Clasificación de Rangos de Precios de Teléfonos Móviles Mediante Aprendizaje Automático Basado en Características.

Nicolás Lozano Mazuera, Sebastián Urquijo Buitrago, y Juan Manuel García Ortiz
Facultad de ingeniería y Ciencia básicas
Universidad Autónoma de Occidente
Cali, Colombia
(Nicolas.lozano_m - Sebastian.urquijo - Juan_man.garcia) @uao.edu.co

Resumen: En este trabajo se abordaron técnicas de Machine Learning con el objetivo de clasificar los rangos de precios para teléfonos móviles de una base de datos con más 2.000 registros de teléfonos, de acuerdo con sus características diversas se desarrollaron tres modelos de aprendizaje automático supervisado (Regresión Logística, Random Forest con Selección de Características y Random Forest optimizado con GridSearchCV.) Se realizó un preprocesamiento y entrenamiento de los modelos se identificaron los modelos más efectivos y las características de mayor impacto. Este estudio ofrece una herramienta predictiva robusta para optimizar estrategias de precios y guiar a los consumidores en el mercado de telefonía móvil.

I. INTRODUCCIÓN

Este trabajo tiene como objetivo desarrollar un modelo de Machine Learning que determine de manera precisa la clasificación de precios en que se puede ubicar un dispositivo celular de acuerdo con sus diferentes características, permitiendo mejorar las estrategias de precios para lograr tener mejores ventas y así mismo orientar al cliente según sus necesidades y presupuesto.

Para esto se utilizará la base de Datos “Mobile Price” **[1]** la cual contiene más de 2.000 registros de celulares con sus diferentes características (Batería, RAM, Memoria etc..) permitiendo desarrollar un modelo de clasificación de rangos de precios.

II. IMPORTANCIA DEL TRABAJO.

El precio de los productos es el principal aspecto y el más importante a determinar por la empresas [2], por lo que resulta importante tener un buen algoritmo de clasificación que determine el rango de precios en el que se encuentra un celular de acuerdo a sus características [3], basados en algunos estudios [4] el modelo SVM obtuvo uno de los mejores resultados, mientras que la regresión lineal no tuvo resultados favorables, por lo que en este trabajo resulta importante la aplicación de un modelo de clasificación de precios de celulares para mejorar la toma de decisiones estratégicas para los consumidores y vendedores de celulares.

III. DESCRIPCIÓN DEL FENÓMENO / PROCESO MODELADO Y EL PROBLEMA A ABORDAR.

En el contexto actual del mercado de dispositivos móviles, las decisiones estratégicas relacionadas con la fijación de precios son fundamentales para lograr competitividad. Compañías como Apple, Samsung y Xiaomi han consolidado su liderazgo no solo a través de innovación tecnológica, sino mediante la alineación precisa entre las características técnicas de sus productos y su posicionamiento de mercado. Sin embargo, para nuevos actores o empresas emergentes, establecer una política de precios basada en la intuición o experiencia previa puede resultar insuficiente o incluso contraproducente.

El proceso modelado en este estudio corresponde a la relación entre las especificaciones técnicas de un teléfono móvil y su precio en el mercado, bajo el supuesto de que existe una correlación significativa entre estos factores. En otras palabras, se busca predecir a qué rango de precios pertenece un dispositivo móvil nuevo basándose exclusivamente en sus características técnicas cuantificables.

Este tipo de relación es de interés tanto para fabricantes como para consumidores. Desde el punto de vista comercial, contar con una herramienta automatizada que permita ubicar un producto en un rango de precio adecuado permite diseñar estrategias de mercadeo y posicionamiento más eficientes. Por su parte, el consumidor puede acceder a un conjunto de opciones ajustadas a su presupuesto con base en atributos objetivos, evitando decisiones basadas en percepciones subjetivas.

Desde una perspectiva de aprendizaje automático, el fenómeno descrito se formaliza como un problema de clasificación multiclase supervisada, donde un conjunto de variables independientes (features) sirve para predecir una variable dependiente categórica (price_range). Esta variable objetivo agrupa los dispositivos en cuatro categorías de precios: bajo, medio, alto y muy alto.

Este tipo de tareas requiere algoritmos capaces de identificar patrones complejos entre múltiples variables.

Según estudios recientes, modelos como Support Vector Machines (SVM), Random Forests y K-Nearest Neighbors (k-NN) han demostrado ser efectivos para problemas similares en el contexto del mercado tecnológico y retail [6][7].

IV. PROCESO DE OBTENCIÓN / GENERACIÓN DEL CONJUNTO DE DATOS.

El conjunto de datos utilizado en este trabajo fue descargado desde la plataforma Kaggle, una comunidad de ciencia de datos reconocida por su aporte al desarrollo de modelos predictivos mediante competencias y repositorios colaborativos. El dataset titulado “Mobile Price Classification”, contiene 2.000 registros de teléfonos móviles con 20 variables predictoras (features) y una variable objetivo categórica.

Las variables incluyen características como:

battery_power (capacidad de la batería en mAh),

ram (memoria RAM en MB),

px_height y px_width (resolución de la pantalla),

four_g, blue, dual_sim (tecnologías soportadas),

entre otras.

La variable objetivo es price_range, que toma valores enteros de 0 a 3, representando los rangos de precios mencionados. Este dataset fue generado artificialmente para propósitos educativos, lo cual asegura un balance entre las clases, la ausencia de valores faltantes y una calidad estructural que facilita su uso para el entrenamiento de modelos de clasificación.

V. DESCRIPCIÓN DEL PROBLEMA DE MACHINE LEARNING

El presente estudio aborda un problema de clasificación supervisada. El objetivo principal es desarrollar un modelo capaz de predecir la categoría en los rangos de precios de un teléfono móvil basándose en el conjunto de características técnicas y físicas.

En el contexto del aprendizaje automático, esto se traduce en aprender una función de mapeo $f: (X) = Y$ donde:

- F: Es la función objetivo o modelo que el algoritmo de aprendizaje automático intenta aprender.
- X: Representa el espacio de entrada. Donde están las posibles variables de entrada o características (features) que se utilizan para hacer una predicción.
- Y: Representa el espacio de salida. Consiste en el conjunto de todas las posibles variables de salida o etiquetas (labels) que el modelo intenta predecir.

En nuestro Caso Para X se representa en el vector de características de un teléfono móvil como (Battery_power, RAM, Px_width, Px_height, etc.) y Y es la variable objetivo-categórica, denominada Price_range.

A. Tipo de Problema

Es un problema de clasificación multiclase, ya que la variable objetivo Price_range puede tomar uno de cuatro valores discretos y ordenados que representan diferentes niveles de costo, por ejemplo: (Bajo, Medio, Alto, Muy alto). Muchos algoritmos de clasificación pueden aplicarse directamente, tratando las clases como categorías distintas. Sin embargo, en este conjunto de datos se muestra una distribución balanceada para cada clase.

B. Tarea de Aprendizaje Por Realizar

La tarea de aprendizaje automático consiste en entrenar uno o varios modelos utilizando un conjunto de datos etiquetados (Supervisado), donde cada instancia (teléfono móvil) tiene un conjunto de atributos y una etiqueta de **Price_range** conocida. El modelo aprenderá a identificar patrones y relaciones entre las características del teléfono y su respectivo rango de precios. Una vez entrenado el modelo en una primera instancia con el 70% de los datos, el modelo deberá ser capaz de generalizar este conocimiento para predecir con precisión el rango de precios de teléfonos móviles con el 100% o el restante 30% de datos, basándose únicamente en sus características. Este tipo de predicciones son importantes tanto para los consumidores, ya que les permite entender de manera rápida bajo su presupuesto o necesidad en tipo de celulares o rango de celulares tiene disponible, pero también resulta útil para los fabricantes o vendedores ya que pueden generarse estrategias de precios de acuerdo con el rango de celulares o movimientos en el mercado.

VI. PREPROCESAMIENTO DE DATOS

El preprocesamiento de datos es una de las etapas fundamentales en el desarrollo de modelos de Machine Learning, ya que la calidad y formato de los datos de entrada impactan directamente el rendimiento y la precisión del modelo. Para el dataset "Mobile Price Classification", que cuenta con más de 2.000 registros y no se presentan datos faltantes y considerando los hallazgos del Análisis Exploratorio de Datos (EDA), se proponen las siguientes metodologías de preprocesamiento.

A.Manejo de Valores Atipicos (Outliers)

Con base en el Análisis Exploratorio de Datos (EDA), se implementó un preprocesamiento de valores atípicos. Para las variables front_camera_res (Resolución de Cámara) y px_height (Altura de Píxeles), se identificaron valores elevados mediante el Rango Intercuartílico (IQR). Considerando que estos podrían ser datos legítimos de dispositivos de gama alta, se optó por una técnica de capping (winsorización) en lugar de su eliminación. Adicionalmente, se detectaron y corrigieron valores irreales de cero en px_height y screen_width (Ancho de la pantalla) mediante un capping inferior.

B. Escalado de Características (Feature Scaling)

Muchos algoritmos de Machine Learning son sensibles a la escala de las características de entrada. En nuestro dataset, características como la Ram y Battery_power tienen magnitudes significativamente diferentes a otras como Clock_speed. Por lo que se aplicará la Estandarización (StandardScaler). Esta técnica transforma los datos para que tengan una media que va de 0 y una desviación estándar de 1. Para una característica X, el valor estandarizado Z se calcula como:

$$X^{'} = \frac{x-\mu}{\sigma}$$

donde μ es la media de la característica y σ es su desviación estándar. Esto se aplicará a todas las características numéricas continuas.

C.Selección de Características

Se abordó la selección de características con el objetivo de identificar las variables más influyentes para la clasificación del rango de precios y reducir la dimensionalidad del modelo. Se inició este proceso mediante un análisis de correlación entre las variables numéricas y la variable objetivo price_range. Este análisis permitió una evaluación preliminar de la relevancia predictiva de cada característica, destacando aquellas con una fuerte asociación lineal con el rango de precios, como la memoria RAM, así como aquellas con una correlación aparentemente baja, como la velocidad del reloj (clock_speed) y el número de núcleos (n_cores). Esta etapa es fundamental para guiar la eliminación de predictores redundantes o irrelevantes, buscando mejorar la eficiencia computacional, la interpretabilidad y el rendimiento predictivo del modelo final al mitigar el riesgo de sobreajuste.

C.1 Aplicación del Análisis de Componentes Principales (PCA)

En este estudio, se aplicaría el Análisis de Componentes Principales (PCA) al conjunto de características numéricas previamente preprocesadas (escaladas y con valores atípicos tratados). Esta técnica transformaría dichas variables originales en un conjunto dimensionalmente reducido de componentes principales ortogonales, los cuales capturan la mayor parte de la varianza de los datos. Posteriormente, estos componentes se emplearían como variables predictoras de entrada para el modelo de clasificación de rangos de precio, en sustitución de las características numéricas originales. El número óptimo de componentes a retener se determinará con base en el porcentaje de varianza acumulada explicada, utilizando un umbral preestablecido (95%) para asegurar la preservación de la información más relevante.

C.**2. Aplicación de Métodos de Selección Univariados (Filter Methods)**

La selección de características mediante métodos univariados se llevaría a cabo evaluando la relación de cada predictor con la variable objetivo price_range de manera independiente. Para las características numéricas (ram, battery_power), se emplearía la prueba **ANOVA F-test** (utilizando la función f_classif de la librería scikit-learn) para cuantificar la significancia de las diferencias de sus medias entre las distintas categorías de price_range. En el caso de las características categóricas (bluetooth, four_g), transformadas previamente mediante one-hot encoding, se aplicaría la prueba Chi-cuadrado (χ^2) (función chi2 de scikit-learn) para medir la dependencia estadística con la variable objetivo. Con base en las puntuaciones o los p-valores resultantes, se procedería a seleccionar un subconjunto de las k características más predictivas para la construcción del modelo de clasificación.

C.**3. Aplicación de Métodos de Selección Basados en Modelos (Wrapper o Embedded)**

Para la selección de características mediante enfoques basados en modelos, se explorarán dos estrategias. Primero, un método embedded, cómo el algoritmo Random Forest, el cual se entrenaría con el conjunto completo de características pre procesadas, incluyendo aquellas obtenidas mediante ingeniería de características (front_camera_present, screen_aspect_ratio). El atributo feature_importances_ derivado de este modelo proporcionaría una métrica de la relevancia de cada predictor, permitiendo la selección de un subconjunto optimizado. Como alternativa, se podría implementar un método wrapper, tal como la Selección Recursiva de Características (RFE), utilizando un algoritmo de clasificación base (Regresión Logística o SVM). RFE reduciría iterativamente el conjunto de predictores, eliminando los menos influyentes en cada paso, hasta alcanzar un número predefinido de características o un rendimiento óptimo en la validación.

VII. ENTRENAMIENTO Y EVALUACIÓN DE MODELOS.

Con el conjunto de datos preprocesado se realizó una partición de los datos en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%) utilizando la función train_test_split de Scikit-learn previo a estandarizar los valores (StandardScaler) y entrenar distintos modelos. Los modelos implementados fueron: Regresión Logística, Random Forest, y Random Forest optimizado mediante GridSearchCV.

Regresión Logística

Se entrenó un modelo de regresión logística multiclase (solver='lbfgs') sobre las variables numéricas estandarizadas. Este modelo obtuvo un accuracy del 97.25% sobre el conjunto de prueba, con un desempeño uniforme entre las cuatro clases de la variable objetivo. El análisis del reporte de clasificación indicó valores elevados de precisión y recall, evidenciando la capacidad del modelo para capturar patrones generalizables en los datos.

Random Forest con Selección de Características (RFE)

Se aplicó la técnica Recursive Feature Elimination (RFE) utilizando Regresión Logística como modelo base para seleccionar las cinco características más relevantes. Estas fueron: battery_power, px_height, px_width, ram y weight. Posteriormente, se entrenó un modelo de Random Forest con 200 árboles (n_estimators=200) sobre este subconjunto reducido, obteniendo un accuracy del 93.5%, con una reducción significativa en la complejidad del modelo sin pérdida crítica de desempeño.

Random Forest con Optimización de Hiperparámetros (GridSearchCV)

Finalmente, se utilizó la técnica de búsqueda en malla (GridSearchCV) para encontrar la mejor combinación de hiperparámetros (n_estimators y max_depth) en un modelo Random Forest. Se evaluaron múltiples combinaciones usando validación cruzada de 5 pliegues. El modelo óptimo encontrado presentó una precisión de validación cruzada del 95.4%, y un desempeño consistente sobre el conjunto de prueba según el reporte de clasificación.

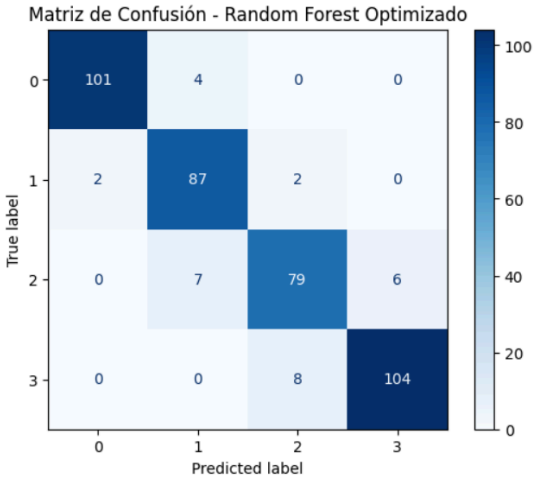

Figura 1: Matriz de Confusion del Random Forest Optimizado

Métricas y Comparación

Las métricas de evaluación utilizadas incluyeron accuracy, precision, recall y f1-score. A continuación se presenta un resumen comparativo de los modelos evaluados:

Modelo	Accuracy (Test)	Comentario
Regresión Logística	97.2%	Mejor desempeño general
Random Forest (con RFE)	93.5%	Buen balance entre simplicidad y rendimiento
Random Forest (GridSearchCV)	95.4%	Buen ajuste con validación cruzada

Los resultados muestran que la Regresión Logística fue el modelo más preciso, a pesar de su simplicidad. Sin embargo, Random Forest mostró una buena capacidad de generalización y permitió interpretar la importancia relativa de las características predictoras.

VIII. REFERENCIAS.

[1] A. iabhishekofficial, “Mobile Price Classification,” Kaggle, [En línea]. Disponible en: [https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification/data]. [Accedido: 7-may-2025].

[2] Smartphone Price Prediction in Retail Industry Using Machine Learning Techniques. In: Sridhar, V., Padma, M., Rao, K. (eds) Emerging Research in Electronics, Computer Science and Technology. Lecture Notes in Electrical Engineering, vol 545. Springer, Singapore. https://doi.org/10.1007/978-981-13-5802-9_34

[3] Algorithms with Feature Selection and Parameter Optimization," 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2021, pp. 483-487

[4] Rodríguez, F. (2024). “Estimación de precios de mercado para celulares usados mediante técnicas de aprendizaje automático”. [Tesis de maestría. Universidad Torcuato Di Tella]. Repositorio Digital Universidad Torcuato Di Tella. https://repositorio.utdt.edu/handle/20.500.13098/12786

[5] "How Many People Own Smartphones? (2024-2029)," Exploding Topics, Abril 2025. [En línea]. Disponible en: https://explodingtopics.com/blog/smartphone-stats. [Accedido: 10-Mayo-2025].

[6] Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.

[7] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Repositorio Público del proyecto: https://github.com/IAproyecto24/MachineLearning.git