

Evaluación Crediticia: Datos y Decisiones

Fernando Martínez Chehda

Temario

- Contexto y audiencia
- Hipótesis
- Análisis exploratorio de datos
- Insights y recomendaciones

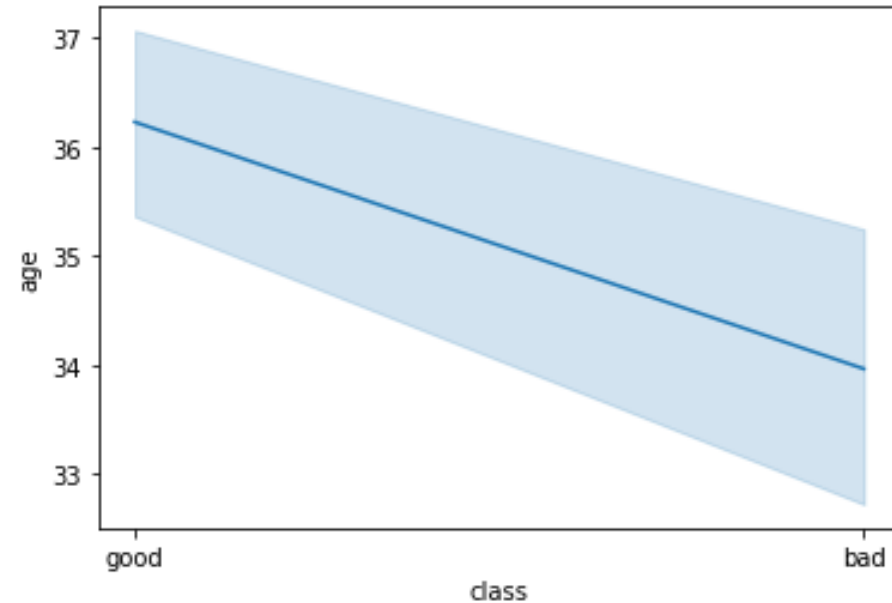
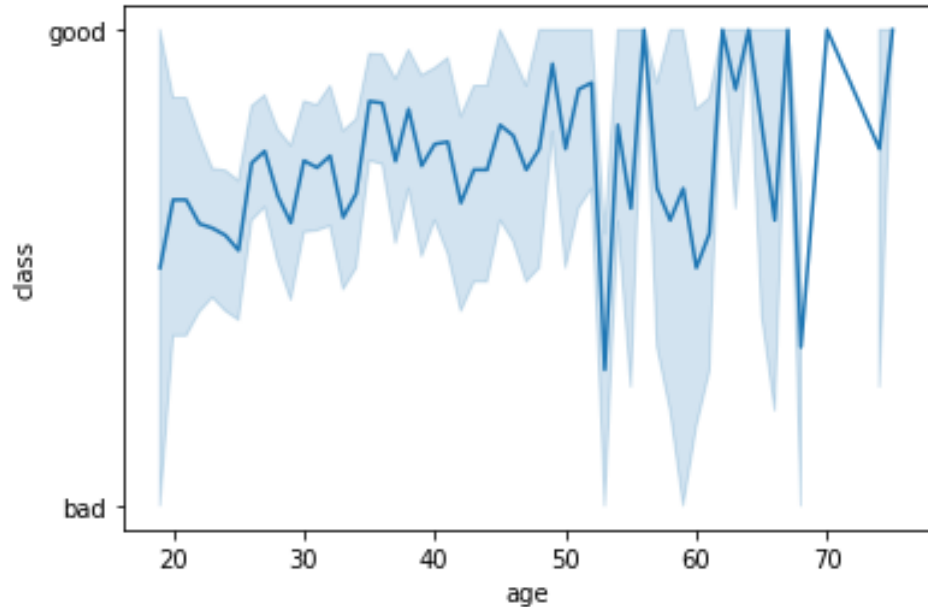
Contexto y audiencia

- Este proyecto tiene como objetivo determinar el riesgo asociado a la concesión de créditos a clientes
- Se utiliza un dataset que contiene información relevante sobre los clientes, como duración del crédito solicitado, ahorros aproximados al momento de la solicitud, propósito del crédito y edad, entre otros.
- El Scoring Crediticio es un modelo matemático que emplea variables financieras y características del cliente para asignar una puntuación que refleja la probabilidad de que dicho cliente cumpla con sus obligaciones crediticias. En base a esta información pública, se intentó tomar condiciones similares para que el proyecto cuente con una dosis de realidad.
- Público objetivo: empresas financieras, ONG, proveedores y destinatarios de crédito, personas interesadas en la ciencia de datos

Hipótesis

<p>A medida que la edad de los solicitantes de crédito disminuye, el riesgo asociado a la entrega del crédito tiende a aumentar.</p>	<p>La estabilidad laboral de una persona influye significativamente en el nivel de riesgo crediticio. Se espera que aquellos con un empleo estable tengan un menor riesgo al ser evaluados para la concesión de un crédito</p>	<p>La cantidad de ahorros de una persona se relaciona directamente con el riesgo de no devolver un crédito. Se anticipa que aquellos individuos con un mayor nivel de ahorros presenten un menor riesgo crediticio.</p>
--	--	---

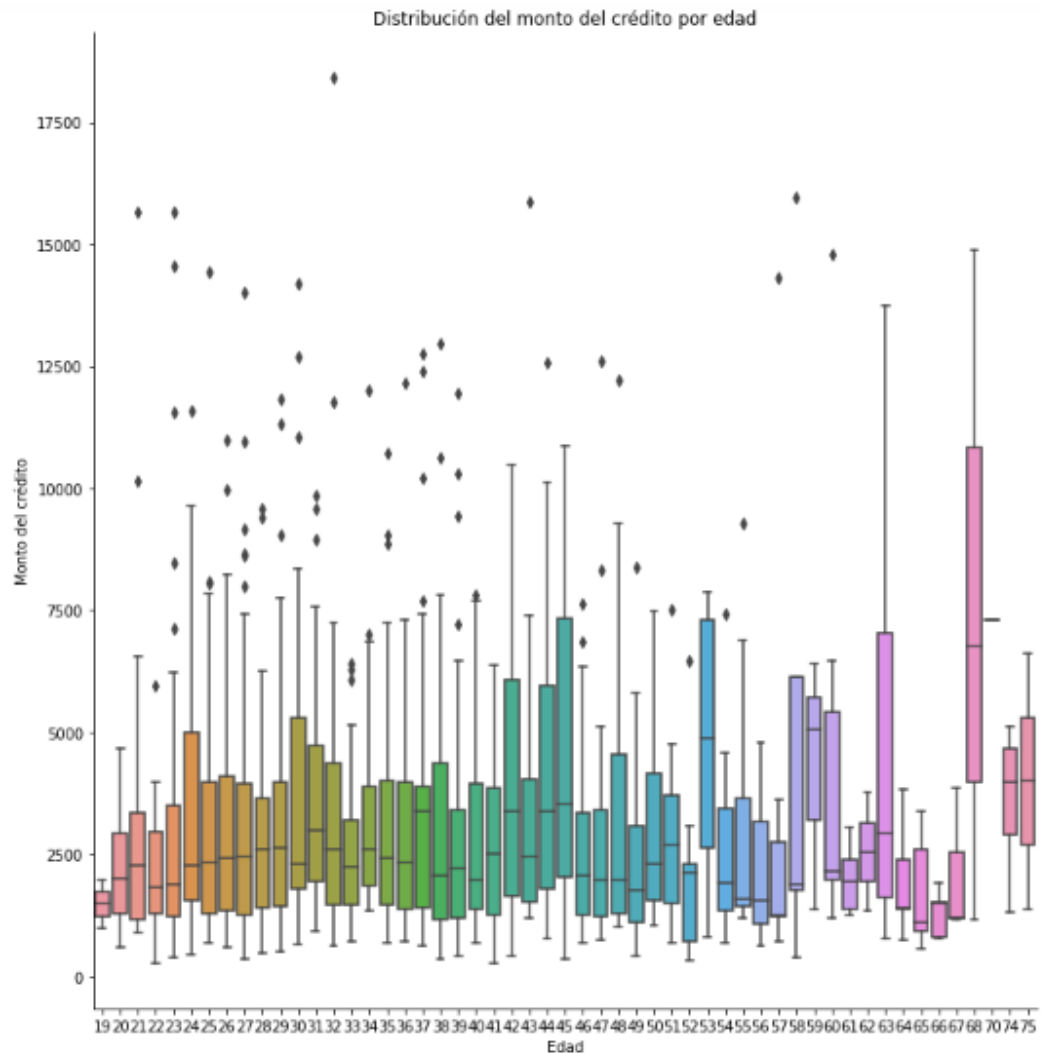
Análisis exploratorio



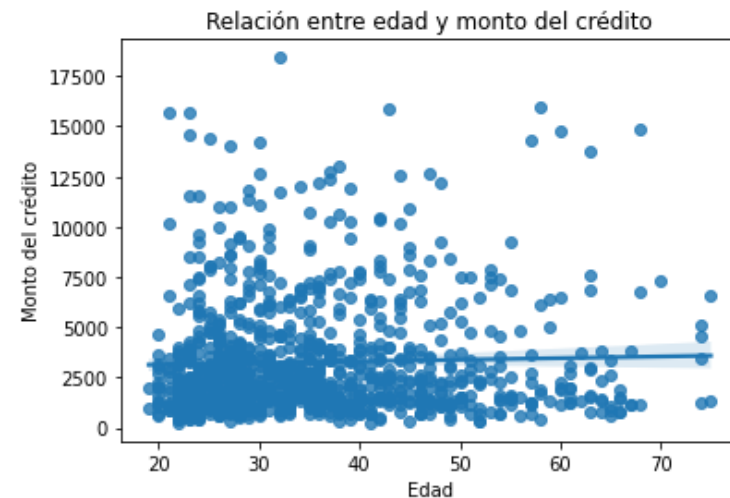
Edad con respecto a clase: Good = riesgo bajo, Bad = riesgo alto

Para poder analizar nuestra primer hipótesis se realizaron estos gráficos, se transformaron los datos y se llegó a las imágenes que se muestran

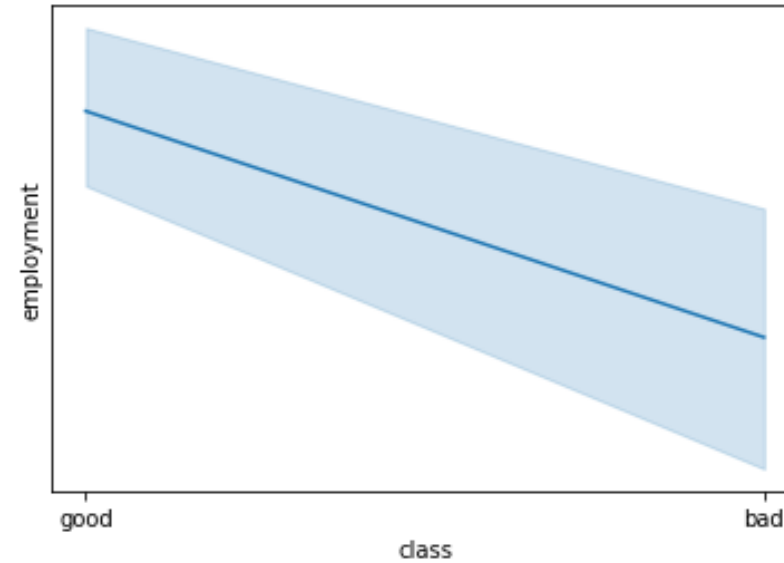
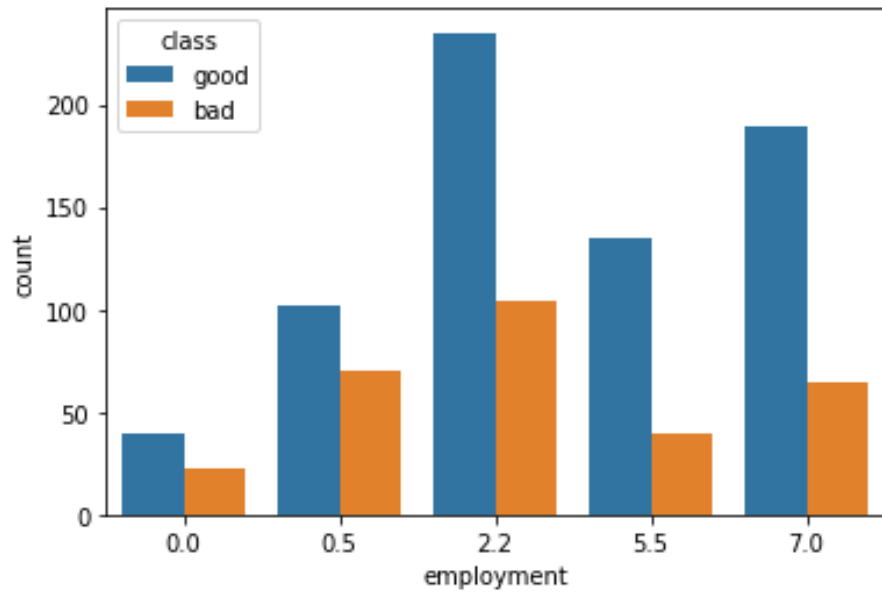
Análisis exploratorio



Aquí quisimos ver la relación entre la edad y el monto del crédito solicitado

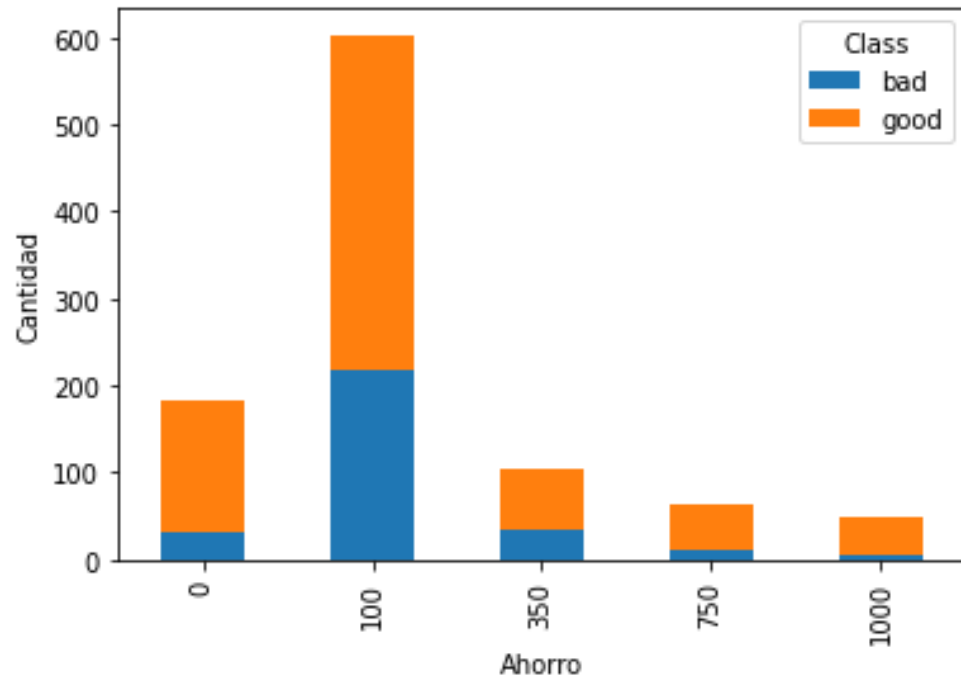


Análisis exploratorio



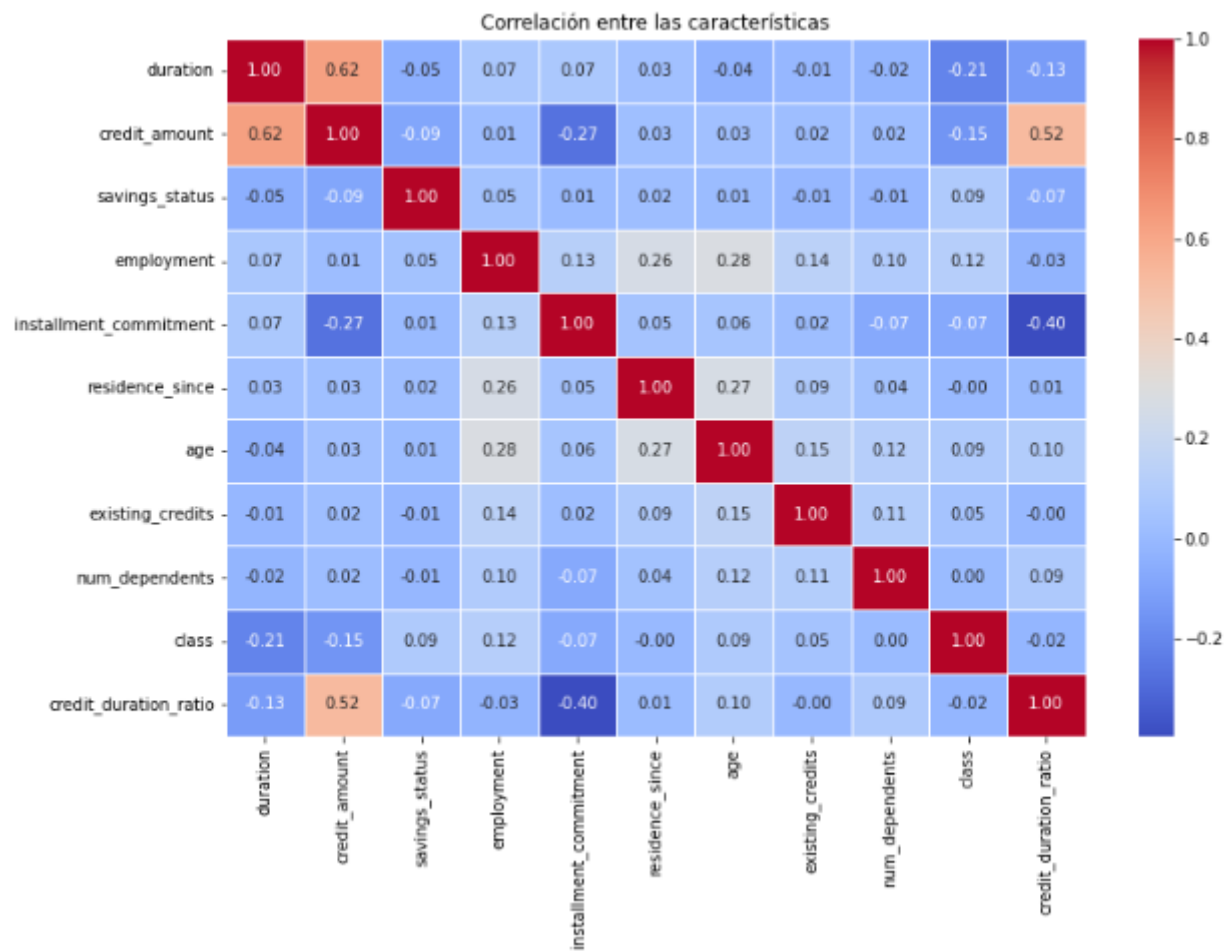
Antigüedad trabajando con respecto al riesgo de entregar crédito

Análisis exploratorio



Cantidad de ahorro con respecto al
riesgo crediticio

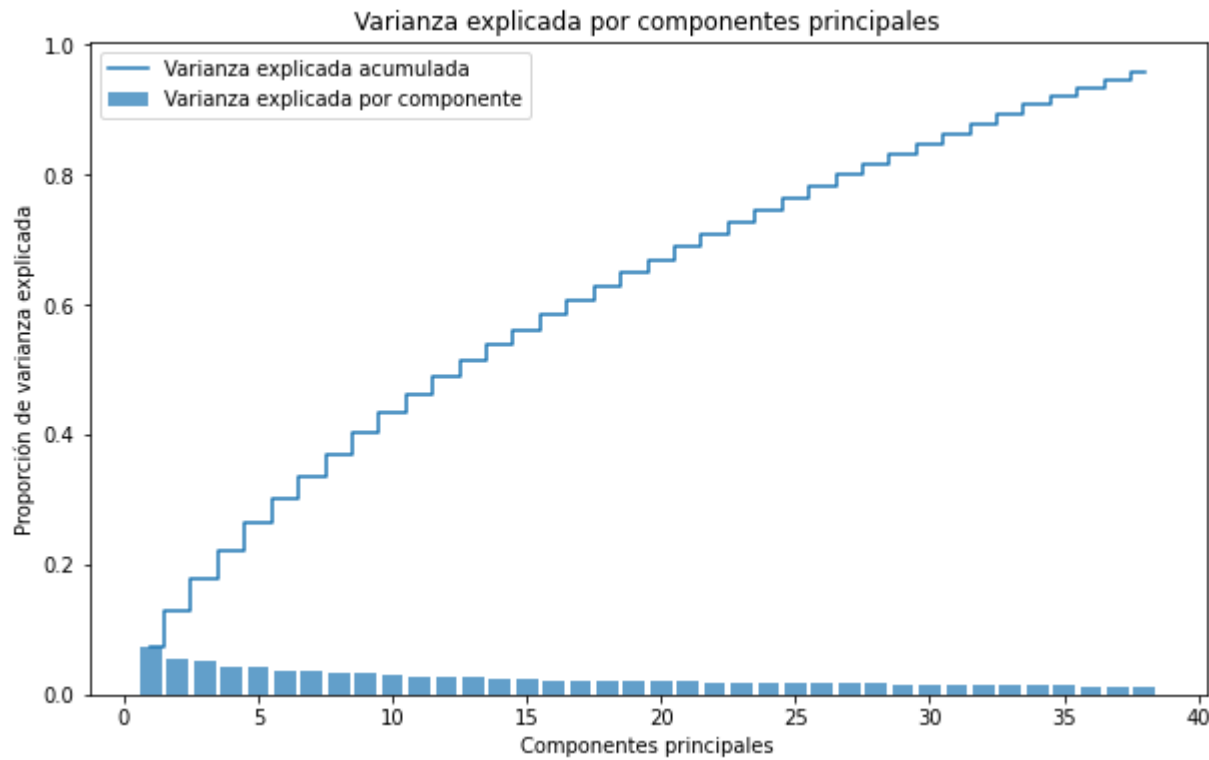
Análisis bivariado



Se observa una relación entre el monto del crédito y la duración, lo cual indicaría que a mayor cantidad de tiempo de vencimiento, mayor el monto del crédito que se solicita.

Modelos de Machine Learning

- Se realizó el análisis de componentes principales mediante PCA.



El gráfico nos resulta muy útil para entender que cerca del componente 38 se alcanza el 95% (o muy cerca de él), lo cual sugiere que con 38 componentes principales se estarían conservando la totalidad de los datos originales.

Modelos de Machine Learning

- Regresión logística

Mejores parámetros: {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}

Precisión del modelo en datos de prueba: 0.995

- Random forest

Mejores parámetros: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}

Precisión del modelo en datos de prueba: 0.9

- Naive Bayes

Precisión del modelo en datos de prueba: 0.895

- Redes neuronales

Mejores parámetros: {'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'learning_rate': 'adaptive', 'solver': 'adam'}

Precisión del modelo en datos de prueba: 0.99375

Cross Validation de modelos

Regresión Logística: Promedio de precisión - 0.9910, Desviación estándar - 0.0037

Random Forest: Promedio de precisión - 0.9040, Desviación estándar - 0.0132

Naïve Bayes: Promedio de precisión - 0.8850, Desviación estándar - 0.0187

Redes Neuronales: Promedio de precisión - 0.9890, Desviación estándar - 0.0037

Modelo seleccionado: Regresión logística:

Motivos: - Precisión

- Se comporta de manera eficaz para variables objetivos de 2 posibilidades
- Computacionalmente no ocupa demasiados recursos.
- Sencilla de interpretar

Modelo final

Precisión del modelo en datos de prueba: 0.995

	precision	recall	f1-score	support
0	0.98	1.00	0.99	59
1	1.00	0.99	1.00	141
accuracy			0.99	200
macro avg	0.99	1.00	0.99	200
weighted avg	1.00	0.99	1.00	200

Matriz de confusión:

```
[[ 59   0]
 [  1 140]]
```

- La precisión general del modelo es del 99.5%
- Presición 0 (clase bad) = 98%
- Presición 1 (clase good) = 100%
- El recall muestra buenos resultados para ambas clases, 98% y 100% para 0 y 1 respectivamente.
- La matriz de confusión muestra que el modelo cometió un solo error al clasificar la clase 0 y ninguno al clasificar la clase 1. Esto indica una clasificación precisa y mínimos falsos positivos/negativos.

Insights y recomendaciones

- Se puede apreciar una clara tendencia que indica que a medida que la edad aumenta, el riesgo asociado a la entrega de un crédito disminuye.
- La edad y el monto del crédito solicitado no tienen una relación muy clara.
- A mayor tiempo trabajado, menor riesgo de entregar un crédito.
- No se observa una notoria diferencia entre las cantidades de ahorro y en la clase en la que divide este dataset entre buen riesgo y mal riesgo.
- Recomendaciones: utilizar un dataset más amplio con mayor cantidad de casos y características
- Para casos de machine learning se recomienda utilizar regresión logística por el tipo de datos que tenemos en el dataset, ya que queremos relacionar muchas características y contrastarlas con la clave “class”.