Proyecto de estadística

Hamilton Smith Gómez Osorio
Universidad EAFIT
Medellín, Colombia

Verónica Mendoza Iguarán
Universidad EAFIT
Medellín, Colombia

Pablo Alberto Osorio Marulanda
Universidad EAFIT
Medellín, Colombia

Mariana Uribe Orrego
Universidad EAFIT
Medellín, Colombia

29 de mayo de 2020

1. Introducción

En la búsqueda de afianzar los conociminetos adquiridos en el curso de estadística multivariante de la Universidad EAFIT enfocado en la analítica de datos, se plantea en este trabajo un desarrollo práctico en el cual, por medio de un análisis estadístico a una base de datos, se podrá genarar discusión respecto al comportamiento de la muestra de la población a analizar, a su vez que se plantean conclusiones respecto a lo observado. En este caso, se estará trabajando con una base de datos llamada 'Pay for Play: Are Baseball Salaries Based on Performance?' que se encontró en la página Journal of Statistics Education, la cual considera como población de interés el conjunto de jugadores de las Grandes Ligas de Béisbol que jugaron al menos un juego en las temporadas de 1991 y 1992, excluyendo los lanzadores. Este conjunto de datos contiene los salarios de 1992 para esa población, junto con las medidas de rendimiento para cada jugador desde 1991 y contiene cuatro variables categóricas que indican la libertad de cada jugador para moverse a otros equipos y se tiene 337 observaciones.

Las variables que se encuentran son :

- Salario en miles de dólares
- Promedio de bateo
- Porcentaje en base (OBP)
- Número de carreras
- Número de hits
- Número de dobles
- Número de triples
- Número de home runs

- Números de carreras bateadas(RBI)
- Número de caminatas
- Número de ponches- strike outs
- Número de bases robadas
- Número de errores
- Indicador de eligibilidad de agencia libre
- Indicador de agente libre en 1991/2
- Indicador de elegibilidad de arbitraje
- Indicador de arbitraje en 1991/2
- Nombre del jugador

2. Análisis exploratorio de los datos

A continuación se dará una explicación del trabajo que se llevó a cabo con esta base de datos y de los resultados que se obtuvieron a partir del análisis de ella. Para la implementación se utilizó el sowtfare de matlab.

Al comenzar el análisis a los datos, se percibió que la primera variable tenía una dimensión más grande que las demás, por lo que, para evitar distorsionar los resultados, se decidió dividir esta variable por diez, y así esta tendrá como unidad diez miles de dólares.

Lo primero que se realizó fueron los estadísticos descriptivos a los datos sin tener en cuenta variables binarias, tales como el vector de medias, vector de desviaciones estándar, vector de asimetrías, vector de curtosis, vector de coeficientes de variación (es decir, la desviación estándar sobre la media), vector de medianas, vector de medas, vector de medas sobre medianas. Los valores encontrados se presentan en la siguiente tabla.

	Salario	Promedio de bateo	Porcentaje en base	# de carreras	# de hits	# dobles	# triples	# home runs	# de carreras bateadas	# de caminatas	# de ponches	# de bases robadas	# de errores
Media	124.8528	0.257824926	0.323973294	46.6973	92.83	16.6736	2.33828	9.097923	44.0207715	35.0178042	56.7062	8.24629	6.7715
Desviación estándar	124.0013	0.039546336	0.047132057	29.0202	51.9	10.452	2.54334	9.289934	29.5594062	24.8424735	33.8288	11.6648	5.9275
Asimetría	1.157791	-0.43595325	-0.711796668	0.3988	0.147	0.61372	1.7395	1.169022	0.60353835	0.92635751	0.68384	2.57003	1.3027
Curtosis	3.644234	7.224025198	6.642620263	2.2908	2.022	2.80032	6.78874	3.772737	2.57373675	3.63915576	3.03157	11.2619	4.5065
Coeficiente variación	0.99318	0.153384457	0.145481304	0.62145	0.559	0.62686	1.0877	1.021105	0.67148769	0.70942408	0.59656	1.41455	0.8754
Mediana	74	0.26	0.323	41	91	15	2	6	39	30	49	4	5
Meda	59.3	0.021	0.029	23	42	7	1	5	20	16	23	4	3
Meda/mediana	0.801351	0.080769231	0.089783282	0.56098	0.462	0.46667	0.5	0.833333	0.51282051	0.53333333	0.46939	1	0.6

Figura 1: Tabla de estadísticos relevantes

Luego se les realizó el plotmatrix que permite ver el comportamiento entre variables y se presenta en la siguiente gráfica.

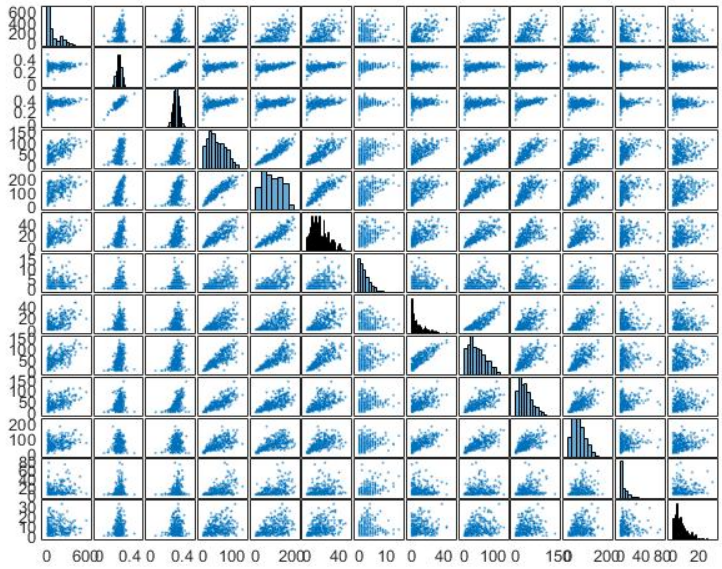


Figura 2: Plotmatrix de la base de datos completa.

Se puede observar que entre algunas variables el comportamiento parece ser lineal, lo cual puede reflejar dependencia lineal entre ellas.

Se calculó las distancias euclídeas y de mahalanobis para los datos. Los valores de las distancias euclídeas se presentan en el siguiente histograma.

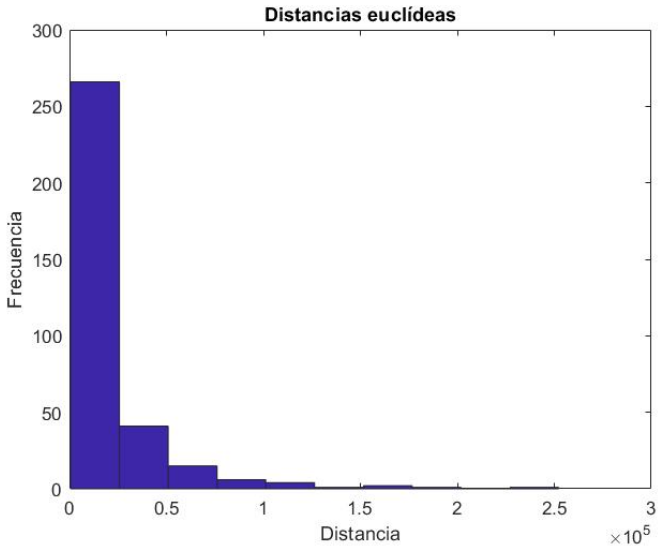


Figura 3: Histograma de las distancias euclideas.

Los valores de las distancias de mahalanobis se presesentan en el siguiente histograma.

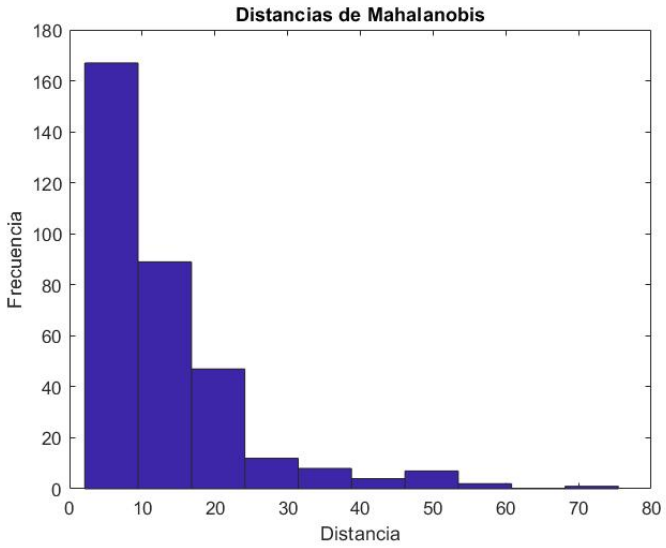


Figura 4: Histograma de las distancias de Mahalanobis.

Cuando se obtiene la distancia de mahalanobis se procede a encontrar los outliers con la prueba de chi cuadrado inversa con una probabilidad de 0.99 y 13 grados de libertad, y se obtienen 25 outliers. Ya que se tiene una dimensión alta, es decir 13 variables,no se puede visualizar los outliers en un solo gráfico, por ende estos se mostrarán en dimensión dos de las variables de # de hits y # de homeruns.

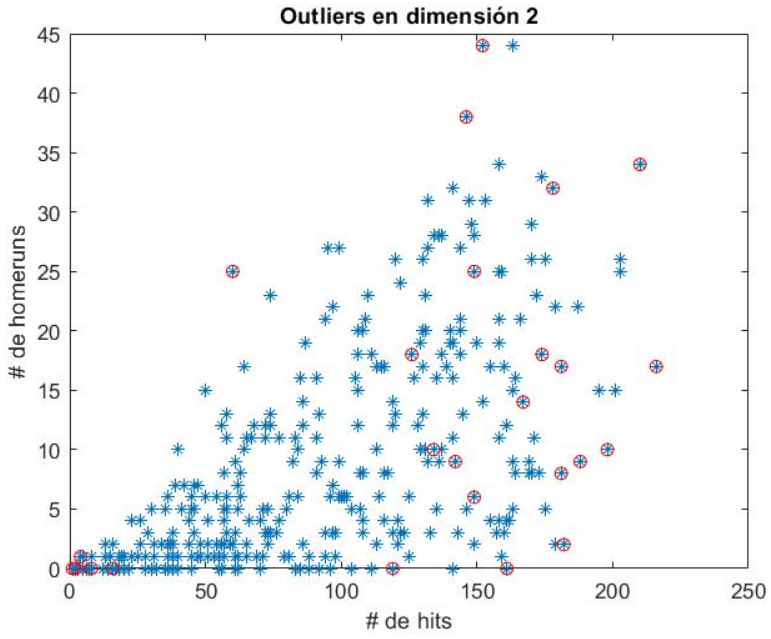


Figura 5: Outliers en dimensión 2.

A la base de datos sin outliers se procede hacerle el mismo procedimiento anterior, de encontrar los estadísticos descriptivos, graficar el plotmatrix y encontrar las distancias euclídeas y de mahalanobis, y teniendo esto se pretende hacer un análisis más exhaustivo que permita generar conclusiones sobre el comportamiento de los datos.

	Salario	Promedio de bateo	Porcentaje en base	# de carreras	# de hits	# dobles	# triples	# home runs	# de carreras bateadas	# de caminatas	# de ponches	# de bases robadas	# de errores
Media	118.6173	0.255823718	0.320958333	44.0865	89.88	16.109	2.14744	8.746795	42.9134615	33.0128205	55.4904	6.85256	6.7019
Desviación estándar	117.3487	0.035673191	0.042481865	26.5903	49.06	9.85592	2.21745	8.843585	28.4611013	22.2808692	31.9421	8.62868	5.7216
Asimetría	1.140281	-0.680858256	-0.603236554	0.3892	0.176	0.60471	1.3465	1.121508	0.60504407	0.78900738	0.61787	2.05615	1.2189
Curtosis	3.446876	5.072213552	4.596068098	2.257	2.036	2.85931	4.70433	3.531146	2.52871921	3.14072465	2.74464	8.06775	4.041
Coeficiente variación	0.989305	0.139444423	0.13235944	0.60314	0.546	0.61183	1.03261	1.011066	0.66322082	0.67491565	0.57563	1.25919	0.8537
Mediana	72.75	0.258	0.321	40	85.5	14.5	2	5.5	38	30	49	3.5	5
Meda	59	0.02	0.028	20	40	7	1	5	20	15	23	4	3
Meda/mediana	0.810997	0.07751938	0.087227414	0.5	0.468	0.48276	0.5	0.909091	0.52631579	0.5	0.46939	1.14286	0.6

Figura 6: Tabla de estadísticos relevantes para la base de datos sin outliers.

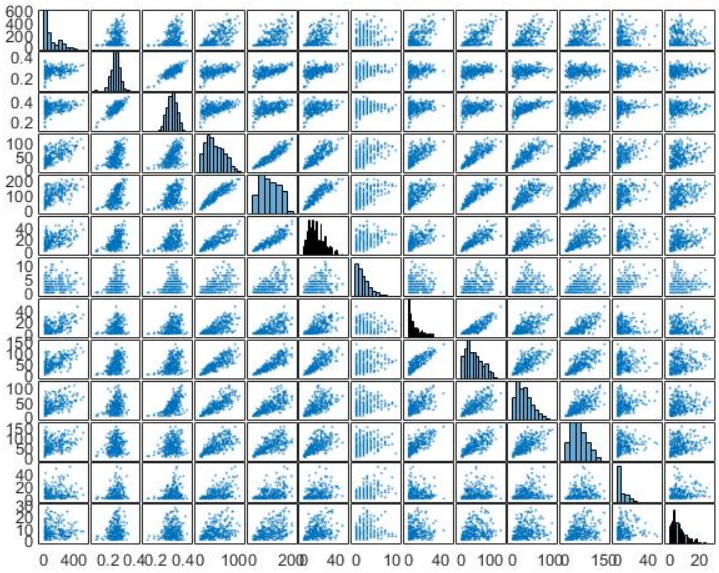


Figura 7: Plotmatrix de la base de datos sin outliers.

Se puede observar en el plotmatrix que se sigue viendo comportamientos lineales entre algunas variables, lo que quiere decir que es probable que una se pueda explicar en términos de la otra, y se puede percibir una dependencia lineal.

Se presentan los histogramas de las distancias euclídeas y de mahalanobis de los datos sin los outliers establecidos.

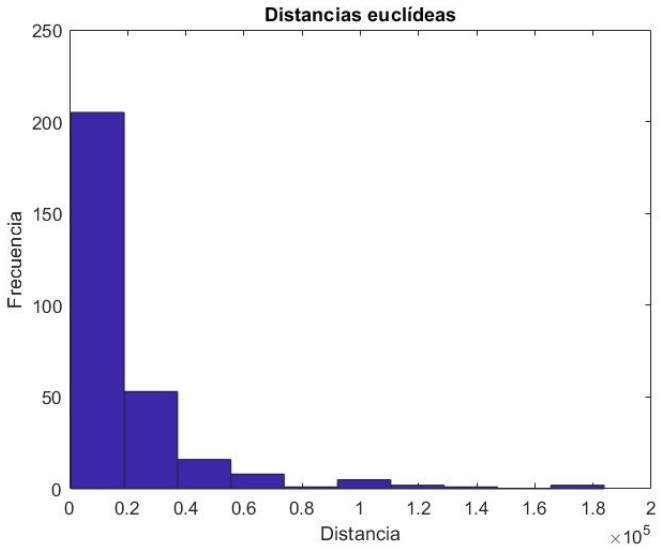


Figura 8: Distancias euclídeas para los datos sin outliers.

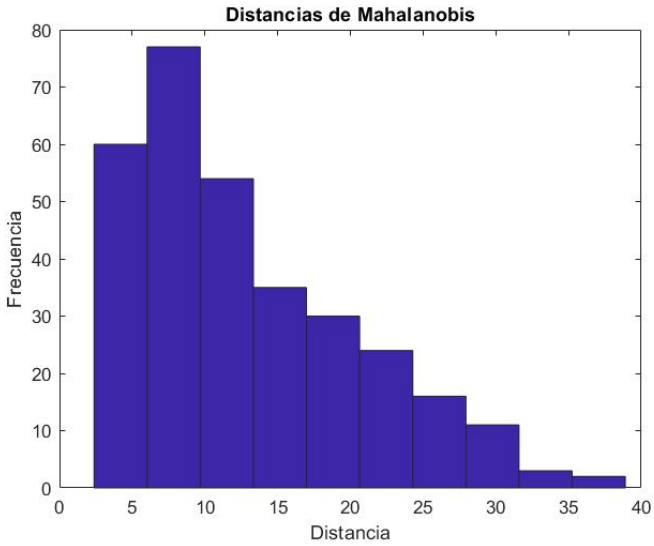


Figura 9: Distancias de mahalanobis para los datos sin outliers.

Para conocer un poco más de lo que se mostró en la figura del plotmatrix, se presenta la matriz de correlación entre variables.

1	0.30049068	0.33320026	0.63872853	0.62272327	0.56846595	0.23961242	0.60584851	0.6739772	0.55639852	0.4289122	0.23324184	0.1046161
0.30049068	1	0.78682255	0.48043257	0.54996854	0.48627653	0.28803879	0.24159742	0.41669676	0.29591449	0.1253213	0.2500885	0.1778664
0.33320026	0.78682255	1	0.52663047	0.47302889	0.41203692	0.22282101	0.31759366	0.4194462	0.61251184	0.23188412	0.25463034	0.09568438
0.63872853	0.48043257	0.52663047	1	0.93436069	0.84134213	0.54271662	0.69552742	0.85792875	0.8232227	0.70492323	0.51646963	0.34105116
0.62272327	0.54996854	0.47302889	0.93436069	1	0.89545066	0.53484166	0.63028427	0.87195361	0.72451188	0.67545945	0.43980973	0.42192602
0.56846595	0.48627653	0.41203692	0.84134213	0.89545066	1	0.40518162	0.63291506	0.83502975	0.64645249	0.63622956	0.32262683	0.31538104
0.23961242	0.28803879	0.22282101	0.54271662	0.53484166	0.40518162	1	0.12931225	0.33065631	0.32022431	0.32097312	0.54125536	0.23638041
0.60584851	0.24159742	0.31759366	0.69552742	0.63028427	0.63291506	0.12931225	1	0.87191869	0.62140711	0.76701409	0.06553841	0.10138507
0.6739772	0.41669676	0.4194462	0.85792875	0.87195361	0.83502975	0.33065631	0.87191869	1	0.73369633	0.76399616	0.21005343	0.26290966
0.55639852	0.29591449	0.61251184	0.8232227	0.72451188	0.64645249	0.32022431	0.62140711	0.73369633	1	0.67118699	0.33681567	0.21542942
0.4289122	0.1253213	0.23188412	0.70492323	0.67545945	0.63622956	0.32097312	0.76701409	0.76399616	0.67118699	1	0.24434504	0.27561497
0.23324184	0.2500885	0.25463034	0.51646963	0.43980973	0.32262683	0.54125536	0.06553841	0.21005343	0.33681567	0.24434504	1	0.18361743
0.1046161	0.1778664	0.09568438	0.34105116	0.42192602	0.31538104	0.23638041	0.10138507	0.26290966	0.21542942	0.27561497	0.18361743	1

Figura 10: Matriz de correlación entre variables.

En esta matriz se puede evidenciar que algunas correlaciones entre variables son bastante altas y sobrepasan el 0.8. Tales como, la correlación entre la variable 4 y 5, es decir entre la variable # de carreras y # de hits, se puede evidenciar que estas dos variables en el mundo del beisbol tienen mucho que ver, por ende es congruente que en la matriz de correlación tengan un valor tan alto.

A continuación, se presenta los coeficientes de correlación múltiple entre cada variable y el resto, es decir los R^2, los cuales presentan el porcentaje en que son explicadas cada variable en términos de las otras. Se puede observar que las variables # de carreras, # de hits y # de carreras bateadas son las variables que son más explicadas en términos de las otras.

Salario	0.53210934
Promedio de bateo	0.87830436
Porcentaje en base	0.88591867
# de carreras	0.95113032
# de hits	0.95934552
# dobles	0.82505372
# triples	0.50374094
# home runs	0.89532149
# de carreras bateadas	0.94214394
# de caminatas	0.89528157
# de ponches	0.76121279
# de bases robadas	0.53213985
# de errores	0.29546901

Figura 11: Coeficientes de correlación múltiple entre cada variable y el resto.

Al realizar el proceso anterior, se procede a realizar estadísticos descriptivos adicionales que permiten conocer la variabilidad de los datos.

Primero se presenta la matriz de covarianzas.

13770.7196	1.25791496	1.66106728	1993.04834	3584.77484	657.476243	62.3508162	628.739445	2250.99861	1454.77695	1607.71753	236.172335	70.2421531
1.25791496	0.00127258	0.0011924	0.45571948	0.96242805	0.17097104	0.02278492	0.07621889	0.42307151	0.23520162	0.14280055	0.07698036	0.0363042
1.66106728	0.0011924	0.00180471	0.59488465	0.98577934	0.17251902	0.02099009	0.11931739	0.50714429	0.57976259	0.31465715	0.09333789	0.02325764
1993.04834	0.45571948	0.59488465	707.043934	1218.78256	220.492147	32.0000618	163.555744	649.271179	487.72236	598.725915	118.49833	51.8876144
3584.77484	0.96242805	0.98577934	1218.78256	2406.45056	432.939711	58.1792605	273.434486	1217.40072	791.892283	1058.40233	186.164791	118.425643
657.476243	0.17097104	0.17251902	220.492147	432.939711	97.1392118	8.85526424	55.1659453	234.234541	141.960013	200.296871	27.4373403	17.7849988
62.3508162	0.02278492	0.02099009	32.0000618	58.1792605	8.85526424	4.91709951	2.53584385	20.8681054	15.8212548	22.7345412	10.3562124	2.99907247
628.739445	0.07621889	0.11931739	163.555744	273.434486	55.1659453	2.53584385	78.2089929	219.460333	122.443771	216.667976	5.00113365	5.13007049
2250.99861	0.42307151	0.50714429	649.271179	1217.40072	234.234541	20.8681054	219.460333	810.034288	465.264779	694.553828	51.5852708	42.8133502
1454.77695	0.23520162	0.57976259	487.72236	791.892283	141.960013	15.8212548	122.443771	465.264779	496.437134	477.681796	64.7543079	27.4636409
1607.71753	0.14280055	0.31465715	598.725915	1058.40233	200.296871	22.7345412	216.667976	694.553828	477.681796	1020.29573	67.3458447	50.3717227
236.172335	0.07698036	0.09333789	118.49833	186.164791	27.4373403	10.3562124	5.00113365	51.5852708	64.7543079	67.3458447	74.454077	9.06523621
70.2421531	0.0363042	0.02325764	51.8876144	118.425643	17.7849988	2.99907247	5.13007049	42.8133502	27.4636409	50.3717227	9.06523621	32.737231

Figura 12: Matriz de covarianzas.

A esta matriz se le calculó sus valores propios y el porcentaje de variabilidad que explica cada uno.

Valores propios	Porcentaje de variabilidad explicada
9.36E-05	4.80086E-07
0.00158333	8.12029E-06
2.352171171	0.012063381
7.656615871	0.039267836
16.15583157	0.082857043
23.93535774	0.122755239
26.17739515	0.134253786
59.99429518	0.307687653
137.3837197	0.70458823
200.0823162	1.026145204
473.3720622	2.427743143
2379.259492	12.20230634
16172.06997	82.94032354

Figura 13: Valores propios y porcentaje de variabilidad que explica cada uno.

Y se muestran los vectores asociados a cada autovalor, cada uno de los vectores es una columna.

3.50E-06	-5.68E-05	4.80E-05	0.00572283	0.00462084	-0.00138675	-0.00422212	-0.01659354	-0.04860451	0.01037617	-0.0297619	-0.4120852	0.90910986
0.7613109	-0.64838518	4.95E-05	-0.00116364	-0.00038912	0.00040353	0.00011738	9.30E-05	0.00025662	0.00014418	0.00073035	0.00021125	0.00010046
-0.64838601	-0.76130846	-0.00069695	-0.00149567	-0.0002291	0.00027159	0.00026186	-0.00017735	0.00057977	-0.00123754	0.00035684	0.00024061	0.0001298
-2.30E-05	-0.00013635	0.08845951	0.20864602	-0.03668801	0.10847464	-0.58040964	0.59392321	-0.0803852	-0.3013014	0.14027812	0.32288202	0.15726977
-0.00036047	0.00077958	0.00322586	-0.14508178	-0.18999914	0.05571885	0.10586894	-0.31897886	-0.17122435	0.22098572	0.52415486	0.62619213	0.28595938
3.61E-05	-1.23E-05	-0.06767017	-0.04795549	0.90127376	0.38574905	0.00522319	-0.02690441	0.05609582	0.08263135	0.07270301	0.11514752	0.05247669
0.000426	0.00011082	-0.97240146	0.20618205	-0.06138109	0.00193545	0.01152109	0.05042958	-0.066712	-0.00582219	0.02545419	0.01775398	0.00542302
0.00012378	0.0017468	-0.19142245	-0.88218534	-0.08280301	0.02013562	-0.15769368	0.23499924	0.24587214	0.0749841	-0.15782146	0.06735145	0.04722422
-8.78E-05	-0.00036409	0.02593235	0.28357364	0.02112039	-0.24814331	0.25950241	0.24210352	0.72551937	0.2564069	-0.12361366	0.30556533	0.17367555
0.00097187	0.00104872	-0.04238121	-0.07429003	0.03956427	-0.04218876	0.15827674	-0.32642517	0.2254147	-0.85293025	-0.12819581	0.22363757	0.11394259
2.85E-05	-0.00091239	0.02262259	0.06532159	-0.00619693	0.04979523	-0.0051183	-0.0797193	-0.3789746	0.16457695	-0.79344494	0.40944537	0.13569613
4.77E-05	0.00017668	0.02823761	-0.13732614	0.14806139	-0.27454041	0.6338917	0.52990627	-0.4017115	-0.15956521	0.10262745	0.05181178	0.01949708
0.00011386	-0.00054815	-0.0399978	-0.05711072	0.34006236	-0.83349144	-0.36418853	-0.18551918	-0.10781127	0.04481036	0.04532417	0.04463202	0.00772348

Figura 14: Vectores propios asociados a cada autovalor.

Luego se calculan las siguientes varianzas y desviaciones.

Varianza total	19498.4409
Varianza media	1499.88007
Varianza generalizada	8.1163E+14
Desviación típica generalizada	28489040.7
Varianza efectiva	14.0240519
Desviación efectiva	3.74487008

Figura 15: Estadísticos relevantes.

3. Análisis de regresión

Ya con este análisis previo a toda la base de datos se busca realizar un modelo de regresión. Para llevarlo a cabo, se debe tener una variable respuesta, en este caso se realizará dos modelos de dos variables distintas, primero con la variable salario que, aunque no es una de las variables más explicadas por las demás, como se evidenció en la tabla de R^2, se efectuará ya que la base de datos cuestiona si los salarios de los jugadores se basan en su actuación en

el juego. Para el segundo modelo se utilizará una variable deportiva, que puede ser explicada por las demás variables deportivas, y en este caso es la variable home runs.

Para encontrar el modelo de regresión se utilizó el comando fitlm de matlab el cual permite conocer el p-valor del f-estadístico que dice si existe alguna variable significativa entre las variables que se tienen como explicativas, da a conocer el R^2 del modelo y dice si cada variable es significativa, donde esto último se puede evidenciar con su p-valor y su t-estadístico. Dado que inicialmente se calcula el modelo con todas las variables, el resultado de la regresión no muestra que todas sean significativas, por ende se deben ir eliminando una por una hasta llegar a que todas sean significativas.

Al realizar este proceso de buscar las variables explicativas para el salario, se eliminan la variable $\#3, \#6, \#7, \#9, \#12$ y $\#13$ y al final se obtiene el siguiente modelo que se muestra en la herramienta de matlab así:

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	91.255	40.929	2.2296	0.026499
x1	-340.32	171.45	-1.985	0.048033
x2	1.205	0.18417	6.5428	2.5371e-10
x3	6.9758	0.85552	8.1539	9.1681e-15
x4	0.9374	0.32735	2.8636	0.0044776
x5	-1.5469	0.27118	-5.7042	2.7567e-08

Number of observations: 312, Error degrees of freedom: 306
Root Mean Squared Error: 82
R-squared: 0.52, Adjusted R-Squared: 0.512
F-statistic vs. constant model: 66.2, p-value = 1.07e-46

Figura 16: Modelo de regresión.

El resultado final de la regresión es:
Salario$= 91{,}255 - 340{,}32x_1 + 1{,}205x_4 + 6{,}9758x_7 + 0{,}9374x_9 - 1{,}5469x_{10}$

Donde las variables explicativas son:

x_1= promedio de bateo x_4= número de hits

x_7=Número de home runs x_9=Número de caminatas x_{10}=Número de ponches

Este modelo explica la variable salario en un 51 %.

El segundo modelo de regresión se hizo tomando como variable respuesta la variable deportiva del número de home runs. Para este análisis se realizaron nuevamente los cálculos de los datos atípicos para poder proseguir con el modelo. Entonces se calcularon las distancias euclídeas y de mahalanobis que se muestran en las figuras:

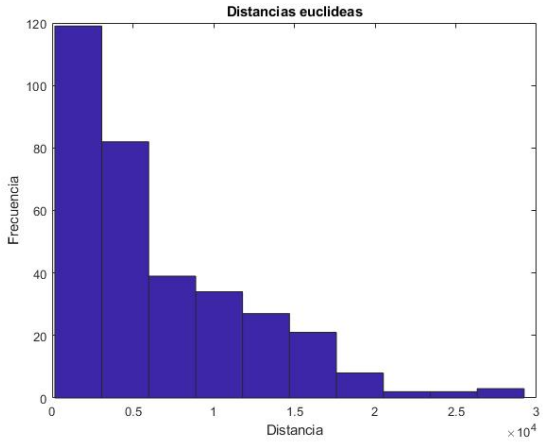


Figura 17: Histograma de distancias euclideas.

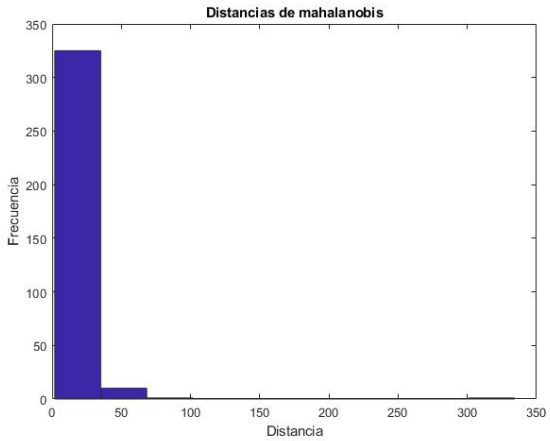


Figura 18: Histograma de distancias de mahalanobis.

Se detectaron 20 outliers con el mismo proceso realizado para la variable salario. El grafico muestra los outliers en dimensión dos del número de carreras contra número de hits.

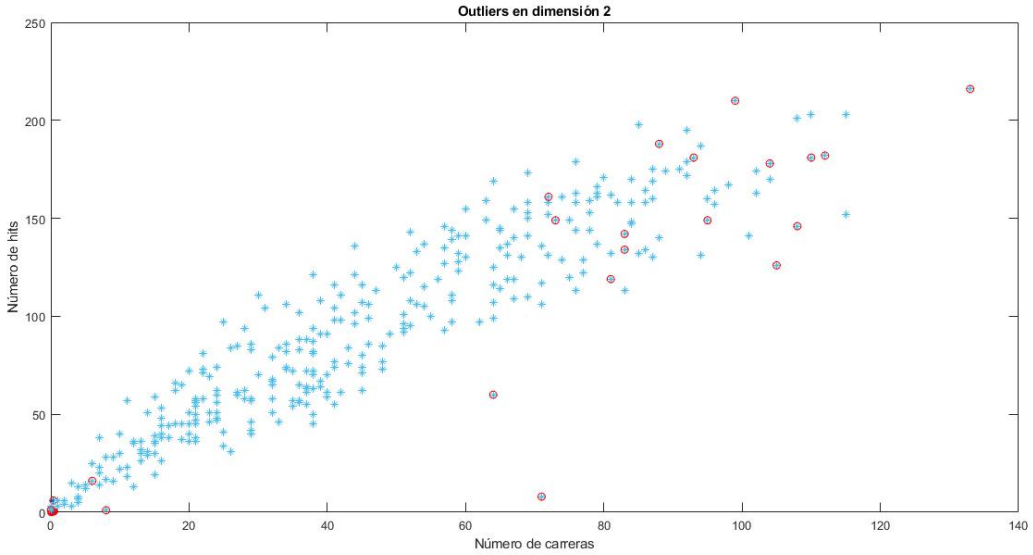


Figura 19: Outliers en dimensión 2.

También se calcularon algunos estadísticos con la base de datos sin outliers para concocer la variabilidad de los mismos. Se muestran en la siguiente tabla:

	Promedio de bateo	Porcentaje en base	# de carreras	# de hits	# de dobles	# de triples	# de home runs	# de carreras	# de caminatas	# de ponches	# de bases rodeadas	# de errores
Media	0.2567	0.3219	44.6404	90.6309	16.3249	2.1451	8.8801	43.2618	33.4006	55.6404	6.9306	6.7508
Desviación estandar	0.0363	0.0431	27.2715	49.8477	10.1135	2.2157	9.0273	28.8978	22.7573	32.3527	8.7548	5.7311
Asimetría	-0.6058	-0.5363	0.3989	0.1657	0.6054	1.3444	1.1833	0.609	0.7746	0.6374	2.0066	1.183
Curtosis	5.0224	4.5795	2.2743	2.0257	2.8152	4.6824	3.8744	2.5589	3.0584	2.8375	7.6788	3.944
Coeficiente de variación	0.1412	0.1338	0.6109	0.55	0.6195	1.0329	1.0166	0.668	0.6813	0.5815	1.2632	0.849
Mediana	0.259	0.322	40	86	15	2	6	38	30	49	3	5
Meda	0.02	0.028	21	40	7	1	5	20	16	23	4	3
Meda/mediana	0.0772	0.087	0.525	0.4651	0.4667	0.5	0.8333	0.5263	0.5333	0.4694	1.3333	0.6

Figura 20: Tabla de estadísticos relevantes

Se realizó el modelo regresión para explicar la variable respuesta home-runs, nuevamente con el comando fitlm donde se eliminaron 5 variables para finalmente obtener el modelo con las 7 variables más representativas.

```
Estimated Coefficients:
                   Estimate        SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)     -4.6708       1.595    -2.9283      0.003661
    x1               11.567      5.3161     2.1758      0.030325
    x2              0.19413    0.023208     8.3651    2.0933e-15
    x3             -0.14162    0.010896    -12.997    4.1609e-31
    x4              -0.5677    0.099892    -5.6832    3.0578e-08
    x5              0.31765    0.014833     21.415     5.146e-63
    x6            -0.085991    0.016133      -5.33    1.8981e-07
    x7             0.078066   0.0090601     8.6165    3.6234e-16


Number of observations: 317, Error degrees of freedom: 309
Root Mean Squared Error: 3.03
R-squared: 0.889,  Adjusted R-Squared 0.887
F-statistic vs. constant model: 355, p-value = 1.12e-143
```

Figura 21: Modelo de regresión que explica el salario.

Home runs $= -4{,}6708 + 11{,}567x_2 + 0{,}19413x_3 - 0{,}14162x_4 - 0{,}5677x_6 + 0{,}31765x_7 - 0{,}085991x_8 + 0{,}078066x_9$

Donde las variables explicativas son:

x_2=Porcentaje en base x_3=Número de carreras x_4=Número de hits

x_6=Número de triples x_7=Número de carreras bateadas

x_8=Número de caminatas x_9=Número de ponches

Este modelo explica la variable home runs en un 89 %.

3.1. Análisis de residuales

Es importante detallar que para una de las variables respuesta que se está evualuando, en este caso la variable del salario, el R-squared de la regresión tiene un valor muy pequeño. Para aumentar este valor, se hace un analisis de outliers especializado para la regresión. Esto es, quitar aquellos datos que otorgan los peores residuales. Para ello, se quitaron de la muestra el 10 % de los peores residuales, y se recalculaba el modelo cada vez que se quitaba uno de ellos. Así obtenemos que, sacando el 10 % de los peores errores y recalculando el modelo solo es posible alcanzar un R-squared de 0.564 con 281 observaciones. El modelo es el siguiente:

	Estimate	SE	tStat	pValue
(Intercept)	122.12	30.696	3.9783	8.8692e-05
x1	-443.97	128.17	-3.464	0.00061652
x2	1.5661	0.12863	12.175	1.3349e-27
x3	5.4687	0.68067	8.0342	2.729e-14
x4	-1.7792	0.20778	-8.5631	7.7268e-16

Number of observations: 282, Error degrees of freedom: 277
Root Mean Squared Error: 59.8
R-squared: 0.57, Adjusted R-Squared 0.564
F-statistic vs. constant model: 91.7, p-value = 1.48e-49

Figura 22: Modelo de regresión lineal para la base de datos sin el 10 % de los peores residuales

Que implica lo siguiente:

$$\text{Salario} = 122{,}12 - 443{,}97x_1 + 1{,}5661x_4 + 5{,}4687x_7 - 1{,}7792X x_{10}$$

Donde,

- x_1=Promedio de bateo
- x_4=Número de hits
- x_7=Número de home runs
- x_{10}=Número de strike-outs

Con el fin de aumentar el porcentaje de explicación para la variable home runs, también se hizo un analisis de residuales, con el que se logra explicar el 92 % con las mismas 7 variables

del modelo de regresión obtenido antes y con 287 observaciones. El modelo se muestra a continuación:

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-3.5409	1.2486	-2.8359	0.0049042
x1	9.7181	4.1666	2.3324	0.02039
x2	0.1948	0.018755	10.386	1.4065e-21
x3	-0.15881	0.0088769	-17.891	3.2847e-48
x4	-0.5811	0.085358	-6.8078	6.0607e-11
x5	0.31941	0.012401	25.756	1.0296e-75
x6	-0.063775	0.013219	-4.8244	2.3137e-06
x7	0.069491	0.0072814	9.5437	7.2926e-19

Number of observations: 287, Error degrees of freedom: 279
Root Mean Squared Error: 2.33
R-squared: 0.921, Adjusted R-Squared 0.919
F-statistic vs. constant model: 467, p-value = 5.07e-150

Figura 23: Modelo de regresión lineal que explica la variable home runs en términos deportivos.

El nuevo modelo es:

$$\text{Home runs} = -3{,}5409+9{,}7181x_2+0{,}1948x_3-0{,}15881x_4-0{,}5811x_6+0{,}31941x_7-0{,}063775x_8+0{,}069491x_9$$

Donde,

- x_2=Porcentaje en base(OBP)
- x_3=Número corridas
- x_4=Número de hits
- x_6=Número de triples
- x_7=Número de homeruns bateados en (RBI)
- x_8=Número de caminatas
- x_9=Número de strike-outs

3.2. Componentes principales

Se realiza un proceso de componentes principales, donde primero se realizó de todas las variables. Estos componentes se llevan a cabo con los vectores y autovalores encontrados anteriormente, y se evidencia que con los dos primeros componentes principales se explicaría alrededor del 95 % de la variabilidad de los datos. El primer componente principal sería como una media ponderada de todas las variables, es decir de sus atributos deportivos y de su salario, dándole más peso al salario, y el segundo vector es una media ponderada que contrapone el salario con todas las otras variables dándole más peso al # de hits de cada jugador.

	Componente1	Componente2
	0.909109865	-0.412085205
	0.000100457	0.00021125
	0.000129803	0.000240608
	0.157269774	0.322882022
	0.285959378	0.626192128
	0.052476694	0.11514752
	0.005423021	0.017753982
	0.047224221	0.067351446
	0.173675555	0.305565335
	0.11394259	0.223637567
	0.135696127	0.409445374
	0.01949708	0.05181178
Valor Propio asociado	16172.06997	2379.259492
Porcentaje de variabilidad explicada	0.829403235	0.122023063

Figura 24: Tabla con las dos primeras componentes principales.

También,se realiza componentes principales con todas las variables excepto con la primera variable respuesta que es el Salario. A continuación se presentan los dos primeros componentes que explican alrededor del 91 % de variabilidad . El primer componente refleja una media ponderada de todas las variables deportivas dándole más peso a la variable # de hits, y el segundo una media ponderada que contrapone algunas variables con otras, principalmente a la variable # de ponches, es decir, strike-outs, con la variable # de hits.

	Componente1	Componente2
	0.000241041	-0.000714104
	0.000295679	-0.000359474
	0.369719616	-0.130055563
	0.690262157	-0.487819634
	0.126804104	-0.066965535
	0.015731687	-0.022643044
	0.097226438	0.148746229
	0.385049258	0.113463777
	0.262967362	0.125604904
	0.374269595	0.824069612
	0.051052403	-0.094680028
	0.031093828	-0.035898893
Valor Propio asociado	4752.40351	479.9830845
Porcentaje de variabilidad explicada	0.829403235	0.083800006

Figura 25: Tabla con las dos primeras componentes principales sin la variable respuesta Salario.

Se realizó un modelo de regresión para que estos dos componentes explicaran la variable respuesta Salario, y se obtuvo un modelo que explicaba en un 42 % a la variable respuesta Salario.

Salario$= -19{,}751 + 1{,}0988 * z_1 - 0{,}46195 * z_2$

z_1= proyección del componente 1

z_2= proyección del componente 2

Componentes principales para los datos sin el 10 % de los peores residuales

Ahora, se hace componentes principales con la base de datos para las variables explicativas de Salario y se obtienen que la variabilidad puede ser explicada por 2 componentes, donde una refleja el 87 % de la variabilidad, y la otra el 23 % de la misma. Para este caso, el modelo obtenido es el siguiente:

	Estimate	SE	tStat	pValue
(Intercept)	93.465	3.9692	23.548	3.5969e-68
x1	1.0274	0.073348	14.007	4.3121e-34
x2	-1.2819	0.19591	-6.543	2.8896e-10

Number of observations: 281, Error degrees of freedom: 278
Root Mean Squared Error: 66.5
R-squared: 0.462, Adjusted R-Squared 0.458
F-statistic vs. constant model: 119, p-value = 3.53e-38

Figura 26: Modelo de regresión lineal para los componentes principales sin el 10 % de los peores residuales

De esta manera, el modelo que obtenemos es

$$\text{Salario} = 93{,}465 + 1{,}0274z_1 - 1{,}2819z_2$$

Donde,

- z_1=Proyección del componente 1
- z_2=Proyección del componente 2

Sin embargo, aunnque el modelo sea estadísticamente válido, este solo explica el 45.8 % del porcentaje total de datos.

Esto implica que, en este caso, es preferible hacer un análisis con los datos en sí que con el vector proyectado de las componentes, pues con el primero se obtiene una mejor aproximación del mismo.

También se realizó componentes principales para el conjunto de datos que contiene todas las variables menos el salario y que busca explicar la variable home runs. Se realizó para los datos con y sin el 10 % de los peores residuales para comparar los resultados.

Para el primer caso, el modelo obtenido es el siguiente:

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-4.4652	0.59091	-7.5566	4.5683e-13
x1	0.098953	0.0040576	24.387	1.9412e-74
x2	-0.1441	0.012709	-11.338	3.304e-25

Number of observations: 317, Error degrees of freedom: 314
Root Mean Squared Error: 4.98
R-squared: 0.697, Adjusted R-Squared 0.695
F-statistic vs. constant model: 362, p-value = 3.32e-82

Figura 27: Modelo de regresión lineal para los componentes principales para la variable explicada home runs

El modelo que se obtiene es

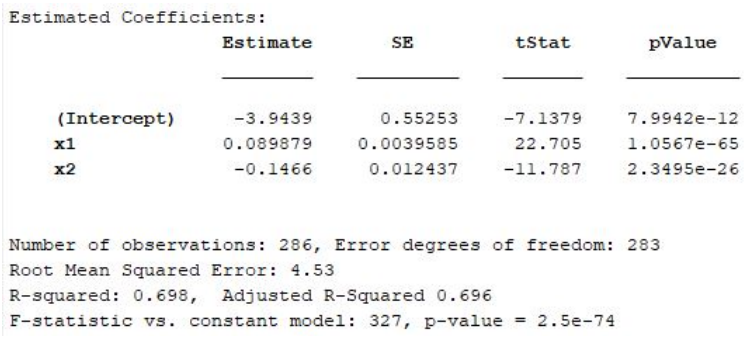
$$\text{Home runs} = -4{,}4652 + 0{,}090953z_1 - 0{,}1441z_2$$

Donde,

- z_1=Proyección del componente 1
- z_2=Proyección del componente 2

Este modelo solo explica el 70 % a los datos.

Sin el 10 % de los peores residuales



```
Estimated Coefficients:
                   Estimate       SE         tStat       pValue
                   ________    _________    _______    _________

    (Intercept)     -3.9439      0.55253    -7.1379    7.9942e-12
    x1             0.089879    0.0039585     22.705    1.0567e-65
    x2              -0.1466     0.012437    -11.787    2.3495e-26


Number of observations: 286, Error degrees of freedom: 283
Root Mean Squared Error: 4.53
R-squared: 0.698,  Adjusted R-Squared 0.696
F-statistic vs. constant model: 327, p-value = 2.5e-74
```

Figura 28: Modelo de regresión lineal para los componentes principales sin el 10 % de los peores residuales

El modelo que se obtiene es

$$\text{Home runs} = -3{,}94392 + 0{,}089879z_1 - 0{,}1466z_2$$

Donde,

- z_1=Proyección del componente 1
- z_2=Proyección del componente 2

Este modelo también explica el 70 % a los datos.

Finalmente se concluye que es preferible hacer un análisis con los datos en sí que con los vectores proyectados de las componentes, pues con el primero se obtiene una mejor porcentaje explicado del mismo.

4. Clasificación

Para este punto en el análisis estadístico realizado en la base de datos que se ha trabajado, se hace importante buscar una manera de analizar los jugadores respecto a aquellos que tengan un comportamiento similar para las diferentes características de los mismos, es decir, identificar grupos de individuos en la muestra con comportamientos similares. Para esto será muy importante la clasificación supervisada y no supervisada que se haga de los idividuos respecto a las cuatro variables binarias disponibles. Estas variables funcionarán como etiquetas para los jugadores y se buscará que aquellos que se dicen pertenecer a un

mismo grupo tengan comportamientos similares respecto a los valores en las variables y sus etiquetas, ya que se está trabajando con cierta información limitada. Sin embargo, se debe tener en cuenta que puede ser mejor clasificar a los individuos respecto a otras características que describe a la población y no tener conocimiento de ello, por lo que los hallazgos aquí descritos siempre pueden variar respecto a nuevos estudios que se hagan.

4.1. Clasificación no supervisada

Este tipo de clasificación se basa en un proceso de asignación a cierto número de grupos predeterminado que, por medio de un algoritmo, utiliza unos centros iniciales para cada grupo para así definir la pertenencia de cada individuo a determinado grupo. Esta asignación se hace respecto a la menor distancia, o norma, que se tenga en comparación a la que poseen para los otros grupos. En este caso se han utilizado los métodos de K-medias y K-medoides, para diferentes normas como lo es la norma 1, norma 2 o euclídea, norma infinito y la distancia de Mahalanobis en el proceso de clasificación.

Para el método de k-medias se utilizaron las tres normas mencionadas y se hizo el proceso de asignación para dos grupos, los cuales fueron posteriormente comparados con cada una de las cuatro variables binarias y se calculó un porcentaje de aciertos, como se muestra en la siguiente tabla.

Porcentaje de acierto para el Método k-medias			
variable binaria	**Norma 1**	**Norma 2**	**Norma infinito**
Free agency eligibility	0.7244	0.7628	0.7692
Free agency eligibility 1991/2	0.5929	0.6506	0.6827
Arbitration eligibility	0.6186	0.6506	0.6442
Arbitration eligibility 1991/2	0.6058	0.6827	0.7083

En esta tabla se puede ver que la clasificación realizada por el método k-medias con norma infinito es la que mayor relación tiene con la primera variable binaria, es decir, que se puede establecer una distinción en un 76.92 % acertada de los jugadores respecto a la libre elección de agencia.

Se realizó este mismo proceso con el método k-medoides, método más robusto que garantiza que los centros escogidos sean siempre un individuo de la muestra, y se utilizó la norma 2 y la distacia de mahalanobis para este proceso, como se muestra a continuación.

Porcentaje de acierto para el Método k-medoides		
variable binaria	**Norma 2**	**Distancia Mahalanobis**
Free agency eligibility	0.766	0.5417
Free agency eligibility 1991/2	0.6603	0.6667
Arbitration eligibility	0.6474	0.5545
Arbitration eligibility 1991/2	0.6987	0.6795

Se puede ver que los resultados para la norma 2, en ciertos casos, presentan un mejor comportamiento que la norma 1 y algunos valores de la norma 2 del método anterior, esto quiere decir que este método puede presentar una clasificación mucho más acertada en ciertas variables que el clasificador de k-medias antes mencionado, teniendo un mejor comportamiento para la primera variable binaria. Sin embargo, se observa que la distancia de Mahalanobis no tuvo un comportamiento satisfactorio en la mayoría de los casos en esta asignación, por lo que no es recomendable su uso para este caso.

Se puede observar que los resultados anteriores permiten tener un conocimiento general del comportamineto del clasficador, sin embargo, en la búsqueda de tener una mayor confiabilidad en la clasificación, en donde se le dará un mayor peso a aquellos individuos con etiqueta 1 que sean asignados correctamente, se tendrán en cuentra otrás métricas como lo son la proporción confiable de las clasificaciones verdaderas (precisión), la proporción de asignaciones correctas respecto a todas las reales que se tenían para esta etiqueta (recall) y el F1 score, el cual es una medida que tiene en cuenta los dos valores anteriores para dar una medida más generalizada.

Las matrices de confusión y diferentes métricas de precisión para las clasificaciones con mejores resultados fueron las siguientes

Figura 29: Matriz de confusión para k-medias con norma infinito

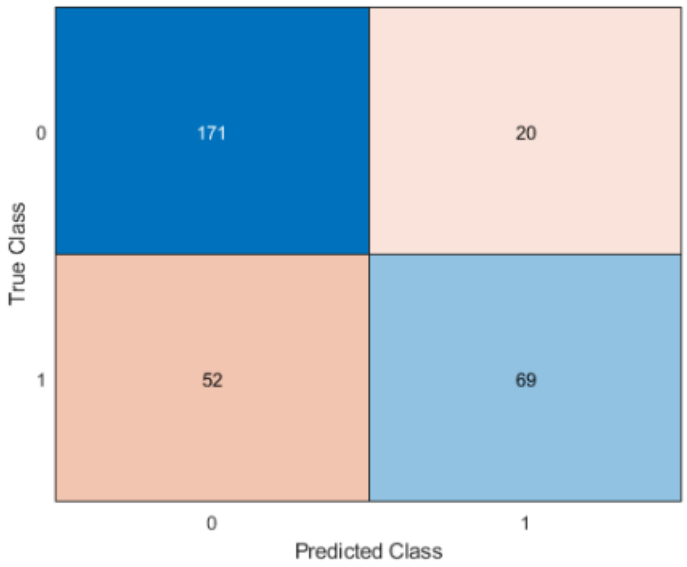


Figura 30: Métricas de precisión

Accuracy	Precision	Recall	F1
0.769	0.775	0.570	0.657

Figura 31: Matriz de confusión para k-medoides con norma 2

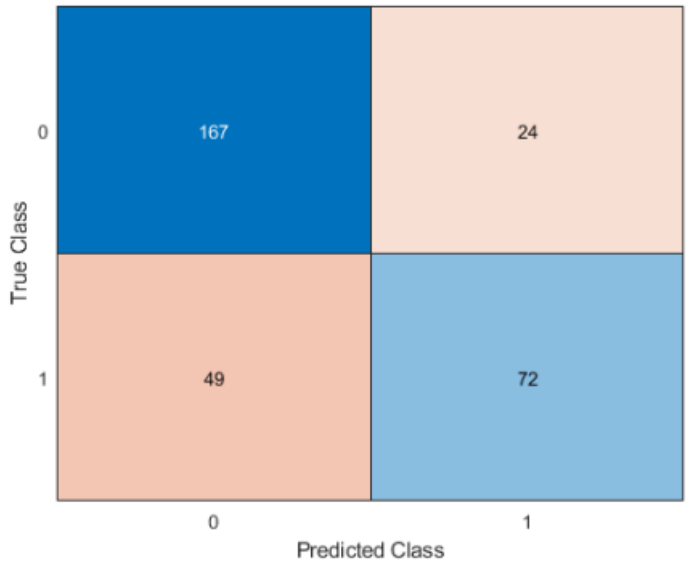


Figura 32: Métricas de precisión

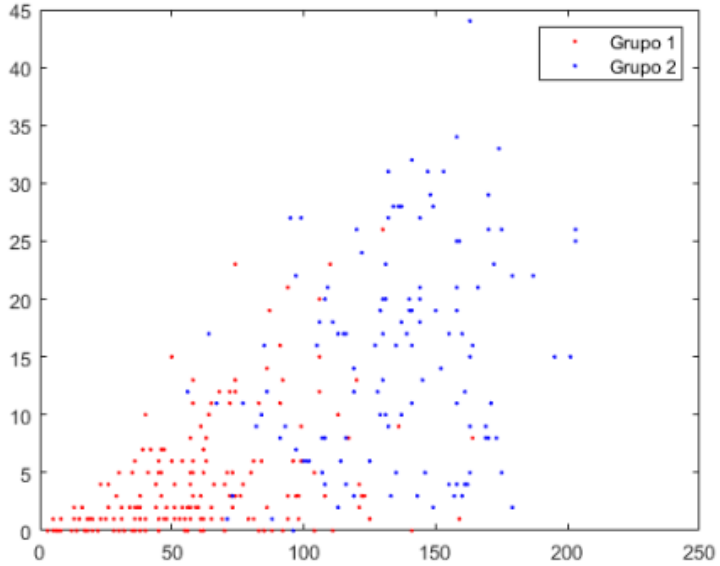
Accuracy	Precision	Recall	F1
0.764	0.750	0.595	0.664

Para ambos casos se puede ver en las matrices de confusión que la diagonal principal, aquella en donde el método logró asignar correctamente a los individuos, presenta un número mayor que en los otros dos recuadros, y además, se puede ver que la diferencia en los números de aciertos entre los dos métodos es muy baja. Además, para las diferentes métricas analizadad, se puede ver que se tiene una buena precisión, sin embargo, el valor de recall, este que indica la proporción de individuos reales con etiqueta uno clasificados correctamente, es muy baja en ambos casos, lo que significa que para ambos métodos se tiene un porcentaje bajo de aciertos para aquellos individuos que cumplen ser elegibles de manera libre por alguna agencia.

Para este punto no se puede hablar de manera concreta sobre qué clasificador supervisado se recomienda, pues podemos ver que k-medoides tiene un porcentaje mejor de aciertos (accuracy) en comparación a k-medias, pero a la vez presenta un recall menor que en el anterior mencionado. Por lo que queda a discusión en primera instancia, respecto al criterio que se desee priorizar y la importancia que se le de a la complejidad computacional que representa, cuál es el método a utilizar.

En el siguiente gráfico se puede evidenciar, para dos de las variables, dado que no es posible visualizarlas todas, el comportamiento de los dos grupos clasificados por el método k-medias. Es claro que existe cierto tipo de distinción entre ambos grupos, sin embargo, estos no están del todo separados por lo que algunos se encuentran en zonas donde ambos grupos se mezclan.

Figura 33: Clasificación en 2 grupos para las variables Home Runs y Hits



4.1.1. Cálculo de grupos en la muestra

Para el cálculo del número de grupos aún no se tiene una fundamentación generalizada por la cual se defina ese número de grupos óptimos para hacer la clasificación, sin embargo, se suele utilizar el criterio de la suma de cuadrados de las desviaciones en los grupos (SCDG), definida como se muestra en la siguiente fórmula [**?**]

$$\sum_{g=1}^{G}\sum_{j=1}^{p}\sum_{i=1}^{n_g} x_{ijg} - \overline{x}_{jg}{}^2$$

La cual es incorporada en el test F, definido como

$$F = \frac{[SCDG(G) - SCDG(G+1)](n+G-1)}{SCDG(G+1)}$$

En donde se busca añadir un nuevo grupo en la población siempre y cuando este estadístico F se mantenga por encima de 10.

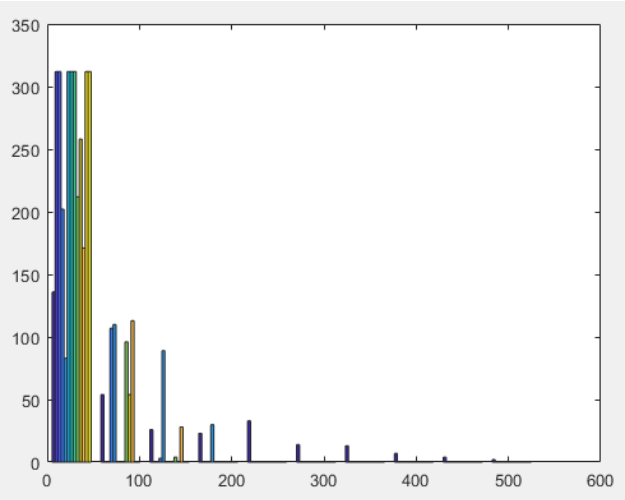
Implementando un algoritmo el cual nos permitiera calcular el número adecuado de grupos siguiendo esta fórmula, se obtuvo un valor de 5 grupos luego de repetir el proceso mil veces. A continuación se observan los resultados de las pruebas.

Figura 34: Tabla de frecuencias para los diferentes grupos

Value	Count	Percent
1	0	0.00%
2	0	0.00%
3	103	10.30%
4	138	13.80%
5	243	24.30%
6	175	17.50%
7	133	13.30%
8	87	8.70%
9	50	5.00%
10	32	3.20%
11	17	1.70%
12	8	0.80%
13	7	0.70%
14	4	0.40%
15	1	0.10%
16	1	0.10%
17	0	0.00%
18	1	0.10%

Se puede ver cómo en mayor proporción el proceso arroja 5 grupos para la base de datos tratada. Para retificar esta información se ha generado el histograma de los datos, donde se pueden ver cierto tipo de distinción en las muestras para determinados intervalos.

Figura 35: Histograma de frecuencias



4.2. Clasificación supervisada para dos grupos

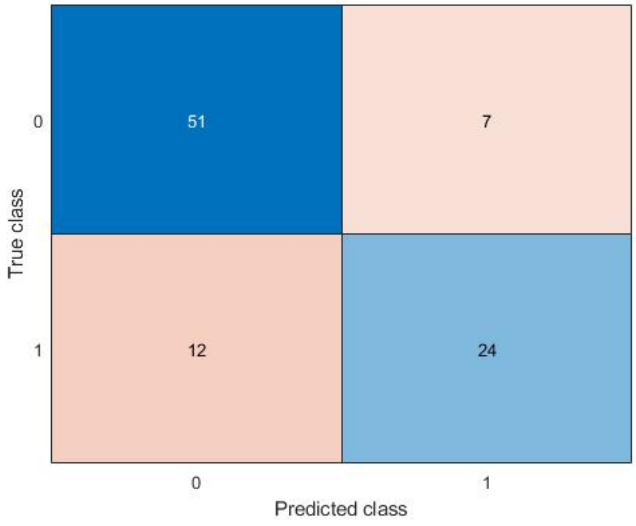
Para este tipo de clasificación se debe conocer las clases o grupos existentes, se tomarán 2 grupos para hacer el estudio. Este se realiza con la variable binaria Indicador de eligibilidad de agencia libre, pues es la que presentó mejores resultados en la clasifcación no supervisada. Donde el grupo 1 etiquetado con un 0 significa que el jugador no puede ser elegido libremente por los equipos dado que cumple con ciertas características y el grupo 2 etiquetado con un 1 significa que si puede ser elegido libremente pues cuenta con las características necesarias. Los datos utilizados fueron los que no contienen los 25 outliers y que incluye 13 variables.

4.2.1. Clasificador Linear

Se toma el 70 % de los datos para hacer el entrenamiento, donde inicialmente se seleccionan los primeros registros correspondientes a esta proporción y posteriormente se verá el proceso de selección aleatoria para comparar los resultados.

Al realizarlo de la primera forma con el clasificador linear de fisher se obtiene la siguiente matriz de confusión

Figura 36: Matriz de confusión con el clasificador Linear



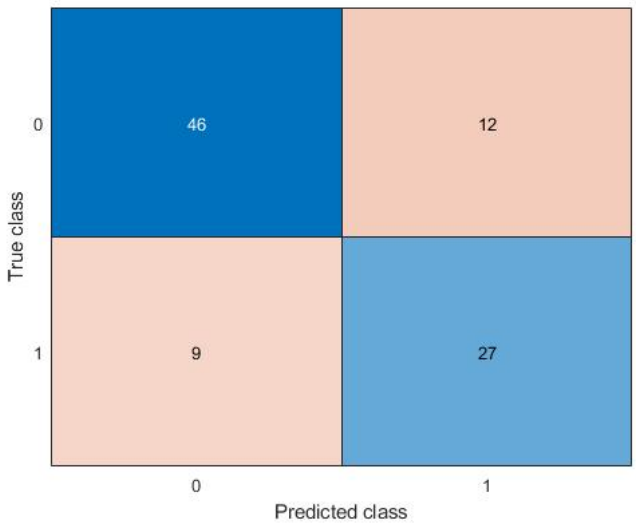
El clasificador dijo que 7 individuos pertenecen al grupo 2 cuando en realidad pertenecen al 1, de un total de 58, y dijo que 12 individuos pertenecen al grupo 1 cuando en realidad pertenecen al 2, de un total de 36.

Se miden las tasas de error donde EC1 corresponde a probabilidad de clasificar un dato de la población 1 en la población 2; EC2 la probabilidad de clasificar un dato de la población 2 en la población 1 y TGE la tasa global de error.

- $EC1$=0.1579
- $EC2$=0.3412
- TGE=0.2294

Tomando los datos para entrenamiento aleatoriamente se obtiene la matriz de confusión que se muestra a continuación:

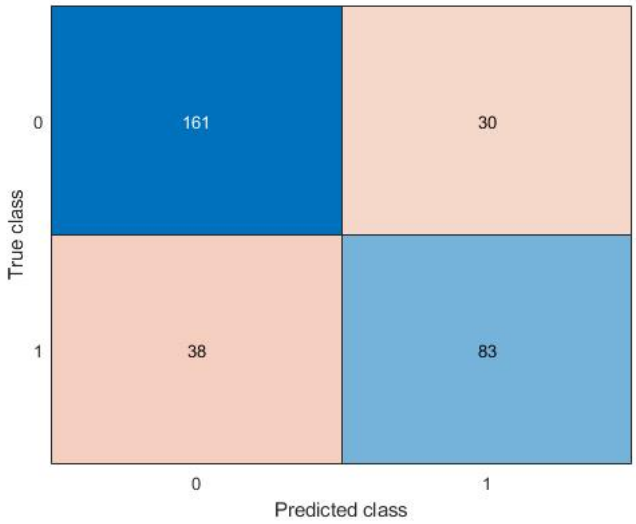
Figura 37: Matriz de confusión con entrenamiento aleatorio



Al realizar la medida de precisión de las pruebas se obtienen los valores 0.7979 y 0.7766 para el clasificador linear tomando como datos de entrenamiento el primer 70 % y tomándolo aleatoriamente respectivamente.

Ahora se realiza validación cruzada con este método lineal que consiste en una prueba fuera de la muestra para evaluar los resultados de aplicar este clasificador a la base de datos, y esta consiste en entrenar con todos los datos sin el que voy a predecir. Se obtiene la matriz de confusión que se muestra:

Figura 38: Matriz de confusión con Validación cruzada



Para esta se obtuvo una precisión de 0.7821.
Las tasas de error para la validación cruzada son las siguientes

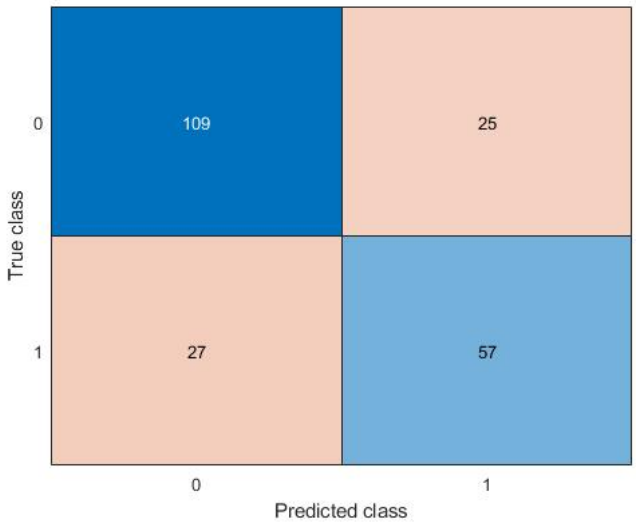
- $EC1$=0.1579
- $EC2$=0.3223

- TGE=0.2219

4.2.2. Clasificador cuadrático

Con este clasificador se obtuvo la siguiente matriz de confusión

Figura 39: Matriz de confusión para el Clasificador cuadrático

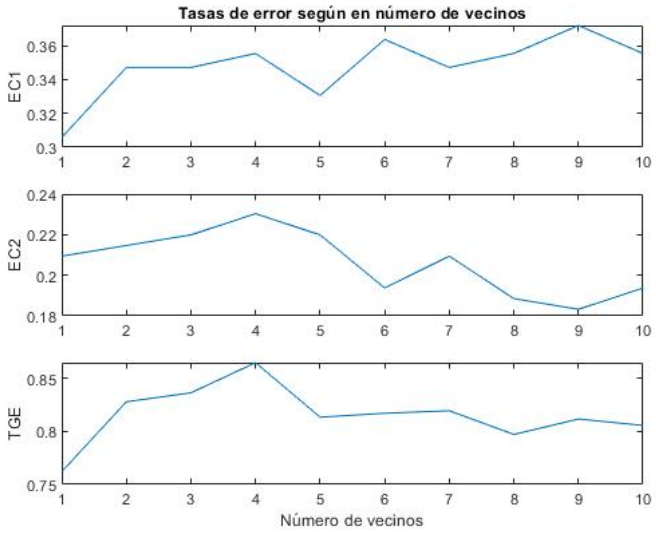


El porcentaje de acierto obtenido es 0.7615.

4.2.3. Clasificador KNN

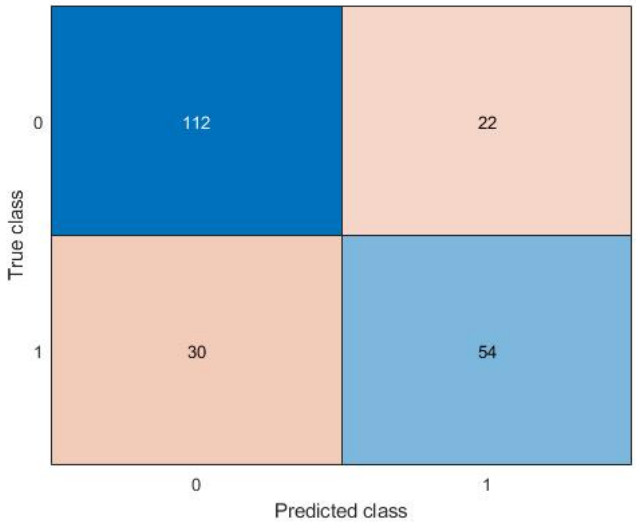
Para este clasificador se realizó un análisis del número optimo de vecinos a tomar, se llegó a que este número es de 5 pues, de esta manera se obtiene un mejor porcentaje de acierto dado que con 6 vecinos el porcentaje de acierto empieza a disminuir aunque luego vuelve a aumentar pero ya podría estar ocurriendo sobreestimación. Este cálculo de vecinos se hizo con las tasas de error mencionadas en los métodos anteriores, esto para diferentes números de vecinos. Los resultados de las tasas de error se muestran en la siguiente figura

Figura 40: Tasas de error para diferentes números de vecinos con el método KNN



Se obtuvo la siguiente matriz de confusión:

Figura 41: Matriz de confusión para el Clasificador KNN

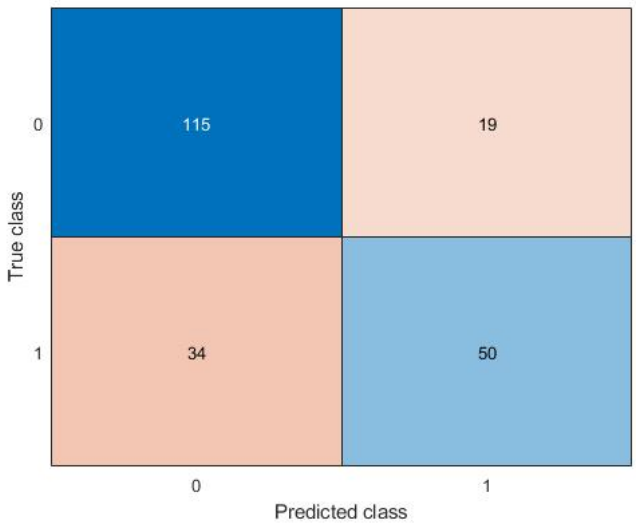


El porcentaje de acierto obtenido es 0.7615.

4.2.4. Clasificador Máquina vector soporte

Para este clasificador se obtuvo la siguiente matriz de confusión

Figura 42: Matriz de confusión para el Clasificador Máquina vector soporte



El porcentaje de acierto obtenido es 0.7569 con una tasa de error generalizado de 0.2234 que se calculó haciendo primero validación cruzada con el comando de matlab crossval y luego con el comando kfoldLoss.

El mejor clasificador para la base de datos resulta ser el clasificador linear pues presenta un mayor porcentaje de acierto de acuerdo a la medida de precisión traza sobre total que fue del 79,8 % de acierto. Sin embargo también se calcularon otras métricas para evidenciar la precisión de los métodos, con esto se sigue observando que para esta base de datos, el clasificador lineal es la mejor opción. Los resultados se muestran en la siguiente figura

Figura 43: Diferentes métricas de precisión

	Linear	Linear entrenamiento aleatorio	Linear cross validation	Cuadrático	KNN	MVS
Accuracy	0.798	0.777	0.782	0.761	0.761	0.757
Precision	0.774	0.692	0.735	0.695	0.711	0.725
Recall	0.667	0.750	0.686	0.679	0.643	0.595
F1	0.716	0.720	0.709	0.687	0.675	0.654

4.3. Clasificación supervisada para cada uno de los dos grupos

El proceso anterior ha permitido hacer un análisis general de la clasificación para dos grupos presentes en la población, sin embargo, en esta parte el objetivo es estudiar el comportamiento individual de cada uno de los registros que quedaron en esos dos grupos, es decir, se busca hacer una clasificación supervisada en aquellos individuos que el casificador de k-medias asignó en el grupo 1 y aquellos que están en el grupo 2.

Luego de obtener la clasificación en dos grupos por el método K-medias, se aplicó cada uno de los clasificadores supervisados utilizados anteriormente, obteniendo así las siguientes matrices de confusión, donde al lado izquierdo se encuentrán los valores reales y en la parte de arriba los clasificados.

Figura 44: Matriz de confusión para el Clasificador lineal con 2 grupos

	Grupo 1	
	0	1
0	40	6
1	5	13

	Grupo 2	
	0	1
0	2	5
1	5	19

Figura 45: Matriz de confusión para el Clasificador lineal con 2 grupos

	Grupo 1	
	0	1
0	41	8
1	11	4

	Grupo 2	
	0	1
0	1	8
1	1	21

Figura 46: Matriz de confusión para el Clasificador SVM con 2 grupos

	Grupo 1	
	0	1
0	48	2
1	9	5

	Grupo 2	
	0	1
0	8	1
1	0	22

Posteriormente, para realizar un análisis más detallado del comportamiento de la clasificación, se calcularon las diferentes métricas de precisión, recall y F1 score mencionadas anteriormente.

Figura 47: Métricas de precisión para los 3 tipos de clasificadores

	Clasificador KNN			
	Accuracy	Precision	Recall	F1
Grupo 1	0.7031	0.3333	0.2667	0.2963
Grupo 2	0.7097	0.7241	0.9545	0.8235

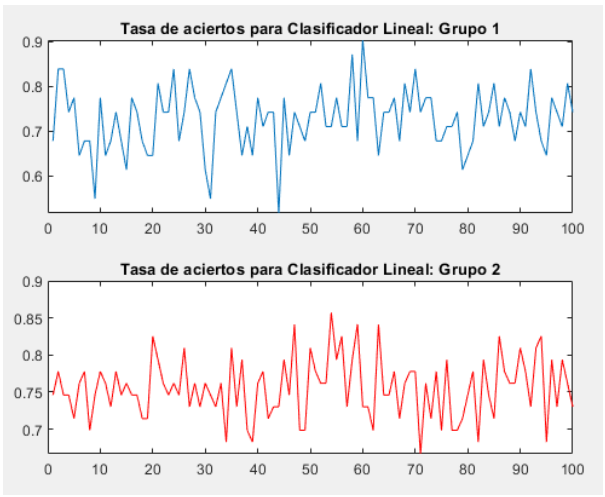
	Clasificador Lineal			
	Accuracy	Precision	Recall	F1
Grupo 1	0.8281	0.6842	0.7222	0.7027
Grupo 2	0.6774	0.7917	0.7917	0.7917

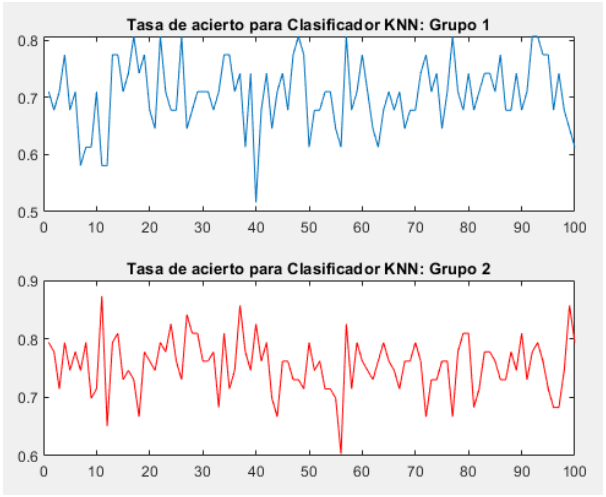
	Clasificador SVM			
	Accuracy	Precision	Recall	F1
Grupo 1	0.8281	0.7143	0.3571	0.4762
Grupo 2	0.9677	0.9565	1.0000	0.9778

En estas tablas se puede ver coóo se alcanza un mayor porcentaje de aciertos para ambos grupos con el clasificador de máquina vector soporte, pero a su vez, la medida de recall, a aquella que le estamos dando más peso dado que se busca tener el mayor porcentaje de aciertos para la etiqueta 1, no presenta un buen rendimiento en este clasificador para el grupo 1. Por eso, y dado a que estamos haciendo pruebas basadas en la escogencia de datos de entrenamiento y validación aleatorios, repetiremos este proceso cien veces, en donde se calculen las dos méticas de interés anteriores para cada iteración, y se grafiquen estos resultados, para obtener así un valor promedio de ocurrencia para dichas métricas, con los

cuales se pueda buscar ua conclusión más acertada respecto al método de clasificacíón más adecuado a utilizar.

Los resultados de este proceso se muestran a continuación





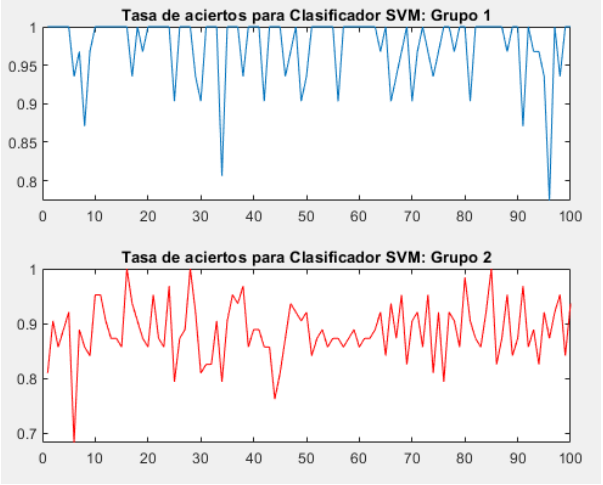
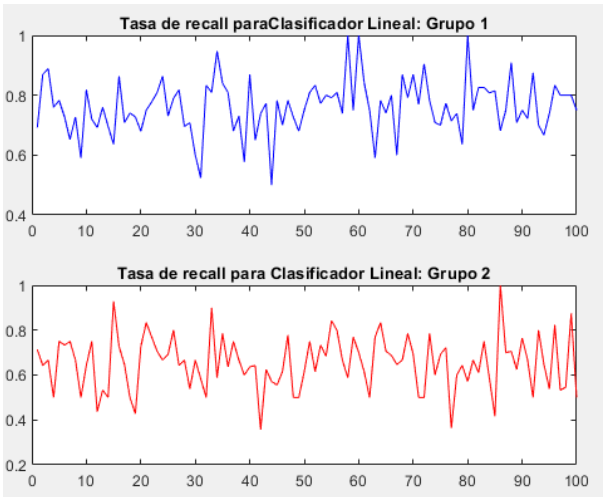
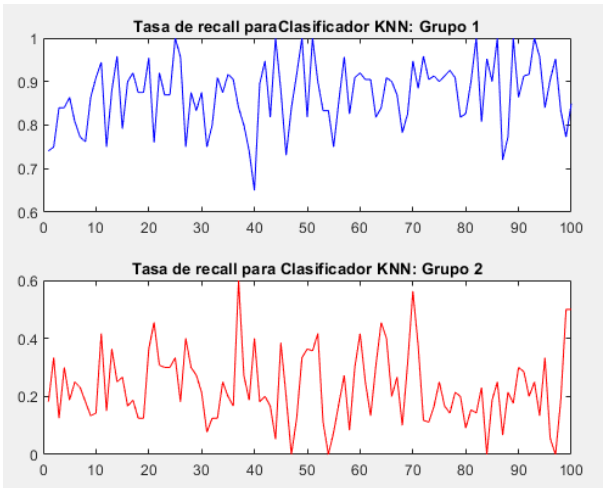


Figura 48: Tasa de aciertos para cien iteraciones





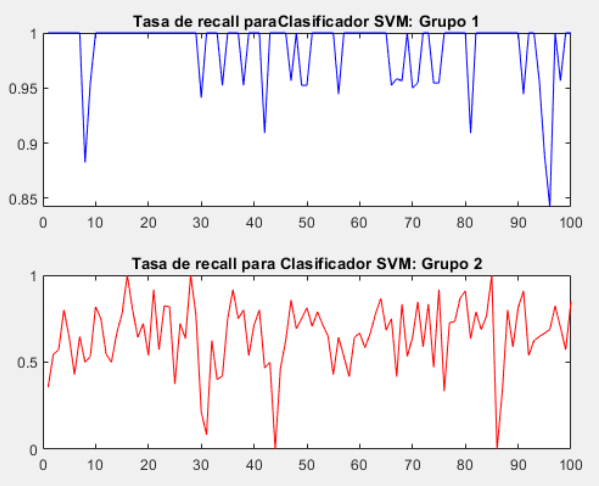


Figura 49: Tasa de recall para cien iteraciones

Se puede ver que rara ambas métricas se tiene una amplia variación en los valores para los diferentes conjuntos de individuos que se utilicen en el entrenamiento y validación, por esto, se realizó adicionalmente una tabla de resultados para las medias de los porcentajes de acierto y de recall, en donde se busca tener una valoración del comportamient en cada clasificador para poder concluir sobre los mismos. Esta tabla se muestran a continuación

	Grupo 1			Grupo 2		
	Lineal	KNN	SVM	Lineal	KNN	SVM
% Recall medio	0.7616	0.8704	0.9819	0.6542	0.2316	0.6503
% Aciertos totales	0.7274	0.7048	0.9632	0.7556	0.7549	0.8863

Figura 50: Frecuencias medias para los aciertos y el recall

Aquí se puede ver para ambas métricas que en la mayoría de casos el clasificador Máquina Vector Soporte tuvo un mejor desempeño medio, donde para el primer grupo se alcanzan porcentajes por encima del 90 % y con valores satisfactorios para le segundo grupo, en comparación al comportamiento de los demás.

4.4. Regresión logística

Para este apartado se desarrollará el modelo de regresón logistica, usando los distintos indicadores de la base de datos como variable respuesta. De esta manera es posible, a través de una función que evalúe una probabilidad binomial, evaluar si un dato pertenece o no a una categoría. El criterio para esta evaluación es que si un dato tiene mas de 70 % de probabilidad a pertencer a una categoría, entonces es asignada a esta.
Para este desarrollo se entrenó el estimador con el 70 % de la muestra elegido de una manera aleatoria. En el desarrollo, además, se debe considerar que se evalúa la aproximación de los datos de entrenamiento y de los datos de validación.

Análisis para el primer indicador
El primer indicador se refiere a elegibilidad de agencia libre

Figura 51: Matriz de confusión para los datos de validación

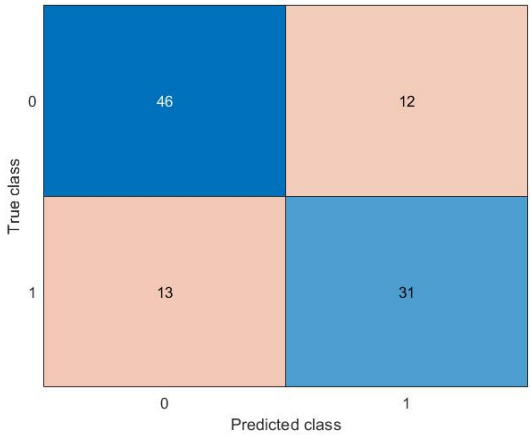
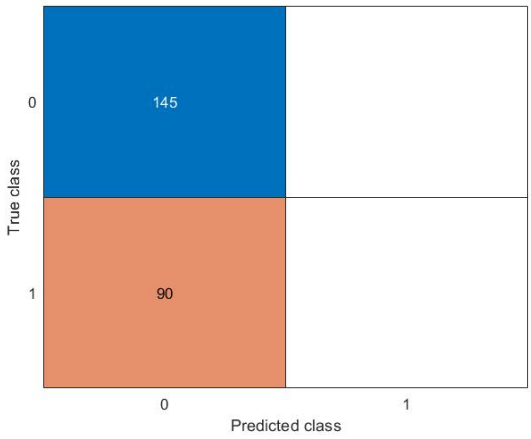


Figura 52: Métricas de precisión

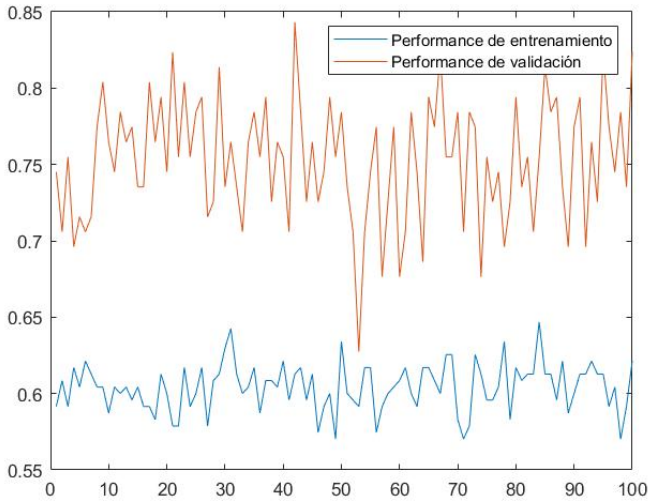
Accuracy	Precision	Recall	F1
0.7549	0.7209	0.7045	0.7126

Figura 53: Matriz de confusión para los datos de entrenamiento



Con esto en mente se puede verificar que, la medida de performarce para los datos de validación es de 0.7549 y para los datos de entrenamiento es de 0.6170. Sin embargo esto fue con una muestra aleatoria. Se visualiza la estabilidad del algoritmo con varias corridas.

Figura 54: Gráfica de los valores de performance con 100 corridas



Con una media para los valores de testeo de 0.7526 y para los valores de entrenamiento de 0.6042.
Aunque el algortimo es estable en cuanto a sus valores, no es el indicador que nos ofrece la mejor medida de precisión.

Análisis para el segundo indicador
El segundo indicador se refiere a elegibilidad de agencia libre para las temporadas.

Figura 55: Matriz de confusión para los datos de validación

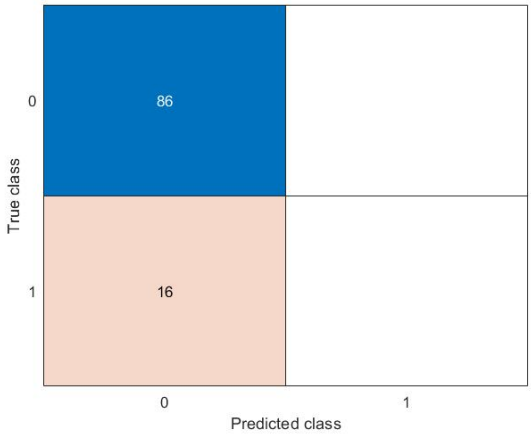
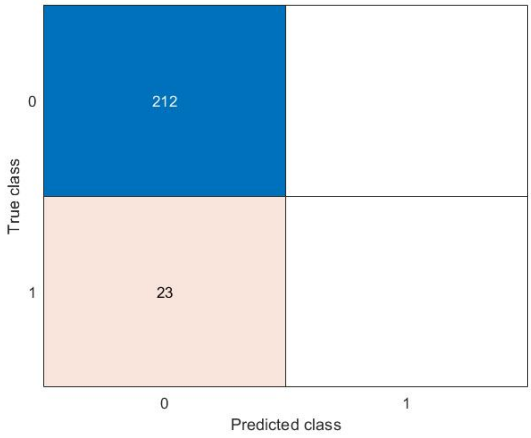
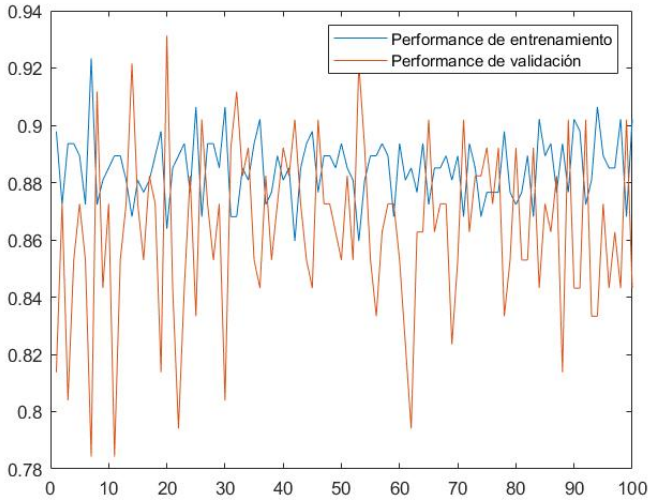


Figura 56: Matriz de confusión para los datos de entrenamiento



Con esto en mente podemos verificar que, la medida de performarce para los datos de validación es de 0.8431 y para los datos de entrenamiento es de 0.9021. Sin embargo esto fue con una muestra aleatoria. Veamos la estabilidad del algoritmo con varias corridas.

Figura 57: Gráfica de los valores de performance con 100 corridas



Con una media para los valores de testeo de 0.8631 y para los valores de entrenamiento de 0.8849.

Para este caso el algoritmo permanece estable para los valores en medidas altas para ambos casos. Es decir, es un buen modelo para este indicador.

Análisis para el tercer indicador

El tercer indicador se refiere al arbitraje

Figura 58: Matriz de confusión para los datos de validación

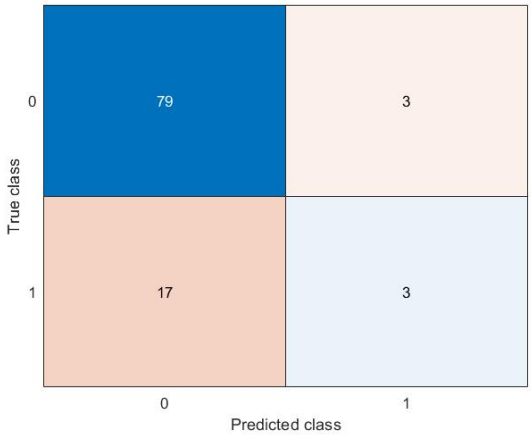
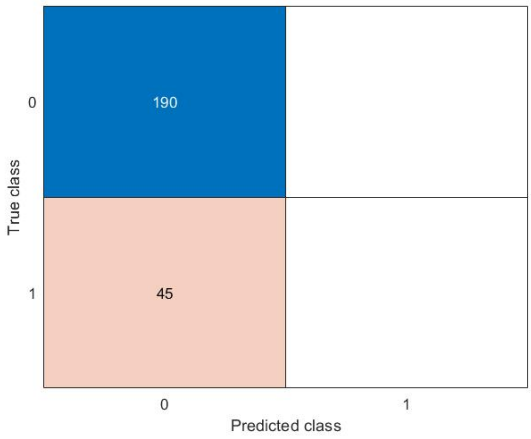


Figura 59: Métricas de precisión

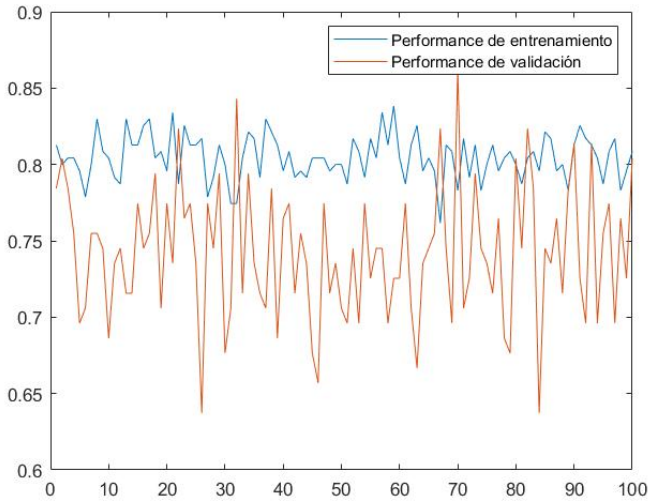
Accuracy	Precision	Recall	F1
0.8039	0.5000	0.1500	0.2308

Figura 60: Matriz de confusión para los datos de entrenamiento.



Con esto en mente se puede verificar que, la medida de performance para los datos de validación es de 0.8039 y para los datos de entrenamiento es de 0.8085. Sin embargo esto fue con una muestra aleatoria. Se visualiza la estabilidad del algoritmo con varias corridas.

Figura 61: Gráfica de los valores de performance con 100 corridas.



Con una media para los valores de testeo de 0.7423 y para los valores de entrenamiento de 0.8046
Para este caso el algoritmo permanece estable para los valores en medidas altas para ambos casos. Es decir, es un buen modelo para este indicador.

Análisis para el cuarto indicador

El cuarto indicador se refiere a arbitraje para el segundo semestre

Figura 62: Matriz de confusión para los datos de validación

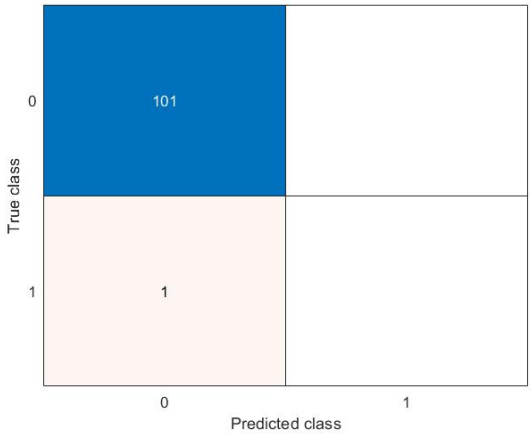
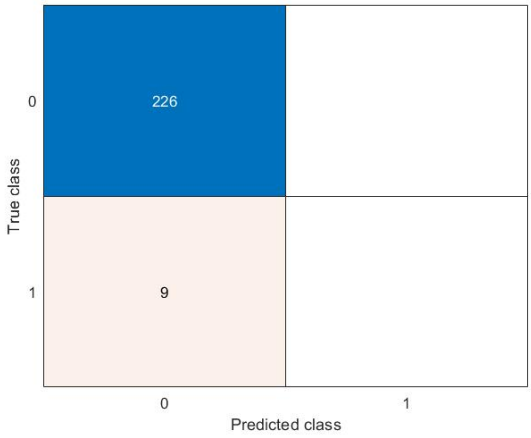
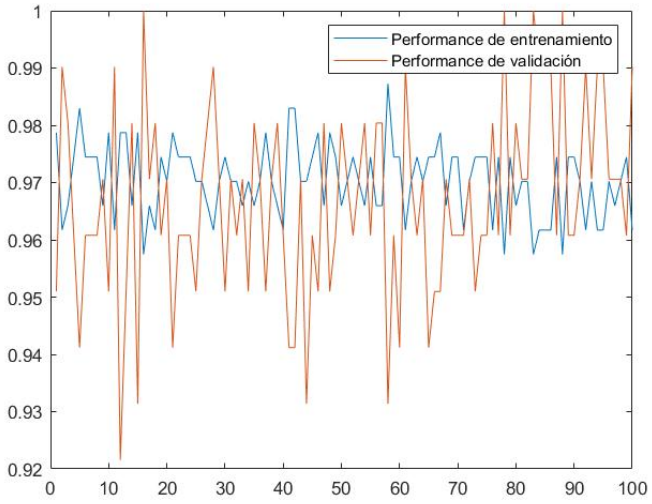


Figura 63: Matriz de confusión para los datos de entrenamiento



Con esto en mente se puede verificar que, la medida de performance para los datos de validación es de 0.9902 y para los datos de entrenamiento es de 0.9617. Sin embargo esto fue con una muestra aleatoria. Se visualiza la estabilidad del algoritmo con varias corridas.

Figura 64: Gráfica de los valores de performance con 100 corridas



Con una media para los valores de testeo de 0.9668 y para los valores de entrenamiento de 0.9704

Para este caso el algoritmo permanece estable para los valores en medidas muy altas para ambos casos. Es decir, es un buen modelo para este indicador. Aunque este indicador, en cuanto a la muestra no sea muy representativo, poderlo modelar es un recurso muy apropiado.

Con los resultados obtenidos se puede decir, que a partir de la modelación logística se obtienen valores muy buenos para los jugadores que no alcanzan el arbitraje en estas temporadas, y para los jugadores que no alcanzan la elegibilidad de agencia libre en estas temporadas, todo esto contando con una buena estabilidad. Esto significa que la clasificación es apropiada para determinar aquellos jugadores que tengan el indicador cero en su etiqueta, pero para aquellos que tienen su etiqueta de 1, punto en el que nos estuvimos centrando anteriormente, se evidencia que la presición y recall en la mayoría de casos es nula.

5. Resultados y conclusiones

Se puede concluir que algunas de las variables deportivas más representativas en la base de datos pueden ser modeladas por la combinación lineal de otras variables deportivas presentes, sin embargo, para aquellas variables que no son deportivas, es recomendable hacer análisis mucho más detallados sobre las variables para identificar un modelo, ya sea lineal robusto o no lineal, por el cual se pueda tener un R^2 mayor al momento de predecir su comportamiento. Además, se hace evidente que para clasificar los individuos según sus etiquetas es muy importante conocer la naturaleza de estas e identificar las prioridades que se tienen al momento de clasificar a los jugadores. Se pudo ver que los clasificadores supervisados que mejor se comportan para la clasificación de individuos que cumplen estar disponibles para ser elegidos por las agencias, aquellos identificados con la etiqueta uno, son el Discriminante Lineal para clasificación generalizada en dos grupos y Máquina Vector soporte, haciendo un análisis intragrupos. Por el contrario, si lo que se desea es estimar aquellos jugadores los cuales estan etiquetados con ceros, lo más recomendable es hacer una regresión logística para cada indicador y así se obtendrán mejores resultados.

Referencias

[1] D. Peña. Análisis de datos multivariantes, Ene. 2002.

[2] Peter J. Rousseeuw and Mia Hubert. Anomaly detection by robust statistics. *WIREs Data Mining Knowl Discov*, pages 1–14, 2018.

[3] Pay for play: Are baseball salaries based on performance?