Master en Data Science (Kschool) – 4ª Edición

# MACHINE LEARNING FOR SPORTS BETS

*13 de noviembre de 2017*

Autor:
Ignacio San Juan Cisneros

Agradecimientos

A mi mujer y mi madre por darme soporte todos estos meses para poder realizar el Master.

# Índice

# 1. Objetivo del Proyecto:

El objetivo del proyecto es hacer un análisis detallado de algunas técnicas de Machine Learning para conocer sus características, ventajas y su forma de actuar. Para ello vamos a tratar de hacer un modelo con datos de partidos de Tenis con buscando por un lado de maximizar el acierto en la previsión del resultado y en ampliaciones posteriores de este trabajo de maximizar el retorno de una hipotética inversión (ROI) a través de apuestas deportivas.

# 2. Los Datos:

Los datos proceden de dos fuentes:

- OnCourt DataBase
- GitHub Jeff Sackmann

Estas nos proporcionan datos de los partidos desde hace más de 100 años. Evidentemente las estadísticas deportivas han ido mejorando con el tiempo y las más recientes contienen un mayor nivel de detalle. Este ha sido uno de los motivos por los que hemos elegido datasets a partir de año 2000 para el trabajo. También porque el propio deporte va cambiando con el tiempo (torneos, superficies, jugadores, material, etc.) y eso puede distorsionar el modelo.

Antes de utilizar la información, la contrastamos en otras fuentes, fundamentalmente en la web de la ATP y otras webs de resultados deportivos.

# 3. Metodología:

En la carpeta 'The Story' están todos los notebooks que componen el trabajo. Estos notebooks siguen el orden descrito en 'Contents'.

Se ha tratado la información paso a paso desde la primera carga en el punto 3 hasta la aplicación del modelo sobre unos hipotéticos datos futuros, los de 2017.

El trabajo se ha realizado en básicamente en Python, utilizando algunas de las librerías más usuales. La única librería que no viene instalada a través de Anaconda es graph Viz

Se hace un análisis estadístico detallado basado en análisis numérico de indicadores y un análisis visual de algunas de las características de los modelos, así como comparativas entre ellas.

# 4. El estudio:

Para cumplir el objetivo del proyecto he hecho un estudio profundo de los métodos de clasificación más populares:

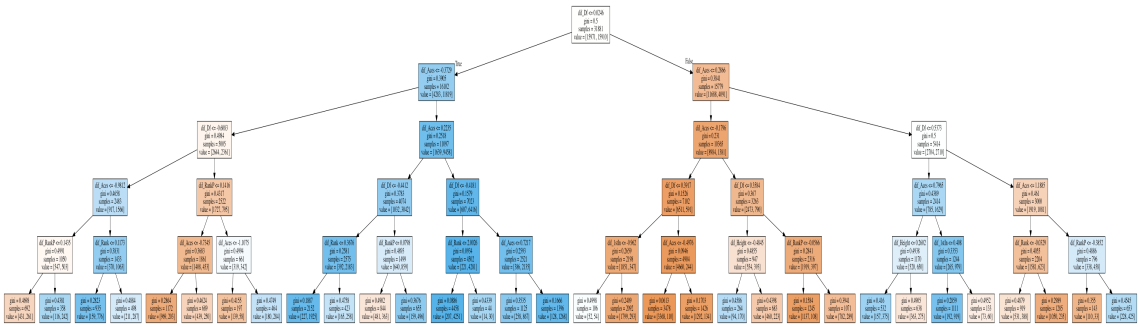
- Regresión Logística

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.466
Model:                            OLS   Adj. R-squared:                  0.466
Method:                 Least Squares   F-statistic:                     3709.
Date:                Mon, 13 Nov 2017   Prob (F-statistic):               0.00
Time:                        21:04:26   Log-Likelihood:                -17519.
No. Observations:               42509   AIC:                         3.506e+04
Df Residuals:                   42498   BIC:                         3.515e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4999      0.002    282.073      0.000       0.496       0.503
dif_Rank      -0.0315      0.002    -16.462      0.000      -0.035      -0.028
dif_RankP      0.0612      0.002     31.564      0.000       0.057       0.065
dif_Height    -0.0966      0.002    -43.040      0.000      -0.101      -0.092
dif_Age       -0.0296      0.002    -16.582      0.000      -0.033      -0.026
w_hand         0.0031      0.002      1.726      0.084      -0.000       0.007
```

```
l_hand           0.0006      0.002       0.327       0.743       -0.003
0.004
dif_Minutes      0.0418      0.002      21.745       0.000        0.038
0.046
dif_Aces         0.2479      0.003      98.312       0.000        0.243
0.253
dif_Df          -0.2089      0.002    -111.096       0.000       -0.213
-0.205
dif_1stIn       -0.0332      0.002     -16.712       0.000       -0.037
-0.029
==============================================================================
========
Omnibus:                     1136.811   Durbin-Watson:
3.077
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
560.487
Skew:                           0.004   Prob(JB):
1.96e-122
Kurtosis:                       2.438   Cond. No.
2.50
==============================================================================
========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

- KN Neighbours
- Support Vector Machine
- Decisión Tree

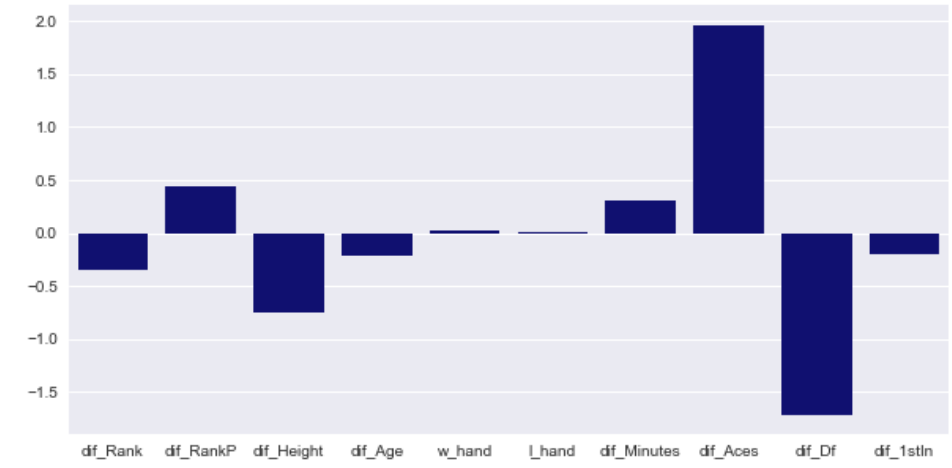


Así como algunos métodos derivados de estos modelos:

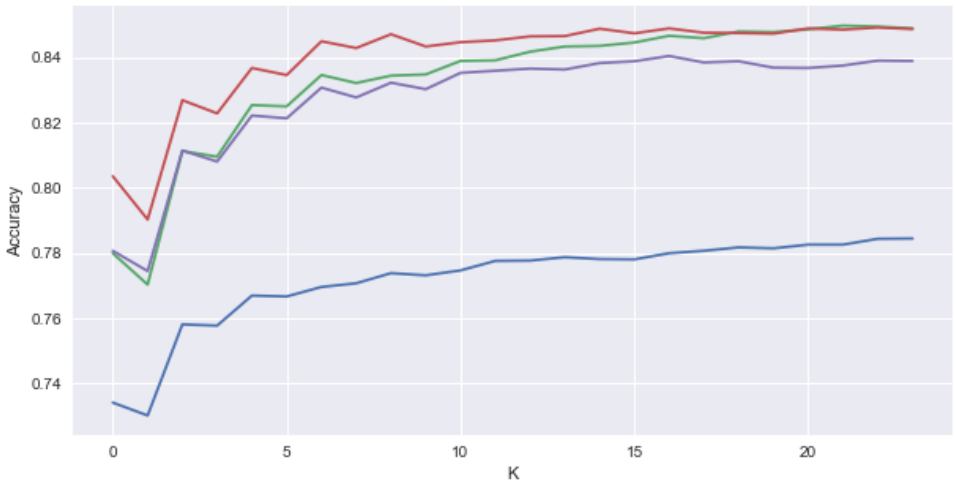
- Boosting
- Voting Clasificador

De los parámetros de estos modelos:

- Coeficientes



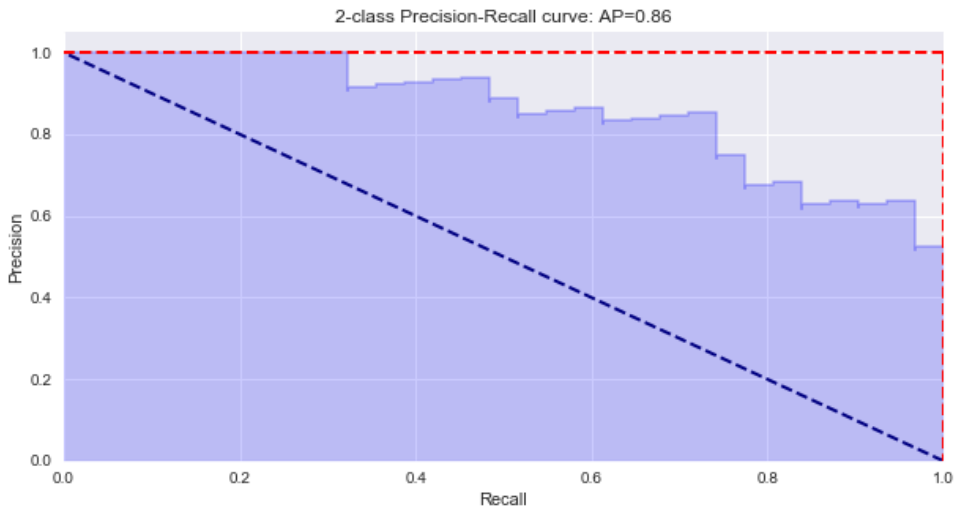

- Correlación
- Profundidad
- Número de iteraciones




- Número de estimadores

Y de la medición de su efectividad:

- Accuracy
- Cross Validation
- Bagging
- Precision y Recall




## 5. Conclusiones:

El resultado del estudio es el modelo con un mayor índice de accuracy en la predicción de los resultados de partidos futuros.

El modelo resultante ha sido optimizado para que su ajuste sea óptimo.

## 6. Aplicación:

Nada de esto tendría sentido sin su aplicación sobre un escenario más allá del test. En el último capítulo se aplica el modelo sobre los partidos de este año analizando las diferencias de rendimiento obtenidas.

# 7. Revisiones posteriores del trabajo:

Han quedado algunos puntos en lo que se podría ahondar este trabajo. En una revisión posterior es fundamental incidir en el punto más importante y complejo del trabajo: el tratamiento de la información.

Existe la posibilidad de crear nuevas variables que aporten mucha información al modelo. Para poder incluirlas en el modelo hay que evitar la colinearidad predominante manipulando las variables. Este es quizá el punto más complejo del proyecto.

El último punto en el que se debería incidir en una futura revisión es la implementación de algún modelo que no haya sido incluido, como una Red de Neuronas, modelo que a priori puede adaptarse bien a las características de nuestro problema.