Ejercicio 1: Movistar

Antonio J. Perán, Roberto García, Damián Bornás

Índice

Descripción del conjunto de datos

El conjunto de datos consta de 21 variables y 7032 observaciones después de eliminar algunas cuyas variables tomaban valores perdidos. Cada observación representa un contrato firmado por un cliente con Movistar, añadiendo además una variable que informa de si este cliente mantiene o no el contrato en la actualidad con la empresa. La lista de variables es la siguiente:

- customerID: Esta variable es un identificador interno de la compañía para el cliente. Como para nuestro análisis es innecesaria, esta variable será eliminada del conjunto de datos.
- gender: El sexo del titular del contrato, que toma los valores `Male` y `Female`.
- SeniorCitizen: Variable que toma los valores `yes` o `no` dependiendo de si el cliente está o no jubilado.
- Partner: Variable que toma los valores `yes` o `no` en función de que el cliente tenga o no pareja.
- Dependents: Variable que toma los valores `Yes` o `No` en función de que el cliente tenga o no familiares dependientes a su cargo.
- tenure: Variable numérica que recoge el número de meses de duración del contrato.
- PhoneService: Variable que toma los valores `Yes` o `No` dependiendo de si el cliente contrata o no el servicio de telefonía.
- MultipleLines: Variable que toma los valores `Yes`, `No` o `No phone service` dependiendo de si el cliente contrata o no múltiples líneas telefónicas o si no ha contratado el servicio de telefonía.
- InternetService: Variable que toma los valores `DSL`, `Fiber Optic` o `No` dependiendo del tipo de internet contratado o si no ha contratado este servicio.
- OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV y StreamingMovies: Variables que toman los valores `Yes`, `No` o `No internet service` dependiendo de si el cliente contrata o no este servicio o si no lo contrata debido a que no contrató el servicio de internet, ya que estos servicios están ligados al servicio de internet.
- Contract: Esta variable recoge el tipo de contrato firmado. Toma los valores `Month-to-month` para un contrato mensual, `One year` para un contrato anual y `Two year` para un contrato bianual.
- PaperlessBilling: Esta variable toma los valores `Yes` o `No` dependiendo de si el cliente ha pedido facturación sin papel o con papel. La facturación sin papel podría ser vía correo electrónico u otras vías.
- PaymentMethod: Esta variable recoge el método de pago empleado en el cobro de la factura. Los valores toma son `Bank transfer (automatic)`, `Credit card (automatic)`, `Electronic check` y `Mailed check`.
- MonthlyCharges: Esta variable recoge el pago mensual que realiza el cliente por mantener el contrato.
- TotalCharges: Esta variable recoge el dinero total que ha desembolsado el cliente mientras ha mantenido vigente el contrato. Por tanto, esta variable está fuertemente relacionada con las variables `tenure` y `MonthlyCharges`, ya que `TotalCharges = MonthlyCharges*tenure`. De hecho el coeficiente de correlación entre `TotalCharges` y `MonthlyCharges*tenure` es 0.9996, lo que reafirma lo anterior. Por esta razón esta variable será eliminada del conjunto de datos.
- Churn: Esta variable es la que nos indica si el cliente ha roto o no su contrato con Movistar. Toma el valor `Yes` si el cliente se ha ido y el valor `No` si sigue con la compañía.

Apartado 1

Movistar nos pide que realicemos un análisis de perfiles de posibles desertores utilizando los datos contenidos en el dataset `Telco-Customer-Churn.csv` con el objetivo de evitar una fuga de usuarios con este perfil a compañias como Vodafone. También nos pide indentificar el perfil de clientes fieles a la compañía. Todo esto en base a características interpretables de los clientes, de ahí que una buena forma de abordar el problema sean los árboles de clasificación.

Metodología

Análisis exploratorio

Antes de aplicar cualquier modelo es conveniente familiarizarnos con los datos mediante un análisis exploratorio. Los cuadros 1, 2, 3 muestran análisis descriptivos de frecuencias para las variables categóricas y el cuadro 4 muestra la media y desviación típica para las variables numéricas agrupando siempre según la variable `Churn`.

Cuadro 1: Tabla de descriptivos para variables sociodemográficas

		Churn					
		No				Yes	
		n	perc	n	perc	n	perc
gender	Female	3483	50	2544	49,3	939	50,2
	Male	3549	50	2619	50,7	930	49,8
SeniorCitizen	No	5890	84	4497	87,1	1393	74,5
	Yes	1142	16	666	12,9	476	25,5
Partner	No	3639	52	2439	47,2	1200	64,2
	Yes	3393	48	2724	52,8	669	35,8
Dependents	No	4933	70	3390	65,7	1543	82,6
	Yes	2099	30	1773	34,3	326	17,4
Contract	Month-to-month	3875	55	2220	43,0	1655	88,6
	One year	1472	21	1306	25,3	166	8,9
	Two year	1685	24	1637	31,7	48	2,6
PaperlessBilling	No	2864	41	2395	46,4	469	25,1
	Yes	4168	59	2768	53,6	1400	74,9
PaymentMethod	Bank transfer (automatic)	1542	22	1284	24,9	258	13,8
	Credit card (automatic)	1521	22	1289	25,0	232	12,4
	Electronic check	2365	34	1294	25,1	1071	57,3
	Mailed check	1604	23	1296	25,1	308	16,5

Cuadro 2: Tabla de descriptivos para variables relacionadas con el servicio telefónico.

		Churn					
		No				Yes	
		n	perc	n	perc	n	perc
PhoneService	No	680	9,7	510	9,9	170	9,1
	Yes	6352	90,3	4653	90,1	1699	90,9
MultipleLines	No	3385	48,1	2536	49,1	849	45,4
	No phone service	680	9,7	510	9,9	170	9,1
	Yes	2967	42,2	2117	41,0	850	45,5

Cuadro 3: Tabla de descriptivos para variables relacionadas con los servicios de internet.

				Churn			
				No		Yes	
		n	perc	n	perc	n	perc
InternetService	DSL	2416	34	1957	38	459	25
	Fiber optic	3096	44	1799	35	1297	69
	No	1520	22	1407	27	113	6
OnlineSecurity	No	3497	50	2036	39	1461	78
	No internet service	1520	22	1407	27	113	6
	Yes	2015	29	1720	33	295	16
OnlineBackup	No	3087	44	1854	36	1233	66
	No internet service	1520	22	1407	27	113	6
	Yes	2425	34	1902	37	523	28
DeviceProtection	No	3094	44	1883	36	1211	65
	No internet service	1520	22	1407	27	113	6
	Yes	2418	34	1873	36	545	29
TechSupport	No	3472	49	2026	39	1446	77
	No internet service	1520	22	1407	27	113	6
	Yes	2040	29	1730	34	310	17
StreamingTV	No	2809	40	1867	36	942	50
	No internet service	1520	22	1407	27	113	6
	Yes	2703	38	1889	37	814	44
StreamingMovies	No	2781	40	1843	36	938	50
	No internet service	1520	22	1407	27	113	6
	Yes	2731	39	1913	37	818	44

Cuadro 4: Tabla de descriptivos para variables numéricas.

		Churn			
		No		Yes	
	n	mean	sd	mean	sd
tenure	7032	37,65	24,08	17,98	19,53
MonthlyCharges	7032	61,31	31,09	74,44	24,67

Las tablas son útiles para determinar con precisión los valores exactos de las variables, pero es difícil apreciar patrones entre tantos números. Una mejor visualización de las posibles relaciones entre las variables de nuestro conjunto de datos y la variable `Churn` puede verse en la figura 1. En ella se muestra la distribución de las variables consideradas tanto para aquellos clientes que siguen con Movistar como para los que no. Para cada variable categórica se muestra un diagrama de barras representando la frecuencia relativa, en tantos sobre uno; y para cada variable numérica un gráfico de cajas o boxplot.

En la figura 1 pueden verse características notables de aquellos clientes que rescinden el contrato con la empresa. Estos son personas que contratan mayoritariamente fibra óptica (variable `InternetService`) con un contrato mensual (variable `Contract`) y cuya duración media de contrato es de 18 meses (variable `tenure`), es decir, apenas superan el año de contrato, realizando el pago mediante Electronic check (variable `PaymentMethod`).

También se aprecian otras posibles relaciones más débiles como que son personas más frecuentemente solteras (variable `Partner`), no contratan soporte técnico ni seguridad online (variables `TechSupport` y `OnlineSecurity`) y asumen un gasto mensual más alto.

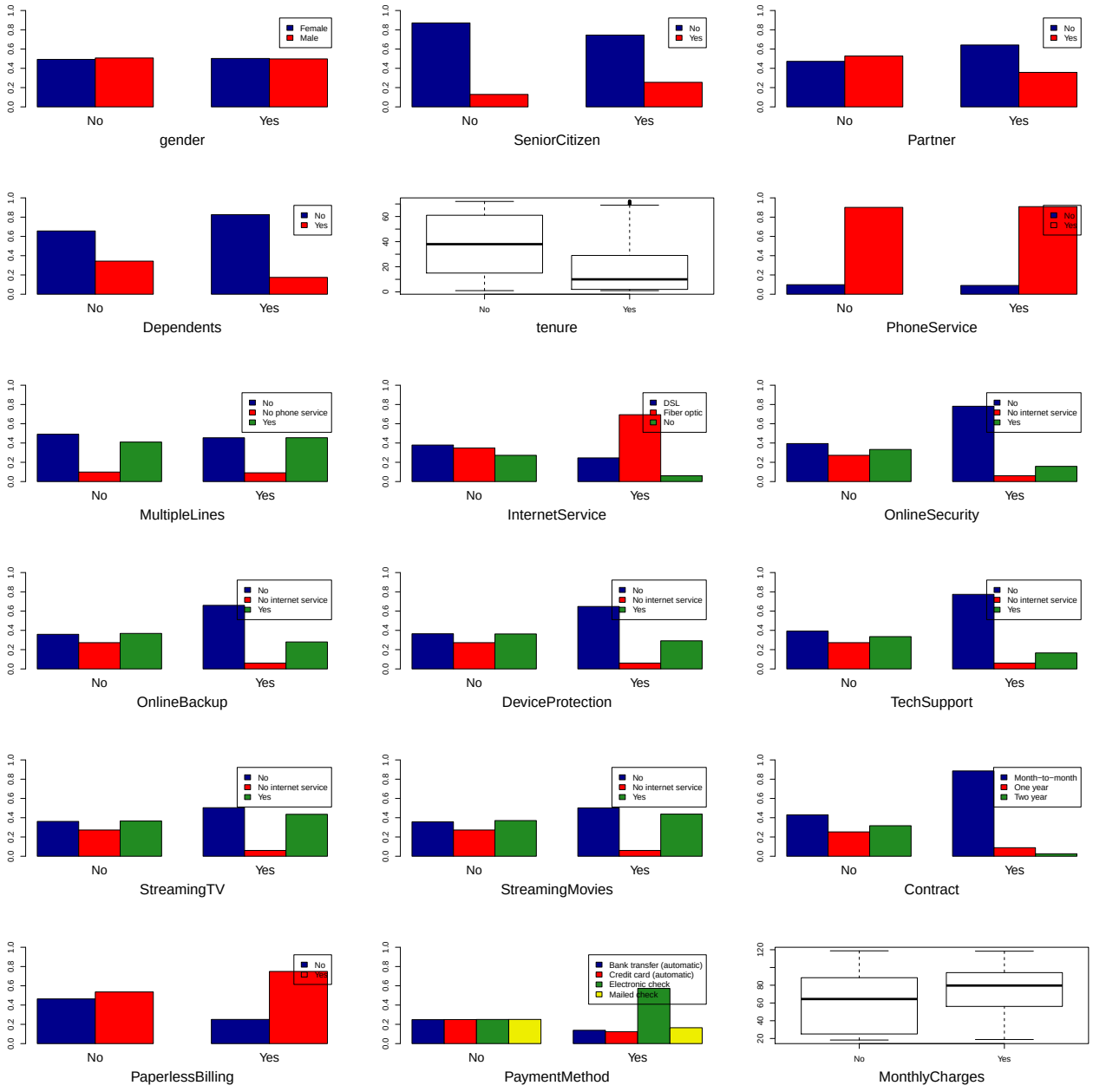

Figura 1: Distribución de las variables del conjunto de datos agrupando según Churn.

Contrastes

Para validar si las relaciones que vimos anteriormente son reales o se deben a la aleatoriedad de los datos se realizan contrastes de hipótesis cuyos resultados pueden consultarse en los cuadros 5 y 7. Para variables categóricas empleamos el test χ^2 y para las numéricas el test t-student. Además calculamos también el tamaño del efecto para cuantificar esta relación, en el caso de variables categóricas empleamos el estadístico V de Cramer y para variables numéricas el estadístico d de Cohen.

Según los resultados que se muestran en el cuadro 5 podemos ver que el género del cliente no tiene relación ninguna con la variable `Churn`, ni el hecho de que contrate o no línea telefónica. Todas las demás resultan tener una relación estadísticamente significativa pero con tamaños del efecto bajos. Como aventurábamos antes, es la variable `Contract` la que tiene una mayor relación con la variable `Churn`.

En el cuadro 7 podemos ver los resultados de las pruebas t de Student para muestras independientes. Ya que el tamaño muestral es tan grande, podemos asumir normalidad, pero no homocedasticidad, pues unos test previos indicaron que las distribuciones no tienen igual varianza. Los resultados indican que existe una relación significativa entre las variables, destacando un gran tamaño del efecto en la variable `tenure`.

Cuadro 5: Resultados de las pruebas χ^2.

	chi2	df	pvalue	vcramer
gender	0.01	1	0.490	0.01
SeniorCitizen	0.15	1	0.000	0.15
Partner	0.15	1	0.000	0.15
Dependents	0.16	1	0.000	0.16
PhoneService	0.01	1	0.350	0.01
MultipleLines	0.04	2	0.004	0.04
InternetService	0.32	2	0.000	0.32
OnlineSecurity	0.35	2	0.000	0.35
OnlineBackup	0.29	2	0.000	0.29
DeviceProtection	0.28	2	0.000	0.28
TechSupport	0.34	2	0.000	0.34
StreamingTV	0.23	2	0.000	0.23
StreamingMovies	0.23	2	0.000	0.23
Contract	0.41	2	0.000	0.41
PaperlessBilling	0.19	1	0.000	0.19
PaymentMethod	0.30	3	0.000	0.30

Cuadro 7: Resultados de las pruebas t de Student.

	t	df	pvalue	dcohen
tenure	-34.97	4045.51	0	0.86
MonthlyCharges	18.34	4139.67	0	0.44

Árboles de clasificación inferenciales[1]

Para abordar el problema de la clasificación vamos a probar diversos algoritmos de aprendizaje automático, para lo que debemos dividir la muestra en dos conjuntos: uno de entrenamiento y otro de test, donde el 80 % de la muestra original será destinado al entrenamiento y el 20 % para el test. en la muestra de entrenamiento tenemos 4109 clientes que continúan con la empresa y 1516 desertores. En la muestra de test tenemos 1054 clientes que continúan con la empresa y 353 desertores.

El primer método que vamos a utilizar son los árboles de clasificación infenciales, un algoritmo a medio camino entre los árboles de clasificación tradicionales y el RandomForest. En este tipo de árboles la selección de variables en cada nodo se realiza mediante contrastes estadísticos quedándonos con aquella variable que arroja el p-valor más bajo.

En la figura 2 puede verse el error cometido en cada árbol entrenado así como la sensibilidad y especificidad del modelo. Los distintos modelos han sido entrenados en función del hiperparámetro `minsplit` que controla el tamaño del árbol asignando un tamaño mínimo a un nodo (número mínimo de observaciones) para que este vuelva a ser dividido. Como se observa en la figura 2 podemos ver que los modelos tienen, en general, un error en torno al 20 %; una sensibilidad media/baja, esto es, mala capacidad para detectar clientes que se van a ir realmente y una especificidad alta, esto es, la capacidad del modelo para detectar clientes que no son van a ir realmente.

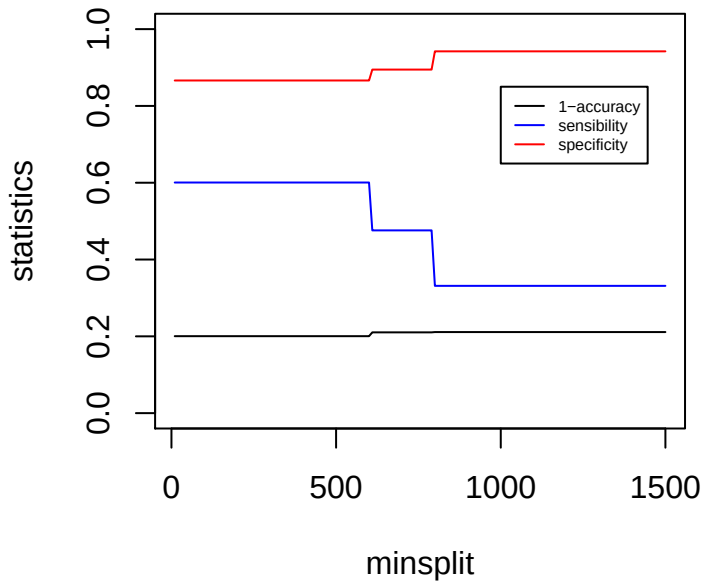

Figura 2: Precisión de los distintos modelos de árboles de clasificación inferenciales generados.

En vista de los resultados anteriores el modelo elegido es el que tiene como valor del hiperparámetro minsplit igual a 600, ya que aunmentar un poco la especificidad supone sacrificar mucho la sensibilidad. Elegimos, entonces, el modelo más simple con mejores resultados en el gráfico anterior.

A continuación mostramos la matriz de confusión resultante de aplicar este modelo de clasificación al conjunto de test. Como podemos ver, 913 clientes son clasificados correctamente como clientes que continúan con la empresa y 212 como desertores. Por otro lado 141 sujetos son clasificados como desertores cuando en realidad no lo son y también 141 sujetos son clasificados como clientes fidedignos cuando en realidad no lo son.

[1] Kabacoff (2015)

```
##       Predicted
## Actual  No Yes
##    No  913 141
##    Yes 141 212
```

En la figura 3 puede verse el gráfico del árbol resultante con `minsplit=600` en el que se han abreviado las etiquetas de las variables para una mejor visualización. El árbol tiene una precisión de 0.7996, una sensibilidad de 0.601 y especificidad de 0.866.

Tal y como muestran las probabilidades de los nodos terminales, son los clientes de los nodos 17, 23, 24 y 27 los que tienen una probabilidad más alta de desertar, siendo mucho más notable en los nodos 23 y 24, y dudoso en los restantes. Por tanto, el perfil más pronunciado de clientes que se van son aquellos que tienen un contrato mensual con fibra óptica y que no superan el año de duración del contrato (nodos 23 y 24), siendo el primer mes (nodo 23) crucial para mantener al cliente, que quizá abandona por no estar satisfecho con la calidad del servicio de fibra óptica.
Por otro lado, aquellos clientes con contrato mensual con fibra óptica que superan el año es más probable que se queden (nodos 28 y 29), salvo aquellos que pagan mediante `Electronic Check`(nodo 27). También tenemos clientes con contratos mensuales con `DSL` o sin internet que es probable que se queden (nodos 16, 19 y 20), salvo aquellos que no contratan soporte técnico y abandonan la empresa en los primeros cinco meses (nodo 17).
Finalmente, aquellos que adquieren un contrato anual o bianual con Movistar tienen todos unas altas probabilidades de ser clientes fidedignos, sobretodo con contratos bianuales (nodos 5, 6, 8, 9, 11 y 12).

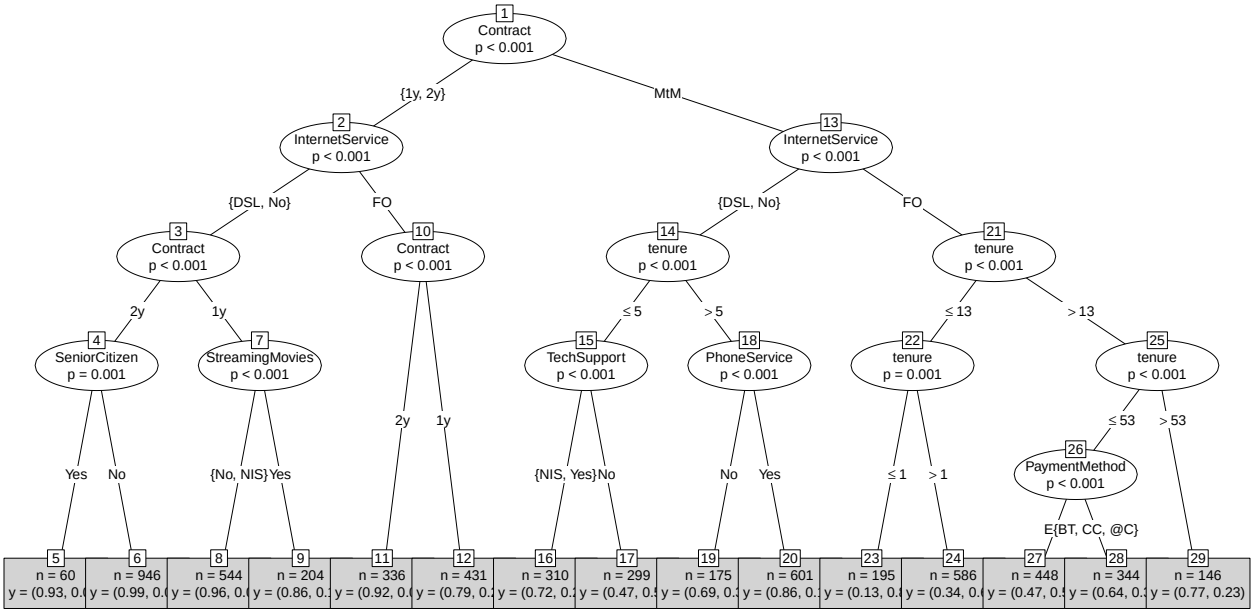

Figura 3: Gráfico del arbol de clasificación inferencial con minsplit=600.

Árboles de clasificación ordinarios[2]

Ahora vamos a probar la clasificación con un modelo más sencillo que el anterior para ver si obtenemos los mismos resultados. En este caso, el modelo utilizado son los árboles de clasificación ordinarios. Con el paquete `rpart` de R realizamos una validación cruzada con distintos tamaños para el árbol obteniendo los siguientes

[2]Kabacoff (2015)

resultados de error de validación cruzada (`xerror`) en función de un parámetro de complejidad del árbol `cp`. Siguiendo la regla de `one-standar-deviation` y viendo que solo tenemos dos modelos, el modelo elegido es el de `cp = 0.01`.

```
##           CP nsplit rel error    xerror       xstd
## 1 0.05837731      0 1.0000000 1.0000000 0.02195115
## 2 0.01000000      3 0.7816623 0.7941953 0.02029145
```

El gráfico del árbol resultante puede verse en la figura 4. Los resultados son consistentes e idénticos a los reportados por el método anterior, aunque este resultado es mucho más sencillo e interpretable, a costa de desequilibrar sensibilidad y especificidad.

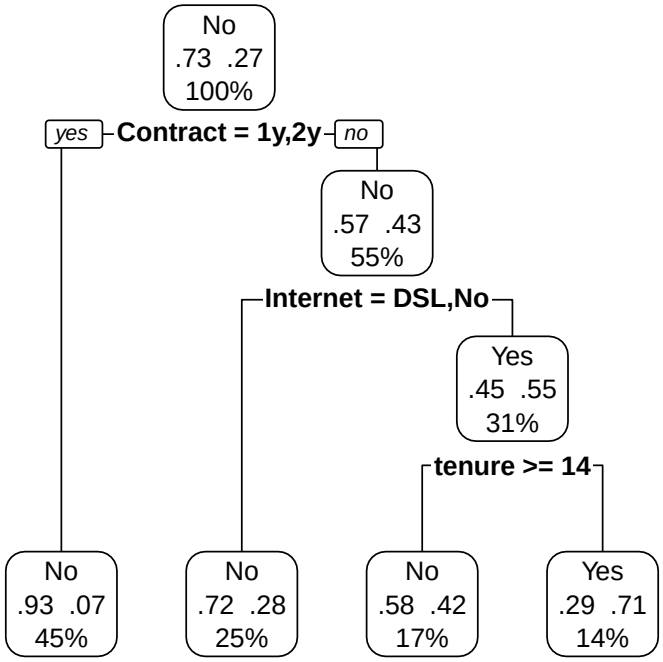

Figura 4: Gráfico del arbol de clasificación ordinal con cp=0.01.

A continuación se muestra la matriz de confusión resultante de aplicar el algoritmo de clasificación al conjunto de test. Como podemos ver, en este caso 993 clientes son clasificados correctamente como clientes que continúan con la empresa y 117 como desertores. Por otro lado 61 sujetos son clasificados como desertores cuando en realidad no lo son y 236 sujetos son clasificados como clientes fidedignos cuando en realidad no lo son. La precisión de este modelo es de 0.789, la sensibilidad toma el valor de 0.331 y la especificidad el valor de 0.942.

```
##       Predicted
## Actual  No Yes
##    No  993  61
##    Yes 236 117
```

Random Forest[3]

Finalmente aplicaremos el algoritmo de aprendizaje automático `RandomForest` con árboles de clasificación ordinaria, más complejo que los dos anteriores, con vistas de buscar consistencia con los resultados obtenidos anteriormente y hacernos una idea de la importancia de cada una de las variables.

La matriz de confusión surgida de evaluar el algoritmo de clasificación obtenido en el conjunto de test es se muestra a continuación. Como podemos ver, en este caso 964 clientes son clasificados correctamente como clientes que continúan con la empresa y 161 como desertores. Por otro lado 90 sujetos son clasificados como desertores cuando en realidad no lo son y 192 sujetos son clasificados como clientes fidedignos cuando en realidad no lo son. La precisión de este modelo es de 0.8, la sensibilidad toma el valor de 0.456 y la especificidad el valor de 0.915.

```
##       Predicted
## Actual  No Yes
##    No  964  90
##    Yes 192 161
```

Finalmente mostramos una tabla de las variables asociadas a un índice de importancia que reporta el algoritmo. En ella podemos ver que las variables más importantes son `tenure`, `MonthlyCharges` y `Contract`, consistente con los resultados anteriores salvo `MonthlyCharges` que no aparece en los modelos desarrollados anteriormente.

Cuadro 9: Tabla de importancia de las variables del modelo de RandomForest.

	MeanDecreaseGini
gender	53.77
SeniorCitizen	41.04
Partner	46.07
Dependents	40.17
tenure	421.7
PhoneService	9.679
MultipleLines	49.47
InternetService	78.11
OnlineSecurity	94.35
OnlineBackup	55.35
DeviceProtection	46.52
TechSupport	84.89
StreamingTV	41.27
StreamingMovies	40.27
Contract	180.1
PaperlessBilling	50.34
PaymentMethod	129.3
MonthlyCharges	369.2

[3]Kabacoff (2015)

Resultados

En vista de lo obtenido anteriormente para cada uno de los algoritmos de clasificación considerados, el más idóneo y elegido finalmente es el de árboles de clasificación inferenciales, debido a su buen balance sensibilidad-especificidad y la facilidad de interpretación. Por tanto, según el árbol de la figura 3 tenemos los siguientes perfiles:

- **Desertores:** El perfil de los desertores, al menos el que detectan los algoritmos empleados, ya que la especificidad de los modelos ha resultado ser baja, es claro. Los desertores son clientes que adquieren un contrato mensual de fibra óptica y cuya relación con Movistar no supera el año de duración, siendo crucial el primer mes de contrato, en el que muchos clientes abandonan quizá insatisfechos con el servicio de fibra óptica recibido.
- **No desertores:** El perfil de los clientes fidedignos es un poco variopinto, pero principalmente son clientes que superan el primer año de contrato con la empresa, ya sea via contrato anual o bianual o via contrato mensual con fibra óptica. Estos últimos clientes quizá dudosos en un principio con la calidad del servicio de fibra óptica, de ahí el contrato mensual, pero finalmente satisfechos. También son no desertores aquellos que contratan `DSL` o servicio telefónico de manera mensual habiendo permanecido en la empresa más de 5 meses.

Apartado 2

El siguiente objetivo es detectar agrupamientos dentro de los grupos ya preestablecidos de desertores y no desertores. El método que emplearemos para abordar este problema es un conocido algoritmo de clustering denominado `Partitioning around medoids (PAM)` del que se dispone en el paquete `cluster`.

Metodología

El algoritmo de clustering `Partitioning around medoids (PAM)`[4] nos proporciona como resultado K grupos donde K es un hiperparámetro que damos nosotros a priori y donde el centro de cada grupo es una observación del conjunto de datos. En los algoritmos de clustering es crucial el concepto de distancia, ya que su funcionamiento está basado completamente en la similitud o diferencia habida entre las observaciones del conjunto de datos. En nuestro caso usaremos la métrica de gower, útil en el caso en que el conjunto de datos está compuesto por una mixtura de variables categóricas y numéricas.

Por otro lado, para medir la calidad de cada modelo usaremos la media de los índices de Silhouette de todas las observaciones del conjunto de datos. Este índice se calcula para cada observación y mide cómo de bien ''cae" una observación en el cluster que le es asignado con un valor entre -1 y 1, siendo preferibles valores próximos a 1.

No desertores

En la figura 5 podemos ver un gráfico en el que se muestra la media de los índices de Silhouette frente al número de clusters o grupos creados. En vista de estos resultados el mejor modelo es aquel con dos grupos, esto es, $K = 2$.

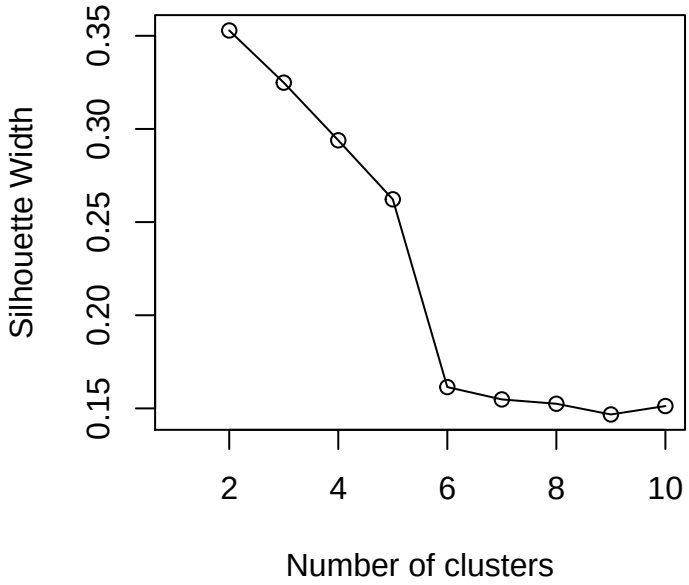

Figura 5: Gráfico del Silhouette width frente al número de clusters para los no desertores.

La tabla siguiente muestra cuantos clientes asocia el algoritmo `PAM` a cada uno de los dos grupos. Como podemos ver, el primer grupo consta de 3554 observaciones y el segundo de 1620 observaciones.

```
##
```

[4] Kabacoff (2015)

```
##    1    2
## 3554 1620
```

La figura 6 muestra una descripción gráfica de las características de cada uno de los dos grupos. Como se aprecia rápidamente, las variables referentes a los servicios de internet han dominado el agrupamiento, por lo que los dos grupos que el algoritmo ha distinguido principalmente son usuarios que tienen cualquier servicio de internet contratado y usuarios que no. Debido a la dominancia de estas variables, para ahondar un poco más en la estructura de los no desertores, nos planteamos dividir este mismo subgrupo en usuarios que contratan algún tipo de internet y usuarios que no y aplicar en estos dos subgrupos el algoritmo de clustering.

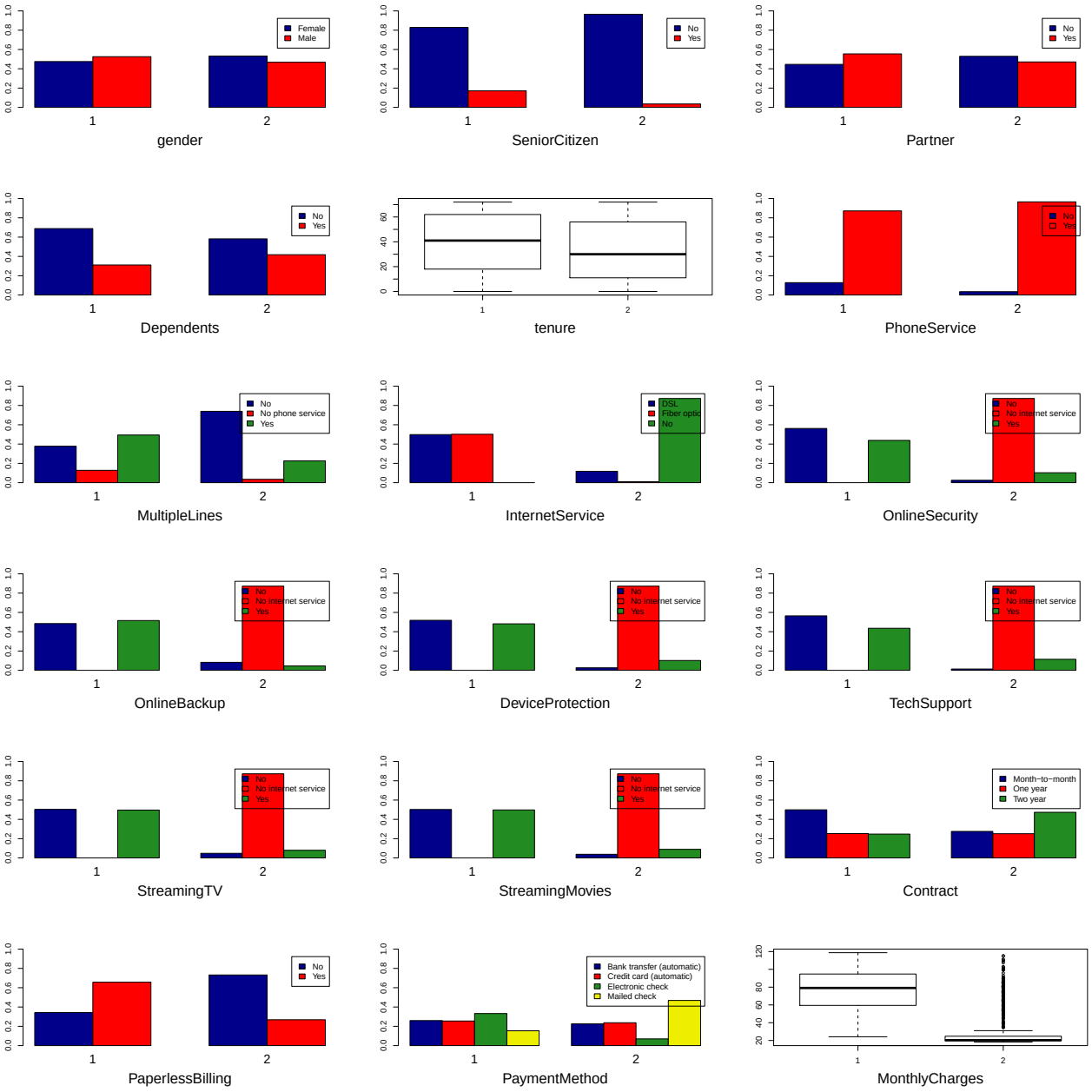

Figura 6: Gráfico descriptivo de las características de ambos grupos generados entre los no desertores.

No desertores que contratan algún servicio de internet

El procedimiento a realizar aquí es el mismo que el anterior. En la figura 7 se muestra la calidad de los distintos modelos posibles en función del silhouette width. De nuevo, el mejor modelo es aquel con K=2.

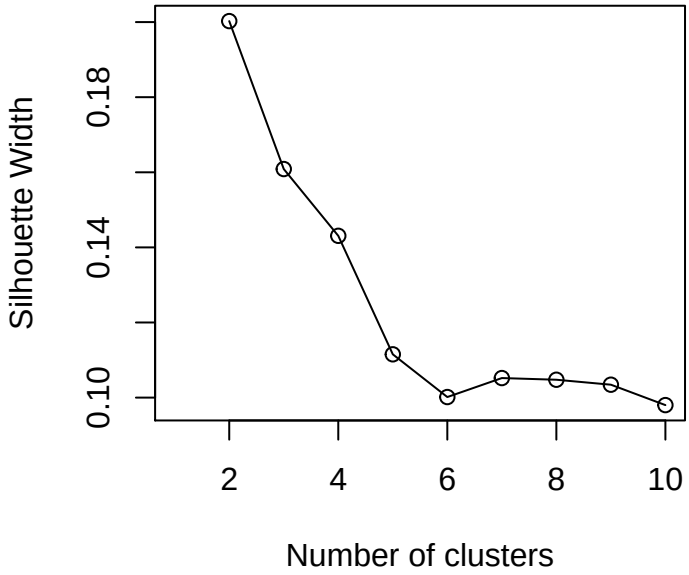

Figura 7: Gráfico del Silhouette width frente al número de clusters para los no desertores con internet.

La tabla siguiente muestra que 1978 observaciones forman el grupo 1 y 1783 observaciones forman el grupo 2.

```
##
##    1    2
## 1978 1783
```

Finalmente, en la figura 8, podemos ver gráficamente las diferencias en las características de ambos grupos. También hay similidaridades, como que casi todos contratan servicio telefónico, son clientes no jubilados y contratan de forma equilibrada tanto DSL como fibra óptica. También prefieren factura electrónica. Los dos perfiles que podemos extraer de aquí son:

- **Perfil 1**: El primer perfil está formado por personas solteras que llevan poco tiempo en Movistar (`tenure`) cuyo contrato es de tipo mensual y no contratan ninguno de los demás servicios asociados a internet. También es más frecuente en este grupo no haber contratado múltiples líneas de teléfono.
- **Perfil 2**: El segundo perfil es el cliente ideal. Está formado por clientes con pareja que contratan toda la cartera de servicios que ofrece Movistar y tienen, mayormente, un contrato bianual.

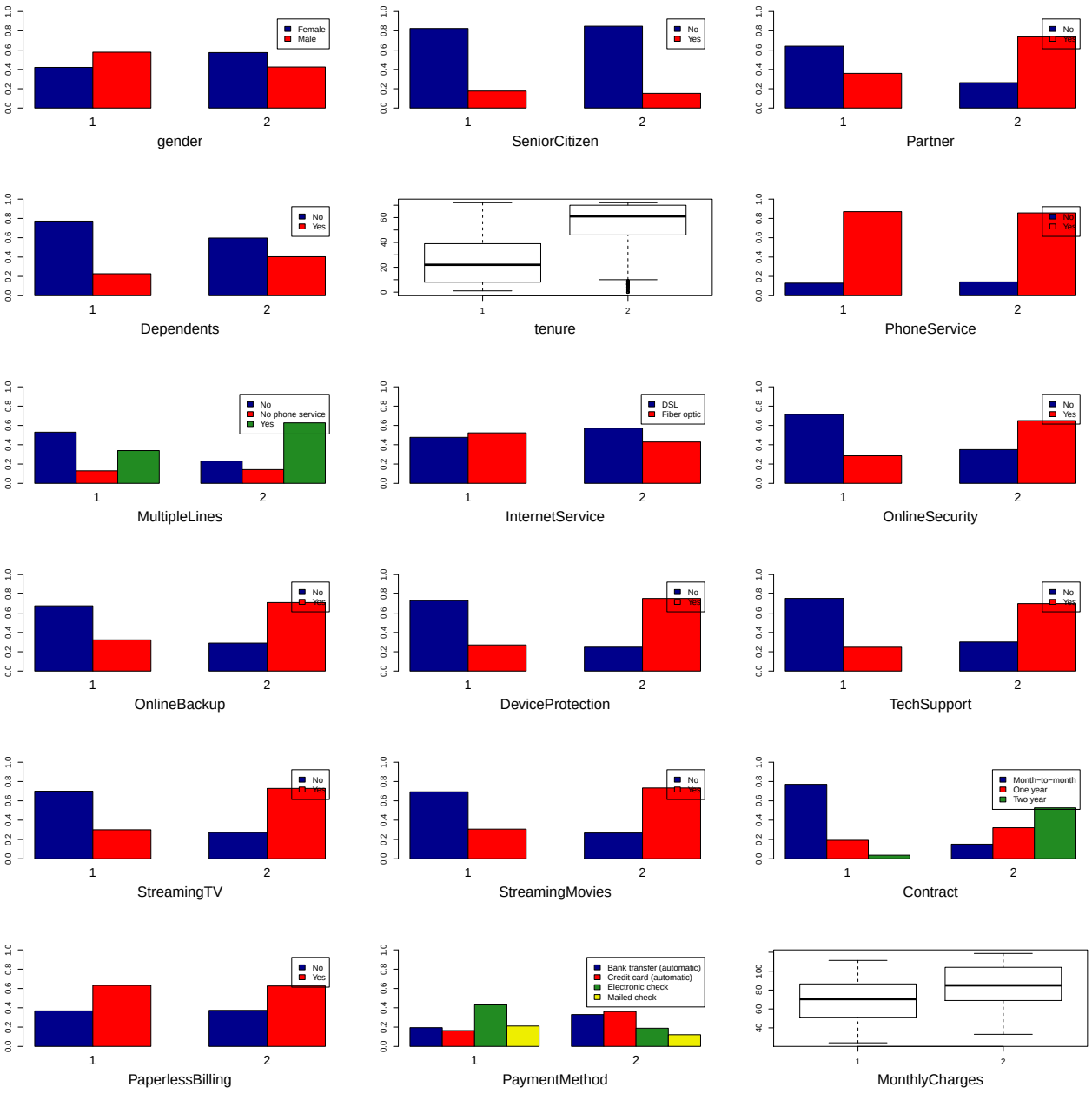

Figura 8: Gráfico descriptivo de las características de ambos grupos generados entre los no desertores con internet.

No desertores que no contratan ningún servicio de internet

En la figura 9 podemos ver el índice de silhouette, un indicador de validez del modelo, en función del número de cluster escogidos. De nuevo, el mejor número de grupos a escoger es dos.

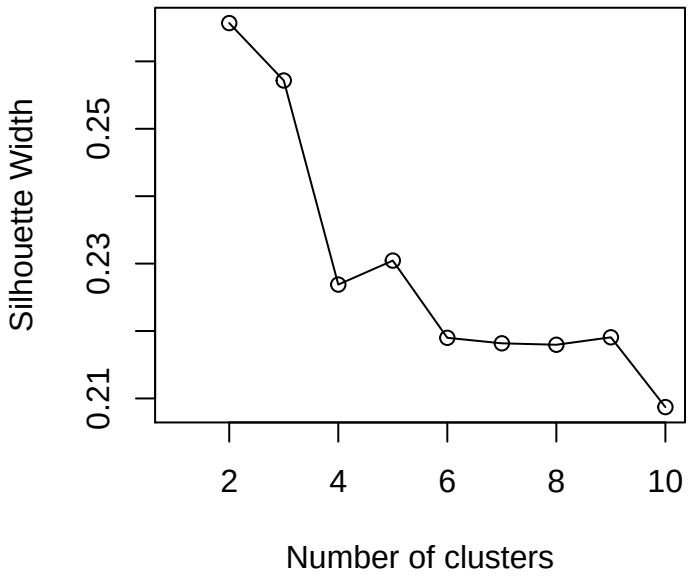

Figura 9: Gráfico del Silhouette width frente al número de clusters para los no desertores sin internet.

La tabla siguiente muestra que 692 observaciones forman el grupo 1 y 721 observaciones forman el grupo 2.

```
##
##   1   2
## 692 721
```

En la figura 10 se muestran gráficamente las características de los grupos generados. Podemos ver que ambos tienen en común son clientes no jubilados que contratan servicio telefónico, aunque no suelen contratar líneas múltiples, y que prefiere la factura en papel. Las diferencias entre los dos perfiles las discutimos a continuación:

- **Perfil 3**: En los clientes del tercer perfil encontramos usuarios mayormente hombres, solteros, con contrato de telefonía (con líneas múltiples poco frecuentes) y sin familiares al cargo. LLevan poco tiempo con Movistar, alrededor de un año de media, y el tipo de contrato que mantienen es mensual.
- **Perfil 4**: En los clientes que conforman el cuarto perfil encontramos usuarios mayormente mujeres, con pareja, con contrato de telefonía y con familiares al cargo, de ahí que en este grupo sea algo más notable el contrato de líneas múltiples, pero sigue sin ser mayoritario. Además son clientes que llevan mucho tiempo con Movistar (más de dos años) y recurre, la gran mayoría, a un contrato bianual.

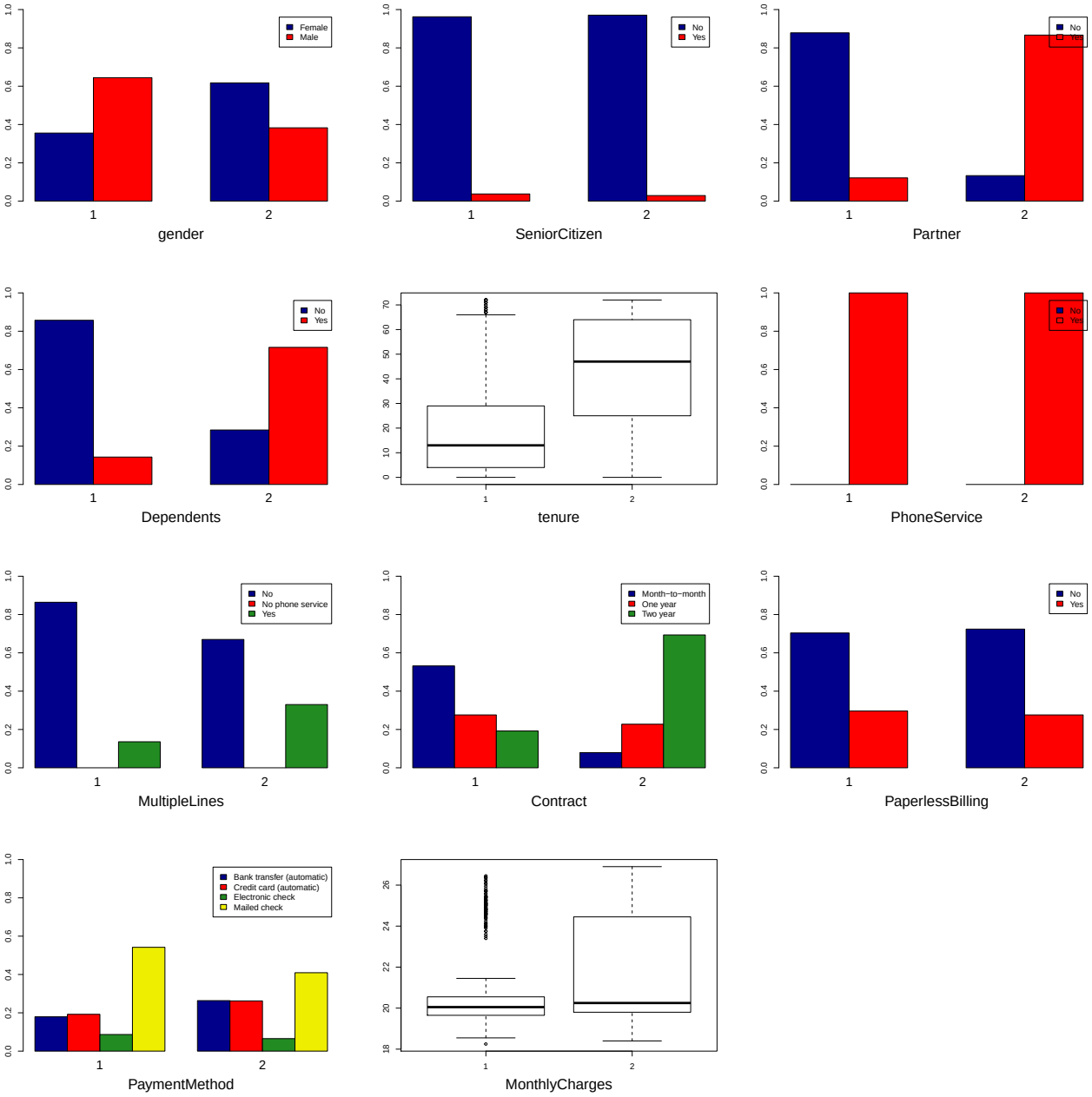

Figura 10: Gráfico descriptivo de las características de ambos grupos generados entre los no desertores sin internet.

Desertores

En la figura 11 se muestra cómo de válido es cada uno de los modelos considerados en función del número de clusters elegidos utilizando como criterio el tamaño de silhouette, que no es más que la media de estos índices sobre todo el conjunto de datos. En este caso, quizá por el número distinto de observaciones, la variable `InternetService` y las asociadas a esta no dominan sobre las demás, obteniendo así como mejor resultado, con diferencia, un modelo con 3 clusters.

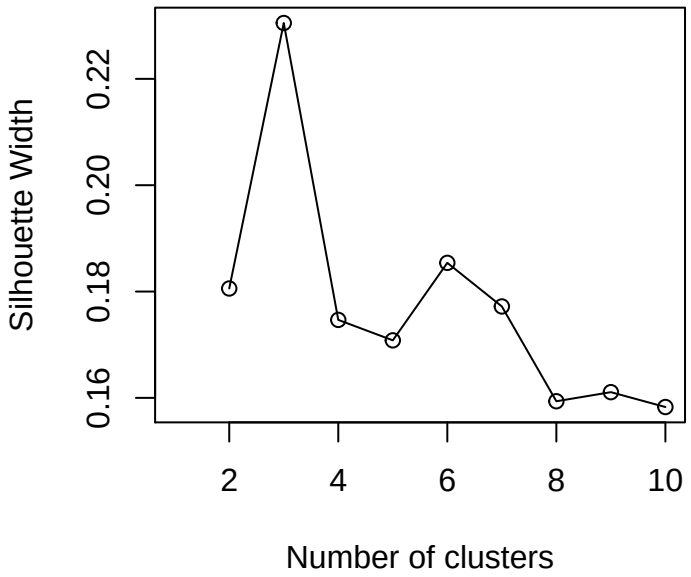

Figura 11: Gráfico del Silhouette width frente al número de clusters para los desertores.

La tabla siguiente muestra que 1001 observaciones forman el grupo 1 y 752 observaciones forman el grupo 2 y 116 observaciones forman el grupo 3.

```
##
##    1    2    3
## 1001  752  116
```

La figura 12 muestra gráficamente las distintas características de los grupos generados por el algoritmo de clutering `PAM` para $K = 3$. Los perfiles que podemos deducir de esta figura son los siguientes:

- **Perfil 5:** Los clientes que conforman en quinto perfil son clientes mayoritariamente solteros sin parientes a su cargo. Las mujeres son ligeramente más numerosas que los hombres en este grupo. Llevan poco tiempo en la empresa y no suelen contratar líneas múltiples de telefonía. Todos ellos contratan servicio de internet y prefieren fibra óptica, aunque también hay un grupo numeroso de clientes que contratan DSL. Son clientes que asumen un gasto mensual moderado, prefieren la factura electrónica y no contratan ninguno de los demás servicios asociados a internet. Todos tienen, además, un contrato de tipo mensual y el método de pago preferido es Electronic Check.
- **Perfil 6:** En el perfil 6 encontramos clientes mayoritariamente con pareja sin dependientes al cargo que llevan una media de dos años. Los hombres son ligeramente más numerosos que las mujeres en este grupo. Se decantan, sin duda, por la fibra óptica y contratan líneas múltiples. No suelen contratar servicios de internet asociados al soporte técnico o la seguridad pero sí contratan televisión y películas en streaming. Asumen un gasto mensual muy alto y prefieren la factura electrónica pagada siempre mediante Electronic check.

- **Perfil 7:** Este perfil es el de los que no contratan el servicio de internet. Son clientes que llevan poco tiempo con Movistar y asumen un gasto mensual bajo. Son personas solteras, sin dependientes al cargo que prefiren la factura en papel pagada mediante Mailed check.

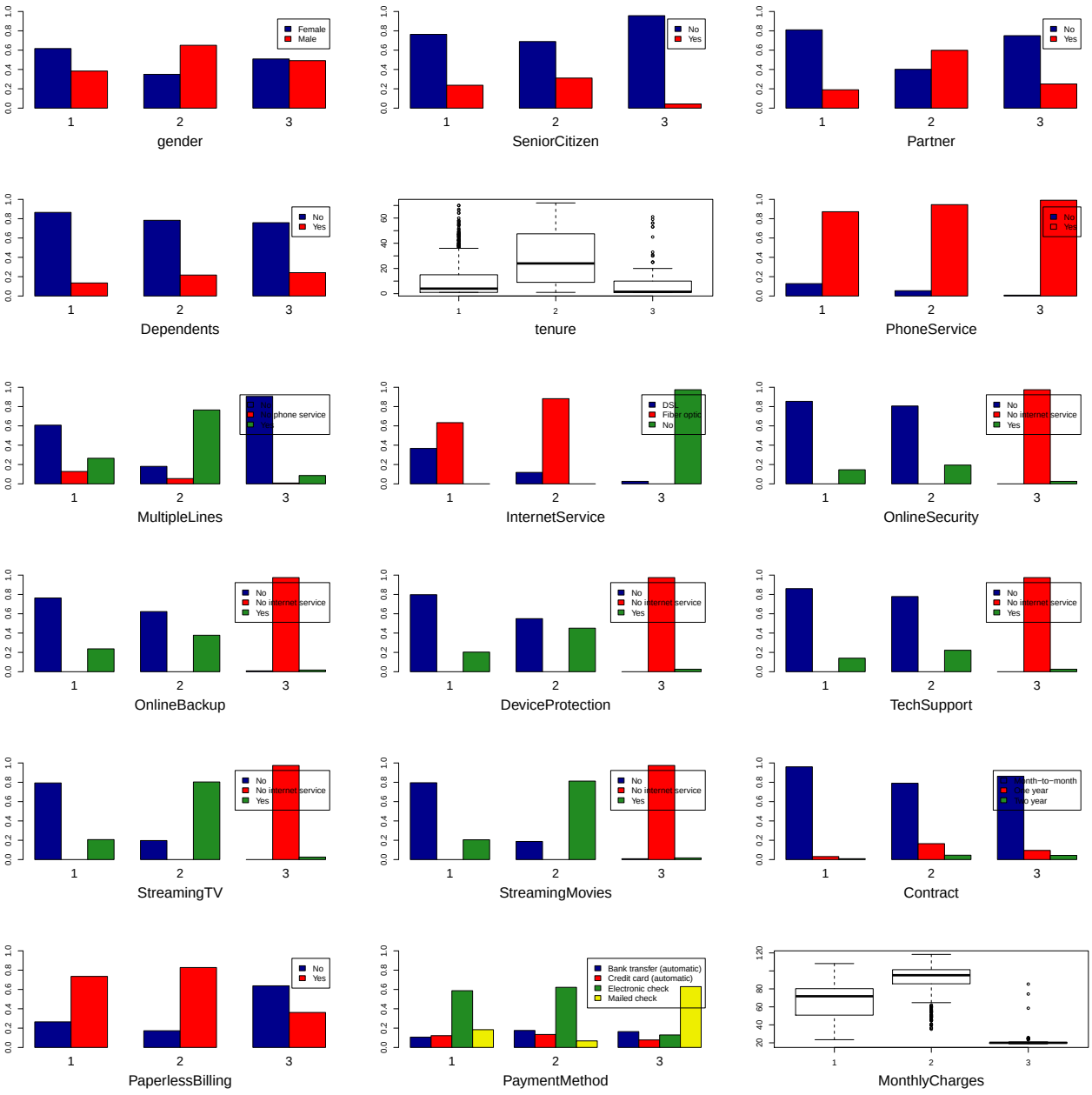

Figura 12: Gráfico descriptivo de las características de los tres grupos generados entre los desertores.

Resultados

A continuación se muestran los siete perfiles obtenidos así como su características.

No desertores

Con servicio internet

- **Perfil 1:**(n=1978) El primer perfil está formado por personas solteras que llevan poco tiempo en Movistar (`tenure`) cuyo contrato es de tipo mensual y no contratan ninguno de los demás servicios asociados a internet. También es más frecuente en este grupo no haber contratado múltiples líneas de teléfono. No tienen una preferencia clara sobre DSL o fibra óptica y sí prefieren factura electrónica.
- **Perfil 2:**(n=1783) El segundo perfil es el cliente ideal. Está formado por clientes con pareja que contratan toda la cartera de servicios que ofrece Movistar y tienen, mayormente, un contrato bianual. No tienen una preferencia clara sobre DSL (aunque este es más frecuente) o fibra óptica y sí prefieren factura electrónica.

Sin servicio a internet

- **Perfil 3:**(n=692) En los clientes del tercer perfil encontramos usuarios mayormente hombres, solteros, con contrato de telefonía (con líneas múltiples poco frecuentes) y sin familiares al cargo. LLevan poco tiempo con Movistar, alrededor de un año de media, y el tipo de contrato que mantienen es mensual. Prefieren factura en papel.
- **Perfil 4:**(n=721) En los clientes que conforman el cuarto perfil encontramos usuarios mayormente mujeres, con pareja, con contrato de telefonía y con familiares al cargo, de ahí que en este grupo sea algo más notable el contrato de líneas múltiples, pero sigue sin ser mayoritario. Además son clientes que llevan mucho tiempo con Movistar (más de dos años) y recurre, la gran mayoría, a un contrato bianual. Prefieren factura en papel.

Desertores

- **Perfil 5:**(n=1001) Los clientes que conforman en quinto perfil son clientes mayoritariamente solteros sin parientes a su cargo. Las mujeres son ligeramente más numerosas que los hombres en este grupo. Llevan poco tiempo en la empresa y no suelen contratar líneas múltiples de telefonía. Todos ellos contratan servicio de internet y prefieren fibra óptica, aunque también hay un grupo numeroso de clientes que contratan DSL. Son clientes que asumen un gasto mensual moderado, prefieren la factura electrónica y no contratan ninguno de los demás servicios asociados a internet. Todos tienen, además, un contrato de tipo mensual y el método de pago preferido es Electronic Check.
- **Perfil 6:**(n=752) En el perfil 6 encontramos clientes mayoritariamente con pareja sin dependientes al cargo que llevan una media de dos años. Los hombres son ligeramente más numerosos que las mujeres en este grupo. Se decantan, sin duda, por la fibra óptica y contratan líneas múltiples. No suelen contratar servicios de internet asociados al soporte técnico o la seguridad pero sí contratan televisión y películas en streaming. Asumen un gasto mensual muy alto y prefieren la factura electrónica pagada siempre mediante Electronic check.
- **Perfil 7:**(n=116) Este perfil es el de los que no contratan el servicio de internet. Son clientes que llevan poco tiempo con Movistar y asumen un gasto mensual bajo. Son personas solteras, sin dependientes al cargo que prefiren la factura en papel pagada mediante Mailed check.

Apartado 3

El último objetivo es diseñar una oferta personalizada a cada uno de los grupos obtenidos anteriormente estimando el coste que supondría dicha campaña. Para ello, en primero lugar, debemos estimar los precios de cada uno de los servicios ofrecidos.

Estimación del precio de los servicios

Para estimar el precio de cada uno de los servicios ofrecidos por Movistar tan solo es necesario desarrollar un modelo lineal en el que la variable dependiente es `MontlyCharges` y las variables independientes o predictoras son cada uno de los servicios que ofrece Movistar, tomando el valor de 1 si el cliente contrata este servicio y 0 si no. Cada uno de los coeficientes del modelo será, entonces, el precio del servicio asociado.

Los coeficientes del modelo lineal calculado se muestran en la siguiente tabla. Como podemos ver, todos son estadísticamente significativos, salvo el término independiente, el cual interesa que sea cero. El modelo explica casi el total de la varianza de los datos, $r^2 = 0{,}99$ lo que concuerda con nuestra hipótesis. De este modo, podemos tomar estos coeficientes como precios aproximados de los servicios.

```
##
## Call:
## lm(formula = MonthlyCharges ~ ., data = gastos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2527 -0.6132 -0.0059  0.6003  4.7973
##
## Coefficients:
##                     Estimate Std. Error t value            Pr(>|t|)
## (Intercept)         -0.09653    0.05364    -1.8              0.0719 .
## PhoneServiceYes     20.05241    0.04783   419.2 <0.0000000000000002 ***
## MultipleLinesYes     5.01220    0.02815   178.0 <0.0000000000000002 ***
## OnlineSecurityYes    5.01319    0.03083   162.6 <0.0000000000000002 ***
## OnlineBackupYes      4.99131    0.02915   171.3 <0.0000000000000002 ***
## DeviceProtectionYes  5.02077    0.03046   164.8 <0.0000000000000002 ***
## TechSupportYes       5.02892    0.03148   159.8 <0.0000000000000002 ***
## StreamingTVYes       9.96866    0.03171   314.4 <0.0000000000000002 ***
## StreamingMoviesYes   9.96270    0.03175   313.8 <0.0000000000000002 ***
## FiberOpticYes       50.00642    0.03924  1274.4 <0.0000000000000002 ***
## DSLYes              25.04904    0.04261   587.9 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.026 on 7032 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9988
## F-statistic: 6.055e+05 on 10 and 7032 DF,  p-value: < 0.00000000000000022
```

Campaña de incentivos

El coste será estimado en términos de pérdidas, esto es, en términos generales,

$$Coste = n \sum_{i}^{10} \theta(i) P_i(D_i) t_i$$

donde $\theta(i)$ vale 1 o 0 dependiendo de si ese servicio es ofertado o no, P_i es el precio del servicio i, D_i es el descuento aplicado al servicio i, t_i el tiempo de aplicación del descuento y n el número de clientes en el grupo.

Ofertas por grupos

No desertores

Con servicio internet

- **Perfil 1:**(n=1978) El objetivo para este grupo sería conseguir un contrato anual o bianual, ya que los clientes con este tipo de contrato tienen una probabilidad de deserción menor. Como son personas solteras no tiene sentido ofrecerles líneas múltiples, así que lo más idóneo sería, por ejemplo, ofrecer televisión y películas en streaming por 5 € mensuales durante seis meses a cambio de un contrato anual Coste: 177205.39 €
- **Perfil 2:**(n=1783) El segundo perfil era el cliente ideal. Era aquel que tenía todo contratado con contratos mayormente anuales y bianuales. Aunque es muy poco probable perder este cliente también hay que cuidarlo. Una buena oferta sería fibra óptica al precio de DSL durante seis para clientes que tengan contratado DSL como premio a su fidelidad. Coste 160196.38 €

Sin servicio a internet

- **Perfil 3:**(n=692) Ya que este grupo son clientes solteros que llevan poco tiempo con nosotros, y en vista de que la gente que contrata fibra óptica llevando poco tiempo con Movistar suele romper el contrato, una buena oferta para este grupo sería DSL por 10€ mensuales durante seis meses con la condición de pasar a contrato anual. Coste: 62483.63 €
- **Perfil 4:**(n=721) Ya que este grupo de clientes suelen ser clientes con pareja y familiares al cargo, además de clientes con contratos bianuales, una buena oferta para premiar su fidelidad y que podrían compartir en familia sería fibra óptica, películas y televisión en streaming al 85 % durante un año. Coste: 90765.25 €

Desertores

- **Perfil 5:**(n=1001) Estos clientes tenían contratado tanto internet como líneas móviles. Una buena oferta para retenerlos en la compañía podría haber sido televisión y películas en streaming gratis durante 6 meses a cambio de un contrato anual. Coste: 119707.75 €
- **Perfil 6:**(n=752) Estos clientes tienen prácticamente todo contratado, pero de forma mensual. Una opción para retenerlos podría ser ofertarles un descuento del 30 % en su tarifa mensual actual durante 6 meses a cambio de un contrato anual. Coste: 142147.6 €
- **Perfil 7:**(n=116) Este era el perfil de aquellas personas con contrato mensual sin internet. Una buena oferta que quizá podría retenerlos en la compañía es buen internet a precio inmejorable. Fibra óptica al 50 % durante seis meses a cambio de un contrato anual. Coste: 17402.23 €

Coste Total: 769908.23 €

Bibliografía

Kabacoff, Robert I. 2015. *R in Action: Data Analysis and Graphics with R Second Edition*. MANNING.