Proyecto 2 - Ciencia de Datos

Clasificador de pacientes con diabetes

Antecedentes:

Dado un set de datos compuesto por informaciones de un determinado grupo de personas evaluadas, se pretende clasificar o diagnosticar en datos futuros cuando una persona puede tener diabetes o no.

Por lo tanto este es un ejercicio de clasificación

Fuente de datos: https://www.kaggle.com/datasets/prosperchuks/health-dataset

Para este ejercicio estaremos utilizando exclusivamente los datos del set de Diabetes

Descripción del set de datos

El set de datos en el que se basa este ejercicio está compuesto por 18 columnas y 70,692 filas.

En una revisión inicial del conjunto de datos se observa que los mismos están balanceados, y estandarizados. Así mismo no se observan mayores retos con datos faltantes o datos duplicados.

Exploración de datos

#	Column	Non-Null Count	Dtype
0	Age	70692 non-null	float64
1	Sex	70692 non-null	float64
2	HighChol	70692 non-null	float64
3	CholCheck	70692 non-null	float64
4	BMI	70692 non-null	float64
5	Smoker	70692 non-null	float64
6	HeartDiseaseorAttack	70692 non-null	float64
7	PhysActivity	70692 non-null	float64
8	Fruits	70692 non-null	float64
9	Veggies	70692 non-null	float64
10	HvyAlcoholConsump	70692 non-null	float64
11	GenHlth	70692 non-null	float64
12	MentHlth	70692 non-null	float64
13	PhysHlth	70692 non-null	float64
14	DiffWalk	70692 non-null	float64
15	Stroke	70692 non-null	float64
16	HighBP	70692 non-null	float64
17	Diabetes	70692 non-null	float64

Observación: En el set de datos en estudio no se observan valores nulos, adicionalmente no se visualizan datos con formato object, por lo tanto, se descartan datos inconsistentes como por ejemplo (Cat, cat)

Diccionario de datos

Age: 13-level age category *_AGEG5YR see codebook)* 1 = 18-24 9 = 60-64 13 = 80 or older
Sex: patient's gender *(1: male; 0: female)*.
HighChol: *(0 = no high cholesterol 1 = high cholesterol)*
CholCheck: *(0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years)*
BMI: Body Mass Index
Smoker: Have you smoked at least 100 cigarettes in your entire life? *([Note: 5 packs = 100 cigarettes] 0 = no 1 = yes)*
HeartDiseaseorAttack: coronary heart disease (CHD) or myocardial infarction (MI)*(0 = no 1 = yes)*
PhysActivity: physical activity in past 30 days - not including job *(0 = no 1 = yes)*
Fruits: Consume Fruit 1 or more times per day *(0 = no 1 = yes)*
Veggies: Consume Vegetables 1 or more times per day *(0 = no 1 = yes)*
HvyAlcoholConsump: (adult men >=14 drinks per week and adult women>=7 drinks per week) *(0 = no 1 = yes)*
GenHlth: Would you say that in general your health is: *(scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor)*
MentHlth: days of poor mental health scale 1-30 days
PhysHlth: physical illness or injury days in past 30 days scale 1-30
DiffWalk: Do you have serious difficulty walking or climbing stairs? *(0 = no 1 = yes)*
Stroke: you ever had a stroke. *(0 = no, 1 = yes)*
HighBP: *(0 = no high, BP 1 = high BP)*
Diabetes: *(0 = no diabetes, 1 = diabetes)*

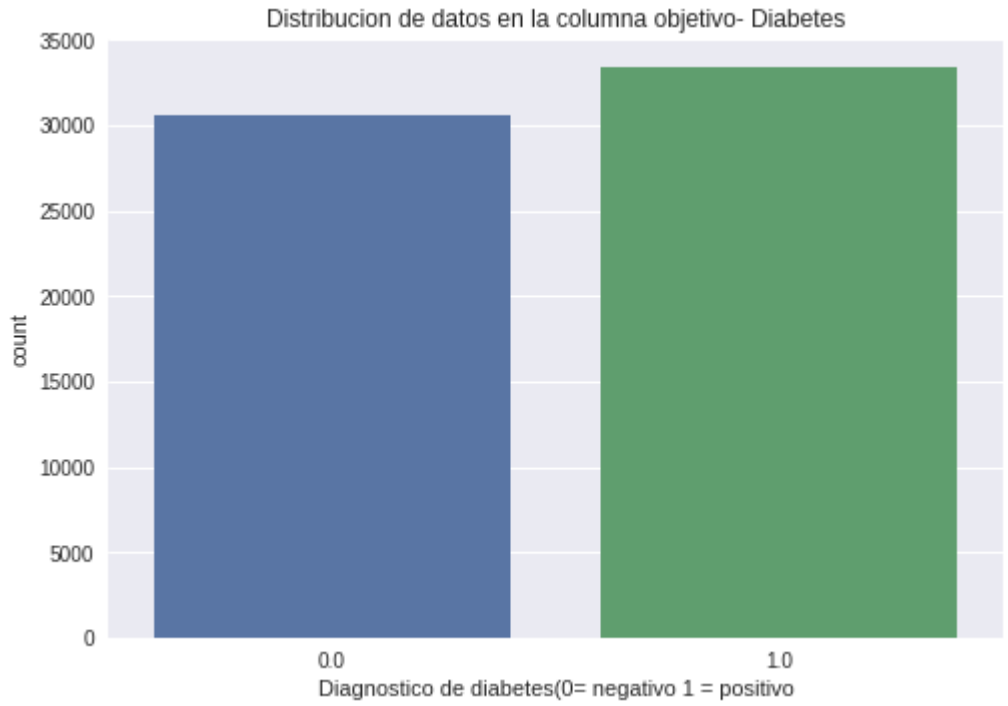
Verificación de datos duplicados

Número de filas duplicadas 6672

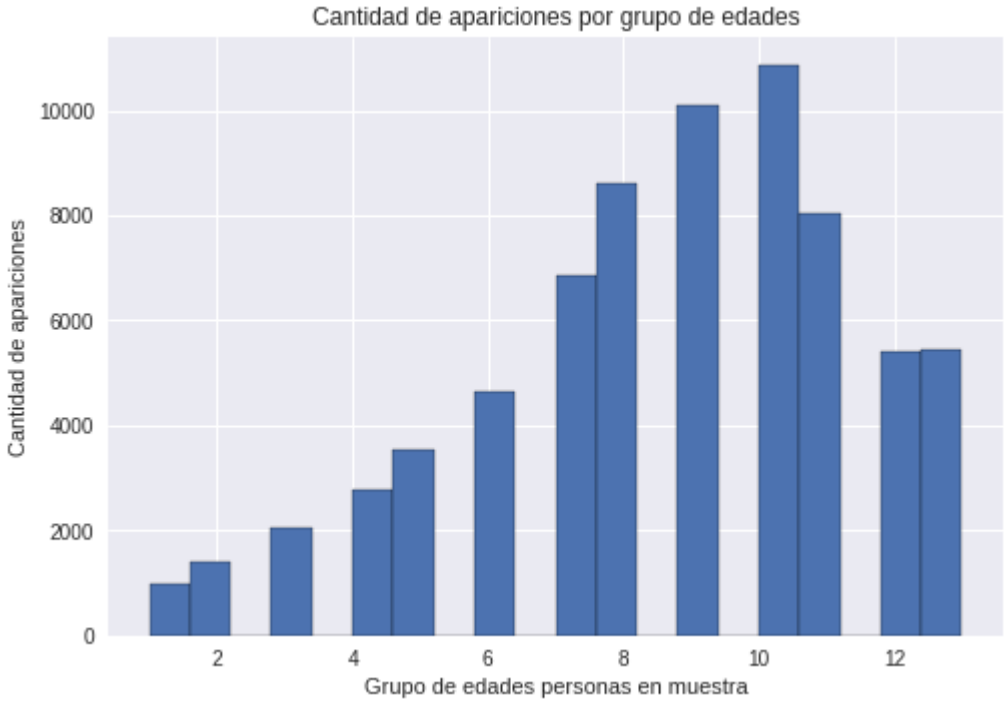
Observaciones: Al verificar el set de datos se observa que los campos que pudieran considerarse como nominales ya están adecuados, además de que son campos de dos valores (0 y 1) o true o false. Por lo que entendemos que no es necesario realizar One- hot encoder.

Así mismo al verificar las estadísticas para cada columna se observa que los datos están adecuadamente normalizados. No se requiere ejecutar un proceso adicional sobre los mismos. Se descarta la existencia de datos atípicos, ya que las columnas observadas se encuentran dentro de los valores esperados y especificados en el diccionario.

Exploración de datos

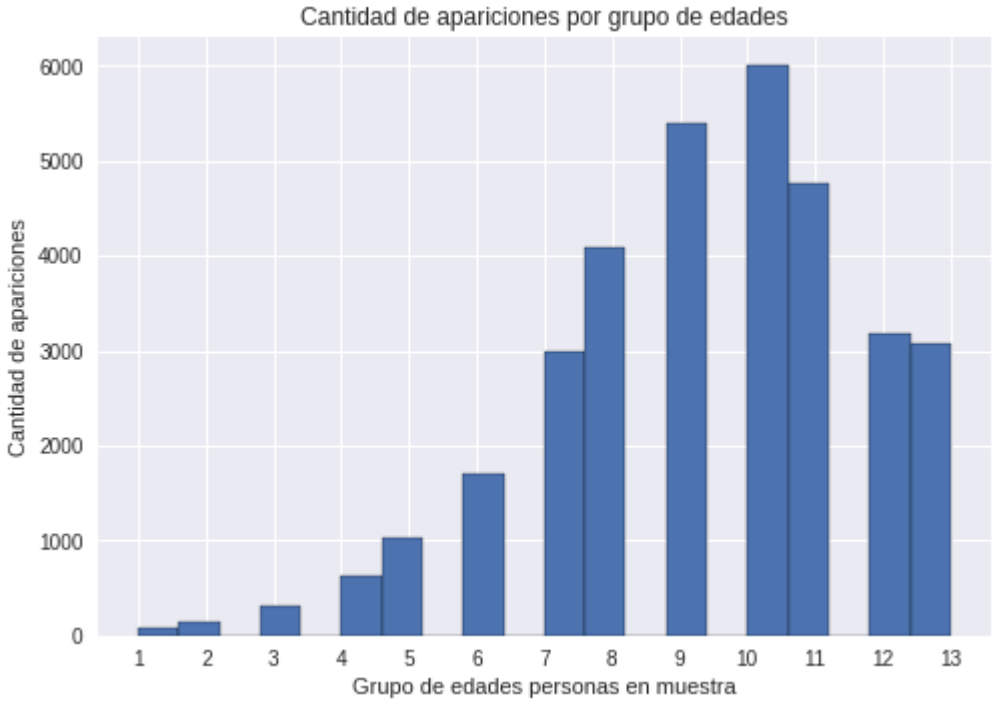


Observaciones: Se observa una buena distribución de datos. aprox. 45%/55%

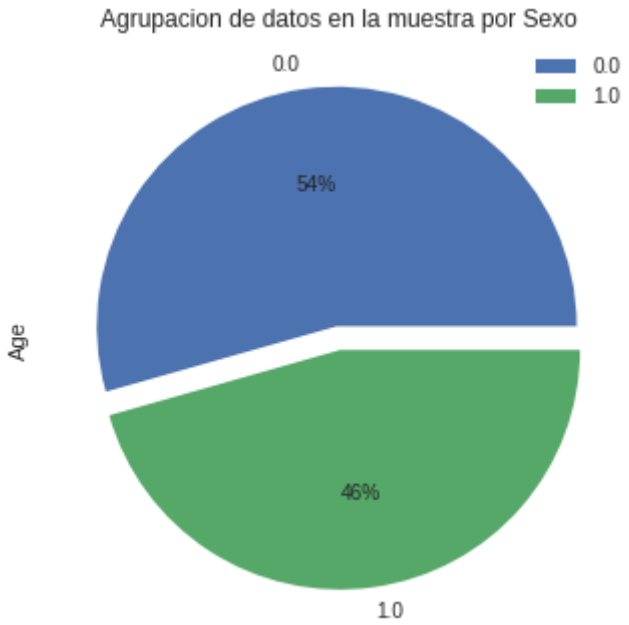


Observaciones: Se verifica que la mayor cantidad de personas en el set de datos esta agrupada en los rangos de edad 9 y 10 equivalentes a las edades 60-70 anos

---



Observaciones: Se verifica que la mayor cantidad de personas en el set de datos está agrupada en los rangos de edad 9 y 10 equivalentes a las edades 60-70 años.

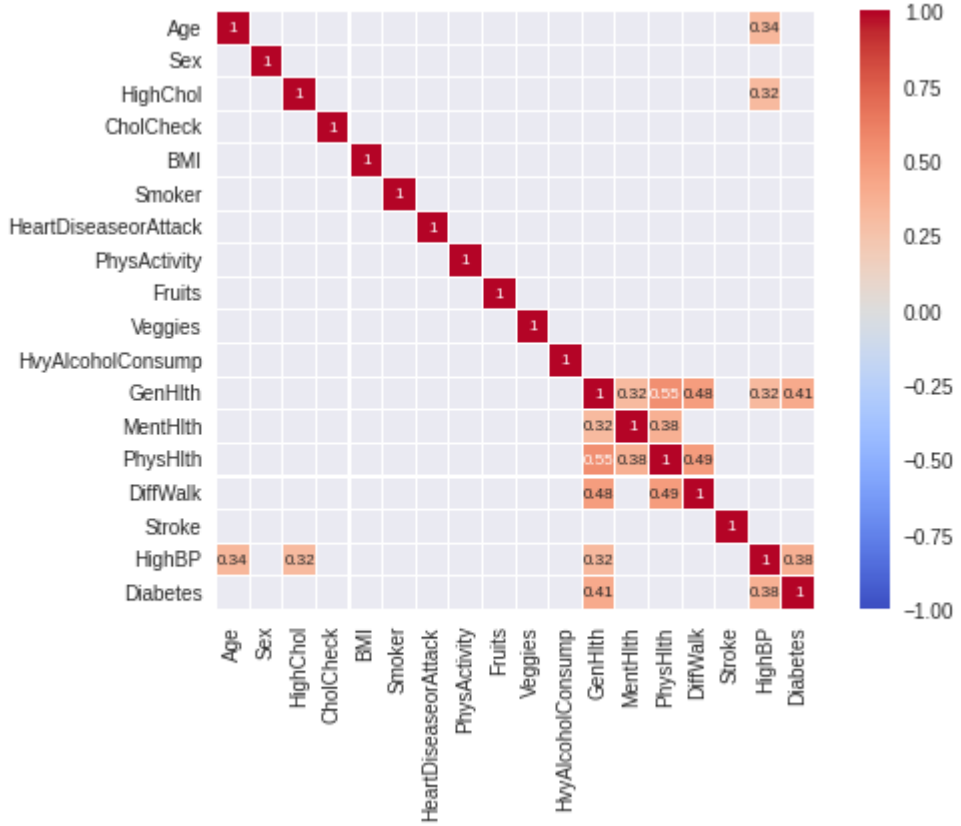


** Observaciones:** El set de datos presenta que el 54% de las personas en referencia pertenecen al sexo femenino, mientras que el 46% corresponde al sexo masculino

---

Observaciones: Se observa que el 48% de las personas senaladas con diabetes corresponden al sexo masculino, mientras que el 52% corresponde a personas del sexo femenino.

Creación de mapa de calor para identificar posibles correlaciones



Observaciones: Los diferentes cuadros de calor indican correlación entre los renglones de salud, como es de esperarse, la salud general está influenciada por la salud mental y salud física. Las personas que presentan dificultad para caminar impactan en la salud física.

También se observa que la alta presión arterial está correlacionada con la edad y el colesterol

A través del mapa de calor no se puede establecer una fuerte correlación entre la diabetes y otros elementos como la alimentación, alcoholismo o tabaquismo. Más allá del valor HighBP(Alta presión arterial) el cual presenta una correlación moderada con la diabetes así como la salud general.

Conclusiones exploración de datos:

De los distintos gráficos destacados más arriba podemos interpretar los siguientes resultados:

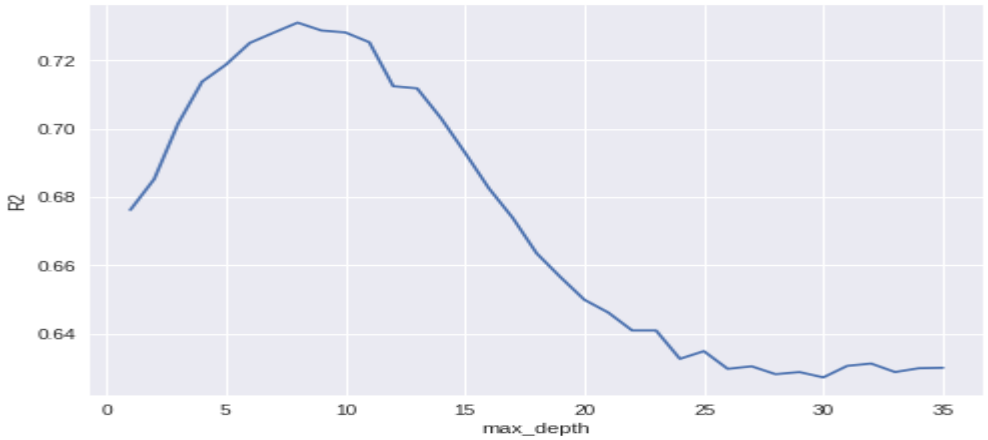
- El set de datos presentado presenta una muestra bien distribuida para un ejercicio de creación de modelos de aprendizaje automático.
- Existe una mayor incidencia de diabetes en las personas entre edades de 60-70 anos.
- Existe una mayor incidencia de diabetes en las personas de sexo femenino.
- Se observa una posible correlación entre la diabetes y la alta presión arterial. Asimismo otros factores de salud están relacionados directamente, como por ejemplo, la relación que existe entre la dificultad de subir escaleras con la salud general. La salud general está relacionada con la salud física y mental.

Modelado de Datos

Determinación de parámetros para configuración de los modelos a implementar

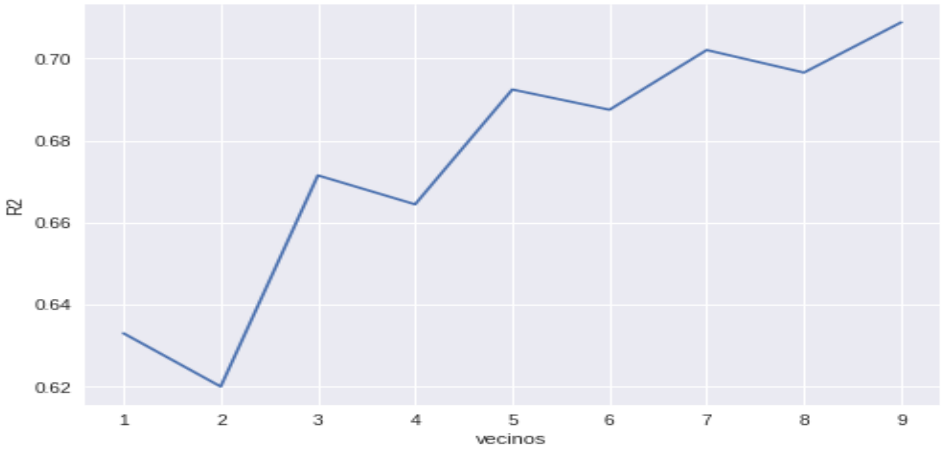
En este bloque estaremos identificando previamente algunos ajustes en los hiperparametros para modelos de árbol de decisión, KNN, bosque aleatorio, con el fin de lograr identificar previamente una configuración adecuada para la configuración en última instancia de cada modelo.

```
#Arbol de decision
```



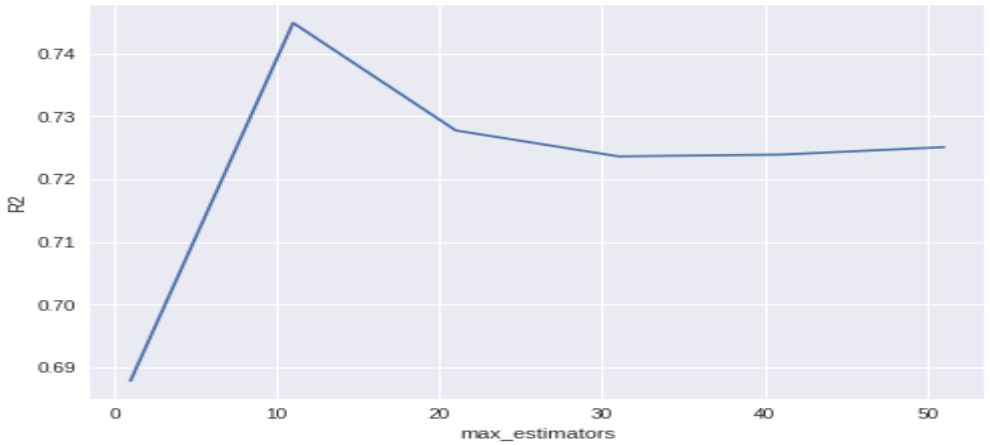
Configuración propuesta: Para el árbol de decisión el máximo rendimiento está en 8 y 9. Por lo que utilizaremos esta configuración en el modelo.

```
#KN Vecinos
```



Configuración propuesta: Se pretende configurar el KNN con un número de 9 u 11 vecinos

```
# Random Forest
```



Configuracion propuesta: Se utilizara estimadores 11 y max_depth de 11

Desarrollo de los modelos

Logisitic Regression

Desempeño del modelo con datos de entrenamiento (Diabetes)

- Accuracy: 0.7478
- F1 score: 0.747727
- Precision: 0.748224
- Recall: 0.747826

Desempeño del modelo con datos de prueba (Diabetes)

- Accuracy: 0.7467
- Fl score: 0.7466
- Precisión: 0.7470
- Recall: 0.7467

==========

Decision Tree

Desempeño del modelo con datos de entrenamiento (Diabetes)

- Accuracy: 0.7515
- F1 score: 0.751043
- Precision: 0.753399
- Recall: 0.751504

Desempeño del modelo con datos de prueba (Diabetes)

- Accuracy: 0.7411
- Fl score: 0.7405
- Precisión: 0.7433
- Recall: 0.7411

==========

Random Forest

Desempeño del modelo con datos de entrenamiento (Diabetes)

- Accuracy: 0.7806
- F1 score: 0.780273
- Precision: 0.782223
- Recall: 0.780588

Desempeño del modelo con datos de prueba (Diabetes)

- Accuracy: 0.7449
- Fl score: 0.7444
- Precisión: 0.7468
- Recall: 0.7449

==========

K-Nearest Neighbors

Desempeño del modelo con datos de entrenamiento (Diabetes)

- Accuracy: 0.7719
- F1 score: 0.771556
- Precision: 0.773632
- Recall: 0.771912

Desempeño del modelo con datos de prueba (Diabetes)

- Accuracy: 0.7322
- Fl score: 0.7317
- Precisión: 0.7342
- Recall: 0.7322

==========

Conclusión-Selección de modelo

Al comparar el rendimiento de los distintos modelos antes configurados se observa un rendimiento similar, no obstante para fines de este ejercicio estaremos seleccionado el algoritmo de **Bosque Aleatorio** ya que el mismo presenta facilidades de configuración que otros modelos no disponen, por lo que en un futuro es posible, que al modificar otras configuraciones podamos alcanzar un rendimiento superior al presentado que es de 0.74