GRADO EN INGENIERÍA DE ORGANIZACIÓN INDUSTRIAL

ESCUELA DE INGENIERÍAS INDUSTRIALES
UNIVERSIDAD DE MÁLAGA

TRABAJO FIN DE GRADO

Aplicación de técnicas de minería de datos sobre registros de operaciones del sistema de alquiler público de bicicletas de Madrid (BiciMAD)

MAXIMILIANO GÁMEZ LÓPEZ

JUNIO 2020

Tutor Dr. Juan Carlos Rubio Romero

Cotutora Dra. María Martínez Rojas

Área de conocimiento Organización de empresas
Departamento de Economía y administración de empresas

UNIVERSIDAD DE MÁLAGA

ESCUELA DE INGENIERÍAS INDUSTRIALES

Contenido:

Documentación disponible en:

https://github.com/maxgamezlopez/TFG

UNIVERSIDAD DE MÁLAGA

ESCUELA DE INGENIERÍAS INDUSTRIALES

1. Introducción

- La presencia de tecnologías de información y la transformación digital juegan un **papel esencial en las organizaciones y sus procesos**. El uso de infraestructuras interconectadas y con mayor presencia de sensores supone la disponibilidad de grandes conjuntos de datos (*«Big Data»*).

Diagrama de jerarquía del conocimiento o cadena DIKW. Elaboración propia a partir de [1]

- Cada vez más ventajas competitivas en las organizaciones se basan en la jerarquía del conocimiento. Se asocia a conceptos como la inteligencia de negocios o *«Business Intelligence»*.

- El proceso de **minería de datos consiste en la exploración e identificación de patrones ocultos en grandes conjuntos de datos**. El conjunto de técnicas pueden integrarse en algoritmos que conformen máquinas de aprendizaje o *«Machine Learning»*. Cuando existe toma de decisiones con base en la percepción del entorno, se puede hablar de inteligencia artificial.

- Algunos **ejemplos de aplicación**: diagnóstico temprano de enfermedades, recomendaciones de compra, reconocimiento facial o sistemas avanzados de mantenimiento preventivo son algunos ejemplos de aplicación.

UNIVERSIDAD DE MÁLAGA

2. Objetivo y alcance

ESCUELA DE INGENIERÍAS INDUSTRIALES

Objetivo:

Aplicación de técnicas de minerías de datos con fin de analizar registros de datos de uso del servicio BiciMAD (abril 2017 – junio 2019), así como de obtener información de utilidad para su operación.

Queda dentro del alcance:

- Introducción y contextualización.
- Adquisición, preparación y procesamiento de datos.
- Aplicación de metodología, herramientas y técnicas propias de la minería de datos.
- Justificación de metodología empleada.
- Obtención de un modelo predictivo de la demanda del servicio BiciMAD.
- Definición de líneas de trabajo y de mejora complementarias a la desarrollada.

Queda fuera del alcance:

- Justificación y desarrollo matemático de las herramientas empleadas.
- Optimización del rendimiento de algoritmos a la arquitectura computacional disponible.
- Análisis y cálculo de costes asociados.
- Estudio de viabilidad y de recursos necesarios para la implantación del modelo.
- Integración del modelo resultante.

UNIVERSIDAD DE MÁLAGA

ESCUELA DE INGENIERÍAS INDUSTRIALES

3. Minería de datos

La **minería de datos se integra** como parte del **proceso de obtención del conocimiento** de bases de datos.

Este comprende etapas desde el acceso a los datos hasta la obtención de conocimiento útil.

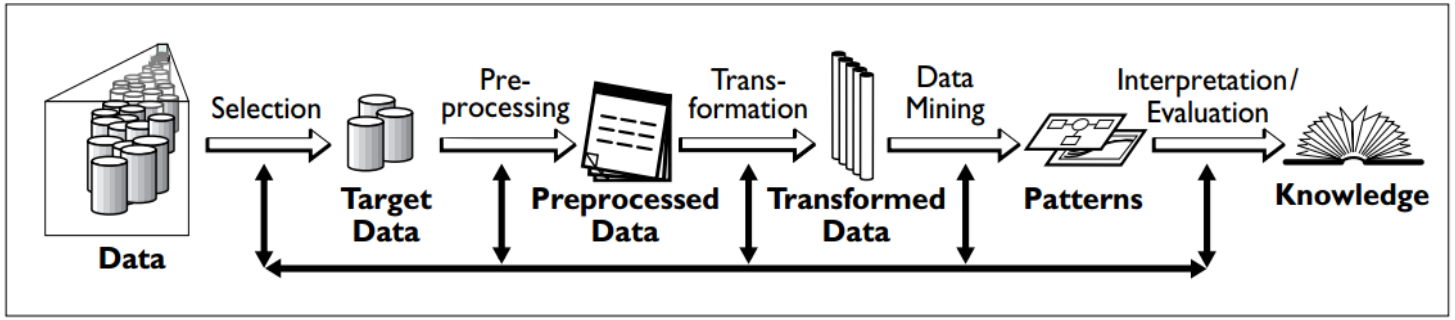


Diagrama de obtención de conocimiento de datos o KDD. Fuente: [2]

Los datos empleados pueden proceder de registros almacenados o flujo de datos a tiempo real. Sistemas ERP, BBDD, telemetría, lectura de sensores o comunicaciones entre máquinas (M2M) sirven como ejemplos.

Los grandes conjuntos de datos quedan caracterizados por las denominadas «siete uves»:

Volumen **V**elocidad **V**ariedad **V**alor

Variabilidad **V**eracidad **V**olatilidad

UNIVERSIDAD DE MÁLAGA

ESCUELA DE INGENIERÍAS INDUSTRIALES

3. Minería de datos

Los datos pueden contenerse de forma:

```
route_id,route_short_name,route_long_name,route_type,route_url
1,1,Parque del Sur - Alameda Principal - San Andrés,3,http://www.emtmalaga.es/emt-mobile/informacionLinea.html?codLinea=1
2,2,Alameda Principal - Ciudad Jardín,3,http://www.emtmalaga.es/emt-mobile/informacionLinea.html?codLinea=2
3,3,Puerta Blanca - Alameda Principal - El Palo (Olias),3,http://www.emtmalaga.es/emt-mobile/informacionLinea.html?codLinea=3
4,4,Paseo del Parque - Cruz Humilladero - Cortijo Alto,3,http://www.emtmalaga.es/emt-mobile/informacionLinea.html?codLinea=4
5,5,Alameda Principal - Guadalmar - Parque de Ocio,3,http://www.emtmalaga.es/emt-mobile/informacionLinea.html?codLinea=5
7,7,Parque Litoral - Alameda Principal - Carlinda,3,http://www.emtmalaga.es/emt-mobile/informacionLinea.html?codLinea=7
8,8,Alameda Principal - Colonia Santa Inés - Clínico,3,http://www.emtmalaga.es/emt-mobile/informacionLinea.html?codLinea=8
```

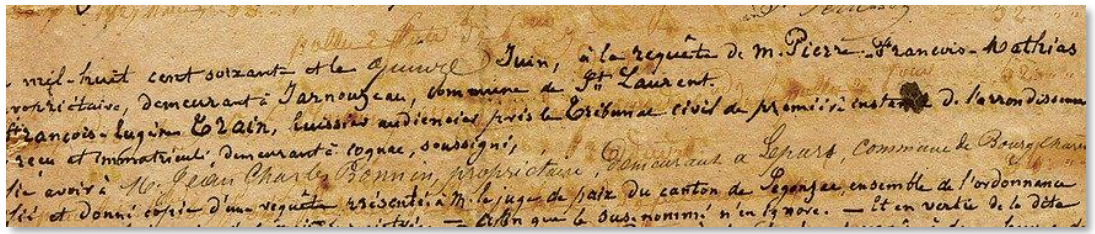
Estructurada

Datos estructurados en formato CSV correspondiente datos GTFS de EMT Málaga. Fuente: [3]

```
"source": {
    "type": "URL",
    "value": "https://datosabiertos.malaga.eu/recursos/aparcamientos/ocupappublicosmun/ocupappublicosmunfiware.json"
},
"totalSpotNumber": {
    "type": "Integer",
    "value": 344
},
"location": {
    "type": "geo:json",
    "value": {
        "type": "Point",
        "coordinates": [
            -4.4165168,
            36.7224312
        ]
```

Semiestructurada

Datos semiestructurados en formato JSON correspondiente ocupación de aparcamiento públicos del Ayuntamiento de Málaga. Fuente: [3]



No estructurada

Datos estructurados en formato manuscrito. Fuente: [4]

UNIVERSIDAD DE MÁLAGA

ESCUELA DE INGENIERÍAS INDUSTRIALES

3. Minería de datos

Secuencia de procesos para el análisis de datos almacenados:

Diagrama de secuencia de procesos para el análisis de datos almacenados o estáticos. . Fuente: [5]

Modelos de aprendizaje:

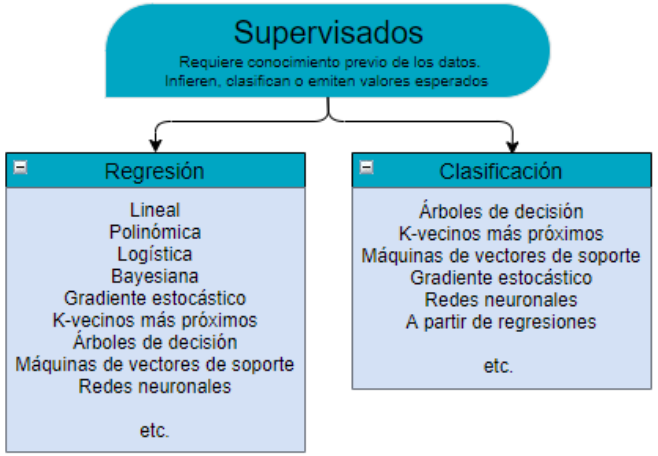


Diagrama de modelos de aprendizaje supervisados y no supervisados, ejemplos de usos y de técnicas empleadas. Fuente: elaboración propia.

Este TFG ha hecho uso de herramientas disponibles en la librería *Sklearn*, combinada con otras librerías de *Python* como *Pandas* o *Matplotlib*.

scikit learn

python

scikit-learn.org

UNIVERSIDAD DE MÁLAGA

3. Minería de datos

ESCUELA DE INGENIERÍAS INDUSTRIALES

Modelo supervisado: predice el valor de una variable objetivo (Y) a partir de variables atributos (X).

	_id	year	month	day	weekday_int	meteo_ tmed	prec	tmin	tmax	dir	velmedia	racha	presMax	presMin
0	Y	X												
1														
2														
3														
…														
808														

Esquema explicativo sobre la segregación del conjunto de datos de este TFG en variable objetivo (Y) y variables de entrenamiento (X) Fuente: elaboración propia.

Para evitar sobreajuste del modelo, la metodología contempla la validación cruzada:

Evaluación y validación del modelo con registros diferentes a empleados para el desarrollo el modelo.

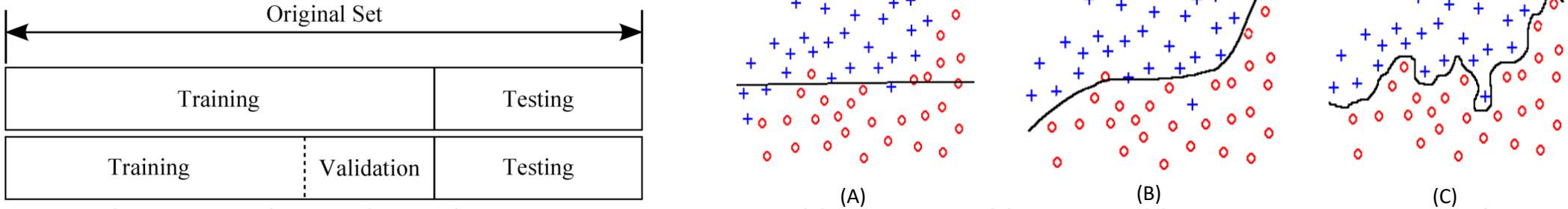


Diagrama de técnica de validación cruzada (izquierda) y diagrama que representa subajuste (A), ajuste adecuado (B) y sobreajuste (C). Fuente: recursos originales de Prof. Kichun Lee [6]

UNIVERSIDAD DE MÁLAGA

ESCUELA DE INGENIERÍAS INDUSTRIALES

3. Minería de datos

Los modelos supervisados y no supervisados pueden trabajar de forma conjunta.

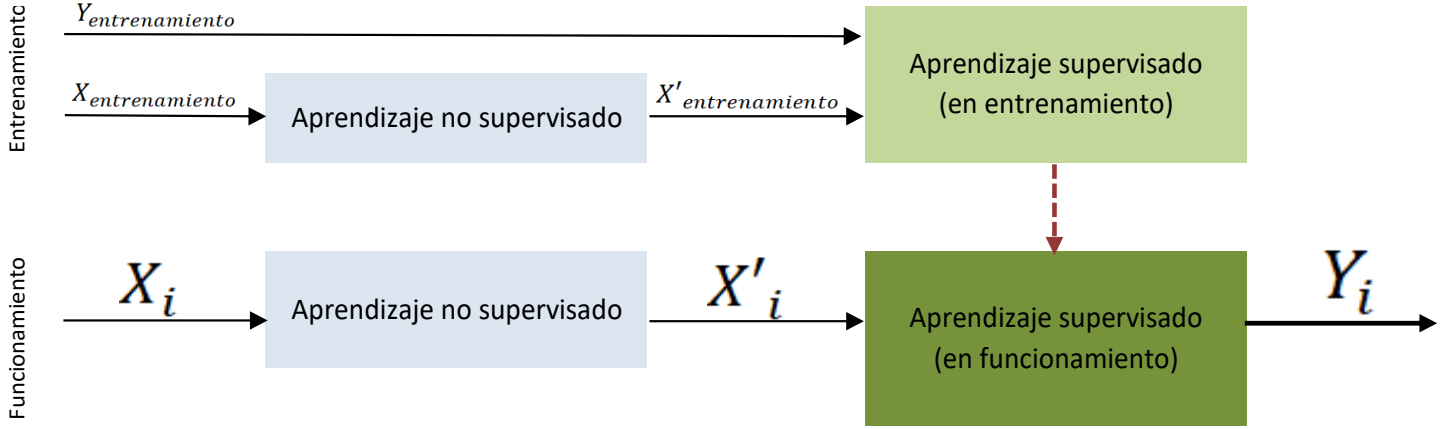
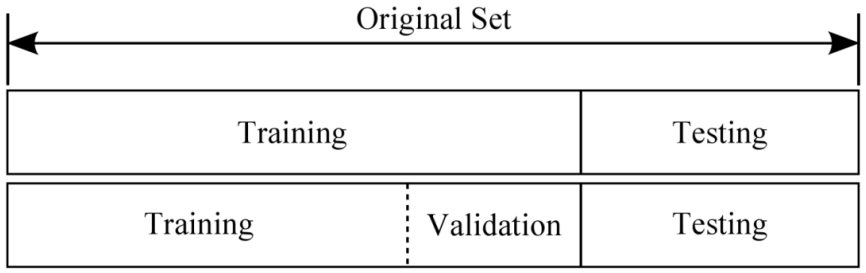


Diagrama de configuración de modelos supervisados y no supervisados ensayada en este TFG. Fuente: elaboración propia.

Para evitar sobreajuste del modelo, la metodología contempla la validación cruzada:

Evaluación y validación del modelo con registros diferentes a empleados para el desarrollo el modelo.



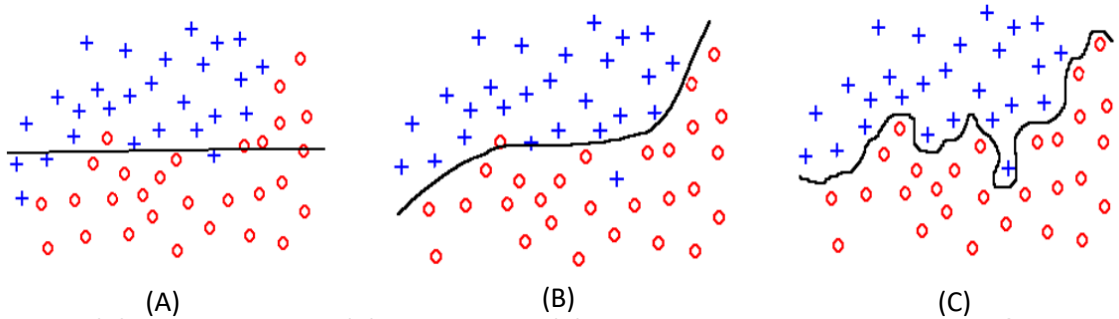


Diagrama de técnica de validación cruzada (izquierda) y diagrama que representa subajuste (A), ajuste adecuado (B) y sobreajuste (C). Fuente: recursos originales de Prof. Kichun Lee [6]

UNIVERSIDAD DE MÁLAGA

ESCUELA DE INGENIERÍAS INDUSTRIALES

4. CRISP-DM

Modelo de estándar abierto para la aplicación de minería de datos en procesos industriales:

- Creado en 1996 como parte del proyecto (ESPRIT). de la Unión Europea.

- Modelo más extendido a día de hoy.

- Carácter cíclico e iterativo, compatible PDCA.

- Etapas interdependientes:
  - Compresión del negocio / procesos.
  - Comprensión de los datos.
  - Preparación de los datos.
  - Confección de un modelo.
  - Evaluación del modelo.
  - Implementación del modelo.

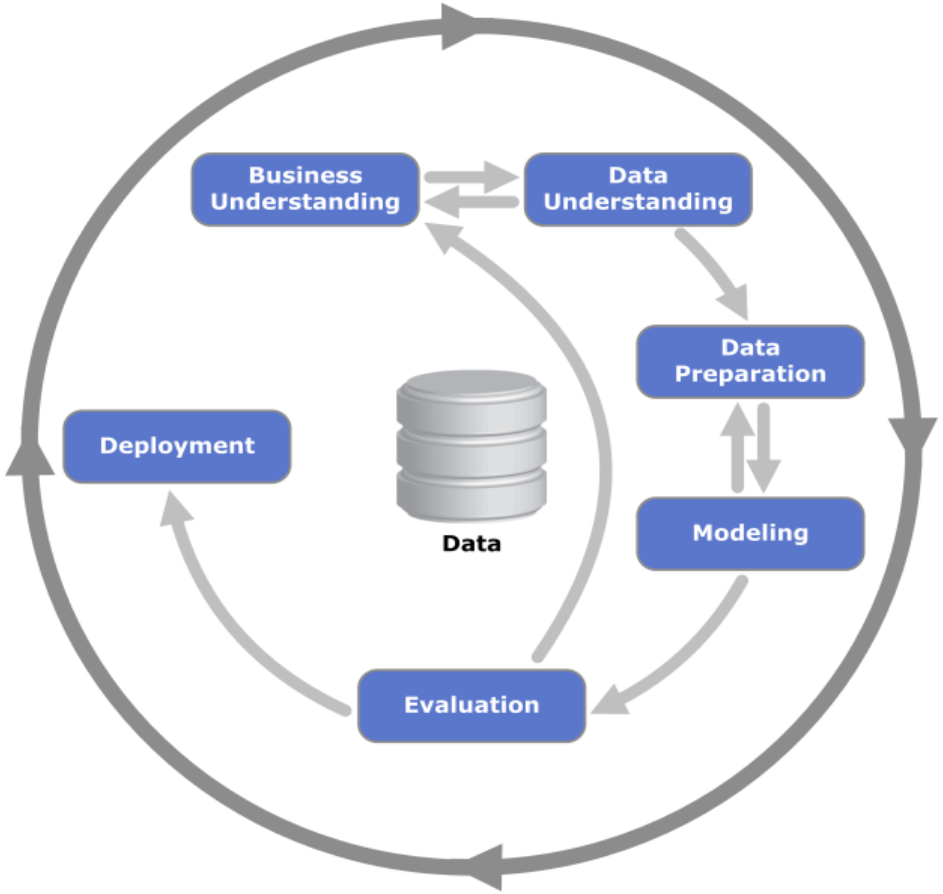


Diagrama de proceso descrito en estándar CRISP-DM. Fuente: Recurso de K. Jensen a partir de [7]

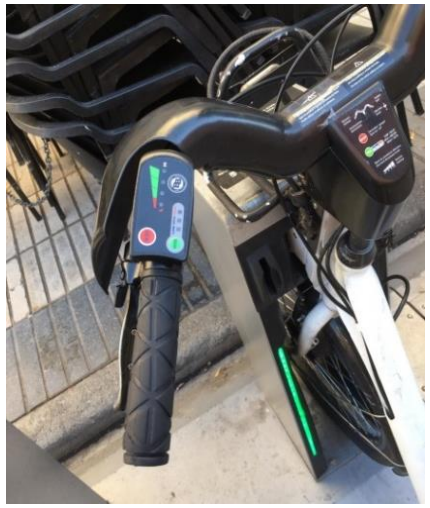
UNIVERSIDAD DE MÁLAGA

5. Servicio BiciMAD

ESCUELA DE INGENIERÍAS INDUSTRIALES

La comprensión del negocio y los procesos se hace necesario el conocer de la naturaleza y funcionamiento del servicio de BiciMAD.

- Sistema de alquiler de bicicleta compartida de pago por tiempo de uso.
- Creado en 2014 por el Ayuntamiento de Madrid. Gestionado por EMT Madrid.
- Retirada y depósito de bicicletas en estaciones de la red BiciMAD.
- Extensión de la red:
  165 estaciones
  4116 estacionamientos.
  2028 bicicletas eléctricas
- La red requiere un equilibrio que garantice disponibilidad de bicicletas en estaciones de origen y de estacionamientos libres en el destino.


Fotografías de la estación 145 ubicada en Calle Ortega y Gasset 87.
Fuente: elaboración de terceros con cesión de derechos.

UNIVERSIDAD DE MÁLAGA

5. Servicio BiciMAD

ESCUELA DE INGENIERÍAS INDUSTRIALES

EMT Madrid ofrece registros de operaciones en su web de datos abiertos en formato JSON:

- Registros de desplazamientos:

  Datos agregados sobre desplazamientos producidos, fecha y hora, base de estacionamiento de origen y destino, duración del trayecto, trayecto empleado, tipo de usuario (ocasional o abonado), rango de edad del usuario, etc.

  Datos M2M provenientes de bicicletas y estaciones.

- Situación de estaciones:

  Registros, por horas, de estado y nivel de ocupación de las estaciones. M2M provenientes de estaciones.

UNIVERSIDAD DE MÁLAGA

ESCUELA DE INGENIERÍAS INDUSTRIALES

6. Adquisición y procesamiento de datos.

Se decide complementar los registros de BiciMAD con datos de otras fuentes:

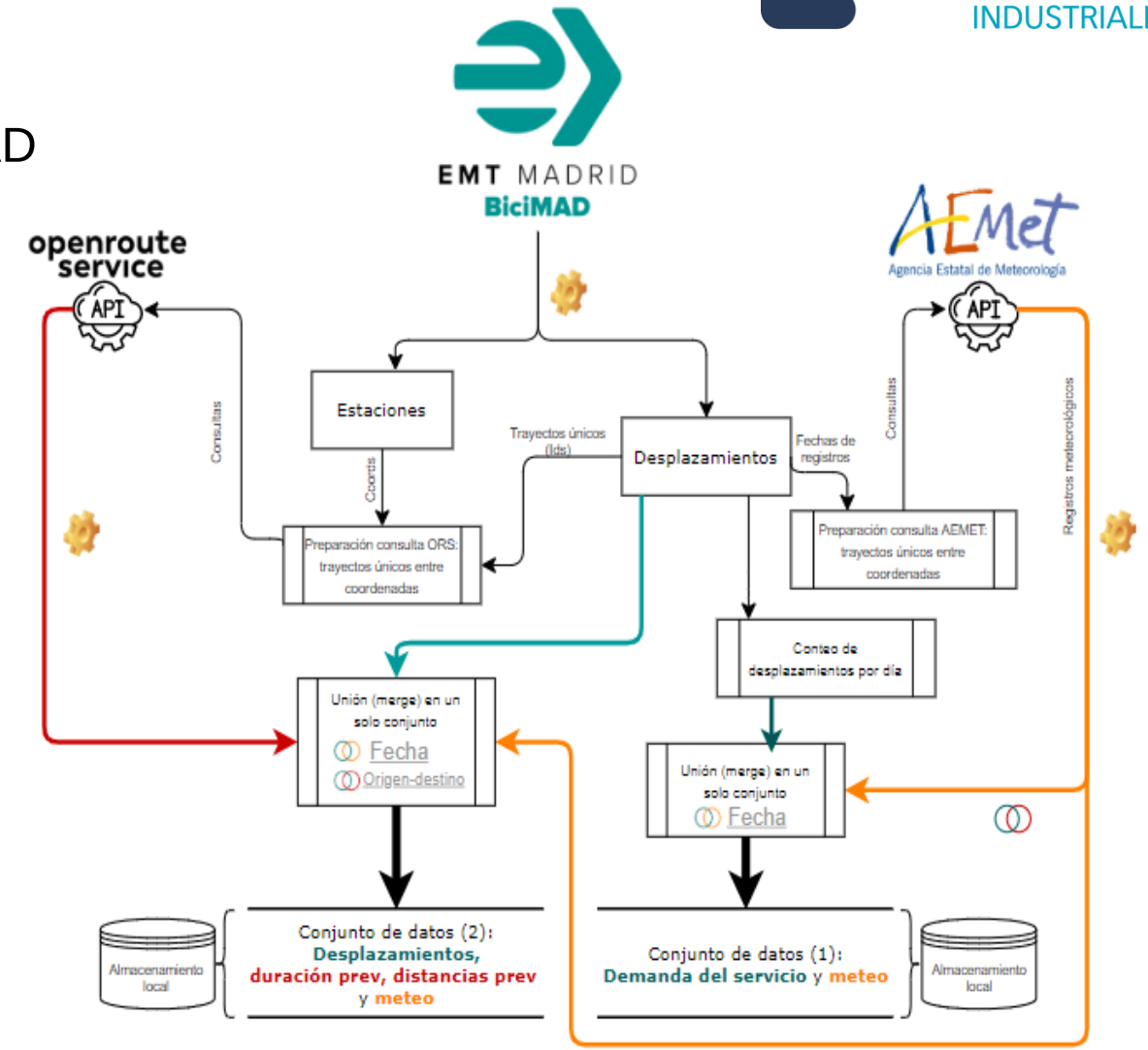
- Distancia y duración prevista para cada desplazamiento.
  (API navegación *Openroute Service)*

- Meteorología registrada por fechas.
  (API AEMET).

Tras convertir, filtrar, procesar y unificar los datos, se obtienen dos conjuntos de datos en formato CSV:

- Desplazamientos, duración y distancia prevista y meteo.
  (9 millones de registros)

- Número de registros por días (demanda) y meteo.
  (829 registros)

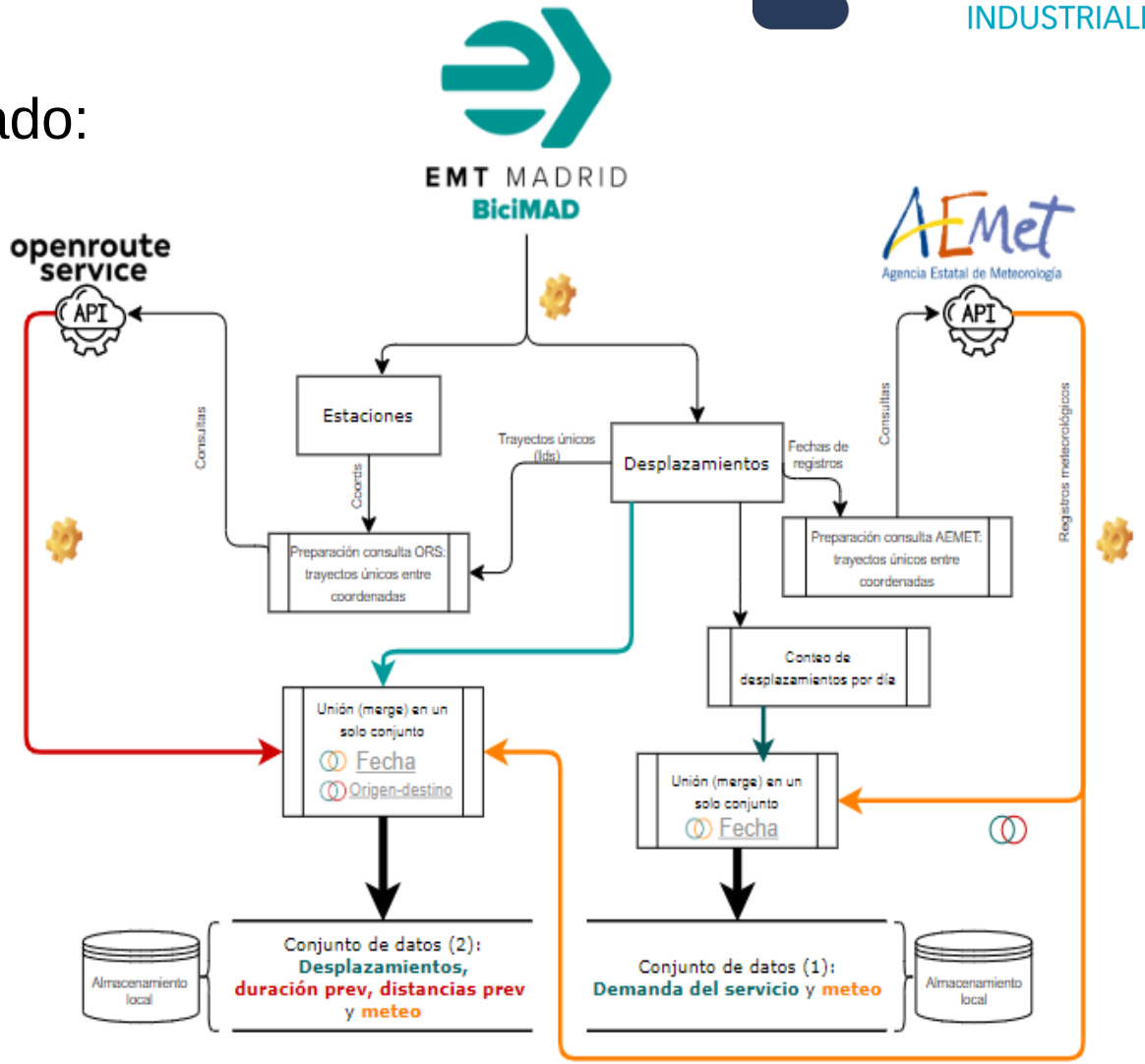


Esquema del proceso de adquisición de datos de las fuentes Openrouteservice, BiciMAD y AEMET. Fuente: elaboración propia.

6. Adquisición y procesamiento de datos.

El procesamiento de los datos ha conllevado:

- Creación de consultas a APIs.
- Conversión de JSON a CSV.
- Filtrado y subsanación de registros anómalos.
- Creación de variables auxiliares
- Corrección de separadores decimales.
- Selección de variables de interés.
- Imputación de valores a registros nulos.
- Unificación de datos en un solo conjunto.
- Almacenamiento en local.



Esquema del proceso de adquisición de datos de las fuentes Openrouteservice, BiciMAD y AEMET. Fuente: elaboración propia.

UNIVERSIDAD DE MÁLAGA

6. Adquisición y procesamiento de datos.

ESCUELA DE INGENIERÍAS INDUSTRIALES

De los dos conjuntos de datos obtenidos se encuentran **multitud de aplicaciones** y líneas de trabajo de utilidad para la operación del servicio:

1. Desplazamientos, duración y distancia prevista y meteo: clasificación de tipo de usuarios y/o estaciones, predicción de la duración/distancia, fator de distancia y tiempo reales frente a previstos,…
2. Número de registros por días (demanda) y meteo: pronóstico de la demanda con base en la meteorología.

Se opta por el desarrollo de la segunda línea, correspondiente al pronóstico del demanda del servicio con base en los registros meteorológicos. Se enuncia la siguiente hipótesis:

« H_1 - La meteorología y determinados factores de estacionalidad influyen en la demanda experimentada por el sistema BiciMAD »

UNIVERSIDAD DE MÁLAGA

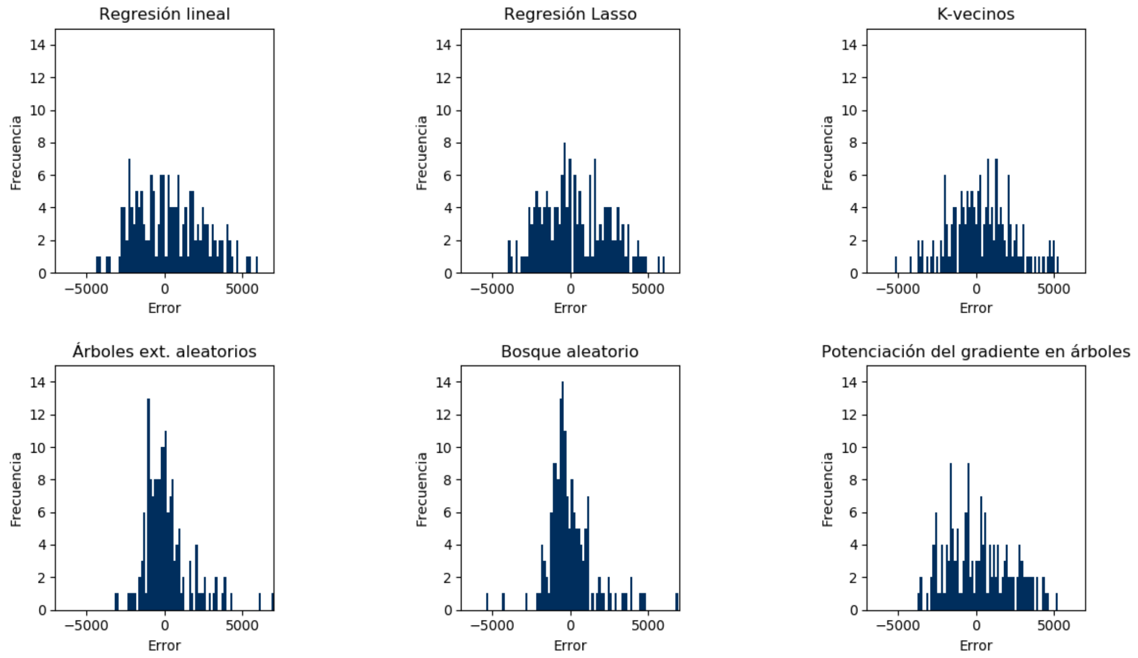
ESCUELA DE INGENIERÍAS INDUSTRIALES

7. Selección de modelo.

Tras un análisis descriptivo de los datos, se ha procedido ensayar diferentes técnicas de regresión para evaluar, a través del coeficiente de determinación R^2 y el error absoluto, su idoneidad en la predicción del número de desplazamientos por día con base al resto de variables.

Validación cruzada: 80% entrenamiento 20% evaluación

$$Error\ Absoluto = Valor\ predicho - Valor\ registrado$$



Comparativa de errores absolutos obtenidos con diferentes modelos. Fuente: elaboración propia.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

y_i : valor registrado

$\bar{y}$: promedio registros

$\hat{y}_i$: valor predicho

```
>>> R2.head()
                   0              1         2              3         4  \
0  Linear regression  LogRegression     Ridge  BayesianRidge     Lasso
1           0.389673       -0.26143  0.378486       0.372885  0.374911

            5                           6                             7  \
0  ElasticNet  PassiveAggressiveRegressor  Supported Vector Machine Reg
1    0.222474                   -0.187036                   -0.00192486

          8              9           10        11        12
0      K-nn  Random Forest  Extra trees       GPR       GBR
1  0.440013       0.664845     0.685455  0.193884  0.461068
>>>
```

Comparativa de coeficientes de determinación obtenidos con diferentes modelos. Valores negativos representan una determinación peor que la aleatoriedad.
Fuente: elaboración propia.

Idealmente, el modelo debe registrar:

- R^2 próximo a +1
- Distribución de error absoluto:
  - simétrica
  - centrada en 0
  - mínima dispersión

7. Selección de modelo.

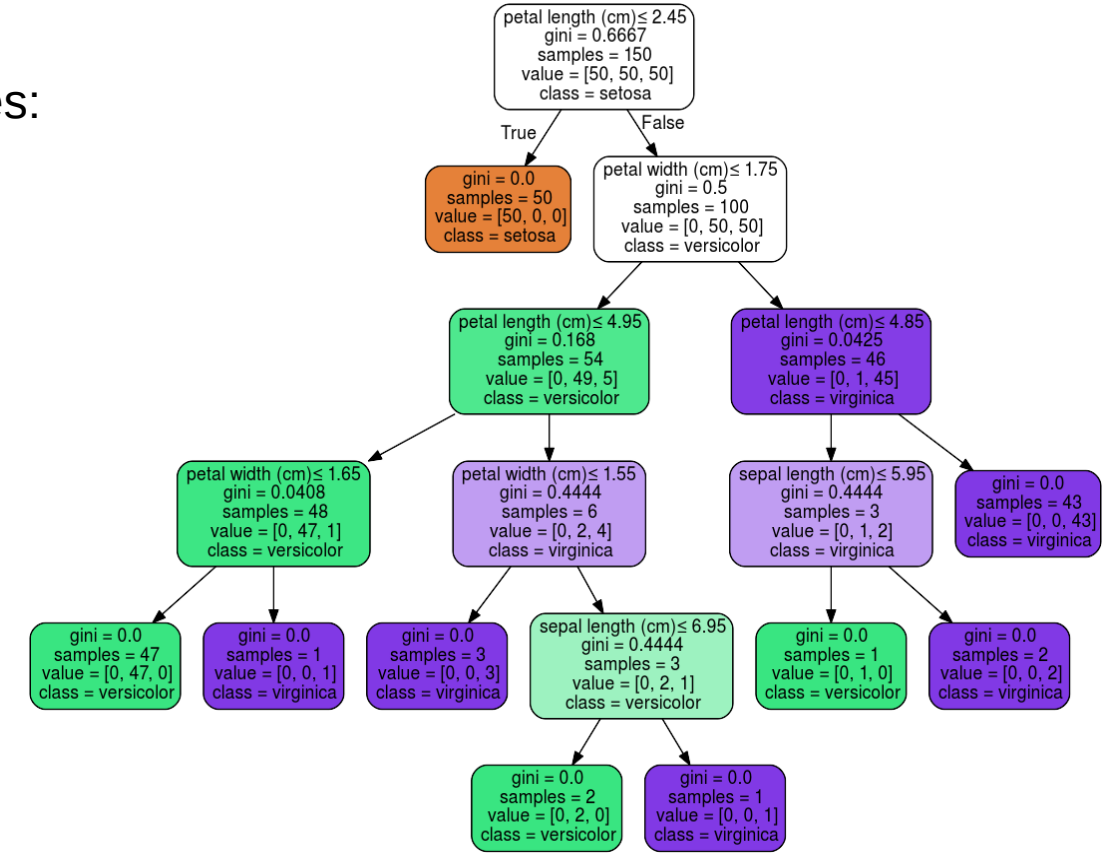
De los modelos probados, **«bosque aleatorio»** (*Random Forest*) y **«árboles extremadamente aleatorios»** (*Extra trees*) son los que mejores resultados obtienen. Se desestima el uso de PCA.

Ambos modelos resultan parecidos y con muchas similitudes:

- Modelos no lineales y no paramétricos.
- Aprendizaje en conjunto (*ensemble learning*) basado en árboles de decisión.
- Registros agrupados por nodos y clasificados mediante condiciones lógicas.
- Compensación de sesgo y varianza construyendo multitud de árboles (bosque) de decisión. Se considera como salida el resultado más frecuente (clasificación) o promedio de los resultados (regresión).

Principal diferencia:

- Bosque aleatorio define límites donde mayor caída de heterogeneidad se produce en cada clasificación.
- Árboles extremadamente aleatorios los establece arbitrariamente.



Representación gráfica de árbol de clasificación de especímenes de flor iris.
Fuente: scikit-learn.org [8].

7. Selección de modelo.

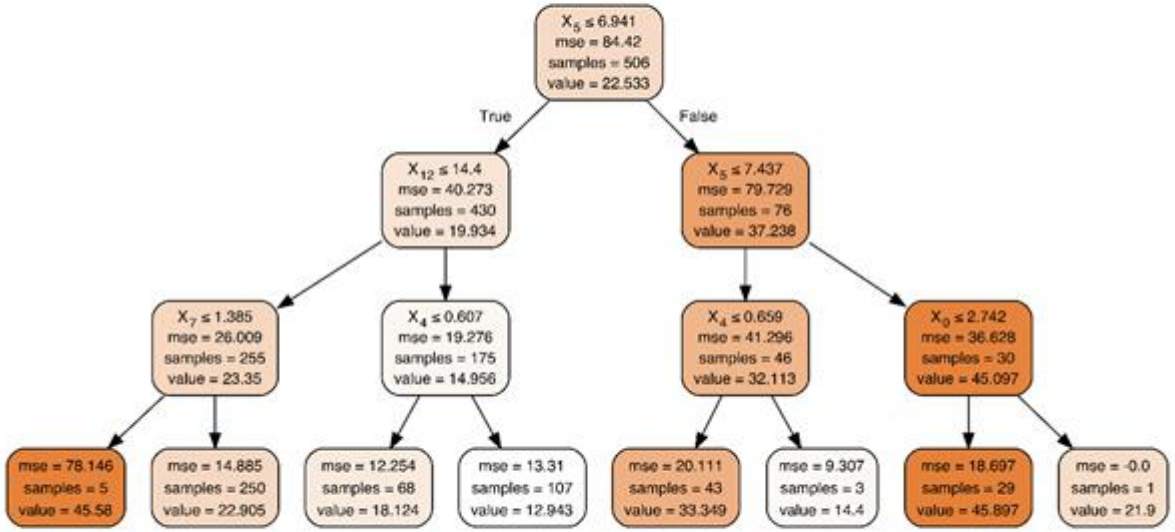
De los modelos probados, **«bosque aleatorio»** (*Random Forest*) y **«árboles extremadamente aleatorios»** (*Extra trees*) son los que mejores resultados obtienen. Se desestima el uso de PCA.

Ambos modelos resultan parecidos y con muchas similitudes:

- Modelos no lineales y no paramétricos.
- Aprendizaje en conjunto (*ensemble learning*) basado en árboles de decisión.
- Registros agrupados por nodos y clasificados mediante condiciones lógicas.
- Compensación de sesgo y varianza construyendo multitud de árboles (bosque) de decisión. Se considera como salida el resultado más frecuente (clasificación) o promedio de los resultados (regresión).

Principal diferencia:

- Bosque aleatorio define límites donde mayor caída de heterogeneidad se produce en cada clasificación.
- Árboles extremadamente aleatorios los establece arbitrariamente



Representación gráfica de árbol de regresión de precio de la vivienda en Boston. Fuente: scikit-learn.org [9].

UNIVERSIDAD DE MÁLAGA

8. Definición de hiper-parámetros.

ESCUELA DE INGENIERÍAS INDUSTRIALES

Una vez se considera el desarrollo de ambos modelos, debe completarse la configuración del mismo a través de la definición de los **hiper-parámetros**.

- En el caso de la librería empleada (*Sklearn*) ambos comparten todos sus hiper-parámetros. (Tabla 8.1)

- Algunos ejemplos:
  - ***n_estimators*** define el número de árboles de decisión.
  - ***criterion*** el indicador que define la heterogeneidad en los nodos (mse, mae,…).
  - ***max_depth*** establece el número máximo de niveles de nodos.
  - ***min_simples_split*** define el mínimo de registros en un nodo para que pueda volver a ser dividido.
  - ***min_impurity_decrease*** exige un mínimo descenso en la heterogeneidad.

RandomForestReressor()	ExtraTreeRegressor()
n_estimators	*n_estimators*
criterion	*criterion*
max_depth	*max_depth*
min_samples_split	*min_samples_split*
min_samples_leaf	*min_samples_leaf*
min_weight_fraction_leaf	*min_weight_fraction_leaf*
max_features	*max_features*
max_leaf_nodes	*max_leaf_nodes*
min_impurity_decrease	*min_impurity_decrease*
min_impurity_split	*min_impurity_split*
bootstrap	*bootstrap*
oob_score	*oob_score*
n_jobs	*n_jobs*
random_state	*random_state*
verbose	*verbose*
warm_start	*warm_start*
ccp_alpha	*ccp_alpha*
max_samples	*max_samples*

Hiper-parámetros configurables en regresores de bosque aleatorio (izq.) y árboles ext.aleatorios (dcha.) en *Sklearn*. Fuente: scikit-learn.org [10] [11].

UNIVERSIDAD DE MÁLAGA

8. Definición de hiper-parámetros.

ESCUELA DE INGENIERÍAS INDUSTRIALES

La definición de los valores óptimos se realiza mediante una **búsqueda secuencial**, en la cual se evalúan para diferentes combinaciones de hiper-parámetros. Se emplea validación cruzada ***K-fold*** **con k=5**.

Selección de modelo

Hiper-parámetros con valor ya definido

Declaración de hiper-parámetros a definir y los valores a combinar

Ejecución de la búsqueda secuencial con validación cruzada (CV=5)

Entrenamiento del modelo con el mejor resultado de la búsqueda secuencial

```
modelo =ExtraTreesRegressor(max_features=None, max_depth=None, bootstrap=True, warm_start=True)

parametros = {'min_samples_leaf': [2,3,4,5,6],
              'criterion' : ['mse','mae'],
              'oob_score' : ['False','True'],
              'min_impurity_decrease' : [0.0, 0.01, 0.05, 0.1, 0.2, 0.3],
              'n_estimators' : [20,100,1000,2000,5000,10000],
              'ccp_alpha' : [0.1, None]}

grid = GridSearchCV(estimator=modelo, param_grid = parametros, cv = 5, n_jobs=-1, verbose=2)
grid.fit(X, y)
```

Captura del código empleado para la búsqueda secuencial , donde se declaran posibles configuraciones. Fuente: elaboración propia.

La ejecución de la búsqueda secuencial se ve limitada por el alto consumo de recursos computacionales.

Por ello se reduce a un número muy limitado de valores posibles.

UNIVERSIDAD DE MÁLAGA

8. Definición de hiper-parámetros.

ESCUELA DE INGENIERÍAS INDUSTRIALES

Como resultado de la búsqueda se registra la combinación de hiper-parámetros que obtiene una mejor evaluación del modelo.

```
Mejor configuración del estimador entre todos los parámetros ensayados:
RandomForestRegressor(bootstrap='False', ccp_alpha=0.1, criterion='mse',
                      max_depth=None, max_features=10, max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.3,
                      min_impurity_split=None, min_samples_leaf=3,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=100, n_jobs=None, oob_score='False',
                      random_state=None, verbose=0, warm_start=True)
>>> print("\n El mejor índice de regresión ensayado:\n",
...           grid.best_score_)

 El mejor índice de regresión ensayado:
0.6642818951762648
```

```
Mejor configuración del estimador entre todos los parámetros ensayados:
ExtraTreesRegressor(bootstrap=True, ccp_alpha=0.1, criterion='mse',
                    max_depth=None, max_features=None, max_leaf_nodes=None,
                    max_samples=None, min_impurity_decrease=0.3,
                    min_impurity_split=None, min_samples_leaf=2,
                    min_samples_split=2, min_weight_fraction_leaf=0.0,
                    n_estimators=100, n_jobs=None, oob_score='False',
                    random_state=None, verbose=0, warm_start=True)
>>> print("\n El mejor índice de regresión ensayado:\n",
...           grid.best_score_)

 El mejor índice de regresión ensayado:
0.6447825948190913
```

Capturas de pantalla de valores de hiper-parámetros resultantes de la búsqueda secuencial en bosque aleatorio (izq.) y árboles ext. aleatorios (dcha.). Fuente: elaboración propia.

Se hace esencial el empleo de validación cruzada con *K-folds* para evitar sobreajuste de los hiperparámetros.

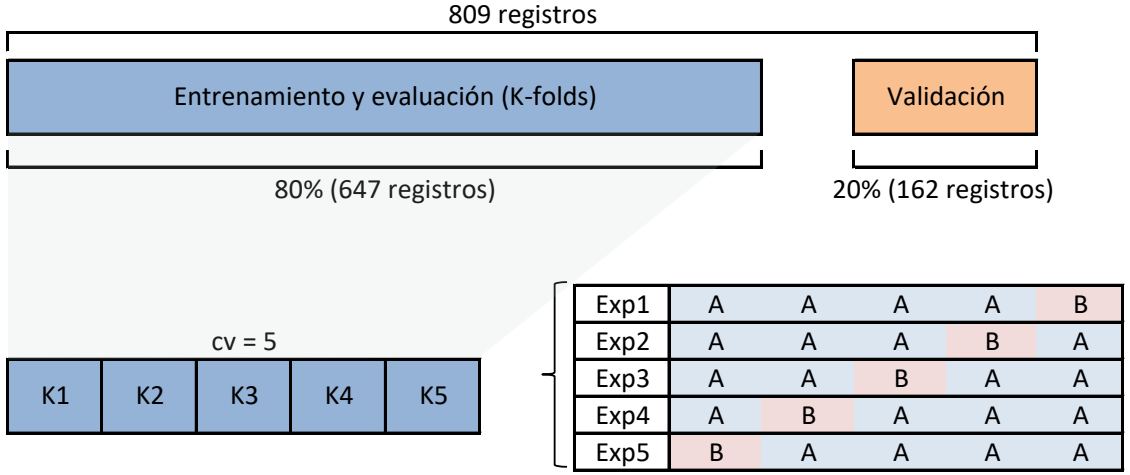


Diagrama del proceso de validación cruzada y subconjuntos de registros empleados durante las fases de definición de parámetros y validación de modelos. Fuente: elaboración propia

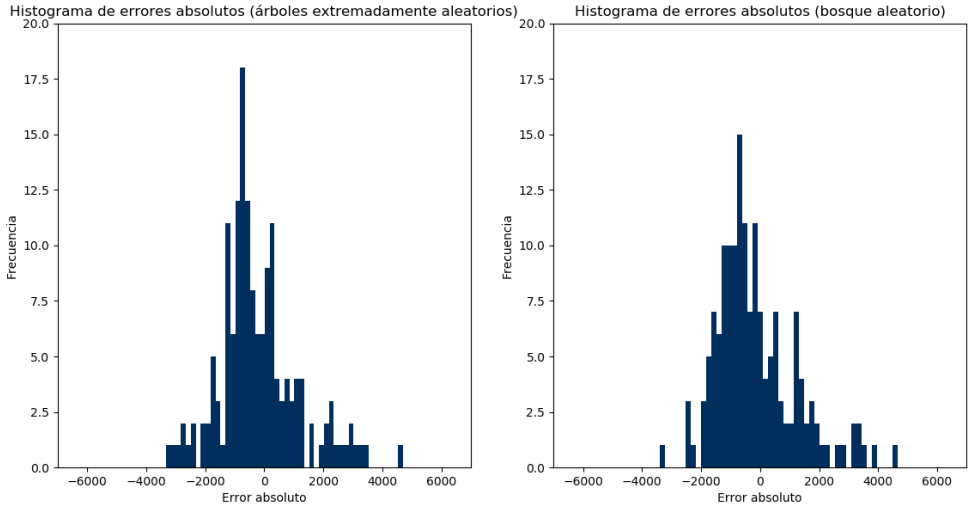
- Definición de hiper-parámetros mediante entrenamiento y evaluación de un 80% de registros.
- La evaluación de cada combinación se produce como el promedio de cada experimento que combina subconjuntos de entrenamiento (A) y de evaluación (B).
- Se reserva 20% para la validación.

9. Validación del modelo.

- Esta etapa permite tener datos más realistas sobre el modelo: se mide el desempeño prediciendo los valores del 20% restante que no se ha empleado para su entrenamiento ni definición de parámetros.

Modelo	R^2	RECM	Error absoluto				
			medio	Q1	Q2	Q3	Desviación estándar
Árboles extremadamente aleatorios	0,65	2234	-327	-1072	-530	428	2287
Bosque aleatorio	0,4	2216	-305	-947	-512	347	2201

Tabla resumen de indicadores del desempeño árboles ext. Aleatorios y bosque aleatorio durante la validación. Fuente: elaboración propia.



Histogramas de error absoluto del modelo de árboles ext. Aleatorios (izq.) y bosque aleatorio (dcha.). Fuente: elaboración propia.

- Un menor desempeño que en la selección del modelo se debe a que ha sido entrenado con menor número de registros (64% de registros en lugar del 80% de la selección de modelo).

9. Validación del modelo.

- A pesar de registrar menor coeficiente de determinación (R^2=0,4) y una distribución de error absoluto similar, se opta por el modelo de árboles extremadamente aleatorios. Entre otros motivos, diferentes ejecuciones del modelo de bosque aleatorio sobre el mismo conjunto evidencia inestabilidad al fluctuar R^2 de forma significativa.

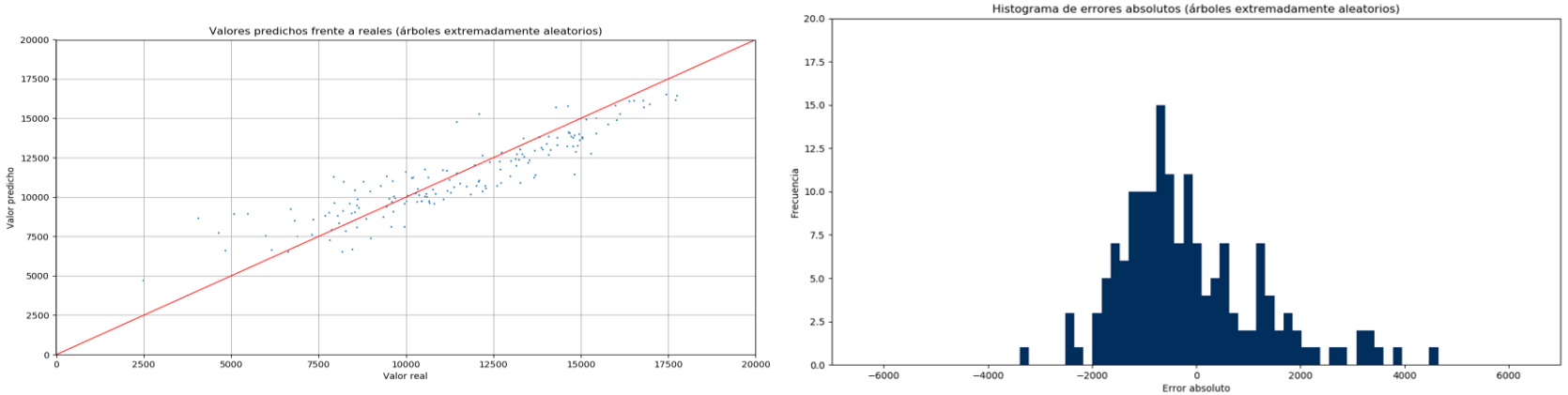


Gráfico que representa valores predichos para los registros de la validación frente a reales (izq,) e histograma de errores absolutos para los mismos registros(dcha.). Nótese se trata de una ejecución diferente a la imagen anterior. Fuente: elaboración propia.

- Se obtiene un modelo que predice registros obteniendo un error absoluto concentrándose entorno al cero (preciso). Sin embargo se muestra cierto sesgo a la izquierda (inexacto).

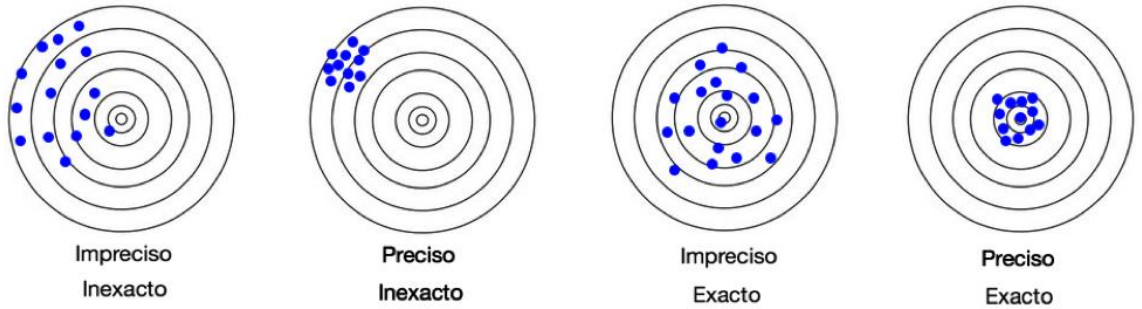


Diagrama explicativo de precisión y exactitud. Fuente: recurso propio de Óscar Moreno [12]

UNIVERSIDAD DE MÁLAGA

9. Validación del modelo.

ESCUELA DE INGENIERÍAS INDUSTRIALES

- Aquellas predicciones con raíz de error cuadrático (REC) mayor a 3000 desplazamientos, 7 de las 10 se identifican como fechas con peculiaridades que pueden afectar el uso del servicio BiciMAD.

Registros con error > 3.000 (durante validación)

Fecha	Real	Predicho	Error absoluto	REC	Posible causa	Posible efecto
jueves 12/10/2017	12.083	15.218	3.135	3.135	Día de la Hispanidad	Festivo, cortes de tráfico
viernes 3/11/2017	14.801	10.686	-4.115	4.115		
viernes 30/3/2018	4.043	8.661	4.618	4.618	Festivo: Viernes Santo	
sábado 31/3/2018	4.637	7.999	3.362	3.362	Festivo: Sábado de Gloria	
miércoles 30/5/2018	12	10.107	10.095	10.095		
miércoles 15/8/2018	7.936	11.447	3.511	3.511	Festividad Virgen de la Paloma	Festivo, cortes de tráfico
lunes 24/12/2018	5.080	9.053	3.973	3.973	Víspera de Navidad	Baja demanda
lunes 31/12/2018	5.472	9.070	3.598	3.598	Nochevieja	Baja demanda, cortes trafico
sábado 1/6/2019	11.448	15.996	4.548	4.548		
jueves 13/6/2019	32.205	16.384	-15.821	15.821	Ampliación de red: 22 estaciones y 486 bicicletas	Aumento capacidad del sistema
viernes 14/6/2019	30.764	16.180	-14.584	14.584		

Análisis de posibles causas que motivan desviaciones de superiores a 3000 desplazamientos/día. Fuente: elaboración propia.

- Para su implantación real, se propone constante **reentrenamiento, evolución y rediseño** del modelo en ciclos de mejora continua.
- El modelo actual solo contempla **factores meteorológicos y de estacionalidad**. Se propone incluir variables específicas en relación a festividades, grandes eventos o cualquier otra circunstancia previsible que pueda repercutir en la demanda.
- **Otras variables**, como demanda eléctrica prevista, pueden resultar indicadores de la actividad económica prevista [13].
- Inclusión de **históricos de demanda** del mismo día en años anteriores.
- **Mejora de capacidades computacionales:** máquinas dedicadas, máquinas virtuales, cálculo en la nube, distribución paralela de la carga (*Apache Hadoop*).
- **Variables que contemplen cambios en tendencia de uso** y en dinámicas sociales, como el producido a causa del COVID-19.

10. Conclusiones.

Con carácter específico de este Trabajo Fin de Grado:

- Se ha completado la **obtención de un modelo predictivo del pronóstico de la demanda** que, aunque solo contempla factores estacionales y meteorológicos, **se aproxima a los valores realmente registrados**. En consecuencia y a la luz de los resultados obtenidos, se valida la hipótesis propuesta:

> *«H_1 - La meteorología y determinados factores de estacionalidad influyen en la demanda experimentada por el sistema BiciMAD»*

- **Se establecen líneas de mejoras** propuestas para el desarrollo del modelo que debe encontrarse en revisión constante.

- **Otros factores** como festividades, grandes eventos deportivos, culturales y sociales, así como cambios en los hábitos de consumo, **deben contemplarse como nuevas variables en el conjunto**.

- Los resultados obtenidos del Análisis de Componentes Principales (PCA), así como los modelos matemáticos que destacan, **son indicios de dinámicas del servicio de carácter no lineal**.

10. Conclusiones.

Con carácter genérico en el ámbito de las organizaciones:

- La jerarquía del conocimiento se hace **cada vez más necesaria**, a todos los niveles de la organización.
- La obtención del conocimiento, en adición a la transformación digital, propicia la **aparición de nuevos productos y servicios** antes no posibles, así como de procesos en línea con la Cuarta Revolución Industrial.
- Esta realidad establece un nuevo **agravio entre las organizaciones** a causa del volumen de sus operaciones, grado de automatización o de la disponibilidad de recursos a emplear en implementación de proyectos de este tipo.
- Los procesos de aprendizaje automatizado e inteligencia artificial establecen nuevos **paradigmas y dilemas éticos** que las sociedades modernas deben afrontar.
- En una sociedad globalizada e interconectada, donde acontecimientos puntuales pueden implicar cambios de orden mundial, se hace más necesarias las técnicas de aprendizaje automático para el **reconocimiento de nuevas dinámicas** y la adaptación de las organizaciones.

UNIVERSIDAD DE MÁLAGA

ESCUELA DE INGENIERÍAS INDUSTRIALES

Gracias por su atención

Título Aplicación de técnicas de minería de datos sobre registros de operaciones del sistema de alquiler público de bicicletas de Madrid (BiciMAD)

Autor Maximiliano Gámez López

Convocatoria junio 2020

Tutor Dr. Juan Carlos Rubio Romero

Cotutora Dra. María Martínez Rojas

Área de conocimiento Organización de empresas
Departamento de Economía y administración de empresas

UNIVERSIDAD DE MÁLAGA

ESCUELA DE INGENIERÍAS INDUSTRIALES

Bibliografía

[1] J. Rowley, "The wisdom hierarchy: Representations of the DIKW hierarchy," *J. Inf. Sci.*, vol. 33, no. 2, pp. 163–180, 2007.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[3] Ayuntamiento de Málaga, "Lineas y horarios bus - Google Transit - Conjuntos de datos - Datos abiertos." [En línea]. Disponible: https://datosabiertos.malaga.eu/dataset/lineas-y-horarios-bus-google-transit [Accedido: 20-May-2020].

[4] "Imagen gratis manuscrito," *Pixabay*. [En línea]. Disponible: https://pixabay.com/es/illustrations/papel-viejo-vintage-textura-2133481/ [Accedido: 26-Jun-2020].

[5] Cisco Networking Academy, "IoT Fundamentals: Connecting Things." .

[6] K. Lee, "Data Mining for Business Intelligence." .

[7] IBM Software Group, "IBM SPSS Modeler CRISP-DM Guide." 2016.

[8] Scikit Learn, "1.10. Decision Trees — scikit-learn 0.23.1 documentation." [En línea]. Disponible: https://scikitlearn.org/stable/modules/tree.html#tree [Accedido: 28-May-2020].

[9] T. Parr and P. Grover, "How to visualize decision trees," *explained.ai*. [En línea]. Disponible: https://explained.ai/decision-tree-viz/ [Accedido: 28-May-2020].

[10] "3.2.4.3.4. sklearn.ensemble.ExtraTreesRegressor — scikit-learn 0.23.1 documentation," *Scikit Learn*. [En línea]. Disponible: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html [Accedido: 28-May-2020].

[11] "3.2.4.3.2. sklearn.ensemble.RandomForestRegressor — scikit-learn 0.23.1 documentation," *Scikit Learn*. [En línea]. Disponible: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html [Accedido: 28-May-2020].

[12] Ó. Moreno-Díaz, "Exactitud y precisión," *Ministerio de Educación y Formación Profesional - INTEF*. [En línea]. Disponible: http://formacion.intef.es/pluginfile.php/246707/mod_resource/content/1/exactitud_y_precisin.html [Accedido: 04-Jun-2020].

[13] Red Eléctrica de España, "Demanda eléctrica y actividad económica: ¿Cambio de paradigma?," 2019.