Análisis de los jugadores de las cinco grandes ligas europeas de fútbol durante la temporada 2019-2020

Asignatura: Análisis, Monitorización y Diagnóstico de Procesos Multivariantes

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones (UPV)

Autores: Javier Porcel Marí
José Nicolás Granero Moreno

Contenido

Resumen

Este trabajo consiste en analizar diferentes datos referentes a los jugadores que participaron en la temporada de fútbol 2019-2020 en las cinco grandes ligas europeas (Premier League en Inglaterra, Ligue 1 en Francia, Bundesliga en Alemania, Serie A en Italia y La Liga en España) mediante técnicas aprendidas en la asignatura “Análisis, Monitorización y Diagnóstico de Procesos Multivariantes”. Con estas técnicas, se buscará realizar un análisis exploratorio de los datos con la técnica del PCA. Y mediante la técnica PLS, se intentará ver qué relación existe entre las variables de los jugadores y las posiciones de éstos, además de comparar con técnicas de clasificación vistas en otras asignaturas.

1. Introducción

Hoy en día se vive una gran revolución digital en la que el análisis de datos forma parte de la toma de decisiones de la gran mayoría de organizaciones volviendo a éstas más competitivas. Entre este tipo de organizaciones, se encuentran los clubs de fútbol. Concretamente, el análisis de datos pasa a ser una buena herramienta para formar plantillas de jugadores bien balanceadas como para aprovechar buenas oportunidades que se ofrecen en el mercado de fichajes.

De estas nuevas necesidades del fútbol actual, surge la motivación de este trabajo en el que se busca realizar un análisis profundo mediante diferentes técnicas, tanto mediante técnicas de aprendizaje no supervisado como el PCA como mediante técnicas de clasificación como el SIMCA y el PLS-DA, de las características que poseen los diferentes jugadores que participaron en la temporada futbolística de las cinco grandes ligas europeas en los años 2019-2020. Estas cinco ligas de fútbol analizadas son: Premier League (Inglaterra), Ligue 1 (Francia), Bundesliga (Alemania), Serie A (Italia) y La Liga (España).

Ilustración 1. Logos de las cinco grandes ligas europeas.

Concretamente con el PCA, se buscará realizar un análisis exploratorio de los datos y con las técnicas predictivas (SIMCA y PLS-DA) se tratará de clasificar a los jugadores por posición. Además de las técnicas supervisadas vistas en la asignatura de “Análisis, Monitorización y Diagnóstico de Procesos Multivariantes”, se utilizarán también técnicas vistas en las asignaturas de “Simulación y Redes Neuronales” y “Minería de Datos”. Estas técnicas adicionales engloban redes neuronales, support vector machine, árbol de clasificación, random forest, vecinos más próximos, Naive Bayes, Bagging y Boosting. Estas últimas técnicas que quedan fuera de la asignatura requerirán un PCA previo pues trabajarán con los scores del mismo y serán comparadas con las técnicas que sí son propias de la asignatura.

2. Material y Métodos

2.1. Descripción de la base de datos

La base de datos que se analiza en este trabajo está formada por un conjunto de datos relativos a los jugadores de fútbol que juegan en las principales ligas de fútbol europeas en la temporada 2019-2020. Toda la información se ha obtenido de la siguiente página web: www.fbref.com (página web en la que se documentan estadísticas, resultados e historia de más de 100 competiciones de equipos tanto masculinos como femeninos).

Así pues, las variables que proporciona esta base de datos, junto con una explicación de la misma, son:

- **RL**: número de identificación del jugador.
- **Jugador**: nombre y primer apellido del jugador.
- **Pais**: nacionalidad del jugador.
- **Posc**: posición en la que juega el jugador (pudiendo ser PO para portero, DF para defensa, CC para centrocampista y DL para delantero).
- Equipo: club de fútbol en el que juega el jugador.
- **Comp**: liga en la que juega el jugador.
- **Edad**: edad del jugador en años.
- Nacimiento: año de nacimiento del jugador.
- **PJ**: número de partidos jugados por el jugador.
- **Titular**: número de partidos jugados desde el inicio por el jugador.
- **Min**: minutos jugados por el jugador.
- **90s**: minutos jugados por el jugador divididos entre 90 ya que 90 minutos equivale a un partido en términos de duración.
- **G_TP**: goles que ha marcado el jugador en toda la temporada por cada 90 minutos jugados descontando los penaltis.
- **Dis**: disparos que realiza el jugador por cada 90 minutos jugados descontando los penaltis.
- **DaP**: disparos a puerta que realiza el jugador por cada 90 minutos jugados descontando los penaltis.
- **Dist**: distancia promedio a la que el jugador dispara medida en yardas.
- **npxG**: goles esperados que no sean de penalti por cada 90 minutos jugados.
- **np:G_xG**: goles marcados por el jugador por cada 90 minutos jugados que no sean de penalti menos los goles esperados por el jugador por cada 90 minutos que no sean de penalti.
- **Pases_Comp.**: pases completados por el jugador por cada 90 minutos jugados.
- **Pases_Int.**: pases intentados por el jugador por cada 90 minutos jugados.
- **Dist.tot.:** distancia promedio, en yardas, que han recorrido los pases completados en cualquier dirección por cada 90 minutos jugados.
- **Dist._prg.:** distancia promedio, en yardas, que han recorrido los pases completados hacia la meta del oponente por cada 90 minutos jugados.

- **Cortos_Cmp.:** pases completados a una distancia de entre 5 y 15 yardas por el jugador por cada 90 minutos jugados.
- **Cortos_Int..:** pases intentados a una distancia de entre 5 y 15 yardas por el jugador por cada 90 minutos jugados.
- **Medios_Cmp.:** pases completados a una distancia de entre 15 y 30 yardas por el jugador por cada 90 minutos jugados.
- **Medios_Int.:** pases intentados a una distancia de entre 15 y 30 yardas por el jugador por cada 90 minutos jugados.
- **Largos_Cmp.:** pases completados a una distancia de más de 30 yardas por el jugador por cada 90 minutos jugados.
- **Largos_Int.:** pases intentados a una distancia de más de 30 yardas por el jugador por cada 90 minutos jugados.
- **Ass**: asistencias realizadas por el jugador por cada 90 minutos jugados.
- **xA**: asistencias esperadas por el jugador por cada 90 minutos jugados.
- **A_xA**: asistencias realizadas por el jugador por cada 90 minutos jugados menos las asistencias esperadas en ese lapso de tiempo.
- **Tkl**: número de entradas realizadas por el jugador por cada 90 minutos jugados.
- **TklG**: número de entradas ganadas por el jugador por cada 90 minutos jugados.
- **Tkl_def**: número de entradas realizadas por el jugador por cada 90 minutos jugados en la zona defensiva (1/3 del campo más cercano a la portería del equipo).
- **Tkl_cent**: número de entradas realizadas por el jugador por cada 90 minutos jugados en el mediocampo (1/3 del campo más cercano al centro del campo).
- **Tkl_ataq**: número de entradas realizadas por el jugador por cada 90 minutos jugados en la zona ofensiva (1/3 del campo más cercano a la portería rival).
- **Presion**: número de veces que se aplica presión al jugador del equipo rival que está recibiendo, llevando o soltando la pelota por cada 90 minutos jugados.
- **Pres_exito**: número de veces que se aplica presión al jugador del equipo rival que está recibiendo, llevando o soltando la pelota por cada 90 minutos jugados y se consigue recuperar el balón tras cinco segundos de presión.
- **Pres_def**: número de veces que se aplica presión al jugador del equipo rival que está recibiendo, llevando o soltando la pelota por cada 90 minutos jugados en la zona defensiva (1/3 del campo más cercano a la portería del equipo).
- **Pres_cent**: número de veces que se aplica presión al jugador del equipo rival que está recibiendo, llevando o soltando la pelota por cada 90 minutos jugados en el mediocampo (1/3 del campo más cercano al centro del campo).

- **Pres_ataq**: número de veces que se aplica presión al jugador del equipo rival que está recibiendo, llevando o soltando la pelota por cada 90 minutos jugados en la zona ofensiva (1/3 del campo más cercano a la portería rival).
- **Int**: número de intercepciones por cada 90 minutos jugados.
- **Tkl+Int**: número de entradas más intercepciones por cada 90 minutos jugados.
- **Desp.:** número de despejes por cada 90 minutos jugados.
- **TA**: número de veces que el jugador es penalizado con una tarjeta amarilla por cada 90 minutos jugados.
- **TR**: número de veces que el jugador es penalizado con una tarjeta roja por cada 90 minutos jugados.
- **AereoG**: número de duelos aéreos ganados por cada 90 minutos jugados.
- **AereoP**: número de duelos aéreos ganados por cada 90 minutos jugados.

La variable “Dist” posee catorce valores faltantes que han sido imputados mediante la técnica de regresión de los k vecinos más próximos (en este caso k = 10) utilizando la distancia euclídea.

Las variables “Edad” y “Nacimiento” aportan la misma información a la base de datos por lo que solo se utilizará la variable “Edad” en las diferentes técnicas de este trabajo.

A pesar de que la base de datos original posee cuatro posiciones (porteros PO, defensas DF, centrocampistas CC y delanteros DL), se trabajará eliminando a los porteros por lo que se trabajará solo con defensas, centrocampistas y delanteros. Esto se debe a que debido a la función totalmente diferenciadora que realizan los porteros las variables que los definen son completamente distintas a las que definen al resto de los jugadores.

Además, se ha decidido eliminar del análisis a todos aquellos jugadores que no hayan jugado durante la temporada un mínimo de 900 minutos lo que equivaldría a 10 partidos de fútbol jugados ya que se considera que los datos que pertenecen a esos jugadores no son lo suficiente robustos.

Para matizar las variables Tkl_def, Tkl_cent, Tkl_ataq, Pres_def, Pres_cent y Pres_ataq, se muestran mediante un diagrama las tres zonas a las que se refieren estas seis variables (es decir, se ha divido el campo en tres rectángulos iguales de modo que Tkl_def y Pres_def se han medido en el rectángulo más cercano a la portería que hay que defender; Tkl_cent y Pres_cent se han medido en el rectángulo correspondiente al centro del campo; y Tkl_ataq y Pres_ataq se han medido en el rectángulo más cercano a la portería rival).

Cabe mencionar para completar la explicación del siguiente diagrama que suponiendo que el jugador analizado perteneciera al equipo rojo, las tres zonas que aparecen en el dibujo serían de abajo a arriba, la zona defensiva, la zona del medio campo y la zona ofensiva.

Ilustración 2. Campo de fútbol dividido en tres zonas.

Por último, es necesario explicar a qué se refieren las variables npxG y xA cuando se refieren a goles y asistencias esperadas, aunque estos valores no se calculen en este trabajo, sino que se obtienen directamente de la fuente. Estos dos valores se obtienen mediante la suposición de que todas las ocasiones de gol poseen una cierta probabilidad de gol para todos los jugadores que realicen el disparo. Así pues, está probabilidad dependerá de la distancia a la portería, el ángulo respecto a la portería, parte del cuerpo con la que se realiza el remate, tipo de asistencia (pase corto, pase en profundidad, pase al hueco, centro, regate, balón parado), etc.

De este modo, se asocia unos goles y unas asistencias esperadas para cada jugador según las oportunidades que haya tenido. Aquel jugador que sea capaz de batir los valores esperados con sus valores reales será considerado un jugador eficaz pues aprovecha las oportunidades que se le presentan.

A modo ilustrativo y como ejemplo, se muestra un diagrama donde se observan las diferentes probabilidades de anotar un tanto cuando un jugador realiza un remate de cabeza a partir de un centro al área:

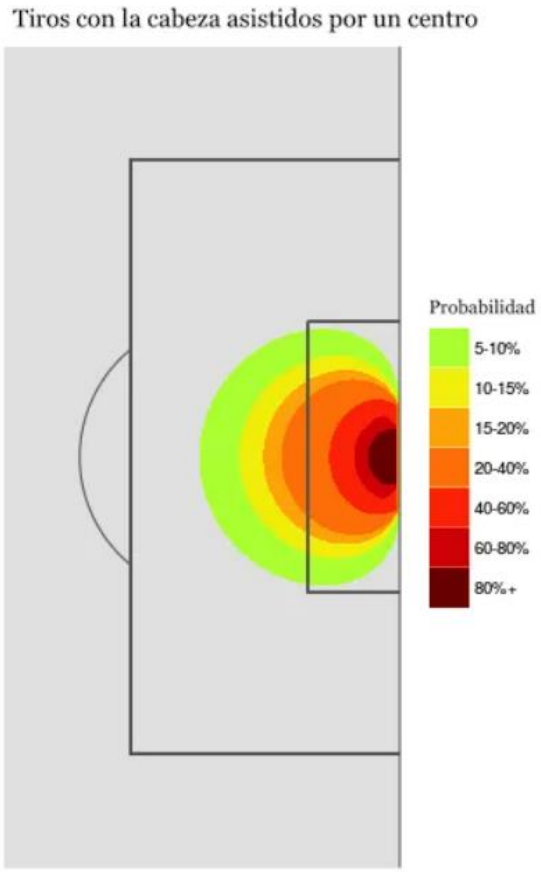


Ilustración 3. Probabilidad de gol al realizar un tiro con la cabeza asistido por un centro según la zona de remate.

Una vez aclarada toda la información que se encuentra en la base de datos, se explican en los siguientes apartados las diferentes técnicas de minería de datos que se han utilizado en este trabajo y se muestran los resultados obtenidos en cada una de ellas.

2.2. PCA

La primera técnica que se utiliza en este trabajo es el análisis de componentes principales o PCA por sus siglas en inglés (Principal Component Analysis). Esta técnica consiste en condensar la información contenida en las K variables primitivas (en nuestro caso K es igual a 41 variables que están correlacionadas entre sí) en un número reducido de A nuevas variables o componentes que ya no son observables o medibles directamente y que ya no están correlacionadas entre sí, pues están definidas como combinaciones lineales de las variables primitivas.

De este modo, se consigue que se manifiesten tanto las relaciones que existen entre individuos como las que existen entre variables. Para transformar los datos de los N individuos y de las K variables en las A componentes principales, se necesita encontrar aquellas proyecciones que permitan reducir la

dimensionalidad tratando de deformar al mínimo la nube de puntos. Para aclarar este concepto mejor se introduce la siguiente expresión:

$$\mathbf{x}_i = \hat{\mathbf{x}}_i + \mathbf{e}_i = t_i\mathbf{p} + \mathbf{e}_i$$

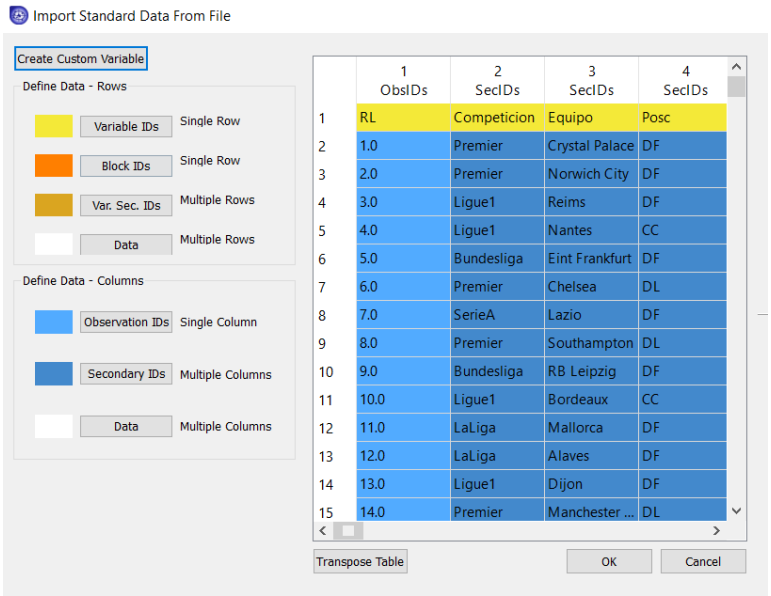
donde $\mathbf{x_i}$ corresponde al vector de datos originales para un individuo de dimensión 1 x K, $\hat{\mathbf{x}}_\mathbf{i}$ corresponde a la estimación del vector anterior con la nueva base, $\mathbf{e_i}$ corresponde al error que se comete en la estimación, t_i corresponde a los scores (los scores corresponden a los datos transformados a la nueva base) y **p** corresponde al vector de loadings que no es más que la nueva base expresada en la base original.

De este modo, la proyección óptima que nos proporciona las A componentes principales sigue la siguiente función objetivo:

$$min\ \mathrm{SCR} = \left\{ \sum_{i=1}^{N} \mathbf{e}_i^T \mathbf{e}_i \right\}$$

Para llevar a cabo esta optimización, existe un algoritmo muy potente llamado algoritmo NIPALS. Para este trabajo, se utiliza el software ASPEN PRO MV que ya nos proporciona las componentes principales de los datos introducidos sin necesidad de reparar en los métodos de resolución.

En primer lugar, se necesita cargar los datos en el ASPEN PRO MV con los valores faltantes ya imputados (la imputación de los datos se encuentra explicada en el subapartado anterior). Con el fin de realizar un buen análisis, es necesario marcar a las variables Competición, Equipo, Posc, País y Jugador como Secondary ID por su naturaleza puramente categórica.



	1 ObsIDs	2 SecIDs	3 SecIDs	4 SecIDs
1	RL	Competicion	Equipo	Posc
2	1.0	Premier	Crystal Palace	DF
3	2.0	Premier	Norwich City	DF
4	3.0	Ligue1	Reims	DF
5	4.0	Ligue1	Nantes	CC
6	5.0	Bundesliga	Eint Frankfurt	DF
7	6.0	Premier	Chelsea	DL
8	7.0	SerieA	Lazio	DF
9	8.0	Premier	Southampton	DL
10	9.0	Bundesliga	RB Leipzig	DF
11	10.0	Ligue1	Bordeaux	CC
12	11.0	LaLiga	Mallorca	DF
13	12.0	LaLiga	Alaves	DF
14	13.0	Ligue1	Dijon	DF
15	14.0	Premier	Manchester ...	DL

Ilustración 4. Captura del ASPEN PRO MV cuando se marcan las variables categóricas como Secondary ID.

Una vez cargados los datos, se excluye a la variable Nacimiento por ser redundante al haber ya una variable que indica la edad y se realiza un escalado y un centrado de los datos previamente a la obtención de las componentes principales para las 41 variables de interés mediante la opción UV para el escalado y MC para el centrado.

Vars Included: 41 # Vars Excluded: 1

	Variable name	Mean	StdDev	Min / Max	Centering	Scaling	Custom	Transfor
	Edad	25.9715	3.95277	16 / 41	MC	UV	1.0	None
	Nacimiento	1992.68	3.95059	1977 / 2002	MC	UV	1.0	None
	PJ	25.7436	6.69629	11 / 38	MC	UV	1.0	None
	Titular	21.3905	7.41829	7 / 38	MC	UV	1.0	None
	Minutos	1894.83	637.047	901 / 3420	MC	UV	1.0	None
	s90	21.0541	7.0788	10 / 38	MC	UV	1.0	None
	G_TP	0.124149	0.158418	0 / 1.1	MC	UV	1.0	None
	Dis	1.24332	0.973357	0 / 6.05	MC	UV	1.0	None
	DaP	0.418714	0.409964	0 / 2.92	MC	UV	1.0	None

Ilustración 5. Captura del ASPEN PRO MV cuando se excluye a la variable "Nacimiento" del análisis.

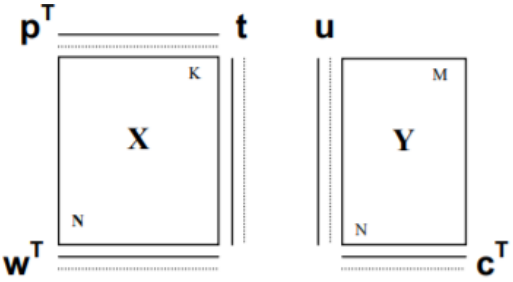
Posteriormente, ya en el apartado de resultados, se obtendrán el número de componentes principales apropiado justificándose esta elección además de realizarse la validación del modelo PCA.

2.3. PLS-DA

La técnica PLS discriminante o PLS-DA trata de predecir un bloque de datos Y a partir de un bloque de datos X al obtener unas componentes que maximicen la covarianza entre las variables predictoras y las variables a predecir (llamadas componentes PLS y que además de ser útiles en la predicción también sirven para encontrar relaciones entre los conjuntos X y Y); y al realizar una regresión sobre estas componentes.

Al igual que con el PCA, lo primero que se debe hacer es escalar los datos para que tengan media cero y centrarlos para que tengan varianza unitaria. Posteriormente se procede a obtener las componentes PLS de manera que mediante una proyección de X se aproximen bien tanto X como Y, y que mediante esta misma proyección se maximice la correlación con Y.

Concretamente, el algoritmo PLS utiliza las siguientes expresiones matemáticas:



$$t_1 = Xw_1;\ t_2 = (X - t_1 p_1^T) w_2;\ \ldots$$

$$X = 1\,\bar{x}^T + TP^T + E$$

$$Y = 1\,\bar{y}^T + UC^T + G = 1\,\bar{y}^T + TC^T + F$$

$U = T + H$ (relación interna)

donde T y P^T corresponden a los scores y a los loadings de la matriz X al realizar un PCA tal y como se ha explicado en el subapartado anterior, W^T corresponde a las diferentes componentes PLS de la matriz X escritas en la base original, C^T corresponde a las diferentes componentes PLS de la matriz Y escritas en la base original y la U corresponde a los valores de la matriz Y utilizando como base las componentes PLS de la matriz Y.

La primera relación que se observa en la imagen corresponde a como se relacionan los scores y los loadings de la matriz X con los vectores w. La segunda relación es exactamente la misma que se ha comentado en el subapartado anterior. Finalmente, la última relación hace referencia a como se relacionan con la matriz Y tanto U como C^T que se definen como se ha explicado en el párrafo anterior.

Una vez explicado el fundamento teórico, se procede a explicar cuál es el procedimiento para realizar una regresión PLS-DA con el software ASPEN PRO MV.

En primer lugar, se introducen los datos completos (habiéndose ya realizado la imputación de valores faltantes mediante la técnica knn) y se seleccionan las variables categóricas para etiquetarlas como Secondary ID al igual que se había hecho con el PCA para posteriormente guardar todos los datos en el mismo bloque X y se excluye del análisis la variable Nacimiento ya que ya existe la variable Edad y sería redundante (como ya sucedía en el PCA).

Para definir el bloque Y con las tres variables dummy, una para cada posición, se va a la pestaña Model y se escoge la opción Set New Model As. Posteriormente se escoge el apartado Observations y donde se muestra la sección Set Classes se escoge by Secondary ID, Posc y finalmente se marca la opción Create PLS-DA model:

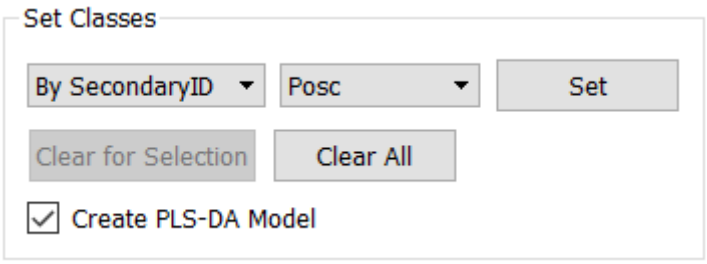


Ilustración 6. Captura del ASPEN PRO MV seleccionándose a la variable "Posc" como bloque Y de un PLS-DA.

2.4. SIMCA

Como método para clasificar a los jugadores por posiciones, se utiliza la técnica SIMCA que consiste en realizar un PCA con los datos de entrenamiento, perteneciendo estos a una sola clase (o posición en nuestro caso) para posteriormente introducir en el PCA los datos de validación de manera que se detecten aquellos datos del set de validación que no pertenezcan a la clase con la que se ha entrenado el modelo ya que se clasificarían como observaciones atípicas en el gráfico SPE-X o como observaciones extremas en el gráfico del T^2 de Hotelling.

De este modo, se necesita realizar tres veces la técnica SIMCA, una vez para cada posición (defensas, centrocampistas y delanteros) pues cada uno de los SIMCA trataría de predecir si un jugador juega en una determinada posición o no lo hace.

2.5. Otras técnicas

Además de las técnicas que ya se han explicado, en este trabajo se utilizan otras técnicas que no son propias de la asignatura siendo estas: máquina de soporte vectorial, Naive Bayes, vecinos más próximos, árbol de clasificación, random forest, bagging, boosting y redes neuronales. Todas estas técnicas se compararán entre ellas y con las anteriores mediante un hold out.

3. Resultados

3.1. PCA

Una vez se han introducido los datos en el ASPEN PRO MV tal y como se ha explicado en el apartado de Material y Métodos, se procede a calcular las componentes principales mediante la opción Auto Fit. Aunque el software calcula 17 componentes principales nos quedaremos solo con las 10 primeras de ellas utilizándose como criterio de selección que el valor propio asociado a cada componente principal sea mayor o igual que uno (la décima componente no posee un valor propio mayor o igual que la unidad, pero debido a su extrema proximidad a la unidad se va a admitir).

Con estas 10 componentes principales se gráfica el SPE-X (Square Prediction Error) y el T2 de Hotelling:

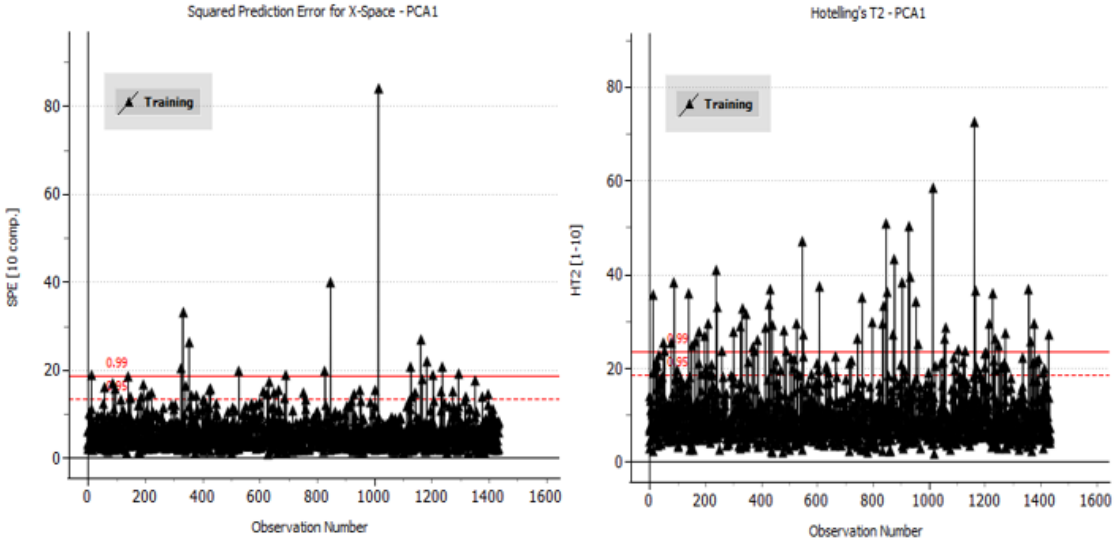


Ilustración 7. Gráficos SPE-X y T^2 de Hotelling.

Tanto en un gráfico como en el otro, se observan dos datos que resaltan sobre el resto de manera notable.

En la gráfica del SPE-X, resalta como dato atípico el dato con ID 1017, que corresponde al jugador Joao Pedro, delantero brasileño que juega en el Cagliari. Para analizar que variables están rompiendo la estructura de correlación en este jugador se visualiza el gráfico de contribuciones:

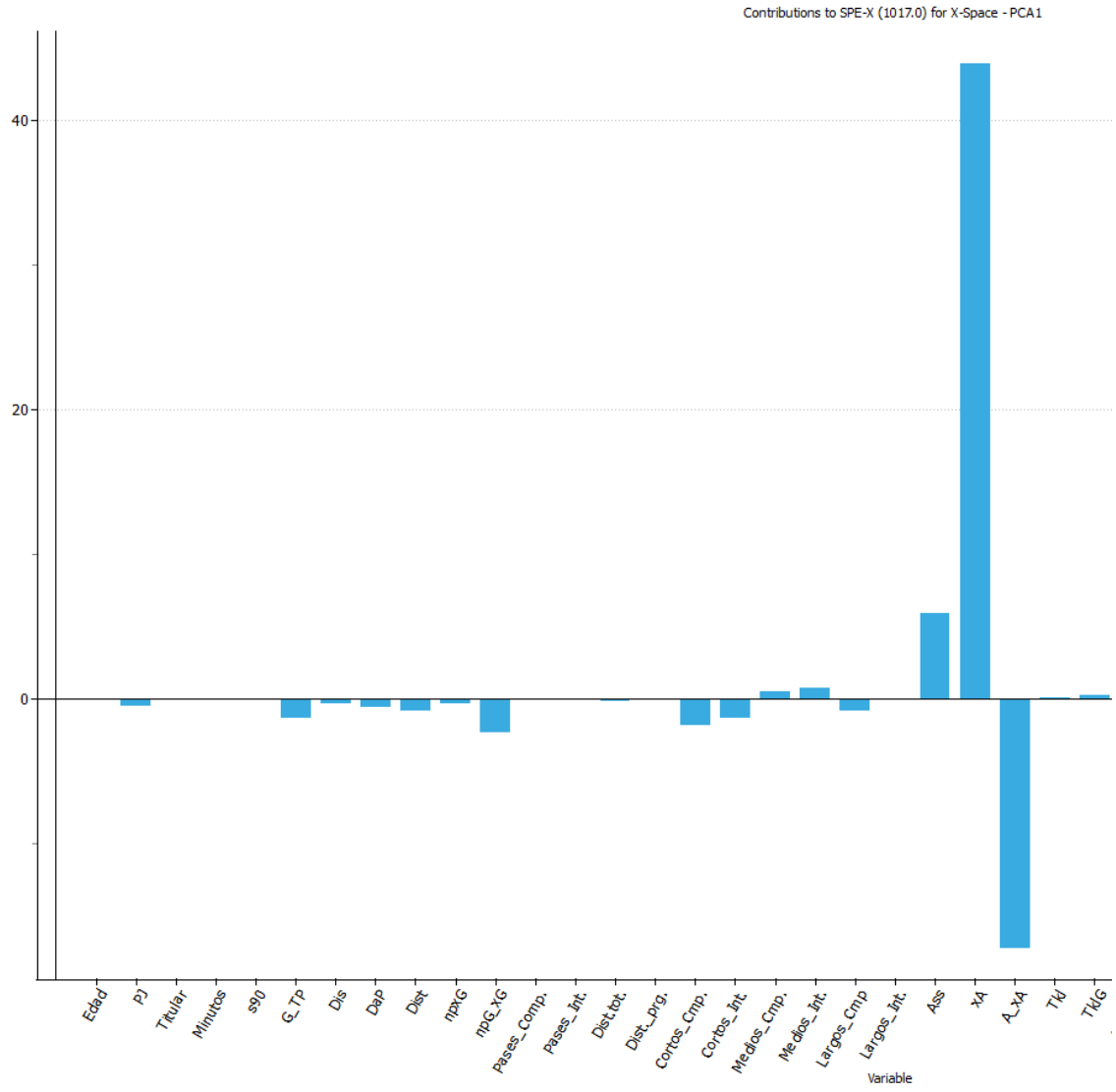


Ilustración 8. Gráfico de contribuciones para el jugador Joao Pedro.

De esta gráfica, se extrae que el jugador Joao Pedro rompe la estructura de correlación ya que ha realizado menos asistencias de lo que cabría suponer, tiene un número de asistencias esperadas muchísimo mayor a lo que cabría suponer y la diferencia entre asistencias y asistencias esperadas es menor de lo que habría que esperar.

Por otro lado, tenemos un dato extremo en el T^2 de Hotelling. Concretamente, el dato extremo corresponde al jugador con ID 1166 Alexis Sánchez, delantero chileno que juega en el Inter de Milán. Su gráfico de contribución es el siguiente:

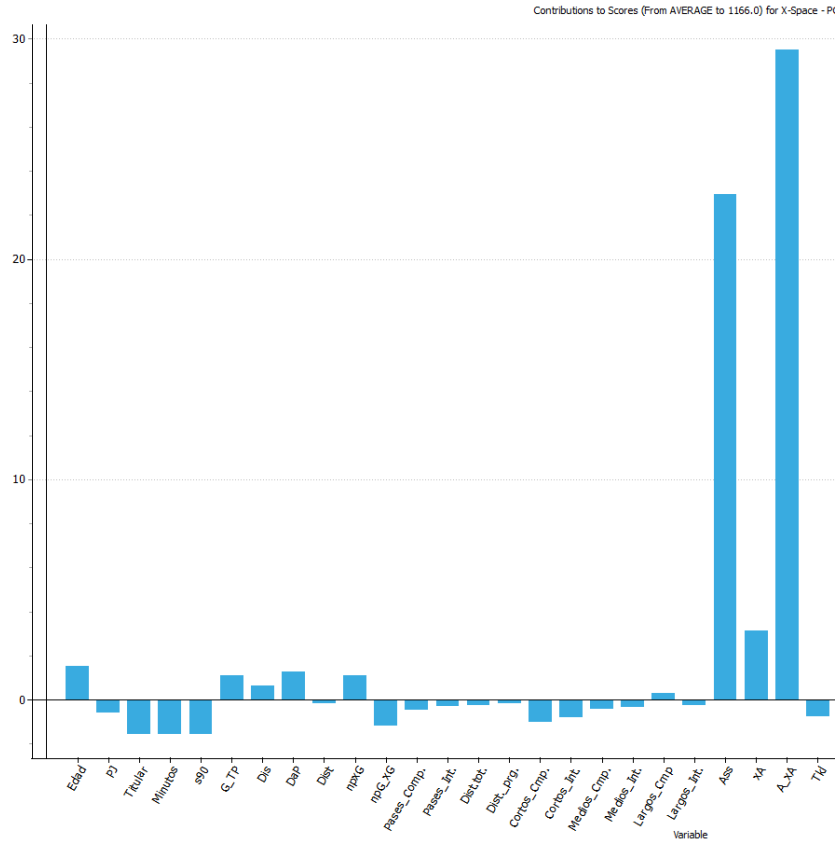


Ilustración 9. Gráfico de contribuciones para el jugador Alexis Sánchez.

En el gráfico de contribución, se observa que es un valor extremo debido a que posee muchas más asistencias que las esperadas.

Para finalizar el análisis de los valores atípicos y extremos, se representa un scatter plot con las asistencias frente a la diferencia de asistencias y asistencias esperadas para vislumbrar donde la posición de estos dos jugadores en la nube de puntos:

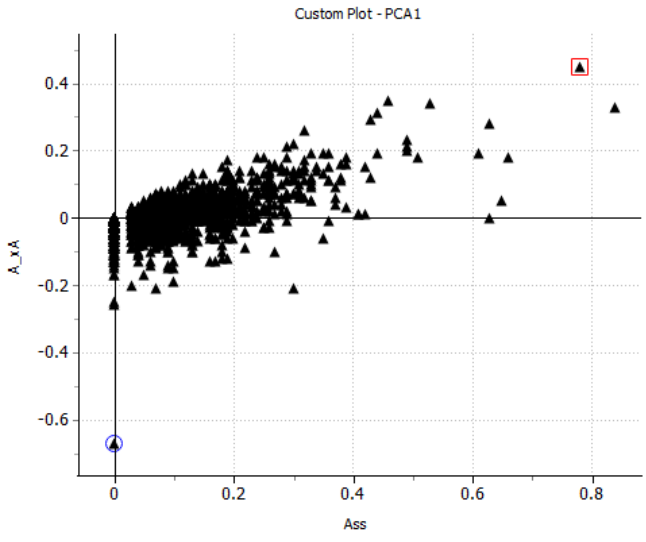


Ilustración 10. Custom plot en el que se observa el dato atípico (círculo azul) y el dato extremo (cuadrado rojo).

donde el triángulo con el círculo azul corresponde al valor atípico y el triángulo con el cuadrado rojo corresponde al valor extremo. Este gráfico visualiza de una manera muy clara como el valor atípico se aleja de la forma ovalada de la distribución mientras que el valor extremo se sitúa al límite derecho de la distribución.

Para mejorar el PCA, se considera adecuado extraer estos dos jugadores del análisis y se realiza un nuevo PCA.

En este nuevo PCA, nos quedamos nuevamente con 10 componentes principales siguiendo el mismo criterio que anteriormente. Concretamente, estos son los parámetros de lo que se demostrará que es el modelo PCA definitivo:

Model 2	
Name	PCA2
Type	PCA
A	10
N	1437
K (tot...	41
Stand...	1
Batch ...	0
Stand...	0
Total ...	0.86526 | 0.86109
1	0.31650 | 0.31447
2	0.15993 | 0.15941
3	0.10724 | 0.10656
4	0.08876 | 0.08917
5	0.04126 | 0.03887
6	0.03683 | 0.03464
7	0.03486 | 0.03690
8	0.03097 | 0.03202
9	0.02567 | 0.02560
10	0.02323 | 0.02345

Ilustración 11. Captura del ASPEN PRO MV donde se resumen las características del PCA definitivo.

Para comprobar que efectivamente el modelo es el definitivo, se vuelven a revisar los gráficos SPE-X y el T2 de Hotelling, observándose que no es necesario eliminar ningún jugador más:

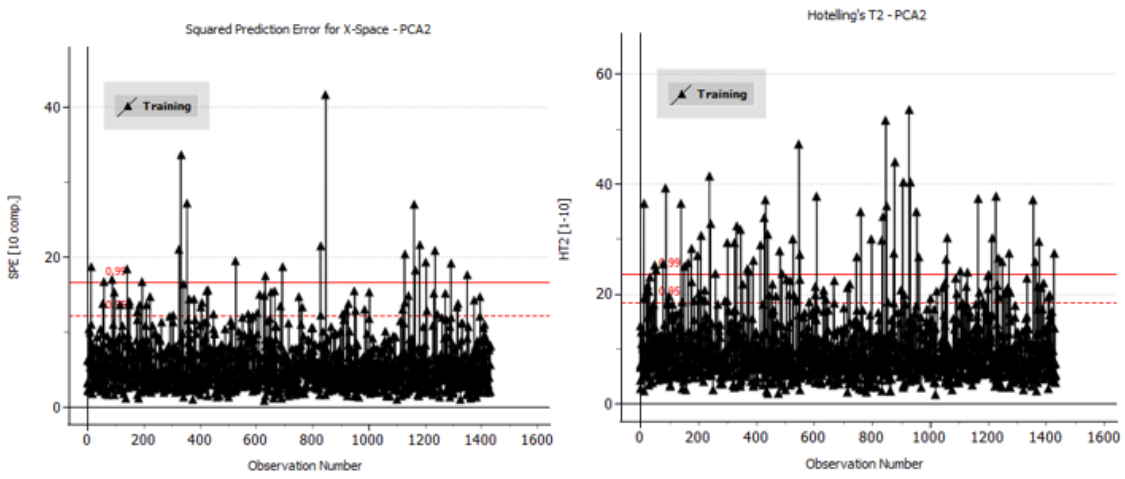


Ilustración 12. Gráficas SPE-X y T^2 de Hotelling para el PCA definitivo.

Una vez se tiene el PCA validado, se muestra el R^2 Variable Summary donde se muestra un gráfico de barras explicando que R^2 de cada variable explica cada componente:

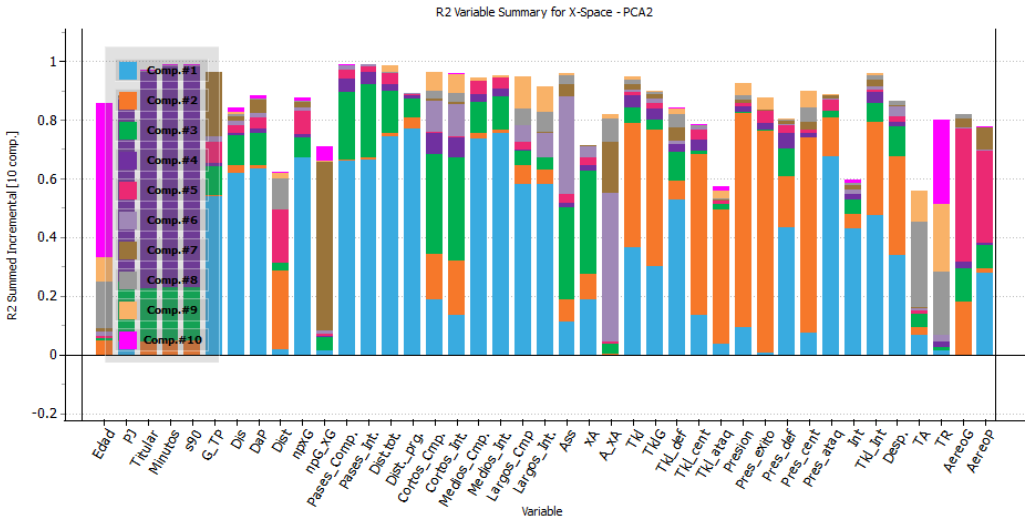


Ilustración 13. Gráfica donde se muestra que porcentaje de la variabilidad de cada variable explica cada componente.

Además, se grafican en primer lugar el score plot y el loading plot para la primera y segunda componente que son las que poseen mayor poder explicativo:

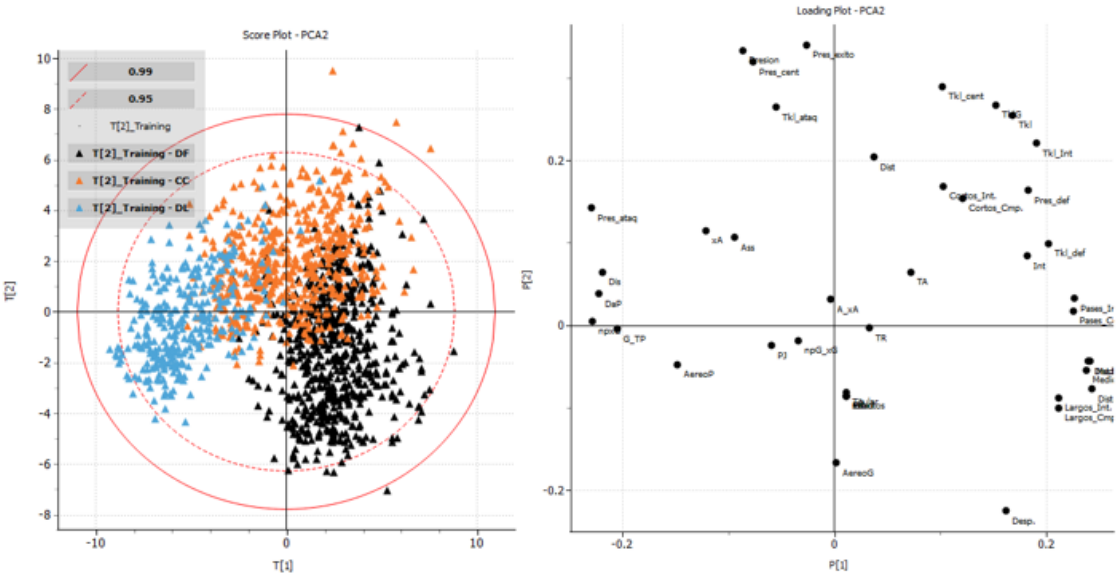


Ilustración 14. Score plot y loading plot de las dos primeras componentes.

En el score plot, se observa cómo se distribuyen las diferentes posiciones perteneciendo el cuadrante inferior derecho y parte del superior derecho a los defensas, el cuadrante superior derecho y parte del cuadrante superior izquierdo a los centrocampistas, y el cuadrante superior izquierdo y parte del cuadrante inferior izquierdo a los delanteros. Además, las variables que aparecen en el loading plot se corresponde con las zonas anteriormente descritas. Los defensas destacan por los despejes, los duelos aéreos ganados, los pases largos, las

tarjetas tanto rojas como amarillas, intercepciones, entradas en defensa y presión en defensa. Los centrocampistas destacan en la distancia de los tiros, pases cortos, entradas en el centro del campo, presiones con éxito y asistencias. Finalmente, los delanteros destacan en goles menos penaltis, goles esperados, disparos a puerta, balones aéreos perdidos y en presión cerca de la portería rival.

Por otro lado, se pueden explorar el resto de las componentes principales para ver si se pueden encontrar otras relaciones entre las variables que sean de interés. Por ejemplo, si se representan la componente 1 contra la 10, y la componente 3 contra la 4 se obtienen los siguientes loading plot:

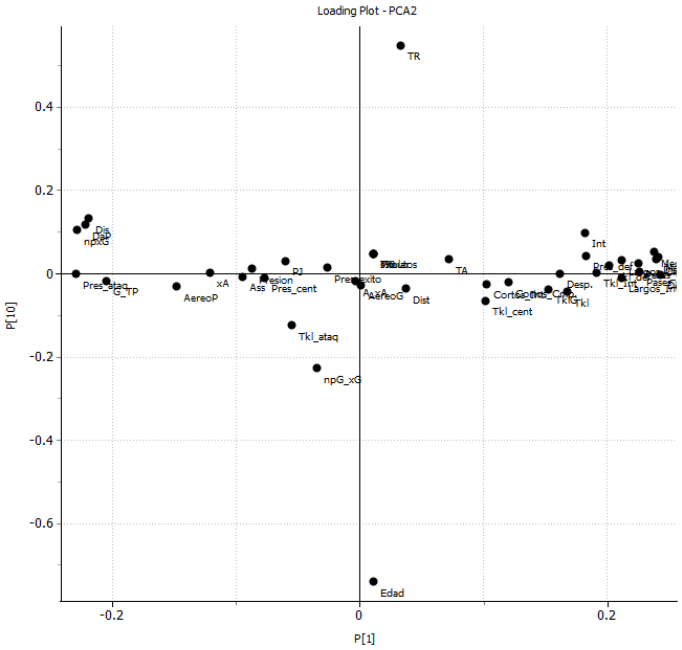


Ilustración 15. Loading plot P[1] vs P[10].

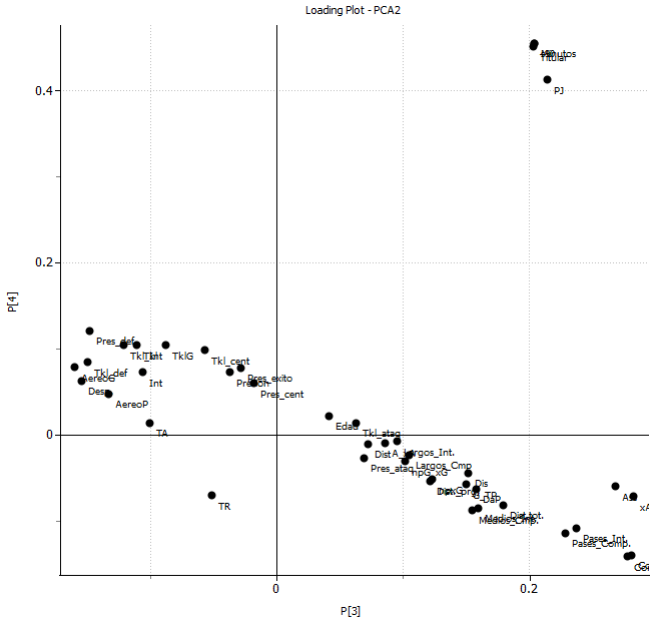


Ilustración 16. Loading plot P[3] vs P[4]

En la primera gráfica, se observa claramente como las tarjetas rojas y la edad se encuentran correlacionadas negativamente de manera que se interpreta que cuanto más joven es un jugador mayor agresividad y menor prudencia posee en el campo por lo que acaba teniendo mayor número de tarjetas rojas mientras que los jugadores de mayor edad son más precavidos y pacíficos.

En la segunda gráfica, se observa como las variables partidos jugados, minutos y partidos jugados como titular se encuentran correlacionados positivamente lo cual es bastante lógico.

3.2. PLS-DA

Una vez se han introducido los datos tal y como se ha explicado en el apartado de Material y Métodos, se realizan los siguientes pasos para validar el modelo:

- Se observan los gráficos SPE-X y T^2 de Hotelling para eliminar del modelo a los valores atípicos y los valores extremos.
- Se eliminan las variables un VIP menor a 0,8 y también aquellas que en el apartado Coeffs muestran unos coeficientes muy pequeños para las tres posiciones a predecir.
- Se vuelve a repetir el primer y segundo paso hasta que ya no queda ningún jugador ni ninguna variable que eliminar del modelo.

De este modo se eliminan cinco jugadores: Troy Deeney (delantero inglés del Watford con ID 334), Kylian Mbappe (delantero francés del PSG con ID 850), Renato Sanches (centrocampista portugués del Lille con ID 1165), Joao Pedro (delantero brasileño del Cagliari con ID 1017) y Alexis Sánchez (delantero chileno del Inter de Milán con ID 1166). Estos dos últimos jugadores son los mismos que se extrajeron del PCA.

Además, se han eliminado las siguientes quince variables (además de la variable Nacimiento por los motivos que ya se han expuesto): Edad, PJ, Titular, Minutos, s90, Dist, npG_xG, Cortos_Cmp., Cortos_Int., Ass, A_xA, Int, TA, TR, AereoG.

Una vez escogidas las observaciones y las variables con las que se va a realizar el PLS-DA se clica sobre Auto Fit, y se va probando a cambiar el número de componentes PLS para averiguar qué número de componentes PLS es el adecuado. En este caso, seguiremos el criterio de quedarnos con el número de componentes PLS que maximice el Q^2.

Siguiendo el criterio mencionado anteriormente, se obtienen 17 componentes PLS con las que se tiene un R^2 del 73,123% de R^2 y un Q^2 del 72,321%.

Model 6	
Name	PLS_DA6
Type	PLS
A	17
N	1434
K (tot...	26
Stand...	1
Batch ...	0
Stand...	1
Total ...	0.73123 | 0.72321
1	0.35404 | 0.35269
2	0.20829 | 0.20741
3	0.07301 | 0.07134
4	0.04501 | 0.04430
5	0.01997 | 0.01970
6	0.01172 | 0.01140
7	0.00864 | 0.00847
8	0.00283 | 0.00213
9	0.00452 | 0.00322
10	0.00099 | 0.00126
11	0.00056 | 0.00071
12	0.00053 | 0.00006
13	0.00012 | 0.00003
14	0.00031 | 0.00004
15	0.00015 | 0.00011
16	0.00008 | 0.00003
17	0.00045 | 0.00033

Ilustración 17. Captura del ASPEN PRO MV donde se resumen las características del PLS-DA.

Respecto a las gráficas que nos son de interés, se tiene el siguiente loading plot que nos muestra que variables se asocian más a según qué posición (las posiciones se encuentran marcadas con un cuadrado rojo) reflejándose la misma reflexión que se observaba con el score plot y el loading plot del PCA.

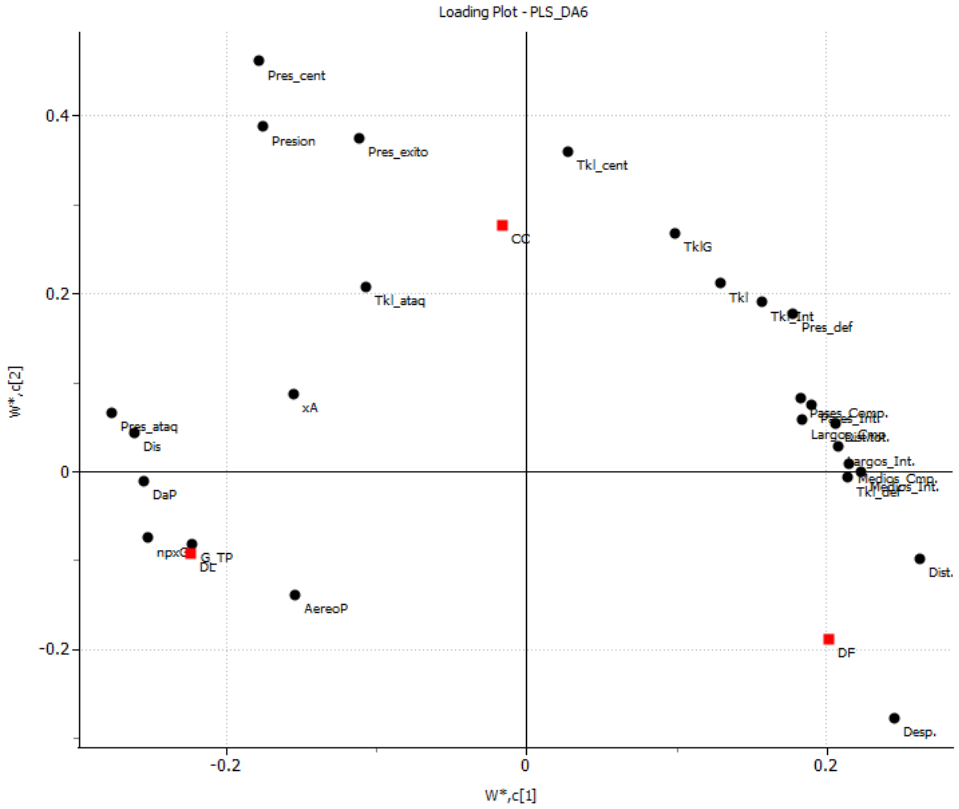


Ilustración 18. Loading plot del PLS-DA.

Finalmente, se muestran las siguientes gráficas que muestran como las componentes PLS explican tanto las posiciones como las variables mediante la opción R^2 Variable Summary.

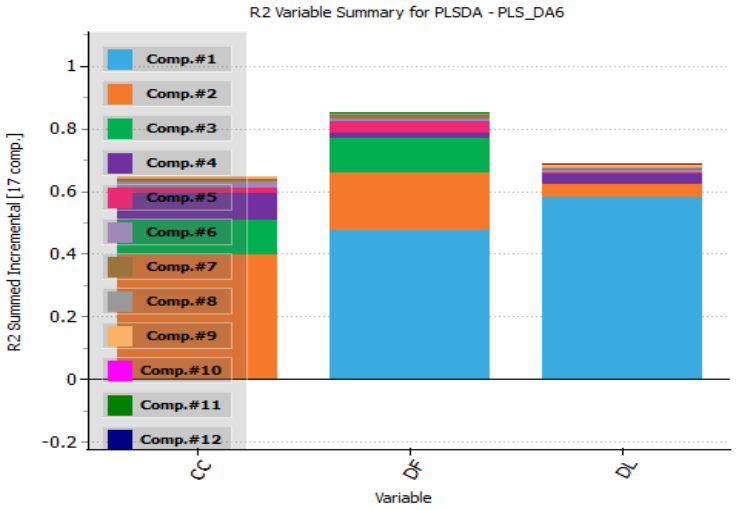


Ilustración 19. Porcentaje de variabilidad explicada para cada posición por cada componente PLS.

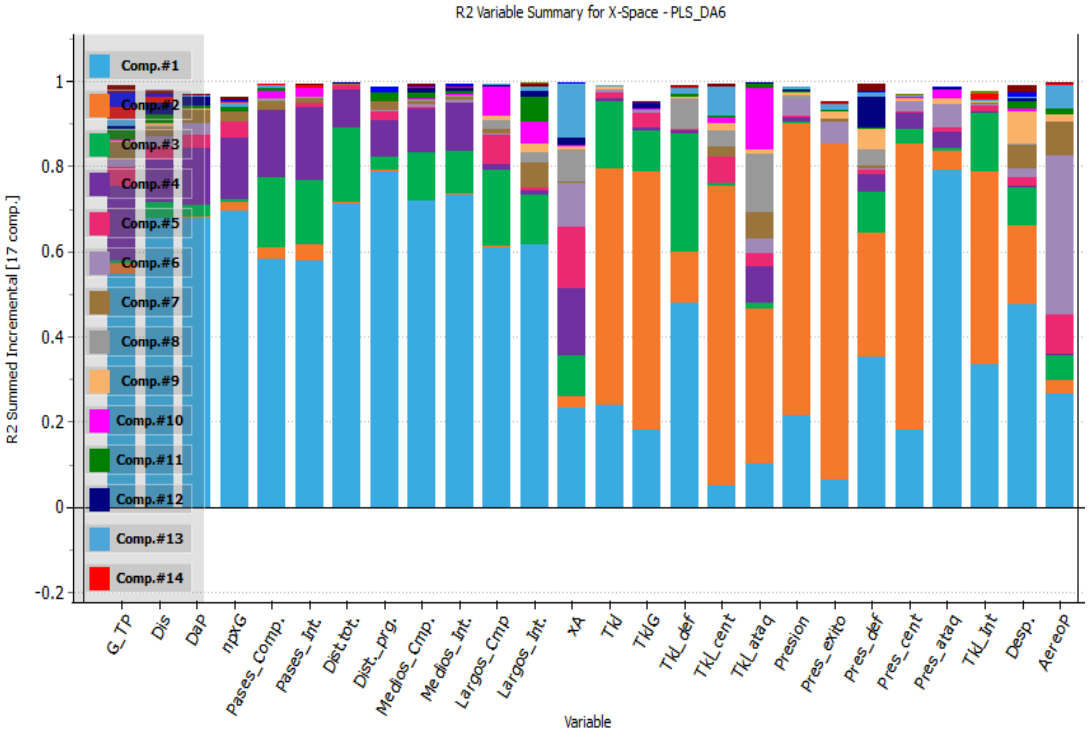


Ilustración 20. Porcentaje de variabilidad explicada para cada variable por cada componente PLS.

3.3. Comparación entre técnicas de clasificación

En este apartado, se ha realizado una comparación entre técnicas de clasificación vistas en la asignatura de Minería de Datos, Redes Neuronales y los modelos PLS-DA y Simca. Para ello, se ha realizado un hold out dividiendo la muestra entre un 75% para entrenar los modelos y un 25% para validarlos,

eligiendo las muestras de entrenamiento y validación de manera aleatoria. Previamente a los modelos de clasificación, se ha realizado un PCA como en un apartado anterior, utilizando también el software Aspen Pro-MV, para detectar datos atípicos y extremos previos a la clasificación y para asegurarse que los datos en la muestra de validación siguen la misma estructura de correlación y no hay ningún dato extremo. En el caso de los modelos Simca y PLS-DA, esto no es necesario ya que el PLS-DA esto lo tiene implícito y con el método Simca, su clasificación gira en torno a ver si la observación queda dentro de los límites del gráfico de error en la predicción y del gráfico de la T^2 de Hotelling.

En el conjunto de entrenamiento, se ha eliminado un dato siendo éste un dato atípico, se ha descartado únicamente ese porque considerando que hay más de 1000 observaciones los elementos que están por encima del límite de control del 99% no eran muy elevados pudiendo pertenecer a ese 1%, exceptuando el dato que se ha eliminado de la muestra. Y en el conjunto de validación, se ha decidido no tener en cuenta otro dato, siguiendo el mismo criterio que en el conjunto de entrenamiento, ya que era un dato extremo con un valor muy elevado en el gráfico de la T^2 de Hotelling. Posteriormente, se han realizado los métodos de clasificación con variables latentes habiéndose decidido por 10, ya que habían 9 cuya variabilidad explicada era mayor al 10%y la décima estaba relativamente cerca de ese valor. A continuación, se muestran los gráficos de los errores en las predicciones y el gráfico de la T^2 de Hotelling con la muestra de entrenamiento, sin el dato dicho anteriormente, y el set de predicción.

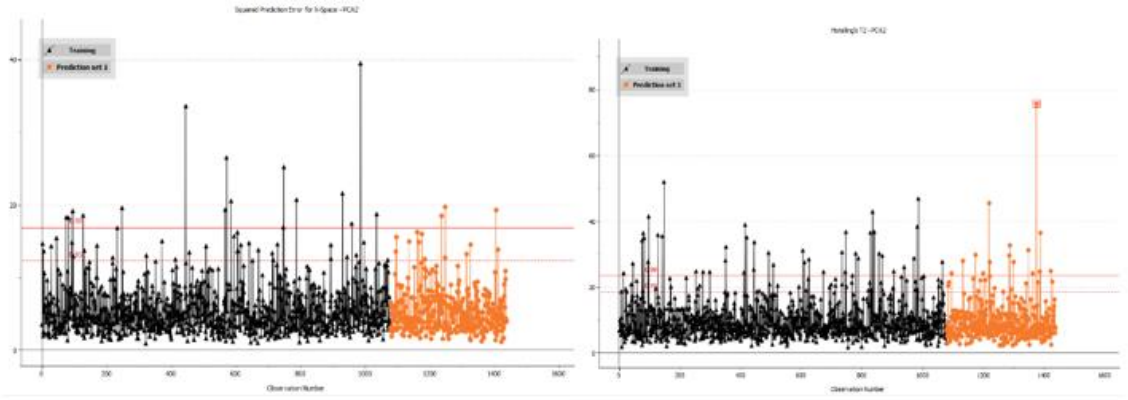


Ilustración 21. Gráficas SPE-X y T² de Hotelling para el PCA previo.

Después, se han evaluado los modelos con las 10 componentes principales y se han validado con los datos de validación, obteniendo los scores con los loadings de las 10 componentes, para acabar obteniendo los siguientes árboles, las importancias de las componentes principales del random forest y las tasas de acierto de los modelos.

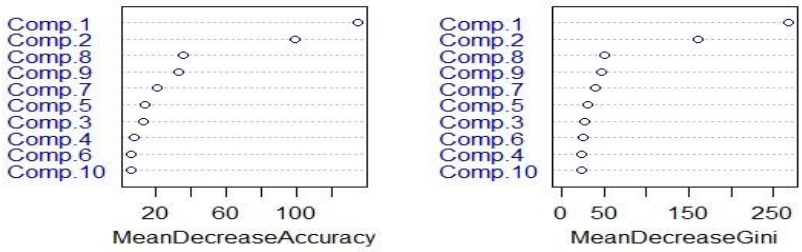


Ilustración 22. Importancias de las variables según el algoritmo random forest.

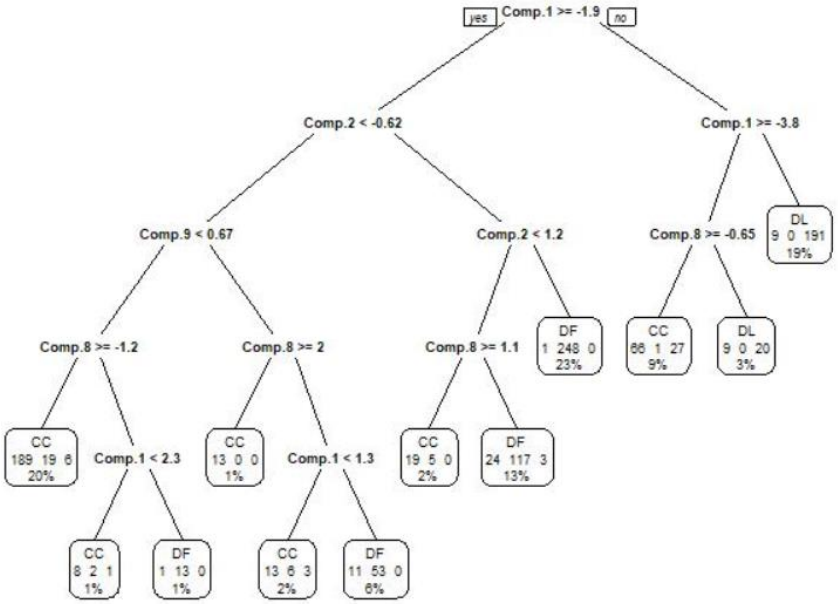


Ilustración 23. Árbol de clasificación podado con el parámetro se=1.

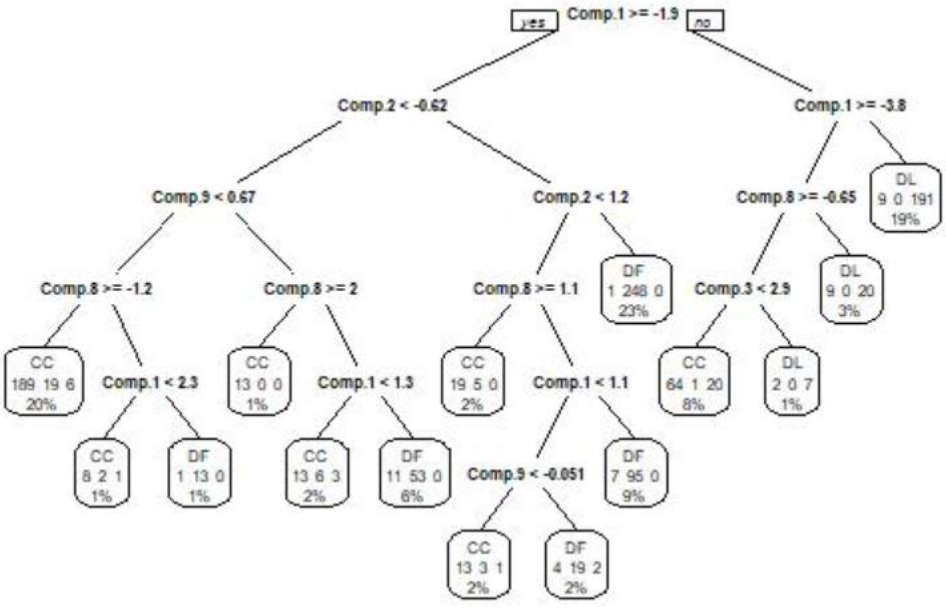


Ilustración 24.Árbol de clasificación podado con el parámetro se=0,5.

Se observa en esos métodos que las componentes más importantes son la primera y la segunda.

Método	Árbol1	Árbol2	KNN	N.Bayes	N.B.Lapl.	SVM1
% acierto	83,06%	84,44%	84,72%	88,61%	88,61%	85,83%

Método	SVM2	RF	NN	Bagging	Boosting
% acierto	89,17%	87,22%	89,44%	84,44%	87,22%

Tabla 1. Tasas de acierto para técnicas que no son propias de la asignatura y que trabajan con los scores de un PCA previo.

En el caso del Simca, se ha realizado un modelo PCA para cada una de las variables a clasificar (defensas, centrocampistas y delanteros). En los modelos, se han obtenido 11, 10 y 9 componentes principales respectivamente y viendo los gráficos SPE-X y T^2 de Hotelling se ha decidido no eliminar ninguna observación para el modelo de entrenamiento. Para clasificar los datos de validación, se han pasado por el modelo los datos de validación y se han graficado éstos en los gráficos SPE-X y T^2 de Hotelling, las observaciones que han quedado dentro de los límites de control al 99% para ambos gráficos se han considerado como pertenecientes a la clase con la que se ha entrenado el modelo y en caso contrario, se ha considerado que no pertenecen a la clase. A continuación, se muestran los datos en los gráficos citados anteriormente.

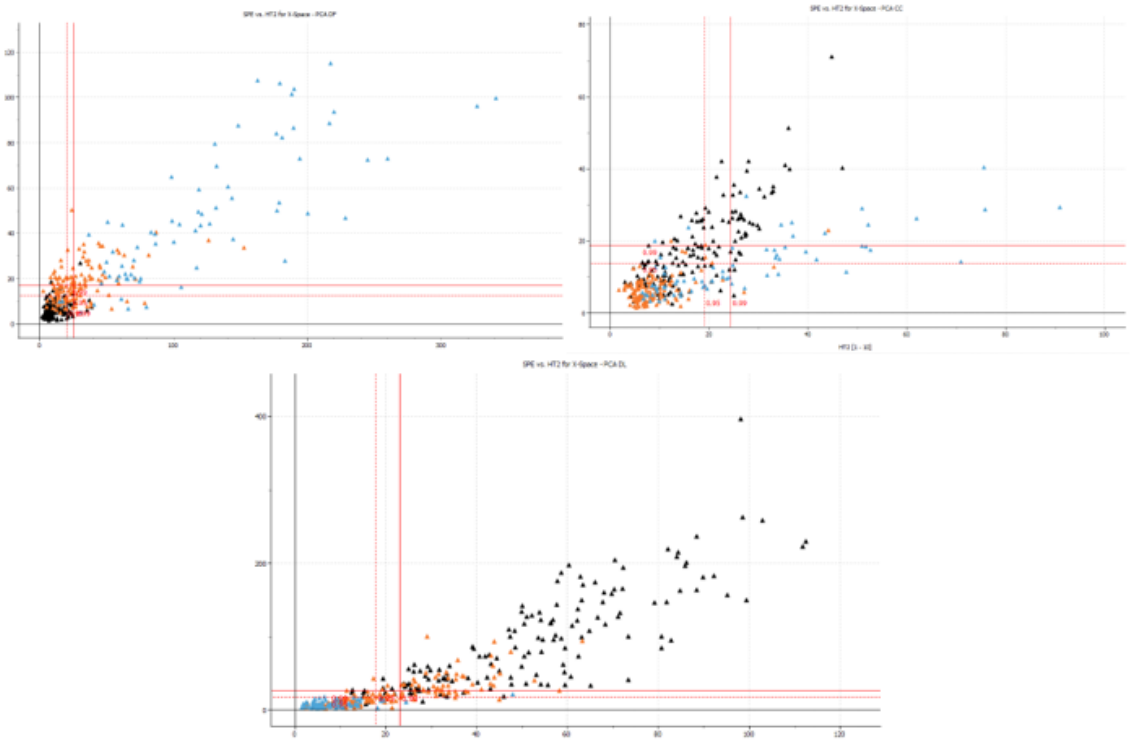


Ilustración 25. Gráfico SPE-X vs T^2 de Hotelling para los datos de validación para cada uno de los SIMCA (DF en la esquina superior izquierda, CC en la esquina superior derecha y DL en la parte inferior).

Como se puede apreciar, hay datos que están bien clasificados y otros que no. En el caso del modelo para los defensas, existen 145 bien clasificados y 6 mal clasificados, para las clases restantes hay 55 mal clasificados y 154 bien clasificados. En el caso de los centrocampistas, hay 135 centrocampistas bien clasificados de 140 y hay 97 mal clasificados de las otras clases de 220. Para finalizar, en el caso de los delanteros, hay 67 bien clasificados de las 69 observaciones y 94 mal de las 291 restantes. Con lo cual, los modelos tienen una sensibilidad de 72,5%, 58% y 49% y una especificidad de 96,25%, 96,09% y 98,99%, teniendo una tasa de acierto total de un 83%, 71,6% y 73% respectivamente.

En el modelo PLS-DA, se ha realizado con 25 variables quitando las que tenían un VIP menor a 0,8 y después eliminando aquellas que tuvieran coeficientes no significativos para los tres modelos. Después, se han eliminado 3 observaciones, un centrocampista (por ser un dato atípico) y dos delanteros (por ser datos extremos), y en el modelo final se han obtenido 11 componentes PLS, ya que son el número de componentes que maximiza el estadístico Q^2. A continuación, se ha evaluado el conjunto de validación con los gráficos de residuos y T^2:

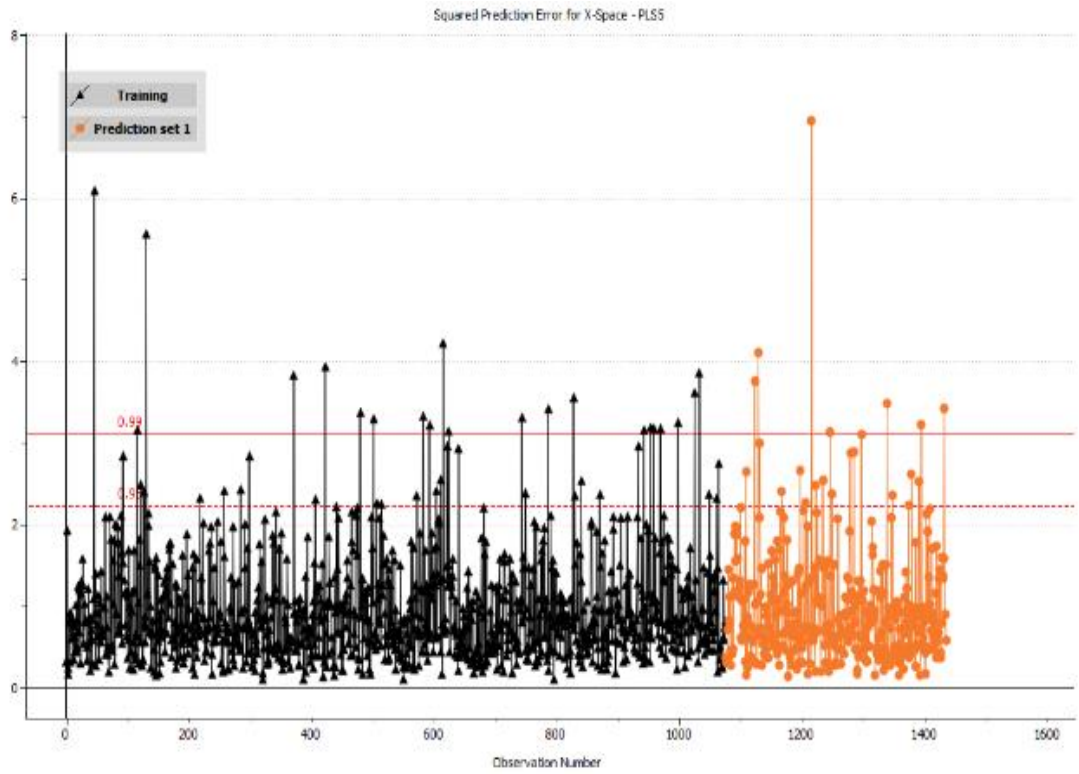


Ilustración 26. Gráfico SPE-X para el PLS-DA con datos de entrenamiento y validación.

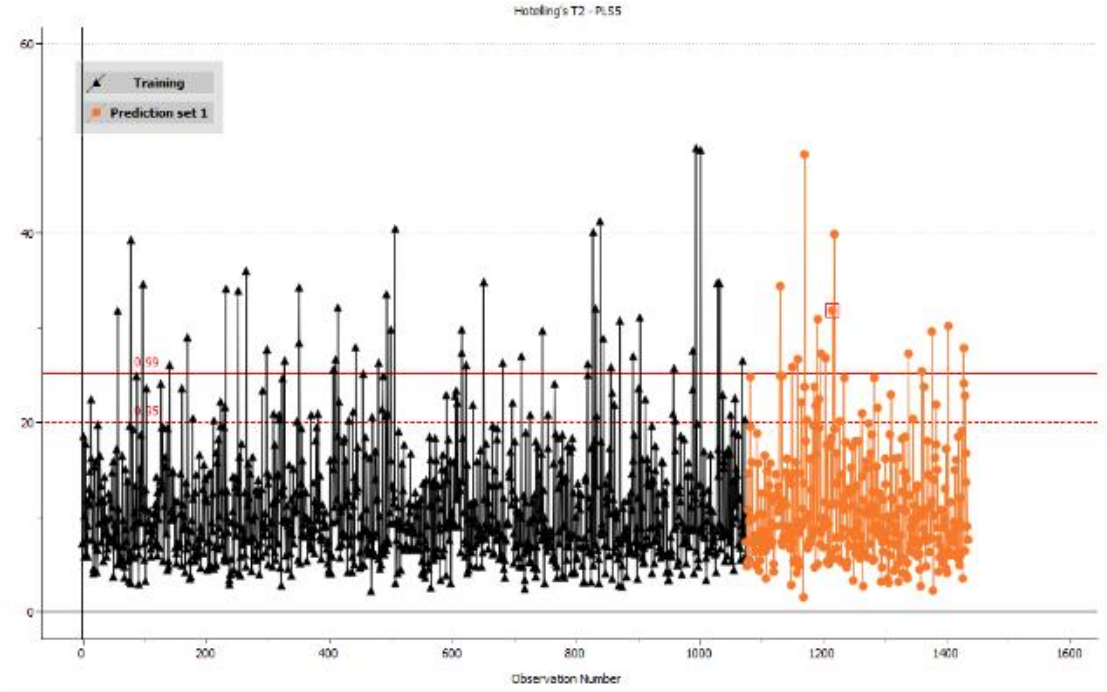


Ilustración 27. Gráfico T^2 de Hotelling para el PLS-DA con datos de entrenamiento y validación.

Como se puede ver, hay algunas observaciones del conjunto de validación por encima de los límites del 99%. Sin embargo, al no ser éstas muy elevadas, se decide dejarlas para el análisis. Una vez realizado esto, se precede a realizar el análisis de lo predicho en contraposición a lo observado y se llega a la siguiente matriz de confusión con los siguientes gráficos:

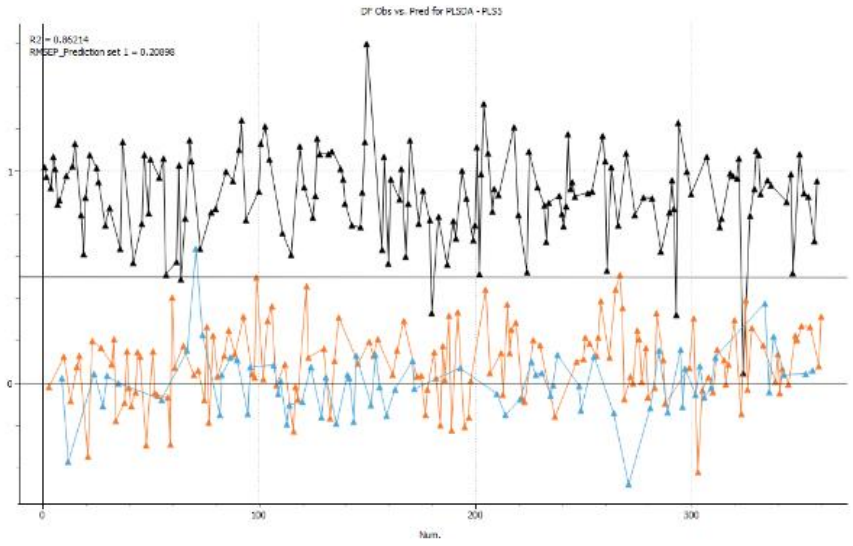


Ilustración 28. Observado vs Predicho para los defensas en el PLS-DA.

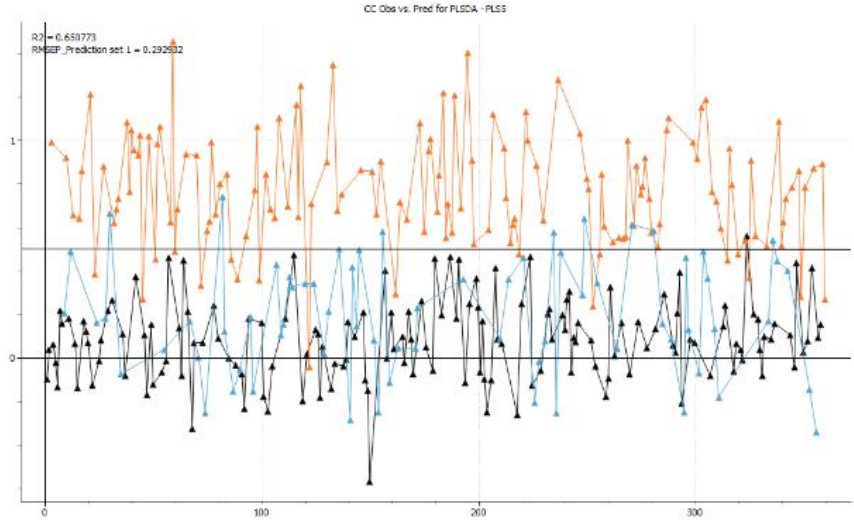


Ilustración 29. Observado vs Predicho para los centrocampistas en el PLS-DA.

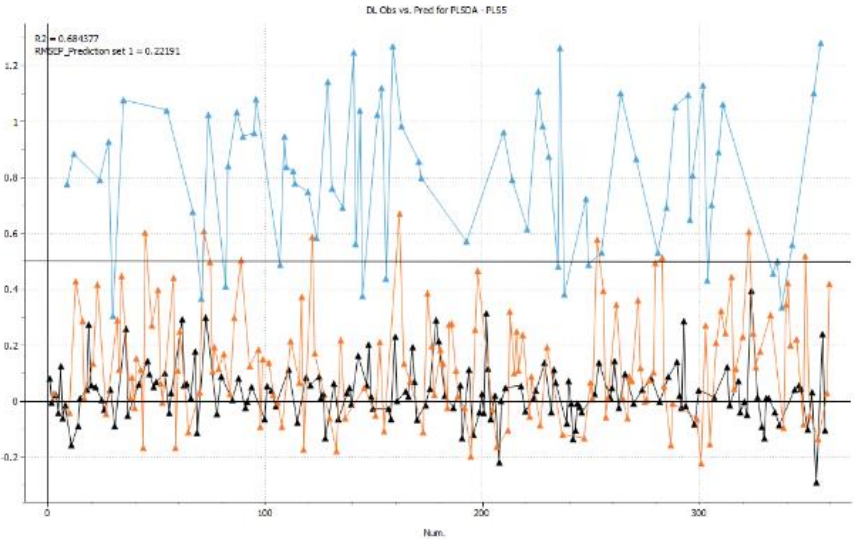


Ilustración 30. Observado vs Predicho para los delanteros en el PLS-DA.

	DF	CC	DL	No clasificados
DF	147	1	0	3
CC	0	120	9	11
DL	1	8	55	5

Tabla 2. Matriz de confusión para el SIMCA.

Calculando la tasa de acierto para el Simca, se observa que ésta es del 89,44%.

4. Conclusión

En conclusión, se ha conseguido realizar un análisis exploratorio de los datos mediante el PCA viendo cuáles son las correlaciones entre las variables y obteniendo las componentes principales utilizando el criterio anteriormente citado. También, se ha visto por la técnica de PLS discriminante cuáles son las variables más importantes a la hora de predecir las clases. A continuación, se ha comparado esta técnica con otras vistas en diferentes asignaturas en variables latentes, esto ha sido utilizando un PCA previo a los análisis para obtener los scores, mediante el método hold out. Para finalizar, se han obtenido las tasas de acierto de todos los métodos y se ha hallado que los métodos con una mejor tasa de acierto son las Redes Neuronales y el PLS-DA. Como para nuestro propósito nos interesa la discriminación tanto como la clasificación, llegamos a la conclusión de que el PLS-DA es el mejor método.

5. Bibliografía

- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Jacob Kaplan (2020). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. R package version 1.6.3. https://CRAN.R-project.org/package=fastDummies
- Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart
- Torgo, L. (2016). Data Mining with R, learning with case studies, 2nd editionChapman and Hall/CRC. URL: http://ltorgo.github.io/DMwR2
- Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
- David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2020). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-4. https://CRAN.R-project.org/package=e1071

- Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. URL: http://www.jstatsoft.org/v11/i09/
- A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22. Andrea Peters and Torsten Hothorn (2019). ipred: Improved Predictors. R package version 0.9-9. https://CRAN.R-project.org/package=ipred
- Stefan Fritsch, Frauke Guenther and Marvin N. Wright (2019). neuralnet: Training of Neural Networks. R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet
- Alfaro, E., Gamez, M. Garcia, N.(2013). adabag: An R Package for Classification with Boosting and Bagging. Journal of Statistical Software, 54(2), 1-35. URL http://www.jstatsoft.org/v54/i02/.
- Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01
- Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.
- [Estadísticas de la temporada 2019-2020 para las 5 grandes ligas europeas]. (s.f.). FBREF.
  URL: https://fbref.com/es/comps/Big5/2019-2020/stats/jugadores/Estadisticas-2019-2020-Las-5-grandes-ligas-europeas
  URL: https://fbref.com/es/comps/Big5/2019-2020/shooting/jugadores/Estadisticas-2019-2020-Las-5-grandes-ligas-europeas
  URL: https://fbref.com/es/comps/Big5/2019-2020/passing/jugadores/Estadisticas-2019-2020-Las-5-grandes-ligas-europeas

  URL: https://fbref.com/es/comps/Big5/2019-2020/defense/jugadores/Estadisticas-2019-2020-Las-5-grandes-ligas-europeas

  URL: https://fbref.com/es/comps/Big5/2019-2020/misc/jugadores/Estadisticas-2019-2020-Las-5-grandes-ligas-europeas

6. Anexo

Los ficheros utilizados son:

- Base de datos entera y con los valores faltantes imputados -> “Dataset_MOD_imputados.xlsx”
- Base de datos con los datos de entrenamiento y con los valores faltantes imputados -> “Dataset_MOD_entrenamiento.xlsx”
- Base de datos con los datos de validación y con los valores faltantes imputados -> “Dataset_MOD_validacion.xlsx”
- Base de datos con los datos de entrenamiento en una hoja de Excel y los datos de validación en otra; y con los valores faltantes imputados -> “Datos entrenamiento y validacion.xlsx”
- Script en R que separa los datos en una muestra de entrenamiento y otra de validación de manera aleatoria -> “Partición en conjuntos de validación.R”
- Script en R que ejecuta diferentes modelos predictivos, incluye PCA previo -> “Modelos predicción.R”
- Proyecto de R -> “Proyecto final MOD.Rproj”
- Archivo de ASPEN PRO MV con el PCA -> “PCAdefinitivo.pmvx”
- Archivo de ASPEN PRO MV con el PLS-DA -> “PLS_DA_definitivo.pmvx”
- Archivo de ASPEN PRO MV con el PCA con los datos de entrenamiento y con el set de validacion -> “Proyecto MOD PCA valid.pmvx”
- Archivo de ASPEN PRO MV con el PLS-DA con los datos de entrenamiento y se valida con el set de validacion -> “Proyecto MOD PLS valid.pmvx”
- Archivo de ASPEN PRO MV con el Simca para cada posición -> “Proyecto MOD Simca.pmvx”