Actividad integradora 2

Jacobo Hirsch Rodriguez

2024-11-19

Bibliotecas

```
# Cargamos todas las librería en la lista "librerias"
librerias = c('tidyverse','broom','ISLR','GGally','modelr','cowplot','rlang','modelr','tibble','Metrics

for (lib in librerias){
  library(lib,character.only=TRUE)}
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
##
##
## Attaching package: 'modelr'
##
##
## The following object is masked from 'package:broom':
##
##     bootstrap
##
##
##
## Attaching package: 'cowplot'
##
##
## The following object is masked from 'package:lubridate':
##
##     stamp
##
##
```

```
##
## Attaching package: 'rlang'
##
##
## The following objects are masked from 'package:purrr':
##
##     %@%, flatten, flatten_chr, flatten_dbl, flatten_int, flatten_lgl,
##     flatten_raw, invoke, splice
##
##
##
## Attaching package: 'Metrics'
##
##
## The following object is masked from 'package:rlang':
##
##     ll
##
##
## The following objects are masked from 'package:modelr':
##
##     mae, mape, mse, rmse
##
##
##
## Attaching package: 'mice'
##
##
## The following object is masked from 'package:stats':
##
##     filter
##
##
## The following objects are masked from 'package:base':
##
##     cbind, rbind
##
##
## Loading required package: lattice
##
##
## Attaching package: 'caret'
##
##
## The following objects are masked from 'package:Metrics':
##
##     precision, recall
##
##
## The following object is masked from 'package:purrr':
##
##     lift
```

Leyendo los datos para el dataset que usaremos para training y validacion:

```
M = read.csv("Titanic.csv")
str(M)
```

```
## 'data.frame':    1309 obs. of  12 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Survived   : int  0 1 0 0 1 0 1 0 1 0 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Mr. Thomas Francis
##  $ Sex        : chr  "male" "female" "male" "male" ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : chr  "" "" "" "" ...
##  $ Embarked   : chr  "Q" "S" "Q" "S" ...
```

leyendo los datos para el testing

```
M_test = read.csv("Titanic_test.csv")
M_test
```

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
892	3	Kelly, Mr. James	male	34.50	0	0	330911	7.8292		Q
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.00	1	0	363272	7.0000		S
894	2	Myles, Mr. Thomas Francis	male	62.00	0	0	240276	9.6875		Q
895	3	Wirz, Mr. Albert	male	27.00	0	0	315154	8.6625		S
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.00	1	1	3101298	12.2875		S
897	3	Svensson, Mr. Johan Cervin	male	14.00	0	0	7538	9.2250		S
898	3	Connolly, Miss. Kate	female	30.00	0	0	330972	7.6292		Q
899	2	Caldwell, Mr. Albert Francis	male	26.00	1	1	248738	29.0000		S
900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18.00	0	0	2657	7.2292		C
901	3	Davies, Mr. John Samuel	male	21.00	2	0	A/4 48871	24.1500		S
902	3	Ilieff, Mr. Ylio	male	NA	0	0	349220	7.8958		S
903	1	Jones, Mr. Charles Cresson	male	46.00	0	0	694	26.0000		S
904	1	Snyder, Mrs. John Pillsbury (Nelle Stevenson)	female	23.00	1	0	21228	82.2667	B45	S
905	2	Howard, Mr. Benjamin	male	63.00	1	0	24065	26.0000		S
906	1	Chaffee, Mrs. Herbert Fuller (Carrie Constance Toogood)	female	47.00	1	0	W.E.P. 5734	61.1750	E31	S
907	2	del Carlo, Mrs. Sebastiano (Argenia Genovesi)	female	24.00	1	0	SC/PARIS 2167	27.7208		C
908	2	Keane, Mr. Daniel	male	35.00	0	0	233734	12.3500		Q

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
909	3	Assaf, Mr. Gerios	male	21.00	0	0	2692	7.2250		C
910	3	Ilmakangas, Miss. Ida Livija	female	27.00	1	0	STON/O2. 3101270	7.9250		S
911	3	Assaf Khalil, Mrs. Mariana (Miriam")"	female	45.00	0	0	2696	7.2250		C
912	1	Rothschild, Mr. Martin	male	55.00	1	0	PC 17603	59.4000		C
913	3	Olsen, Master. Artur Karl	male	9.00	0	1	C 17368	3.1708		S
914	1	Flegenheim, Mrs. Alfred (Antoinette)	female	NA	0	0	PC 17598	31.6833		S
915	1	Williams, Mr. Richard Norris II	male	21.00	0	1	PC 17597	61.3792		C
916	1	Ryerson, Mrs. Arthur Larned (Emily Maria Borie)	female	48.00	1	3	PC 17608	262.375	B57 B59 B63 B66	C
917	3	Robins, Mr. Alexander A	male	50.00	1	0	A/5. 3337	14.5000		S
918	1	Ostby, Miss. Helene Ragnhild	female	22.00	0	1	113509	61.9792	B36	C
919	3	Daher, Mr. Shedid	male	22.50	0	0	2698	7.2250		C
920	1	Brady, Mr. John Bertram	male	41.00	0	0	113054	30.5000	A21	S
921	3	Samaan, Mr. Elias	male	NA	2	0	2662	21.6792		C
922	2	Louch, Mr. Charles Alexander	male	50.00	1	0	SC/AH 3085	26.0000		S
923	2	Jefferys, Mr. Clifford Thomas	male	24.00	2	0	C.A. 31029	31.5000		S
924	3	Dean, Mrs. Bertram (Eva Georgetta Light)	female	33.00	1	2	C.A. 2315	20.5750		S
925	3	Johnston, Mrs. Andrew G (Elizabeth Lily" Watson)"	female	NA	1	2	W./C. 6607	23.4500		S
926	1	Mock, Mr. Philipp Edmund	male	30.00	1	0	13236	57.7500	C78	C
927	3	Katavelas, Mr. Vassilios (Catavelas Vassilios")"	male	18.50	0	0	2682	7.2292		C
928	3	Roth, Miss. Sarah A	female	NA	0	0	342712	8.0500		S
929	3	Cacic, Miss. Manda	female	21.00	0	0	315087	8.6625		S
930	3	Sap, Mr. Julius	male	25.00	0	0	345768	9.5000		S
931	3	Hee, Mr. Ling	male	NA	0	0	1601	56.4958		S
932	3	Karun, Mr. Franz	male	39.00	0	1	349256	13.4167		C
933	1	Franklin, Mr. Thomas Parham	male	NA	0	0	113778	26.5500	D34	S
934	3	Goldsmith, Mr. Nathan	male	41.00	0	0	SOTON/O.Q. 3101263	7.8500		S
935	2	Corbett, Mrs. Walter H (Irene Colvin)	female	30.00	0	0	237249	13.0000		S
936	1	Kimball, Mrs. Edwin Nelson Jr (Gertrude Parsons)	female	45.00	1	0	11753	52.5542	D19	S
937	3	Peltomaki, Mr. Nikolai Johannes	male	25.00	0	0	STON/O 2. 3101291	7.9250		S
938	1	Chevre, Mr. Paul Romaine	male	45.00	0	0	PC 17594	29.7000	A9	C
939	3	Shaughnessy, Mr. Patrick	male	NA	0	0	370374	7.7500		Q

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
940	1	Bucknell, Mrs. William Robert (Emma Eliza Ward)	female	60.00	0	0	11813	76.2917	D15	C
941	3	Coutts, Mrs. William (Winnie Minnie" Treanor)"	female	36.00	0	2	C.A. 37671	15.9000		S
942	1	Smith, Mr. Lucien Philip	male	24.00	1	0	13695	60.0000	C31	S
943	2	Pulbaum, Mr. Franz	male	27.00	0	0	SC/PARIS 2168	15.0333		C
944	2	Hocking, Miss. Ellen Nellie" "	female	20.00	2	1	29105	23.0000		S
945	1	Fortune, Miss. Ethel Flora	female	28.00	3	2	19950	263.0000	C23 C25 C27	S
946	2	Mangiavacchi, Mr. Serafino Emilio	male	NA	0	0	SC/A.3 2861	15.5792		C
947	3	Rice, Master. Albert	male	10.00	4	1	382652	29.1250		Q
948	3	Cor, Mr. Bartol	male	35.00	0	0	349230	7.8958		S
949	3	Abelseth, Mr. Olaus Jorgensen	male	25.00	0	0	348122	7.6500	F G63	S
950	3	Davison, Mr. Thomas Henry	male	NA	1	0	386525	16.1000		S
951	1	Chaudanson, Miss. Victorine	female	36.00	0	0	PC 17608	262.3750	B61	C
952	3	Dika, Mr. Mirko	male	17.00	0	0	349232	7.8958		S
953	2	McCrae, Mr. Arthur Gordon	male	32.00	0	0	237216	13.5000		S
954	3	Bjorklund, Mr. Ernst Herbert	male	18.00	0	0	347090	7.7500		S
955	3	Bradley, Miss. Bridget Delia	female	22.00	0	0	334914	7.7250		Q
956	1	Ryerson, Master. John Borie	male	13.00	2	2	PC 17608	262.3750	B57 B59 B63 B66	C
957	2	Corey, Mrs. Percy C (Mary Phyllis Elizabeth Miller)	female	NA	0	0	F.C.C. 13534	21.0000		S
958	3	Burns, Miss. Mary Delia	female	18.00	0	0	330963	7.8792		Q
959	1	Moore, Mr. Clarence Bloomfield	male	47.00	0	0	113796	42.4000		S
960	1	Tucker, Mr. Gilbert Milligan Jr	male	31.00	0	0	2543	28.5375	C53	C
961	1	Fortune, Mrs. Mark (Mary McDougald)	female	60.00	1	4	19950	263.0000	C23 C25 C27	S
962	3	Mulvihill, Miss. Bertha E	female	24.00	0	0	382653	7.7500		Q
963	3	Minkoff, Mr. Lazar	male	21.00	0	0	349211	7.8958		S
964	3	Nieminen, Miss. Manta Josefina	female	29.00	0	0	3101297	7.9250		S
965	1	Ovies y Rodriguez, Mr. Servando	male	28.50	0	0	PC 17562	27.7208	D43	C
966	1	Geiger, Miss. Amalie	female	35.00	0	0	113503	211.5000	C130	C
967	1	Keeping, Mr. Edwin	male	32.50	0	0	113503	211.5000	C132	C
968	3	Miles, Mr. Frank	male	NA	0	0	359306	8.0500		S
969	1	Cornell, Mrs. Robert Clifford (Malvina Helen Lamson)	female	55.00	2	0	11770	25.7000	C101	S
970	2	Aldworth, Mr. Charles Augustus	male	30.00	0	0	248744	13.0000		S
971	3	Doyle, Miss. Elizabeth	female	24.00	0	0	368702	7.7500		Q
972	3	Boulos, Master. Akar	male	6.00	1	1	2678	15.2458		C

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
973	1	Straus, Mr. Isidor	male	67.00	1	0	PC 17483	221.7792	C55 C57	S
974	1	Case, Mr. Howard Brown	male	49.00	0	0	19924	26.0000		S
975	3	Demetri, Mr. Marinko	male	NA	0	0	349238	7.8958		S
976	2	Lamb, Mr. John Joseph	male	NA	0	0	240261	10.7083		Q
977	3	Khalil, Mr. Betros	male	NA	1	0	2660	14.4542		C
978	3	Barry, Miss. Julia	female	27.00	0	0	330844	7.8792		Q
979	3	Badman, Miss. Emily Louisa	female	18.00	0	0	A/4 31416	8.0500		S
980	3	O'Donoghue, Ms. Bridget	female	NA	0	0	364856	7.7500		Q
981	2	Wells, Master. Ralph Lester	male	2.00	1	1	29103	23.0000		S
982	3	Dyker, Mrs. Adolf Fredrik (Anna Elisabeth Judith Andersson)	female	22.00	1	0	347072	13.9000		S
983	3	Pedersen, Mr. Olaf	male	NA	0	0	345498	7.7750		S
984	1	Davidson, Mrs. Thornton (Orian Hays)	female	27.00	1	2	F.C. 12750	52.0000	B71	S
985	3	Guest, Mr. Robert	male	NA	0	0	376563	8.0500		S
986	1	Birnbaum, Mr. Jakob	male	25.00	0	0	13905	26.0000		C
987	3	Tenglin, Mr. Gunnar Isidor	male	25.00	0	0	350033	7.7958		S
988	1	Cavendish, Mrs. Tyrell William (Julia Florence Siegel)	female	76.00	1	0	19877	78.8500	C46	S
989	3	Makinen, Mr. Kalle Edvard	male	29.00	0	0	STON/O 2. 3101268	7.9250		S
990	3	Braf, Miss. Elin Ester Maria	female	20.00	0	0	347471	7.8542		S
991	3	Nancarrow, Mr. William Henry	male	33.00	0	0	A./5. 3338	8.0500		S
992	1	Stengel, Mrs. Charles Emil Henry (Annie May Morris)	female	43.00	1	0	11778	55.4417	C116	C
993	2	Weisz, Mr. Leopold	male	27.00	1	0	228414	26.0000		S
994	3	Foley, Mr. William	male	NA	0	0	365235	7.7500		Q
995	3	Johansson Palmquist, Mr. Oskar Leander	male	26.00	0	0	347070	7.7750		S
996	3	Thomas, Mrs. Alexander (Thamine Thelma")"	female	16.00	1	1	2625	8.5167		C
997	3	Holthen, Mr. Johan Martin	male	28.00	0	0	C 4001	22.5250		S
998	3	Buckley, Mr. Daniel	male	21.00	0	0	330920	7.8208		Q
999	3	Ryan, Mr. Edward	male	NA	0	0	383162	7.7500		Q
1000	3	Willer, Mr. Aaron (Abi Weller")"	male	NA	0	0	3410	8.7125		S
1001	2	Swane, Mr. George	male	18.50	0	0	248734	13.0000	F	S
1002	2	Stanton, Mr. Samuel Ward	male	41.00	0	0	237734	15.0458		C
1003	3	Shine, Miss. Ellen Natalia	female	NA	0	0	330968	7.7792		Q
1004	1	Evans, Miss. Edith Corse	female	36.00	0	0	PC 17531	31.6792	A29	C
1005	3	Buckley, Miss. Katherine	female	18.50	0	0	329944	7.2833		Q
1006	1	Straus, Mrs. Isidor (Rosalie Ida Blun)	female	63.00	1	0	PC 17483	221.7792	C55 C57	S
1007	3	Chronopoulos, Mr. Demetrios	male	18.00	1	0	2680	14.4542		C
1008	3	Thomas, Mr. John	male	NA	0	0	2681	6.4375		C

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1009	3	Sandstrom, Miss. Beatrice Irene	female	1.00	1	1	PP 9549	16.7000	G6	S
1010	1	Beattie, Mr. Thomson	male	36.00	0	0	13050	75.2417	C6	C
1011	2	Chapman, Mrs. John Henry (Sara Elizabeth Lawry)	female	29.00	1	0	SC/AH 29037	26.0000		S
1012	2	Watt, Miss. Bertha J	female	12.00	0	0	C.A. 33595	15.7500		S
1013	3	Kiernan, Mr. John	male	NA	1	0	367227	7.7500		Q
1014	1	Schabert, Mrs. Paul (Emma Mock)	female	35.00	1	0	13236	57.7500	C28	C
1015	3	Carver, Mr. Alfred John	male	28.00	0	0	392095	7.2500		S
1016	3	Kennedy, Mr. John	male	NA	0	0	368783	7.7500		Q
1017	3	Cribb, Miss. Laura Alice	female	17.00	0	1	371362	16.1000		S
1018	3	Brobeck, Mr. Karl Rudolf	male	22.00	0	0	350045	7.7958		S
1019	3	McCoy, Miss. Alicia	female	NA	2	0	367226	23.2500		Q
1020	2	Bowenur, Mr. Solomon	male	42.00	0	0	211535	13.0000		S
1021	3	Petersen, Mr. Marius	male	24.00	0	0	342441	8.0500		S
1022	3	Spinner, Mr. Henry John	male	32.00	0	0	STON/OQ. 369943	8.0500		S
1023	1	Gracie, Col. Archibald IV	male	53.00	0	0	113780	28.5000	C51	C
1024	3	Lefebre, Mrs. Frank (Frances)	female	NA	0	4	4133	25.4667		S
1025	3	Thomas, Mr. Charles P	male	NA	1	0	2621	6.4375		C
1026	3	Dintcheff, Mr. Valtcho	male	43.00	0	0	349226	7.8958		S
1027	3	Carlsson, Mr. Carl Robert	male	24.00	0	0	350409	7.8542		S
1028	3	Zakarian, Mr. Mapriededer	male	26.50	0	0	2656	7.2250		C
1029	2	Schmidt, Mr. August	male	26.00	0	0	248659	13.0000		S
1030	3	Drapkin, Miss. Jennie	female	23.00	0	0	SOTON/OQ 392083	8.0500		S
1031	3	Goodwin, Mr. Charles Frederick	male	40.00	1	6	CA 2144	46.9000		S
1032	3	Goodwin, Miss. Jessie Allis	female	10.00	5	2	CA 2144	46.9000		S
1033	1	Daniels, Miss. Sarah	female	33.00	0	0	113781	151.5500		S
1034	1	Ryerson, Mr. Arthur Larned	male	61.00	1	3	PC 17608	262.3750	B57 B59 B63 B66	C
1035	2	Beauchamp, Mr. Henry James	male	28.00	0	0	244358	26.0000		S
1036	1	Lindeberg-Lind, Mr. Erik Gustaf (Mr Edward Lingrey")"	male	42.00	0	0	17475	26.5500		S
1037	3	Vander Planke, Mr. Julius	male	31.00	3	0	345763	18.0000		S
1038	1	Hilliard, Mr. Herbert Henry	male	NA	0	0	17463	51.8625	E46	S
1039	3	Davies, Mr. Evan	male	22.00	0	0	SC/A4 23568	8.0500		S
1040	1	Crafton, Mr. John Bertram	male	NA	0	0	113791	26.5500		S
1041	2	Lahtinen, Rev. William	male	30.00	1	1	250651	26.0000		S
1042	1	Earnshaw, Mrs. Boulton (Olive Potter)	female	23.00	0	1	11767	83.1583	C54	C
1043	3	Matinoff, Mr. Nicola	male	NA	0	0	349255	7.8958		C
1044	3	Storey, Mr. Thomas	male	60.50	0	0	3701	NA		S
1045	3	Klasen, Mrs. (Hulda Kristina Eugenia Lofqvist)	female	36.00	0	2	350405	12.1833		S
1046	3	Asplund, Master. Filip Oscar	male	13.00	4	2	347077	31.3875		S

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1047	3	Duquemin, Mr. Joseph	male	24.00	0	0	S.O./P.P. 752	7.5500		S
1048	1	Bird, Miss. Ellen	female	29.00	0	0	PC 17483	221.7792	C97	S
1049	3	Lundin, Miss. Olga Elida	female	23.00	0	0	347469	7.8542		S
1050	1	Borebank, Mr. John James	male	42.00	0	0	110489	26.5500	D22	S
1051	3	Peacock, Mrs. Benjamin (Edith Nile)	female	26.00	0	2	SOTON/O.Q. 3101315	13.7750		S
1052	3	Smyth, Miss. Julia	female	NA	0	0	335432	7.7333		Q
1053	3	Touma, Master. Georges Youssef	male	7.00	1	1	2650	15.2458		C
1054	2	Wright, Miss. Marion	female	26.00	0	0	220844	13.5000		S
1055	3	Pearce, Mr. Ernest	male	NA	0	0	343271	7.0000		S
1056	2	Peruschitz, Rev. Joseph Maria	male	41.00	0	0	237393	13.0000		S
1057	3	Kink-Heilmann, Mrs. Anton (Luise Heilmann)	female	26.00	1	1	315153	22.0250		S
1058	1	Brandeis, Mr. Emil	male	48.00	0	0	PC 17591	50.4958	B10	C
1059	3	Ford, Mr. Edward Watson	male	18.00	2	2	W./C. 6608	34.3750		S
1060	1	Cassebeer, Mrs. Henry Arthur Jr (Eleanor Genevieve Fosdick)	female	NA	0	0	17770	27.7208		C
1061	3	Hellstrom, Miss. Hilda Maria	female	22.00	0	0	7548	8.9625		S
1062	3	Lithman, Mr. Simon	male	NA	0	0	S.O./P.P. 251	7.5500		S
1063	3	Zakarian, Mr. Ortin	male	27.00	0	0	2670	7.2250		C
1064	3	Dyker, Mr. Adolf Fredrik	male	23.00	1	0	347072	13.9000		S
1065	3	Torfa, Mr. Assad	male	NA	0	0	2673	7.2292		C
1066	3	Asplund, Mr. Carl Oscar Vilhelm Gustafsson	male	40.00	1	5	347077	31.3875		S
1067	2	Brown, Miss. Edith Eileen	female	15.00	0	2	29750	39.0000		S
1068	2	Sincock, Miss. Maude	female	20.00	0	0	C.A. 33112	36.7500		S
1069	1	Stengel, Mr. Charles Emil Henry	male	54.00	1	0	11778	55.4417	C116	C
1070	2	Becker, Mrs. Allen Oliver (Nellie E Baumgardner)	female	36.00	0	3	230136	39.0000	F4	S
1071	1	Compton, Mrs. Alexander Taylor (Mary Eliza Ingersoll)	female	64.00	0	2	PC 17756	83.1583	E45	C
1072	2	McCrie, Mr. James Matthew	male	30.00	0	0	233478	13.0000		S
1073	1	Compton, Mr. Alexander Taylor Jr	male	37.00	1	1	PC 17756	83.1583	E52	C
1074	1	Marvin, Mrs. Daniel Warner (Mary Graham Carmichael Farquarson)	female	18.00	1	0	113773	53.1000	D30	S
1075	3	Lane, Mr. Patrick	male	NA	0	0	7935	7.7500		Q
1076	1	Douglas, Mrs. Frederick Charles (Mary Helene Baxter)	female	27.00	1	1	PC 17558	247.5208	B58 B60	C
1077	2	Maybery, Mr. Frank Hubert	male	40.00	0	0	239059	16.0000		S
1078	2	Phillips, Miss. Alice Frances Louisa	female	21.00	0	1	S.O./P.P. 2	21.0000		S

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1079	3	Davies, Mr. Joseph	male	17.00	2	0	A/4 48873	8.0500		S
1080	3	Sage, Miss. Ada	female	NA	8	2	CA. 2343	69.5500		S
1081	2	Veal, Mr. James	male	40.00	0	0	28221	13.0000		S
1082	2	Angle, Mr. William A	male	34.00	1	0	226875	26.0000		S
1083	1	Salomon, Mr. Abraham L	male	NA	0	0	111163	26.0000		S
1084	3	van Billiard, Master. Walter John	male	11.50	1	1	A/5. 851	14.5000		S
1085	2	Lingane, Mr. John	male	61.00	0	0	235509	12.3500		Q
1086	2	Drew, Master. Marshall Brines	male	8.00	0	2	28220	32.5000		S
1087	3	Karlsson, Mr. Julius Konrad Eugen	male	33.00	0	0	347465	7.8542		S
1088	1	Spedden, Master. Robert Douglas	male	6.00	0	2	16966	134.5000	E34	C
1089	3	Nilsson, Miss. Berta Olivia	female	18.00	0	0	347066	7.7750		S
1090	2	Baimbrigge, Mr. Charles Robert	male	23.00	0	0	C.A. 31030	10.5000		S
1091	3	Rasmussen, Mrs. (Lena Jacobsen Solvang)	female	NA	0	0	65305	8.1125		S
1092	3	Murphy, Miss. Nora	female	NA	0	0	36568	15.5000		Q
1093	3	Danbom, Master. Gilbert Sigvard Emanuel	male	0.33	0	2	347080	14.4000		S
1094	1	Astor, Col. John Jacob	male	47.00	1	0	PC 17757	227.5250	C62 C64	C
1095	2	Quick, Miss. Winifred Vera	female	8.00	1	1	26360	26.0000		S
1096	2	Andrew, Mr. Frank Thomas	male	25.00	0	0	C.A. 34050	10.5000		S
1097	1	Omont, Mr. Alfred Fernand	male	NA	0	0	F.C. 12998	25.7417		C
1098	3	McGowan, Miss. Katherine	female	35.00	0	0	9232	7.7500		Q
1099	2	Collett, Mr. Sidney C Stuart	male	24.00	0	0	28034	10.5000		S
1100	1	Rosenbaum, Miss. Edith Louise	female	33.00	0	0	PC 17613	27.7208	A11	C
1101	3	Delalic, Mr. Redjo	male	25.00	0	0	349250	7.8958		S
1102	3	Andersen, Mr. Albert Karvin	male	32.00	0	0	C 4001	22.5250		S
1103	3	Finoli, Mr. Luigi	male	NA	0	0	SOTON/O.Q. 3101308	7.0500		S
1104	2	Deacon, Mr. Percy William	male	17.00	0	0	S.O.C. 14879	73.5000		S
1105	2	Howard, Mrs. Benjamin (Ellen Truelove Arman)	female	60.00	1	0	24065	26.0000		S
1106	3	Andersson, Miss. Ida Augusta Margareta	female	38.00	4	2	347091	7.7750		S
1107	1	Head, Mr. Christopher	male	42.00	0	0	113038	42.5000	B11	S
1108	3	Mahon, Miss. Bridget Delia	female	NA	0	0	330924	7.8792		Q
1109	1	Wick, Mr. George Dennick	male	57.00	1	1	36928	164.8667		S
1110	1	Widener, Mrs. George Dunton (Eleanor Elkins)	female	50.00	1	1	113503	211.5000	C80	C
1111	3	Thomson, Mr. Alexander Morrison	male	NA	0	0	32302	8.0500		S
1112	2	Duran y More, Miss. Florentina	female	30.00	1	0	SC/PARIS 2148	13.8583		C

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1113	3	Reynolds, Mr. Harold J	male	21.00	0	0	342684	8.0500		S
1114	2	Cook, Mrs. (Selena Rogers)	female	22.00	0	0	W./C. 14266	10.5000	F33	S
1115	3	Karlsson, Mr. Einar Gervasius	male	21.00	0	0	350053	7.7958		S
1116	1	Candee, Mrs. Edward (Helen Churchill Hungerford)	female	53.00	0	0	PC 17606	27.4458		C
1117	3	Moubarek, Mrs. George (Omine Amenia" Alexander)"	female	NA	0	2	2661	15.2458		C
1118	3	Asplund, Mr. Johan Charles	male	23.00	0	0	350054	7.7958		S
1119	3	McNeill, Miss. Bridget	female	NA	0	0	370368	7.7500		Q
1120	3	Everett, Mr. Thomas James	male	40.50	0	0	C.A. 6212	15.1000		S
1121	2	Hocking, Mr. Samuel James Metcalfe	male	36.00	0	0	242963	13.0000		S
1122	2	Sweet, Mr. George Frederick	male	14.00	0	0	220845	65.0000		S
1123	1	Willard, Miss. Constance	female	21.00	0	0	113795	26.5500		S
1124	3	Wiklund, Mr. Karl Johan	male	21.00	1	0	3101266	6.4958		S
1125	3	Linehan, Mr. Michael	male	NA	0	0	330971	7.8792		Q
1126	1	Cumings, Mr. John Bradley	male	39.00	1	0	PC 17599	71.2833	C85	C
1127	3	Vendel, Mr. Olof Edvin	male	20.00	0	0	350416	7.8542		S
1128	1	Warren, Mr. Frank Manley	male	64.00	1	0	110813	75.2500	D37	C
1129	3	Baccos, Mr. Raffull	male	20.00	0	0	2679	7.2250		C
1130	2	Hiltunen, Miss. Marta	female	18.00	1	1	250650	13.0000		S
1131	1	Douglas, Mrs. Walter Donald (Mahala Dutton)	female	48.00	1	0	PC 17761	106.4250	C86	C
1132	1	Lindstrom, Mrs. Carl Johan (Sigrid Posse)	female	55.00	0	0	112377	27.7208		C
1133	2	Christy, Mrs. (Alice Frances)	female	45.00	0	2	237789	30.0000		S
1134	1	Spedden, Mr. Frederic Oakley	male	45.00	1	1	16966	134.5000	E34	C
1135	3	Hyman, Mr. Abraham	male	NA	0	0	3470	7.8875		S
1136	3	Johnston, Master. William Arthur Willie" "	male	NA	1	2	W./C. 6607	23.4500		S
1137	1	Kenyon, Mr. Frederick R	male	41.00	1	0	17464	51.8625	D21	S
1138	2	Karnes, Mrs. J Frank (Claire Bennett)	female	22.00	0	0	F.C.C. 13534	21.0000		S
1139	2	Drew, Mr. James Vivian	male	42.00	1	1	28220	32.5000		S
1140	2	Hold, Mrs. Stephen (Annie Margaret Hill)	female	29.00	1	0	26707	26.0000		S
1141	3	Khalil, Mrs. Betros (Zahie Maria" Elias)"	female	NA	1	0	2660	14.4542		C
1142	2	West, Miss. Barbara J	female	0.92	1	2	C.A. 34651	27.7500		S
1143	3	Abrahamsson, Mr. Abraham August Johannes	male	20.00	0	0	SOTON/O 3101284	27.9250		S
1144	1	Clark, Mr. Walter Miller	male	27.00	1	0	13508	136.7792	C89	C
1145	3	Salander, Mr. Karl Johan	male	24.00	0	0	7266	9.3250		S
1146	3	Wenzel, Mr. Linhart	male	32.50	0	0	345775	9.5000		S
1147	3	MacKay, Mr. George William	male	NA	0	0	C.A. 42795	7.5500		S
1148	3	Mahon, Mr. John	male	NA	0	0	AQ/4 3130	7.7500		Q

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1149	3	Niklasson, Mr. Samuel	male	28.00	0	0	363611	8.0500		S
1150	2	Bentham, Miss. Lilian W	female	19.00	0	0	28404	13.0000		S
1151	3	Midtsjo, Mr. Karl Albert	male	21.00	0	0	345501	7.7750		S
1152	3	de Messemaeker, Mr. Guillaume Joseph	male	36.50	1	0	345572	17.4000		S
1153	3	Nilsson, Mr. August Ferdinand	male	21.00	0	0	350410	7.8542		S
1154	2	Wells, Mrs. Arthur Henry (Addie" Dart Trevaskis)"	female	29.00	0	2	29103	23.0000		S
1155	3	Klasen, Miss. Gertrud Emilia	female	1.00	1	1	350405	12.1833		S
1156	2	Portaluppi, Mr. Emilio Ilario Giuseppe	male	30.00	0	0	C.A. 34644	12.7375		C
1157	3	Lyntakoff, Mr. Stanko	male	NA	0	0	349235	7.8958		S
1158	1	Chisholm, Mr. Roderick Robert Crispin	male	NA	0	0	112051	0.0000		S
1159	3	Warren, Mr. Charles William	male	NA	0	0	C.A. 49867	7.5500		S
1160	3	Howard, Miss. May Elizabeth	female	NA	0	0	A. 2. 39186	8.0500		S
1161	3	Pokrnic, Mr. Mate	male	17.00	0	0	315095	8.6625		S
1162	1	McCaffry, Mr. Thomas Francis	male	46.00	0	0	13050	75.2417	C6	C
1163	3	Fox, Mr. Patrick	male	NA	0	0	368573	7.7500		Q
1164	1	Clark, Mrs. Walter Miller (Virginia McDowell)	female	26.00	1	0	13508	136.7792	C89	C
1165	3	Lennon, Miss. Mary	female	NA	1	0	370371	15.5000		Q
1166	3	Saade, Mr. Jean Nassr	male	NA	0	0	2676	7.2250		C
1167	2	Bryhl, Miss. Dagmar Jenny Ingeborg	female	20.00	1	0	236853	26.0000		S
1168	2	Parker, Mr. Clifford Richard	male	28.00	0	0	SC 14888	10.5000		S
1169	2	Faunthorpe, Mr. Harry	male	40.00	1	0	2926	26.0000		S
1170	2	Ware, Mr. John James	male	30.00	1	0	CA 31352	21.0000		S
1171	2	Oxenham, Mr. Percy Thomas	male	22.00	0	0	W./C. 14260	10.5000		S
1172	3	Oreskovic, Miss. Jelka	female	23.00	0	0	315085	8.6625		S
1173	3	Peacock, Master. Alfred Edward	male	0.75	1	1	SOTON/O.Q. 3101315	13.7750		S
1174	3	Fleming, Miss. Honora	female	NA	0	0	364859	7.7500		Q
1175	3	Touma, Miss. Maria Youssef	female	9.00	1	1	2650	15.2458		C
1176	3	Rosblom, Miss. Salli Helena	female	2.00	1	1	370129	20.2125		S
1177	3	Dennis, Mr. William	male	36.00	0	0	A/5 21175	7.2500		S
1178	3	Franklin, Mr. Charles (Charles Fardon)	male	NA	0	0	SOTON/O.Q. 3101314	7.2500		S
1179	1	Snyder, Mr. John Pillsbury	male	24.00	1	0	21228	82.2667	B45	S
1180	3	Mardirosian, Mr. Sarkis	male	NA	0	0	2655	7.2292	F E46	C
1181	3	Ford, Mr. Arthur	male	NA	0	0	A/5 1478	8.0500		S
1182	1	Rheims, Mr. George Alexander Lucien	male	NA	0	0	PC 17607	39.6000		S
1183	3	Daly, Miss. Margaret Marcella Maggie" "	female	30.00	0	0	382650	6.9500		Q
1184	3	Nasr, Mr. Mustafa	male	NA	0	0	2652	7.2292		C
1185	1	Dodge, Dr. Washington	male	53.00	1	1	33638	81.8583	A34	S

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1186	3	Wittevrongel, Mr. Camille	male	36.00	0	0	345771	9.5000		S
1187	3	Angheloff, Mr. Minko	male	26.00	0	0	349202	7.8958		S
1188	2	Laroche, Miss. Louise	female	1.00	1	2	SC/Paris 2123	41.5792		C
1189	3	Samaan, Mr. Hanna	male	NA	2	0	2662	21.6792		C
1190	1	Loring, Mr. Joseph Holland	male	30.00	0	0	113801	45.5000		S
1191	3	Johansson, Mr. Nils	male	29.00	0	0	347467	7.8542		S
1192	3	Olsson, Mr. Oscar Wilhelm	male	32.00	0	0	347079	7.7750		S
1193	2	Malachard, Mr. Noel	male	NA	0	0	237735	15.0458	D	C
1194	2	Phillips, Mr. Escott Robert	male	43.00	0	1	S.O./P.P. 2	21.0000		S
1195	3	Pokrnic, Mr. Tome	male	24.00	0	0	315092	8.6625		S
1196	3	McCarthy, Miss. Catherine Katie” ”	female	NA	0	0	383123	7.7500		Q
1197	1	Crosby, Mrs. Edward Gifford (Catherine Elizabeth Halstead)	female	64.00	1	1	112901	26.5500	B26	S
1198	1	Allison, Mr. Hudson Joshua Creighton	male	30.00	1	2	113781	151.5500	C22 C26	S
1199	3	Aks, Master. Philip Frank	male	0.83	0	1	392091	9.3500		S
1200	1	Hays, Mr. Charles Melville	male	55.00	1	1	12749	93.5000	B69	S
1201	3	Hansen, Mrs. Claus Peter (Jennie L Howard)	female	45.00	1	0	350026	14.1083		S
1202	3	Cacic, Mr. Jego Grga	male	18.00	0	0	315091	8.6625		S
1203	3	Vartanian, Mr. David	male	22.00	0	0	2658	7.2250		C
1204	3	Sadowitz, Mr. Harry	male	NA	0	0	LP 1588	7.5750		S
1205	3	Carr, Miss. Jeannie	female	37.00	0	0	368364	7.7500		Q
1206	1	White, Mrs. John Stuart (Ella Holmes)	female	55.00	0	0	PC 17760	135.6333	C32	C
1207	3	Hagardon, Miss. Kate	female	17.00	0	0	AQ/3. 30631	7.7333		Q
1208	1	Spencer, Mr. William Augustus	male	57.00	1	0	PC 17569	146.5208	B78	C
1209	2	Rogers, Mr. Reginald Harry	male	19.00	0	0	28004	10.5000		S
1210	3	Jonsson, Mr. Nils Hilding	male	27.00	0	0	350408	7.8542		S
1211	2	Jefferys, Mr. Ernest Wilfred	male	22.00	2	0	C.A. 31029	31.5000		S
1212	3	Andersson, Mr. Johan Samuel	male	26.00	0	0	347075	7.7750		S
1213	3	Krekorian, Mr. Neshan	male	25.00	0	0	2654	7.2292	F E57	C
1214	2	Nesson, Mr. Israel	male	26.00	0	0	244368	13.0000	F2	S
1215	1	Rowe, Mr. Alfred G	male	33.00	0	0	113790	26.5500		S
1216	1	Kreuchen, Miss. Emilie	female	39.00	0	0	24160	211.3375		S
1217	3	Assam, Mr. Ali	male	23.00	0	0	SOTON/O.Q. 3101309	7.0500		S
1218	2	Becker, Miss. Ruth Elizabeth	female	12.00	2	1	230136	39.0000	F4	S
1219	1	Rosenshine, Mr. George (Mr George Thorne”)”	male	46.00	0	0	PC 17585	79.2000		C
1220	2	Clarke, Mr. Charles Valentine	male	29.00	1	0	2003	26.0000		S
1221	2	Enander, Mr. Ingvar	male	21.00	0	0	236854	13.0000		S
1222	2	Davies, Mrs. John Morgan (Elizabeth Agnes Mary White)	female	48.00	0	2	C.A. 33112	36.7500		S
1223	1	Dulles, Mr. William Crothers	male	39.00	0	0	PC 17580	29.7000	A18	C

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1224	3	Thomas, Mr. Tannous	male	NA	0	0	2684	7.2250		C
1225	3	Nakid, Mrs. Said (Waika Mary" Mowad)"	female	19.00	1	1	2653	15.7417		C
1226	3	Cor, Mr. Ivan	male	27.00	0	0	349229	7.8958		S
1227	1	Maguire, Mr. John Edward	male	30.00	0	0	110469	26.0000	C106	S
1228	2	de Brito, Mr. Jose Joaquim	male	32.00	0	0	244360	13.0000		S
1229	3	Elias, Mr. Joseph	male	39.00	0	2	2675	7.2292		C
1230	2	Denbury, Mr. Herbert	male	25.00	0	0	C.A. 31029	31.5000		S
1231	3	Betros, Master. Seman	male	NA	0	0	2622	7.2292		C
1232	2	Fillbrook, Mr. Joseph Charles	male	18.00	0	0	C.A. 15185	10.5000		S
1233	3	Lundstrom, Mr. Thure Edvin	male	32.00	0	0	350403	7.5792		S
1234	3	Sage, Mr. John George	male	NA	1	9	CA. 2343	69.5500		S
1235	1	Cardeza, Mrs. James Warburton Martinez (Charlotte Wardle Drake)	female	58.00	0	1	PC 17755	512.3292	B51 B53 B55	C
1236	3	van Billiard, Master. James William	male	NA	1	1	A/5. 851	14.5000		S
1237	3	Abelseth, Miss. Karen Marie	female	16.00	0	0	348125	7.6500		S
1238	2	Botsford, Mr. William Hull	male	26.00	0	0	237670	13.0000		S
1239	3	Whabee, Mrs. George Joseph (Shawneene Abi-Saab)	female	38.00	0	0	2688	7.2292		C
1240	2	Giles, Mr. Ralph	male	24.00	0	0	248726	13.5000		S
1241	2	Walcroft, Miss. Nellie	female	31.00	0	0	F.C.C. 13528	21.0000		S
1242	1	Greenfield, Mrs. Leo David (Blanche Strouse)	female	45.00	0	1	PC 17759	63.3583	D10 D12	C
1243	2	Stokes, Mr. Philip Joseph	male	25.00	0	0	F.C.C. 13540	10.5000		S
1244	2	Dibden, Mr. William	male	18.00	0	0	S.O.C. 14879	73.5000		S
1245	2	Herman, Mr. Samuel	male	49.00	1	2	220845	65.0000		S
1246	3	Dean, Miss. Elizabeth Gladys Millvina" "	female	0.17	1	2	C.A. 2315	20.5750		S
1247	1	Julian, Mr. Henry Forbes	male	50.00	0	0	113044	26.0000	E60	S
1248	1	Brown, Mrs. John Murray (Caroline Lane Lamson)	female	59.00	2	0	11769	51.4792	C101	S
1249	3	Lockyer, Mr. Edward	male	NA	0	0	1222	7.8792		S
1250	3	O'Keefe, Mr. Patrick	male	NA	0	0	368402	7.7500		Q
1251	3	Lindell, Mrs. Edvard Bengtsson (Elin Gerda Persson)	female	30.00	1	0	349910	15.5500		S
1252	3	Sage, Master. William Henry	male	14.50	8	2	CA. 2343	69.5500		S
1253	2	Mallet, Mrs. Albert (Antoinette Magnin)	female	24.00	1	1	S.C./PARIS 2079	37.0042		C
1254	2	Ware, Mrs. John James (Florence Louise Long)	female	31.00	0	0	CA 31352	21.0000		S
1255	3	Strilic, Mr. Ivan	male	27.00	0	0	315083	8.6625		S
1256	1	Harder, Mrs. George Achilles (Dorothy Annan)	female	25.00	1	0	11765	55.4417	E50	C
1257	3	Sage, Mrs. John (Annie Bullen)	female	NA	1	9	CA. 2343	69.5500		S
1258	3	Caram, Mr. Joseph	male	NA	1	0	2689	14.4583		C

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1259	3	Riihivouri, Miss. Susanna Juhantytar Sanni" "	female	22.00	0	0	3101295	39.6875		S
1260	1	Gibson, Mrs. Leonard (Pauline C Boeson)	female	45.00	0	1	112378	59.4000		C
1261	2	Pallas y Castello, Mr. Emilio	male	29.00	0	0	SC/PARIS 2147	13.8583		C
1262	2	Giles, Mr. Edgar	male	21.00	1	0	28133	11.5000		S
1263	1	Wilson, Miss. Helen Alice	female	31.00	0	0	16966	134.5000	E39 E41	C
1264	1	Ismay, Mr. Joseph Bruce	male	49.00	0	0	112058	0.0000	B52 B54 B56	S
1265	2	Harbeck, Mr. William H	male	44.00	0	0	248746	13.0000		S
1266	1	Dodge, Mrs. Washington (Ruth Vidaver)	female	54.00	1	1	33638	81.8583	A34	S
1267	1	Bowen, Miss. Grace Scott	female	45.00	0	0	PC 17608	262.3750		C
1268	3	Kink, Miss. Maria	female	22.00	2	0	315152	8.6625		S
1269	2	Cotterill, Mr. Henry Harry" "	male	21.00	0	0	29107	11.5000		S
1270	1	Hipkins, Mr. William Edward	male	55.00	0	0	680	50.0000	C39	S
1271	3	Asplund, Master. Carl Edgar	male	5.00	4	2	347077	31.3875		S
1272	3	O'Connor, Mr. Patrick	male	NA	0	0	366713	7.7500		Q
1273	3	Foley, Mr. Joseph	male	26.00	0	0	330910	7.8792		Q
1274	3	Risien, Mrs. Samuel (Emma)	female	NA	0	0	364498	14.5000		S
1275	3	McNamee, Mrs. Neal (Eileen O'Leary)	female	19.00	1	0	376566	16.1000		S
1276	2	Wheeler, Mr. Edwin Frederick" "	male	NA	0	0	SC/PARIS 2159	12.8750		S
1277	2	Herman, Miss. Kate	female	24.00	1	2	220845	65.0000		S
1278	3	Aronsson, Mr. Ernst Axel Algot	male	24.00	0	0	349911	7.7750		S
1279	2	Ashby, Mr. John	male	57.00	0	0	244346	13.0000		S
1280	3	Canavan, Mr. Patrick	male	21.00	0	0	364858	7.7500		Q
1281	3	Palsson, Master. Paul Folke	male	6.00	3	1	349909	21.0750		S
1282	1	Payne, Mr. Vivian Ponsonby	male	23.00	0	0	12749	93.5000	B24	S
1283	1	Lines, Mrs. Ernest H (Elizabeth Lindsey James)	female	51.00	0	1	PC 17592	39.4000	D28	S
1284	3	Abbott, Master. Eugene Joseph	male	13.00	0	2	C.A. 2673	20.2500		S
1285	2	Gilbert, Mr. William	male	47.00	0	0	C.A. 30769	10.5000		S
1286	3	Kink-Heilmann, Mr. Anton	male	29.00	3	1	315153	22.0250		S
1287	1	Smith, Mrs. Lucien Philip (Mary Eloise Hughes)	female	18.00	1	0	13695	60.0000	C31	S
1288	3	Colbert, Mr. Patrick	male	24.00	0	0	371109	7.2500		Q
1289	1	Frolicher-Stehli, Mrs. Maxmillian (Margaretha Emerentia Stehli)	female	48.00	1	1	13567	79.2000	B41	C
1290	3	Larsson-Rondberg, Mr. Edvard A	male	22.00	0	0	347065	7.7750		S
1291	3	Conlon, Mr. Thomas Henry	male	31.00	0	0	21332	7.7333		Q
1292	1	Bonnell, Miss. Caroline	female	30.00	0	0	36928	164.8667	C7	S

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1293	2	Gale, Mr. Harry	male	38.00	1	0	28664	21.0000		S
1294	1	Gibson, Miss. Dorothy Winifred	female	22.00	0	1	112378	59.4000		C
1295	1	Carrau, Mr. Jose Pedro	male	17.00	0	0	113059	47.1000		S
1296	1	Frauenthal, Mr. Isaac Gerald	male	43.00	1	0	17765	27.7208	D40	C
1297	2	Nourney, Mr. Alfred (Baron von Drachstedt")"	male	20.00	0	0	SC/PARIS 2166	13.8625	D38	C
1298	2	Ware, Mr. William Jeffery	male	23.00	1	0	28666	10.5000		S
1299	1	Widener, Mr. George Dunton	male	50.00	1	1	113503	211.5000	C80	C
1300	3	Riordan, Miss. Johanna Hannah" "	female	NA	0	0	334915	7.7208		Q
1301	3	Peacock, Miss. Treasteall	female	3.00	1	1	SOTON/O.Q. 3101315	13.7750		S
1302	3	Naughton, Miss. Hannah	female	NA	0	0	365237	7.7500		Q
1303	1	Minahan, Mrs. William Edward (Lillian E Thorpe)	female	37.00	1	0	19928	90.0000	C78	Q
1304	3	Henriksson, Miss. Jenny Lovisa	female	28.00	0	0	347086	7.7750		S
1305	3	Spector, Mr. Woolf	male	NA	0	0	A.5. 3236	8.0500		S
1306	1	Oliva y Ocana, Dona. Fermina	female	39.00	0	0	PC 17758	108.9000	C105	C
1307	3	Saether, Mr. Simon Sivertsen	male	38.50	0	0	SOTON/O.Q. 3101262	7.2500		S
1308	3	Ware, Mr. Frederick	male	NA	0	0	359309	8.0500		S
1309	3	Peter, Master. Michael J	male	NA	1	1	2668	22.3583		C

Las variables son:

- *Name:* Nombre del pasajero
- *PassengerId:* Ids del pasajero
- *Survived:* Si sobrevivió o no (No = 0, Sí = 1)
- *Ticket:* Número de ticket
- *Cabin:* Cabina en la que viajó
- *Pclass:* Clase en la que viajó (1 = 1era, 2 = 2da, 3 = 3ra)
- *Sex:* Masculino o Femenino (male/female)
- *Age:* Edad
- *SibSp:* Número de hermanos/conyuge a bordo
- *Parch:* Número de padres/hijos a bordo
- *Fare:* Tarifa que pagó
- *Embarked:* Puerto de embarcación (C = Cherbourg, Q = Queenstown, S = Southampton)

Preparación de la base de datos

Ajustando las variables

Variables de interés: Quita aquellas que de entrada no tengan que ver con la sobrevivencia del pasajero. Por ejemplo: Quitar variables 4, 9 y 11 (define si hay más)

Variables categóricas que deben aparecer como factores: define qué variables aparecerán como factores Por ejemplo: Survived, Pclass, Sex y Embarked (define si hay más)

```
# Eliminar variables irrelevantes
M1 <- M[,c(-4, -9, -11)]
M1_test <- M_test[, !colnames(M_test) %in% c("Name", "Ticket", "Cabin")]



#Transformar a factores:
for(var in c('Survived','Pclass','Embarked','Sex'))
  M1[,var] <-as.factor(M1[,var])

for(var in c('Pclass', 'Embarked', 'Sex')) {
  M1_test[,var] <- factor(M1_test[,var], levels = levels(M1[,var]))
}
```

Análisis de datos faltantes

Detectar si hay espacios vacíos en lugar de datos:

```
V = matrix(NA,ncol=1,nrow=9)
for(i in c(1:9)){
  V[i,] <- sum(with(M1,M1[,i])=="")}
V
```

0
0
0
0
NA
0
0
NA
NA

Ninguna variable contiene espacios vacíos a excepción de las variables 5 (Age), 8 (Fare) y 9 (Embarked) tienen datos faltantes.

Para contar los datos faltantes:

```
N = apply(X=is.na(M1),MARGIN = 2,FUN = sum)
P = round(100*N/length(M1[,2]),2)
NP = data.frame(as.numeric(N),as.numeric(P))
row.names(NP)= c("PassengerId", "Survived", "Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "Embarked
```

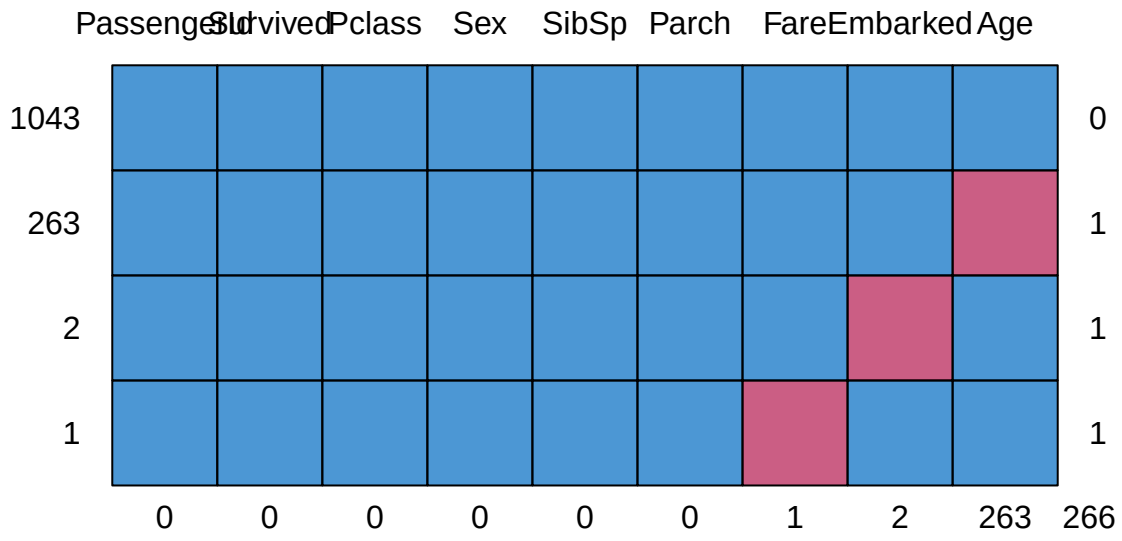
```
names(NP)=c("Número","Porcentaje")
t(NP)
```

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
Número	0	0	0	0	263.00	0	0	1.00	2.00
Porcentaje	0	0	0	0	20.09	0	0	0.08	0.15

En edad hay muchos datos faltantes, el 20% de los datos.

Observemos el patrón de los datos faltantes:

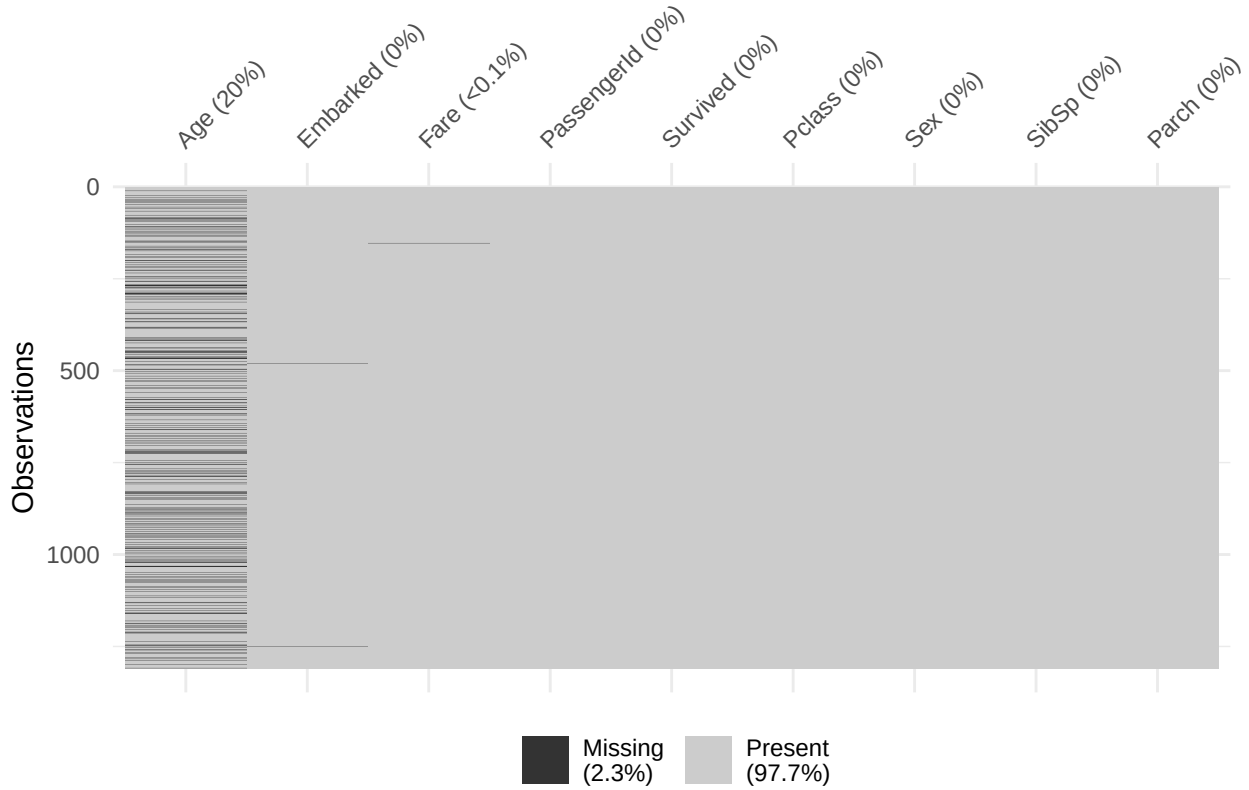
```
md.pattern(M1)
```



	PassengerId	Survived	Pclass	Sex	SibSp	Parch	Fare	Embarked	Age	
1043	1	1	1	1	1	1	1	1	1	0
263	1	1	1	1	1	1	1	1	0	1
2	1	1	1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	0	1	1	1
	0	0	0	0	0	0	1	2	263	266

Todos los datos faltantes son de distintos pasajeros (observaciones), por lo tanto, si se eliminan los NA, se eliminarían 266 observaciones y nos quedaríamos con 1043 observaciones.

```
vis_miss(M1,sort_miss = TRUE)
```



Análisis sobre datos faltantes

Medidas con datos faltantes

```
summary(M1[,-1])
```

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0:815	1:323	female:466	Min. : 0.17	Min. :0.0000	Min. :0.000	Min. : 0.000	C :270
1:494	2:277	male :843	1st Qu.:21.00	1st Qu.:0.0000	1st Qu.:0.000	1st Qu.: 7.896	Q :123
NA	3:709	NA	Median :28.00	Median :0.0000	Median :0.000	Median : 14.454	S :914
NA	NA	NA	Mean :29.88	Mean :0.4989	Mean :0.385	Mean : 33.295	NA's: 2
NA	NA	NA	3rd Qu.:39.00	3rd Qu.:1.0000	3rd Qu.:0.000	3rd Qu.: 31.275	NA
NA	NA	NA	Max. :80.00	Max. :8.0000	Max. :9.000	Max. :512.329	NA
NA	NA	NA	NA's :263	NA	NA	NA's :1	NA

Medidas sin datos faltantes

```r
M2 = na.omit(M1)
M2_test = na.omit(M1_test)
summary(M2[,-1])
```

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0:628	1:282	female:386	Min. : 0.17	Min. :0.0000	Min. :0.0000	Min. : 0.00	C:212
1:415	2:261	male :657	1st Qu.:21.00	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.: 8.05	Q: 50
NA	3:500	NA	Median :28.00	Median :0.0000	Median :0.0000	Median : 15.75	S:781
NA	NA	NA	Mean :29.81	Mean :0.5043	Mean :0.4219	Mean : 36.60	NA
NA	NA	NA	3rd Qu.:39.00	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.: 35.08	NA
NA	NA	NA	Max. :80.00	Max. :8.0000	Max. :6.0000	Max. :512.33	NA

proporcion de mujeres/hombres de M1 = 0.552 proporcion de mujeres/hombres de M2 = 0.587

proporcion de num de sobrevivientes 0 / 1 de M1:1.64 proporcion de num de sobrevivientes 0 / 1 de M2:1.64

como podemos ver los cambios no fueron significativos eliminando los valores faltantes. Los valores estadisticos descriptivos con los valores faltantes se hicieron con los datos que había disponibles, únicamente pudieron cambiar las proporciones y sus derivados se obtuvo la proporción de mujeres por cantidad de hombres para ambos datasets y esta no se vio muy afectada. De la misma forma se obtuvo la razón de no sobrevivientes por sobrevivientes y a los primeros dos decimales es identica. Esto indica que no habrá una influencia significativa o de impacto sobre el modelo resultante.

Análisis descriptivo

Se recomienda analizar dividiendo la base de datos entre los que sobrevivieron y los que no. Usa:

- Medidas
- Gráficos

Partición. Entrenamiento y prueba

Se toma el 70% de la muestra como entrenamiento y el 30% para prueba.

```r
M_indice <- createDataPartition(M2$Survived, p = .7, list = FALSE, times = 1)

M_train <- M2[ M_indice,]  %>% as_tibble()
M_valid <- M2[-M_indice,] %>% as_tibble()
```

Proporciones de sobrevivientes en las tres bases de datos

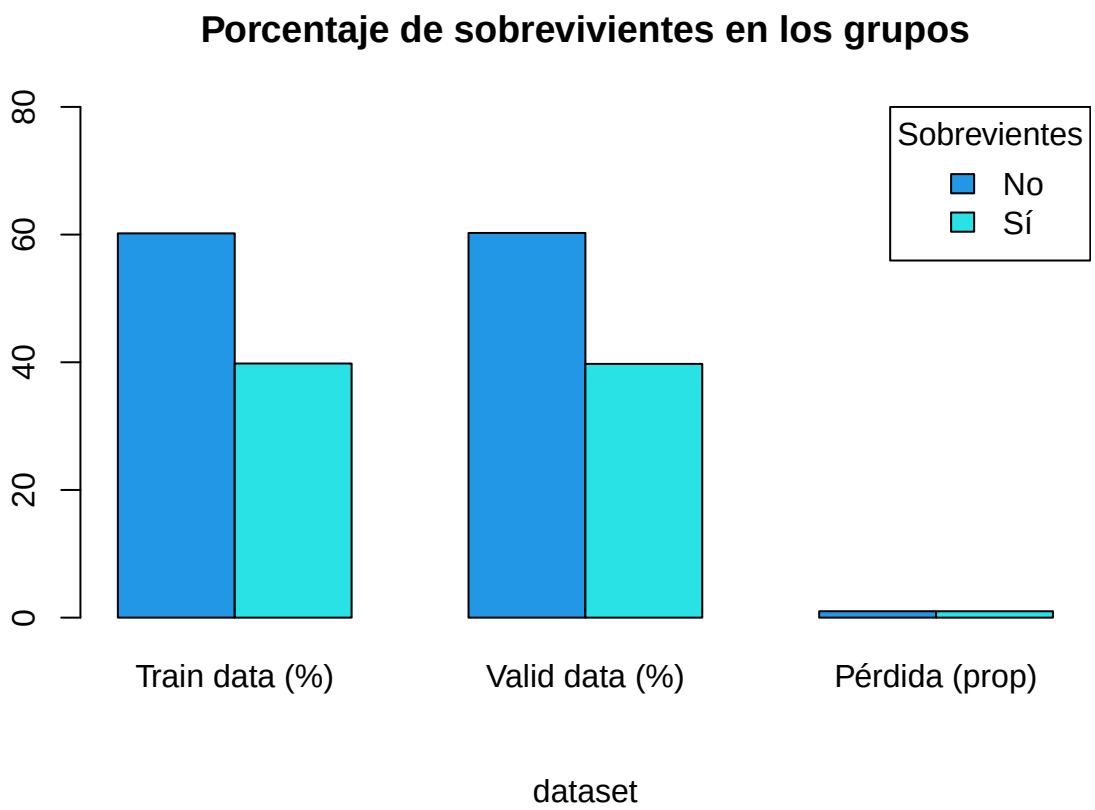
proporción de sobrevivientes entrenamiento y de pruebas

```r
t2c_train = 100*prop.table(table(M_train[,2]))
t2s_valid = 100*prop.table(table(M_valid[,2]))
t2s_complete = 100*prop.table(table(M2[,2]))
t2p_split = c(t2s_valid[1]/t2c_train[1],t2s_valid[2]/t2c_train[2])
t2_split = data.frame(as.numeric(t2c_train),as.numeric(t2s_valid),as.numeric(t2p_split))
row.names(t2_split) = c("Murió","Sobrevivió")
names(t2_split) = c("Train data (%)","Valid data (%)","Pérdida (prop)")
round(t2_split,2)
```

	Train data (%)	Valid data (%)	Pérdida (prop)
Murió	60.19	60.26	1
Sobrevivió	39.81	39.74	1

- Calcula la proporción de sobrevivientes en cada base de datos: Entrenamiento, prueba y completa. Haz una tabla comparativa
- Haz un gráfico de barras que te ayude a comparar las tres bases de datos. Auxíliate del código:

```r
barplot(as.matrix(t2_split), col=4:5, beside=TRUE, main="Porcentaje de sobrevivientes en los grupos", su
legend("topright",legend = c("No","Sí"), title = "Sobrevientes",fill = 4:5)
```



La proporción en el data set se mantiene con una proporción casi perfecta, esto es útil por que hará que nuestro modelo no esté sesgado. Se puede inferir si dividimos los datos y su proporción se mantiene intacta, al momento de unirlos la proporción será la misma ya que estaremos añadiendo por igual a ambas categorías, por ende su proporción se mantiene.

Modelación (entrenamiento)

Comienza con el modelo completo, incluyendo las variables categóricas (factores). Aplica el comando *step* para poder encontrar el mejor modelo.

step utiliza el criterio de Aikaike (AIC) para definir el mejor modelo, sin embargo también proporciona la desviación residual del modelo completo. Un menor AIC y una menor *Deviance* indicarán un mejor modelo.

Corremos el modelo con todas las variables menos el modelo objetivo, la familia de modelos es binomial por que la variable objetivo es de caracter binario.

```
A = glm(Survived ~ . , data = M_train, family = "binomial")
```

```
step(A, direction="both", trace=1 )
```

```
## Start:  AIC=569.43
## Survived ~ PassengerId + Pclass + Sex + Age + SibSp + Parch +
##     Fare + Embarked
##
##               Df Deviance    AIC
## - Embarked     2   548.69 566.69
## - Fare         1   547.46 567.46
## - Parch        1   547.74 567.74
## <none>             547.43 569.43
## - PassengerId  1   551.28 571.28
## - SibSp        1   554.32 574.32
## - Age          1   569.81 589.81
## - Pclass       2   589.00 607.00
## - Sex          1   883.33 903.33
##
## Step:  AIC=566.69
## Survived ~ PassengerId + Pclass + Sex + Age + SibSp + Parch +
##     Fare
##
##               Df Deviance    AIC
## - Fare         1   548.82 564.82
## - Parch        1   549.00 565.00
## <none>             548.69 566.69
## - PassengerId  1   552.32 568.32
## + Embarked     2   547.43 569.43
## - SibSp        1   556.18 572.18
## - Age          1   571.78 587.78
## - Pclass       2   594.31 608.31
## - Sex          1   886.45 902.45
##
## Step:  AIC=564.82
## Survived ~ PassengerId + Pclass + Sex + Age + SibSp + Parch
##
##               Df Deviance    AIC
## - Parch        1   549.05 563.05
## <none>             548.82 564.82
## - PassengerId  1   552.36 566.36
## + Fare         1   548.69 566.69
## + Embarked     2   547.46 567.46
```

```
## - SibSp        1   556.18 570.18
## - Age          1   571.95 585.95
## - Pclass       2   622.32 634.32
## - Sex          1   888.33 902.33
##
## Step:  AIC=563.05
## Survived ~ PassengerId + Pclass + Sex + Age + SibSp
##
##               Df Deviance    AIC
## <none>             549.05 563.05
## - PassengerId  1   552.55 564.55
## + Parch        1   548.82 564.82
## + Fare         1   549.00 565.00
## + Embarked     2   547.75 565.75
## - SibSp        1   557.83 569.83
## - Age          1   571.95 583.95
## - Pclass       2   622.35 632.35
## - Sex          1   896.85 908.85

##
## Call:  glm(formula = Survived ~ PassengerId + Pclass + Sex + Age + SibSp,
##     family = "binomial", data = M_train)
##
## Coefficients:
## (Intercept)  PassengerId      Pclass2      Pclass3      Sexmale          Age
##   5.1324877   -0.0005484   -1.6630334   -2.4820045   -3.6973064   -0.0422832
##       SibSp
##  -0.3985859
##
## Degrees of Freedom: 730 Total (i.e. Null);  724 Residual
## Null Deviance:       982.8
## Residual Deviance: 549.1     AIC: 563.1
```

El mejor modelo del AIC es el que tiene un valor de 542.19, en el que se utilizan las variables Pclass2, Pclass3, Sex, Age y SibSp

La ultima variable que elimino antes de encontrar el modelo fue “fare”.

```
summary(A)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = "binomial", data = M_train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.3503716  0.6428608   8.323  < 2e-16 ***
## PassengerId -0.0005803  0.0002975  -1.951   0.0511 .
## Pclass2     -1.5249978  0.3562835  -4.280 1.87e-05 ***
## Pclass3     -2.3527827  0.3702445  -6.355 2.09e-10 ***
## Sexmale     -3.7383011  0.2560126 -14.602  < 2e-16 ***
## Age         -0.0422521  0.0092630  -4.561 5.08e-06 ***
## SibSp       -0.3734438  0.1459732  -2.558   0.0105 *
## Parch       -0.0815642  0.1449797  -0.563   0.5737
```

```
## Fare         0.0004972  0.0027686   0.180   0.8575
## EmbarkedQ   -0.3889554  0.5849908  -0.665   0.5061
## EmbarkedS   -0.3241441  0.2932707  -1.105   0.2690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 982.80  on 730  degrees of freedom
## Residual deviance: 547.43  on 720  degrees of freedom
## AIC: 569.43
##
## Number of Fisher Scoring iterations: 5
```

Modelo B

- Prueba el modelo incluyendo la última variable que eliminó el comando *step*.
- Indica cuáles son las variables que incluye.
- Interpreta la significancia global (de todo el modelo) y la individual (de cada una de las variables)

El siguiente modelo incluye las variables sex, plclass, age, sibsp y fare

```
B = glm(formula = Survived ~ Sex + Pclass + Age + SibSp + Fare , family = "binomial", data = M_train)
summary(B)
```

```
##
## Call:
## glm(formula = Survived ~ Sex + Pclass + Age + SibSp + Fare, family = "binomial",
##     data = M_train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.6554596  0.5403502   8.616  < 2e-16 ***
## Sexmale     -3.6697559  0.2453991 -14.954  < 2e-16 ***
## Pclass2     -1.6196483  0.3443619  -4.703 2.56e-06 ***
## Pclass3     -2.3954195  0.3603024  -6.648 2.96e-11 ***
## Age         -0.0410575  0.0091197  -4.502 6.73e-06 ***
## SibSp       -0.3788277  0.1398729  -2.708  0.00676 **
## Fare         0.0002822  0.0026122   0.108  0.91397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 982.80  on 730  degrees of freedom
## Residual deviance: 552.54  on 724  degrees of freedom
## AIC: 566.54
##
## Number of Fisher Scoring iterations: 5
```

Una de las variables del modelo no es significativa, la que se había eliminado originalmente en el primer modelo (fare), con un AIC de 543. ## Modelo C

- Prueba el modelo tal como te lo recomendó el comando *step*.
- Indica cuáles son las variables que incluye.
- Interpreta la significancia global (de todo el modelo) y la individual (de cada una de las variables)

El siguiente modelo incluye las variables sex, plclass, age y sibsp

```
C = glm(formula = Survived ~  Sex + Pclass + Age + SibSp, family = "binomial", data = M_train)
summary(C)
```

```
##
## Call:
## glm(formula = Survived ~ Sex + Pclass + Age + SibSp, family = "binomial",
##     data = M_train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.679589   0.492345   9.505  < 2e-16 ***
## Sexmale     -3.671485   0.244869 -14.994  < 2e-16 ***
## Pclass2     -1.636265   0.308163  -5.310 1.10e-07 ***
## Pclass3     -2.415406   0.309318  -7.809 5.77e-15 ***
## Age         -0.041108   0.009111  -4.512 6.42e-06 ***
## SibSp       -0.376201   0.137714  -2.732   0.0063 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 982.80  on 730  degrees of freedom
## Residual deviance: 552.55  on 725  degrees of freedom
## AIC: 564.55
##
## Number of Fisher Scoring iterations: 5
```

#Modelo D

para este decidí implementar un random forest con las variables que originalmente recomendó el comando step

```
library(randomForest)
```

```
## randomForest 4.7-1.2
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
D <- randomForest(Survived ~ Sex + Pclass + Age + SibSp ,  data = M_train, ntree = 4000, mtry = 2, simp
print(D)
```

```
##
## Call:
##  randomForest(formula = Survived ~ Sex + Pclass + Age + SibSp,      data = M_train, ntree = 4000, mt
##                Type of random forest: classification
##                      Number of trees: 4000
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 14.23%
## Confusion matrix:
##     0   1 class.error
## 0 399  41  0.09318182
## 1  63 228  0.21649485
```

Comparación entre los modelos:

La comparación entre los tres modelos muestra que el Random Forest (con ntree = 4000, mtry = 2 y maxnodes = 20) tiene un Error OOB del 12.86%, logrando una mejor predicción general que los modelos GLM, aunque aún presenta una tasa de error del 19.93% para la clase 1 (sobrevivió) y 8.18% para la clase 0 (no sobrevivió). Por otro lado, el modelo de glm sin Fare tiene un aic de 542.19, todas las variables predictoras (Sex, Pclass, Age, SibSp) son estadísticamente significativas, y es más simple de interpretar. Finalmente, el modelo de glm con Fare, con un AIC de 543.2, no mejora respecto al modelo sin Fare ya que esta variable no es significativa (p = 0.326), y el desempeño general es similar.

Análisis de los modelos B , C y D

Resumen de los indicadores importantes de los modelos B , C y D

Elabora una tabla comparativa

```
modelos_b_c_tabla <- data.frame(

  aic = c(B$aic, C$aic),
  residual_deviance = c(B$deviance, C$deviance),
  null_deviance = c(B$null.deviance, C$null.deviance)
)

modelos_b_c_tabla
```

aic	residual_deviance	null_deviance
566.5405	552.5405	982.7966
564.5522	552.5522	982.7966

podemos ver que el null_deviance es mayor en ambos modelos , por lo que el intercepto no es suficiente para poder explicar los datos en ambos modelos, pero hace falta calcular el valor de chi cuadrado para poder ver si mejora el ajuste del modelo significativamente

Cálculo de la Desviación explicada ($pseudor^2$)

Calcula la desviación explicada para cada modelo. Recuerda que es igual a:

pseudo r^2 = 1-Desviación residual/Desviación nula

Desviación explicada para el modelo B

```
pseudo_r2_b <- 1 - ( (B$deviance) / (B$null.deviance) )
pseudo_r2_b
```

```
## [1] 0.4377875
```

Desviación explicada para el modelo C

```
pseudo_r2_c <- 1 - ( (C$deviance) / (C$null.deviance) )
pseudo_r2_c
```

```
## [1] 0.4377756
```

Para obtener la proporcion de datos explicados por el random forest podemos hacer lo siguiente, no hay forma explicita de hacer el equivalente a un modelo linear generalizado ya que no tenemos predictores que podamos quitar para comparar un modelo con y sin estos, pero podemos medir que tan bien se hace la predicción

```
pseudo_r2_rf <- 1 - (D$err.rate[nrow(D$err.rate), "OOB"] / var(as.numeric(as.character(M_train$Survived
pseudo_r2_rf
```

```
##       OOB
## 0.4070603
```

la variabilidad explicada por ambos modelos es casi identica, los valores resultantes indican que ambos modelos explican el 46% de los datos, sin embargo el equivalente en random forest es ligeramente mejor

Prueba de razón de verosimilitud

H_0 : El modelo con predictores explica mejor la variable respuesta: $log(\frac{p}{1-p})$ que el modelo nulo
H_1 : El modelo nulo explica mejor la variable respuesta: $log(\frac{p}{1-p})$ (la probabilidad es constante)

Se calcula el estadístico de χ^2 para la razón de verosimilutud a partir de las *Deviance* de los modelos.

para el modelo B

```
anova(B, test="Chisq")
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	730	982.7966	NA
Sex	1	353.8633397	729	628.9332	0.0000000
Pclass	2	50.7761556	727	578.1571	0.0000000
Age	1	17.6353047	726	560.5218	0.0000268

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
SibSp	1	7.9695482	725	552.5522	0.0047571
Fare	1	0.0117014	724	552.5405	0.9138583

para el modelo C

```
anova(C, test="Chisq")
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	730	982.7966	NA
Sex	1	353.863340	729	628.9332	0.0000000
Pclass	2	50.776156	727	578.1571	0.0000000
Age	1	17.635305	726	560.5218	0.0000268
SibSp	1	7.969548	725	552.5522	0.0047571

Para ambos modelos se puede observar que el que tiene más impacto es la variable sex, ya que tiene amyor disminucion en la devianza, por otra parte en ambos modelos todos son significativos a excepción de fare que su p-value es de 0.317 (mayor a 0.05).

```
# Importancia de las variables
importance(D)
```

	MeanDecreaseGini
Sex	148.252855
Pclass	25.081236
Age	26.033848
SibSp	8.344747

se confirma que la variable con mayor importancia es el sexo de la persona para saber si sobrevivirá o no.

Comparación entre los modelos B y C

Se pueden comparar los modelo B y C para ver si hay una diferencia significativa entre ambos con la misma razón de verosimilitud utilizando el comando ANOVA y la prueba LR.

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```r
anova(B,C,test="LR")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
724	552.5405	NA	NA	NA
725	552.5522	-1	-0.0117014	0.9138583

Dado que el modelo más simple explica los datos tan bien como el modelo más complejo a diferencia es una pequeña mejor de 0.99 (530.19 - 529.20), se recomienda usar el Modelo 2 porque es menos complejo y no incluye predictores innecesarios.

Por otro lado la variabilidad explicada del randomforest era un poco mejor comparandolas con las de los glm, por lo que decidí utilizarlo como el modelo para utilizar en la fase de testing.

Modelo Seleccionado

hacemos las predicciones con el modelo seleccionado, para ello lo evaluamos con el dataset de testing, que previamente preparamos elimiando los valores nulos (si es que tenía) y convirtiendo a factor las mismas variables convertidas que para el modelo de training, para luego hacer la comparación con val

```r
# Predicciones en el conjunto de validación
rf_pred_prob_valid <- predict(D, newdata = M_valid, type = "prob")[, 2]  # Probabilidades para clase "1
rf_pred_valid <- ifelse(rf_pred_prob_valid > 0.5, 1, 0)  # Clasificación binaria usando umbral 0.5
```

Matriz de confusión

```r
# Matriz de confusión
library(caret)
conf_matrix <- confusionMatrix(factor(rf_pred_valid), factor(M_valid$Survived))

# Imprimir la matriz de confusión
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 169  26
##          1  19  98
##
##                Accuracy : 0.8558
##                  95% CI : (0.8118, 0.8928)
##     No Information Rate : 0.6026
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6959
##
##  Mcnemar's Test P-Value : 0.3711
##
```

```
##             Sensitivity : 0.8989
##             Specificity : 0.7903
##          Pos Pred Value : 0.8667
##          Neg Pred Value : 0.8376
##              Prevalence : 0.6026
##          Detection Rate : 0.5417
##    Detection Prevalence : 0.6250
##       Balanced Accuracy : 0.8446
##
##        'Positive' Class : 0
##
```

calculamos las metricas de rendimiento

```
accuracy <- conf_matrix$overall["Accuracy"]
cat("Exactitud (Accuracy):", accuracy, "\n")
```

```
## Exactitud (Accuracy): 0.8557692
```

```
sensitivity <- conf_matrix$byClass["Sensitivity"]
cat("Sensibilidad (Recall):", sensitivity, "\n")
```

```
## Sensibilidad (Recall): 0.8989362
```

```
specificity <- conf_matrix$byClass["Specificity"]
cat("Especificidad (TNR):", specificity, "\n")
```

```
## Especificidad (TNR): 0.7903226
```

```
precision <- conf_matrix$byClass["Pos Pred Value"]
cat("Precisión (PPV):", precision, "\n")
```

```
## Precisión (PPV): 0.8666667
```

El modelo tiene un desempeño sólido, con una exactitud (accuracy) de 83.65%, lo que indica que predice correctamente el 83.65% de las observaciones, con un intervalo de confianza del 95% entre 79.07% y 87.58%. Su sensibilidad de 86.70% muestra que identifica correctamente la mayoría de los casos positivos, mientras que su especificidad de 79.03% refleja un desempeño ligeramente inferior al identificar los casos negativos. Además, la precisión de 86.24% indica que, de todas las predicciones positivas, el 86.24% son correctas.

Curva ROC

```
# Cargar las librerías
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'
```

```
## The following object is masked from 'package:Metrics':
##
##     auc

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

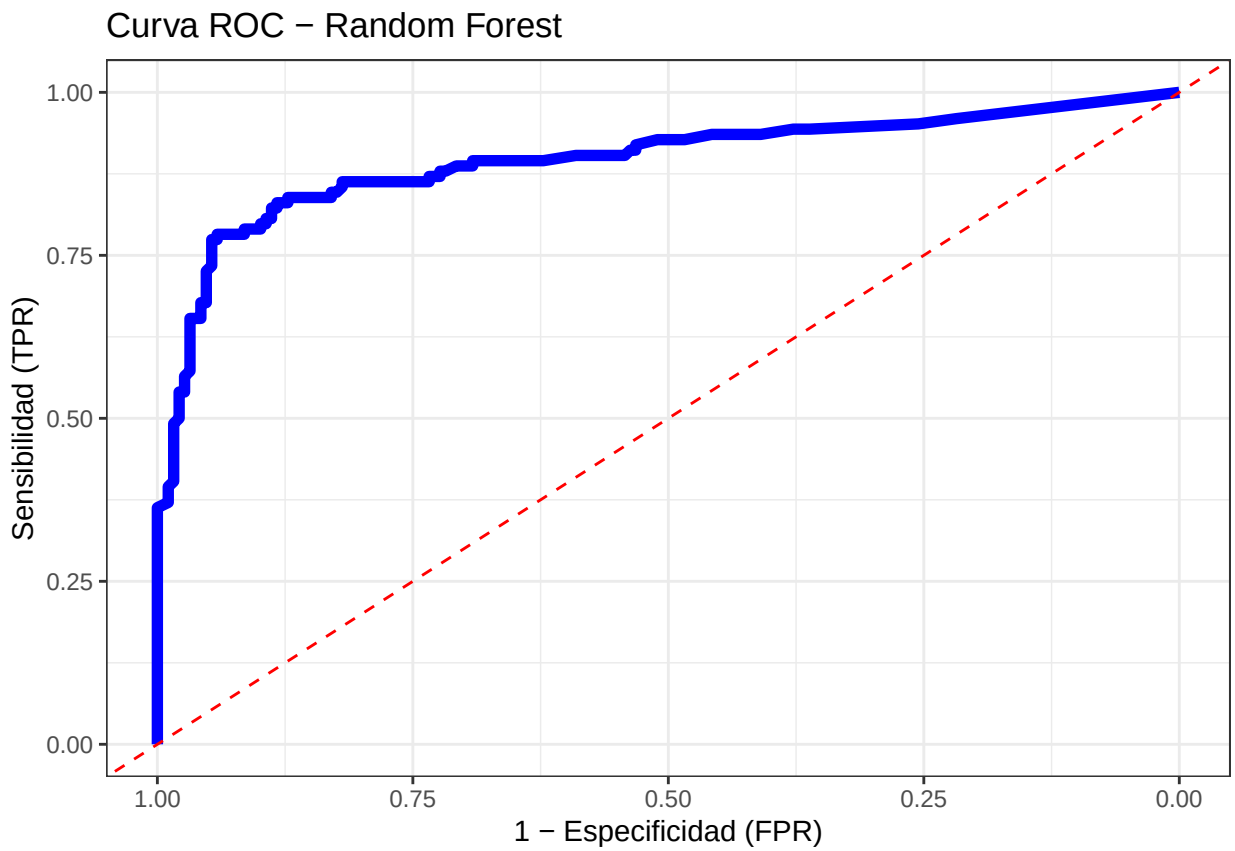
```
library(ggplot2)

# Predicciones del modelo Random Forest (probabilidades)
rf_pred_prob_valid <- predict(D, newdata = M_valid, type = "prob")[, 2]

# Crear el objeto ROC
ROC <- roc(response = M_valid$Survived, predictor = rf_pred_prob_valid)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```
# Imprimir el resultado de la curva ROC
print(ROC)
```

```
##
## Call:
## roc.default(response = M_valid$Survived, predictor = rf_pred_prob_valid)
##
## Data: rf_pred_prob_valid in 188 controls (M_valid$Survived 0) < 124 cases (M_valid$Survived 1).
## Area under the curve: 0.8964
```

```
# Graficar la curva ROC
ggroc(ROC, color = "blue", size = 2) +
  geom_abline(slope = 1, intercept = 1, linetype = 'dashed', color = "red") +
  labs(title = "Curva ROC - Random Forest", x = "1 - Especificidad (FPR)", y = "Sensibilidad (TPR)") +
  theme_bw()
```



La curva ROC generada para el modelo Random Forest refleja un excelente desempeño, con un Área Bajo la Curva de 0.8895, lo que indica que el modelo tiene una capacidad de discriminación muy buena entre las clases positivas y negativas. La gráfica muestra una curva que se aproxima al vértice superior izquierdo, lo cual es indicativo de una alta sensibilidad (verdaderos positivos) y una baja tasa de falsos positivos para la mayoría de los umbrales. Comparada con la línea diagonal roja, que representa un clasificador aleatorio, la curva está consistentemente por encima, confirmando que el modelo supera significativamente el azar

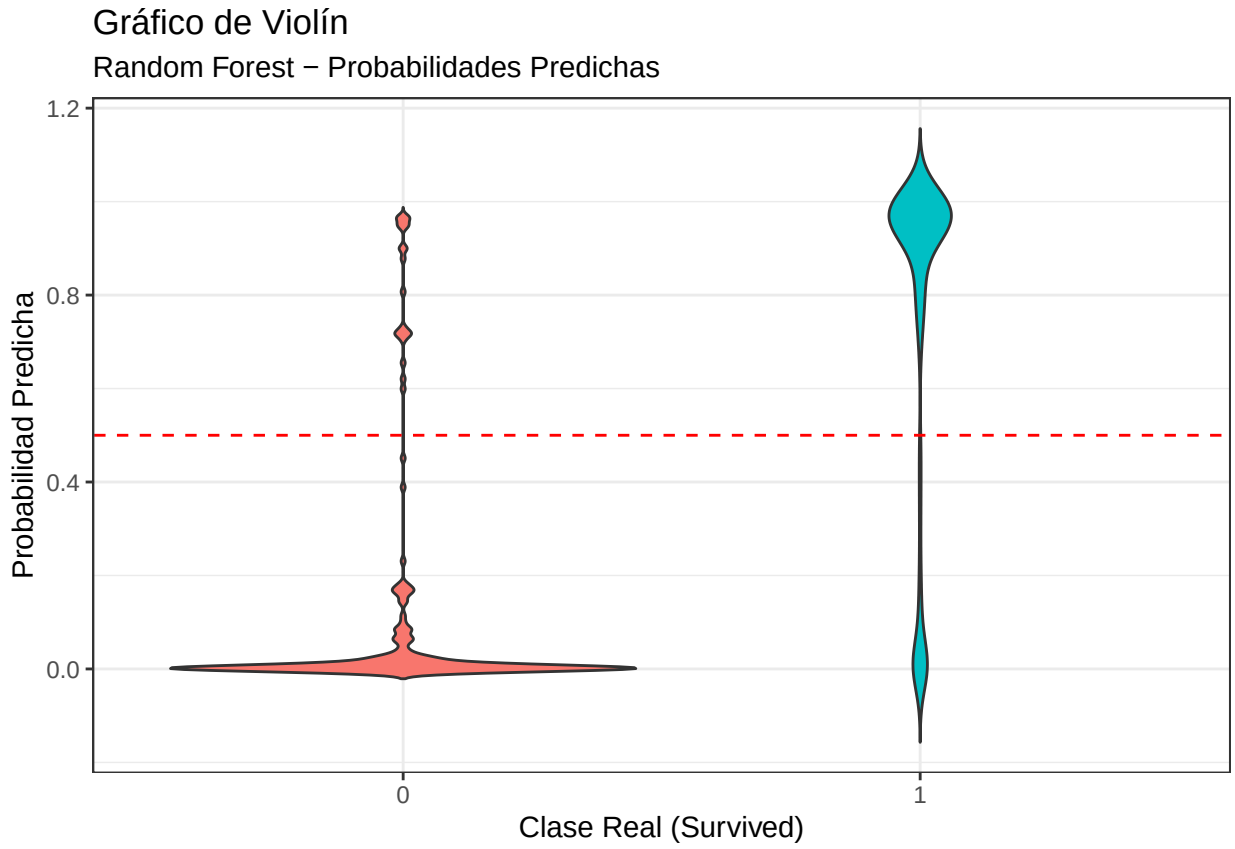
Gráfico de violín

Se crea la base de datos para el gráfico, se usan las predicciones ya elaboradas para el gráfico ROC y las clasificaciones originales (*train$M_Survived*).

```
# Crear la base de datos para el gráfico
v_d = data.frame(Survived = M_valid$Survived, pred = rf_pred_prob_valid)

# Graficar el violín
library(ggplot2)
ggplot(data = v_d, aes(x = Survived, y = pred, group = Survived, fill = factor(Survived))) +
  geom_violin(trim = FALSE) + # Gráfico de violín
  geom_abline(aes(intercept = 0.5, slope = 0), color = "red", linetype = "dashed") + # Línea de referen
  theme_bw() +
  guides(fill = FALSE) +
  labs(
    title = 'Gráfico de Violín',
    subtitle = 'Random Forest - Probabilidades Predichas',
    x = 'Clase Real (Survived)',
```

```
    y = 'Probabilidad Predicha'
  )
```

```
## Warning: The ‘<scale>‘ argument of ‘guides()‘ cannot be ‘FALSE‘. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



La distribución para la clase “0” (No sobrevivió) está predominantemente concentrada cerca de 0, lo que indica que el modelo clasifica correctamente la mayoría de los casos negativos con alta confianza. Sin embargo, existe una pequeña proporción de casos negativos con probabilidades más altas, cercanas o incluso por encima del umbral de 0.5, lo que sugiere que el modelo confunde algunos de estos casos como positivos (falsos positivos). Por otro lado, para la clase “1” (Sobrevivió), las probabilidades predichas están concentradas en torno a 1, lo que refleja una clasificación precisa para la mayoría de los casos positivos.

Validación

Elección de un umbral de clasificación optimo.

Elección del umbral de clasificación (punto de corte)

Se trabaja con la base de datos de validación (*M_valid*) y se realiza el gráfico de la Exactitud, Sensibilidad, Especificidad y Precisión para distintos valores del umbral de clasificación. Se siguen los siguientes pasos:

1. Predicción en los datos de validación con el modelo elegido (en el ejemplo, el B)
2. Se definen los umbrales de clasificación: irán desde 0.05 hasta 0.95.
3. Se definen las métricas de la matriz de confusión para cada umbral de clasificación
4. Se prepara el conjunto de datos: se quitan los NA y se agrega la columna de umbrales de clasificaición
5. Se le da un formato a la base de datos para que pueda ser graficada más fácilmente.

Generación de base de datos para graficar

```
# Obtener las probabilidades predichas por el modelo Random Forest
pred_val = predict(D, newdata = M_valid, type = 'prob')[, 2]  # Probabilidades de la clase "1"
clase_real = M_valid$Survived

# Crear el dataframe para guardar las métricas
datosV = data.frame(accuracy = NA, recall = NA, specificity = NA, precision = NA)

# Iterar sobre los umbrales de clasificación
for (i in 5:95) {
  clase_predicha = ifelse(pred_val > i / 100, 1, 0)  # Clasificaciones basadas en el umbral

  # Crear matriz de confusión
  cm = table(clase_predicha, clase_real)

  # Verificar si la matriz de confusión es válida (2x2) antes de calcular métricas
  if (all(dim(cm) == c(2, 2))) {
    # Accuracy: Proporción de correctamente predichos
    datosV[i, 1] = (cm[1, 1] + cm[2, 2]) / sum(cm)
    # Recall: Tasa de positivos correctamente predichos
    datosV[i, 2] = cm[2, 2] / (cm[1, 2] + cm[2, 2])
    # Specificity: Tasa de negativos correctamente predichos
    datosV[i, 3] = cm[1, 1] / (cm[1, 1] + cm[2, 1])
    # Precision: Tasa de bien clasificados entre los clasificados como positivos
    datosV[i, 4] = cm[2, 2] / (cm[2, 1] + cm[2, 2])
  } else {
    datosV[i, ] = NA  # Si no es válida, llenamos con NA
  }
}

# Limpiar el dataframe de valores NA
datosV = na.omit(datosV)
datosV$umbral = seq(0.05, 0.95, 0.01)  # Agregar la columna de umbrales
```

Formato de datos

- Se crea la variable *métrica* que será una variable categórica para las métricas (Exactitud, Sensibilidad, Especificidad y Precisión)
- Los valores de las métricas se ponen en una sola columna.

- Se identifican las métricas para los distintos umbrales con la variable 'umbral'.

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

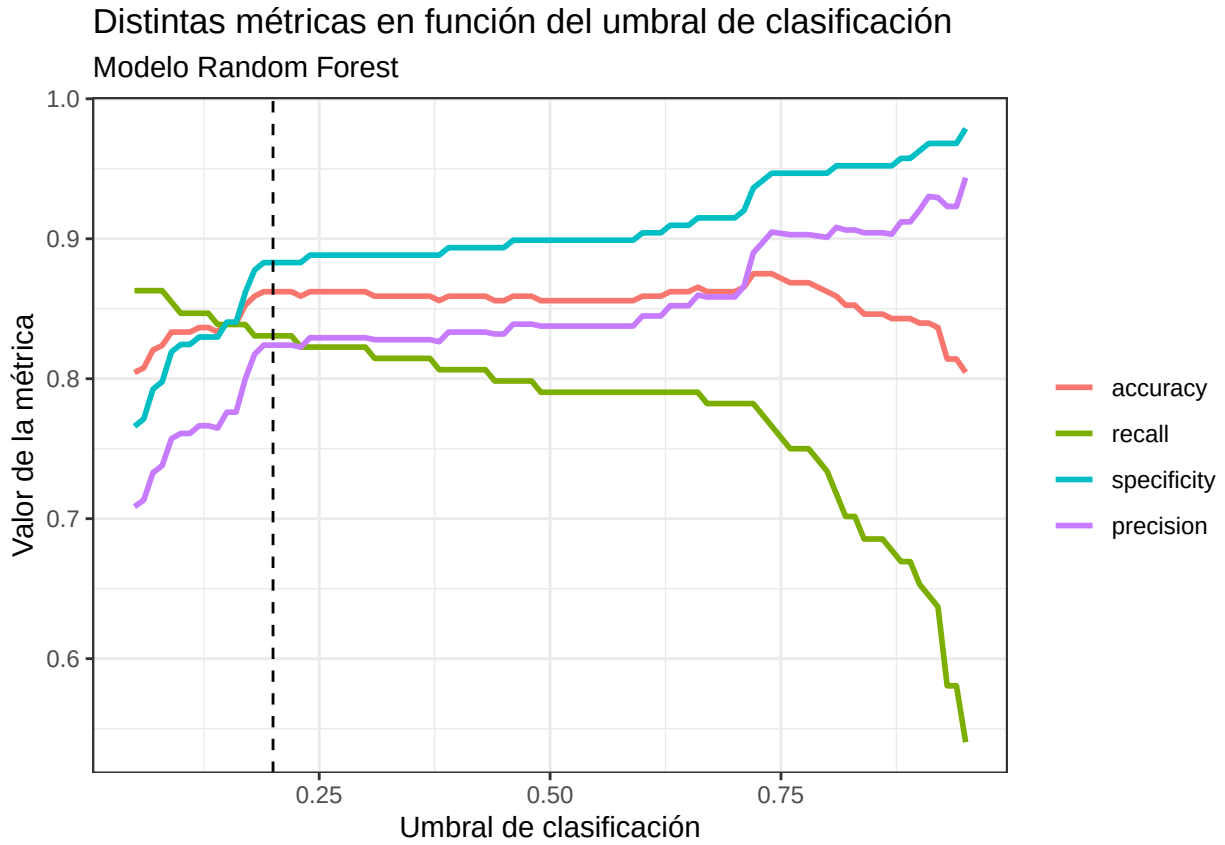
```
# Transformar el dataframe a formato largo
datosV_m <- reshape2::melt(datosV, id.vars = c('umbral'))

# Renombrar la columna que representa las métricas
colnames(datosV_m)[2] <- 'Metrica'
```

Gráfica

En la gráfica se define cuál es el mejor umbral de clasificación dependiendo de cuál métrica es más importante en el contexto del problema (Exactitud, Sensibilidad, Especificidad o Precisión). Si no hay una métrica de preferencia, se opta por escoger el máximo valor de que pueden tener estás métricas en conjunto. En cualquier caso da valores a u para mover el umbral de clasificación y observar como se comporta con respecto a las métricas.

```
library(ggplot2)

# Define el umbral de referencia (modifícalo según tu criterio)
u = 0.20

# Generar el gráfico con ggplot2
ggplot(data = datosV_m, aes(x = umbral, y = value, color = Metrica)) +
  geom_line(size = 1) +  # Línea para cada métrica
  theme_bw() +  # Tema blanco y negro
  labs(
    title = 'Distintas métricas en función del umbral de clasificación',
    subtitle = 'Modelo Random Forest',
    color = "",  # Etiqueta para la leyenda
    x = 'Umbral de clasificación',
    y = 'Valor de la métrica'
  ) +
  geom_vline(xintercept = u, linetype = "dashed", color = "black")  # Línea vertical en el umbral selec
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



El umbral óptimo se encuentra alrededor de 0.25, ya que ofrece un buen balance entre métricas clave: Recall se mantiene alta, asegurando la detección de la mayoría de los positivos; Specificity es aceptable, minimizando falsos positivos; y Accuracy alcanza su valor máximo (~0.8), reflejando un buen desempeño. ## Matriz de confusión con el umbral de clasificación optimo

De acuerdo al umbral seleccionado, calcula la matriz de confusión y las métricas obtenidas. Indica si mejora la predicción con respecto al umbral de u = 0.5, que es el que se maneja por default.

```
library(vcd)
```

```
## Loading required package: grid
```

```
##
## Attaching package: 'vcd'
```

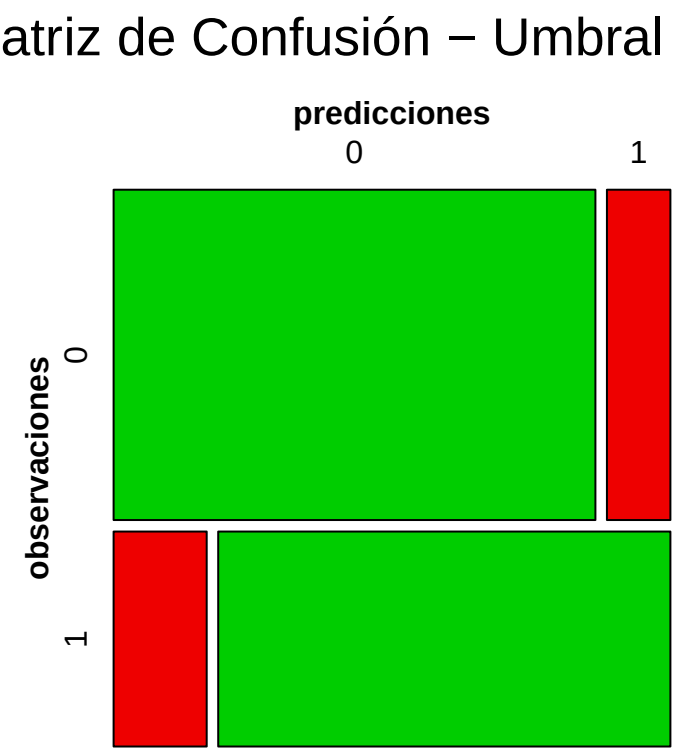
```
## The following object is masked from 'package:ISLR':
##
##     Hitters
```

```
# Definir el umbral seleccionado (cambia a 0.25 o el que desees evaluar)
umbral_seleccionado <- 0.25

# Generar las predicciones usando el umbral seleccionado
prediccionesV <- ifelse(rf_pred_prob_valid > umbral_seleccionado, yes = 1, no = 0)

# Crear la matriz de confusión
```

```r
M_Cv <- table(prediccionesV, M_valid$Survived, dnn = c("observaciones", "predicciones"))
print(M_Cv)
```

```
##              predicciones
## observaciones   0   1
##             0 167  22
##             1  21 102
```

```r
# Graficar el mosaico
mosaic(
  M_Cv,
  shade = TRUE,
  colorize = TRUE,
  gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)),
  main = paste("Matriz de Confusión - Umbral =", umbral_seleccionado)
)
```



La matriz de confusión resultante para un umbral de 0.25 refleja un buen desempeño general del modelo, con un alto número de verdaderos positivos y verdaderos negativos (áreas verdes predominantes).

Conclusiones

El análisis revela que las personas con mayor probabilidad de sobrevivir eran mujeres, de clase alta (primera clase), jóvenes y con pocos acompañantes. Estos hallazgos se fundamentan en los coeficientes del modelo

de regresión logística. Estas conclusiones fueron corroboradas con los resultados del modelo, donde mujeres y personas de clases altas fueron correctamente clasificadas como sobrevivientes con mayor frecuencia, alineándose con las prioridades y desigualdades de la época.

El umbral óptimo para clasificar la sobrevivencia fue 0.25, ya que maximiza la sensibilidad (Recall), minimizando los errores al identificar sobrevivientes. Este umbral detecta la mayoría de los casos positivos, aunque aumenta los falsos positivos, una decisión apropiada en este contexto donde es preferible sobreestimar a los sobrevivientes.