Análisis de los factores de riesgo para el colesterol alto, hipertensión y diabetes por regiones en Argentina

Alumna: Diana Yelós

CONTENIDO

1. Introducción
2. Objetivo
3. Dataset
4. Procesamiento y revisión de datos
5. Análisis mediante aplicación de técnicas de aprendizaje supervisado
6. Análisis mediante aplicación de técnicas de aprendizaje no supervisado
7. Conclusiones

1. INTRODUCCIÓN

Las enfermedades crónicas No transmisibles: **hipertensión, colesterol y diabetes** junto con la enfermedades cardio y cerebrovasculares, cánceres y enfermedades respiratorias representan un 73.4% de las muertes anuales en la Argentina, el 52% de años de vida perdidos por muerte prematura y 76% de años de vida ajustados por discapacidad (fuente).

En gran porcentaje todas estas enfermedades comparten los mismos factores de riesgo (que explican 3 de cada 4 muertes) y pueden ser prevenibles mediante comportamientos personales y sociales adecuados.

- Alimentación
- Sobrepeso
- Inactividad física
- Tabaco
- Alcohol, estrés, etc.

2. OBJETIVO GENERAL

Dado que la Argentina es un país extenso y con una gran diversidad en todo su territorio, se presentan muchas particularidades en cuanto a la cultura de cada región: diferencias en la alimentación, tipo de trabajo, forma de vida, ingresos, educación, etc.

El objetivo de este trabajo será determinar si estas características son lo suficientemente distintivas como para que puedan existir diferencias en la importancia de los factores de riesgo del colesterol, la hipertensión y diabetes por región. De cumplirse la hipótesis, sería posible optimizar la prevención de las mismas de acuerdo a cada una de las regiones.

3. DATASET

Datos obtenidos de la "4ta Encuesta Nacional de Factores de Riesgo (ENFR)" - desarrollada entre septiembre y diciembre de 2018 por INDEC, DPE, Min. salud y desarrollo social de la nación.

PASO 1: AUTORREPORTE

- Vivienda y hogar
- Situación laboral
- Salud general
- Actividad física
- Tabaco
- Hipertensión arterial
- Peso corporal
- Alimentación
- Colesterol
- Consumo de alcohol
- Diabetes
- etc…

100% (29224)

PASO 2: MEDICIONES ANTROPOMÉTRICAS

- Presión arterial
- Peso
- Talla
- Perímetro de la cintura

56.7% (16577)

PASO 3: MEDICIONES BIOQUÍMICAS

- Glucemia capilar
- Colesterol total

35.4% (10355)

Manual de uso de la base de datos de usuario 4° ENFR

4. PROCESAMIENTO Y REVISIÓN DE DATOS

4.1 Procesamiento de columnas

- **Eliminación de columnas:**
  - Información duplicada para el análisis (provincia y región)
  - Información NO relevante en la determinación de las causas de la enfermedad (fecha de medición del colesterol)

- **Reconversión o eliminación de nulos:**
  - Nulos esperados debido a condiciones de preguntas previas: se convertían a “0” o algún valor que permitiera interpretarlos como tales.
  - Nulos por error o faltantes: Se eliminaron luego de haber procesado todo el dataset (10%)

- **Codificación de variables categóricas:** OneHotEncoder()

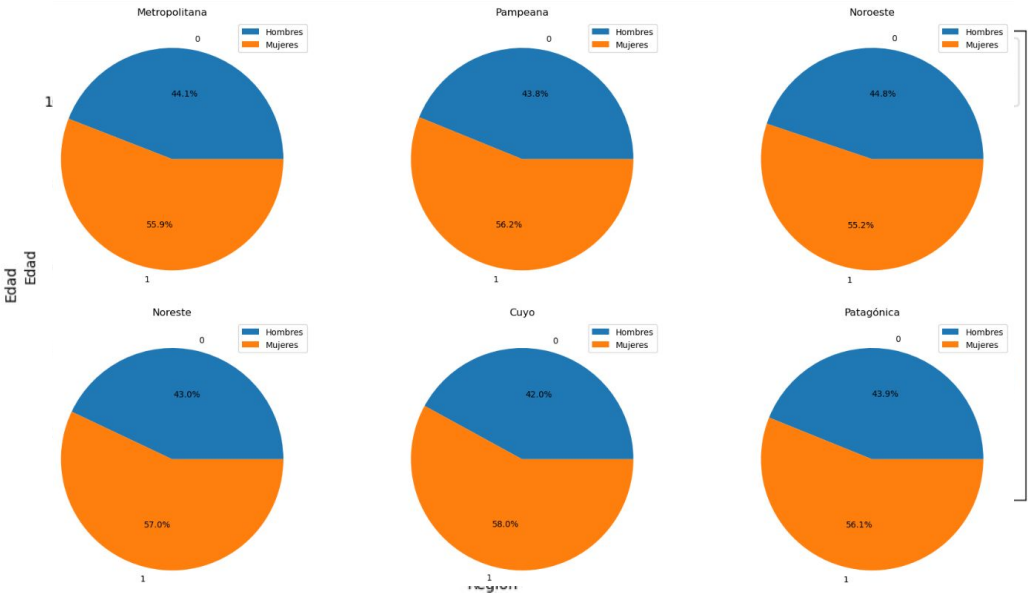
- **Escalado de variables numéricas continuas:** StandardScaler()

4. PROCESAMIENTO Y REVISIÓN DE DATOS

4.2 Análisis exploratorio de los datos

- Determinar existencia de diferencias regionales en las ECNT
- Analizar equilibrio en dataset luego del procesamiento

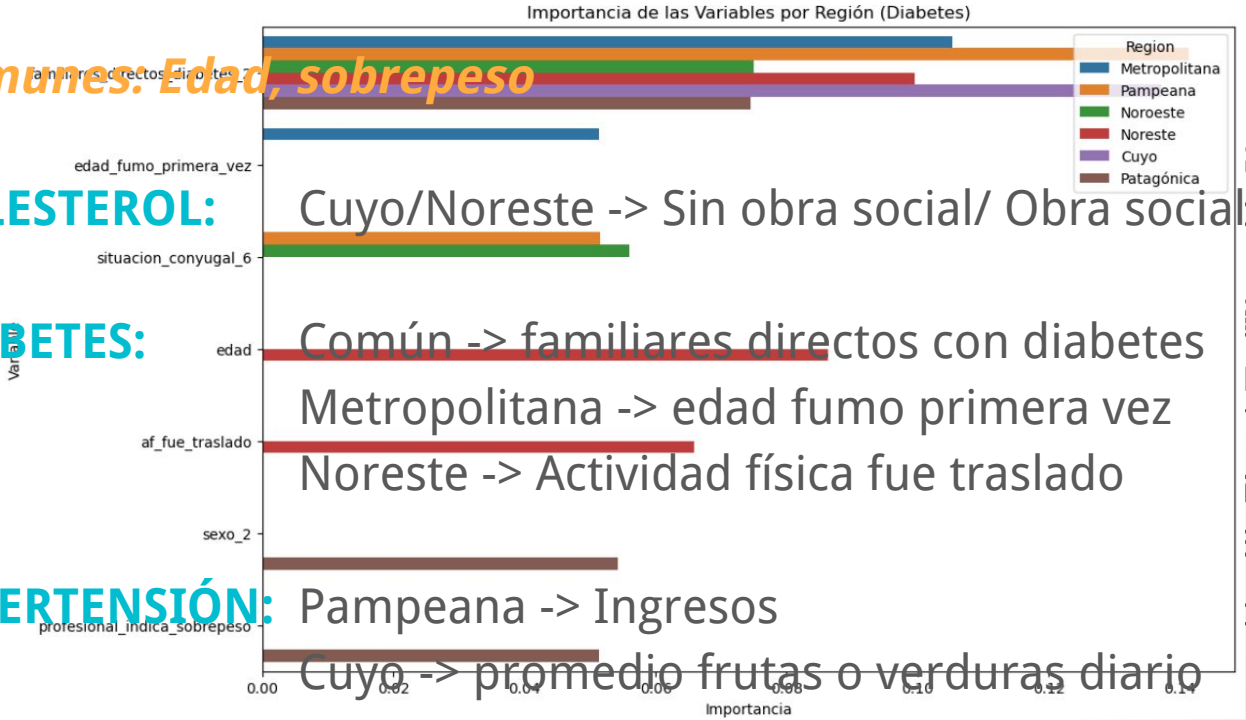
Δ8.66%



5.1 ANÁLISIS MEDIANTE APLICACIÓN DE TÉCNICAS DE APRENDIZAJE SUPERVISADO - ÁRBOLES DE DECISIÓN

Se calcularo
y por REGI
métricas:

- Exactitu
- Precisi
- Sensib
- F1-scor
- Matriz



Comunes: Edad, sobrepeso

COLESTEROL: Cuyo/Noreste -> Sin obra social/ Obra social

DIABETES: Común -> familiares directos con diabetes
Metropolitana -> edad fumo primera vez
Noreste -> Actividad física fue traslado

HIPERTENSIÓN: Pampeana -> Ingresos
Cuyo -> promedio frutas o verduras diario

bilidad
ilares para
le una misma

unos falsos +
le métricas
tes para las
y sensibilidad
ación
similares
e hiper
isión y sensib.

5.2 ANÁLISIS MEDIANTE APLICACIÓN DE TÉCNICAS DE APRENDIZAJE SUPERVISADO - REGRESIÓN LOGÍSTICA

Comunes: Edad

Se calcularon para ENFERMEDAD y por REGIÓN las siguientes métricas:

- Exactitud
- Precisión
- Sensibilidad
- F1-score
- Matriz de confusión

COLESTEROL: Cuyo -> baño
Metropolitana/Noreste -> Obra social

DIABETES: Común -> fuma cigarrillos/Metropolitana/Cuyo -> baño
Cuyo -> Obra social/actividad física

HIPERTENSIÓN: Común -> Peso
Metropolitana -> baño/nivel instrucción

COLESTEROL: Exactitud, sensibilidad y precisión **similares** para todas las regiones dentro de una misma enfermedad

Métricas 70-75. falsos + y falsos -

Diabetes: Exactitud, precisión y sensibilidad **considerablemente diferentes** para las distintas enfermedades

Modelo muy sólido, buena predicción

Hipertensión: Exactitud, precisión y sensib. (70)

Tiene algunos falsos + y falsos -

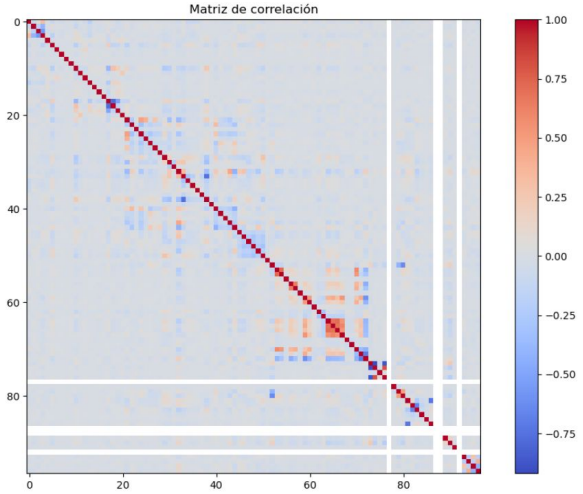
2. Se observaron resultados de métricas considerablemente diferentes para las distintas enfermedades
3. Se observaron resultados levemente mejores luego de la optimización de hiper parámetros con GridSearch ***(mejor que modelo anterior)***

6. ANÁLISIS MEDIANTE APLICACIÓN DE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO

97 columnas iniciales

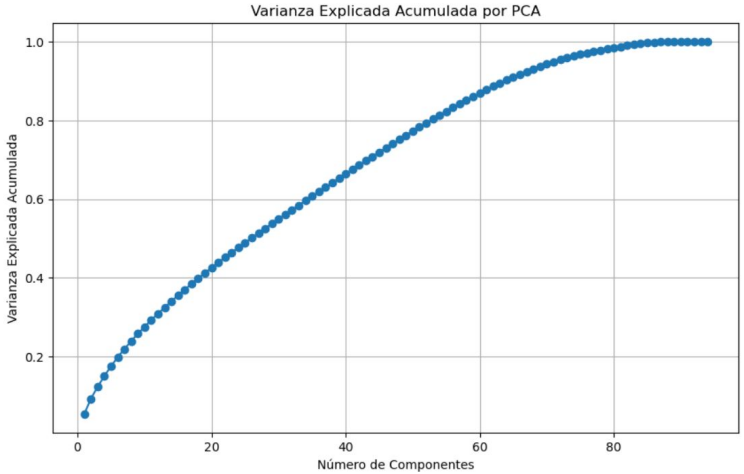
REDUCCIÓN DE DIMENSIONALIDAD

1. MATRIZ DE CORRELACIÓN



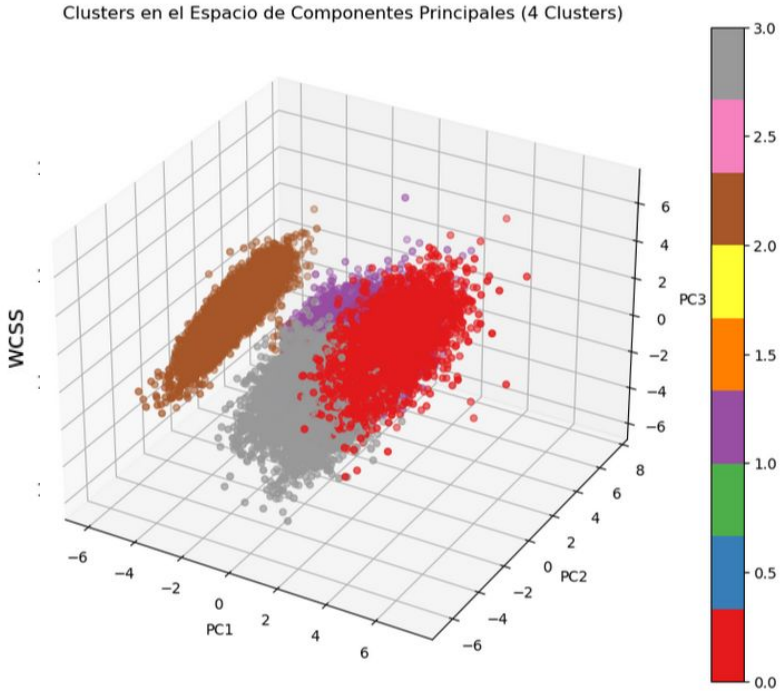
THR>0.75 -> Eliminamos 3 columnas

2. COMPONENTES PRINCIPALES (PCA)

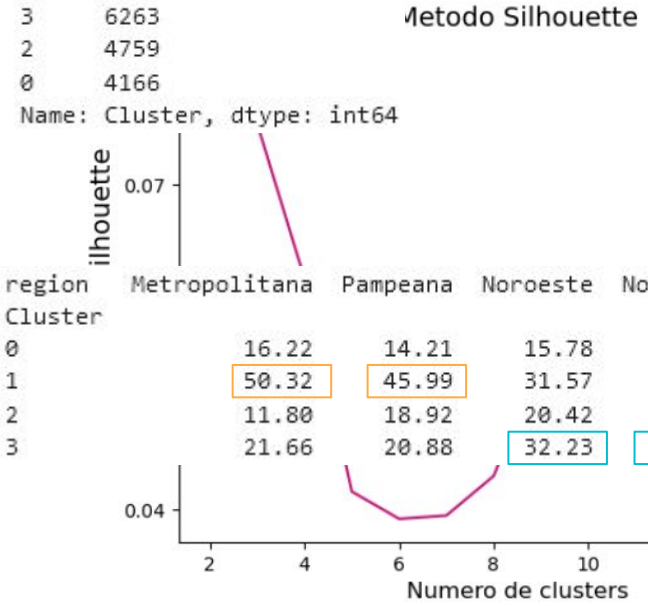


Varianza = 80% -> Quedan 53 columnas

6.1 ANÁLISIS MEDIANTE APLICACIÓN DE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO - KMEANS

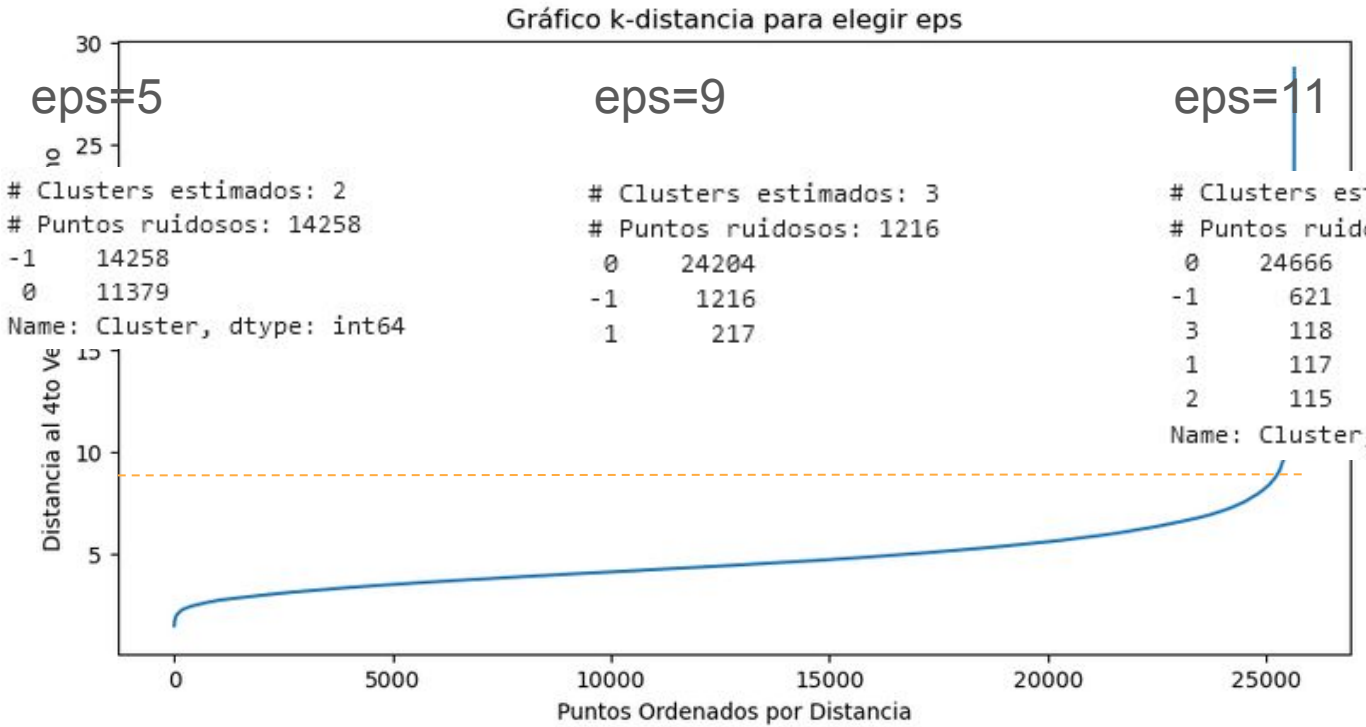



```
1    10449
3     6263
2     4759
0     4166
Name: Cluster, dtype: int64
```



region	Metropolitana	Pampeana	Noroeste	Noreste	Cuyo	Patagónica
Cluster						
0	16.22	14.21	15.78	10.68	27.20	19.88
1	50.32	45.99	31.57	28.13	39.14	45.41
2	11.80	18.92	20.42	25.01	14.41	17.84
3	21.66	20.88	32.23	36.18	19.25	16.87

6.2 ANÁLISIS MEDIANTE APLICACIÓN DE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO - DBSCAN



eps=5

```
# Clusters estimados: 2
# Puntos ruidosos: 14258
-1    14258
 0    11379
Name: Cluster, dtype: int64
```

eps=9

```
# Clusters estimados: 3
# Puntos ruidosos: 1216
 0    24204
-1     1216
 1      217
```

eps=11

```
# Clusters estimados: 5
# Puntos ruidosos: 621
 0    24666
-1      621
 3      118
 1      117
 2      115
Name: Cluster, dtype: int64
```

7. CONCLUSIONES

- En todas las enfermedades y para todas las regiones consideradas se observan como factores muy fuertes la EDAD y el SOBREPESO

- Las métricas para la Diabetes y por lo tanto, los resultados obtenidos se observan bastante robustos tanto en el caso del árbol de decisión y en regresión logística (es llamativa la no aparición del factor genético en la regresión logística).

- Se aconsejaría probar otro tipo de ajustes para el colesterol e hipertensión que presentaran mejores métricas.

- A partir de clusterización con KMEANS se puede inferir que las 6 regiones no son la mejor forma de dividir la población, tal vez 2 regiones: Norte y Sur
  Se observa que en general las características observadas pertenecían a más de una región, lo que reafirma la conclusión obtenida mediante KMEANS