APRENDIZAJE AUTOMÁTICO

Evaluación Parcial

Etapa 2: Descripción del Dataset y Origen

Profesor Lic. Martin Mirabete
Alumna: Demari Monica Valeria

- **En esta segunda entrega, debe tener acceso al dataset que utilizará en su proyecto de Aprendizaje Automático.**
- **Proporcione una descripción completa del dataset, incluyendo la cantidad de instancias, características (columnas), tipos de datos, y cualquier información relevante.**
- **Informe sobre el origen del dataset, es decir, de dónde provienen los datos. Esto puede incluir la fuente, la fecha de adquisición y cualquier proceso de recopilación o preprocesamiento que haya realizado.**
- **Se recomiendo utilizar dataset de dominio público (ipiec, datos.gob.ar, etc).**
- **Esta entrega debe ser realizada en el repositorio GIT, incluyendo archivos de dataset y documentos.**

NOMBRE DEL DATASET

Para el desarrollo del proyecto de Aprendizaje Automático utilizaré el archivo Excel llamado “dataset_jcyd_ok.xlsx”

INFORME SOBRE EL ORIGEN DEL DATASET

Con el fin de avanzar en el proyecto de optimización de la asignación y clasificación de docentes en el nivel secundario, presenté una nota a la Secretaría de Gestión Educativa del Ministerio de Educación con el propósito obtener el set de datos en posesión de la Junta de Clasificación y Disciplina de Educación Secundaria, específicamente aquellos relativos a los títulos de los aspirantes inscriptos en la ciudad de Río Grande.

La secretaria de Gestión Educativa, Profesora Silvina Solohaga autorizó mi solicitud, con la condición de garantizar la debida reserva de los datos individuales y particulares de los docentes. Conforme a esta autorización, el equipo informático, bajo la supervisión de los miembros de la Junta, proporcionó el archivo Excel denominado "dataset_jcyd_ok.xlsx", que recopila la totalidad de la información hasta el último día de inscripción, fijado el 30 de junio de 2024.

DESCRIPCIÓN DEL DATASET

En el Jupyter "Descripción del dataset", se procedió cargar el dataset para dar inicio del proyecto de Aprendizaje Automático, y como información podemos destacar que este set está conformado por 52936 registros y 16 columnas.

El DataFrame contiene varias columnas y muestra el conteo de entradas no nulas (non-null) para cada una:

Orden	Nombre de columna	Entradas	Tipo de datos	Descripción
1.	idespacio	52,936	tipo entero (int64).	Clave de identidad del Espacio curricular
2.	Idtitulo	52,936	tipo objeto (string).	Clave de identidad del Título
3.	idtitulo_real	52,936	tipo entero (int64).	Clave de identidad del Título conformado entre los datos del título y la casa de estudio
4.	Idcasaestudio	52,936	tipo objeto (string)	Clave de identidad de la casa emisora del título
5.	Carácter	52,936	tipo objeto (string).	Nombre que se clasifican los títulos en el espacio curricular
6.	idNivel_espacio	52,935	tipo flotante (float64) - una entrada es nula.	Clave de identidad del Nivel del Espacio curricular o cargo
7.	desc_espacio	52,935	tipo objeto (string) - una entrada es nula.	Descripción del Cargo o Espacio curricular
8.	tipo_espacio	52,935	tipo objeto (string) - una entrada es nula.	Tipo de cargo o espacio curricular
9.	ciudad_espacio	52,935	tipo flotante (float64) - una entrada es nula.	La ciudad que se encuentra habilitado el cargo o espacio curricular
10.	resolucion_espacio	52,935	tipo objeto (string) - una entrada es nula.	Resolución del plan de estudio del título relacionado al cargo o espacio curricular
11.	titulo	52,811	tipo objeto (string) - 125 entradas son nulas.	Nombre del título
12.	resolucion	50,811	tipo objeto (string) - 2,125 entradas son nulas.	Resolución del plan de estudio del título

13.	nombre	52,695	tipo objeto (string) - 241 entradas son nulas.	nombre de la casa de estudio que emitió el título
14.	facultad	14,125	tipo objeto (string) - 38,811 entradas son nulas.	facultad de la casa de estudio que emitió el título
15.	provincia	50,367	tipo objeto (string) - 2,569 entradas son nulas.	provincia de la casa de estudio que emitió el título
16.	ciudad	42,042	tipo objeto (string) - 10,894 entradas son nulas.	ciudad de la casa de estudio que emitió el título

Por lo detallado en la tabla, el DataFrame tiene dos tipos de datos flotantes, dos enteros y doce objetos (cadenas de texto)

Las columnas que el tipo de datos es numérico se identifican como claves ID, por lo que es irrelevante proceder a hacer las estadísticas descriptivas de dichas columnas.

Las estadísticas descriptivas de las columnas categóricas presentan información sobre la cantidad de datos y la variedad de categorías en cada columna. De acuerdo a lo brindado en el Jupiter Notebook, puedo resaltar las siguientes observaciones más relevantes a mi investigación:

- Se distinguen tres categorías únicas, Docente “D”, habilitante “H” y supletorio “S”, siendo que la que mas prevalece es la habilitante con 20355 casos.
- En la columna descripción de espacios curriculares o cargos, se registran 468 descripciones únicas, destacando "Preceptor/a" como la más frecuente con 4648 casos de un total de 52935 registros.
- Con respecto a los títulos, figuran 2375 títulos únicos, y el más común es "PROFESOR/A DE EDUCACIÓN FÍSICA".

Como observación de la data set, se detecta que existen valores nulos, como así también valores faltantes, razón por la cual, en la próxima instancia se debería realizar un análisis más avanzado como limpieza de datos o correlaciones