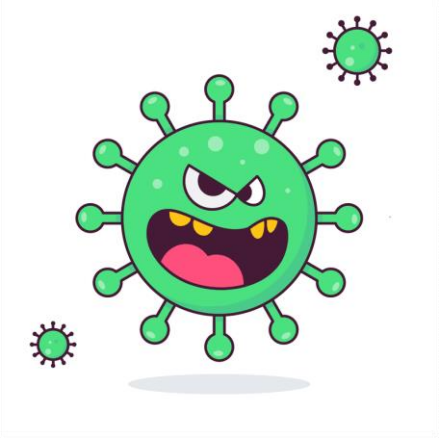
Análisis COVID19 en la CDMX

Alfredo Nájera Nájera

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 la Ciudad de México.

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

Registros
1.9 M

Columnas
40

Construcción de una TAD

Calidad de datos

- Cruce con catálogos
- Duplicados
- Completitud
- Normalización de variables

Análisis Exploratorio

Hallazgos relevantes de las variables

Ingeniería de variables

- Variables dummies
- Variable objetivo

Detección de valores extremos

Representaron **0.56%** de la muestra total, se eliminaron

Valores ausentes

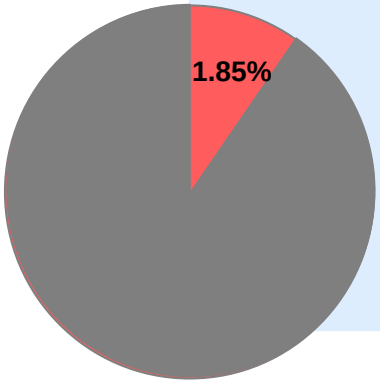
Se imputaron valores mediante la moda

Reducción de dimensiones

- Alta correlación
- Multicolinealidad

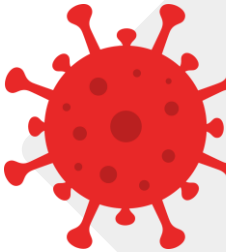
Variable Objetivo

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 (muere)**.



Proporción de pacientes fallecidos

Modelación



Se utilizaron los siguientes modelos para poder cumplir los objetivos planteados al inicio:

Modelación supervisada

Modelos de clasificación*:

- Árbol de decisión
- XGBoost
- Regresión Logística
- Naive Bayes
- Red Neuronal

Modelación no supervisada

Clusterización:

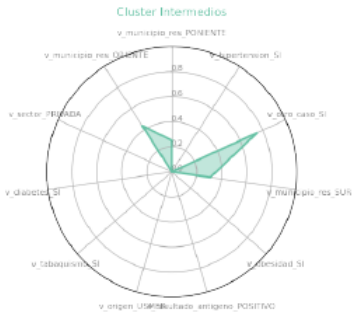
- K-mediodes
- Mezclas Gaussianas
- Clustering Jerárquico Aglomerativo

*Se hicieron los modelos con la muestra de la variable objetivo balanceada y sin balancear

Clustering

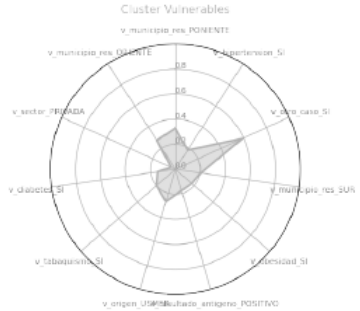
Utilizando la tabla analítica de datos que contiene las variables **dummies**, se realizaron varios modelos de aprendizaje no supervisando y no se logró obtener grupos que fueran de utilidad para el contexto requerido. Los grupos se partieron principalmente por las alcaldías, mismos que se pudieron obtener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

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



(a) Cluster 0

(b) Cluster 1



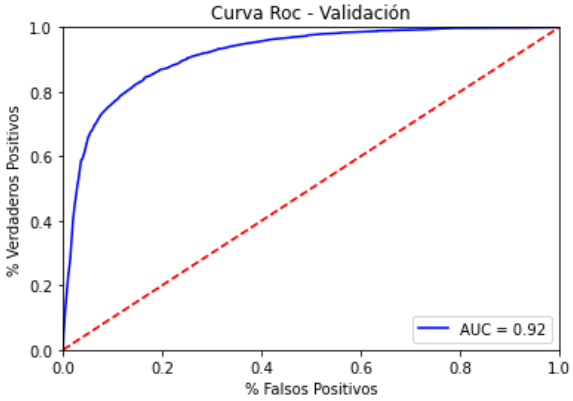
(c) Cluster 2

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 siete variables.

Modelo	AUC	ACC	F1
Sin Balanceo - Entrenamiento	0.923	0.963	0.223
Sin Balanceo - Validación	0.915	0.963	0.234
Undersample - Entrenamiento	0.925	0.844	0.846
Undersample - Validación	0.916	0.836	0.279

Curva Roc - Validación



Scoring

Pregunta	Respuesta	Puntos
¿Qué edad tienes?	[2,23]	233
	[24,31]	168
	[32,39]	124
	[40,47]	81
	[48,56]	48
	[57,87]	-15
¿Padeces diabetes?	NO	61
	SI	12
¿Estás embarazada?	NO	77
	NO APLICA	31
	SI	272
¿Tienes hipertensión?	NO	59
	SI	26
¿Has estado en contacto con otra persona que tiene COVID 19?	NO	39
	SI	71
¿Padeces de insuficiencia renal crónica?	NO	54
	SI	-34
¿A qué sector acudes por servicios médicos?	IMSS	-21
	ISSSTE	-56
	OTRO	-10
	PRIVADA	39
	SSA	116

Probabilidad por score

Score	% Sobrevivencia	% Fallo
[2,173.6)	51.06%	48.94%
[173.6,345.2)	85.08%	14.92%
[345.2,516.8)	98.75%	1.25&
[516.8, 688.4)	99.93%	0.07%

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** se logró realizar un modelo que permitiera agrupar a los tipos de pacientes que fueron atendidos.