Análisis Supervisado y No Supervisado del Boston Housing mediante GMM y XGBoost

Marco Rivera [Universidad Autónoma de Nuevo León | *marco.riveracrz@uanl.edu.mx*]

Resumen.

En este trabajo se analizan los datos de **Boston Housing** mediante dos enfoques de aprendizaje: (1) un modelo de **agrupamiento no supervisado** basado en *Gaussian Mixture Models* (GMM), y (2) un modelo de **aprendizaje supervisado** utilizando *XGBoost* para predecir el valor mediano de las viviendas. El modelo GMM permite identificar estructuras en los datos, mientras que XGBoost ofrece una predicción precisa del valor de las viviendas, mostrando menor error.

Palabras clave: Aprendizaje supervisado, Aprendizaje no supervisado, GMM, XGBoost, Boston Housing
Keywords: Supervised learning, Unsupervised learning, GMM, XGBoost, Boston Housing

1 Introducción

El análisis de datos mediante técnicas de **aprendizaje automático** puede abordarse desde dos enfoques principales: el **aprendizaje no supervisado**, que busca descubrir patrones sin etiquetas, y el **aprendizaje supervisado**, cuyo objetivo es predecir una variable a partir de otras observadas.

En el primer caso, el análisis de agrupamiento permite identificar estructuras en los datos (*unsupervised learning*). Entre los métodos más utilizados se encuentran k-means, DBSCAN y los **Modelos de Mezcla Gaussiana** (GMM), los cuales modelan explícitamente la distribución de cada grupo y asignan pertenencia probabilística a las observaciones. En este trabajo se aplican GMM a los datos de *Boston Housing*, determinando el número óptimo de grupos mediante el criterio **BIC**.

En el segundo enfoque, de tipo **supervisado**, se busca predecir el **valor mediano de las viviendas** a partir de variables socioeconómicas y ambientales. Para ello, se emplea el algoritmo **XGBoost**, que combina múltiples árboles de decisión para minimizar el error de predicción mediante optimización de gradiente. Se evalluó el desempeño de este mediante métricas como **RMSE** y $\mathbf{R^2}$.

De esta forma, el presente estudio integra ambos enfoques —supervisado y no supervisado— para analizar el conjunto de datos *Boston Housing*, permitiendo tanto la identificación de patrones en los datos como la predicción precisa del valor mediano de las viviendas.

2 Descripción de los datos

La base de datos de Boston, disponible en R, contiene 506 observaciones y 14 variables socioeconómicas y ambientales, entre ellas:

- `rm`: número promedio de cuartos por vivienda.
- `lstat`: porcentaje de población de bajo estatus.
- `medv`: valor medio de las viviendas (variable respuesta en contextos de regresión).

En este estudio se seleccionaron `rm` y `lstat` como variables de entrada para el agrupamiento, dado que muestran contrastes claros en la literatura y permiten visualización bidimensional.

3 Antecedentes

Li, Dong y Hua (2007) propusieron un modelo de mezcla gaussiana en subespacios con *feature saliency* local, seleccionando el número de grupos mediante **MML** Li *et al*. [2007]. Este enfoque fue aplicado específicamente al conjunto de datos *Boston Housing*, y permite identificar tanto la cantidad óptima de grupos como las variables más relevantes por grupo. Fang et al. (2021) discuten el papel de MML como criterio de selección de modelos bayesiano, cercano en espíritu al BIC Fang *et al*. [2021].

Por otro lado, en el ámbito del **aprendizaje supervisado**, Ding (2024) comparó distintos modelos de regresión —entre ellos el **Multiple Linear Regression (MLR)**, **Random Forest (RF)** y **Extreme Gradient Boosting (XGBoost)**— para predecir el valor mediano de las viviendas en el conjunto *Boston Housing*. El estudio concluyó que el modelo **XGBoost** presentó el mejor desempeño al obtener el menor **RMSE**, lo que evidencia su capacidad para capturar relaciones no lineales y reducir el error de predicción en comparación con los otros modelos. Estos resultados respaldan su selección en el presente trabajo como el algoritmo principal para el enfoque supervisado Ding [2024].

4 Metodología

El procedimiento seguido se resume en los siguientes pasos:

1. **Selección de variables.** De las 14 variables disponibles en el conjunto de datos *Boston Housing*, se eligieron `rm` (número promedio de cuartos) y `lstat` (porcentaje de población de bajo estatus). Esta decisión se tomó debido a que ambas variables presentan una relación contrastante y facilitan la visualización bidimensional de los grupos.
2. **Preprocesamiento.** Las variables fueron **estandarizadas** con `StandardScaler`, de modo que cada una tuviera media cero y desviación estándar uno.
3. **Modelado con GMM.** Se entrenaron modelos de mezcla gaussiana (GMM) con diferentes números de componentes ($K = 1, \ldots, 9$), utilizando la implementación de `scikit-learn`. El algoritmo estima los parámetros mediante *Expectation-Maximization* (EM), asignando

pertenencia probabilística de cada observación a los grupos.

El modelo de mezcla gaussiana busca maximizar la log-verosimilitud de los datos, dada por:

$$\max_{\alpha,\mu,\Sigma} \sum_{j=1}^{n} \ln \left(\sum_{i=1}^{K} \alpha_i \frac{1}{2\pi^{n/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x_j-\mu_i)^T \Sigma_i^{-1}(x_j-\mu_i)} \right)$$

donde μ_i y Σ_i son los parámetros de la i-ésima componente gaussiana, α_i son los coeficientes de mezcla con $\sum_{i=1}^{k} \alpha_i = 1$, y d es la dimensión de los datos Zhang [2017].

4. **Selección del número de grupos.** Para determinar K se calculó el **Bayesian Information Criterion (BIC)** para cada modelo:

$$\text{BIC} = -2 \log \hat{L} + k \log n,$$

donde $\hat{L}$ es la verosimilitud máxima del modelo, k es el número de parámetros y n el tamaño de la muestra. Se eligió el valor de K que minimiza el BIC, en este caso $K = 3$. Este criterio se considera una aproximación práctica al **Minimum Message Length** (MML).

5. **Asignación y visualización.** Con $K = 3$, se ajustó el modelo definitivo y se asignó a cada observación su grupo correspondiente. Finalmente, se realizó una representación gráfica de los datos en el plano coloreando los puntos según el grupo obtenido.
6. **Modelado matemático del XGBoost.** El algoritmo **XGBoost** se basa en el método de *gradient boosting*, donde el modelo se construye de manera aditiva combinando múltiples árboles de decisión.

   En cada iteración t, la predicción se actualiza mediante:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i),$$

donde $f_t(x_i)$ representa el nuevo árbol ajustado sobre los errores residuales de la iteración anterior.

El modelo minimiza una función de pérdida regularizada:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \Omega(f_t),$$

donde $l(\cdot)$ mide el error de predicción y $\Omega(f_t)$ penaliza la complejidad del modelo para evitar el sobreajuste Lai *et al*. [2021].

7. **División del conjunto de datos.** Los datos se separaron en dos subconjuntos: un **80 % para entrenamiento** y un **20 % para prueba**. La variable dependiente fue `medv`, mientras que las trece variables restantes se emplearon como regresoras.
8. **Evaluación del modelo.** El desempeño del modelo se midió utilizando dos métricas principales: el **Coeficiente de Determinación** (R^2) y la **Raíz del Error Cuadrático Medio** (**RMSE**). El primero evalúa la proporción de la variabilidad de la variable dependiente explicada por el modelo, mientras que el segundo cuantifica la magnitud promedio del error de predicción.

   El **Coeficiente de Determinación ajustado** se define como:

$$R^2 = 1 - \frac{n-1}{n-k-1}(1-\bar{R}^2),$$

donde n representa el número de observaciones, k el número de variables regresoras y $\bar{R}^2$ el coeficiente de determinación no ajustado. Este ajuste corrige el incremento artificial de R^2 cuando se añaden nuevas variables al modelo.

Por su parte, el **RMSE** se calcula mediante:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}},$$

donde y_i son los valores observados, $\hat{y}_i$ los valores predichos y n el número total de observaciones. Un menor valor de RMSE indica un mejor ajuste del modelo Ding [2024].

9. **Diseño de experimentos para XGBoost.** El ajuste de hiperparámetros de XGBoost se planteó como un diseño discreto donde los factores son hiperparámetros del modelo y los niveles son los valores explorados en la grilla. Cada tratamiento corresponde a una combinación específica de niveles, evaluada mediante validación cruzada aleatoria (*ShuffleSplit*).

Tabla 1. Factores y niveles en el Diseño de Experimentos de XGBoost.

Factor	Niveles
`n_estimators`	{400, 700}
`max_depth`	{3, 5}

Se evaluaron cuatro combinaciones: $\{(400,3),(400,5),(700,3),(700,5)\}$. Se usó ShuffleSplit ($n_{\text{splits}} = 50$, `test_size`=0.2) con MAE promedio en CV; el mejor tratamiento se reentrenó y evaluó en el conjunto de prueba (20 %) reportando RMSE y R^2.

5 Resultados

La Figura 1 muestra los valores del **BIC** obtenidos para diferentes números de grupos ($K = 1, \ldots, 9$). Se observa un decrecimiento inicial seguido de un mínimo en $K = 3$, tras el cual el criterio comienza a aumentar nuevamente. De esta manera, se determinó que el número óptimo de grupos es $K = 3$.

Una vez seleccionado $K = 3$, se ajustó el modelo definitivo de mezcla gaussiana. La Figura 2 presenta la partición de los datos en el plano (`rm`, `lstat`), donde cada color corresponde a un grupo. Se observa que los grupos se diferencian principalmente por el **número de cuartos promedio** y el **porcentaje de población de bajo estatus**.

Además del análisis no supervisado, se implementó un modelo de **aprendizaje supervisado** utilizando el algoritmo **XGBoost** (incorporando el GridSearch) para predecir el valor mediano de las viviendas. El modelo se evaluó mediante dos métricas de desempeño: la **Raíz del Error Cuadrático Medio (RMSE)** y el **Coeficiente de Determinación** (R^2). Los resultados obtenidos fueron los siguientes:

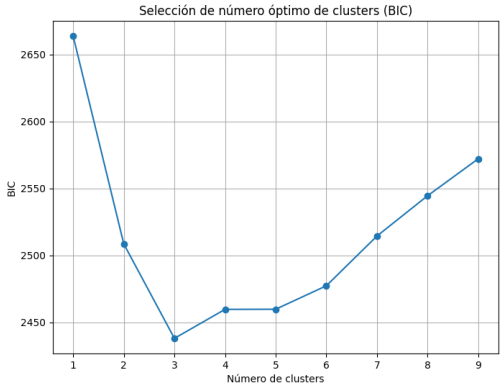

Figura 1. Selección del número óptimo de grupos mediante BIC. El mínimo se alcanza en $K = 3$.

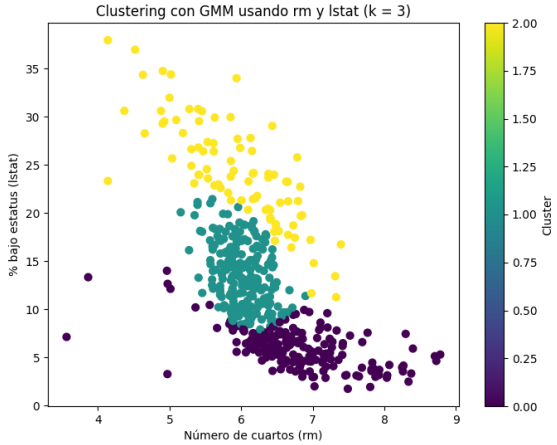

Figura 2. Agrupamiento con GMM en el plano (`rm`, `lstat`) con $K = 3$.

Estos valores indican un buen desempeño predictivo; el modelo logra errores bajos (RMSE) y valores de R^2 entre 0.78 y 0.89 según el clúster.

Las figuras 3, 4 y 5 muestran la comparación entre los valores reales y los valores predichos por el modelo. La cercanía de los puntos a la línea diagonal sugiere que el modelo logra capturar adecuadamente la relación entre las variables predictoras y la variable dependiente.

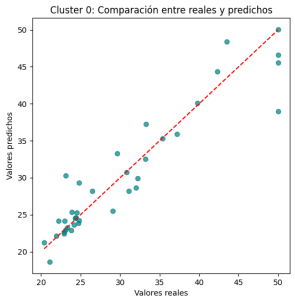

Figura 3. Comparación entre valores reales y predichos por el modelo XGBoost del grupo 0.

6 Conclusiones y discusión

El presente estudio integró enfoques de **aprendizaje no supervisado** y **supervisado** aplicados al conjunto de datos

Tabla 2. Métricas de desempeño del modelo XGBoost

Métrica	Grupo 0	Grupo 1	Grupo 2
R^2	0.89	0.78	0.82
RMSE	3.05	2.75	1.98

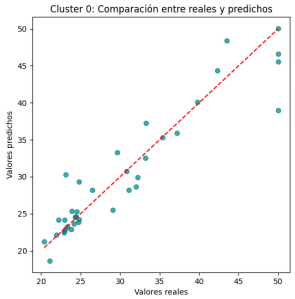

Figura 4. Comparación entre valores reales y predichos por el modelo XGBoost del grupo 1.

Boston Housing.

Por un lado, el modelo de **Mezcla Gaussiana (GMM)** permitió identificar estructuras subyacentes en los datos, revelando tres agrupaciones principales ($K = 3$) determinadas mediante el criterio **BIC**.

Este análisis confirmó que los GMM ofrecen una descripción probabilística más rica que métodos como k-means, al capturar la forma gaussiana de los agrupamientos y considerar la covarianza entre variables.

El análisis de agrupamiento reveló una relación **inversa** entre el número promedio de cuartos (`rm`) y el porcentaje de población de bajo estatus (`lstat`).

Los vecindarios con un mayor número de cuartos tienden a presentar una menor proporción de población de bajo estatus, lo que sugiere que el modelo logra capturar adecuadamente el gradiente socioeconómico implícito en los datos.

Por otro lado, el modelo **XGBoost** fue implementado como técnica de **aprendizaje supervisado** para predecir el **valor mediano de las viviendas** a partir de variables socioeconómicas y ambientales. Los resultados mostraron un excelente ajuste, con un coeficiente de determinación de R^2 entre 0.78 y 0.89 según el clúster y un error cuadrático medio (RMSE) entre 1.98 y 3.05, lo que evidencia la capacidad del modelo para capturar relaciones no lineales entre las variables.

En conjunto, ambos enfoques demuestran el potencial del **aprendizaje automático** para analizar y modelar datos complejos. Mientras que el GMM permite descubrir patrones latentes sin supervisión, el XGBoost ofrece un marco predictivo robusto que complementa el análisis exploratorio con una estimación precisa del valor de las viviendas.

Referencias

Ding, H. (2024). Predicting boston housing price using machine learning models. In *Proceedings of the 2024 2nd International Conference on Management Innovation and Economy Development (MIED 2024)*, pages 439–444.

Fang, Z., Dowe, D. L., Peiris, S., and Rosadi, D. (2021). Minimum message length in hybrid arma and lstm model forecasting. *Entropy*, 23(12):1601. DOI: 10.3390/e23121601.

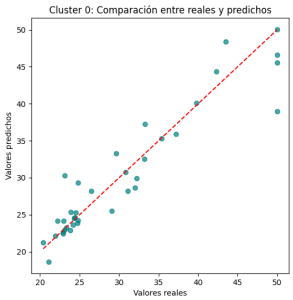

Figura 5. Comparación entre valores reales y predichos por el modelo XGBoost del grupo 2.

Lai, S. B. S., Shahri, N. H. N. B. M., Mohamad, M. B., Rahman, H. A. B. A., and Rambli, A. B. (2021). Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data. *Mathematics and Statistics*, 9(3):379–385.

Li, Y., Dong, M., and Hua, J. (2007). A gaussian mixture model to detect clusters embedded in feature subspace. *Communications in Information and Systems*, 7(4):337–352.

Zhang, Y. (2017). Gaussian mixture model clustering for high-dimensional data. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.290522.