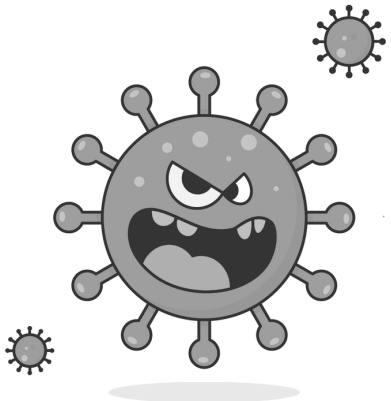
Análisis COVID19 en México

André Marx Puente Arévalo

Objetivo

Modelar la mortalidad de un paciente que fue confirmado con COVID19 y otros padecimientos o características del mismo y lograr encontrar semejanzas en las variables que permitan **generar grupos** por tipo de pacientes, mediante la creación de una tabla que contenga la información necesaria y procesada.



Conjunto de Datos

Se extrajo de la página del gobierno mexicano el conjunto de datos, el cual, contiene los registros diarios de **pacientes** que fueron **atendidos por posible COVID19** de manera diaria en toda la República Mexicana.

Cuenta con contenido desagregado por sexo, edad, nacionalidad, padecimientos asociados, entre otros.

Registros
2.5 M

Columnas
38

Calidad de Datos

Duplicados y Completitud

Ningún duplicado, 100% completitud

Cruce con católogos

Las variables originalmente contenían puros valores numéricos

Normalización y Anális Exploratorio

Homologar variables y resumir la información

Detección Extremos

Representaron **0.32%** de la muestra total, se eliminaron

Completitud

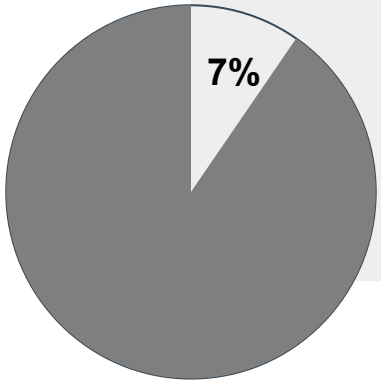
Se imputaron valores mediante la moda

Ingeniería de Variables

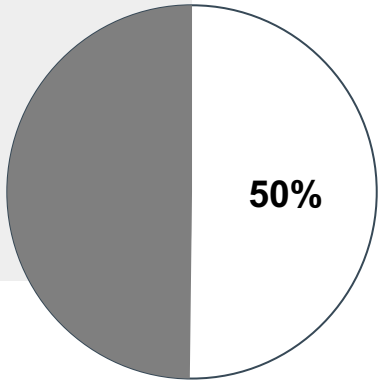
Transformación entrópica, variables dummies y variable objetivo

Variable Objetivo

Se utilizó la variable que indicaba la fecha de defunción, donde si tenía el valor “9999-99-99” toma el valor de **0 (sobrevive)**, de lo contrario **1 (fallo)**.

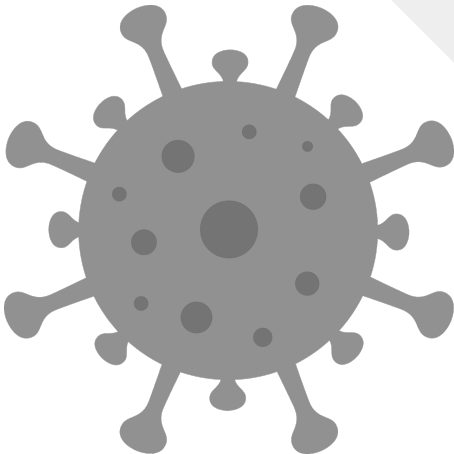


Sin Balanceo



Undersample

Aprendizaje Supervisado



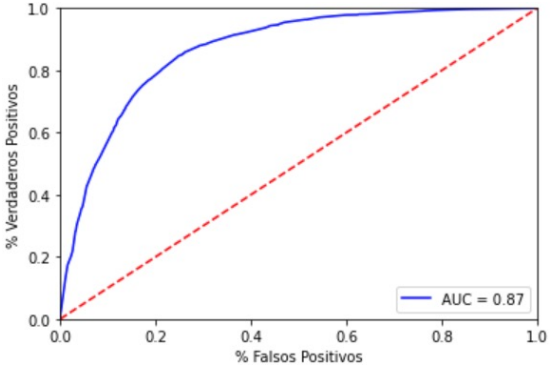
Nota: Se modeló con pacientes que salieron positivo a COVID19.

Scoring

Dado que se cuentan con prácticamente puras variables discretas, se realizó un modelo de Scoring utilizando la muestra sin balanceo y con balanceo.
El modelo que se ajustó fue uno de clasificación binario llamado **Regresión Logística** y sólo se utilizaron seis variables.

Modelo	AUC	ACC	F1
Sin Balanceo - Entrenamiento	0.861	0.931	0.108
Sin Balanceo - Validación	0.869	0.932	0.107
Undersample - Entrenamiento	0.859	0.786	0.786
Undersample - Validación	0.869	0.789	0.445

Curva Roc - Validación



Scoring

Pregunta	Respuesta	Puntos
¿Qué edad tienes?	[0 - 26]	279
	[27 - 34]	257
	[35 - 41]	182
	[42 - 49]	116
	[50 - 59]	48
	[60 - 89]	-52
¿A qué sector acudes por servicios médicos?	Estatal	43
	IMSS	39
	IMSS Bienestar	68
	ISSSTE	13
	Otro	111
	PEMEX	18
	Privada	94
	SEDENA	4
	SEMAR	82
	SSA	98
¿Padeces de diabetes?	No	83
	Sí	43
¿Tienes hipertensión?	No	81
	Sí	43
¿Has estado en contacto con otra persona que tiene COVID?	No	131
	Sí	35
¿Padeces de insuficiencia renal crónica?	No	75
	Sí	-46

Probabilidad por score

Score	% Sobrevivencia	% Fallo
[1, 203)	57.40%	42.60%
[203, 404)	80.99%	19.01%
[404, 606)	96.73%	3.27%
[606, 807)	99.50%	0.50%

Otros Modelos

En lugar de la trasformación entrópica, se utilizaron variables "**dummies**". Aplicando reducción de dimensiones (principalmente correlación con el objetivo y poca varianza) se consiguieron **35** variables, se obtuvieron los siguientes performances:

Modelo	AUC	ACC	F1
Árbol de Decisión	0.4994	0.9009	0
XGBoost	0.4963	0.9009	0
Regresión Logística	0.4985	0.9009	0
Naive Bayes	0.4982	0.8861	0.0299
Red Neuronal	0.4993	0.9009	0

Sin Balanceo

Modelo	AUC	ACC	F1
Árbol de Decisión	0.4994	0.6543	0.149
XGBoost	0.5012	0.5194	0.1644
Regresión Logística	0.4994	0.4952	0.1654
Naive Bayes	0.5011	0.2019	0.1786
Red Neuronal	0.4951	0.4341	0.1677

Undersample

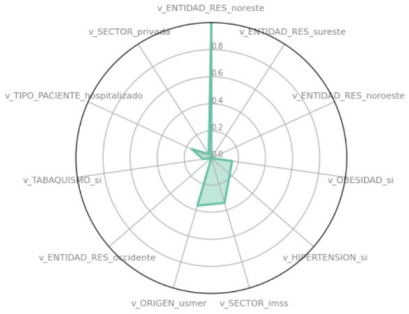
Nota: Los scores mostrados en las tablas son con el conjunto de validación.

Aprendizaje No Supervisado

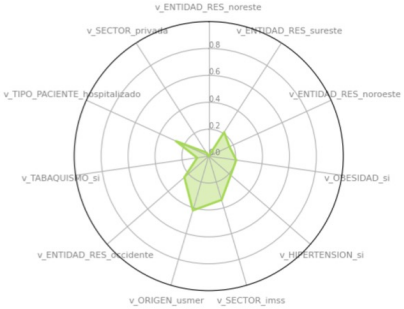
Clustering

Utilizando la tabla analítica de datos que contiene las variables **dummies**, se realizaron varios modelos de aprendizaje no supervisando y no se logró obtener grupos que fueran de utilidad para el contexto requerido. Los grupos se partieron principalmente por los estados de la república, mismos que se pudieron obtener sin necesidad de hacer un modelo.

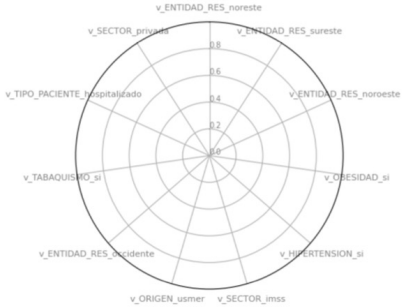
Los modelos probados fueron K – mediodes, DBSCAN, Mezclas Gaussianas y clustering jerarquico.



(a) Cluster 0



(b) Cluster 1



(c) Cluster 2

Conclusión

Con la información con la que se cuenta actualmente, fue **posible** generar un modelo de machine learning que permitiera **medir la mortalidad** de un paciente contagiado de COVID19. Sin embargo, **no** fue se logró realizar un modelo que permitiera agrupar a los tipos de pacientes que fueron atendidos.