Taller Machine Learning

John González

Marzo 2021

El problema que se pretende resolver consiste en el pronóstico del desempeño de los estudiantes que presentaron la prueba Saber Pro en el año 2019 en Bogotá. En otras palabras, a partir de las características de los estudiantes se pretende estimar su desempeño en la prueba.

Tareas a desarrollar

1. **Descargue** los **datos** disponibles en este enlace.
2. **Realice un listado de las variables**. Si su tabla se llama *df*, puede usar el comando:

```
list(df.columns)
```

   para listarlas, observará que hay en total 106 variables.

3. **Tome una sola variable de salida**, esta puede ser seleccionada de cualquiera de aquellas que comiencen con: *MOD*, *PUNT* o *PERCENTIL*.
4. **Tome como variables de entrada (predictivas) máximo diez** que comiencen con: *ESTU*, *FAMI* o *INST*.
   Nota 1: Se recomienda máximo diez variables para que la actividad no se extienda más de la cuenta, sin embargo, puede tomar las que considere necesarias (incluso todas o solo una).
   Nota 2: No utilice como variables de entrada cualquiera que mida el desempeño del estudiante (dadas en el ítem anterior), porque en un caso práctico usted no dispondrá de estas antes de la prueba.
5. **Realice un análisis descriptivo** de las variables seleccionadas en el ítem anterior. Se recomienda realizar este análisis:

   (a) realizando un análisis individual de cada variable.

   (b) confrontando cada variable de entrada con la de salida.

   Nota: Este análisis debe incluir posibles problemas por falta de simetría, datos atípicos o desbalance de clases, además escriba explícitamente su posible corrección.

6. Si es necesario, **transforme las variables** de acuerdo al análisis anterior. Para esto puede utilizar i) el `pipeline` de `sklearn`, ii) el `mapper` de `sklearn-pandas` o iii) simplemente construir una nueva tabla con los datos transformados.

7. Realice al menos dos modelos de pronóstico, evalúe el desempeño de cada uno y compárelos. Nota: ¿Qué tan bueno es su mejor modelo?

8. Bono: Intente realizar un modelo de ensamble, posiblemente esto mejorará el desempeño.

9. Bono: Realice el pronóstico de usted mismo como estudiante. Nota: Alimente el pronóstico con sus propios datos cuando era estudiante de último semestre de la universidad, si no estudió en Bogotá, simule que lo hizo.

Que se debe entregar

Comparta un notebook en formato `.ipynb` con el desarrollo de la actividad. En cada resultado (gráfico o valor numérico) escriba explícitamente su análisis. Este análisis se tendrá muy en cuenta en el momento de la calificación.

Si lo desea, puede trabajar en grupo de máximo 3 estudiantes, comparta una sola entrega en la plataforma y escriba los integrantes.

Fuente

Los datos se descargaron directamente de la página del icfes mediante consulta en el aplicativo `ftpicfes` (también están disponibles en datos.abiertos.gov.co) y se filtraron usando `ESTU_PRGM_MUNICIPIO == 'BOGOTÁ D.C.'`