Sistema de Big Data Analítico para el tenis mundial

Federico sartori - 212602

Contenido

[Introducción 2](#_Toc56874920)

[Descripción del Proyecto 3](#_Toc56874921)

[Planteamiento del Problema 3](#_Toc56874922)

[Objetivo General 3](#_Toc56874923)

[Objetivos específicos 3](#_Toc56874924)

[Alcance 3](#_Toc56874925)

[Limitaciones 3](#_Toc56874926)

[Metodología 3](#_Toc56874927)

[Arquitectura de la Solución 3](#_Toc56874928)

[Ingeniería de Atributos 4](#_Toc56874929)

[all\_matches.csv 4](#_Toc56874930)

[all\_tournaments.csv 6](#_Toc56874931)

[all\_players.csv 6](#_Toc56874932)

[Herramientas Utilizadas 7](#_Toc56874933)

[Herramientas y Justificación por Capa 7](#_Toc56874934)

[Ingesta de Datos 7](#_Toc56874935)

[Almacenamiento de Datos 7](#_Toc56874936)

[Procesamiento de Datos 7](#_Toc56874937)

[Visualización de Datos 7](#_Toc56874938)

[Conclusiones 8](#_Toc56874939)

# 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.

# 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?

## Objetivo General

Generar un motor de búsqueda de Big Data. Este motor ayudará a responder las preguntas de la sección anterior, y además le brindará una interfaz de visualización al usuario.

## 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

## Limitaciones

## Metodología

# Arquitectura de la Solución

Limpieza de datos: Spark + Jupyter Notebook

Ingesta: Logstash

Motor de Búsqueda: Elasticsearch

Visualización: Kibana

# Ingeniería de Atributos

La ingeniería de atributos será realizada mediante una Jupyter Notebook con Python y PySpark. Cada notebook terminará generando un archivo ejecutable de Python (.py) que genera el procesamiento de los datos y crea nuevos archivos con los nuevos atributos.

Dentro de la carpeta dataset/raw se encontrarán los archivos sin procesar, y dentro de dataset/processed se encontrarán los archivos ya procesados.

### 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.

En este dataset, tomamos las siguientes decisiones:

* Eliminar columnas innecesarias (descripto en la siguiente tabla).
* Eliminar columnas que contenían datos que podían obtenerse con un join con otro dataset. (player\_name y opponent\_name).
* Eliminar registros que tenían muchas columnas nulas.
* Modificar tipos de columnas para mantener una coherencia
  + Columnas con valores “f” y “t” se transforman a columnas booleanas
  + Columnas con números como string se transforman a columnas numéricas
* Eliminamos caracteres de strings mal formados (“ ’/’ ”)

|  |  |  |
| --- | --- | --- |
| **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 |

### all\_tournaments.csv

El dataset all\_tournaments 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.

En este dataset, tomamos las siguientes decisiones:

* Eliminar columnas innecesarias (descripto en la siguiente tabla).
* No eliminar registros con columnas nulas ya que puede generar inconsistencias con el dataset de partidos.
  + En caso de existir fecha de comienzo nula, generamos una fecha a partir del año de disputa
  + En caso de existir fecha de finalización nula, generamos una fecha igual a la fecha de comienzo
* Eliminamos caracteres de strings mal formados (“ ’/’ ”)
* Modificar tipos de columnas para mantener una coherencia
  + Columnas con fechas como string se transforman a columnas de tipo Date

|  |  |  |
| --- | --- | --- |
| **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, además de que solamente posee 2 columnas.

# Herramientas Utilizadas

## Herramientas y Justificación por Capa

### Ingesta de Datos

### Almacenamiento de Datos

### Procesamiento de Datos

### Visualización de Datos

# Conclusiones