CURSO: ANALÍTICA PREDICTIVA EN R.

PROYECTO FINAL: PREDICCIÓN DE DIABETES EN UN GRUPO DETERMINADO.

ALUMNO: CHRISTIAN FARNAST.

INSTRUCTOR: DAVID ZARRUK.

AGOSTO 2021.

PARTE 1: ANALISIS EXPLORATORIO DE DATOS.

En este primer avance, el objetivo es hacer el análisis exploratorio de los datos. En particular, se debe encontrar:

i. ¿Cuántos datos tenemos para nuestro análisis?

```
28
29  dim(diabetes)
30  # disponemos de 614 filas y 9 columnas
31
32  names(diabetes)
33  # los nombres de las variables son:
34  # "num_embarazos"       "plasma"            "presion_diastolica"
35  # "grosor_piel"         "insulina"          "bmi"
36  # "diabetes_pedigree"   "edad"              "diabetes"
37
```

ii. Encontrar estadísticas descriptivas de todas las variables de la base de datos. Entre ellas, deben estar: media, desviación estándar, mínimo, máximo, percentiles 25, 50 y 75.

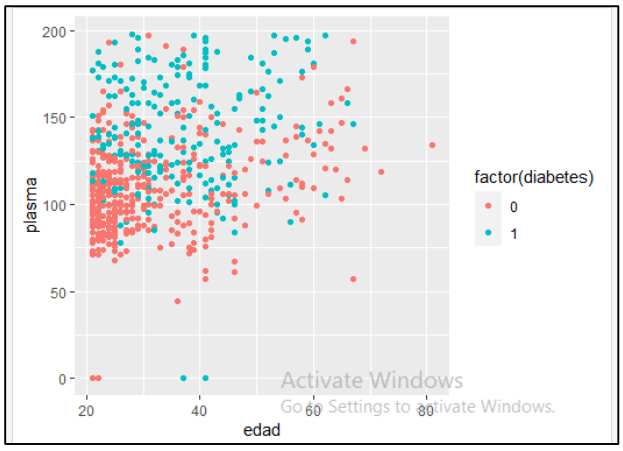
Con la función de funModeling, profiling_num(), se obtienen la media, desviación standard, y los percentiles 25, 50 y 75.

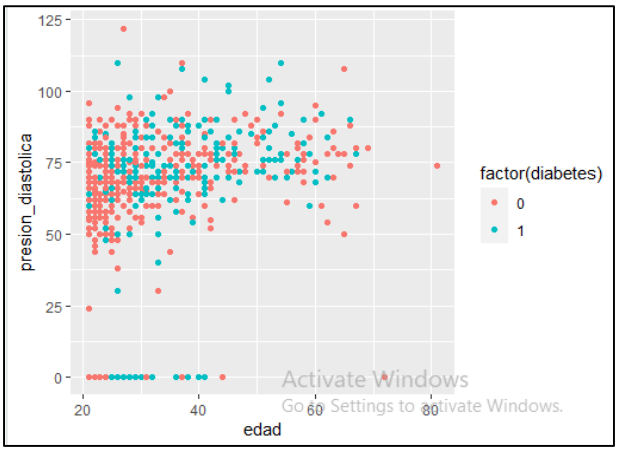
```
46  profiling_num(diabetes)
47  #             variable        mean     std_dev   p_25
48  # 1      num_embarazos   3.8013029   3.3461914  1.000
49  # 2             plasma 120.4478827  32.7038676 99.000
50  # 3 presion_diastolica  68.5228013  19.6080343 64.000
51  # 4        grosor_piel  20.6563518  15.9581677  0.000
52  # 5            insulina  80.2996743 116.6440526  0.000
53  # 6                bmi  31.8903909   7.9752690 27.100
54  # 7  diabetes_pedigree   0.4811792   0.3368657  0.248
55  # 8               edad  33.1970684  11.7728054 24.000
56  # 9           diabetes   0.3469055   0.4763735  0.000
57  #
58  #
59  #                                       p_50    p_75
60  # 1      num_embarazos                 3.0000   6.000
61  # 2             plasma                116.0000 141.000
62  # 3 presion_diastolica                 70.0000  80.000
63  # 4        grosor_piel                 23.0000  32.000
64  # 5            insulina                37.0000 126.000
65  # 6                bmi                 32.0000  36.500
66  # 7  diabetes_pedigree                  0.3865   0.647
67  # 8               edad                 29.0000  40.000
68  # 9           diabetes                  0.0000   1.000
69
```

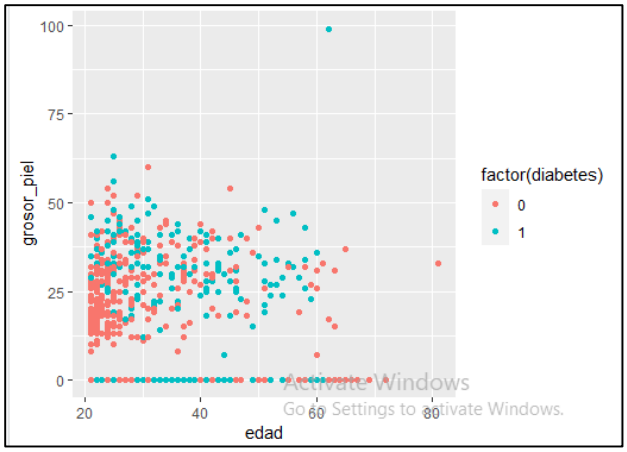
iii. Hacer un análisis exploratorio para ver qué variables pueden ser las mejores al momento de predecir si una persona tiene diabetes. Para esto, debes generar:

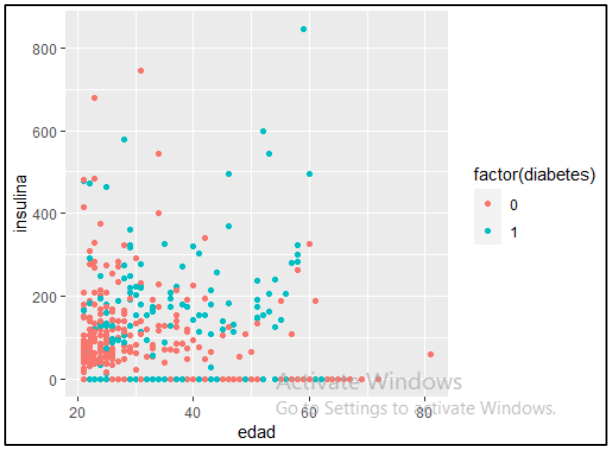
i. Gráficos de dispersión de las variables 2-8 contra la variable edad. ¿Cuáles variables parecieran ser buenas para explicar la edad de un abalón?

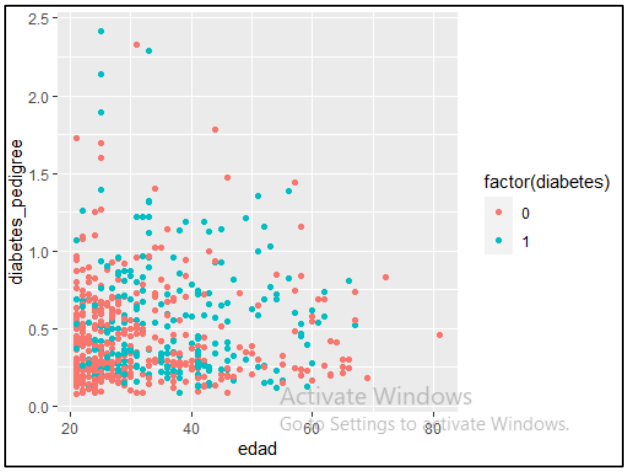
```
ggplot(diabetes, aes(x=edad, y=plasma, color = factor(diabetes))) +
  geom_point()*

ggplot(diabetes, aes(x=edad, y=num_embarazos, color = factor(diabetes))) +
  geom_point()

ggplot(diabetes, aes(x=edad, y=presion_diastolica, color = factor(diabetes))) +
  geom_point()*

ggplot(diabetes, aes(x=edad, y=grosor_piel, color = factor(diabetes))) +
  geom_point()*

ggplot(diabetes, aes(x=edad, y=bmi, color = factor(diabetes))) +
  geom_point()*

ggplot(diabetes, aes(x=edad, y=insulina, color = factor(diabetes))) +
  geom_point()*

ggplot(diabetes, aes(x=edad, y=diabetes_pedigree, color = factor(diabetes))) +
  geom_point()*

```

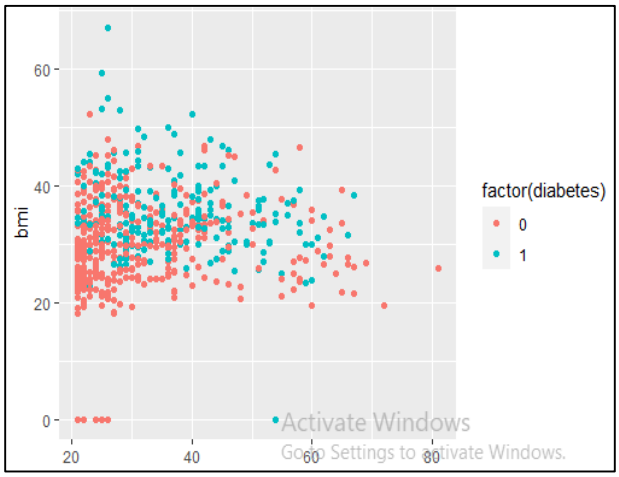












Los gráficos mostrados en asterisco, muestran una dispersión que puede mostrar una correlación entre las variables.

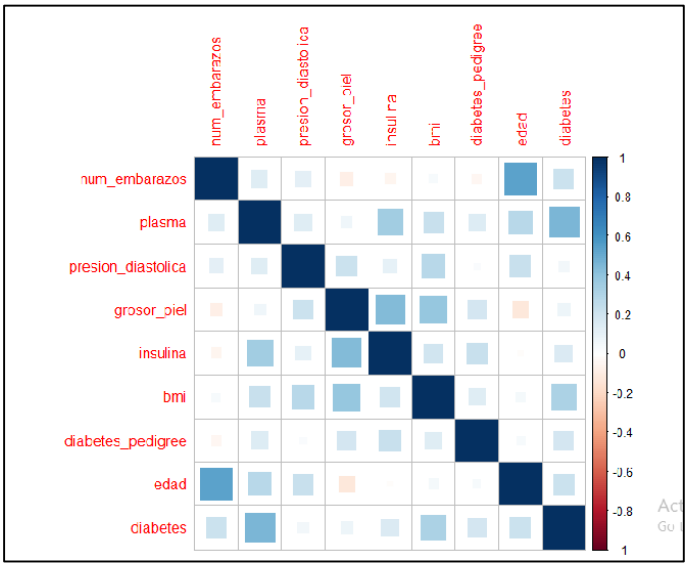
ii. Gráfico de calor para ver las correlaciones existentes entre variables. ¿Lo que se observa en el gráfico de calor concuerda con lo encontrado en los gráficos de dispersión?, ¿Cuáles variables tienen correlaciones positivas o negativas más fuertes con respecto a la variable edad?

Con la función corrplot de corrr, se programa:

→ corrplot(cor(diabetes), method = 'square')

Obteniendo el siguiente mapa:

Mapa gráfico n°1: relación de todas las variables en estudio



Las variables según el mapa de correlación con la edad son:

1. plasma
2. presion_diastolica
3. grosol_piel
4. num_embarazos
5. diabetes

Por lo tanto, lo que muestra este mapa se relaciona con lo mostrado en los gráficos de dispersión.

iii. Gráficos de cajas y bigotes para todas las variables categóricas o discretas. ¿Cuáles variables son categóricas?, ¿Qué categorías tienen mayor edad en promedio?

El dataset diabetes_train no tiene variables categóricas *per se*, por lo que hubo que hacer transformaciones a factor para responder estas preguntas.

Al ser las variables num_embarazos, plasma, presión_diastolica y diabetes, numéricas y con muchos datos salvo diabetes (tiene valores binarios 0 y 1, solamente), se convirtieron en grupos más pequeños para poder hacer mejores observaciones. Para esto último, se ocupó la función equal_freq(), de funModeling.

```
156 diabetes$num_embarazos1 <- equal_freq(diabetes$edad, n_bins= 3)
157 diabetes$plasma1 <- equal_freq(diabetes$plasma, n_bins= 8)
158 diabetes$presion_diastolica1 <- equal_freq(diabetes$presion_diastolica, n_bins= 3)
159 diabetes$edad1 <- equal_freq(diabetes$edad, n_bins= 5)
```

Ya con grupos más pequeños y convertidos en factor por la función equal_freq, se pueden lograr gráficos de caja y bigote.

Se consideraron aquellas variables que tienen correlación con la edad:

1. num_embarazos (no tomada en cuenta para efectos de diabetes. Sin correlación con esta última)

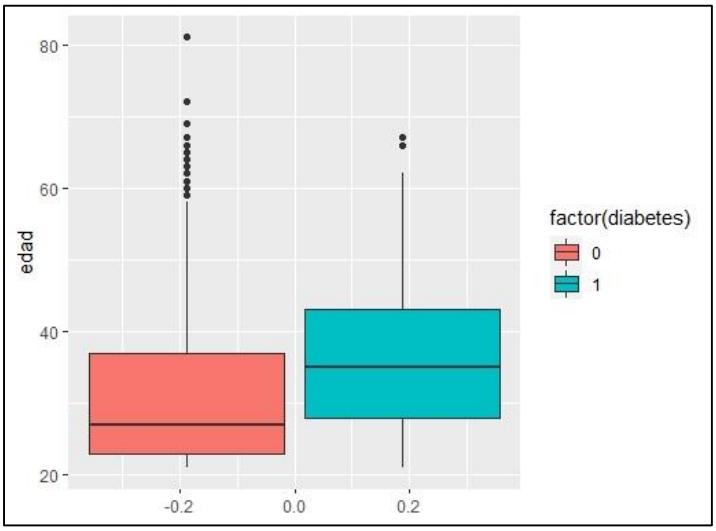
2. plasma

3. presion_diastólica

4. diabetes

1. Relación edad-diabetes

Grafico n°1: relación edad-diabetes

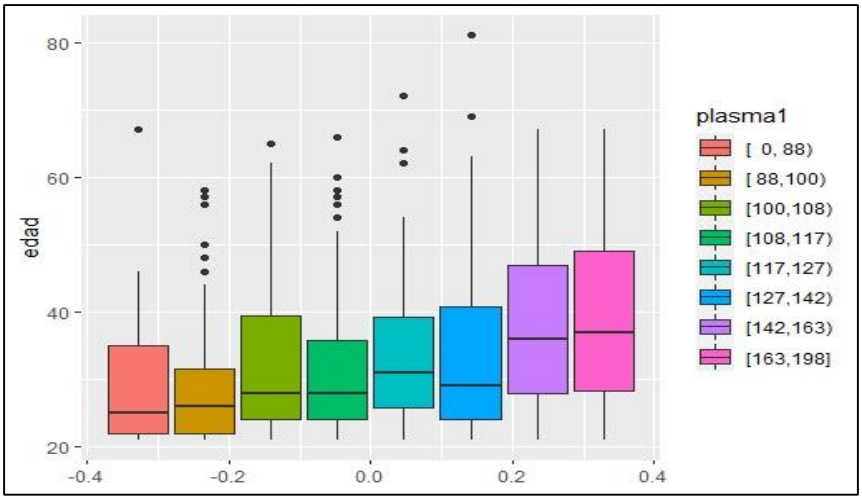



```
149  G1 <- ggplot(diabetes, aes(x=edad, fill = factor(diabetes))) +
150    geom_boxplot() +
151    coord_flip()
```

El gráfico muestra que en promedio la enfermedad se presenta alrededor de los 35 años promedio.

2. Relación edad-plasma

Grafico n°2: relación edad-plasma

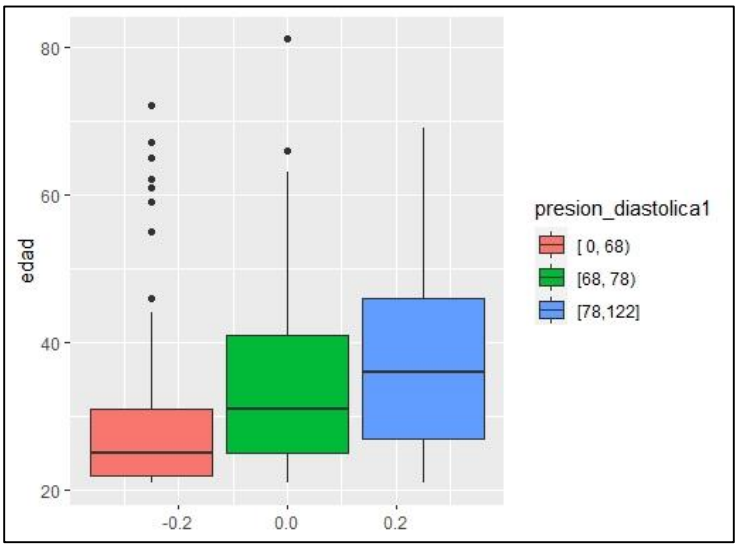


```
G2 <- ggplot(diabetes, aes(x=edad, fill = plasma1)) +
  geom_boxplot() +
  coord_flip()
```

El gráfico n°2 muestra que las cajas 3 y 4 (contándolas de izquierda a derecha), desarrollan enfermedad metabólica previa a diabetes (índices sobre 100 mg/dL de azúcar en plasma hasta 120-126 mg/dL de azúcar en plasma) bajo la línea de los 30 años en promedio versus los grupos 5 y 6 (de izquierda a derecha) que desarrollaron diabetes a los 30 años en promedio. Los grupos 7 y 8 desarrollaron la enfermedad sobre la línea de los 35 años en promedio.

3. Relación edad-presión diastólica

Grafico n°4: relación edad– presión diastólica

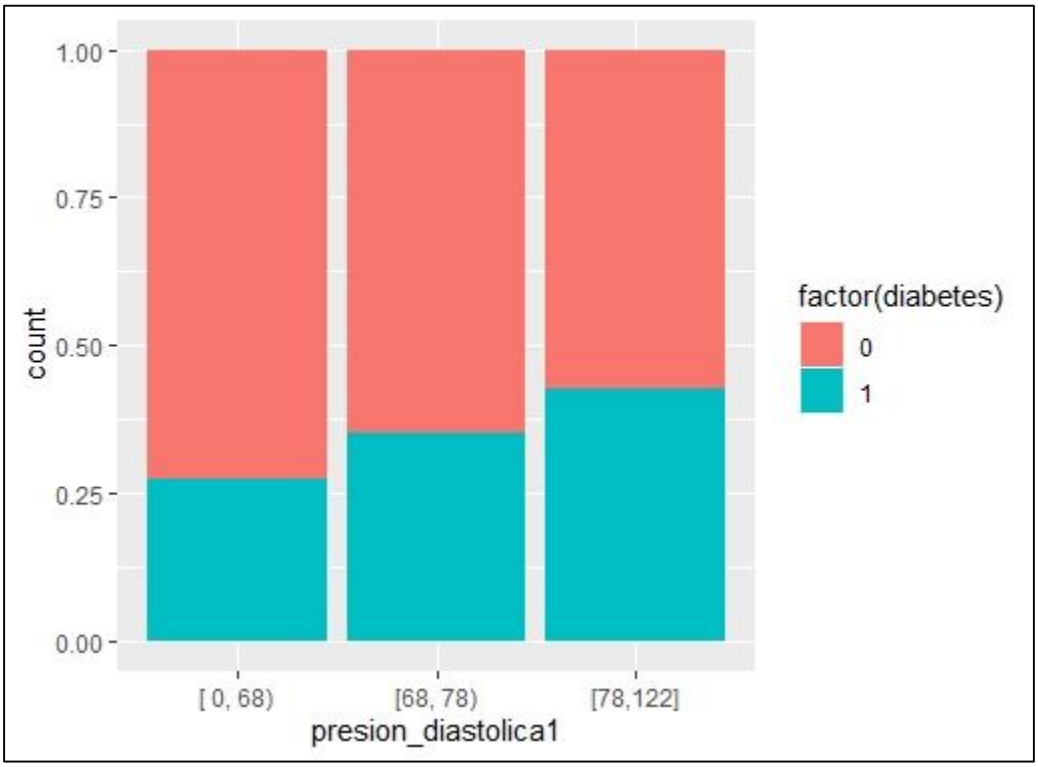



```
172  G3 <- ggplot(diabetes, aes(x=edad, fill = presion_diastolica1)) +
173    geom_boxplot() +
174    coord_flip()
175
```

En este grafico se muestra que conforme aumenta la edad, la presión diastólica del corazón aumenta. Esto puede deberse por el hecho de la perdida de funcionalidad del corazón por envejecimiento natural. Sin embargo seria de utilidad, ver correlación entre diabetes y presión diastólica, como se muestra a continuación.

4. Relación diabetes-presión diastólica

Grafico n°4: relación diabetes-presión diastólica




```
G4 <- ggplot(diabetes, aes(presion_diastolica1, fill = factor(diabetes))) +
  geom_bar(position = 'fill')
```

El grafico 4 muestra que a mayor presión diastólica, los casos de diabetes aumentan. En el caso de la presión entre 0 a 67 años, pueden ser casos *outliers*, por lo que requiere una mejor revisión. En cuanto a la presión diastólica entre los 68 a los 77 años, se debe chequear por ejemplo que tipo de diabetes es: tipo 1 o insulinorequiriente o tipo 2 o no insulinorequiriente.

PARTE 2: DETERMINACION DE MATRIZ DE CONFUSION, ACCURACY, RECALL , PRECISIÓN Y F1-SCORE

Para esta segunda entrega, se debe analizar la base de datos "diabetes_test.csv". Esta base de datos contiene el conjunto de validación, sobre el que se medirá qué tan buen desempeño tiene un modelo particular.
Esta base de datos contiene exactamente las mismas variables que la base de datos de la primera entrega, además de dos nuevas columnas llamadas modelo_2 y modelo_3. Estas variables son la predicción de que un usuario tenga diabetes utilizando dos modelos entrenados por otra persona.
En este segundo avance, el objetivo es que midas el desempeño de cada modelo para predecir la variable diabetes y escoger cuál modelo es mejor entre los dos propuestos. Se desea saber:

1. Computar la matriz de confusión para cada modelo (haz una matriz separada separado para el modelo_2 y para el modelo_3). ¿Qué modelo tiene mayor tasa de falsos negativos?, ¿y de falsos positivos?
2. Calcular las métricas relevantes para un modelo de clasificación:
   a. Accuracy:
   b. Recall
   c. Precisión
   d. F1-score
3. ¿Qué métrica se cree que es la más relevante para escoger el "mejor" modelo? Justifica tu respuesta.
4. De acuerdo a tu respuesta en el punto 3, ¿qué modelo escogerías?.

5. ¿Qué tan bueno es el mejor modelo? Es decir, describe qué significa que el modelo que escogiste tenga ese nivel de precisión y de recall.

Respuestas:

Modelo 2:

```
214
215  #Resultados modelo 2:
216  ######################
217
218  #matriz de confusión
219
220  #           Reference
221  # Prediction  0  1
222  #          0 87 12
223  #          1  6 49
224  #Sensitivity : 0.9355  <- tasa falsos positivos: 1-0.9355 = 0.0645 (6.5%)
225  #Specificity : 0.8033 <- tasa falsos negativos: 1-0.8033 = 0.1967 (20%)
226  #recall: 0.80
227  #precision: 0.89
228  #f1 score:0.85
229
```

Modelo 3:

```
249  #Resultados modelo 3:
250  ######################
251
252  #matriz de confusión
253
254  #           Reference
255  # Prediction  0  1
256  #          0 77 22
257  #          1 13 42
258  #Sensitivity : 0.8556 <- tasa falsos positivos: 1-0.8556 = 0.1444 (14%)
259  #Specificity : 0.6562 <- tasa falsos negativos: 1-0.6562 = 0.3438 (34%)
260  #recall: 0.65
261  #precision: 0.76
262  #f1 score:0.70
```

Tabla resumen n°1: Tasa de falsos positivos y negativos, precisión, recall y f1-score por modelo

Tasa / Modelo	Tasa falsos positivos	Tasa falsos negativos	Recall	Precision	F1-score	Accuracy (exactitud)
Modelo 2	6.5%	20%	0.80	0.89	0.85	88%
Modelo 3	14%	34%	0.65	0.76	0.70	77%

De acuerdo a estos valores, el modelo 2 es más específico y sensitivo pues ofrece menor cantidad de tasa en falsos positivos y negativos, es decir, este modelo detecta menos cantidad de pacientes no tienen diabetes cuando el modelo arrojó que si tenía y menos cantidad de pacientes (falsos positivos) que el modelo predijo que no tenían la enfermedad cuando en realidad si la tienen (falsos negativos). En cuanto a los parámetros Recall (cantidad que el modelo de machine learning es capaz de identificar), Precisión (capacidad de repetir un valor una *n* cantidad de veces*)*, Accuracy (medida que determina que tan cerca del valor verdadero está un resultado) y F1-score ("promedio" de la precisión y el recall), el modelo 2 tiene mejores índices que el modelo número 3. F1-score, al ser un promedio armónico entre recall y precisión y según lo expuesto por el profesor Zarruk en clases, este es un parámetro a considerar para elegir qué modelo es mejor que otro. Con los valores de precisión y recall del segundo modelo, se logran un mejor resultado con respecto al tercero pues los datos tienen mejor balance. Los datos test son un 20% de total, por lo que para el modelo 2 fue oportuno este Split para tener mejores resultados que el tercer modelo (habría que estudiar si un Split de 70/30 u 75/25 ofrece mejores scores que el segundo modelo). En consecuencia y tomando en consideración lo expuesto anteriormente, el modelo 2 es mejor que el número 3.

TERCERA PARTE: PROYECTO FINAL

El objetivo es construir modelos para predecir si una persona padece diabetes o no y concluir cuál es el mejor modelo.

Los puntos a entregar son:

1. De acuerdo con el análisis exploratorio de la primera entrega, ¿cuáles variables parecieran ser las que más información tienen para predecir si una persona tiene diabetes?

Según el análisis exploratorio, las variables que mayor información pueden entregar para predecir diabetes son:

→ Código: corrplor(cor(df_train))

- ✓ num_embarazos.
- ✓ plasma
- ✓ insulina
- ✓ bmi
- ✓ diabetes_pedigree
- ✓ edad

2. Toma los datos del conjunto de entrenamiento diabetes_train.csv. Entrena 3 distintos modelos de predicción con estos datos. Recuerda que las variaciones a tus modelos pueden ser:
   2.1. Utiliza distintos modelos: modelo de probabilidad lineal univariado, modelo de probabilidad lineal multivariado, regresión logística, Random Forests con distintos parámetros.

2.2. Utiliza distintas variables en cada modelo. Por ejemplo, en un modelo puedes utilizar num_embarazos, plasma, presion_diastolica y grosor_piel. En otro modelo puedes utilizar todas las variables.

2.3. Puedes construir nuevas variables.

3. Tomar los datos del conjunto de validación diabetes_test.csv. Para cada uno de los modelos del punto 2., hacer la matriz de confusión.
4. Calcular las métricas de accuracy, precisión, recall y F1-score y agrégalas en una tabla que permita comparar el desempeño de los tres modelos. ¿Qué modelo es mejor para predecir si una persona tiene diabetes? ¿Qué tan bueno es el mejor modelo?

Para responder las preguntas 2 a la 4, se hicieron 3 modelos de predicción en una primera instancia considerando todas las variables versus diabetes y en una segunda, variables que tienen una mayor correlación con la variable a predecir. Los modelos son:

- Regresion lineal multivariada
- Regresión logística
- Random forest

a) Modelos de predicción diabetes v/s todas las variables.

	Accuracy	Recall	Precision	F1-score
Lineal multivariante	0.76	0.74	0.52	0.61
Logistica	0.77	0.75	0.55	0.63
Random Forest	0.77	0.73	0.55	0.63

b) Modelos de predicción diabetes v/s plasma + bmi + insulina + edad + num_embarazos + diabetes_pedigree

	Accuracy	Recall	Precision	F1-score
Lineal multivariante	0.77	0.76	0.53	0.62
Logistica	0.77	0.73	0.55	0.63
Random Forest	0.72	0.61	0.55	0.58

<u>Modelos de diabetes versus todas las variables:</u>

Las métricas calculadas para los 3 modelos, son bastante parejos y casi no hay diferencias entre estos. Sin embargo, la regresión logística saca una pequeña ventaja en recall en comparación con random forest e igualando en f1-score a este modelo.

Por cuanto en esta primera medición, se impone como mejor modelo la regresión logística

Modelos de diabetes versus variables más correlacionadas:

Tal cual como ocurrió anteriormente y, para este caso, las métricas están bastante parejas y no hay diferencias significativas entre ellas. La regresión logística, vuelve a ser el mejor modelo comparando con el modelo lineal multivariado y random forest. Considerando f1-score, la regresión logística vuelve a ser el mejor modelo predictivo.

En resumen, la regresión logística en ambas situaciones es el mejor modelo, infiriendo que tomando en cuenta un Split de 80% de entrenamiento y 20% de prueba, este modelo es mejor predictor que la regresión lineal multivariada y random forest.

Para nuevos cálculos, se hace presente mencionar que estos dos últimos modelos necesiten de splits adecuados para ellos, hablando de un 75/25 o 70/30 para lograr mejores métricas para saber si son mejores, iguales o menores que la regresión logística.

Para acceder al código escrito para este trabajo acceder a: www.github.com/cfarnast.

Muchas gracias.