UNIVERSIDAD DE PIURA

FACULTAD DE INGENIERÍA

PYTHON PARA EL ANÁLISIS DE DATOS 2

Informe Final

Grupo 13

Gutierrez Lozano, Gonzalo Raúl
Soto Rodríguez, Fernando Daniel
Torres Cabrera, Alisson Xiomara
Vega Sanz, Victor Alejandro

Profesor:
Ing. Pedro Rotta.

Lima, febrero 2022

ÍNDICE

1. INTRODUCCIÓN

En el siguiente trabajo se analizará el dataset escogido de la página Kaggle, que contiene estadísticas internacionales de ajedrez en cada país. Para el desarrollo de esta investigación, hemos utilizado diferentes librerías que nos han permitido hacer códigos para analizar la data y programar el mejor modelo para que nuestros resultados sean precisos. Para el caso de estudio, nos centraremos en predecir la cantidad de mujeres ajedrecistas que hay en cualquier país, dependiendo de la cantidad de jugadores que haya en el mismo.

2. ANÁLISIS DEL PROBLEMA

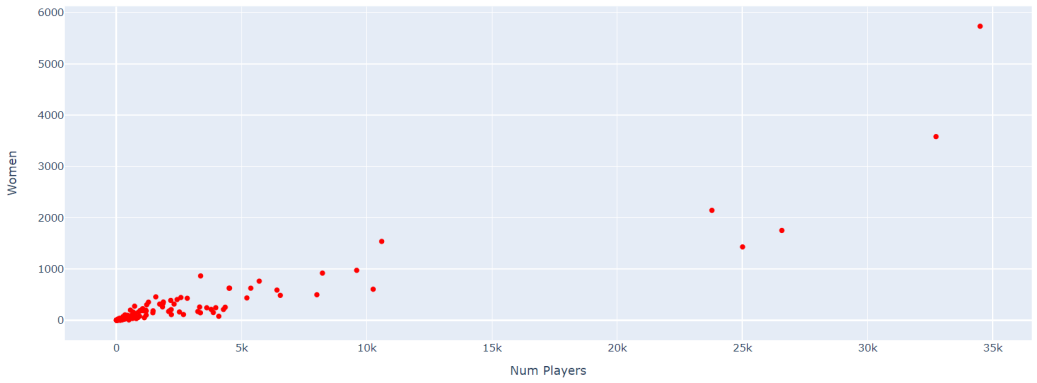
2.1. Planteamiento del problema

De un determinado número de jugadores ajedrecistas en un país, se desea elaborar un modelo capaz de predecir el número de mujeres ajedrecistas que existen en el mismo.

2.2. Características de la data

Se cuenta con 190 filas con información de la cantidad de ajedrecistas que existen en diversos países, sin embargo, se observaron dos filas de las que se desconocía su país de origen, debido a esto y a que la cantidad es muy pequeña en comparación a toda la data existente, se decidió eliminarlas.

Dentro de las características en cada columna, se eligió trabajar con las siguientes: Num Players y Women, según el problema planteado anteriormente.

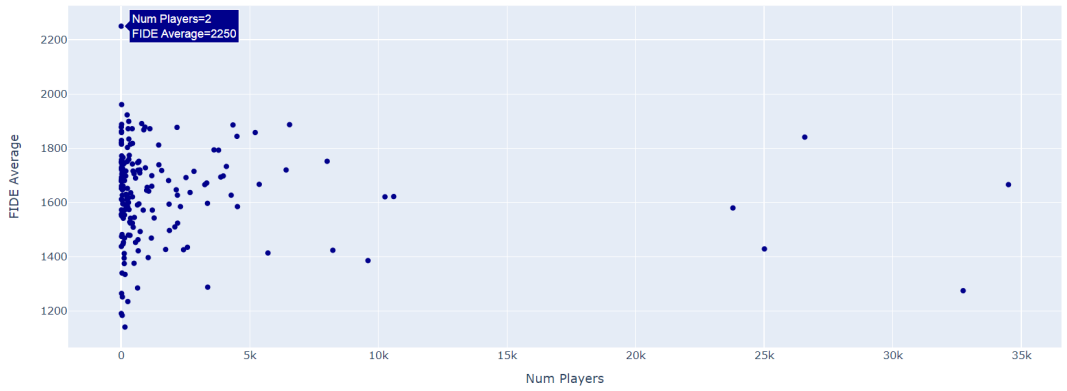


Gráfica 1

2.3. Importancia de la data

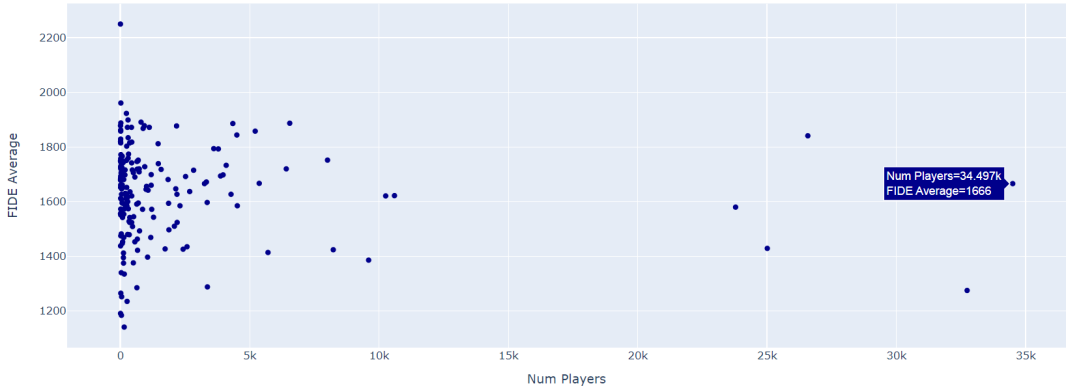
El poder contar con una cantidad de datos superior a 50 ayuda a realizar un mejor análisis estadístico, de modo que las predicciones que resulten del modelo elegido serán más cercanas a la realidad.

Con solo analizar la data se hace muy notorio que a pesar de que un país pequeño como Cambodia que cuenta con solo dos jugadores, resulta tener un puntaje promedio (FIDE) mayor al de una potencia como Rusia, la cual posee un número de ajedrecistas considerable. Por lo tanto, es importante observar que un mayor número de jugadores no asegura una mejor calidad.



Gráfica 2

En esta gráfica se muestra que Cambodia, con 2 jugadores alcanza un puntaje promedio de 2250 puntos.



Gráfica 3

En esta gráfica se muestra que Rusia, con 34497 jugadores alcanza un puntaje promedio de tan solo 1666 puntos.

3. ANÁLISIS DE RESULTADOS

El modelo que usamos fue Random Forest para regresión, en donde ”X” es el número de jugadores y “y” representa a la cantidad de mujeres ajedrecistas. Se decidió que los cuatro modelos a realizar serían usando el mismo hiper-parámetro “random state”, con el valor de 50, pero con “n_estimators” distintos.

3.1. Modelo 1

Se realizó un Random Forest con “n_estimators=100”.

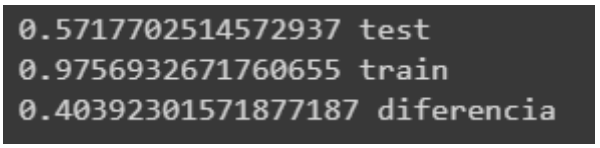


Imagen 1. Métricas para el modelo 1.

Las métricas usadas fueron el r2_score para test y para train, dando como resultado el de la imagen 1. Se descubrió que tiene un sobreajuste.

3.2. Modelo 2

Se realizó un Random Forest con “n_estimators=200”.

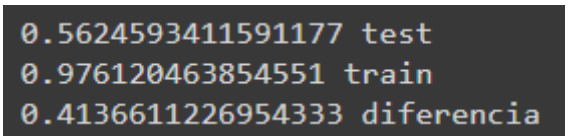


Imagen 2. Métricas para el modelo 2.

Las métricas usadas fueron el r2_score para test y para train, dando como resultado el de la imagen 2. Se descubrió que tiene un sobreajuste.

3.3. Modelo 3

Se realizó un Random Forest con “n_estimators=400”.

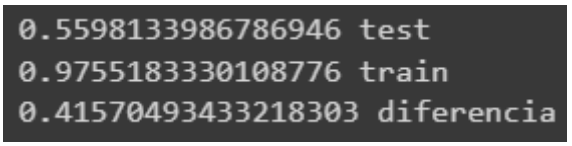


Imagen 3. Métricas para el modelo 3.

Las métricas usadas fueron el r2_score para test y para train, dando como resultado el de la imagen 3. Se descubrió que tiene un sobreajuste.

3.4. Modelo 4

Se realizó un Random Forest con “n_estimators=800”.

```
0.5554222213997386 test
0.973297804855216 train
0.4178755834554774 diferencia
```

Imagen 4. Métricas para el modelo 4.

Las métricas usadas fueron el r2_score para test y para train, dando como resultado el de la imagen 4. Se descubrió que tiene un sobreajuste.

3.5. Modelo Seleccionado

De todos los modelos, se escogió el modelo 1 como el mejor, porque su diferencia entre r2_score de test y train es el menor. Cabe precisar que lo ideal sería que esta diferencia sea aún mucho menor, pero para efectos prácticos de este informe se aceptó este nivel de eficiencia

Esta diferencia de r2_score se traduce como mejor precisión y por eso en el momento de realizar la predicción se escogió al modelo 1, dando como predicción la imagen 5.

```
Ingresa el valor la cantidad de jugadores: 1150
La cantidad de jugadoras es:  [177.78]
```

Imagen 5. Predicción del modelo 1.

A manera de comparación, se predijo el mismo valor X para los modelos 2, 3 y 4, dando valores similares pero menos precisos.

4. CONCLUSIONES

- Mediante más datos se tengan, los resultados del modelo y por lo tanto la toma de decisiones que puedan tomar los usuarios, será más precisa. Entre las decisiones que se pueden llevar a cabo se encuentra que si se observa que la cantidad de mujeres ajedrecistas y representantes de sus países está decayendo o creciendo, se pueden realizar o seguir realizando programas de incentivación hacia ellas.
- A pesar de existir un nivel de sobreajuste en el modelo, al momento de introducir valores para predecir en la función, se obtiene un valor dentro de un rango aceptable.
- Las gráficas de dispersión nos permite definir si existe una relación directa o inversamente proporcional entre 2 categorías.
- Cuando se introduce una cantidad de 10k o más como el número de jugadores, la cantidad de jugadoras predicha solo es un poco cercana a lo que se podría esperar como resultado, pero esto se puede explicar mediante la gráfica de dispersión (Gráfica 1) realizada en donde para cantidades superiores a este valor, los puntos se encuentran mucho más dispersos (hay menos data). Contrario a lo que sucede en el punto anterior, para valores de jugadores menores a 10k, la data se encuentra más concentrada y esto conlleva a que el número de jugadoras predicho sea más preciso.