Predicción de Ingresos con Machine Learning

Asier Rodríguez | Bootcamp de Data Science

PROYECTO INDIVIDUAL

Problema de Negocio

El Desafío

Una cadena de librerías con múltiples tiendas y canales de venta necesita predecir sus ingresos mensuales para optimizar la planificación financiera y la toma de decisiones operativas.

Objetivo del Proyecto

Desarrollar un modelo de machine learning capaz de predecir con precisión los ingresos netos mensuales (net_revenue), capturando tendencias, estacionalidad y diferencias entre tiendas y canales.

Dataset y Variables

Fuente de Datos

Dataset de ventas y payroll de una cadena de librerías

- Periodo: 2019–2024
- Entrenamiento principal: desde 2020
- Más de 1.000 observaciones mensuales

Variable Objetivo

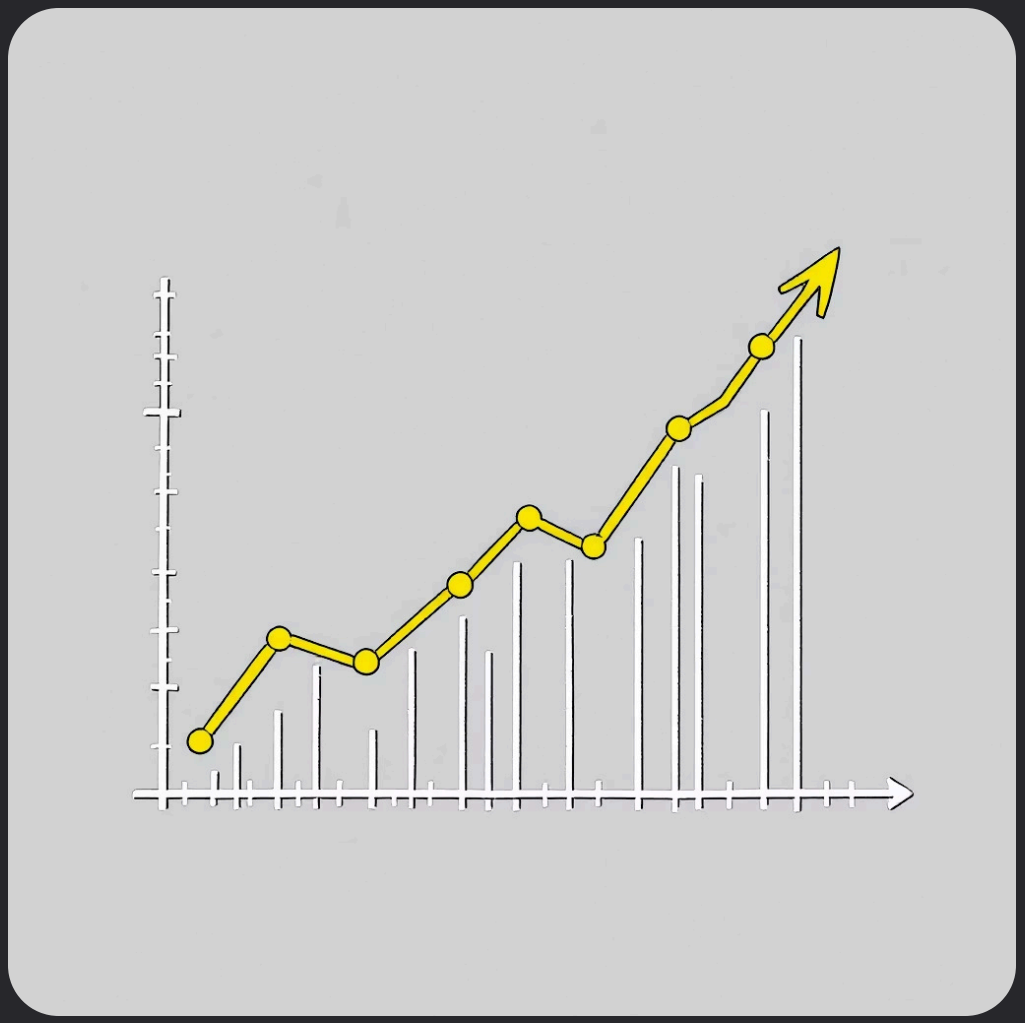
net_revenue: ingresos netos mensuales

Variables Principales

- date
- store_id
- channel (IN_STORE / ONLINE)
- transactions
- payroll (análisis complementario)

Exploratory Data Analysis

HALLAZGOS CLAVE



Tendencia Creciente

Los ingresos muestran una tendencia ascendente desde 2020, con recuperación post-COVID

Estacionalidad Clara

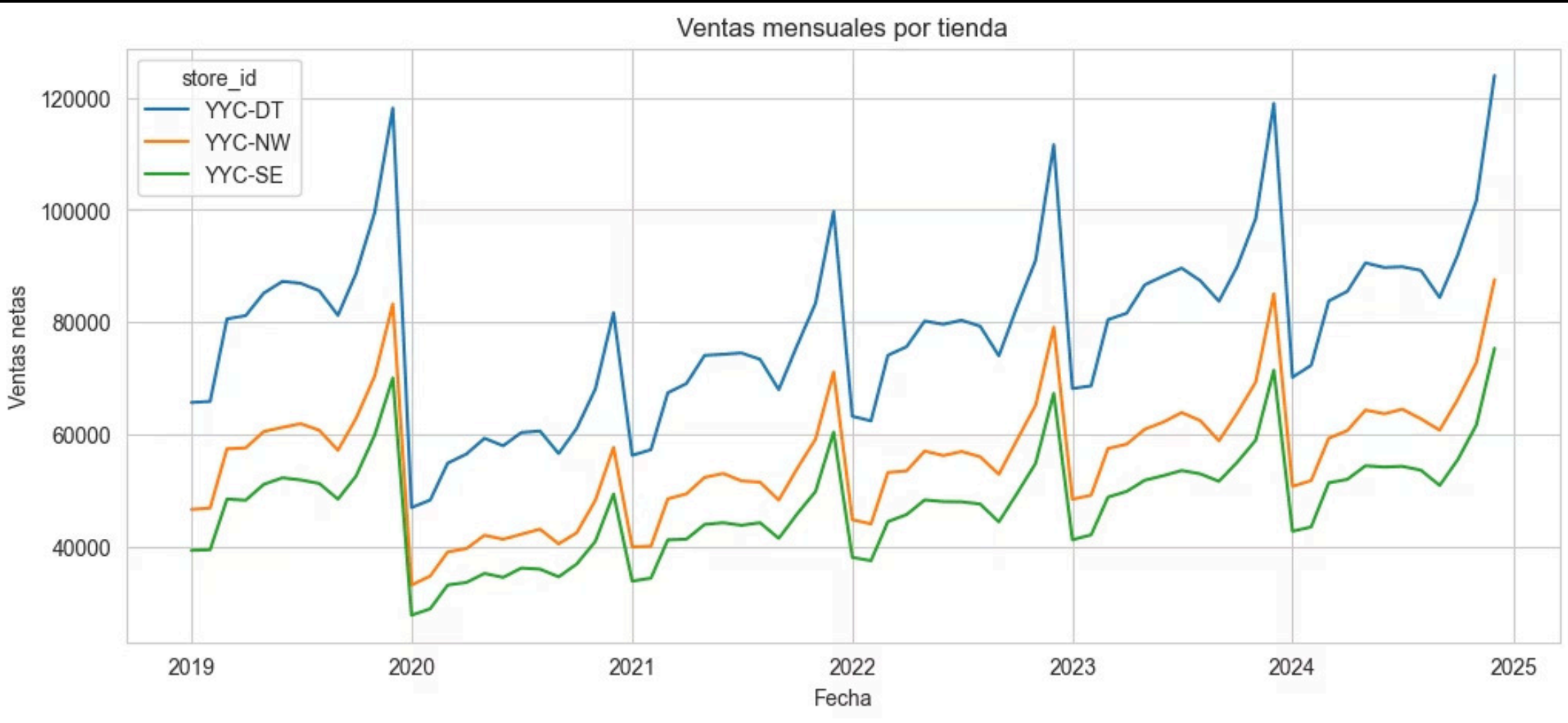
Picos consistentes en septiembre (vuelta al cole) y diciembre (Navidad)

Impacto del COVID-19

Efecto evidente en 2020 con posterior recuperación

Diferencias Entre Canales

Comportamiento diferenciado entre tiendas físicas y canal online



Feature Engineering

Features Temporales

Lags de ingresos:

- 1, 3, 6 y 12 meses

Rolling means:

- 3, 6 y 12 meses

Features Calculadas

Avg_ticket:

net_revenue / transactions

Indica el valor medio por transacción

Encoding Categórico

One-hot encoding de:

- store_id
- channel
- month

Modelos Evaluados

SARIMA

Modelo benchmark clásico de series temporales para capturar estacionalidad y tendencia

Regresión Lineal

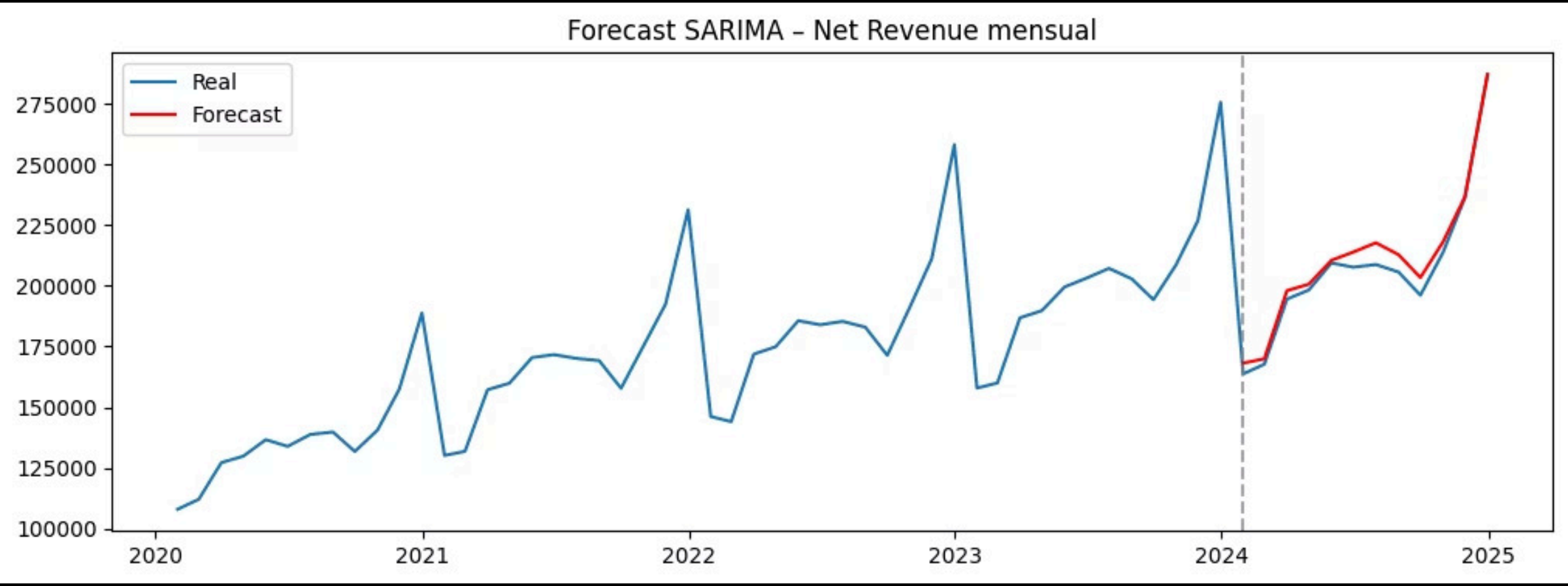
Modelo lineal con features temporales, interpretable y estable

Random Forest

Ensemble de árboles de decisión para capturar relaciones no lineales

XGBoost

Gradient boosting avanzado, potente para predicciones complejas



Estrategia de Validación

01

Validación Temporal Estricta

Se respeta el orden cronológico de los datos para evitar data leakage y garantizar resultados realistas

02

Conjunto de Test

Los últimos 12 meses se reservan como conjunto de test para evaluar el rendimiento del modelo

03

Métricas de Evaluación

MAE (Mean Absolute Error) y **RMSE** (Root Mean Squared Error) para medir precisión

Resultados del Modelo Final

Modelo Seleccionado

Regresión Lineal con features temporales

Mejor Error en Test

Logra el mejor equilibrio entre precisión y estabilidad en el conjunto de validación

Estabilidad

Predicciones consistentes sin sobreajuste, rendimiento robusto en datos nuevos

Interpretabilidad

Coeficientes claros que permiten entender qué variables impactan más en los ingresos

El modelo captura efectivamente la tendencia y estacionalidad de los ingresos, siendo apto para producción y planificación financiera.

Puntuaciones de los modelos

Sarima

- RMSE → 4879
- MAE → 4002

RandomForest

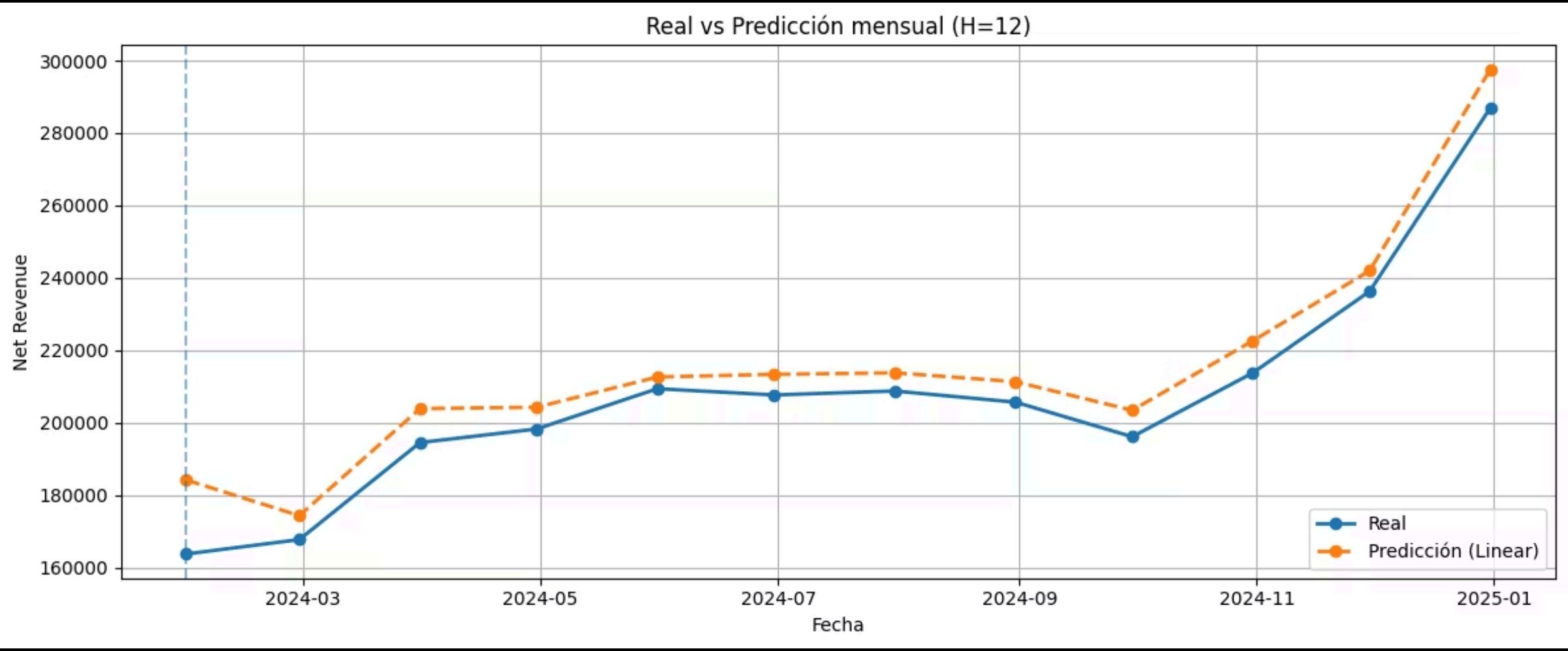
- RMSE → 5559
- MAE → 3870

LinearRegression

- RMSE → 2663
- MAE → 1803

XGBoost

- RMSE → 4048
- MAE → 2742



Análisis de Payroll

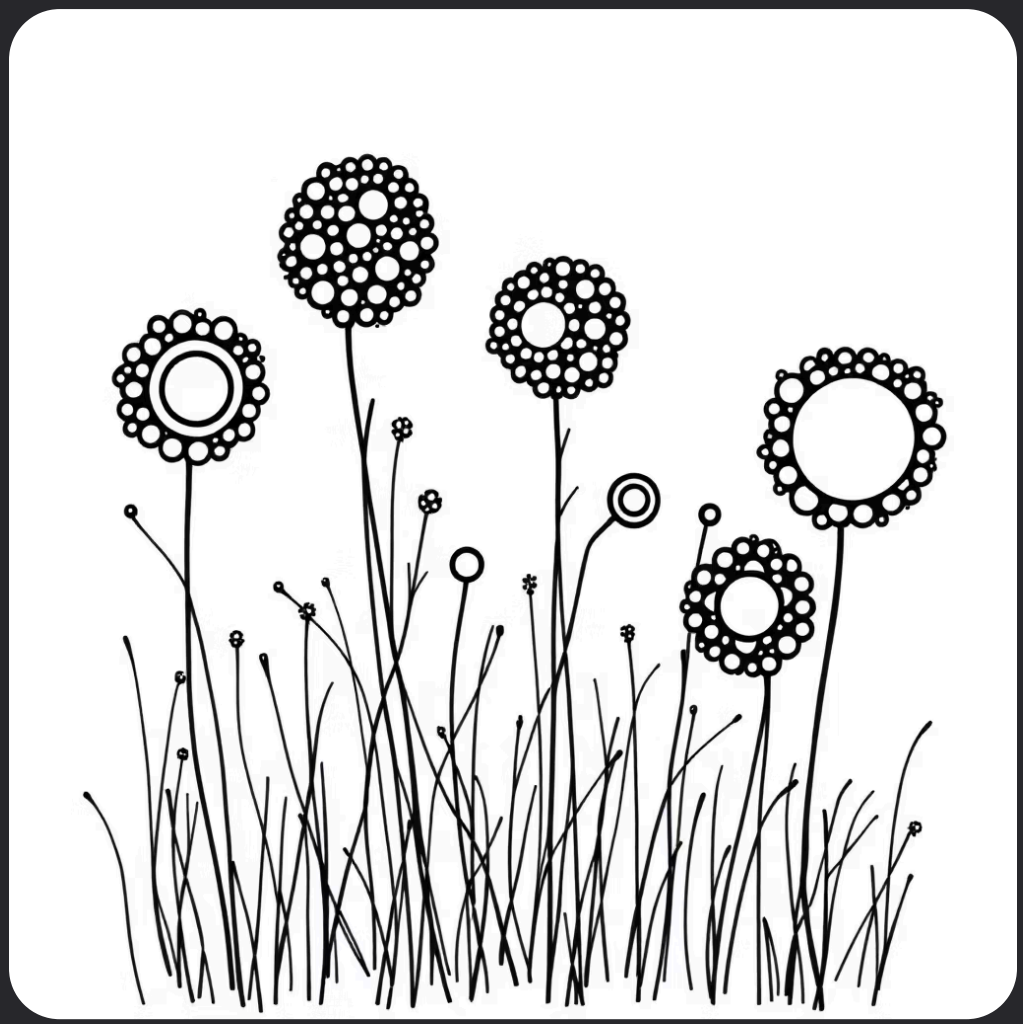
ANÁLISIS NO SUPERVISADO

Isolation Forest

Detección de anomalías en payroll

Identifica meses con gastos salariales inusuales que requieren revisión o investigación adicional

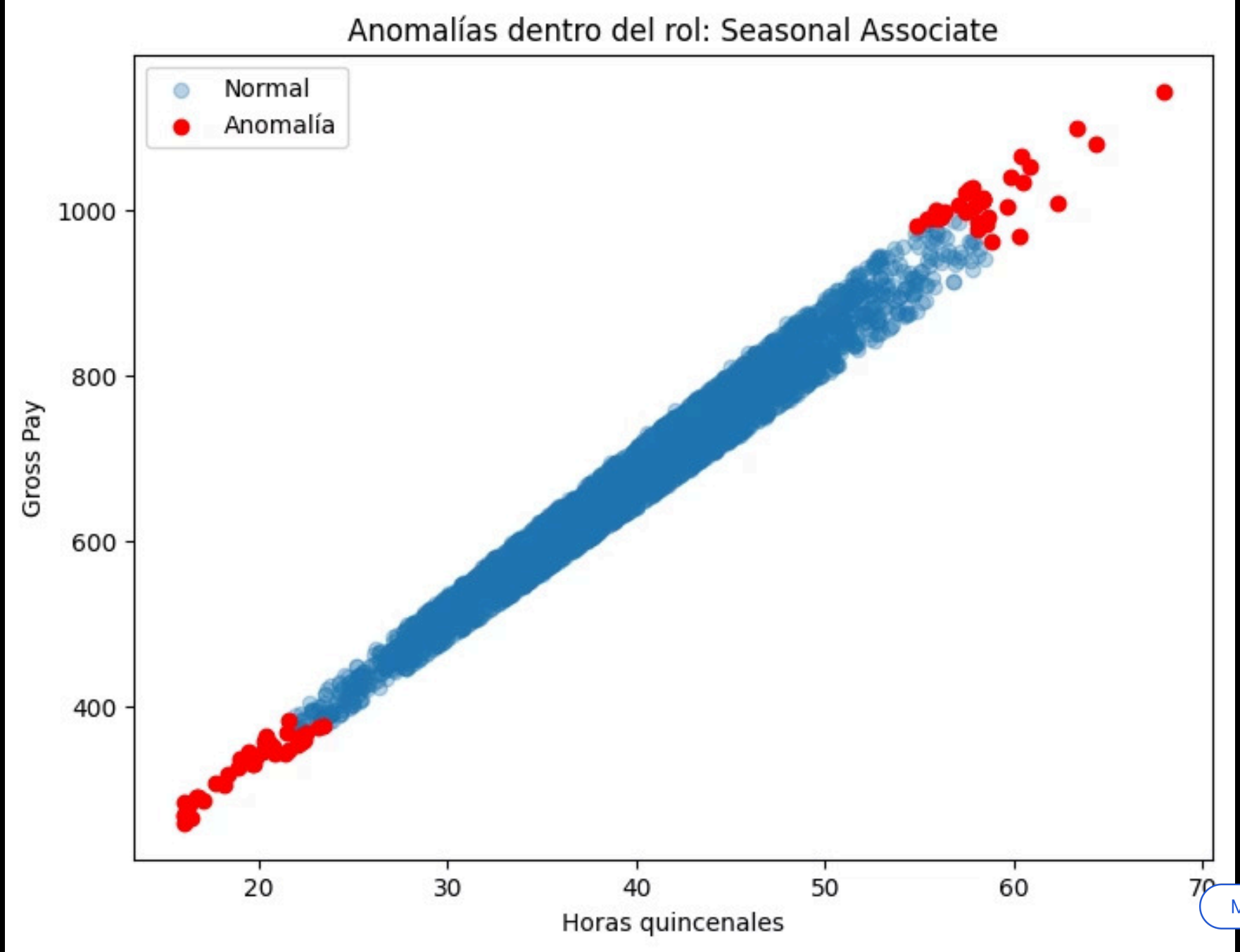
KMeans Clustering

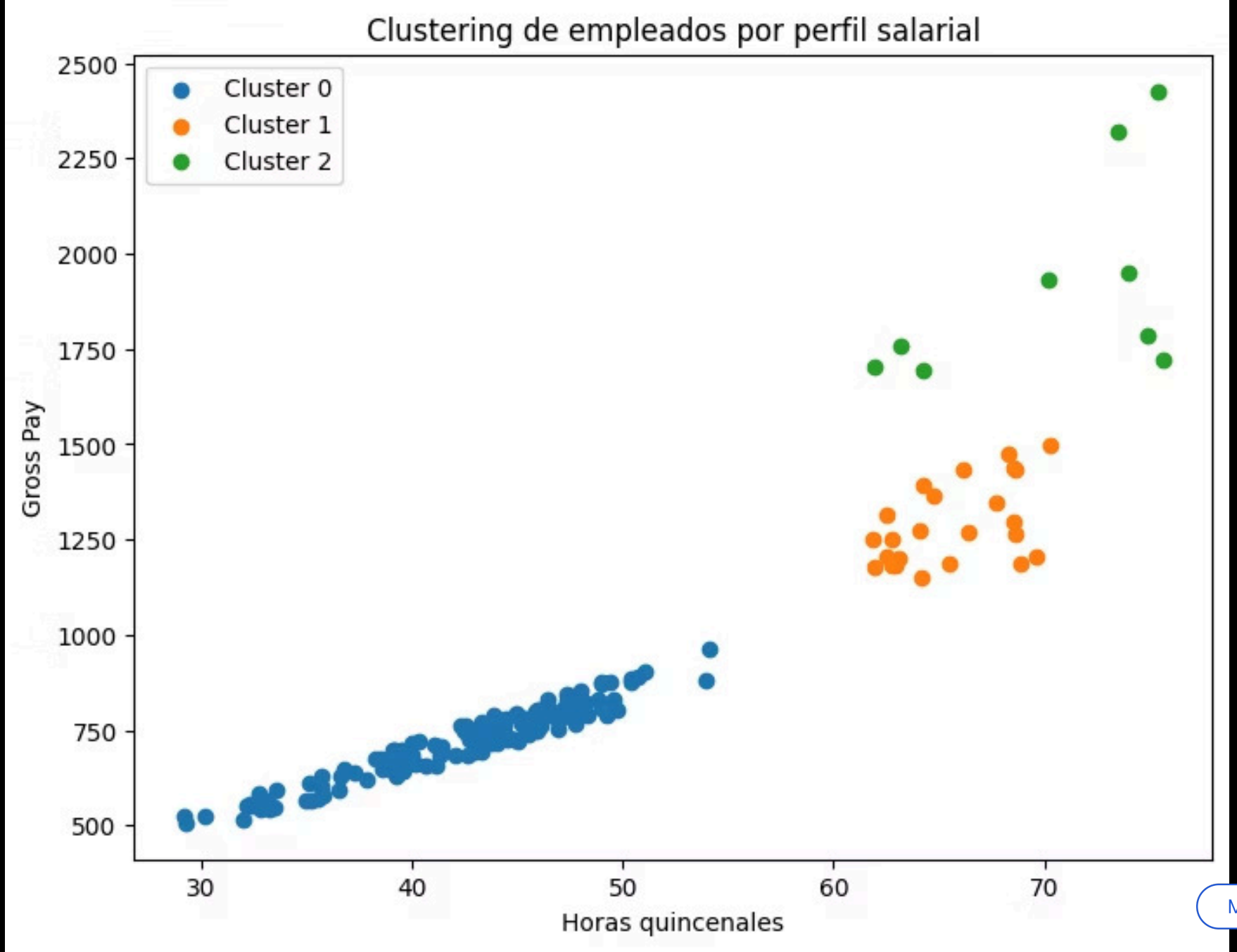


Segmentación salarial en 3 clusters

Agrupa meses según patrones de payroll para apoyar decisiones de gestión de personal

Este análisis complementario no se usa para predicción directa, sino para identificar patrones y apoyar decisiones operativas





Conclusiones y Próximos Pasos

Logros del Proyecto

- Modelo captura tendencia y estacionalidad con precisión
- Alta interpretabilidad para stakeholders
- Preparado para apoyar planificación financiera
- Útil para decisiones operativas informadas

Limitaciones

- No incluye variables externas (economía, marketing, competencia)
- Scope limitado a datos históricos internos

Mejoras Futuras

- Incorporar variables externas (indicadores económicos, campañas)
- Explorar modelos más complejos (LSTM para patrones profundos)
- Automatización del pipeline y despliegue en producción