Universidad del Valle de Guatemala
Facultad de Ingeniería
Minería de Datos

Proyecto #2

Entrega final

Autor/Carnet:

María José Villafuerte 22129
Fabiola Contreras 22787

Catedrático:

Mario Barrientos

Sección 20

Fecha:

04/05/2025

Índice

I. Introducción

El presente informe constituye la culminación del proyecto de análisis predictivo desarrollado para InmoValor S.A., con el objetivo de optimizar sus estrategias de valoración inmobiliaria mediante técnicas de Machine Learning. En un mercado inmobiliario caracterizado por su complejidad y múltiples variables influyentes, la capacidad de realizar estimaciones precisas del valor de propiedades representa una ventaja competitiva fundamental para cualquier empresa del sector.

Este proyecto se ha desarrollado utilizando el conjunto de datos "House Prices - Advanced Regression Techniques" de Kaggle, que comprende información detallada sobre 1,460 propiedades residenciales, con 81 variables descriptivas que abarcan desde características físicas como superficie y número de habitaciones, hasta factores cualitativos como la calidad de los acabados y la ubicación. Este conjunto de datos, por su complejidad, proporciona un escenario ideal para evaluar y comparar el rendimiento de diversos algoritmos predictivos.

El desarrollo del proyecto ha seguido una metodología estructurada en cinco fases principales. En primer lugar, se realizó un Análisis Exploratorio de Datos (EDA), lo cual permitió identificar patrones iniciales, correlaciones y el comportamiento de variables clave que influyen en el precio de las propiedades. Esta etapa es fundamental para comprender la estructura del conjunto de datos y orientar los pasos siguientes del análisis. Luego, se llevó a cabo una segmentación mediante técnicas de clustering, con el objetivo de identificar grupos naturales dentro del mercado inmobiliario. Estos grupos comparten características comunes y presentan tendencias de precios similares, lo que permitió una comprensión más granular del comportamiento del mercado.

Se consideraron siete familias de algoritmos: Regresión Lineal, Árboles de Decisión y Random Forest, Naive Bayes, K-Nearest Neighbors (KNN), Regresión Logística, Support Vector Machines (SVM) y Redes Neuronales Artificiales (RNA). La diversidad de enfoques permitió comparar distintos métodos de aprendizaje automático en función de su capacidad para predecir con precisión el precio de venta de las propiedades.

Finalmente, se realizó un análisis comparativo del rendimiento de estos modelos utilizando métricas relevantes para el negocio inmobiliario, para poder identificar la solución más adecuada según el contexto del problema y los datos disponibles. Este enfoque ha permitido desarrollar herramientas predictivas de alta precisión, y generar conocimientos valiosos sobre los factores determinantes del valor inmobiliario en el mercado analizado, conocimientos que pueden traducirse en ventajas estratégicas para InmoValor S.A.

II. Resumen de Análisis Exploratorio

El conjunto de datos analizado se compone de información sobre viviendas, con dos archivos principales: un conjunto de entrenamiento con 1,460 registros y 81 columnas, y un conjunto de prueba con 1,459 registros y 80 columnas (faltando únicamente la variable objetivo SalePrice). El dataset contiene 35 variables numéricas (de tipo int64) y 43 variables categóricas (de tipo object), las cuales caracterizan diferentes aspectos de las propiedades residenciales

Una inspección preliminar revela características importantes del mercado inmobiliario estudiado:

- Las casas fueron construidas entre 1872 y 2010, con un año medio de construcción de 1971
- La mayoría de las propiedades cuentan con 2-3 dormitorios y 1-2 baños completos
- El área habitable promedio (GrLivArea) es de 1,515 pies cuadrados

Se identificaron valores faltantes en varias columnas, como LotFrontage (259 valores nulos), MasVnrType (872 valores nulos) y PoolQC (1,453 valores nulos). Sin embargo, estos valores nulos generalmente corresponden a características que "no aplican" para ciertas propiedades, más que a información realmente perdida.

La variable objetivo SalePrice muestra las siguientes características:

- Rango amplio desde $34,900 hasta $755,000
- Precio medio de venta: $180,921
- Precio mediano (percentil 50): $163,000
- La diferencia entre media y mediana ($17,921) confirma una distribución asimétrica positiva
- El percentil 75 ($214,000) está considerablemente más cercano a la mediana que al valor máximo, indicando la presencia de valores extremos en el límite superior

El análisis de correlación identifica las variables numéricas con mayor influencia en el precio:

1. **OverallQual** (Calidad general de la vivienda): correlación de 0.79 con SalePrice
    - Muestra una relación casi exponencial con el precio
    - Las casas con calificaciones más bajas (1-3) tienen precios por debajo de $100,000
    - Las propiedades con calificaciones altas (9-10) superan los $400,000
2. **GrLivArea** (Área habitable sobre nivel del suelo): correlación de 0.71
    - Relación lineal positiva clara
    - Mayor dispersión en propiedades de gran tamaño
3. **Características del garaje**:
    - GarageCars: correlación de 0.64
    - GarageArea: correlación de 0.62
4. **Superficies adicionales**:
    - TotalBsmtSF (Área total del sótano)
    - 1stFlrSF (Área del primer piso)
5. **YearBuilt** (Año de construcción):
    - Casas más recientes (1990-2010) tienen precios significativamente más altos
    - Propiedades de 1890 muestran valores relativamente altos para su antigüedad
    - A partir de 1950 se observa un incremento gradual en precios a medida que las construcciones son más recientes

6. **YearRemodAdd vs. YearBuilt**:
    - Las casas remodeladas muestran mayor variabilidad en precios
    - Presentan más valores atípicos en el extremo superior
    - Sin embargo, la remodelación por sí sola no garantiza un aumento sustancial en el precio mediano

Entre las variables categóricas, destaca especialmente:
Neighborhood (Vecindario): Muestra diferencias significativas en precios

- Los vecindarios con valores más altos son NoRidge, NridgHt y StoneBr
- Existe una estratificación clara de precios según la ubicación

Otras variables categóricas importantes incluyen ExterQual (Calidad del exterior) y KitchenQual (Calidad de la cocina).

El análisis exploratorio permite identificar las variables más predictivas del precio de viviendas:

1. La calidad general (OverallQual) es el predictor más importante, con una relación casi exponencial con el precio
2. El área habitable (GrLivArea) muestra una fuerte correlación positiva
3. Las características del garaje, superficies habitables y antigüedad son factores significativos
4. La ubicación (Neighborhood) genera diferencias sustanciales en el valor de propiedades similares
5. La distribución asimétrica de SalePrice y la presencia de valores atípicos sugieren la conveniencia de aplicar una transformación logarítmica antes del modelado predictivo

Estos hallazgos proporcionan una base sólida para la selección de variables y el desarrollo de modelos predictivos robustos del precio de viviendas

III. Clustering

Como parte del análisis exploratorio del proyecto de valoración para InmoValor S.A., también se aplicaron técnicas de clustering para segmentar el mercado inmobiliario y detectar patrones relevantes que influyen en el precio de las propiedades. Estos fueron los principales hallazgos:

- **Calidad y Tamaño**: Las propiedades grandes y de alta calidad alcanzan precios hasta cinco veces mayores que las pequeñas de baja calidad, con un efecto multiplicador entre ambas variables más que aditivo.
- **Época y Remodelación**: Las viviendas contemporáneas remodeladas presentan los valores medianos más altos. Las remodelaciones también elevan considerablemente el valor de viviendas históricas.
- **Tipología y Estilo**: Las casas unifamiliares de dos pisos registran el precio promedio más alto (~$226,000), destacándose como segmento premium.
- **Ubicación**: La zona geográfica es el principal diferenciador de precio, con diferencias de hasta 300% entre barrios. Las zonas residenciales de baja densidad en barrios valorizados concentran los precios más altos.
- **Amenidades**: Las propiedades con múltiples amenidades (chimeneas, terrazas amplias, garajes grandes) alcanzan promedios de ~$250,000, superando ampliamente a los segmentos estándar y básicos.

- **Tamaño del terreno**: Se observa una relación creciente entre superficie del lote y precio, especialmente marcada en terrenos excepcionalmente grandes (>50,000 sq ft).

Estas segmentaciones permitieron identificar variables clave para el modelado predictivo (como OverallQual, Neighborhood, GrLivArea, entre otras), y brindaron una base sólida para estimaciones más precisas y segmentadas en el proceso de valoración.

IV. Modelos de regresión lineal

Modelo de Regresión

El documento analiza exclusivamente modelos de regresión lineal para la predicción de precios de viviendas. Se desarrollaron tres modelos con diferentes niveles de complejidad:

1. Modelo Univariado
   - Variables utilizadas: Únicamente 'OverallQual' (calidad general en escala 1-10)
   - Ecuación: price_pred = 45304.8124*area -95782.2317
   - Métricas: $R^2 = 0.64$, MSE = 2,350,108,583.96, RMSE = 48,477.92
   - Resultados: Capacidad predictiva limitada, con heterocedasticidad visible en los residuales

2. Modelo Multivariable Completo (26 variables)
   - Variables utilizadas: Todas las variables numéricas disponibles (LotArea, OverallQual, YearBuilt, GrLivArea, GarageCars, etc.)
   - Métricas: $R^2 = 0.77$, MSE = 1,580,045,490.15, RMSE = 39,749.79
   - Problemas identificados:
     - Multicolinealidad entre grupos de variables relacionadas (componentes del sótano, áreas habitables, características del garaje)
     - Señales de overfitting al incluir variables redundantes
     - Persistencia de heterocedasticidad

3. Modelo Optimizado (9 variables)
   - Variables seleccionadas: 'OverallQual', 'OverallCond', 'YearBuilt', 'BsmtFullBath', 'TotalBsmtSF', 'GrLivArea', 'GarageCars', 'WoodDeckSF', 'FullBath'
   - Métricas: $R^2 = 0.82$, MSE = 1,174,827,939.01, RMSE = 34,275.76
   - Ventajas:
     - Mayor poder predictivo que modelos anteriores (32% mejor que el modelo univariado)
     - Elimina problemas de multicolinealidad al evitar variables redundantes
     - Mejor equilibrio entre simplicidad y capacidad predictiva
     - Mayor interpretabilidad y estabilidad

Variables con Mayor Impacto

- OverallQual: +$27,330 por cada punto adicional en calidad
- GrLivArea: +$26,161 por cada pie cuadrado adicional
- GarageCars: +$11,137 por cada espacio para auto
- YearBuilt: Efecto positivo significativo en precio

Limitaciones Identificadas

- Todos los modelos tienden a subestimar propiedades de alto valor
- Persistencia de heterocedasticidad incluso en el modelo optimizado
- Posible necesidad de transformaciones no lineales para mejorar predicciones

A partir de este punto para los modelos de regresión, se seleccionaron las siguientes características clave:

- Numéricas: OverallQual (calidad general), GrLivArea (área habitable), GarageCars (capacidad del garaje), YearBuilt (año de construcción), TotalBsmtSF (superficie del sótano), FullBath (baños completos)
- Categóricas: Neighborhood (vecindario), ExterQual (calidad exterior), KitchenQual (calidad de cocina), BsmtQual (calidad del sótano)

Los datos se dividieron en conjuntos de entrenamiento (80%) y prueba (20%), y se aplicó preprocesamiento mediante:

- Estandarización para variables numéricas
- Codificación one-hot para variables categóricas

Modelo de Clasificación

El documento analizado no contiene información sobre modelos de clasificación. El estudio se enfoca exclusivamente en modelos de regresión lineal para predecir la variable continua del precio de las viviendas (SalePrice).

V. Modelo de Árboles de decisión

Modelo de Regresión

El análisis comparativo de varios modelos de árboles de decisión para predecir precios de viviendas reveló que:

- El árbol de decisión con profundidad 7 fue el mejor modelo de regresión simple, con:
  - R^2 de 0.8371 (explica el 83.71% de la variabilidad en precios)
  - MAE de 23,038.76
  - MSE de 1,119,462,689.92
- Este modelo superó significativamente al modelo de regresión lineal (R^2 de 0.8174), demostrando que los árboles de decisión capturan mejor las relaciones no lineales presentes en los datos inmobiliarios.
- Sin embargo, el Random Forest con profundidad 7 superó a ambos modelos anteriores:
  - R^2 de 0.8787 (explica el 87.87% de la variabilidad)
  - MAE de 19,573.69
  - MSE de 831,844,212.04
- Las visualizaciones confirman que Random Forest presentó una distribución de errores más concentrada alrededor de cero y con menor dispersión hacia los extremos.

Modelo de Clasificación

Al desarrollar este modelo se estableció, desde este punto utilizar las mismas características que en los modelos de regresión, pero la variable objetivo se transformó en categorías discretas basadas en terciles:

- Casas Económicas: Primer tercil de precios
- Casas Intermedias: Segundo tercil de precios
- Casas Caras: Tercer tercil de precios

Para la clasificación de propiedades en categorías de precio (Económicas, Intermedias, Caras), basadas en los terciles de la distribución ($139,685 y $190,000):

- El mejor árbol de decisión simple fue el de profundidad 7:
    - Exactitud global de 76.48%
    - Mejor desempeño en la categoría "Cara" (F1-score de 0.8374)
    - Mayor dificultad en la categoría "Intermedia" (F1-score de 0.6803)
- La matriz de confusión reveló que, de 438 propiedades evaluadas, 322 (73.5%) fueron correctamente categorizadas, con confusiones principalmente entre categorías adyacentes.
- El Random Forest mejoró estos resultados:
    - Exactitud global de 80.14%
    - Mejora notable en la categoría "Intermedia" (F1-score de 0.72)
    - Mejor identificación de viviendas "Caras" (F1-score de 0.87)

VI. Modelo de Bayes Ingenuo

Modelo de Regresión

El modelo de Naive Bayes aplicado a la regresión para predecir precios inmobiliarios mostró un desempeño considerablemente deficiente. Los resultados indicaron:

- R^2: 0.6356 (solo explica alrededor del 63.56% de la varianza)
- MAE: 31,814.55
- RMSE: 49,301.62

Estos errores de predicción elevados evidencian que el modelo tiene dificultades para hacer predicciones precisas. La visualización de errores reflejó una dispersión notable de puntos a medida que los valores reales aumentaban, sugiriendo un problema de sesgo.

Causas del Bajo Rendimiento

El principal problema identificado fue la violación del supuesto de independencia de Naive Bayes. El mapa de correlación reveló relaciones significativas entre variables clave:

- "OverallQual" y "GrLivArea" (0.72)
- "YearBuilt" y "OverallCond" (0.87)
- "FullBath" y "GrLivArea" (0.80)

Estas fuertes correlaciones impiden que el modelo capture correctamente las relaciones entre variables, llevando a suposiciones incorrectas sobre la distribución de los datos.

Optimización del Modelo de Regresión

Se implementó una optimización de hiperparámetros mediante GridSearchCV, ajustando el parámetro 'var_smoothing' a 1e-06, lo que mejoró:

- R^2: aumentó de 0.6416 a 0.6753
- MAE: disminuyó de 31,293.24 a 29,775.87 (mejora del 4.9%)

El modelo optimizado predijo con mayor precisión los valores en el rango medio ($100,000-$300,000), aunque continuó subestimando las propiedades de alto valor (>$300,000).

Modelo de Clasificación

Se implementó un modelo de Naive Bayes Gaussiano para clasificar propiedades en tres categorías de precio (Económica, Intermedia, Cara). El modelo inicial obtuvo:

- Accuracy, precision, recall y F1-score: 72.1%

La matriz de confusión reveló que el modelo clasificaba bien las propiedades "Económicas" (121 correctas de 145) e "Intermedias" (123 correctas de 146), pero tenía mayor dificultad con las "Caras" (solo 72 correctas de 147).

Validación Cruzada y Selección de Características

La implementación de validación cruzada de 5 folds con características clave seleccionadas (OverallQual, GrLivArea, YearBuilt, TotalBsmtSF, GarageCars) mejoró significativamente el rendimiento:

- Exactitud promedio: 78.70%
- Baja desviación estándar: 0.0109

El modelo simplificado mostró mejoras notables en la categoría "Intermedia" (136 clasificaciones correctas) y redujo las confusiones entre categorías extremas.

Optimización de Hiperparámetros

La optimización mediante GridSearchCV, ajustando 'var_smoothing' a 1e-05, elevó aún más el rendimiento:

- Accuracy: aumentó de 72.15% a 76.48%
- Mejoras en la identificación de propiedades "Caras" (92 vs 72 clasificaciones correctas)
- Reducción de errores graves "Cara→Económica" (de 29 a solo 12 casos)

VII. Modelo de KNN

Modelo de Regresión

El modelo optimizado de regresión KNN demostró ser una herramienta eficaz para la predicción precisa de precios inmobiliarios con las siguientes características:

- Configuración óptima: 4 vecinos, ponderación por distancia y métrica Manhattan (p=1)
- Rendimiento: R^2 de 0.8664, explicando aproximadamente el 87% de la variabilidad en los precios
- Error de predicción: RMSE promedio de $32,747 con una desviación estándar de $2,426
- Error absoluto medio: $19,854
- Tiempo de ejecución: Extremadamente rápido (0.0159s)

La validación cruzada con 5 folds mostró una notable estabilidad del modelo, con valores de RMSE consistentes entre $29,223 y $35,179. El análisis de hiperparámetros reveló que la combinación de

distancia Manhattan con el algoritmo 'ball_tree' y ponderación por distancia proporcionó resultados óptimos.

Modelo de Clasificación

El modelo de clasificación KNN categoriza las propiedades en tres segmentos de precio: "Económica", "Intermedia" y "Cara", con umbrales en $139,699 y $190,000:

- Configuración optimizada: 12 vecinos, algoritmo 'ball_tree', distancia Manhattan (p=1) y ponderación por distancia
- Precisión global: 75% (tras optimización de hiperparámetros)
- Rendimiento por categoría:
    - Propiedades "Económicas": Precisión de 0.82, Recall de 0.75
    - Propiedades "Intermedias": Precisión de 0.65, Recall de 0.66
    - Propiedades "Caras": Precisión de 0.78, Recall de 0.84
- Tiempo de ejecución: 0.0050s

VIII. Modelo de Regresión logística

Modelo de Regresión Logística

El análisis de regresión logística evaluó la capacidad de clasificar propiedades inmobiliarias en tres categorías: casas caras, económicas e intermedias. Los resultados muestran un desempeño distintivamente diferente para cada categoría:

Desempeño por Categoría:

- Casas Caras: Excelente desempeño con 283 verdaderos negativos y 120 verdaderos positivos. El AUC de 0.98 indica una capacidad excepcional para identificar propiedades de alta gama. Un modelo posterior (60%) mostró 242 verdaderos negativos y 156 verdaderos positivos con AUC de 0.96.
- Casas Económicas: Muy buen rendimiento con 249 verdaderos negativos y 136 verdaderos positivos. El AUC de 0.95 confirma alta confiabilidad para identificar propiedades económicas.
- Casas Intermedias: Rendimiento deficiente con 285 verdaderos negativos pero solo 1 verdadero positivo y 143 falsos negativos. El AUC de 0.61 indica que el modelo apenas supera la clasificación aleatoria.

Optimización del Modelo:

- Se detectó un problema de multicolinealidad entre variables predictoras como GrLivArea/FullBath (0.63) y OverallQual/GarageCars (0.60).
- El proceso de tuneo con GridSearchCV identificó los mejores hiperparámetros: C=1.0, penalty=l2, solver=newton-cg (accuracy promedio 0.905749).
- El ajuste del umbral de clasificación de 0.5 a 0.46 mejoró levemente las métricas (accuracy 0.9110, F1-score 0.8895).
- No se detectó overfitting significativo, con diferencias pequeñas entre métricas de entrenamiento y prueba (accuracy 0.8914 vs 0.8790).

Eficiencia Computacional:

- El modelo original fue extremadamente eficiente: 36,763 llamadas a funciones en 0.066 segundos.
- El modelo original superó al tuneado en términos de AIC/BIC (235.69/264.26 vs 243.72/272.29) y eficiencia (2,675 llamadas en 0.033s vs 3,324 llamadas en 0.066s).

Otros modelos de Clasificación

Se evaluaron diversos modelos de clasificación para comparar su rendimiento en la identificación de propiedades caras vs. no caras:

Árbol de Decisión:

- Logró clasificar correctamente 238 casos negativos y 138 positivos
- Cometió 37 falsos negativos y 25 falsos positivos
- Tiempo de entrenamiento: 0.036 segundos
- Fortaleza: Facilidad de interpretación

Random Forest:

- Clasificó correctamente 241 casos negativos y 147 positivos
- Registró 22 falsos positivos y 28 falsos negativos
- Tiempo de entrenamiento: 2.41 segundos (el más lento)
- Fortaleza: Mayor robustez y menor tendencia al sobreajuste

Naive Bayes:

- Identificó correctamente 214 casos negativos y 156 positivos
- Mayor número de falsos positivos (49) pero pocos falsos negativos (19)
- Tiempo de entrenamiento: 0.020 segundos (el más rápido)
- Debilidad: Supuesto de independencia entre variables que no siempre se cumple

IX. Modelo de SVM (Máquinas de vectores de soporte)

Modelo de regresión

El modelo SVR (Support Vector Regression) mostró un rendimiento deficiente:

- RMSE (Error Cuadrático Medio): ~80,000 (el más alto entre los modelos comparados)
- R^2: 0.12 (extremadamente bajo)
- MAE (Error Absoluto Medio): ~49,000 (el más elevado)
- Comportamiento: Las predicciones se agruparon horizontalmente alrededor de $180,000 independientemente del precio real, subestimando sistemáticamente las propiedades más caras (>$200,000)

Modelo de clasificación

El análisis comparativo mostró que el SVM con kernel lineal fue el mejor modelo de clasificación SVM, con las siguientes características:

- Precisión: 0.8059 (80.59%)
- Parámetros óptimos: C = 10
- Fortalezas: Demostró un rendimiento sólido especialmente en las categorías "Económica" (130 aciertos) e "Intermedia" (123 aciertos)

- Debilidades: Presentó dificultades para clasificar correctamente la categoría "Cara" (100 aciertos con 47 clasificaciones erróneas)
- Comparación con otros kernels: El kernel lineal presentó el menor sobreajuste (6.39%) comparado con los kernels polinomial (9.78%) y RBF (9.13%)
- Eficiencia: Aunque no es el más rápido en procesamiento, mostró el mejor equilibrio general con 350 clasificaciones correctas, superando al RBF (347) y al Polinomial (339)

X. Modelo de RNA (Red Neuronal Artificial)

Modelo de Regresión

El modelo de regresión se enfoca en predecir el precio exacto de venta de las casas utilizando técnicas de redes neuronales.

Arquitecturas Implementadas

Se implementaron dos modelos de regresión con diferentes arquitecturas:

Modelo 1 (ReLU):

- Dos capas ocultas con 10 y 5 neuronas
- Activación ReLU
- Regularización alpha=0.001
- Batch size=32
- Learning rate=0.001

Modelo 2 (tanh):

- Tres capas ocultas con 20, 10 y 5 neuronas
- Activación tanh
- Regularización alpha=0.0001
- Batch size=64
- Learning rate=0.01

Resultados y Evaluación

La evaluación de los modelos se realizó mediante métricas de regresión (RMSE y R^2) y un análisis de residuos. El Modelo 1 con activación ReLU mostró mejor rendimiento, presentando:

- Menor error cuadrático medio
- Mayor coeficiente de determinación (R^2)
- Mejor generalización en datos de prueba

El análisis de las curvas de aprendizaje no mostró evidencia significativa de sobreajuste en el Modelo 1, con una diferencia entre el R^2 de entrenamiento y prueba menor al 10%.

Modelo de Clasificación

El modelo de clasificación categoriza las casas en tres niveles de precio: Económica, Intermedia y Cara, basándose en terciles del precio de venta.

Arquitecturas Implementadas

Se implementaron dos modelos de clasificación con diferentes arquitecturas:

Modelo 1:

- Dos capas ocultas (5, 3)
- Activación tanh
- Tiempo de procesamiento: ~0.90 segundos

Modelo 2:

- Una capa oculta (10)
- Activación ReLU
- Tiempo de procesamiento: ~0.69 segundos

Resultados y Evaluación

La evaluación de los modelos se realizó mediante métricas de clasificación como accuracy, precision, recall y F1-score, junto con matrices de confusión.

Resultados clave:

- El Modelo 2 (ReLU, una capa de 10 neuronas) mostró mejor rendimiento general con un accuracy del 80.82% vs 80.36% del Modelo 1
- El Modelo 2 también presentó mejor F1-score global, especialmente en la categoría "Económica"
- El Modelo 2 fue más eficiente computacionalmente (menor tiempo de procesamiento)
- Ambos modelos mostraron dificultad para clasificar correctamente las casas de la categoría "Intermedia"

XI. Tabla de resumen – Modelos de Regresión

Nombre del modelo	Métricas	Posición de mejor algoritmo (1 el mejor, 7 el menos recomendado)	Eficiencia temporal
Modelos de regresión lineal	R^2: 0.82 MSE: 1,174,827,939.01 RMSE: 34,275.76	5	0.0481 segundos
Modelo de Árboles de decisión	R^2:0.8371 MAE:23,038.76 MSE:1,119,462,689.92	4	0.0287s
Random Forest	R^2: 0.8787 MAE: 19,573.69 MSE: 831,844,212.04	1	4.4865s
Modelo de Bayes Ingenuo	R^2: 0.6753 MAE: 29,775.87 RMSE: 46,886.78	6	0.0189 segundos
Modelo de KNN	R^2: 0.8664 MAE: 19,854 RMSE: 32,747	3	0.0159
Modelo de SVM (Máquinas de vectores de soporte)	R^2: 0.12 MAE: 49,000 RMSE: 80,000	7	0.0255 segundos
Modelo de RNA (Red Neuronal Artificial)	Modelo con ReLU R^2: ~0.85 RMSE: ~33,000	2	2.5 segundos

XII. Tabla de resumen – Modelos de Clasificación

Nombre del modelo	Métricas	Posición de mejor algoritmo (1 el mejor, 6 el menos recomendado)	Eficiencia
Modelo de Árboles de decisión	Accuracy: 0.7215 Precision: 0.7196 Recall: 0.7215 F1-Score: 0.7183	6	0.0287 segundos
Random Forest	Accuracy: 0.7785 Precision: 0.7750 Recall: 0.7785 F1-Score: 0.7759	4	1.8488 segundos
Modelo de Bayes Ingenuo	Accuracy: 76.48% Precision: 76.70% Recall: 76.48% F1-score: 76.39%	5	0.0189 segundos
Modelo de KNN	Accuracy: 0.7877 Precision: 0.7904 Recall: 0.7877 F1-Score: 0.7887	3	0.01599 segundos
Modelo de Regresión logística	Accuracy: 0.9087 Precision: 0.8814 Recall: 0.8914 F1-score: 0.8864 AUROC: 0.9602	Opción óptima para escenarios de **decisión binaria**.	0.023 segundos
Modelo de SVM (Máquinas de vectores de soporte)	Accuracy: 0.7991 Precision: 0.7958 Recall: 0.7991 F1-Score: 0.7957	2	0.0345 segundos
Modelo de RNA (Red Neuronal Artificial)	Modelo ReLU, topología 10: Accuracy: 80.82% Precision: 80.82% Recall: 80.82% F1-score: 80.82%	1	0.69 segundos

XIII. Conclusiones

Entonces, ¿qué significa todo esto? ¿qué implicaciones tiene para la empresa?

Rendimiento Comparativo de Modelos

1. **Para predicción exacta de precios (regresión)**:
    - El algoritmo **Random Forest** demostró ser el modelo más efectivo, con un R^2 de 0.8787 y un error absoluto medio (MAE) de apenas $19,573, lo que representa aproximadamente el 10.8% del precio medio de venta ($180,921).
    - Le sigue de cerca el modelo de **KNN**, con un R^2 de 0.8664 y un MAE de $19,854, destacándose además por su extraordinaria eficiencia computacional (0.0159 segundos).
    - Las **Redes Neuronales** también mostraron un rendimiento notable (R^2 ~0.85), aunque con mayor costo computacional.
    - Los modelos de **Regresión Lineal optimizados** (R^2 de 0.82) ofrecen un equilibrio atractivo entre precisión, interpretabilidad y eficiencia.
2. **Para categorización de propiedades (clasificación)**:
    - La **Regresión Logística** mostró un rendimiento excepcional para clasificación binaria (casas caras vs. no caras) con una exactitud del 90.87% y un área bajo la curva ROC de 0.96, aunque por su naturaleza está limitada a escenarios de decisión binaria.
    - Para la clasificación multiclase (económica, intermedia, cara), las **Redes Neuronales** y el modelo **KNN** destacaron con exactitudes del 80.82% y 78.77% respectivamente, ofreciendo mayor versatilidad.
    - El **SVM** con kernel lineal también mostró un desempeño sólido (79.91%) manteniendo un buen equilibrio entre precisión y eficiencia computacional.
    - Todos los modelos presentaron dificultades para clasificar correctamente las propiedades de rango intermedio, sugiriendo la necesidad de reevaluar los límites entre categorías o incorporar variables adicionales para este segmento específico.

Factores Determinantes del Valor Inmobiliario

El análisis transversal de todos los modelos confirma que las variables más influyentes en la determinación del precio son:

1. **Calidad general de la vivienda** (OverallQual): Variable con mayor poder predictivo en todos los modelos, con un impacto exponencial en el precio.
2. **Ubicación** (Neighborhood): Factor crítico que puede generar diferencias de hasta 300% entre propiedades similares ubicadas en diferentes zonas.
3. **Superficie habitable** (GrLivArea): Fuerte correlación positiva con el precio, con efecto multiplicativo cuando se combina con alta calidad.
4. **Características del garaje**: Especialmente el número de plazas (GarageCars), con impacto significativo en todos los modelos.
5. **Antigüedad y renovación**: Las propiedades recientes o renovadas muestran valores significativamente superiores.

Implicaciones para el Negocio

1. **Recomendaciones para valoración inmobiliaria**:
    - Implementar un sistema híbrido que combine **Random Forest** para predicciones numéricas precisas y **Redes Neuronales** para segmentación multiclase de propiedades.
    - Utilizar **Regresión Logística** específicamente para clasificaciones binarias de alto valor estratégico (ej. identificar propiedades premium o detectar oportunidades de inversión).
    - Mantener **KNN** como alternativa eficiente para valoraciones rápidas con precisión aceptable, aprovechando su excepcional velocidad de ejecución.
    - Incorporar un modelo de **Regresión Lineal** optimizado para proporcionar explicabilidad a clientes y stakeholders cuando la interpretabilidad sea prioritaria.
2. **Oportunidades de negocio identificadas**:
    - Especialización en la valoración de propiedades de gama alta, donde la mayoría de los modelos mostraron mayor precisión.
    - Desarrollo de productos específicos para detectar propiedades infravaloradas en el mercado, aprovechando la capacidad predictiva de los modelos.
    - Creación de un índice de potencial de revalorización basado en las variables clave identificadas.
3. **Limitaciones y consideraciones**:
    - Todos los modelos tienden a subestimar las propiedades de precio extraordinariamente alto (outliers).
    - Los modelos requieren recalibración periódica para adaptarse a las tendencias cambiantes del mercado.
    - La clasificación de propiedades de rango intermedio sigue siendo un desafío que requiere investigación adicional.

XIV. Enlaces

Repositorio del proyecto: https://github.com/Fabiola-cc/InmoValor_SA
Documento final: Proyecto # 2, Documento Final .docx

Entrega Modelos RNA: https://github.com/Fabiola-cc/InmoValor_SA/tree/main/Análisis/Entrega_7