

Evaluación - R y Python

Luz Frias y Carlos Gil Bellosta

Introducción

Para evaluar la asignatura de programación en R y Python es necesario entregar un proyecto en R y otro en Python, por grupo de TFM. Cada proyecto se entregará al profesor que cursó la asignatura.

Para este proyecto, podéis elegir una de estas dos opciones:

- Avanzar con vuestro TFM (recomendable)
- Utilizar algún conjunto de datos de vuestro interés de la [web de datos abiertos de Madrid](#). Puede ser interesante hacerlo así si aún no tienes los datos para tu TFM, o son de carácter privado.

La parte de R y de Python pueden ser iguales o tener diferencias, a vuestra elección.

Criterios de evaluación

La evaluación se centrará en estos puntos. Si algún TFM o conjunto de datos no requiere nada o poco esfuerzo en algún punto, no se penalizará, pero el resto de puntos cobrarán mayor peso:

1. Obtención de los datos

En algunos casos será simple, p.e. si los datos están en CSV. En otros puede ser más complejo, p.e. si se obtienen vía API o web scraping.

2. Tratamiento y limpieza

El conjunto de datos debe quedar adecuado para su uso. Esto quiere decir, que tenga nombres de columna representativos y usables (minúsculas, sin caracteres extraños, ...) y que los tipos de dato sean los adecuados (fechas, números, ... según aplique).

Según el objetivo perseguido por cada proyecto, puede que sea interesante la eliminación de filas con muchos nulos o la imputación de estos.

3. Enriquecimiento de datos

Según cada caso, puede tener sentido incorporar información de otras fuentes que enriquezcan el conjunto de datos principal. Esto, generalmente, implica obtener, tratar y limpiar la nueva fuente, y su cruce con los datos originales.

4. Análisis exploratorio

Deberá presentarse en un informe (Rmarkdown o notebook de jupyter). Deberás plantear una serie de preguntas básicas sobre tu dataset y combinar texto explicando los pasos con los gráficos obtenidos.

Algunas posibles preguntas son:

- ¿Cuál es la distribución de ciertas variables?
- ¿Cómo se relacionan estas variables entre sí?
- ¿Cómo evoluciona una variable a lo largo del tiempo?

- ¿Cómo se distribuyen geográficamente (en un mapa)?

Tu informe debe poder ejecutarse hasta el final sin errores en una sesión o kernel nuevos. Asegúrate de comprobarlo antes de enviar el proyecto. Un informe que no termina de ejecutarse porque salta un error supone el suspenso en la práctica.

Ten cuidado también con las rutas a ficheros. Deben ser relativas, y apuntar a ficheros contenidos en tu entregable. Si tienes una ruta del estilo `c:/users/pepa/blabla.dat`, fallará en otro ordenador y será un suspenso.

Ejemplos

Como ejemplos, puedes fijarte en:

- [Estructura salarial en España](#)
- [Tiroteos en EEUU](#)

Fíjate especialmente en cómo se combinan los textos explicativos y los gráficos. La extensión no es relevante.

Opcional

GitHub

Os animamos a aprender lo básico sobre git y a crear y subir vuestro proyecto a un repositorio en GitHub (u otra alternativa como GitLab, Bitbucket, ...).

Si se hace así, deberás compartir con el profesor la url del repositorio. Recuerda que debe haber un repositorio diferente por proyecto (es decir, uno para R y otro para Python).

En este caso, se asignará +1 punto extra.

Entregable

El proyecto final entregado debe contener lo siguiente:

- El código
- Los datos, si aplica.
- Una descripción de cómo ejecutar el proyecto desde cero. Puede que sea un único informe (Rmarkdown o notebook de jupyter) o que implique ejecutar ciertos pasos antes (p.e. `01_lee_datos.[R|py]`, `02_limpia.[R|py]`, ...)
- Una breve descripción del conjunto de datos utilizado y objetivo final. Puede estar incluido al principio del análisis exploratorio, o en un documento de texto externo.
- El listado de alumnos que formáis el grupo.

Si utilizas `git`, el entregable debe ser únicamente la url del repositorio. Los datos deben estar también en él, y las descripciones y listado de alumnos deben estar detallados en el `README.md`.

Plazo de entrega

El plazo de entrega tiene como fecha límite el domingo 17-feb a las 23:59.