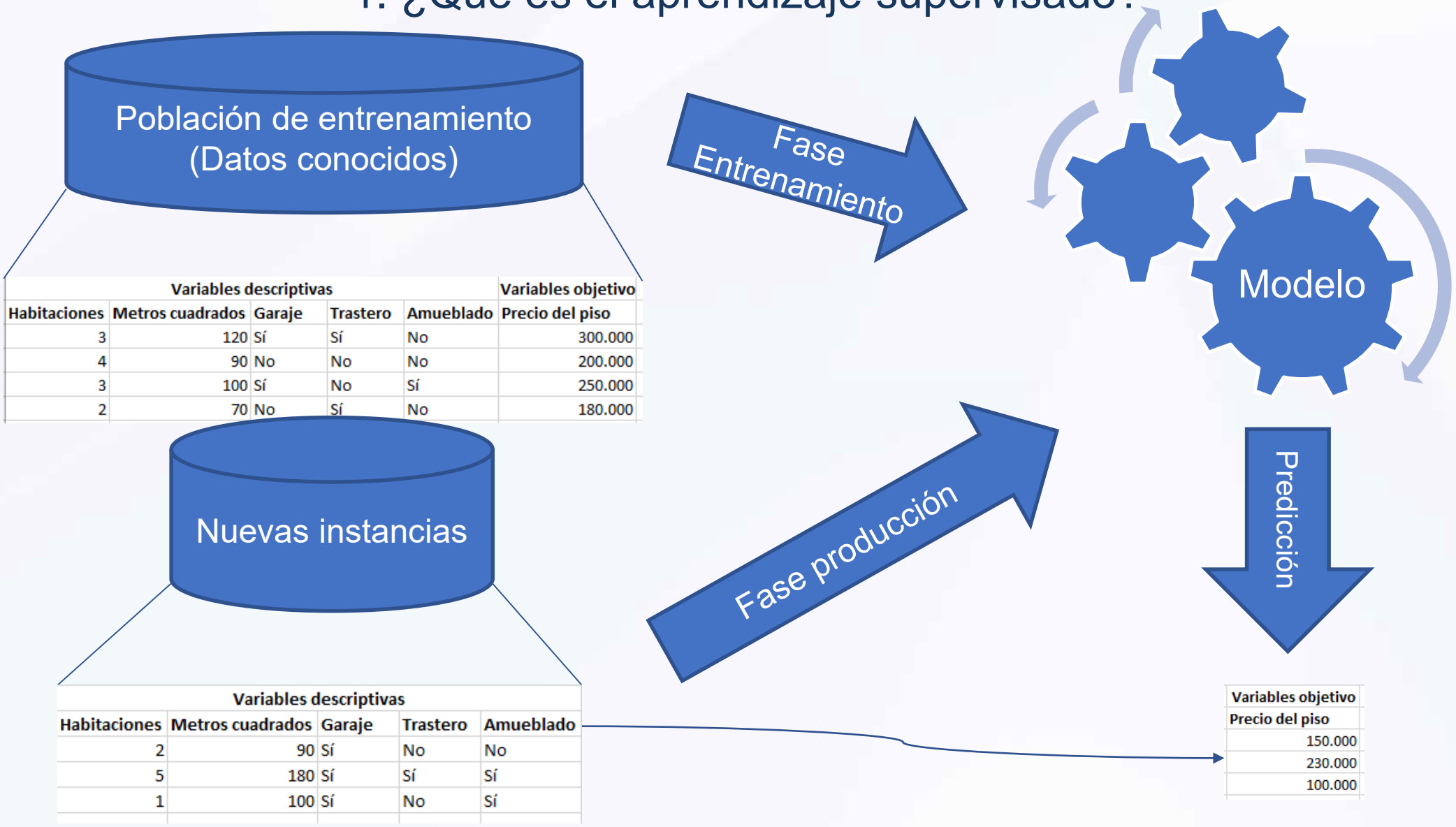
PROYECTO NUDGE

Febrero 2021

Índice

1. ¿Qué es el aprendizaje supervisado?
2. Introducción al proyecto
3. Desarrollo del proyecto
4. Resultados del entrenamiento
5. Predicciones del modelo

1. ¿Qué es el aprendizaje supervisado?



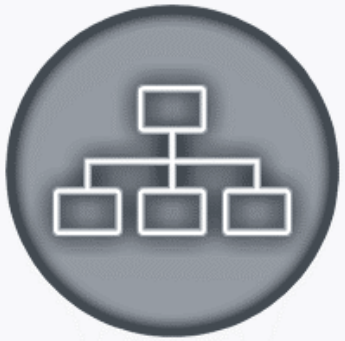
Variables descriptivas					Variables objetivo
Habitaciones	Metros cuadrados	Garaje	Trastero	Amueblado	Precio del piso
3	120	Sí	Sí	No	300.000
4	90	No	No	No	200.000
3	100	Sí	No	Sí	250.000
2	70	No	Sí	No	180.000

Variables descriptivas				
Habitaciones	Metros cuadrados	Garaje	Trastero	Amueblado
2	90	Sí	No	No
5	180	Sí	Sí	Sí
1	100	Sí	No	Sí

Variables objetivo
Precio del piso
150.000
230.000
100.000

2. Introducción al proyecto

- Numerosas actuaciones de control regularizan **modificaciones incorrectas** del **borrador** precargado en **RentaWeb**.
- Esto origina **costes de tramitación** y **pérdidas de recaudación** para la AEAT.

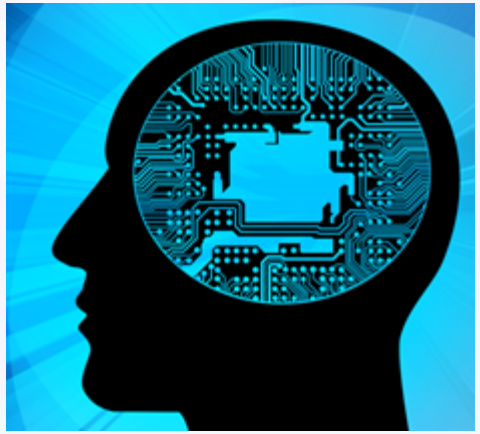


- Por ello, se nos ha propuesto la creación de un **sistema predictivo** para **categorizar** a los **contribuyentes** con mayor probabilidad de **equivocarse**.

2. Introducción al proyecto

Datos

Aprendizaje



Sistema predictivo

- El sistema sería **entrenado** partiendo de los resultados de las **actuaciones de control** realizadas en el pasado, y usaría los **datos fiscales** y otra información conocida de los contribuyentes para establecer los perfiles. Tendría dos usos importantes:
    1. **Prevenir**, mediante mensajes personalizados, **modificaciones incorrectas** en RentaWeb (fase de **asistencia**).
    2. **Guiar** la **selección** de expedientes para tramitar en la aplicación gestora (fase de **control**).

- Es una forma de **rentabilizar** el **conocimiento implícito** en las actuaciones de la **AEAT**.

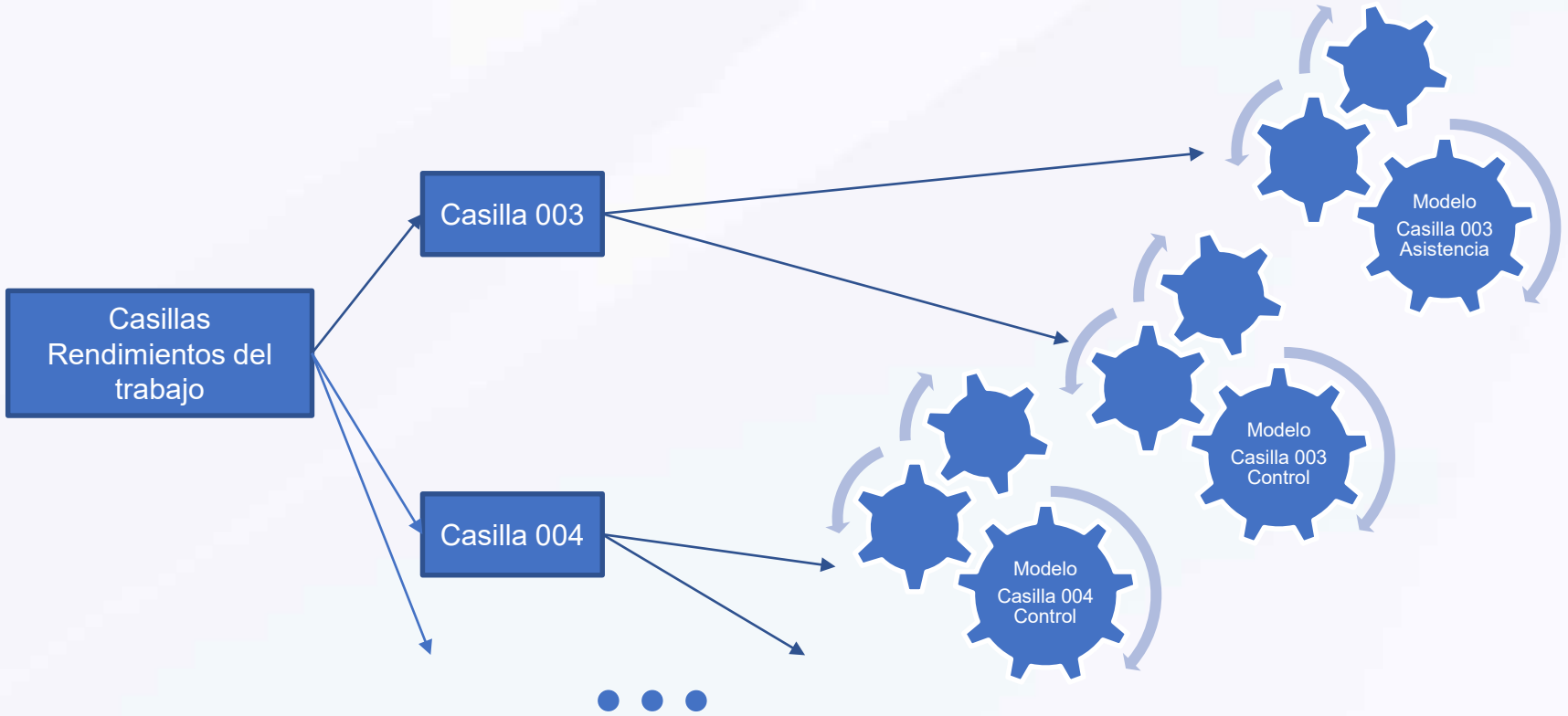
2. Introducción al proyecto. Algunos números.

- Datos del ejercicio 2018 para las tres casillas con más modificaciones del bloque de rendimientos del trabajo.

Casilla	Modificaciones		Comprobaciones	
[003] Retribuciones dinerarias	Número	Importe	Modificación correcta	Modificación errónea
	94.510	498.068.220,22 €	18.918	16.032
[013] Cotizaciones	Número	Importe	Modificación correcta	Modificación errónea
	78.906	115.822.173,82 €	10.493	3.279
[596] Retenciones	Número	Importe (*)	Modificación correcta	Modificación errónea
	78.505	76.590.374,98 €	7.444	3.927

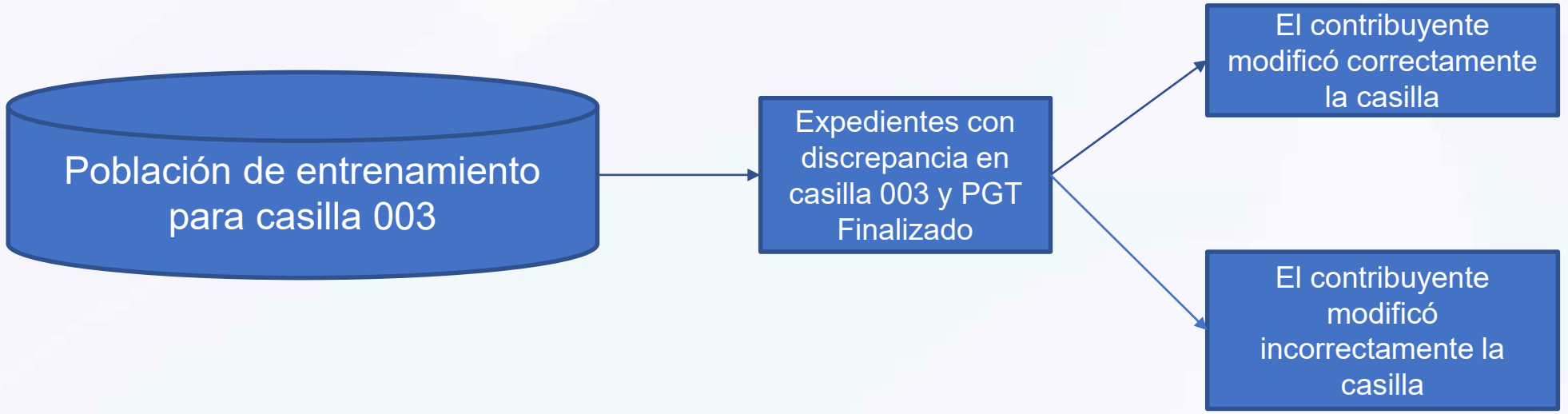
(*) Este importe es en cuota. Los dos anteriores en base.

3. Desarrollo del proyecto. Proyectos Nudge.



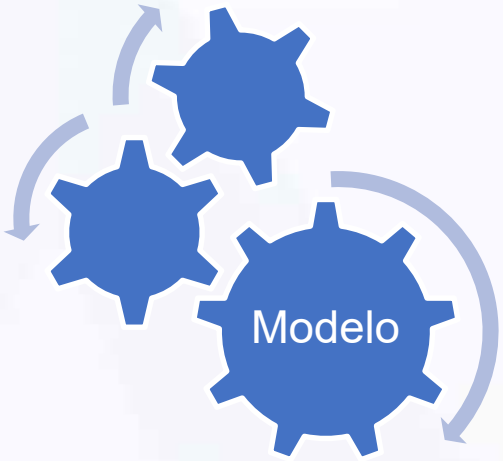
- De momento, tenemos modelos para las 6 casillas de Rendimientos del Trabajo (casillas [003], [004], [005], [006], [0013] y [0596]). En total hemos construido **6 modelos predictivos**.
- Además, se han entrenado dos modelos comunes, que predicen el acierto o error en cualquiera de las casillas del grupo de rendimientos del trabajo.

3. Desarrollo del proyecto. Población de entrenamiento.



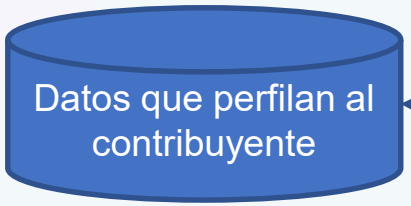
La población de entrenamiento es única para cada casilla. Y común a los modelos de control y asistencia.

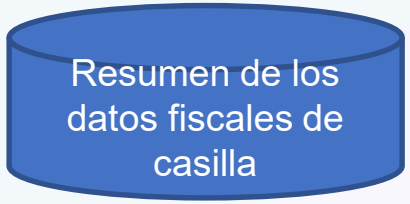
3. Desarrollo del proyecto. Fuentes de Datos.

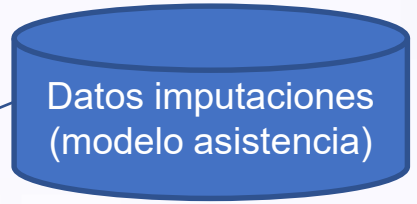


Unas 500 variables en total

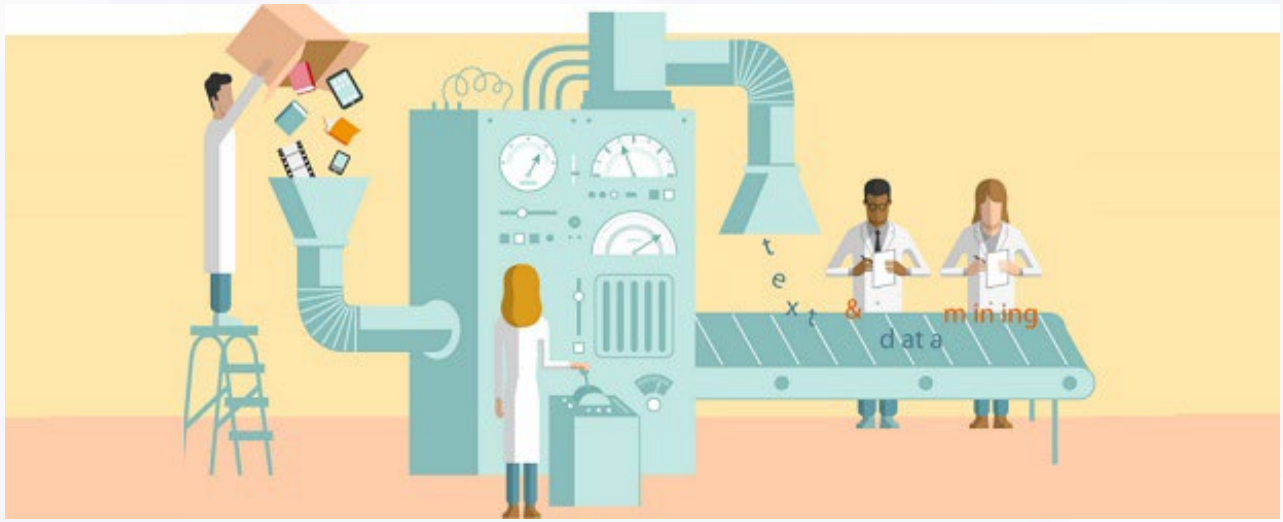
Variables descriptivas						
Perfil contribuyente			Fuentes de la casilla			Variable objetivo
Tiene AAEE	Volumen Rentas	...	Número pagadores	Suma Percepciones	...	Clasificación
Sí	60.000	...	2	30.000	...	Modificación correcta
No	70.000	...	3	50.000	...	Modificación errónea
No	55.000	...	1	45.000	...	Modificación correcta







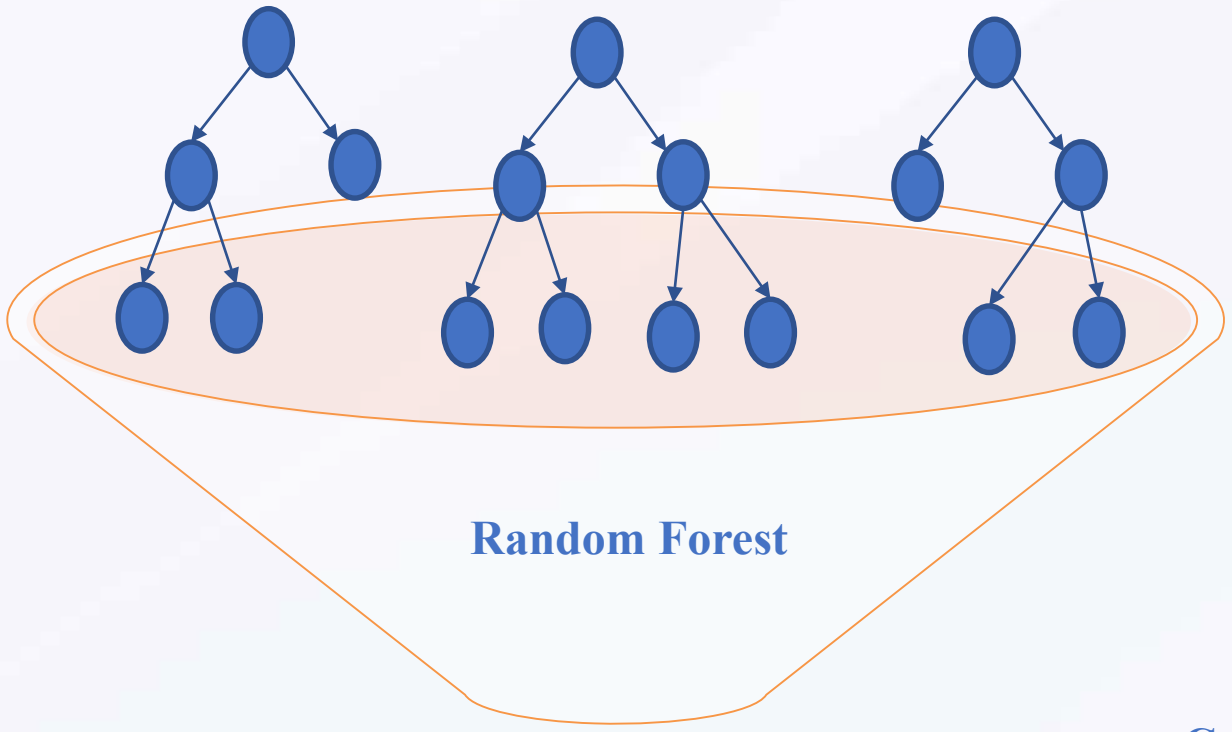
3. Desarrollo del proyecto. Preparación de los datos.



El **objetivo** es procesar los datos anteriores para mejorar la calidad del modelo final. Algunas de las técnicas empleadas son:

- **Generar estadísticos** cuando los datos no están agrupados (caso del 190).
- **Convertir** datos no numéricos a numéricos.
- **Enriquecer** el conjunto de variables con nuevos datos calculados a partir de los anteriores.
- **Eliminar** datos irrelevantes.

3. Desarrollo del proyecto. Entrenamiento del modelo.

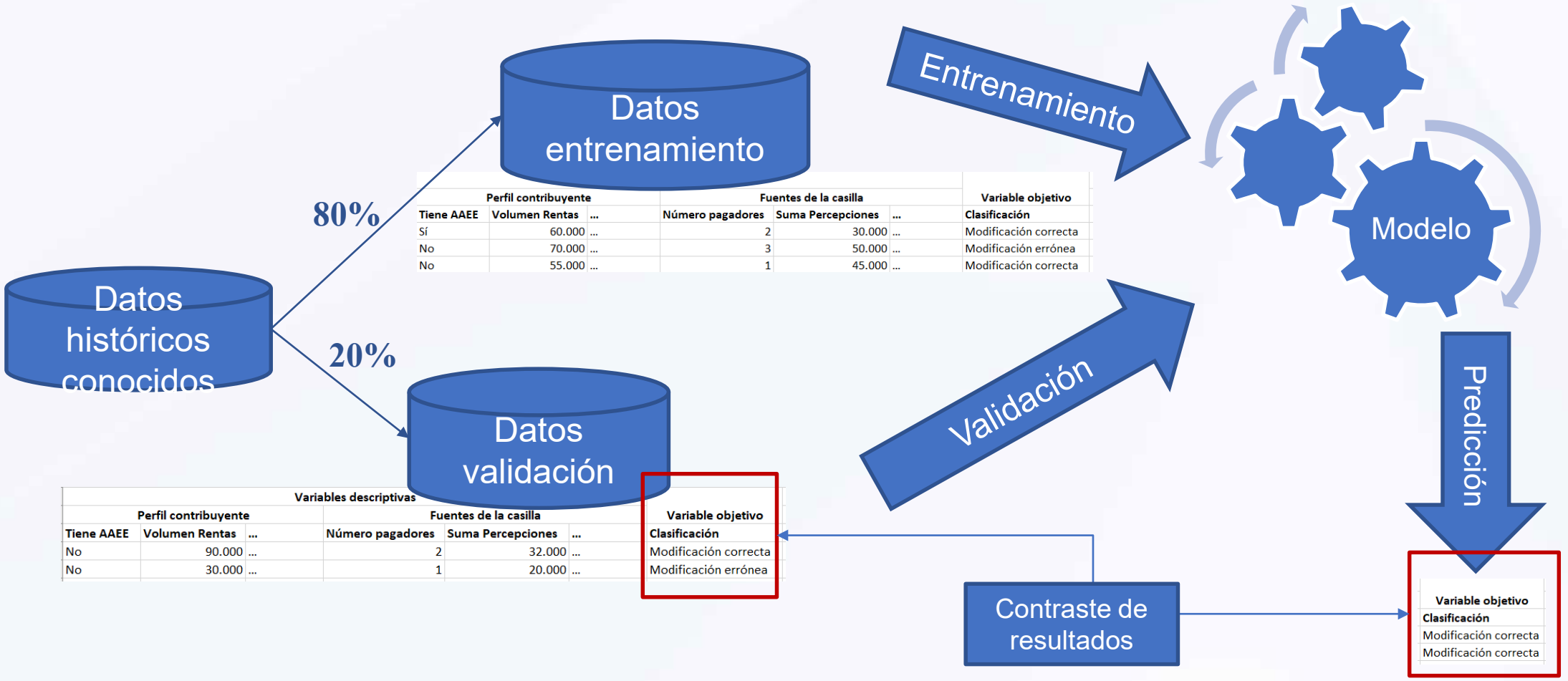


- Se han usado algoritmos de la familia de **random forest**, que son modelos que entrenan numerosos árboles de decisión y los ensamblan en un único modelo final.

- Después de entrenar el modelo, se realiza una **calibración de probabilidad** (es otro modelo de regresión que se superpone al modelo ya existente) para que el output del modelo tenga una interpretación directa como un valor de probabilidad.

4. Resultados del entrenamiento. Calidad de los modelos.

¿Cómo medir la calidad de los modelos?



3. Resultados del entrenamiento. Calidad modelo 003 Asistencia.

		Valor predicho		
		Modificación correcta	Modificación incorrecta	Total
Valor real	Modificación correcta	7.074	1.928	9.002
	Modificación incorrecta	2.463	5.317	7.780
	Total	9.537	7.245	16.782

Cobertura
5317/7780 = 68%

Precisión
5317/7245 = 73%

- Con este umbral de decisión (0,5), se logra 73% de precisión (cuando el modelo predice que el contribuyente se equivoca, acierta el 73% de las veces) y 68% de cobertura (de todas las equivocaciones del contribuyente, el modelo ha recuperado el 68%). El porcentaje de acierto total es del 74%.

- Se puede conseguir mayor precisión a costa de disminuir la cobertura. Por ejemplo, con un umbral de decisión en 0,8, se tendrían:
  - 86% de precisión.
  - 39% de cobertura.

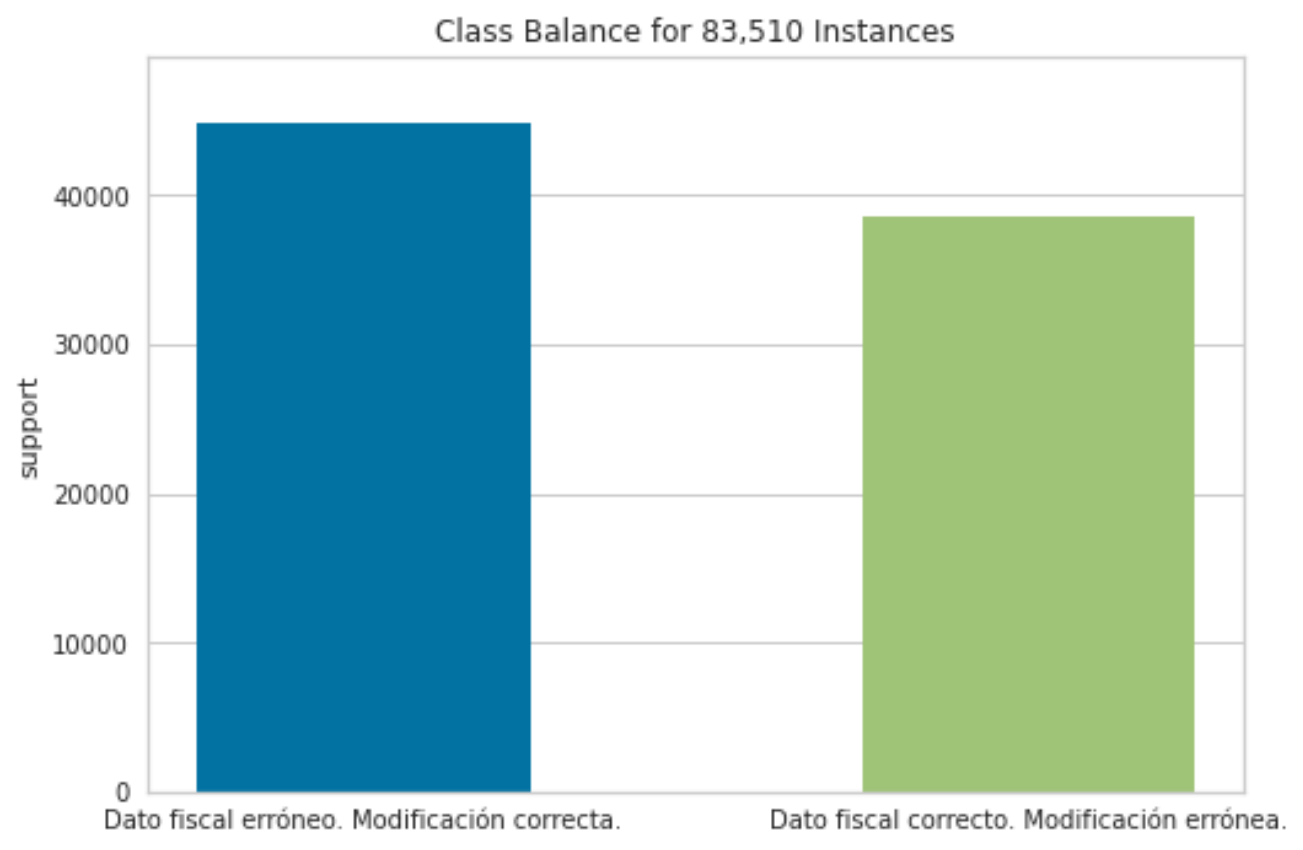
4. Resultados del entrenamiento. Resumen de calidad de los modelos.

- En rojo los casos en los que el modelo entrenado errores en cualquiera de las 6 casillas ha tenido mejor puntuación que el modelo específico. El área bajo la curva ROC es una métrica que se puede interpretar como una puntuación entre 0 y 1. La multiplicamos por 10 para facilitar la interpretación como una nota.

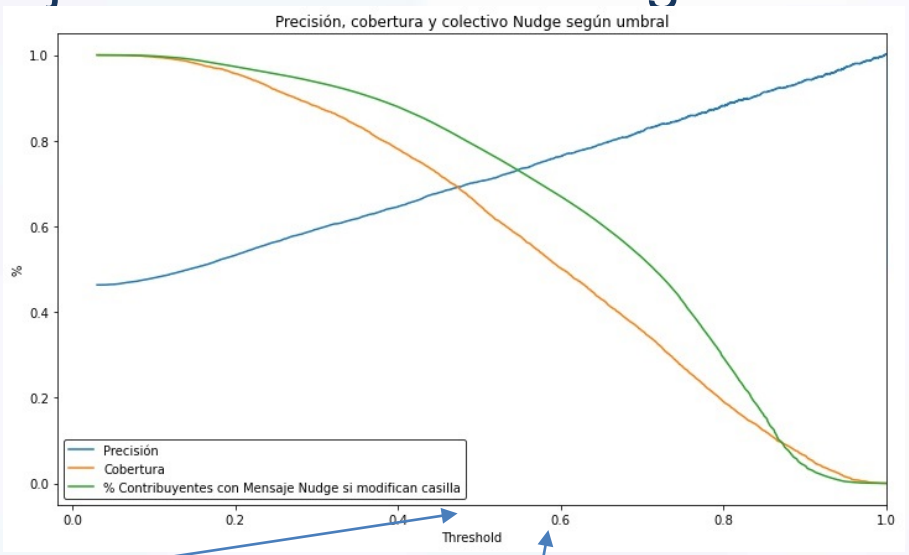
Casilla	Tamaño población entrenamiento	Porcentaje modificaciones erróneas del contribuyente	Área bajo curva ROC x 10
[003]	83.510	46,30	7,8
[004]	2.378	62,57	7,8 / 8,4
[005]	693	70,42	7,2 / 9,4
[006]	977	42,48	7,1 / 8,5
[0013]	33.017	27,26	7,6
[0596]	24.544	38,73	7,7

5. Algunos detalles del modelo 003.

- Distribución inicial de la población de entrenamiento.



5. Predicciones realizadas por el modelo. Zújar de asistencia Nudge.

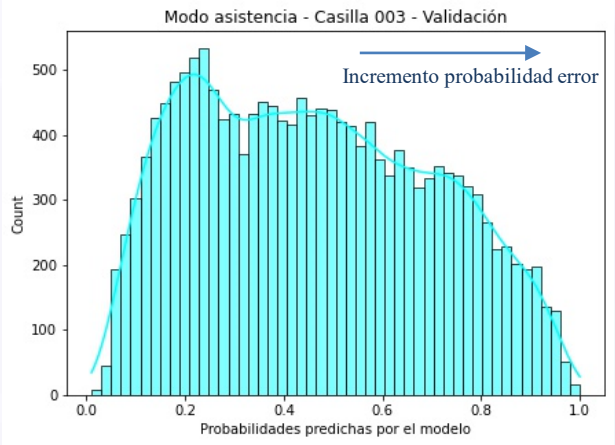
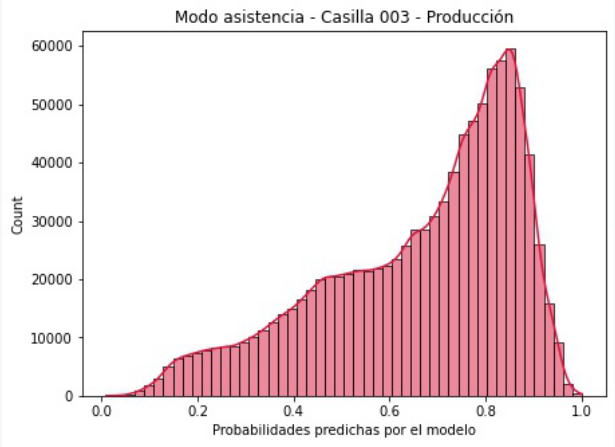


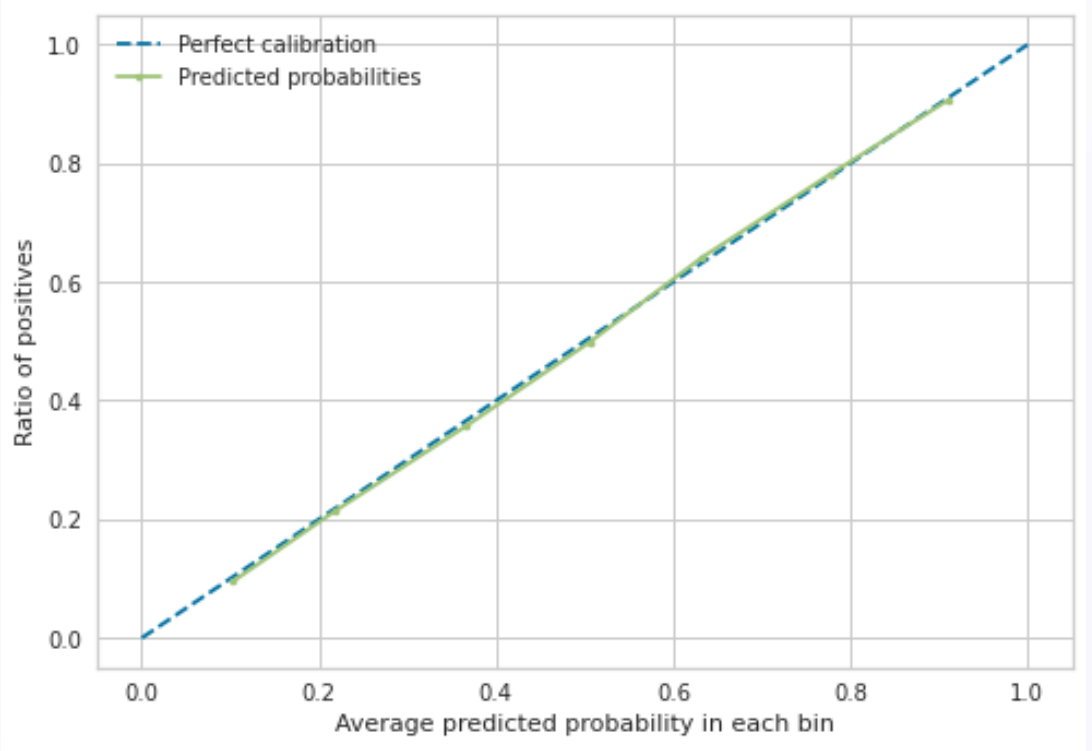
NIF identificado	Probabilidad casilla 003	Mensaje umbral 0.5 casilla 003	Precisión umbral 0.5 casilla 003	Cobertura umbral 0.5 casilla 003	Mensaje umbral 0.6 casilla 003
05118375R	0,5139	S	70,50	64,92	N
05118628R	0,8552	S	70,50	64,92	S
05118899L	0,6698	S	70,50	64,92	S

- Con este Zújar el Departamento de Gestión podrá definir los colectivos en función del número de contribuyentes a los que se quiera mostrar el mensaje y de la precisión deseada.

- Además, se generarán columnas con la clasificación del contribuyente a diferentes umbrales, para que el Departamento de Gestión pueda decidir en función de:
  - ¿A cuántos contribuyentes se les pondrá un mensaje?
  - ¿Qué probabilidad hay de acertar?

5. Predicciones realizadas por el modelo. Histograma de probabilidades.

	Asistencia
Conjunto validación	
Producción	



Gráfica de calibrado

5. Predicciones realizadas por el modelo. Interpretación histograma.

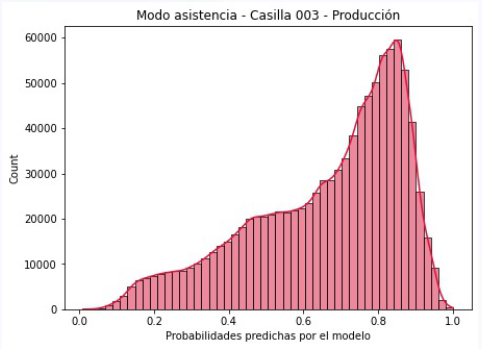
Conjunto Producción Asistencia: censo general de contribuyentes

99.2% contribuyentes que no modificarán la casilla 003

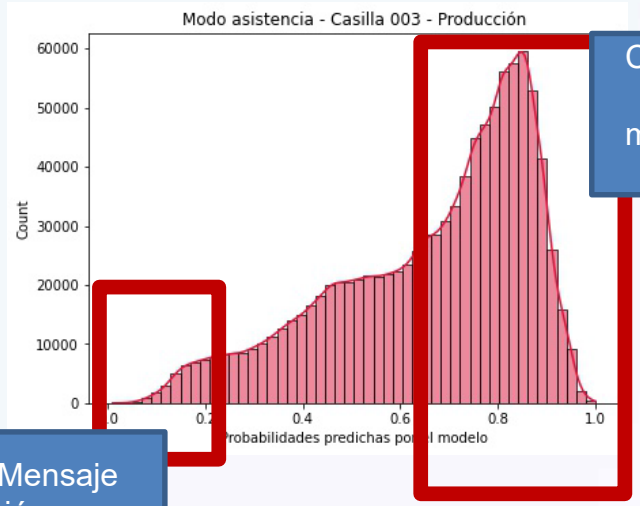
0,8% contribuyentes que modificarán la casilla 003

En este conjunto, es **muy probable** que el contribuyente se **equivoque**, como refleja la predicción del modelo.

Podemos asumir que la mayoría de contribuyentes de este conjunto no modifican la casilla porque no lo necesitan: el dato fiscal es correcto.



Mejorar la asistencia para reducir errores en las presentaciones



Campaña 2020: Mensaje Nudge en esta región para evitar modificaciones erróneas cuando el dato fiscal es correcto

¿Campaña 2021: Mensaje para sugerir revisión por posible dato fiscal erróneo?