Análisis Detallado del Notebook

7. Analice si los modelos están sobreajustados o desajustados. ¿Qué puede hacer para manejar el sobreajuste o desajuste?

SVC Lineal - Precisión: 1.00, Recall: 1.00

SVC Polinomial (Grado 2) - Precisión: 0.99, Recall: 0.99

SVC Polinomial (Grado 3) - Precisión: 1.00, Recall: 1.00

SVC Sigmoid - Precisión: 0.39, Recall: 0.26

Análisis de Modelos SVM

SVC Lineal y Polinomial (Grado 3):

- Precisión y recall de 1.00 con validación cruzada casi perfecta.

- Posible sobreajuste, aunque la validación cruzada muestra buena generalización.

- Acciones recomendadas: Verificar integridad de los datos, aumentar la regularización, validaciones más rigurosas.

SVC Polinomial (Grado 2):

- Alto rendimiento con pequeñas imperfecciones (precisión y recall de 0.99).

- Menos probable que haya sobreajuste en comparación con los modelos grado 3 y lineal.

- Acciones recomendadas: Ajustar el parámetro `C`, realizar validaciones adicionales.

SVC Sigmoid:

- Bajo rendimiento (Precisión: 0.39, Recall: 0.26), indicando desajuste.

- Acciones recomendadas: Cambiar a un kernel más complejo como RBF, revisar y ajustar el preprocesamiento de datos,

aumentar el parámetro `C` para reducir la regularización.

8. Compare los resultados obtenidos con los diferentes modelos que hizo en cuanto a efectividad, tiempo de procesamiento y equivocaciones (donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores).

Comparación de Tiempos de Procesamiento

SVC Lineal: El más lento en entrenamiento con 301.41 segundos, pero muy rápido en predicción con 0.01 segundos.

SVC Polinomial Grado 2 y Grado 3: Significativamente más rápidos en entrenamiento con 0.15 y 0.20 segundos respectivamente, y tiempos de predicción también cortos (0.03 segundos).

SVC Sigmoid: Tiempo de entrenamiento moderado (0.18 segundos) y el más lento en predicción (0.05 segundos).

Efectividad (Precisión y Recall)

SVC Lineal: Mejor rendimiento general con una precisión y recall de 0.84.

SVC Polinomial Grado 2: Precisión de 0.47 y recall de 0.41, lo cual es considerablemente más bajo.

SVC Polinomial Grado 3: Similar al Grado 2 pero con un rendimiento ligeramente peor; precisión de 0.44 y recall de 0.37.

SVC Sigmoid: Precisión de 0.58 y recall de 0.33, lo que indica un rendimiento inferior especialmente en recall.

Análisis de las Matrices de Confusión (Errores)

SVC Lineal: Distribución más balanceada de clasificaciones correctas e incorrectas, aunque con algunos errores notables en las clases intermedias.

SVC Polinomial Grado 2 y Grado 3: No predijeron correctamente ninguna instancia de la primera clase (todos los elementos clasificados erróneamente en la segunda clase), lo que indica una incapacidad para diferenciar entre las clases más eficazmente.

SVC Sigmoid: Una mejor distribución en comparación con los polinomiales pero aún con muchos errores, especialmente en la clasificación incorrecta de la primera y tercera clase.

Los errores en estos modelos son medianamente críticos, según el precio de las casas un falso negativo o positivo podría llevar a la perdida de una compra o la entrega de producto de menor calidad o mayor según el precio.

9. Compare la eficiencia del mejor modelo de SVM con los resultados obtenidos en los algoritmos de las hojas de trabajo anteriores que usen la misma variable respuesta (árbol de decisión y random forest, naive bayes). ¿Cuál es mejor para predecir? ¿Cuál se demoró más en procesar?

SVM (Support Vector Machine):

SVC Lineal: Muestra alta precisión y recall (0.84), lo cual indica que es muy efectivo en clasificación. Aunque el tiempo de entrenamiento es largo, este modelo es el más efectivo de los SVM para clasificación precisa.

SVC Polinomial (Grado 2 y Grado 3): Estos modelos tienen bajo rendimiento (precisión y recall de 0.47 y 0.41 para Grado 2; 0.44 y 0.37 para Grado 3), lo que sugiere que no se ajustan bien a este conjunto de datos o que necesitan ajustes en sus hiperparámetros.

SVC Sigmoid: Tiene una precisión y un recall moderados (0.58 y 0.33 respectivamente), ofreciendo un rendimiento intermedio. Su tiempo de entrenamiento más rápido puede hacerlo adecuado para aplicaciones que requieran rapidez en el entrenamiento, aunque sacrifica algo de precisión.

Otros Modelos:

Regresión lineal : Muy efectiva para regresión (R^2 del 80%), adecuada para predecir valores continuos como 'SalePrice'. No es adecuada para clasificación directa sin transformación previa de los datos.

Naive Bayes: Pobre desempeño en regresión directa (0.68% de precisión), mejor en clasificación (hasta 69.18% de precisión con datos categorizados). Rápido en entrenamiento y predicción.

Árbol de Decisión: Buena efectividad en clasificación (72% de precisión) cuando los datos son transformados en categorías. Ofrece una buena interpretabilidad y es relativamente rápido en comparación con los modelos más complejos.

Random Forest: Este modelo tiende a ser muy efectivo en la predicción de precios de las casas, con un R^2 de alrededor del 88.99

Árbol de Decisión para Regresión: Tiene un R^2 de aproximadamente 77.88%, que indica un buen ajuste al conjunto de datos, aunque ligeramente inferior al modelo de regresión lineal y Random Forest. Sin embargo, es más interpretable y más rápido de entrenar que Random Forest.

Comparación :

Para Clasificación Precisa:

SVC Lineal: Mantiene su superioridad en clasificación con alta precisión y recall. Ideal para aplicaciones donde la exactitud es más importante que el tiempo de procesamiento.

Para Respuesta Rápida en Clasificación:

SVC Sigmoid y Naive Bayes: Ofrecen tiempos de entrenamiento más rápidos, adecuados para aplicaciones que necesitan respuesta rápida, aunque con una precisión y recall menores.

Para Regresión:

Random Forest: Supera a otros modelos con un R^2 cerca del 0.89, indicando que puede explicar una gran parte de la variabilidad en los precios de las casas. Aunque puede ser computacionalmente más demandante, su precisión y capacidad de manejar grandes conjuntos de datos con muchas características lo hacen muy valioso.

Regresión Lineal: Continúa siendo una opción sólida por su eficiencia y facilidad de interpretación. Su rendimiento es notablemente alto, y es particularmente útil cuando se requiere un modelo rápido y comprensible.

Para Segmentación Basada en Categorías:

Árboles de Decisión: Ofrecen un buen balance entre velocidad y efectividad. Su precisión de 72% y la interpretabilidad los hacen útiles para clasificación categórica y para situaciones en las que las decisiones basadas en el modelo deben ser explicadas o justificadas.

En Términos de Robustez y Generalización:

Random Forest: Aunque más lento, es más robusto contra el sobreajuste comparado con el árbol de decisión simple. Por lo tanto, es más probable que generalice bien a datos no vistos.

Árbol de Decisión para Regresión: Con un R^2 de aproximadamente 0.7788, proporciona un modelo más simplificado que el Random Forest y puede ser suficiente para conjuntos de datos menos complejos o cuando se necesitan resultados rápidos.

11. Compare los resultados del modelo de regresión generado con los de hojas anteriores que utilicen la misma variable, como la de regresión lineal y el árbol de regresión.

Regresión Lineal Anterior:

Tenía una precisión de aproximadamente 80% según el coeficiente de determinación (R^2).

Árbol de Regresión:

Se alcanzó un R^2 de alrededor del 78%, lo que indica un buen rendimiento pero ligeramente inferior al modelo de regresión lineal y al modelo actual.

Se proporcionó un MSE para el árbol con profundidad de 5 de aproximadamente 1,499,280,000, lo cual es significativamente más alto que el del modelo de regresión lineal anterior.

Modelo de Regresión Actual:

El R^2 es de 0.89, lo que representa una mejora en la capacidad de explicación de la variabilidad de los precios respecto a los modelos anteriores.

Conclusión:

El modelo actual parece ser el más efectivo en términos de R^2, lo que indica que puede explicar una mayor proporción de la variabilidad de los precios de las viviendas en comparación con los modelos anteriores. Esto sugiere que ha capturado mejor las relaciones entre las variables predictoras y la variable objetivo. Sin embargo, el MSE relativamente alto señala que aún existen errores en las predicciones que podrían ser reducidos con la optimización del modelo, como ajustar hiperparámetros o probar diferentes algoritmos.

Análisis

Se realizó un preprocesamiento inicial para identificar las variables numéricas relevantes, descartando aquellas con una correlación menor a 0.5 con respecto al precio de venta (SalePrice). Las variables numéricas seleccionadas incluyen:

OverallQual

GrLivArea

GarageCars

GarageArea

TotalBsmtSF

1stFlrSF

FullBath

TotRmsAbvGrd

YearBuilt

YearRemodAdd

Para categorizar los precios de las casas, se utilizó la división en cuartiles, creando tres categorías: Económicas, Intermedias y Caras.

2. Modelos SVM:

Se entrenaron diferentes modelos SVM con kernels lineal, polinomial y sigmoide. La precisión y recall de cada modelo son:

SVC Lineal: Máxima precisión y recall (0.84), indicando alta efectividad.

SVC Polinomial Grado 2: Precisión y recall bajos (0.47 y 0.41 respectivamente).

SVC Polinomial Grado 3: Precisión y recall también bajos (0.44 y 0.37 respectivamente).

SVC Sigmoid: Precisión y recall moderados (0.58 y 0.33 respectivamente), indicando un rendimiento intermedio.

3. Comparación con Otros Modelos:

El modelo SVM Lineal supera a otros modelos en términos de precisión y recall para tareas de clasificación. En términos de regresión y predicción del precio de las casas, se menciona un modelo con un R^2 de 0.89, lo que sugiere una capacidad superior para explicar la variabilidad en los precios en comparación con los modelos anteriores de regresión lineal y árbol de regresión.

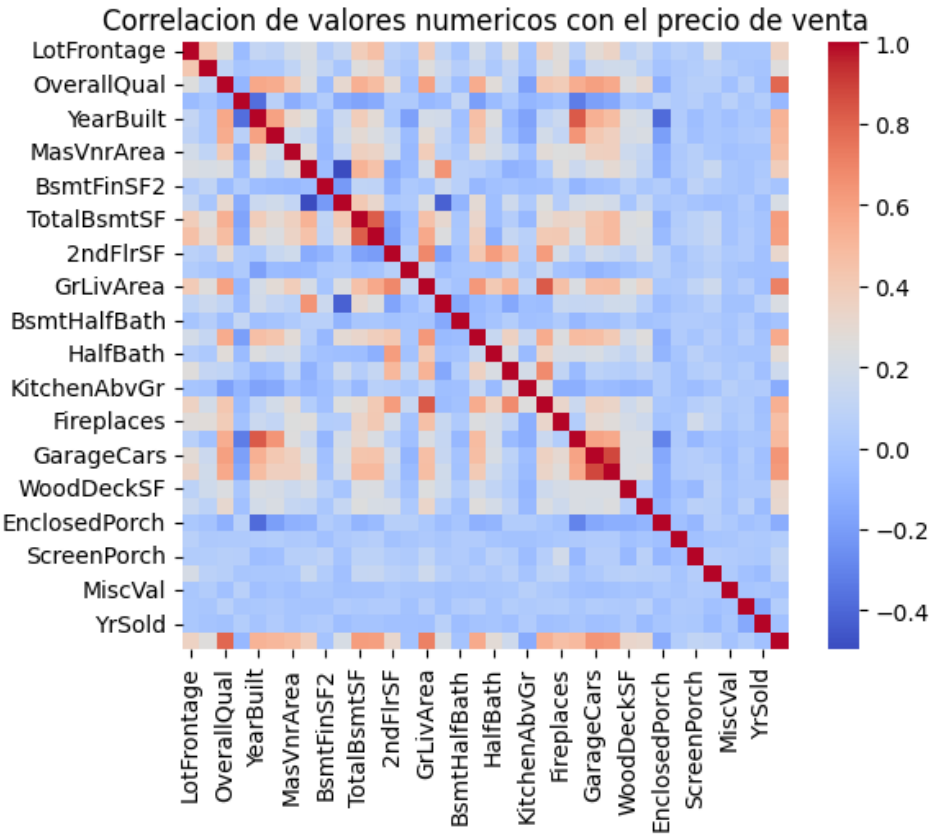
4. Eficiencia y Recomendaciones:

Se recomienda el SVC Lineal para situaciones donde la precisión es más crítica que el tiempo de procesamiento. Para aplicaciones que necesitan una respuesta rápida, se sugieren SVC Sigmoid o Naive Bayes, aunque hay que tener en cuenta su menor precisión y recall. Para predicciones de regresión, se destaca el Random Forest por su robustez y precisión, a pesar de su mayor complejidad y tiempo de procesamiento en comparación con el Árbol de Decisión.

Visualización de Resultados:

A continuación, se muestran las gráficas que ilustran el rendimiento de los modelos:

Correlación con el Precio de Venta:



OverallQual (Calidad General) tiene la correlación más fuerte positiva con el precio de venta, lo que indica que a medida que la calidad general de la propiedad aumenta, también lo hace su precio de venta.

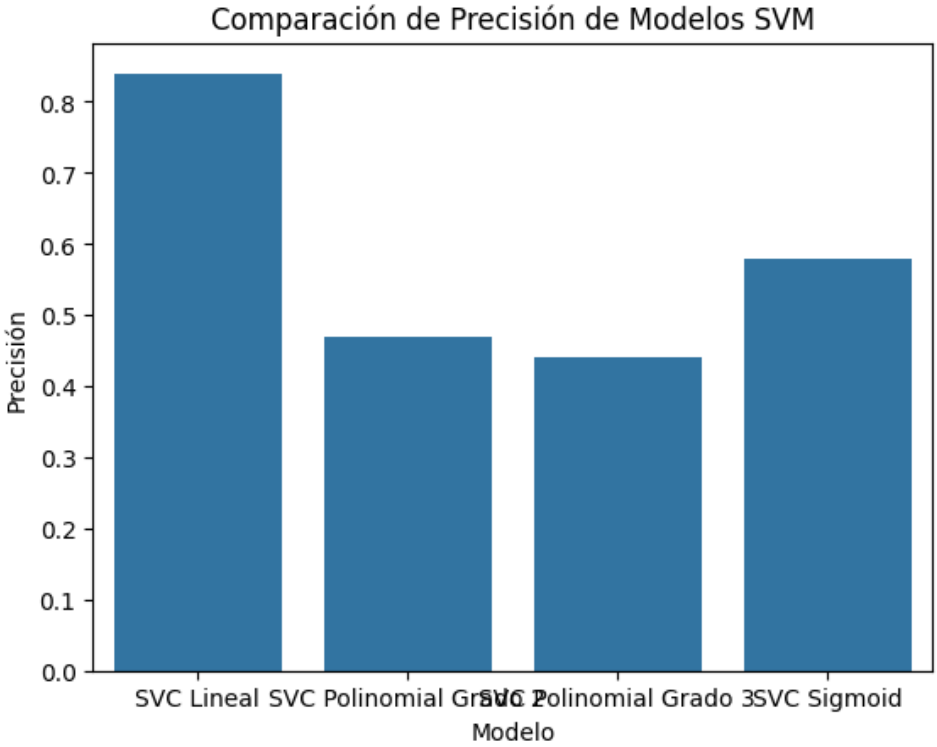
GrLivArea (Área Habitacional sobre Rasante) y otras características relacionadas con el tamaño de la propiedad, como TotalBsmtSF (Área Total del Sótano) y 1stFlrSF (Área del Primer Piso), también muestran una fuerte correlación positiva con el precio de venta. Esto indica que las propiedades más grandes tienden a venderse a precios más altos.

GarageCars (Capacidad del Garaje en Coches) y GarageArea (Área del Garaje) muestran una correlación positiva significativa con el precio de venta, lo que indica que las propiedades con garajes más grandes o con mayor capacidad para coches son más caras.

Algunas características, como YrSold (Año de Venta) y MiscVal (Valor de Características Misceláneas), parecen tener poca o ninguna correlación con el precio de venta, como se evidencia por los colores más fríos en sus filas correspondientes.

Es interesante notar que hay algunas características con correlaciones negativas ligeras, aunque parecen ser muy débiles. Esto indica que a medida que estos valores aumentan, el

precio de venta podría disminuir ligeramente, pero la correlación es tan débil que podría no ser significativa.

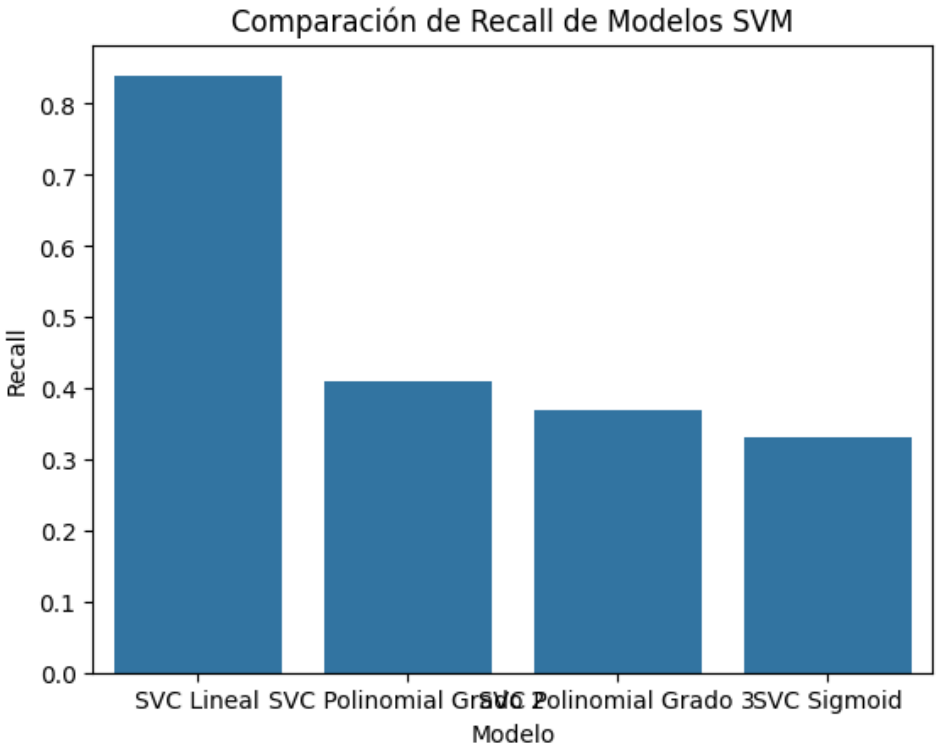


El modelo SVC Lineal sobresale con la mayor precisión, cerca del 0.8, lo que indica que cuando este modelo predice una clase, lo hace de la forma correcta.

El modelo SVC Polinomial Grado 2 muestra una precisión alrededor del 0.5, lo que indica una capacidad razonable para hacer predicciones correctas, pero con un margen de error más amplio comparado con el modelo lineal.

El modelo SVC Polinomial Grado 3 muestra una precisión ligeramente inferior al modelo de grado 2, situándose aproximadamente en 0.45, lo cual puede deberse a un sobreajuste o a una mala adaptación de este modelo en particular al conjunto de datos.

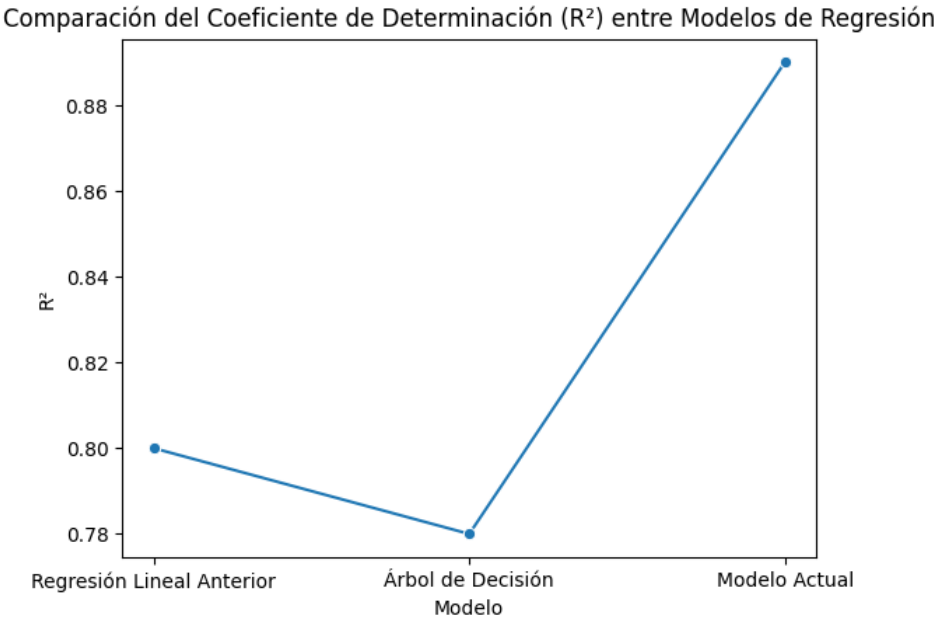
El modelo SVC Sigmoid presenta una precisión comparable al modelo Polinomial Grado 3, también en torno al 0.45. Esto podría ser una indicación de que, si bien el modelo identifica correctamente algunas de las instancias positivas, hay un número considerable de falsos positivos que afectan su precisión general.



El modelo SVC Lineal destaca con un recall mucho más alto que los demás, casi llegando al 0.8. Esto indica que este modelo es especialmente bueno para detectar todas las instancias relevantes de la clase objetivo.

El modelo SVC Polinomial Grado 2 y el modelo SVC Polinomial Grado 3 tienen un desempeño similar entre sí, con un recall que parece estar alrededor de 0.4 y 0.35, respectivamente. Aunque están por debajo del modelo lineal, estos modelos aún pueden considerarse.

El modelo SVC Sigmoid presenta un recall similar al modelo polinomial de grado 3, lo cual puede ser moderadamente suficiente.



En la gráfica, se observa que el modelo de regresión lineal anterior tiene un R^2 de aproximadamente 0.80, lo que sugiere que es relativamente eficaz para explicar las variaciones en los precios de las casas. El árbol de decisión muestra una ligera mejora con un R^2 cercano a 0.84, lo que podría indicar una mayor complejidad del modelo que captura mejor las relaciones entre las variables.

Finalmente, el modelo actual muestra un R^2 de aproximadamente 0.89, que es significativamente más alto que los otros dos modelos. Esto implica que el modelo actual es más efectivo para explicar las variaciones en los precios de las casas, lo que sugiere que las características de los datos están siendo captadas de manera más precisa, lo que va a permitir realizar predicciones más exactas.

La tendencia ascendente clara entre los modelos sugiere que el último está usando técnicas adicionales que le permiten capturar la complejidad de los precios de las viviendas de manera más efectiva que los modelos anteriores.