Informe 4 Actividad 5

Data Science with Python

Build and Evaluate Models

Edwin Spencer

Contexto

Este conjunto de datos incluye 41.586 resultados de partidos de fútbol internacionales desde el primer partido oficial en 1972 hasta 2019. Los partidos van desde la Copa Mundial de la FIFA hasta la Copa FIFI Wild hasta partidos amistosos regulares. Los partidos son estrictamente internacionales masculinos completos y los datos no incluyen los Juegos Olímpicos o los partidos donde al menos uno de los equipos era el equipo B de la nación, el U-23 o un equipo seleccionado de la liga.

Nota sobre los nombres de equipo y país:
Para los equipos locales y fuera, se ha utilizado el nombre actual del equipo. Por ejemplo, cuando en 1882 un equipo que se hacía llamar Irlanda jugó contra Inglaterra, en este conjunto de datos, se llama Irlanda del Norte porque el equipo actual de Irlanda del Norte es el sucesor del equipo de Irlanda de 1882. Esto se hace para que sea más fácil rastrear el historial y las estadísticas de los equipos.

Para los nombres de países, se utiliza el nombre del país en el momento del partido. Entonces, cuando Ghana jugó en Accra, Gold Coast en la década de 1950, a pesar de que los nombres del equipo local y el país no coinciden, fue un partido en casa para Ghana. Esto está indicado por la columna neutral, que dice FALSO para esos partidos, lo que significa que no estaba en un lugar neutral.

Fuente de la información

Los datos se recopilan de varias fuentes, entre ellas Wikipedia, fifa.com, rsssf.com y sitios web de asociaciones de fútbol individuales.

Definición de Framework y EDA

Justamente, con el conocimiento en la actividad anterior, El marco de trabajo a utilizar será el BADIR ya que, según la necesidad del negocio, es la herramienta que más se ajusta para cumplir con los objetivos planteados

Objetivos que atender

Como parte los objetivos planteados y a resolver se encuentran los siguientes:

- ¿Quién es el mejor equipo de todos los tiempos?
- ¿A representado una ventaja jugar en casa?
- ¿Qué métodos probados podemos usar para descubrir más información y por qué?

Plan de análisis

Justamente este es uno de los puntos por el cual se definió utilizar el marco de trabajo BADIR, ya que con la información (DataSet) que se cuenta hasta el momento nos permite poder identificar claramente los objetivos del negocio, además de que con base a esto se puede definir una hipótesis y poder definir un plan de trabajo para poder cumplir con los objetivos planteados.

Colección de datos

Con respecto a los datos, se cuenta con datos históricos desde 1972 hasta el 2019 los cuales representan una cantidad amplia de registros (41.586) además de contener variables importantes que nos ayudarán a poder los objetivos planteados anteriormente, a partir de acá se procederá a realizar una depuración y validación de los datos, para obtener resultados más certeros.

Estructura del dataSet

- fecha - fecha del partido
- home_team - el nombre del equipo local
- away_team - el nombre del equipo visitante
- home_score: puntaje del equipo local de tiempo completo que incluye tiempo extra, sin incluir penales
- away_score: puntaje del equipo visitante de tiempo completo que incluye tiempo extra, sin incluir penales
- torneo - el nombre del torneo
- ciudad - el nombre de la ciudad / pueblo / unidad administrativa donde se jugó el partido
- country - el nombre del país donde se jugó el partido
- neutral: columna VERDADERO / FALSO que indica si el partido se jugó en un lugar neutral

Gestión de los datos

Como parte de la administración o gestión de los datos, se debe realizar varias acciones importantes para poder aprovechar de la mejor manera la información contenida en el set de datos, las acciones a ejecutar a los datos serían las siguientes:

- Depuración
- Transformación
- Reducción
- Discretización

Set de datos

Considero que con la información que se comparte en el set de datos es suficiente para poder realizar el análisis que se solicita, más sin embargo si se quisiera analizar con mayor detalle el comportamiento de los partidos si sería importante poder conocer los minutos de cada gol, tarjetas, tanto amarillas como rojas para conocer la afectación que tiene esto en los partidos, entre otros datos.

Preparación y exploración de los Datos

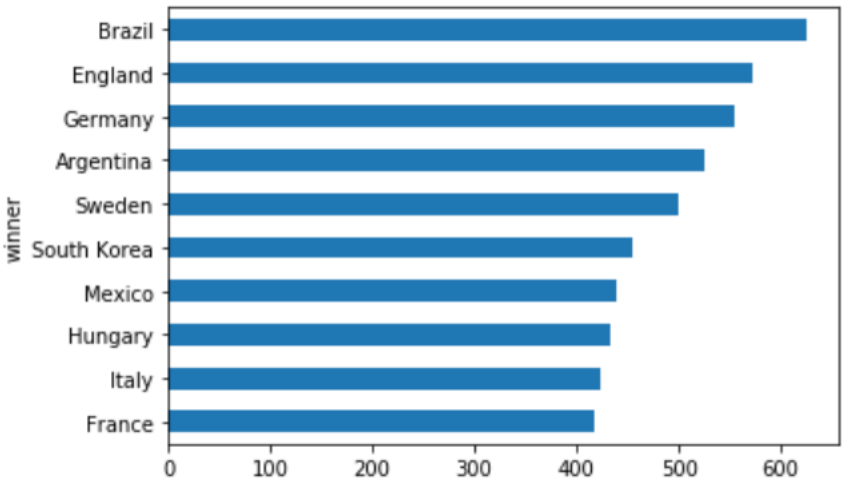
Acciones ejecutadas

- Entendimiento del dataset a través de los comandos Head(), describe(), info(), columns() y dtypes
- Procedí a crear 5 nuevas variables
    - Home_away_winner: Donde 1 significa que el ganador fue el equipo de casa, 2 el ganador fue el equipo visitante y 3 que el resultado fue un empate.
    - Winner: para conocer el nombre del equipo ganador
    - Losser: Para conocer el nombre del equipo perdedor
    - Tournament:Para identificar si el partido correspondía o no a juego de un campeonato
    - ID: Un numero consecutivo para identificarlo como ID.
- Validación de los datos nulos a través del comando isnull().any() .
- Además se realizó la verificación de si existen valores duplicados con el comando duplicated().any() sin encontrar ningún valor duplicado.

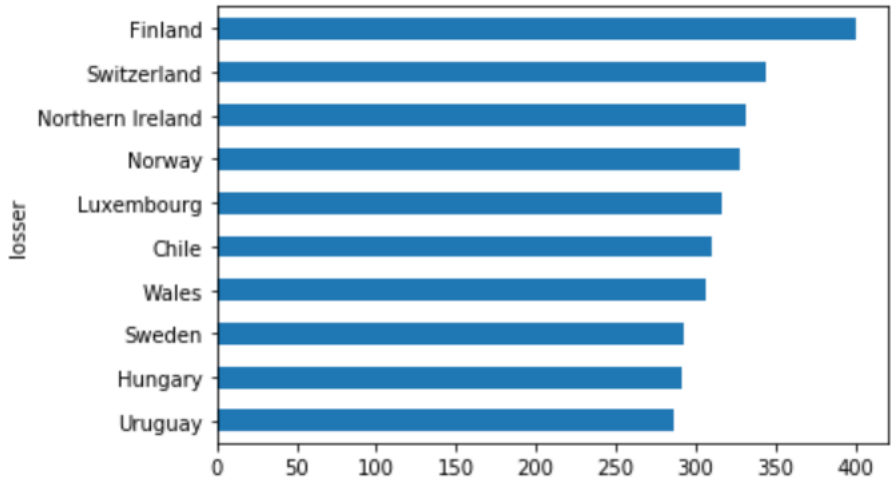
Creación y evaluación de modelos

¿Quién es el mejor equipo de todos los tiempos?

- Como parte del análisis podemos observar en la siguiente gráfica como se refleja que el equipo ganador de todos los tiempos a sido Brazil, además se refleja el top 10 de los equipos ganadores en toda la historia.

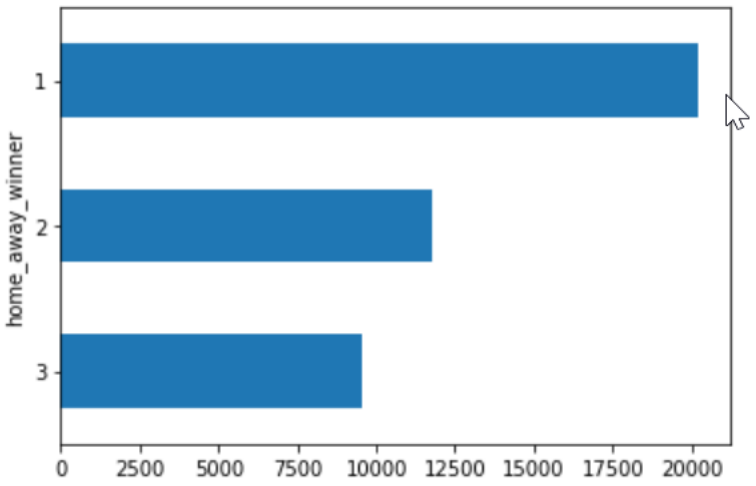


- Además, quise reflejar cual es la selección nacional que más partidos a perdido en la historia el cual es Finlandia, con un aproximada de 400 partidos, de igual forma, presento el top 10 de esta categoría:



¿A representado una ventaja jugar en casa?

- Analizando los datos podemos observar como realmente jugar en casa representa una ventaja, ya que en más de 20000 partidos el equipo casa fue el ganador y en aproximadamente 12500 fue el equipo visitante el vencedor y por último en menos de 10000 partidos se presento un empate.



¿Qué métodos probados podemos usar para descubrir más información y por qué?

Como parte del proceso se evaluaron 3 modelos Random Forest, Regresión Linea y SVR los cuales, posterior a los entrenamientos nos da los siguientes resultados:

- Random Forest

```
In [373]: print(cross_val_score(modelRF, train_features,train_labels))
          [0.99865968 0.99965188 0.99982328]
```

- SVR

```
In [374]: print(cross_val_score(modelSVR, train_features,train_labels))
          [0.28658623 0.29034574 0.28639949]
```

- Regresión lineal

```
In [375]: print(cross_val_score(modelLR, train_features,train_labels))
          [0.28658623 0.29034574 0.28639949]
```

A partir de estos resultados, tomo la decisión de utilizar el modelo Random Forest para correr el set de pruebas.

```
In [376]: modelRF.score(train_features,train_labels)
Out[376]: 0.9999921239415687
```

```
In [378]: rmse = sqrt(mean_squared_error(test_labels, predictions))
```

```
In [379]: predRsquared = r2_score(test_labels,predictions)
```

```
In [380]: print('R Squared: %.3f' % predRsquared)
          print('RMSE: %.3f' % rmse)
          R Squared: 1.000
          RMSE: 0.017
```