INFORME CLASIFICACIÓN SPEED DATING

1.- Preprocesamiento de los datos

El dataset original del ejercicio de clasificación consta de 8378 filas y 62 columnas o variables.
En principio, no debería de haber outliers ya que todas las variables se encuentran dentro de un rango definido para cada una de estas por lo que no voy a eliminar valores atípicos.
Por otro lado, para convertir todas las variables en numéricas convierto la variable género en binaria.
En cuanto a limpiar el dataset de valores faltantes, primero elimino las variables expected_num_interested_in_me y expected_num_matches porque tienen un alto número de valores faltantes y estos valores no influyen en principio si una cita hará match o no. Por último elimino las filas con más de dos valores faltantes y compruebo el porcentaje de valores faltantes por instancia. Como el porcentaje más alto no llega a 6, elimino también todas las filas que quedan con NA's.

Exploro los datos y me doy cuenta de que el dataset está muy desbalanceado y esto puede generar problemas ya que un modelo que genere la predicción de que ninguna cita hará match ya tiene una precisión de aproximadamente 80%. Modifico los datos para que el ratio match – no match esté balanceado y divido el conjunto de datos en train y test para entrenar el modelo con un conjunto de datos y validarlo con el otro para evitar problemas de sobreajuste.
Con todo esto el dataset final se me queda en 1986 filas y 60 columnas, el cual escalamos con el Z score (restando la media y dividiendo por la desviación típica)

2.- Técnicas de reducción de la dimensionalidad.

Una buena forma de simplificar el modelo es quedarnos solamente con las variables más importantes o que más decidan el veredicto final de match o no match. Para hacer esto uso varias técnicas como stepwise, regresión de mejores subsets, selección de variables por regresión lasso o ridge y selección de variables por correlación.

El primer modelo es un modelo lineal con todas las variables, con el segundo modelo lo que hago es optimizar el AIC del modelo 1 mediante la técnica stepwise en ambas direcciones y con el segundo modelo hago lo mismo pero con el BIC el cual penaliza aún más la complejidad del modelo (es por esto que el modelo 3 tiene aún menos variables que el modelo 2)
Luego hago uso de la regresión de mejores subconjuntos pero debido a la potencia de mi ordenador sólo pude poner como número máximo de variables 10, sería interesante ver cuantas y qué características devolvería esta técnica si pudiese hacer esta regresión con las 59 variables (60 variables que tengo en mi dataset final menos la variable objetivo match).
Me devuelve que el conjunto de variables más óptimo es 10 entre ellas funny_o, attractive_partner, shared_interests_o y, curiosamente, art.

También hago uso de la selección de variables por LASSO y me quedo con las 15 variables más relevantes para el estudio de 'match' y por RIDGE el cual me da valores similares.

En cuanto al modelo de selección de variables por las más correladas, hago un filtro de las variables cuya correlación en valor absoluto con match sea de 0.075 o más y me quedo con 17 variables.

```
##        nmodelos numero_vars    AIC_s    BIC_s        R2
## 1       modelo1          59 3305.904 3624.258 0.3690530
## 2       modelo2          31 3267.042 3439.266 0.3745020
## 3       modelo3          10 3295.047 3357.674 0.3517979
## 4 modelo_fuerza          10 3293.476 3356.103 0.3525434
## 5  modelo_lasso          15 3299.092 3387.813 0.3522277
## 6   modelo_corr          17 3305.045 3404.205 0.3503368
```

En la imagen anterior muestro una comparación de todos estos modelos con diferentes métricas.

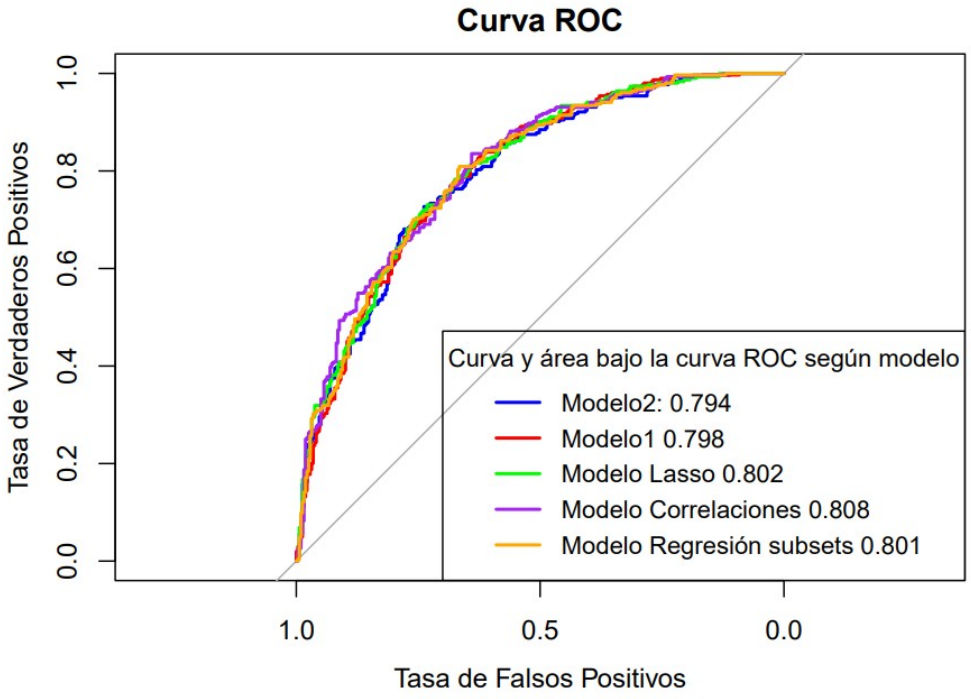
3.- Modelos de regresión logística y cálculo de predicciones

Procedo a entrenar modelos de regresión logística usando los diferentes conjuntos de variables que he obtenido con el ejercicio previo. Escalo los datos entre 0 y 1 ya que el modelo de regresión logística así lo pide.
Además calculo algunas métricas como el R^2 de Efron y de Nagelkerke que miden la bondad del ajuste. Otras métricas que utilizo son la precisión y el área bajo la curva ROC.
Hago las predicciones de cada modelo sobre el conjunto de datos de test para evitar problemas de overfitting y comparo los resultados:

nmodelos <chr>	efron <dbl>	nagel <dbl>	aic_log <dbl>	precision <dbl>
modelo2	0.4261223	0.5276441	1268.591	70.20934
modelo1	0.4353259	0.5382869	1306.441	68.43800
modelo lasso	0.3895787	0.4867410	1304.184	71.65862
modelo corr	0.3820614	0.4830714	1314.088	71.65862
modelo fuerza	0.3863797	0.4821303	1301.597	71.81965

Los resultados son bastante similares en todas las métricas, estando el modelo 1 con todas las variables un poco por debajo en precisión.

Comparo también los resultados del área bajo la curva ROC.



Los valores del área bajo la curva son muy similares también estando un poco por encima de los demás el modelo de regresión logística de las variables más correladas.

4.- Modelos Naive Bayes y cálculo de predicciones

Entreno dos modelos Naive Bayes con las dos selecciones de variables más grandes ya que este modelo es útil cuando este número es grande y no funcionan bien los métodos multivariantes.
Vuelvo a hacer las predicciones sobre el subconjunto de validación.

5.- Comparación de modelos y conclusión final.

Comparo los dos mejores modelos de regresión logística con los dos modelos Naive Bayes

```
##          modelo_usado numero_vars metodo_extraccion_vars precision_final
## 1 Regresión Logística          17          Correlaciones        71.65862
## 2 Regresión Logística          15                  Lasso        71.65862
## 3         Naive Bayes          59    Todas las variables        72.62480
## 4         Naive Bayes          31               Stepwise        71.81965
##   auc_final
## 1     0.808
## 2     0.802
## 3     0.799
## 4     0.803
```

Todos los modelos son muy similares entre sí, destacando un poco por encima en precisión el Naive Bayes con todas las variables aunque es el que menos área bajo la curva ROC tiene.

Con todos estos datos encima de la mesa, finalmente el modelo que escogería sería el modelo de regresión logística de 17. Tiene una precisión muy buena como todas las demás y la mejor medida de AUC. Además es un modelo muy intuitivo al ser una regresión logística de las variables más correladas con la variable objetivo match.

Resumen final justificado para los directivos no técnicos

Con este estudio lo que hemos intentado ha sido predecir si dos personas dentro de nuestro programa de citas rápidas van a hacer un match (si ambos quieren seguir adelante con una segunda cita)
Para esto hemos usado los datos de los últimos años y analizado y comparado varios modelos estadísticos.
Los resultados entre modelos son muy parecidos pero el modelo final es un modelo sencillo que hace uso de las variables más relacionadas con la variable final match.
Este modelo se puede usar para intentar juntar a las parejas que previamente ya tendrían más probabilidades de hacer match, con el objetivo de agilizar el proceso de las citas. Esto puede mejorar la experiencia general de los participantes así como su satisfacción y, en última instancia, la reputación de la empresa.
Además, se podría reducir la cantidad de personas que no hacen match y esto podría incrementar la retención de clientes y atraer a otros nuevos.
Por último, si se logra mejorar la tasa de éxito de los eventos, se podría aumentar la cantidad de eventos que la empresa puede realizar y en consecuencia, aumentar su rentabilidad.