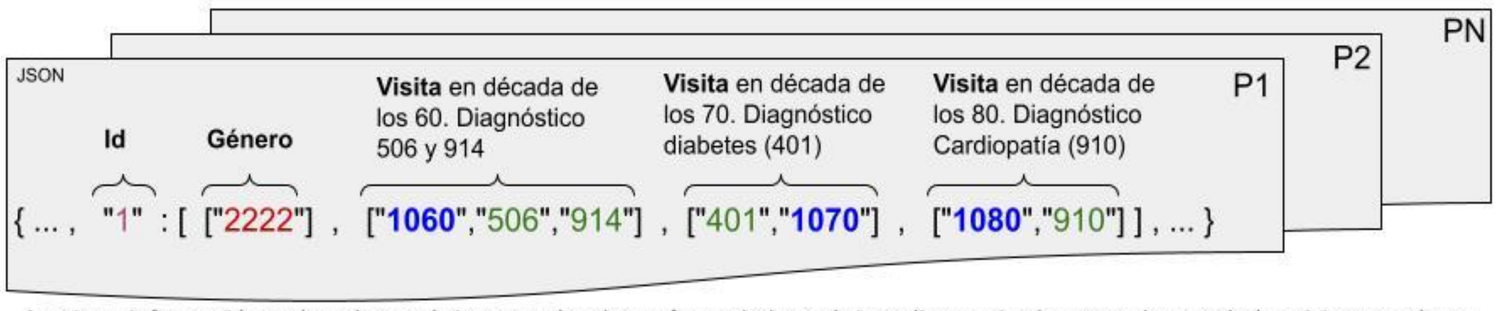
OBJETIVO

En este trabajo se desarrolla un sistema de predicción de patologías médicas utilizando datos secuenciales de diagnósticos de pacientes diabéticos. La meta principal es mejorar la precisión en la detección temprana de diversas enfermedades, lo que podría tener un impacto significativo en la toma de decisiones clínicas y en la prevención de complicaciones graves. Este estudio se centra en el análisis de datos médicos temporales generados por una GAN basada en datos reales obtenidos por la BPS, permitiendo capturar las dinámicas y patrones que preceden a la aparición de patologías.

Dataset pacientes diabéticos

	Cantidad
Pacientes	1.215.974
Códigos Patologías	83

Dataset proporcionado por uno de los retos propuestos por el congreso anual **CAMDA** (Critical Assessment of Massive Data Analysis). En este reto, se proporciona un dataset sintético que trazan las trayectorias de diagnóstico de pacientes diabéticos. Estos registros fueron creados mediante Autoencoders Duales Adversarios entrenados con datos de 1,2 millones de pacientes reales de la Base Poblacional de Salud (BPS) del Sistema Andaluz de Salud.



Contiene información ordenada en el tiempo sobre las enfermedades crónicas diagnosticadas en cada una de las visitas que haya realizado cada paciente. Además, también hay información sobre las décadas en las que se han hecho cada visita.

PACIENTES ELIMINADOS:
- Pacientes sin diabetes
- Pacientes con la patología objetivo antes del diagnóstico de diabetes.
- Pacientes que tienen menos de X (configurable) patologías antes de diabetes.

METODOLOGÍA

PROCESAMIENTO DE LA INFORMACIÓN — 1

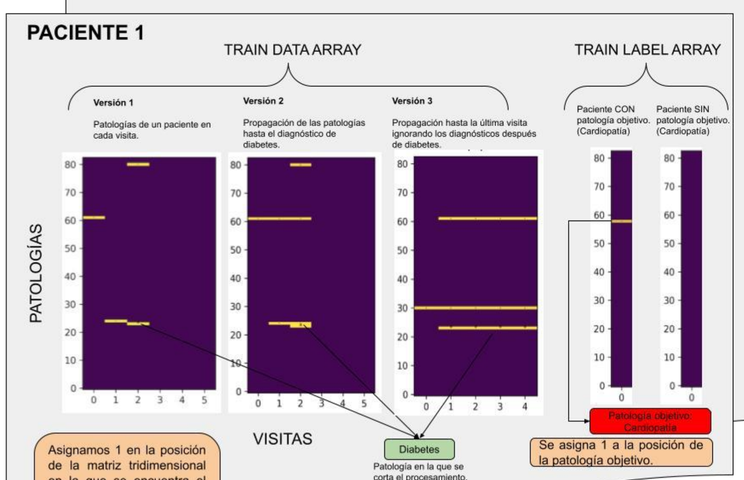
PACIENTE N

PACIENTE 2

PACIENTE 1

TRAIN DATA ARRAY

TRAIN LABEL ARRAY



Asignamos 1 en la posición de la matriz tridimensional en la que se encuentra el paciente, la visita y la patología correspondiente. Todo lo demás se le asigna 0.

PROCESO DE ENTRENAMIENTO — 2

Del dataset de entrada se un subconjunto configurable para pacientes de entrenamiento

INPUT
1. TrainData
2. TrainLabel
3. Epochs
4. hiddenLayerSize

RNN - Recurrent Neural Network

Input (None, input_size)
Capa de entrada que acepta secuencias de longitud variable.

Mejor comprensión del contexto — **Bidirectional LSTM** (nº : hidden_layer_size)
Permite capturar dependencias tanto pasadas como futuras
Activación: tanh

Mejora la convergencia del modelo — **Batch Normalization**
Normaliza las activaciones de la capa anterior para acelerar el entrenamiento y mejorar la estabilidad.

Reduce sobreajuste — **Dropout** (Rate 0.3)
Regulariza el modelo al apagar aleatoriamente un porcentaje de las neuronas durante el entrenamiento

Enfoca la atención del modelo en pasos de tiempo importantes — **Attention Weights Calculation**
Dense Layer: activation softmax
Calcula los pesos de atención para cada paso de tiempo en la secuencia

Extrae información relevante de la secuencia — **Context Vector Calculation**
Dot(lstm_layer, attention_weights) : Activation tanh
Obtiene un vector de contexto ponderado por los pesos de atención.

Preparación del vector de contexto para la predicción final. — **Context Vector Reduction**
Dense (hidden_layer_size, activation relu)
Batch Normalization, Dropout (0.3) y Fatten
Reduce el vector de contexto a una dimensión manejable para la capa de salida.

Output Layer
Dense (1, activation = 'Sigmoid')

PROCESO DE PREDICCIÓN POR PACIENTE — 3

Return Modelo rnn entrenada

PROCESO DE VALIDACIÓN — 4

PACIENTE N

PACIENTE 2

PACIENTE 1

0 <= **PREDICCIÓN** <= 1

Representa probabilidad de tener patología objetivo

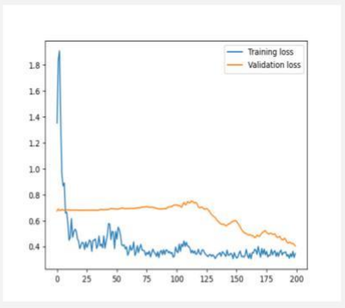
Del dataset de entrada, se utiliza un subconjunto de datos configurable para pacientes de test aplicando a cada uno la misma versión utilizada en los datos de entrenamiento. Se pasan los diagnósticos hasta diabetes (incluido).

Ejecución en Local

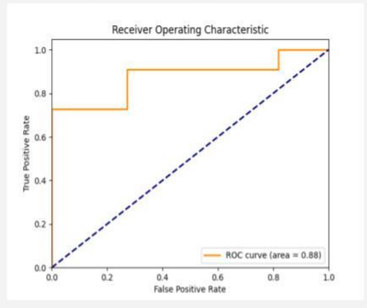
Muestra utilizada		
Totales		Patología objetivo
Total Pacientes	222	110
Total Entrenamiento	180	90
Total Test	22	11

Configuración de ejecución	
Datos de entrada	Versión 3
BatchSize	32
HiddenLayerSize	64
Épocas	200

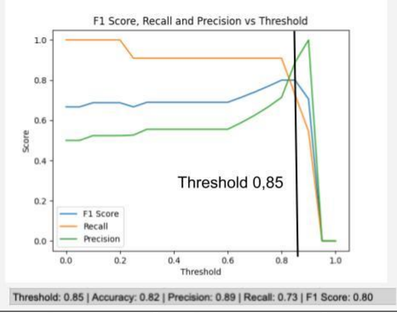
Gráfica que representa el proceso de entrenamiento de la RNN



Gráfica que representa la capacidad de clasificación que tiene el modelo. En este caso la capacidad indica que es buena.



Gráfica que representa el Threshold a elegir en base a la especificidad y la sensibilidad.



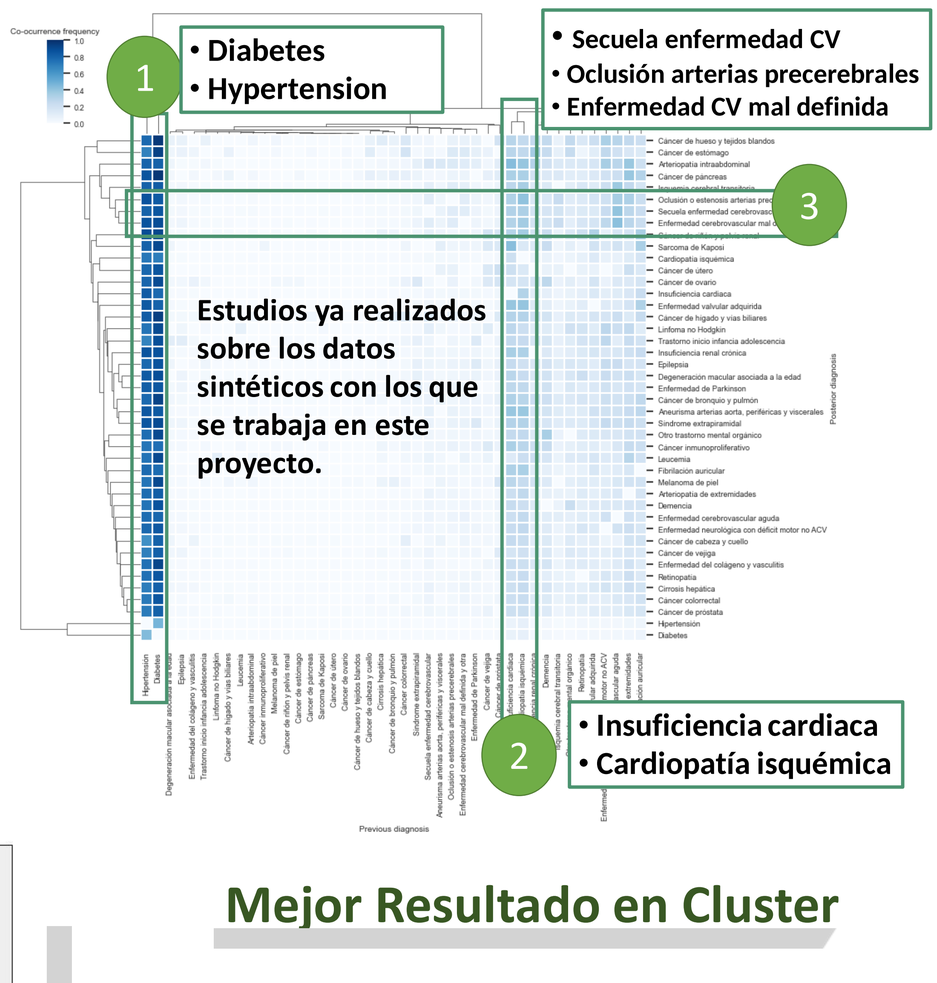
Threshold: 0.85 | Accuracy: 0.82 | Precision: 0.89 | Recall: 0.73 | F1 Score: 0.80

Tabla dónde se pueden ver las predicciones y los resultados.

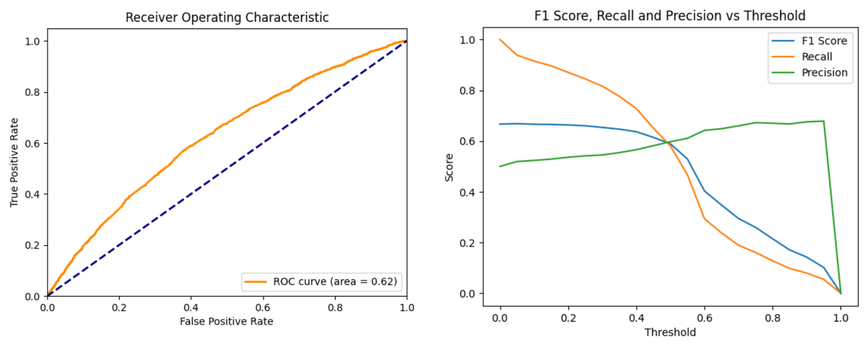
Confusion Matrix	
10	1
3	8

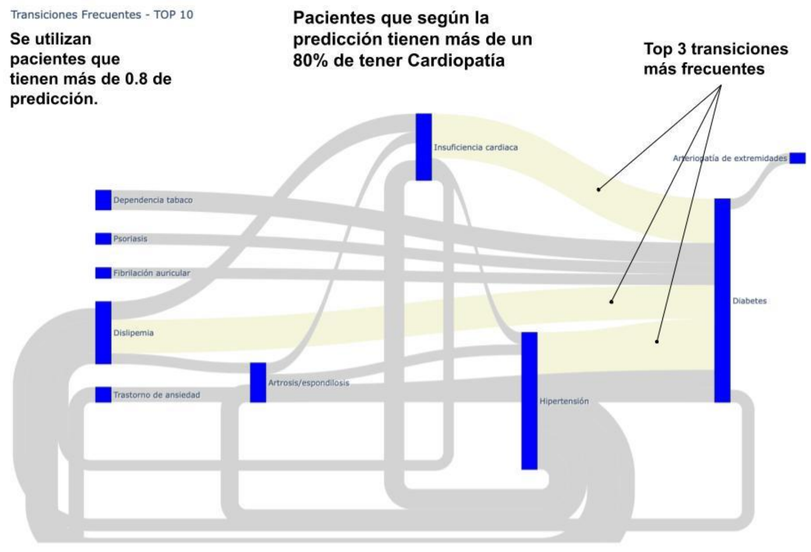
Paciente Test	Predicción	Cardiopatía(real)	Resultado aplicando Threshold (0.85)
1	0,056	0	0
2	0,901	1	1
3	0,895	1	1
4	0,81	1	0
5	0,805	0	0
6	0,667	0	0
7	0,288	0	0
8	0,901	1	1
9	0,901	1	1
10	0,237	0	0
11	0,901	1	1
12	0,901	1	1
13	0,895	1	1
14	0,823	1	0
15	0,24	1	0
16	0,611	0	0
17	0,702	0	0
18	0,857	0	1
19	0,752	0	0
20	0,901	1	1
21	0,825	0	0
22	0,848	0	0

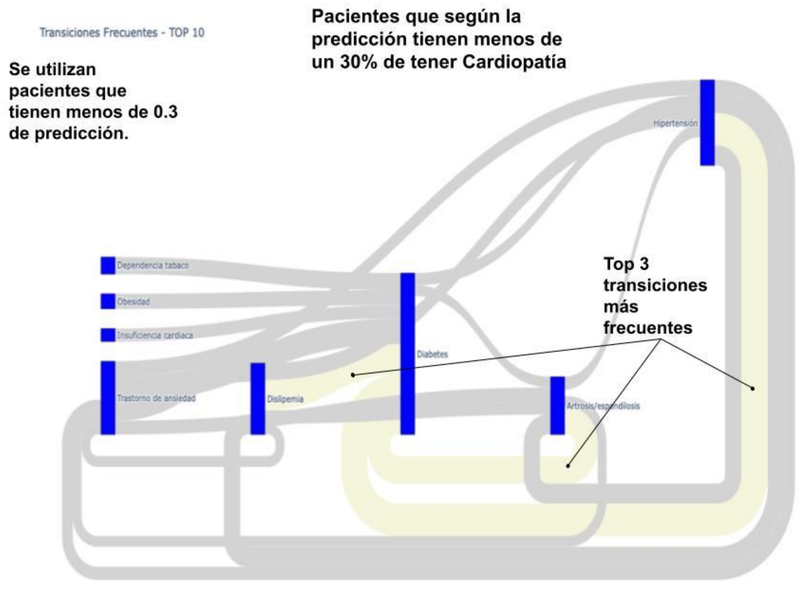
Secuencias de co-ocurrencias más frecuentes en dataset



Mejor Resultado en Cluster







Pruebas funcionales Dataset completo (Cluster)

Ejecución	Versión	Numero Patologias previas a Diabetes	Total Pacientes	Pacientes Entrenamiento	Pacientes Test	Balanceada patología objetivo	Batchsize	Épocas	Hidden Layer Size	ROC	F1-Score
1	1	1	65718	52576	13142	si	64	600	64	0,55	0,60
2	2	3	318807	255045	63762	no	64	600	64	0,55	0,10
3	2	5	7724	6950	774	si	128	600	128	0,58	0,67
4	2	3	62254	56028	6226	si	128	600	128	0,56	0,67
5	3	1	177692	159922	17770	si	128	600	128	0,59	0,67
6	3	2	111162	100044	11118	si	128	600	128	0,61	0,67
7	3	3	62254	56028	6226	si	128	600	128	0,62	0,67
8	3	4	32420	29178	3242	si	128	600	128	0,6	0,67

CONCLUSION

Los resultados muestran que todavía hace falta ajustar el modelo para el dataset completo. De todas formas, para este tipo de problemas parece que la RNN se comporta bien para muestras en las que hay un número significativo de patologías antes de Diabetes. Por otra parte, puede que la patología objetivo elegida no tenga ningún patrón dentro del dataset con lo que habría que probar con otras patologías objetivo.

Contact jiramirez@fidesol.org

Affiliations: