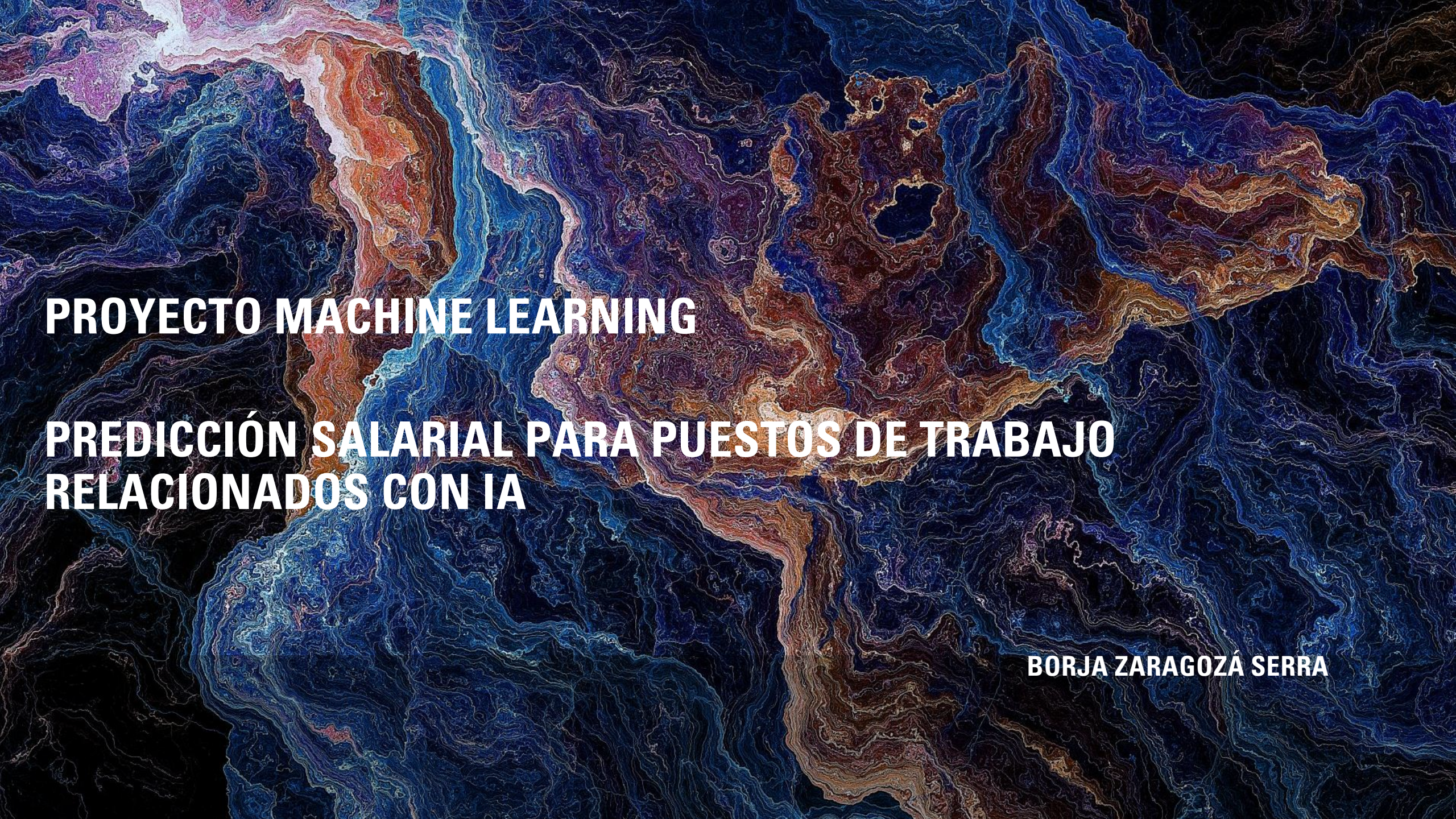

PROYECTO MACHINE LEARNING

PREDICCIÓN SALARIAL PARA PUESTOS DE TRABAJO RELACIONADOS CON IA

BORJA ZARAGOZÁ SERRA

- Presentación del problema de negocio
- Objetivo técnico y enfoque
- Dataset y análisis exploratorio
- Modelado y arquitectura final
- Resultados y conclusiones

¿POR QUÉ PRECEDIR EL SALARIO EN PUESTOS IA?

- El sector de la Inteligencia Artificial está en constante crecimiento, pero la información salarial es poco transparente, especialmente en nuevas tecnologías y perfiles híbridos.
- Esto genera incertidumbre y desigualdad salarial, especialmente para personas que están buscando empleo o cambiando de sector.
- Predecir el salario de una oferta de trabajo en IA a partir de sus características (experiencia requerida, localización, modalidad, etc.) ayuda a los candidatos a tomar decisiones informadas y negociar mejor sus condiciones.
- También permite a reclutadores y plataformas de empleo detectar desajustes, ajustar presupuestos y mejorar la conversión de ofertas públicas.

¿QUÉ TAREA RESOLVEMOS Y CÓMO LA EVALUAMOS?

- El problema se aborda como una tarea de **regresión supervisada**, cuyo objetivo es predecir el salario asociado a una oferta laboral del sector IA
- La métrica utilizada es el **Error Absoluto Medio** (MAE) porque:
  - Penaliza los errores de forma lineal, lo cual es más interpretable y equilibrado en términos prácticos
  - Es menos sensible a *outliers* extremos que el RMSE, lo que resulta útil en salarios con alta variabilidad
  - Refleja directamente el desfase medio en dólares, lo que es más fácil de entender y comunicar
- Se descarta un problema de clasificación, por ejemplo (en rangos salariales) porque:
  - Limita la resolución del problema a categorías predefinidas, perdiendo granularidad
  - No responde bien a la necesidad de ofrecer una estimación concreta del salario
  - Una mala clasificación puede llevar a errores significativos si luego se encadena con una regresión

ANÁLISIS EXPLORATORIO Y CARACTERÍSTICAS DEL DATASET

Origen del dataset

- Kaggle: 'Global AI Job Market and Salary Trends 2025'
- Muestra: 15.000 ofertas de empleo
- Formato: CSV

Características del dataset

- 19 *features* o columnas
- Variables mayoritariamente categóricas de baja cardinalidad
- El target es una variable numérica continua que muestra alta dispersión y asimetría positiva

Distribución del target

- Rango: 32.519 USD hasta 399.095 USD
- Media: 115.349 USD, con fuerte sesgo hacia salarios bajos (concentración entre 50K y 150K)
- Presencia de *outliers*, justificación adicional para usar **MAE** en lugar de RMSE

Ingeniería de variables

- Codificación de variables categóricas con baja cardinalidad (3 – 4 valores únicos)
- One-Hot Encoding individual por país y agrupación por región
- Agregados estadísticos (medias y medianas) por grupo relevante
- Nuevas variables: plazo de la oferta, binarias por tipo de contrato

MINI EDA Y SELECCIÓN DE FEATURES

Primera selección de features

- Análisis visual bivariante
- Variables numéricas
  Correlación de Pearson (0,3)
- Variables Categóricas
  Evaluados con ANOVA (p-valor < 0,05)

Selección de features con otros modelos

- Selección por KBest
- Selección mediante Eliminación Recursiva
- Selección Secuencial de Features
- Todas las features

MODELO Y OPTIMIZACIÓN

Validación cruzada y selección de modelo

+ Se evalúan las 5 listas con 5 modelos diferentes: Decision Tree, Random Forest, XGBoost, LightGBM y CatBoost

+ Modelo baseline: Decisión Tree al ser un modelo interpretable y rápido, ideal como referencia inicial

Mejor combinación encontrada

LightGBM con la lista seleccionada mediante Selección Secuencial de Features (SFS)

→ Mejor rendimiento según validación cruzada (MAE más baja)

Optimización de hiperparámetros

· *GridSearch*: búsqueda exhaustiva

· *Optuna*: búsqueda bayesiana y eficiente. Mejores resultados para *Optuna*, aunque con diferencia pequeña

Entrenamiento y evaluación

- Escalado logarítmico del target durante el entrenamiento y predicciones devueltas a escala original mediante transformación exponencial
- Evaluación final del modelo frente al conjunto de test con un MAE de 14.497,66 USD

CONCLUSIONES

➔ Se ha conseguido un **MAE** final de aproximadamente 14,5K USD, valor sólido teniendo en cuenta la gran dispersión y heterogeneidad del salario en el sector IA (salarios entre 32K y 400K)

➔ El modelo elegido, **LightGBM**, ha demostrado ser el más eficaz gracias a su capacidad para manejar datos categóricos y gran cantidad de features con buen rendimiento.

➔ **La transformación logarítmica** del target ha permitido reducir el impacto de valores extremos y mejorar la estabilidad del modelo

➔ El uso de ***Optuna*** para la optimización ha sido clave para reducir la métrica final en test

➔ Las features con mayor relevancia han sido:

- La experiencia en años
- El tamaño de la compañía