PROPUESTA DE PROYECTO
ANÁLISIS PREDICTIVO DEL IMPACTO DE COVID-19

Javier Pérez Vargas
Mario Ruiz Vaquett

CONTENIDO

01 Introducción

02 Objetivos y KPI

03 Acceso a los datos

04 Análisis Exploratorio

05 Técnicas de Preprocesamiento

06 Modelado y Predicción

07 Métricas y Consecución de Metas

08 Conclusiones

INTRODUCCIÓN

IMPACTO COVID-19

- Expansión del virus
- Estado de alarma
- Saturación hospitalaria
- Fiebre, fatiga, olfato...

PRUEBAS PCR

- Detección del virus
- Alta fiabilidad
- Fragmentos de ADN
- Control de brotes

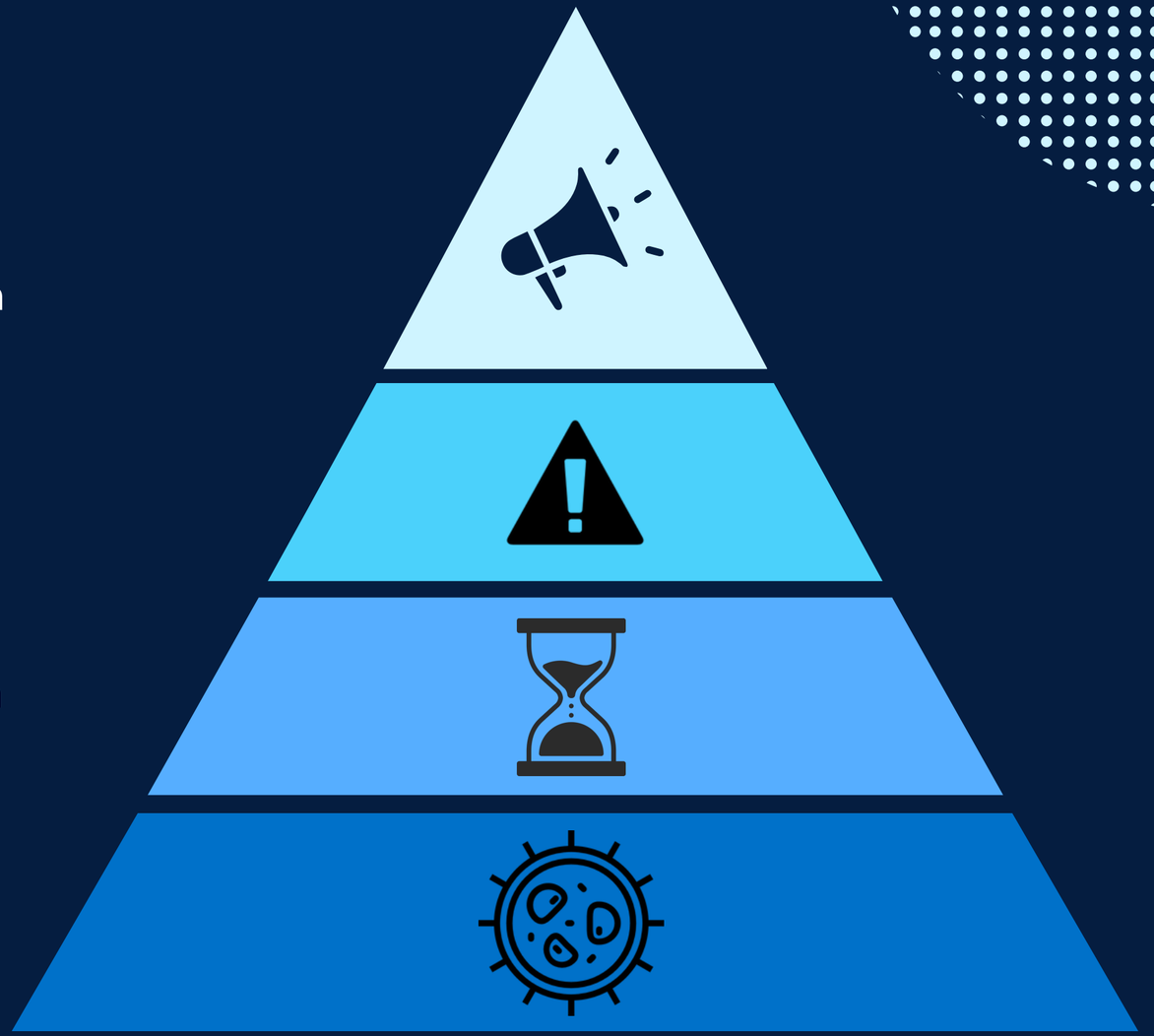
OBJETIVOS

01 Identificar factores que influyen en un resultado positivo de las PCR.

KPI Reducir la saturación hospitalaria mediante la rápida identificación de infectados.

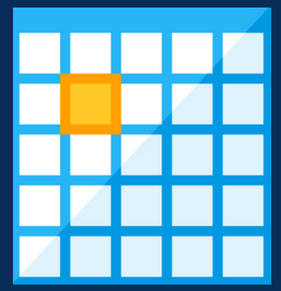
02 Desarrollar un modelo predictivo para clasificar la probabilidad de infección por COVID-19.

KPI Precisión (Accuracy) para medir el equilibrio general de predicciones correctas.



ACCESO A LOS DATOS

QUERY



supabase

pandas

CÓMO SON LOS DATOS

DATOS IDENTIFICATIVOS

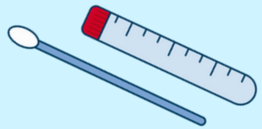
OBSERVACIONES

SÍNTOMAS CLÍNICOS

CONDICIONES PREEXISTENTES

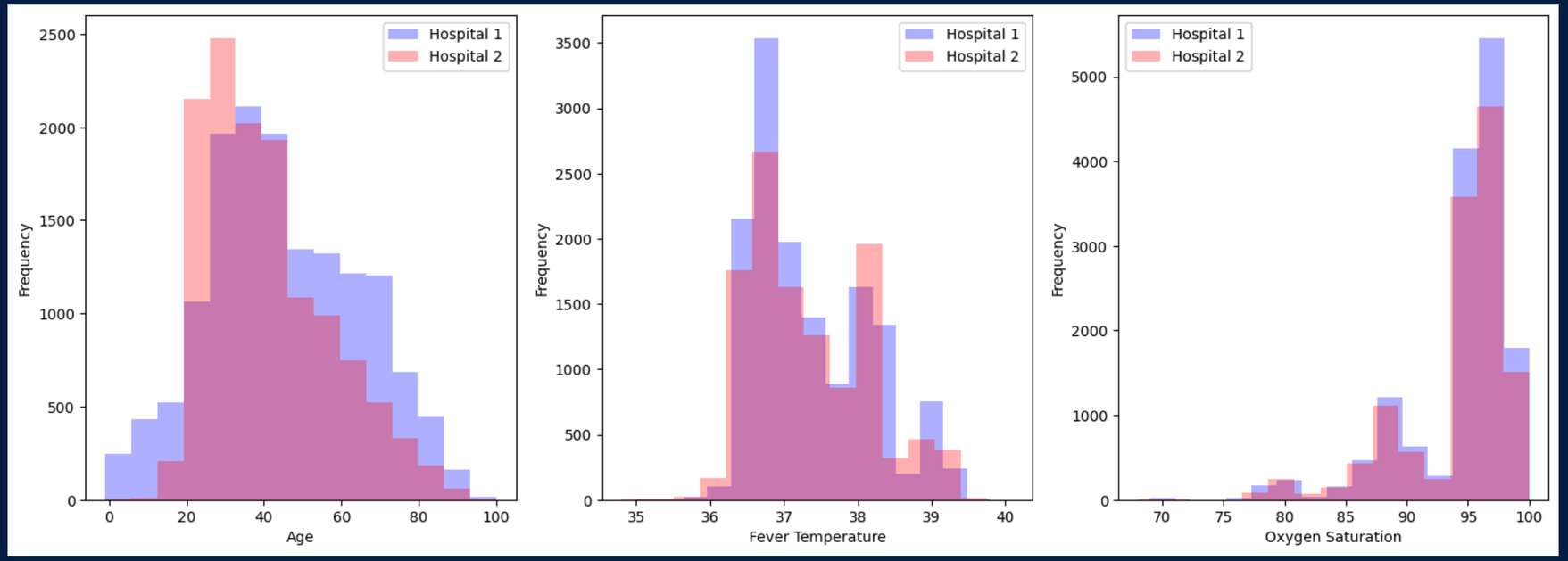
HÁBITOS Y RIESGOS

DIAGNÓSTICO PCR



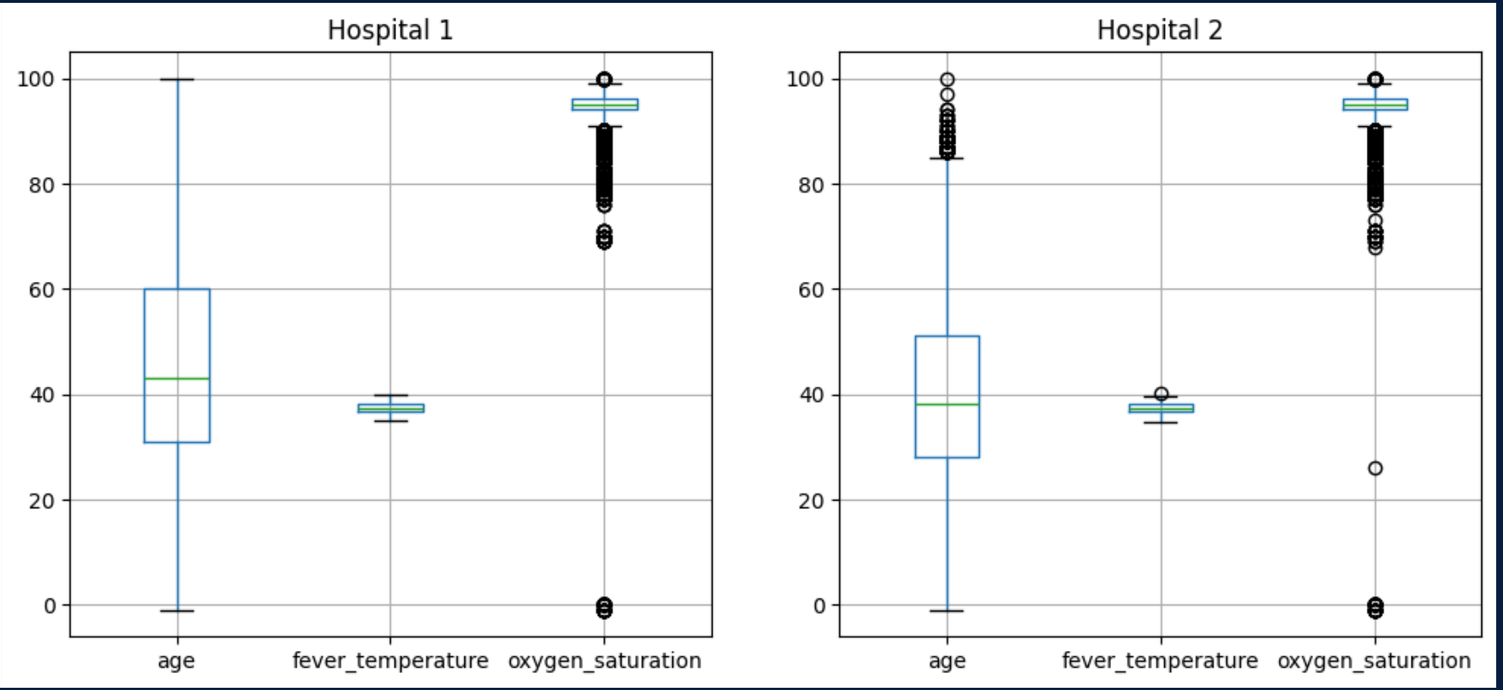
ANÁLISIS EXPLORATORIO

HISTOGRAMAS



ANÁLISIS EXPLORATORIO

BOXPLOTS



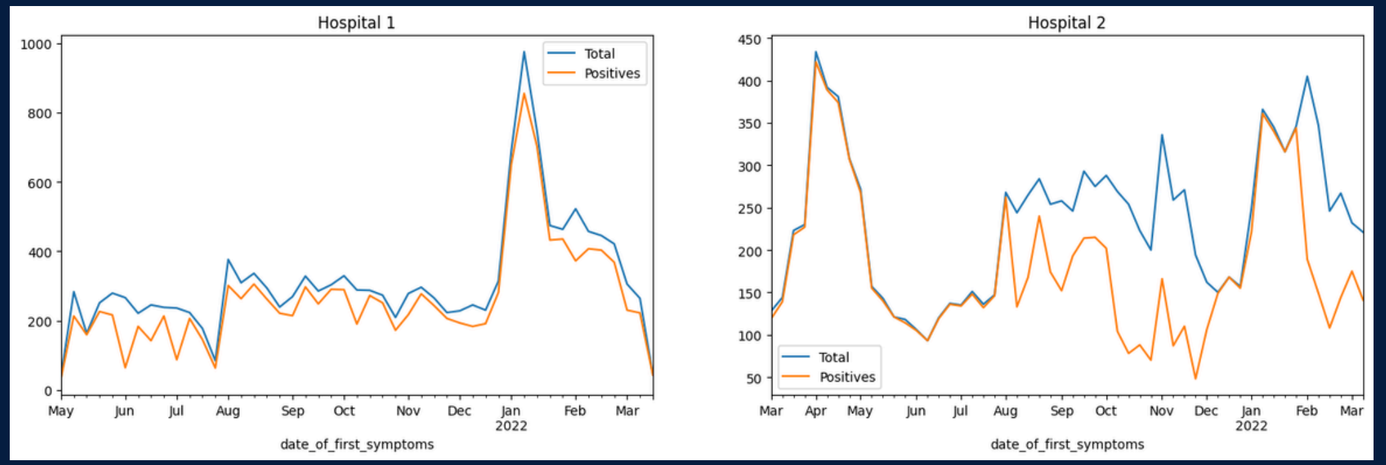
ANÁLISIS EXPLORATORIO

CORRELACIÓN

	Headache | Asthma	Headache | Hypertension	Asthma | Hypertension
Hospital 1	0.9397	0.9392	0.9391

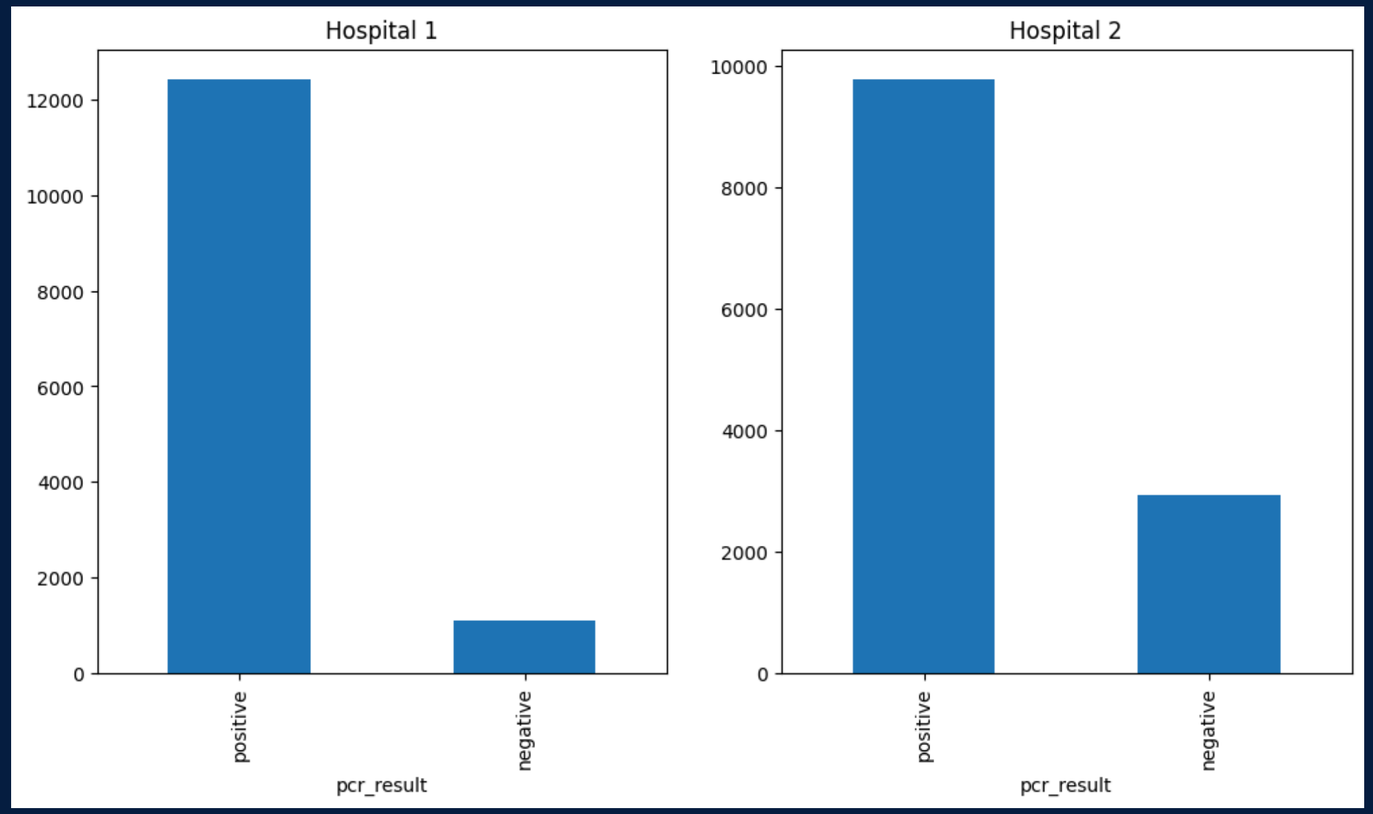
	Muscle aches | Fatigue malaise	Age | Hypertension	Abdominal Pain | Diarrhoea
Hospital 2	0.3987	0.3309	0.2936

EVOLUCIÓN TEMPORAL



ANÁLISIS EXPLORATORIO

BALANCEO DE PCR_RESULT



¿PROBLEMA?

TÉCNICAS DE PREPROCESAMIENTO

- Redundancia
- Tratamiento IDs
- Tratamiento Nulos
- Incoherencias
- Transformación

REDUNDANCIA

HOSPITAL 1

date_of_ first_symptoms

basvurutahiri

¿dofs == basvurutahiri?

SÍ

Solución: Eliminar basvurutahiri

HOSPITAL 2

date_of_ first_symptoms

admission date

¿dofs == admission date?

SÍ

Solución: Eliminar admission date

***basvurutahiri = fecha de solicitud**

TRATAMIENTO DE IDS

HOSPITAL 1

patient_ID

patient_ID1

¿patient_ID == patient_ID1?

NO

Dos ‘patient_ID’ distintos para el mismo ‘patient_ID1'

Solución: Eliminar patient_ID1

HOSPITAL 2

patient_id

TRATAMIENTO DE NULOS

HOSPITAL 1

1720 valores nulos

fever_temperature	468
oxygen_saturation	4
chronic_kidney_disease	7
obesity	22
liver_disease	6
asplenia	22
chronic_neurological_disorder	2
chronic_hematologic_disease	2
aids_hiv	2
diabetes_mellitus_type_1	3
diabetes_mellitus_type_2	2
rheumatologic_disorder	2
dementia	2
pcr_result	1176

* 5 valores de -1 en **'age'**

HOSPITAL 2

1411 valores nulos

fever_temperature	1219
oxygen_saturation	4
history_of_fever	5
bleeding	36
other_symptoms	36
pcr_result	33

* 65 valores de 0/-1 en **'oxygen_saturation'**
* 1 valor de -1 en **'age'**

Solución

- **Variables continuas: Imputadas con la media.**
- **Variables binarias: Imputadas con 0.**
- **Variable objetivo: Eliminadas.**

INCOHERENCIAS EN LOS DATOS

Edad

patient__id	nationality	age	gender	date_of_first_symptoms
712249	T.C.	28	K	2021-06-29 00:00:00
712249	T.C.	28	K	2021-09-25 00:00:00
712249	T.C.	30	K	2022-01-13 00:00:00
712249	T.C.	28	K	2022-02-03 00:00:00

HOSPITAL 1

2021-05-01 → 2022-03-14

10 MESES Y 14 DÍAS

Solución: Mínimo de la edad

HOSPITAL 2

2021-03-01 → 2022-03-13

12 MESES Y 11 DÍAS

Solución: Mínimo de la edad y suma a aquellos que hayan cumplido un año.

Nacionalidad

patient_id	country_of_residence	age	sex
99454396	T.C.	23	K
99454396	Pakistan	23	K
99454396	T.C.	23	K
99454396	T.C.	23	K

Solución: Poner T.C. en las incoherencias

*T.C. = Türkiye Cumhuriyeti = República de Turquía

TRANSFORMACIONES

pcr_result

- negative → 0
- positive → 1

Label Encoder

gender

- female → 0
- male → 1

Label Encoder

nationality

- T.C.
- Azerbaijan
- Other

One Hot Encoder

CREACIÓN DE CARACTERÍSTICAS

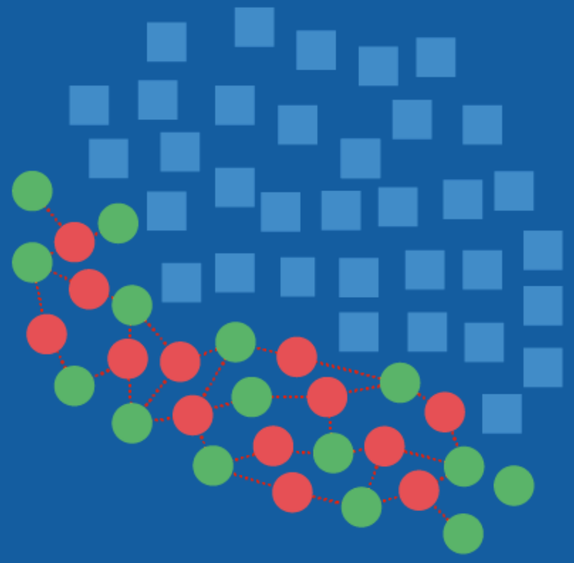
season → date of first symptoms → Spring, Summer, Fall, Winter

previous positives → Suma de previos positivos del paciente

OVERSAMPLING

SMOTE (SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE)

ORIGINAL DATASET



GENERATING SAMPLES

RESAMPLED DATASET

MODELADO Y PREDICCIÓN

Reg. Logística

Accuracy: 79.63%
Precisión: 84%
Recall: 80%
F1-Score: 81%

SVM

Accuracy: 78.57%
Precisión: 83%
Recall: 79%
F1-Score: 80%

Random Forest

Accuracy: 87.17%
Precisión: 88%
Recall: 87%
F1-Score: 87%

XGBoost

Accuracy: 87.96%
Precisión: 87%
Recall: 87%
F1-Score: 87%

***Métricas (WEIGHTED) de modelos sobre el conjunto de validación**

MEJOR MODELO

87%

De **ACCURACY** en nuestro modelo determinado por **XGBoost** en datos de TEST.

Confusion Matrix

246	**157**
184	2037

Otras métricas:

- Recall: 87%
- F1-Score: 87%
- Precisión: 87%

MEJOR MODELO

KPI Reducir la saturación hospitalaria mediante la rápida identificación de infectados.

La solución permite hacer una criba inicial para desaturar las urgencias hospitalarias.

KPI Precisión (Accuracy) para medir el equilibrio general de predicciones correctas.

El accuracy del modelo permite distinguir, en un alto porcentaje, los positivos y negativos.

IMPORTANCIA DE VARIABLES

- Saturación de oxígeno en sangre
- Temperatura de fiebre
- Tos
- History of fever
- Dificultad para respirar

CONCLUSIONES

- Se mejoró la calidad de los datos mediante imputación y resolución de inconsistencias.
- El modelo **XGBoost** alcanzó un **accuracy del 87%**, cumpliendo los objetivos planteados.
- Peligro del oversampling en el dominio médico
- Posible mejora de la sensibilidad del modelo.
- El modelo demuestra potencial para reducir la saturación hospitalaria en escenarios reales.

GRACIAS A TODOS

Javier Pérez Vargas
Mario Ruiz Vaquett