

MÉTODOS Y APLICACIONES DE ANALÍTICA I - EXAMEN FINAL

Pontificia Universidad Javeriana

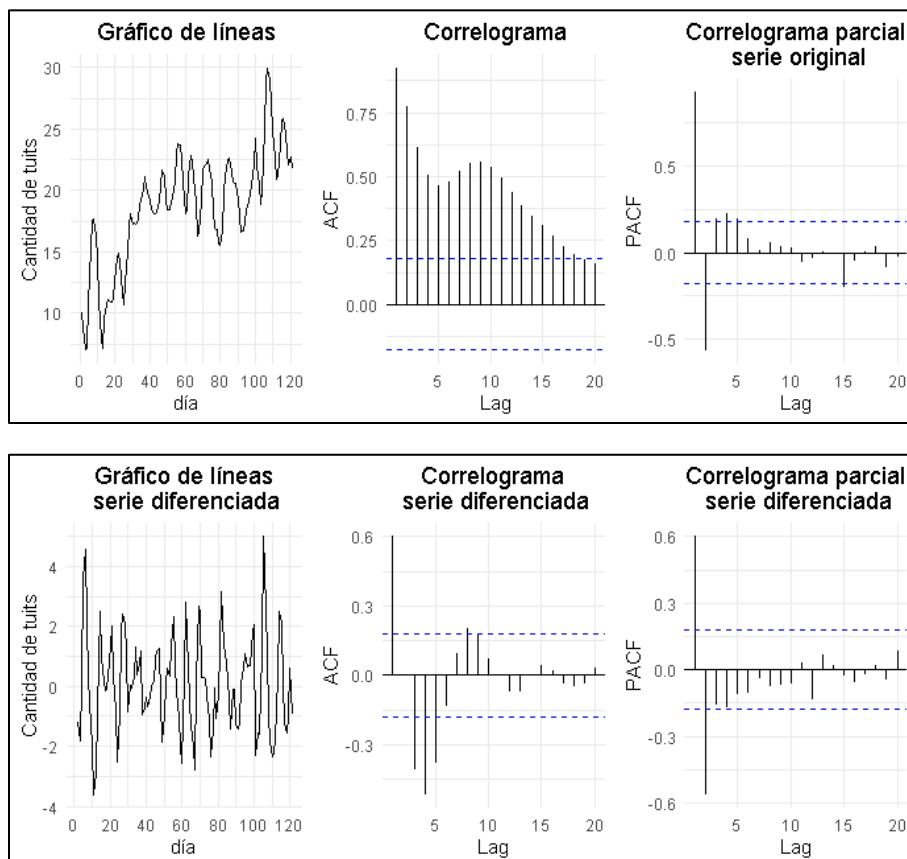
Junio 3 de 2022

Reglas del examen: Se pueden consultar materiales e Internet, pero está completamente prohibida la comunicación entre estudiantes o con personas externas durante el examen, de forma presencial o virtual (incluidos foros). Favor responder a continuación de la pregunta. Una vez diligenciado el cuestionario deben guardarlo en formato PDF y subirlo como respuesta a la asignación en Brightspace. La hora máxima de entrega es a las 5:05 pm.

NOMBRE: Laura Juliana Mora Páez

Caso 1: (20 puntos) Las redes sociales se han ido convirtiendo paulatinamente en una herramienta para la comunicación política alrededor del mundo. Colombia no es ajena a ello y, desde las elecciones a la presidencia de 2010, estos nuevos medios de comunicación han tomado protagonismo en las contiendas electorales tanto a nivel nacional como regional y local. En campaña, las redes son en un medio para convocar a eventos de todo tipo, sirven para dar a conocer las plataformas de los candidatos y difundir sus ideas y pueden ser utilizadas como escenario para el debate directo, de forma horizontal, entre candidatos y simpatizantes. Por estas y otras razones se hace importante analizar cómo las redes sociales intervienen de forma activa en las campañas políticas.¹ Con lo anterior en mente, durante la actual temporada electoral, una consultora fue contratada para analizar el comportamiento de la campaña de uno de los candidatos a la presidencia en redes sociales. El interés recae en la cantidad de tuits diarios que dicho candidato escribe en su cuenta oficial de Twitter. Para hacer el análisis, se tomaron los 120 registros entre el 31 de enero y el 31 de mayo de este año (hasta mayo 17 se usaron para train y el resto para test).

Gráficos de diagnóstico



¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.

Suavización exponencial

ETS(A,Ad,N)

Call:ets(y = serie_tuits, model = "zzz", opt.crit = "mse", ic = "aic")

Smoothing parameters:

alpha = 0.9983
beta = 0.9965
phi = 0.8

Initial states:

l = 7.3654
b = 1.8326

sigma: 1.3743

AIC AICC BIC
664.1305 664.8674 680.9053

Training set error measures:

	ME	RMSE	MAE	MPE
Training set	0.005413882	1.345625	1.048024	0.372276

	MAPE	MASE	ACF1
Training set	6.382951	0.830176	0.2624311

Augmented Dickey-Fuller Test
data: residuals(modelo_suav_tuits)
Dickey-Fuller = -8.307, Lag order = 4,
p-value = 0.01
alternative hypothesis: stationary

Resultados en la base de validación:

	ME	RMSE	MAE
Training set	0.04679141	7.48455	3.647332

	MPE	MAPE	MASE	ACF1
Training set	-0.07306702	13.61659	0.8265	-0.4560294

Auto ARIMA

ARIMA(2,1,2) with drift

Coefficients:

	ar1	ar2	ma1	ma2	drift
1.	1.2832	-0.6663	-0.5112	-0.2552	0.1044
s.e.	0.0866	0.0860	0.1178	0.1221	0.0578

sigma^2 estimated as 1.036: log likelihood=-170.69
AIC=353.38 AICc=354.12 BIC=370.1

Training set error measures:

	ME	RMSE	MAE
Training set	0.009815858	0.9920609	0.7729097

	MPE	MAPE	MASE
Training set	-0.0524107	4.599221	0.6122483

	ACF1
Training set	0.01026672

¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.

```

Augmented Dickey-Fuller Test
data: residuals(modelo_tuits)
Dickey-Fuller = -4.571, Lag order = 4,
p-value = 0.01
alternative hypothesis: stationary

```

Resultados en la base de validación

	ME	RMSE	MAE	MPE
Training set	0.2052186	1.031987	0.8078012	1.012193
	MAPE	MASE	ACF1	
Training set	4.935688	0.6423	0.005468141	

1. (6) Haga un detallado análisis descriptivo del comportamiento de la serie de la cantidad de tuits diarios publicados por el candidato. Analice, además, cómo la modelación de esta serie podría usarse en la campaña electoral. Explique.

Rta: Inicialmente se observa una tendencia de acenso en la cantidad de tweets con el paso del tiempo, es decir que con el tiempo el candidato va interactuando más con sus seguidores y futuros votantes, posiblemente generándoles confianza y tratando de empatizar más con ellos a través de su presencia en redes. Por otro lado con los gráficos de correlación se observa que en la gran mayoría de casos se sobrepasa el umbral esperado, indicándonos una posible relación del hoy con los pasos anteriores, sin embargo tras realizar las transformaciones ya no todos los casos sobrepasan este umbral, más si se puede destacar que los primeros días tienen una correlación.

Adicionalmente, a nivel de cómo se puede usar para la campaña, por parte del equipo del candidato pueden predecir los días donde el candidato este más activo y llevar un control de esto es decir que esta actividad sea en pro de conectar con la gente y así poder ganar más adeptos y no que se sea porque se está metiendo en una polémica y viendo afectada su imagen ya que esto lo podría llevar a la perdida de las elecciones y de sus votantes; por otro parte se podría realizar una análisis extra en relación a la cantidad de tweets y que tanto sube en las encuestas y sus seguidores.

Por otra parte en las campañas de otros candidatos pueden tomar provecho de esta actividad del candidato en cuestión para que ellos también fortalezcan sus redes los días donde tal vez este candidato este ausente o para compartir su contenido destacando las ventajas que tendría votar por ellos tratando de convencer a los seguidores de dicho candidato.

2. (6) Con base en las métricas presentadas, ¿cuál sería el modelo más adecuado para predecir la cantidad de tuits diarios escritos en la cuenta de Twitter del candidato? Explique su modelo escogido en términos del contexto electoral.

Rta: Para este caso se busca en el modelo que se incurra en el menor error posible para esto se tienen en cuenta las métricas de RMSE MAE MAPE y MASE en las bases de validación (test) de cada uno de los modelos; con esto en mente se puede observar que para el modelo auto ARIMA se obtienen valores menores que para el modelo de Suavización Exponencial, es decir nos quedamos con el ARIMA donde se tiene un error promedio absoluto de 0.8078012, un RMSE que si bien es más sensible a los outliers le

¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.

da más peso a los errores grandes para este caso es un poco más de una unidad (1.031987) mientras que el del primer algoritmo son más de 7 unidades; adicionalmente se tiene un promedio porcentual de los errores de 4.935688 y un MASE de 0.6423 y al ser este menor que 1 nos indica que es mejor que un ingenuo en un 64.23% ya que solo se incluye el 35,77% de error del ingenuo.

Ahora bien a nivel del contexto se observa que la cantidad de tweets que el candidato genere hoy se ven influenciados por la cantidad de tweets que se generaron hace dos pasos, es decir hace dos días, al igual que el error en la predicción que se puede llegar a tener en cuantos tweets se harán en x días involucra el error cometido hace dos días. Por otra parte con las métricas expuestas anteriormente se puede observar que el error no es tan grande y que el sistema puede estar atinándole varias veces, entonces si este modelo lo utiliza la competencia pueden decir “Bueno el candidato x va a hacer esto, mejor yo me le adelanto y hago el doble así conecto más”.

Basado en los resultados de los modelos presentados, comente las siguientes afirmaciones indicando si las considera pertinentes/ciertas para el contexto (señale puntos cuantitativos específicos de los resultados que soporten su respuesta en cada caso):

3. (4) La campaña del candidato contendor afirma: “*los registros indican que el candidato analizado está disminuyendo la fuerza de su campaña en redes sociales a medida que se acerca la fecha de elecciones*”.

Rta: Para este caso estamos viendo una serie que presenta estacionalidad, donde la principio podemos ver (según la gráfica) una tendencia al acenso luego se ve como que se acaba esa tendencia, sin embargo tras realizar las transformaciones si seguimos viendo que hay fechas donde hay picos es decir una mayor cantidad de tweets y luego puede volver a caer, como se nos indica en el valor del ARIMA (2,1,2) se depende en la predicción de los dos pasos anteriores, es por esto que se debería tomar un poco más de contexto para decir “*que se está disminuyendo la fuerza de su campaña en redes*” por la forma en como esto se ve involucrado puede que hayan sido días donde el contexto haya tenido más influencia y dentro de un par de días vuelva y se dispare la fuerza de la campaña en redes.

4. (4) El Community Manager de la campaña dice: “*no vale la pena tener un modelo que pronostique la cantidad de publicaciones pues yo he fijado el número de tuits de acuerdo con el día de la semana*”.

Rta: Teniendo en cuenta que el MASE para este modelo es menor a 1 (0.6423) se nos indica que el modelo tiene un mejor comportamiento que un modelo ingenuo, como sería el fijarse en los tweets de cada día, es decir que nos va mejor con el modelo que pronostique que con solo a ojo humano.

¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.

Caso 2: (30 puntos) La base de datos “caso1.xlsx” contiene el registro de 7629 clientes actuales de un supermercado con las siguientes variables:

- Genero: género declarado por la persona
- Estadocivil: estado civil declarado por la persona
- Edad: Edad declarada por la persona
- ViviendaPropia: información traída de una base externa sobre si la persona posee o no vivienda propia
- Compra: el resultado de compra a una oferta de seguros de hogar (contenidos de la casa y reparaciones).

Usted debe desarrollar un modelo de analítica para predecir la posible compra de futuros clientes de los que se posee la misma información. El desarrollo del modelo incluye:

- a) (5) Un proceso de entendimiento de los datos
- b) (10) El desarrollo adecuado de un modelo acorde con lo visto en curso
- c) (5) La valoración del modelo con las métricas y técnicas adecuadas
- d) (10) Una explicación de las características detectadas por el modelo de potenciales clientes del seguro que debe ser entregada al área de ventas.

Como primer paso se tiene el entendimiento de los datos, se cuenta con una base de datos de 7629 registros, con 5 atributos donde el número 5, es decir el atributo “compra” es el que se busca predecir. Adicionalmente solo cuenta con variables de tipo carácter y para el caso de la variable “edad” se cuenta con algunos datos perdidos.

```
> ## revisar la estructura del objeto que se tiene y los diferentes
> ## tipos de datos mediante el comando str
> str(caso)
tibble [7,629 x 5] (S3:tbl_df/tbl/data.frame)
$ Genero      : chr [1:7629] "FEMENINO" "FEMENINO" "MASCULINO" "FEMENINO" ...
$ estadocivil : chr [1:7629] "VIUDO" "CASADO" "SOLTERO" "SOLTERO" ...
$ edad        : chr [1:7629] "51 - 60 años" "41 - 50 años" "41 - 50 años" "31 - 40 años" ...
$ ViviendaPropia: chr [1:7629] "NO" "SI" "NO" "NO APLICA" ...
$ compra      : chr [1:7629] "no" "no" "no" "si" ...
>
>
> summary(caso)
   Genero      estadocivil      edad      ViviendaPropia      compra
Length:7629    Length:7629    Length:7629    Length:7629    Length:7629
Class :character Class :character Class :character Class :character Class :character
Mode  :character Mode  :character Mode  :character Mode  :character Mode  :character
> |
```

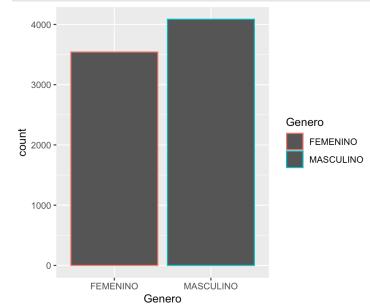
Por su parte el balance de los datos en el atributo a predecir, se observa que al rededor 67% son correspondientes al valor “no”, mientras que el resto (32%) al valor “si”, que para este caso es el más importante ya se quiere predecir “la posible compra de futuros clientes”.

```
> balance <- table(caso$compra)
> prop.table(balance)

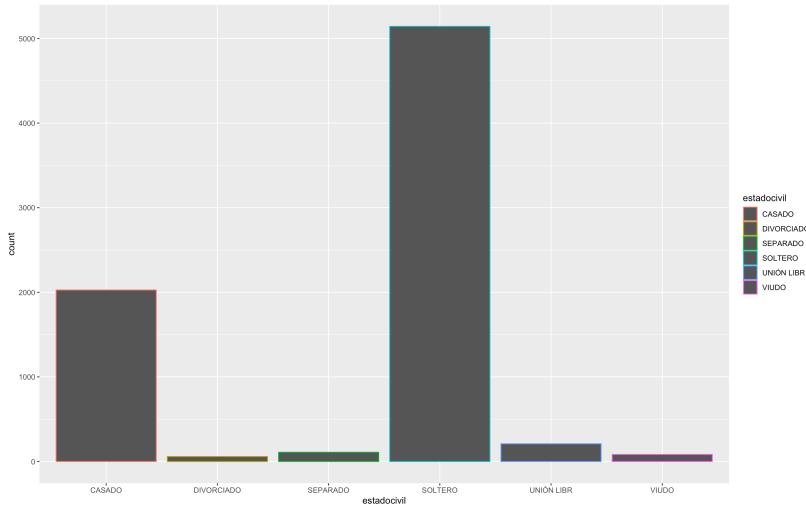
          no       si
0.6709366 0.3290634
```

Posteriormente se pasa a explorar los datos a través de diferentes diagramas de barras mostrados a continuación. Para el caso del atributo Genero se observa que los dos géneros tienen alta representatividad en la base de datos.

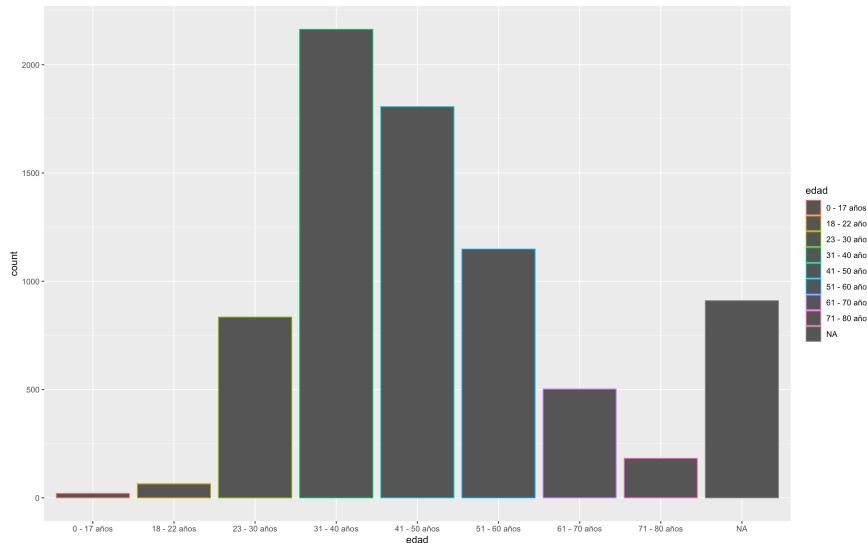
¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.



Para el siguiente atributo a evaluar que fue estadocivil se observa que puede tomar 5 valores, casado, divorciado separado, soltero, unión libre o viudo; adicionalmente se observa que hay una alta presencia de solteros en el data set, seguido de casados, mientras que hay muy pocos divorciados y viudos.

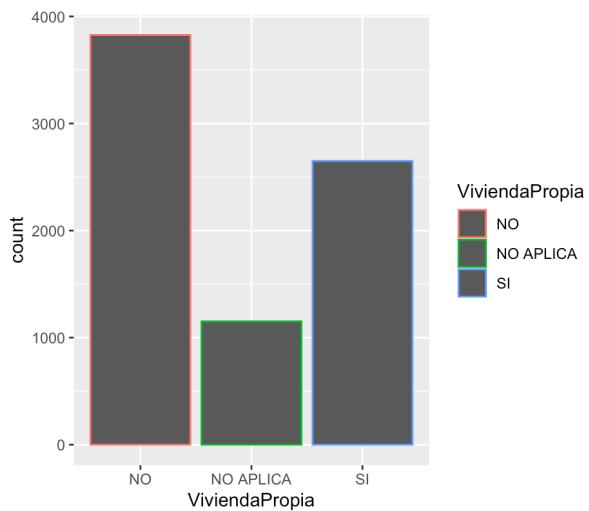


Posteriormente se gráfica el atributo edad, el cual se observa que esta dado en rangos donde los que tienen mayor representatividad en la base de datos son de 31-40 años y de 41 – 50 años, mientras que son muy pocos los datos que hay de las edades más jóvenes, es decir de 0-17 años y de 18 – 22 años.

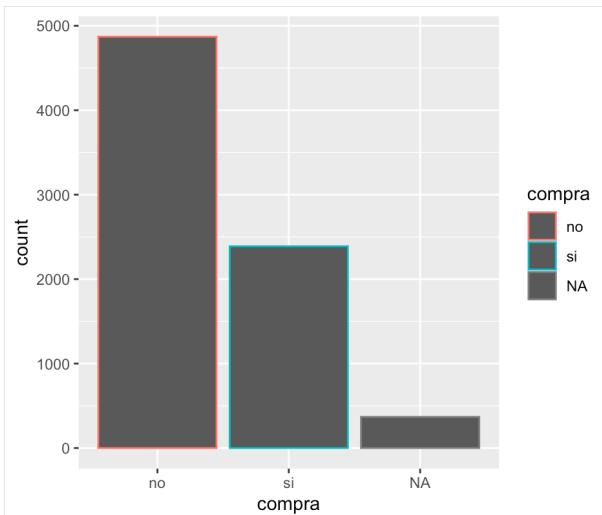


El siguiente atributo a analizar fue el de vivienda propia, en esta caso se observa que la gran mayoría no tienen una vivienda propia.

¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.



La última variable que se grafico fue la de compra donde se observa como se menciono anteriormente la gran mayoría de estos datos son de personas que no compraron.



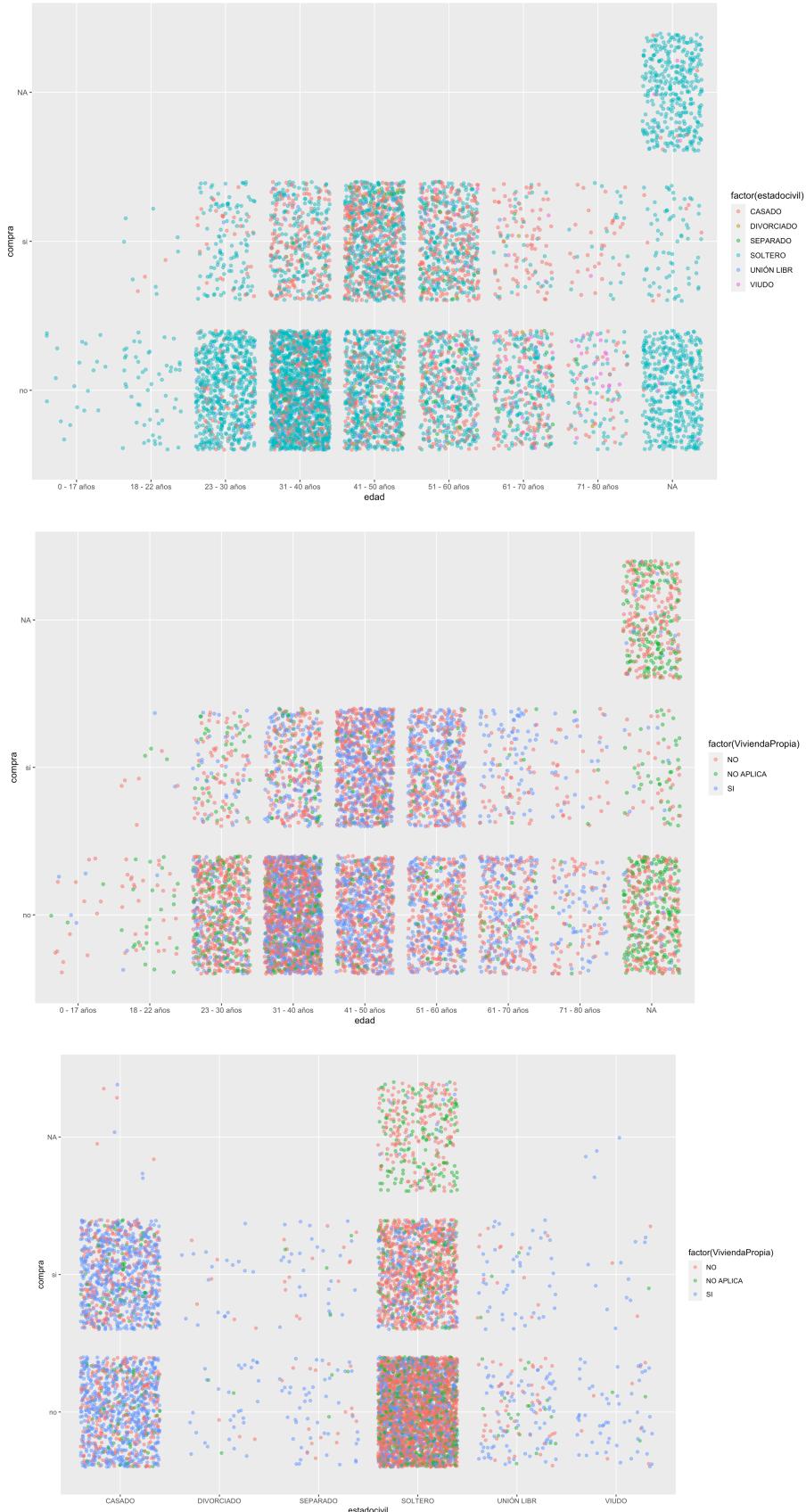
Posteriormente se observa el comportamiento de los atributos con la variable a predecir, como se ven las gráficas a continuación, para este caso se destacan las siguientes observaciones:

- Los solteros, sin importar su genero y edad, no tienden a comprar.
- Las personas casadas si tienden a comprar.
- Las personas en los rangos de edad de 31- 40 años, sin importar si tienen vivienda propia o no, presentan una mayor concentración en no comprar.

¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.



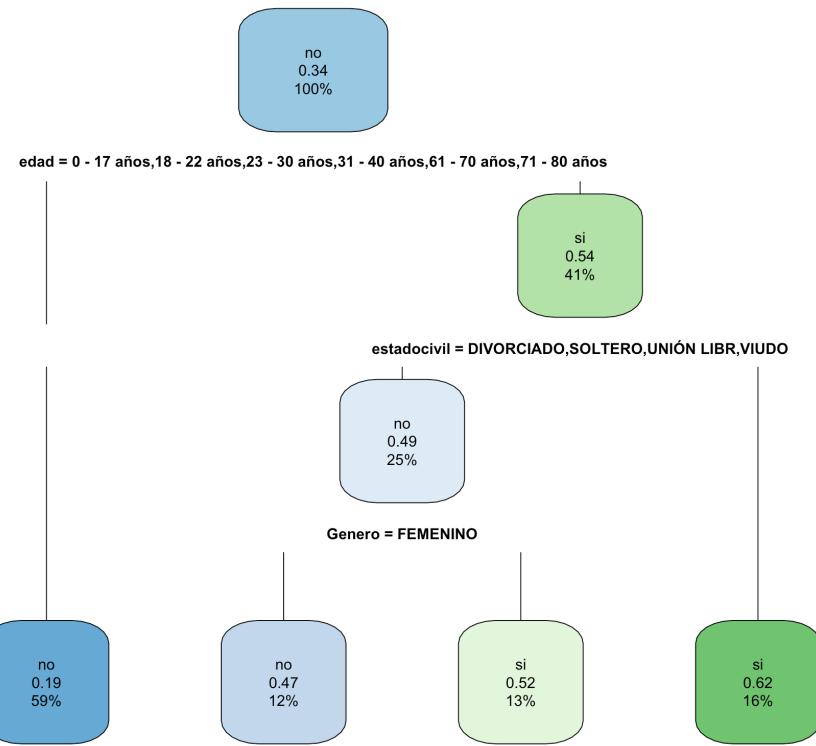
¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.



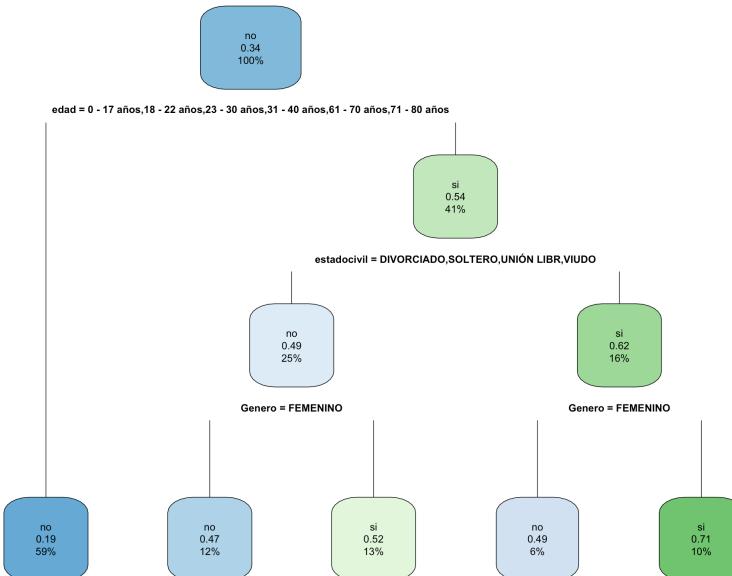
Posteriormente se decide pasar los diferentes atributos a tipo factor con el fin de facilitar el proceso al utilizar un árbol para realizar las predicciones. Y se realiza la partición del data set en dos uno para el train y otro para el test.

¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.

Se generan 2 árboles, el primero ilustrado a continuación no se le especifico nada en sus parámetros más que tenía que predecir la compra.



Sin embargo para el segundo árbol si se le especifican más cosas, como que tenía que tener un minbucket de 5, un minisplit de 30 y un cp de 0.0041, el árbol obtenido se ilustra a continuación.



Por otro lado, se quiso saber cuáles variables tenían mayor importancia para el segundo modelo, estas se destacan en la imagen a continuación, como se puede observar la más importante es la edad, la cual es la primera con la que se realiza la primera partición, donde descarta a todos los menores de 40 años, es decir que menciona que ninguno de estos es probable que compre; posteriormente la siguiente partición la realiza según

¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.

el estado civil de la persona, sin embargo la importancia de esta no es tan alta como la de la edad, en tercer lugar se tiene el género y finalmente si se cuenta con vivienda propia.

Variable importance		Genero	ViviendaPropia
edad	estadocivil		
80	11	6	2

Se decide obtener la matriz de confusión y las estadísticas para cada uno de los árboles. A continuación se muestran los resultados de estadísticas en test para cada uno de los árboles.

```
> conf1
Confusion Matrix and Statistics

Reference
Prediction  no  si
no 997 278
si 260 276

Accuracy : 0.7029
95% CI : (0.6813, 0.7239)
No Information Rate : 0.6941
P-Value [Acc > NIR] : 0.2150

Kappa : 0.294

McNemar's Test P-Value : 0.4636

Sensitivity : 0.7932
Specificity : 0.4982
Pos Pred Value : 0.7820
Neg Pred Value : 0.5149
Prevalence : 0.6941
Detection Rate : 0.5505
Detection Prevalence : 0.7040
Balanced Accuracy : 0.6457

'Positive' Class : no
```

Ilustración 1. Estadísticas árbol

```
> conf2
Confusion Matrix and Statistics

Reference
Prediction  no  si
no 1050 320
si 207 234

Accuracy : 0.709
95% CI : (0.6875, 0.7298)
No Information Rate : 0.6941
P-Value [Acc > NIR] : 0.08783

Kappa : 0.2733

McNemar's Test P-Value : 1.067e-06

Sensitivity : 0.8353
Specificity : 0.4224
Pos Pred Value : 0.7664
Neg Pred Value : 0.5306
Prevalence : 0.6941
Detection Rate : 0.5798
Detection Prevalence : 0.7565
Balanced Accuracy : 0.6289

'Positive' Class : no
```

Ilustración 2. Estadísticas árbol con parámetros

Como se puede observar para ambos casos se obtienen estadísticas similares, por una parte vemos que la sensibilidad es alta, es decir que de los que tenía que predecir como positivos predice la gran mayoría en un caso 79,32% y en el otro 83,53% de estos positivos , sin embargo su especificidad no es muy alentadora, es decir

¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.

la probabilidad de que diga que algo va a pasar y pase es tan solo del 49,82% y 42,22%, como en este caso el modelo toma como clase positiva el no, por sus estadísticas este va a decir en muchos casos que no se va a comprar cuando puede que si compre. Finalmente vemos que se tiene un accuracy del 70,29% y 70,9% donde nos indica el que del total de datos a predecir el 70,29% y 70,9%, en cada modelo, está bien predicho, si consideramos que en las curvas ROC y en un ingenuo usual el chance de que se clasifique bien está en un 50-50, podemos decir que se está por encima de estos y a pesar de las falencias que se tienen no están mal.

Finalmente a modo de recomendación para el negocio según las características más importantes en los modelos, es enfocarse en las ventas para las personas de mayor edad, esto puede ser debido a que

1. Tienen mayores ingresos y pueden cubrir los costos del adquirir un inmueble o están ya buscando estabilidad para su hogar, echar raíces o fallecer, mientras que en el caso de los jóvenes muchos apenas están iniciando su vida laboral por ende no tienen los ingresos que cubran los costos de adquirir un inmueble, sus prioridades son otras como el viajar.
2. Los mayores suelen contar con un historial crediticio más amplio que también les permite tener la facilidad de adquirir un crédito para financiar sus hogares.

Otra variable importante es el estado civil siempre y cuando las personas sean de mayor edad , el cual como se vio en las gráficas, las personas que están solas no tienden a comprar, esto puede estar ligado a las prioridades que tienen, se recomienda investigar el porqué de esto, para poder generar propuestas de compra para las personas solitarias.

Finalmente se tiene en cuenta el Género, donde si se es mujer se tiende a comprar más, cabe resaltar que para este atributo también se debe tener en cuenta que sea una mujer de mayor edad, donde puede llegar a influir su estado civil más no en un 100%. Para concluir se puede decir que el grupo focal podrían ser mujeres, con cualquier estado civil mayores a 40 años.

¹ Palacio, E. (2016). Las redes sociales: un aliado para las campañas políticas. Revista de Estudiantes de Ciencia Política, 8, 20-33.