



Tecnológico de Monterrey

Entregable E3

–
Minería de Datos

Mónica González - A01735626
Mariana Rico - A01735770
Alberto Muro - A01734046

Clasificación Bayesiana

El paso subsiguiente consiste en emplear la clasificación bayesiana para desarrollar un modelo predictivo que determine la categoría a la que pertenecen las viviendas. Esta clasificación se fundamenta en los clusters generados durante el análisis de clusters. Las categorías asignadas reflejan los niveles socioeconómicos a los que cada casa está orientada, abarcando segmentos como medio bajo, medio alto, alto bajo y alto alto. Este enfoque proporciona una herramienta valiosa para asignar de manera precisa y eficiente las viviendas a sus respectivos grupos socioeconómicos, de modo que podamos conocer a qué segmento dirigirnos cuando tengamos un nuevo inmueble con ciertas características.

Limpieza y Categorización

Para lograr este objetivo, previamente realizamos una limpieza de la base de datos, dado que algunos strings presentaban espacios que, de manera incorrecta, eran interpretados como la creación de una nueva clase. Este proceso de limpieza garantizó la coherencia de los datos y evitó distorsiones en la clasificación. Posteriormente, durante el análisis de clusters, se generó una nueva columna que asigna a cada casa a su correspondiente grupo socioeconómico. Esta columna recién creada se erige como nuestra variable dependiente en el modelo de clasificación bayesiana, sirviendo como base fundamental para la predicción precisa de las categorías socioeconómicas de las viviendas.

Por otro lado, nuestras variables dependientes son:

- Cocina Equipada
- Gimnasio
- Amueblado
- Alberca
- Terraza
- Elevador
- Número de baños
- Número de recamaras
- Lugares de estacionamiento

Los datos de estas variables nos ayudarán a predecir la clase de los nuevos registros.

Cálculo de Modelos

Para hacer el cálculo de modelos, debimos seguir los siguientes pasos:

- Creación de sets de prueba y entrenamiento: se dividió la base de datos, de modo que el 20% de los registros aleatorios nos sirvieran como base de prueba para evaluar el modelo.
- Entrenamiento de modelo: Se entrenó con el Naive Bayes Multinomial y posteriormente se ingresaron los datos.

- Evaluación del modelo

Evaluación de Modelos

Puntajes

Los puntajes obtenidos para nuestro modelo fueron:

```
Comparación de Métricas:  
Clasificador: Naive Bayes Multinomial  
Exactitud: 0.67  
Puntaje F1: 0.62  
Precisión: 0.58  
Sensibilidad (Recall): 0.67
```

Imagen1. Resultados Modelo Naive-Bayes Multinomial

Los resultados de las métricas para el clasificador Naive Bayes Multinomial revelan un desempeño general moderado en la tarea de clasificación. La exactitud del 67% indica que el modelo acierta en la categorización de las instancias casi dos tercios del tiempo. El puntaje F1, que combina precisión y sensibilidad, se sitúa en 0.62, sugiriendo un equilibrio aceptable entre ambas métricas. Sin embargo, la precisión del 58% indica que, de las instancias clasificadas como positivas, un poco más de la mitad son verdaderamente positivas. La sensibilidad, o recall, del 67% destaca la capacidad del modelo para identificar la mayoría de las instancias positivas. En conjunto, estos resultados indican que el Naive Bayes Multinomial es capaz de realizar la tarea de clasificación, aunque con ciertas limitaciones en términos de precisión. Se sugiere una evaluación más detallada para comprender las características específicas de los errores,

Matriz de Confusión

Por otro lado, tenemos la Matriz de Confusión, la cual revela un desempeño alentador del modelo, ya que la mayor parte de sus datos se encuentran en los cuadrantes correspondientes a las clasificaciones correctas.

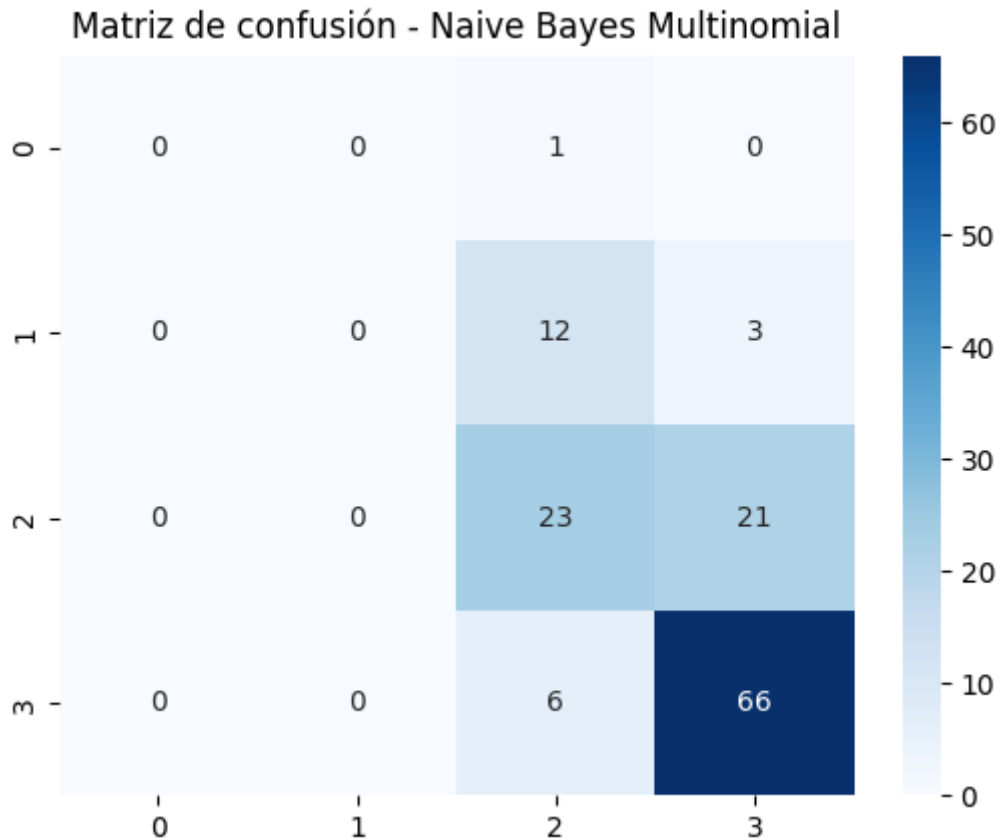


Imagen 2. Matriz de Confusión - Modelo Naive-Bayes Multinomial

Predicción

Finalmente tenemos una prueba de predicción del modelo, donde generamos una función que nos permita predecir el nivel socioeconómico de nuevos registros con el modelo entrenado anteriormente.

Después, generamos un nuevo registro con las siguientes características:

Cuenta con:

- Cocina equipada
- Alberca
- Terraza
- Elevador
- Tres baños
- Tres recamaras
- Dos lugares de estacionamiento

Por otro lado, no cuenta con gimnasio ni viene amueblado.

Con esta prueba, obtuvimos la siguiente respuesta:

```
Probabilidad por clase 0: 0.0285
Probabilidad por clase 1: 0.3180
Probabilidad por clase 2: 0.5845
Probabilidad por clase 3: 0.0690
Clase Predicha: MedioAlto
```

Imagen3. Probabilidades de clase de acuerdo a modelo de Clasificación Bayesiana

Esto nos indica que (de acuerdo a la probabilidad más alta), la casa puede pertenecer a personas con un nivel socioeconómico medio alto.

Conclusiones

En conclusión, el enfoque de clasificación bayesiana, respaldado por la generación de clusters y la inclusión de variables socioeconómicas clave, ofrece una herramienta efectiva para asignar con precisión viviendas a sus respectivos grupos socioeconómicos. La limpieza y categorización previa de la base de datos fueron cruciales para garantizar la coherencia en la clasificación. Aunque el modelo Naive Bayes Multinomial presenta métricas moderadas, su capacidad para predecir niveles socioeconómicos es alentadora, con la matriz de confusión validando su desempeño. La evaluación detallada de métricas y la prueba de predicción demuestran la utilidad práctica del modelo para identificar el nivel socioeconómico de nuevas viviendas. Este enfoque integrado, desde la limpieza de datos hasta la evaluación de modelos y predicciones, proporciona un marco sólido para la toma de decisiones informada en el ámbito inmobiliario.