UNIVERSIDAD INTERNACIONAL DE LA RIOJA

Laboratorio: Regresión y clasificación con árboles y random forests

Aprendizaje Automático

MÁSTER UNIVERSITARIO EN INTELIGENCIA ARTIFICIAL

2019-2020

FEDERICO DAMIAN ESTEBANEZ

DIEGO PEDRO GONZÁLEZ GONZÁLEZ

JOSÉ MARÍA ZAZO MARTÍN

07/06/2019

Asignatura	Datos del alumno	Fecha
Aprendizaje Automático	Federico Damian Estebanez, Diego Pedro González González, José María Zazo Martín	07/06/2019

Contenido

Asignatura	Datos del alumno	Fecha
Aprendizaje Automático	Federico Damian Estebanez, Diego Pedro González González, José María Zazo Martín	07/06/2019

Introducción

En este trabajo se busca comprobar la eficiencia del sistema de aprendizaje automático por Random Forest, mediante la biblioteca de Python adecuada, comparando los resultados con otras técnicas de aprendizaje automático como son los árboles de decisión.

Como datos para el entrenamiento y prueba, se utilizará el conjunto de Capital Bike Sharing contiene información relacionada con el programa de bicicletas compartidas de Washington DC en los años 2011 y 2012.

El conjunto de datos contiene valores agrupados por instantes, en los que se detalla la fecha, la temporada, el año, el mes, la hora, si es día festivo, diario o finde semana, condiciones meteorológicas, temperatura ambiente, sensación térmica, humedad, velocidad del viento y el número de bicicletas alquiladas, desglosado en usuarios casuales, registrados y la suma total de ambos.

A través de este conjunto de datos, se intentará estimar bajo unas condiciones de entrada, cual será el número de bicicletas alquiladas, y se clasificará como número de ventas altas si son más de 20 a la hora o como numero de ventas bajas en caso de que no se alcance dicha cifra.

Flujo de trabajo

Para alcanzar el estimador y el clasificador más adecuado, se recorrerán los siguientes pasos:

1. Tratamiento de los datos de entrada, determinando qué datos son útiles y necesarios para el aprendizaje y cuales se pueden descartar.
2. Selección de las variables de salida, utilizando solo aquellas que sean necesarias en cada caso.
3. Tratamiento de aquellas variables de entrada que sean categóricas.
4. División del conjunto de datos en conjuntos de prueba y de examen.
5. Ejecución de los algoritmos de regresión de Random Forest, buscando cuales son los parámetros en los que obtenemos mejor resultado.
6. Estudio de los resultados del regresor.
7. Comparativa entre los resultados de los clasificadores mediante Random Forest y Árboles de Decisión

8. Ejecución de los algoritmos de clasificación de Random Forest, mediante el uso de parámetros óptimos.
9. Estudio de los resultados del clasificador.
10. Comparativa entre los resultados de los regresores mediante Random Forest y Árboles de Decisión.

Variables de entrada

La selección y el tratamiento de las variables de entrada es prácticamente el paso más importante para crear modelos de aprendizaje lo más preciso posible, por lo que en este paso se detallarán las decisiones tomadas sobre las variables de entrada.

- instant: El índice del conjunto de datos. Dado que es una variable que no aporta ninguna información práctica, será eliminado**.**
- dteday: La fecha. Esta variable no tiene información relevante para realizar ninguna predicción. La información útil está desglosada en otros campos.
- season: Temporada del año. Esta variable categórica indica si estamos en 1: Primavera, 2: Verano, 3: Otoño o 4: Invierno.
- yr: Año. Esta información no puede ser utilizada para la regresión.
- mnth: Mes del año. Otra variable categórica.
- hr: Hora del día. Una variable categórica. Dado que las horas se pueden interpretar como variables continuas, no las someteremos a ningún proceso.
- holiday: Si el día es festivo o no. Variable categórica y binaria, por lo que no es necesario ninguna adaptación.
- weekeday; Dia de la semana. De domingo a lunes, se trata como variable categórica."
- weathersit: Condiciones meteorológicas. Puede tomar 4 valores, relacionados como 1: Despejado, 2: Bruma/Nublado, 3: Lluvias/Nevadas ligeras, 4: Tormentas/Nieve y Niebla. Estas variables son discretas e exclusivas entre si.
- temp: temperatura en Celsius. Variable de tipo decimal con coma flotante, que se utilizará como variable de entrada al sistema.
- atemp: Sensación térmica. Como la variable anterior, se eliminará ya que es totalmente dependiente de temp.
- hum: Humedad.Otra variable que desecharemos por no influir demasiado en la muestra y estar representada en la variable weathersit.

- windspeed: Velocidad del viento. Eliminamos esta variable, ya que interpretamos que su valor queda representado dentro de la variable weathersit.

Variables de salida

Las variables de salida son la respuesta del sistema, por lo que es necesario tener claro cuál es el objetivo del regresor o clasificador.

- casual: Número de usuarios no habituales del sistema de alquiler de bicicletas. Para este trabajo, no es necesario evaluar esta característica, por lo que se elimina.
- registered: Número de usuarios registrados del sistema de alquiler de bicicletas. Para este trabajo, no es necesario evaluar esta característica, por lo que se elimina. Esta variable junto con casual son dependientes de cnt
- cnt: Cuenta total de bicicletas alquiladas en esa franja horaria como suma de las casuales y las registradas. **Variable objetivo** de nuestro trabajo. A partir de esta variable crearemos el vector de variables de salida de entrenamiento y ensayo del regresor.
- sales: **Nueva variable creada para el clasificador**. Si el valor de cnt es mayor de 20, la variable tendrá un valor “high” de alto número de ventas, y en el caso de alquilar menos de 20 bicicletas durante la franja, tendrá un valor “low”, de bajas ventas. El vector que conforma estos resultados será la variable respuesta del clasificador. Por tanto es un label usado en la clasificación

Asignatura	Datos del alumno	Fecha
Aprendizaje Automático	Federico Damian Estebanez, Diego Pedro González González, José María Zazo Martín	07/06/2019

Algoritmos de regresión

La ejecución del algoritmo de regresión mediante Random Forest se realiza utilizando la biblioteca **RandomForestRegressor(RFR) y DecisionTreeRegressor (DTR)** de **sklearn.ensemble**. Para configurar adecuadamente el regresor, será necesario adecuar los siguientes parámetros:

RandomForestRegressor

- n_stimators: El número de árboles que conformarán el bosque.
- max_features: El número máximo de entradas a considerar cuando buscamos la mejor división.

DecisionTreeRegressor

- Criterion: Calidad de una partición
- Max_leaf_nodes: Medida de crecimiento óptimo del árbol
- Min_samples_leaf : número mínimo de muestras
- Min_samples_split El número mínimo de muestras necesario para dividir un nodo interno.

Esta selección de parámetros se realiza mediante la función **GridSearchCV** de la biblioteca **sklearn.model_selection** que busca para un determinado estimador, los resultados obtenidos con diferentes parámetros de entrada:

- `Best parameters RFR: {'max_features': 'log2', 'n_estimators': 200}`
- `Best parameters DTR : {'criterion': 'mse', 'max_depth': 8, 'max_leaf_nodes': 100, 'min_samples_leaf': 20, 'min_samples_split': 10}`

Algoritmo de clasificación

Para ejecutar el algoritmo de clasificación a través de Random Forest y Arbol de Decisión se utilizará las funciones **RandomForestClassifier(RFC) y DecisionTreeClassifier(DTC)** de **sklearn.ensemble**. Los parámetros escogidos mediante GridSearCV han sido

- `Best parameters : {'max_features': 'log2', 'n_estimators': 300}`
- `Best parameters DTC: {'criterion': 'entropy', 'max_depth': 8, 'max_leaf_nodes': 100, 'min_samples_leaf': 20, 'min_samples_split': 10}`

Asignatura	Datos del alumno	Fecha
Aprendizaje Automático	Federico Damian Estebanez, Diego Pedro González González, José María Zazo Martín	07/06/2019

Resultados de la regresión

Para determinar el éxito del regresor, se utilizan las siguientes métricas:

- Precisión, utilizando la media de la validación cruzada.
- Error absoluto, comparando los valores esperados y los predichos.

Además, es importante conocer que entradas han sido más importantes a la hora de generar el bosque de árboles de decisión, lo que obtendremos mediante la función **feature_importances_** del regresor.

En la siguiente tabla podemos comparar los valores obtenidos por nuestro regresor de Random Forest, comparado con un sencillo árbol de decisiones.

Medida	Random Forest	Árbol de Decisión
Accuracy	**0.8620**	**0.7927**
neg_mean_squared_error	**41.1052**	**54.5103**

Se observa que el método de regresión por **Random Forest** ha sido más preciso que el árbol de decisión simple.

Resultados de la clasificación

Las métricas para evaluar nuestros algoritmos de clasificación se muestran a continuación:

RFC:

Matriz de confusión, en la que se muestran las predicciones y los valores reales agrupados por valores, detectando el numero de falsos acierto por variable de salida.

Precisión general del clasificador, cual ha sido la tasa de éxito total.

Real \ Predicción	High	Low
High	2840	58
Low	58	520

Valor-F1, en el que ponderamos la precisión y la exhaustividad de los aciertos de cada variable según la fórmula:

$$F_1 = 2 \cdot \frac{Precisión \cdot Exhaustividad}{Precisión + Exhaustividad}$$

Medida	High	Low
F1	0. 9800	0. 8997

DTC:

Matriz de confusión

Real \ Predicción	High	Low
High	2849	49
Low	78	500

Valor-F1 :

Medida	High	Low
F1	0.9782	0.8873

Comparación

Los resultados comparados entre nuestro algoritmo y el árbol de decisión se muestran a continuación:

Medida	Random Forest	Árbol de Decisión
Accuracy_scored	0. 9672	0. 9635

Se observa que, aunque no significativamente, el método de regresión por **Random Forest** ha sido más preciso que el árbol de decisión simple.

Conclusión

Para el desarrollo de estos ensayos, se han probado diferentes técnicas y se han obtenido muy diversos valores. Cada cambio en el algoritmo ofrecía resultados variantes y tiempos de ejecución diferentes. En nuestro caso, hemos seleccionado el procesado de datos que explicamos en este informe, aportan los mejores resultados y que incluimos adjunto.

El resultado de los ensayos con diferentes valores se puede ver en el ANEXO I, y el código se puede estudiar, ya que está alojado en la plataforma (pedir acceso si se desea ver) kaggle.com.

Asignatura	Datos del alumno	Fecha
Aprendizaje Automático	Federico Damian Estebanez, Diego Pedro González González, José María Zazo Martín	07/06/2019

Bibliografía

Universidad Internacional de la Rioja. (2019). Aprendizaje Automático. Tema 7: Regresión y clasificación con árboles de decisión. Material no publicado.

Universidad Internacional de la Rioja. (2019). Aprendizaje Automático. Tema 8: Combinación de clasificadores: Bootstrapping, bagging y boosting. Material no publicado.

Universidad Internacional de la Rioja. (2019). Aprendizaje Automático. Tema 9: Aprendizaje supervisado: Regresión y clasificación con random forests. Material no publicado.

Scikit Learn Random Forest Regressor documentation. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Scikit Learn Random Forest Classifier documentation. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Scikit Learn Grid Search documentation. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Asignatura	Datos del alumno	Fecha
Aprendizaje Automático	Federico Damian Estebanez, Diego Pedro González González, José María Zazo Martín	07/06/2019

ANEXO I

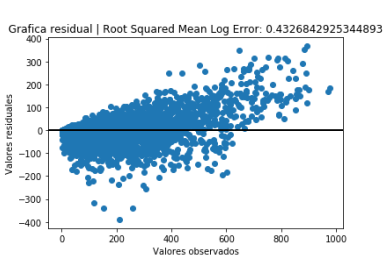
En este anexo se incluyen los resultados de las pruebas realizadas con los modelos de predicción regresivos, variando el tratamiento de las variables de entrada y salida.(normalizando o convirtiendo a dummies) Se ha decidido añadir por su **valor pedagógico**. Además, se ha comparado con otro algoritmo de regresión ExtraTreesRegressor

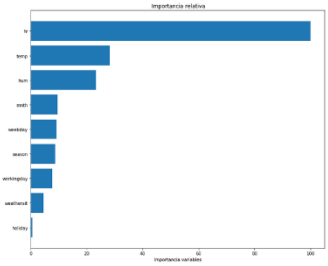
Asignatura	Datos del alumno	Fecha
Aprendizaje Automático	Federico Damian Estebanez, Diego Pedro González González, José María Zazo Martín	07/06/2019

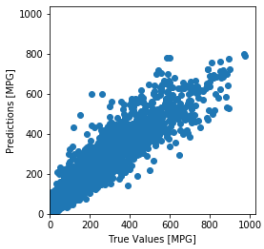
Los algoritmos han sido llamados sin ningún tipo de tratamiento de datos (Fork of Fork of kerneled5a09f9df)

Regresion: Random Forest

```
{'max_features': 'log2', 'n_estimators': 250}
0.8624535225668574
Precisión               :0.861381425741744
error medio absoluto    :41.05291716840649
```

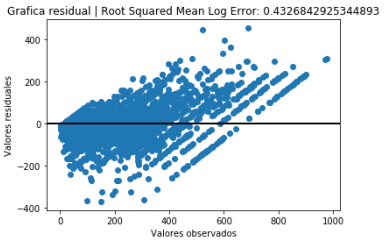


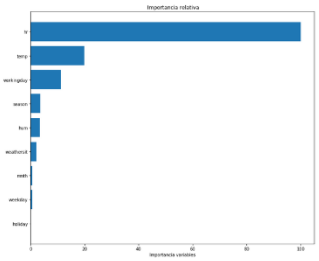


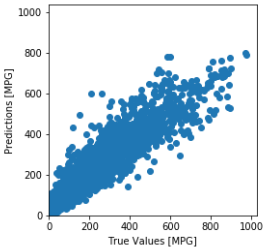


Regresion Decision Tree

```
{'criterion': 'mse', 'max_depth': 8, 'max_leaf_nodes': 100,
'min_samples_leaf': 20, 'min_samples_split': 10}
0.7927130050760155
Precisión               :0.7927129517143351
error medio absoluto    :54.51033904709948
```

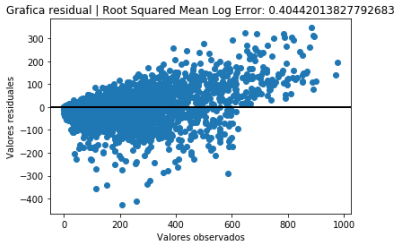


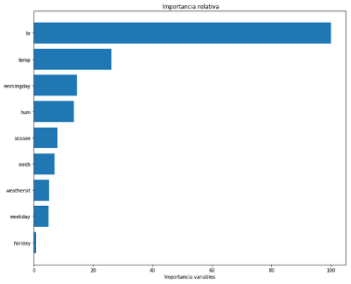


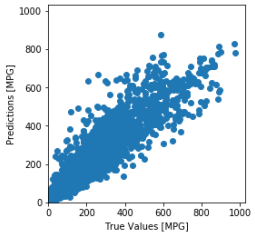


ExtraTreesRegressor

```
{'etr__max_depth': 20, 'etr__n_estimators': 500,
'pca__n_components': 8}
Precisión               :0.8618478107826277
error medio absoluto    :42.99854344039635
```







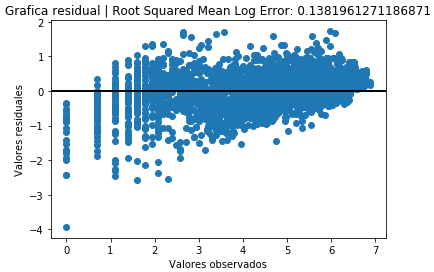
Comparativa models = [DecisionTree, RandomForest,ExtraTrees] Regressors

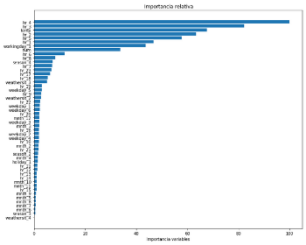
```
-8476.61975972062
-5075.020571098685
-5112.6064168541225
```

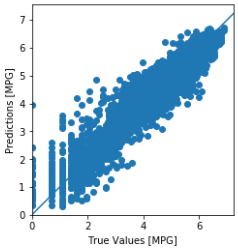
Todas las variables categóricas han sido pasadas a dummy. La variable cnt ha sido normalizada(kerneled5a09f9df)

Regresion: Random Forest

```
{'max_features': 'auto', 'n_estimators': 300}
0.8910537207354357
Precisión              :0.8901239608920449
error medio absoluto   :0.3550837398893295
```

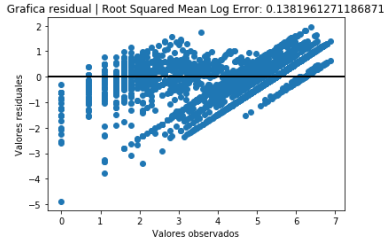


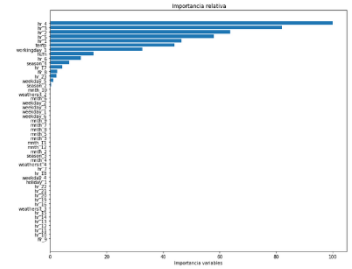


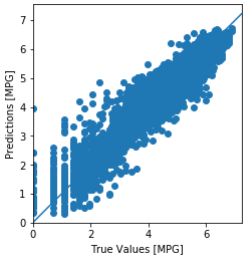


Regresion Decision Tree

```
{'criterion': 'mse', 'max_depth': 8, 'max_leaf_nodes': 100,
'min_samples_leaf': 20, 'min_samples_split': 10}
0.7608638119788496
Precisión              :0.7608645860354196
error medio absoluto   :0.5490769737142512
```

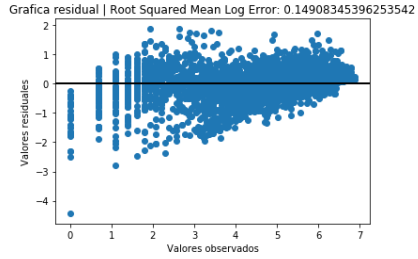


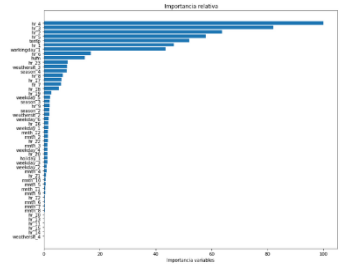


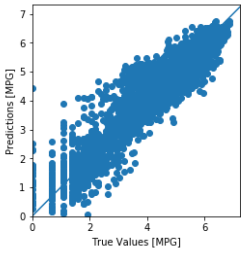


ExtraTreesRegressor

```
{'etr__max_depth': 20, 'etr__n_estimators': 500,
'pca__n_components': 31}
Precisión              :0.8719765377429829
error medio absoluto   :0.3951030044002766
```







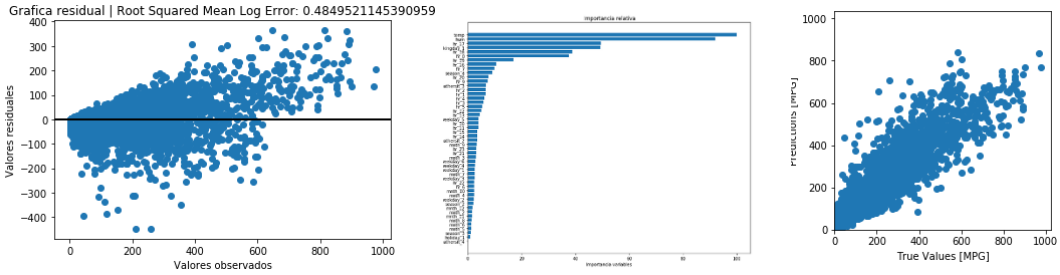
Comparativa

```
-0.44345689566218194
-0.2622120232911613
-0.2549211266172657
```

Todas las variables categóricas han sido pasadas a dummy. La variable cnt NO se normaliza(Fork of kerneled5a09f9df)

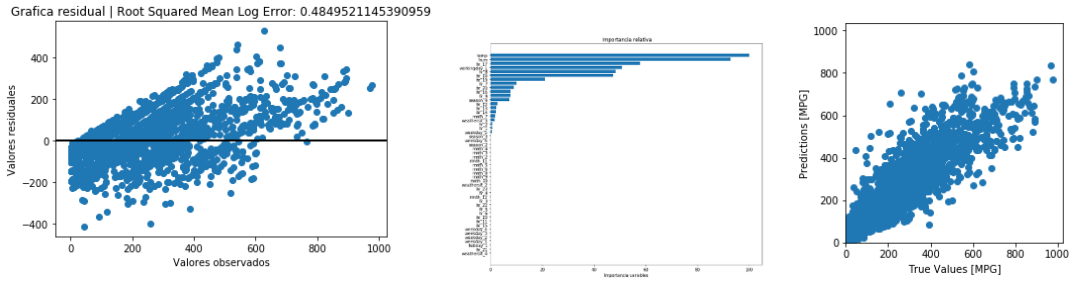
Regresion: Random Forest

```
{'max_features': 'auto', 'n_estimators': 150}
0.8206841258782187
Precisión              :0.8203124416071191
error medio absoluto   :49.51971162849106
```



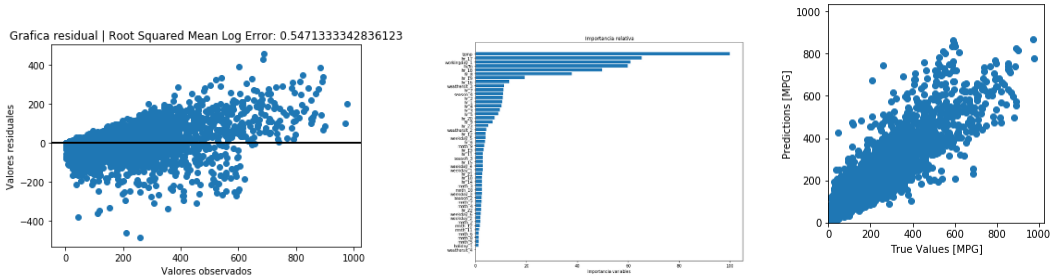
Regresion Decision Tree

```
{'criterion': 'mse', 'max_depth': 8, 'max_leaf_nodes': 100,
'min_samples_leaf': 20, 'min_samples_split': 10}
0.6392803591609012
Precisión              :0.6392815046379758
error medio absoluto   :78.67025002931177
```



ExtraTreesRegressor

```
{'etr__max_depth': 20, 'etr__n_estimators': 100,
'pca__n_components': 31}
Precisión              :0.8081287720733427
error medio absoluto   :53.266255268651435
```



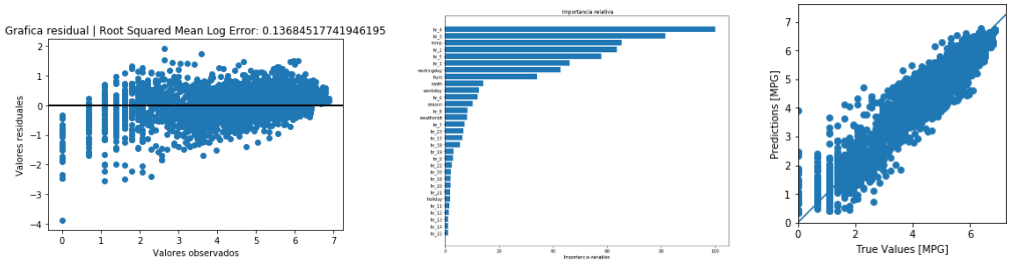
Comparativa

```
-10587.920039762581
-6351.712894265621
-6460.479656071419
```

Solo la variable categórica hr ha sido pasadas a dummy. La variable cnt ha sido normalizada(kernel5aa30d61c1)

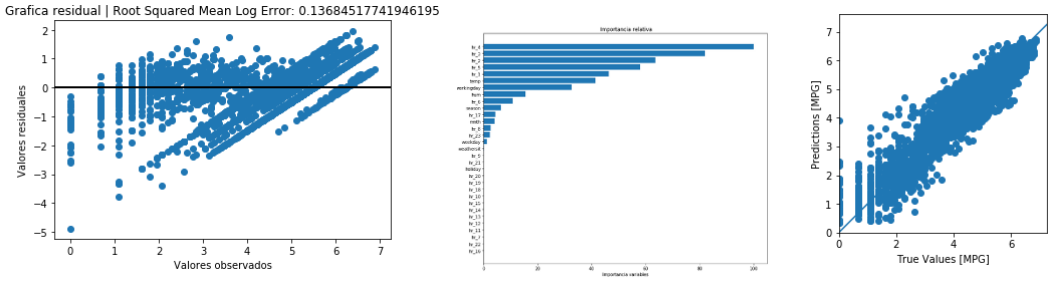
Regresion: Random Forest

```
{'max_features': 'auto', 'n_estimators': 200}
0.8910362548026324
Precisión               :0.8974076893636866
error medio absoluto    :0.34954679118290904
```



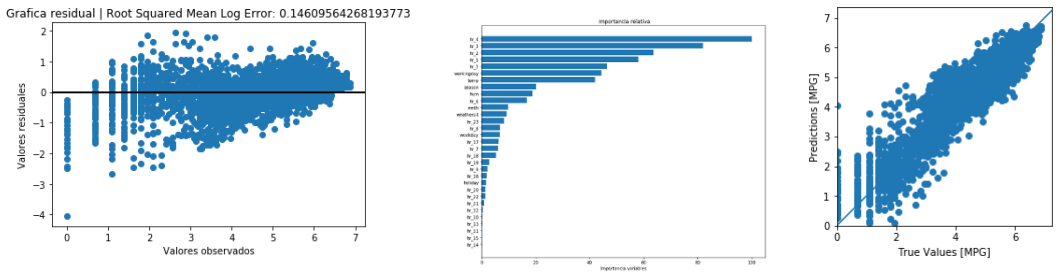
Regresion Decision Tree

```
{'criterion': 'mse', 'max_depth': 8, 'max_leaf_nodes': 100,
'min_samples_leaf': 20, 'min_samples_split': 10}
0.761279797725795
Precisión               :0.7612805920395482
error medio absoluto    :0.5480504720072049
```



ExtraTreesRegressor

```
{'etr__max_depth': 20, 'etr__n_estimators': 500,
'pca__n_components': 31}
Precisión               :0.87542163724157
error medio absoluto    :0.3856035981859412
```



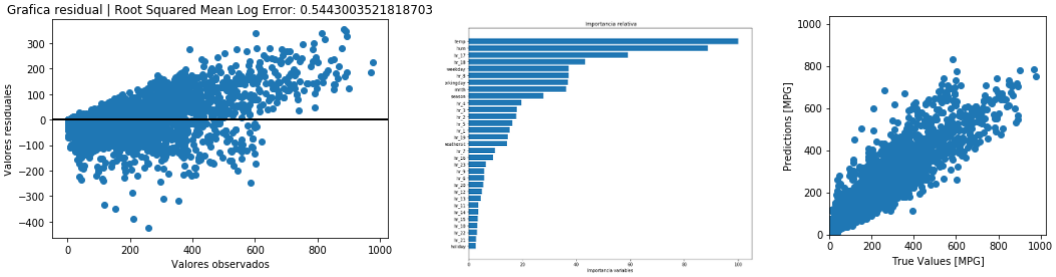
Comparativa

```
-0.44182064349317607
-0.2539996265787164
-0.24163468947721153
```

La variable hr ha sido pasadas a dummy. La variable cnt NO ha sido normalizada([Fork of kernel5aa30d61c1]())

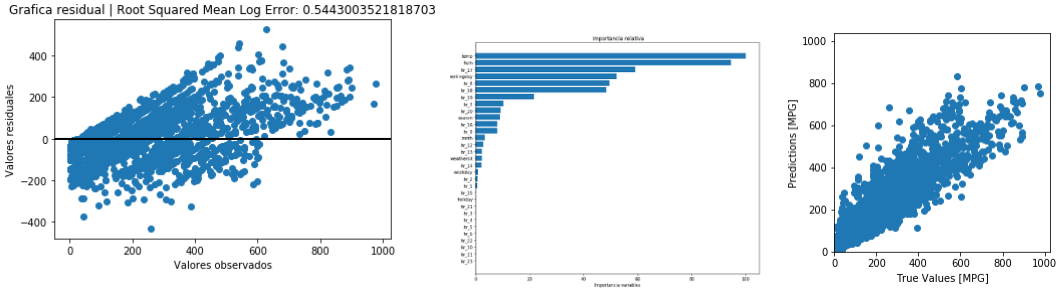
Regresion: Random Forest

```
{'max_features': 'sqrt', 'n_estimators': 300}
0.8216343647430522
Precisión                :0.8283895216589674
error medio absoluto     :50.261003492200494
```



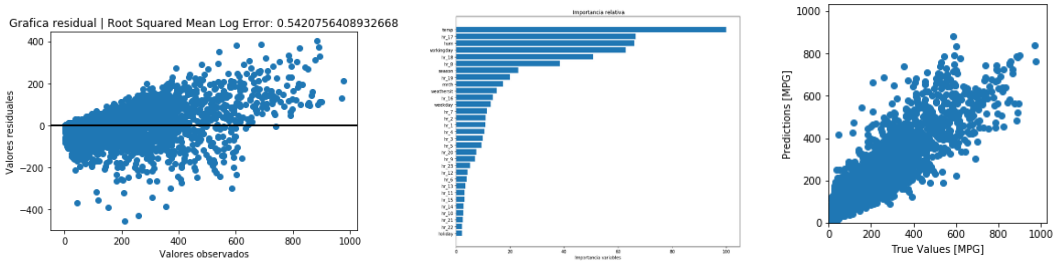
Regresion Decision Tree

```
{'criterion': 'mse', 'max_depth': 8, 'max_leaf_nodes': 100,
'min_samples_leaf': 20, 'min_samples_split': 10}
0.6384625831592826
Precisión                :0.6384636883456687
error medio absoluto     :78.52444348408991
```



ExtraTreesRegressor

```
{'etr__max_depth': 20, 'etr__n_estimators': 500,
'pca__n_components': 31}
Precisión                :0.81496185993184
error medio absoluto     :52.49949393312872
```



Comparativa

```
-10322.577959609309
-6342.6763815755985
-6294.546677170542
```

Normalización cnt