

DEFINE TU ESTRATEGIA

Cómo definir la estrategia de tu proyecto

Estos son los pasos para definir la estrategia:

1. Plantear el problema y sus objetivos
2. Nivel de la pirámide dónde nos enfocamos
3. Proyecto tipo que te encuentras
4. Definir la estructura de tu tabla de datos y los grupos de variables
5. Definir las variables con detalle
6. Define la estructura entradas - salida si aplica

En esta hoja de trabajo vas a definir la idea de tu proyecto pasando por 6 pasos que utilizo en las sesiones estratégica de mis alumnos y clientes.

Se trata de ver el problema de forma global y tratar de llevarlo a una estructura entradas – salida. Si no se puede tratar de identificar otro proyecto tipo que puedas aplicar.

¡A por ello!

1. Plantear el problema y sus objetivos

Contesta a las siguientes preguntas para definir el problema de tu proyecto:

¿Qué problema quieres resolver? Explícalo con palabras sencillas para que te todo el mundo lo pueda entender. Cuando más te alejes del lenguaje técnico mejor.

Se quiere conocer si una serie de hábitos de vida y de condicionantes personales y de salud están relacionadas con el hecho de sufrir o no un ictus.

El objetivo principal es reconocer patrones y predecir el ictus a partir de una serie de hábitos de vida y condicionantes personales y de salud. También se va a comparar los hábitos de vida y condicionantes personales y de salud en función del género, la edad y de si han sufrido o no un ictus.

¿Por qué es importante resolver este problema?

Una vez conocidos los factores de riesgo de padecer un ictus, se puede trabajar la prevención de aquellos que pueden ser revertidos por el paciente, como los hábitos de vida.

Por otro lado, también puede ser de ayuda a la hora de planificar pruebas diagnósticas con más frecuencia en aquellos pacientes con mayor riesgo de ictus e incluso iniciar tratamiento preventivo.

A nivel de salud pública, es importante planificar estrategias tanto enfocadas en la prevención como en la detección de un ictus para poder actuar en la mayor brevedad posible, ya que el tiempo de respuesta es un factor que predice la recuperación del paciente.

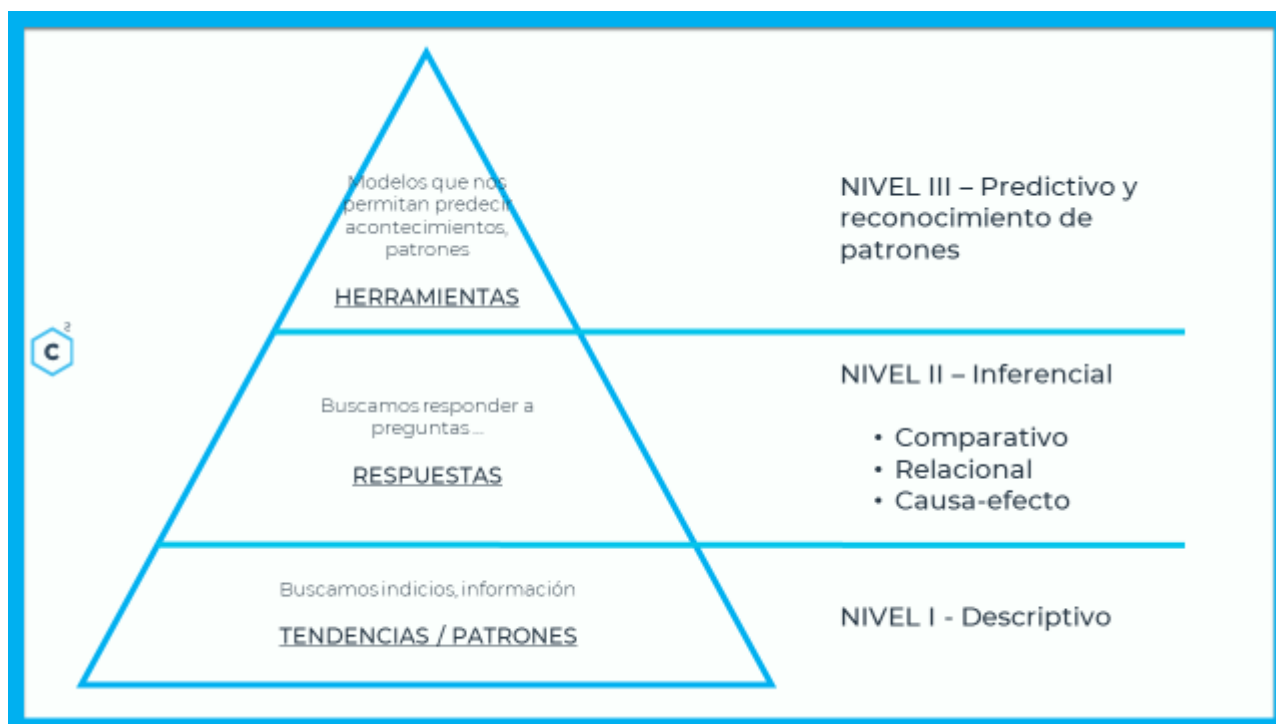
¿Qué se ha trabajado hasta hoy sobre ello? ¿Qué otros proyectos has podido ver que se parezcan al tuyo? Lista temas que has encontrado ya sea dentro de tu empresa o fuera en investigaciones o aplicaciones que hayas podido ver en la red.



- Se han estudiado algunos factores de riesgo, tanto factores personales, hábitos de vida, condicionantes de salud y predisposición genética.
- Tratamientos que pueden ayudar a prevenir ictus en personas con riesgo de padecerlo (antidiabéticos, antitrombóticos, etc.).

2. Nivel de la pirámide dónde nos enfocamos

Ya sabes que hay 3 niveles de extracción de información de los datos y quiero que pienses en cada uno de estos tres niveles para enfocar tu proyecto y tus esfuerzos.



Usa la tabla para hacerlo:

ETAPA DE ANÁLISIS DE DATOS	¿QUÉ VAS A HACER EN CADA ETAPA? Escribe lo que se te ocurra sin demasiados filtros.
Descriptivo. El objetivo es lograr explicar los datos con gráficos y estadísticos descriptivos que nos permitan encontrar tendencias y empezar a sacar valor de los datos.	Describir las variables en función del género, en función de la edad (categorizada en grupos) y de si han sufrido o no un ictus mediante estadísticos descriptivos y gráficos en función de la naturaleza de las variables.
Inferencial. Recuerda que aquí puedes comparar, puedes relacionar variables o incluso crear modelos entradas - salida que nos permiten obtener patrones clave de los datos.	Calcular los diferentes pruebas estadísticas para valorar si las relaciones tienen o no una significación estadística. Y, analizar mediante un modelo de regresión logística qué variables pueden predecir el ictus.



Reconocimiento de patrones. Puedes solucionar problemas de: • Clasificación • Regresión • Clustering - agrupado de datos • Reglas de asociación • Reducir dimensiones • Predicción de series temporales	Mediante Machine Learning, entrenar un modelo con algoritmos supervisados de clasificación para predecir el ictus.
---	--

Marca qué etapa es la más importante para tu proyecto;

- o Descriptivo – si solo quiero ver patrones a nivel gráfico
- o Inferencial – si pretendo contestar hipótesis de investigación y comprender qué factores influyen en una determinada respuesta
- o **Reconocimiento de patrones – si pretendo crear una herramienta capaz de predecir, encontrar grupos similares, reducir dimensiones o reglas de asociación.**

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

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



4. Definir la estructura de tu tabla de datos y los grupos de variables

FEATURES

Ejemplo:
CARACTERÍSTICAS
MAYORES

ALGORITMO

CLASE

Ejemplo:
SI ENFERMEDAD
NO ENFERMEDAD

VARIABLES ENTRADA
Características

SALIDA
Clase

SEX	AGE	HYP	HEART	EVER	WORK	RES	AVG	BMI	SMOK	STROKE
1	55	1	1	1	1	1	120	25	1	1
2	60	0	0	0	0	0	110	22	0	0
3	65	1	1	1	1	1	130	28	1	1
4	70	0	0	0	0	0	100	20	0	0
5	75	1	1	1	1	1	140	30	1	1
6	80	0	0	0	0	0	90	18	0	0
7	85	1	1	1	1	1	150	32	1	1
8	90	0	0	0	0	0	80	15	0	0
9	95	1	1	1	1	1	160	35	1	1
10	100	0	0	0	0	0	70	12	0	0

- 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!!!

Características clave de tu tabla de datos

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

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)



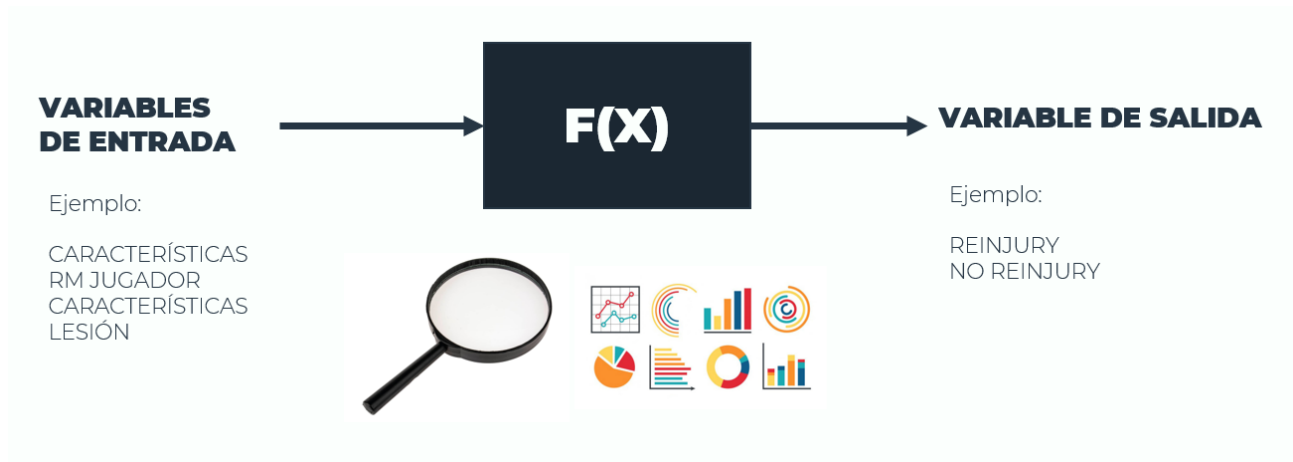
	• Nivel de glucosa (avg_glucose_level) • Índice de masa corporal (bmi) • Tabaquismo (smoking_status)
Lista las variables de entrada que crees que pueden tener un impacto importante	• Edad • Sexo • Hipertensión • Enfermedad cardíaca • Nivel de glucosa • Índice de masa corporal • Tabaquismo

Recomendable: dibujar la tabla de datos y pensar en una estructura entradas – salida.

5. Define tus variables con detalle

	Nombre Variable	Código Cómo la escribes en R o Python o SPSS	Grupo. Relativa a... o si es la respuesta	Tipo Variable	Tipo variable en R /Python	Unidades Cualitativa indica los campos Cuantitativa indica las unidades	Explicación
1	Identificador	id		Númerica	Númerica		Identificación del paciente
2	Sexo	gender	Sociodemográficas	Nominal	Factor	Male, Female and Others	Sexo del paciente
3	Edad	age	Sociodemográficas	Númerica	Númerica	años	Edad del paciente
4	Hipertensión	hypertension	Salud	Nominal binaria	Factor	0 = "No HTA"; 1 = "HTA"	Indica si el paciente tiene o no hipertensión
5	Enfermedad cardíaca	heart_disease	Salud	Nominal binaria	Factor	0 = "No enfermedad cardíaca"; 1 = "Enfermedad cardíaca"	Indica si el paciente tiene o no enfermedad cardíaca
6	Ha estado casado	ever_married	Sociodemográficas	Nominal binaria	Factor	Yes; No	Indica si ha estado casado alguna vez
7	Tipo de trabajo	work_type	Sociodemográficas	Nominal	Factor	Children; Govt_job; Never_worked; Private; Self-employed	Tipo de trabajo que realiza
8	Entorno rural/urbano	Residence_type	Sociodemográficas	Nominal binaria	Factor	Rural; Urban	Indica si reside en un entorno urbano o rural
9	Nivel de glucosa	avg_glucose_level	Salud	Númerica	Númerica	mg/dl	Nivel medio de glucosa en sangre
10	Índice de masa corporal	bmi	Salud	Númerica	Númerica	Kg/m ²	Índice de masa corporal
11	Tabaquismo	smoking_status	Salud	Nominal	Factor	Formerly smoked; Never smoked; Smokes; Unknown	Indica si la persona es o ha sido fumadora
12	Ictus	stroke	Respuesta	Nominal binaria	Factor	0 = "No ictus"; 1 = "Ictus"	Si ha padecido o no un ictus

6. Dibuja el esquema entradas – salida



Dibuja tu esquema:



Puedes aplicar este guion a tus datos o a datos abiertos de kaggle por ejemplo:

- <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>

Usa este guion siempre que te enfrentes a un nuevo proyecto y te ayudará a enfocarte en cómo puedes sacar valor de tus datos con un proceso de 6 pasos.

¡Deseando ver tu estrategia ya definida!

1 fuerte abrazo
Jordi Ollé