Contagios y defunciones Covid-19 en la CDMX de Febrero 2020 a Febrero 2021

Resumen ejecutivo

Lo que se buscará con el modelo propuesto será obtener la probabilidad de mortalidad por Covid-19 de una persona basándose a grandes rasgos en dos factores: su edad y su colonia de residencia. El modelo utiliza técnicas de análisis de datos y aprendizaje automático para identificar patrones y relaciones entre los datos de entrada.

La edad es un factor importante debido a que se ha demostrado empíricamente que es una variable con mucho peso para esta predicción, cosa que se confirma con los propios datos. Por otro lado, la residencia se considera un indicador del nivel socioeconómico, el cual puede estar relacionado con la calidad de vida y la disponibilidad de recursos de atención médica.

Se busca que el modelo sea utilizado por las autoridades de salud pública para planificar y asignar recursos de atención médica de manera más efectiva, especialmente en áreas con mayores necesidades.

Objetivos específicos del proyecto

- Realizar una predicción de la probabilidad de que un individuo falleciera de Coivd-19 durante las fechas de Febrero de 2020 a Febrero de 2021, basados en la edad y datos relacionados con la colonia de residencia del individuo.

- Agrupar a los individuos en conjuntos de personas con características similares, en particular que tengan algún rasgo económico social en común.

Descripción de la base

La base presentada contiene todas las pruebas de Covid-19 realizadas en la Ciudad de México, de febrero del 2020 a febrero del 2021. Cada prueba trae variables intrínsecas de la persona como edad, sexo, ocupación, etc. Además, contiene factores del Índice de Desarrollo Social de la colonia en la que habita, como CAEj (indicador de falta de servicio de luz), TELj (indicador de falta de teléfono), etc, los cuales fueron reducidos por un procedimiento de componentes principales a 3 componentes. De igual forma los datos contienen el propio Índice de Desarrollo Social de la colonia, el cual está calculado a partir de los demás factores que se redujeron. El diccionario de las variables se encuentra anexo al final del documento.
Finalmente la base tiene un etiquetado para detectar personas cuya prueba fue positiva o negativa y otra etiqueta para para conocer qué personas fallecieron y cuales sobrevivieron al contagio.
Acotaremos el análisis solamente para aquellas personas contagiadas.

También es relevante mencionar que la base fue construida a partir de 3 distintas.

- Base de IDS: estudio levantado por INEGI en 2020 que obtiene el Índice de Desarrollo Social por manzana, esta información se proceso para hacer el promedio por colonia.
- Base de Pruebas covid: datos de todas las pruebas covid realizadas en la cdmx.
- Base defunciones: Todas las defunciones de la CDMX, esta base fue filtrada con base al análisis experto de doctores especialistas en el tema para mantener solamente las muertes de interés.

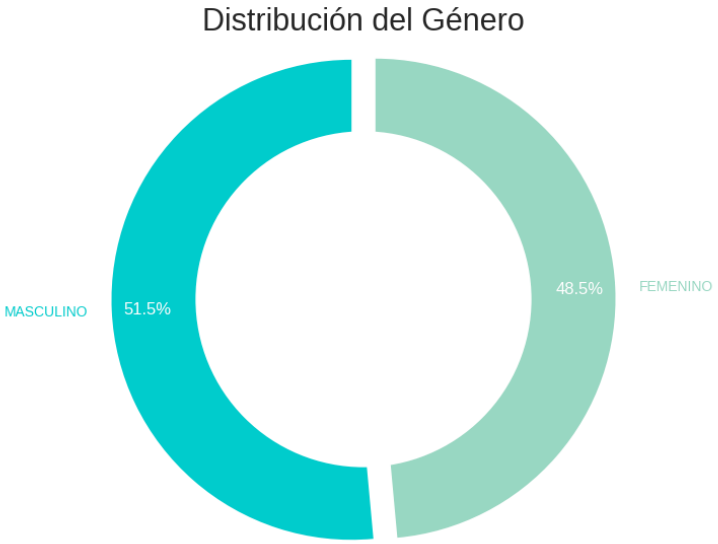
La base de Pruebas e IDS se unieron a partir de la dirección (Colonia) de la persona mientras que la de Pruebas y Defunciones por el CURP.

Por la forma de ensamble de esta base la mayoría de las variables no son homogéneas y presentan problemas en la calidad de datos, por lo que la mayoría de ellas se desechan para el análisis.

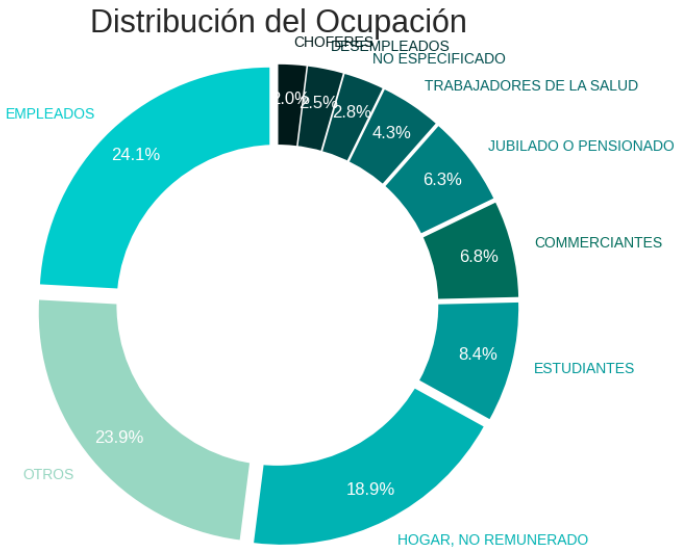
Análisis exploratorio de datos

Discretas

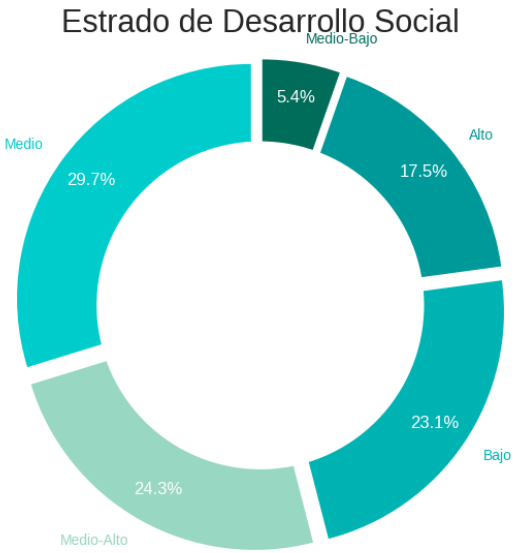
Primeramente diferenciamos nuestras variables categóricas quedándonos con Sexo, Ocupación y Estrato del Índice de Desarrollo Social. Observando la distribución del género podemos observar el mismo porcentaje de hombre y mujeres en la población.



Se observa una alta cardinalidad (77) en la variable original de Ocupación, realizando una recategorización, se dejan 10 categorías para trabajar. Sin embargo al realizar los cálculo del Information Value, se obtiene que la variable es sobre predictiva, esto debido a que la base tiene dos insumos para esta variable uno de la base de defunciones y otro de la base de pruebas, cada una con su catálogo, por lo que esta diferenciación por sí sola discrimina nuestra variable objetivo.

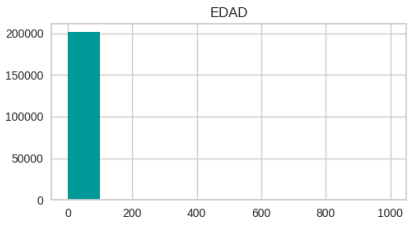


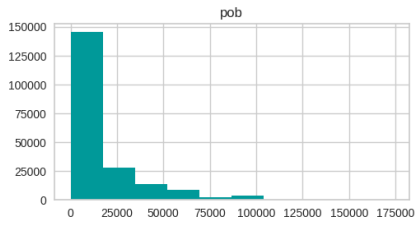
Finalmente, observando esta última distribución, podemos afirmar que la mayoría de la población en la CDMX se encuentra en los estratos medios, y solamente un 17.5% alcanza un estrato alto, por lo que en general podemos esperar muchos grupos de colonias pobres o con alguna deficiencia de desarrollo en sus localidades. Cabe recalcar que la mitad de la población se encuentra repartida solamente en los estratos Medio y Bajo.

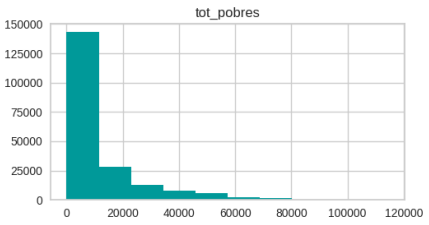


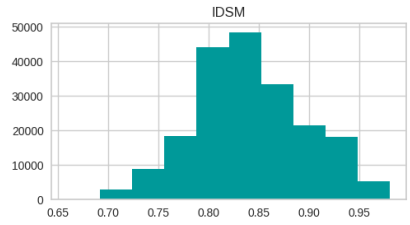
Continuas

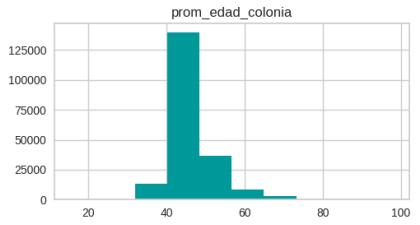
Por otro lado tenemos nuestras variables continuas, que quitando los factores del IDS a los que se les aplicará PCA, obtenemos las distribuciones mostradas.

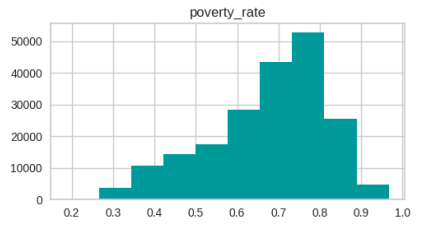


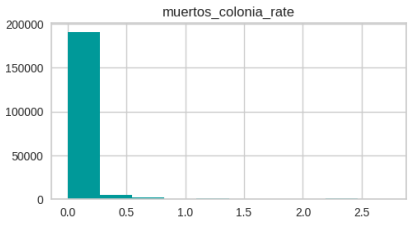






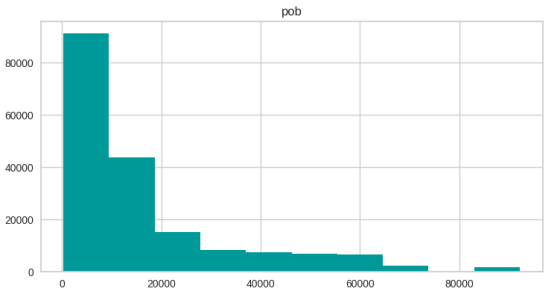


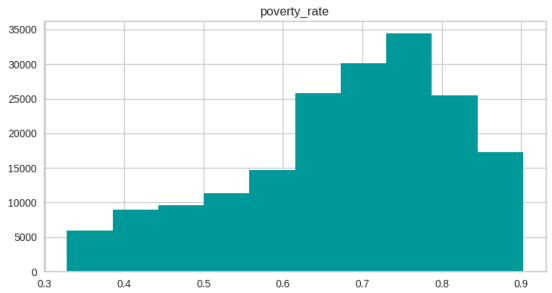


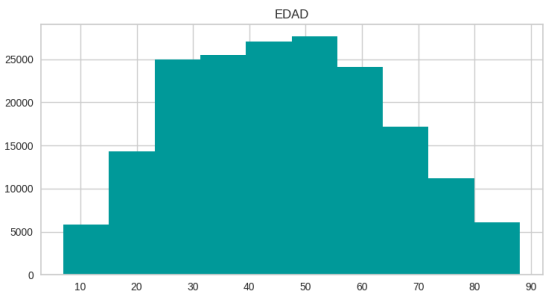


Observamos también las correlaciones de las variables

Usando el algoritmo de Varclushi nos quedamos únicamente con las variables de población, poverty_rate y edad, esta última variable al obtener su IV, confirmamos que es una variable muy potente. Tomaremos estas variables quitando sus valores extremos, con lo que su distribución quedará de la siguiente manera.



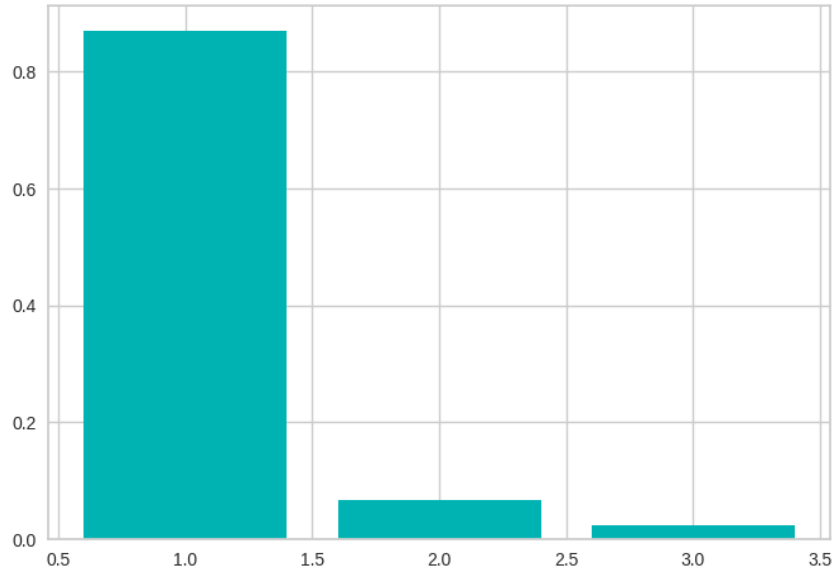




Reducción de dimensiones a factores de IDS

Finalmente, realizaremos el análisis de componentes principales sobre los factores del índice de Desarrollo Social, este se hará con 3 componentes que explican la varianza de la siguiente manera.

Componente	Varianza explicada
1	86.93%
2	6.64%
3	2.53%

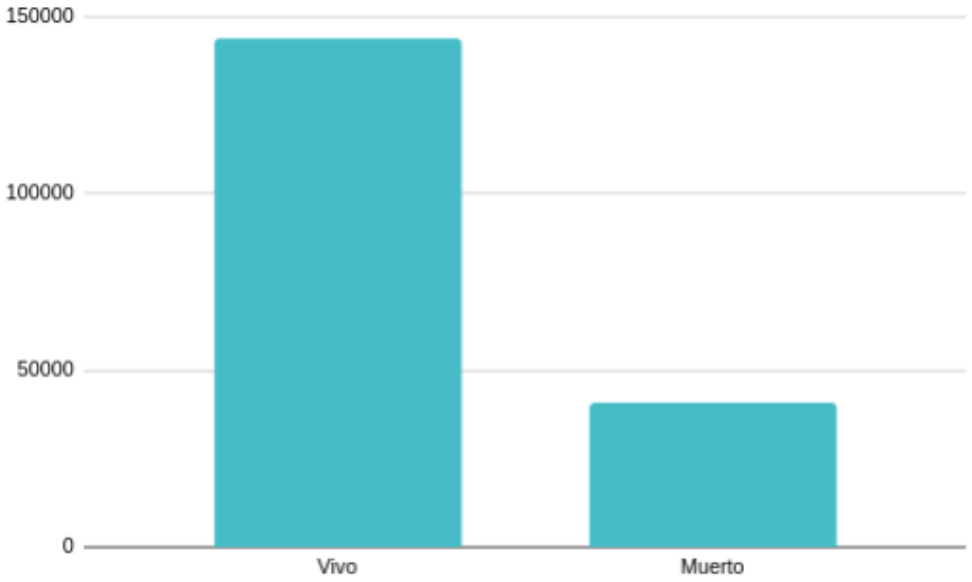


Modelación supervisada (Regresión logística)

Se opta por usar una regresión logística por la poca cantidad de recursos necesarios para su implementación, pensando en que en el sector gobierno no hay las mejores infraestructuras para implementar algoritmos complejos. Además de que esta brinda una clara interpretación a partir de sus coeficientes.

Undersampling/Oversampling

Lo primero a notar antes de realizar la modelación es que hay un desbalance en las clases



Estado	No. personas
Vivo	143460
Muerto	40418

Para resolver el problema se usa la técnica de Smote sampling con la estrategía de minority.

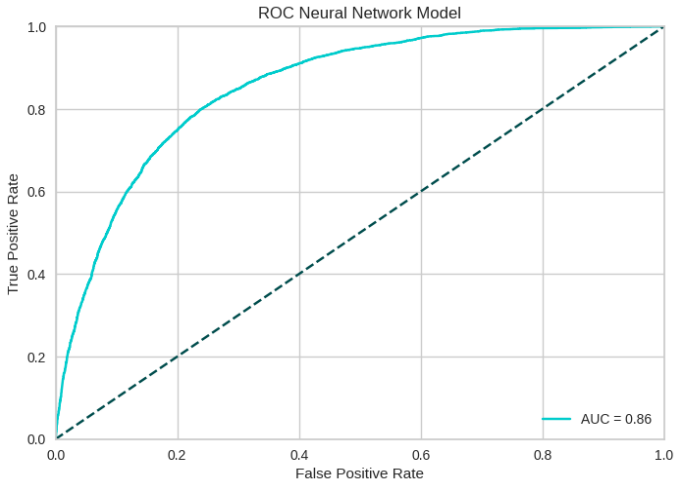
Arquitectura

Una vez resuelto el problema de desbalance se crea la gradilla de hiper-parámetros y se ajusta el modelo a través de una búsqueda por toda la gradilla, la cual obtiene una regresión lineal sin penalización con el solver ‘lbfgs’ como mejor arquitectura para el modelo. A continuación se enlistan todos los hiperparámetros de la arquitectura.

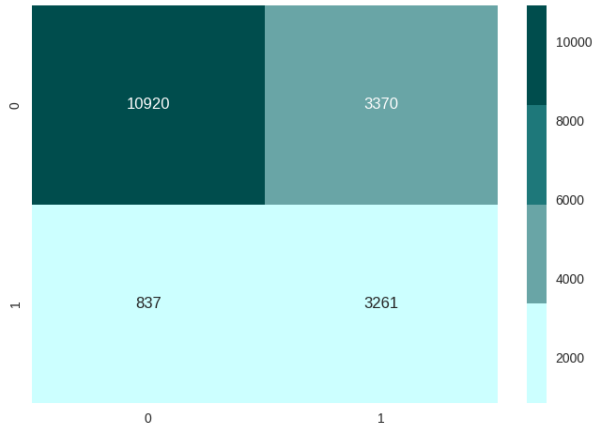
C: 1.0, **class_weight**: None, **dual :** False, **fit_intercept**: True, **intercept_scaling**:1, **l1_ratio**: None ,**max_iter**: 100, **multi_class**: 'auto', **n_jobs**: None, **penalty**: none, **random_state**: None, **solver**: lbfgs, **tol**: 0.0001, **verbose**: 0, **warm_start**: False.

Métricas

Para los conjuntos tanto de validación como de entrenamiento se obtiene un accuracy score de **.86**, por lo que el modelo está ajustado de muy buena manera, sin presentar algún sobreajuste para los datos.



La matriz de confusión acumula la mayoría de los registros en la diagonal lo que es una buena señal, sin embargo por el desbalance de las clases, en números absolutos los aciertos de la categoría 1 (Muerto) son casi iguales a los errores de la categoría 0 (Vivo). Este desbalance también impacta en algunas de las demás métricas como por ejemplo el Precision, en el que vemos que la categoría Muerto sufre una baja importante debido a este fenómeno por lo que es engañoso dejarse guiar por esta métrica en este caso, si observamos el promedio ponderado de la precisión podremos observar como la métrica vuelve a subir ya que este número está ya ajusta ese desbalance.



Valor	Precision	Recall	f1-score	Support
Vivo	0.93	0.76	0.84	14,290
Muerto	0.49	0.8	0.61	4,098
Accuracy			0.77	18,388
Weighted avg	0.83	0.77	0.79	18,388

Conclusión

En general las métricas obtenidas son lo suficientemente aceptables como para considerar que el modelo opera de forma correcta y realiza predicciones acertadas. Para obtener la probabilidad de fallecimiento solamente es necesaria la probabilidad "cruda" de la predicción.

Analizando más a detalle los resultados, a través de los coeficientes obtenidos, observamos que la edad se comporta como se esperaba y entre mayor sea, mayor será la probabilidad de muerte de la persona, también el género tiene cierta injerencia, siendo que los hombres tienen una mayor probabilidad de fallecer. Por otro lado, observemos que entre más alta sea la tasa de pobreza de la colonia a la que pertenece el individuo también será más probable que fallezca.

Variable	Coeficiente
x_SEXO	0.435505
x_EIDS	0.078503
x_poblacion	0.109982
x_poverty_rate	-0.170229
x_edad	1.762574

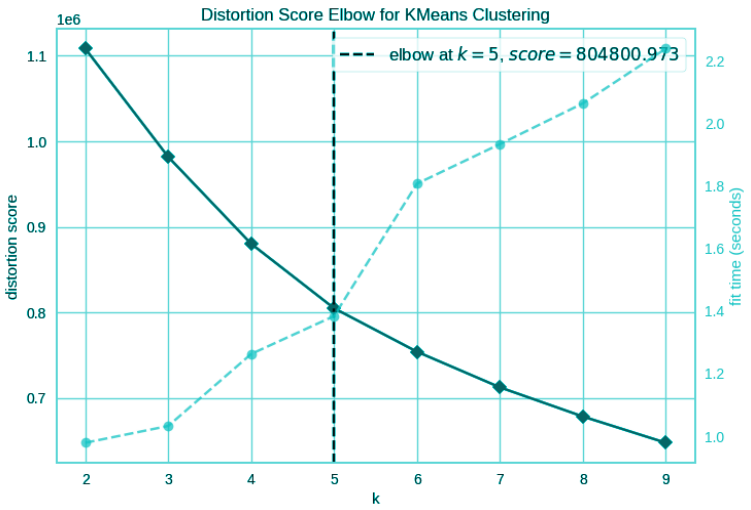
Esta predicción podría servir a manera de cálculo para la población en general para conocer su riesgo a manera de cautela.

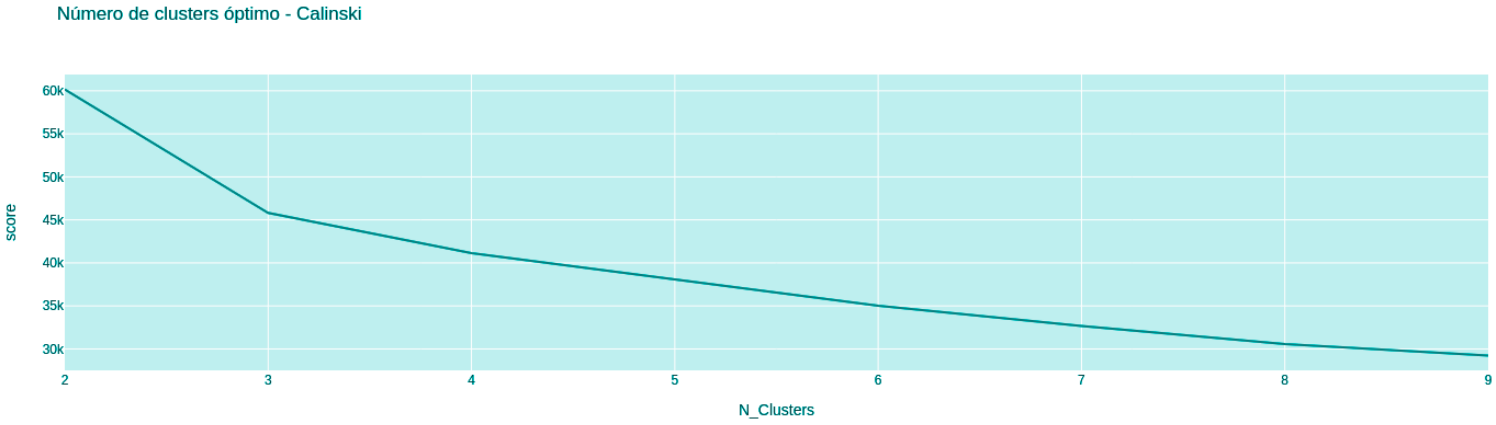
Por parte de las autoridades, podrán usar este análisis para detectar las colonias de riesgo, ya sea porque el nivel de desarrollo de esa colonia es deficiente o porque la edad promedio de las personas de la localidad sea mayor y con esa información buscar elaborar una estrategia para la mejora de los servicios de la colonia.

Modelación no supervisada (k-means)

Número de clusters

Partiremos de escalar los datos para poder utilizar los algoritmos de clustering y realizaremos las pruebas de codo y calinski para determinar el número de clusters con los que trabajaremos. Obteniendo los siguientes gráficos

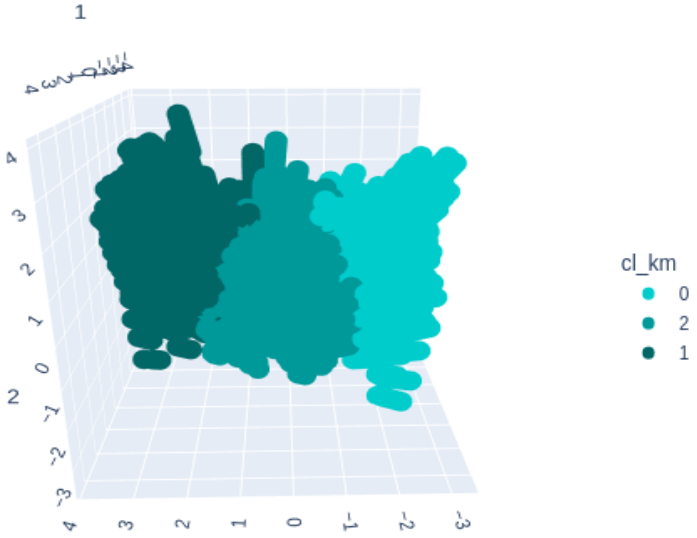




La prueba del codo nos indica que el número de clusters óptimos serían 5, mientras que Calinski arroja 3 ó 4 clusters como el score máximo. Para la toma de decisión se modeló con 4 y 5 clusters para conocer cuál opción se ajustaba mejor a las necesidades.

Clusterización

- **3 Clusters:** Observemos que con esta opción la visualización es muy buena, nuestros clusters están bien separados y son densos. Sin embargo el problema viene al momento de realizar el perfilamiento ya que aunque algunas características entre grupos son discriminatorias no todas las variables llegan a tomar importancia como lo hacen para el caso de 4 y 5 clusters que se revisarán posteriormente.



Perfilamiento

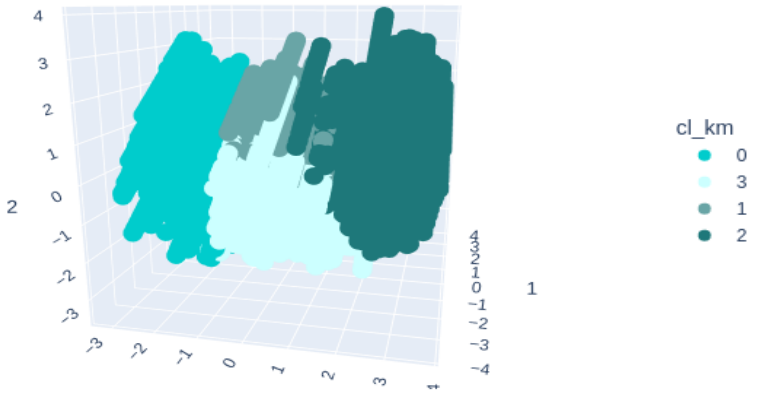
cl_km	x_SEXO	x_EIDS	x_poblacion	x_poverty_rate	x_edad	registros	proporcion_muertos
0	0.508948	1.178906	18196.576956	0.820247	45.594088	49730	0.239212
1	0.510475	4.784093	13660.492068	0.464177	46.843896	37757	0.192997
2	0.516770	3.370356	16291.188181	0.688806	46.215279	96391	0.220301

Cluster 0: Personas que habitan en una colonia muy poblada, con un IDSM **bajo**, son el grupo más **joven** y tienen una **alta** tasa de defunciones.

Cluster 1: Personas que habitan en una colonia muy poblada, con un IDSM **alto**, son el grupo con el promedio de edad más **alto** y con una **baja** tasa de defunciones.

Cluster 2: Personas que habitan en una colonia muy poblada, con un IDSM **medio**, y tienen una **alta** tasa de defunciones.

- **4 Clusters:** Observemos que con esta opción, aunque la visualización no es tan clara como con 3, nuestros clusters se siguen viendo bien separados.Checando el perfilamiento vemos que algunas variables son mucho más explicativas y ayudan a la interpretación de los clusters de mejor forma.



Perfilamiento

cl_km	x_SEXO	x_EIDS	x_poblacion	x_poverty_rate	x_edad	registros	proporcion_muertos
0	0.512013	1.089120	19324.737273	0.823922	45.580006	45534	0.243401
1	0.000000	3.360395	15299.664283	0.685030	45.574982	51746	0.162447
2	0.512673	4.914445	14829.869091	0.446202	46.960300	31839	0.193379
3	1.000000	3.363502	15471.825161	0.685565	46.784711	54759	0.269764

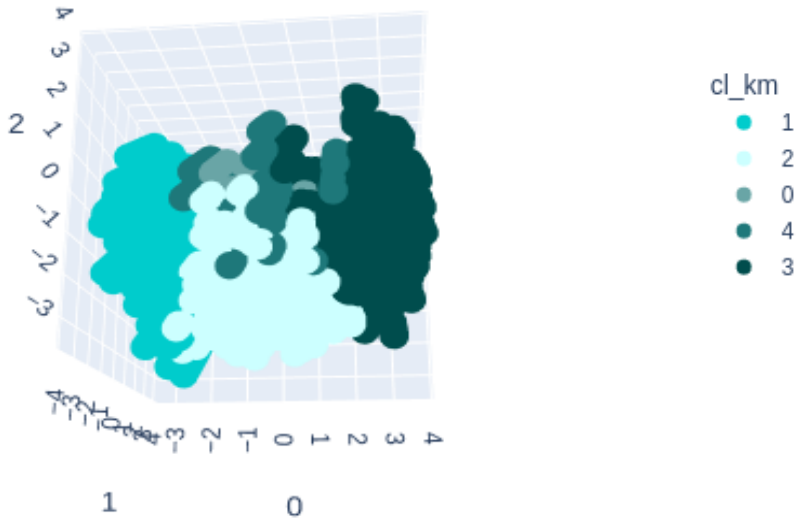
Cluster 0: Personas que habitan en una colonia poblada, con un IDSM **bajo**, son un cluster más **joven** y tienen una **alta** tasa de defunciones.

Cluster 1:Mujeres que habitan en una colonia poblada, con un IDSM **medio**, son un cluster más **joven** y tienen una **baja** tasa de defunciones.

Cluster 2: Personas que habitan en una colonia poblada, con un IDSM **alto**, son el grupo más **viejo** y tienen una **mediana** tasa de defunciones.

Cluster 3: Hombres que habitan en una colonia muy poblada, con un IDSM **medio**, son un cluster viejo y una **alta** tasa de defunciones.

- **5 Clusters:** Observemos que de las 3 visualizaciones es la más revuelta, aún así se alcanza a notar cierta separación en los clusters. Por otro lado el perfilamiento aporta



Perfilamiento

cl_km	x_SEXO	x_EIDS	x_poblacion	x_poverty_rate	x_edad	registros	proporcion_muertos
0	0.000000	3.325735	9745.951274	0.687025	45.392363	46197	0.156460
1	0.509838	1.083789	18702.834683	0.828407	45.645913	44218	0.244606
2	1.000000	3.323245	9952.028632	0.688312	46.515981	48932	0.261608
3	0.514275	4.912464	13196.833851	0.442309	47.092591	30262	0.194567
4	0.515593	3.594926	57989.206321	0.650142	47.250543	14269	0.258252

Cluster 0: Mujeres que habitan en una colonia poco poblada, con un IDSM **medio**, de edad un poco más **joven** que el promedio y con una **baja** tasa de defunciones.

Cluster 1: Personas que habitan en una colonia poblada, con un IDSM **muy bajo**, de edad un poco más **joven** que el promedio y una **alta** tasa de defunciones.

Cluster 2: Hombres que habitan en una colonia poco poblada, con un IDSM **medio**, tienen la más **alta** tasa de defunciones.

Cluster 3: Personas que habitan en una colonia poblada, con un IDSM **alto**, de edad un poco más **grande** que el promedio y una tasa **moderada** de defunciones.

Cluster 4: Personas que habitan en una colonia muy poblada, con un IDSM **medio**, de edad un poco más **grande** que el promedio y una **alta** tasa de defunciones.

Conclusión

En general las opciones tienen aspectos positivos que pudieran ser de utilidad para distintos enfoques, para lo requerido se usará el caso de 5 cluster en el que se pueden describir cada uno de ellos de manera resumida de la siguiente forma:

Cluster 0: Mujeres jóvenes que habitan en una colonia con desarrollo deficiente.

Cluster 1: Personas jóvenes que habitan en localidades con muy mal desarrollo.

Cluster 2: Hombres que habitan en una colonia con desarrollo deficiente.

Cluster 3: Personas mayores que habitan en zonas con excelente desarrollo.

Cluster 4: Personas que habitan en localidades grandes con desarrollo deficiente.

Basados en estos clusters se puede observar un claro impacto de la variable IDS, en conjunto con la edad y el sexo en el porcentaje de muertes que hay en cada cluster. El **cluster 3** de las personas que habitan en una localidad con un buen índice de desarrollo, a pesar de ser el grupo con la mayor edad, contra intuitivamente el porcentaje de muertes es de los más bajos, lo que claramente es un impacto del contexto socioeconómico del individuo.

Mientras que en los **clusters 0** y **cluster 2,** ambos clusters son de personas con el mismo IDS y solo un año de edad de diferencia en promedio, a pesar de la poca diferencia que hay entre ellos en estas variables, el género provoca que el aumento de muertes porcentuales entre clusters sea de más del 10%, probablemente debido a que en la sociedad mexicana el hombre es el principal proveedor de ingresos en una familia por lo que se ve más expuesto al salir a laborar, y no goza con los privilegios del **cluster 3** que quizá pudo laborar desde su hogar y cuidarse una vez contagiado o acceder a un diagnóstico oportuno y un tratamiento adecuado, por lo que también se convierte en una variable importante a considerar.

Con todo esto se recomendaría a la institución pública pertinente, atender las necesidades de cada grupo según sean las deficiencias presentadas, prestando principal atención al **cluster 2,** en el que se recomienda hacer inversión en pruebas y protocolos dentro de los empleos que provean de un diagnóstico oportuno a las personas de este cluster y se den los permisos y tiempos suficientes para una pronta recuperación del trabajador. Y para las personas del **Cluster 1** se recomienda una estrategia integral de colocación de módulos de salud en las colonias de los individuos pertenecientes a este cluster, ya que se identificó una deficiencia grave en el entorno de estas personas lo que imposibilita el diagnóstico y tratamiento.

Anexo: Diccionario de Variables

x_SEXO: Sexo del individuo (0 : Mujer , 1 : Hombre).
x_EIDS: Nivel de desarrollo social de la colonia donde habita el individuo. (1 : Bajo , 2 : Medio Bajo, 3 : Medio, 4 : Medio Alto, 5 : Alto)
x_EDAD: Edad del individuo.
x_tot_pobres: Número de pobres en la colonia donde habita el individuo.
x_prom_edad_colonia: Promedio de edad de las personas que viven en la colonia donde habita el
individuo.
x_muertos_colonia_rate: Porcentaje de muertes de personas contagiadas contagiadas que habitan en
la colonia donde habita el individuo.
poverty_rate: Porcentaje de pobres que habitan en la colonia donde habita el individuo.
población: Población total de la colonia donde habita el individuo.
ISDM: Índice de Desarrollo Social de la colonia donde habita el individuo.
x_componente_1 : Primer componente obtenido de reducir los factores del IDSM.
x_componente_2 : Segundo componente obtenido de reducir los factores del IDSM.
x_componente_3 : Tercer componente obtenido de reducir los factores del IDSM.

Componente	Descripción
CCEVj	Calidad y espacio de la vivienda
CAEj	Carencia Energética
TELj	Teléfono
IVj	Internet
CBDj	Bienes durables
rei	Educación
CASSI	Seguridad Social
CASI	Salud
nbi	NBI