Sistema de Big Data Analítico para el tenis mundial

Federico sartori - 212602

Contenido

[Introducción 2](#_Toc55336134)

[Descripción del Proyecto 3](#_Toc55336135)

[Planteamiento del Problema 3](#_Toc55336136)

[Objetivo General 3](#_Toc55336137)

[Objetivos específicos 3](#_Toc55336138)

[Alcance 3](#_Toc55336139)

[Limitaciones 3](#_Toc55336140)

[Arquitectura de la Solución 3](#_Toc55336141)

[Herramientas Utilizadas 4](#_Toc55336142)

# Introducción

En el presente documento se analizará un conjunto de datos de más de cuatro millones de registros de partidos de tenis de ATP e ITF, tanto singles como dobles. Son datos recopilados desde el 1950 hasta el 2018.

El dataset fue obtenido mediante el siguiente link: https://www.kaggle.com/ehallmar/a-large-tennis-dataset-for-atp-and-itf-betting . El mismo posee 7 archivos, de los cuales me centraré en 3 de ellos:

* All\_matches.csv: posee 4.3 millones de registros sobre partidos de tenis disputados.
* All\_players.csv: posee 314.000 registros sobre jugadores de tenis. Será útil para obtener cada jugador del primer archivo con su información completa.
* All\_tournments.csv: posee 36.500 registros sobre torneos de tenis. Será útil para obtener cada torneo del primer archive con su información completa.

Levantar datos en spark con jupyter y limpralos.

Indice jugadores, partidos, torenaos y luego hacer la relación.

# Descripción del Proyecto

## Planteamiento del Problema

El problema consiste en tratar grandes volúmenes de datos con baja latencia.

Hoy en día, el mundo de las apuestas en tiempo real es cada vez más grande. Cada vez existen más personas que realizan apuestas, y el constante arreglo de partidos en el tenis es un comprobante de este.

Además, el tenis también se caracteriza por ser un deporte emocional, en donde la mentalidad de la persona juega un factor muy importante.

Debido a esto, planteo responder las siguientes preguntas, que son de utilidad a la hora de realizar una apuesta en tiempo real:

* ¿Cuántas veces un jugador ganó un partido, luego de ganar el primer set?
* ¿Cuántas veces un jugador ganó un partido, luego de ganar un tie-break?
* ¿Qué probabilidad posee un jugador de ganar el primer set en los distintos pisos?

Resolver las preguntas.

## Objetivo General

juntos

## Objetivos específicos

* Analizar conjunto de datos de los partidos de tenis
* Construir ingesta de datos a través de Logstash e indexando en Elasticsearch.
* Crear y generar reportes en Kibana a través de índices.

## Alcance

juntos

## Limitaciones

juntos

# Arquitectura de la Solución

Limpieza de datos: Spark + Jupyter Notebook

Ingesta: Logstash

Motor de Búsqueda: Elasticsearch

Visualización: Kibana

# Ingeniería de Atributos

### all\_matches.csv

El dataset all\_matches.csv contiene muchos atributos, de los cuales solo serán relevantes algunos de ellos. Debido a esto, es que tomaremos solo los atributos necesarios y el resto los eliminaremos. Además, luego de un análisis en Jupyter Notebook, tomamos la decisión de eliminar algunos cambos extra que podían resolverse mediante un “join” con otro dataset (player\_name y opponent\_name). También, algunos campos no eran tan relevantes, y existían muchos valores nulos.

|  |  |  |
| --- | --- | --- |
| **Atributo** | **Descripción** | **Acción tomada** |
| start\_date | Fecha de comienzo del partido. | Mantener |
| end\_date | Fecha de finalización del partido. | Eliminar |
| location | País de disputa del partido. | Eliminar |
| court\_surface | Superficie de la cancha. | Mantener |
| prize\_money | Premio al ganador. | Eliminar |
| currency | Moneda del premio al ganador. | Eliminar |
| year | Año. | Mantener |
| player\_id | Id del jugador. | Mantener |
| player\_name | Nombre del jugador. | Eliminar |
| opponent\_id | Id del oponente. | Mantener |
| opponent\_name | Nombre del oponente. | Eliminar |
| tournament | Nombre del torneo. | Mantener |
| round | Ronda del partido. | Eliminar |
| num\_sets | Sets jugados. | Mantener |
| sets\_won | Sets ganados por el jugador. | Mantener |
| games\_won | Games ganados por el jugador. | Mantener |
| games\_against | Games perdidos por el jugador. | Mantener |
| tiebreaks\_won | Tiebreaks ganados por el jugador. | Mantener |
| tiebreaks\_total | Tiebreaks totales del partido. | Mantener |
| serve\_rating | Puntaje de saque del jugador (combinando todos sus saques) | Eliminar |
| aces | Aces del jugador. | Eliminar |
| double\_faults | Doble faltas del jugador. | Eliminar |
| first\_serve\_made | Primeros saques correctos del jugador. | Eliminar |
| first\_serve\_attempted | Primeros saques realizados del jugador. | Eliminar |
| first\_serve\_points\_made | Puntos ganados en el primer saque por el jugador. | Eliminar |
| first\_serve\_points\_attempted | Puntos jugados en el primer saque por el jugador. | Eliminar |
| second\_serve\_points\_made | Puntos ganados en el segundo saque por el jugador. | Eliminar |
| second\_serve\_points\_attempted | Puntos ganados en el segundo saque por el jugador. | Eliminar |
| break\_points\_saved | Break points salvados por el jugador. | Eliminar |
| break\_points\_against | Break points en contra del jugador. | Eliminar |
| service\_games\_won | Puntos ganados por el jugador en su servicio. | Eliminar |
| return\_rating | Porcentaje de devoluciones de saque del jugador. | Eliminar |
| first\_serve\_return\_points\_made | Puntos ganados luego de devolver el primer saque por parte del jugador. | Eliminar |
| first\_serve\_return\_points\_attempted | Puntos jugados luego de devolver el primer saque por parte del jugador. | Eliminar |
| second\_serve\_return\_points\_made | Puntos ganados luego de devolver el segundo saque por parte del jugador. | Eliminar |
| second\_serve\_return\_points\_attempted | Puntos jugados luego de devolver el segundo saque por parte del jugador. | Eliminar |
| break\_points\_made | Break points ganados por el jugador. | Eliminar |
| break\_points\_attempted | Break points intentados por el jugador. | Eliminar |
| return\_games\_played |  | Eliminar |
| service\_points\_won | Puntos ganados en su servicio por el jugador. | Eliminar |
| service\_points\_attempted | Puntos jugados en su servicio por el jugador. | Eliminar |
| return\_points\_won | Puntos ganados por el jugador luego de devolverle el saque. | Eliminar |
| return\_points\_attempted | Puntos jugados luego de devolverle el saque al jugador. | Eliminar |
| total\_points\_won | Puntos ganados por el jugador. | Eliminar |
| total\_points | Total de puntos en el partido. | Eliminar |
| duration | Duración del partido. | Eliminar |
| player\_victory | Si gano o no el jugador. | Mantener |
| retirement | Si se retiro o no el jugador. | Mantener |
| seed | Clasificación del jugador dentro del torneo. | Eliminar |
| won\_first\_set | Si el jugador gano el primer set o no. | Mantener |
| doubles | Si el partido es de dobles o no. | Mantener |
| masters | Número de master del torneo. | Eliminar |
| round\_num | Ronda del partido. | Eliminar |
| nation | País donde se disputa el partido. | Eliminar |

Además, eliminaremos aquellos datos donde existan valores nulos. Este estudio lo realizaremos en la Jupyter Notebook.

### all\_tournaments.csv

El dataset all\_tournaments contiene algunos atributos innecesarios, y por eso los eliminaremos.

|  |  |  |
| --- | --- | --- |
| **Atributo** | **Descripción** | **Acción tomada** |
| year | Año en que se jugó. | Mantener |
| tournament | Nombre del torneo. | Mantener |
| start\_date | Fecha de Comienzo | Mantener |
| end\_date | Fecha de finalización. | Mantener |
| location | País en que se juega. | Mantener |
| court\_surface | Superficie de la cancha. | Mantener |
| prize\_money | Premio al ganador. | Eliminar |
| currency | Moneda del premio al ganador. | Eliminar |
| masters | Número de master del torneo. | Eliminar |

‘

### all\_players.csv

No realizaremos ninguna acción sobre el dataset para mantener integridad de los datos, y además, solamente posee 2 columnas.

# Herramientas Utilizadas