Clasificacion de personas de riesgo en Argentina y predicción de su presión sistólica

Benedetto, Matias – Fatur, Tomas– Hara, Tobias

UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Objetivos

- En el siguiente trabajo utilizamos distintos modelos de clasificación y regresion aplicados a la Encuesta Nacional de Factores de Riesgo (ENFR) con el objetivo de clasificar a personas de riesgo y predecir su presion sistolica.
- El mismo fue realizado en el mes de octubre del 2020 sobre los datos de Septiembre y Diciembre de 2018. La Encuesta Nacional de Factores de Riesgo es realizada por el INDEC y Proporciona información válida, confiable y oportuna sobre factores de riesgo (como consumo de tabaco, alcohol, alimentación, actividad física, entre otros), procesos de atención en el sistema de salud y principales ENT en la población argentina (hipertensión, diabetes, obesidad y otras)

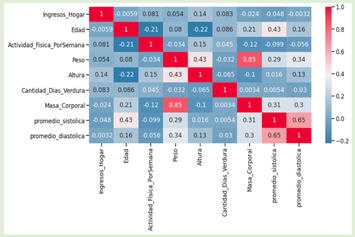

Figure 1. Correlacion lineal entre las variables

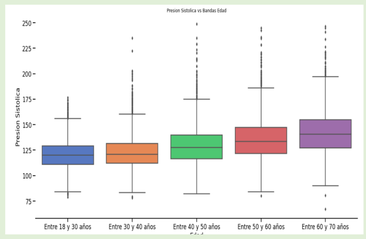

Figure 2. Presion sistolica según la Edad

Dataset and Pre-Processing

- 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.
- Del dataset mencionado, se optó por conservar las siguientes features:
  - Provincia
  - Tipo de vivienda
  - Tipo de hogar
  - Ingresos en el hogar
  - AUH
  - Sexo
  - Edad
  - Situacion conyugal
  - Nivel de instrucción
  - Condicion de actividad
  - Salud general
  - Cobertura de salud
  - Actividad fisica por semana
  - Barreras en la actividad fisica
  - Si fuma o no cigarrillo
  - Cantidad
  - Cantidad de veces que fue diagnosticado con presion alta
  - Peso
  - Altura
  - Masa corporal
  - Cantidad de dias a la semana que come verduras
  - Tipo de alimentacion
  - Colesterol
  - Si bebio alcohol
  - Si ha sido diagnosticado con diabetes
  - Presion medida de la presion sistolica
  - Presion medida de la presion diastolica

Luego de escoger las features y realizar la limpieza correspondiente, terminamos trabajando con un dataset de 15.912 samples y 27 features.

Para poder encontrar un modelo de clasificador adecuado, se creo la label ”Riesgo” según si la persona es de riesgo o no. Una persona es de riesgo si tiene o ha tenido diabetes o colesterol; o bien si su indice de masa corporal es mayor a 30 (obesidad) o si su presion sistolica es mayor o igual a 140 mmHG

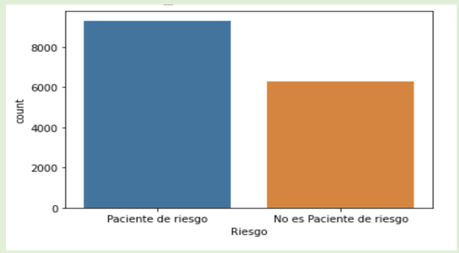

Figure 3. Cantidad de personas encuestadas que son pacientes de riesgo

Materiales y metodos

Fueron utilizados dos tipos de aprendizaje supervisado:

1. CLASIFICACIÓN

Se buscó predecir si una persona será paciente de riesgo en base a las features:

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

Modelos utilizados para la predicción

- Logistic Regression: clasificador lineal procedido de una función de activación Sigmoid. A cada Sample, le asigna una probabilidad de pertenecer a cada clase y la clasifica.
- SVM: Clasifica encontrando un hiperplano separador que maximice el margen entre las clases.
- Naive Bayes: Clasificador basado en el teorema de Bayes.
- KNN: Clasifica cada nuevo dato en el grupo que corresponda, según tenga K vecinos más cerca de un grupo o del otro

2. REGRESIÓN

Se buscó predecir la presión sistólica de una persona a partir de las siguientes features:

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

KNN: Se determinan los K vecinos que se encuentran mas cercanos por distancia euclídea

SVR: Este modelo busca construir la función lineal (hiperplano) que mejor se ajuste a los datos. Determina un margen como función de costo.

Resultados encontrados en la clasificacion de personas de riesgo

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.

Aquí, 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.

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

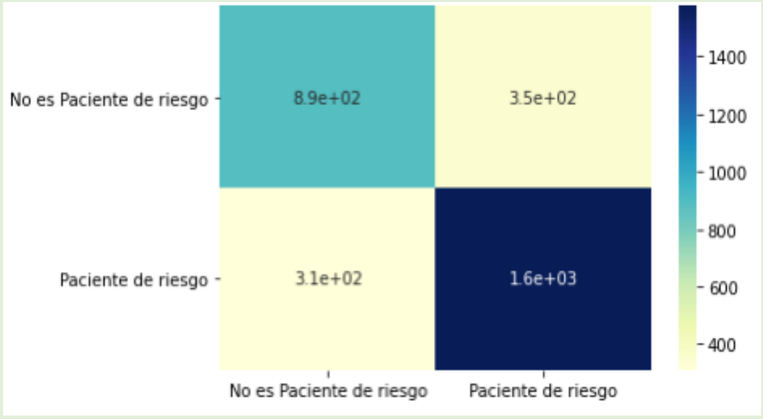

Figure 4. Confusion Matrix for SVM

Resultados encontrados en la regresión

Medidas	KNN	SVR
R^2	0,238448	0.241406
MSE	334.434778	333.135454
MAE	13.561749	13.290861

Puede observarse que el modelo SVR se ajusta mejor al modelo ya que presenta un menor error (MSE y MAE). Esto se ve reflejado en un mayor valor de R^2 obtenido

Resultado y Conclusiones

CLASIFICACION

El modelo SVM es el que presentó una mejor performance prediciendo con un 79% de precisión. Mejorando el accuracy a por lo menos un 90%, creemos que se podría tener un acercamiento hacia conocer si una persona es susceptible a tener algún tipo de enfermedad no transmitible, dada la vida que lleva y sus carácterísticas.

REGRESIÓN

A partir del modelo SVR, y en base a las features previamente descriptas, se puede llegar a estimar la presión sistólica de una persona.

Futuras Mejoras

- Sería útil incorporar más enfermedades no transmisibles detectadas en la encuesta nacional de factores de riesgo (como por ejemplo, distintos tipos de cáncer).
- Buscar los mejores hiperparámetros utilizando gridsearch en clasificacion.