Modelo Predictivo de Precios Agropecuarios en Colombia (2013-2021) mediante Machine Learning

Shara Pérez Estrada y Cristian Camilo Orozco Aguirre

Talento Tech

Inteligencia Artificial  
Talento Tech, Universidad de Caldas, Universidad de Antioquia

Manizales, Caldas

Fecha 22 de marzo 2025

****Contenido****

Resumen 4

Introducción 5

Planteamiento del Problema 6

Alcance del Proyectos 7

Justificación 8

Objetivos 8

Objetivo General 8

Objetivos Específicos 8

Metodología 9

Obtención y carga de datos: 9

Limpieza y preprocesamiento de datos: 9

Análisis exploratorio de datos (EDA): 10

Selección y entrenamiento de modelos: 10

Evaluación de modelos: 10

Generación de predicciones y visualización de resultados: 11

Importación de librerías 11

Descripción de la Base de Datos 13

Entendimiento de los datos 14

Análisis de los Datos del EDA 16

Las variables clave en este análisis son: 18

Limpieza de datos 22

Modelado de Datos 23

¿Qué es Gradient Boosting? 24

¿Quién usa LightGBM? 24

Ejemplo Práctico: Uso de LightGBM en Finanzas 25

Preparación de los datos 25

Encontrar el mejor modelo 26

Descripción de algunas de las herramientas utilizadas por FLAML: 26

Resultados 27

Evaluación del rendimiento del modelo 27

Nuestros resultados 29

Hiperparámetros 30

Interpretación general 32

Comparación de valores reales vs predichos 32

Importancia de las variables 34

Conclusiones 35

Líneas Futuras 36

Mejoras para el Modelo Predictivo 36

Posibles Usos Futuros del Modelo 38

Referencias 41

Anexos 44

****Resumen****

**El documento presenta un estudio que desarrolla un modelo predictivo para anticipar los precios mayoristas del sector agropecuario en Colombia entre 2013 y 2021, utilizando técnicas de Machine Learning. Los autores, Shara Pérez Estrada y Cristian Camilo Orozco Aguirre, emplearon datos del DANE, realizando una rigurosa limpieza y estandarización de más de 500.000 registros para garantizar su calidad y precisión.**

**El estudio seleccionó LightGBM como el modelo más eficiente, por su rapidez y capacidad para manejar grandes volúmenes de datos y variables categóricas sin pre procesamiento adicional. Durante la metodología, se aplicaron técnicas como codificación de variables, análisis exploratorio de datos y evaluación mediante métricas como R², RMSE y MAE, alcanzando un coeficiente de determinación de 0.97, lo que evidencia un alto nivel de precisión.**

**Se identificó que las variables más influyentes en la predicción son el tipo de producto y el grupo agropecuario, mientras que la región tuvo menor impacto. La investigación también señala la importancia de detectar valores atípicos y transformar los datos para mejorar la estabilidad del modelo.**

**En sus conclusiones, el estudio resalta el potencial del modelo para optimizar decisiones en la cadena agroindustrial, al permitir a productores, comerciantes y entidades públicas anticipar tendencias de precios y reducir riesgos. Finalmente, propone líneas futuras como la incorporación de variables climáticas y económicas, modelos más complejos como redes neuronales, y la automatización del modelo para actualizaciones en tiempo real, ampliando así su aplicabilidad a políticas públicas, exportaciones y desarrollo de productos financieros.**

****Introducción****

El sector agropecuario colombiano enfrenta desafíos significativos debido a la volatilidad de los precios de los productos agrícolas. La variabilidad en los costos se debe a múltiples factores, incluyendo la oferta y la demanda, las condiciones climáticas y las políticas gubernamentales. En este contexto, predecir el comportamiento de los precios se convierte en una herramienta clave para productores, comerciantes y responsables de la formulación de políticas.

El uso de modelos predictivos basados en técnicas de Machine Learning permite identificar patrones en los datos históricos y generar proyecciones sobre la evolución de los precios. A través del análisis de datos recopilados y debidamente tratados entre los años 2013 y 2021, este estudio busca comprender las tendencias del mercado agropecuario en Colombia y proporcionar una base analítica para mejorar la toma de decisiones.

Uno de los principales beneficios de los modelos predictivos es su capacidad para reducir la incertidumbre en la comercialización de productos agropecuarios. Al conocer las posibles variaciones de precios con antelación, los productores pueden planificar mejor su producción y distribución, evitando pérdidas económicas y maximizando sus ganancias. Asimismo, los comerciantes pueden gestionar inventarios de manera más eficiente, reduciendo costos de almacenamiento y evitando escasez o sobreoferta.

El presente estudio se basa en datos provenientes del Departamento Administrativo Nacional de Estadística (DANE), una fuente confiable y oficial en Colombia. La base de datos utilizada cuenta con aproximadamente 500.000 registros, lo que permite mejorar la precisión del modelo predictivo. Se seleccionaron variables clave que influyen en la variabilidad de los precios, como el año, el mes, el sector agropecuario, el producto, la región y el municipio.

Este documento presenta la metodología utilizada en la construcción del modelo predictivo, detallando el proceso de limpieza y validación de datos, las bibliotecas de Python empleadas y la interpretación de los resultados obtenidos. Finalmente, se discutirán las implicaciones del modelo en la toma de decisiones y posibles mejoras futuras para optimizar su precisión y aplicabilidad.

Planteamiento del Problema

El sector agropecuario en Colombia se enfrenta a una alta volatilidad en los precios de sus productos debido a múltiples factores como políticas gubernamentales, demanda del mercado y fluctuaciones económicas. La falta de herramientas analíticas que permitan prever estos cambios genera incertidumbre tanto para productores como para comerciantes, dificultando la planificación y afectando la rentabilidad del sector.

La ausencia de modelos predictivos confiables impide la toma de decisiones informadas y estratégicas en la producción y comercialización de bienes agrícolas. Esto lleva a escenarios en los que los productores pueden vender a precios desfavorables o acumular inventarios innecesarios. Por otro lado, los consumidores también pueden verse afectados por el encarecimiento repentino de ciertos productos básicos.

Por lo tanto, surge la necesidad de desarrollar un modelo basado en datos históricos y técnicas de Machine Learning que permita predecir los precios agropecuarios en Colombia, brindando información clave para la toma de decisiones en la cadena de suministro agroindustrial. Según ScienceDirect Topics (s.f.), el aprendizaje automático es una subdisciplina de la inteligencia artificial que permite a los sistemas aprender automáticamente a partir de datos, identificar patrones y tomar decisiones sin estar programados explícitamente para cada tarea. La implementación de este modelo contribuirá a la reducción de riesgos asociados a la variabilidad del mercado y permitirá mejorar la planificación en el sector agropecuario.

Alcance del Proyectos

El presente estudio abarca el análisis de datos de precios agropecuarios en Colombia entre los años 2013 y 2021, con el propósito de desarrollar un modelo predictivo que permita identificar patrones y tendencias en la variabilidad de los precios. Se centra en la aplicación de técnicas de Machine Learning para la predicción.

El modelo se enfocará en analizar la relación entre las variables clave seleccionadas, tales como año, mes, producto, región y precio, con el objetivo de detectar comportamientos estacionales y factores que influyen en la variabilidad de los costos. Asimismo, se integrarán procesos de validación de datos para asegurar la fiabilidad del conjunto de información utilizado en la predicción.

Finalmente, este estudio tiene un alcance limitado a los datos disponibles en el DANE y a la precisión de los modelos utilizados. Aunque se han aplicado técnicas avanzadas de optimización, la inclusión de variables adicionales como factores climáticos o políticas económicas, por ejemplo, podría mejorar la capacidad predictiva en futuras investigaciones. Se espera que este modelo sirva como una base sólida para la toma de decisiones estratégicas y la planificación en el sector agropecuario en Colombia.

Justificación

Este estudio justifica la implementación de modelos de Machine Learning, específicamente LightGBM, como herramientas clave para analizar la variabilidad de los precios agropecuarios en Colombia. El uso de modelos supervisados permite identificar patrones y tendencias en los datos históricos del DANE, optimizando la predicción de precios y facilitando la toma de decisiones estratégicas en el sector agropecuario. La combinación de estas técnicas garantiza un análisis robusto y confiable de los factores que influyen en la comercialización de productos agrícolas, proporcionando una solución efectiva para mejorar la planificación y reducir la incertidumbre en el mercado.

Objetivos

Objetivo General

Desarrollar un modelo predictivo basado en técnicas de Machine Learning que permita analizar la evolución de los precios agropecuarios en Colombia entre 2013 y 2021.

Objetivos Específicos

- Recolectar, depurar y estandarizar datos históricos de precios agropecuarios en Colombia, asegurando su calidad y consistencia.

-Seleccionar e implementar el modelo de Machine Learning supervisado con el mejor rendimiento para el conjunto de datos

- Evaluar la precisión, el ajuste y aplicabilidad del modelo predictivo para la estimación de tendencias futuras.

Metodología

Para garantizar la precisión y calidad de los datos utilizados en el modelo predictivo, se realizó un proceso riguroso de validación y limpieza. La base de datos, proveniente del DANE, fue sometida a una depuración exhaustiva para eliminar inconsistencias y errores en la digitación de precios y demás variables.

Obtención y carga de datos:

* + Se recopiló información histórica de precios de productos agropecuarios desde 2013 hasta 2021 en distintas ciudades de Colombia.
  + Los datos fueron importados y procesados utilizando bibliotecas como pandas y numpy.

Limpieza y preprocesamiento de datos:

* + Se estandarizó el formato de cada documento, lo que incluye la eliminación de columnas y homogeneización de títulos.
  + Se eliminaron valores nulos y duplicados para asegurar la calidad de la información.
  + Dividieron columnas que contenían dos variables juntas, como por ejemplo ciudades junto con departamento y las fechas, haciendo separación de mes y año.
  + Se transformaron variables categóricas en variables numéricas mediante técnicas de codificación.
  + Descarte de base de datos con errores en la digitación de precios
  + Eliminación de dos archivos de excel, correspondientes a los años 2022 y 2023
  + Estandarización de nombres de mercados por región.
  + Eliminación de tildes y conversión de palabras a mayúsculas.

Análisis exploratorio de datos (EDA):

* + Se realizaron visualizaciones con Pandas**,** matplotlib y seaborn para identificar tendencias y patrones en los precios a lo largo del tiempo de manera rápida y visual.
  + Se calcularon estadísticas descriptivas para comprender la distribución de los datos.

Selección y entrenamiento de modelos:

* + Se probaron varios modelos de aprendizaje automático para la predicción de precios, y LightGBM (LGBM).
  + Se dividieron los datos en conjuntos de entrenamiento y prueba para evaluar el rendimiento de los modelos.
  + Se ajustaron hiperparámetros utilizando técnicas de optimización.

Evaluación de modelos:

* + Se utilizaron métricas como el error cuadrático medio (MSE), el coeficiente de determinación (R²) y el error absoluto medio (MAE) para comparar el desempeño de los modelos.
  + Se observó que el modelo LGBM obtuvo los mejores resultados en términos de precisión.

Generación de predicciones y visualización de resultados:

* + Se generaron gráficos para visualizar las predicciones y compararlas con los valores reales.
  + Se interpretaron los resultados para extraer conclusiones sobre las variaciones de precios en diferentes regiones y periodos.

Este enfoque permitió obtener un modelo predictivo que captura tendencias en los precios agropecuarios y facilita la toma de decisiones basada en datos.

Importación de librerías

- Pandas:  Se empleó para la manipulación y limpieza de los datos, permitiendo leer, transformar y gestionar la base de datos de precios agropecuarios.  
- Seaborn y Matplotlib: Estas bibliotecas se usaron para la visualización de datos, facilitando la exploración de tendencias y patrones en los precios mediante gráficos.  
- Numpy: Utilizada para la optimización de cálculos numéricos y operaciones matemáticas dentro del modelo predictivo.  
- Sweetviz: Herramienta aplicada en el análisis exploratorio de datos (EDA), generando informes automáticos con resúmenes estadísticos y gráficos.  
- Category\_encoders: Se utilizó para la codificación de variables categóricas, transformando datos no numéricos en formatos compatibles con modelos de Machine Learning.  
- Google Colab: Entorno en el que se ejecutó el código, permitiendo la implementación y prueba del modelo predictivo con acceso a herramientas en la nube.  
- OS, RE y Math: Se emplearon para tareas auxiliares como la gestión de archivos, la manipulación de cadenas de texto y cálculos matemáticos avanzados.Se continuo con la limpieza de los datos, se homogeneizaron las variables en cada columna, con funciones que toman de referencia el municipio se completó la columna de departamento y se creó la columna región.

Descripción de la Base de Datos

Nuestra base de datos consta de 5 documentos de Excel que al ser concatenados, limpiados y estructurados posee una dimensión de 505377 filas y 8 columnas.

**Tabla 1**

*Descripción de las variables*

|  |  |  |
| --- | --- | --- |
| Variable | **Tipo** | **Descripción** |
| Mes | Ordinal | Indica el mes en el que se registró el precio del producto. |
| Año | Ordinal | Año en el que se registró el precio del producto agropecuario. |
| Grupo de Productos | No Ordinal | Clasificación general del producto dentro del sector agropecuario. |
| Producto | No Ordinal | Nombre específico del producto agropecuario analizado. |
| Ciudad | No Ordinal | Ciudad donde se registró el precio del producto. |
| Departamento | No Ordinal | Departamento de Colombia donde se tomó el registro de precios. |
| Región | No Ordinal | Zona geográfica general donde se registró el precio del producto. |
| Precio por Kg | Numérica continua | Valor del producto agropecuario por kilogramo en el mercado. |

Ordinal:

Son variables categóricas que tienen un orden o jerarquía específica. Por ejemplo, los niveles educativos (primaria, secundaria, universidad) o los meses del año (enero, febrero, marzo…). Aunque no representan valores numéricos, su orden sí es significativo. (FreeCodeCamp, 2021; Superprof, s.f.)

No ordinal:

También conocidas como categóricas nominales, estas variables representan categorías sin un orden específico. Ejemplos incluyen el color de un producto (rojo, azul, verde) o el nombre de una ciudad (Bogotá, Medellín, Cali). No hay una jerarquía en estas categorías. (FreeCodeCamp, 2021; Superprof, s.f.)

***Numérica***:

Son variables que representan cantidades medibles y pueden usarse en operaciones matemáticas. Se dividen en:

* *Discretas:* Toman valores enteros (por ejemplo, el número de productos vendidos).
* *Continuas:* Pueden tomar cualquier valor dentro de un rango (por ejemplo, el precio en kilogramos de un producto). (FreeCodeCamp, 2021; Superprof, s.f.)

Entendimiento de los datos

Antes de la construcción del modelo, es fundamental comprender la estructura y distribución de los datos utilizados en el análisis. Esto permite identificar patrones, correlaciones y posibles problemas que puedan afectar la calidad de las predicciones.

Se realizó un análisis exploratorio de los datos (EDA) para evaluar la consistencia de la información. Se examinaron las variables clave: mes, año, grupo de productos, producto, ciudad, departamento, región y precio por kilogramo, verificando su distribución y relación con el precio.

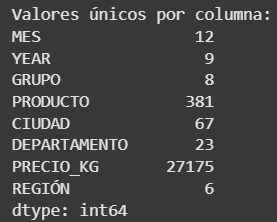
Entre los aspectos analizados, se incluyen:

* Distribución de los datos: Se identificaron tendencias en los precios a lo largo del tiempo y diferencias entre regiones y productos.
* Valores atípicos y datos faltantes: Se evaluó la existencia de valores extremos o registros incompletos que podrían sesgar el modelo.
* Correlación entre variables: Se determinó qué factores tienen mayor influencia en la variabilidad de los precios agropecuarios.

Este proceso permitió garantizar que los datos estuvieran limpios y estructurados correctamente antes de aplicarlos al modelo de Machine Learning.

**Figura 1**

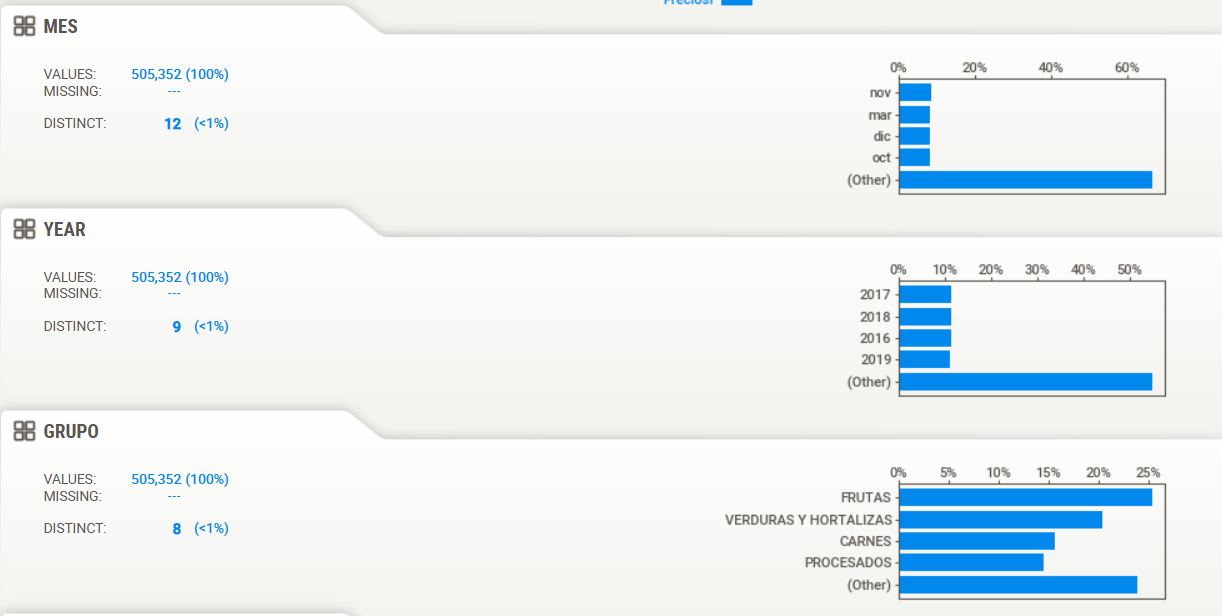
*Valores únicos por columna*



Análisis de los Datos del EDA

**Figura 2**

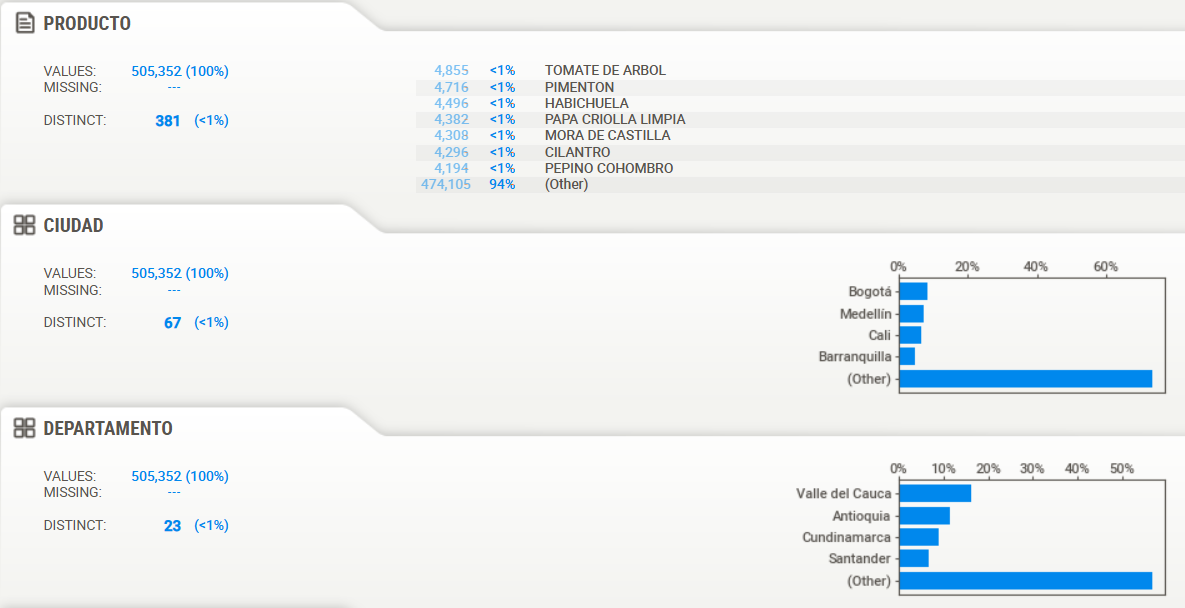
*Graficas de barras para MES, YEAR y GRUPO*

****

Nota: Las gráficas presentadas se generaron como resultado de ejecutar un modelo de análisis de datos (Exploratory Data Analysis, EDA) en Google Colab. Este modelo permite explorar y visualizar los datos de manera interactiva para identificar patrones, tendencias y características clave.

**Figura 3**

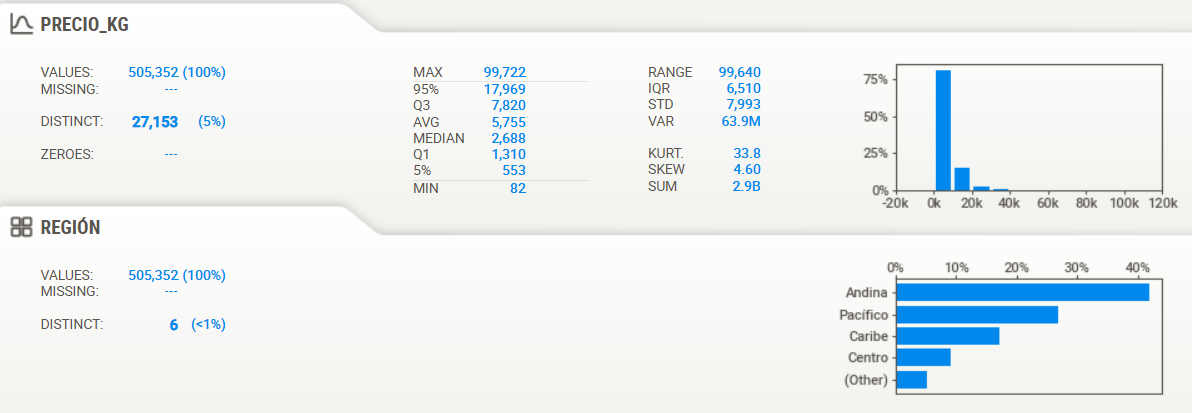
*Graficas de barras para PRODUCTO, CIUDAD y DEPARTAMENTO*

****

Nota: Las gráficas presentadas se generaron como resultado de ejecutar un modelo de análisis de datos (Exploratory Data Analysis, EDA) en Google Colab. Este modelo permite explorar y visualizar los datos de manera interactiva para identificar patrones, tendencias y características clave.

**Figura 4**

*Graficas de barras para PRECIO\_KG y REGION*

****

Nota: Las gráficas presentadas se generaron como resultado de ejecutar un modelo de análisis de datos (Exploratory Data Analysis, EDA) en Google Colab. Este modelo permite explorar y visualizar los datos de manera interactiva para identificar patrones, tendencias y características clave.

Análisis Exploratorio de Datos (EDA) del Mercado Agropecuario en Colombia

El análisis exploratorio de datos (EDA) es un paso fundamental para comprender la estructura de los datos y detectar patrones clave antes de aplicar modelos predictivos. En este estudio, contamos con 500,000 registros, y nos enfocamos en las variables más relevantes para la predicción de precios en el sector agropecuario colombiano.

Las variables clave en este análisis son:

1. Precios
2. Productos
3. Grupos Agropecuarios
4. Factores Temporales (Mes y Año)
5. Factores Geográficos (Ciudad, Departamento y Región)

A continuación, se presenta el análisis detallado de cada una:

Distribución de Precios

La variable precio muestra una amplia variabilidad, con rangos muy distintos entre productos:

* Precios altos: Se encuentran hasta $99,800, representando casos excepcionales dentro de la base de datos.
* Precios medios: Se ubican alrededor de $2,600, lo que representa la mayoría de los productos en condiciones normales del mercado.
* Precios bajos: Llegan hasta $500, lo que indica productos de bajo costo o vendidos en unidades pequeñas.

Distribución y Errores en los Datos:

* La mayoría de los productos tienen precios entre $500 y $12,000, con una distribución que sugiere un comportamiento concentrado en ese rango.
* Los productos representan entre el 0.5% y el 0.2% de la base de datos, lo que significa que la mayoría tiene un peso relativamente bajo en el total de registros.
* El 0.1% de los datos muestra valores atípicos posiblemente causados por errores de digitación, lo que podría generar ruido en el modelo. Es necesario realizar una limpieza de datos para corregir estos valores erróneos antes de entrenar el modelo.

Implicaciones para el Modelo:

* Dada la dispersión de precios, es recomendable aplicar transformaciones estadísticas (como escalado logarítmico) para mejorar la estabilidad del modelo.
* Se debe implementar una detección y corrección de outliers, asegurando que valores digitados incorrectamente no afecten las predicciones.
* La segmentación de productos según rangos de precios podría ayudar a mejorar la precisión del modelo y evitar sesgos en productos con menor representación.

Distribución de Productos

* La base de datos contiene una gran diversidad de productos, cada uno representando entre 0.5% y 1% del total de registros.
* No hay un producto dominante, lo que indica una distribución homogénea en la cantidad de registros por producto.
* Esto sugiere que el modelo deberá manejar una amplia gama de productos y considerar enfoques que permitan capturar tendencias individuales o agrupar productos con comportamientos similares.

Distribución por Grupos Agropecuarios

Los productos están clasificados en 8 grupos principales, con la siguiente distribución:

**Tabla 2**

*Distribución por Grupos Agropecuarios*

|  |  |
| --- | --- |
| **Grupo** | **Porcentaje de Datos** |
| **Frutas** | 25% |
| **Verduras y Hortalizas** | 20% |
| **Carnes** | 16% |
| **Procesados** | 15% |
| **Tubérculos, Raíces y Plátanos** | 8% |
| **Granos y Cereales** | 6% |
| **Pescado** | 5% |
| **Lácteos y Huevos** | 4% |

Interpretación:

* Frutas y verduras dominan el mercado con un 45% de los datos, lo que indica que su comportamiento influirá fuertemente en las tendencias de precios generales.
* Carnes y procesados representan otro 31%, lo que los convierte en grupos de gran relevancia para la predicción.
* Grupos con menor presencia (granos, pescados, lácteos y tubérculos) pueden requerir estrategias específicas en el modelo para mejorar su precisión.

Factores Temporales: Mes y Año

El análisis de estas variables permite identificar tendencias estacionales:

* Año: Se observa un crecimiento en la cantidad de registros en los años más recientes, lo que puede indicar una mayor disponibilidad de datos con el tiempo.
* Mes: La distribución de datos es uniforme, lo que sugiere que el comportamiento de los precios puede estar influenciado por estacionalidad y ciclos productivos de cada grupo agropecuario.
* Será clave analizar la variación de precios mes a mes para detectar patrones estacionales que puedan mejorar la precisión del modelo.

Factores Geográficos: Ciudad, Departamento y Región

* La base de datos contiene información distribuida en varias ciudades y departamentos, con algunas regiones concentrando un mayor porcentaje de registros.
* Las diferencias geográficas pueden influir en los precios debido a factores como acceso a mercados, costos de transporte y condiciones climáticas.
* Se observa que ciertas ciudades tienen una mayor representación en la base de datos, lo que podría sesgar las predicciones si no se equilibran correctamente.

Unificación y estandarización

* 1. Inicialmente, los datos estaban en múltiples archivos de Excel que estaban determinados por años. Se descartaron algunos y se estableció un formato único para que fuera más fácil encontrar las variables, dar estructura y orden.
  2. Se eliminaron columnas innecesarias, algunas vacías y otras que tenían datos no tan relevantes como las plazas de mercado y se homogeneizaron los títulos.
  3. Dos archivos de Excel (2022 y 2023) fueron eliminados ya que estaban mal digitados y generaban una mala interpretación en el modelo.

Limpieza de datos

1. Se estandarizó el formato de cada documento, lo que incluye la eliminación de columnas y homogeneización de títulos.
2. Se dividieron columnas que contenían dos variables juntas, como por ejemplo ciudades junto con departamento y las fechas, haciendo separación de mes y año.
3. Descarte de base de datos con errores en la digitación de precios, eliminación de dos archivos de Excel, correspondientes a los años 2022 y 2023
4. Se transformaron variables numéricas en categóricas mediante técnicas de codificación (los meses se convirtieron en palabras.
5. Eliminación de tildes y conversión de palabras a mayúsculas, por ejemplo, en la unificación de la variable grupo, “(FRUTA), (fruta) se dejó solo la que estaba en mayúscula”
6. Se completó la variable departamento y se creó la variable región para estructurar mejor la información y zonificar.

Modelado de Datos

Para este proyecto se seleccionó el modelo LightGBM (Light Gradient Boosting Machine). Según la documentación oficial de LightGBM (s.f.) LightGBM es un algoritmo de Machine Learning desarrollado por Microsoft que utiliza la técnica de Gradient Boosting para mejorar la precisión de los modelos predictivos. Se basa en la combinación de múltiples árboles de decisión débiles en un modelo más fuerte, optimizando el aprendizaje iterativo.

LightGBM es rápido, eficiente en memoria, maneja datos categóricos sin preprocesamiento adicional y ofrece alta precisión gracias a su estrategia de crecimiento “leaf-wise”. Se usa en aplicaciones como predicción de precios, detección de fraudes, sistemas de recomendación, diagnóstico médico y análisis de sentimientos. (LightGBM Documentation, s.f.)

El conjunto de modelos que FLAML podría probar para encontrar el mejor ajuste para tu conjunto de datos incluye principalmente los modelos basados en árboles (XGBoost, LightGBM, CatBoost, Random Forest), así como modelos lineales, KNN, redes neuronales y otros métodos de ensamblaje.

¿Qué es Gradient Boosting?

Según DeepAI (s.f.), Gradient Boosting es una técnica de aprendizaje supervisado utilizada para mejorar la precisión de modelos de predicción combinando múltiples árboles de decisión débiles en un modelo más fuerte. Se basa en la idea de corregir los errores de modelos anteriores mediante la optimización iterativa de sus predicciones. LightGBM es una implementación optimizada de Gradient Boosting, lo que le permite entrenar modelos de manera más rápida y eficiente con grandes volúmenes de datos.

¿Quién usa LightGBM?

LightGBM es utilizado por diversas empresas y organizaciones en el mundo, incluyendo:

* Microsoft: Para tareas de predicción y optimización de servicios.
* Alibaba: En sistemas de recomendación y detección de fraudes. (Alibaba Cloud, 2019)
* Uber: Para optimizar rutas y estimar tiempos de llegada. (Uber Engineering, 2019)
* Google: En motores de búsqueda y clasificación de anuncios.
* Investigadores y científicos de datos: En proyectos de predicción y clasificación.

Ejemplo Práctico: Uso de LightGBM en Finanzas

Towards Data Science (s.f.) explica que un caso de uso importante de LightGBM ha sido en el mercado financiero para la predicción de precios de acciones. Empresas como JP Morgan y Goldman Sachs han utilizado modelos de boosting, incluyendo LightGBM, para analizar grandes volúmenes de datos de mercado y hacer predicciones sobre la volatilidad de activos financieros.

Por ejemplo, un modelo de LightGBM puede ser entrenado con datos históricos de acciones, tasas de interés y noticias económicas para prever movimientos del mercado y tomar decisiones de inversión automatizadas con alta precisión.

LightGBM se ha consolidado como una herramienta poderosa para el análisis predictivo en diversos sectores, destacándose por su eficiencia, velocidad y precisión. Su aplicabilidad en problemas reales sigue creciendo, siendo una elección clave en la inteligencia artificial moderna.

Preparación de los datos

Se prepara los datos para que un modelo de Machine Learning pueda aprender a predecir el PRECIO\_KG.

1-Se limpian los datos eliminando valores nulos.

2- en nuestro caso, Se separa la variable objetivo (PRECIO\_KG).

3-Se usa HyperTransformer para transformar los datos automáticamente.

4- Se comparan los datos antes y después de la transformación.

5-Se dividen los datos en entrenamiento (80%) y prueba (20%).

Encontrar el mejor modelo

Utilizamos **FLAML (Fast and Lightweight AutoML)** para resolver un problema de **regresión**, donde se busca encontrar el mejor modelo predictivo para los datos de entrenamiento, es una herramienta de AutoML (Automated Machine Learning), que permite automatizar el proceso de selección, entrenamiento y ajuste de modelos.

Descripción de algunas de las herramientas utilizadas por FLAML:

Según Chi Wang, Mingwei Chen y Jianfeng Li (2021), FLAML es una biblioteca de AutoML diseñada para producir modelos precisos con bajo costo computacional, automatizando tanto la selección de algoritmos como el ajuste de hiperparámetros y explica algunos de ellos:

X\_train: Son las características (variables independientes) del conjunto de entrenamiento.

y\_train: Son las etiquetas o valores dependientes (lo que queremos predecir).

task="regression": Se indica que el problema es de regresión (es decir, que la variable de salida es continua, no categórica).

time\_budget=300: FLAML tiene un límite de 300 segundos (5 minutos) para encontrar el mejor modelo. Este valor se puede ajustar según el tiempo que quieras invertir.

metric="r2": La métrica de evaluación que FLAML usará para evaluar los modelos. En este caso, se usa el coeficiente de determinación R², que mide qué tan bien el modelo se ajusta a los datos. Se pueden utilizar otras métricas como RMSE o MAE dependiendo de tus necesidades.

verbose=1: Establece el nivel de detalles que se imprimen durante el proceso de entrenamiento.

FLAML entrena y buscar el mejor modelo, es el modelo con el mejor rendimiento según la métrica definida, En nuestro caso **LGNM**, con el modelo optimizado, ahora se puede hacer predicciones sobre el conjunto de prueba X\_test. Estas predicciones se almacenan en y\_pred.

Resultados

Evaluación del rendimiento del modelo

Se importan las funciones de sklearn.metrics:

* **r2\_score:** Mide qué tan bien el modelo se ajusta a los datos. (Montgomery, Peck, & Vining, 2012).
* **mean\_squared\_error:** Calcula el error cuadrático medio. (James, Witten, Hastie, & Tibshirani, 2013)
* **mean\_absolute\_error:** Mide la diferencia absoluta promedio entre los valores reales y predichos. (James, Witten, Hastie, & Tibshirani, 2013)

El Coeficiente de Determinación R²

Mide qué porcentaje de la variabilidad de los datos es explicada por el modelo.

Fórmula del R² Score:

Donde:

son los valores verdaderos (de y\_test).

son las predicciones del modelo (y\_pred).

es el valor promedio de los datos reales.

Interpretación de R²:

R² = 1 → Predicciones perfectas.

R² ≈ 0 → El modelo no explica nada (es como predecir el promedio).

R² < 0 → El modelo es peor que predecir el promedio (¡muy malo!). (Montgomery, Peck, & Vining, 2012).

RMSE (Root Mean Squared Error)

Mide el error promedio en las predicciones en la misma unidad que los datos.

Fórmula del RMSE:

Donde n es el número total de observaciones.

Interpretación de RMSE:

Valor bajo → Mejor modelo (menor error).

Valor alto → Peor modelo (mayor error). (James, Witten, Hastie, & Tibshirani, 2013)

Error absoluto medio (MAE)

Mide la diferencia absoluta promedio entre los valores reales y predichos.Menos sensible a valores extremos en comparación con RMSE.

Entre menor sea, mejor es el modelo.

Fórmula del MAE:

(James, Witten, Hastie, & Tibshirani, 2013)

Nuestros resultados

* **R² Score: 0.9797579657060155**
* **RMSE: 1319625.4715020116**
* **MAE: 695.3894**

El modelo predictivo desarrollado con LightGBM obtuvo un coeficiente de determinación R² de 0.9797, lo cual indica que el 97.97% de la variabilidad del precio por kilogramo de los productos agropecuarios es explicada por las variables seleccionadas en el modelo. Este valor se encuentra muy próximo a 1, lo que evidencia un excelente ajuste entre las predicciones generadas por el modelo y los datos reales observados.

El RMSE es una medida que calcula la raíz cuadrada del promedio de los errores al cuadrado entre los valores reales y los predichos. Esta métrica penaliza con mayor severidad los errores grandes, por lo que es especialmente útil para detectar desviaciones extremas. En este caso, el valor de RMSE = 1.319.625,47 indica que, en promedio, las predicciones del modelo se desvían de los valores reales en esa magnitud.

Un RMSE más bajo representa un modelo más preciso. Dado que este modelo fue aplicado a un conjunto de precios agropecuarios donde algunos valores superan los $99.000 por kilogramo, este resultado puede considerarse inadecuado en relación con la escala del problema y la dispersión natural de los precios.

Por su parte, el MAE calcula la media de los errores absolutos entre los valores reales y los predichos, sin elevarlos al cuadrado. Este valor es más robusto frente a los valores atípicos, lo que lo convierte en una métrica confiable para evaluar el error promedio real del modelo. En este estudio, el MAE fue de 695,39, lo que sugiere que, en promedio, las predicciones del modelo difieren de los precios reales en menos de $700 por kilogramo.

Hiperparámetros

Los **mejores hiperparámetros** del modelo seleccionado por **FLAML** después del proceso de optimización automática, se muestran a continuación, además los desarrolladores de LightGBM (s.f.-b) recomiendan y describen las técnicas específicas para el ajuste de hiperparámetros, con el fin de lograr un balance entre rendimiento y sobreajuste:

n\_estimators: 366. Número de árboles en el modelo. Más árboles pueden mejorar la precisión, pero aumentan el tiempo de entrenamiento y el riesgo de sobreajuste. En este caso, 366 árboles fueron suficientes para lograr un buen rendimiento.

*num\_leaves: 8.* Cantidad máxima de hojas por árbol de decisión. Un valor bajo genera modelos más simples y evita el sobreajuste, pero puede reducir la capacidad de aprendizaje. Aquí, 8 hojas por árbol indican un modelo relativamente simple.

*min\_child\_samples: 2*. Cantidad mínima de muestras que un nodo debe tener para dividirse. Valores bajos hacen que el modelo sea más flexible (posible sobreajuste). Aquí, con 2 muestras, el modelo está permitiendo divisiones en nodos con muy pocos datos, lo que indica un modelo bastante detallado.

*learning\_rate: 0.3519*. Controla cuánto se ajustan los pesos en cada iteración. Valores bajos (< 0.1) hacen el modelo más preciso, pero más lento.0.35 es relativamente alto, lo que sugiere que el modelo aprende rápido, pero con el riesgo de no converger bien.

*log\_max\_bin: 10.* Se relaciona con el número de bins usados para dividir las características continuas. LightGBM usa histogramas para acelerar el entrenamiento. Un valor mayor puede permitir capturar más detalles en los datos.

*colsample\_bytree: 0.8646.* Porcentaje de características usadas en cada árbol. Aquí, el modelo usa el 86.5% de las características en cada árbol, lo que ayuda a reducir el sobreajuste mientras mantiene una buena capacidad predictiva.

*reg\_alpha: 0.0045 (L1 regularization).* Regularización L1: penaliza valores altos en los coeficientes para mejorar la generalización. Un valor bajo significa que apenas está afectando el modelo.

*reg\_lambda: 0.0971 (L2 regularization).* Regularización L2: evita que los coeficientes sean demasiado grandes y reduce la varianza. 0.097 es un valor bajo, lo que indica que el modelo no necesita mucha regularización.

Interpretación general

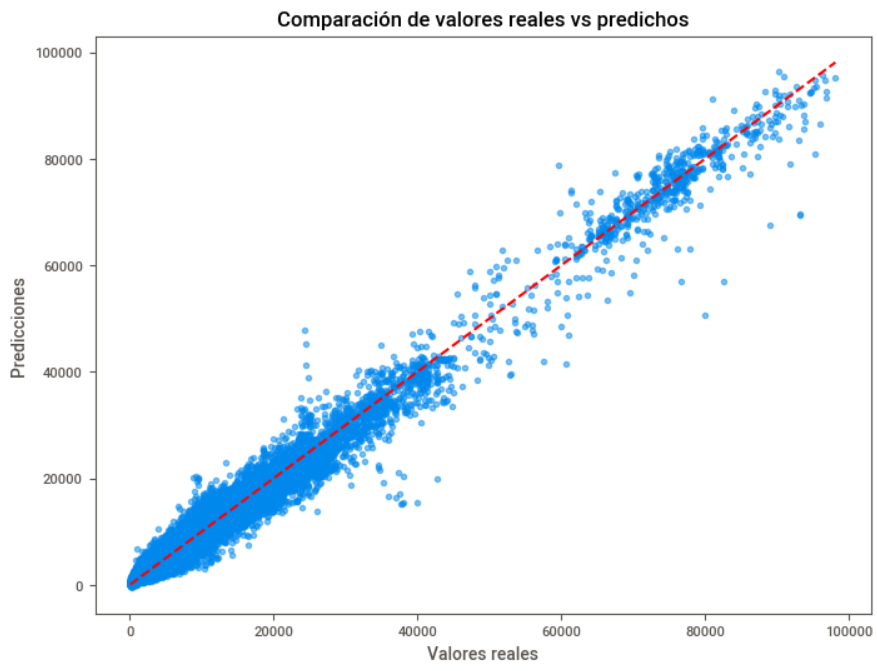
* El modelo está configurado para aprender rápidamente (learning\_rate: 0.35) y construir árboles relativamente pequeños (num\_leaves: 8) con pocos datos en cada nodo (min\_child\_samples: 2).
* El número de árboles (n\_estimators: 366) sugiere que el modelo busca un equilibrio entre complejidad y rendimiento.
* Las regularizaciones (reg\_alpha y reg\_lambda) son bajas, lo que indica que el modelo no necesita mucha penalización para evitar el sobreajuste.
* Se usa el 86.5% de las características en cada árbol (colsample\_bytree), lo que ayuda a la diversidad de los árboles sin perder demasiada información.

Comparación de valores reales vs predichos

Se crea un gráfico de dispersión, donde los valores en x representan los valores reales, y los valores en Y representan los valores predichos por el modelo. En la siguiente imagen podemos visualizar como se traza una línea roja discontinua (--) que representa la predicción perfecta, si el modelo fuera perfecto, todos los puntos deberían estar sobre esta línea, en cambio, si los puntos están muy dispersos, el modelo no está haciendo buenas predicciones

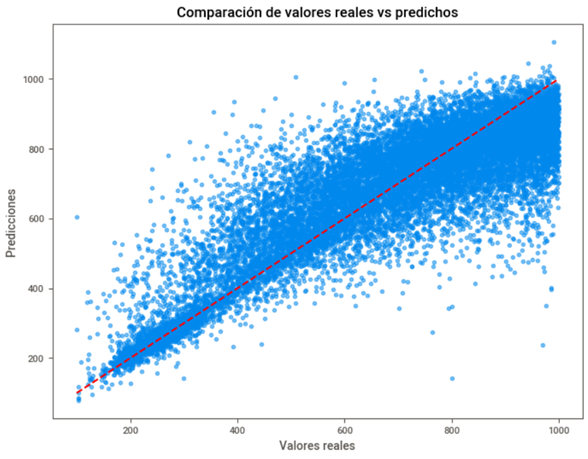
**Figura 5**

*Comparación de valores reales vs predichos*

****

**Figura 6**

*Comparación de valores reales vs predichos (sin eliminación de archivos)*



Nota: Esta gráfica corresponde a la comparación de valores reales vs predichos contemplando los archivos correspondientes a los años 2022 y 2023, los cuales fueron eliminados posteriormente.

Como puede observarse en la figura 6 la gráfica presenta mayor dispersión y los datos no siguen la tendencia de la línea de referencia, tal como lo hace la distribución de los datos en la figura 5, lo cual determina que el modelo tuvo un ajuste considerablemente mayor al eliminar dichos archivos.

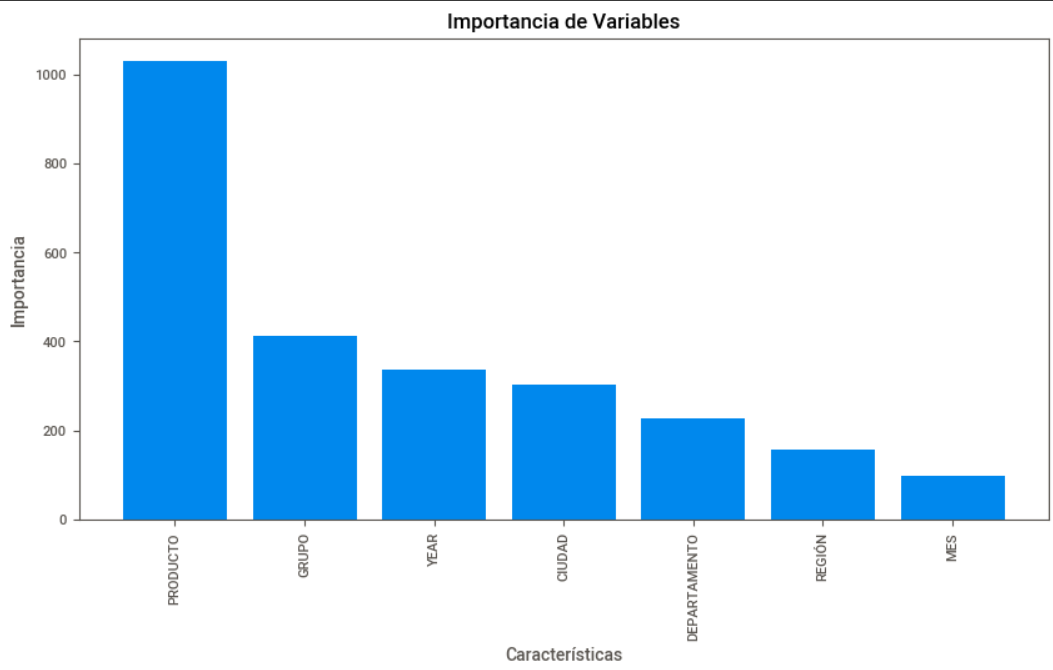
Importancia de las variables

Se usa numpy para manejar y ordenar la importancia de las características, genera un gráfico de barras para visualizar la importancia de las variables en el modelo entrenado.

Un valor más alto significa que la característica es más relevante para hacer predicciones.

**Figura 7**

*Importancia de Variables*



Como puede observarse en la gráfica, la característica con mayor importancia en la predicción es “PRODUCTO”, y como es de esperarse una de las que poseen menor importancia es “REGIÓN” ya que esta categoría fue creada posteriormente tomando de referencia la categoría “CIUDAD”

Conclusiones

**Efectividad del modelo**: LightGBM demostró ser una herramienta eficaz para la predicción de precios agropecuarios en Colombia, logrando un buen ajuste en los datos analizados.

**Importancia de la limpieza y depuración de datos:** La calidad del modelo estuvo directamente relacionada con la limpieza de datos, lo que permitió mejorar la precisión de las predicciones al eliminar valores nulos, duplicados y estructurar mejor la información.

**Variables clave en la predicción:** Se identificó que las variables más influyentes en el modelo fueron el precio, el tipo de producto y su grupo, además de factores geográficos y temporales.

**Técnicas de transformación y selección de datos:** La aplicación de técnicas como Label Encoding y normalización mejoró el desempeño del modelo, mientras que la selección adecuada de datos ayudó a reducir sesgos y optimizar la capacidad predictiva.

**Oportunidades de mejora y expansión:** Aunque el modelo tiene un buen ajuste, métricas como el error cuadrático medio y el error absoluto medio sugieren la necesidad de una revisión más profunda de la calidad de los datos. Además, existe potencial para incorporar nuevas variables y mejorar la adaptación del modelo a diferentes escenarios económicos y climáticos.

Líneas Futuras

Mejoras para el Modelo Predictivo

* *Incorporación de Variables Externas (Datos Exógenos):* Además de los precios históricos de los productos agropecuarios, puedes incorporar datos exógenos que puedan influir en los precios, como **condiciones climáticas**, **política gubernamental** (subsidios, cambios en los aranceles), **eventos internacionales** (tendencias globales), o **tipo de cambio**. Estos factores son cruciales para el mercado agropecuario y pueden mejorar la precisión del modelo.
* *Ampliación del Conjunto de Datos:* Si la base de datos es limitada en tiempo o cantidad de productos, sería ideal ampliarla, incorporando más productos agropecuarios, más años de datos, o más fuentes (mercados locales, bases de datos gubernamentales, datos privados). Una mayor variedad de datos puede ayudar a entrenar modelos más robustos.
* *Optimización de Algoritmos:*
* **Ajuste de Hiperparámetros:** Bergstra y Bengio (2012) proponen técnicas como **Grid Search** o **Random Search** para optimizar los hiperparámetros de los algoritmos de machine learning que estás utilizando.
* **Modelos más complejos:** Si estás utilizando un modelo lineal, podrías probar modelos más complejos como **Random Forest (Breiman, 2001)**, **XGBoost (Chen & Guestrin, 2016)**, o **Redes Neuronales Recurrentes (RNN)** (Rumelhart, Hinton, & Williams, 1986) para series temporales. Los **LSTM (Long Short-Term Memory) (Hochreiter & Schmidhuber, 1997)** también son buenos para predecir series temporales con dependencias a largo plazo.
* *Validación y Regularización:* Kohavi (1995) destaca la validación cruzada **(cross-validation)** como una técnica efectiva para evaluar el desempeño de los modelos sin sesgos significativos.Además, Tibshirani (1996) introdujo el método Lasso como una forma de regularización que realiza selección de variables y mejora la predicción al imponer una penalización L1. Hoerl y Kennard (1970) propusieron la regresión Ridge para reducir la multicolinealidad y mejorar la estabilidad de los modelos lineales. Así que se implementa **regularización** (como Lasso o Ridge) para evitar el sobreajuste (overfitting) si estás trabajando con un conjunto de datos pequeño.
* *Detección de Estacionalidad:* Dado que los precios de los productos agropecuarios tienden a ser estacionales, asegúrate de que el modelo esté capturando patrones estacionales. Puedes incluir variables indicadoras de las estaciones o usar modelos como **ARIMA (Box & Jenkins, 1976)**  o **SARIMA** (Hyndman & Athanasopoulos, 2018) que son específicos para series temporales.
* *Incorporación de Técnicas de Ensemble:* Dietterich (2000) señala que las técnicas de ensemble como el **bagging** y **boosting**, permiten mejorar la precisión y estabilidad del modelo al combinar múltiples modelos predictivos. Esto podría dar lugar a mejores resultados en la predicción de precios.
* *Automatización del Modelo:* Implementa un sistema automatizado que permita el entrenamiento periódico del modelo con datos nuevos, ajustando continuamente los parámetros a medida que se obtienen nuevos datos, para mantener su precisión.
* *Interpretabilidad del Modelo:* Es crucial que puedas interpretar los resultados del modelo y explicar las predicciones. Usa técnicas como **SHAP** o **LIME** para interpretar modelos complejos y hacer que las predicciones sean más comprensibles para los stakeholders. Lundberg y Lee (2017) introdujeron SHAP como un método unificado basado en teoría de juegos para explicar las salidas de cualquier modelo de aprendizaje automático. Ribeiro, Singh y Guestrin (2016) propusieron LIME como una técnica para explicar predicciones individuales de modelos complejos mediante aproximaciones locales interpretables.
* *Modelos Multi-objetivo:* Si tu modelo solo predice precios futuros, podrías ampliarlo a un enfoque multi-objetivo que no solo prediga precios, sino también factores como el **riesgo** asociado, **la oferta y la demanda** en diferentes zonas, o incluso la **sostenibilidad** de los precios.

Posibles Usos Futuros del Modelo

* *Optimización de Precios en Tiempo Real:* El modelo predictivo podría ayudar a los agricultores, distribuidores y comerciantes a **optimizar precios en tiempo real**, basándose en las predicciones del modelo para maximizar los márgenes de ganancia y evitar pérdidas por fluctuaciones inesperadas.
* *Sistemas de Recomendación:* Se puede desarrollar un sistema de recomendación basado en el precio y las predicciones de demanda para sugerir a los agricultores qué productos cultivar, cuándo vender, o cómo diversificar su oferta de productos según las tendencias del mercado.
* *Planificación y Gestión de Inventarios:* Puedes usar el modelo para ayudar a los distribuidores y minoristas a gestionar mejor sus **inventarios**, previendo cuándo los precios estarán más altos o bajos y ajustando sus compras y ventas en consecuencia.
* *Análisis de Riesgos para Agricultores:* Además de predecir precios, el modelo podría integrarse con otros sistemas de gestión para identificar posibles **riesgos** asociados con la volatilidad de los precios y ayudar a los agricultores a tomar decisiones informadas sobre seguros y coberturas frente a riesgos económicos.
* *Modelos de Predicción para Mercado de Exportaciones:* La herramienta podría adaptarse para predecir precios de productos destinados a la exportación, ayudando a identificar oportunidades o amenazas en los mercados internacionales.
* *Desarrollo de Políticas Públicas:* El modelo puede ser utilizado por el gobierno o por organizaciones no gubernamentales para **desarrollar políticas públicas** que protejan a los agricultores de las fluctuaciones de precios o para implementar subsidios de manera más eficiente.
* *Determinación de Costos de Producción:* Puedes aplicar el modelo para analizar la relación entre los precios y los costos de producción, ayudando a los productores a ajustar sus costos y mejorar la rentabilidad.
* *Predicción de Demanda Regional:* Utilizando datos históricos de precios y otros factores, podrías predecir no solo los precios futuros, sino también las **demandas regionales**, lo cual sería útil para orientar la producción hacia áreas donde se prevea mayor necesidad de ciertos productos.
* *Desarrollo de Productos Financieros:* Con la capacidad de prever precios y riesgos, podrías colaborar en el diseño de **productos financieros** como contratos de futuros, seguros agrícolas, o derivados para proteger a los agricultores de la volatilidad en los precios.
* *Integración con Internet de las Cosas (IoT):* Integrando este modelo con sensores IoT que monitorean las condiciones del clima o el estado de los cultivos en tiempo real, podrías crear un sistema predictivo que permita tomar decisiones proactivas sobre precios y cosechas.

Referencias

Alibaba Cloud. (2019). How Alibaba uses machine learning to enhance its e-commerce platform. <https://www.alibabacloud.com/blog/how-alibaba-uses-machine-learning-to-enhance-its-e-commerce-platform_594430>

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305. <https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf>

Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control. Holden-Day.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>

DeepAI. (s.f.). Gradient boosting definition. <https://deepai.org/machine-learning-glossary-and-terms/gradient-boosting>

Dietterich, T. G. (2000). Ensemble methods in machine learning. En Multiple Classifier Systems (pp. 1–15). Springer. <https://doi.org/10.1007/3-540-45014-9_1>

FreeCodeCamp. (2021). Tipos de datos en estadística: Tipos de datos nominales, ordinales, de intervalo y de razón explicados con ejemplos. <https://www.freecodecamp.org/espanol/news/tipos-de-datos-en-estadistica-tipos-de-datos-nominales-ordinales-de-intervalo-y-de-razon-explicados-con-ejemplos/>

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. <https://doi.org/10.1080/00401706.1970.10488634>

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2nd ed.). OTexts. <https://otexts.com/fpp2/>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R (Vol. 112). Springer. <https://doi.org/10.1007/978-1-4614-7138-7>

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. En Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 (pp. 1137–1143). Morgan Kaufmann Publishers Inc.

LightGBM Developers. (s.f.-b). Parameters tuning. LightGBM Documentation. <https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html>

LightGBM Documentation. (s.f.). Welcome to LightGBM’s documentation! <https://lightgbm.readthedocs.io/>

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. En Advances in Neural Information Processing Systems (Vol. 30). <https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (5th ed.). Wiley. <https://www.wiley.com/en-us/Introduction+to+Linear+Regression+Analysis%2C+5th+Edition-p-9780470542811>

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. En Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. <https://doi.org/10.1145/2939672.2939778>

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536. <https://doi.org/10.1038/323533a0>

ScienceDirect Topics. (s.f.). Machine learning – An overview. <https://www.sciencedirect.com/topics/computer-science/machine-learning>

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>

Towards Data Science. (s.f.). All you need to know about Gradient Boosting Algorithm – Part 1 (Regression). Medium. <https://medium.com/towards-data-science/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502>

Uber Engineering. (2019). Michelangelo: Uber’s machine learning platform. <https://eng.uber.com/michelangelo-machine-learning-platform>

Wang, C., Chen, M., & Li, J. (2021). FLAML: A fast and lightweight AutoML library. Microsoft Research. <https://www.microsoft.com/en-us/research/publication/flaml-a-fast-and-lightweight-automl-library/>

Anexos

**Anexo A**

**Base de Datos**

[Microdatos DANE - Precios Agropecuarios](https://microdatos.dane.gov.co/index.php/catalog/776/get-microdata) Contiene los datos utilizados en el estudio sobre precios agropecuarios en Colombia.

**Anexo B**

**Diccionario de Datos:**

[Microdatos DANE - Diccionario de Datos](https://microdatos.dane.gov.co/index.php/catalog/776/data-dictionary) Explica el significado de cada variable dentro de la base de datos.

**Anexo C**

**Material Relacionado:**

[Microdatos DANE - Material Relacionado](https://microdatos.dane.gov.co/index.php/catalog/776/related-materials) Documentos adicionales sobre la recopilación y procesamiento de datos.

**Anexo D**

**Descripción Estadística:**

[Microdatos DANE - Descripción del Estudio](https://microdatos.dane.gov.co/index.php/catalog/776/study-description)  Explica la metodología, alcance y estructura de los datos recopilados