

Vamos a empezar a pensar en el tipo de proyecto que estamos trabajando.

Ahora ya los conoces y puedes trabajar con los siguientes:

- Proyecto 1 - descriptivo / dashboard
- Proyecto 2 – exploratorio/investigativo:
 - Comparativo
 - Relacional
 - Causal
- Proyecto 3 – clasificador
- Proyecto 4 – regresión
- Proyecto 5 – clustering
- Proyecto 6 – reducción dimensional
- Proyecto 7 – reglas de asociación
- Proyecto 8 – series temporales

Te propongo los siguientes datasets:

- <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
- <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>
- <https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists>
- <https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement>
- <https://www.kaggle.com/datasets/ayessa/salary-prediction-classification>

Elige alguno de ellos y ubícalo en uno o máximo dos proyectos tipo.

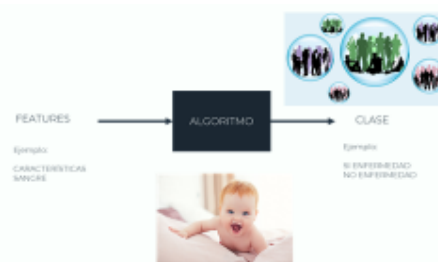
1. ¿En qué tipo de proyecto te estás enfrentando? Marca máximo dos tipos de proyectos.

- Proyecto 1 – exploratorio/investigativo:
 - Comparativo
 - Relacional
 - Causal
- Proyecto 2 – clasificador
- Proyecto 3 – regresión
- Proyecto 4 – clustering
- Proyecto 5 – reducción dimensional
- Proyecto 6 – reglas de asociación
- Proyecto 6 – series temporales

¡Excelente!

El siguiente paso es pensar en cómo va a ser la tabla de datos.

Sigue con el mismo dataset y trata de realizar el siguiente ejercicio:



SLY	SLAM	SLAE	SLAF	SLAG	SLAH	SLAY	SLBG	SLBH	SLBI	SLBM	SLBN	UNKNOWN
0.79	16.66	15.11	14.63	14.75	15.28	17.15	16.52	15.31	15.45	23.88	17.52	NAO
0.71	12.47	16.68	15.11	16.86	12.46	18.96	12.17	13.25	13.31	21.11	16.48	NAO
0.14	16.66	15.00	15.51	12.16	11.36	18.94	12.14	11.63	15.69	23.91	15.52	SPN,Negative
0.85	15.06	16.68	15.00	11.14	13.61	17.24	12.14	11.11	13.52	17.81	16.96	SPN,Negative
0.92	15.36	16.11	16.16	15.49	15.86	16.46	11.51	16.66	12.00	19.13	19.11	SPN,Negative
0.48	16.19	15.87	15.11	15.48	19.19	16.40	12.10	12.14	16.26	21.46	17.88	SPN,Negative
0.48	16.19	15.47	15.31	12.45	19.19	16.40	12.10	12.14	16.26	21.46	17.88	SPN
0.72	14.34	15.33	16.66	19.16	11.36	17.96	19.17	19.41	14.39	14.62	18.35	SPN,Negative
0.48	15.33	16.17	17.14	17.83	12.87	19.37	11.47	11.21	14.72	26.76	26.82	SPN
0.48	15.33	16.17	17.14	15.50	12.87	19.37	11.47	11.21	14.72	26.76	26.82	SPN,Negative
0.83	12.49	15.87	15.19	16.66	16.68	18.36	11.63	15.87	15.76	19.81	17.43	SPN,Negative
0.98	10.17	15.54	25.30	30.19	11.14	27.25	26.89	15.37	19.34	31.25	26.79	UNKNOWN
0.22	14.37	17.88	16.80	17.73	9.17	22.80	13.33	10.34	15.48	16.88	15.11	SPN,Negative
0.25	12.71	15.32	14.64	14.55	11.96	19.09	10.35	11.11	15.94	20.75	15.15	SPN,Negative
0.31	15.63	15.14	15.64	19.50	11.14	19.28	19.58	16.81	17.55	22.19	17.45	SPN,Negative
0.28	16.30	14.77	16.08	12.18	11.52	17.73	11.83	14.66	16.51	19.71	17.11	SPN,Negative
0.87	16.79	16.74	17.11	15.16	9.11	21.76	15.36	19.56	15.58	14.62	16.30	NAO
0.83	12.10	12.30	12.31	9.11	9.96	20.45	10.71	10.37	18.33	21.36	20.11	NAO,Neutral

- Variables de entrada = características que nos ayudarán a identificar la clase. Pueden ser numéricas como cualitativas.
- Variable de salida = CLASE (cualitativa)
- Qué significa cada observación = fila
- Cómo las etiquetamos (agrupamos previamente) = clase
- **¡¡¡El proyecto más común!!!**

2. Te planteo que piensas en **la tabla de datos**, que la dibujes y que ordenes las ideas de la siguiente manera:

Características clave de tu tabla de datos

Responde a las siguientes preguntas para plantear correctamente la estructura de tu clasificador.

Tu variable respuesta o variables respuesta

Stroke

Tus variables de entrada. Features/Inputs

Gender, age, hypertension, heart_disease,
ever_married, work_type, Residence_type,
avg_glucose_level, bmi, smoking_status

¿Qué son las filas u observaciones?

Cada uno de los registros es un paciente

¿Qué grupos de variables o factores vas a incluir? Estos factores son los que nos darán información de tu respuesta.	Características sociodemográficas: • Sexo (gender) • Edad (age) • Ha estado casado (ever_married) • Tipo de trabajo (work_type) • Entorno rural/urbano (Residence_type) Condicionantes de salud: • Hipertensión (hypertension) • Enfermedad cardíaca (heart_disease) • Nivel de glucosa (avg_glucose_level) • Índice de masa corporal (bmi) • Tabaquismo (smoking_status)
Lista las variables de entrada que crees que pueden tener un impacto importante	• Edad • Sexo • Hipertensión • Enfermedad cardíaca • Nivel de glucosa • Índice de masa corporal • Tabaquismo

¡Gran trabajo!

Ahora ya conoces el tipo de proyecto que estás trabajando.

También conoces cómo están estructuradas las variables, cuál es la respuesta y las variables de entrada.

Fíjate que nos estamos enfocando realmente bien

¡Seguimos por buen camino!

Celébralo como más te guste.

Salgo a correr y hago ejercicio para celebrar esta hoja de trabajo ;)

1 abrazo y hasta pronto. Brindo por tus éxitos :)

Jordi Ollé