

End-to-end ML

Guillem Sitges i Puy

Evaluación

Proyecto de evaluación del módulo de aprendizaje supervisado

1 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:

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.

MICROSOFT MALWARE PREDICTION

 El ejercicio puntuable de Supervised Machine Learning se entregará, **5 Febrero 2025 a las 23.59h**

 La entrega la podéis realizar en el Campus, cualquier duda me podéis contactar por email: sitgesguillem@gmail.com o slack

 Se espera la entrega de un Notebook con el formato .ipynb y el siguiente nombre:

- **0924_SupML_GrupoX**
- Se puede hacer una sola entrega por grupo, **los grupos son los mismos que en el TFM. No se permiten cambios.**

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).

2 MICROSOFT MALWARE PREDICTION



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



CORRECCIÓN

50%

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



AUTOMATIZACIÓN

25%

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



EXPLICABILIDAD

25%

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

¡SUERTE!

Guillem Sitges i Puy

Evaluación

Proyecto de evaluación del módulo de aprendizaje supervisado