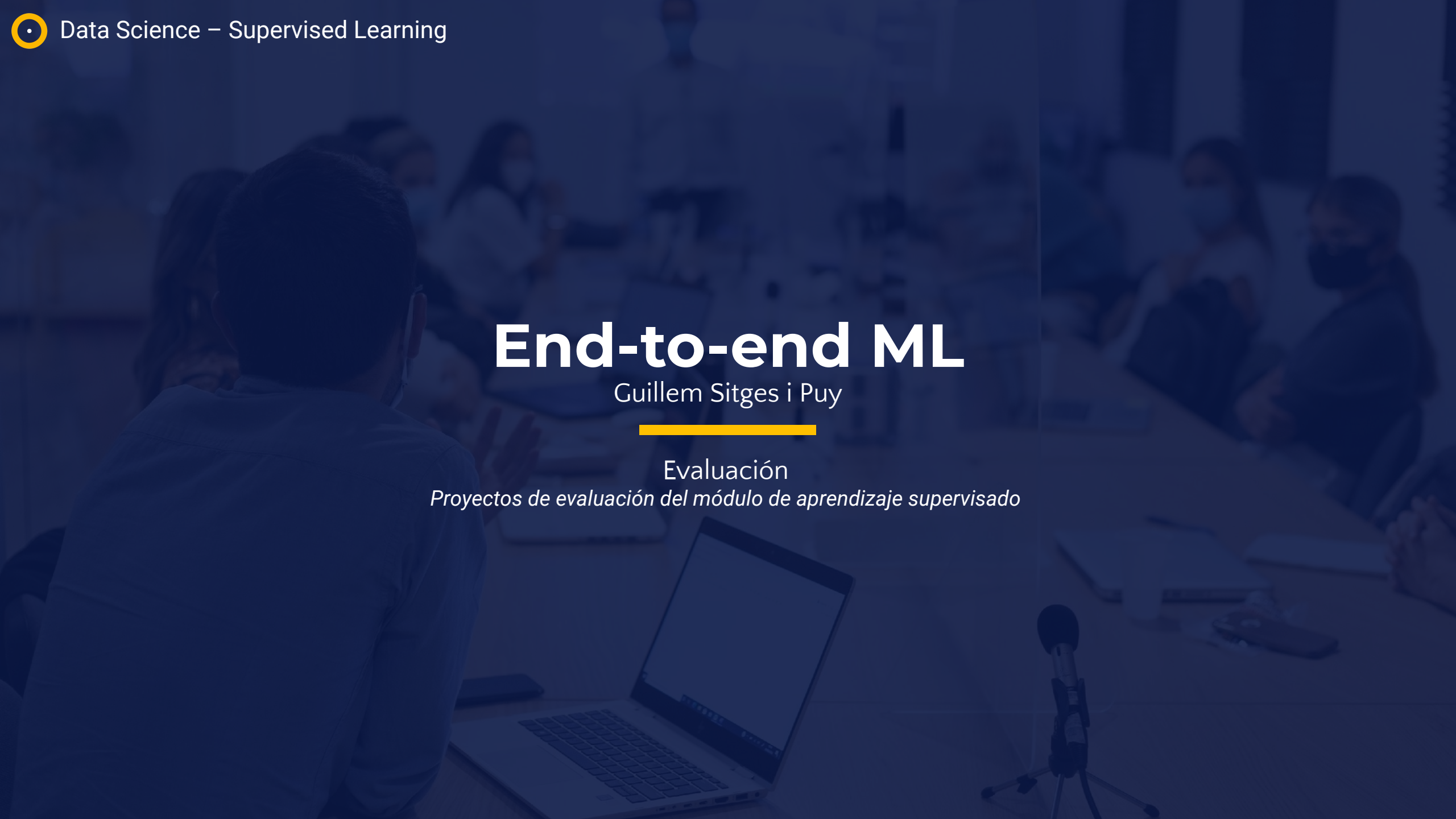


End-to-end ML

Guillem Sitges i Puy

Evaluación

Proyectos de evaluación del módulo de aprendizaje supervisado

1 EJERCICIO OBLIGATORIO: MICROSOFT MALWARE PREDICTION

El objetivo de este ejercicio es estimar la probabilidad de que una máquina con Sistema Operativo Windows se vea infectada por algún tipo de malware, en base a las distintas propiedades de la máquina. Los datos para este ejercicio se encuentran en la siguiente url:

https://www.dropbox.com/s/sxl5bpi2620p496/sample_mmp.csv

Y se han obtenido muestreando el dataset original de la competición de Kaggle Microsoft Malware Prediction (https://www.kaggle.com/c/microsoft-malwareprediction), y se basan en las características obtenidas en la solución de endpoint Windows Defender. Cada fila del dataset corresponde a una máquina única, identificada por el campo MachineIdentifier. El target es la variable HasDetections, que indica que se ha detectado Malware en la máquina.

Se solicita:

1. Desarrollar el ML Canvas para este problema, suponiendo que nuestro modelo se usará para implementarlo en la aplicación Windows Defender dando aviso al usuario cuando su máquina supere un cierto umbral de probabilidad de ser infectada.

2. Desarrollar un Notebook con nuestra propuesta de modelo para resolver el problema. El Notebook debe contener todas las etapas de la ML Checklist debidamente comentadas (se valorará la claridad), y ejecutar sin problemas para obtener el modelo resultado. En concreto, debe realizarse la exploración de datos (se valorará el desarrollo de visualizaciones interesantes), el preprocesamiento, el modelado mediante un Decision Tree (opcionalmente, explorar otros algoritmos) y la evaluación.

1 EJERCICIO OBLIGATORIO: MICROSOFT MALWARE PREDICTION

El ejercicio puntuable de Supervised Machine Learning se entregará, como máximo, el domingo 27/02/21 a las 23:59h.
La entrega la podéis realizar por email: **sitgesguillem@gmail.com**

Se espera la entrega de un Notebook con el formato .ipynb y el siguiente nombre: **1121_SupML_{Nombre y Apellido}**

En caso de haber modificado el dataset de .csv original previo a cargar los datos en el Notebook, se deberá adjuntar también el .csv (link de Drive o Github).

¿En qué debo centrar mis esfuerzos para conseguir un buen proyecto?

CORRECIÓN

- Detectar la tarea correctamente
- Uso adecuado de algoritmo / algoritmos
- Transformación de variables distintas
- Partición correcta del dataset
- Métricas de evaluación adecuadas
- Conclusiones acuradas

AUTOMATIZACIÓN

- Uso de funciones en Python.
- Eficiencia del código: evitar for loops si es posible

EXPLICABILIDAD

- Visualizaciones chulas
- Plots variados
- Razonamiento de negocio
- Diferentes formas de explicar resultado del modelo: feature importance, árbol pintado,...

2 EJERCICIO OPCIONAL: BMW PRICING

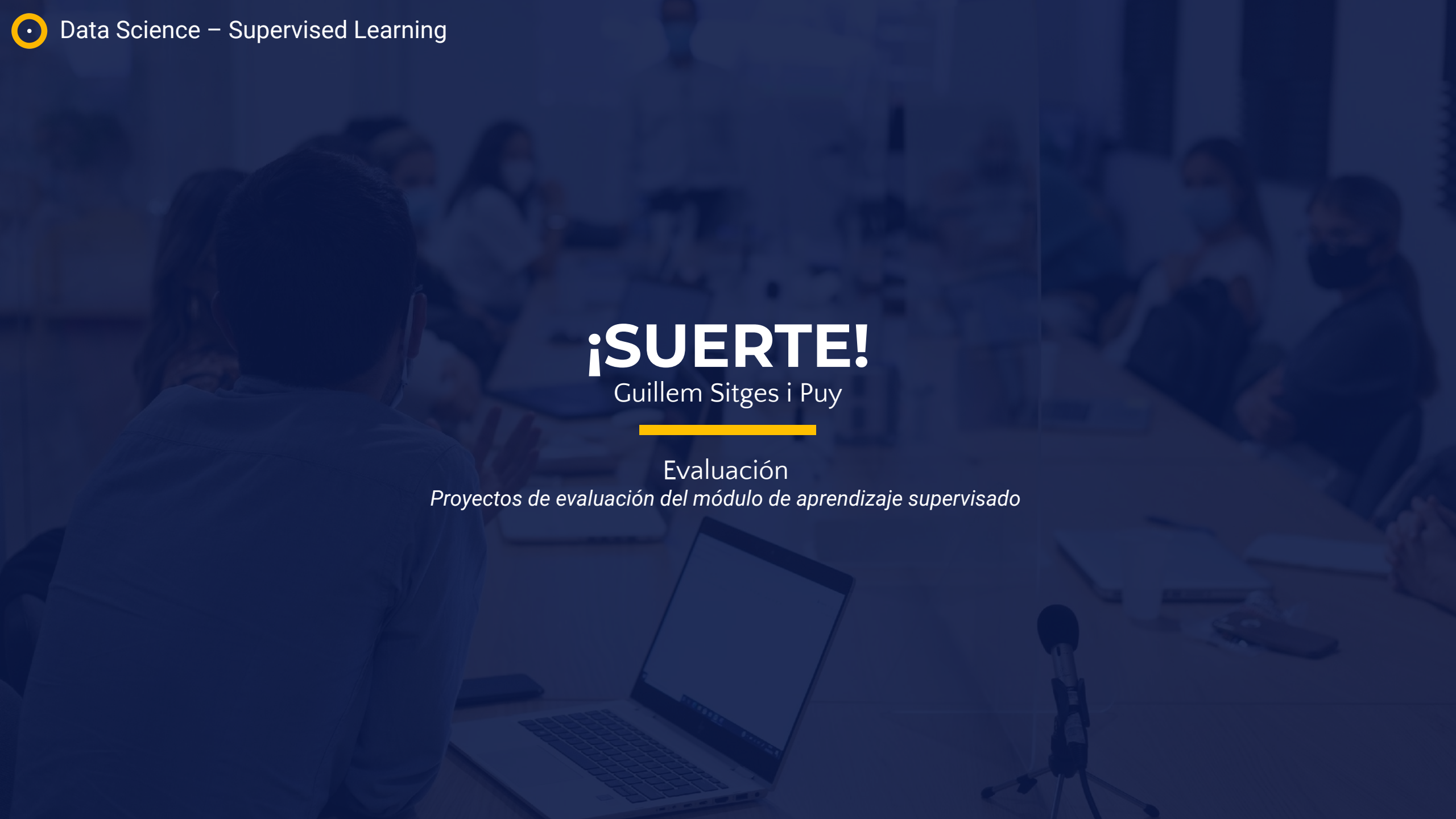
El objetivo de este ejercicio es estimar el precio de venta de un vehículo dadas sus características. Los datos para este ejercicio se encuentran en la siguiente url:

https://www.dropbox.com/s/di2dnc31k8cega1/bmw_pricing_v2.csv

La primera parte del ejercicio consistirá en pasar esta info a una tabla en SQL y realizar las siguientes consultas:

- Hacer un listado de los vehículos agrupados por el modelo y tipo de gasolina.
- ¿Cuantos vehículos tienen potencia superior a los 150, que sean de tipo de coche convertible y tengan el volante regulable?
- ¿Podrías indicar si hay algún caso que la fecha de registro sea superior a la fecha de venta?, de ser así indicar el precio medio encontrado.
- Indicar el precio medio de los vehículos con aire acondicionado.
- ¿Existe alguna diferencia significativa entre el color del vehículos y el precio? agrupar los vehículos por su color y su precio medio para responder la pregunta.

Tras esto, se pide realizar un modelo predictivo en Python mediante una regresión lineal, incluyendo la exploración (al menos las visualizaciones de evolución temporal del precio y el número de vehículos vendidos por modelo), el preprocesamiento de los datos y la estimación mediante una regresión lineal (opcionalmente, explorar otros algoritmos).



¡SUERTE!

Guillem Sitges i Puy

Evaluación

Proyectos de evaluación del módulo de aprendizaje supervisado