Universidad del Valle de Guatemala

Facultad de Ingeniería

Mineria de Datos

**INFORME**

Gabriel Paz 221087

Diedrich Solis

27 de abril del 2025, Guatemala de la Asunción

**Informe**

**Resumen Ejecutivo**

Este informe presenta un análisis exhaustivo de modelos de clasificación y regresión aplicados al conjunto de datos “House Prices: Advanced Regression Techniques” de Kaggle.

1. **Clasificación:** Se creó la variable categórica PrecioCategoria con tres niveles (Barato, Medio, Caro) y se compararon SVM (kernels lineal, RBF, polinomial) frente a Random Forest, Regresión Logística, KNN, Árbol de Decisión y Naive Bayes.
2. **Regresión:** Se evaluó KNN Regressor, un método de Naive Bayes adaptado a regresión y un SVR afinado, comparándolos además con Regresión Lineal y Bayesian Ridge.
3. **Hallazgos clave:**
   * **Clasificación:** Random Forest mostró la mayor exactitud (82.9 %), mientras que SVM lineal equilibró precisión (82.4 %) y velocidad.
   * **Regresión:** Bayesian Ridge y Regresión Lineal lideraron en R² (~0.82) y menor MSE; el SVR lineal afinado obtuvo R² = 0.586.
4. Se discuten implicaciones prácticas y recomendaciones de uso según la velocidad de cómputo, la precisión deseada y el impacto de errores en cada categoría.

**Introducción**

InmoValor S.A., consultora en valoración inmobiliaria, requiere modelos predictivos para categorizar propiedades (baratas, medias, caras) y estimar precios con precisión. Este proyecto, dividido en seis entregas, aborda:

* Creación y codificación de variables objetivo.
* Desarrollo de pipelines reproducibles de preprocesamiento.
* Entrenamiento y comparación de múltiples algoritmos de clasificación y regresión.
* Análisis de resultados, errores críticos y recomendaciones de mejoras.

**Conjunto de Datos**

* **Origen:** Kaggle “House Prices: Advanced Regression Techniques”.
* **Dimensiones:** 1 460 registros × 81 atributos.
* **Tipos de variables:**
  + Numéricas: áreas, años, conteos de habitaciones, precio de venta (SalePrice).
  + Categóricas: vecindario, tipo de calle, calidad de acabados, etc.
* **Valores faltantes:** Distribuidos en variables como PoolQC, Alley, FireplaceQu.
* **Preprocesamiento inicial:**
  + Imputación de nulos (mediana para numéricas).
  + Eliminación de variables con excesivos nulos o irrelevancia.
  + Selección de atributos numéricos para regresión directa y codificación para clasificación.

**Metodología**

**Preprocesamiento**

1. **Imputación:** Nulos rellenos con mediana (numéricas) o moda (categóricas).
2. **Codificación:**
   * Para SVM y regresores numéricos, se mantuvieron solo columnas numéricas.
   * Para modelos con variables categóricas, se aplicó One-Hot Encoding.
3. **Escalado:** StandardScaler sobre características numéricas.

**División Train/Test**

* **Clasificación:** 70 % entrenamiento, 30 % prueba, semilla fija 221087.
* **Regresión:** 70 % / 30 % con la misma semilla para asegurar comparabilidad entre modelos.

**Clasificación de Precio**

**Creación de la variable PrecioCategoria**

* Se calcularon los percentiles 33 % y 66 % de SalePrice.
* Se asignaron etiquetas:
  + **Barato:** ≤ percentil 33
  + **Medio:** entre 33 y 66
  + **Caro:** > percentil 66

**Modelos SVM y ajuste de hiperparámetros**

* **Algoritmos:** SVM lineal, RBF y polinomial.
* **GridSearchCV:**
  + Lineal: C ∈ {0.1,1,10} → C=0.1
  + RBF: C ∈ {0.1,1,10}, γ ∈ {0.01,0.1,1} → C=1, γ=0.01
  + Polinomial: C ∈ {0.1,1,10}, degree ∈ {2,3,4} → C=10, degree=2

**Comparación interna de SVM**



**Interpretación:**

* Lineal maximiza accuracy/recall.
* Polinomial elevada precision y F1.
* Todas confunden principalmente la categoría “Medio”.

**Comparación con otros clasificadores**



**Conclusión de clasificación:**

* Random Forest lidera en precisión.
* SVM lineal y Regresión Logística ofrecen mejor velocidad manteniendo > 82 % de accuracy.
* Naive Bayes demasiado impreciso para esta tarea.

**Regresión de Precio Continuo**

**KNN Regressor**

* **MAE (test):** 28 153
* **RMSE (test):** 47 681
* **R² (test):** 0.7036

**Naive Bayes “Regresión” por bins**

* **MAE (test):** 77 942
* **RMSE (test):** 98 117
* **R² (test):** –0.2551

**SVR Afinado**

* **Modelo:** SVR(kernel=linear, C=10, ε=0.5)
* **MSE:** 2.89 × 10⁹ → RMSE ≈ 53 767
* **R²:** 0.5856
* **Tiempo CV:** 19 s

**Comparación con Regresores Lineales y Bayesiana**



**Interpretación de regresión:**

* Bayesiana y Lineal explican ~82 % de la varianza con menor error.
* SVR lineal queda rezagado en precisión y velocidad.
* KNN equilibra bien generalización (R² ~0.70) pero no supera a los lineales.

**Conclusiones y Recomendaciones**

1. **Modelos de clasificación:**
   * Para máxima accuracy: **Random Forest**.
   * Para equilibrio precisión/latencia: **SVM Lineal** o **Regresión Logística**.
2. **Modelos de regresión:**
   * **Regresión Bayesiana** y **Regresión Lineal** lideran en explicación de varianza y eficiencia.
   * El **SVR lineal** afinado no supera a los regresores lineales más sencillos.
3. **Mejoras Futuras:**
   * Ingeniería de variables: crear interacciones, transformaciones polinomiales de atributos clave.
   * Ensambles avanzados: **XGBoost**, **LightGBM** para capturar no linealidades.
   * Evaluar métricas de negocio: MAE, median absolute error, intervalos de confianza.

**Referencias**

* Kaggle. House Prices: Advanced Regression Techniques.
* Scikit-Learn: documentación oficial de SVM, SVR, Random Forest, KNN, regresión lineal, BayesianRidge.
* Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.