

PRA2 - Tipología y ciclo de vida de los datos

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset utilizado tiene el título de “Hipotecas”, y contiene una serie de datos de clientes que tiene la sucursal de un banco. Los datos recogidos giran entorno a la toma de decisión del banco de si proporcionar un crédito hipotecario a los clientes o no, dependiendo de algunas variables.

Este dataset resulta interesante porque es un caso práctico que se da en la vida real y que tiene efectos reales sobre las personas. Los datos que se recogen en el dataset son:

- Ingresos, son los ingresos mensuales de la familia cliente
- Gastos comunes, que son los pagos mensuales de luz, agua, gas, etc.
- Pago coche, incluye las cuotas de posibles créditos de coches, y los pagos mensuales en combustible
- Gastos otros, incluye la compra de comida y los bienes necesarios para vivir
- Ahorros, suma de los ahorros dispuestos a utilizar para la compra de una vivienda
- Estado civil, marca el estado civil del cliente según los números:
 - o 0 – Soltero
 - o 1 – Casados
 - o 2 – Divorciados
- Hijos, cantidad de hijos menores y que no ingresan dinero
- Trabajo, marca el tipo de trabajo que ejerce el cliente según los números:
 - o 0 – Sin empleo
 - o 1 – Autónomo
 - o 2 – Empleado
 - o 3 – Empresario
 - o 4 – Pareja de autónomos
 - o 5 – Pareja de empleados
 - o 6 – Pareja autónomo y asalariado
 - o 7 – Pareja empresario y autónomo
 - o 8 – Empresarios los dos o empresario y empleado
- Hipoteca, si no ha sido concedida es 0, y si sí ha sido concedida es 1

Teniendo el clasificador “hipoteca” podemos entrenar un modelo de manera que podremos ayudar a la sucursal del banco prever si un nuevo cliente podrá obtener una hipoteca o no, sin tener que hacer un análisis en profundidad manual como lo podrían estar haciendo hasta ahora.

2. Integración y selección de los datos de interés a analizar

Realizamos un análisis descriptivo del dataset. Podemos comprobar que inicialmente hay 135 casos de petición de hipotecas en las que NO se han concedido, y 67 casos en los que sí se ha concedido.

Visualizamos además los historiogramas de las variables para entender la varianza y distribución de su contenido.

3. Limpieza de los datos

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Hacemos las pruebas para comprobar si hay ceros o elementos vacíos y vemos que no los hay.

En el caso de que los hubieran, eliminaríamos esos casos para que no afecten negativamente en el modelo predictivo.

3.2. Identificación y tratamiento de valores extremos.

Hacemos las pruebas para comprobar que si hay valores extremos y vemos que no los hay.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Consideramos comprar variables numéricas para conocer sus correlaciones con la variable objetivo “hipoteca”.

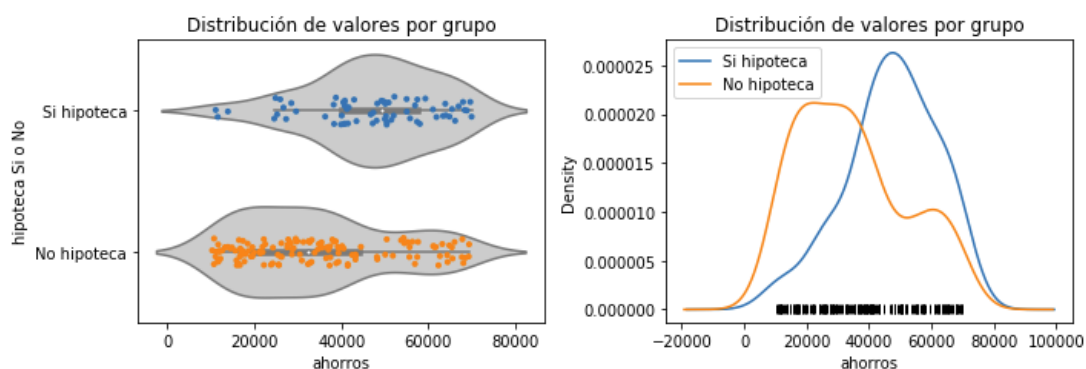
4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Hacemos las comprobaciones de normalidad y homogeneidad, y vemos que no se cumple la homogeneidad. La variable ingresos tiene el resultado del pvalue bajo.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Aplicamos en el código 3 tests.

5. Representación de los resultados a partir de tablas y gráficas.



6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Sí, podemos responder al problema, en la medida de que hemos podido elaborar un modelo predictivo que puede indicar a priori si un nuevo cliente recibirá la concesión de hipoteca o no.

- 7. Código:** Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Adjuntado en el enlace GitHub

Contribuciones	Firma
Alberto Cuevas, Carlos Salas	acuevasgonz, csalas0
Alberto Cuevas, Carlos Salas	acuevasgonz, csalas0
Alberto Cuevas, Carlos Salas	acuevasgonz, csalas0