Universitat Oberta de Catalunya uoc.edu

Práctica 2: Limpieza y análisis de datos

Autores: Diego Álvarez Padrón y Kevin Mateo García

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas. Para hacer esta práctica tendréis que trabajar en grupos de 2 personas. Tendréis que entregar un solo archivo con el enlace al repositorio Git donde se encuentren las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Aunque no se trata del mismo enunciado, los siguientes ejemplos de ediciones anteriores os pueden servir como guía:

- Ejemplo: https://github.com/Bengis/nba-gap-cleaning
- Ejemplo complejo (archivo adjunto).

Importante: si se elige un nuevo dataset es interesante que contenga una amplia variedad de datos numéricos y categóricos para poder realizar un análisis más rico.

Competencias
En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.

- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos
Los objetivos concretos de esta práctica son:
- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.

- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.

- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.

● Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.

● Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.

● Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.

● Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Descripción de la Práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com). Algunos ejemplos de dataset con los que podéis trabajar son:

● Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)
● Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic) El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Nos disponemos a analizar un dataset formado por multitud de estadísticas de los jugadores de fútbol profesional, concretamente los integrantes de FIFA del año 2017. Dicho dataset recoge datos de más de 17.000 jugadores entre los que se encuentran atributos relacionados con la manera de jugar al fútbol, por ejemplo: agresividad, velocidad,capacidad de regate, pierna preferida… Así como datos anatómicos de los jugadores, fechas de inicio en clubs y un puntaje en determinados parámetros del juego como puede ser capacidad de reacción, dribbling, marcaje…

Con todos estos datos, nos ha parecido interesante analizar y sacar conclusiones sobre qué tipo de posición es la más importante dentro de los 11 jugadores que forman el equipo en un partido de fútbol, relaciones entre variables como peso-edad-rendimiento-altura y que conclusiones podemos obtener a la hora de que parámetros influyen más para que un jugador sea convocado con la selección nacional de fútbol de un país.

2. Integración y selección de los datos de interés a analizar.

El dataset es un dataset presente en kaggle, de licencia pública. Concretamente este: https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global
Como añadido hemos incluido una nueva variable categórica al dataset para hacer el análisis más rico, concretamente una variable que de una valoración general a cada jugador a modo de puntuación. Esta variable se llama “clasificación” y tiene los siguientes valores:
"Excelente", "Muy bueno", “Bueno”, “Regular”, “Malo”, “Muy malo”.
Los valores descritos se asignan basándose en el valor que tenga otra variable ya presente, “Rating” que designa una puntuación general a cada jugador siendo 0 el mínimo y 100 el máximo. Se asigna basándose en la puntuación numérica general que tiene cada jugador, Dada por la variable “Rating”. Para esclarecer mejor la regla seguida, podemos verlo en el siguiente código en R:

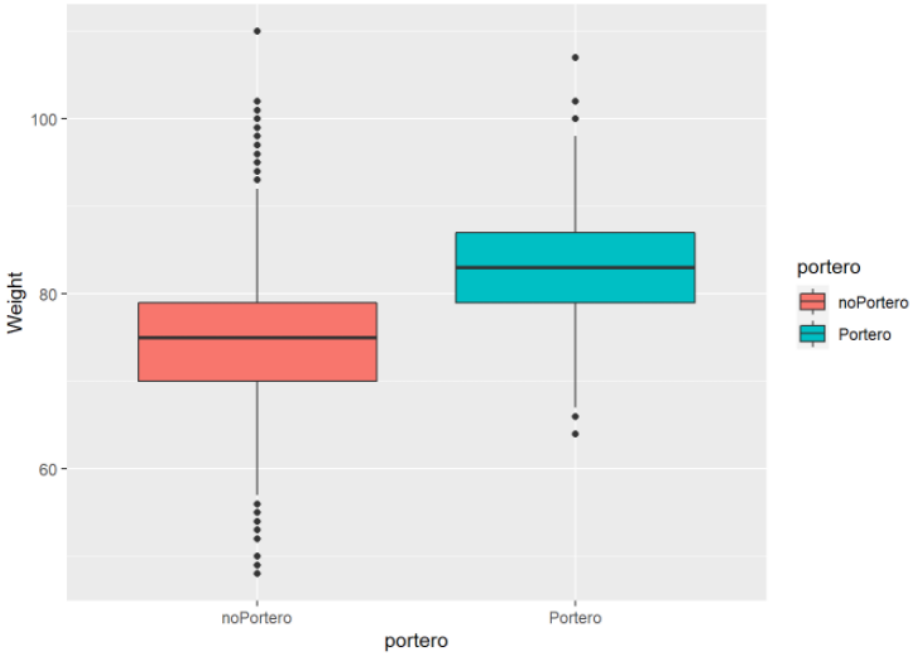
```
fifa$clasificacion[fifa$Rating <= 99 & fifa$Rating >= 90] <- "Excelente"
fifa$clasificacion[fifa$Rating <= 89 & fifa$Rating >= 80] <- "Muy bueno"
fifa$clasificacion[fifa$Rating <= 79 & fifa$Rating >= 70] <- "Bueno"
fifa$clasificacion[fifa$Rating <= 69 & fifa$Rating >= 50] <- "Regular"
fifa$clasificacion[fifa$Rating <= 49 & fifa$Rating >= 40] <- "Malo"
fifa$clasificacion[fifa$Rating <= 39 & fifa$Rating >= 0] <- "Muy malo"
```

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?
Las únicas variables que tienen valores ausentes además de 'National_Position' y 'National_Kit' (los cuales no eliminaremos ya que no son verdaderos missings, sino que simplemente indican que el jugador no ha jugado nunca con el equipo nacional), son las variables 'Club_Kit' y 'Contract_Expiry' las cuáles solo incluye 1 NA cada una y por tanto la pérdida no será un problema mayor, simplemente los dejaremos así.

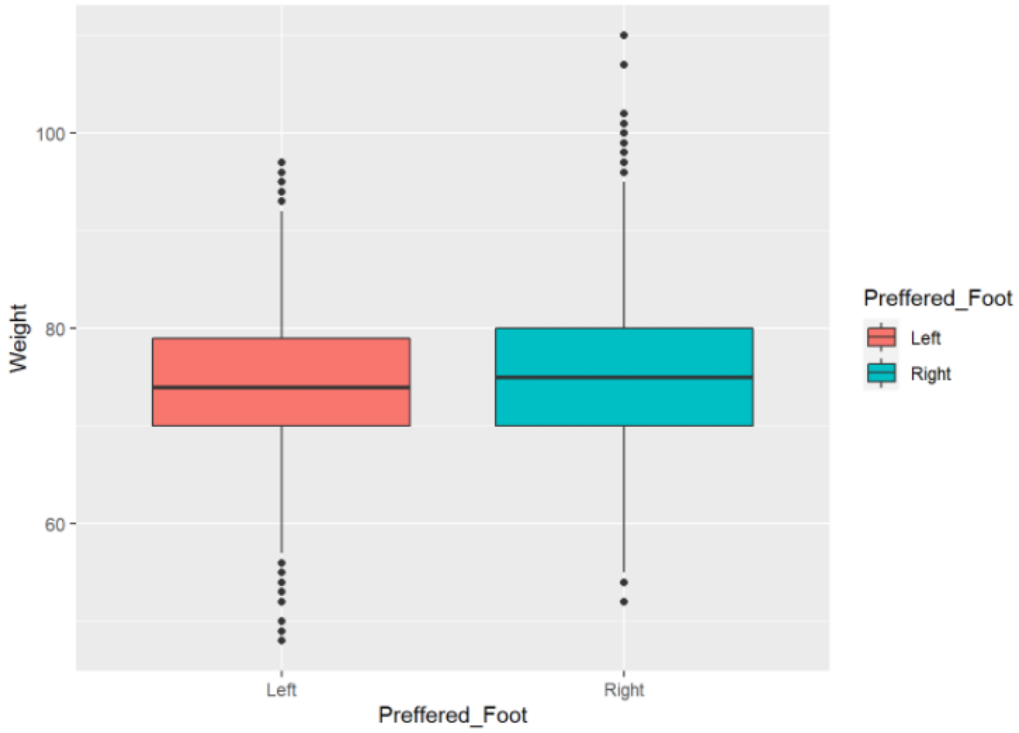
3.2. Identificación y tratamiento de valores extremos.
Hemos tratado (como se podrá ver en el código) de manera visual, valores extremos en relación al peso de los jugadores comparado con otras variables mediante distintos boxplots.

Hemos identificado valores extremos en la relación de variables peso-posición = portero. Detectamos outliers en pesos cercanos a los 100 kilos o superior y en pesos cercanos a los 60 kilos o inferiores.Vemos a raíz de esto que los jugadores que son porteros, tienen un peso más alto que el resto de jugadores:

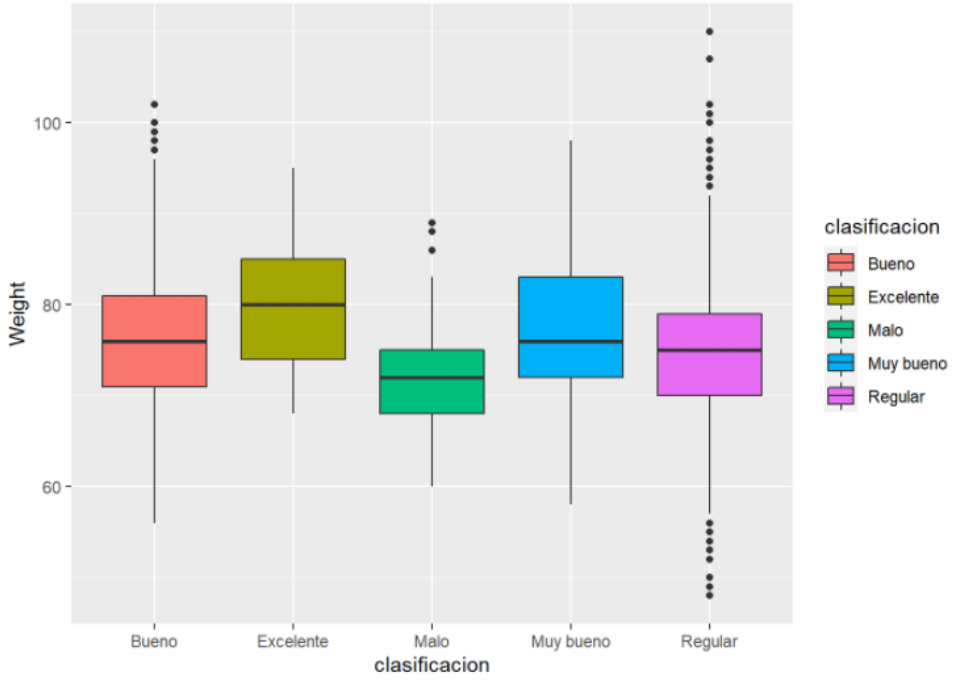


Hemos detectado también valores extremos a la hora de comparar el peso con el pie/pierna preferido para jugar al fútbol, detectando que el pie con el que prefieren lanzar los jugadores

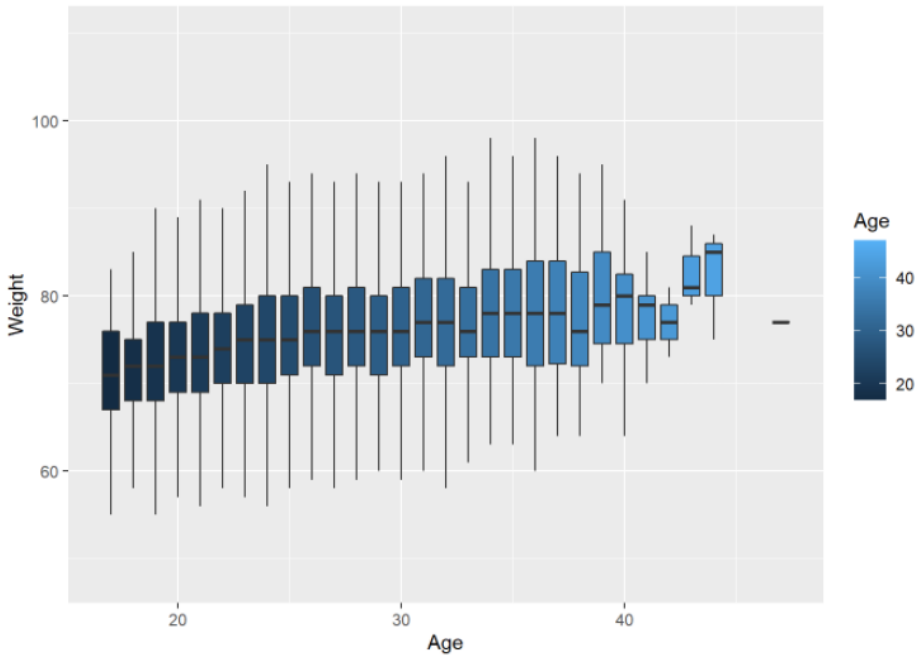
no revela una gran diferencia en cuanto al peso de éstos. Obtenemos valores extremos en la siguiente gráfica:



También nos ha parecido coherente analizar la relación peso-clasificación. Aquí hemos obtenido valores extremos sobre todo en los jugadores que tienen clasificación “regular” y además los jugadores que tienen una mejor calificación (“Excelente” y “Muy bueno”) tienen un peso más elevado que el resto, siendo los calificados como “Malo” los más ligeros. Podemos verlo aqui:



Por último en cuanto a la edad, vemos que según van envejeciendo los jugadores aumentan de peso:



4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

En primer lugar, en nuestro proyecto vemos un resumen de las variables contenidas en él, así como el número de observaciones y las distintas variables que tenemos.

Buscaremos también el número de clubs distintos que hay en estudio y también el número de distintas nacionalidades.

Realizamos análisis descriptivo y tras mostrar un summary del dataset Vemos entonces que tenemos un total de 17588 observaciones, con 54 variables. Además, vemos que tenemos en estudio un total de 634 clubes distintos y 160 nacionalidades diferentes.

Podemos observar aquí las 54 variables:

'data.frame': 17588 obs. of 54 variables:

$ Name : chr "Cristiano Ronaldo" "Lionel Messi" "Neymar" "Luis SuÃ¡rez" ...

$ Nationality : Factor w/ 160 levels "Afghanistan",..: 122 6 20 155 59 139 121 158 143 14 ...

$ National_Position : Factor w/ 28 levels "","CAM","CB",..: 14 25 15 14 6 6 14 24 1 6 ...

$ National_Kit : num 7 10 10 9 1 1 9 11 NA 1 ...

$ Club : Factor w/ 634 levels "1. FC Heidenheim",..: 461 207 207 207 209 364 209 461 364 149 ...

$ Club_Position : Factor w/ 30 levels "","CAM","CB",..: 16 27 16 29 7 7 29 27 29 7 ...

$ Club_Kit : num 7 10 11 9 1 1 9 11 9 13 ...

$ Club_Joining : Factor w/ 1678 levels "","01/01/1993",..: 848 843 852 927 850 850 853 1247 855 1012 ...

$ Contract_Expiry : num 2021 2018 2021 2021 2021 ...

$ Rating : int 94 93 92 92 92 90 90 90 90 89 ...

$ Height : num 185 170 174 182 193 193 185 183 195 199 ...

$ Weight : num 80 72 68 85 92 82 79 74 95 91 ...

$ Preffered_Foot : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2 2 1 2 1 ...

$ Birth_Date : Factor w/ 6063 levels "01/01/1982","01/01/1983",..: 623 2991 630 412 1490 5212 3952 3362 4669 2265 ...

$ Age : int 32 29 25 30 31 26 28 27 35 24 ...

$ Preffered_Position: Factor w/ 292 levels "CAM","CAM/CDM",..: 172 237 157 266 113 113 266 237 266 113 ...

$ Work_Rate : Factor w/ 9 levels "High / High",..: 2 9 3 3 9 9 3 3 8 9 ...

$ Weak_foot : int 4 4 5 4 4 3 4 3 4 3 ...

$ Skill_Moves : int 5 4 5 4 1 1 3 4 4 1 ...

$ Ball_Control : int 93 95 95 91 48 31 87 88 90 23 ...

$ Dribbling : int 92 97 96 86 30 13 85 89 87 13 ...

$ Marking : int 22 13 21 30 10 13 25 51 15 11 ...

$ Sliding_Tackle : int 23 26 33 38 11 13 19 52 27 16 ...

$ Standing_Tackle : int 31 28 24 45 10 21 42 55 41 18 ...

$ Aggression : int 63 48 56 78 29 38 80 65 84 23 ...

$ Reactions : int 96 95 88 93 85 88 88 87 85 81 ...

$ Attacking_Position: int 94 93 90 92 12 12 89 86 86 13 ...

$ Interceptions : int 29 22 36 41 30 30 39 59 20 15 ...

$ Vision : int 85 90 80 84 70 68 78 79 83 44 ...

$ Composure : int 86 94 80 83 70 60 87 85 91 52 ...

$ Crossing : int 84 77 75 77 15 17 62 87 76 14 ...

$ Short_Pass : int 83 88 81 83 55 31 83 86 84 32 ...

$ Long_Pass : int 77 87 75 64 59 32 65 80 76 31 ...

$ Acceleration : int 91 92 93 88 58 56 79 93 69 46 ...

$ Speed : int 92 87 90 77 61 56 82 95 74 52 ...

$ Stamina : int 92 74 79 89 44 25 79 78 75 38 ...

$ Strength : int 80 59 49 76 83 64 84 80 93 70 ...

$ Balance : int 63 95 82 60 35 43 79 65 41 45 ...

```
##  $ Agility          : int  90 90 96 86 52 57 78 77 86 61 ...
##  $ Jumping          : int  95 68 61 69 78 67 84 85 72 68 ...
##  $ Heading          : int  85 71 62 77 25 21 85 86 80 13 ...
##  $ Shot_Power       : int  92 85 78 87 25 31 86 91 93 36 ...
##  $ Finishing        : int  93 95 89 94 13 13 91 87 90 14 ...
##  $ Long_Shots       : int  90 88 77 86 16 12 82 90 88 17 ...
##  $ Curve            : int  81 89 79 86 14 21 77 86 82 19 ...
##  $ Freekick_Accuracy : int  76 90 84 84 11 19 76 85 82 11 ...
##  $ Penalties        : int  85 74 81 85 47 40 81 76 91 27 ...
##  $ Volleys          : int  88 85 83 88 11 13 86 76 93 12 ...
##  $ GK_Positioning   : int  14 14 15 33 91 86 8 5 9 86 ...
##  $ GK_Diving        : int  7 6 9 27 89 88 15 15 13 84 ...
##  $ GK_Kicking       : int  15 15 15 31 95 87 12 11 10 69 ...
##  $ GK_Handling      : int  11 11 9 25 90 85 6 15 15 91 ...
##  $ GK_Reflexes      : int  11 8 11 37 89 90 10 6 12 89 ...
##  $ clasificacion    : Factor w/ 5 levels "Bueno","Excelente",..: 2 2 2 2 2 2 2 2 2 4 ...
```

Tras esto realizaremos estadística descriptiva y visualización, luego estadística inferencial con contraste de hipótesis, modelos de regresión logística y lineal y por último análisis de la varianza. Todo ello acompañado de visualizaciones visibles en el proyecto (visualizaciones y código).

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Hemos realizado un análisis de la varianza (ANOVA) comparando los grupos de variables rating-edad de los jugadores teniendo grupos de edad.

Tras calcular la media de cada uno de los grupos para a continuación comparar la varianza de estas medias, tras realizar un modelo logístico comparando la hipótesis nula y alternativa, podemos descartar la hipótesis nula a favor de la hipótesis alternativa. (ver modelo y adecuación realizado en código apartados 6.3,6.4).

Además hemos realizado un análisis ANOVA multifactorial comenzando por una visualización simple de las variables a estudiar que son la edad y el rating de los jugadores (ver apartado 7 entero)

A continuación, agrupamos por “AgeInt” y portero para calcular las medias en los distintos casos y representamos para su estudio. Calculamos también el modelo,

Como conclusión podemos observar que los porteros, tengan la edad que tengan, tendrán una mejor valoración que los no porteros. Teniendo esta variable más peso a la hora de asignar una valoración en comparación con la edad del jugador en el modelo de regresión logística creado.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

En los apartados 3,4 y 5 hemos realizado contrastes de hipótesis, modelos de regresión lineal y modelos de regresión logística detallando variables utilizadas asi como sus representaciones gráficas y visualizaciones

5. Representación de los resultados a partir de tablas y gráficas.
Ver proyecto en R, cada apartado tiene representación gráfica

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

A lo largo de este estudio, hemos recogido varias conclusiones que resumimos a continuación:
Los porteros tienen un peso más alto que el resto de jugadores.
El pie con el que prefieren lanzar los jugadores no revela una gran diferencia en cuanto al peso de éstos.
Los jugadores que tienen una mejor calificación (“Excelente” y “Muy bueno”) tienen un peso más elevado que el resto, siendo los calificados como “Malo” los más ligeros.
En cuanto a la edad, vemos que según van envejeciendo los jugadores aumentan de peso.
Tras el estudio pertinente, se puede asumir que la variable Weight sigue una distribución normal.
Al igual que habíamos calculado la diferencia de peso de acuerdo a la posición, tras hacer lo propio con la altura, se concluye que los porteros miden 5 cm más que los jugadores de campo.
En el modelo de regresión lineal generado, vemos que el hecho de que el jugador prefiera el pie derecho, tiene una gran importancia a la hora de asignarle una valoración.

En el primer modelo de regresión logística, vemos que la variable portero (que nos indica si el jugador es portero o no) tiene más peso que el resto de variables a la hora de convocar a un jugador a la selección. Aunque este modelo, tras entrenarlo y hacer pruebas de predicciones, vemos que la matriz de confusión nos muestra que el modelo es muy bueno reconociendo los casos negativos, pero que por el contrario, a muchos jugadores nacionales no los clasifica como tal. La precisión media es de un 54.4% por este motivo.

En cuanto al ANOVA de un factor, descartamos la hipótesis nula que afirmaba que todas las medias de la población (medias de los niveles de los factores) son iguales a favor de la hipótesis alternativa, que establece que al menos una de estas medias de la población es diferente.

De acuerdo al ANOVA multifactorial en el que analizamos los pesos que tienen tanto la edad como la posición en la que juegan los jugadores, podemos concluir que los porteros, tengan la edad que tengan, tendrán una mejor valoración que los no porteros. Teniendo esta variable más peso a la hora de asignar una valoración en comparación con la edad del jugador en el modelo de regresión logística creado en este caso.

Contribuciones	Firma
Investigación previa	**DA,KM**
Redacción de las respuestas	**DA,KM**
Desarrollo código	**DA,KM**

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Enlace al repositorio: https://github.com/kmateo/Practica2_Limpieza_y_analisis
Enlace al video explicativo: _____

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

● Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
● Megan Squire (2015). Clean Data. Packt Publishing Ltd.
● Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
● Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
● Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.
● Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.
● Tutorial de Github https://guides.github.com/activities/hello-world.

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

● Los apartados 1, 2 y 6 valen 0,5 puntos.
● Los apartados 3, 5 y 7 valen 2 puntos.
● El apartado 4 vale 2,5 puntos. Se valorará la idoneidad de las respuestas, que deberán ser claras y completas. Las diferentes etapas deberán justificarse y acompañarse del código correspondiente. También se valorará la síntesis y claridad, a través del uso de comentarios, del código resultante, así como la calidad de los datos finales analizados.