Estudio para la mejor elección de tipo de pista en jugadores de la ATP realizado en pyspark

Carlos María Canut Domínguez

Big Data - UPV

**Resumen**

En el presente estudio se va a utilizar el conjunto de datos de los partidos realizados en la ATP desde el año 2010 hasta el 2021. Con esto, se realizará una serie de modelos de clasificación con los que predecir en base a las características de un jugador vencedor y de otro perdedor, además del ranking en la ATP de ambos y el tiempo de duración de partida en minutos, cuál sería el tipo de pista más idónea para que el resultado planteado ocurriera.

**Indice**

1.-Descripción del Proyecto

[Resumen 2](#_Toc74137178)

[Indice 3](#_Toc74137179)

[Descripción del Proyecto 4](#_Toc74137180)

[Análisis de variables 6](#_Toc74137181)

[Edad: 6](#_Toc74137182)

[Posición en ranking ATP: 7](#_Toc74137183)

[Altura: 8](#_Toc74137184)

[Duración de la partida: 8](#_Toc74137185)

[Mano dominante: 9](#_Toc74137186)

[Descripción de los modelos: 11](#_Toc74137187)

[Modelo de Regresión Logística con One-vs-All: 11](#_Toc74137188)

[Modelo de Arboles de decisión: 11](#_Toc74137189)

[Modelo de Bosques Aleatorios: 12](#_Toc74137190)

[Descripción de la Pipeline: 13](#_Toc74137191)

[Descripción de los hiperparametros: 14](#_Toc74137192)

[Modelo de Regresión Logística con One-vs-All: 14](#_Toc74137193)

[Modelo de Arboles de decisión: 14](#_Toc74137194)

[Modelo de Bosques aleatorios: 14](#_Toc74137195)

[Descripción de las métricas para evaluar los modelos: 15](#_Toc74137196)

[Modelo de Regresión Logística con One-vs-All: 15](#_Toc74137197)

[Modelo de Arboles de decisión: 16](#_Toc74137198)

[Modelo de Bosques aleatorios: 16](#_Toc74137199)

[Análisis de los resultados: 17](#_Toc74137200)

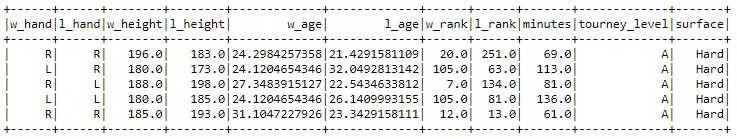
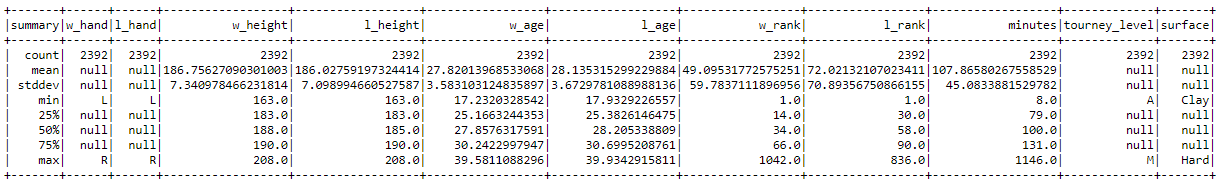
[Referencias 18](#_Toc74137201)

**Descripción del Proyecto**

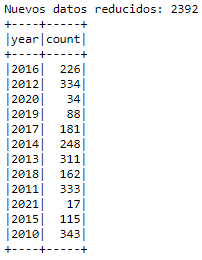
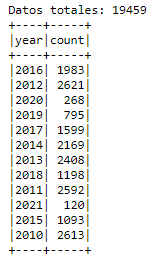
Para este proyecto se ha seleccionado un conjunto de datos con todos los partidos jugados en la ATP con el cual se ha decidido utilizar los siguientes campos:

* Mano dominante del jugador (w\_hand, l\_hand)
* Altura del jugador (w\_height, l\_height)
* Edad del jugador (w\_age, l\_age)
* Posición en la ATP del jugador (w\_rank, l\_rank)
* Duración de la partida en minutos (minutes)
* Rango del torneo (tourney\_level)
* Tipo de pista (Surface)

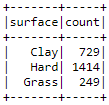
\*\* Los campos con w\_ se refieres a ganadores y con l\_ a perdedores\*\*



Tras decidir que datos se utilizarían, se realizó un muestreo estratificado del conjunto en base al año en el que se jugó el partido:



Finalmente observamos el total de partidas tenemos por cada tipo de pista:

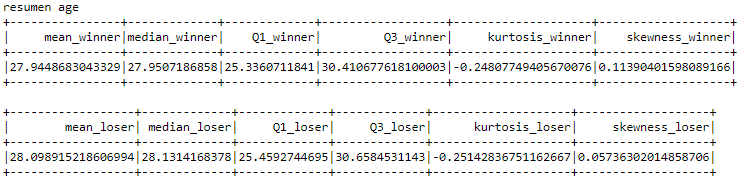


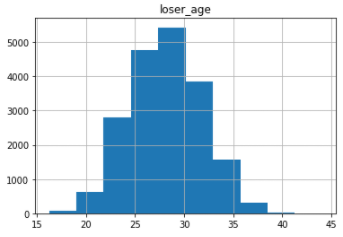
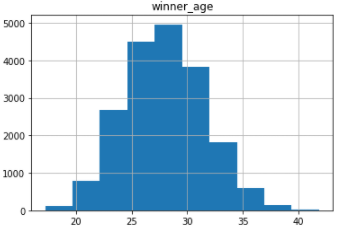
Podemos observar que falta la pista de alfombra, esto es porque al tener pocos registros y para los entrenamientos se decidió eliminarla.

**Análisis de variables**

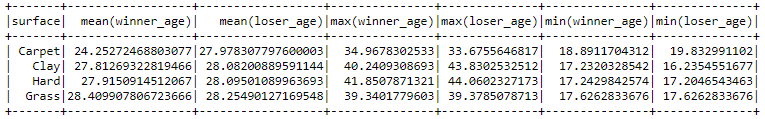
Se ha realizado un análisis variable a variable y también se han observado ciertas variables en relación a otras:

## Edad:



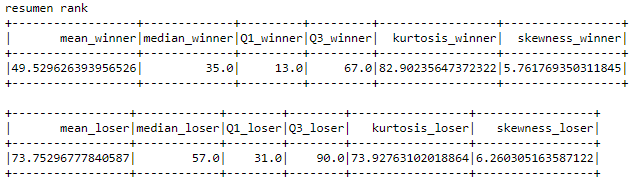


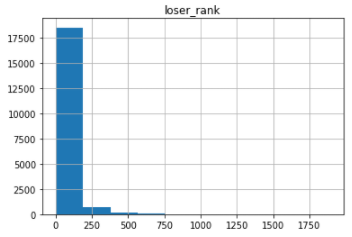
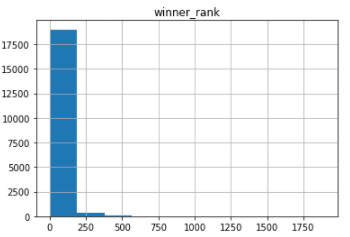
Podemos observar la edad de los jugadores ganadores y de los perdedores se distribuye siguiendo una normal si miramos la asimetría, curtosis y los histogramas, también se puede observar que la edad media de los perdedores suele ser mayor.

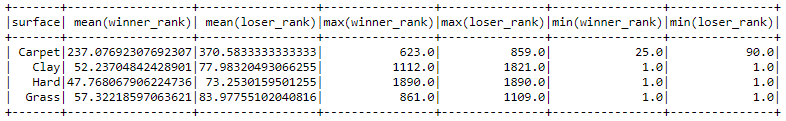


Aquí podemos observar como en las pistas de alfombra es donde la edad tiene un importante factor, ya que los jugadores más jóvenes tienen más posibilidad de ganar observando las medias.

## Posición en ranking ATP:

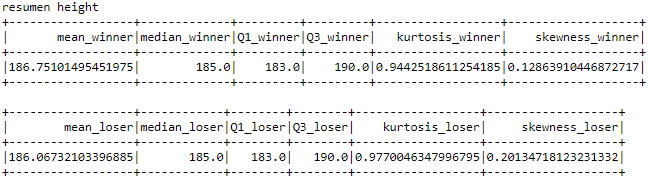


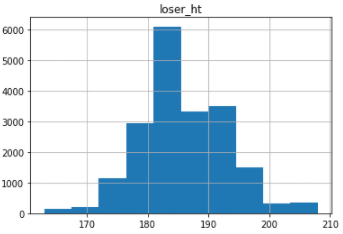
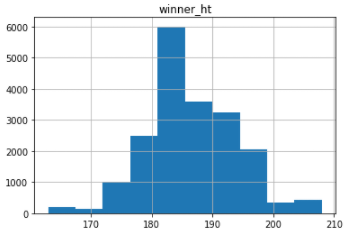


En cuanto a la posición en el ranking de la ATP, tanto jugadores ganadores como perdedores tienen una posición alta en la ATP, siendo la de los ganadores superior como es coherente. 

Tambien se puede observar que en pistas de tierra dura es donde más diferencia de rango podemos observar es en tierra batida.

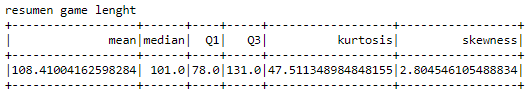
## Altura:

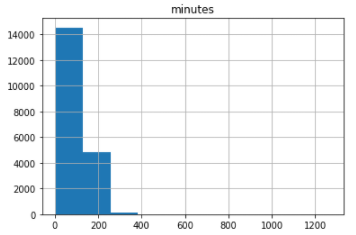




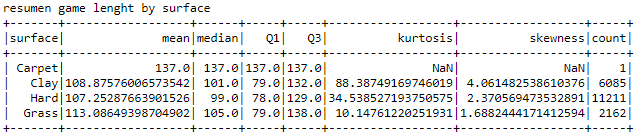
En cuanto a la altura de los jugadores, tanto los ganadores como los perdedores se distribuyen de la misma forma, además se aproxima a una normal.

## Duración de la partida:



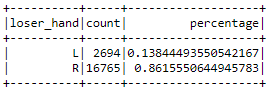
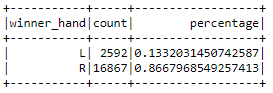


Observando la duración de las partidas, podemos ver que la mayor parte de las partidas dura aproximadamente una hora y 40 minutos.

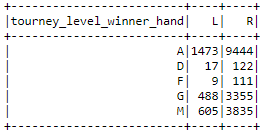


También se puede ver que las partidas más largas son las de césped, pudiendo deberse esto a que la pista de césped es la más rápida (ignorando alfombra ya que solo contamos con 1 registro).

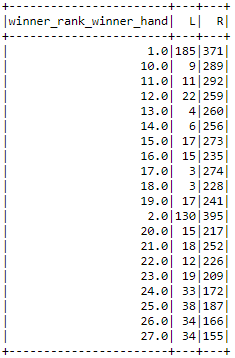
## Mano dominante:



Analizando el tipo de mano que utilizan los jugadores, podemos ver que en ambos casos, los diestros son la mayoría.



También es destacable el hecho de que independientemente de la categoría, se mantiene una diferencia significativa entre zurdos y diestros.



Cabe destacar que en los puestos 1 y 2, contamos con muchas partidas de jugadores zurdos a diferencia de puestos inferiores, este hecho puede que se deba a Rafael Nadal.

**Descripción de los modelos:**

## Modelo de Regresión Logística con One-vs-All:

El clasificador OnevsRest (One-vs-All) es un tipo de clasificador multiclase que funciona con un clasificador binario de base, en nuestro caso una regresión logística.

El funcionamiento se fundamenta en obtener clasificaciones binarias entre todas las clases existentes, así con cada instancia comprueba si la clase es o no la que debería ser en la fase de entrenamiento.

Para hacer las predicciones, realiza una clasificación binaria con cada clase, y la que más confianza tenga es la que elige como resultado.

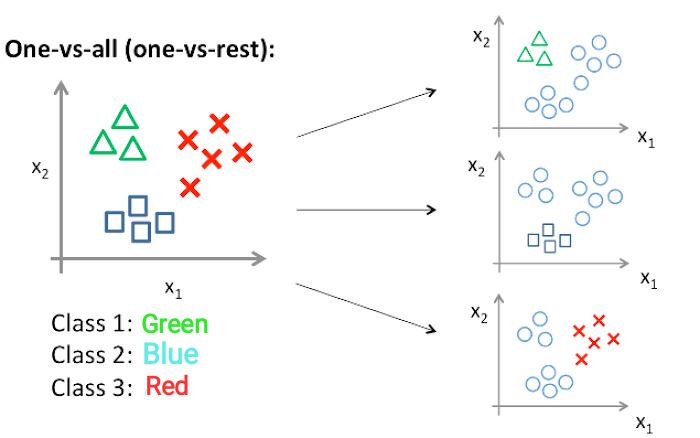


Ilustración One-vs-All. Anton Haugen (1 Marzo 2021).

## Modelo de Arboles de decisión:

Los arboles de decisión funcionan en base a un primer nodo base del que se van generando distintas ramas en base a ciertas características que llevan a otros nodos, de manera que dependiendo de la instancia que estemos prediciendo, llegará hasta un nodo terminal que es el que indica el tipo de clase que se predice.

Diagrama

Descripción generada automáticamente

Ilustración Árbol de decisión. Casper Hansen (15 Septiembre 2021).

## Modelo de Bosques Aleatorios:

Los Bosques aleatorios (Random Forest) hacen uso de los arboles de decisión, de manera que combinan diversos arboles de decisión, esto lo realizan creando cada arbol en base a un vector aleatorio de valores, todos los vectores se generan de manera aleatoria, y posteriormente se promedian entre estos, de manera que se evita el problema de overfitting.

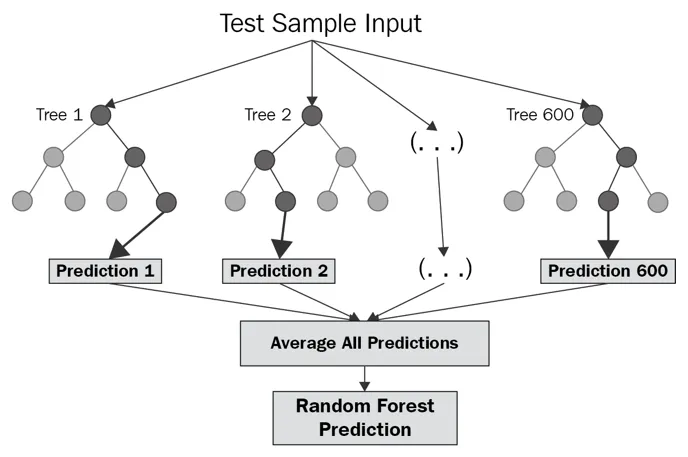


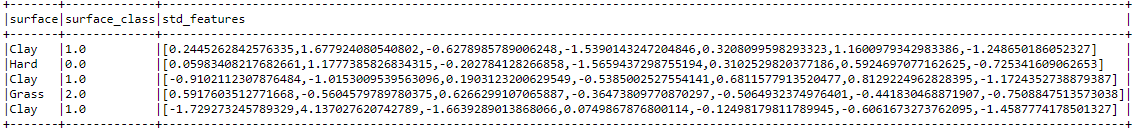
Ilustración Bosque Aleatorio. CFI (2021).

**Descripción de la Pipeline:**

Para la tubería se ha realizado una binarización diferenciando entre jugadores de posiciones más altas (observando la media de los jugadores ganadores) y también con el tiempo de duración medio de los partidos, posteriormente se ha juntado con todas las demás variables predictoras, se ha realizado un PCA con los valores más significativos y se han normalizado todos los valores para el análisis.

Diagrama

Descripción generada automáticamente



Este es el conjunto resultante después de aplicar la tubería.

**Descripción de los hiperparametros:**

## Modelo de Regresión Logística con One-vs-All:

Los hiperparametros a editar de este modelo son:

* regParam 
* elasticNetParam 
* maxIter 

## Modelo de Arboles de decisión:

Los hiperparametros a editar de este modelo son:

* maxDepth 
* maxBins 

## Modelo de Bosques aleatorios:

Los hiperparametros a editar de este modelo son:

* maxDepth: 
* maxBins 
* numTrees 

**Descripción de las métricas para evaluar los modelos:**

Para la evaluación de los modelos, se ha calculado el porcentaje de acierto del modelo, además de mostrar ciertas instancias de la predicción y como reacciona el modelo a estas.

## Modelo de Regresión Logística con One-vs-All:





## Modelo de Arboles de decisión:



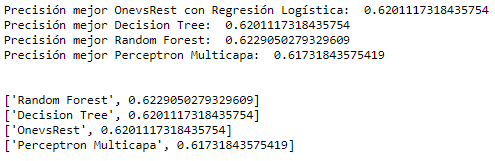


## Modelo de Bosques aleatorios:





**Análisis de los resultados:**



Podemos concluir que aunque el modelo con más precisión sea el de Bosque aleatorio, al ser tan ligeras las diferencias entre los modelos, no se trata de una mejora sustancial.

**Referencias**

JeffSackmann (31 Mayo 2021). <https://github.com/JeffSackmann/tennis_atp>

Spark 2.3.0 (30 Mayo 2014). <https://spark.apache.org/docs/2.3.0/ml-classification-regression.html>

Spark 2.3.0 (30 Mayo 2014). <https://spark.apache.org/docs/2.3.0/ml-pipeline.html>

Spark 2.3.0 (30 Mayo 2014). <https://spark.apache.org/docs/2.3.0/ml-features.html>

Spark 2.3.0 (30 Mayo 2014). <https://spark.apache.org/docs/2.3.0/ml-tuning.html>

Wikipedia (14 Mayo 2021). <https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica>

Wikipedia (15 Febrero 2021). <https://es.wikipedia.org/wiki/%C3%81rbol_de_decisi%C3%B3n>

Wikipedia (5 Octubre 2020). <https://es.wikipedia.org/wiki/Random_forest>

Arya Mohapatra (25 Enero 2020). <https://medium.com/analytics-vidhya/logistic-regression-from-scratch-multi-classification-with-onevsall-d5c2acf0c37c>

CFI (2021). <https://corporatefinanceinstitute.com/resources/knowledge/other/random-forest/>

Casper Hansen (15 Septiembre 2021). <https://mlfromscratch.com/decision-tree-classification/>

Anton Haugen (1 Marzo 2021). <https://antonhaugen.medium.com/introducing-mllibs-one-vs-rest-classifier-402eeab22493>