Clasificación de personas de riesgo y predicción de su presión sistólica

Benedetto Matías; Fatur Tomas; Hara Tobías

***Abstracto*—En el siguiente trabajo realizamos la predicción de la presión sistólica de personas por medio de distintos modelos de regresión aplicados a la Encuesta Nacional de Factores de Riesgo, conociendo ciertas características de los mismos. Además, aplicamos modelos de clasificación con el fin de predecir si una persona es paciente de riesgo o no a través de features conocidas. El mejor resultado obtenido fue utilizando el modelo de Support Vector Machines logrando un accuracy del 79%.**

1. *Introducción*

Este trabajo fue realizado utilizando la Encuesta Nacional de Factores de Riesgo cuyo objetivo es proporcionar información válida, confiable y oportuna sobre factores de riesgo (como consumo de tabaco, alcohol, actividad física), procesos de atención en el sistema de salud y principales enfermedades no transmisibles en la población (hipertensión, diabetes, obesidad y otras). Con este propósito, esta encuesta pretende conocer la situación de cada individuo en cuanto a su situación de salud general.

1. *Descripción del dataset*

El dataset utilizado está compuesto originalmente por 29.224 samples y 287 features. Cada sample representa un individuo encuestado y cada feature la pregunta/medición realizada al individuo con su respectiva respuesta/resultado.
Se decidió mantener las siguientes features con el objetivo de realizar un óptimo análisis exploratorio de datos y aplicación de modelos de aprendizaje (muchas de estas features arrojaban más de la mitad de sus valores nulos). Estas variables elegidas son las que creemos más útiles para lograr el objetivo de clasificar a personas que son pacientes de riesgo y también estimar la presión sistólica de los individuos encuestados

- Provincia
- Tipo de vivienda
- Tipo de hogar
- Ingresos en el Hogar
- AUH
- Sexo
- Edad
- Situación Conyugal
- Nivel de instrucción
- Condición de actividad
- Salud general
- Cobertura de salud
- Actividad física por semana
- Barreras en la actividad física
- Si fuma cigarrillo
- Cantidad de veces que fue diagnosticado con presión alta
- Peso
- Altura
- Masa corporal
- Cantidad de días a la semana que come verduras
- Tipo de alimentación
- Colesterol
- Si bebió alcohol
- Si ha sido diagnosticado con diabetes
- Presión medida de la presión sistólica
- Presión medida de la presión diastólica

Dadas estas features, se procedió a reemplazar los valores numéricos que representan las respuestas de los individuos, por su respectiva categoría en formato de caracteres siguiendo la bibliografía dada por el set de datos.
Por último, se realizó la limpieza correspondiente, eliminando filas con valores nulos. De esta manera, el set de datos a trabajar quedó conformado por 15.912 samples y 27 features

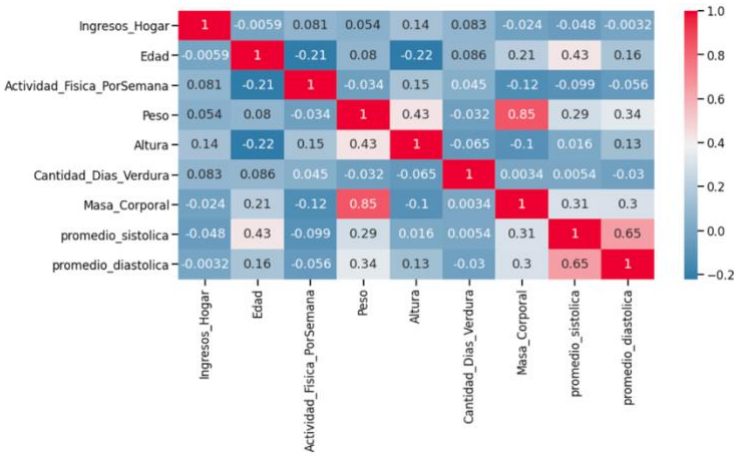
1. *Análisis Exploratorio de datos*

▪ *Matriz de correlación*

Llevamos a cabo esta matriz para entender las relaciones lineales entre pares de variables.

Pudimos sacar algunas conclusiones luego de obtenido el gráfico:

- La actividad física puede ayudar a reducir la presión arterial sistólica: En la matriz podemos observar que la relación lineal entre la cantidad de veces que una persona realiza actividad física y la presión sistólica, es negativa.
- A mayor edad, en general la presión sistólica de las personas aumenta: observamos en la matriz una correlación lineal positiva entre estas variables.
- Observamos también una correlación lineal positiva entre el peso y las presiones sistólicas y diastólicas. Se puede llegar a deducir que personas con sobrepeso presentan presiones más altas.

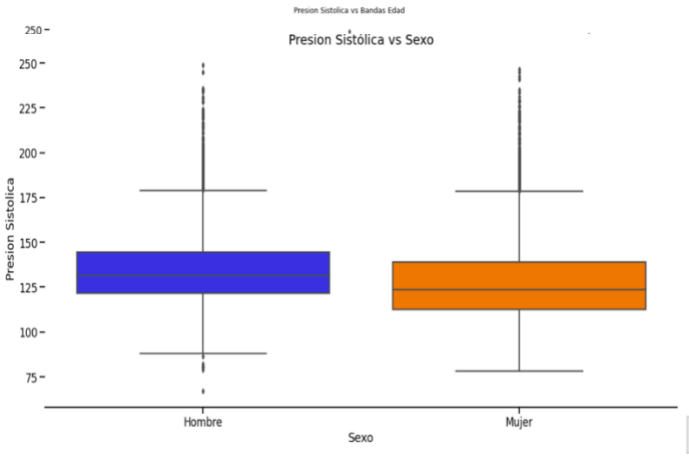


Boxplot- Presión Sistólica vs Bandas de edad

Analizamos cómo variaba la presión sistólica según el rango de edad. Como dijimos previamente, vemos que en rangos etarios más altos, se corresponden presiones sistólicas más altas. Se puede ver por el ancho del Inter Quartile Range que en rangos de edad altos, muchos de los encuestados tienen problemas de hipertensión (presión sistólica mayor a 140 mmHG).

Boxplot-Presión sistólica vs Sexo

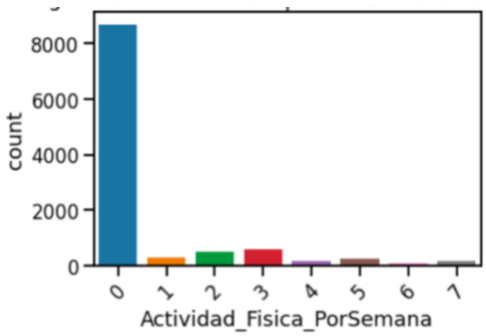
Creímos conveniente analizar si existen diferencias de presiones entre sexos. Encontramos que las personas del sexo femenino tienen presiones que normalmente son bajas a diferencia de los hombres que presentan una presión mayor.



Actividad física por semana en los pacientes de riesgo

Los expertos en salud indican que la actividad física tiene importantes beneficios para la salud y contribuye a prevenir las ENT. Esto se puede ver claramente en el gráfico que analizamos a continuación.

Observamos que casi la totalidad de los pacientes de riesgo no realizan actividad física a la semana.



1. Materiales y métodos

En base a la información del dataset y los resultados obtenidos del Análisis Exploratorio de Datos, se utilizarán modelos de aprendizaje supervisado para, a partir de samples Xi (vector de features) y labels Yi (variable dependiente), encontrar una función f'(x) cuyo output y' sea lo más similar a las y dadas.

En el caso de este trabajo práctico, fueron utilizados dos tipos de aprendizaje supervisado:

- Clasificación, en donde las etiquetas toman valores discretos.
- Regresión, en el cual las etiquetas aprendidas adoptan valores continuos.

4.1. Clasificación

En este caso, la aplicación de modelos de aprendizaje supervisado se utilizan para predecir si

una persona será de riesgo en base a las siguientes features:

- Edad
- Peso
- Cantidad de días a la semana que realiza actividad física
- Sexo
- Si la persona ha fumado cigarrillo

En el caso de las últimas dos variables, debieron ser transformadas a dummies para que adopten valores binarios.

Elegimos estas features tomando como input el heatmap realizado en el análisis exploratorio de datos, viendo las relaciones lineales entre las características y teniendo en cuenta cuales son las que presentan mayor relación. Además creemos que son las que más influyen para poder clasificar si una persona es de riesgo o no.

Label

Para que el modelo clasifique a una persona como paciente de riesgo, la misma debe cumplir con una o más de las siguientes condiciones:

- Diabetes
- Colesterol alto
- Tener un índice de masa corporal (IMC) mayor o igual a 30.
- Tener una presión sistólica mayor o igual a 140mmHg.

En el caso del IMC, se considera el valor a partir del cual la persona tiene obesidad, según información de la OMS

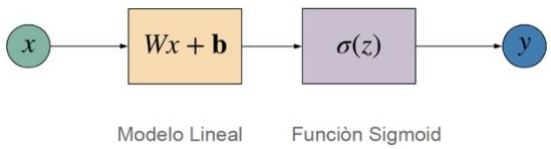
Para esto fue necesario crear una nueva label que agrupe a personas que reúnen alguna de estas condiciones ("Riesgo").

Para realizar las predicciones, se utilizaron los siguientes modelos:

- Logistic Regression
- SVM
- Naive Bayes
- KNN

Logistic regression

Es un clasificador lineal procedido de una función de activación Sigmoid, lo que genera que el output sea binario. A cada Sample, le asigna una probabilidad de pertenecer a cada clase, y, en caso de ser mayor a cierto threshold (0,5), entonces pertenece a esa clase o viceversa.



Logistic Regression, Pattern Recognition, Bishop

Probabilidad de la clase yi dado un X:

Probabilidad de la clase Y_i dado un vector de entrada x.

$$p(y_i|\mathrm{x}) = \sigma(w^T \mathrm{x})$$

Logistic Regression, Pattern Recognition, Bishop

Support Vector Machines

Es un clasificador lineal cuyo objetivo es encontrar un hiperplano separador que maximice el margen entre las clases, el cual está definido por un subconjunto de muestras que se denominan "Support Vectors". Cada muestra mal clasificada es penalizada por un costo C (hiperparámetro)

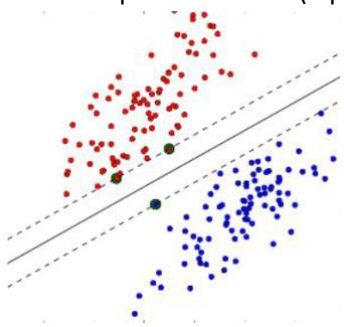


Figura obtenida de los apuntes de cluster AI

Naive Bayes

Clasificador basado en el teorema de Bayes en el cual se presume una "inocente" independencia entre features.

Permite que cada distribución se pueda estimar de manera independiente como una distribución unidimensional, siendo conocido como un método simple.

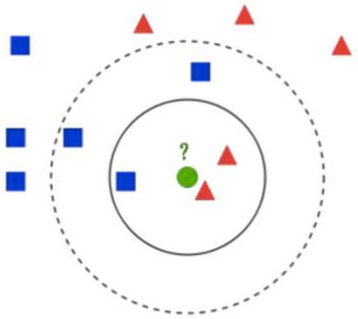
$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

KNN

Modelo que clasifica cada nuevo dato en el grupo que corresponda, según tenga

K vecinos (hiperparámetro) más cerca de un grupo o del otro.

Selecciona la etiqueta y qué más frecuente aparece entre las K clases para clasificar a la muestra en cuestión.



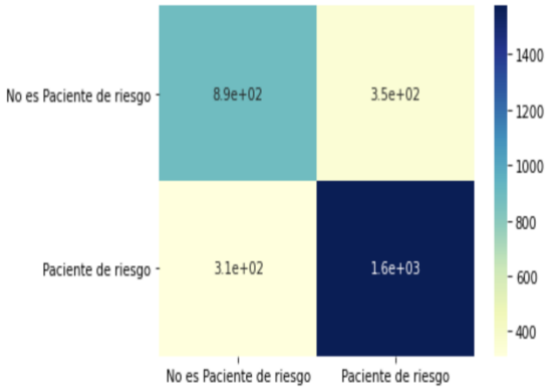
$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Figuras obtenidas de los apuntes de cluster AI

- ***Resultados***

Con un train size que representa el 80% de los datos, el mejor accuracy (TN+TP/Total) obtenido es a través del modelo Support Vector Machines. En este caso, la matriz de confusión arroja los siguientes resultados:

Como se puede visualizar en la imagen superior, hay un total de 893 personas que fueron clasificados correctamente como no de riesgo, mientras que hay 1575 clasificadas correctamente como pacientes de riesgo. En el resto de los casos (aquellos mal clasificados), suman un total de 657 casos.



A continuación mostramos el accuracy obtenido para cada modelo:

Modelo	Accuracy in %
LR	77,152
SVM	78,976
NB	77,376
KNN	78,048

- ***4.2. Regresión***

En cuanto a la regresión, se utilizaron diferentes modelos para poder predecir, en base a diferentes features, la presión sistólica de una persona.

Las features utilizadas son las siguientes:

- Edad
- Masa corporal
- Peso
- Altura
- Sexo (Hombre/mujer) con dummies aplicados.
- Cantidad de días a la semana que realiza actividad física
- Si la persona fue diagnosticada alguna vez con diabetes o colesterol. (Utilizando dummies)
- Si la persona ha fumado alguna vez en su vida cigarrillo.

Se utilizaron los siguientes modelos de regresión de machine learning supervisado:

KNN

En el modelo de KNN Regression, se determinan los K vecinos que se encuentran mas cercanos por distancia euclídea (par a par) en el entrenamiento.

El valor a predecir, es determinado por medio de una interpolación de valores “Y” en los K vecinos más cercanos.

Luego, los pesos W indican cómo se interpolará cada K vecino: uniforme o por distancia.

SVR (Support Vector Regression)

Este modelo busca construir la función lineal (hiperplano) que mejor se ajuste a los datos.

Aquí, se determina un margen (épsilon) como función de costo, y trata de que las muestras caigan dentro del mismo. De esta forma, el modelo busca

maximizar dicho margen con el objetivo de que estén contenidas la mayor cantidad de samples.

Se define además una función de costo que penaliza las muestras que se encuentren fuera de los límites establecidos:

Medidas de calidad del modelo

Para determinar cuán bueno es cada modelo utilizado, se tienen en cuenta las siguientes medidas

R^2: Indica la proporción de la varianza de "y" que explica el modelo.

- *RSS*: Sumatoria de los residuos al cuadrado
- *TSS*: Indica la varianza total de las etiquetas "y".

MAE: Media del error
MSE: Error cuadrático medio

Resultados

Por medio de las features explicadas anteriormente, se buscó estimar el valor de la presión de una persona.

	Model	R2	MSE	MAE
0	SVR	0.241406	333.135454	13.290861
1	KNN	0.238448	334.434778	13.561749

1. Conclusiones

Una vez concluidos ambos análisis y obtenidos los resultados correspondientes, llegamos a diferentes conclusiones.
Para el caso de clasificacion de personas que son de riesgo, encontramos que el modelo de clasificacion que mejor performance presenta es Support Vector Machines, prediciendo con un 79% de precisión a las personas según sea de riesgo o no, conociendo unicamente la edad, el peso, el sexo, la cantidad de días que realiza actividad física a la semana y si alguna vez ha fumado o no cigarrillo.

Creemos que en caso de lograr mejorar el accuracy (por lo menos un 90%), utilizando este modelo de clasificación encontrado, se podría tener un buen acercamiento hacia el conocimiento de la salud de una persona y poder detectar de una manera rápida si es susceptible de tener algún tipo de enfermedad no transmisible dada la vida que lleva.

Con el objetivo de mejorar aún más el accuracy, se planteó la posibilidad de buscar los mejores hiperparámetros utilizando gridsearch, ya que no se pudo realizar debido a limitaciones computacionales. También creemos que sería útil incorporar más enfermedades no transmisibles detectadas en la encuesta nacional de factores de riesgo (como por ejemplo, distintos tipos de cáncer).

Para el caso de la regresión, encontramos que conociendo la edad, masa corporal, peso y altura de una persona; como así también los días que realiza actividad física a la semana, si ha fumado cigarrillo o si ha sido diagnosticado con alguna enfermedad no transmisible (diabetes o colesterol), se puede llegar a estimar la presión sistólica de la misma.

De los modelos de regresión utilizados, el que mejor se ajusta a nuestro objetivo fue el SVR ya que presenta un menor error (MSE y MAE). Esto se ve reflejado en un mayor valor de R^2 obtenido (0.241).

1. Referencias

1 Encuesta nacional de factores de riesgo (https://www.indec.gob.ar/ftp/cuadros/menusuperior/enfr/manual_base_usuario_enfr2018.pdf)

2. Apuntes de Cluster AI.

3. Factores de riesgo para las enfermedades no transmisibles.

(https://scielo.conicyt.cl/scielo.php?pid=s0034-98871999000800017&script=sci_arttext)

4. Scikit-learn. Machine Learning in Python (https://scikit-learn.org/stable/modules/naive_bayes.html)