

SEGUNDO PARCIAL

Esta prueba consiste en la aplicación del proceso completo de Machine Learning a un problema particular. Se ha de realizar en forma domiciliaria, debiendo entregar el trabajo en la tarea indicada más abajo.

Requerimientos generales a ser cumplidos en este trabajo:

Etapas del proceso

- **Análisis del problema planteado – contexto del problema.**
- Identificación del tipo de problema (ej: supervisado / no supervisado, regresión, clasificación, agrupación)
- Análisis de los datasets y preparación de los datos (todas las actividades que sean necesarias y convenientes)
- Respuestas a las preguntas planteadas, con justificación correspondiente.
- Identificación de **más de un (1+)** enfoques de algoritmos de ML para resolver el problema.
 - Descripción de configuración de parámetros para cada modelo/algoritmo (justificación de cada una)
 - NOTA: **¡DEBEN USARSE ALGORITMOS VISTOS DURANTE EL CURSO!**
- **Descripción** del proceso / modelo creado y ejecución (RapidMiner y Python).
 - **Descripción** de las técnicas usadas para entrenamiento / prueba / validación (RapidMiner y Python)
- Análisis de resultados (RapidMiner y Python)
 - Matrices de confusión
 - Predicciones y estadísticas de confianza
 - Curvas ROC y otros parámetros de evaluación pertinentes.
 - Selección de mejor modelo / proceso y justificación.
- Posibles mejoras a evaluar:
 - Ejemplos: Ajustes de parámetros, Otros algoritmos más apropiados, Otras actividades de pre-procesamiento

Se pide documentar y remitir a la tarea correspondiente de la webasignatura:

- Documento con cada uno de estos requerimientos
- Procesos completos de RapidMiner generados, completamente funcionales
- Programas en Python que implementen los mismos procesos generados con RapidMiner
- Planillas comparativas de resultados de procesos en RapidMiner y Python
- Planillas electrónicas eventualmente utilizadas

REMITIR A LA TAREA “SEGUNDO PARCIAL 2023” un archivo (1) comprimido conteniendo todo lo realizado (documentos, modelos, planillas, datasets, etc.), hasta el día martes 5 de diciembre a las 18:00 horas.

EJERCICIO.

Contexto

De 1934 a 1963, San Francisco fue tristemente célebre por albergar a algunos de los criminales más célebres del mundo en la ineludible isla de Alcatraz.

Hoy, la ciudad es más conocida por su escena tecnológica que por su pasado criminal. Pero, con el aumento de la desigualdad económica y la escasez de vivienda, la delincuencia no escasea en la zona de la bahía.

Este dataset ofrece casi 12 años de informes sobre delitos en todos los barrios de San Francisco.

Dada la hora y la ubicación, **debes predecir la categoría de delito que se ha producido.**

Datos e información:

Para este ejercicio se utilizan datos reales de criminalidad de la ciudad de San Francisco, disponibles en

[San Francisco Crime Classification | Kaggle](https://www.kaggle.com/competitions/sf-crime/overview) , <https://www.kaggle.com/competitions/sf-crime/overview>

Se proveen en el portal de Kaggle 2 datasets, uno de entrenamiento y otro de test. Pueden descargarse de

[San Francisco Crime Classification | Kaggle](https://www.kaggle.com/competitions/sf-crime/data), <https://www.kaggle.com/competitions/sf-crime/data>

Información más detallada (y datasets más actualizados, pero mucho más grandes) puede ser vista en el sitio oficial de la ciudad de San Francisco:

[Police Department Incident Reports: Historical 2003 to May 2018 | DataSF | City and County of San Francisco \(sfgov.org\)](https://data.sfgov.org/Police-Department-Incident-Reports-Historical-2003-to-May-2018)