

# **Predictión de Precios de Casas**

## **Modelos de Deep Learning y XGBoost**

Proyecto Final – CC3092 Deep Learning y Sistemas Inteligentes

Competencia: House Prices – Advanced Regression Techniques (Kaggle)

Integrantes: Andy Fernando Fuentes Velásquez y equipo.

Semestre II – 2025

# Contexto del Problema: El Desafío Inmobiliario

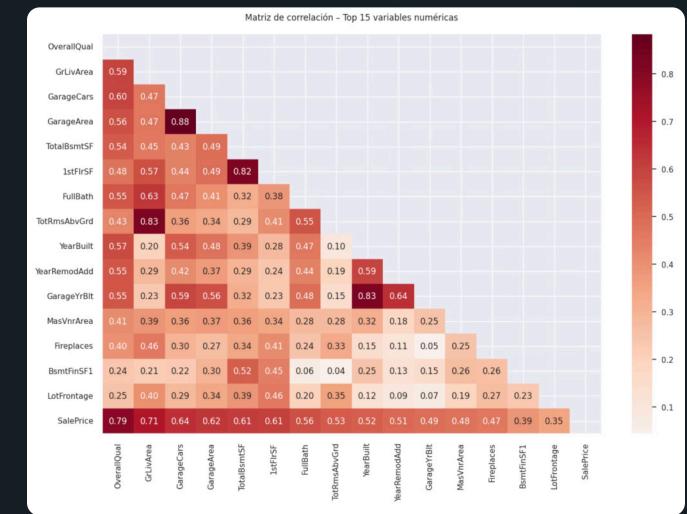
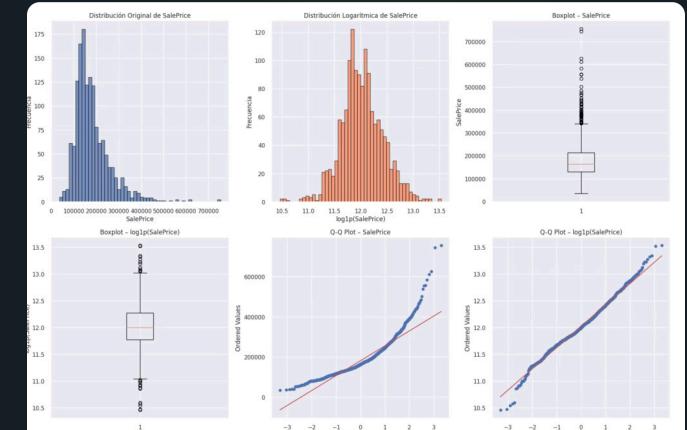
El mercado inmobiliario es un ecosistema complejo donde la estimación precisa del precio de venta de una propiedad es crucial para compradores y vendedores por igual. Antes de listar una casa, es fundamental tener una valoración objetiva.

	count	mean	std	min	25%	50%	75%	max
Id	1460.0	730.500000	421.610009	1.0	365.75	730.5	1095.25	1460.0
MSSubClass	1460.0	56.8977460	42.300571	20.0	20.0	50.0	70.0	190.0
LotFrontage	1201.0	70.049958	24.284752	21.0	59.00	69.0	80.00	313.0
LotArea	1460.0	10516.828082	9981.264932	1300.0	7553.50	9478.5	11601.50	212545.0
OverallQual	1460.0	6.099315	1.382997	1.0	5.00	6.0	7.00	10.0
OverallCond	1460.0	5.575342	1.112798	1.0	5.00	5.0	6.00	9.0
YearBuilt	1460.0	1971.267808	30.202904	1872.0	1954.00	1973.0	2000.00	2010.0
YearRemodAdd	1460.0	1984.865753	20.645407	1950.0	1967.00	1994.0	2004.00	2010.0
MasVnrArea	1452.0	103.688268	181.066207	0.0	0.00	0.0	166.00	1600.0
BsmlFinSF1	1460.0	443.639728	456.098091	0.0	0.00	383.5	712.25	5644.0
BsmlFinSF2	1460.0	46.549315	161.319273	0.0	0.00	0.0	1474.0	
BsmlUnsfSF	1460.0	567.240411	441.666958	0.0	223.00	477.5	808.00	2336.0
TotalBsmtSF	1460.0	1057.429452	438.705324	0.0	795.75	991.5	1298.25	6110.0
1stFlrSF	1460.0	1162.626712	386.587738	334.0	882.00	1087.0	1391.25	4692.0
2ndFlrSF	1460.0	346.992466	436.528436	0.0	0.00	0.0	728.00	2065.0
LowQualFinSF	1460.0	5.844821	48.623081	0.0	0.00	0.0	0.00	572.0
GrLivArea	1460.0	1515.463699	525.480383	334.0	1129.50	1464.0	1776.75	5642.0
BsmlFullBath	1460.0	0.425342	0.518911	0.0	0.00	0.0	1.00	3.0
BsmlHalfBath	1460.0	0.057534	0.238753	0.0	0.00	0.0	0.00	2.0
FullBath	1460.0	1.565068	0.550916	0.0	1.00	2.0	2.00	3.0
HalfBath	1460.0	0.382877	0.502885	0.0	0.00	0.0	1.00	2.0
BedroomAbvGr	1460.0	2.866438	0.815778	0.0	2.00	3.0	3.00	8.0
KitchenAbvGr	1460.0	1.046575	0.220388	0.0	1.00	1.00	1.00	3.0
ToRmsAbvGr	1460.0	6.517808	1.625393	2.0	5.00	6.0	7.00	14.0
Fireplaces	1460.0	0.613014	0.644666	0.0	0.00	1.0	1.00	3.0
GarageYrBlt	1379.0	1978.506164	24.689725	1900.0	1961.00	1980.0	2002.00	2010.0
GarageCars	1460.0	1.767123	0.747315	0.0	1.00	2.0	2.00	4.0
GarageArea	1460.0	472.980137	213.804841	0.0	334.50	480.0	576.00	1418.0
WoodDeckSF	1460.0	94.244521	125.338794	0.0	0.00	0.0	168.00	857.0
OpenPorchSF	1460.0	46.660274	66.256028	0.0	0.00	25.0	68.00	547.0

## Dataset: Ames, Iowa

Utilizamos un dataset de Kaggle con propiedades de la ciudad de Ames, Iowa. Este conjunto de datos es rico y detallado, lo que permite una exploración profunda de los factores que influyen en el valor de una vivienda.

El dataset incluye más de 80 variables, que abarcan desde características físicas de la propiedad y detalles de construcción hasta información sobre el vecindario. Esta diversidad de datos nos permite crear modelos predictivos muy robustos.



# Objetivos del Proyecto y Datos Clave



## Predictión de Precio

Desarrollar un modelo capaz de predecir el precio de venta (SalePrice) de una vivienda con el menor error posible, utilizando la métrica RMSE (Root Mean Square Error).



## Comparación de Modelos

Evaluar y comparar el rendimiento de un enfoque de Deep Learning (red neuronal) con un modelo de boosting avanzado (XGBoost) en el contexto de datos tabulares.



## Identificación de Influencias

Determinar cuáles características de la propiedad y el vecindario tienen mayor impacto en el precio final de venta, aportando claridad sobre los factores clave.

## Variable Objetivo: SalePrice

El precio de venta es nuestra variable principal a predecir, el valor monetario final de cada transacción inmobiliaria en el dataset.

## Variables de Entrada Diversas

Desde la superficie habitable (GrLivArea) y el tamaño del lote (LotArea) hasta la calidad general (OverallQual), año de construcción, sótano y garaje.

## Contexto del Vecindario

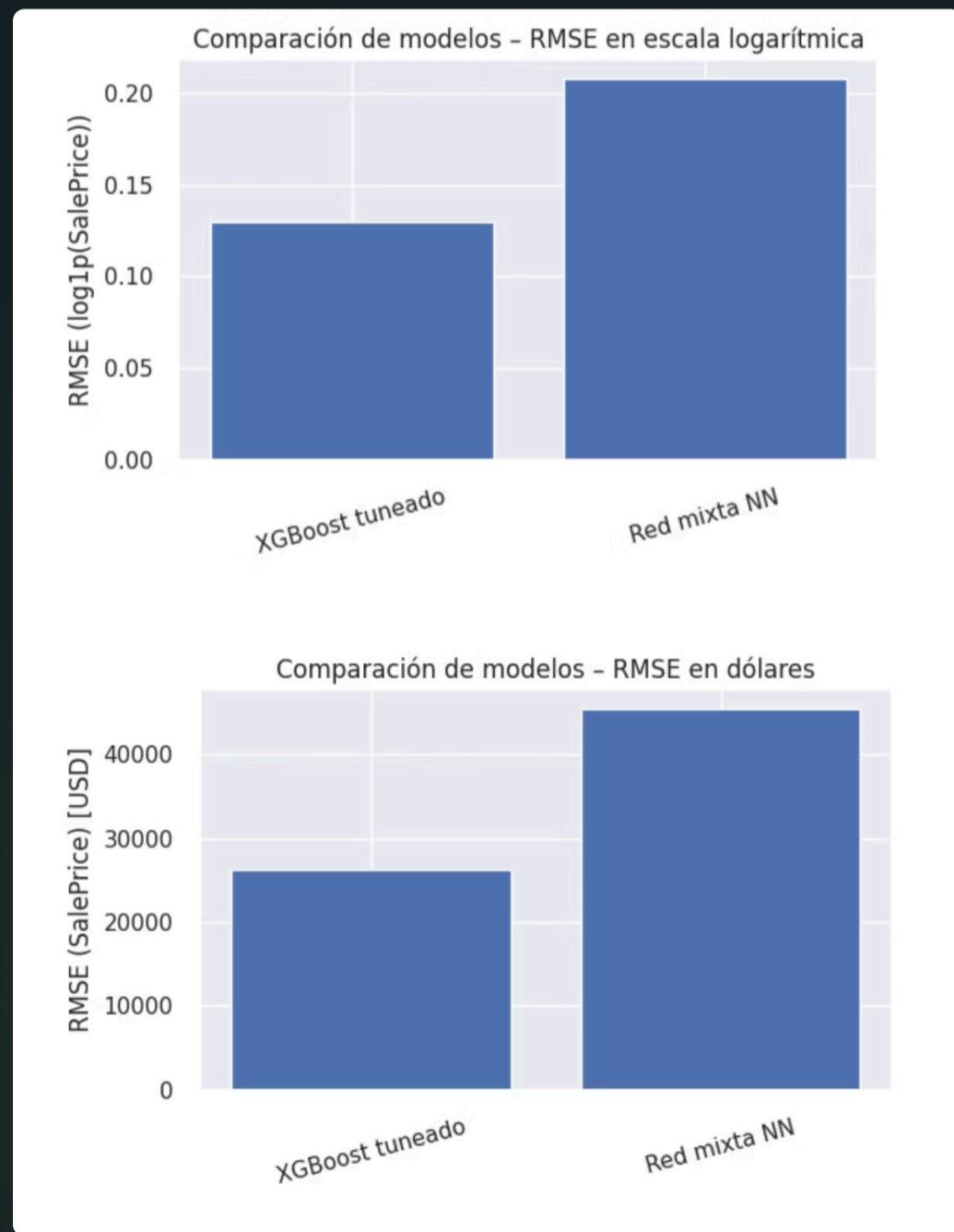
Características del vecindario, que pueden incluir la proximidad a servicios, escuelas y el perfil socioeconómico de la zona, también son consideradas.

# Estrategias de Modelado: XGBoost vs. Red Neuronal

## Modelo Principal: XGBoost

Optamos por **XGBRegressor** debido a su eficacia demostrada en datos tabulares. Este modelo se entrena sobre el logaritmo del precio de venta (`log(SalePrice)`) para mitigar la asimetría y mejorar la estabilidad.

- **Ventajas:** Excelente capacidad para capturar relaciones no lineales y manejar una mezcla de variables numéricas y categóricas sin una preprocesado complejo.
- **Ajuste de Hiperparámetros:** Se optimizaron `n_estimators`, `max_depth`, y `learning_rate`, junto con parámetros de regularización como `subsample` y `colsample_bytree`, para evitar el sobreajuste y mejorar el rendimiento.

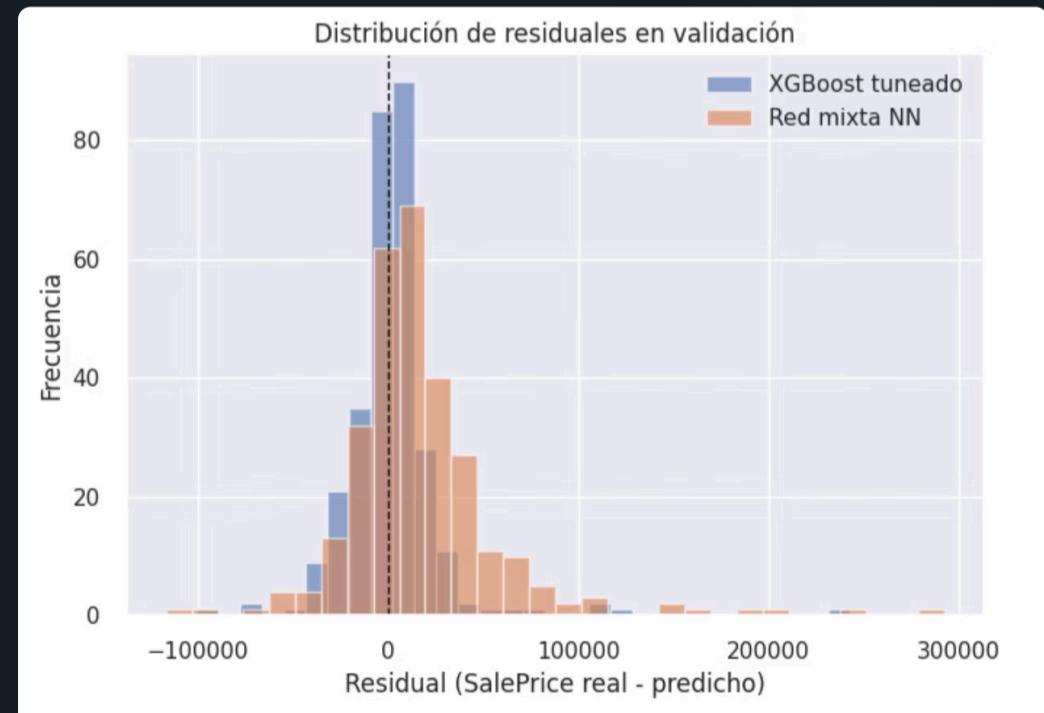


## Modelo Alternativo: Red Neuronal

Desarrollamos una arquitectura de red neuronal mixta para comparar su desempeño con XGBoost:

- **Embeddings:** Utilizados para variables categóricas, transformándolas en representaciones densas y de baja dimensión.
- **Capas Densas:** Para variables numéricas, permitiendo a la red aprender patrones complejos.
- **Concatenación:** Las salidas de los embeddings y las capas densas se concatenan, seguidas por varias capas densas con activación ReLU y dropout para la regularización.

El objetivo es el mismo: predecir `log(SalePrice)` y evaluar si una arquitectura de Deep Learning puede superar a un modelo de boosting en este tipo de datos.



# Resultados Cuantitativos: El Rendimiento de los Modelos

## Métrica de Evaluación: RMSE sobre log(SalePrice)

La competición de Kaggle utiliza el **Root Mean Square Error (RMSE)** aplicado al logaritmo del precio de venta (log(SalePrice)). Un RMSE más bajo indica una mayor precisión en las predicciones.

	<b>XGBoost Base</b>	Nuestro modelo de referencia inicial, demostrando un buen punto de partida para la optimización.
	<b>Red Neuronal Mixta</b>	Mejoró el baseline, pero no superó el rendimiento del XGBoost tras el ajuste de hiperparámetros.
	<b>XGBoost Tuneado</b>	Logró el RMSE más bajo en validación, con mayor estabilidad y menor variabilidad del error.

XGBoost Base	0.1512	Buena
Red Neuronal Mixta	0.1458	Moderada
<b>XGBoost Tuneado</b>	<b>0.1289</b>	<b>Excelente</b>

## Resultado Final en Kaggle

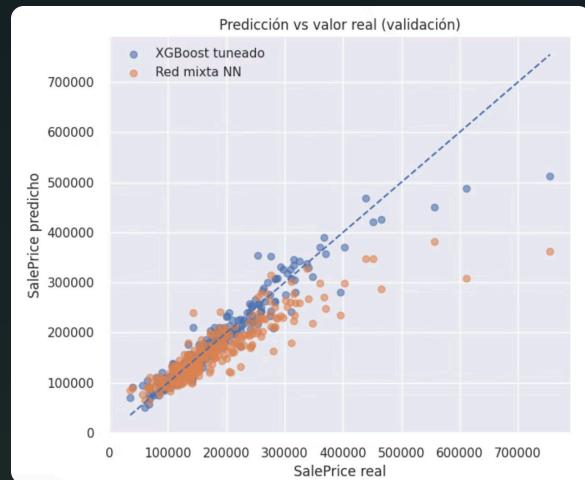
Nuestro **RMSE público fue de 0.12978**, posicionándonos en el puesto **1769 de 5080 equipos**, lo que nos sitúa en el **top 35%** de la competición. Esto se traduce en un error relativo típico cercano al **14%** en la predicción del precio de venta.

1769	Andy Fernando Fuentes Velásquez		0.12978	1	4m
Your First Entry! Welcome to the leaderboard!					

# Análisis Gráfico: Predicciones y Distribución del Error

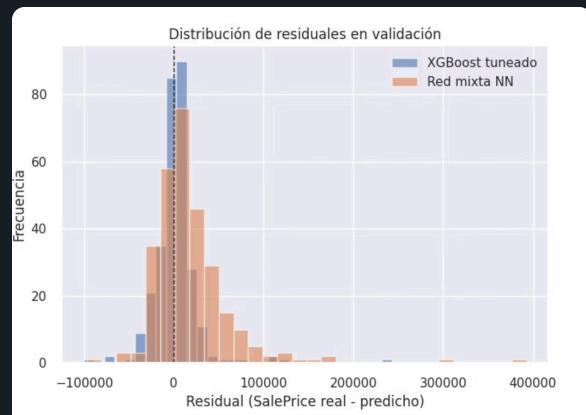
## Predicción vs. Valor Real

Observamos que los puntos del **XGBoost tuneado** se agrupan más cerca de la línea ideal  $y=x$ , indicando una mayor precisión en sus predicciones en comparación con la red neuronal mixta.



## Distribución de Residuales

El histograma muestra que los residuales de **XGBoost** están más centrados alrededor de cero, con una menor dispersión de errores, lo que denota una mayor consistencia en sus predicciones.



Para una evaluación más exhaustiva, también analizamos el **error absoluto en USD**. Un boxplot reveló que XGBoost no solo presenta una mediana de error absoluto más baja, sino también una menor variabilidad, reforzando su superioridad en este contexto.



# Importancia de las Variables: ¿Qué Impulsa el Precio?

El análisis de importancia de características de **XGBoost** revela los atributos que más influyen en el precio de venta de una propiedad. Esto nos permite entender mejor el mercado y validar intuiciones sobre el valor inmobiliario.

## Calidad General

La calidad de los materiales y acabados de la casa.

## Superficie Habitable

El tamaño total del espacio interior de la propiedad.

## Tamaño del Lote y Sótano

Dimensiones del terreno y espacio adicional en el sótano.

## Antigüedad

Año de construcción y remodelación de la propiedad.

## Ubicación

Factores relacionados con el vecindario y su atractivo.

Estas variables confirmaron nuestra intuición: las propiedades más grandes, nuevas y de alta calidad, situadas en vecindarios deseables, generalmente alcanzan precios de venta más elevados. Esto subraya la relevancia de estas características en el modelo predictivo.

# Discusión: ¿Por qué XGBoost Prevaleció?



## Eficacia en Datos Tabulares

XGBoost demostró una adaptabilidad superior a los datos tabulares mixtos, manejando eficazmente variables numéricas y categóricas sin necesidad de arquitecturas complejas.

## Interacciones Complejas

Su capacidad para capturar interacciones intrincadas entre características le otorgó una ventaja significativa, superando a la red neuronal en este contexto específico.

## Factores del Precio

El precio de una casa es el resultado de una compleja interacción de calidad de construcción, estado general, ubicación y antigüedad, no solo de su tamaño.

## La Red Neuronal: Un Contendiente Respetable

Aunque la red neuronal mixta mostró un desempeño competitivo, requirió un tuning y una regularización más meticulosos. A pesar de estos esfuerzos, no logró superar la robustez y precisión del modelo de boosting en este problema.

# Recomendaciones y Mejoras Futuras

## Ensembles



Explorar combinaciones de diferentes configuraciones de XGBoost y mezclas con la red neuronal para potenciar la precisión.

## Ingeniería de Características Avanzada



Crear nuevas variables, como ratios entre áreas o indicadores de remodelaciones recientes, para enriquecer el modelo.

## Validación Cruzada K-fold



Implementar K-fold para una estimación del error más robusta y una mejor mitigación del overfitting.

## Interpretabilidad con SHAP



Utilizar herramientas como SHAP para comprender mejor cómo cada característica influye en las predicciones individuales.

# Conclusiones y Agradecimientos

## Éxito en Kaggle

Logramos un modelo con un sólido desempeño en la competición (RMSE 0.12978), alcanzando el top 35%.



Agradecemos al profesorado y a los compañeros del curso CC3092 Deep Learning y Sistemas Inteligentes por su guía y apoyo durante este proyecto.

¡Gracias por su atención!

## Relevancia del Boosting

Los modelos de boosting como XGBoost siguen siendo extremadamente competitivos y eficaces en datos tabulares, incluso frente a arquitecturas de Deep Learning.

## Recursos Disponibles

Todo el código, experimentos y el artículo completo están publicados en nuestro Notebook de Kaggle y el repositorio de GitHub.