Proyecto Ciencia de datos:

Descubrimiento de insights de peleas de la UFC

Contenido

Introducción

Este proyecto enfocado a las artes marciales mixtas (MMA) de la UFC, busca identificar los principales insights como lo son golpes en la cabeza, cuerpo y piernas con esto se pretende detectar posibles traumatismos. De igual forma se busca proponer un modelo de Machine Learning que identifica si un peleador va a ganar o no su próximo combate. La principal fuente de datos es el histórico de peleas de los luchadores el cual se encuentra de forma publica en la siguiente dirección:

AL hacer una exploración en los datos fuentes de históricos, no existe una forma directa de obtener la información, por lo cual se realizó una investigación sobre cómo obtener los datos.

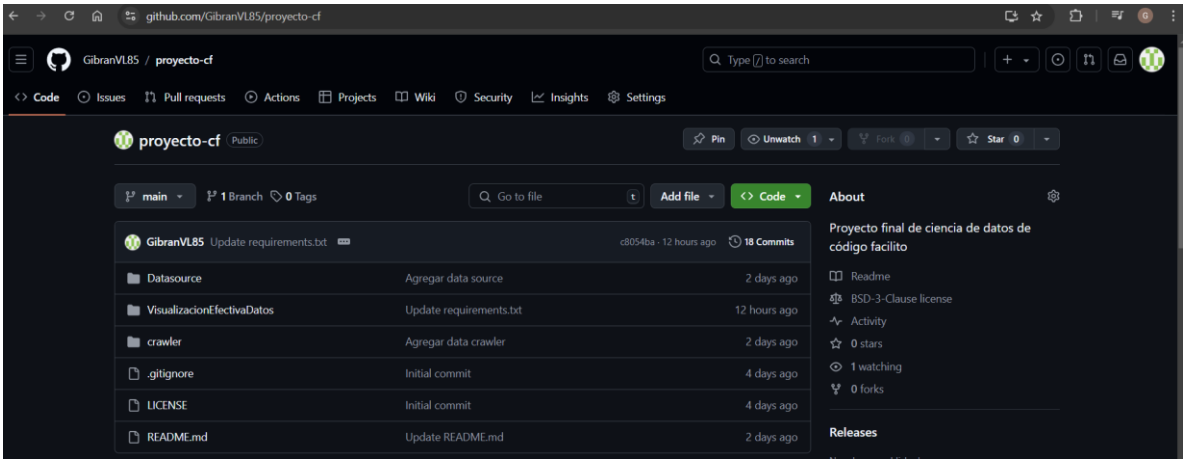
http://www.ufcstats.com/

El proyecto se integra en 3 partes:

- Visualización Efectiva del proyecto
- EDA
- Machine Learning

Cada sección cuenta con su propio archivo de Jupyter Notebooks, los cuales se pueden acceder a través del repositorio GIT del proyecto:

https://github.com/GibranVL85/proyecto-cf



Extracción de datos

Tras una investigación en internet se encontró un scraper que utiliza la librería scrapy de Python.

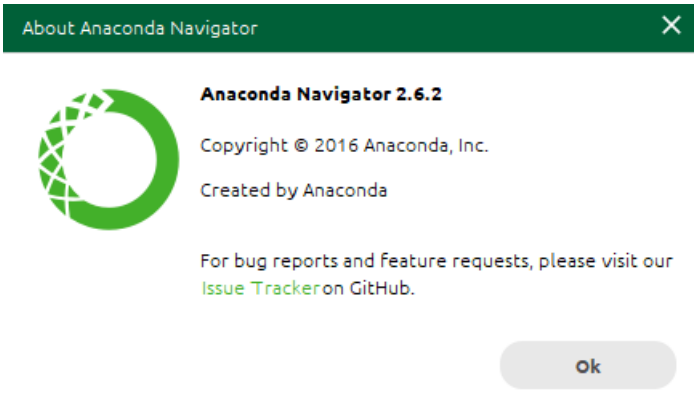
https://github.com/fanghuiz/ufc-stats-crawler

LA idea de utilizar el scraper es obtener la data que esta expuesta en el sitio web de la UFC y disponer de esta para realizar el análisis del proyecto.

https://scrapy.org/

El ambiente se configuro en Windows10 con Anaconda2.6.2 y Python 3.11.7

```
Microsoft Windows [Versión 10.0.22631.4169]
(c) Microsoft Corporation. Todos los derechos reservados.

(base) C:\Users\scoop>python
Python 3.11.7 | packaged by Anaconda, Inc. | (main, Dec 15 2023, 18:05:47) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
```



Para implementar el proyecto se instalo en un ambiente de conda independiente:

```
(base) C:\Users\scoop>conda env list
# conda environments:
#
base                  *  C:\Users\scoop\anaconda3
crawler                  C:\Users\scoop\anaconda3\envs\crawler
estadistica              C:\Users\scoop\anaconda3\envs\estadistica
pandas-env               C:\Users\scoop\anaconda3\envs\pandas-env
pruebas_spyder           C:\Users\scoop\anaconda3\envs\pruebas_spyder
```

Se actualizo la versión de scrapy y date parser, ya que la versión actual tiene conflictos con el proyecto.

Versiones del proyecto

Versiones implementadas

```
requirements.txt
1  dateparser==0.7.2
2  Scrapy==1.7.4
```

Versiones actualizadas

```
datashader                0.16.0          py311haa95532_0
dateparser                1.2.0                    pypi_0    pypi
dav1d                     1.2.1                h2bbff1b_0
debugpy                   1.6.7           py311hd77b12b_0
```

```
scipy                     1.11.4          py311hc1ccb85_0
scrapy                    2.8.0           py311haa95532_0
seaborn                   0.12.2          py311haa95532_0
semver                    2.13.0              pyhd3eb1b0_0
```

Tras actualizar las librerías del proyecto también fue necesario actualizar el proyecto en los siguientes archivos para evitar errores en la ejecución:

spider.py

```
7   def time_to_seconds(time_str):
8       """Convierte una cadena de tiempo 'm:s' a segundos."""
9       minutes, seconds = map(int, time_str.split(':'))
10      return minutes * 60 + seconds
11
```

```
158             n_rev = stats_total[9].css('p ::text').getall()
159             try:
160                 n_rev = [time_to_seconds(i.strip()) for i in n_rev] # type: ignore
161             except ValueError:
162                 # Si ocurre una excepción, probablemente el formato es incorrecto, así que lo dejamos como está
163                 n_rev = [int(i.strip()) for i in n_rev if i.strip().isdigit()]
164
```

```
236 class FightersSpider(scrapy.Spider):
237     name = 'ufcFighters'
238     start_urls = ['http://ufcstats.com/statistics/fighters']
239
240     custom_settings = {
241         'format': 'csv',
242         'URI': 'data/fighter_stats/%(time)s.csv'
243     }
```

```python
class UpcomingFightsSpider(scrapy.Spider):
    name = 'upcoming'
    start_urls = ['http://ufcstats.com/statistics/events/completed']
    time_created = print_time('now')

    custom_settings = {
        'format': 'csv',
        'URI': f'data/upcoming/{time_created}.csv'
    }
```

items.py

```python

import scrapy
from scrapy.loader.processors import Identity, MapCompose, Join
from itemloaders.processors import TakeFirst,Compose

```

```python
    name = scrapy.Field(output_processor=TakeFirst())
    height = scrapy.Field(output_processor=TakeFirst())
    weight = scrapy.Field(output_processor=())
    reach = scrapy.Field(output_processor=TakeFirst())
    stance = scrapy.Field(output_processor=TakeFirst())
    dob = scrapy.Field(output_processor=TakeFirst())
```

Tras las correcciones, se procedió a ejecutar el proyecto:

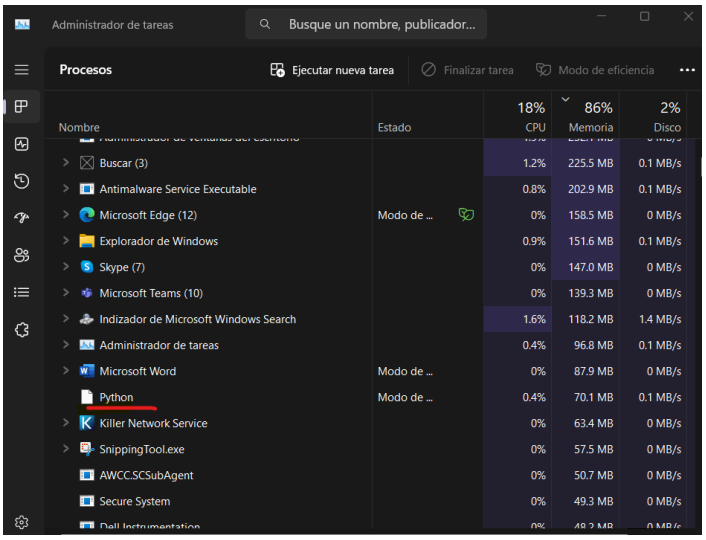
En la ruta del proyecto

```
(base) C:\Users\scoop\anaconda3\envs\crawler>dir
 El volumen de la unidad C es OS
 El número de serie del volumen es: F270-2D7C

 Directorio de C:\Users\scoop\anaconda3\envs\crawler

31/07/2024  11:59 a. m.    <DIR>          .
20/09/2024  06:37 p. m.    <DIR>          ..
20/10/2020  12:01 a. m.             1,872 .gitignore
23/07/2024  07:28 p. m.                 0 .nonadmin
23/07/2024  07:28 p. m.    <DIR>          conda-meta
01/08/2024  07:30 p. m.    <DIR>          data
20/10/2020  12:01 a. m.               100 Dockerfile
23/07/2024  07:28 p. m.    <DIR>          etc
20/10/2020  12:01 a. m.             1,068 LICENSE
01/08/2024  07:30 p. m.    <DIR>          log
20/10/2020  12:01 a. m.               767 Makefile
31/07/2024  11:59 a. m.               140 packages_list.txt
20/10/2020  12:01 a. m.             3,092 README.md
20/10/2020  12:01 a. m.                31 requirements.txt
20/10/2020  12:01 a. m.               259 scrapy.cfg
31/07/2024  11:42 a. m.    <DIR>          ufcStats
               9 archivos          7,329 bytes
               7 dirs  235,113,820,160 bytes libres
```

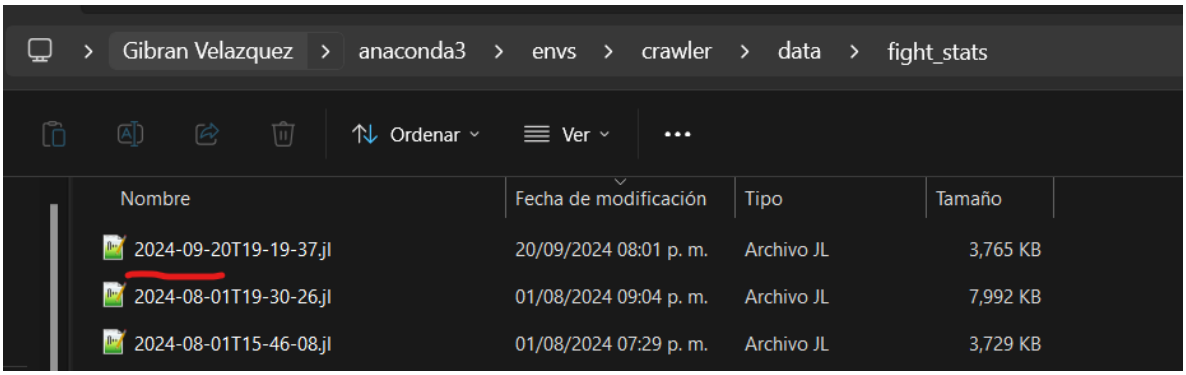
Ejecutar comando scrapy que manda llamar al crawler ufcFights

```
(base) C:\Users\scoop\anaconda3\envs\crawler>scrapy crawl ufcFights
```

Esto empezara a ejecutar un proceso Python



El cual se encarga de extraer la data de la web de UFCStats. Y generara un fichero CSV con la información de las peleas en el directorio con el nombre del crawler.



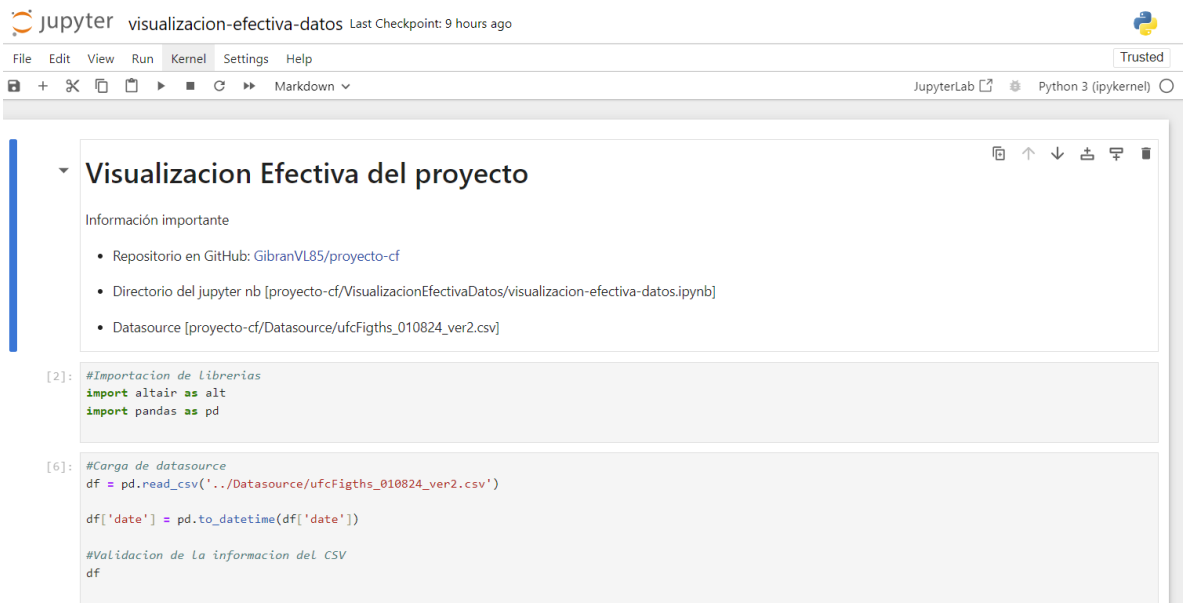
Visualización Efectiva

Para la visualización se consideraron los siguientes puntos:

- Claridad y precisión: Las graficas implementadas fueron seleccionadas con el objetivo de entenderse de manera fácil acorde al dataset analizado.
- Diversidad de visualizaciones: Se seleccionaron diferentes gráficas con diferentes sets para transmitir los hallazgos del dataset.
- Interpretación y presentación: Dentro del fichero de visualización efectiva

Para validar el desarrollo de la visualización efectiva es necesario revisar el archivo visualización-efectiva-datos.ipynb del repositorio de git:

https://github.com/GibranVL85/proyecto-cf/tree/main/VisualizacionEfectivaDatos



Análisis Exploratorio de Datos (EDA)

Para el EDA se realizaron 2 partes:

La primera parte del EDA se realizo como complemento a la visualización efectiva de datos, este complemento, en esta sección se realizó la preparación de datos, limpieza e identificación de insights. El acceso para esta sección es en el notebook de visualización efectiva:

https://github.com/GibranVL85/proyecto-cf/tree/main/VisualizacionEfectivaDatos

La segunda parte, busca realizar un análisis mas exhaustivo de los datos implementando un dataframe más depurado.

https://github.com/GibranVL85/proyecto-cf/tree/main/EDA

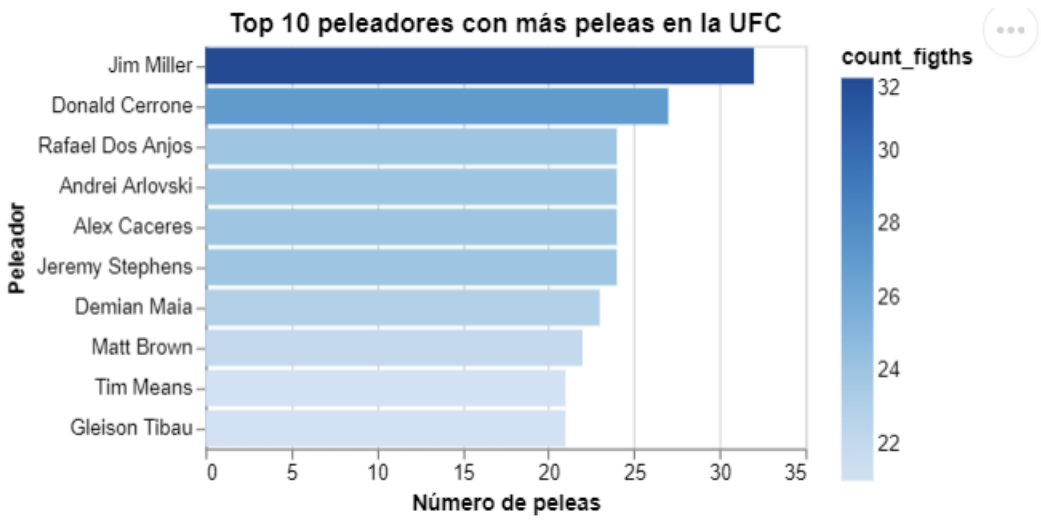
Modelado de Datos

Por cuestiones de tiempo no fue posible completar el modelado.

Hallazgos y conclusiones

Tras realizar la revisión de los datos se puede concluir que un peleador profesional de la UFC recibe muchos golpes a lo largo de su trayectoria como peleador lo cual indica que son sujetos susceptibles de sufrir cualquier tipo de contusiones, daños y/o traumatismos físicos.

Derivado del análisis se puede recomendar a los peleadores considerar análisis periódicos para detectar alguna anomalía física y evitar secuelas que a lo largo del tiempo puedan complicarse.



De igual forma se presentan los insights identificados:

- La mayoría de las peleas terminan por método de decisión unánime, lo cual indica que se realiza una pelea completa. Esto implica mayor desgaste físico en los peleadores. Se recomienda a los peleadores considerar campamentos donde se trabaje el cardio, esto con la finalidad de cubrir el tiempo completo de la pelea sin denotar mayor cansancio.
- El segundo método de decisión es el KO/TKO lo cual señala que la empresa debe estar preparada y tener un equipo médico adecuado para este tipo de finalizaciones, por el lado de los peleadores se les recomienda preparación física para evitar este tipo de finalizaciones.
- La tercera forma de finalización de una pelea es la de submission, esto apunta a recomendar a los peleadores a implementar en su campamento, preparación en técnicas de lucha, jiu-jitsu y zambo para evitar sumisiones.
- LA parte más golpeada en el cuerpo de un peleador es la cabeza, esto indica que los peleadores deben realizar revisiones médicas periódicas o al terminar su pelea, al recibir un mayor daño en la cabeza el peleador se convierte en propenso a sufrir alguna contusión o daño cerebral tras recibir muchos golpes a lo largo de su carrera como peleador.

- Se identificó al peleador Max Holloway como el peleador que más golpes ha recibido en las peleas de la UFC, se recomienda al peleador, una revisión general médica para detectar posibles daños colaterales.
- Se identificaron a los peleadores con más golpes recibidos en las piernas, por lo que se les recomienda realizar análisis médicos para detectar posibles fisuras, hematomas o posibles daños causados por la cantidad de golpes recibidos.
- Se identificaron a los peleadores con más golpes recibidos en el cuerpo a los cuales se les recomienda realizar análisis médicos enfocados en las zonas del hígado, riñones y las costillas siendo estas las zonas más comúnmente golpeadas.