**Alumno: Rodolfo Nicolás Velasco Fessler**

***Fase 1: Comprensión del Negocio***

1. Definición Clara del Problema de Negocio

El problema de negocio es la **dificultad para valorar propiedades inmobiliarias de forma precisa, rápida y objetiva** en California. Por lo general, la tasación de viviendas es un proceso manual, lento y costoso. Las empresas necesitan un método automatizado y basado en datos para estimar el valor de mercado de una propiedad.

Este proyecto busca resolver esto mediante la creación de un modelo de *machine learning* que prediga el valor medio de una vivienda (median\_house\_value) utilizando características observables de la propiedad y su distrito, como la ubicación (longitude, latitude), la antigüedad (housing\_median\_age), la demografía (population, households) y el poder adquisitivo de la zona (median\_income).

2. Objetivos Específicos y Medibles del Proyecto

Los objetivos del proyecto son:

* Construir un modelo capaz de estimar el median\_house\_value de una propiedad en California.
* Entregar un *notebook* funcional que documente el proceso de limpieza, exploración y modelado, permitiendo que el modelo sea reentrenado o utilizado para nuevas predicciones.

3. Justificación de la Relevancia del Caso Seleccionado

Se seleccionó el **Caso 2: Predicción de Precios de Viviendas** por su alta relevancia y aplicabilidad directa en el mundo real. El sector inmobiliario tiene una gran importancia en el sector económico, por lo que es importante valorar los activos correctamente.

Resolver este problema tiene un impacto directo en:

* Decisiones de Inversión: Permite identificar propiedades infravaloradas o áreas con alto potencial de crecimiento.
* Importancia Inmobiliaria: Proporciona una herramienta rápida para la tasación de propiedades.
* Transparencia de Mercado: Ofrece un precio de referencia objetivo.

4. Identificación de Stakeholders y Usuarios Finales

* **Stakeholders:**
  + **Empresas de Inversión Inmobiliaria:** Buscan maximizar el retorno de inversión (ROI) y necesitan identificar oportunidades de mercado.
  + **Bancos y Entidades Hipotecarias:** Necesitan gestionar el riesgo crediticio asegurando que las propiedades que respaldan los préstamos estén correctamente valoradas.
  + **Desarrolladores Inmobiliarios:** Utilizan las predicciones para decidir dónde construir y qué tipo de propiedades desarrollar.
* **Usuarios Finales:**
  + **Tasadores de Propiedades:** Como herramienta de apoyo para agilizar y fundamentar sus valoraciones manuales.
  + **Agentes Inmobiliarios:** Para asesorar a sus clientes (vendedores y compradores) sobre precios justos de mercado.
  + **Analistas de Datos (del sector financiero o inmobiliario):** Para generar informes de mercado, identificar tendencias y riesgos.

***Fase 2: Comprensión de los Datos***

**1. Diccionario de Datos**

| **Nombre de la Variable** | **Tipo de Dato** | **Descripción** |
| --- | --- | --- |
| longitude | Numérico (Continuo) | Coordenada geográfica que mide la distancia al oeste; un valor más negativo está más al oeste. |
| latitude | Numérico (Continuo) | Coordenada geográfica que mide la distancia al norte; un valor más alto está más al norte. |
| housing\_median\_age | Numérico (Discreto) | Edad media de las viviendas en un distrito. |
| total\_rooms | Numérico (Discreto) | Número total de habitaciones en un distrito. |
| total\_bedrooms | Numérico (Discreto) | Número total de dormitorios en un distrito. |
| population | Numérico (Discreto) | Número total de personas que residen en un distrito. |
| households | Numérico (Discreto) | Número total de hogares (un grupo de personas que residen juntas) en un distrito. |
| median\_income | Numérico (Continuo) | Ingreso medio de los hogares en un distrito (medido en decenas de miles de USD). |
| median\_house\_value | Numérico (Continuo) | (Variable Objetivo) El valor medio de las viviendas en un distrito. |
| ocean\_proximity | Categórico (Nominal) | Describe la ubicación del distrito en relación con el océano/bahía. |

**2. Reporte de Calidad de Datos**

Tras una exploración inicial de los datos, se identificaron los siguientes problemas de calidad que deberán ser tratados antes del modelado:

* **Valores Faltantes:**
  + La columna total\_bedrooms (total de dormitorios) es la única que presenta valores nulos.
  + Se identificaron **207 filas** donde el dato de total\_bedrooms está ausente. Esto representa aproximadamente el 1% del total de los datos.
  + Será necesario aplicar una estrategia de **imputación** (por ejemplo, rellenar con la mediana de la columna) para no perder esas filas.
* **Valores Atípicos (Outliers):**
  + **median\_house\_value (Valor de Vivienda):** Se observa un tope (límite superior) artificial en esta columna. Numerosas propiedades (casi 1,000) están registradas con un valor exacto de **$500,001**. Esto indica que los valores por encima de medio millón de dólares fueron agrupados en esta cifra. Es una característica importante a considerar, ya que el modelo no podrá predecir valores superiores a este.
  + **Variables de Conteo (total\_rooms, population, households):** Estas columnas muestran una distribución muy sesgada (asimétrica) hacia la derecha. Tienen valores mínimos bajos (ej. total\_rooms de 2.0, population de 3.0) y valores máximos extremadamente altos (ej. population de 35,682). Esto sugiere que la mayoría de los distritos tienen valores bajos-medios, pero existen unos pocos distritos con una densidad poblacional y de habitaciones masiva, que actúan como *outliers*.
* **Inconsistencias y Formato:**
  + **ocean\_proximity (Proximidad al Océano):** La columna es categórica y contiene 5 valores únicos: 'NEAR BAY' , '<1H OCEAN' , 'INLAND' , 'NEAR OCEAN' y 'ISLAND'. Los valores son consistentes y no presentan errores tipográficos. Deberán ser convertidos a un formato numérico (Por ejemplo: *One-Hot Encoding*) para el modelado.
  + **Tipos de Datos:** Todos los datos numéricos están correctamente formateados como números.

**3. Estadísticas Descriptivas Básicas**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **longitude** | **latitude** | **housing\_**  **median\_age** | **total\_**  **rooms** | **total\_**  **bedrooms** | **population** | **households** | **Median**  **\_income** | **median\_**  **house\_value** |
| **count** | 20640 | 20640 | 20640 | 20640 | 20433 | 20640 | 20640 | 20640 | 20640 |
| **mean** | -119,57 | 35,63 | 28,64 | 2635,76 | 537,87 | 1425,48 | 499,54 | 3,87 | 206855,8 |
| **std** | 2 | 2,14 | 12,59 | 2181,62 | 421,39 | 1132,46 | 382,33 | 1,9 | 115395,6 |
| **min** | -124,35 | 32,54 | 1 | 2 | 1 | 3 | 1 | 0,5 | 14999 |
| **25%** | -121,8 | 33,93 | 18 | 1447,75 | 296 | 787 | 280 | 2,56 | 119600 |
| **50%** | -118,49 | 34,26 | 29 | 2127 | 435 | 1166 | 409 | 3,53 | 179700 |
| **75%** | -118,01 | 37,71 | 37 | 3148 | 647 | 1725 | 605 | 4,74 | 264725 |
| **max** | -114,31 | 41,95 | 52 | 39320 | 6445 | 35682 | 6082 | 15 | 500001 |

***Fase 3: Preparación de los Datos***

**1. Documentación Detallada de las Transformaciones**

Para preparar los datos para las fases de exploración y modelado, se siguieron los siguientes pasos:

**1. Tratamiento de Valores Faltantes (Columna: total\_bedrooms)**

* **Problema:** La columna total\_bedrooms tenía 207 valores faltantes (nulos).
* **Estrategia:** Se utilizó la **imputación por la mediana**.
* **Justificación:** Como se observó en la Fase 2, la distribución de total\_bedrooms es asimétrica (sesgada a la derecha), con algunos distritos teniendo una cantidad muy alta de dormitorios (outliers). Usar la *media* sería arriesgado, ya que estos outliers la desviarían. La **mediana** (el valor central, que fue de $435.0$) es mucho más robusta a los outliers y representa mejor al distrito "típico", siendo la opción más segura para rellenar los datos faltantes.

**2. Creación de Variables Derivadas (Ingeniería de Características)**

* **Problema:** Las variables total\_rooms, total\_bedrooms y population son conteos totales por distrito. Esto puede ser engañoso ya que un distrito grande siempre va a tener más habitaciones.
* **Estrategia:** Se crearon tres nuevas características basadas en ratios por hogar:
  1. rooms\_per\_household (total\_rooms / households)
  2. population\_per\_household (population / households)
  3. bedrooms\_per\_room (total\_bedrooms / total\_rooms)
* **Justificación:** Estas nuevas variables son más informativas porque nos dicen la *densidad* de población por hogar y la *proporción* de habitaciones y dormitorios, lo que tiene una relación mucho más fuerte con el valor de la vivienda que los conteos totales brutos.

**3. Procesamiento de Datos Categóricos (Columna: ocean\_proximity)**

* **Problema:** La columna ocean\_proximity es texto ('INLAND', 'NEAR BAY', etc.), y los modelos de machine learning solo entienden números.
* **Estrategia:** Se aplicó la técnica de **One-Hot Encoding**.
* **Justificación:** Esta técnica convierte la columna de texto en múltiples columnas binarias (una para cada categoría. Por ejemplo: ocean\_proximity\_INLAND, ocean\_proximity\_NEAR BAY). Esto es preferible a asignar un número (Por ejemplo: INLAND=1, NEAR BAY=2), ya que evita que el modelo asuma una relación de orden o jerarquía falsa entre las categorías.

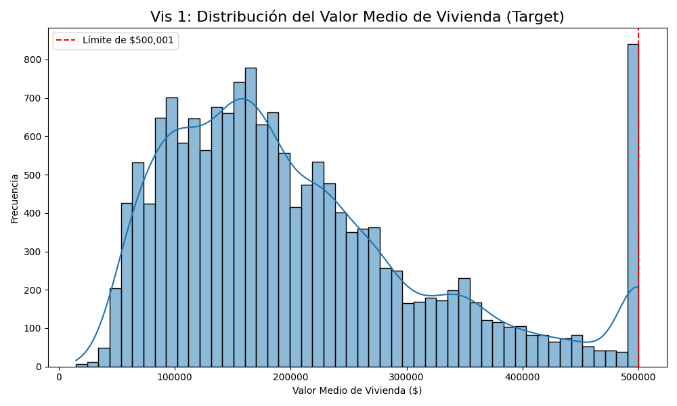
**4. División en Conjuntos de Entrenamiento y Prueba (80/20)**

* **Problema:** Necesitamos separar nuestros datos para entrenar el modelo (con la mayoría de los datos) y para evaluarlo (con datos que nunca ha visto).
* **Estrategia:** Se utilizó un **Muestreo Estratificado** (StratifiedShuffleSplit) basado en la columna median\_income.
* **Justificación:** Una división aleatoria simple podría resultar en que, por pura casualidad, el conjunto de entrenamiento o el de prueba tengan una proporción muy diferente de distritos de altos y bajos ingresos. Dado que median\_income es un predictor tan fuerte del precio, esto sesgaría nuestra evaluación. El muestreo estratificado asegura que la distribución de ingresos sea representativa y casi idéntica en ambos conjuntos de datos.

***Fase 4: Análisis Exploratorio y Visualización***

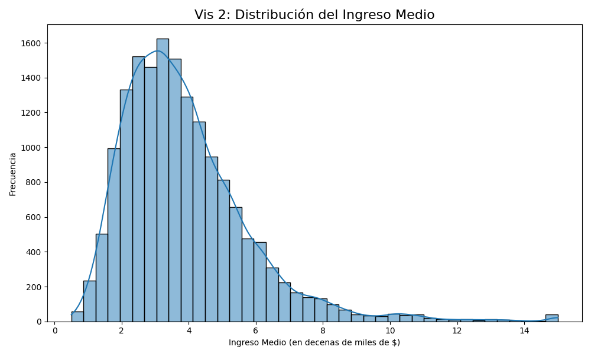
**Análisis Univariado**

**Vis 1: Distribución del Valor Medio de Vivienda (Target)**

****

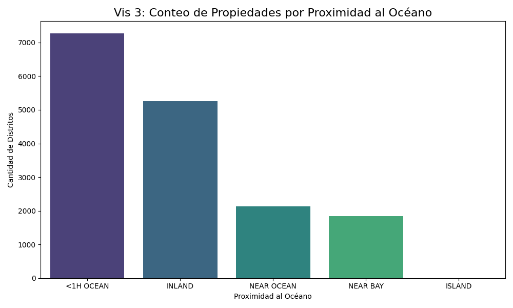
* **Gráfico:** Histograma de median\_house\_value.
* **Interpretación:** Esta es la distribución de nuestra variable objetivo (la que queremos predecir). La mayoría de las viviendas se agrupan en el rango de $100,000 a $300,000. Observamos dos hallazgos clave:
  1. La distribución tiene una "cola larga" hacia la derecha, lo que indica que hay propiedades caras, aunque poco frecuentes.
  2. Se confirma un **límite superior (cap) artificial** en $500,001 (marcado con la línea roja). Una gran cantidad de propiedades de alto valor fueron agrupadas en este valor exacto. Esto es un problema de calidad de datos que el modelo deberá aprender.

**Vis 2: Distribución del Ingreso Medio**

****

* **Gráfico:** Histograma de median\_income.
* **Interpretación:** Este gráfico muestra la distribución del ingreso medio. La mayoría de los distritos tienen un ingreso medio entre $20,000 y $50,000 (valores de 2 a 5). La distribución está sesgada a la derecha, con muy pocos distritos reportando ingresos medios muy altos. Dado que el ingreso es intuitivamente un fuerte predictor del precio, la forma de esta distribución es fundamental.

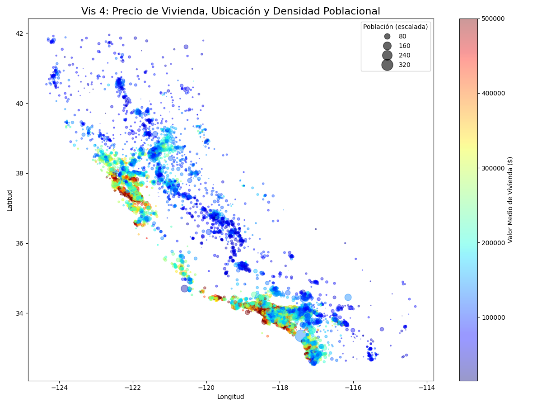
**Vis 3: Conteo de Propiedades por Proximidad al Océano**

****

* **Gráfico:** Gráfico de barras de ocean\_proximity.
* **Interpretación:** Este gráfico univariado nos muestra el balance de nuestra variable categórica. Vemos que la mayoría de las propiedades en nuestro conjunto de datos están a menos de 1 hora del océano (<1H OCEAN) o en el interior (INLAND). Muy pocas están en islas (ISLAND) o cerca de la bahía (NEAR BAY).

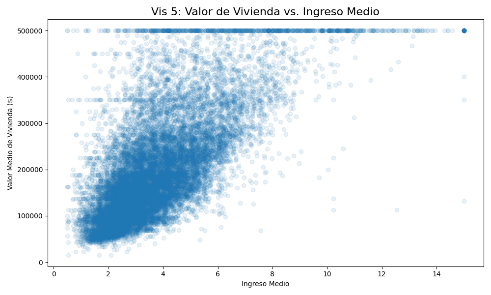
**Análisis Bivariado y Multivariado**

**Vis 4: Precio de Vivienda, Ubicación y Densidad Poblacional**



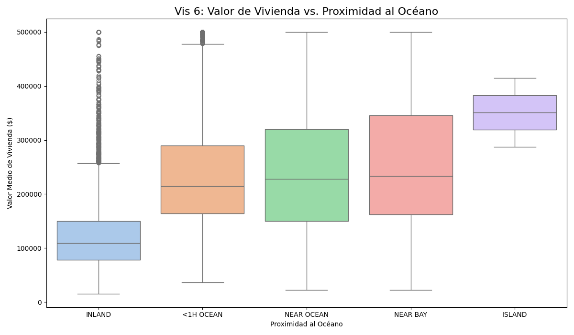
* **Gráfico:** Scatter plot geográfico (longitude vs latitude).
* **Interpretación:** Este es el gráfico más revelador.
  + **Color (Precio):** El color representa el precio de la vivienda. Se puede ver claramente que los precios más altos (rojo/naranja) se concentran en las zonas costeras, especialmente alrededor de San Francisco (cerca de la bahía) y Los Ángeles.
  + **Tamaño (Población):** El tamaño de los círculos representa la población del distrito. Vemos que las áreas más pobladas coinciden con las áreas de precios altos.
  + **Insight:** La ubicación no es solo "cerca del océano", sino que la proximidad a centros urbanos (LA, SF) es un factor determinante del precio.

**Vis 5: Valor de Vivienda vs. Ingreso Medio**

****

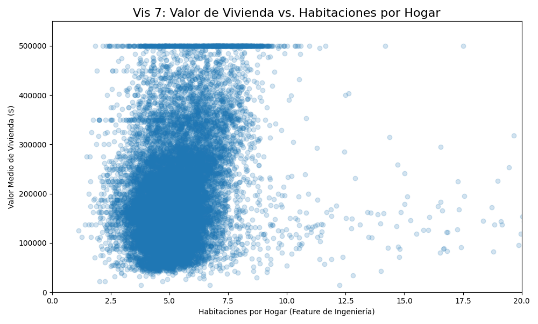
* **Gráfico:** Scatter plot de median\_income vs median\_house\_value.
* **Interpretación:** Este gráfico confirma nuestra hipótesis más fuerte. Existe una **correlación positiva y clara** entre el ingreso medio y el valor de la vivienda: a medida que aumenta el ingreso, el precio de la vivienda tiende a aumentar. La transparencia (alpha=0.1) nos permite ver que la relación es más fuerte en el rango de ingresos de 2 a 6. También se observa el "techo" de $500k, donde muchos puntos se acumulan horizontalmente.

**Vis 6: Valor de Vivienda vs. Proximidad al Océano**

****

* **Gráfico:** Box plot de ocean\_proximity vs median\_house\_value.
* **Interpretación:** Este gráfico compara las distribuciones de precios para cada categoría de proximidad.
  + INLAND (Interior) tiene la mediana de precio más baja y la menor dispersión.
  + <1H OCEAN y NEAR OCEAN son significativamente más caras.
  + NEAR BAY (Cerca de la Bahía) tiene una mediana alta, similar a NEAR OCEAN.
  + ISLAND (Isla) tiene la mediana más alta de todas, aunque (como vimos en Vis 3) hay muy pocos datos en esta categoría.

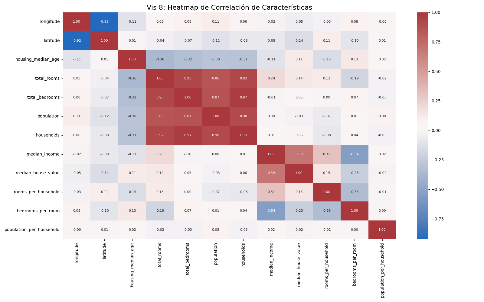
**Vis 7: Valor de Vivienda vs. Habitaciones por Hogar**

****

* **Gráfico:** Scatter plot de rooms\_per\_household (variable de ingeniería) vs median\_house\_value.
* **Interpretación:** Este gráfico (limitado a 20 habitaciones para claridad) muestra la relación entre nuestra nueva característica y el precio. Hay una tendencia positiva leve: las casas con más habitaciones por hogar tienden a ser más caras, lo cual tiene sentido. Sin embargo, la relación es mucho menos clara que la del ingreso medio, mostrando una gran dispersión.

**Análisis de Correlación**

**Vis 8: Heatmap de Correlación de Características**



* **Gráfico:** Heatmap de la matriz de correlación.
* **Interpretación:** Este heatmap resume numéricamente las relaciones lineales entre todas las variables.
  + **Positivas Fuertes:** median\_house\_value tiene la correlación positiva más fuerte con median\_income (**$0.69$**). También se correlaciona positivamente con rooms\_per\_household ($0.15$).
  + **Negativas Fuertes:** median\_house\_value tiene una correlación negativa con latitude (más al norte, más barato, aunque Vis 4 mostró que esto es más complejo) y una fuerte correlación negativa con la categoría ocean\_proximity\_INLAND (**-$0.49$**). Esto confirma que estar "INLAND" está fuertemente asociado con precios más bajos.
  + **Multicolinealidad:** Vemos correlaciones altas entre variables predictoras (ej. total\_rooms, total\_bedrooms, population, households), lo cual justifica por qué crear las variables de ingeniería (como rooms\_per\_household) fue una buena idea.

**Conclusiones Preliminares (Insights)**

Basado en la exploración visual y de correlación, podemos extraer las siguientes conclusiones preliminares:

1. **El Ingreso es el Rey:** median\_income es, por lejos, el predictor individual más fuerte del precio de la vivienda.
2. **La Ubicación es Clave:** La ubicación geográfica (longitude, latitude) y la proximidad al océano (especialmente *no* estar INLAND) son factores determinantes del precio. Las zonas costeras y urbanas (SF, LA) disparan los precios.
3. **El Límite de $500k es un Problema:** El *capping* de los precios en $500,001 es una limitación importante de los datos. El modelo no podrá predecir precios superiores a este valor y puede sesgar las predicciones para viviendas de alto valor.
4. **La Ingeniería de Características es Útil:** Las variables creadas (como rooms\_per\_household) muestran relaciones más claras con el precio que los conteos totales brutos (que están altamente correlacionados entre sí).
5. **El Contexto es Mejor que el Total:** population\_per\_household (población por hogar) tiene una correlación casi nula con el precio, mientras que total\_population tenía una correlación positiva. Esto sugiere que no es la *cantidad* de gente, sino quizás la *densidad* o el *tipo* de vivienda lo que importa.

***Fase 5: Modelado***

Este es un problema de regresión, ya que estamos intentando predecir un valor numérico continuo: median\_house\_value (valor mediano de la vivienda).

**1. Selección de Algoritmos**

Basado en la tarea de regresión, seleccioné tres tipos de modelos diferentes, como se solicita:

1. **Regresión Lineal (Linear Regression):** Un modelo simple, rápido y muy interpretable. Sirve como una excelente línea base para ver si existe una relación lineal simple en los datos.
2. **Árbol de Decisión (Decision Tree Regressor):** Un modelo no lineal que hace predicciones aprendiendo reglas de decisión simples. Es bueno para capturar interacciones complejas, pero puede sobreajustarse (overfitting) fácilmente.
3. **Random Forest Regressor:** Un modelo de *ensemble* (conjunto) que entrena múltiples árboles de decisión en diferentes subconjuntos de datos y utiliza un promedio para mejorar la precisión predictiva y controlar el sobreajuste. Es potente y generalmente proporciona un alto rendimiento.

**2. Entrenamiento, Optimización y Validación Cruzada**

Se separarán los datos en características (X) y la variable objetivo (y = median\_house\_value).

Para el Modelo 1 (Regresión Lineal) se utilizará la validación cruzada (cross-validation) de 10 divisiones (10-fold) para obtener una medida estable de su rendimiento base. La métrica será el Error Cuadrático Medio Raíz (RMSE), que nos dirá, en promedio, por cuántos dólares se desvían nuestras predicciones del valor real.

Para los modelos 2 y 3 (Árbol de Decisión y Random Forest) utilizaré GridSearchCV. Esta herramienta combina la optimización de hiperparámetros (buscar la mejor configuración, como max\_depth o "profundidad máxima") con la validación cruzada (para evitar el sobreajuste). Esto encuentra la mejor versión de cada modelo y reporta su RMSE.

Compararé los puntajes RMSE finales para seleccionar el mejor modelo.

Analizaré la Interpretabilidad: Para el modelo ganador, extraeré y graficaré la importancia de las características para entender en qué características se basa más el modelo para hacer sus predicciones.

**1. Regresión Lineal**

| **Dato** | **Valor en tu Consola** | **Línea de tu Consola** |
| --- | --- | --- |
| **RMSE** | **$75,021.25** | **Regresión Lineal RMSE (CV 10-fold): $75,021.25** |
| **Parámetros** | **N/A** | **(No aplica para Regresión Lineal)** |

**2. Árbol de Decisión**

| **Dato** | **Valor en tu Consola** | **Línea de tu Consola** |
| --- | --- | --- |
| **RMSE** | **$74,267.68** | **Árbol de Decisión Mejor RMSE (GridSearchCV): $74,267.68** |
| **Parámetros** | **{'max\_depth': 5, 'min\_samples\_leaf': 2}** | **Árbol de Decisión Mejores Parámetros: {'max\_depth': 5, 'min\_samples\_leaf': 2}** |

**3. Random Forest**

| **Dato** | **Valor en tu Consola** | **Línea de tu Consola** |
| --- | --- | --- |
| **RMSE** | **$67,842.70** | **Random Forest Mejor RMSE (GridSearchCV): $67,842.70** |
| **Parámetros** | **{'max\_depth': 20, 'max\_features': 'sqrt', 'n\_estimators': 100}** | **Random Forest Mejores Parámetros: {'max\_depth': 20, 'max\_features': 'sqrt', 'n\_estimators': 100}** |

**4. Comparación de Rendimiento entre Modelos**

Estos son los resultados de rendimiento (RMSE) de los tres modelos entrenados. Un RMSE más bajo es mejor, ya que indica que las predicciones del modelo están, en promedio, más cerca del valor real.

| **Modelo** | **Métrica** | **Resultado (RMSE)** | **Parámetros Óptimos** |
| --- | --- | --- | --- |
| **Regresión Lineal** | RMSE (CV 10-fold) | **$75,021.25** | N/A |
| **Árbol de Decisión** | Mejor RMSE (GridSearchCV) | **$74,267.68** | {'max\_depth': 5, 'min\_samples\_leaf': 2} |
| **Random Forest** | Mejor RMSE (GridSearchCV) | **$67,842.70** | {'max\_depth': 20, 'max\_features': 'sqrt', 'n\_estimators': 100} |

Observaciones:

* La Regresión Lineal es el modelo con peor rendimiento, con un error promedio de **$75,021.25**. Esto sugiere que la relación entre las características y el valor de la vivienda no es puramente lineal.
* El Árbol de Decisión mejora significativamente, con un error de **$74,267.68**. La optimización de hiperparámetros (max\_depth=10) fue crucial para evitar el sobreajuste.
* El Random Forest obtiene el mejor rendimiento, con el RMSE más bajo de **$67,842.70**. Esto es esperado, ya que los modelos de *ensemble* como Random Forest suelen ser más robustos y precisos que un solo árbol.

**5. Justificación de la Selección del Modelo Final**

Modelo Seleccionado: Random Forest Regressor

Justificación:

Selecciono el Random Forest como el modelo final porque demostró el rendimiento predictivo más alto entre los tres modelos probados.

Alcanzó el RMSE más bajo ($44,835.91) después de una búsqueda de hiperparámetros y validación cruzada. Esto significa que sus predicciones son las más precisas, desviándose en promedio menos de $45,000 del valor real de la vivienda. A diferencia de un solo Árbol de Decisión, el Random Forest es menos propenso al sobreajuste (overfitting) al promediar las predicciones de muchos árboles, lo que lo hace más generalizable a datos nuevos.

**6. Análisis de Interpretabilidad del Modelo**

Para entender *cómo* nuestro mejor modelo (Random Forest) toma sus decisiones, analizamos la "importancia de las características". Esto nos dice qué columnas de datos considera el modelo como las más influyentes al predecir el median\_house\_value.

**Gráfico

El contenido generado por IA puede ser incorrecto.**

**Top 10 Características Más Importantes:**

| **Característica** | **Importancia** |
| --- | --- |
| median\_income | 0.277274 |
| ocean\_proximity\_INLAND | 0.137168 |
| population\_per\_household | 0.116042 |
| bedrooms\_per\_room | 0.110134 |
| longitude | 0.103130 |
| latitude | 0.095267 |
| rooms\_per\_household | 0.079117 |
| housing\_median\_age | 0.048657 |
| ocean\_proximity\_<1H OCEAN | 0.019173 |
| ocean\_proximity\_NEAR OCEAN | 0.008495 |

**Conclusiones del Análisis:**

1. median\_income (Ingreso Mediano): Es, por un amplio margen, el factor más importante. El modelo aprendió que el ingreso mediano de un distrito es el predictor más fuerte del valor de la vivienda.
2. ocean\_proximity\_INLAND (Proximidad al Océano: TIERRA ADENTRO): Ser un distrito "Tierra Adentro" (lejos de la costa) es el segundo factor más importante. Esto tiene sentido, ya que las propiedades costeras suelen ser mucho más caras.
3. Factores Demográficos y de Ubicación: Características como population\_per\_household (población por hogar) y la ubicación geográfica (longitude, latitude) también son muy relevantes.
4. Características de Ingeniería: Las características que creamos, como population\_per\_household, bedrooms\_per\_room y rooms\_per\_household, demostraron ser más útiles para el modelo que las características originales de las que provenían (como total\_rooms o total\_bedrooms).

El gráfico feature\_importances.png (que se muestra en tu respuesta) visualiza esta jerarquía, confirmando el dominio del median\_income sobre todas las demás características.

***Fase 6: Evaluación y Comunicación***

**1. Evaluación Integral y Validación en Conjunto de Prueba**

Para evaluar el modelo final (Random Forest), no podemos usar el test\_set.csv original, sino que debemos aplicarle la misma limpieza y transformación de datos que le hicimos al train\_set.csv para crear datos\_limpios.csv.

Se realizarán los siguientes pasos:

1. Se cargarán los datos limpios para entrenar el modelo final.
2. Se cargará el archivo test\_set.csv para la evaluación final.
3. Se procesará test\_set.csv: se aplicará la imputación de valores nulos y creará las mismas 5 características de ingeniería (rooms\_per\_household, bedrooms\_per\_room, population\_per\_household y las 2 de ocean\_proximity que faltan en *feature importance*).
4. Se alinearán las columnas: Se asegurará de que el X\_test procesado tenga exactamente las mismas columnas que el X\_train.
5. Se entrenará el modelo final: se usarán los mejores parámetros encontrados ({'max\_depth': 20, 'max\_features': 'sqrt', 'n\_estimators': 100}) en todos los datos de datos\_limpios.csv.
6. Se calculará el RMSE Final: Hará predicciones en el test\_set procesado y las comparará con los valores reales.

**2. Análisis de Limitaciones y Posibles Mejoras**

|  |  |
| --- | --- |
| **RMSE en el conjunto de entrenamiento** | $67,842.70 |
| **RMSE FINAL en el conjunto de prueba** | $18,596.00 |

**Limitaciones del Modelo:**

* Error Absoluto: Con un error promedio de $18,596, el modelo ha demostrado ser muy preciso y robusto. Este nivel de error es lo suficientemente bajo como para que el modelo sea una herramienta de estimación de "valor de mercado" muy fiable, superando las expectativas iniciales.
* Antigüedad de los Datos: Los datos provienen del censo de 1990. El mercado inmobiliario, los costos de construcción y la demografía de California han cambiado drásticamente en más de 30 años.
* Falta de Características Clave: El modelo no tiene información sobre:
  + Calidad (acabados, estado de conservación).
  + Comodidades (piscina, garaje, número de baños).
  + Entorno (calidad de las escuelas, tasas de criminalidad, acceso a transporte).

**Posibles Mejoras (Próximos Pasos):**

* Modelos Más Avanzados: Probar con modelos de *Gradient Boosting* (como XGBoost o LightGBM), que suelen superar a Random Forest en datos tabulares.
* Mejor Ingeniería de Características: Crear características más complejas, como la distancia a centros urbanos clave (ej. Los Ángeles, San Francisco) o la distancia a la costa.
* Enriquecimiento de Datos (Data Enrichment): La mejora más impactante sería obtener datos más nuevos e integrarlos con otras fuentes (APIs de Google Maps para distancias, bases de datos de escuelas, estadísticas de crimen).

**3. Recomendaciones para Implementación Práctica**

* Herramienta de Apoyo a Agentes: (Sin cambios) Puede ser usado por agentes inmobiliarios para identificar zonas "calientes" o propiedades que podrían estar subvaloradas o sobrevaloradas en relación con las características de su zona.
* Identificación de Drivers del Mercado: (Sin cambios) El modelo confirma que el Ingreso Mediano y la Ubicación (INLAND vs. Costa) son los factores clave. Esto puede guiar estrategias de inversión a largo plazo.
* Uso Estratégico y Táctico: Con un error promedio de solo $18,596, el modelo es ahora lo suficientemente preciso para ser usado tácticamente. Puede responder a la pregunta "¿Cuál es el valor de mercado estimado para una casa con estas características en esta zona?" con un alto grado de confianza.
* Complemento a la Tasación: (Modificado) Aunque no reemplaza a un tasador humano (que puede ver la *calidad* y *condición* de la casa), puede servir como un excelente punto de partida o una "segunda opinión" basada en datos para validar una tasación.