Universitat Oberta de Catalunya

# Análisis predictivo de datos abiertos sobre el uso turístico del servicio de alquiler compartido de bicicletas de Nueva York

## Una perspectiva desde la Ciencia de Datos

uoc.edu

Carlos E. Jiménez Gómez
Trabajo Final de Máster
Máster Universitario en Ciencia de Datos

# Índice

Motivación
Objetivos y preguntas a responder
Metodología
Estado del arte
Técnicas utilizadas
Resultados
Conclusiones

# Motivación

# Motivación

- Problemáticas en el uso/demanda del servicio público de alquiler compartido de bicicletas
- El turismo como área estratégica en Nueva York: >60 millones visitantes en 2017
- En general, análisis predictivos en la literatura sin segmentación de usuario
- Aprovechamiento de datos abiertos
- Especial interés de las administraciones en *Smart Government*

# Objetivos y preguntas a responder

# Objetivos

- Análisis de datos abiertos sobre el uso del servicio público de alquiler de bicicletas de Nueva York (usuarios turistas)
- Análisis predictivo comparativo de modelos bajo un enfoque de *machine learning*. Especial interés en función de precipitaciones y eventos (festivos)
- Mostrar la necesidad estratégica de construir una organización pública orientada al dato e integrar la perspectiva de la ciencia de datos en su ecosistema organizacional

# Preguntas a responder

- ¿Podemos considerar los eventos -identificados con días festivos y fines de semana- como variables sustanciales en la predicción de la demanda diaria de bicicletas por usuarios no suscritos al servicio público de alquiler de bicicletas?

- ¿El alquiler y la demanda de bicicletas por parte de los usuarios turistas (no suscritos o *Customer*), se rige por los mismos patrones que el de los usuarios suscritos (*Subscriber*)?

- ¿Qué algoritmo y, en base a éste, que modelo arroja mejores resultados en la predicción de la demanda diaria de bicicletas por parte de los usuarios no suscritos?

- ¿A partir de un análisis de uso de bicicletas segmentado por tipo de usuario y de la predicción sobre el mismo, que propuestas podrían aportar valor a la gestión de recursos de la ciudad, la planificación y al diseño de políticas públicas en las áreas de transporte y turismo?

# Metodología

# Metodología

## Seguimiento de las fases propias de un proyecto de Ciencia de Datos

1. Análisis del ámbito, problemáticas y preguntas a responder
2. Identificación, recogida y almacenamiento de conjuntos de datos
3. Preprocesado, preparación de datos y análisis inicial, incluyendo visualización preliminar
4. Visualizaciones y análisis visual
5. *Machine learning*:
    - Entrenamiento y test de modelos clasificación
    - Predicción de los modelos y análisis comparativo
6. Análisis de resultados
7. Conclusiones y respuestas a las preguntas planteadas

# Estado del arte

# Estado del arte

## Análisis - Machine Learning

- Martins et al., (2015); Feng y Wang (2017). Random Forest
- Thu et al., (2017); Liu et al. (2015). Red Neuronal Artificial
- Datta (2014). AdaBoost
- Nekkanti (2017). J48
- Nekkanti (2017); Martins et al. (2015). Utilizan Naive Bayes y Red Bayesiana
- No se profundiza en la segmentación por tipo de usuario
- Detalles generales sobre la preparación de datos

## Organizacional - estratégico:

- Kaplan et al. (2014). Ante un hipotético escenario de vacaciones, interés de los potenciales turistas en estos servicios
- Vogel et al. (2011). Patrones de comportamiento en los sistemas de uso de bicicleta para mejora de planificación estratégica y operativa.

# Técnicas utilizadas

# Técnicas utilizadas (I)

## Preprocesado, preparación y almacenamiento de datos (R, PostgreSQL)

***Conjunto de datos para visualización:***

- Julio 2017: 1.735.599 observaciones de  21 variables
- Segmentación/granularidad:
  - 2 tipos de usuarios
  - Diaria
  - Semanal
  - Mensual

# Técnicas utilizadas (II)

## Preprocesado, preparación y almacenamiento de datos (R, PostgreSQL)

***Conjunto de datos para machine learning (I)***:

- Abril a septiembre 2017: más de 10 millones de observaciones de 21 variables
    - Agrupadas por día en 183 observaciones y 6 variables:
        - Demanda diaria (clase a predecir, con 3 niveles: Baja, Media, Alta por método K-medias)
        - Día (del año)
        - Precipitación
        - Laborable/festivo
        - Mes
        - Día de la semana

# Técnicas utilizadas (III)

## Preprocesado, preparación y almacenamiento de datos (R, PostgreSQL)

***Conjunto de datos para machine learning (II)***:

- Segmentación para usuario *Customer*
- 3 conjuntos para entrenamiento y test (variación tipo variable día)
- 1 conjunto con nuevos datos para predicción
- Validación cruzada con 10 particiones

# Técnicas utilizadas (IV)

## Análisis georreferenciado, mapas y visualización de características (Carto)

## Análisis comparativo machine learning y predicción (Weka)

- J48
- Random Forest
- AdaBoostM1
- Red Neuronal Artificial
- Naive Bayes
- Red Bayesiana

## Búsqueda hiperparámetros y mejor clasificador (AutoWeka)

- JRip

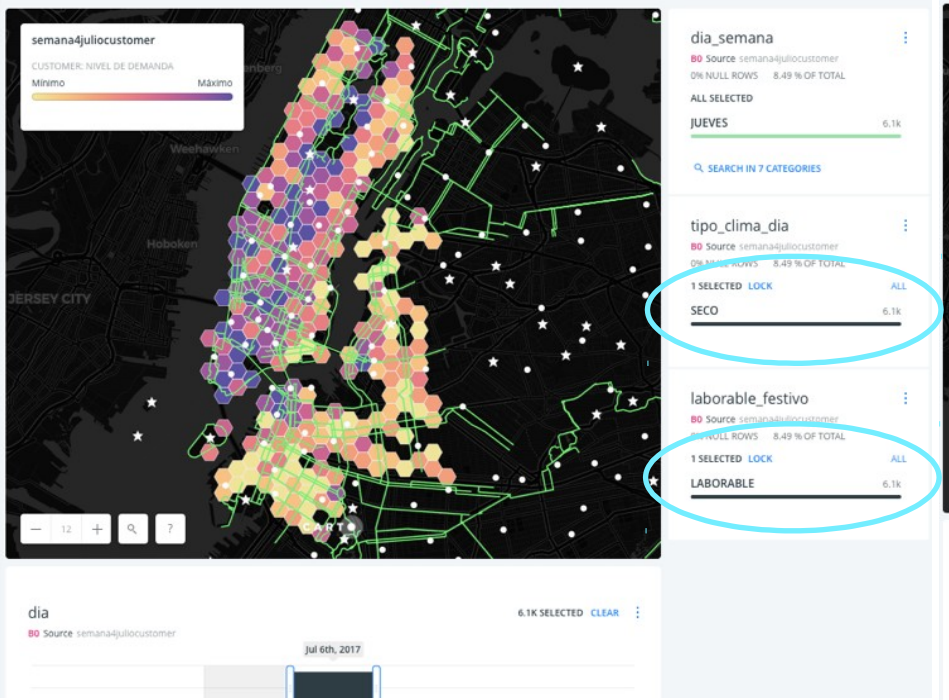
## Evaluación de clasificación correcta (Weka)

- AUC ROC
- F *measure*

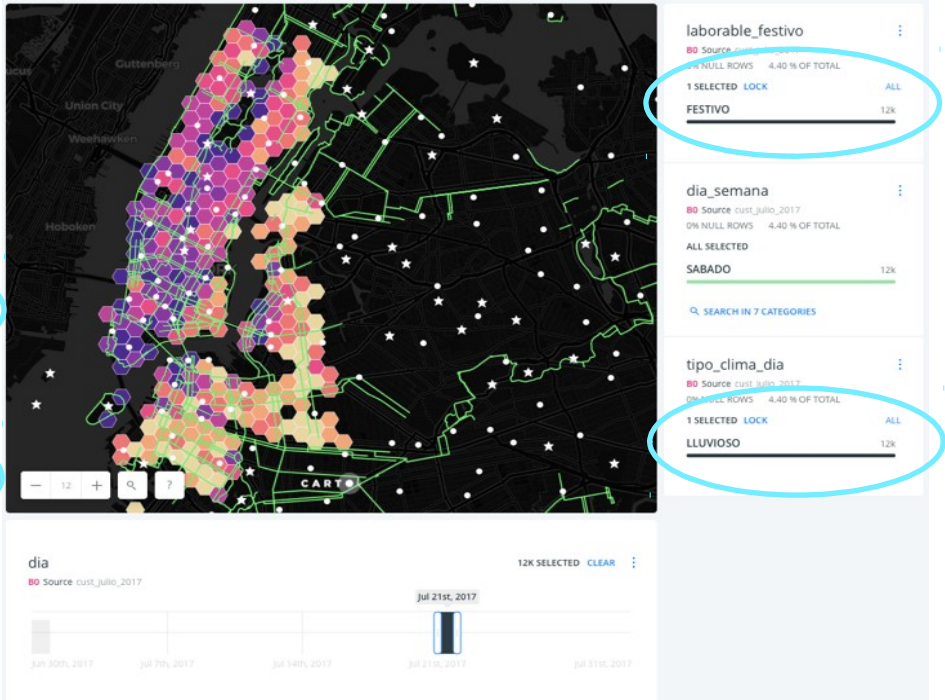
# Resultados

# Visualizaciones (I)

**6.100**

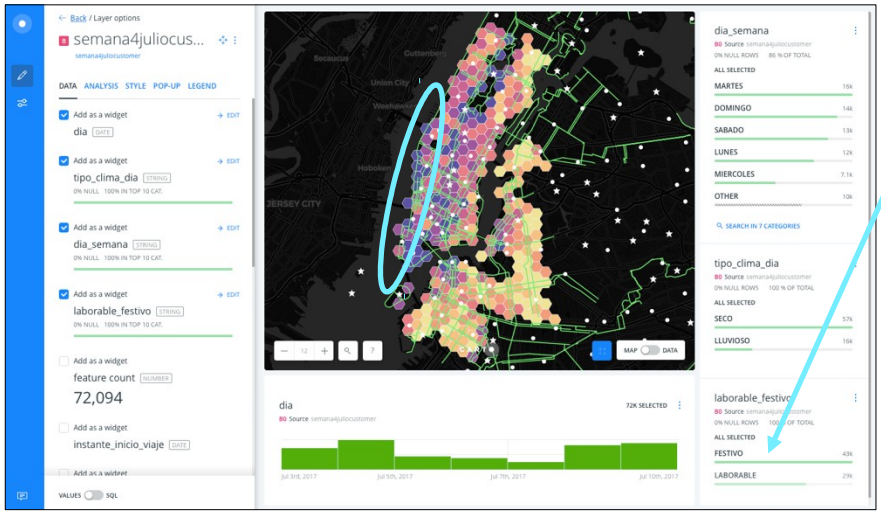


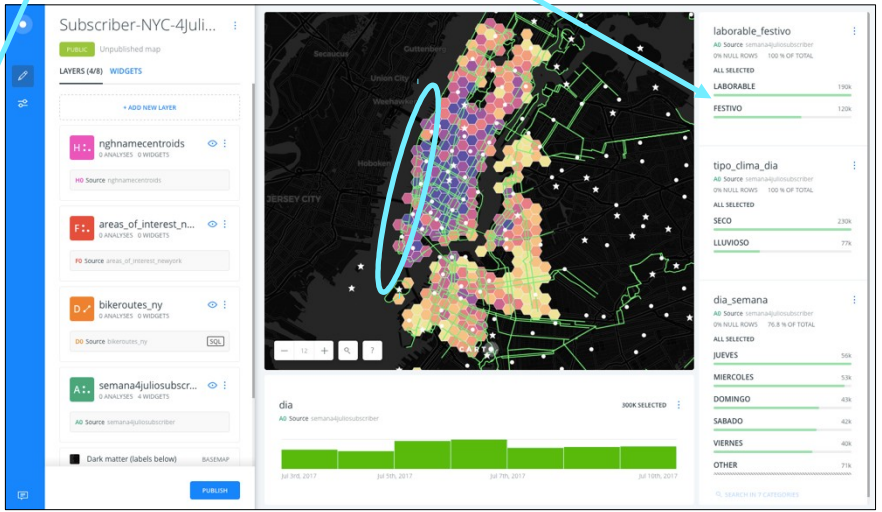

**12.000**



# Visualizaciones (II)

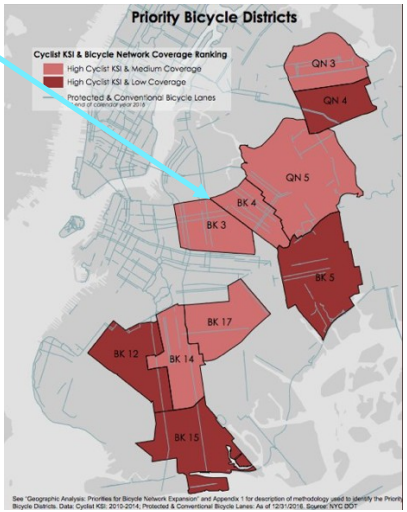
- Diferentes densidades (misma zona) y preferencia días festivos (*Customer*) vs. laborables (*Subscriber*)



# Visualizaciones (III)

- Identificación de puntos de mayor afluencia turística: valor para decisiones en la priorización de estrategias y planificación

# Influencia del preprocesado y preparación de datos: tipo de datos entrenamiento y test (% correcto)

- Variación en el tipo de un atributo (tipo de la variable “día”)

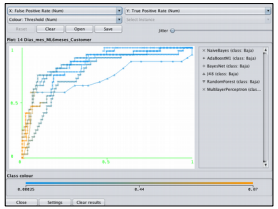
A partir de hiperparámetros y clasificador óptimos

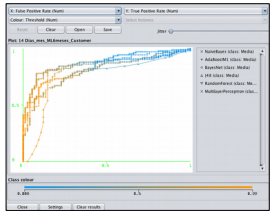
| | J48 | AdaBoostM1 | Random Forest | Red Neuronal | Red Bayesiana | Naive Bayes | JRip(*) |
|---|---|---|---|---|---|---|---|
| Conjunto datos variable "día" incluye tipo fecha | 80.3 | 77.11 | 75.76 | 77.62 | --- | --- | --- |
| Conjunto datos variable "día" nominal (solo formato fecha) | 80.6 | 80.46 | 66.36 * | 77.67 | 78.35 | 78.56 | --- |
| Conjunto datos variable "dia" agrupada (rango 1 a 31) | 82.21 | 75.6 | 71 13 * | 75.77 * | 77.74 | 79.32 | 84.153 |

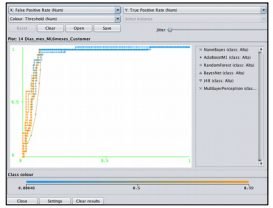
*Dataset* para la predicción

Impacto de preparación y preprocesado de datos

# Modelos: clasificación (*F measure* y AUC ROC media ponderada)



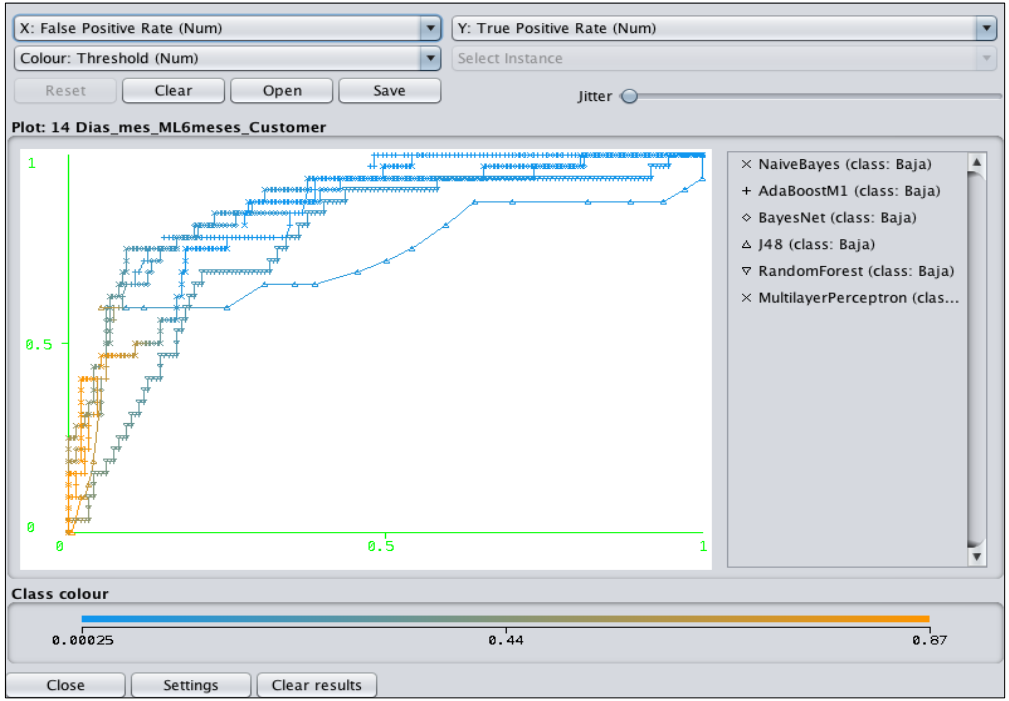




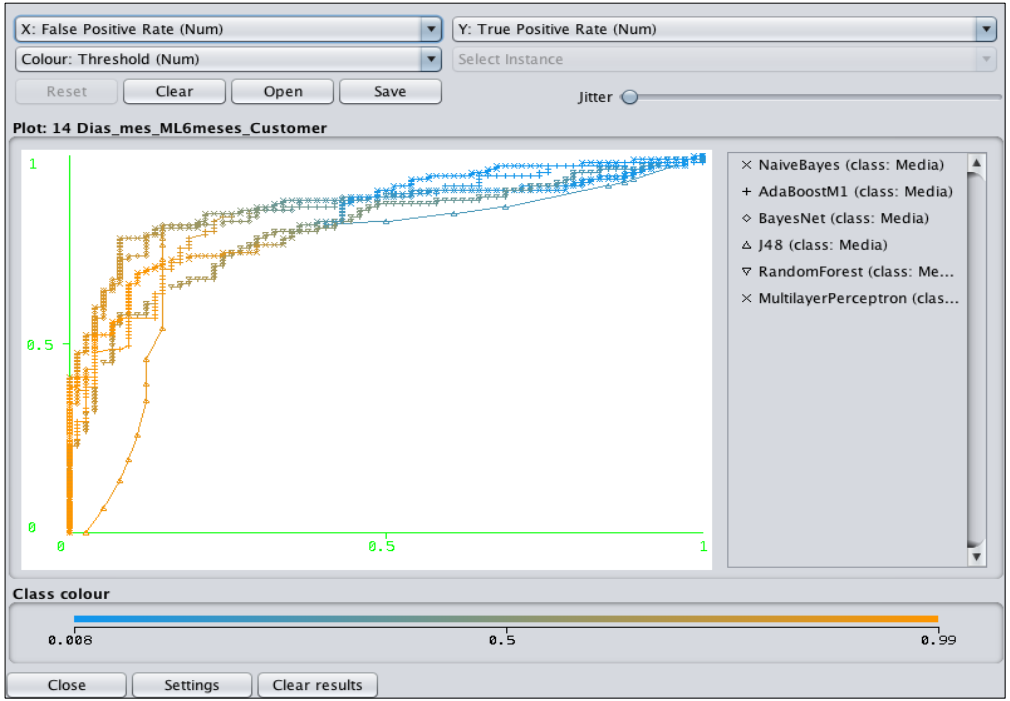
| *F measure* | J48 | AdaBoostM1 | Random Forest | Red Neuronal | Red Bayesiana | Naive Bayes | JRip(*) |
|---|---|---|---|---|---|---|---|
| Conjunto datos variable "día" nominal (solo formato fecha) | 0.67 | 0.67 | - | 0.55 | 0.55 | 0.54 * | --- |
| Conjunto datos variable "dia" agrupada (rango 1 a 31) | 0.67 | 0.62 | 0.24 * | 0.53 | 0.52 * | 0.57 | 0.838 |

| | J48 | AdaBoostM1 | Random Forest | Red Neuronal | Red Bayesiana | Naive Bayes | JRip(*) |
|---|---|---|---|---|---|---|---|
| % Clasificación correcta | 81.4208 | 77.0492 | 70.4918 | 71.5847 | 77.0492 | 78.1421 | 84.153 |
| % Clasificación incorrecta | 18.5792 | 22.9508 | 29.5082 | 28.4153 | 22.9508 | 21.8579 | 15.847 |
| Precisión media pond. | 0.821 | 0.768 | 0.641 | 0.719 | 0.759 | 0.772 | 0.84 |
| AUC ROC media pond. | 0.786 | 0.855 | 0.824 | 0.869 | 0.873 | 0.88 | 0.887 |

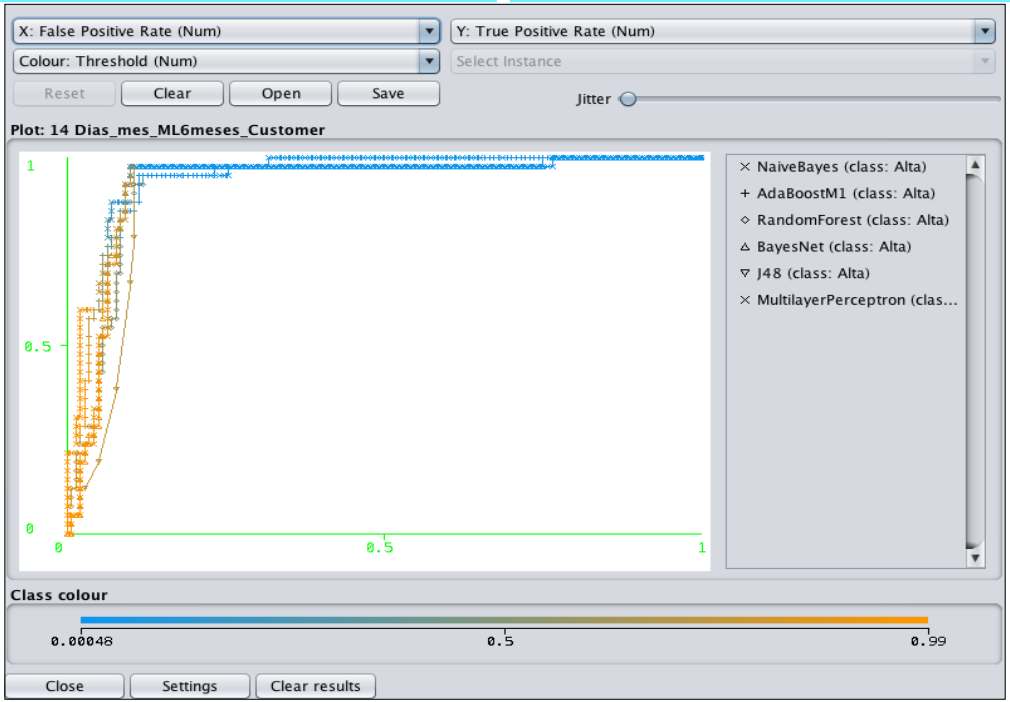
# Modelos: clasificación. AUC ROC demanda Baja



# Modelos: clasificación. AUC ROC demanda Media

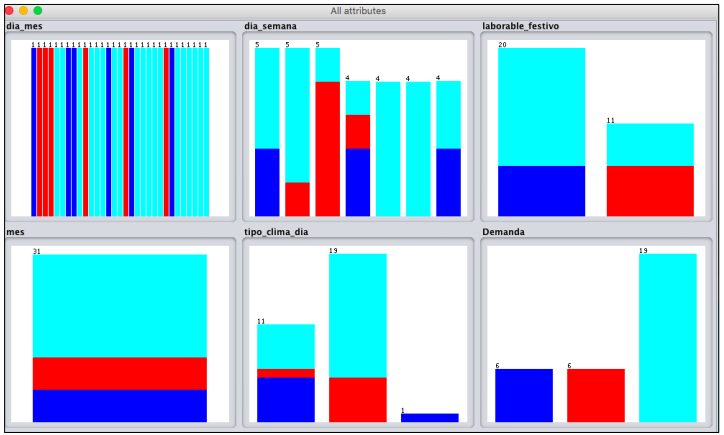


# Modelos: clasificación. AUC ROC demanda Alta



# Predicción de los modelos sobre nuevos datos

- 1.380.110 observaciones en un mes (julio 2016) agrupadas por día en:
- 31 observaciones, y 6 variables
- No existieron días con etiqueta “Lluvia ocasional” para el atributo *tipo_clima_dia*




| | Número de Aciertos | Tasa de Aciertos | Número de Errores | Tasa de Error |
|---|---|---|---|---|
| Red Neuronal | 25 | 0.8065 | 6 | 0.1935 |
| J48 | 24 | 0.7742 | 7 | 0.2258 |
| JRip | 23 | 0.7419 | 8 | 0.2581 |
| AdaBoostM1 | 22 | 0.7097 | 9 | 0.2903 |
| Random Forest | 21 | 0.6774 | 10 | 0.3226 |
| Red Bayesiana | 21 | 0.6774 | 10 | 0.3226 |
| Naive Bayes | 21 | 0.6774 | 10 | 0.3226 |

# Conclusiones

# Conclusiones (I)

- Patrones diferentes entre los distintos tipos de usuarios
- Los eventos identificados con días festivos es una variable sustancial e incluso más determinante clasificando que las precipitaciones
- Importancia del preprocesado y proceso de preparación de conjuntos de datos de entrenamiento y test: condicionante de resultados
- Red Neuronal Artificial (ANN) como mejor modelo predictor, teniendo en cuenta asimismo:
  - Pocas diferencias con significación estadística en la clasificación
  - F más concluyente que AUC ROC para resultados entre clasificación y predicción ()
  - Peores en la clasificación correcta: Random Forest y Red Bayesiana, frente a J48.

# Conclusiones (II)

- Importancia de incorporar el proceso y la metodología de la Ciencia de Datos dentro de la propia organización pública, para el alineamiento con los objetivos organizacionales
- Necesaria orientación al dato para el avance hacia *Smart Government*
- Los resultados podrían ser útiles para la toma de decisiones en cuanto
  - Planificar la potenciación del uso de zonas específicas con nuevas rutas
  - Incorporar estaciones en zonas donde antes no existían
  - Reducir la presencia de las mismas donde su uso es muy bajo
  - Utilizar puntos de información en lugares de preferencia
- Estrategia para favorecer de un modo selectivo una mayor afluencia de usuarios a determinadas zonas.
- Importancia de que las Administraciones Públicas incorporen una estrategia de apertura de datos transversal, siguiendo unos estándares y requisitos mínimos, que permitan aprovechar adecuadamente los datos.

Universitat Oberta
de Catalunya

@estratic
carlosjg@uoc.edu
www.estratic.com
www.linkedin.com/in/carlosejimenez