Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

MAESTRÍA EN ECONOMÍA APLICADA

TRABAJO FINAL DE MAESTRÍA

Aplicación del web scrapping como herramienta para la medición y proyección del Índice de Precios al Consumidor en Bolivia

Employing web scraping for measuring and projecting the Consumer Price Index in Bolivia

AUTOR: OSCAR CUENTAS SANDY

DIRECTOR: GABRIEL MONTES ROJAS

SEPTIEMBRE 2024

Dedicatoria

A Dios, por acompañarme y guiarme a lo largo de todo este proceso.
A mi madre, por impulsarme a perseguir mis sueños, apoyarme en los momentos de incertidumbre y estar siempre a mi lado.
A mi hermana, por ser mi motivación diaria para ser una mejor persona.
A mi padre, mis abuelos y hermano, por su apoyo incondicional.

Agradecimientos

A la Universidad de Buenos Aires, por permitirme formar parte de su prestigiosa institución.
A mis profesores, por sus valiosas enseñanzas.
A Gabriel, por sus consejos y guía invaluables durante el desarrollo del trabajo.
A mis compañeros, por la colaboración y apoyo durante toda la cursada.
A Pablo, por compartir sus conocimientos y motivarme en la elaboración de este trabajo.

Resumen

Este trabajo tiene como objetivo demostrar que la aplicación de las nuevas técnicas tecnologías para el procesamiento de datos como lo es el web scrapping son una gran ayuda para la medición de indicadores económicos como el Índice de Precios al Consumidor, para ello se toma como muestra un país en vías de desarrollo con una economía precarizada como lo es Bolivia. Durante el trabajo se fueron resolviendo diferentes obstáculos presentados tanto por el contexto socioeconómico del país como en la programación de los códigos para la extracción de los datos en el lenguaje de programación de Python en su entorno de Google Colab.

En la actual situación angustiante que atraviesa la economía boliviana donde después de muchos años con los niveles más bajos de inflación a nivel mundial la tendencia parece cambiar radicalmente con un incremento sostenido de los precios, el uso de nuevas técnicas que faciliten el acceso a la información en frecuencia semanal permite una mayor transparencia en los hogares bolivianos que necesitan conocer estos datos a tiempo real para poder planificar sus diferentes egresos futuros.

El trabajo cumple con el objetivo propuesto y va en línea con los resultados de investigaciones anteriores relacionadas a la temática, se logra superar los obstáculos presentados y la recolección de datos realizada por el web scrapping presenta las mismas tendencias que las reportadas por el Instituto Nacional de Estadística, además de poder anticiparse a futuras tendencias inflacionarias realizando pronósticos a corto plazo utilizando modelos de aprendizaje automático.

Palabras clave: E31, E37, C82, C88

Abstract

This work aims to demonstrate that the application of new data processing technologies, such as web scraping, is highly beneficial for measuring economic indicators like the Consumer Price Index (CPI). The study focuses on a developing country with a precarious economy, specifically Bolivia. Throughout the project, various challenges were addressed, both those arising from the country's socioeconomic context and those encountered in programming the data extraction codes using Python in the Google Colab environment.

In the current distressing situation facing the Bolivian economy, where, after many years of having one of the lowest inflation rates worldwide, the trend seems to be changing drastically with a sustained increase in prices, the use of new techniques that facilitate access to information on a weekly basis provides greater transparency for Bolivian households. These households need to know this information in real time to plan their future expenses.

The work meets its proposed objectives and aligns with the results of previous research on the subject. The obstacles encountered were successfully overcome, and the data collection through web scraping reflects the same trends reported by the National Institute of Statistics. Additionally, it offers the capability to anticipate future inflationary trends by making short-term forecasts using machine learning models

Keywords: E31, E37, C82, C88

I. Introducción

La medición del Índice de Precios al Consumidor (IPC) tiene un rol fundamental en todos los países, conocer cómo se encuentra la variación de precios de los diferentes bienes y servicios de una economía permite a los gobiernos reformular e implementar diferentes políticas públicas que beneficien al Estado y la sociedad en su conjunto, asimismo este valor publicado mensualmente por los institutos de estadística nacionales le otorga a las familias la posibilidad de planificar y organizar su futuro a corto y mediano plazo sobre como podrá administrar sus ingresos, ya sea que los destine al consumo o al ahorro. De igual forma las empresas requieren conocer el valor de este indicador para realizar diferentes proyecciones sobre los futuros ingresos y egresos que podrían llegar a suceder, además de también tomar decisiones financieras respecto a su capital.

Los países con alta inflación padecen de problemas como la pobreza, la baja inversión, la disminución del consumo y la imposibilidad de poder importar bienes y servicios del extranjero, entre otros, por ello la detección temprana de un incremento sostenido del indicador es fundamental para preservar la estabilidad de la sociedad y prevenir lo que se conoce como una de las peores crisis socioeconómicas denominada comúnmente como hiperinflación.

En la historia moderna se reconoce al menos cincuenta y ocho episodios de hiperinflación en el mundo, de los cuales Bolivia aparece entre los países que sufrieron una de las más profundas en los períodos comprendidos de 1982 a 1985, este terrible episodio vivido hace ya más de cuatro décadas dejo un trauma en la ciudadanía que como primer refugio y temor a vivir un evento de las mismas proporciones acudieron en los años posteriores al uso masivo del dólar como reserva de valor, esta acción se la identifico como una dolarización de facto que imponía un esquema de bimonetarismo implícito entre la moneda doméstica (el boliviano) y la moneda extranjera (el dólar). Si bien la ciudadanía había logrado preservar su poder adquisitivo, el mismo iba en desmedro de la economía en su conjunto afectando las inversiones y la producción nacional por la desconfianza existente en el boliviano.

Esta constante incertidumbre no permitió que la economía boliviana tuviera un crecimiento significativo en los años posteriores a diferencia de sus vecinos que posterior a eventos similares iban despegando, dejando replegado al país ubicado en el centro de Sudamérica. Es por lo que en las elecciones nacionales del 2005 la ciudadanía daba su voto de confianza en un partido diferente al poder político tradicional que venía con la promesa de terminar la incertidumbre y de impulsar la “bolivianización” con la preferencia de la

moneda doméstica por encima de la divisa extranjera, tras años de conflictos sociales recién en 2011 a partir de tomar una política de régimen cambiario fijo la inflación en Bolivia descendería a niveles inferiores a los históricos.

Esta situación permitió a la ciudadanía volver a confiar en el boliviano y por ende generar un crecimiento económico sostenido que era visto y admirado por la comunidad internacional donde siempre se destacaba el éxito de la inflación baja en comparación a las grandes economías que sufrían largos periodos inflacionarios.

Esta estabilidad económica y social existió por ocho años entre la gestión de 2011 hasta finales del 2019 donde por conflictos sociales se interrumpiría la gestión del gobierno con la renuncia del presidente, seguido posteriormente por la pandemia del coronavirus en 2020, ambos eventos fueron el principio de lo que ha sido una crisis socioeconómica que se vive hasta la actualidad en la cual los factores principales son la escasez de combustibles y dólares que han traído al presente los recuerdos de un pasado complejo que exacerba las expectativas en la ciudadanía.

En este contexto el tener una base de datos del IPC que refleje la realidad diaria de la variación de alimentos se hace una necesidad imperativa para la sociedad, sin embargo, los datos mensuales por el Instituto Nacional de Estadística (INE) no permiten este cometido. Ante ello, la aplicación de nuevas tecnologías puede ser la solución para devolver la transparencia a los bolivianos.

En línea con esta premisa, los avances tecnológicos experimentados en los últimos años con relación al manejo, procesamiento, recolección y análisis de datos a niveles masivos comúnmente conocidos como *Big data* han revolucionado la forma de trabajar en todos los campos académicos. Esta evolución en el ámbito computacional está proporcionalmente relacionada con la digitalización de la vida cotidiana, el uso del internet se ha convertido prácticamente en un insumo de primera necesidad para los individuos, hoy en día no se puede vivir sin tener un smartphone o laptop para interactuar y relacionarse con el resto del mundo, las redes sociales, los juegos en línea y la posibilidad de consumir bienes y servicios desde la comodidad del hogar permitieron cambiar los hábitos de vida cotidiana de una forma nunca antes vista en la historia.

Por estas razones, este trabajo explora el uso de nuevas tecnologías como lo es el web scrapping para la recolección de datos y su potencial aplicación para mejorar la vida de los bolivianos con una información más rápida con relación a los cambios en los precios de los diferentes bienes y servicios consumidos por la sociedad.

El presente trabajo presenta un primer apartado con literatura relacionada al tema, la misma aún se encuentre en un estado incipiente, pero trae consigo información valiosa sobre el procesamiento y recolección de *Big data* en el campo de diferentes variables económicas. Posteriormente en la sección de metodología se menciona la complejidad de las diferentes técnicas de recolección en economías aún muy precarizadas como lo es el caso de Bolivia, se explica cómo se resolvió estas dificultades para lograr los objetivos propuestos, a partir de esta explicación se presentan los resultados encontrados en el trabajo y finalmente se encuentra la sección de conclusiones con sus recomendaciones respectivas para futuros trabajos relacionados.

II. Literatura

La técnica del web scrapping en los últimos años ha logrado notoriedad por sus grandes beneficios en el procesamiento y recolección de datos en lenguajes de programación como lo son Python o R, está herramienta proveniente del campo computacional en un principio fue diseñada para almacenar grandes cantidades de información extraídas de los sitios webs solamente en el rubro de la programación de algoritmos y comparación de precios para el mejoramiento constante de las empresas, sin embargo, su amplia transversalidad despertó en los economistas la posibilidad de explorar esta técnica para el análisis de distintas variables macroeconómicas.

Esta evolución histórica entre el paso del web scrapping como herramienta de programación a una herramienta para economistas data de los últimos años, por lo tanto, en este contexto es importante explorar y comentar los pocos trabajos relacionados en la temática.

La revolución tecnológica dio el salto entre la década de los 90 y los comienzos del nuevo siglo, la aparición de computadoras cada vez con mayor capacidad de procesamiento y recolección de datos, además de la expansión masiva en el uso del internet fueron los factores que causaron el boom de la tecnología. Es por lo que los primeros estudios relacionados al tema datan de principios del siglo XXI, en primer lugar, se tiene el trabajo de (Morton, Zettelmeyer, & Silva-Risso, 2001), en este los autores hacen un estudio sobre el efecto del uso del internet en las recomendaciones del servicio de autos en el precio de un concesionario de automóviles en el estado de California-Estados Unidos. Las diferentes páginas webs que ofrecían este servicio reflejaban información detallada sobre los potenciales autos a elegir, donde incluían los precios de facturación en tiempo real por temporada y sus diferentes condiciones individuales para quienes visitasen el sitio web. Según los autores, los datos reflejados de los autos de forma online eran mucho más

completos y precisos que los reflejados de forma offline en folletos, trípticos o volantes que se utilizaban mayormente para conseguir clientes en los negocios de esa época. De esta forma en el trabajo los autores construyen una base de datos a partir de un sitio web especifico llamado Autobytel.com durante el año 1999, fue uno de los primeros estudiantes con una gran cantidad de observaciones, específicamente más de dos millones que incluían toda la información relacionada al cliente, el auto elegido, la fecha en que el pedido fue realizado, el negocio al cual la página enviaba la respectiva recomendación y el tiempo que transcurría en el que el cliente se interesaba por el vehículo recomendado. Entre las diferentes conclusiones de este trabajo, se destaca en como los servicios online empezaban a desplazar a diferentes metodologías tradicionales de venta en los comercios, incluso se aprecia como existía el proceso de destrucción creativa en esos años, cientos de personas que trabajan repartiendo la información de los negocios de manera tradicional en formatos físicos de papel pronto se verían reemplazadas por trabajadores expertos en el sector de la programación para hacer el mismo trabajo a un menor costo, por el hecho de que un solo programador a nivel nacional podría reemplazar a cientos de empleados distribuidos en todos los estados del respectivo país.

Posteriormente en el trabajo de (Baye & Morgan, 2004) se realiza por primera vez un estudio comparativo de precios internacionales, en este los autores se enfocan en el impacto que existió debido a la introducción del Euro en los precios online de los diferentes negocios o empresas minoristas, ellos observan los datos que se tenían registrados previo a la integración monetaria posterior adopción del Euro como moneda única y los datos posteriores, así como también separan la muestra en países integrantes y no integrantes de la Unión Europea, después de controlar por diferentes variables el trabajo demuestra que el impacto del Euro fue un aumento en el promedio de los precios en la Eurozona de entre un 3% t un 7%. Era el primer estudio que hacia un análisis comparativo de precios utilizando el web scrapping de la página de datos Kelkoo.

Si bien los beneficios de esta técnica se empezaron a ver a principios del siglo, los mismos eran únicamente relacionados con la posibilidad de conseguir almacenar información a gran escala de clientes o consumidores para el uso de las distintas empresas en mejorar sus productos y precios en el mercado, sería recién a finales de la década de los 2000 cuando el trabajo realizado por (Lünnemann & Wintr, 2006), donde se haría un estudio de la rigidez de precios entre Estados Unidos y Europa al investigar el comportamiento de los precios online. De esta forma, los autores determinaron que, al contrario de los datos provenientes del Índice de Precios del Consumidor, los precios online no cambian en la

misma proporción para países de Europa en comparación a Estados Unidos. En este trabajo, por primera vez se utiliza el web scrapping para hacer una comparativa con relación a la medición del IPC de manera tradicional (encuestas de hogares y de consumo) y los precios online observados, las conclusiones que se encuentran son interesantes por como se demuestra que el cambio promedio en los precios online, si bien es relativamente grande, es menor que aquellos reportados por los institutos nacionales.

Ya entre la década de 2010 a 2020 se tiene a (Cavallo & Rigobon, 2016) donde específicamente utilizan el *Big data* para mejorar estadísticas de variables económicas, en este caso nuevamente la del IPC. En el mismo los autores construyen una base de datos con precios online de diferentes países con una frecuencia diaria, de esta forma evitando los sesgos que se pueden llegar a distorsionar por la rigidez de los precios y sus relatividades internacionales. Aplicando la técnica del *scrapping* de precios de los diferentes productos de las empresas comercializadoras más significativas para cada región, replicando el IPC de cada miembro de la muestra seleccionada. De esta forma llegan a la conclusión de que el Índice construido con datos online es muy parecido al desarrollado por los Institutos de Estadística Nacionales de cada país con la metodología tradicional.

En el mismo sentido de medir un Índice de Precios al Consumo con datos online el trabajo de (Aparicio & Bertolotto, 2020), no solo demuestra los beneficios del web scrapping en este cometido, sino también demuestra que las bases de datos construidas con *Big data* son más eficientes y útiles para pronosticar tendencias inflacionarias futuras con hasta un mes de anticipación, mostrando una mayor precisión frente a los desarrollados por Bloomberg, sin duda este trabajo da un gran paso en la consolidación de las nuevas tecnologías para el análisis de variables económicas no solo en tiempo presente, sino también a futuro.

En esta línea un trabajo bastante interesante para la región latinoamericana es el de (Orlandi & Osovi Conti, 2018), donde los autores plantean un cambio en el paradigma de la recolección de variables económicas de manera tradicional por las nuevas técnicas brindadas por la tecnología como lo es el web scrapping. En este trabajo con la ayuda de la *Big data* se calcula paridad de poder de compra entre Argentina y diversos países latinoamericanos, si bien los autores logran construir su base de datos y hacer la respectiva comparación, ellos hacen una importante contribución en el ámbito de las diferentes limitaciones que tiene el uso de este tipo de técnicas en países con diversas complejidades como lo son los países latinoamericanos, el difícil acceso de información en el caso de países que son precarizados

o la complejidad de comparación entre regiones por sus diferentes costumbres representan un obstáculo para la aplicación del web scrapping en la región latinoamericana.

De igual forma (Llada & Aromí, 2023) utilizan la técnica para extraer información de la red social X (antes Twitter) con relación a las búsquedas de las palabras relacionadas a la inflación y el tipo de cambio, con ello logran construir un indicador del nivel de atención asignado a la inflación en las discusiones públicas, esta variable permite darles mayor robustez y precisión a los modelos de proyección de la inflación, demostrando de esta forma el potencial beneficio de incluir variables recolectadas por los sitios web para el tratamiento de diferentes variables macroeconómicas.

Con relación al avance existente en los Institutos Nacionales de Estadística, a nivel mundial los mismos ya se encuentran utilizando esta técnica para la recolección de información y pertinente medición de diferentes indicadores. En el caso del indicador relacionado a la temática de este trabajo, los diferentes institutos han encontrado la forma de combinar la forma de recolección tradicional mediante encuestas a los consumidores y empresas para conocer las variaciones del IPC, no obstante, han permitido apoyarse en la recolección de masivas cantidades de datos vía web scrapping para un mejor análisis de la información recibida y sobre todo un control de frecuencia diaria para advertir con anticipación posibles tendencias inflacionarias.

En el caso de la región, la pandemia del COVID-19 ha permitido a los Institutos Nacionales de Estadística latinoamericanos acelerar este proceso de integración a los avances tecnológicos con relación a sus pares, países como Brasil y Uruguay están cada vez más avanzados en esta temática, pero sin lugar a dudas el país que destaca por su gran avance y actual aplicación de web scrapping para la medición del IPC es Chile quien en la actualidad se encuentra recolectando de manera mixta los datos pertinentes, así como lo hacen las económicas avanzadas de Europa y América del Norte.

En el caso de Bolivia no existe ningún trabajo relacionado con el uso del web scrapping para la medición de índices de variables económicas, así como tampoco existen avances o intención de aplicar estas técnicas por el Instituto Nacional de Estadística boliviano, por ello este trabajo se presenta como el primero en su tipo.

III. Metodología

Construcción del Índice de Precios del Consumidor por el Instituto Nacional de Estadística

La metodología implementada en el trabajo sigue los pasos del Instituto Nacional de Estadística de Bolivia para evitar distorsiones o sesgos que puedan perjudicar los resultados,

para ello se divide el Índice de Precios del Consumidor en doce divisiones para clasificar los diferentes productos, las mismas son: Alimentos y Bebidas No Alcohólicas, Bebidas Alcohólicas y Tabaco; Prendas de Vestir y Calzados; Vivienda y Servicios Básicos; Muebles, Bienes y Servicios Domésticos; Salud; Transporte; Comunicaciones; Recreación y Cultura; Educación; Alimentos y Bebidas Consumidos Fuera del Hogar y Bienes y Servicios Diversos. Estas divisiones son similares en todas los Institutos de países vecinos, sin embargo, es importante mencionar que en la división 11 de Alimentos y Bebidas Consumidos Fuera del Hogar si existen algunas variaciones con relación a las medidas en esta categoría por otros países, en la misma se hace énfasis en los diferentes platos (desayuno, almuerzo, cena y platos a la carta) que se pueden llegar a consumir en restaurantes o mercados locales, omitiendo por completo los servicios de hotelería y alojamiento que si son medidos comúnmente y llama la atención que el INE de Bolivia no lo haga, perdiendo de esta forma un producto importante en la medición del índice. Estas doce divisiones presentan 77 grupos, 153 clases, 397 productos y 513 variedades.

En cuanto a la recolección de datos el Instituto Nacional de Estadística divide al país en seis ciudades capitales de los departamentos (Chuquisaca, Oruro, Potosí, Tarija, Beni y Pando), y concentra las ciudades más grandes del eje troncal en: Conurbación La Paz (Nuestra Señora de La Paz, El Alto, Viacha, Achocalla), la Región Metropolitana Kanata (Cercado, Quillacollo, Tiquipaya, Vinto, Sipe Sipe, Sacaba, Colcapirhua y la Conurbación Santa Cruz (Santa Cruz de la Sierra, Warnes, La Guardia, Cotoca). Para definir la participación de un producto o no dentro del índice los criterios de selección están basados en los datos recogidos por las encuestas de hogares donde se toma en cuenta la participación en el gasto, para que un producto pueda ser considerado el mismo debe representar, al menos, un gasto mayor o igual al 0,06%, también se toma en cuenta la Frecuencia de adquisición, los mismos para ser seleccionados deben tener una frecuencia de consumo de, al menos, un valor mayor o igual al 8%, por ultimo a los productos seleccionados con los dos criterios anteriormente mencionados se les exige que los mismos deben encontrarse por lo menos en tres de cinco quintiles de gasto. Además de estos criterios, otros considerados complementarios son que los productos seleccionados deben ofrecer garantías de su permanencia en el mercado, los mismos también deben cumplir características y especificaciones que posibiliten el seguimiento de precios y el producto seleccionado debe ser representativo en la evolución de precios del conjunto de productos de la canasta básica. Con relación a las ponderaciones el IPC de base 2016 se calcula como un promedio ponderado de las variaciones de precios de los bienes y servicios incluidos en el índice, Se

consideran solo los gastos de consumo final de los hogares urbanos para la determinación de las respectivas ponderaciones por producto. No se consideran el autoconsumo, el auto suministro, el salario en especie, ni tampoco el valor imputado por el uso de la vivienda propia. Las ponderaciones se calculan en forma proporcional al gasto total de consumo. Estas se calculan a partir de la Encuesta de Presupuestos Familiares (EPF) y se actualizan cuando dejan de reflejar adecuadamente la estructura de consumo de la población de referencia. La ponderación de un bien o servicio es proporcional al gasto realizado en el mismo respecto del gasto total. De igual forma como se calcula las ponderaciones para los diferentes productos, se realiza la ponderación de las ciudades en el índice nacional, para ello, se calcula tomando el gasto total de todos los hogares de cada una de las ciudades, sobre el gasto total de todos los hogares de todas las ciudades, y de esta forma se construye la tasa de ponderación de la ciudad en el índice nacional.

Con relación a la muestra de establecimientos, la misma se subdivide en la selección de las áreas comerciales y los establecimientos. Para las áreas comerciales se toman las que cumplan los criterios de representatividad, los mismos son que exista al menos uno o más mercados significativamente concurridos por demandantes y oferentes, que la población que forma parte de un área comercial tenga comportamiento y hábitos homogéneos, por último, que sea significativo en el volumen de ventas que realiza. Para el caso de establecimientos informantes de precios se consideran en primer lugar los mercados, en el caso de estos su tamaño, el número de puestos que tienen y la afluencia de compradores, además se verifica que el tiempo de permanencia en el mercado sea como mínimo de un año de funcionamiento, con relación a negocios minoristas se toma en cuenta que los dueños, gerentes y administradores tengan predisposición a dar información sobre los productos a ser cotizados en su establecimiento. Dentro de estos establecimientos los considerados son los puestos móviles, fijos, Kioscos, tiendas de barrio o almacenes, supermercados o micro mercados, locales especializados, servicios básicos y transporte.

Los establecimientos pueden ir variando de manera constante por cambios en los comportamientos de los consumidores.

Con respecto a la recolección de los datos por los operadores, se realizan aplicando el método directo donde el trabajador del INE se presenta formalmente al lugar para hacer el control de los precios en los productos requeridos y el método indirecto donde el trabajador no se identifica y simula ser un cliente que quiere comprar algún bien o solicitar un determinado servicio, en ambos casos posterior a la recolección de datos los mismos

deben ser transferidos en el mismo día al sitio web de la institución para su respectivo control.

En la recolección de precios, se presentan diferentes problemas como el ser el desabastecimiento de un producto o la estacionalidad de estos, para ello se establecen precios referenciales que ocupen esos espacios vacíos o dependiendo la situación se aplica una imputación de precios para replicar la variación de un producto similar en la del producto faltante.

Construcción del Índice de Precios del Consumidor vía web scrapping

En la construcción de la base de datos para replicar el IPC nos enfrentamos a diferentes obstáculos tanto en la programación de los diferentes códigos para la extracción de información de los productos como por la gran proporción de economía informal en el país.

Es importante entender el contexto al respecto, Bolivia es el país con la mayor proporción de economía informal en el mundo, aproximadamente diferentes estudios establecen que entre el 80% y 85% de las personas se dedican a la informalidad, por esta razón son varios los establecimientos que los trabajadores del INE tienen que visitar de manera personal para conocer las distintas variaciones, desde tiendas pequeñas instaladas en los mismos domicilios hasta grandes mercados con puestos de venta ubicados en el suelo, sin duda es un escenario bastante atípico en la actualidad a nivel mundial llegando solamente a ser comparable con lo que se vive en países del continente africano.

Esta situación presenta un gran obstáculo para la recolección de precios vía web scrapping a nivel nacional, debido a que los pocos supermercados existentes en las ciudades no suben sus precios a sitios de internet para poder hacer la extracción correspondiente. Es por ello que para la construcción del índice se tomará únicamente a los departamentos del eje troncal (Cochabamba, Santa Cruz) con las cadenas de supermercados más grandes que tengan disponibilidad en los datos, en el caso de Santa Cruz se utilizó los datos extraídos por el sitio web de Amarket.com que es una página con amplia cantidad de productos y constante actualización de datos, para el caso de Cochabamba se utilizó la cadena de supermercado IC Norte.

Otro problema relacionado fue la imposibilidad de acceder a la información de divisiones enteras por su nula disponibilidad de datos. En el caso de Vivienda y Servicios Básicos, la misma presenta dos obstáculos, en primer lugar el servicio de arrienda de la vivienda en Bolivia es diferente al resto del mundo porque no solamente se tiene contratos por alquiler sino también por anticrético, a está particularidad se le suma que los contratos

no son públicos en las páginas de agencias en bienes raíces no se encuentra la información y la única respuesta que se puede obtener es haciendo una consulta individual directa a un agente, algo que el web scrapping no puede realizar.

En el caso de Transporte, en primer lugar, las variaciones en los precios no existen en su mayoría porque la gasolina está regulada por el Estado y mantiene un precio fijo inalterable lo cual evita variaciones en los precios de transporte público, en segundo lugar, para el caso de los transportes interdepartamentales de igual forma no se cuenta con un sitio web que registre los precios para su extracción de datos.

En el caso de las divisiones de Comunicaciones y Educación, los diferentes productos que componen el índice no son publicados, las agencias que controlan el servicio telefónico no tienen un sitio web donde se pueda extraer el dato de los costos que se incurren por los contratos pre o post pago. Asimismo, los colegios y universidades de carácter privado no hacen pública su mensualidad, en ambos casos la única forma de obtener la información es consultar de manera personal directamente a los establecimientos mencionados.

En el caso de Salud, las farmacias a nivel nacional no cuentan con un sitio web que actualice periódicamente los precios de sus medicamentos, como tampoco se puede sacar la información de la web con relación a los costos de consulta medica privada.

Por último, para la división de Alimentos y Bebidas Consumidos Fuera del Hogar se presenta un problema de información por parte del Instituto Nacional de Estadística que no proporciona de manera pública cuales son los mercados locales o restaurantes tomados en cuenta para hacer el seguimiento de las distintas comidas diarias en el IPC, por lo tanto, es imposible realizar el *scrapping* de precios.

Sin duda este escenario presenta un obstáculo complejo a la hora de realizar la réplica del IPC con la técnica del web scrapping como una alternativa de información semanal para la ciudadanía, por ello, se tomó la decisión de excluir las categorías mencionadas y realizar la comparación con sus variaciones mensuales del INE.

De igual forma, se resolvieron los problemas presentados en la programación del código vía Python con Google Colab, la complicación presentada fue que los datos que están disponibles en los sitios web para la recolección de datos no están ordenados de manera uniforme, para cada supermercado se presenta una estructura HTML diferente, por lo cual se debe personalizar cada código para la extracción de datos.

Para uniformar esta extracción se establecieron nombres de columnas iguales para todos los establecimientos de la siguiente forma: ID, Categoría, Nombre del producto, Procedencia, Fecha y Precio.

En el caso de la columna ID, se estudió la composición de cada estructura HTML para encontrar el identificador de cada producto y así evitar que el código extraiga productos repetidos que podrían están en más de una categoría.

Para la columna Categoría, los productos no estaban ordenados por los grupos que tiene cada división e incluso algunos supermercados no tenían categorías para poder dividir los productos, por lo tanto, se estudió la estructura HTML producto por producto de cada supermercado para crear un diccionario en Python que permita al código identificar de manera automática a que grupo pertenecen de acuerdo con su nombre del producto.

Para la columna Nombre del producto, se analizó la estructura de cada establecimiento para encontrar el identificador y extraer la información requerida.

Para la columna Procedencia, los supermercados no contaban con la división entre productos de origen nacional e importado, por lo tanto, se creó un diccionario de palabras clave que el código pueda reconocer automáticamente del nombre de la marca y así clasificarlas por su fuente.

Para la columna Fecha, se incluyó una función que registre automáticamente el día en el cual se está haciendo la consulta y de esta forma ordenar la base de datos.

Para la columna Precio, se analizó la estructura de cada establecimiento para encontrar el identificador y extraer la información requerida.

Posteriormente a obtener los datos diarios y semanales de los productos, se procedió a realizar el código para medir las variaciones de los productos en sintonía con el trabajo que realiza el Instituto Nacional de Estadística, para ello se llevó a cabo un merge en dos pasos para unificar los resultados, primero a los resultados por supermercado, para lo cual se le pidió al código que según el número de archivos que se vaya a utilizar realice el merge vertical identificando los productos por su ID, en el caso de productos que en algún período no registrasen un precio se le ordenó al código que repita el último dato conocido, además se incluyó la función de medir la variación en sus pesos respectivos, para ello con la información disponible por el INE se tomaron las ponderaciones respectivas para cada mes y se las incluyó en un diccionario donde el código tomaría como base 100 los datos de la primera fecha introducida y a partir de ahí mediría las variaciones para cada división.

Posteriormente se unifican las tablas de merge extraídas de cada supermercado en un solo Excel, en este paso el código repite la estructura del merge individual para cada establecimiento.

En la sección de anexos se presentan los códigos implementados para el desarrollo del trabajo.

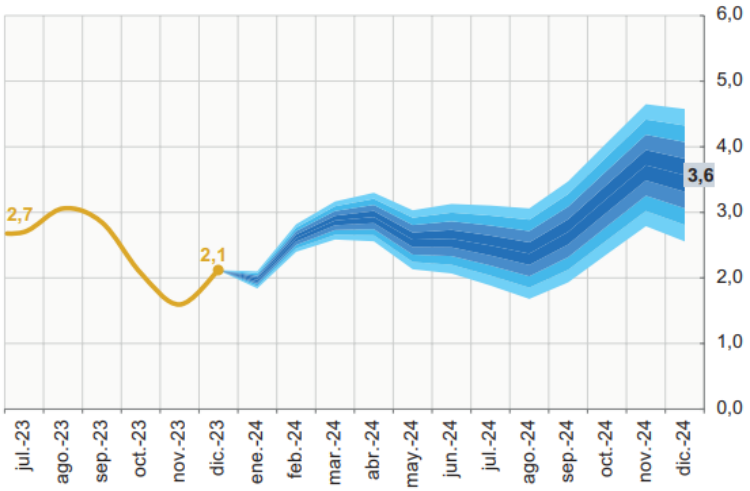
IV. Resultados

Primero se obtuvieron los resultados por departamento, en general se observa un incremento sostenido de varios productos considerados para el cálculo del Índice de Precios al Consumidor, si bien estos incrementos son todavía muy pequeños en comparación con otros países vecinos, es importante recalcar que en la última década Bolivia ha sido uno de los países con la inflación más baja a nivel mundial, incluso en la etapa de la pandemia la inflación se mantuvo en niveles mínimos en contraste con el resto del mundo.

El éxito de estos resultados ha sido en gran parte alcanzado por el tipo de cambio fijo con el dólar que mantiene el gobierno desde 2011, la subvención permanente a los combustibles y un incremento inaudito en los ingresos por exportaciones de gas que permitieron robustecer las Reservas Internacionales Netas (RIN). Sin embargo, la bonanza en la economía boliviana ha terminado, la reducción drástica de las RIN, los ingresos reducidos al mínimo por la exportación de gas y un aparato estatal demasiado grande son factores que ocasionan déficits gemelos para el país sudamericano que observa como el período de estabilidad parece estar llegando a su fin.

Hasta la fecha trabajada se puede afirmar que en esta gestión después de muchos años las proyecciones de inflación interanual por el Banco Central de Bolivia serán sobrepasadas con una tendencia creciente y constante.

Figura N°1 Inflación interanual observada y proyectada por el Banco Central de Bolivia (expresado en porcentaje)

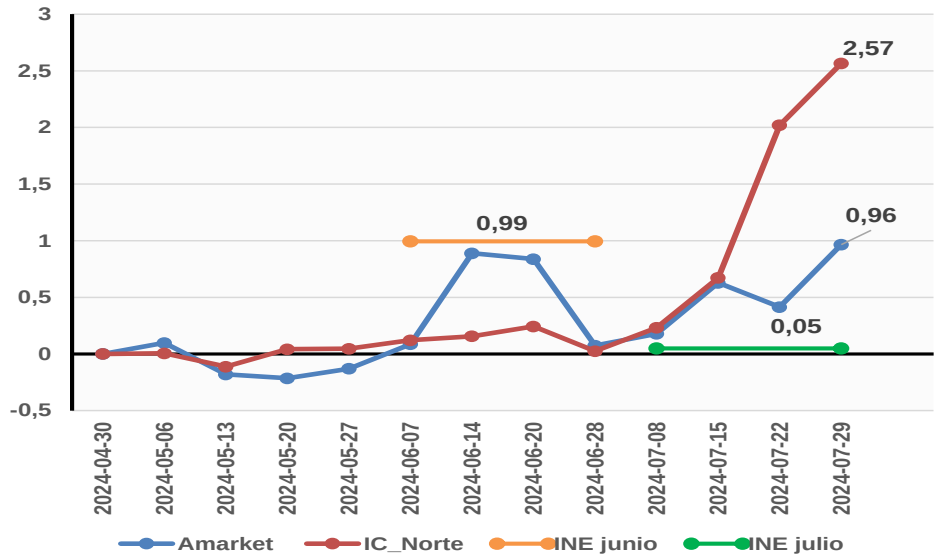


Fuente: Informe de política monetaria del BCB.

En la división de Alimentos por departamentos observamos que existen ciertos bienes que tienen diferentes comportamientos de una región a otra, esto debido al origen de la producción de estos y su traslado directo de productor a consumidor que reducen los costos de transacción.

Los resultados observados que toman mayo como mes de referencia reflejan una tendencia alcista cada vez mayor en la categoría que mayor peso tiene en el cálculo del IPC, si bien en junio los resultados fueron en línea con los publicados por el INE, para el mes de julio los mismos ya presentan una variación alcista a la reportada por el instituto responsable.

Figura N°2 Inflación mensual observada para Alimentos y Bebidas no alcohólicas (expresado en porcentaje)



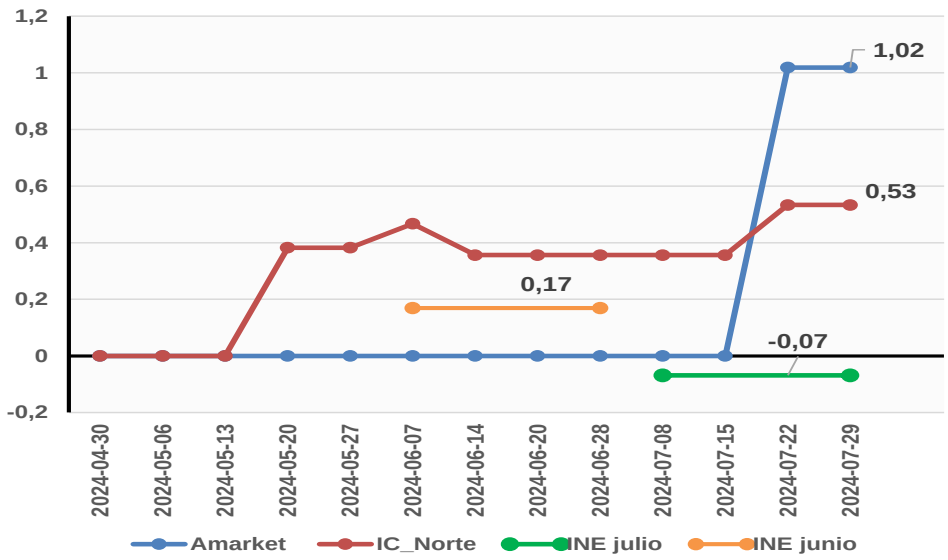
En el caso del departamento de Cochabamba debido a su amplia extensión de valles la producción principal es de frutas como el durazno, la chirimoya y frutilla, y verduras como la cebolla, zanahoria y tomate que logran otorgar una presión inflacionaria baja a estos alimentos generalmente volátiles por factores climatológicos, un ejemplo de ello es como para el departamento de Cochabamba la inflación producida por las heladas en la cosecha del tomate para los meses de mayo y junio fue mínima en comparación con las otras regiones del país.

En el caso del departamento de Santa Cruz, por su extensa área tropical goza de las mejores condiciones para la producción de productos pecuarios, avícolas y todos los derivados de los bovinos, en los últimos años el departamento ha tomado el protagonismo

de productor principal en diferentes bienes de la canasta familiar por su amplia migración de personas que se dedican a la agricultura y los incentivos del departamento para los productores, este factor permite al departamento tener la mayor cantidad de productos transferidos directamente al consumidor, por ello también se observa por la frecuencia diaria de recolección de datos que si los precios incrementan en Santa Cruz en los posteriores días los mismos se incrementarán en los otras regiones, de esta forma el *Big data* permite anticiparse y obtener tendencias inflacionarias precisas para futuras proyecciones.

En la división de Bebidas alcohólicas y tabaco para los departamentos se observa un incremento sostenido, lo cual refleja un mayor costo que están padeciendo los importadores de estos insumos, en comparación a los presentados por el INE esta categoría también presenta resultados de mayor presión inflacionaria, si bien los mismos habían mantenido el mismo nivel de precios en Santa Cruz durante mayo y junio, esto se debió al hecho de que tenían un stock importante para la venta, una vez que el mismo fue vendido en su totalidad debieron importar para el mes de julio, padeciendo de esta forma los problemas relacionados a la escases de divisas en el sistema financiero.

Figura N°3 Inflación mensual observada para Bebidas alcohólicas y tabaco (expresado en porcentaje)

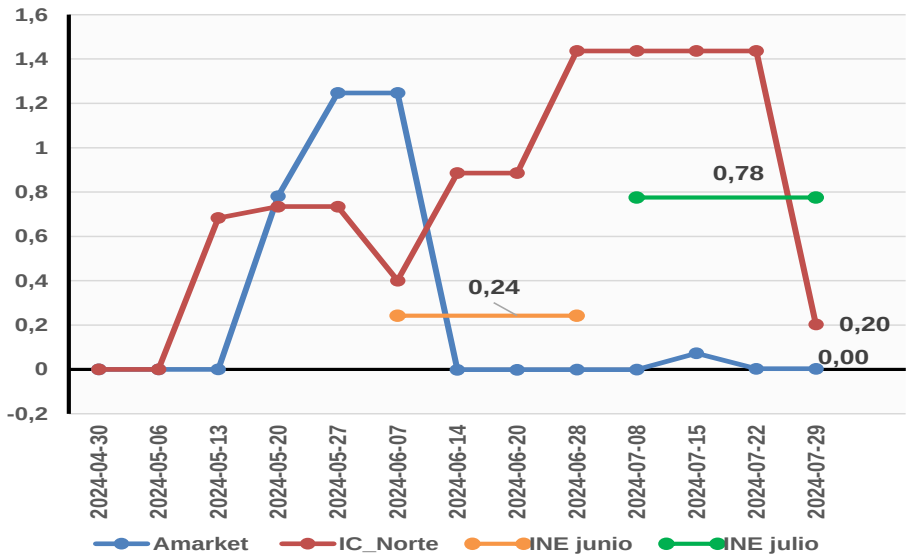


Para las divisiones restantes en su mayoría se tratan de bienes de origen importado, en los mismos se registra un incremento constante en distintos productos, esto va relacionado con la actual escasez de dólares al tipo de cambio oficial (6,96Bs/$us) que sufre el país por la cual todos los productos transportados desde otros países deben ser adquiridos con dólares

conseguidos del mercado paralelo que desde en los periodos comprendidos del trabajo paso de 8,50Bs/$us a 14,70Bs/$us, prueba inobjetable del incremento generalizado en el índice de precios al consumidor. Por ello en el caso de productos y servicios para el hogar los mismos presentaron una mayor tendencia alcista en los últimos meses que van en línea con los reportados por el INE.

Figura N°4 Inflación mensual observada para Muebles, bienes y servicios domésticos

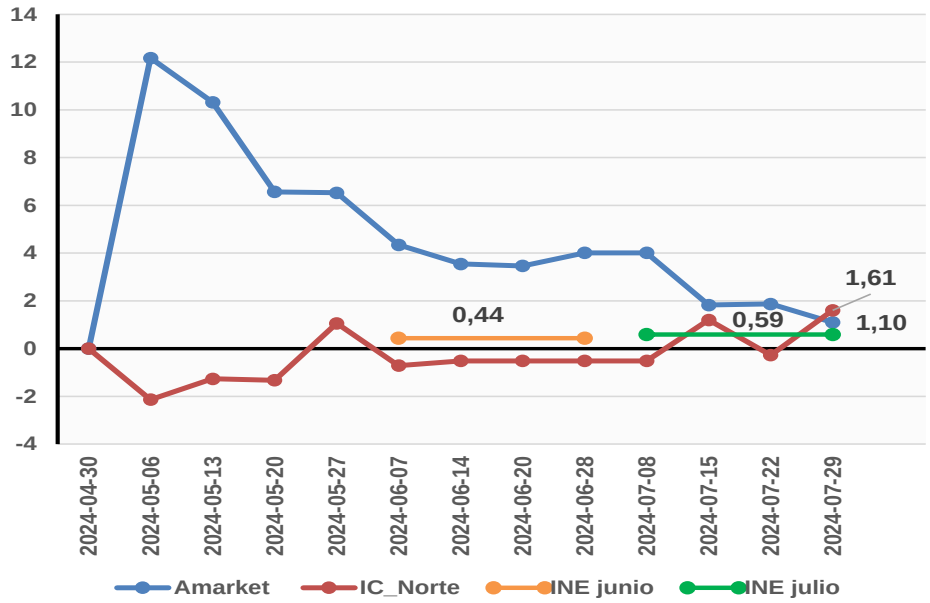
(expresado en porcentaje)



En el caso de la vestimenta, esta categoría presento un salto bastante alto en el mes de mayo, esto puede deberse a los conflictos sociales que generaron especulación en el rubro, para los posteriores meses las variaciones reportadas fueron convergiendo a las publicadas por el instituto de estadística.

Figura N°5 Inflación mensual observada para Prendas de vestir y calzados

(expresado en porcentaje)



Para el caso da la categoría de recreación y cultura, los mismos van presentando un incremento mayoritario, este hecho se debe al incremento de los precios en las categorías anteriormente mencionadas de primera necesidad, reflejando de esta forma una inflación de segunda vuelta menor por el aumento de precios en alimentos y ropa (figura 6). Mismo hecho se refleja para la categoría de bienes y servicios diversos (figura 7).

Figura N°6 Inflación mensual observada para Recreación y cultura (expresado en porcentaje)

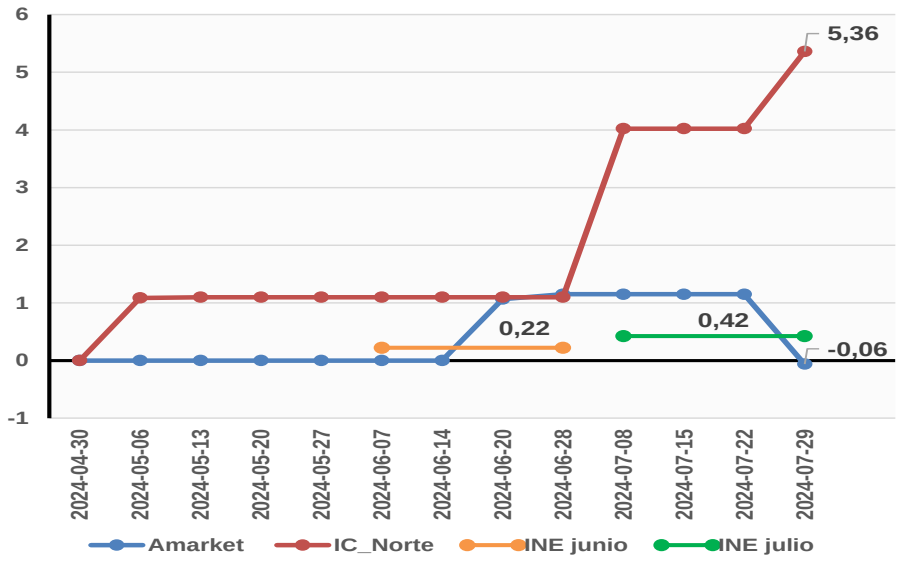


Figura N°7 Inflación mensual observada para Bienes y Servicios diversos

(expresado en porcentaje)

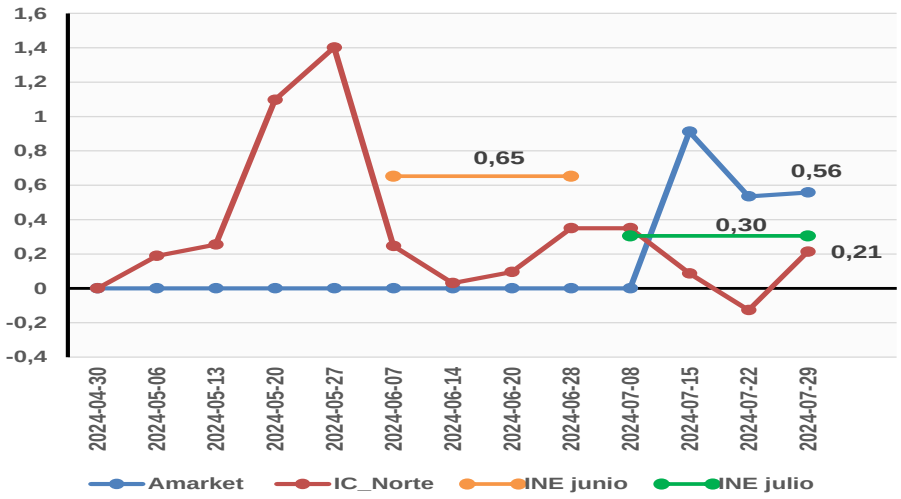
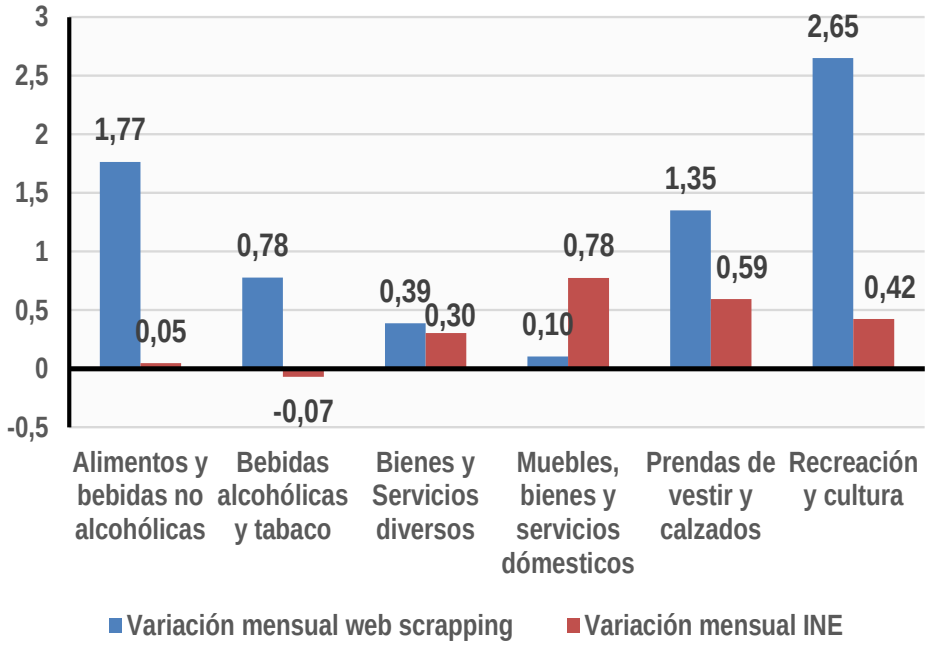


Figura N°8 Variación mensual observada para el mes de julio 2024

(expresado en porcentaje)



En el seguimiento diario realizado con Python se reporta un mayor incremento en Alimentos y bebidas no alcohólicas en comparación al observado por el INE de 1,72pp. En el caso de la categoría Bebidas alcohólicas y tabaco se reporta un mayor incremento en comparación al observado por el INE de 0,85pp, en Prendas de vestir y calzadas se observa que el aumento de precios es superior en 0,76pp y de igual forma para Recreación y cultura

en 2,23pp. Para el INE existió un mayor incremento en la categoría Muebles, bienes y servicios domésticos, en el caso de Bienes y Servicios diversos los resultados son similares para ambas metodologías (ver figura 8).

El Instituto Nacional de Estadística reportó una variación mensual de 0,47% en el nivel de precios para el mes de julio. El seguimiento realizado vía web scrapping presenta una variación mensual de 0,86% para el mes de julio.

De manera general para el IPC los resultados obtenidos aplicando la técnica del web scrapping son, en la mayoría de categorías, superiores a los reportados por el Instituto Nacional de Estadística, lo cual puede deberse a dos razones: la primera por el hecho de tomarse únicamente los departamento del eje troncal como muestra que genere un sesgo con relación a quizás una menor inflación reportada en las otras regiones, que al ser departamentos más pequeños tienen una menor demanda de productos importados que son de los más inflacionarios, la segunda razón puede deberse al hecho de que debido a la economía informal los puestos de ventas ambulantes en las calles considerados por el INE para la recolección de información pueden registrar menores variaciones de precios a los observados en las cadenas de supermercados, sin embargo, las tendencias en todas las categorías y de manera agregada son las mismas que las reportadas por la institución, por lo tanto se demuestra la eficiencia de poder aplicar esta técnica para el recogido de información pertinente a la inflación, además de su potencial valor de anticipar tendencias inflacionarias por su masiva cantidad de datos en frecuencia diaria.

Pronóstico a corto plazo

Con el objetivo de probar la utilidad de la extracción de datos se realizaron diferentes modelos para pronosticar la inflación del mes de agosto, la intención de ello fue demostrar la utilidad del *Big data* con el web scrapping para anticipar tendencias inflacionarias a muy corto plazo, en el caso de este trabajo se decidió que las proyecciones fueran a un mes.

Con esa finalidad se utilizaron tres tipos de modelos de Machine Learning:

- **Ensemble Learning**: Random Forest, Gradient Boosting y XGBoost
- **Regresión Lineal Regularizada**: Regresión Ridge y Elastic Net
- **Redes Neuronales**: LSTM (Long Short-Term Memory)

Random Forest

Es un modelo de Ensemble Learning que combina múltiples arboles de decisión para mejorar la precisión y reducir el sobreajuste, para ello utiliza el método de bootstrap

(muestras aleatorias con reemplazo) para entrenar cada árbol en un subconjunto diferente de conjunto de datos. En cada nodo de un árbol se considera un subconjunto aleatorio de características para la división, lo que ayuda a introducir diversidad entre los árboles, para la predicción final se obtiene por un promedio de las predicciones de todos los árboles.

Este enfoque de aleatorización, tanto en la selección de datos como en la de características, permite que el modelo capture diferentes patrones en el conjunto de datos y sea menos propenso a errores específicos de un único árbol. Como resultado, el Random Forest logra una mayor robustez y generalización en comparación con los modelos individuales, siendo capaz de manejar grandes volúmenes de datos y mitigar problemas como el sobreajuste. Además, la naturaleza paralela de los árboles de decisión permite que el modelo sea eficiente y escalable en aplicaciones prácticas.

Gradient Boosting

Es una técnica de boosting que constuye un modelo fuerte combinando varios modelos débiles (como árboles de decisión pequeños) de manera secuencial. Cada nuevo modelo se ajusta a los errores residuales de los modelos anteriores, es decir, intenta corregir los errores de las predicciones previas y usa el gradiente descendente para minimizar la función de pérdida, ajustando los parámetros del modelo en cada iteración para mejorar el rendimiento.

A medida que se agregan nuevos modelos, cada uno se centra en reducir los errores cometidos por los modelos anteriores, lo que permite al conjunto aprender de manera progresiva y mejorar la precisión. Al final del proceso, todas las predicciones individuales se combinan para generar una predicción final más precisa. Este enfoque iterativo permite que el Gradient Boosting sea muy eficaz en la captura de relaciones complejas dentro de los datos, pero también lo hace más susceptible al sobreajuste si no se controlan adecuadamente parámetros como la profundidad de los árboles, la tasa de aprendizaje y el número de iteraciones.

XGBoost (Extreme Gradient Boosting)

Es un algoritmo de aprendizaje automático basado en árboles de decisión que ha ganado popularidad debido a su alto rendimiento y eficacia en competiciones de ciencia de datos. Su metodología se basa en el concepto de boosting, en el que se entrenan múltiples modelos débiles (en este caso, árboles de decisión) de manera secuencial. Cada nuevo árbol

se ajusta para corregir los errores residuales del modelo anterior, lo que mejora progresivamente el rendimiento del conjunto.

XGBoost optimiza este proceso de boosting mediante técnicas avanzadas como la regularización para prevenir el sobreajuste, el manejo eficiente de datos faltantes y la paralelización del proceso de entrenamiento para reducir el tiempo de cómputo. Además, utiliza una estrategia de búsqueda de umbrales para la selección de las divisiones en los nodos de los árboles y aplica el método de "shrinkage" o tasa de aprendizaje, que ajusta la contribución de cada árbol al modelo final, lo que permite un control más fino y preciso en la construcción del modelo. Estas características hacen que XGBoost sea altamente eficiente, escalable y capaz de manejar grandes volúmenes de datos complejos.

Regresión Ridge

También conocida como regresión de penalización L2, es una técnica de regularización utilizada para abordar problemas de multicolinealidad en regresión lineal. Su objetivo principal es mejorar la estabilidad y la generalización del modelo cuando hay alta colinealidad entre las variables predictoras. Ridge añade una penalización a la función de costo tradicional de la regresión lineal que es proporcional al cuadrado de la magnitud de los coeficientes del modelo. Al añadir una penalización a la magnitud de los coeficientes, Ridge ayuda a evitar el sobreajuste del modelo a los datos de entrenamiento. Esto se traduce en una mejor generalización y un rendimiento más robusto en datos no vistos.

La penalización aplicada por Ridge reduce los coeficientes de las variables menos relevantes hacia cero, aunque no los elimina completamente, lo que significa que todas las variables permanecen en el modelo, pero con una influencia reducida. Este enfoque es especialmente útil cuando se trabaja con conjuntos de datos donde el número de variables predictoras es grande en comparación con el número de observaciones. A diferencia de la regresión Lasso, que puede llevar algunos coeficientes a cero y, por lo tanto, realizar selección de variables, Ridge es más adecuado cuando se espera que todas las variables tengan alguna influencia en la respuesta. La cantidad de penalización se controla mediante un hiperparámetro conocido como lambda (λ), que determina la fuerza de la regularización: un valor más alto de λ conduce a coeficientes más pequeños, mientras que un valor más bajo permite que el modelo se ajuste más a los datos originales.

Elastic Net

Es una técnica de regresión lineal que combina las penalizaciones L1 de Lasso y L2 de Ridge en la función de perdida, en este caso la función de perdida de Elastic Net es la suma del error cuadrático y una combinación ponderada de las penalizaciones L1 y L2.

Al combinar dos tipos de penalizaciones, Elastic Net ofrece una mayor robustez y capacidad de generalización en comparación con los métodos que utilizan solo una forma de regularización, lo que reduce el riesgo de sobreajuste y mejora la performance en datos nuevos.

Es especialmente útil en situaciones donde hay una alta correlación entre las variables predictoras o cuando el número de variables es mucho mayor que el número de observaciones. La penalización L1 (de Lasso) tiende a forzar algunos coeficientes a cero, lo que efectúa una selección de variables, mientras que la penalización L2 (de Ridge) distribuye el sesgo entre todas las variables, manteniéndolas en el modelo pero reduciendo su influencia. Elastic Net, al combinar ambas penalizaciones, logra un equilibrio que permite tanto la selección de variables como la mitigación de problemas de multicolinealidad. Los hiperparámetros lambda (λ) y alpha (α) controlan la intensidad y la mezcla de las penalizaciones: λ regula la fuerza total de la regularización, mientras que α determina la proporción entre L1 y L2. Esta flexibilidad hace que Elastic Net sea una opción preferida en problemas de regresión donde se necesita un control preciso sobre el ajuste del modelo y la selección de características.

LSTM (Long Short-Term Memory)

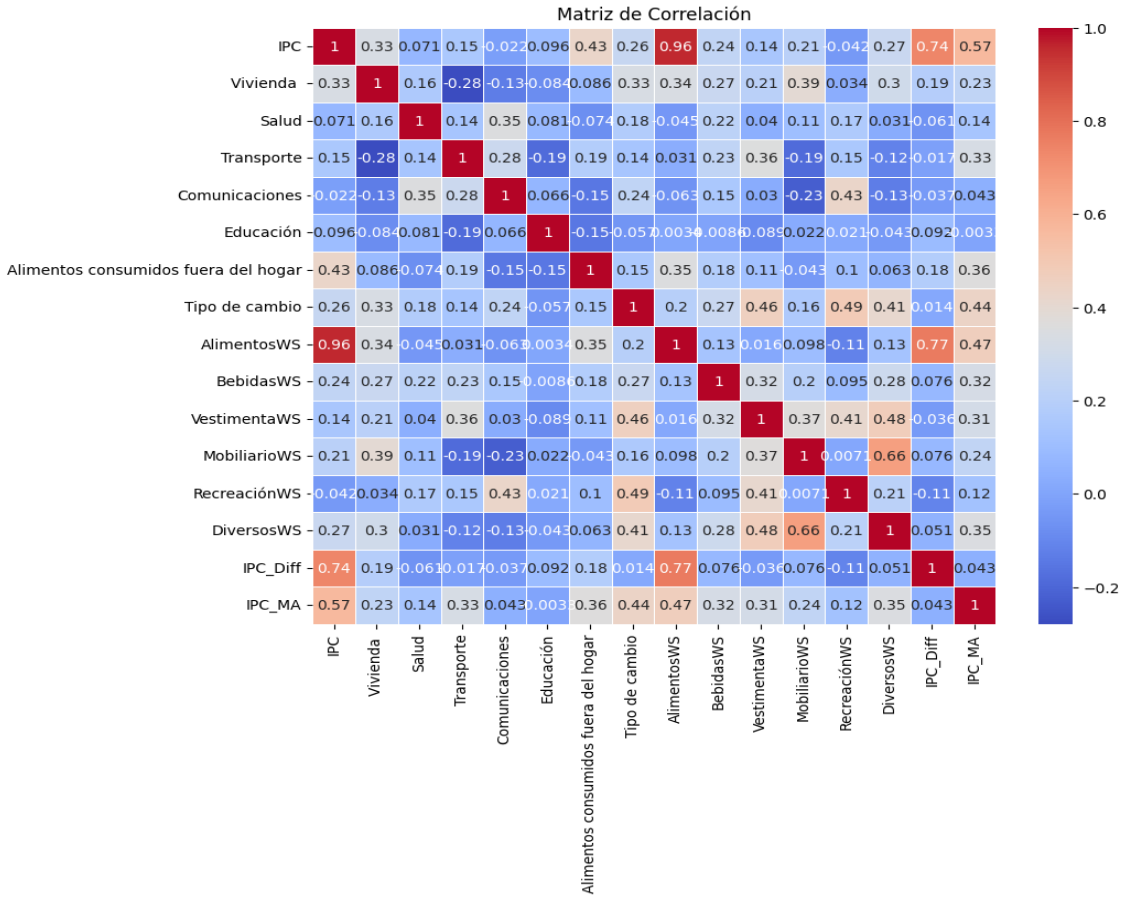
Este modelo de memoria a corto y largo plazo es un tipo de Red Neuronal Recurrente (RNN) diseñada para aprender dependencias a largo plazo en secuencias de datos. Estos modelos usan celdas de memoria y puertas (entrada, salida y olvido) para controlar el flujo de información a lo largo del tiempo, lo que les permite retener información relevante y olvidar la irrelevante, su estructura ayuda a superar el problema del desvanecimiento del gradiente que afecta a las RNN tradicionales. Es bastante flexible en su aplicación, desde la predicción de series temporales, el análisis de sentimiento, la generación de texto y cualquier problema que implique datos secuenciales.

Además, las celdas LSTM son capaces de manejar secuencias de longitud variable y pueden aprender patrones temporales complejos que se extienden a lo largo de intervalos prolongados. Este enfoque permite que las LSTM se destaquen en tareas donde es crucial capturar la dinámica a largo plazo de los datos, como en la predicción de eventos futuros

basados en un historial extenso o en la comprensión de la estructura subyacente de series de tiempo. Las puertas de entrada, salida y olvido dentro de la celda LSTM se activan mediante funciones sigmoides, lo que permite al modelo decidir de manera precisa qué información conservar, actualizar o descartar en cada paso temporal. Esta capacidad de manipular y preservar información durante largas secuencias hace que las LSTM sean una herramienta poderosa en la modelización de datos secuenciales complejos, superando a menudo a otras arquitecturas de redes neuronales en escenarios donde la memoria a largo plazo es crítica.

Para el tratamiento de los datos, los del Índice de Precios al Consumidor y sus diferentes divisiones fueron transformados a diferencias logarítmicas, de igual forma se hizo para el Tipo de Cambio, esto con el objetivo de estandarizar la muestra y obtener resultados más eficientes. El tamaño de la muestra fue de 79 meses (enero-2018 a julio-2024), además se implementó rezagos para las variables altamente correlacionadas, las elegidas fueron: Alimentos y bebidas sin alcohol; Alimentos y bebidas consumidos fuera del hogar; Transporte; Muebles, bienes y servicios domésticos; Bienes Diversos y Tipo de cambio debido a su alta correlación (ver figura 9).

Figura N°9 Matriz de correlación entre variables

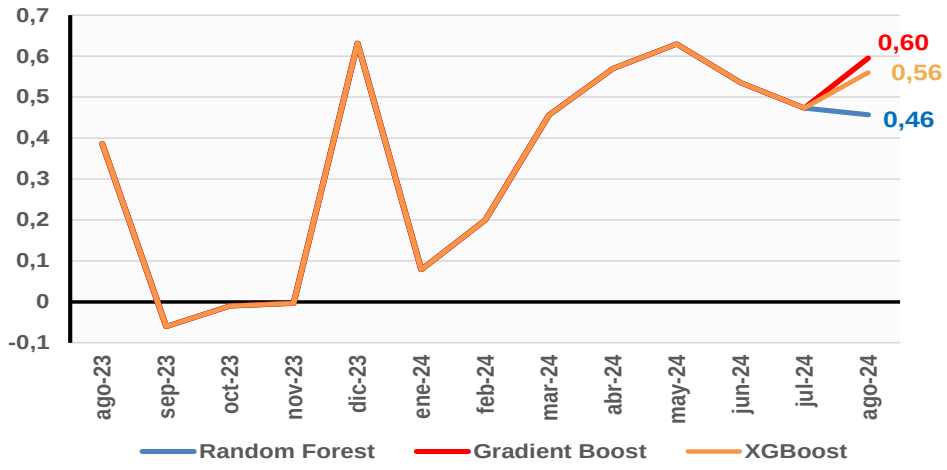


En el caso de los AlimentosWS son los insumos recolectados vía web scrapping que muestran una alta correlación con el Índice de Precios al Consumidor, esta categoría tiene una relación directa con los Alimentos y bebidas consumidos fuera del hogar y el Transporte por impactar directamente en sus precios y tarifas, en el caso de la categoría Vivienda no se tomó en cuenta porque debido a su complejidad anteriormente explicada, esta tiende a variar anualmente y no sería de un uso práctico para el pronóstico a un mes

Todas las variables son iguales hasta el mes de abril 2024, para los últimos periodos en las categorías disponibles los datos fueron reemplazados por los obtenidos con la extracción del web scrapping, en el caso del tipo de cambio, la misma toma en cuenta los datos recolectados en las casas de cambio del eje troncal.

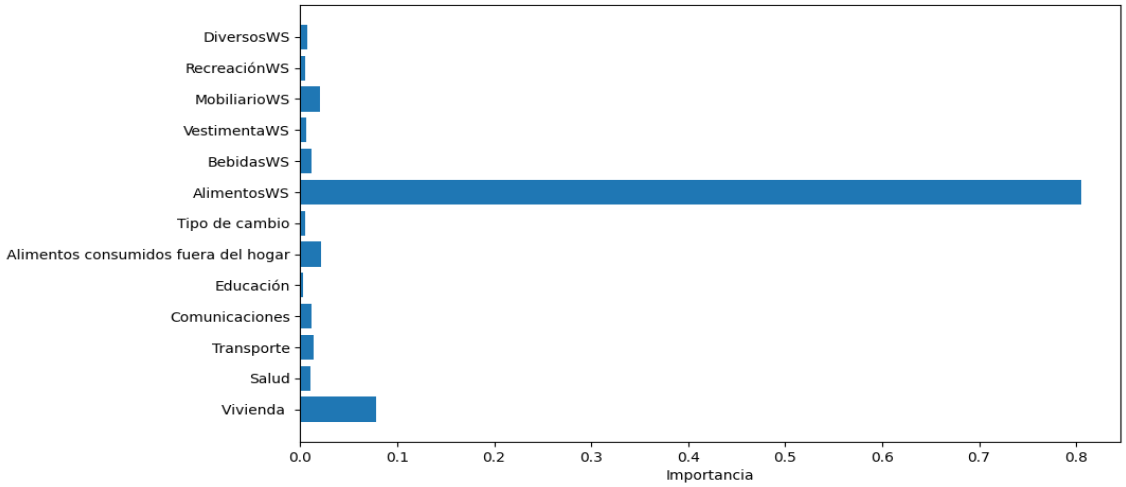
Las proyecciones de inflación para el mes de agosto mantienen su tendencia alcista, en este caso los modelos de Ensemble Learning son más conservadores y pronostican una variación mensual del IPC entre 0.46% y 0.60% ambos con un MAE (Error absoluto medio) de 0,001; demostrando así su eficacia entre los valores reales y los valores pronosticados.

Figura N°10 Pronóstico de la variación intermensual agosto 2024 con Ensemble Learning (expresado en porcentaje)



Estos modelos consideran a los Alimentos y bebidas no alcohólicas como la característica más importante a la hora de construir los diferentes modelos, le sigue la variable de Vivienda y se puede visualizar como el tipo de cambio empieza a incidir debido a su reciente flexibilización en el mercado paralelo, afectando de esta forma sobre todo a los bienes importados.

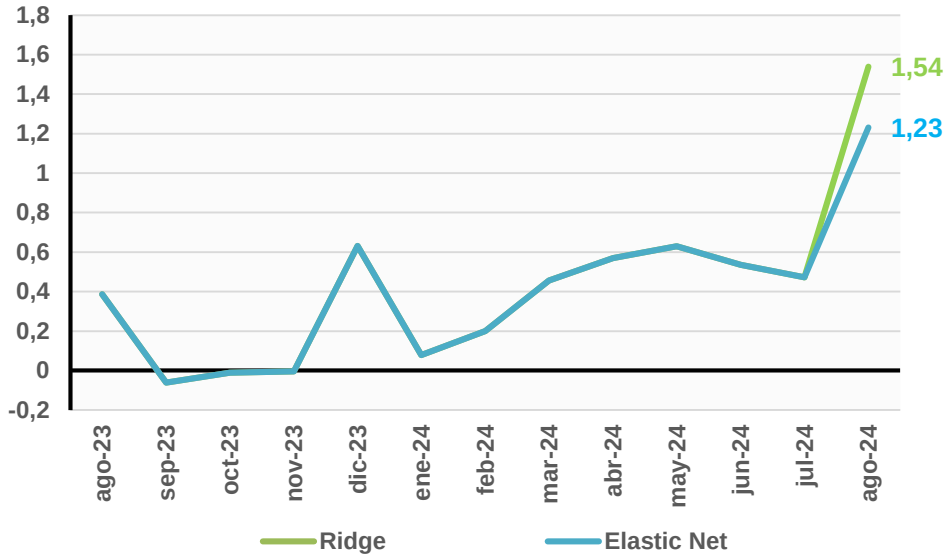
Figura N°11 Importancia de las variables con Ensemble Learning



Para los modelos de Regresión Lineal Regularizada la variación se estima por encima del 1,23% y llega a 1.54% con una mejor penalización en los datos, para estos casos existe una reducción de los coeficientes a valores cercanos a cero, por ende captura de mejor manera la tendencia alcista que presenta la inflación en los últimos meses, por ello la regresión Ridge proyecta un escenario más alto, sin embargo, la regresión Elastic Net al combinar las penalizaciones L1 y L2 proyecta una variación mensual menor. En ambos casos los resultados del MAE son de 0,001.

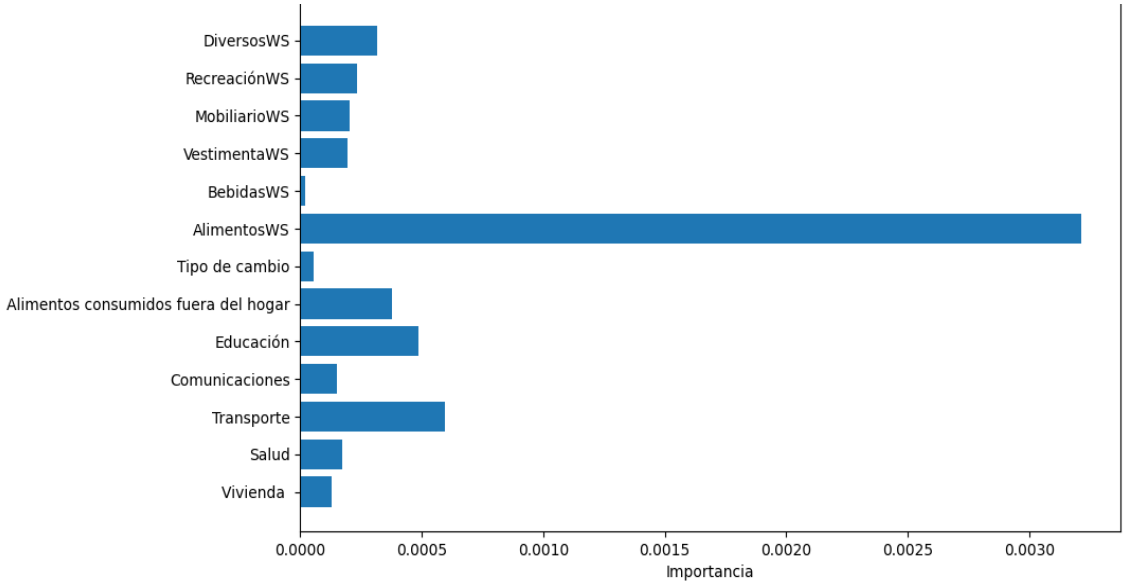
Figura N°12 Pronóstico de la variación intermensual agosto 2024 con Regresión Lineal Regularizada

(expresado en porcentaje)



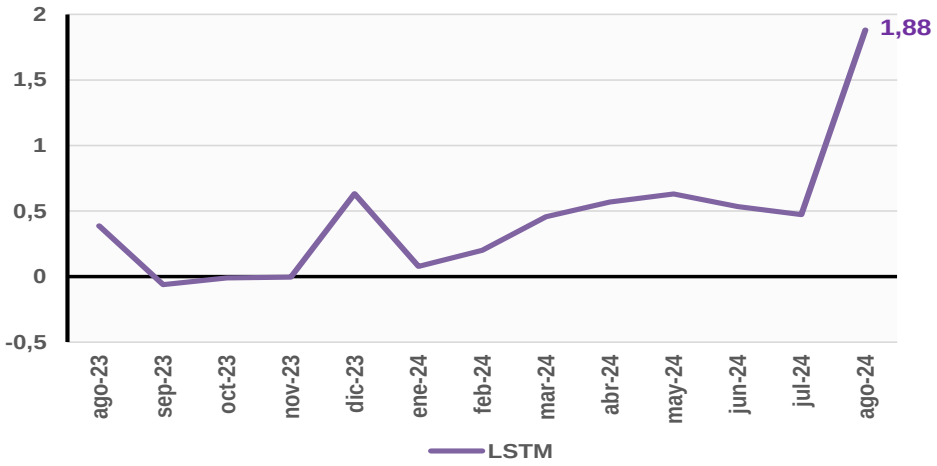
En estos modelos de igual forma se observa como la característica más importante del IPC es la categoría de Alimentos y bebidas no alcohólicas, posteriormente las diferentes categorías si tienen una pequeña reducción de sus coeficientes acorde a la importancia que le dan los modelos en su procesamiento, destaca la variable del Transporte por sus recientes variaciones debido a la escasez de combustibles, así mismo se observa que las categorías con bienes predominantemente importado empiezan a tener mayor incidencia en el Índice de Precios al Consumidor.

Figura N°13 Importancia de las variables con Regresión Lineal Regularizada



Por último, el modelo de red neuronal es el que captura una mayor variación mensual en el mes de agosto de 1,88%, su capacidad para detectar secuencias es de gran utilidad y su eficacia al igual que los otros modelos es bastante alta, las características para almacenar información a largo plazo y desechar aquello que no considera importante permite que el modelo entienda con mejor precisión la dinámica que tiene el IPC con relación a sus categorías. presenta un MAE de 0,001.

Figura N°14Pronóstico de la variación intermensual agosto 2024 con red neuronal (expresado en porcentaje)



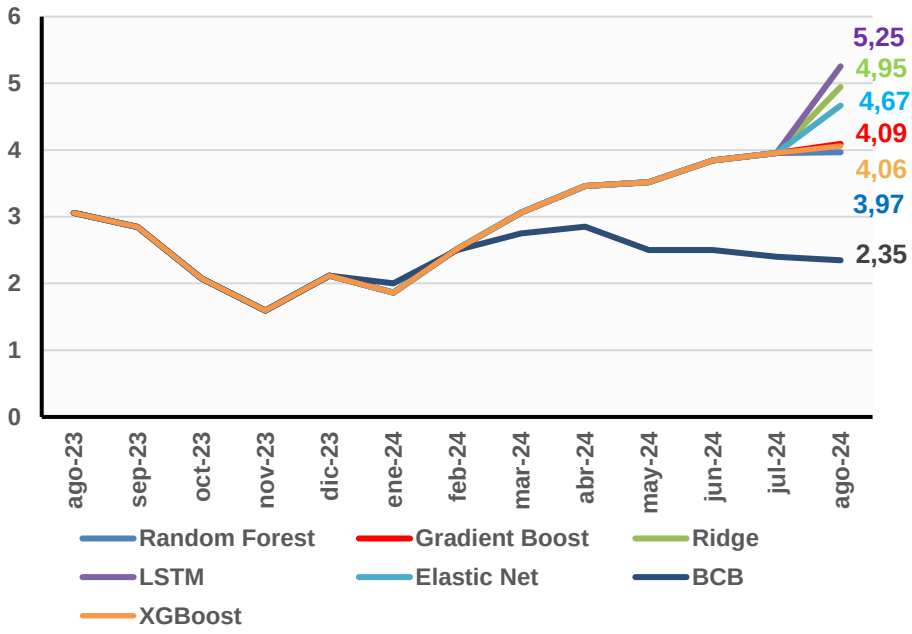
Para todos los modelos analizados, la inflación interanual supera las estimaciones del Banco Central de Bolivia por al menos un 1,62%, situando la variación entre el 3,97% y el 5.25%. Este dato indica una situación socioeconómica compleja, especialmente en el corto plazo para el nivel de precios.

Los modelos de Ensamble (Random Forest, Gradient Boost y XGBoost) proyectan una inflación menor en comparación a los demás, esto sucede por su aprendizaje histórico de las variables, al tener en la última década niveles muy bajos de inflación los modelos tienden a suavizar las futuras variaciones interanuales.

En el caso de los modelos de Regularización (Ridge y Elastic Net) su capacidad de penalizar valores atípicos de la muestra le hace mas sensible a percibir futuras tendencias inflacionarias, sin embargo, para esta muestra todas las variables tienen importancia en el aprendizaje automático, por ello cuando las penalizaciones pueden llegar a cero, como es el caso de Elastic Net, la proyección tiene a suavizarse.

Por último, para el caso del modelo de red neuronal LSTM, su procesamiento de la información le permite ser más preciso en el anticipo a las ultimas tendencias inflacionarias y descartar periodos de mayor estabilidad que perjudiquen el aprendizaje automático.

Figura N°12 Pronóstico de la variación interanual agosto 2024
(expresado en porcentaje)



Fuente: Informe de política monetaria 2024 (BCB).

V. Conclusiones

En el presente trabajo se intentó comparar el Índice de Precios al Consumidor reportado por el Instituto Nacional de Estadística de Bolivia aplicando la técnica del web scrapping con el lenguaje de programación de Python en su entorno de Google Colab.

La construcción de la base de datos para el trabajo presentó diferentes desafíos a nivel de la recolección de información como de la programación del código, la primera debido a que Bolivia es un país en vías en desarrollo, con una economía totalmente precarizada y la mayor proporción de trabajo en la informalidad en el mundo, estos factores hacen difícil conseguir la extracción de datos en canales online de sitios web como ser cadenas de supermercados o negocios minoristas, por lo cual se debió recortar la muestra para fines metodológicos. En este sentido también se tropezó con la nula disponibilidad de datos en todos los productos de divisiones del IPC, por esta razón se decidió excluir las divisiones con valores inexistentes.

En el caso de la programación del código los problemas pasaron por la nula uniformidad de los sitios web, lo cual llevó a personalizar cada código para los diferentes establecimientos, al tener diferentes estructuras HTML se decidió arbitrariamente considerar cinco columnas para el trabajo (ID, Categoría, Nombre del producto, Procedencia, Fecha y Precio), en el caso de la categoría y la procedencia de los diferentes bienes o servicios se creó un diccionario de palabras clave para identificar a donde pertenecían y de donde era su origen.

Pese a las diferentes dificultades expuestas, el trabajo consiguió resultados satisfactorios y alentadores para futuras investigaciones. A nivel económico los datos obtenidos reflejan la angustiante situación que atraviesa el país con una tendencia constante y creciente de los precios tanto de origen nacional por factores climatológicos (heladas, sequías) como también y en mayor medida de los productos importados por la escasez de dólares que no permite conseguir dólares al tipo de cambio oficial, sino a un nivel superior en el mercado paralelo.

En cuanto a la aplicación del web scrapping y por ende el uso de *Big data*, se demostró satisfactoriamente el éxito de esta técnica para medir la variación de los diferentes precios que componen el IPC, los resultados reflejan la misma tendencia que los reportados por el INE y además permiten anticipar futuras tendencias inflacionarias a corto plazo por la cadena de producción y distribución de los alimentos utilizando modelos de aprendizaje automático para pronosticar la variación intermensual a un mes de distancia.

El trabajo se suma y va en línea con investigaciones anteriores y relacionadas a la temática, el uso de nuevas técnicas tecnológicas para la recolección de información de alta frecuencia en la actualidad es una herramienta indispensable para todos los institutos de estadística nacionales.

La mayoría de los países con economías desarrolladas en la actualidad ya aplican estas técnicas en su recolección de información para distintos indicadores económicos, en el caso de los países de la región, Brasil y Uruguay tienen avances importantes en el tema y Chile se muestra como el único país con uso de las técnicas al mismo nivel que países de América del Norte o Europa.

Se recomienda al Instituto Nacional de Estadística de Bolivia empezar a tomar este camino para bridar a la población información transparente y de frecuencia semanal para una mejor planificación en la economía de los hogares bolivianos, para cumplir este cometido se debe realizar un plan de acción coordinado con diferentes instituciones del gobierno que también permitan formalizar en mayor medida el comercio en el país para la digitalización de la información pertinente en los canales adecuados y su posterior procesamiento por servidores especializados en manejar esta cantidad masiva de datos con frecuencia diaria.

Para futuras investigaciones con fines académicos se recomienda ampliar el horizonte de la muestra para verificar que los resultados obtenidos se mantienen con los reportados en este trabajo.

REFERENCIAS BIBLIOGRÁFICAS

Aparicio, D., & Bertolotto, M. (2020). Forecasting inflation with online prices. *International Journal of Forecasting*, 232-247.

Baye, M., & Morgan, J. (2004). Brand and Price Advertising in Online Markets. *Competition Policy Center, Working Paper Series*.

Bertolotto, M. (2016). Matching Distortion and Mean-Reversion Properties of the Real exchange Rates. *Universidad de San Andrés*.

Cavallo, A. (2016). Scraped Data and Sticky Prices. *MIT & NBER*.

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives, 30*(2), 151-178.

Cavallo, A., Diewert, E., Feenstra, R., Inklaar, R., & Timmer, M. (mayo de 2018). Using Online Prices for Measuring Real Consumption across Countries. *AEA Papers and Proceedings*, 483-497.

Cavallo, A., Neiman, B., & Rigobon, R. (2014). Currency unions, product introductions, and the real exchange rate. *The Quaterly Journal of Economics*, 529-595.

Deaton, A., & Heston, A. (2010). Undestanding PPPs and PPP-based National Accounts. *American Economic Journal*, 1-35.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science, 346*. doi:10.1126/science.1243089

Ellison, G., & Ellison, S. F. (2005). Lessons About Markets from the Internet. *Journal of Economic Perspectives, 19*(2), 139-158.

Instituto Nacional de Estadística Bolivia (INE). (s.f.). Obtenido de https://www.ine.gob.bo/

Llada, M., & Aromí, D. (2023). Pronóstico de la inflación con Twitter. *Económica, 69*. doi:https://doi.org/10.24215/18521649e031

Lünnemann, P., & Wintr, L. (2006). Are internet prices sticky? *BCL Working papers*.

Morton, F. S., Zettelmeyer, F., & Silva-Risso, J. (2001). Internet Car Retailing. *The Journal of Industrial Economics*, 501-519.

Orlandi, J. I., & Osovi Conti, M. N. (julio de 2018). *Repositorio digital San Andrés.* (Universidad de San Andrés, Ed.) Obtenido de http://hdl.handle.net/10908/17034

Polidoro, F., Giannini, R., Lo Conte, R., Mosca, S., & Rossetti, F. (2015). Web scraping techniques to collect data on consumer electronic and airfares for Italian HICP compilation. *Statistical Journal of the LAOS*, 165-17.

Taylor, L., Schroeder, R., & Meyer, E. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? *Big data & Society*, 1-10.

ANEXO °1 Código para extracción de datos cadena amarket.

```
# AMARKET
import requests
from bs4 import BeautifulSoup
from openpyxl import Workbook
from datetime import datetime
import re

def extraer_info_productos(url, archivo_salida,
palabras_clave_nacionales):
    try:
        wb = Workbook()
        ws = wb.active
        ws.append(['ID', 'Categoría', 'Nombre', 'Procedencia',
'Fecha', 'Precio'])  # Agregamos 'Procedencia' como encabezado

        # Lista para almacenar las partes únicas de los IDs de
productos ya capturados
        partes_ids_capturados = []

        # Realizar la solicitud GET a la página web
        response = requests.get(url)

        # Verificar si la solicitud fue exitosa (código de estado
200)
        if response.status_code == 200:
            # Utilizar BeautifulSoup para analizar el HTML de la
página
            soup = BeautifulSoup(response.content, 'html.parser')

            # Encontrar todos los enlaces a las colecciones de
productos
            enlaces_colecciones = soup.find_all('a',
href=re.compile(r'/collections/'))

            # Diccionario para mapear IDs de colecciones a categorías
            categorias = {}

            # Iterar sobre cada enlace a la colección de productos
            for enlace in enlaces_colecciones:
                url_coleccion = url + enlace['href']  # Construir la
URL completa de la colección
                nombre_coleccion = enlace.text.strip()  # Obtener el
nombre de la colección

                # Mapear el ID de la colección al nombre de la
categoría
```

```
                match = re.search(r'/collections/([^/]+)',
enlace['href'])
                if match:
                    id_coleccion = match.group(1)
                    categorias[id_coleccion] = nombre_coleccion

                # Realizar la solicitud GET a la colección de
productos
                response_coleccion = requests.get(url_coleccion)

                # Verificar si la solicitud a la colección fue
exitosa (código de estado 200)
                if response_coleccion.status_code == 200:
                    # Utilizar BeautifulSoup para analizar el HTML de
la colección de productos
                    soup_coleccion =
BeautifulSoup(response_coleccion.content, 'html.parser')

                    # Encontrar todos los productos en la colección
                    productos = soup_coleccion.find_all('li',
class_='productgrid--item')  # Ajusta el selector según la estructura
real de la página web

                    # Iterar sobre cada producto en la colección
                    for producto in productos:
                        # Extraer el ID del producto
                        id_producto = producto.get('data-product-
quickshop-url', 'ID no disponible')

                        # Extraer la parte única del ID (por ejemplo,
los últimos 4 dígitos)
                        parte_unica_id = id_producto[-4:]

                        # Verificar si la parte única del ID ya ha
sido capturada
                        if parte_unica_id not in
partes_ids_capturados:
                            partes_ids_capturados.append(parte_unica_
id)

                            # Extraer la categoría del producto
                            categoria = categorias.get(id_coleccion,
'Categoría no disponible')

                            # Extraer el nombre del producto
                            nombre_element = producto.find('h2',
class_='productitem--title')
                            nombre = nombre_element.text.strip() if
nombre_element else 'Nombre no disponible'
```

```
                            # Determinar la procedencia del producto
                            procedencia = 'Nacional' if
any(palabra.lower() in nombre.lower() for palabra in
palabras_clave_nacionales) else 'Importado'

                            # Extraer el precio del producto y
convertirlo a número con formato decimal
                            precio_element = producto.find('div',
class_='price__current')
                            precio_raw = precio_element.text.strip()
if precio_element else 'Precio no disponible'

                            # Extraer solo el número del precio y
reemplazar la coma por un punto
                            precio_number = precio_raw.split("Bs")[-
1].replace('.', '').replace(',', '.')

                            # Convertir el precio a formato decimal
                            precio = float(precio_number)

                            # Obtener la fecha actual
                            fecha_actual =
datetime.now().strftime('%Y-%m-%d')

                            # Escribir la información del producto en
una fila del archivo Excel
                            ws.append([id_producto, categoria,
nombre, procedencia, fecha_actual, precio])

            # Guardar el archivo Excel después de procesar todas las
colecciones
            wb.save(archivo_salida)
        else:
            print(f"Error al obtener la página {url}:
{response.status_code}")
    except Exception as e:
        print(f"Error al procesar la respuesta de la página: {e}")

# URL de la página web que quieres analizar
url = 'https://amarket.com.bo/'

# Nombre del archivo Excel de salida
archivo_salida = 'precios_05_08_24.xlsx'

# Lista de palabras clave para productos nacionales
palabras_clave_nacionales = ["Paceña", "Farmacorp", "Pil",
"Victoria", "Punto Verde", "Soy", "Biogurt", "Cocinero", "Bloch",
"Caisy", "Aguai", "Bio", "De Zero", "Big", "Fino", "Bragantina",
```

```
"Buena Vista", "Cafe Gourmet", "Andean", "Belen Pan", "Amor Tw",
"Chocoleo", "Delocious", "Divertiloops", "Sofia", "Fridosa", "Chuy",
"Merida", "Iglu", "La Sierra", "Santa Fe", "Bonle", "Chiqui", "Del
Campo", "Delizia", "Chicolac", "Guapurutu", "Guaruja", "Chiquitano",
"San German", "San Javier", "Suiza", "Huevo", "Nutrirolon",
"Chiquidrink", "Clara Bella", "Greco", "Ghee", "Regia", "Reyna",
"Santa Elena", "Aranjuez", "Activade", "Aloe", "Premium", "Coco",
"Durazno", "Lima", "Limon", "Mandarina", "Maya", "Acelga", "Ajo",
"Berenjena", "Carote", "Cebollita", "Choclo", "Habas", "Jengibre",
"Locoto", "Papa", "Alimentarte", "Apita", "Huavi", "Mican", "Alka",
"Calmadol", "Chaki", "Curadil", "Digestan", "Escudo", "Punto Chef",
"Podium"]

# Llamar a la función para extraer la información de los productos y
guardarla en el archivo Excel
extraer_info_productos(url, archivo_salida,
palabras_clave_nacionales)
```

ANEXO °2 Código para extracción de datos cadena IC norte.

```
import requests
import datetime
import pandas as pd

# Diccionario de palabras clave para marcas nacionales
palabras_clave_nacionales = ["Paceña", "Farmacorp", "Pil",
"Victoria", "Punto Verde", "Soy", "Biogurt", "Cocinero", "Bloch",
"Caisy", "Aguai", "Bio", "De Zero", "Big", "Fino", "Bragantina",
"Buena Vista", "Cafe Gourmet", "Andean", "Belen Pan", "Amor Tw",
"Chocoleo", "Delocious", "Divertiloops", "Sofia", "Fridosa", "Chuy",
"Merida", "Iglu", "La Sierra", "Santa Fe", "Bonle", "Chiqui", "Del
Campo", "Delizia", "Chicolac", "Guapurutu", "Guaruja", "Chiquitano",
"San German", "San Javier", "Suiza", "Huevo", "Nutrirolon",
"Chiquidrink", "Clara Bella", "Greco", "Ghee", "Regia", "Reyna",
"Santa Elena", "Aranjuez", "Activade", "Aloe", "Premium", "Coco",
"Durazno", "Lima", "Limon", "Mandarina", "Maya", "Acelga", "Ajo",
"Berenjena", "Carote", "Cebollita", "Choclo", "Habas", "Jengibre",
"Locoto", "Papa", "Alimentarte", "Apita", "Huavi", "Mican", "Alka",
"Calmadol", "Chaki", "Curadil", "Digestan", "Escudo", "Punto Chef",
"Podium"]

def obtener_datos(url):
    try:
        respuesta = requests.get(url)
        respuesta.raise_for_status()  # Lanza una excepción para
errores HTTP
        datos = respuesta.json()
        return datos
    except requests.exceptions.RequestException as e:
        print("Error al obtener los datos:", e)
        return None

def extraer_info_productos(datos):
    productos = []
    if datos:
        for producto in datos:
            nombre_producto = producto.get('productName', 'Nombre no
disponible')
            precio_producto = obtener_precio(producto)
            id_producto = producto.get('productId', 'ID no
disponible')
            categoria_producto = obtener_categoria(producto)
            procedencia_producto = obtener_procedencia(producto)
            fecha_actual = obtener_fecha_actual()
            productos.append({'ID': id_producto, 'Categoría':
categoria_producto, 'Fecha': fecha_actual, 'Nombre': nombre_producto,
'Precio': precio_producto, 'Procedencia': procedencia_producto})
```

```
    return productos

def obtener_precio(producto):
    sellers = producto.get('items', [])
    for item in sellers:
        commertial_offer = item.get('sellers',
[])[0].get('commertialOffer', {})
        precio = commertial_offer.get('Price')
        if precio is not None:
            return precio
    return 'Precio no disponible'

def obtener_categoria(producto):
    categorias = producto.get('categories', [])
    if categorias:
        return categorias[0].split('/')[-2]
    return 'Categoria no disponible'

def obtener_procedencia(producto):
    marca_producto = producto.get('brand', 'Marca no disponible')
    for palabra in palabras_clave_nacionales:
        if palabra.lower() in marca_producto.lower():
            return "Nacional"
    return "Importado"

def obtener_fecha_actual():
    fecha_actual = datetime.datetime.now().strftime('%Y-%m-%d')
    return fecha_actual

def guardar_en_excel(productos):
    df = pd.DataFrame(productos)
    df.to_excel('icnorte_05_08_24.xlsx', index=False)

def main():
    categorias_subcategorias = {
        "lacteos": ["leches", "yogurt/yogurt-bebible", "mantequillas-
y-margarinas"],
        "automotriz": ["limpiadores", "siliconas-y-ceras"],
        "limpieza-y-detergentes": ["limpieza-de-ropa",
"ambientadores-e-insecticidas", "limpieza-de-cocina", "limpieza-de-
banos", "limpieza-del-hogar", "limpieza-para-calzados"],
        "abarrotes-y-despensa": ["aceites-y-vinagres/aceite-vegetal",
"arroz", "granos", "azucar-y-endulzantes", "fideos-y-pastas",
"comidas-instantaneas"],
        "frutas-y-verduras": ["verduras"],
        "carnes-mariscos-y-pescado": ["cortes-por-kg-",
"hamburguesas", "menudencias", "pescado", "pollo"],
        "bebes": ["panales", "toallas-humedas", "alimento-de-bebe",
"cuidado-y-aseo-del-bebe"],
```

```
        "cuidado-personal": ["cuidado-oral", "jabones-y-perfumeria",
"cuidado-del-cabello", "afeitado", "cuidado-corporal"],
        "bebidas-cervezas-y-licores": ["aguas", "jugos", "isotonicos-
y-energizantes", "gaseosas", "cervezas", "vinos", "licores-y-
destilados"],
        "panaderia-y-reposteria": ["panes", "reposteria",
"empanadas", "masas-tipicas", "bizcochos"],
        "mascotas": ["perros", "gatos"],
        "fiambres-y-quesos": ["fiambres", "embutidos", "quesos",
"huevos"]
    }

    productos_totales = []

    for categoria, subcategorias in categorias_subcategorias.items():
        for subcategoria in subcategorias:
            url_api =
f"https://www.icnorte.com/api/catalog_system/pub/products/search/{cat
egoria}/{subcategoria}"
            print("API:", url_api)
            datos = obtener_datos(url_api)
            productos = extraer_info_productos(datos)
            if productos:
                productos_totales.extend(productos)

    guardar_en_excel(productos_totales)

if __name__ == "__main__":
    main()
```

ANEXO °3 Código para el merge de datos de amarket por mes

```
import pandas as pd
from google.colab import files

# Función para cargar archivos Excel
def cargar_archivos(num_archivos):
    archivos = []
    for i in range(num_archivos):
        print(f"precios_{i + 1}:")
        uploaded = files.upload()
        archivos.append(pd.read_excel(list(uploaded.keys())[0]))
    return archivos

# Función para fusionar los DataFrames verticalmente
def fusionar_verticalmente(archivos):
    df_final = pd.concat(archivos, ignore_index=True)
    return df_final

# Función para calcular los índices
def calcular_indices(df):
    # Encontrar la fecha base (la primera fecha)
    fecha_base = df.columns[4]  # La quinta columna es la de Fecha

    # Calcular el índice para la fecha base (siempre 100)
    df['Indice'] = 100

    # Iterar sobre las fechas posteriores y calcular los índices
    fechas_posteriores = df.columns[5:]  # Las columnas posteriores a
la fecha base son las fechas posteriores
    for fecha in fechas_posteriores:
        # Calcular el índice para la fecha actual en función de la
fecha base
        df['Indice'] = (df['Precio'] /
df.groupby('ID')['Precio'].transform('first')) * 100

        # Llenar los valores faltantes con el índice de la fecha base
        df['Indice'].fillna(100, inplace=True)

    return df

# Función para asignar categoría INE
def asignar_categoria_INE(categoria):
    if categoria in categorias_ine:
        return categorias_ine[categoria]
    else:
        return 'Otro'

# Función para asignar peso
```

```
def asignar_peso(categoria_ine):
    return pesos_categorias_ine.get(categoria_ine, 0)  # Si no se
encuentra la categoría INE, se asigna un peso de 0

# Cargar los archivos Excel
num_archivos = int(input("Ingrese el número de archivos de precios:
"))
archivos = cargar_archivos(num_archivos)

# Fusionar los DataFrames verticalmente
df_resultados = fusionar_verticalmente(archivos)

# Calcular los índices
df_resultados = calcular_indices(df_resultados)

# Diccionario de categorías INE
categorias_ine = {
    'Alimentos y bebidas no alcohólicas': ['Fideos y pastas', 'Arroz
Legumbres y Semillas', 'Salsas Sazonadores y Aderezos', 'Azúcaresy
Edulcorantes', 'Harinas', 'Aceites y Vinagre', 'Sal Pimienta y
Especias', 'Café Té y Mates', 'Panadería y Repostería', 'Comidas
Preparadas', 'Granos y Cereales', 'Res', 'Pollo', 'Hamburguesas',
'Fiambres y Embutidos', 'Cerdo', 'Pescados', 'Leches', 'Quesos',
'Huevos', 'Yogurt', 'Mantequilla y Margarinas', 'Aguas Jugos y
Gaseosas', 'Energizantes e Isotónicas', 'Hielo',
    'Frutas', 'Verduras', 'Frutos Secos y Semillas', 'Chocolates',
'Caramelos y Golosinas', 'Helados', 'Galletas', 'Saladitos',
'Abarrotes y Despensas', 'Frutas y Verduras', 'Leche evaporada',
'Mezclas Lacteas', 'Leche Fresca', 'Leche de Soya', 'Leche en Polvo',
'Leche Fresca', 'Yogurt Bebible', 'Margarinas', 'Mantequillas',
'Mantequillas y Margarinas', 'Aceite Vegetal', 'Arroz Extra', 'Arroz
Superior', 'Otros Granos', 'Quinua', 'Frijol', 'Frutos Secos',
'Quinua', 'Azucar', 'Endulzantes', 'Fideos Largos', 'Fideos Cortos',
    'Purés', 'Sopas', 'Ramen', 'Verduras', 'Cortes por Kg. ',
'Menudencias', 'Pescado', 'Agua de Mesa', 'Agua Saborizada',
'Nectar', 'Refrescos', 'Isotónicos y Energizantes', 'Gaseosas',
'Panes', 'Repostería', 'Empanadas', 'Masas Típicas', 'Bizcochos',
'Fiambres', 'Embutidos', 'huevos', 'lacteos', 'panaderia-y-
reposteria', 'despensa-y-abarrotes', 'canasta-basica', 'dulces',
'congelados', 'chocolates-1', 'carnes'],
    'Bebidas alcohólicas y tabaco': ['Vinos Cervezas y Licores',
'Cigarrillos', 'Cervezas Nacionales', 'Vino Tinto', 'Vino Blanco',
'bebidas-y-licores'],
    'Prendas de vestir y calzados': ['Mamá y Bebé', 'Ropa y
Accesorios', 'Pañales', 'Compotas', 'Baño del Bebé', 'Talco para
Bebé', 'Licores y Destilados', 'bebes'],
    'Muebles, bienes y servicios dómesticos': ['Electro Hogar',
'Limpieza del hogar', 'Herramientas', 'Limpiadores', 'Siliconas y
Ceras', 'Detergentes', 'Suavizantes', 'Insecticidas',
```

```
'Ambientadores', 'Papel Toalla y Servilletas', 'Sacagrasa y Otros
Limpiadores', 'Lavavajillas', 'Papel Film', 'Papel Aluminio',
'Otros', 'Accesorios de Limpieza', 'Limpiadores y Desinfectantes',
'Limpiadores de Muebles', 'Desinfectantes', 'Limpiadores de Vidrios',
'Ceras y Limpiadores de Piso', 'Betun Lustrador', 'Cepillos',
'Escobillas', 'Accesorios', 'Repelentes',
    'limpieza-del-hogar'],
    'Recreación y cultura': ['Electrónicos y Computación',
'Mascotas', 'Escolar y Librería', 'Regalos y Juguetes', 'Alimentos
para Mascotas', 'Perros', 'Gatos', 'mascotas'],
    'Bienes y Servicios diversos': ['Jabones y Colonias',
'Sanitizador', 'Belleza y Cuidado Personal', 'Pasta Dental', 'Hilo
Dental y Otros', 'Cepillos de Dientes', 'Toallas Húmedas', 'Papel
Higiénico', 'Desodorantes y Antitranspirantes', 'Shampoo',
'Tratamiento para Cabello', 'Tintes para Cabello', 'Maquinas de
Afeitas', 'Lociones y Bálsamos', 'Espumas y Geles', 'Cremas
Corporales', 'Protectores Solares', 'cuidado-personal']
}

# Diccionario de pesos
pesos_categorias_ine = {
    'Alimentos y bebidas no alcohólicas': 0.4888721510526270,
    'Bebidas alcohólicas y tabaco': 0.0158874798095131,
    'Prendas de vestir y calzados': 0.1366007862440250,
    'Muebles, bienes y servicios dómesticos': 0.1097916535788970,
    'Recreación y cultura': 0.1123645293336670,
    'Bienes y Servicios diversos': 0.1364833999812700
}


# Función para asignar categoría INE a cada fila
def asignar_categoria_INE(categoria):
    if isinstance(categoria, str):  # Verificar si la categoría es
una cadena
        for categoria_ine, palabras_clave in categorias_ine.items():
            for palabra_clave in palabras_clave:
                if palabra_clave.lower() in categoria.lower():
                    return categoria_ine
    return 'Otro'


df_resultados['Categoria_INE'] =
df_resultados['Categoría'].apply(asignar_categoria_INE)

# Aplicar la función para asignar el peso a cada categoría según la
categoría INE
df_resultados['Peso'] =
df_resultados['Categoria_INE'].apply(asignar_peso)
```

```
# Calcular la columna 'Valor'
df_resultados['Valor'] = df_resultados['Indice'] *
df_resultados['Peso']

# Calcular el promedio de la columna "Valor" por "Categoria_INE" en
cada fecha
df_promedio_indice_por_categoria = df_resultados.groupby(['Fecha',
'Categoria_INE'])['Indice'].mean().reset_index()

# Fusionar el promedio de valores por categoría con el DataFrame
principal
df_resultados = pd.merge(df_resultados,
df_promedio_indice_por_categoria, on=['Fecha', 'Categoria_INE'],
suffixes=('', '_promedio'))

# Calcular el promedio de la columna "Valor" por "Categoria_INE" en
cada fecha
df_promedio_valor_por_categoria = df_resultados.groupby(['Fecha',
'Categoria_INE'])['Valor'].mean().reset_index()

# Fusionar el promedio de valores por categoría con el DataFrame
principal
df_resultados = pd.merge(df_resultados,
df_promedio_valor_por_categoria, on=['Fecha', 'Categoria_INE'],
suffixes=('', '_promedio'))


# Agregar columna "Supermercado"
df_resultados['Supermercado'] = 'Amarket'

# Guardar los resultados en un archivo Excel
nombre_archivo_excel = "4Amarket_tablaindice.xlsx"
df_resultados.to_excel(nombre_archivo_excel, index=False)
print(f"Resultados guardados en el archivo {nombre_archivo_excel}.")
```

ANEXO °4 Código para el merge de datos de IC norte.

```
import pandas as pd
from google.colab import files

# Función para cargar archivos Excel
def cargar_archivos(num_archivos):
    archivos = []
    for i in range(num_archivos):
        print(f"precios_{i + 1}:")
        uploaded = files.upload()
        archivos.append(pd.read_excel(list(uploaded.keys())[0]))
    return archivos

# Función para fusionar los DataFrames verticalmente
def fusionar_verticalmente(archivos):
    df_final = pd.concat(archivos, ignore_index=True)
    return df_final

# Función para calcular los índices
def calcular_indices(df):
    # Encontrar la fecha base (la primera fecha)
    fecha_base = df.columns[4]  # La quinta columna es la de Fecha

    # Calcular el índice para la fecha base (siempre 100)
    df['Indice'] = 100

    # Iterar sobre las fechas posteriores y calcular los índices
    fechas_posteriores = df.columns[5:]  # Las columnas posteriores a
la fecha base son las fechas posteriores
    for fecha in fechas_posteriores:
        # Calcular el índice para la fecha actual en función de la
fecha base
        df['Indice'] = (df['Precio'] /
df.groupby('ID')['Precio'].transform('first')) * 100

        # Llenar los valores faltantes con el índice de la fecha base
        df['Indice'].fillna(100, inplace=True)

    return df

# Función para asignar categoría INE
def asignar_categoria_INE(categoria):
    if categoria in categorias_ine:
        return categorias_ine[categoria]
    else:
        return 'Otro'

# Función para asignar peso
```

```
def asignar_peso(categoria_ine):
    return pesos_categorias_ine.get(categoria_ine, 0)  # Si no se
encuentra la categoría INE, se asigna un peso de 0

# Cargar los archivos Excel
num_archivos = int(input("Ingrese el número de archivos de precios:
"))
archivos = cargar_archivos(num_archivos)

# Fusionar los DataFrames verticalmente
df_resultados = fusionar_verticalmente(archivos)

# Calcular los índices
df_resultados = calcular_indices(df_resultados)

# Diccionario de categorías INE
categorias_ine = {
    'Alimentos y bebidas no alcohólicas': ['Fideos y pastas', 'Arroz
Legumbres y Semillas', 'Salsas Sazonadores y Aderezos', 'Azúcaresy
Edulcorantes', 'Harinas', 'Aceites y Vinagre', 'Sal Pimienta y
Especias', 'Café Té y Mates', 'Panadería y Repostería', 'Comidas
Preparadas', 'Granos y Cereales', 'Res', 'Pollo', 'Hamburguesas',
'Fiambres y Embutidos', 'Cerdo', 'Pescados', 'Leches', 'Quesos',
'Huevos', 'Yogurt', 'Mantequilla y Margarinas', 'Aguas Jugos y
Gaseosas', 'Energizantes e Isotónicas', 'Hielo',
    'Frutas', 'Verduras', 'Frutos Secos y Semillas', 'Chocolates',
'Caramelos y Golosinas', 'Helados', 'Galletas', 'Saladitos',
'Abarrotes y Despensas', 'Frutas y Verduras', 'Leche evaporada',
'Mezclas Lacteas', 'Leche Fresca', 'Leche de Soya', 'Leche en Polvo',
'Leche Fresca', 'Yogurt Bebible', 'Margarinas', 'Mantequillas',
'Mantequillas y Margarinas', 'Aceite Vegetal', 'Arroz Extra', 'Arroz
Superior', 'Otros Granos', 'Quinua', 'Frijol', 'Frutos Secos',
'Quinua', 'Azucar', 'Endulzantes', 'Fideos Largos', 'Fideos Cortos',
    'Purés', 'Sopas', 'Ramen', 'Verduras', 'Cortes por Kg. ',
'Menudencias', 'Pescado', 'Agua de Mesa', 'Agua Saborizada',
'Nectar', 'Refrescos', 'Isotónicos y Energizantes', 'Gaseosas',
'Panes', 'Repostería', 'Empanadas', 'Masas Típicas', 'Bizcochos',
'Fiambres', 'Embutidos', 'huevos', 'lacteos', 'panaderia-y-
reposteria', 'despensa-y-abarrotes', 'canasta-basica', 'dulces',
'congelados', 'chocolates-1', 'carnes'],
    'Bebidas alcohólicas y tabaco': ['Vinos Cervezas y Licores',
'Cigarrillos', 'Cervezas Nacionales', 'Vino Tinto', 'Vino Blanco',
'bebidas-y-licores'],
    'Prendas de vestir y calzados': ['Mamá y Bebé', 'Ropa y
Accesorios', 'Pañales', 'Compotas', 'Baño del Bebé', 'Talco para
Bebé', 'Licores y Destilados', 'bebes'],
    'Muebles, bienes y servicios dómesticos': ['Electro Hogar',
'Limpieza del hogar', 'Herramientas', 'Limpiadores', 'Siliconas y
Ceras', 'Detergentes', 'Suavizantes', 'Insecticidas',
```

```
'Ambientadores', 'Papel Toalla y Servilletas', 'Sacagrasa y Otros
Limpiadores', 'Lavavajillas', 'Papel Film', 'Papel Aluminio',
'Otros', 'Accesorios de Limpieza', 'Limpiadores y Desinfectantes',
'Limpiadores de Muebles', 'Desinfectantes', 'Limpiadores de Vidrios',
'Ceras y Limpiadores de Piso', 'Betun Lustrador', 'Cepillos',
'Escobillas', 'Accesorios', 'Repelentes',
    'limpieza-del-hogar'],
    'Recreación y cultura': ['Electrónicos y Computación',
'Mascotas', 'Escolar y Librería', 'Regalos y Juguetes', 'Alimentos
para Mascotas', 'Perros', 'Gatos', 'mascotas'],
    'Bienes y Servicios diversos': ['Jabones y Colonias',
'Sanitizador', 'Belleza y Cuidado Personal', 'Pasta Dental', 'Hilo
Dental y Otros', 'Cepillos de Dientes', 'Toallas Húmedas', 'Papel
Higiénico', 'Desodorantes y Antitranspirantes', 'Shampoo',
'Tratamiento para Cabello', 'Tintes para Cabello', 'Maquinas de
Afeitas', 'Lociones y Bálsamos', 'Espumas y Geles', 'Cremas
Corporales', 'Protectores Solares', 'cuidado-personal']
}

# Diccionario de pesos
pesos_categorias_ine = {
    'Alimentos y bebidas no alcohólicas': 0.4888721510526270,
    'Bebidas alcohólicas y tabaco': 0.0158874798095131,
    'Prendas de vestir y calzados': 0.1366007862440250,
    'Muebles, bienes y servicios dómesticos': 0.1097916535788970,
    'Recreación y cultura': 0.1123645293336670,
    'Bienes y Servicios diversos': 0.1364833999812700
}


# Función para asignar categoría INE a cada fila
def asignar_categoria_INE(categoria):
    if isinstance(categoria, str):  # Verificar si la categoría es
una cadena
        for categoria_ine, palabras_clave in categorias_ine.items():
            for palabra_clave in palabras_clave:
                if palabra_clave.lower() in categoria.lower():
                    return categoria_ine
    return 'Otro'


df_resultados['Categoria_INE'] =
df_resultados['Categoría'].apply(asignar_categoria_INE)

# Aplicar la función para asignar el peso a cada categoría según la
categoría INE
df_resultados['Peso'] =
df_resultados['Categoria_INE'].apply(asignar_peso)
```

```
# Calcular la columna 'Valor'
df_resultados['Valor'] = df_resultados['Indice'] *
df_resultados['Peso']

# Calcular el promedio de la columna "Valor" por "Categoria_INE" en
cada fecha
df_promedio_indice_por_categoria = df_resultados.groupby(['Fecha',
'Categoria_INE'])['Indice'].mean().reset_index()

# Fusionar el promedio de valores por categoría con el DataFrame
principal
df_resultados = pd.merge(df_resultados,
df_promedio_indice_por_categoria, on=['Fecha', 'Categoria_INE'],
suffixes=('', '_promedio'))

# Calcular el promedio de la columna "Valor" por "Categoria_INE" en
cada fecha
df_promedio_valor_por_categoria = df_resultados.groupby(['Fecha',
'Categoria_INE'])['Valor'].mean().reset_index()

# Fusionar el promedio de valores por categoría con el DataFrame
principal
df_resultados = pd.merge(df_resultados,
df_promedio_valor_por_categoria, on=['Fecha', 'Categoria_INE'],
suffixes=('', '_promedio'))


# Agregar columna "Supermercado"
df_resultados['Supermercado'] = 'IC_Norte'

# Guardar los resultados en un archivo Excel
nombre_archivo_excel = "5IC_Norte_tablaindice.xlsx"
df_resultados.to_excel(nombre_archivo_excel, index=False)
print(f"Resultados guardados en el archivo {nombre_archivo_excel}.")
```

ANEXO °5 Código para control de datos intersemanal

```
import pandas as pd
from google.colab import files

# Cargar los datos de las dos series de precios desde tu computadora
print("precios_xx_07_24:")
uploaded = files.upload()
archivo1 = list(uploaded.keys())[0]

print("precios_xx_07_24:")
uploaded = files.upload()
archivo2 = list(uploaded.keys())[0]

# Leer los archivos Excel y cargar los datos en dataframes de pandas
df1 = pd.read_excel(archivo1)
df2 = pd.read_excel(archivo2)

# Fusionar los dataframes en uno solo, utilizando el ID como clave de
fusión
df_merged = pd.merge(df1, df2, on='ID', suffixes=('_serie1',
'_serie2'))

# Calcular la variación porcentual de precio para cada producto
df_merged['Variacion_Porcentual'] = ((df_merged['Precio_serie2'] -
df_merged['Precio_serie1']) / df_merged['Precio_serie1']) * 100

# Leer la columna de procedencia y determinar si el producto es
Importado o Nacional
df_merged['Procedencia'] =
df_merged['Procedencia_serie1'].apply(lambda x: 'Importado' if x ==
'Importado' else 'Nacional')

# Filtrar solo los productos que tuvieron un aumento de precio
productos_aumentados = df_merged[df_merged['Variacion_Porcentual'] >
0]

# Imprimir los productos que aumentaron de precio y la variación
porcentual
print("Productos que aumentaron de precio:")
print(productos_aumentados[['Nombre_serie1', 'Variacion_Porcentual',
'Procedencia']])

# Calcular el promedio de la variación de precios solo para los
productos que aumentaron de precio
promedio_importados_aumentados =
productos_aumentados[productos_aumentados['Procedencia'] ==
'Importado']['Variacion_Porcentual'].mean()
```

```
promedio_nacionales_aumentados =
productos_aumentados[productos_aumentados['Procedencia'] ==
'Nacional']['Variacion_Porcentual'].mean()

print(f"Promedio de variación de precios para productos importados
que aumentaron de precio: {promedio_importados_aumentados:.2f}%")
print(f"Promedio de variación de precios para productos nacionales
que aumentaron de precio: {promedio_nacionales_aumentados:.2f}%")
```