Máster en Ingeniería Software - Cloud, Datos y Gestión TI
Fundamentos de Ingeniería de Datos (FID)

Análisis y predicción de resultados del Mundial de Fútbol de Catar 2022

Escuela Técnica Superior de Ingeniería Informática

Carlos Núñez Arenas
Mariano Manuel Torrado Sánchez
Alejandro Santisteban Corchos
José Antonio Zamudio Amaya

Índice general

Índice general

1. Introducción

Alcance y objetivos

Github: https://github.com/joszamama/qatar-wc-predictor

Índice general

2. Contexto del problema

Mundial de Fútbol

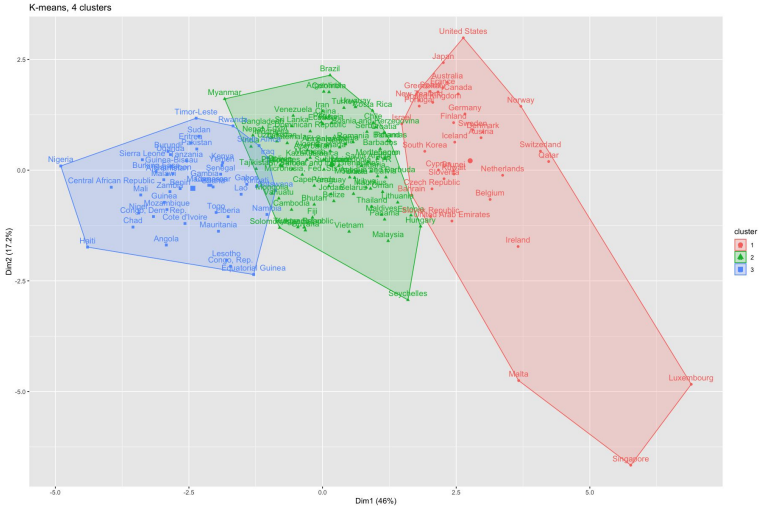
FIFA WORLD CUP
Qatar2022

2. Contexto del problema

Dos tipos de algoritmos

Aprendizaje supervisado: predicción del resultado de los partidos

Aprendizaje no supervisado: clustering de los equipos en base a estadísticas



2. Contexto del problema

Dataset inicial

FIFA World Cup 2022 ⚽🏆

International soccer matches and team strengths (1993-2022)

date <date>	home_team <chr>	away_team <chr>	home_team_continent <chr>	away_team_continent <chr>
1993-08-08	Bolivia	Uruguay	South America	South America
1993-08-08	Brazil	Mexico	South America	North America
1993-08-08	Ecuador	Venezuela	South America	South America
1993-08-08	Guinea	Sierra Leone	Africa	Africa
1993-08-08	Paraguay	Argentina	South America	South America
1993-08-08	Peru	Colombia	South America	South America
1993-08-08	Zimbabwe	Eswatini	Africa	Africa
1993-08-09	Guinea	Sierra Leone	Africa	Africa
1993-08-11	Faroe Islands	Norway	Europe	Europe
1993-08-11	Sweden	Switzerland	Europe	Europe

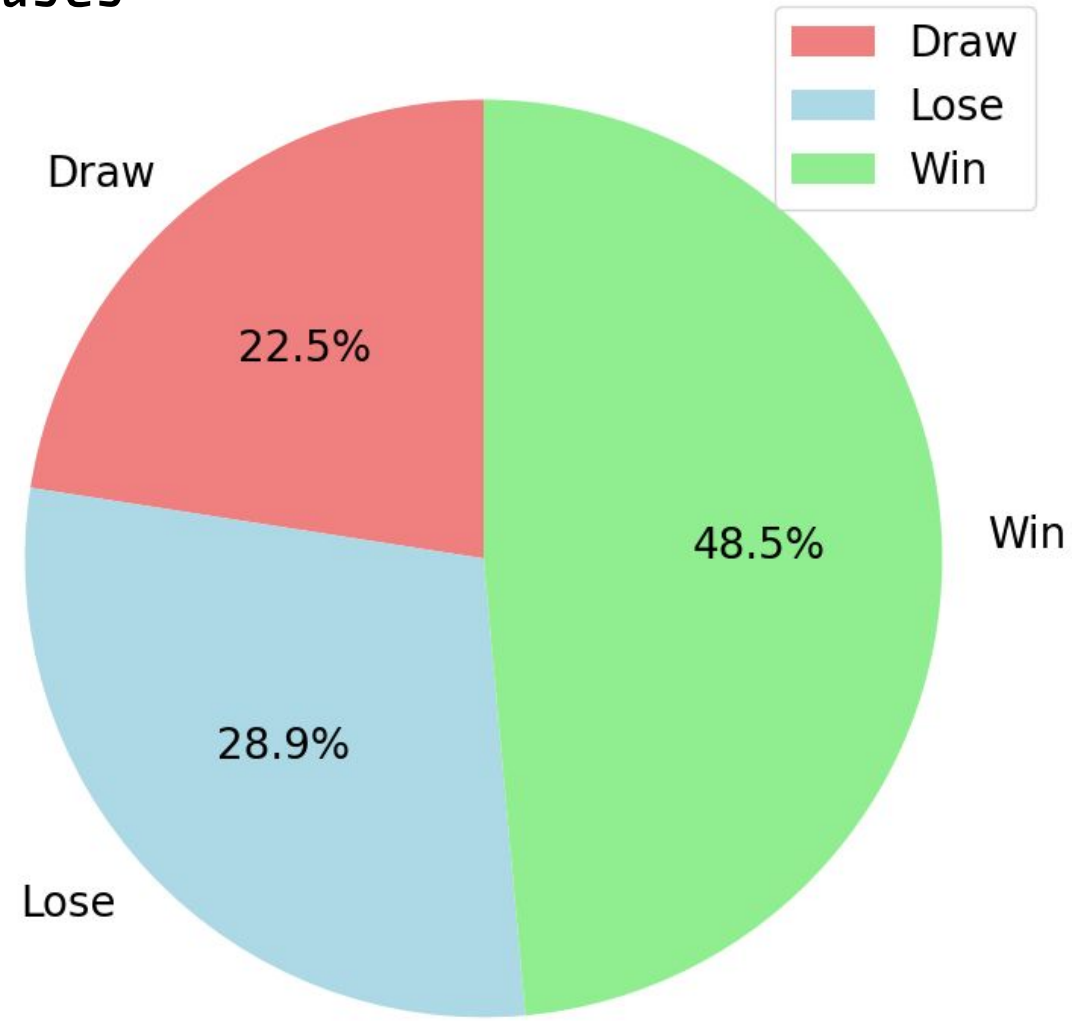
1-10 of 23,921 rows | 1-5 of 25 columns

Previous 1 2 3 4 5 6 … 100 Next

https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022

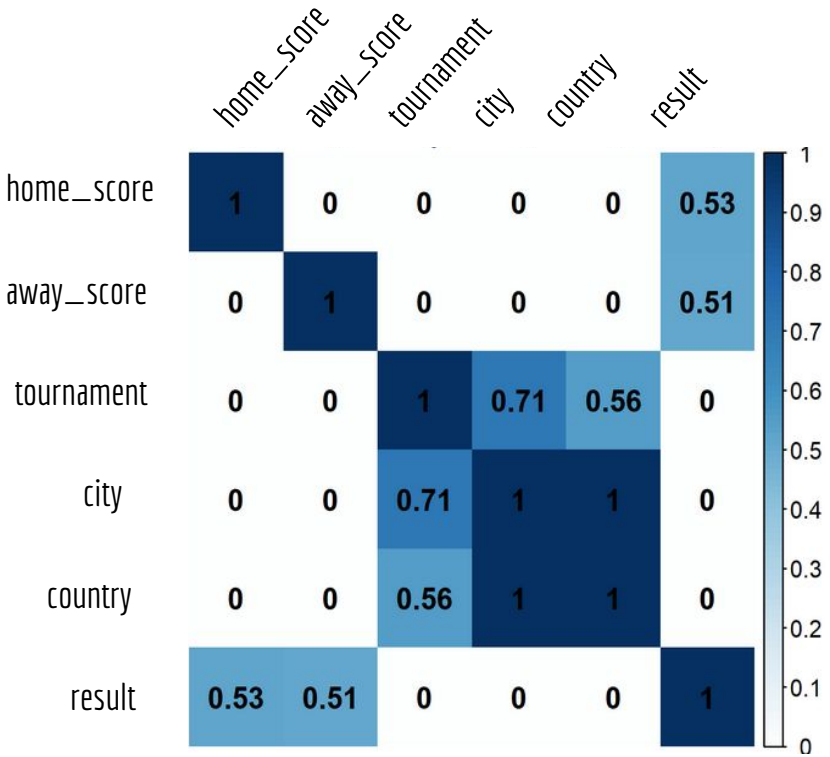
2. Contexto del problema

Balanceo de clases

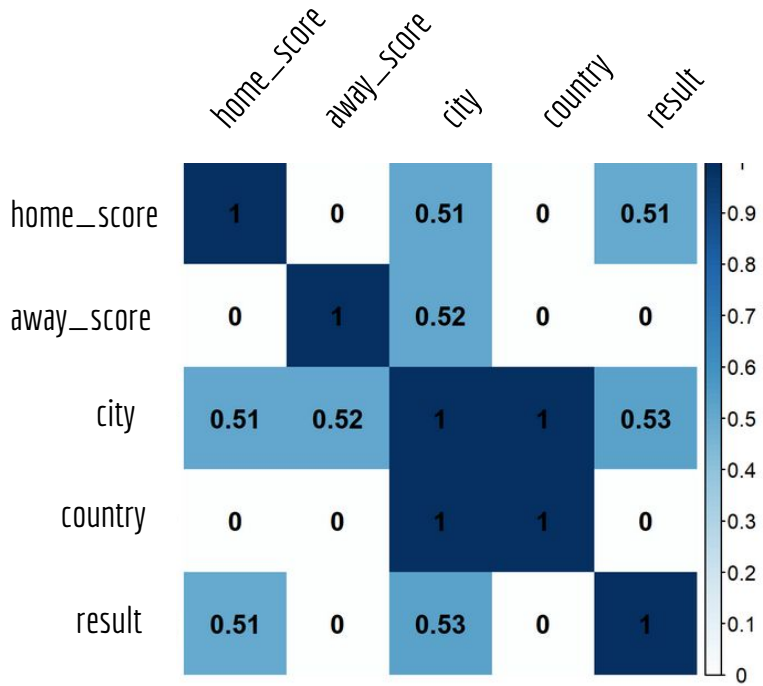


2. Contexto del problema

Correlación de atributos



Partidos oficiales

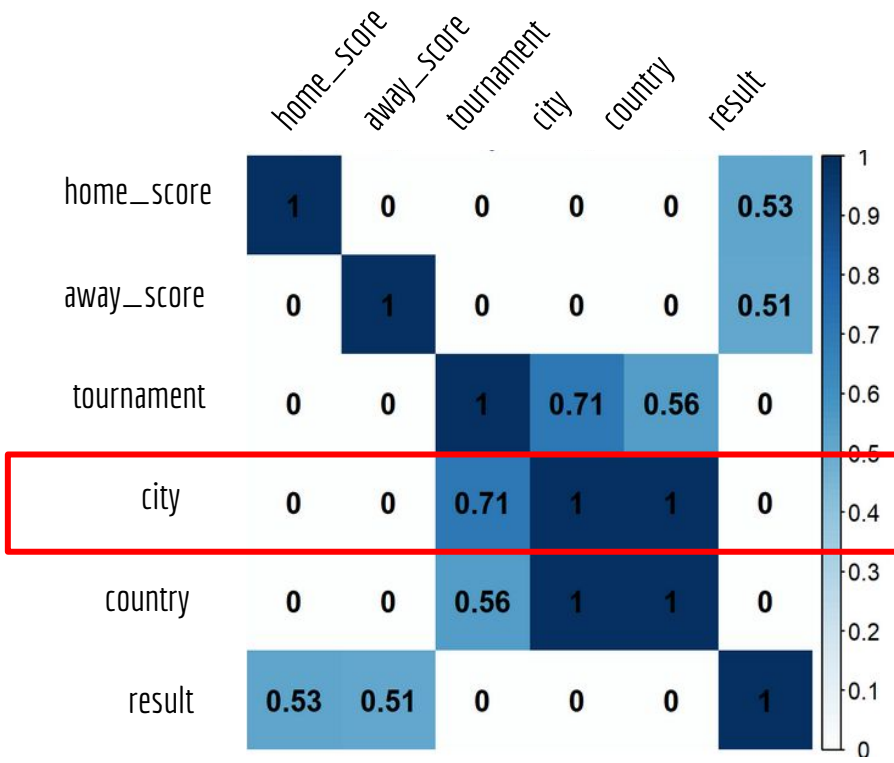


Partidos amistosos

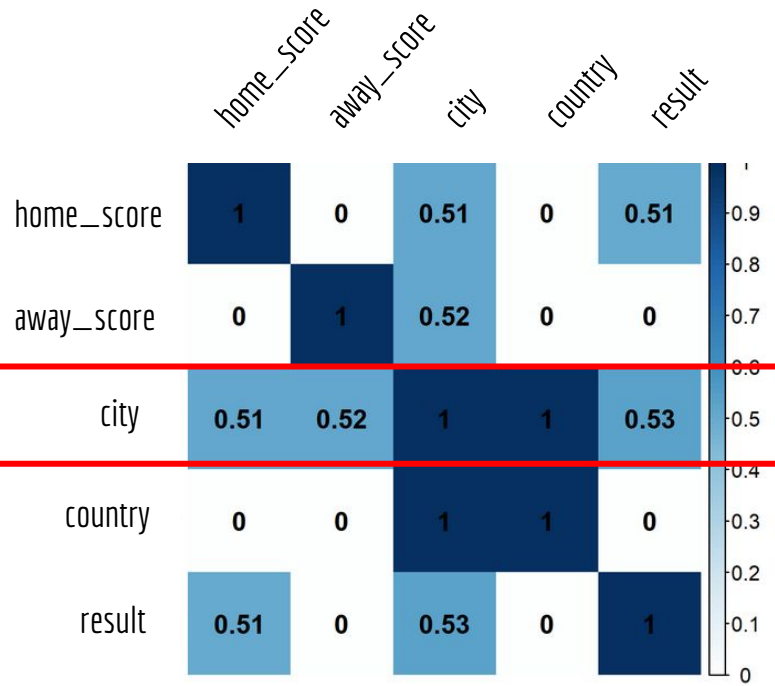
* Se muestran sólo las variables fuertemente correladas (corr › 0.5)

2. Contexto del problema

Correlación de atributos



Partidos oficiales



Partidos amistosos

07 - 31

* Se muestran sólo las variables fuertemente correladas (corr › 0.5)

Índice general

3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

3. Aprendizaje supervisado

Preselección de atributos

home_team <chr>	away_team <chr>	home_team_continent <chr>	away_team_continent <chr>
Equatorial Guinea	South Africa	Africa	Africa
Botswana	Zimbabwe	Africa	Africa
Gabon	Burkina Faso	Africa	Africa
Tunisia	Sudan	Africa	Africa
Nigeria	Angola	Africa	Africa
Senegal	Sudan	Africa	Africa
Tunisia	Côte d'Ivoire	Africa	Africa
Angola	Sierra Leone	Africa	Africa
Zambia	Namibia	Africa	Africa
Oman	Congo DR	Asia	Africa

1-10 of 9,198 rows | 1-4 of 14 columns

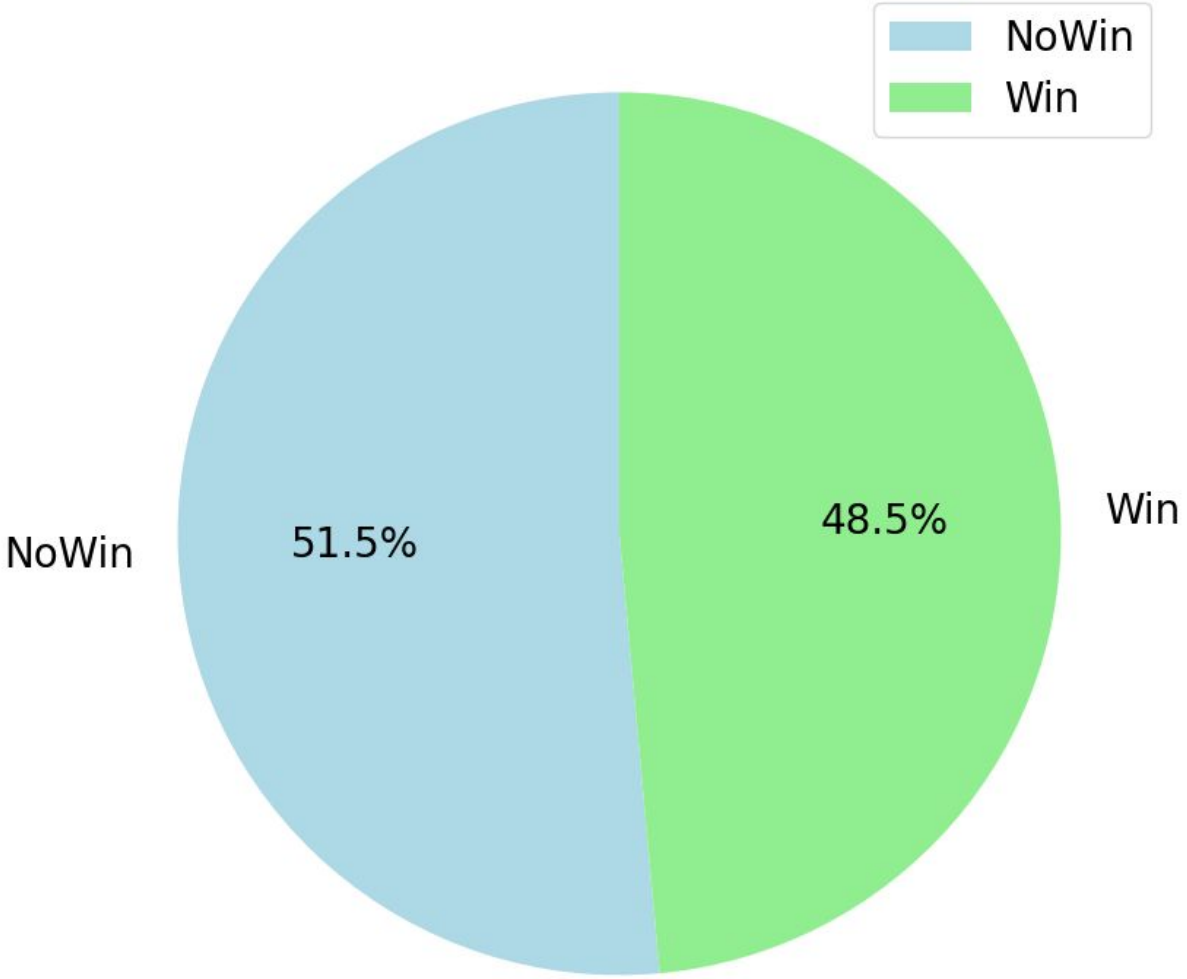
Previous 1 2 3 4 5 6 ... 100 Next

3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

3. Aprendizaje supervisado

Corrección del balanceo de clases

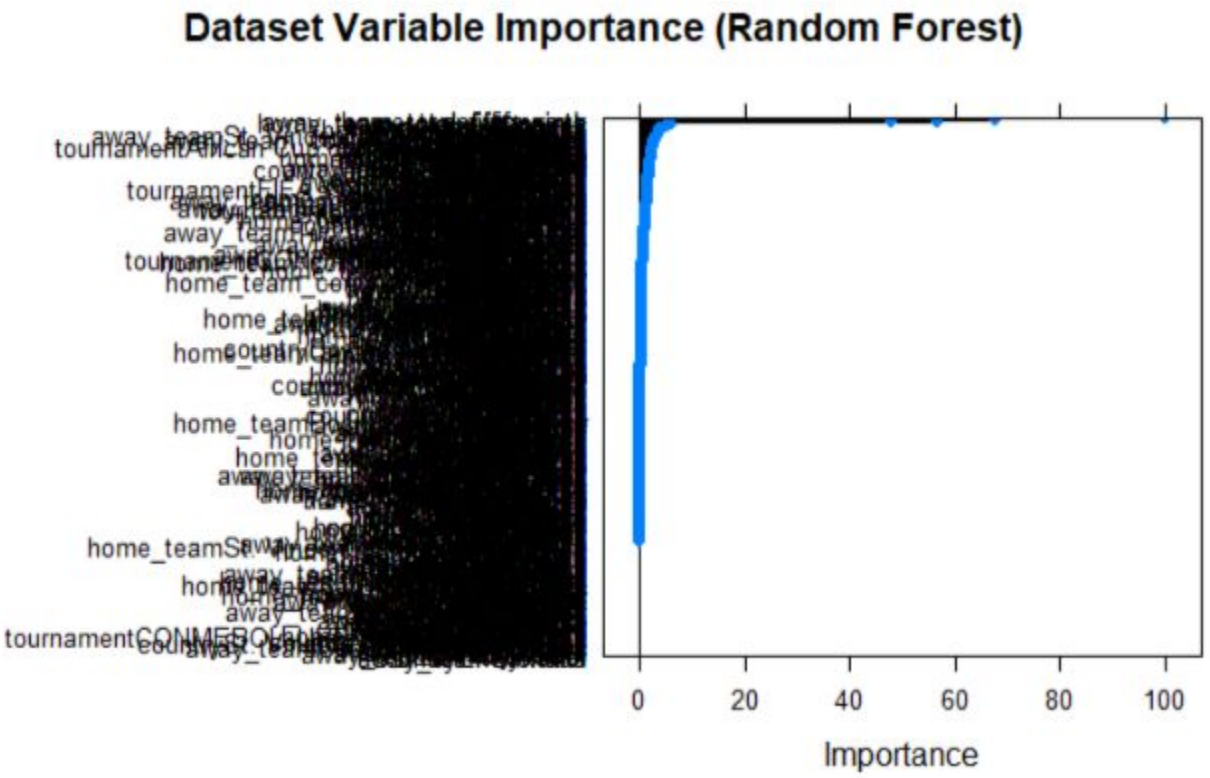


3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

3. Aprendizaje supervisado

Evaluación de los atributos - Random Forest



3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

3. Aprendizaje supervisado

Codificación Label Encode y re-evaluación

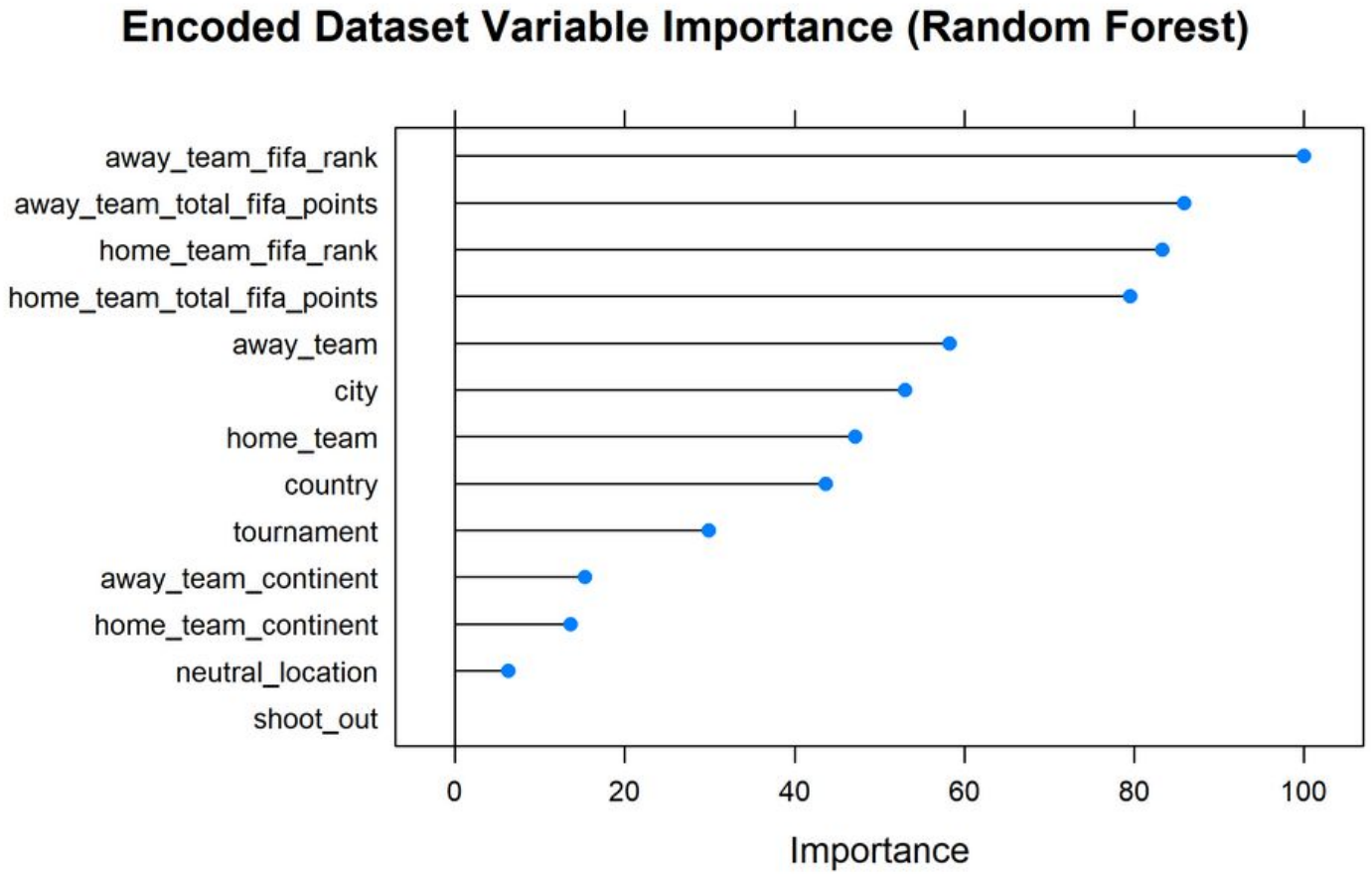
home_team <dbl>	away_team <dbl>	home_team_continent <dbl>	away_team_continent <dbl>	home_team_fifa_rank <dbl>
64	174	1	1	150
27	211	1	1	95
73	32	1	1	77
194	181	1	1	59
165	181	1	1	44
144	47	2	1	85
73	181	1	1	77
105	204	2	2	99
189	143	2	3	124
64	112	1	1	151

1-10 of 6,440 rows | 1-5 of 14 columns

Previous 1 2 3 4 5 6 ... 100 Next

3. Aprendizaje supervisado

Codificación Label Encode y re-evaluación

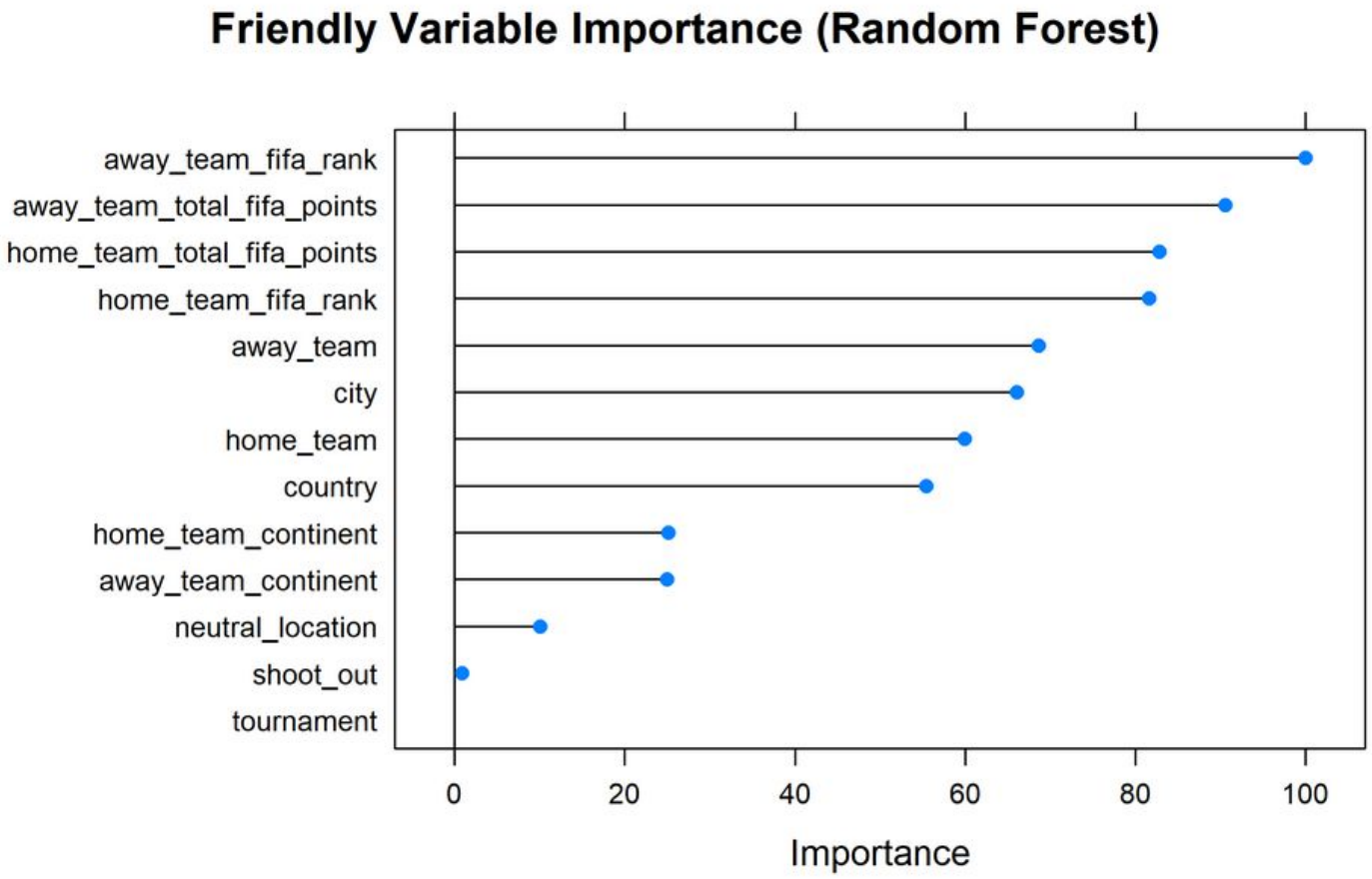


3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

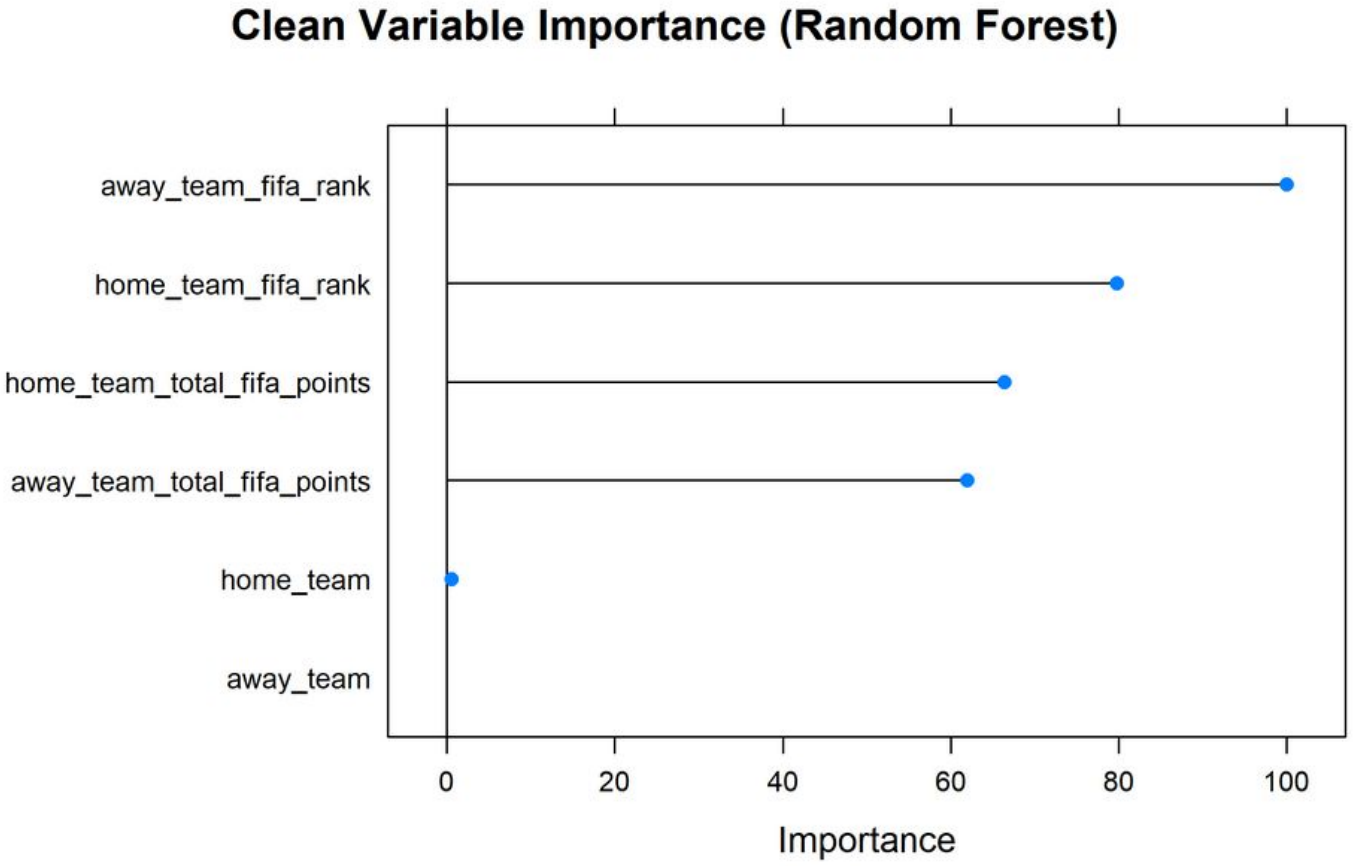
3. Aprendizaje supervisado

Influencia de la correlación en la selección de atributos



3. Aprendizaje supervisado

Influencia de la correlación en la selección de atributos



3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

3. Aprendizaje supervisado

Selección final de atributos

```
> summary(supervised_international_matches)
  home_team          away_team         home_team_continent away_team_continent home_team_fifa_rank away_team_fifa_rank home_team_total_fifa_points
 Length:9198        Length:9198        Length:9198         Length:9198         Min.   :  1.00      Min.   :  1.00      Min.   :   0.0
 Class :character   Class :character   Class :character    Class :character    1st Qu.: 35.00      1st Qu.: 40.00      1st Qu.: 365.0
 Mode  :character   Mode  :character   Mode  :character    Mode  :character    Median : 76.00      Median : 80.00      Median : 815.0
                                                                               Mean   : 82.62      Mean   : 86.46      Mean   : 818.6
                                                                               3rd Qu.:124.00      3rd Qu.:128.00      3rd Qu.:1247.8
                                                                               Max.   :211.00      Max.   :211.00      Max.   :2164.0
 away_team_total_fifa_points  tournament           city              country           neutral_location  shoot_out         home_team_result
 Min.   :   0.0              Length:9198        Length:9198        Length:9198        Mode :logical    Length:9198        Length:9198
 1st Qu.: 354.2              Class :character   Class :character   Class :character   FALSE:6730       Class :character   Class :character
 Median : 778.0              Mode  :character   Mode  :character   Mode  :character   TRUE :2468       Mode  :character   Mode  :character
 Mean   : 797.9
 3rd Qu.:1219.0
 Max.   :2164.0
```

3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

3. Aprendizaje supervisado

Entrenamiento de múltiples modelos

```
# Usar Naive-Bayes con los datos sin LabelEncoding.

supervised_naive_bayes <- train(home_team_result ~., data = supervised_training_set,
            method = "naive_bayes",
            trControl = supervised_control,
            metric="ROC")

# Luego probar con los datos ya codificados con los métodos de KNN y SVC

supervised_knn <- train(home_team_result ~., data = supervised_training_set_encoded,
            method = "knn",
            trControl = supervised_control,
            metric = "ROC",
            tuneGrid = data.frame(k = seq(11,85,by = 2)))


supervised_svm <- train(home_team_result ~., data = supervised_training_set_encoded,
            method = "svmLinear",
            trControl = supervised_control,
            preProcess = c("center","scale"),
            metric = "ROC",
            tuneGrid = expand.grid(C = seq(0.0001, 5, length = 25)))
```

3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

3. Aprendizaje supervisado

Evaluación de modelos

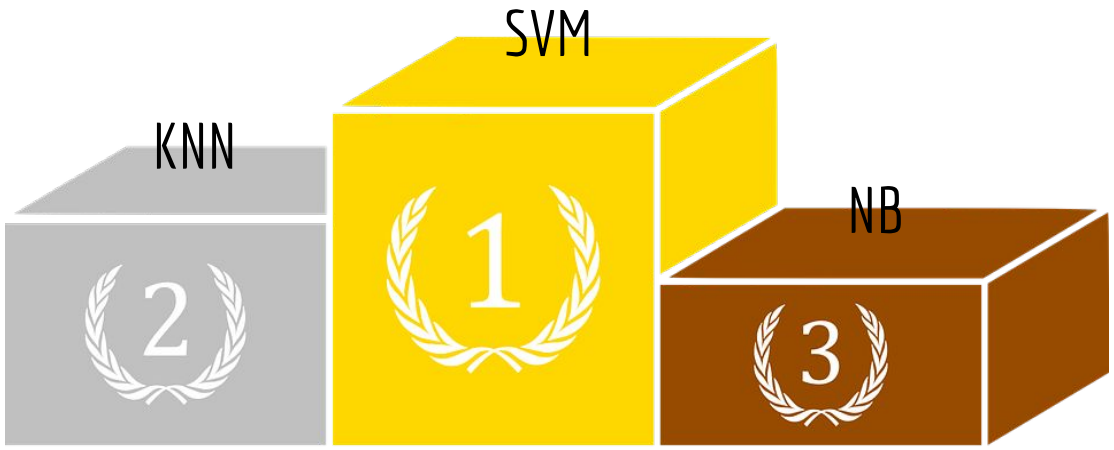
```
Models: NB, KNN, SVM
Number of resamples: 9

ROC
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
NB  0.7011568 0.7057580 0.7235435 0.7194141 0.7296536 0.7350087    0
KNN 0.7502912 0.7549656 0.7559645 0.7578538 0.7644398 0.7660866    0
SVM 0.7501832 0.7580664 0.7630947 0.7641330 0.7701709 0.7801118    0

Sens
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
NB  0.0000000 0.2889785 1.0000000 0.6987754 1.0000000 1.0000000    0
KNN 0.6599462 0.6787634 0.6841398 0.6926523 0.7137097 0.7177419    0
SVM 0.6854839 0.6935484 0.7190860 0.7177419 0.7432796 0.7540323    0

Spec
         Min.   1st Qu.     Median      Mean   3rd Qu.      Max. NA's
NB  0.0000000 0.0000000 0.01291248 0.3274350 0.9167862 1.0000000    0
KNN 0.6599713 0.6814921 0.68723099 0.6872310 0.6944046 0.7116212    0
SVM 0.6398852 0.6585366 0.66571019 0.6673043 0.6743185 0.7073171    0
```

3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

3. Aprendizaje supervisado

Selección del mejor clasificador



3. Aprendizaje supervisado

1. Preselección de atributos
2. Corrección del balanceo de clases
3. Evaluación de los atributos - Random Forest
4. Codificación Label Encode y re-evaluación
5. Influencia de la correlación en la selección de atributos
6. Selección final de atributos
7. Entrenamiento de múltiples modelos
8. Evaluación de modelos
9. Selección del mejor clasificador
10. Predicción de los resultados del mundial de Qatar 2022

3. Aprendizaje supervisado

Predicción de los resultados del mundial de Qatar 2022

home_team <chr>	away_team <chr>	group <chr>	home_team_result_pred <fctr>
Qatar	Ecuador	Group A	Lose
England	IR Iran	Group B	Draw
Senegal	Netherlands	Group A	Draw
USA	Wales	Group B	Draw
Argentina	Saudi Arabia	Group C	Win
Denmark	Tunisia	Group D	Draw
Mexico	Poland	Group C	Lose
France	Australia	Group D	Draw
Morocco	Croatia	Group F	Lose
Germany	Japan	Group E	Win

1-10 of 64 rows

Previous **1** 2 3 4 5 6 7 Next

3. Aprendizaje supervisado

Predicción de los resultados del mundial de Qatar 2022

Confusion Matrix and Statistics

Prediction	Reference Draw	Lose	Win
Draw	6	9	10
Lose	4	10	14
Win	0	2	9

Statistics by Class:

	Class: Draw	Class: Lose	Class: Win
Sensitivity	0.60000	0.4762	0.2727
Specificity	0.64815	0.5814	0.9355
Pos Pred Value	0.24000	0.3571	0.8182
Neg Pred Value	0.89744	0.6944	0.5472
Precision	0.24000	0.3571	0.8182
Recall	0.60000	0.4762	0.2727
F1	0.34286	0.4082	0.4091
Prevalence	0.15625	0.3281	0.5156
Detection Rate	0.09375	0.1562	0.1406
Detection Prevalence	0.39062	0.4375	0.1719
Balanced Accuracy	0.62407	0.5288	0.6041

3. Aprendizaje supervisado

Predicción de los resultados del mundial de Qatar 2022

Índice general

4. Aprendizaje no supervisado

1. Preprocesamiento
2. Imputación de valores
3. Entrenamiento de múltiples modelos
4. Evaluación de los modelos
5. Selección del mejor modelo
6. Visualización de los clústeres

4. Aprendizaje no supervisado

1. Preprocesamiento
2. Imputación de valores
3. Entrenamiento de múltiples modelos
4. Evaluación de los modelos
5. Selección del mejor modelo
6. Visualización de los clústeres

4. Aprendizaje no supervisado

Preprocesamiento

home_team	away_team	home_team_continent	away_team_continent	home_team_fifa_rank	away_team_fifa_rank	home_team_total_fifa_points	away_team_total_fifa_points	home_team_score	away_team_score	country	neutral_location	home_team_result	home_team_goalkeeper_score	away_team_goalk
Malaysia	Bangladesh	Asia	Asia	154	188	1035	903	4	1	Malaysia	FALSO	Win		
Bahrain	Turkmenistan	Asia	Asia	89	134	1262	1117	1	0	Malaysia	VERDADERO	Win		
India	Hong Kong	Asia	Asia	106	147	1174	1053	4	0	India	FALSO	Win	64.0	
Myanmar	Singapore	Asia	Asia	152	158	1044	1012	2	6	Kyrgyz Republic	VERDADERO	Lose		
El Salvador	USA	North America	North America	74	15	1331	1633	1	1	El Salvador	FALSO	Draw	61.0	77.0
Jamaica	Mexico	North America	North America	64	9	1378	1658	1	1	Jamaica	FALSO	Draw	76.0	80.0
Kyrgyz Republic	Tajikistan	Asia	Asia	95	114	1218	1159	0	0	Kyrgyz Republic	FALSO	Draw		
Afghanistan	Cambodia	Asia	Asia	150	171	1049	966	2	2	India	VERDADERO	Draw		
Indonesia	Nepal	Asia	Asia	159	168	1001	978	7	0	Kuwait	VERDADERO	Win		
Kuwait	Jordan	Asia	Asia	146	91	1059	1259	0	3	Kuwait	FALSO	Lose		
Costa Rica	New Zealand	North America	Oceania	31	101	1503	1206	1	0	Qatar	VERDADERO	Win	88.0	68.0
Palestine	Philippines	Asia	Asia	100	133	1208	1117	4	0	Mongolia	VERDADERO	Win		72.0
Uzbekistan	Thailand	Asia	Asia	83	111	1286	1167	2	0	Uzbekistan	FALSO	Win		66.0
Maldives	Sri Lanka	Africa	Asia	156	205	1025	842	1	0	Uzbekistan	VERDADERO	Win		
Mongolia	Yemen	Asia	Asia	186	151	911	1046	2	0	Mongolia	FALSO	Win		
Haiti	Guyana	North America	South America	90	174	1261	961	6	0	Dominican Republic	VERDADERO	Win	66.0	51.0
Moldova	Andorra	Europe	Europe	180	153	932	1040	2	1	Moldova	FALSO	Win	65.0	
Turkey	Lithuania	Europe	Europe	43	138	1461	1092	2	0	Turkey	FALSO	Win	79.0	71.0
Luxembourg	Faroe Islands	Europe	Europe	94	124	1229	1137	2	2	Luxembourg	FALSO	Draw	69.0	
Liechtenstein	Latvia	Europe	Europe	192	135	895	1105	0	2	Liechtenstein	FALSO	Lose		65.0
Korea Republic	Egypt	Asia	Africa	29	32	1519	1500	4	1	Korea Republic	FALSO	Win	75.0	
Japan	Tunisia	Asia	Africa	23	35	1553	1499	0	3	Japan	FALSO	Lose	73.0	
Chile	Ghana	South America	Africa	28	60	1526	1387	0	0	Japan	VERDADERO	Lose	79.0	74.0
Romania	Montenegro	Europe	Europe	48	70	1446	1342	0	3	Romania	FALSO	Lose	77.0	65.0
Netherlands	Wales	Europe	Europe	10	18	1658	1588	3	2	Netherlands	FALSO	Win	81.0	74.0
Germany	Italy	Europe	Europe	12	6	1650	1723	5	2	Germany	FALSO	Win	90.0	89.0
England	Hungary	Europe	Europe	5	40	1761	1466	0	4	England	FALSO	Lose	83.0	85.0
Poland	Belgium	Europe	Europe	26	2	1544	1827	0	1	Poland	FALSO	Lose	87.0	89.0
Bosnia and Herzegovina	Finland	Europe	Europe	59	57	1388	1406	3	2	Bosnia and Herzegovina	FALSO	Win	76.0	83.0
Ukraine	Republic of Ireland	Europe	Europe	27	47	1535	1449	1	1	Poland	VERDADERO	Draw	75.0	75.0
Armenia	Scotland	Europe	Europe	92	39	1245	1472	1	4	Armenia	FALSO	Lose	52.0	77.0
Honduras	Canada	North America	North America	82	38	1289	1479	2	1	Honduras	FALSO	Win		76.0
Nicaragua	Bahamas	North America	North America	144	201	1062	858	4	0	Nicaragua	FALSO	Win		
Trinidad and Tobago	St. Vincent and the Grenadines	South America	North America	103	175	1203	960	4	1	Trinidad and Tobago	FALSO	Win	56.0	
Morocco	Liberia	Africa	Africa	24	149	1551	1050	2	0	Morocco	FALSO	Win	82.0	
Australia	Peru	Oceania	South America	42	22	1462	1562	0	0	Qatar	VERDADERO	Win	77.0	74.0
S�o Tom� and Pr�ncipe	Nigeria	Africa	Africa	183	30	917	1504	0	10	Morocco	VERDADERO	Lose		74.0
Sierra Leone	Guinea-Bissau	Africa	Africa	108	115	1173	1158	2	2	Guinea	VERDADERO	Draw		
Azerbaijan	Belarus	Europe	Europe	129	93	1127	1243	2	0	Azerbaijan	FALSO	Win		
Kazakhstan	Slovakia	Europe	Europe	125	45	1134	1454	2	1	Kazakhstan	FALSO	Win		81.0
Albania	Estonia	Europe	Europe	66	110	1371	1169	0	0	Albania	FALSO	Draw	80.0	
Iceland	Israel	Europe	Europe	63	76	1380	1305	2	2	Iceland	FALSO	Draw	70.0	70.0
Guatemala	Dominican Republic	North America	North America	118	155	1147	1029	2	0	Guatemala	FALSO	Win		

4. Aprendizaje no supervisado

Preprocesamiento

team	fifa_rank	total_fifa_points	goalkeeper_score	mean_defense_score	mean_offense_score	mean_midfield_score
Tunisia	35	1499	64.3781512605042	72.22488479262672	71.39493087557602	72.77281105990784
Ghana	60	1387	68.21134020618557	73.4908256880734	74.53807339449541	78.92660550458716
Belgium	2	1827	82.44329896907216	81.2360824742268	81.72319587628864	81.53298969072165
Wales	18	1588	73.81656804733728	74.66568047337279	74.3414201183432	78.3526627218935
USA	15	1633	81.07829181494662	75.00427046263346	76.0338078291815	77.09928825622777
Mexico	9	1658	79.18849840255591	76.98466453674123	79.17284345047923	77.7884984025559
Croatia	16	1621	78.67632850241546	78.8096618357488	80.08550724637682	81.20241545893721
Canada	38	1479	71.99319727891157	68.62585034013605	71.236690647482	72.93401360544217
Serbia	25	1547	76.43820224719101	80.24662921348315	78.25786516853931	80.30449438202248
Portugal	8	1674	81.80973451327434	82.89159292035399	84.93362831858407	83.59026548672567
Poland	26	1544	82.10502283105023	75.83470319634704	79.27488584474885	76.73013698630137
Germany	12	1650	89.05371900826447	84.6888429752066	83.60206611570247	85.79669421487601
France	3	1789	86.86808510638298	84.08127659574467	85.7	86.15872340425533
Spain	7	1709	88.78481012658227	85.50928270042195	85.8746835443038	87.1278481012658
Cameroon	37	1480	77.68780487804878	77.87170731707317	78.71317073170732	76.86439024390242
Netherlands	10	1658	83.23287671232876	80.73470319634704	85.47853881278539	83.70913242009134
England	5	1761	83.24074074074075	84.69305555555556	85.11898148148147	84.51111111111112
Brazil	1	1832	86.26122448979592	85.92040816326531	86.53102040816326	85.34367346938775
Denmark	11	1653	79.66494845360825	78.64845360824742	77.47835051546392	78.99742268041237
Uruguay	13	1635	79.36018957345972	79.64644549763034	83.57725118483413	78.51469194312797
Ecuador	46	1452	71.45086705202313	70.73036649214659	75.36473988439306	74.93815028901734
Switzerland	14	1635	80.17948717948718	78.14717948717949	76.52717948717948	78.71846153846155
Costa Rica	31	1503	78.62439024390244	69.99879032258065	71.24524714828897	70.52016129032258
Morocco	24	1551	72.08673469387755	75.72857142857143	77.05612244897958	76.47193877551021
Argentina	4	1765	80.69396551724138	82.98189655172415	88.26034482758621	84.42801724137932
Australia	37	1486	78.35960591133005	72.60344827586206	74.43054187192119	74.16798029556651
Korea Republic	29	1522	73.84824902723736	72.6466926070039	75.08949416342413	75.06031128404669
Senegal	18	1587	72.35828877005348	77.5144385026738	79.81818181818181	76.35026737967914
Japan	23	1549	70.38150289017341	72.69239130434782	72.84112903225807	76.91156716417912
Saudi Arabia	53	1433	70.91803278688525	71.3360655737705	70.51967213114754	73.08524590163934
IR Iran	21	1572	68.36065573770492	68.97241379310346	70.98275862068965	69.9951219512195
Qatar	46	1437				

4. Aprendizaje no supervisado

1. Preprocesamiento
2. Imputación de valores
3. Entrenamiento de múltiples modelos
4. Evaluación de los modelos
5. Selección del mejor modelo
6. Visualización de los clústeres

4. Aprendizaje no supervisado

Imputación de valores

qatar [28]

	team	fifa_rank	total_fifa_points	goalkeeper_score	mean_defense_score	mean_offense_score	mean_midfiel
1	Qatar	46	1437	NA	NA	NA	

Table Chart 1 row

qatar [33]

	team	fifa_rank	total_fifa_points	goalkeeper_score	mean_defense_score	mean_offense_score	mean_midfiel
1	Qatar	46	1437	78.6763	77.5144	78.2579	

Table Chart 1 row

```
# Interpolar valores nulos del dataset

qatar = teams_data[nrow(teams_data), ]
bagMissing <- preProcess(teams_data, method = "medianImpute")
teams_data[nrow(teams_data), ] <- predict(bagMissing, qatar)
rownames(teams_data) <- teams_data$team
```

4. Aprendizaje no supervisado

1. Preprocesamiento
2. Imputación de valores
3. Entrenamiento de múltiples modelos
4. Evaluación de los modelos
5. Selección del mejor modelo
6. Visualización de los clústeres

4. Aprendizaje no supervisado

Entrenamiento de múltiples modelos

```
# Normaliza las columnas que quieres usar para agrupar los equipos
teams_data_normalized <- scale(teams_data[, colnames(teams_data)[-1]])

# k-medias con k = 3 clústeres y 10 iteraciones
kmeans_clustering <- kmeans(teams_data_normalized, 3, nstart = 10)

# C-medias con k=3 clústeres utilizando distancia euclídea
Cmeans_euclidean_clustering <- fanny(teams_data, diss=FALSE, k=3,
                                     metric="euclidean", stand=FALSE)

# C-medias con k=3 clústeres utilizando distancia manhattan
Cmeans_sqeuclidean_clustering <- fanny(teams_data, diss=FALSE, k=3,
                                     metric="SqEuclidean", stand=FALSE)
```

4. Aprendizaje no supervisado

1. Preprocesamiento
2. Imputación de valores
3. Entrenamiento de múltiples modelos
4. Evaluación de los modelos
5. Selección del mejor modelo
6. Visualización de los clústeres

4. Aprendizaje no supervisado

Evaluación de los modelos

```
rank <- team_cluster$team_cluster

# Compute the ARI between the kmeans and FCM clusters
kmeans_ari <- ARI(kmeans_clustering$cluster, rank)
cmeans_e_ari <- ARI(Cmeans_euclidean_clustering$cluster, rank)
cmeans_sq_ari <- ARI(Cmeans_sqeuclidean_clustering$cluster, rank)

# Compute the AMI between the kmeans and FCM clusters
kmeans_ami <- AMI(kmeans_clustering$cluster, rank)
cmeans_e_ami <- AMI(Cmeans_euclidean_clustering$cluster, rank)
cmeans_sq_ami <- AMI(Cmeans_sqeuclidean_clustering$cluster, rank)

# Compute the Silhouette for every model
suppressWarnings({
  kmeans_silhouette <- mean(as.data.frame(
            silhouette(kmeans_clustering$cluster, dist(team_cluster)))$sil_width)
  cmeans_e_silhouette <- mean(as.data.frame(
            silhouette(Cmeans_euclidean_clustering))$sil_width)
  cmeans_sq_silhouette <- mean(as.data.frame(
            silhouette(Cmeans_sqeuclidean_clustering))$sil_width)
})

cat("Comparativa de ARIs")
```

4. Aprendizaje no supervisado

Evaluación de los modelos

Kmeans ARI = 0.366336633663366
Cmeans ARI con distancia euclidea = 0.485193621867882
Cmeans ARI con distancia euclidea cuadrática = 0.450759707371975

Kmeans AMI = 0.386837891325094
Cmeans AMI con distancia euclidea = 0.545896378875695
Cmeans AMI con distancia euclidea cuadrática = 0.51420268909183

Kmeans Silhouette mean = 0.344353505291005
Cmeans Silhouette mean con distancia euclidea = 0.561040592165388
Cmeans Silhouette mean con distancia euclidea cuadrática = 0.750412203967254

4. Aprendizaje no supervisado

1. Preprocesamiento
2. Imputación de valores
3. Entrenamiento de múltiples modelos
4. Evaluación de los modelos
5. Selección del mejor modelo
6. Visualización de los clústeres

4. Aprendizaje no supervisado

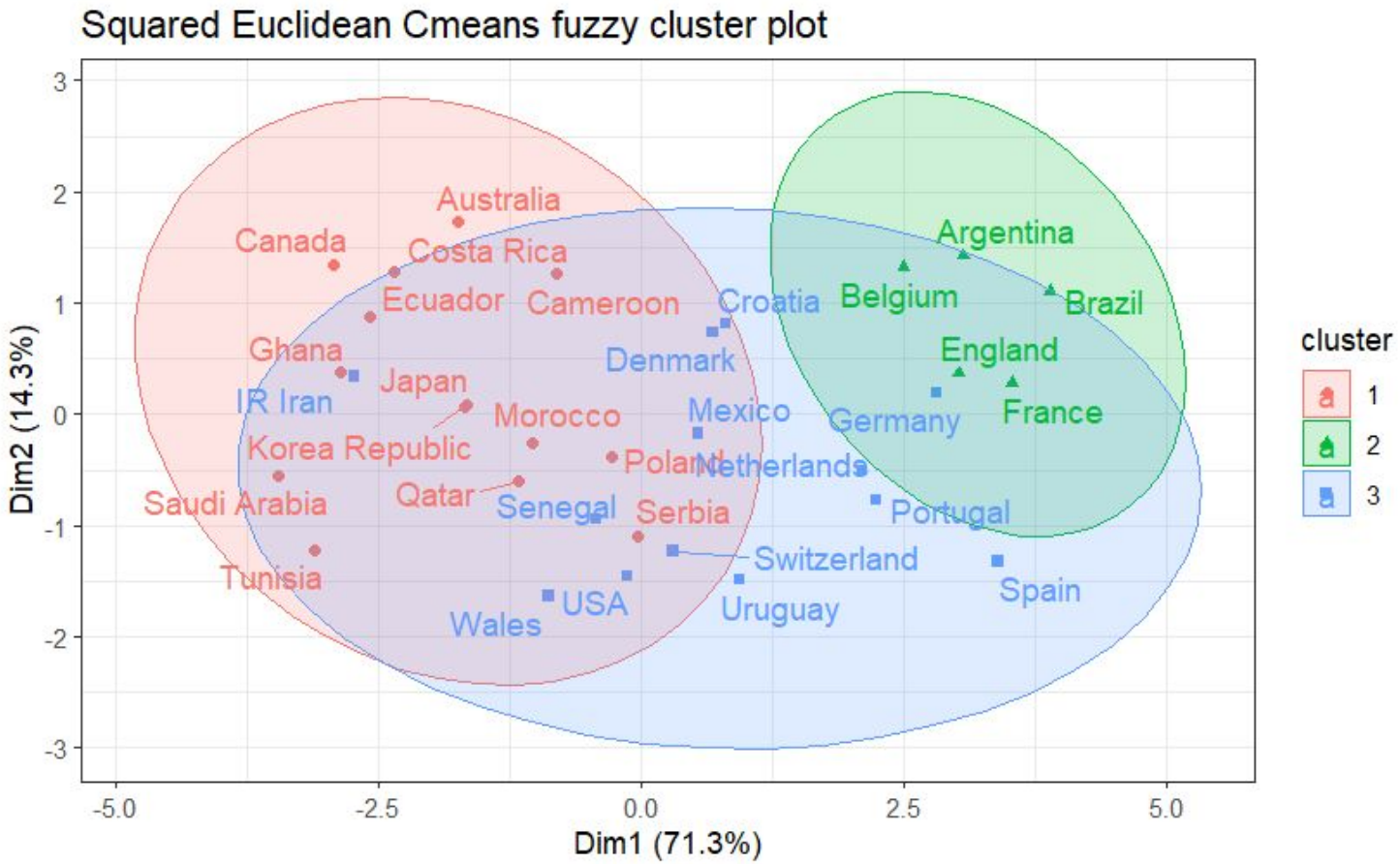
Selección del mejor modelo

4. Aprendizaje no supervisado

1. Preprocesamiento
2. Imputación de valores
3. Entrenamiento de múltiples modelos
4. Evaluación de los modelos
5. Selección del mejor modelo
6. Visualización de los clústeres

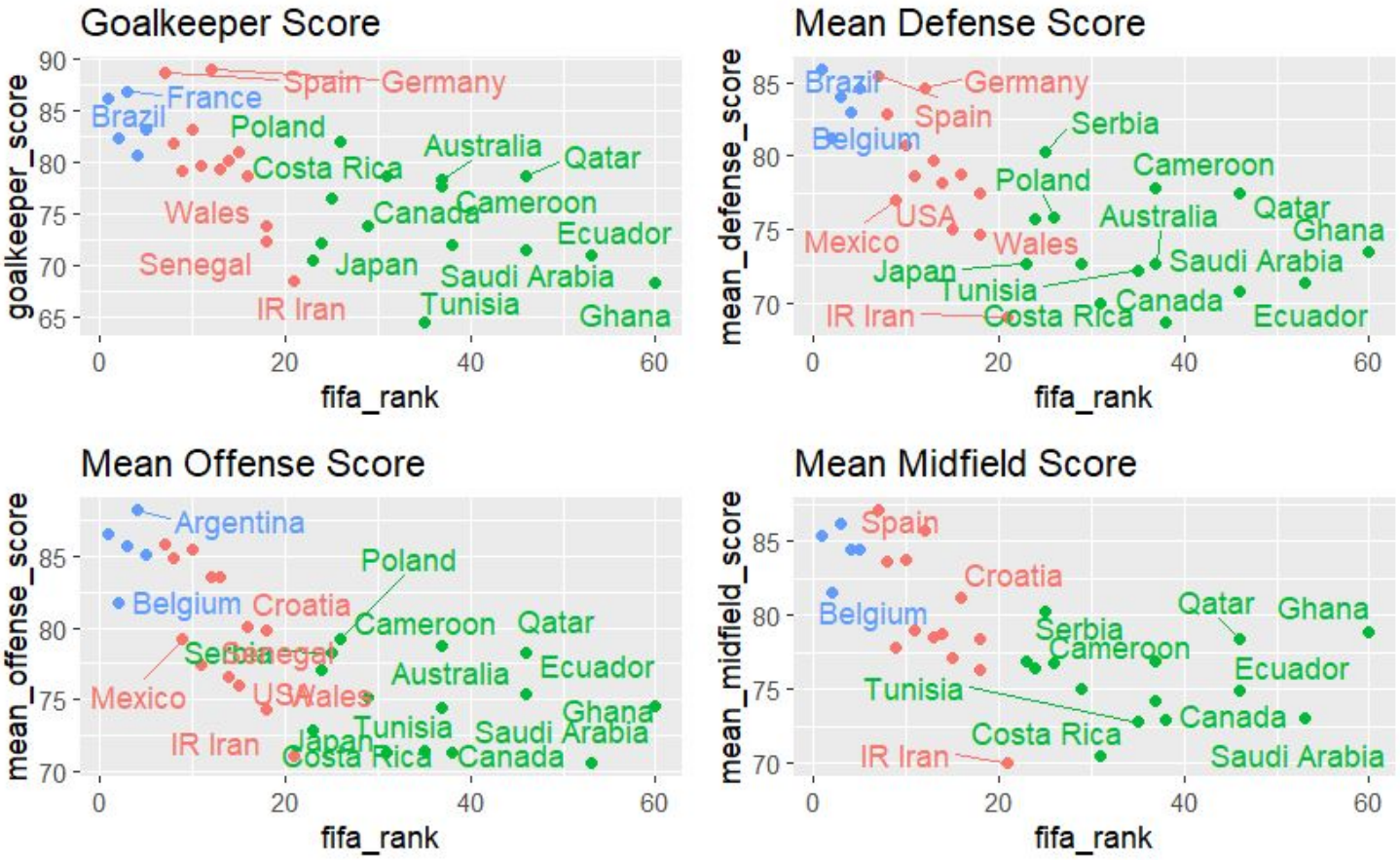
4. Aprendizaje no supervisado

Visualización de los clústeres



4. Aprendizaje no supervisado

Visualización de los clústeres



Índice general

5. Conclusiones

Conclusiones

Es crucial trabajar bien el dataset y estructurar los datos de manera adecuada antes de pasar a implementar algoritmos de aprendizaje.

Puede no existir un “mejor” modelo. Dependiendo de en qué casos, puede interesarnos más un clasificador que otro, incluso cuando en la evaluación haya obtenido peor puntuación.

Máster en Ingeniería Software - Cloud, Datos y Gestión TI
Fundamentos de Ingeniería de Datos (FID)

Análisis y predicción de resultados del Mundial de Fútbol de Catar 2022

Escuela Técnica Superior de Ingeniería Informática

Carlos Núñez Arenas
Mariano Manuel Torrado Sánchez
Alejandro Santisteban Corchos
José Antonio Zamudio Amaya