Predicción de resultados de Fútbol

Recuperación de Información y Recomendaciones en la Web

Integrantes:

Juan Pablo Pascual 4.569.366-6

Nicolás Sanguinetti 4.712.015-4

Kevin Quincke 4.789.663-6

Miguel Langone 4.651.185-1

Docente:

Libertad Tansini

Índice

1. Introducción

En los últimos años las casas de apuestas han crecido de manera muy rápida, impulsadas principalmente por las aplicaciones web y mobile, llegando a un público mucho más amplio. Ante esta creciente demanda, los jugadores desean sacar el mejor partido de sus apuestas con el fin de ganar más dinero.
A este aumento se le suma las redes sociales que ayudan a distribuir información de los partidos, así como de información referente a casi cualquier equipo de fútbol del mundo.

2. Problema y Motivación

Dado que las casas de apuestas son tantas y cada una tiene sus propios algoritmos para evaluar qué equipo puede ganar, sumandole que el fútbol suele ser muy impredecible, hace muy difícil saber qué detalles son relevantes a la hora de predecir un resultado.
O dicho de otro modo, se deben incluir cientos o hasta miles de variables que pueden hacer posible un resultado específico en un partido de fútbol.
Para ello se decidió crear un sistema capaz de “predecir” en algún porcentaje, los posibles resultados de fútbol, y de esa manera que cualquier usuario sepa a qué equipo puede apostar en una casa de apuesta.
Se decidió realizando Web Scraping con métodos de aprendizaje automáticos, el sistema hablado anteriormente.
Algunos de los principales obstáculos a pasar son:

- Saber qué variables incluir en el programa
- Qué tipos de métodos de aprendizaje automático utilizar
- De qué fuentes extraer la información relevante
- Cómo mostrar de manera amigable dichos resultados

3. Enfoque de la solución

Para poder plantear un enfoque a la solución primero se tuvieron que destacar que funcionalidades deseábamos incluir, así como también que herramientas serian útiles para poder realizarlo.

Primero que nada establecimos una arquitectura del sistema la cual será constituida por 4 módulos principales y una base de datos:

- Módulo Presentación
- Módulo Publicador
- Módulo de recuperación de datos históricos

- Módulo de recuperación de valores de apuestas
- Módulo de aprendizaje automático
- Base de datos

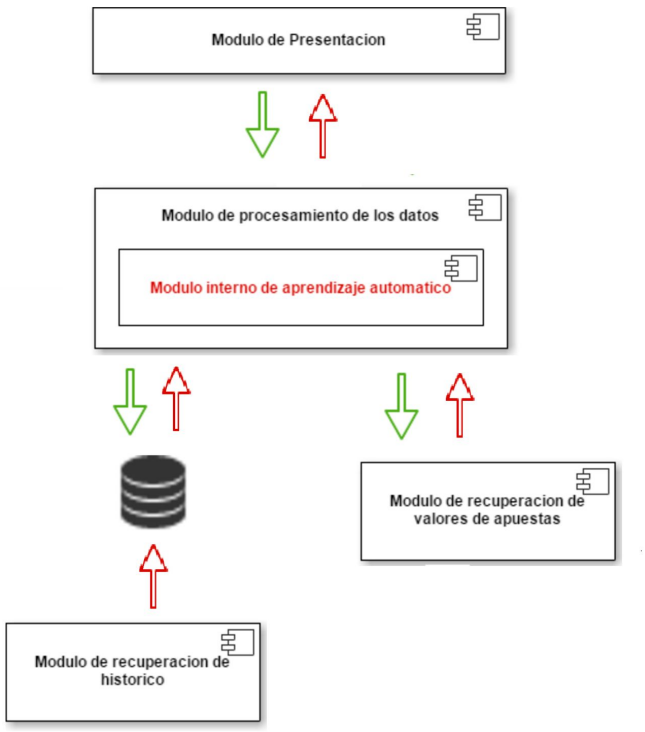
La solución planteada por el equipo fue que el módulo encargado de recabar información de los datos históricos de un partido guarde toda la información encontrada dentro de la base de datos para un par de equipos en específico.

Una vez recabada dicha información, se entrenará a 3 algoritmos de aprendizaje automático y se los evaluará para poder elegir el que mejor performance tiene.

Al finalizar el procesamiento estos datos quedan cargados en la base de datos disponibles para que la página web consuma, a partir de determinados servicios publicados por el Módulo Publicador, toda la información necesaria para el usuario.

4. Diseño, Implementación

El diseño se llevó a cabo siguiendo el siguiente diagrama de despliegue:



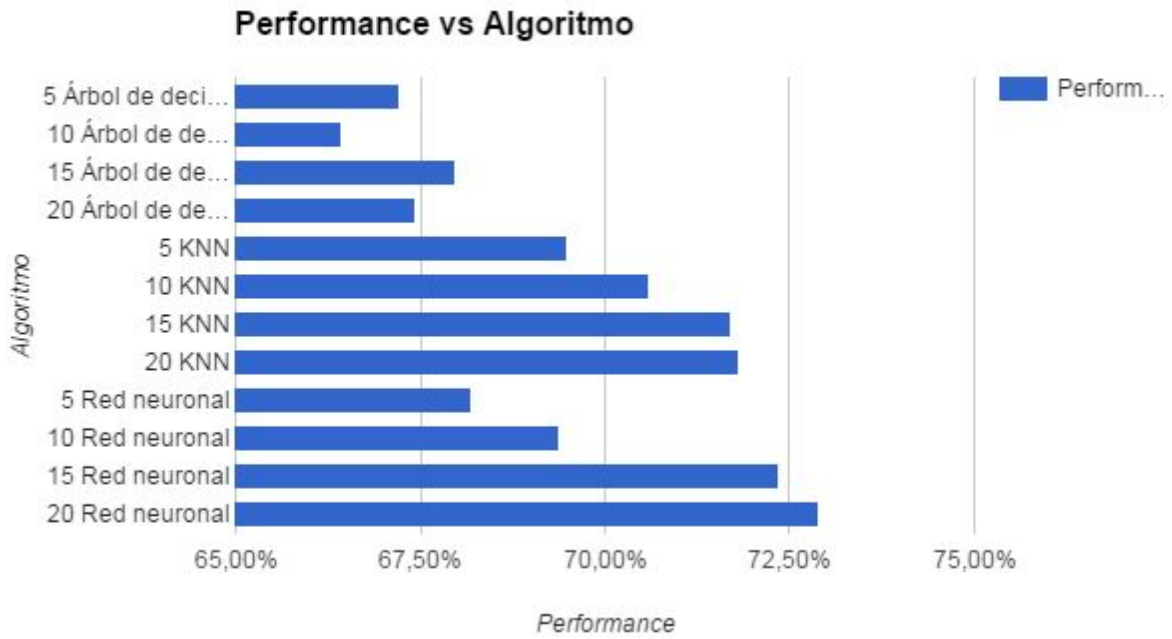
Modulo de Presentacion
La presentación del proyecto será realizada mediante una página web, la cual fue desarrollada en HTML con Javascript.

Dicha implementación cuenta, por un lado con dos .jsp que se encargan de la presentación y realizan los llamados al servidor remoto. Y por otro lado se cuenta con servlets que atienden las peticiones de cambios de páginas y redirigen entre dichas páginas.
La presentación en nuestro caso no se considero de suma importancia dada la gran cantidad de lógica que dicho programa contiene.
En la cual se listaran los próximos partidos para la liga española de fútbol y cuando el usuario selecciona un partido este obtendrá información relevante de los equipos que juegan el partido y los datos de la predicción del algoritmo de aprendizaje automático.

Módulo de recuperación de datos históricos
Este módulo es de suma importancia ya que sin los datos históricos no se podría entrenar al algoritmo para intentar predecir un partido.
Para obtener los datos el módulo realiza web scraping sobre la página www.soccerway.com, ya que es una web que contiene un gran conjunto de datos históricos de partidos de fútbol, luego con los datos obtenidos los guarda en la base de datos, para que estos estén accesibles para el módulo de aprendizaje.
Para implementar este módulo se decidió utilizar Python con ayuda de la librería Beautiful Soup para realizar el scrapping.

Módulo de recuperación de valores de apuestas
Este módulo es muy similar al anterior dado que lo que se busca es extraer de una página web de apuestas, como la mayoria de las paginas de apuestas en sus web utilizan flash, se decidio extraer la informacion de la pagina www.oddsportal.com que no es ninguna casa de apuestas, pero recopila informacion de ellas y las exibe de una manera mas facil de extraer con tecnicas de Scrapping.
Una consideración importante es que los valores de las apuestas en este caso son únicos y además cambian constantemente, por lo tanto no sería productivo guardar los valores en una base de datos ya que estos pueden cambiar en cualquier momento en la página web porque solamente se tiene un posible resultado a futuro en esa web. Como en el módulo anterior, también se utilizará Python y Beautiful Soup para la extracción de la información, pero además se utiliza DryScrape para cargar la web ya que las tablas de la web se cargan mediante JavaScript.

Módulo de aprendizaje automático
Este módulo se encarga de intentar predecir en base a los resultados anteriores, que equipo tiene mayor probabilidad de ganar en un partido dado o si el caso más probable es un empate
Los tres algoritmos seleccionados para realizar el entrenamiento fueron KNN, Árboles de decisión y redes neuronales.
Para el entrenamiento se obtuvieron los resultados de los partidos de la liga española desde el año 2006 al 2016 (aproximadamente unos 6100 partidos) y se realizó un preprocesamiento de los datos para poder entrenar las distintas configuraciones de los algoritmos seleccionados. Una de las variables más influyentes a la hora de predecir resultados fue la cantidad de partidos que se consideraban “hacia atrás” en el tiempo. A continuación se muestra una gráfica con los resultados obtenidos en el entrenamiento.



Para medir la performance de cada algoritmo se evaluaron aproximadamente 600 partidos y se le permitió realizar 5 tipos de predicciones a cada algoritmo: 1, X, 2, 1X, X2 que se corresponden con “gana el local”, “hay empate”, “gana el visitante”, “gana el local o hay empate”, “gana el visitante o hay empate”. En base a esto y a los resultados reales de los partidos, se obtuvo que la red neuronal fue el algoritmo con mayor performance de los tres, particularmente en la configuración que considera 20 partidos hacia atrás con un porcentaje de acierto de casi 73%.

Una vez obtenido el “predictor” de resultados se utilizaron las probabilidades de ocurrencia de cada una de las predicciones y los datos obtenidos de las casas de apuestas para el producto de la probabilidad de cada uno de los escenarios por la cuota obtenida de la casa de apuestas. Este valor se lo denomina “coeficiente” y representa la ganancia esperada de la predicción realizada.

Finalmente estos datos son enviados al módulo de presentación y se muestran en la tabla principal.

5. Funcionalidades y uso

Las funcionalidades del sistema son principalmente 3:

- Obtener próximos encuentros
- Obtener datos históricos de un par de equipos
- Proveer estadísticas y predicciones de resultados

Por fuera de las funcionalidades, su principal uso es el de brindar a los usuarios de dicho sistema la facilidad de decisión a la hora de apostar.

6. Evaluación y resultados

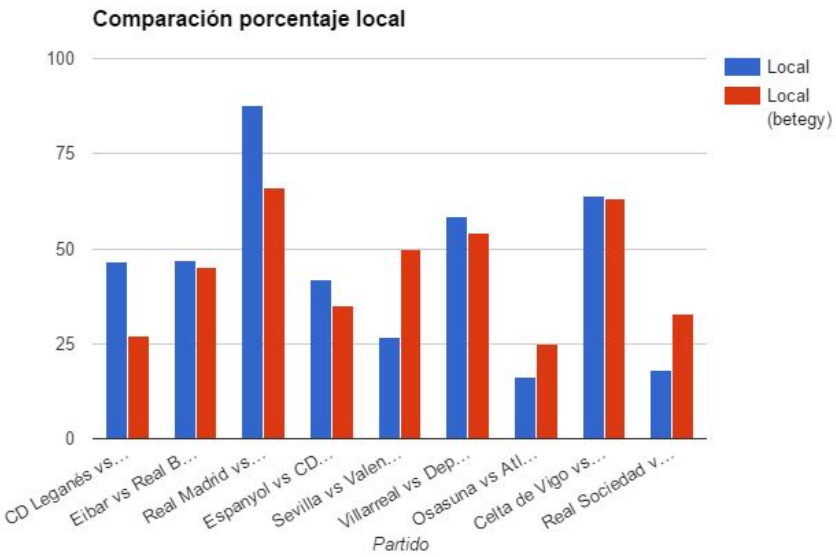
Para realizar la evaluación del producto, es necesario tener una referencia clara de contra qué otro sistema debemos compararnos. Para ello elegimos una web que mantiene porcentajes de las posibilidades de ganar de un par de equipos y la posibilidad de empate de ellos.

La web seleccionada fue la siguiente:
https://betegy.com/predictions/la_liga

La siguiente tabla muestra las diferencias entre las probabilidades de los resultados posibles de la web contra los obtenidos luego de ejecutar el aprendizaje automático:

Partido	Local	Local (betegy)	Empate	Empate (betegy)	Visita	Visita (betegy)
CD Leganés vs Osasuna	46,69	27	27,85	31	25,46	42
Eibar vs Real Betis	47,08	45	28,42	30	24,51	25
Real Madrid vs Sporting Gijón	87,79	66	7,91	23	4,3	11
Espanyol vs CD Leganés	41,87	35	30,84	33	27,28	32
Sevilla vs Valencia	26,71	50	31,15	28	42,13	22
Villarreal vs Deportivo Alavés	58,54	54	24,05	27	17,41	19
Osasuna vs Atlético Madrid	16,28	25	30,18	29	53,54	46
Celta de Vigo vs Granada	64	63	20,17	24	15,83	13
Real Sociedad vs Barcelona	17,92	33	31,65	34	50,42	33

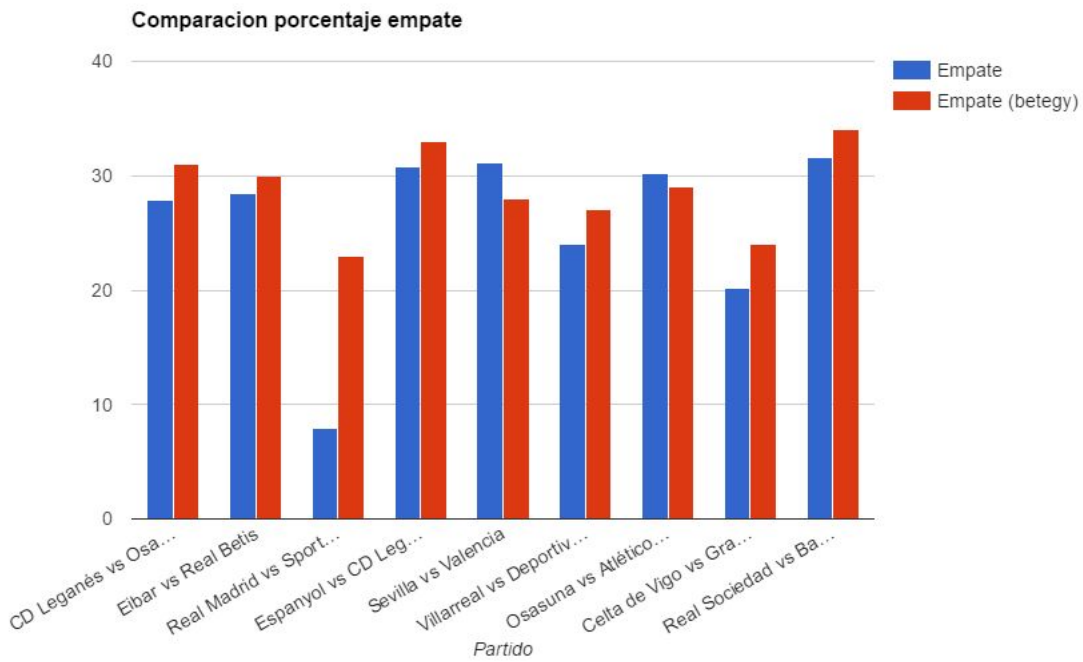
Primero veamos las probabilidades obtenidas a partir del aprendizaje realizado para los cuadros que juegan de local:



Como podemos observar, en determinados partidos como el del “Celta de Vigo vs Granada” el pronóstico es prácticamente el mismo que el de la casa de predicciones “Betegy”.

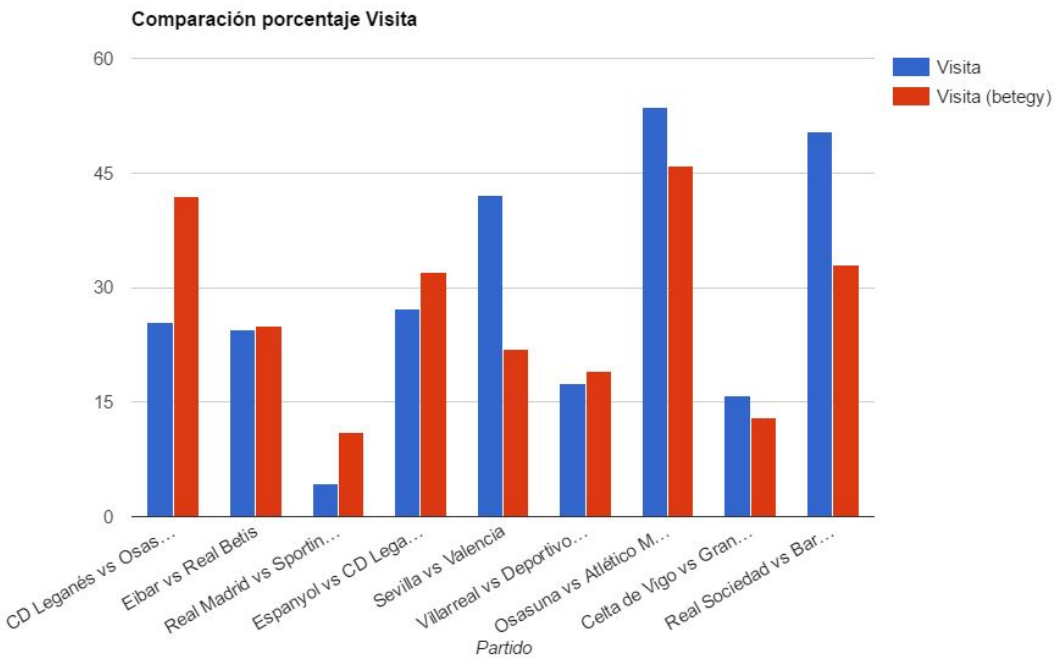
En otros casos podemos ver como el “Real Madrid vs Sporting Gijón” en el algoritmo pronostica en un 87% de que gana el Real Madrid y en la casa de probabilidades estima un poco menos.

Aquí presentamos las gráficas asociadas a la probabilidades de empate:



En este caso, como podemos ver, otra vez el algoritmo en determinadas ocasiones fue casi exacto en relación a la casa de probabilidades.

Por último, mostramos las probabilidades del cuadro visitante:



Este cuadro quizás es el que tiene más diferencia con respecto a la casa de apuestas, pero aún así los resultados no se diferencian en más de un 10% en ningún caso.

7. Conclusiones

Luego de realizar las pruebas podemos concluir que Beautiful Soup es una herramienta muy poderosa que permite navegar en las webs y obtener información de ellas.

Se obtuvo conocimiento muy valioso de los algoritmos de aprendizaje automático y los poderosos que pueden llegar a ser a la hora de procesar muchos datos.

Uno de los problemas que se observó es que muchas paginas realizan abreviaciones de los equipos (ej: Dep. La Coruna que representa a Deportivo La Coruña) y al extraer información de más de una página se necesita procesar los datos para saber cuales hacen referencia al mismo equipo. Además hay que tener en cuenta que al hacer web scraping, el cambio en la estructura del HTML puede afectar el parseo realizado, y esto puede resultar en un cambio del sistema.

A pesar de que las probabilidades obtenidas de algunos partidos difieren con respecto a las obtenidas en “Betegy”, creemos que se obtuvieron muy buenos resultados considerando que se logró entrenar solamente con los partidos de la liga española desde el 2000 al 2016.

Quizás si se hubiese contado con más tiempo para realizar el entrenamiento se hubiesen obtenido resultados aún mejores.

Finalmente concluímos que se obtuvo un prototipo aceptable, que cumple con nuestros principales objetivos, y el cual podría ser la base de un proyecto de mayor escala.

8. Trabajo Futuro

Como trabajo futuro cabe la posibilidad de procesar los datos con más variables que afectan a los partidos como pueden ser los jugadores de cada equipo, jugadores expulsados y lesionados.

Agregar más ligas al sistema sería algo muy deseable, ya que al tratarse de un prototipo sólo se utilizó la liga Española como ejemplo.

Otra posibilidad es la de recabar datos de más de una casa de apuesta para además de predecir el resultado poder recomendar en qué casa de apuesta conviene apostar.

Como último punto quizá no tan relevante en el funcionamiento del sistema pero si para los usuarios, sería deseable mejorar la interfaz de la página web y hacer una aplicación mobile para que los usuarios puedan consultar desde sus teléfonos celulares.