Capítulo 1

Selección de Variables

Este capítulo presenta un análisis referente a la elección de las mejores variables para alimentar el proceso de entrenamiento y predicción de los modelos de capítulos posteriores. Se llevan a cabo dos análisis independientes, desde dos puntos de vista diferentes. El primer análisis estudia, mediante procedimientos estadísticos, las correlaciones y asociaciones que existen entre las covariables (regresores exógenos) y el *target*. Para ello, se utilizan diferentes técnicas estadísticas. Por otro lado, se realiza también un análisis basado en *Machine Learning*, observando la importancia que las diferentes covariables tienen a la hora de ajustar un modelo estático.

1.1. Análisis estadístico de la importancia de las variables

A la hora de realizar una predicción mediante cualquier modelo predictivo, es de crucial importancia realizar una selección previa de las variables. Esto se debe a que el potencial predictivo de un conjunto de datos depende directamente de las distintas características que introduzca. Un dataset con pocas variables no podrá realizar predicciones complejas y precisas, mientras que un dataset con demasiadas variables tendrá dificultades para generalizar el resultado obtenido (sobreajuste). Normalmente, una combinación de las variables del dataset que deje fuera algunas de las características con menos importancia dará un resultado con la suficiente complejidad como para ser útil, y por otro lado será capaz de generalizar el resultado a muestras que no se encuentren en el conjunto de entrenamiento.

Esta primera sección del capítulo explora las relaciones que existen entre las distintas variables del dataset de parquímetros de Seattle, con el objetivo de ver, exclusivamente, cómo se relacionan los distintos regresores exógenos con la variable de predicción (porcentaje de ocupación de un parquímetro en concreto, a una hora determinada. En el capítulo de EDA ya se han estudiado algunos parámetros similares a lo que se muestra en este capítulo, si bien el objetivo aquí es definir un conjunto de variables para alimentar a los modelos predictivos de capítulos posteriores.

A lo largo de este capítulo se hace referenica en varias ocasiones a la **prueba U de Mann-Whitney**. Esta prueba, entre otras aplicaciones, se utiliza para establecer si hay asociación entre una variable continua (el porcentaje de ocupación en nuestro caso) y una variable binaria. La prueba U de Mann-Whitney es un test no paramétrico (no impone ninguna condición a la distribución de los datos), basado en suma de rangos. En ella, se calcula el estadístico U, definido como:

$$U = \text{mín}(U_1, U_2)$$

U_1 y U_2 son se calculan como sigue:

$$U_i = R_i - \frac{n_i(n_i+1)}{2},\ i \in \{1,2\}\ ,$$

donde n_1 es el número de muestras que corresponden a uno de los dos valores de la variable binaria bajo estudio, mientras que n_2 es el número de muestras restantes. R_i es la suma de los rangos correspondiente a cada muestra.

1.1.1. Relevancia de los intervalos horarios

A continuación, se muestran los resultados de aplicar la prueba U a los 30 Element Key seleccionados para análisis. Para ello, se segmenta el dataset por horas, definiendo la hipótesis nula como sigue:

H_0 : Encontrarse en el rango horario h_i tiene la misma distribución estadística que no encontrarse en él

Los resultados se muestran en las figuras - , e indican que, para un nivel de significación del 99.5 %, la gran mayoría de los intervalos horarios son significativos, aunque la distribución concreta depende de cada Element Key.

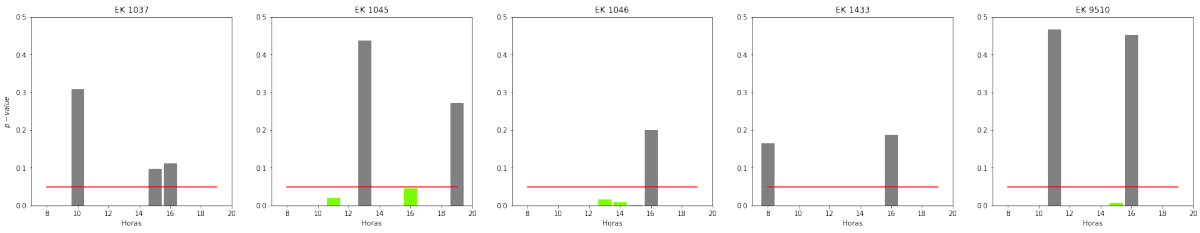

Figura 1.1: Relevancia horas top 30 EK (1 de 6)

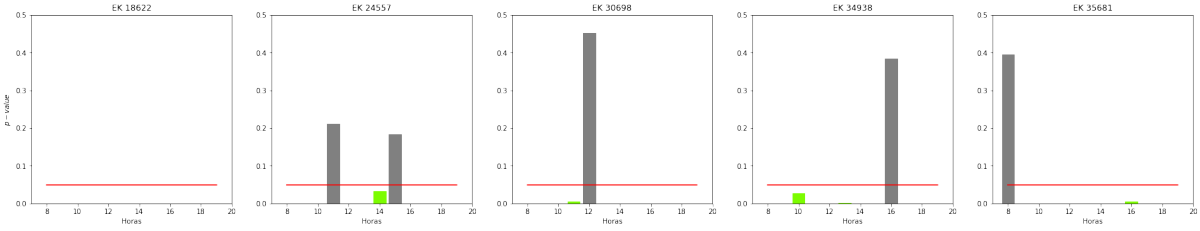

Figura 1.2: Relevancia horas top 30 EK (2 de 6)

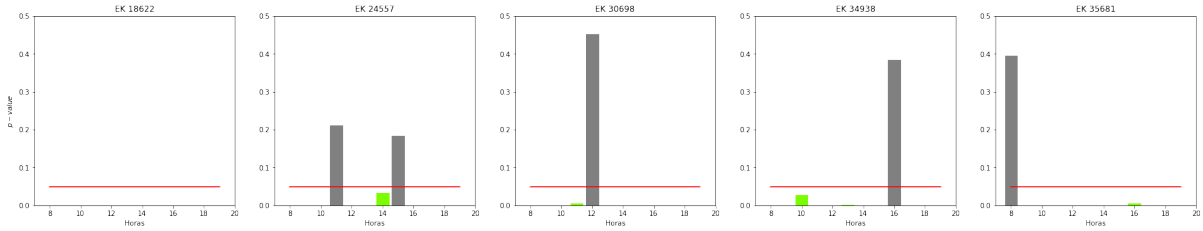

Figura 1.3: Relevancia horas top 30 EK (3 de 6)

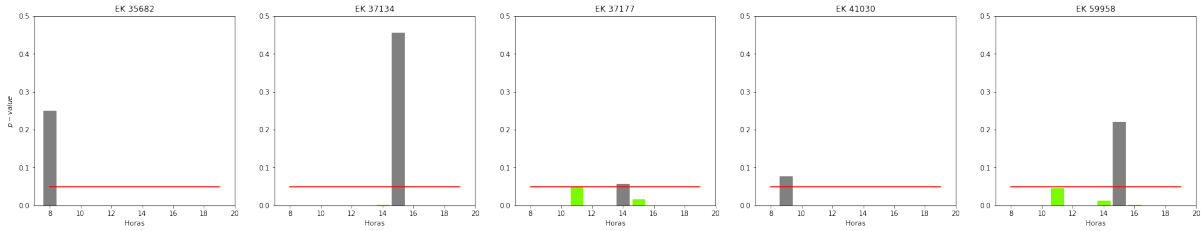

Figura 1.4: Relevancia horas top 30 EK (4 de 6)

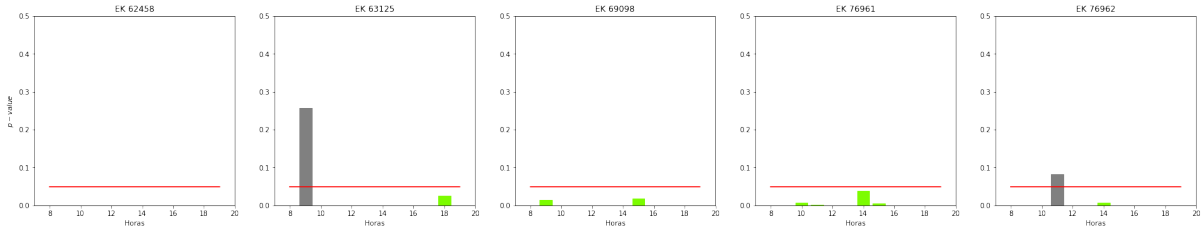

Figura 1.5: Relevancia horas top 30 EK (5 de 6)

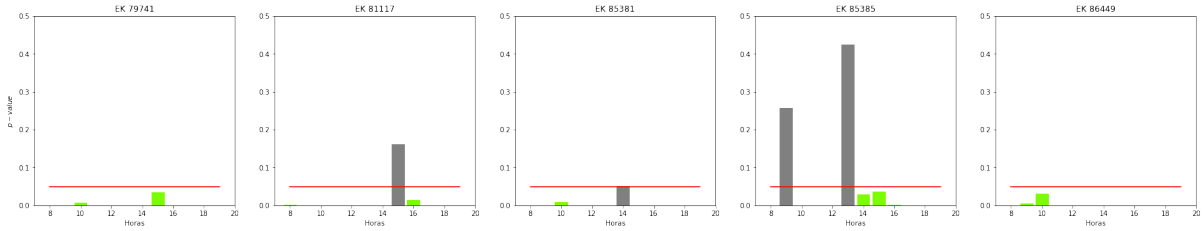

Figura 1.6: Relevancia horas top 30 EK (6 de 6)

1.1.2. Relevancia del día de la semana

En esta sección se lleva a cabo un análisis equivalente al de la sección anterior, pero estudiando la significación estadística del día de la semana en el que se realiza la predicción. Para ello, se utiliza la prueba U de Mann-Whitney, pues estamos comparando de nuevo una variable con varios niveles y una variable continua. En este caso, los intervalos corresponden con el día de la semana: Lunes

(L), Martes (L), miércoles (X), Jueves (J), Viernes (V) y Sábado (S). Los domingos quedan fuera del análisis al ser días no operativos de los parquímetros. Los resultados se presentan en las figuras 1.7 - 1.12. Del mismo modo que en la sección anterior, se presentan gráficos de barras donde se puede observar el nivel de significación asociado a cada día de la semana y a cada EK. El código de colores es idéntico al de la sección anterior.

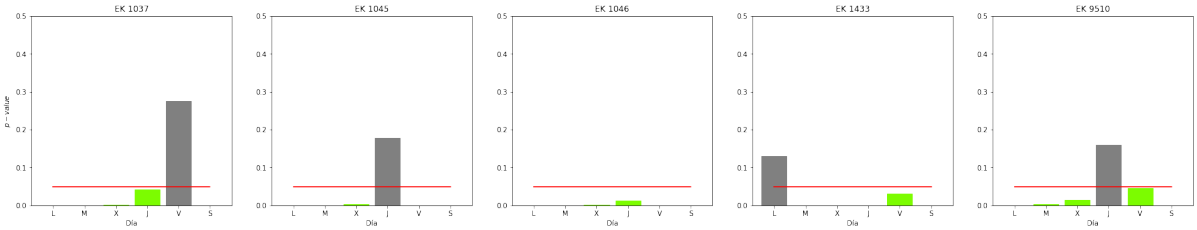

Figura 1.7: Relevancia weekday top 30 EK (1 de 6)

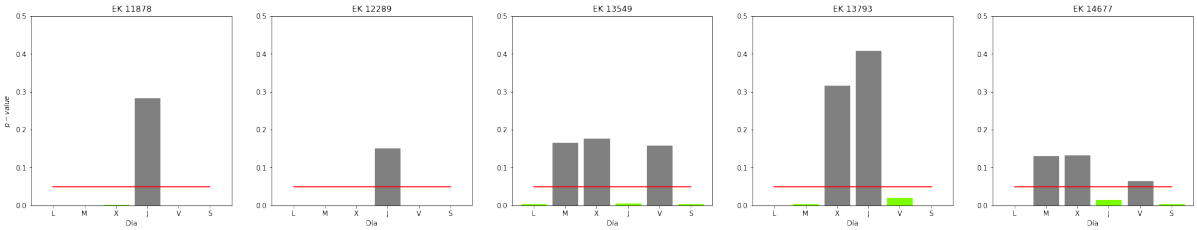

Figura 1.8: Relevancia weekday top 30 EK (2 de 6)

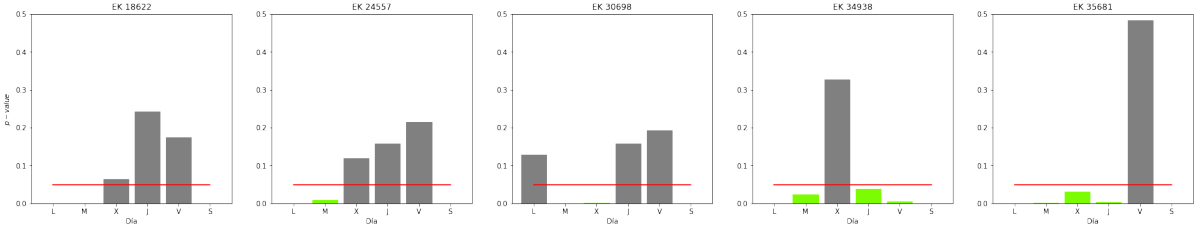

Figura 1.9: Relevancia weekday top 30 EK (3 de 6)

La relevancia del día de la semana es menor que la de la hora del día, pues hay más columnas en gris para el día de la semana. Aun así, los resultados muestran que el día de la semana es relevante para la predicción, aunque menos que la hora del día.

1.1.3. Regresores exógenos

Esta sección analiza, mediante varios métodos, la importancia de los regresores exógenos del dataset. Dado que hay distintos tipos de variables, realizaremos una prueba diferente para cada una de ellas. En concreto:

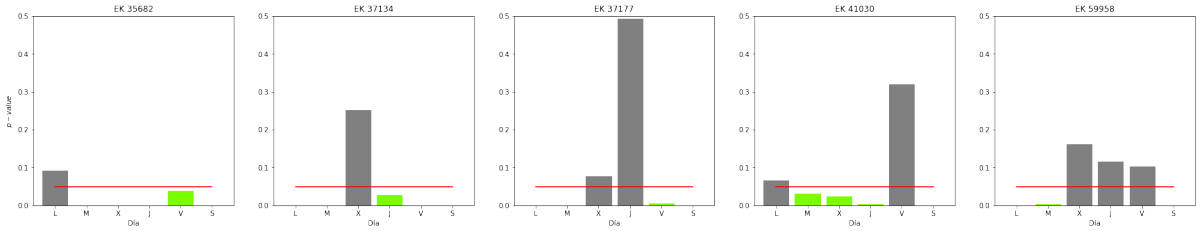

Figura 1.10: Relevancia weekday top 30 EK (4 de 6)

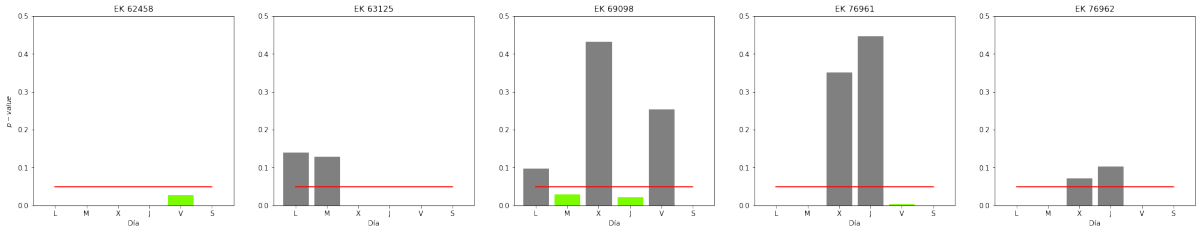

Figura 1.11: Relevancia weekday top 30 EK (5 de 6)

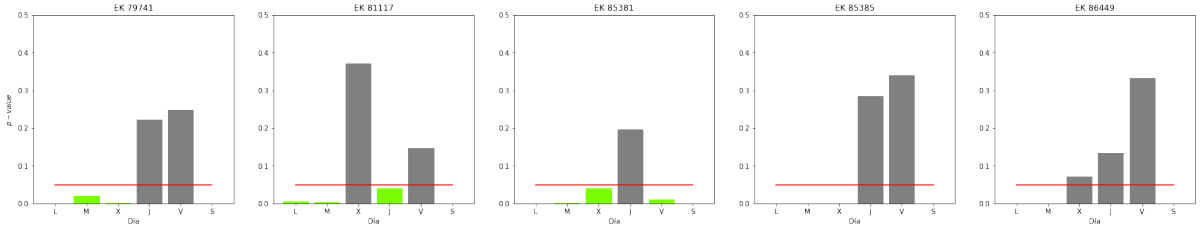

Figura 1.12: Relevancia weekday top 30 EK (6 de 6)

- Variables continuas: se compara una variable continua con el target (variable continua). Test: Correlación de Pearson
- Variables binarias: se compara una variable binaria con el target (variable continua). Test: U Mann-Whitney

Variables continuas

Son lecturas de sensores acerca de las condiciones atmosféricas y de temperatura. Se les aplica un test de correlación de Pearson con respecto al porcentaje de ocupación (target). Utilizaremos las siguientes reglas para calibrar el significado del coeficiente de correlación (ρ):

- $0{,}0 \leq |\,\rho\,| < 0{,}3$: correlación *débil*
- $0{,}3 \leq |\,\rho\,| < 0{,}7$: correlación *moderada*
- $0{,}7 \leq |\,\rho\,| \leq 1{,}0$: correlación *fuerte*

Si la correlación es moderada o fuerte, el signo de ρ nos indicará el sentido de la correlación (directa o inversa). El nivel de significación se establece en $\alpha = 0{,}05$. En la Figura se presenta un gráfico de barras con los resultados del cálculo de la correlación para un Element Key en concreto. Se marcan en gris las variables cuya significación asintótica es menor que el nivel de significación.

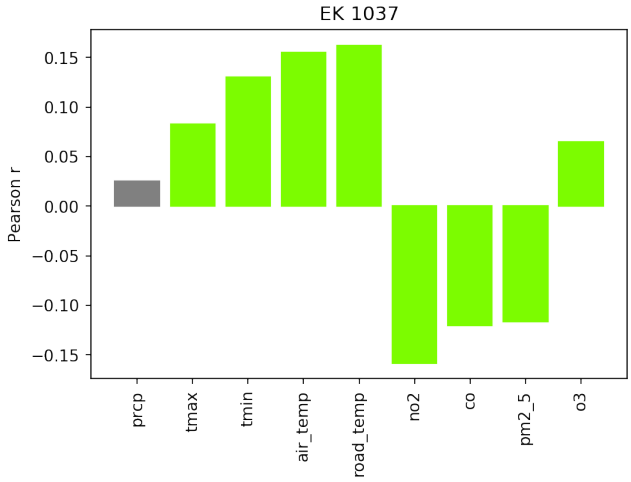

Figura 1.13: Regresores exógenos EK1037

En las Figuras 1.14 y 1.15 se muestra la distribución del coeficiente de correlación de Pearson de cada uno de los regresores exógenos con el porcentaje de ocupación, a lo largo del dataset formado por los 30 EK seleccionados. Además, se representa también su significación asintótica, para calibrar el nivel de significación que tiene el resultado. Como conclusión se establece que no existe una correlación demasiado elevada para ningún regresor exógeno. Esto se debe a la complejidad del fenómeno bajo estudio (distribución del flujo de personas que buscan sitio en un parquímetro). Solo existe un grado de significación compatible con $\alpha = 0{,}05$ para algunas de las variables, que son aquellas para las que, en la Figura 1.15 tienen un pico en la distribución cerca de 0.

1.1.4. Variables binarias

Este análisis únicamente es relevante si se tienen en cuenta varios Element Keys. Si solo se analiza un EK aislado, entonces todas estas variables son constantes y por tanto no aportan información. Por ello, para este cálculo, se consideran todos los EK. Tomando el dataset de Element Keys seleccionados, solo es distinta de cero la variable 'Punto de interés'. Además, esta variable obtiene un nivel de significación del orden de 10^{-53} bajo una prueba U de Mann-Whitney, lo que la convierte en una variable relevante, pero, tal y como se ha comentado, solo en el caso de que se consideren localizaciones geográficas diferentes.

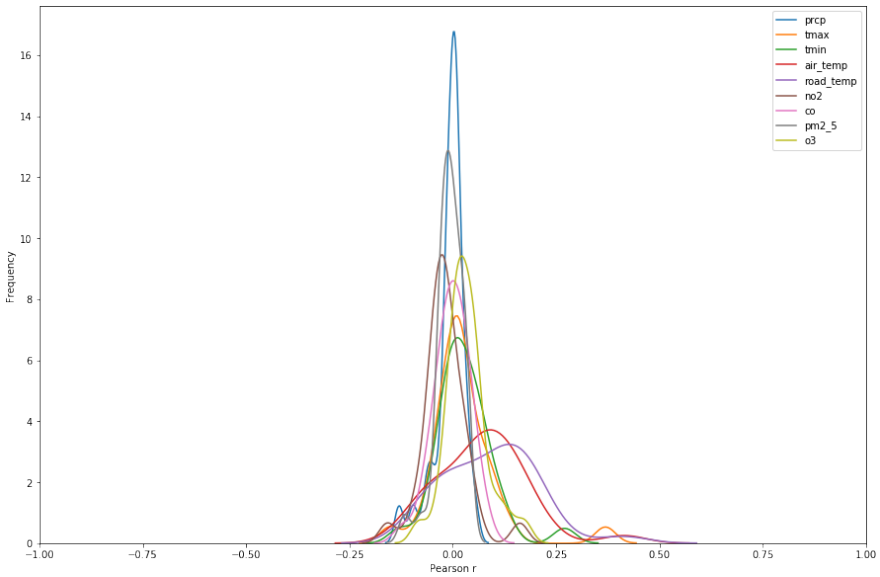

Figura 1.14: Distribución del coeficiente de correlación de Pearson en el top 30

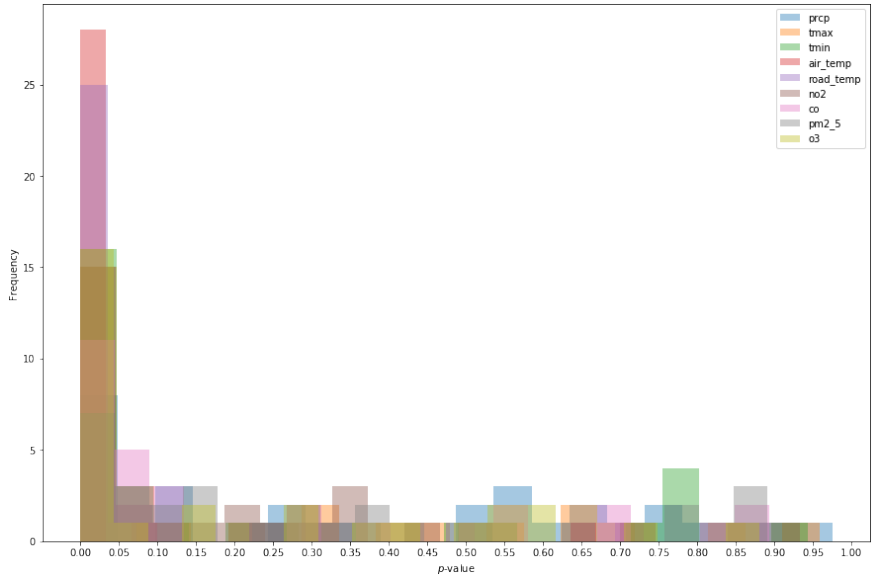

Figura 1.15: Significación asintótica de la correlación del top 30

1.2. Importancia de variables basada en métodos de Machine Learning

En esta sección se presenta un análisis de la importancia de las variables según métodos de Machine Learning. Para ello, se utiliza el concepto de **importancia de permutación**. Este concepto

consiste en evaluar la diferencia que existe en la métrica utilizada para el análisis (AUC, precisión, ...) cuando se elimina una de las variables del dataset. Tras llevar a cabo este análisis, se obtienen diferentes pesos para cada una de las variables, clasificándose en términos de su importancia. Debe tenerse cuidado con este método para evaluar la importancia pues, en primer lugar, necesita de un modelo entrenado que dé buenos resultados y, en segundo lugar, es vulnerable a la multicolinealidad.

Por otro lado, es de crucial relevancia mecionar que la interpretación de los resultados de este análisis debe realizarse con cuidado. Esto se debe a que la estructura de los datos bajo análisis es de tipo serie espacio-temporal. Es decir, cada registro está relacionado con el anterior. Por lo tanto, al entrenar los modelos de ML, y separarlos en entrenamiento y validación, se va a romper la estructura de serie de los datos, con lo que el rendimiento del modelo puede ser bajo. Con ello, los resultados obtenidos no van a ser generalizables en sentido estricto, pero sí se va a obtener una medida de cómo son de relevantes las variables, similar a una correlación, pero basándonos en modelos más complejos que el coeficiente de Pearson. Para evaluar la importancia de permutación, se ha utilizado la implementación de ELI5, que dispone de un *wrapper* para los modelos entrenados con *sklearn*. Se han entrenado dos modelos, basados Random Forest y Gradient Boosting respectivamente.

1.2.1. Entrenamiento de los modelos

El entrenamiento de los modelos se ha realizado según los siguientes pasos:

1. Discretización de la variable target (porcentaje de ocupación) en 5 intervalos, para que encaje con los modelos de clasificación aceptados por la implementación de ELI5. El número de intervalos se ha obtenido de un compromiso entre rendimiento del modelo y capacidad de cómputo.

2. Discretización de la variable (latitud, longitud) en 25 cuadrantes.

3. Eliminación del identificador de cada EK.

4. Separación en entrenamiento y validación.

5. Fitting.

6. Análisis de resultados.

El modelo basado en Random Forest se ha entrenado con 75 estimadores, sin limitación en la profundidad de los nodos ni en su número de hojas. Del mismo modo, el modelo basado en Gradient Boosting utiliza también 75 árboles, sin penalizar la profundidad de los mismos. Aunque el hecho de no limitar la profundidad de los árboles puede llevar a sobreajustar el conjunto de datos, en este caso se ha realizado así para aumentar la capacidad predictiva de ambos modelos. Los resultados se evalúan mediante el ratio de verdaderos positivos (TPR) y el de falsos negativos (FPR), presentándolos en la curva ROC, y calculando el área bajo la curva (AUC), en la Figura 1.16. Los dos modelos tienen, en concreto, el valor de AUC que se muestra en la tabla 1.1.

Se observa que el modelo basado en Random Forest es algo superior al de Gradient Boosting, con los parámetros seleccionados para su entrenamiento. Sin embargo, los dos modelos tienen dificultades para superar el valor de 0.5, que correspondería a una clasificación aleatoria. Esto se debe a lo que se comentó previamente acerca de la estructura de serie de los datos.

Modelo	AUC
Random Forest	0.57
Gradient Boosting	0.54

Tabla 1.1: AUC modelos Permutation Importance

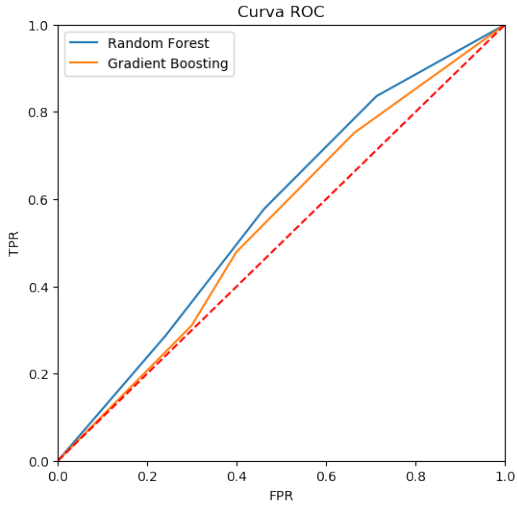

Figura 1.16: Validación de los modelos utilizados para la importancia de permutación

1.2.2. Permutation Importance

Para concluir esta sección, se muestran los pesos de las variables más relevantes según los dos modelos estudiados. Estos resultados se presentan en la Figura 1.17.

Weight	Feature
0.0412 ± 0.0027	road_temp
0.0200 ± 0.0013	air_temp
0.0179 ± 0.0007	poi
0.0175 ± 0.0006	longitude_0
0.0125 ± 0.0011	longitude_2
0.0107 ± 0.0007	latitude_5
0.0103 ± 0.0009	latitude_1
0.0095 ± 0.0016	longitude_1
0.0065 ± 0.0021	tmax
0.0050 ± 0.0004	latitude_0
0.0036 ± 0.0005	latitude_3
0.0022 ± 0.0016	longitude_5
0.0019 ± 0.0011	latitude_2
0.0013 ± 0.0013	latitude_4
0.0011 ± 0.0005	longitude_4
0.0003 ± 0.0013	longitude_3
0 ± 0.0000	event
0 ± 0.0000	soccer
-0.0001 ± 0.0001	basket
-0.0002 ± 0.0001	baseball

Random Forest

Weight	Feature
0.0656 ± 0.0025	road_temp
0.0337 ± 0.0024	tmax
0.0239 ± 0.0011	poi
0.0224 ± 0.0015	latitude_1
0.0192 ± 0.0009	longitude_0
0.0180 ± 0.0017	longitude_2
0.0167 ± 0.0014	longitude_1
0.0116 ± 0.0008	longitude_5
0.0114 ± 0.0007	latitude_0
0.0106 ± 0.0008	latitude_5
0.0062 ± 0.0006	latitude_4
0.0061 ± 0.0008	latitude_2
0.0047 ± 0.0009	tmin
0.0042 ± 0.0004	longitude_3
0.0036 ± 0.0006	air_temp
0.0030 ± 0.0012	latitude_3
0.0022 ± 0.0012	no2
0.0014 ± 0.0004	longitude_4
0.0012 ± 0.0004	o3
0.0009 ± 0.0004	prcp

Gradient Boosting

Figura 1.17: Pesos importancia de permutación