Defunciones Argentina

Análisis de la tendencia de fallecimientos por provincia, género y rango etario.

Trouchot Juan Cruz, Tkaczyszyn Vanesa y Amaro Schneider Florencia.

Universidad Tecnológica Nacional

Abstract

En este paper aplicaremos K nearest neighbor regression a un set de datos categóricos con evolución temporal muy amplio con la finalidad de ver la tendencia de estos a lo largo del tiempo, compararemos este método con otros métodos de regresión y sacaremos una conclusiones en base de los datos antes y después de aplicada la regresión.

Keywords
Defunciones, causas, provincia, rango de edad, cie10.

---

1. INTRODUCCIÓN

En el presente informe se realizará un análisis de la tendencia de las defunciones en la Argentina. Se mostrará en forma ordenada la evolución en el tiempo de las razones de muerte por provincia, género y rango etario. Finalmente, se buscará predecir su evolución en el futuro próximo.

1.1 Primeros Pasos

Para comenzar, se inició con una búsqueda de información brindada en su gran mayoría por la página oficial del gobierno de la República Argentina. Allí se consiguió la información relevante y necesaria para nuestro problema planteado.

Toda esta información hallada debió ser depurada y consolidada de forma acorde para luego poder comenzar con un análisis exploratorio de datos (EDA), cuyo objetivo fue encontrar potenciales inconvenientes en los datos y visualizar la información

---

2. Descripción del dataset

Desde distintas fuentes, se consiguieron los siguientes datasets:

2.1 Datasets obtenidos

2.1.1 Defunciones anuales por sexo y por provincia 2010-2017

Fuente: https://datos.gob.ar/dataset/salud-defunciones-ocurridas-registradas-republica-argentina

A continuación se menciona cómo está establecido el dataset.

- Samples: **379625**
- Features: **6**

⇒ Año
⇒ Género de la víctima
⇒ Provincia donde ocurrió la defunción
⇒ Tipo de muerte
⇒ Cantidad de muertos en ese año
⇒ Muerte materna

Todos estos features contaban con una versión en código numérico o alfanumérico y una columna adjunta con la traducción de dicho código.

Debido a que los datos incluidos en “tipos de muerte” son extremadamente numerosos y diversos, se decidió agregar manualmente el motivo de muerte como nomenclador general que engloba los tipos de muertes incluidos en las features. Por lo tanto: “motivo de muerte” pasó a ser una nueva feature. Éstos se obtuvieron de un manual llamado CIE10.

Se decidió eliminar la columna muerte materna, ya que nuestro objetivo desde el inicio es enfocarnos en las muertes en general y este feature contiene nans en todas las causas que no incluyan muertes maternas

Cabe destacar que decidimos no eliminar ningún outlier debido a que estos representan muertes y es un tipo de información que no deseamos eliminar.

Para concluir esta parte, limpiamos Nans.

2.1.2 Población argentina por género y provincia 2002-2025

Fuente:
https://sitioanterior.indec.gob.ar/nivel4_default.asp?id_tema_1=2&id_tema_2=24&id_tema_3=119

A continuación se mencionara cómo está establecido el dataset.

- Features: **3**

⇒ Año
⇒ Provincia
⇒ Cantidad de población por Género

Para poder unir este dataset con el anterior, fue necesario omitir la información de 2002 hasta 2009 y de 2018 a 2025, ya que el primer dataset comienza con información de 2010 y finaliza con 2017.

Aprovechando esta nueva información se agregó una nueva feature llamada **población** que involucra la división entre la cantidad de muertos en ese año y cantidad de población por género. Esto nos da como resultado la cantidad de muertes per cápita por provincia.

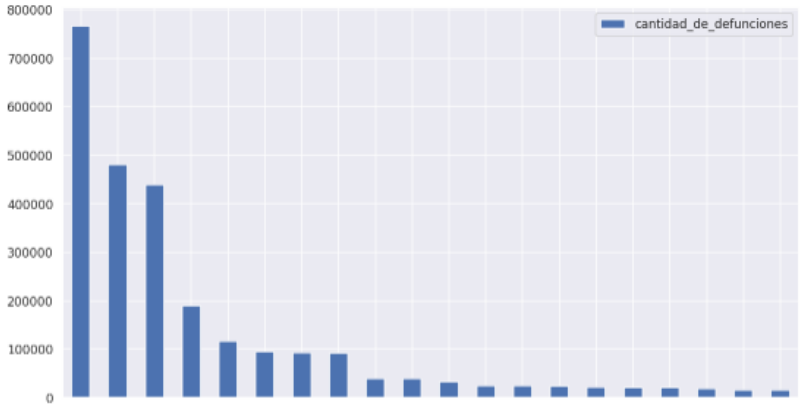
Debimos depurar toda la información para evitar incurrir en errores. Se decidió quitar aquellos campos que aparecían como no definidos de: Género de la víctima, Provincia donde ocurrió la defunción, Rango Etario y Causa de Muerte.

2.2 Datasets no utilizados finalmente

En nuestra búsqueda de información, encontramos otros Datasets que no fueron utilizados finalmente ya que la información de los mismos no se adaptaba bien a nuestro dataset de base, ya que o no contaban con un desagregado por provincia o dependían linealmente de alguna de las variables, entre ellos estaban la geolocalización de cada provincia, el PBI per cápita y el presupuesto en salud.

Por último intentamos encontrar la cantidad de hospitales por provincia, pero esta información no nos fue posible encontrar, se pudo hallar la información de un año específico por provincia, pero nuestro Dataset es de 8 añoS

3. EDA: exploración y análisis de los datos



3.1 Gráficos de barra y torta

Cantidad de muertes totales para cada motivo de muerte por año.

1-Enfermedades del sistema circulatorio

2-Tumores (neoplasias) malignos

3-Enfermedades del sistema respiratorio

4-Síntomas, signos y hallazgos anormales clínicos y de laboratorio no clasificados en otra parte

5-Enfermedades del sistema digestivo

6-Enfermedades endocrinas, nutricionales y metabólicas

7-Ciertas enfermedades infecciosas y parasitarias

8-Enfermedades del sistema genitourinario

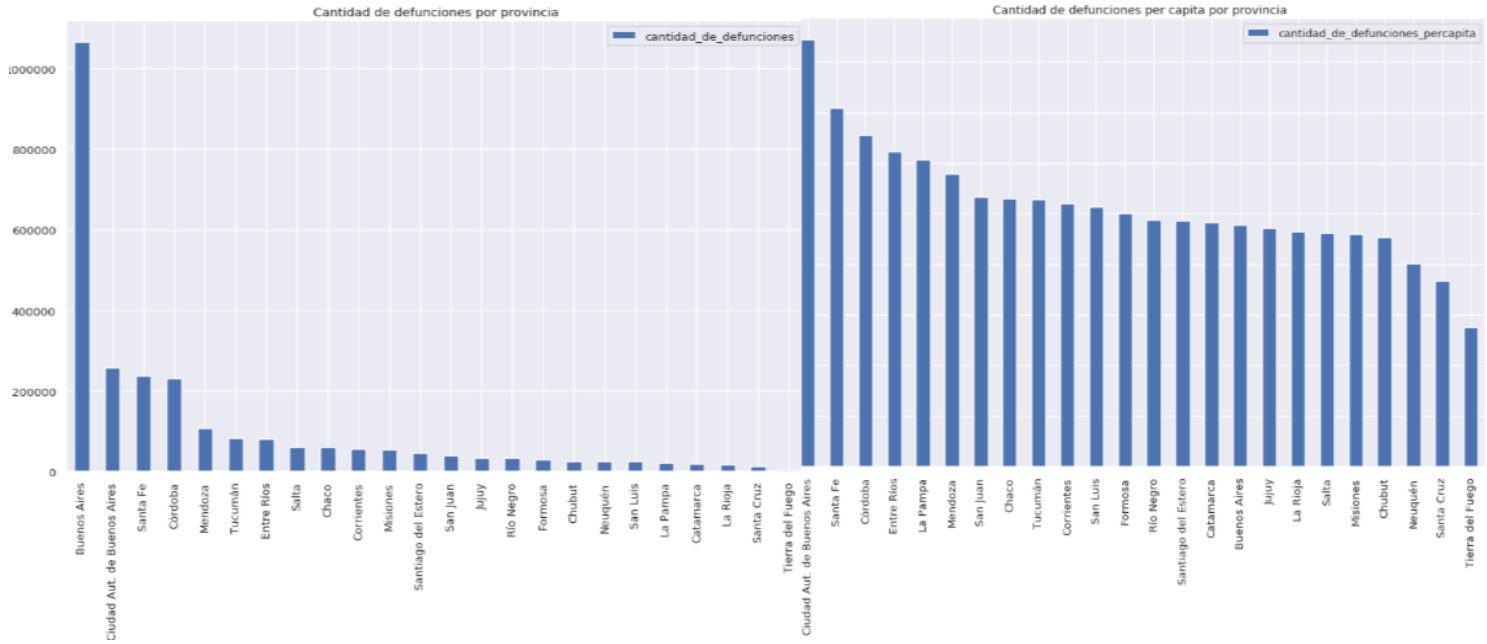
9-Accidentes de transporte

10-Enfermedades del sistema nervioso

11-Ciertas afecciones originadas en el periodo neonatal

12-Eventos de intencion no determinada

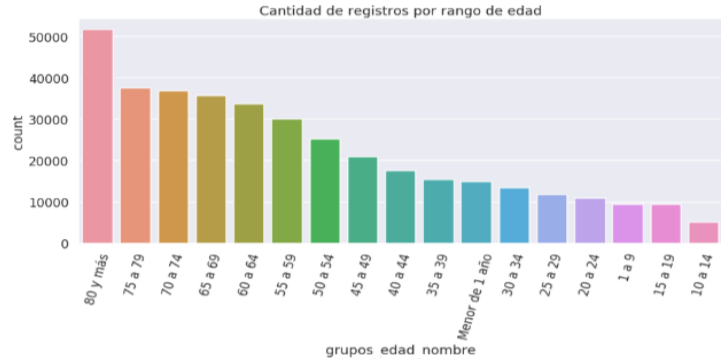
13-Lesiones autoinflingidas intencionalmente



Podemos apreciar que la provincia de Buenos Aires es la que registra mayor cantidad de muertes, esto va de la mano con ser la provincia con mayor cantidad de habitantes.

Ajustando las muertes per cápita Buenos Aires deja de ser un dato relevante, y pasa ser La Ciudad Autónoma de Buenos Aires la zona con más muertes. De todas formas, sabemos que la cantidad que trabaja en capital es mucho mayor a la que vive efectivamente en ésta, lo cual influye en los muertos que hay en la misma, aumentándolos.

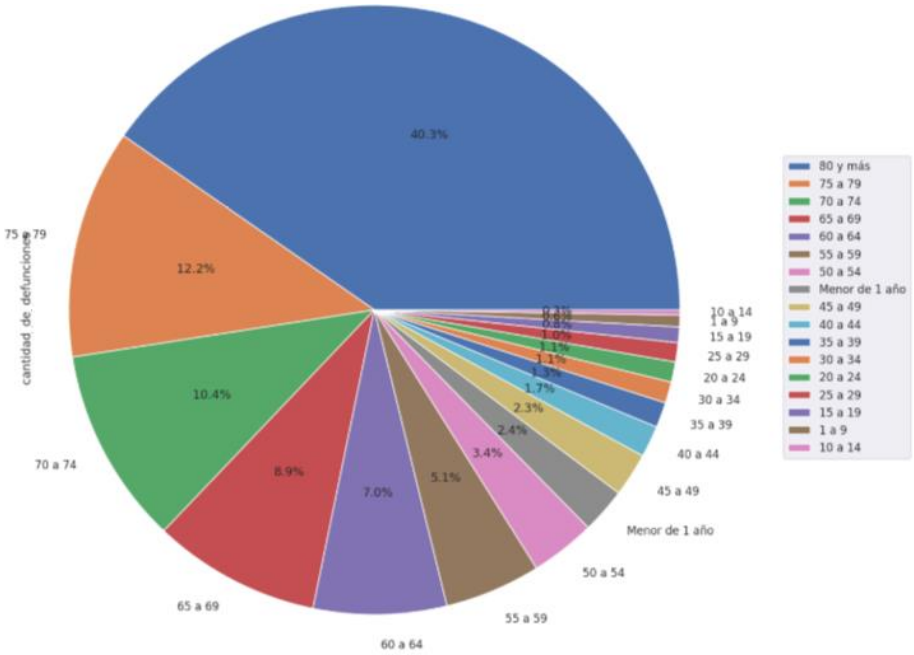
Importante es ver que Santa Fe, Córdoba y Entre Ríos presenten tantas muertes con respecto a su población



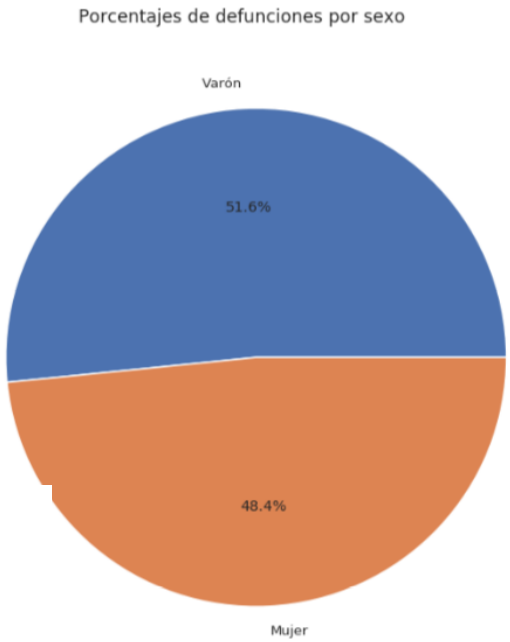
El gráfico muestra los registros de las distintas muertes, no muestra cantidad de defunciones, sino la variabilidad de razones por las que puede morir una persona dependiendo de su rango de edad.

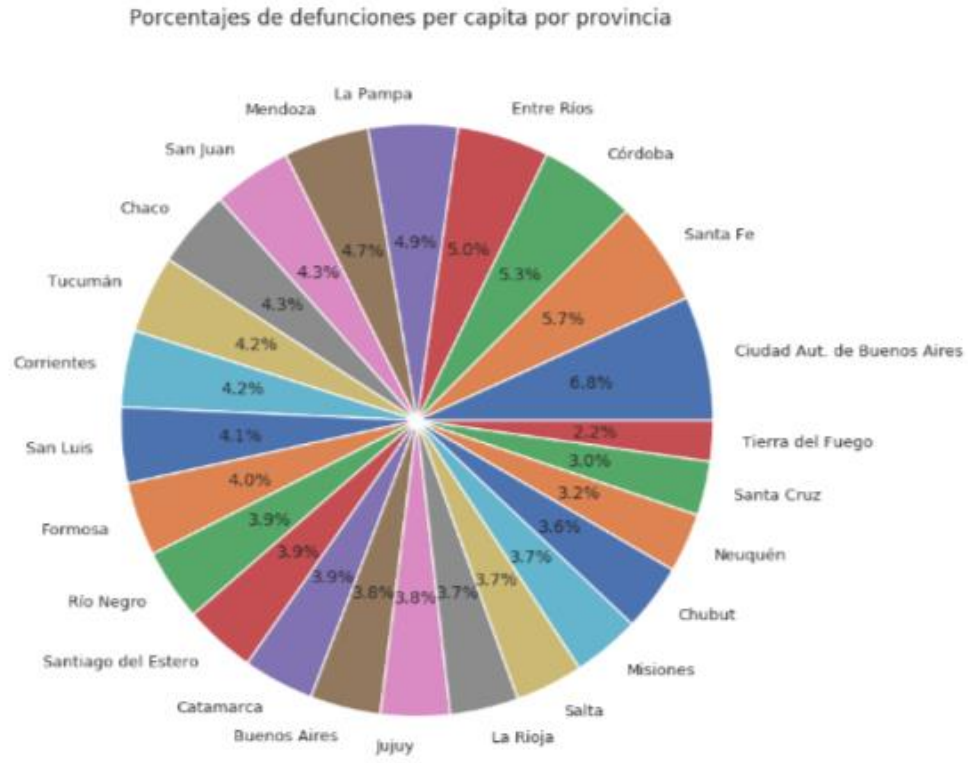
Lo interesante de este último gráfico es que, si bien la gente de 80 años o más es aquella que tiene el mayor número de registros, lo cual era esperable, se observa que tenemos que los menores a 1 año pueden morir de muchas más formas que los menores a 35 años. Y que dentro del grupo de menores de 35 años, el rango de 10 a 14 años son los que tienen menor variabilidad en las causas de muerte para los ocho años analizados.

Porcentaje de defunciones por edad



Esta es nuestra primera aproximación para entender cómo está distribuida la cantidad de muertes. Se puede apreciar que a medida que aumenta el rango etario, aumenta la cantidad de muertes con el 40.3% correspondiente al rango 80 o más.





3.2 Información general del Dataset

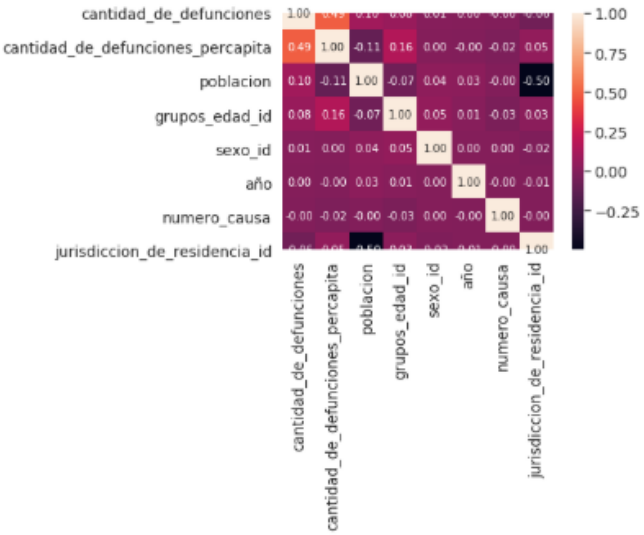
Las principales 10 razones de muerte específicas

Razon	Cantidad
Insuficiencia cardíaca	228.946
Neumonía, organismo no especificado	188.646
Infarto agudo del miocardio	129.242
Otras causas mal definidas y las no especificadas de mortalidad	124.769
Insuficiencia respiratoria, no clasificada en otra parte	85.405
Otras Sepsis	83.057
Tumor maligno de los bronquios y del pulmón	74.533
Accidente vascular encefálico agudo, no especificado como hemorrágico o isquémico	71.402
Diabetes mellitus, no especificada	51.740
Tumor maligno del colon	49.333

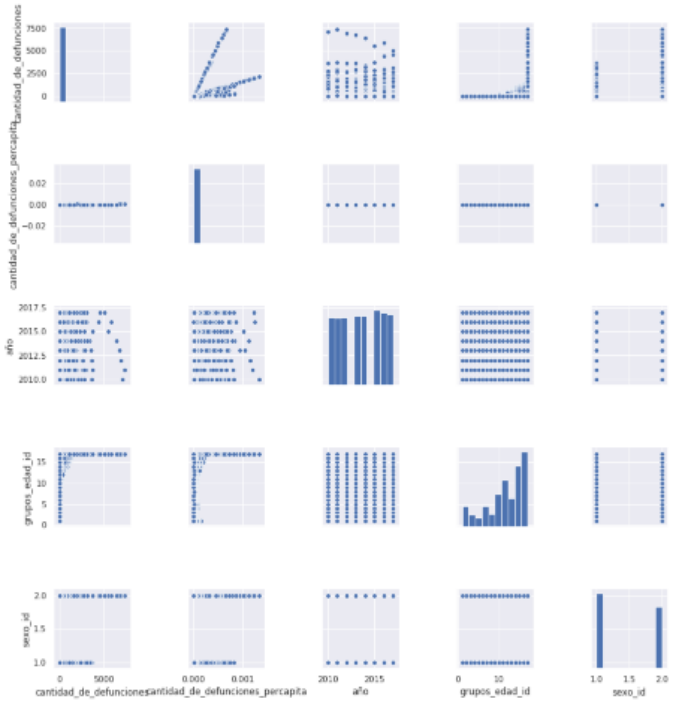
```
#descriptive statistics summary
defunciones['cantidad_de_defunciones'].describe()
```

```
count    379625.000000
mean          6.885827
std          52.982837
min           1.000000
25%           1.000000
50%           1.000000
75%           3.000000
max        7372.000000
Name: cantidad_de_defunciones, dtype: float64
```

Otro dato interesante obtenido de este primer análisis es, como se puede apreciar al describir el dataset, que las causas de muerte son tan variadas que el 75% de éstas son entre 1 a 3 muertes.



Con una matriz de correlación pudimos apreciar que los grupos de edad tienen una leve correlación con las muertes per cápita, y que la población por género y año por provincia tiene correlación inversamente proporcional con el código de la provincia.

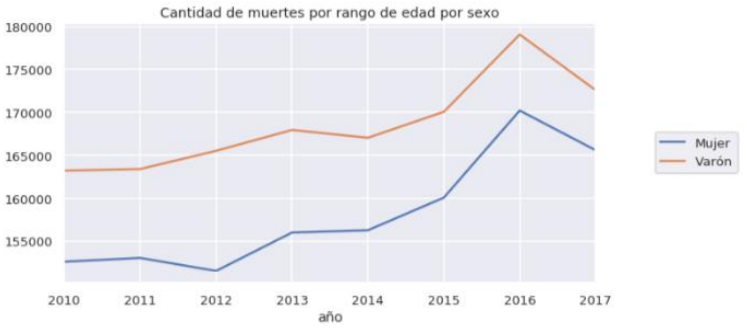


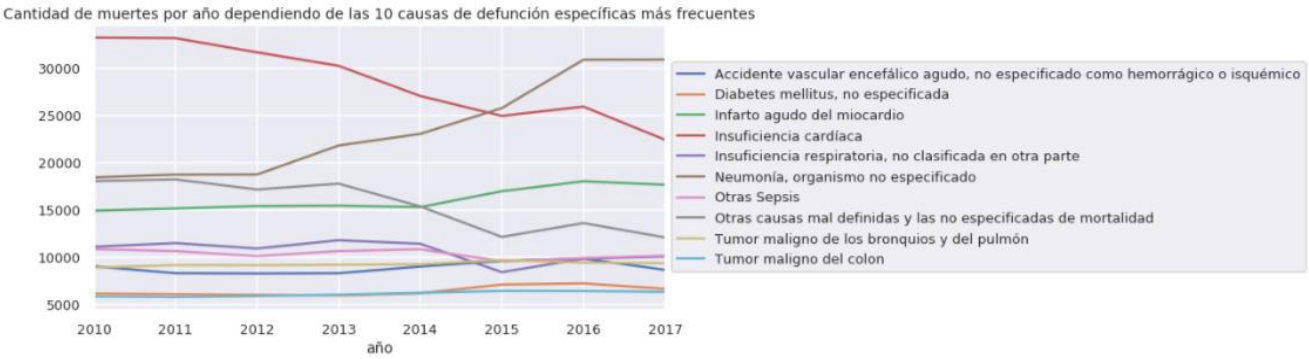
Finalmente, mediante un scatter plot podemos apreciar las relaciones que hay entre las features. La feature género sólo cuenta con hombre mujer, es por eso que solo se separa en dos. Podemos ver la relación grupo de edad y cantidad de defunciones, que se condice con el gráfico anterior, a mayor edad, mayor cantidad de muertes

.3.3 Evoluciones en el tiempo

De este gráfico se puede apreciar que las defunciones totales por año fueron en constante crecimiento hasta el 2016 y luego bajan abruptamente en el 2017.

El comportamiento entre hombres y mujeres es relativamente parecido, y los hombres son los que registran más muertes.

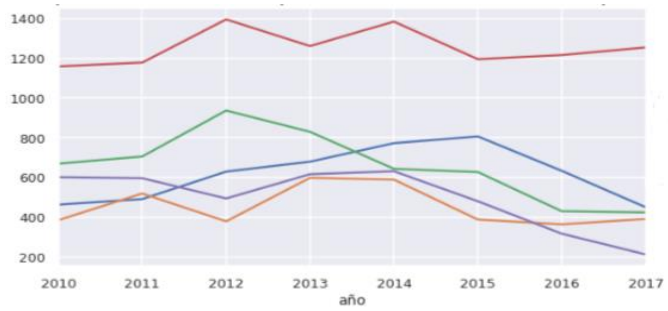




Podemos apreciar a simple vista que la insuficiencia cardiaca va en constante detrimento, las causas mal definidas también van en descenso, lo cual implica una mejora en la carga de los datos originales, por otro lado la neumonía presenta un aumento alarmante a lo largo del tiempo

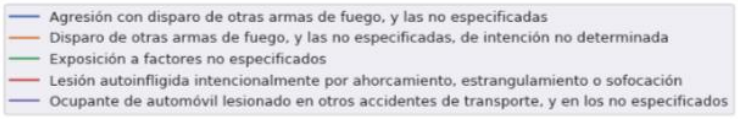
Cantidad de muertes por año dependiendo de las 10 causas generales de defunción más frecuentes

En este gráfico podemos apreciar que las enfermedades del sistema circulatorio son las que más afectan a la población, seguida por los tumores y enfermedades del sistema circulatorio



Cantidad de muertes por año dependiendo de las 5 causas específicas de defunción más frecuentes en personas jóvenes (entre 15 y 34 años)

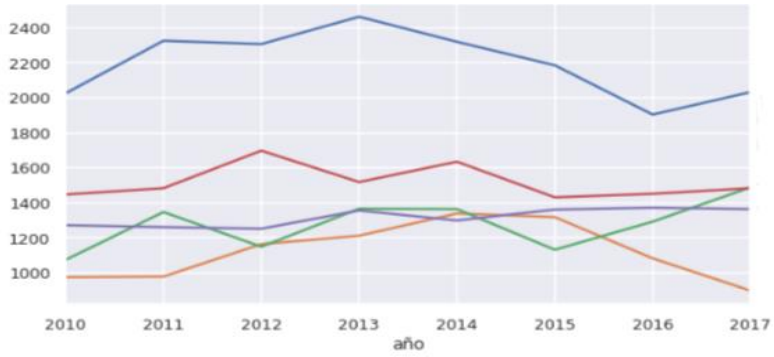
Podemos apreciar que para los jóvenes entre 15 a 34 años la principal causa de muerte es el suicidio por estrangulamiento Seguido últimamente por agresiones con armas de fuego.

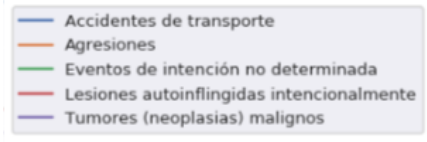


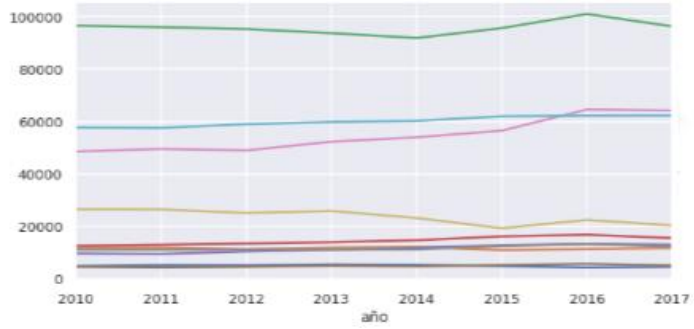
Cantidad de muertes por año dependiendo de las 5 causas generales de defunción más frecuentes en personas jóvenes (entre 15 y 34 años)

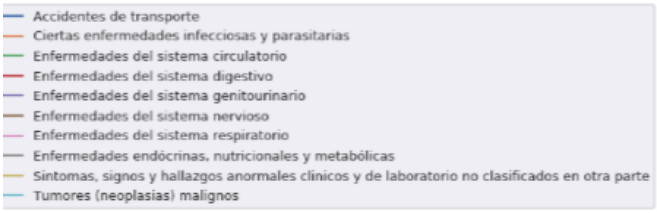
Para este mismo rango su principal causa de muerte general son los accidentes de transporte.

Seguido, como habíamos visto anteriormente, por los suicidios.









---

4. Materiales y Métodos

El proceso se hizo a través de google colab, utilizando las librerías, numpy, pandas, matplotlib.pyplot, seaborn, sklearn (linear_model, svm para el SVR y LinearSVR, neighbors para KNN, decomposition para PCA, metrics y model_selection) y mpl_toolkits.mplot3d

4.1 Agrupación de datos

Una vez que vimos el aspecto en general del dataset, su variabilidad y distribuciones, intentamos predecir la cantidad de muertes a partir de los datos de año, grupo de edad y causa de muerte.
Primeramente agrupamos la información por grupos de edad, año y causa de muerte, y sumamos sus cantidades de defunciones.
A modo de ejemplo, para el año 2010, para el grupo de edad de 1 a 9 años murieron 161 personas en accidentes de transporte (desestimando la provincia ni el género de la víctima).

Esta agrupación logró que el dataset se achicara y cambiara su media, que como habíamos visto en la etapa de exploración era de 6 muertes y el 75% de las muestras no superaba las 4 muertes
Conteo de muestras 3742

Media:	699	Desvío:	3030	Mínimo: 1	Máximo: 50291
25%:	18	50%:	92	75%:	251

Generamos dummies para las causas de muerte y los grupos de edad.
El test size es de 0.01, de esta manera el xtrain es de
3704 samples y 52 features.

4.2 Arendizaje supervisado

Vamos a entrenar un modelo, a traves de entrenamientos buscando minimizar el error de clasificacion.

4.2.1 Cross validation

Para el entrenamiento usaremos este método, el cual usa la porción de dataset que definimos para entrenamiento y lo separa en K folds, para luego iterar K veces.
En cada iteración, una porción se utiliza como validación y el resto como entrenamiento, donde se entrena un modelo con train evaluará el resultado de clasificación con validación. Luego se realizará un promedio de los resultados de exactitud de todas las iteraciones.
El objetivo es buscar que clasificador asegura que nuestro modelo no realiza overfitting (sobreajuste) y clasifica mejor.

4.2.2 Grid-search
Los modelos de clasificación que utilizaremos consistirán de hiper-parámetros que debemos seleccionar, para facilitar el trabajo, esta herramienta enlista los posibles hiper-parámetros y probara todas las posibles combinaciones entre ellas, la que cuente con el mayor Train Accuracy promedio durante el cross-validation será la que usaremos para testear el modelo.

4.3 Metodos de regresion
Una vez determinados los features principales probamos distintos modelos de regresión para predecir las defunciones a través del tiempo

4.3.1 SVR: Support Vector Regression

Este método construye una función lineal determinando un margen como función de costo y trata de que todas las muestras se encuentren dentro de dicho margen.

4.3.2 Regresión lineal

Este método se basa en el cálculo de cuadrados mínimos, calculando parámetros "Beta" asociados a cada feature.

$$\min_{\beta} \|X_w - y\|^2$$

4.3.3 KNN: K Nearest Neighbour

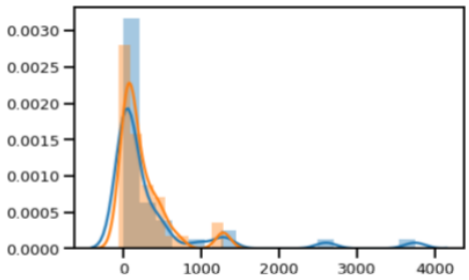
Este método calcula las distancias entre una cantidad K de vecinos más cercanos y predice el Yi interpolando los Y de los K vecinos.
Este método considera pesos para la interpolación.

5.Resultados
5.1 Métodos de regresión
utilizando los métodos descriptos anteriormente obtuvimos los siguientes resultados.

5.1.1 SVR: Support Vector Regression
Elegimos 5 folds para el cross validation
Los mejores parámetros luego del cross validation fueron
C 100
gamma 0.1
kernel lineal

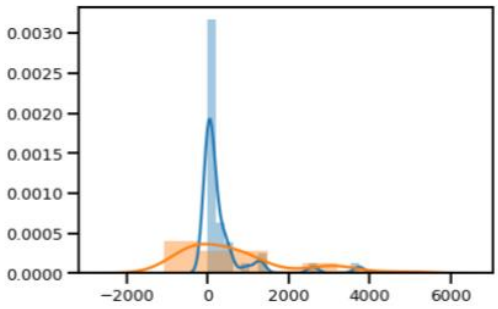
Pero el mejor score fue de 0.067
Root mean squared error de 657.14
Mean absolute error 282.65
Promedio de accuracy 0.2253

Vemos que el método no fue el indicado para predecir los datos



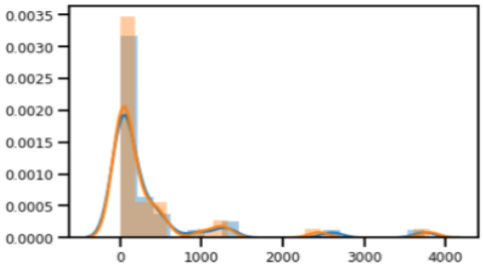
En azul vemos las etiquetas de entrenamiento y en naranja las etiquetas de la comprobación.

5.1.2 Regresión lineal
Error de train: 1122.68
Mean squared error: 1260425.53
Mean absolute error: 786.47



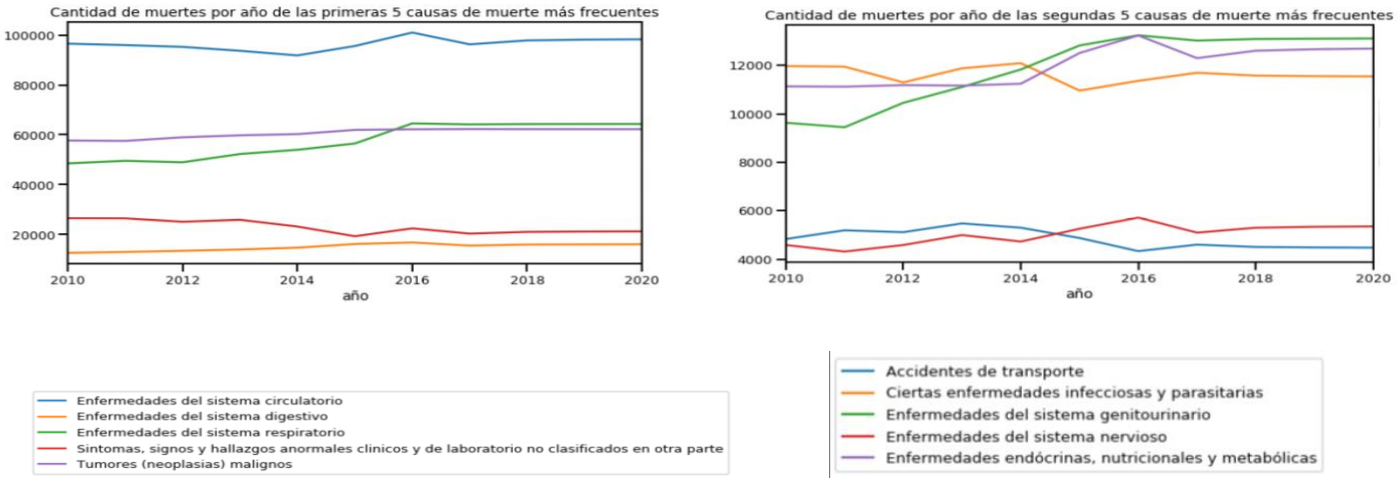
es el método que peor ajusta a los datos, permitiéndonos ver que no es posible ver el problema de forma lineal.

5.1.3 KNN: K Nearest Neighbour
Elegimos 5 folds para el entrenamiento de cross validation
k vecinos: 2
Best score: 0.996
Raíz cuadrada del error cuadrático medio entre ytest (etiqueta real) vs ypred (etiqueta estimada por el modelo): 37.22



Este método es el que mejor ajusta a los datos brindados y es el que utilizaremos para predecir las defunciones a futuro

5.2 Defunciones 2018, 2019 y 2020
Intentaremos predecir, dado el rango de edad y causa de muerte, la cantidad de defunciones para los años 2018, 2019 y 2020 utilizando el método KNN.
Para esto calculamos con dummies.



Podemos apreciar cómo el modelo mantiene la tendencia de los años anteriores y a medida que aumentan los años estas cantidades se mantienen constantes.

cantidad de defunciones por año											
causa muerte	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Enfermedades del sistema circulatorio	96.533	95.956	95.242	93.666	91.811	95.591	100.980	96.256	97.825	98.138	98.274
Enfermedades del sistema digestivo	12.619	12.990	13.454	13.972	14.698	16.173	16.793	15.535	15.950	16.029	16.066
Enfermedades del sistema respiratorio	48.518	49.539	48.954	52.272	53.967	56.478	64.548	64.161	64.284	64.308	64.321
Sintomas, signos y hallazgos anormales clinicos y de laboratorio no clasificados en otra parte	26.519	26.464	25.086	25.873	23.180	19.292	22.451	20.344	21.039	21.181	21.241
Tumores (neoplasias) malignos	57.691	57.532	58.968	59.764	60.252	61.944	62.181	62.272	62.237	62.229	62.225
Accidentes de transporte	4.830	5.189	5.113	5.474	5.297	4.872	4.329	4.598	4.502	4.483	4.473
Ciertas enfermedades infecciosas y parasitarias	11.971	11.946	11.296	11.878	12.091	10.958	11.360	11.695	11.577	11.555	11.544
Enfermedades del sistema genitourinario	9.631	9.446	10.454	11.107	11.830	12.821	13.242	13.023	13.089	13.100	13.110
Enfermedades del sistema nervioso	4.582	4.311	4.582	4.993	4.725	5.253	5.713	5.094	5.296	5.334	5.352
Enfermedades endócrinas, nutricionales y metabólicas	11.133	11.121	11.185	11.165	11.241	12.510	13.237	12.298	12.605	12.666	12.694

---

6. Discusión y conclusiones

Del EDA concluimos que las enfermedades cardio respiratorias son el principal problema que tiene la sociedad argentina, en particular la neumonía.
Otras causas graves son los tumores y accidentes de tránsito.
Los accidentes de tránsito y los suicidios son las principales causas de muerte para los jóvenes.
Y por último Buenos aires, Santa Fe y Córdoba, son las provincias con mayor mortalidad de sus habitantes.
Entre los 3 métodos utilizados concluimos que la regresión lineal era la peor para predecir este tipo de datos, sin embargo el método SVR tuvo una performance muy mala y se ajustó mal a los datos, esto quizá se debió al nivel de variabilidad con la que éstos cuentan.
El método KNN fue el mejor de los tres modelos de regresión y permitió predecir de forma aceptable la evolución de las muertes a lo largo del tiempo. Dos cosas que podemos destacar es que el método prevé un aumento en la cantidad de defunciones a lo largo del tiempo, concordante con la tendencia general de los ultimo 9 años, pero al momento de calcular por causa de muerte estas cantidades no varían significativamente entre el 2018 al 2020, De esta manera consideramos que es un buen método para ver la tendencia de los datos pero no es ideal para predecir de forma exacta el futuro de estos.

---

7. Fuentes

- **Cantidad de defunciones histórico:** https://datos.gob.ar/dataset/salud-defunciones-ocurridas-registradas-republica-argentina
- **Presupuesto en salud:** https://www.presupuestoabierto.gob.ar/sici/datos-abiertos#
- **Proyecciones poblacionales por sexo y provincia:** https://sitioanterior.indec.gob.ar/nivel4_default.asp?id_tema_1=2&id_tema_2=24&id_tema_3=119
- **caracteristicas y localizacion de las provincias:** https://datos.gob.ar/dataset/modernizacion_7/archivo/modernizacion_7.1
- **PBI:** https://datos.gob.ar/dataset/modernizacion_1/archivo/modernizacion_1.17

Libros de referencia:

- The Elements of Statistical Learning Data Mining, Inference, and Prediction - Trevor Hastie Robert Tibshirani Jerome Friedman
- Pattern recognition and machine learning - Christopher M. Bishop
- Hands-On Machine Learning with Scikit-learn and TensorFlow - Aurelien Geron