Tipo de documento: Tesis de maestría

Escuela de Negocios. Master in Management + Analytics

Predicción del precio de azúcar en Argentina a partir de modelos de Machine Learning

Autoría: Sal, Mauricio José

Año: 2024

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

PREDICCIÓN DEL PRECIO DE AZÚCAR EN ARGENTINA A PARTIR DE MODELOS DE MACHINE LEARNING

TESIS

Mauricio José Sal

Mayo 2024

Tutor: Ricardo Di Pasquale

Resumen

Gracias al avance tecnológico vivenciado en la última década, posibilitando contar con una disponibilidad masiva de datos y una mayor facilidad de cómputo, se observaron avances en algoritmos y técnicas de aprendizaje automático. Esto abrió una nueva oportunidad para las organizaciones de contar con cierto tipo de información que antes se consideraban inalcanzable. En Argentina, y sobre todo en el sector agropecuario, el uso de estas herramientas aún tiene un largo camino por recorrer. El objetivo de este trabajo es utilizar técnicas relacionadas con el enfoque estadístico de aprendizaje supervisado (machine learning) para predecir el precio del azúcar en el mercado argentino. El mismo, cuenta con un contexto económico y político que lleva a pensar que responde a la ley de oferta y demanda, donde el precio depende de la cantidad de oferta circulante en el país. Por otro lado, el azúcar proviene del procesamiento de la caña de azúcar producida en los campos argentinos, siendo el clima un factor relevante para la predicción de la oferta de azúcar circulante. Debido a esto, para la realización de este trabajo se combinaron los datos climáticos con datos de producción de azúcar para predecir la producción anual a través de los modelos de XGBoost, Maquinas de Vector Soporte y Redes Neuronales. A partir de esta predicción y en combinación con los precios históricos del azúcar, se utilizaron las técnicas de Regresión Lineal, Regresión Exponencial, PieceWise, Splines, Prophet y XGBoost para la predicción del precio y posterior inferencia sobre su movimiento. Durante la ejecución de ambas predicciones el mejor modelo resultó ser XGBoost, alcanzando un error porcentual en datos no observados sobre la producción anual de 10,88% y un MAPE de 9,50% en la predicción del precio. A pesar de contar con estos resultados, luego de analizar aquel modelo que proporciona el valor que tendrá el precio en el futuro, se pudo observar que el mismo no utiliza a la producción nacional como un factor determinante a la hora de realizar su predicción. Esto, no permite validar desde una perspectiva estadística la hipótesis desde la cual se parte al iniciar este trabajo. Sin embargo, este trabajo demuestra que la aplicación de técnicas de machine learning puede proporcionar con razonable precisión predicciones que resultan de gran utilidad a la hora de tomar decisiones estratégicas.

Abstract

Thanks to the technological advancements experienced in the last decade, enabling massive data availability and easier computing processes, there have been significant improvements in algorithms and machine learning techniques. This has opened up new opportunities for organizations to access certain types of information previously considered unattainable. In Argentina, particularly in the agricultural sector, the use of these tools still has a long way to go. The objective of this work is to utilize techniques related to the statistical approach of supervised learning to predict the price of sugar in the Argentine market. The economic and political context suggests that the market follows the law of supply and demand, where the price depends on the amount of circulating supply in the country. Additionally, sugar comes from the processing of sugarcane produced in Argentine fields, making the climate a relevant factor for predicting the circulating supply of sugar. Therefore, in this study, climate data were combined with sugar production data to predict annual production using XGBoost models, Support Vector Machines, and Neural Networks. Based on this prediction and in combination with historical sugar prices, techniques such as Linear Regression, Exponential Regression, Piecewise, Splines, Prophet, and XGBoost were used to predict the price and subsequently infer its movement. During the execution of both predictions, the best model turned out to be XGBoost, achieving a percentage error on unobserved

data for annual production of 10.88% and a MAPE of 9.50% in price prediction. Despite these results, after analyzing the model that provides the future price value, it was observed that it does not use national production as a determining factor in its prediction. This does not allow us to statistically validate the hypothesis from which this work originated. However, this work demonstrates that the application of machine learning techniques can provide reasonably accurate predictions that are highly useful for making strategic decisions.

Índice

1. Introducción

1.1. Industria sucroalcoholera

La caña de azúcar es un vegetal noble, base de la industria sucroalcoholera, que, con sus productos principales, el azúcar y el alcohol, y otros derivados contribuye al desarrollo de la especie humana.

El azúcar o sacarosa es un disacárido formado por la combinación de glucosa y fructosa. Se obtiene en el reino vegetal mediante la combinación del agua y la luz solar por medio de la clorofila (pigmento verde que se encuentra en las hojas) a través del proceso de la fotosíntesis. Solo las plantas de color verde pueden realizar este fenómeno, por medio del cual se fija la energía solar y se la pone a disposición del hombre y los animales para su consumo. En la Argentina, como en casi toda Sudamérica, sólo se obtiene a partir de la caña de azúcar.

Argentina es un mediano productor en la industria sucroalcoholera que concentra la actividad principalmente en dos regiones del Noroeste de su territorio (NOA), al sur del Trópico de Capricornio: por un lado, la provincia de Tucumán y por otro las provincias de Salta y Jujuy, a las que genéricamente se las denomina Norte. Aunque significativamente menor en volumen, también hay producción de azúcar en las provincias de Santa Fe y Misiones, en el Noreste del territorio nacional.

El cultivo de la caña se realiza a lo largo de los 12 meses del año en una superficie de aproximadamente 390 mil ha, en tanto que la actividad fabril ocupa seis meses, entre mayo y mediados de noviembre.

La introducción del bioetanol para ser mezclado con las naftas (gasolina) para automóviles en 2006, ha significado un desafío para esta industria, que concretó muy importantes inversiones para poder producir y atender la demanda del actual corte en la mezcla (12%, la mitad del cual aporta la industria sucroalcoholera y la mitad la del maíz).

Esta actividad ha cambiado el paradigma de la industria y los productores, ya que permite un redireccionamiento de excedentes que aporta sostenibilidad económica y previsibilidad productiva, que se tradujo en un incremento del área cultivada.

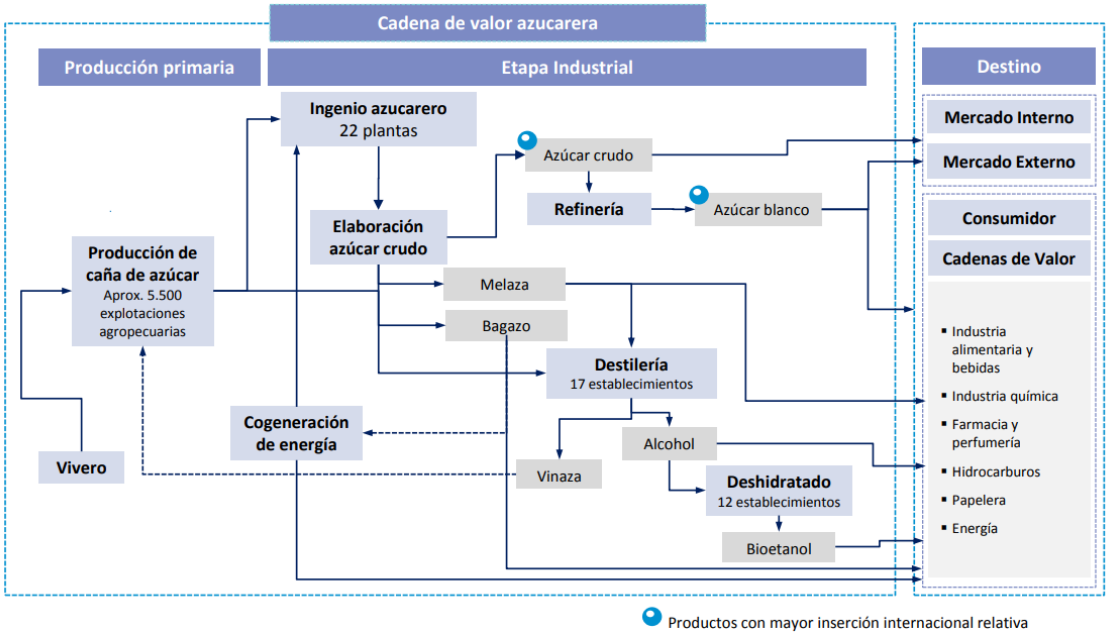
El creciente uso de biocombustibles en el país es la contribución más efectiva para la reducción de los Gases de Efecto Invernadero, compromiso asumido por Argentina en Kyoto y París, luego ratificado por leyes nacionales [1].

La actividad tiene un fuerte impacto socioeconómico en la región del NOA: genera 60.900 puestos de trabajo directos y 140.000 indirectos. En Tucumán la industria sucroalcoholera aporta el 10% del Producto Bruto Provincial mientras que la participación en Jujuy es del 6%. [2]

Las personas que producen la caña de azúcar en tierras argentinas se denominan cañeros. Las empresas encargadas de procesar la caña de azúcar para su posterior conversión en azúcar, se denominan ingenios. Estos operan a través de contratos de maquila, donde los cañeros entregan la producción resultante de su cosecha a los ingenios para que estos produzcan el azúcar. Por medio de este contrato, los cañeros e ingenios acuerdan el porcentaje que corresponderá de la producción resultante a cada parte. Siendo en promedio el cañero participe del 58% de la azúcar obtenida de la producción.

El siguiente diagrama ilustra con claridad el esquema de la cadena de valor de la industria sucroalcoholera:

Figura 1. Cadena de valor de la industria sucroalcoholera



Fuente: https://www.argentina.gob.ar/sites/default/files/sspmicro_cadenas_de_valor_azucar.pdf [3]

1.2. Mercado Azucarero:

El consumo interno de azúcar no ha registrado grandes variaciones en la última década, ya que se trata de un alimento básico para la población y tiene un precio comparativamente reducido. El consumo aparente se ha estabilizado en los últimos tres años alrededor de 1,5 millones de toneladas. Por esta razón, los excedentes exportables guardan gran relación con la producción. [4]

En el mercado internacional opera tanto la comercialización a través de cuotas y acuerdos especiales como en forma de libre mercado. Estados Unidos (EE. UU.) y la Unión Europea (UE) tienen contingentes para el ingreso de azúcar. En la UE, desde el año 2006 se aplican reformas en la regulación para reducir la protección de los productores de azúcar de remolacha.

En el período 2006-2016 la producción mundial de caña de azúcar se incrementó 33%: el promedio fue 1.754 millones de toneladas. Luego se registra un estancamiento de la producción. Los mejoramientos genéticos constituyen un factor relevante de la competitividad a nivel mundial. Se está llevando a cabo un importante proceso de inversión en la búsqueda de potenciar la producción combinada de azúcar y bioetanol.

En 2016, Brasil fue el primer productor con el 41% de la caña mundial. Le siguieron India (18%), China (7%), Tailandia (5%) y Pakistán (3%). Argentina ocupó el puesto 13 con 1,2% del total mundial. Los mayores rendimientos por hectárea se obtuvieron en Guatemala (129 toneladas); Senegal (118); Egipto (115) y Perú (112). Brasil obtuvo 75 toneladas por hectárea y Argentina 66.

En los últimos años, se ha acelerado el proceso de concentración y centralización, tanto de la producción como de las exportaciones. De este modo, se ha acentuado la relevancia

de las zonas del planeta donde los ingenios azucareros son más eficientes y el desplazamiento de las que obtienen bajos rendimientos.[3]

1.3. Política Argentina:

Históricamente, debido a la gran competencia externa existente, sobre todo la proveniente de Brasil, en Argentina se protegieron a los productores locales. El artículo 1° del Decreto 797/92 [5] establece lo siguiente:

"*Las importaciones de cualquier origen o procedencia de las mercaderías de las posiciones arancelarias que se detallan en el Anexo I del presente, estarán sujetas además del derecho 'ad valorem' vigente, al que surge de la aplicación del artículo 673 del Código Aduanero Argentino (Ley 22.415), en los términos del presente decreto.*"

En el anexo al cual se refiere este artículo se detallan las mercaderías de cualquier origen o procedencia de azúcar de caña o de remolacha y sacarosa químicamente pura, en estado sólido. Por otro lado, el artículo 673 del Código Aduanero Argentino [6] mencionado establece lo siguiente:

"*La importación para consumo en las condiciones previstas en este Capítulo podrá ser gravada por el Poder Ejecutivo con un impuesto de equiparación de precios, para cumplir con alguna de las siguientes finalidades: a) evitar un perjuicio real o potencial a las actividades productivas que se desarrollaren o hubieren de desarrollarse en un futuro próximo dentro del territorio aduanero; b) asegurar, para la mercadería producida en el territorio aduanero, precios, en el mercado interno, razonables y acordes con la política económica en la materia; c) evitar los inconvenientes para la economía nacional que pudiere llegar a provocar una competencia fuera de lo razonable entre exportadores al país; d) evitar un perjuicio real o potencial a las actividades del comercio interno o de importación que se desarrollaren en el territorio aduanero, cualquiera fuere el origen de la mercadería objeto de las mismas; e) orientar las importaciones de acuerdo con la política de comercio exterior; f) disuadir la imposición en el extranjero de tributos elevados o de prohibiciones a la importación de mercadería originaria o procedente del territorio aduanero; g) alcanzar o mantener el pleno empleo productivo, mejorar el nivel de vida general de la población, ampliar los mercados internos o asegurar el desarrollo de los recursos económicos nacionales; h) proteger o mejorar la posición financiera exterior y salvaguardar el equilibrio de la balanza de pagos.*"

A partir de esto, se establece un arancel fijo del 20% más un arancel móvil específico en función de las cotizaciones del azúcar blanco en la Bolsa de Londres.

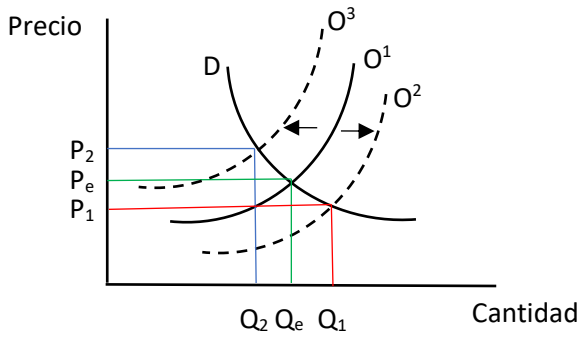
Luego, por el artículo 1° de la Ley N° 25.715 [7], se establece que los aranceles para la importación de las posiciones arancelarias que se detallan en el anexo del presente decreto mantendrán su vigencia mientras no se disponga lo contrario por una ley de la Nación.

1.4. Problema

Dada las políticas estatales establecidas por el estado argentino, el mercado azucarero argentino es un mercado que se encuentra completamente regulado. Esta regulación a las importaciones implica que los ingenios argentinos, fabricas destinadas a la producción de azúcar, sean los encargados de producir todo el azúcar que se consume en el país.

Además, asumiendo una demanda interna constante debido a la poca variación registrada en la última década como se mencionó anteriormente. Y siguiendo la ley de oferta y demanda introducida por la escuela neoclásica, este escenario lleva a pensar que el precio del azúcar depende exclusivamente de la oferta de azúcar circulante en el país, es decir de la producción de los ingenios. A mayor oferta circulante, menor será el precio y viceversa.

Figura 2. Ley de Oferta y Demanda



Como se puede observar en la figura anterior, existen tres escenarios posibles:

- Un primer escenario de equilibrio, en el cual los productores nacionales (oferta 1) producen exactamente la cantidad demandada por el mercado interno.
- Un segundo escenario en el cual los productores nacionales (oferta 2) producen más de lo demandado por el consumo interno. En este escenario, el precio tiende a disminuir y el excedente producido es exportado. En la práctica, es habitual que todos los ingenios comuniquen el stock que tienen en sus almacenes para así poder planificar la cantidad a exportar manteniendo los volúmenes necesarios para abastecer el consumo interno.
- Un tercer escenario en el cual los productores nacionales (oferta 3) producen menos de lo demandado por el consumo interno. En este escenario, el precio tiende a aumentar y el faltante para abastecer la demanda interna será importado de productores externos.

Esta situación, plantea la posibilidad de poder inferir el precio del azúcar a través de la producción resultante de los productores nacionales. Lo que lleva a plantear la siguiente pregunta, ¿Cómo se puede estimar la producción total nacional?

Como se mencionó anteriormente, el azúcar es un producto que surge de la extracción del jugo de la caña de azúcar. Dependiendo de las condiciones específicas de la caña, esta produce un jugo con mayor o menor rendimiento. Entonces, se puede decir que, la producción de azúcar depende pura y exclusivamente de la producción (cantidad) y el rendimiento (calidad) de la caña.

Esta última, proviene de la explotación de los campos del país. Suponiendo que la cantidad de hectáreas producidas de caña de azúcar se mantiene estable durante los años, el factor clave que determina cuanto se producirá y la posible calidad que puede llegar a tener la misma, es el clima.

A partir de esto, se concluye que la producción de azúcar va a depender de las condiciones climáticas que afectan a la producción de caña. Es decir que, a través de datos climáticos es posible determinar la cantidad de azúcar producida en el país, y, por ende, a partir de esta predicción se podrá además determinar el precio del azúcar en el mercado interno.

1.5. Objetivo

Al momento de realizar este trabajo, existen otros trabajos relacionados a la predicción del precio del azúcar en países como México y Brasil.

Peralta González, J. A. (2021), estimó modelos de Box-Jenkins tanto para el precio nacional de México como el internacional. El modelo que mejor pronosticó el precio nacional del azúcar fue un modelo estacional de doce meses, sin componente autorregresivo, de orden uno para el orden de integración y el componente de promedios móviles tanto en el término estacional como en el no estacional. Para el precio internacional en su término no estacional fue de orden uno en su componente autorregresivo y de integración y en su término estacional fue de orden uno en su componente de integración y de promedios móviles. El error de pronóstico (MAPE) fue 5.8% y 7.2%, respectivamente. [8]

Silva RF, Barreira BL, Cugnasca CE, aplicaron modelos de aprendizaje automático para predecir los precios diarios del maíz y el azúcar en Brasil en comparación con el uso de modelos econométricos tradicionales. Los siguientes modelos fueron implementados y comparados: ARIMA, SARIMA, regresión por vectores de soporte (SVR), AdaBoost y redes de memoria a largo y corto plazo (LSTM). Se observó que, los modelos que presentaron los mejores resultados fueron: SVR, un conjunto de los modelos SVR y LSTM, un conjunto de los modelos AdaBoost y SVR, y un conjunto de los modelos AdaBoost y LSTM. Los modelos econométricos presentaron los peores resultados para ambos productos en todas las métricas consideradas (MAE, MSE y R2). [9]

Si bien, en otros países, se comenzaron a realizar algunos pronósticos sobre el precio del azúcar, el marco político y económico del sector azucarero argentino plantea un escenario completamente diferente.

Dada las condiciones microeconómicas en las que se encuentra el mercado azucarero argentino y suponiendo que la cantidad de hectáreas destinados a la producción de caña de azúcar no aumentan, y el consumo interno de azúcar aumenta conforme aumenta la población nacional; comportamiento habitual según conocimiento de expertos, una proyección de la cantidad y calidad de caña que producirá el país aplicando técnicas estadísticas y de machine learning, pueden dar un indicio de hacia dónde se moverá el precio del azúcar.

El objetivo de este trabajo consiste en aplicar técnicas estadísticas y de machine learning sobre datos históricos climáticos de las distintas regiones destinadas a la producción de caña de azúcar en Argentina, para así poder predecir la cantidad de azúcar que se producirá, y luego pronosticar e inferir hacia donde se moverá el precio de esta última.

Dado que, en Argentina, aún no se están aplicando metodologías para la predicción del precio del azúcar, esta información puede resultar de mucha utilidad a la hora de tomar decisiones estratégicas en un ingenio, aumentando el poder de negociación del mismo y fortaleciendo su negocio de compraventa. Contar con información relevante puede cambiar significativamente las dinámicas de negociación, permitiendo a una parte obtener mejores condiciones contractuales. En el caso de un ingenio, la capacidad de predecir precios con precisión podría permitirles negociar contratos de venta y compra más favorables, reduciendo riesgos y aumentando márgenes de beneficio.

1.6. Organización del trabajo

En el capítulo 2 de este trabajo se expondrá el origen de los datos utilizados para el logro del objetivo descripto anteriormente, detallando como se obtuvieron los mismos y un

posterior análisis que validará en primera instancia las hipótesis sobre las que se basa este trabajo. De esta manera, en el capítulo 3 se detallará el marco teórico sobre el cual se sustentaron los procedimientos que se llevaron a cabo a lo largo de este proyecto.

A partir de allí, en los capítulos 4 y 5 se describen de manera exhaustiva los preprocesamientos realizados sobre los datos obtenidos, así como los procesos realizados para el entrenamiento de aquellos modelos estadísticos elegidos y sus consecuentes rendimientos.

Por otra parte, en el capítulo 6 se explicará la elección de los modelos que mejor se ajustan a los problemas planteados y la ejecución de estos en datos no observados, con la finalidad de validar la robustes de los mismos.

Por último, en el capítulo 7 se expone la conclusión a la cual se arribó en función de los procesos efectuados, como también una sección de puntos de mejora para futuras investigaciones.

2. Obtención y comprensión de los datos

2.1. Origen de los datos

Como ya se mencionó en la sección anterior, el objetivo de este trabajo es predecir el precio del azúcar en el mercado interno argentino. Para ello, es necesario contar con tres bases de datos:

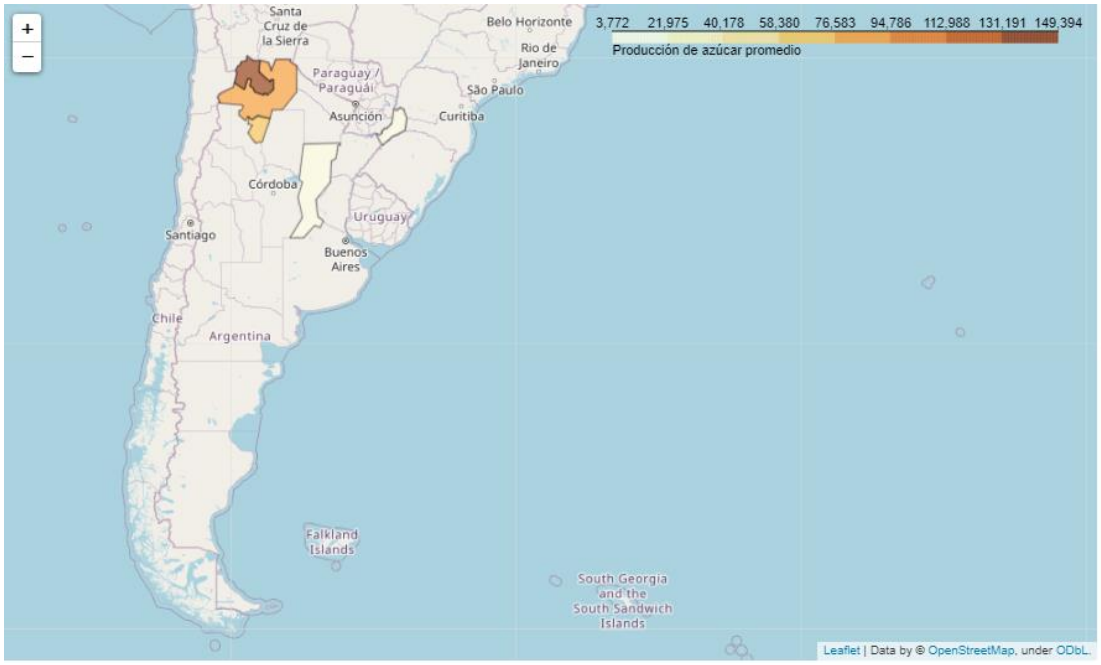
1. Datos Climáticos: Datos relacionados al clima de las regiones que producen caña de azúcar en Argentina. Estos se relacionarán con la base de datos de producción de azúcar.
2. Datos de Producción: Datos relacionados a la producción de azúcar de cada ingenio del país. A través de esta base de datos, relacionada con los datos climáticos, se predecirá la producción nacional de azúcar que luego será relacionada con la base de datos del precio.
3. Datos del Precio: Datos relacionados al precio del azúcar, variable a la cual se quiere arribar.

En base a esto, se procedió con la búsqueda y obtención de las mismas. Desde la página oficial del Centro Azucarero Argentino [2], se obtuvieron los datos anuales tanto de producción de caña, como de azúcar desde 1990 hasta 2020. En el base de datos se pueden observar las siguientes variables de cada año:

- Provincia
- Ingenio
- Azúcar Total Producido (tn)
- Total Caña Molida (tn)
- Rendimiento %
- Total producción azúcar (TMCV)
- Rendimiento %.1

Debido a que no es de esperar que cada productor cañero se vea afectado por el mismo clima, a partir de las variables “Provincia” e “Ingenio”, se determinó cuáles son las regiones de argentinas en donde se produce la caña de azúcar. En los datos obtenidos, el “Litoral” (región argentina conformada por varias provincias) era considerado como una provincia. Es por esto que, en base a la ubicación de los ingenios, se le asignó la provincia correspondiente. El ingenio San Javier corresponde a la provincia de Misiones y los demás ingenios del litoral a la provincia de Santa Fe.

Figura 3. Producción de azúcar promedio por Provincia



Como se puede observar en el gráfico, las provincias de Jujuy, Salta, Tucumán, Santa Fe y Misiones, son las únicas provincias en el país donde se produce azúcar. Es de esperarse que la producción de la caña destinada a la elaboración de azúcar se lleve a cabo en localidades aledañas a los ingenios. De hecho, según un informe del Ministerio de Hacienda de la Nación [3], publicado en junio de 2018, muestra como la producción de caña y de azúcar se distribuye en las provincias arriba mencionadas.

A partir de esto, se solicitó al Servicio Meteorológico Nacional [10], datos históricos diarios desde 1990 hasta la fecha, de todas las localidades pertenecientes a las provincias de Jujuy, Salta, Tucumán y El Litoral (Santa Fe y Misiones). La base de datos obtenida cuenta con las siguientes variables:

1. Fecha: Esta columna representa la fecha en la que se registraron los datos climáticos. Proporciona un marcador temporal para cada conjunto de observaciones y es crucial para analizar las tendencias y variaciones climáticas a lo largo del tiempo.
2. Máxima: Indica la temperatura máxima registrada en un día específico. Es el valor más alto alcanzado durante ese período de tiempo y suele ocurrir durante las horas más cálidas del día, generalmente en la tarde.
3. Mínima: Representa la temperatura mínima registrada en el mismo periodo de tiempo. Es el valor más bajo alcanzado, generalmente durante las horas más frías del día o de la noche.
4. Media: Refiere a la temperatura media, que es el promedio de la temperatura registrada en un día. Proporciona una medida representativa de la temperatura diaria.
5. Precipitación: Esta columna indica la cantidad de precipitación registrada en un periodo específico, generalmente expresada en milímetros. La precipitación puede incluir lluvia, nieve u otras formas de humedad que caen del cielo.
6. Hum. Relat.: Muestra el nivel de humedad relativa del aire en porcentaje. La humedad relativa es la proporción de vapor de agua presente en el aire en comparación con la cantidad máxima que podría contener a una temperatura dada.

7. Nubosidad: Indica el grado de cobertura del cielo por nubes. Puede proporcionar información sobre la cantidad de luz solar que llega a la superficie terrestre.
8. Viento Max. (Dirección): Representa la dirección desde la cual proviene la ráfaga de viento más intensa registrada durante un periodo de tiempo determinado.
9. Viento Max. (Intensidad): Indica la velocidad máxima del viento registrada durante el periodo de tiempo especificado.
10. Viento Medio (Intensidad): Representa la velocidad promedio del viento durante el periodo de tiempo registrado.

Por último, se obtuvieron los datos diarios relacionados al precio del azúcar desde junio 2007 hasta noviembre 2022 desde la página del Centro de Agricultores Cañeros de Tucumán [11]. La base de datos obtenida cuenta con las siguientes variables:

11. Fecha: Fecha diaria en la cual se recopilo el precio del azúcar.
12. 50Kg: Precio del azúcar por 50 kilogramos.
13. 1Kg: Precio del azúcar por 1 kilogramos.
14. s/IVA: Precio del azúcar por 1 kilogramos sin tener en cuenta el impuesto al valor agregado (IVA - 21% en nuestro país).
15. T/C: Tipo de cambio del día relacionado.
16. U$/K: Precio del azúcar por 1 kilogramos expresado en dólares sin tener en cuenta el impuesto al valor agregado (IVA).

Nota: No se pudieron obtener precios anteriores. Los mismos se encontraban en un libro el cual nunca fue provisto.

2.2. Análisis exploratorio de los datos

Una vez obtenidas las bases de datos, se procedió a analizar las mismas con la finalidad de tener una leve comprensión del comportamiento de las variables a predecir. Además, es de esperarse que se deban realizar ciertos ajustes en las bases de datos, ya que es común en la práctica encontrar valores atípicos y es de vital importancia manejar estos con cautela para lograr una predicción más robusta.

Datos de Producción

Lo primero que se observó fue la existencia nombres de ingenios y de provincias los cuales se cargaron con diferentes formatos, llevando a que el lenguaje de programación utilizado los considere como valores distintos. Los valores modificados fueron los siguientes:

- “Tucumán” = “Tucumán”
- “Río Grande” = "Rio Grande"
- “Santa Bárbara” = "Santa Barbara"
- “Tabacal ” = "Tabacal"

Luego, se realizaron los siguientes gráficos para poder visualizar el comportamiento de la producción a lo largo de los años:

Figura 4. Producción de azúcar por año en Argentina

Figura 5. Caña molida por año en Argentina

Figura 6. Rendimiento promedio de caña por año

Como se puede observar, la producción de azúcar tiene un comportamiento similar a la cantidad de caña molida desde 1990 hasta 2017.

En este último año y en los siguientes, debido al aumento en la cantidad de caña molida, se esperaba que la producción aumente. Sin embargo, esta disminuyo drásticamente. Esto se ve explicado por el bajo rendimiento que tuvo la caña durante esos años (7-8%).

Al analizar los gráficos anteriores se puede validar en primera instancia que la producción de azúcar depende tanto de la producción (cantidad) y el rendimiento (calidad) de la caña.

Datos Climáticos:

Dentro de la base de datos del clima existen valores \N en determinadas fechas. Se considera que no se guardaron esos datos en ese instante. Es por esto que se procedió a reemplazar esos valores por la media entre el día anterior y posterior a los cuales se guardó registro. Por ejemplo, en el siguiente extracto de la base de datos, se observa que el 18 de abril de 1990 no se registró la temperatura máxima, por lo que le asignamos un valor de 19,55 correspondiente a la media entre el día anterior (23,7) y el día posterior (15,4).

Tabla 1. Extracto de data set climático

Fecha	Maxima	Minima	Media
17/4/1990	23,70	11,00	14,80
18/4/1990	\N	8,60	14,00
19/4/1990	15,40	9,00	11,60

Datos del Precio:

Al analizar la base de datos del precio, se identificó que las únicas variables a utilizar serán la fecha y el precio por kilogramo de azúcar sin IVA. Además. se observó que existen ciertos días con omisión de datos. El organismo encargado de la recopilación de los datos dispuso asignar ceros a todas las columnas del archivo. Debido a esto, se procedió a eliminar esos valores ya que afectan las fechas de la base de datos al momento de realizar el análisis.

Luego, se realizó un gráfico del precio diario con la finalidad de analizar el movimiento del precio de una manera visual.

Figura 7. Precio de azúcar con irregularidades

A partir del grafico anterior y un posterior análisis de los datos, se puede observar que existen ciertas fechas las cuales fueron cargadas erróneamente. Debido a esto se identificó que:

- En enero de 2009 existían días que se registraron con fecha en julio 2009.
- Ciertos registros de febrero de 2019 fueron cargados con fecha de febrero 2018.
- Registros de enero y febrero de 2022 fueron cargados con fecha de enero y febrero 2021.

Es por esto que se modificaron esas fechas para contar con una base de datos uniforme.

Figura 8. Precio de azúcar sin irregularidades

Correlación:

Una vez que se limpiaron todas las bases de datos de las irregularidades observadas, se realizó un mapa de calor sobre la correlación existente entre el precio promedio anual y la producción de azúcar anual. Este gráfico indicará las relaciones existentes entre las diferentes variables analizadas.

En estadística existen distintas formas de calcular el coeficiente de correlación entre variables. Para este trabajo se utilizó el método de correlación de Spearman.

El coeficiente de correlación de Spearman es una medida no paramétrica de la fuerza y dirección de la asociación entre dos variables clasificadas. Evalúa qué tan bien se puede describir la relación entre dos variables utilizando una función monótona, independientemente de si es lineal o no.

El coeficiente de correlación de Spearman intenta medir:

- Relación Monótona: Evalúa si la relación entre dos variables tiende a aumentar o disminuir juntas, sin requerir que la relación sea lineal. En otras palabras, evalúa si una variable tiende a cambiar consistentemente a medida que la otra variable cambia, incluso si la relación no es estrictamente lineal.
- Fuerza de la Asociación: Cuantifica la fuerza y dirección de la relación monótona entre las variables. Un valor cercano a +1 indica una relación monótona positiva perfecta (a medida que una variable aumenta, la otra también aumenta), un valor cercano a -1 indica una relación monótona negativa perfecta (a medida que una variable aumenta, la otra disminuye), y un valor cercano a 0 indica ninguna relación monótona.
- Análisis Basado en Rangos: Opera en los valores clasificados de las variables en lugar de sus valores brutos. Esto lo hace robusto a los valores atípicos y menos sensible a la distribución específica de los datos en comparación con las medidas de correlación paramétricas como el coeficiente de correlación de Pearson.

Figura 9. Matriz de correlación de variables a predecir

Los puntos más importantes del mapa de calor anterior son los siguientes:

- Existe una fuerte relación positiva (0.74) entre la producción de azúcar y el rendimiento de la caña, indicando que a medida que aumente el rendimiento de la caña, aumentará la producción de azúcar. Al depender el rendimiento exclusivamente de las condiciones climáticas a las cuales se vean afectadas las cañas de azúcar, valida en primera instancia la hipótesis de que el clima es el determinante de la cantidad de azúcar producida en el país.
- Existe una fuerte relación negativa (-0.70) entre la producción de azúcar y el precio de la misma, indicando que a medida que aumenta la producción de azúcar, el precio de esta disminuirá. Esta fuerte relación negativa, valida en primera instancia nuestra hipótesis, basada en la ley de oferta y demanda, de que la oferta de azúcar de los productores nacionales es la principal responsable del precio de la misma.

3. Modelos y métricas de evaluación a utilizar

3.1. Aprendizaje Supervisado

Suponiendo que se observa una variable cuantitativa Y y p diferentes predictores X1, X2, ..., Xp. Se puede asumir que hay alguna relación entre Y y X = (X1, X2, ..., Xp), que puede escribirse de la siguiente forma:

$$Y = f(X) + \varepsilon \quad (1)$$

Aquí, f es alguna función fija pero desconocida de $X1, ..., Xp$, y ε es un término de error aleatorio que es independiente de X y tiene media cero. En esta formulación, f representa la información sistemática que X proporciona sobre Y.

$X1, ..., Xp$ son variables de entrada mientras que Y es una variable de salida. Para los primeros, se utilizarán diferentes nombres como predictores, variables independientes, características o, en ocasiones, solo variables. Para el último, se hará referencia a la respuesta, objetivo o variable dependiente, ya que su valor no es independiente, sino que depende de los valores de $X1, ..., Xp$.

En estadística, aquellos problemas que cuentan con la estructura mencionada recientemente caen dentro del dominio del aprendizaje supervisado. Para cada observación de la(s) medición(es) del predictor x_i, $i = 1, ..., n$, hay una medición de respuesta asociada y_i. Se desea ajustar un modelo que relacione la respuesta con los predictores, con el objetivo de predecir con precisión la respuesta para observaciones futuras (predicción) o comprender mejor la relación entre la respuesta y los predictores (inferencia). [12]

3.1.1. Predicción

En muchas situaciones, un conjunto de entradas X está fácilmente disponible, pero la salida Y no puede obtenerse fácilmente. En este escenario, dado que el término de error se promedia a cero, se puede predecir Y usando:

$$\hat{Y} = \hat{f}(X) \quad (2)$$

donde $\hat{f}$ representa la estimación para f y $\hat{Y}$ representa la predicción resultante para Y. En este contexto, $\hat{f}$ a menudo se trata como una caja negra, en el sentido de que típicamente no es relevante la forma exacta de $\hat{f}$ siempre que produzca predicciones precisas para Y.

La precisión de $\hat{Y}$ como predicción para Y depende de dos cantidades, a las que se denominan error reducible y error irreducible. En general, $\hat{f}$ no será una estimación perfecta para f, y esta inexactitud introducirá algún error. Este error es reducible porque potencialmente se puede mejorar la precisión de $\hat{f}$ utilizando la técnica de aprendizaje estadístico más apropiada para estimar f. Sin embargo, incluso si fuera posible formar una estimación perfecta para f, de modo que la respuesta estimada tomara la forma $\hat{Y} = f(X)$, esta predicción aún tendría algún error. Esto se debe a que, Y también es una función de ε, que, por definición, no se puede predecir usando X. Por lo tanto, la variabilidad asociada con ε también afecta la precisión de las predicciones. Esto se conoce como el error irreducible, porque no importa cuán bien se estime f, no se puede reducir el error introducido por ε. [12]

3.1.2. Inferencia

A menudo las personas están interesadas en comprender la forma en que, *Y* se ve afectada a medida que *X1, ..., Xp* cambian. En esta situación, se desea estimar *f*, pero el objetivo no es necesariamente hacer predicciones para *Y*. En cambio, se desea entender la relación entre *X* e *Y* o, más específicamente, entender cómo *Y* cambia como función de *X1, ..., Xp*. Ahora, $\hat{f}$ no puede tratarse como una caja negra, porque es necesario conocer su forma exacta. En este escenario, uno puede estar interesado en responder las siguientes preguntas:

- ¿Qué predictores están asociados con la respuesta? Es común que solo algunas variables estén significativamente asociadas con *Y*.
- ¿Cuál es la relación entre la respuesta y cada predictor? Algunos predictores pueden tener una correlación positiva con la variable dependiente, mientras que otros pueden tener una correlación negativa. Además, es probable que las magnitudes varíen dentro de un conjunto de diferentes predictores.
- ¿Se puede resumir adecuadamente la relación entre *Y* y cada predictor utilizando una ecuación lineal, o es la relación más complicada? Comúnmente se desea que la correlación sea explicada con precisión por una función lineal, ya que esto proporciona una comprensión fácil del impacto de cada predictor. Sin embargo, la realidad a menudo no es tan simple, y se requieren modelos más flexibles. [12]

3.1.3. Balance entre precisión en la predicción e interpretabilidad del modelo

Dependiendo de si el objetivo final es la predicción, la inferencia, o una combinación de ambos, diferentes métodos para estimar *f* pueden ser apropiados. Los modelos lineales permiten una inferencia relativamente simple e interpretable, pero pueden no ofrecer predicciones tan precisas como algunos otros enfoques. En contraste, algunos de los enfoques altamente no lineales pueden proporcionar predicciones bastante precisas para *Y*, pero esto se logra a expensas de un modelo menos interpretable para el cual la inferencia es más desafiante. [12]

3.2. Problemas de Regresión vs Clasificación

Las variables pueden caracterizarse como cuantitativas o cualitativas (también conocidas como categóricas). Las variables cuantitativas toman valores numéricos. Por ejemplo, la edad, altura o ingreso de una persona. En contraste, las variables cualitativas toman valores de una de las *K* diferentes clases o categorías. Por ejemplo, el género de una persona (varón o mujer) o la marca de un producto comprado (marca A, B, C). En la práctica es habitual referirse a problemas con una respuesta cuantitativa como problemas de regresión, mientras que aquellos que involucran una respuesta cualitativa a menudo se denominan problemas de clasificación. [12]

3.3. Describiendo el problema de este trabajo

Como ya se mencionó anteriormente, el objetivo de este trabajo consiste en aplicar técnicas estadísticas y de machine learning sobre datos históricos climáticos de las distintas regiones destinadas a la producción de caña de azúcar en Argentina, para así poder predecir la cantidad de azúcar que se producirá, y luego así poder pronosticar hacia donde se moverá el precio de esta última.

Si bien, el objetivo final de este trabajo es determinar el precio que tendrá el azúcar en el futuro, por simplicidad vamos a dividir el problema en 2:

- Problema 1: Predicción de la cantidad de azúcar producida a partir de datos climáticos.
- Problema 2: Predicción del precio del azúcar a partir de datos sobre la producción.

Al analizarlos, se puede identificar fácilmente que se está hablando de **problemas de regresión** con la finalidad de **predecir** dentro del dominio del **aprendizaje supervisado**. Son problemas de regresión, ya que en ambos casos se busca determinar un valor cuantitativo (cantidad y precio). Encuadran dentro del aprendizaje supervisado, ya que en ambos casos la finalidad es encontrar el valor de una variable dependiente (problema 1: cantidad producida; y problema 2: precio) a partir de variables independientes (problema 1: datos climáticos; y problema 2: producción de azúcar). Y, si bien el objetivo es predecir esas variables cuantitativas, con la finalidad de brindar información útil para la toma de decisiones estratégicas, además se buscará inferir el movimiento del precio.

3.4. Técnicas estadísticas

En estadísticas ningún método domina a todos los demás sobre todos los posibles conjuntos de datos. En un conjunto de datos particular, un método específico puede funcionar mejor, pero algún otro método puede funcionar mejor en un conjunto de datos similar pero diferente. Por lo tanto, es una tarea importante decidir para cualquier conjunto de datos dado qué método produce los mejores resultados. Seleccionar el mejor enfoque puede ser una de las partes más desafiantes de realizar.

Es por ello, que este trabajo se basa en la comparación de distintas estrategias con la finalidad de encontrar aquella que mejor se adapte a los problemas planteados. Para una mejor compresión, en las siguientes subsecciones se explicará solo el funcionamiento de aquellas que fueron utilizadas.

3.4.1. Precisión del modelo

Para evaluar el rendimiento de un método de aprendizaje estadístico en un conjunto de datos dado, es necesario alguna forma de medir qué tan bien coinciden realmente sus predicciones con los datos observados. Es decir, necesitamos cuantificar en qué medida el valor de respuesta predicho para una observación dada se acerca al valor de respuesta real para esa observación. [12]

3.4.1.1. Error Cuadrático Medio

En el entorno de regresión, la medida más utilizada es el Error Cuadrático Medio o MSE por sus siglas en inglés (Mean Squared Error), el cual se calcula a través de la siguiente función:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2 \qquad (3)$$

donde $\hat{f}(x_i)$ es la predicción para la i-ésima observación. El MSE será pequeño si las respuestas predichas están muy cerca de las respuestas reales y será grande si, para algunas observaciones, las respuestas predichas y reales difieren sustancialmente.

El MSE se calcula utilizando los datos de entrenamiento que se utilizaron para ajustar el modelo, por lo que más precisamente debería referirse como el MSE de entrenamiento. Pero en general, realmente no importa qué tan bien funciona el método en los datos de entrenamiento. Más bien, lo que importa es la precisión de las predicciones que se obtienen cuando se aplica el método a datos de prueba previamente no vistos, que no se utilizaron para entrenar el método de aprendizaje estadístico. Se busca elegir el método que dé el MSE de prueba más bajo, en lugar del MSE de entrenamiento más bajo. [12]

3.4.1.2. Raíz del Error Cuadrático Medio

Una métrica alternativa al MSE, es la Raíz del Error Cuadrático Medio o RMSE por sus siglas en inglés (Root Mean Squared Error). Esta se calcula como la raíz cuadrada del MSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2} \tag{4}$$

A diferencia del MSE, el RMSE se expresa en las mismas unidades que la variable objetivo, lo que facilita su interpretación. Cuanto menor sea el valor del RMSE, mejor será la capacidad predictiva del modelo, ya que indica que las predicciones del modelo están más cerca de los valores reales.

3.4.1.3. Error Porcentual Absoluto Medio

A pesar de que el MSE es una métrica robusta para evaluar la precisión de un modelo, desafortunadamente no es intuitivo para comunicar resultados a la hora de tomar decisiones estratégicas. Esto se da principalmente porque este depende de la escala, un error del 10% en una predicción de alto valor impactará más en el MSE que un error del 10% en una predicción de bajo valor. Por lo tanto, en este trabajo, no solo se concentrará en medir el MSE, sino que también en medir métricas más robustas como el Error Porcentual Absoluto Medio o MAPE por sus siglas en inglés (Mean Absolute Percentage Error). La misma está definida por la siguiente función:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|100 * \frac{y_i - \hat{f}(x_i)}{y_i}\right| \tag{5}$$

Las métricas basadas en porcentajes tienen la ventaja de no tener unidades, y, por lo tanto, se utilizan con frecuencia para comparar el desempeño de los pronósticos entre conjuntos de datos. Sin embargo, también es importante destacar que MAPE, presenta dos principales desventajas:

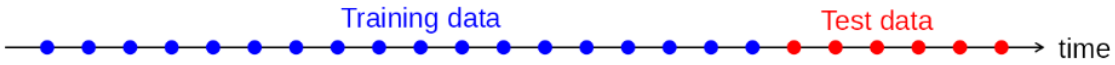
- Tiende a ser infinita o indefinida si $y_i = 0$ para cualquier i en el período de interés, y tiende valores extremos si cualquier y_i está cerca de cero.
- Impone una penalización más severa en los errores negativos que en los errores positivos. [13]

3.4.1.4. Datos de entrenamiento y de testeo

Al elegir modelos, es práctica común separar los datos disponibles en dos partes: datos de entrenamiento (training set) y datos de testeo (test set), donde los datos de entrenamiento se utilizan para estimar cualquier parámetro del modelo y los datos de prueba se utilizan para evaluar su precisión. Debido a que los datos de prueba no se

utilizan en la determinación del modelo, deberían proporcionar una indicación confiable de qué tan bien es probable que el modelo funcione en nuevos datos.

Figura 10. División de base de datos en entrenamiento y testeo



El tamaño del conjunto de prueba suele ser aproximadamente el 20% del total de la muestra, aunque este valor depende de cuán grande sea la muestra. Idealmente, los datos de testeo deberían ser al menos tan grande como el horizonte de pronóstico máximo requerido. Se deben tener en cuenta los siguientes puntos:

- Un modelo que se ajusta bien a los datos de entrenamiento no necesariamente pronosticará bien.
- Un ajuste perfecto siempre se puede obtener utilizando un modelo con suficientes parámetros.
- Sobre ajustar un modelo a los datos es tan malo como no lograr identificar un patrón sistemático en los datos.

A partir de ahora, se hará referencia a error de entrenamiento a aquellas métricas de precisión calculadas con los datos de entrenamiento, y como error de testeo a aquellas métricas de precisión calculadas con los datos de testeo.

En este trabajo, al igual que en cualquier desarrollo de machine learning, el rendimiento de los modelos será evaluado principalmente con el error de testeo, ya que esto representa cómo el modelo se desempeñará en observaciones desconocidas. [13]

3.4.1.5. El Balance Sesgo – Varianza

El MSE esperado en testeo, para un valor dado x_0, puede descomponerse en la suma de tres cantidades fundamentales: la varianza de $\hat{f}(x_0)$, el sesgo al cuadrado de $\hat{f}(x_0)$ y la varianza de los términos de error ε. Es decir,

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\varepsilon) \qquad (6)$$

Esta ecuación informa que, para minimizar el error esperado de testeo, es necesario seleccionar un método de aprendizaje estadístico que simultáneamente logre una baja varianza y un sesgo bajo. Un punto a tener en cuenta es que, la varianza es inherentemente una cantidad no negativa, y el sesgo al cuadrado también es no negativo. Por lo tanto, se puede ver que el MSE esperado de testeo nunca puede ser inferior a $Var(\varepsilon)$, el error irreducible de la ecuación (1).

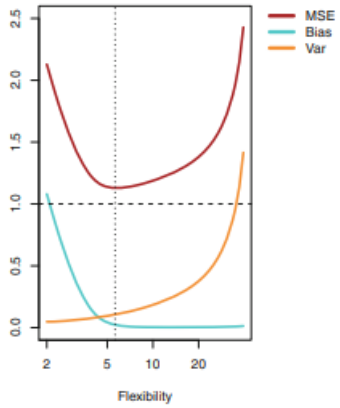
La varianza se refiere a la cantidad por la cual $\hat{f}$ cambiaría si se estima utilizando un conjunto de datos de entrenamiento diferente. Dado que los datos de entrenamiento se utilizan para ajustar el método de aprendizaje estadístico, diferentes conjuntos de datos de entrenamiento resultarán en un $\hat{f}$ diferente. Pero idealmente, la estimación de f no debería variar demasiado entre los conjuntos de entrenamiento. Sin embargo, si un método tiene alta varianza, entonces pequeños cambios en los datos de entrenamiento pueden resultar en grandes cambios en $\hat{f}$. En general, los métodos estadísticos más flexibles tienen una varianza más alta.

Por otro lado, el sesgo se refiere al error que se introduce al aproximar un problema de la vida real, que puede ser extremadamente complicado, por un modelo mucho más

simple. Por ejemplo, la regresión lineal asume que hay una relación lineal entre *Y* y *X1, X2, ..., Xp*. Es improbable que cualquier problema de la vida real tenga una relación lineal tan simple, por lo que realizar una regresión lineal sin duda resultará en algún sesgo en la estimación de *f*. En general, los métodos más flexibles resultan en menos sesgo.

Como regla general, al utilizar métodos más flexibles, la varianza aumentará y el sesgo disminuirá. La tasa relativa de cambio de estas dos cantidades determina si el MSE de prueba aumenta o disminuye. A medida que se aumenta la flexibilidad de una clase de métodos, el sesgo tiende a disminuir inicialmente más rápido que la varianza aumenta. En consecuencia, el MSE de testeo esperado disminuye. Sin embargo, en algún momento, aumentar la flexibilidad tiene poco impacto en el sesgo, pero comienza a aumentar significativamente la varianza. Cuando esto sucede, el MSE de prueba aumenta.

Figura 11. Balance Sesgo – Varianza



La Figura 11 muestra, la forma típica de U obtenida para el MSE. La curva sólida azul representa el sesgo al cuadrado, para diferentes niveles de flexibilidad, mientras que la curva naranja corresponde a la varianza. La línea horizontal punteada representa *Var(ε)*, el error irreducible. Finalmente, la curva roja, correspondiente al MSE del conjunto de testeo, es la suma de estas tres cantidades. A medida que aumenta la flexibilidad, también lo hace la varianza, lo que resulta en un MSE más alto. Esto se llama sobreajuste y es altamente indeseable. Por otro lado, cuando la flexibilidad es baja, el MSE es alto debido a un sesgo alto. Esto se llama subajuste y también es altamente indeseable. La línea de puntos vertical indica el nivel de flexibilidad correspondiente al valor más pequeño del MSE.

Un buen rendimiento en los datos de testeo de un método de aprendizaje estadístico requiere una baja varianza, así como un bajo sesgo al cuadrado. Esto se conoce como un balance porque es fácil obtener un método con sesgo extremadamente bajo, pero alta varianza (por ejemplo, trazando una curva que pase por cada observación de entrenamiento) o un método con varianza muy baja pero alto sesgo (ajustando una línea horizontal a los datos). El desafío radica en encontrar un método para el cual tanto la varianza como el sesgo al cuadrado sean bajos.

Es importante mencionar que el Balance Sesgo – Varianza no está solo presente cuando se calcula la métrica de MSE, sino que es aplicable a otras métricas de precisión estadísticas. [12]

3.4.2. Modelos de Machine Learning

En el marco del aprendizaje automático (Machine Learning), existen numerosos modelos estadísticos que se pueden utilizar tanto para predecir como para inferir una variable.

Durante la realización de este trabajo se utilizaron muchos de ellos, sin embargo, con la finalidad de una mejor compresión sobre el funcionamiento de estos, solo se procederá a describir dos:

- Regresión Lineal: Un modelo sencillo, el cual sirve de base para entender el funcionamiento de los modelos estadísticos.
- XGBoost: Un modelo más flexible que utiliza el método de árboles de decisión basados en la estrategia de boosting, el cual fue aquel que tuvo mejor rendimiento para los problemas planteados en este trabajo.

Para más detalles sobre todos los modelos utilizados durante este trabajo, referirse a la bibliografía tomada como referencia. [12] [13] [14]

3.4.2.1. Regresión Lineal

La regresión lineal es un enfoque muy simple para el aprendizaje supervisado, útil para predecir una respuesta cuantitativa. Aunque puede parecer algo aburrido en comparación con algunos de los enfoques de aprendizaje automático más modernos, la regresión lineal sigue siendo un método estadístico útil y ampliamente utilizado para la inferencia, debido a su alta interpretabilidad y facilidad de comunicación. Dicho esto, la regresión lineal es un excelente punto de partida para comprender qué características clave están involucradas en un problema de negocios no explorado y su impacto potencial positivo o negativo en una variable dependiente.

3.4.2.1.1. Regresión Lineal Simple

La regresión lineal simple es un enfoque muy directo para predecir una respuesta cuantitativa Y en función de una única variable predictora X. Supone que hay aproximadamente una relación lineal entre X e Y. Matemáticamente, se puede escribir esta relación lineal de la siguiente manera:

$$Y \approx \beta_0 + \beta_1 X \tag{7}$$

En la ecuación (7), β_0 y β_1 son dos constantes desconocidas que representan los términos de intercepción y pendiente en el modelo lineal. Juntos, β_0 y β_1 son conocidos como los coeficientes o parámetros del modelo. Una vez que se utilizaron los datos de entrenamiento para producir estimaciones $\widehat{\beta_0}$ y $\widehat{\beta_1}$ para los coeficientes del modelo, se pueden predecir los valores futuros de Y en función de un valor particular de X, mediante el siguiente cálculo:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x \tag{8}$$

donde $\hat{y}$ indica una predicción de Y en función de $X = x$.

El objetivo es obtener estimaciones de coeficientes β_0 y β_1 de tal manera que el modelo lineal se ajuste bien a los datos disponibles, es decir: $\hat{y}_i \approx \widehat{\beta_0} + \widehat{\beta_1} x_i$ para $i = 1, \ldots, n$. En otras palabras, se busca encontrar una intercepción $\widehat{\beta_0}$ y una pendiente $\widehat{\beta_1}$ de tal manera que la línea resultante esté lo más cerca posible de los n puntos de datos. Hay varias formas de medir la cercanía. Sin embargo, de lejos, el enfoque más común implica minimizar el criterio de mínimos cuadrados.

Sea $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ la predicción para Y basada en el i-ésimo valor de X. Entonces $e_i = y_i - \widehat{y_i}$ representa el i-ésimo residuo, es decir, la diferencia entre el i-ésimo valor de respuesta observado y el i-ésimo valor de respuesta que es predicho por el modelo lineal. Se define la suma residual de los cuadrados (RSS) de la siguiente manera:

$$RSS = e_1^2 + e_n^2 + \cdots + e_n^2 \tag{9}$$

3.4.2.1.2. Regresión Lineal Múltiple

La regresión lineal simple es un enfoque útil para predecir una respuesta en función de una sola variable predictora. Sin embargo, en la práctica a menudo se cuenta con más de una variable predictora. La regresión lineal puede extenderse para que pueda acomodar directamente múltiples variables predictoras. Es posible hacer esto asignando a cada variable predictora un coeficiente de pendiente separado en un único modelo. En general, suponiendo que existen p predictores distintos. Entonces, el modelo de regresión lineal múltiple toma la siguiente forma:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{10}$$

donde X_p representa el p-ésimo predictor y β_p cuantifica la asociación entre esa variable y la respuesta. β_p se interpreta como el efecto promedio en Y de un aumento de una unidad en X_p, manteniendo fijos todos los demás predictores.

Como fue el caso en el ajuste de regresión lineal simple, los coeficientes de regresión β_0, β_1, ..., β_p son desconocidos y deben ser estimados. Dados los estimados $\widehat{\beta_0}$, $\widehat{\beta_1}$, ..., $\widehat{\beta_p}$, se puede hacer predicciones usando la siguiente fórmula:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i,1} + \widehat{\beta_2} x_{i,2} + \cdots + \widehat{\beta_p} x_{i,p} \tag{11}$$

Para la estimación de los coeficientes, se puede utilizar el mismo enfoque de mínimos cuadrados utilizado para la regresión lineal:

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \tag{12}$$

3.4.2.2. Métodos basados en arboles

Aunque los modelos de regresión lineal son fáciles de interpretar y, por lo tanto, son apropiados para el análisis de inferencia, su capacidad de predicción es baja en comparación con métodos de aprendizaje automático más sofisticados.

3.4.2.2.1. Arboles de decisión

Los árboles de decisión son un método de aprendizaje automático con una capacidad de predicción baja. Sin embargo, son la base para algoritmos más potentes, como XGBoost. Por lo tanto, se requiere una breve introducción.

Hablando en términos generales, hay dos pasos:

1. Se divide el espacio de predictores, es decir, el conjunto de valores posibles para *X1, X2, ..., Xp*, en *J* regiones distintas y no superpuestas, *R1, R2, ..., RJ*.
2. Para cada observación que cae en la región *Rj*, se realiza la misma predicción, que es simplemente la media de los valores de respuesta para las observaciones de entrenamiento en *Rj*.

Por ejemplo, si en el Paso 1 obtenemos dos regiones, *R1* y *R2*, y la media de respuesta de las observaciones de entrenamiento en la primera región es 10, mientras que la media de respuesta de las observaciones de entrenamiento en la segunda región es 20. Entonces, para una observación dada $X = x$, si $x \in R1$, se predecirá un valor de 10, y si $x \in R2$, se predecirá un valor de 20.

Ahora, ¿Cómo se construyen las regiones R1, ..., RJ? En teoría, las regiones podrían tener cualquier forma. Sin embargo, se elige dividir el espacio de predictores en rectángulos de alta dimensión, o cajas, por simplicidad y para facilitar la interpretación del modelo predictivo resultante. El objetivo es encontrar cajas R1, ..., RJ que minimicen el RSS, dado por:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \widehat{y_{R_1}})^2 \tag{13}$$

donde $\widehat{y_{R_j}}$ es la respuesta media para las observaciones de entrenamiento dentro de la j-ésima caja.

Desafortunadamente, es computacionalmente inviable considerar cada posible partición del espacio de características en J cajas. Por esta razón, se toma un enfoque de arriba hacia abajo y codicioso que se conoce como partición binaria recursiva. El enfoque es de arriba hacia abajo porque comienza en la parte superior del árbol (en ese momento todas las observaciones pertenecen a una sola región) y luego divide sucesivamente el espacio de predictores; cada división se indica mediante dos nuevas ramas más abajo en el árbol. Es codicioso porque en cada paso del proceso de construcción del árbol, se realiza la mejor división en ese paso en particular, en lugar de mirar hacia adelante y elegir una división que conduzca a un árbol mejor en algún paso futuro.

Para realizar la partición binaria recursiva, primero se selecciona el predictor X_j y el punto de corte s de manera que dividir el espacio de predictores en las regiones $R_1(j, s) = \{X|X_j < s\}$ y $R_2(j, s) = \{X|X_j \geq s\}$ conduzca a la mayor reducción posible en el RSS. El RSS queda definido como:

$$RSS = \sum_{i:\, x_i \in R_1(j,s)} (y_i - \widehat{y_{R_1}})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \widehat{y_{R_2}})^2 \tag{14}$$

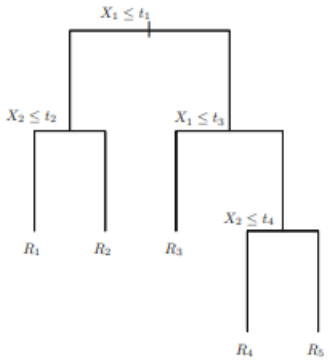
donde $\widehat{y_{R_1}}$ es la respuesta media para las observaciones de entrenamiento en $R_1(j, s)$, y $\widehat{y_{R_2}}$ es la respuesta media para las observaciones de entrenamiento en $R_2(j, s)$.

Luego, se repite el proceso buscando el mejor predictor y el mejor punto de corte para dividir los datos aún más para minimizar el RSS dentro de cada una de las regiones resultantes. Sin embargo, esta vez, en lugar de dividir todo el espacio de predictores, se divide una de las dos regiones identificadas previamente. Ahora se cuenta con tres regiones. Nuevamente, se busca dividir una de estas tres regiones aún más, para minimizar el RSS. El proceso continúa hasta que se cumpla un criterio de detención; por ejemplo, se podría continuar hasta que ninguna región contenga más de cinco observaciones.

Una vez que se han creado las regiones R_1, ..., R_j, se predice la respuesta para una observación de prueba dada utilizando la media de las observaciones de entrenamiento en la región a la que pertenece esa observación de prueba.

En el siguiente gráfico se puede observar un ejemplo de este enfoque con cinco regiones.

Figura 11. Ejemplo de un árbol de decisión de 5 regiones



Se denomina al punto de inicio del árbol como la **raíz**, y a los espacios de predictores finales R_1, ..., R_j, como **nodos terminales** o **hojas**. Los puntos a lo largo del árbol donde se divide el espacio de predictores se denominan **nodos internos**.

3.4.2.2.2. Bagging

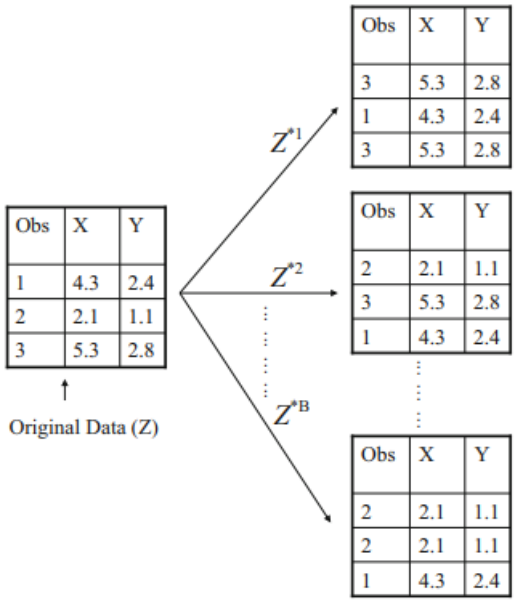
Los árboles de decisión sufren de alta varianza. Esto significa que, se dividen los datos de entrenamiento en dos partes al azar y se ajusta un árbol de decisión a ambas mitades, los resultados que se obtendrán podrían ser bastante diferentes. La agregación bootstrap o bagging, es un procedimiento de propósito general para reducir la varianza de un método de aprendizaje estadístico.

Dado un conjunto de n observaciones independientes Z_1, ..., Z_n, cada una con varianza σ^2, la varianza de la media $\overline{Z}$ de las observaciones se da por σ^2/n. En otras palabras, promediar un conjunto de observaciones reduce la varianza. Por lo tanto, una forma natural de reducir la varianza y, por lo tanto, aumentar la precisión de predicción de un método de aprendizaje estadístico es tomar muchos conjuntos de entrenamiento de la población, construir un modelo de predicción separado utilizando cada conjunto de entrenamiento, y promediar las predicciones resultantes. En otras palabras, se podría calcular $\widehat{f^1}(x)$, $\widehat{f^2}(x)$, ..., $\widehat{f^B}(x)$ usando B conjuntos de entrenamiento separados, y promediarlos para obtener un único modelo de aprendizaje estadístico de baja varianza, dado por:

$$\widehat{f_{avg}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f^b}(x) \tag{15}$$

Por supuesto, esto no es práctico porque generalmente no se cuenta con acceso a múltiples conjuntos de entrenamiento. En cambio, es posible realizar un método llamado bootstrap, en el cual se toman muestras repetidas del conjunto de datos de entrenamiento (único). La siguiente figura muestra un ejemplo con una base de datos reducida de solo 3 observaciones:

Figura 12. Ejemplo del método bootstrap



En este enfoque se generan B conjuntos de datos de entrenamiento bootstrap diferentes. Luego se entrena en el b-ésimo conjunto de datos de entrenamiento bootstrap para obtener $\widehat{f^{*b}}(x)$, y finalmente se promedian todas las predicciones para obtener:

$$\widehat{f_{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f^{*b}}(x) \tag{16}$$

Esto se llama bagging. Si bien puede mejorar las predicciones para muchos métodos de regresión, es particularmente útil para los árboles de decisión. Para aplicar el bagging a los árboles de regresión, simplemente se construyen B árboles de regresión utilizando B conjuntos de datos de entrenamiento bootstrap, y se promedian las predicciones resultantes. Estos árboles se desarrollan en profundidad y no se podan. Por lo tanto, cada árbol individual tiene alta varianza, pero bajo sesgo. El promedio de estos B árboles reduce la varianza. Se ha demostrado que el bagging proporciona mejoras impresionantes en la precisión al combinar cientos o incluso miles de árboles en un solo procedimiento.

3.4.2.2.3. Boosting

Bagging implica crear múltiples copias del conjunto de datos de entrenamiento original utilizando la técnica de bootstrap, ajustar un árbol de decisión separado a cada copia y luego combinar todos los árboles para crear un único modelo predictivo. Es importante destacar que cada árbol se construye en un conjunto de datos bootstrap, independiente de los otros árboles. Por otro lado, el boosting funciona de manera similar, excepto que los árboles crecen de manera secuencial: cada árbol se construye utilizando información de los árboles previamente creados. El boosting no implica muestreo de bootstrap; en cambio, cada árbol se ajusta en una versión modificada del conjunto de datos original.

Al igual que el bagging, el boosting implica combinar un gran número de árboles de decisión $\widehat{f^1}(x)$, ..., $\widehat{f^B}(x)$. En el siguiente algoritmo se observa la descripción del mismo:

Algoritmo 1. Boosting para arboles de Regresión

1. Establecer $\hat{f}(x) = 0$ y $r_i = y_i$ para todos los i en el conjunto de entrenamiento.
2. Para $b = 1, 2, \ldots, B$, repetir:
   a. Ajustar un árbol $\widehat{f^b}$ con d divisiones ($d + 1$ nodos terminales) a los datos de entrenamiento (X, r).
   b. Actualizar $\hat{f}$ agregando una versión reducida del nuevo árbol:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \widehat{f^b}(x)$$

   c. Actualizar los residuos,

$$r_i \leftarrow r_i - \lambda \widehat{f^b}(x_i)$$

3. Generar el modelo potenciado,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \widehat{f^b}(x)$$

A diferencia de ajustar un solo árbol de decisión grande a los datos, lo cual implica ajustar los datos de manera más estricta y potencialmente sobre ajustar, el enfoque de boosting aprende de manera más lenta. Dado el modelo actual, se ajusta un árbol de decisión a los residuos del modelo. Es decir, se ajusta un árbol utilizando los residuos actuales, en lugar del resultado Y, como la respuesta. Luego, se agrega este nuevo árbol de decisión en la función ajustada para actualizar los residuos. Cada uno de estos árboles puede ser bastante pequeño, con solo unos pocos nodos terminales. Al ajustar árboles pequeños a los residuos, se mejora lentamente $\hat{f}$ en áreas donde no funciona bien. El parámetro de contracción λ ralentiza aún más el proceso, permitiendo que más árboles de diferentes formas ataquen los residuos. En general, los enfoques de aprendizaje estadístico que aprenden lentamente tienden a tener un buen desempeño. En el boosting, a diferencia del bagging, la construcción de cada árbol depende fuertemente de los árboles que ya se han creado.

3.4.2.2.4. Hiperparámetros

Cada modelo tiene parámetros de regularización que no se aprenden automáticamente en el proceso de entrenamiento del modelo. En cambio, deben ajustarse en función del problema en cuestión y proporcionarse al algoritmo de entrenamiento. Estos parámetros se llaman hiperparámetros y serán de gran importancia al ajustar modelos basados en árboles.

En Boosting, se utilizan los siguientes [15]:

1. Hiperparámetros que hacen que el modelo sea más flexible, aumentando así la probabilidad de sobreajuste:
   - **nround:** número de árboles.
   - **max_depth:** máxima profundidad que pueden alcanzar los árboles.
   - **eta (λ):** parámetro de contracción.
   - **colsample_bytree:** cantidad de variables a seleccionar aleatoriamente en cada árbol.
   - **subsample:** tamaño de la(s) muestra(s) a extraer en cada árbol.

2. Hiperparámetros que hacen que el modelo sea menos flexible, reduciendo así la probabilidad de sobreajuste:

- **gamma:** reducción mínima del error para realizar una división.
- **min_child_weight:** tamaño mínimo de los nodos terminales.

Con la finalidad de encontrar los mejores hiperparámetros para los problemas de este trabajo, se utilizó lo que se denomina búsqueda aleatoria (Random Search). En una búsqueda aleatoria, los hiperparámetros se eligen aleatoriamente n veces. Después de que el modelo se entrena y se prueba con los n conjuntos de hiperparámetros elegidos al azar, se seleccionan aquellos que ofrecieron el mejor rendimiento.

4. Problema 1: Predicción de la cantidad de azúcar producida a partir de datos climáticos

4.1. Creación de datos de entrenamiento y testeo

Uno de los procesos más importantes durante la elaboración de un modelo estadístico, es la confección de la base de datos la cual será utilizada para entrenar aquel modelo que nos brindará una solución al problema planteado. Es por esto que, gran parte del tiempo dedicado a este trabajo se basó en la creación de estas bases de datos.

Como se mencionó anteriormente, el primer problema se basa en la predicción de la producción anual de azúcar en Argentina. Para ello se obtuvieron, por un lado, los datos climáticos diarios de las regiones que producen caña en el país, y, por otro lado, los datos de producción anual de azúcar en el país a partir de la cantidad de caña molida por ingenio.

Al tener dos bases de datos separadas, y con la desventaja de que una de ellas está basada en registros anuales y la otra en diarios, se planteó la siguiente pregunta. ¿Cómo relacionar ambas bases de datos?

Lo primero a tener en cuenta, es definir la variable dependiente. Es decir, aquella la cual se busca predecir. En este caso, se considera como variable dependiente a la producción de azúcar, ya que, en base al análisis preliminar en la sección 2 de este trabajo, al tener la misma una correlación positiva fuerte con el rendimiento de la caña, y al depender esta última exclusivamente de los datos climáticos, se planteará una relación directa entre la producción de azúcar y ciertas variables climáticas.

Luego, debido a que la variable dependiente esta expresada en unidades anuales en base al tiempo, se procedió a realizar lo que en estadística se denomina ingeniería de atributos. De esta forma, tomando como base los datos con los que se cuenta, se crearon nuevas variables expresadas en la misma unidad de tiempo que la variable dependiente, las cuales formarán parte del conjunto de variables independientes.

Este proceso de creación de nuevas variables se realizó en dos partes. Por un lado, para la creación de las nuevas variables independientes, se consultó a la guía técnica del cañero de la Estación Experimental Obispo Colombres [16]. Por otro lado, se realizaron cálculos matemáticos y estadísticos.

4.1.1. Guía técnica del cañero

El crecimiento y desarrollo de la caña de azúcar se caracteriza por tener cuatro fases:

- Fase de emergencia y establecimiento de la población inicial de tallos: Tradicionalmente, se denomina "**brotación**". Se destaca por la emergencia sucesiva y el mantenimiento temporal de los tallos primarios. Esta fase está condicionada por la temperatura y el comportamiento de las variedades.
- Fase de macollaje y cierre del cañaveral: El "**macollaje**" es una fase de gran importancia en la definición del rendimiento, ya que es el momento en que se establece el número potencial de órganos cosechables (tallos).
  Su principal característica es el rápido aumento de la población total de tallos.
  La intensidad y calidad de la radiación solar incidente ejerce un rol central en la regulación del macollaje. Otros factores destacables son: la temperatura, la disponibilidad de agua y nutrientes (especialmente el nitrógeno), las

características de la variedad, la competencia con las malezas y los efectos de las plagas y enfermedades, entre otros. Esta fase finaliza con el cierre del cañaveral.

- Fase de determinación del rendimiento cultural: El nombre tradicional de esta fase es "**período de gran crecimiento**". En este período se define la producción de caña, al establecerse la población final de tallos molibles y, en gran medida, el peso fresco por tallo. Además, se inicia el almacenamiento de azúcar en los entrenudos, que van completando su desarrollo. En esta fase el cultivo expresa la máxima respuesta a los factores ambientales y de manejo.
  Con el "cierre del cañaveral" (finalización de la fase de macollaje) se desencadena una severa competencia entre los tallos, que ocasiona la muerte de muchos de ellos, provocando una disminución significativa de la población establecida durante el macollaje.
  La fecha de inicio de esta fase, su intensidad y duración, depende de los factores ambientales, que están estrechamente relacionados con la época de plantación o de cosecha del ciclo anterior y por el manejo suministrado al cañaveral.
- Fase de maduración y definición de la producción de azúcar: En esta etapa se define el contenido final de sacarosa en los tallos y la producción de azúcar por unidad de área. Esto va acompañado de una disminución progresiva en el ritmo de crecimiento de los tallos y de la conservación de hojas fotosintetizando, lo que le permite a la caña almacenar una parte importante del azúcar que produce. Esta actividad de las hojas es transitoria, ya que posteriormente las hojas entran en un período de envejecimiento, agudizado por la ocurrencia de bajas temperaturas.

Las fases del ciclo del cultivo ocurren sucesivamente y con una cierta superposición. La duración total del ciclo o de cada fase en particular no es cronológicamente constante, notándose modificaciones entre variedades y por efecto de las condiciones ambientales y del manejo.

4.1.1.1. Requerimientos ambientales para el crecimiento del cultivo

La caña de azúcar se adapta a un amplio abanico de climas tropicales y subtropicales, sin embargo, no tolera temperaturas de congelamiento (bajo 0°C) y el crecimiento prácticamente se detiene por debajo de los 10-12°C. Se desarrolla satisfactoriamente en una variedad de tipos de suelos, aunque los más adecuados son los de textura franca o franco-arcillosa, bien drenados. Al tratarse de un cultivo que extrae grandes cantidades de nitrógeno y potasio del suelo, demanda suelos bien provistos de nutrientes o de alta fertilidad.

En cuanto a la temperatura, cada fase de crecimiento tiene exigencias diferentes. La brotación de las yemas se inicia o activa con temperaturas superiores a 10°C, aunque es lenta hasta los 16-18°C, generalizándose cuando los valores superan los 20°C.

Durante el período de gran crecimiento, el desarrollo vegetativo puede verse afectado por temperaturas inferiores a 16-17°C, siendo óptimos valores entre los 28-35°C. La caña de azúcar puede soportar temperaturas máximas de 45-50°C, aunque esto provoca retrasos en su crecimiento.

La radiación solar, por su incidencia en la fotosíntesis de la planta, es otro factor importante que determina el nivel de crecimiento y la acumulación de materia seca. En general, intensidades crecientes de radiación solar se asocian con aumentos en la producción de caña y de azúcar.

Otro factor decisivo es la disponibilidad de agua. Al ser un cultivo que produce gran cantidad de material vegetal por unidad de superficie, requiere de grandes volúmenes de

agua. El consumo de agua de un cañaveral varía en cada fase de crecimiento, alcanzando los valores máximos durante el llamado "período de gran crecimiento", que tiene lugar entre diciembre y marzo.

En general, en las etapas relacionadas con los procesos de crecimiento del cañaveral (brotación, macollaje y gran crecimiento), para alcanzar altas producciones de caña de azúcar (elevado rendimiento cultural) se necesitan condiciones de luminosidad elevadas, temperaturas entre 30-35º C y buena disponibilidad de agua y nutrientes, especialmente nitrógeno.

Una estación otoño-invernal con baja humedad atmosférica y del suelo, escasas precipitaciones, alta insolación y gran amplitud térmica –con días frescos pero libres de heladas–, resulta óptima para alcanzar elevados contenidos de sacarosa y favorecer una cosecha eficiente y un adecuado transporte de la materia prima.

4.1.2. Ingeniería de atributos

La ingeniería de atributos, también conocida como ingeniería de características, es el proceso de seleccionar y transformar variables o atributos en un conjunto de datos con el objetivo de mejorar el rendimiento y la eficacia de los modelos de aprendizaje automático. Este proceso implica la creación de nuevas características derivadas de las características existentes, la eliminación de características irrelevantes o redundantes, y la transformación de las características de una manera que haga que los datos sean más útiles para el modelo.

La ingeniería de atributos puede incluir tareas como:

- Selección de características: identificar las características más relevantes para el problema en cuestión.
- Creación de nuevas características: combinar o transformar características existentes para crear nuevas que puedan ser más predictivas.
- Transformación de características: normalizar, estandarizar o codificar características categóricas para que sean más adecuadas para el modelo.
- Reducción de dimensionalidad: reducir el número de características mediante técnicas como PCA (Análisis de Componentes Principales) o selección de características.

La ingeniería de atributos es un paso crucial en el proceso de modelado de datos que puede tener un gran impacto en el rendimiento final del modelo de aprendizaje automático.

Como se mencionó anteriormente, para este problema se procedió a la realización de ingeniería de atributos en dos fases:

- Una primera fase, basada en la guía técnica del cañero.
- Una segunda fase, basada en cálculos matemáticos y estadísticos.

4.1.2.1. Primera fase

El proceso de cultivo de caña de azúcar y sus correspondientes fases de crecimiento no serán siempre iguales para todos los cañeros del país en cuanto al tiempo calendario. Para los fines prácticos de este trabajo y basándose en prácticas comunes de la industria, se asumirá que el proceso de cultivo de la caña de azúcar va de julio a junio. Es por esto que se debe considerar que, por ejemplo, la producción de azúcar de 2020 proviene de la caña producida en el periodo que va desde julio 2019 a junio 2020.

Por lo tanto, con fines prácticos y para poder relacionar los datos climáticos con la producción de azúcar, se determinarán los años de julio a junio, es decir, el año 1991 corresponderá a la fecha que va desde 1/7/1990 al 30/6/1991. De esta forma, la base de datos climática estará definida desde 1991 a 2020.

A partir de esto, las fases del ciclo del cultivo estarán definidas en base al tiempo de la siguiente manera:

- Brotación: Julio a agosto.
- Macollaje: Septiembre a octubre.
- Gran Crecimiento: Noviembre a marzo.
- Madurez: Abril a junio.

Una vez definidas las fases de crecimiento en el tiempo, y comprendiendo las necesidades ambientales de cada fase, se procedió a la creación de las siguientes variables para cada una de las etapas en cada año considerado:

- Heladas: Cantidad de heladas durante el período (una helada se considera cuando la temperatura es menor a 0°).
- Detención: Cantidad de días con temperaturas menores a los 10°/12° durante el período (Con este tipo de temperaturas, la caña detiene su crecimiento).
- Optimo: Cantidad de días con temperaturas óptimas. Entre los 30° y 35° durante los periodos de Brotación, Macollaje y Gran Crecimiento; y entre los 12° y 14° en el período de madurez.
- Retrasos: Cantidad de días con temperaturas mayores a los 45° durante el período (Estas son temperaturas que producen retrasos en el crecimiento).
- Agua: Cantidad de agua recibida durante el período. La precipitación mide esta variable. Se realizo una suma de la variable para cada período.
- Max_Acum_Agua: Máxima cantidad de agua acumulada en días consecutivos de lluvia. Esto podría ser un indicio de cuando se producen inundaciones.
- Sin_Agua: Máxima cantidad de días sin agua. Esto podría dar indicios de sequía.
- Vientos_Medio: Cantidad de días con vientos medios de 0 a 1.5, la cual se consideran los vientos óptimos para el crecimiento.
- Vientos_Maximos: Cantidad de días con vientos superiores a 15, lo cual podría producir desprendimientos del cultivo.
- Soleado: Cantidad de días con sol durante el período. Esto se determina a través de la nubosidad. Se puede decir que una nubosidad menor o igual a 3 se considera un día con sol.
- Hum_Rel_Opt: Cantidad de días con Humedad Relativa óptima durante el período. Durante el período de crecimiento rápido, las condiciones de alta humedad (80 - 85%) favorecen una rápida elongación de la caña. Valores moderados, de 45 - 65%, acompañados de una disponibilidad limitada de agua, son beneficiosos durante la fase de maduración.

Al crear 11 nuevas variables para cada fase de crecimiento de la caña, la base de datos ahora estará conformada por 44 variables expresadas en términos anuales con relación al tiempo.

4.1.2.2. Segunda fase

Luego de la creación de las variables climáticas descriptas anteriormente, y con la finalidad de conservar ciertas características estadísticas relacionadas a las variables con las que ya se contaba, se procedió a la creación de variables estadísticas para cada etapa

de la producción de caña de azúcar (Brotación, Macollaje, Gran Crecimiento y Madurez). Las mismas se detallan a continuación:

- Máxima: Valor máximo registrado.
- Mínima: Valor mínimo registrado.
- Media: Promedio de los valores registrados.
- Desvío estándar: Desvío estándar de los valores registrados.
- PCA1: Primer componente principal de los valores registrados.
- PCA2: Segundo componente principal de los valores registrados.

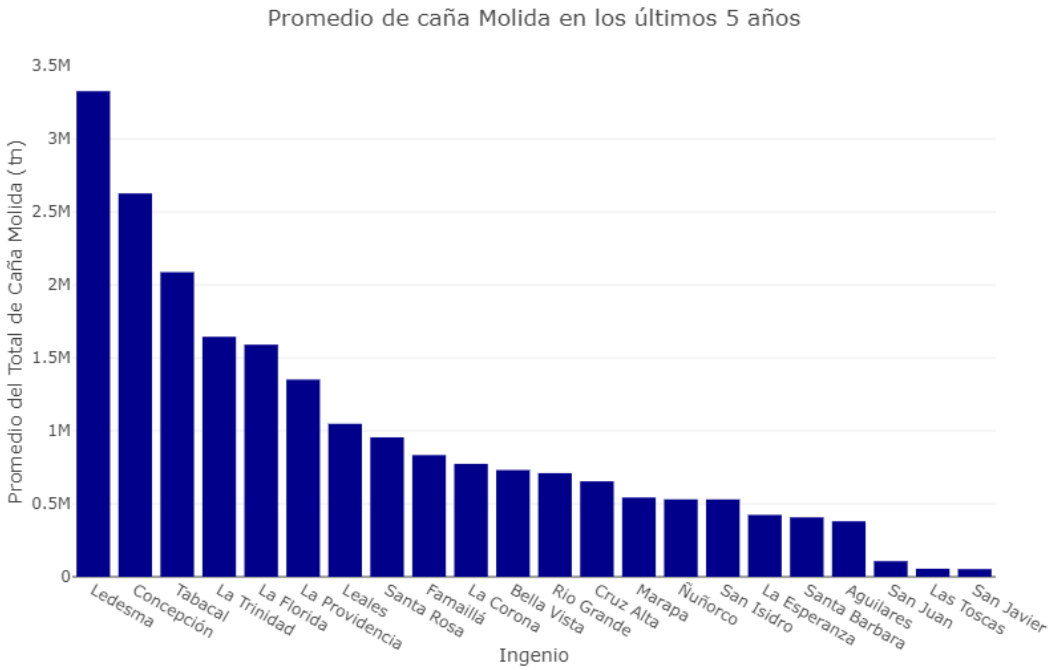
El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica de reducción de la dimensionalidad comúnmente utilizada en aprendizaje automático. Su objetivo es transformar datos de alta dimensionalidad en una representación de menor dimensionalidad mientras se preserva la mayor cantidad de varianza posible. Esto se logra mediante la identificación de los componentes principales, que son las direcciones en los datos que capturan la variabilidad más significativa.

Adicionalmente, el año calendario esta divido de cierta forma el cual existen años de 366 días y años de 365 días. De esta forma, en la base de datos del clima existen años con 366 registros y años con 365. Es por esto que, para captar este escenario, se creó una variable dicotómica llamada “Año bisiesto”, la cual tomará el valor 1 en caso de que se trate de un año bisiesto y 0 en el caso contrario.

Posteriormente, se observó que existen grandes diferencias entra la cantidad de caña de azúcar que recibe cada ingenio, lo que conlleva a una mayor producción de azúcar para aquellos con mayor cantidad de caña. Por otro lado, al momento de realizar la predicción de la producción de cada ingenio, no se cuenta con esta información. Es por esto que, se decidió crear una variable que capte en cierta medida la diferencia de producción que existe entre los distintos ingenios, variable que consideramos será de utilidad para aumentar el rendimiento de la predicción de los modelos. Para ello, le asignamos un numero a cada uno de los ingenios.

El siguiente histograma, muestra la media de caña de azúcar molida de cada ingenio en los últimos 5 años. Se realizó el cálculo tomando en cuenta los últimos 5 años, con la finalidad de captar la información más reciente.

Figura 13. Promedio de caña molida por ingenio en los últimos 5 años



Una vez identificados estos valores, se procedió a asignar un valor del 1 al 100 a cada ingenio, partiendo de la base de que el ingenio que recibió en promedio una mayor cantidad de caña de azúcar (Ledesma), tendrá un valor de 100. De esta forma, quedan conformado los siguientes valores para cada ingenio:

- Ledesma = 100
- Concepción = 79
- Tabacal = 63
- La Trinidad = 49
- La Florida = 48
- La Providencia = 41
- Leales = 32
- Santa Rosa = 29
- Famaillá = 25
- La Corona = 23
- Bella Vista = 22
- Rio Grande = 21
- Cruz Alta = 20
- Marapa = 16
- Ñuñorco = 16
- San Isidro = 16
- La Esperanza = 13
- Santa Barbara = 12
- Aguilares = 11
- San Juan = 3
- Las Toscas = 2
- San Javier = 2
- Inaza S.A. = 0

4.1.3. Separación en datos de entrenamiento, validación y testo

Una vez finalizada la creación de las variables que serán utilizadas para el entrenamiento de los modelos estadísticos, la base de datos queda conformada por 264 columnas que contienen 678 registros con datos que van desde 1991 hasta 2020.

Sin embargo, un último preprocesamiento realizado antes de la separación de los datos para la aplicación de los modelos estadísticos, se basó en eliminar aquellos registros que contaban con una producción de azúcar menor a 1.000 toneladas. Esto se realizó debido a que una producción tan baja solo estaría perjudicando el poder predictivo de los modelos, además de que debido a la constante comunicación e información que se maneja en la industria azucarera, es de esperarse que se conozca con antelación si un ingenio producirá o no. De esta forma, la base de datos queda conformada con 659 registros.

Luego, con la finalidad de poder medir la precisión de los modelos estadísticos, se dividió la misma en tres partes:

- Datos de entrenamiento: El conjunto de datos con los cuales se entrenarán los modelos y estará conformado por todos aquellos registros anteriores al 2017.
- Datos de validación: El conjunto de datos con el cual se medirá la precisión de los modelos y estará conformado por todos aquellos registros que van desde 2017 hasta 2019 inclusive. A partir de los datos de validación se realizará la comparación de los mismos a partir de métricas de medición con la finalidad de seleccionar aquel que se ajuste mejor a al problema planteado.
- Datos de testo: El conjunto de datos con el cual se verificará la verdadera precisión del modelo que mejor rendimiento tuvo en los datos de validación estará conformado por todos aquellos registros pertenecientes al año 2020. A partir de los datos de testeo se podrá inferir sobre el rendimiento que tendría el mejor modelo en datos nunca observados, simulando de esta forma el resultado que se podría obtener en la realidad.

4.2. Ejecución y optimización de modelos

Luego de haber determinado la forma en la cual se procederá a la evaluación de los modelos estadísticos, se procedió al entrenamiento de los mismos.

El objetivo en esta sección no estará relacionada a la elección del mejor modelo, sino a la descripción de cada uno de ellos. A continuación, se presentan los resultados y aquellas características salientes de los mismos.

4.2.1. XGBoost

El modelo de XGBoost (Extreme Gradient Boosting), es un modelo que utiliza los métodos de árboles de decisión basados en la estrategia de boosting.

Para la implementación de este modelo se utilizó la librería “xgboost” versión 1.7.4 [15] de Python definiendo los siguientes parámetros:

- Booster: “gbtree”, lo que define el método de árboles de decisión como base del modelo.
- Objective: “reg:squarederror”, indicándole al modelo que se encuentra frente a un problema de regresión con el objetico de minimizar el error cuadrático medio.

Además, como se mencionó anteriormente, cada modelo estadístico cuenta con hipeparámetros los cuales dependerán del problema específico del que se trate. Para entrenar el modelo de XGBoost, se inició un modelo con hipeparámetros estándares para luego realizar una mejora del poder predictivo a través de técnicas algorítmicas.

Para ello, se definió un algoritmo que pudiera iterar entre las características posibles que podrían tener los hiperparámetros a definir dentro del modelo. Esta forma de optimización de hiperparámetros se basa en establecer para cada uno de ellos una cota máxima y otra mínima de los valores que estos podrían asumir. Una vez determinados estos puntos, el algoritmo ejecutará n modelos combinando de forma aleatoria entre cada uno de los intervalos seleccionados, hasta encontrar la mejor reducción del error (RMSE) entre esas n pruebas. Se ejecutaron 500 modelos, a partir de los siguientes intervalos de hiperparámetros con el objetivo de encontrar la mejor combinación que se ajuste de mejor manera al problema planteado:

- nround: 5 a 1.000
- max_depth: 1 a 10
- eta (λ): 0,0025 a 0,3
- gamma: 0 a 5
- colsample_bytree: 0,7 a 1,0
- min_child_weight: 1 a 10
- subsample: 0,75 a 1,0

Una vez ejecutado el algoritmo, se obtuvo la siguiente tabla con los modelos que mejor rendimiento tuvieron en los datos de validación, juntos con aquellos hiperparámetros óptimos que se ajustan mejor a estos datos.

Tabla 2. Rendimiento de modelos XGBoost

iteracion	nround	max_ depth	eta	gamma	colsample _bytree	subsa mple	min_child _weight	perf_tr	perf_vd
439	82	2	0,22349	2,05148	0,99391	0,89213	5,42927	17.993,80	20.140,86
284	455	8	0,12856	1,20765	0,9005	0,76629	1,23995	2.422,48	21.178,28
220	162	5	0,10258	2,49140	0,70333	0,89773	7,71376	6.977,99	21.301,84
275	419	10	0,17367	3,57254	0,80287	0,83941	8,08936	2.640,50	21.744,23
336	469	1	0,07520	3,25274	0,89337	0,86923	9,94767	26.678,88	21.796,87

Como se puede observar, el modelo entrenado en la iteración número 439 del algoritmo es aquel que tiene un menor RMSE (última columna de la Tabla 2) en los datos de validación. Un punto importante a tener en cuenta es que, en la Tabla 2, se puede visualizar que el RMSE de entrenamiento es similar al de validación. Esto da un indicio de que no existe sobreajuste del modelo.

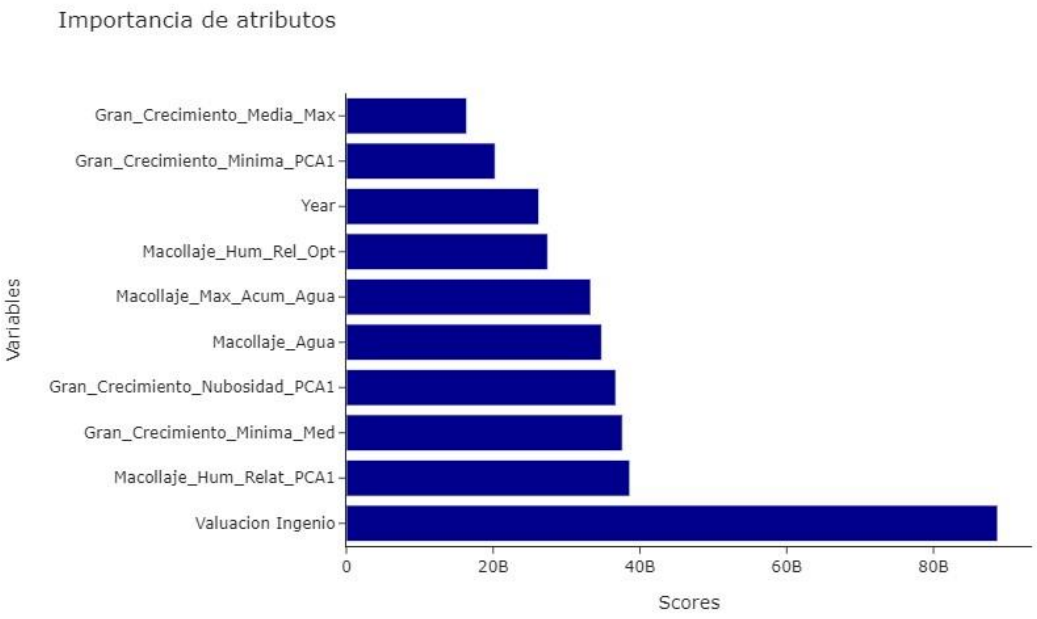
El tiempo estimado que llevo la ejecución de este algoritmo para entrenar el mejor modelo de XGBoost fue de 50 minutos aproximadamente.

Además, calculamos la segunda métrica a tener en cuenta para este trabajo:

MAPE = 29,40 %

Luego, a partir del modelo seleccionado, se analizan aquellas variables que aportan un mayor poder predictivo en el siguiente gráfico:

Figura 14. Importancia de variables de XGBoost



El histograma anterior muestra las 10 variables más importantes. Se puede observar que la variable que más aporta a la predicción del modelo es la valuación de los ingenios. Un detalle no menor a tener en cuenta es que, las variables que se pueden visualizar se crearon tanto en la primera como en la segunda fase de ingeniería de atributos.

4.2.2. Máquinas de Vector Soporte

El modelo de máquinas de vector soporte o SVM por sus siglas en inglés (Support Vector Machines), es un algoritmo de aprendizaje supervisado cuyo objetivo principal es encontrar un hiperplano en un espacio de alta dimensión que se ajuste a los puntos de datos lo más cerca posible. El objetivo es ajustar el hiperplano de manera que se maximice la distancia entre el hiperplano y los puntos de datos más cercanos, mientras se minimiza la desviación de los puntos de datos.

Este modelo, utiliza el concepto de kernel para transformar el espacio de características original en un espacio de mayor dimensión donde los datos pueden ser separados de manera lineal. Esto permite al modelo SVM manejar relaciones no lineales entre las características de entrada y la variable de respuesta.

El modelo de SVM fue implementado con la librería “sklearn” versión 1.0.2 [17] de Python, con el método regression.SVR y el kernel “rbf”, utilizados comúnmente en problemas de regresión.

Por último, al igual que en el modelo anterior, SVM cuenta con un hiperparámetro de regularización, que controla la compensación entre el ajuste del modelo a los datos de entrenamiento y la suavidad de la solución. Un valor más alto significa que se permite un mayor ajuste a los datos de entrenamiento. Por el contrario, un valor más bajo penalizará más fuertemente la complejidad del modelo. De la misma forma que en XGBoost, se realizó un algoritmo que permitiera iterar entre distintos valores del hiperparámetro, con la finalidad de encontrar aquel que sea óptimo para el problema planteado. El intervalo elegido para este hiperparámetro va desde 0,01 a 30.000.000.

Una vez ejecutado el algoritmo y luego de 500 iteraciones, se obtuvo la siguiente tabla:

Tabla 3. Rendimiento de modelos SVM

iteracion	C	perf_tr	perf_vd
369	29.942.960,00	37.602,89	31.046,14
413	29.824.120,00	37.604,58	31.050,36
443	29.823.640,00	37.604,59	31.050,38
200	29.822.960,00	37.604,60	31.050,40
2	29.797.740,00	37.604,96	31.051,30

Como se puede observar, el modelo entrenado en la iteración número 369 del algoritmo es aquel que tiene un menor RMSE (última columna de la Tabla 3) en los datos de validación. Si bien se tomaron valores del hiperparámetro C muy elevados, existiendo así una alta posibilidad de sobre ajustar el modelo, en la Tabla 3 se puede visualizar que el RMSE de entrenamiento es similar al de validación.

El tiempo de ejecución para lo obtención del mejor modelo de SVM fue de 2 minutos aproximadamente.

Por último, se calcula la segunda métrica a tener en cuenta para este trabajo:

MAPE = 36,92 %

4.2.3. Redes Neuronales

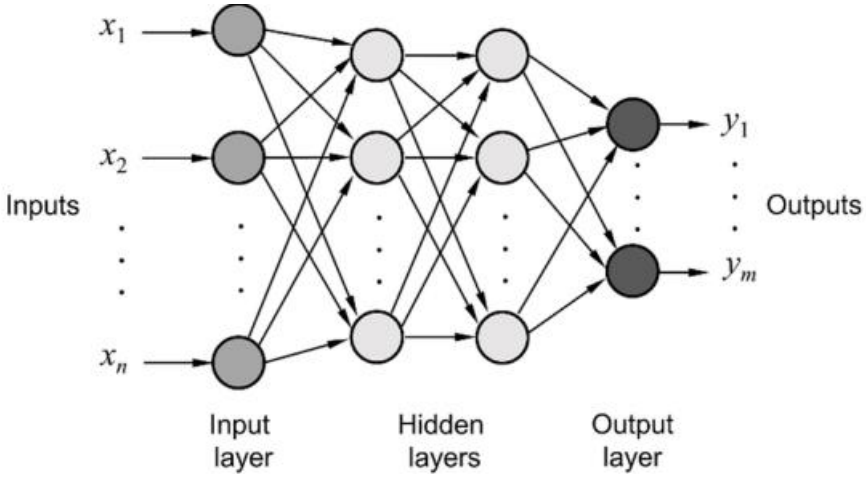
Las redes neuronales son modelos computacionales inspirados en el funcionamiento del cerebro humano. Están compuestas por unidades de procesamiento llamadas neuronas, que están interconectadas y organizadas en capas. Cada neurona realiza una operación simple en los datos de entrada y transmite su resultado a las neuronas de la capa siguiente.

Una red neuronal típica consta de tres tipos principales de capas:

- Capa de entrada: Recibe los datos de entrada y los transmite a la red.
- Capas ocultas: Realizan transformaciones no lineales en los datos. Estas capas son responsables de extraer características complejas de los datos.
- Capa de salida: Produce la salida final de la red neuronal.

Durante el entrenamiento, la red neuronal ajusta los pesos de conexión entre las neuronas para minimizar una función de pérdida, que mide la discrepancia entre las predicciones del modelo y los valores verdaderos de los datos de entrenamiento.

Figura 15. Estructura típica de redes neuronales



Dependiendo de cada problema específico, se define la estructura de la red neuronal. Para este problema, al igual que en los modelos anteriores se busca minimizar el Error Cuadrático Medio a través de la siguiente estructura:

- Una capa de entrada densa con 64 neuronas y una función de activación ReLU.
- Dos capas ocultas densas con 128 y 64 neuronas respectivamente y una función de activación ReLU.
- Una de salida densa lineal con 1 neurona.

Para la definición de la arquitectura del modelo se utilizó la librería keras versión 2.15.0 [18] de Python, con el método sequential. Además, se utilizó la librería tensorflow versión 2.15.0 [19] de Python, para definir la función de activación ReLU y el optimizador RMSprop a utilizar durante el entrenamiento del modelo. Este último es un algoritmo de optimización basado en gradiente para actualizar los pesos de la red neuronal durante el entrenamiento, el cual definimos en un valor de 0,01.

Las métricas obtenidas luego del entrenamiento de la red neuronal fueron las siguientes:

RMSE = 28.416,72

MAPE = 38,02%

Además, el tiempo de ejecución de este entrenamiento fue de 10 segundos aproximadamente.

4.2.4. Comparación de rendimiento de modelos problema 1

Luego del entrenamiento de los modelos anteriores, es necesario compararlos a través de todas las métricas definidas al inicio de este trabajo. A continuación, se puede observar la tabla 4 con esta comparativa:

Tabla 4. Comparativa resultados modelos predicción de la producción. Mejores corridas.

Método	Librería	Tiempo de Ejecución	RMSE Test	MAPE Test
XGBoost	xgboost	50 min.	20.140,86	29,40%
SVM	sklearn.SVR	2 min.	31.046,14	36,92%
Redes Neuronales	keras & tensorflow	10 seg.	28.416,72	38,02%

A través de la tabla comparativa anterior, es fácil identificar que el modelo entrenado de XGBoost es aquel que menor RMSE y MAPE tiene en los datos de validación. Sin embargo, el modelo de redes neuronales cuenta con un tiempo de ejecución mucho menor, siendo además este el segundo con menor RMSE y MAPE. Por último, el modelo de máquinas de vector soporte es el que peor rendimiento tuvo en todas las métricas anteriores.

Para la realización del cálculo de estas métricas, se utilizó la librería "sklearn" versión 1.0.2 [17] de Python con el método metrics.

Un punto importante a tener en cuenta es que, los modelos entrenados no tienen un buen rendimiento promedio en términos porcentuales. XGBoost, aquel que mejor rendimiento tuvo en términos de precisión cuenta con un MAPE del 29,40%. Una de las razones por las cuales los modelos pueden estar arribando a estos valores es la poca cantidad de datos existentes para el entrenamiento de los mismos.

Por otro lado, como se pudo visualizar en análisis anteriores, existe una diferencia considerable en la producción de los distintos ingenios. Si bien, en la figura 14 se pudo observar que la variable de valuación de ingenio es aquella que más poder predictivo aporta al modelo de XGBoost, un valor elevado de la métrica MAPE puede indicarnos que el modelo no capta con claridad la diferencia de producción que existe entre cada uno de los ingenios. Esta situación podría solucionarse separando la base de datos entre los distintos ingenios, o en su defecto en grupos de ingenios con características de producción similares separados por categorías, entrenar un modelo para cada categoría y luego promedia el error de cada uno de ellos. Método para el cual seria necesario una mayor cantidad de registros.

A pesar de esto, dado que, para la predicción del precio del azúcar, objetivo final de este trabajo, solo se toma en cuenta la producción total del país, el error porcentual de cada observación no se considera relevante.

5. Problema 2: Predicción del precio del azúcar a partir de datos sobre la producción

5.1. Creación de datos de entrenamiento y testeo

Al igual que en el problema 1 de este trabajo, para la predicción del precio del azúcar también fue necesario realizar cierto procesamiento en los datos obtenidos.

Un factor importante a tener en cuenta en la predicción del precio es que, Argentina es un país con una inestabilidad económica constante. El fenómeno inflacionario por el que atraviesa el país es un gran determinante en el aumento generalizado de los precios. Es por esto que, para contar con una predicción que sea de mayor utilidad en términos de negocios, se obtuvo desde la página del Banco de la Nación Argentina [20] la cotización del dólar diaria desde 2007 hasta 2023, y se transformó la base de datos en pesos a dólares. Si bien el poder adquisitivo de la moneda estadounidense es superior a la del peso argentino, esto no implica que Estados Unidos no cuente con inflación. Sin embargo, esta última es mucho menor a la observada en Argentina. A pesar de que la base de datos que se obtuvo inicialmente contaba con una variable con los precios en dólares, en la misma existían muchos registros que no contaban con esta información es por esto que se procedió a la transformación recientemente descripta.

Además, debido a que la producción de azúcar del problema anterior esta expresada en toneladas, se multiplicó por 1.000 al precio de la base de datos. Es decir que ahora, el precio esta expresado en dólares por tonelada.

Luego, por practicidad en la realización de los procedimientos que se explicaran a continuación, se calculó el precio promedio por mes. De esta forma, el comportamiento del precio del azúcar se puede visualizar de la siguiente manera:

Figura 16. Precio de Azúcar en Dólares por Tonelada

Se puede observar que el precio lateraliza desde 2007 hasta 2009. A partir de allí, este aumenta llegando a máximos históricos hasta mediados del 2011, donde se produce una ruptura y cambio de tendencia ascendente a descendiente, llegando a mínimos históricos en 2016 (en base a los registros con los que contamos). Luego, se puede identificar que el precio oscila entre los 250 y 500 dólares.

Por último, con la finalidad de unir la producción de azúcar en Argentina con el precio del mismo, se procedió a agrupar la producción de cada ingenio durante todo un año calendario. Además, por conocimiento de la industria, la producción de azúcar va desde mayo a octubre (6 meses), por lo que se asume que la producción de mayo se inserta en el mercado en junio impactando en el precio del producto.

Debido a esto, se dividió la producción anual del país en los 6 meses de producción, quedando conformada una base de datos con inserción de azúcar al mercado únicamente en los meses de junio, julio, agosto, septiembre, octubre y noviembre (al resto de los meses se le asigno un valor de 0 en concepto de producción). La división antes mencionada, se realizó en partes iguales.

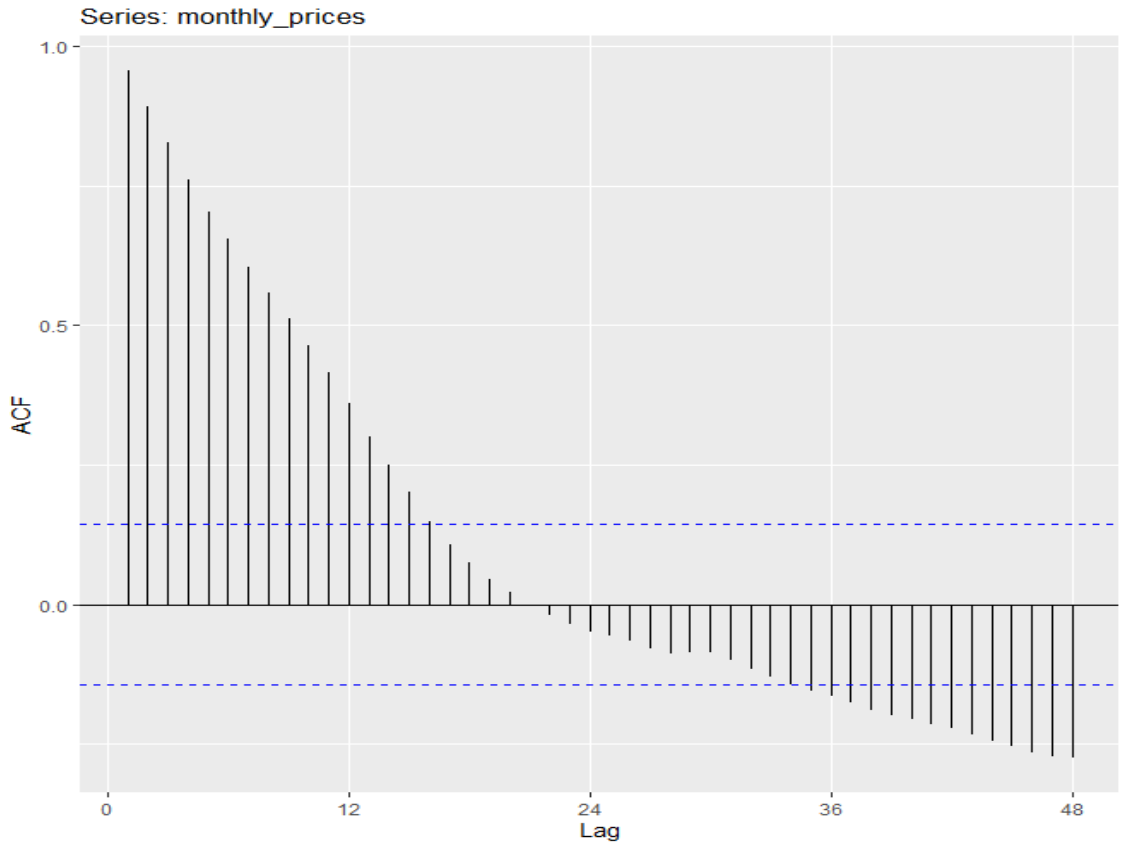
5.1.1. Análisis de autocorrelación

Previo al proceso de ingeniería de atributos, es de utilidad realizar un análisis de autocorrelación en el precio ya que:

- La autocorrelación puede revelar patrones temporales en los datos, lo que significa que el precio en un período de tiempo está relacionado con el precio en períodos anteriores. Esto puede ser útil para detectar estacionalidad, tendencias u otros patrones que pueden influir en el precio futuro.
- Si se encuentran patrones significativos de autocorrelación en las características utilizadas para predecir el precio, esto puede influir en la selección de características. Puede ser beneficioso incluir características que capturen la autocorrelación en el modelo para mejorar su capacidad predictiva.
- Al comprender la autocorrelación en los datos, los modelos pueden ser ajustados para tener en cuenta esta estructura temporal, lo que puede mejorar la precisión de las predicciones. Los modelos que ignoran la autocorrelación pueden subestimar o sobreestimar el precio futuro.

A continuación, se puede observar el grafico de autocorrelación de la serie temporal:

Figura 17. Autocorrelación del precio del azúcar



Los valores de ACF (Autocorrelation Function) en varios períodos representa el grado de correlación entre la serie de tiempo y los valores pasados.

Un valor de 0.98 en el primer período indica una fuerte correlación positiva. Esto sugiere que el precio de un momento determinado está muy correlacionado con el precio del periodo anterior. En otras palabras, existe una fuerte relación lineal entre las observaciones adyacentes.

El suave decrecimiento de los valores de ACF indican una correlación continua en los patrones de la serie de tiempo. Los valores permanecen altos durante muchos períodos antes de decaer gradualmente.

El valor negativo de -0.25 en el período 48, indica una moderada correlación negativa en ese período. Esto sugiere que el precio de un momento determinado, esta correlacionado negativamente con el precio existente en 48 períodos anteriores. Puede existir una relación negativa retrasada entre los precios de estos períodos.

Las bandas de ACF de 0.125 y -0.125 representan los intervalos de confianza. Los periodos cuyos valores de ACF caen dentro de esas bandas se consideran estadísticamente no significativos. Aquellos períodos con valores por fuera de esas bandas se consideran estadísticamente significativos.

5.1.2. Ingeniería de atributos

Al igual que en el problema anterior y luego del análisis de autocorrelación, se procedió a la elaboración de aquellas variables que se consideran ayudaran a aumentar el poder predictivo de los modelos estadísticos.

Las nuevas variables creadas son las siguientes:

- Variables relacionadas a la fecha: Se extrajeron variables a partir de la fecha.
    - Mes
    - Quarter
    - Año
- Variables temporales: Se crearon variables que capturen de cierta forma la temporalidad en los datos. Para ello se transformó la variable Mes con las funciones del seno y coseno, para mapear cada mes a un punto en un círculo unitario. De esta manera, enero (1) y diciembre (12), que tienen solo un mes de diferencia, estarán cerca en el espacio transformado.
    - Mes_Sin
    - Mes_Cos
- Variables históricas: se crearon variables para capturar los valores históricos tanto del precio como de la producción.
    - Lag1_Price
    - Lag2_Price
    - Lag3_Price
    - Lag1_Produc
    - Lag2_Produc
    - Lag3_Produc

  Si bien a partir del análisis de autocorrelación anterior, se podría incluir una mayor cantidad de valores históricos, se considera que 3 son suficientes para aportar poder predictivo a los modelos. Además, no se incluyeron aquellos valores históricos con correlación negativa ya que son muy alejados en el tiempo y no se cuenta con una gran cantidad de registros.

  Para los primeros 3 registros de datos, no se cuenta con los valores históricos de los meses anteriores, por lo que se procedió a asignarle el valor del primer mes de registro, U$D 259.8 para los precios de los 3 meses anteriores. En cuanto a los valores de la producción de azúcar, se le asignó un valor de 0 a los 3 meses anteriores debido a durante los meses de marzo, abril y mayo siempre será ese valor (aún no comienza la producción de azúcar).
- Variables estadísticas: Se crearon variables estadísticas que capturen cierta tendencia y patrones de los datos históricos anteriormente creados.
    - Rolling_Mean_Price: Promedio del precio de los últimos 3 meses.
    - Rolling_Std_Price: Desvío estándar del precio de los últimos 3 meses.

  Solo se crearon variables sobre los precios, ya que los valores de producción no cuentan con estos patrones estadísticos debido a los supuestos aplicados al dividir la producción anual en partes iguales.
- Interacción de variables: Se creó una variable multiplicando el precio y la producción con la finalidad de capturar posibles relaciones.
    - Price_Prod_Interact
- Variable de variación: Se creó una variable que capture la variación que sufrió el precio en el periodo anterior con la finalidad de capturar momentum en los datos.
    - Price_RoC

Un punto importante a tener en cuenta luego de haber realizado el cálculo de las nuevas variables recientemente mencionadas es que, al haber realizado cálculos de datos históricos, al momento de predecir un valor del precio, es necesario calcular nuevamente las variables relacionadas al precio con la finalidad de no tomar valores que no se conocerían en la vida real y así poder predecir nuevamente. No se realizaron nuevos cálculos relacionados a las variables de producción, ya que esta estrategia se basa en

conocer la misma (anteriormente calculada). Para ello se elaboró una función que será utilizada en la predicción de los modelos en los datos de validación y posteriormente en los de testo.

5.1.3. Separación en datos de entrenamiento, validación y testo

Ya que, en el primer problema, se dejó una porción de los datos para predecir la producción de azúcar del 2020, y con la finalidad de definir el mejor modelo para la serie de tiempo de precios, se separó la base de datos de la siguiente manera:

- Datos de entrenamiento: El conjunto de datos con los cuales se entrenarán los modelos y estará conformado por todos aquellos registros hasta junio de 2019.
- Datos de validación: El conjunto de datos con el cual se medirá la precisión de los modelos y estará conformado por todos aquellos registros que van desde julio hasta diciembre de 2019 inclusive. Debido a que el objetivo final de este problema será predecir el precio de los próximos 6 meses, se considera que estos registros son adecuados para el proceso de validación.
- Datos de testeo: El conjunto de datos con el cual se verificará la verdadera precisión del modelo con mejor rendimiento en los datos de validación, y estará conformado por todos aquellos registros pertenecientes al año 2020.

5.2. Ejecución y optimización de modelos

Al igual que en el problema 1, luego de haber realizado los preprocesamientos necesarios, se procedió al entrenamiento de los modelos estadísticos.

El objetivo en esta sección no estará relacionada a la elección del mejor modelo, sino a la descripción de cada uno de ellos. A continuación, se presentan los resultados y aquellas características salientes de los mismos.

5.2.1. Autoregresión

Como ya se describió en la sección 3 de este trabajo, la regresión lineal es un enfoque muy simple para el aprendizaje supervisado, útil para predecir una respuesta cuantitativa. Es un método estadístico utilizado para modelar la relación entre una variable dependiente y una o más variables independientes. El objetivo de la regresión lineal es encontrar la mejor línea recta que se ajuste a los datos observados, de modo que pueda usarse para predecir los valores de la variable dependiente para nuevas observaciones basadas en los valores de las variables independientes.

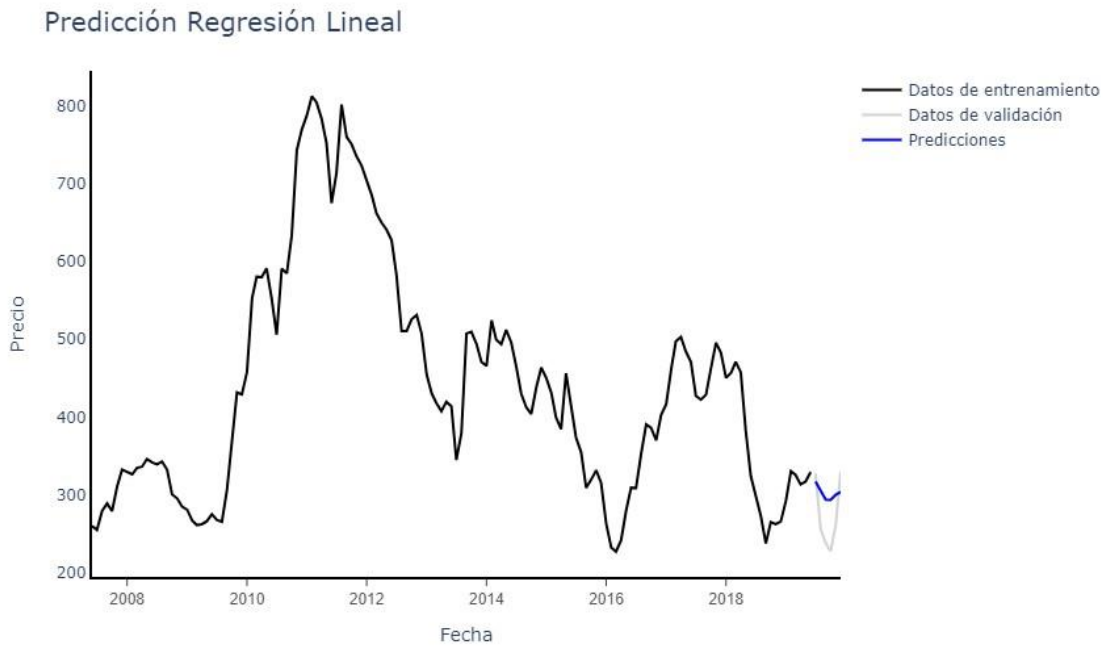
Como es de esperarse para el problema planteado, la variable dependiente será el precio del azúcar en dólares por tonelada tanto para este modelo como para los que se describirán en las siguientes subsecciones.

Ahora bien, en un contexto de series temporales, y como se muestra en la figura 17 es de esperarse cierta relación lineal entre el valor actual de la serie de tiempo y sus valores pasados. Es por esto que, al crearse las variables históricas descriptas en la sección anterior, se entrenó un modelo de autoregresión de orden 3 (AR(3)).

Para el entrenamiento de este modelo, se utilizó la función lm() versión 4.1.2 del lenguaje de programación R. El tiempo de ejecución del entrenamiento de este modelo fue de 1 segundo.

Una vez entrenado el modelo de regresión lineal con los datos de entrenamiento, se procedió a la predicción en los datos de validación. Luego se realizó el siguiente gráfico con la finalidad de visualizar el rendimiento del mismo.

Figura 18. Rendimiento Regresión Lineal



Por último, se realizó el cálculo de las métricas elegidas para la comparación de los modelos elegidos:

RMSE = 45,00

MAPE = 16,25 %

5.2.2. Regresión Exponencial

La regresión exponencial es una técnica estadística utilizada para modelar relaciones entre variables donde se observa un crecimiento o decrecimiento exponencial en los datos. A diferencia de la regresión lineal, que modela relaciones lineales, la regresión exponencial modela relaciones no lineales que se ajustan mejor a una curva exponencial.

En la regresión exponencial, el modelo asume que la variable dependiente varía exponencialmente con respecto a una o más variables independientes.

Al igual que en el modelo de regresión lineal, para el entrenamiento de este modelo, se utilizó la función lm(), con la única diferencia de que a la variable dependiente se le aplico el logaritmo natural a través de la función log(), tratando de captar ese movimiento exponencial en el precio. El tiempo de ejecución del entrenamiento de este modelo fue de 1 segundo.

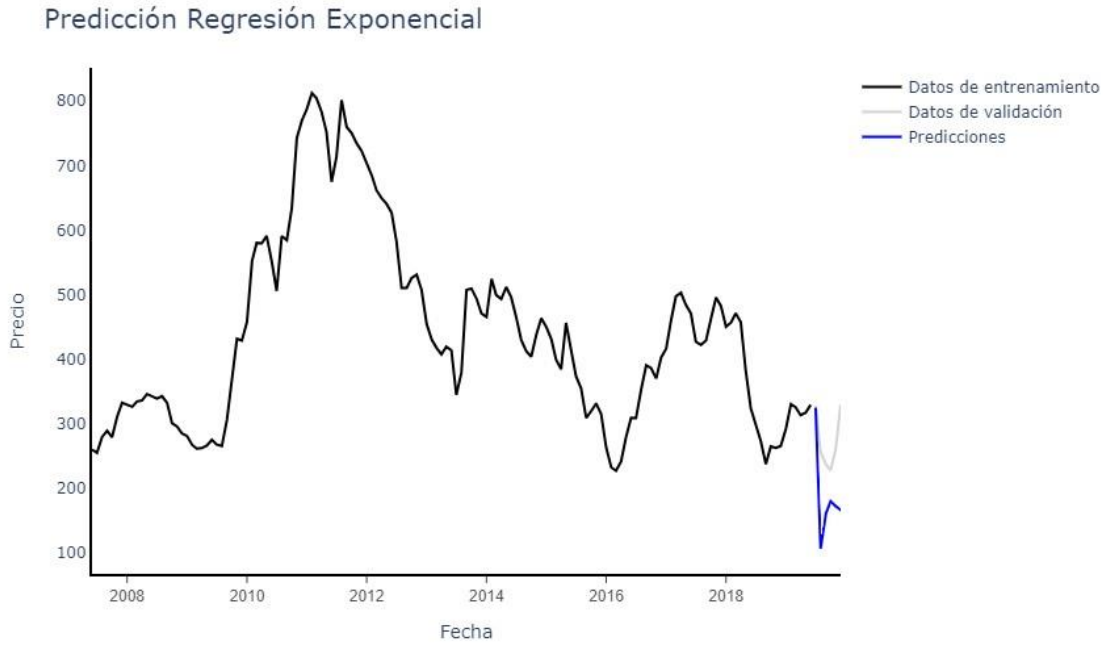
Luego de entrenar el modelo exponencial y predecir en validación, se obtuvieron las siguientes métricas:

RMSE = 104,31

MAPE = 32,71 %

Además, se puede visualizar el rendimiento en el siguiente gráfico:

Figura 19. Rendimiento Regresión Exponencial



5.2.3. PieceWise

El modelo Piecewise, o "por partes" en español, es una técnica utilizada en análisis de regresión para modelar relaciones que cambian de forma abrupta o no lineal a lo largo de diferentes rangos de valores de la variable independiente. En lugar de asumir una relación lineal o exponencial en todo el rango de la variable independiente, el modelo Piecewise divide el rango en segmentos o "piezas" y aplica un modelo de regresión diferente a cada segmento.

La idea detrás del modelo Piecewise es capturar la variación en el comportamiento de los datos a lo largo de diferentes rangos de la variable independiente. Cada segmento puede tener su propia pendiente, intercepto u otra forma funcional de relación entre la variable dependiente y la variable independiente.

Basándose en el gráfico del precio de este problema, se tomó como punto de división el año 2012, ya que es el año en el que se observa una disminución del precio máximo histórico.

La misma función para entrenar los modelos anteriores se utilizó en el modelo de PieceWise, lm(), con la diferencia de que se reemplazó la variable “Fecha” por las variables del tiempo confeccionadas por el punto de división anterior. El tiempo de ejecución del entrenamiento de este modelo fue de 1 segundo.

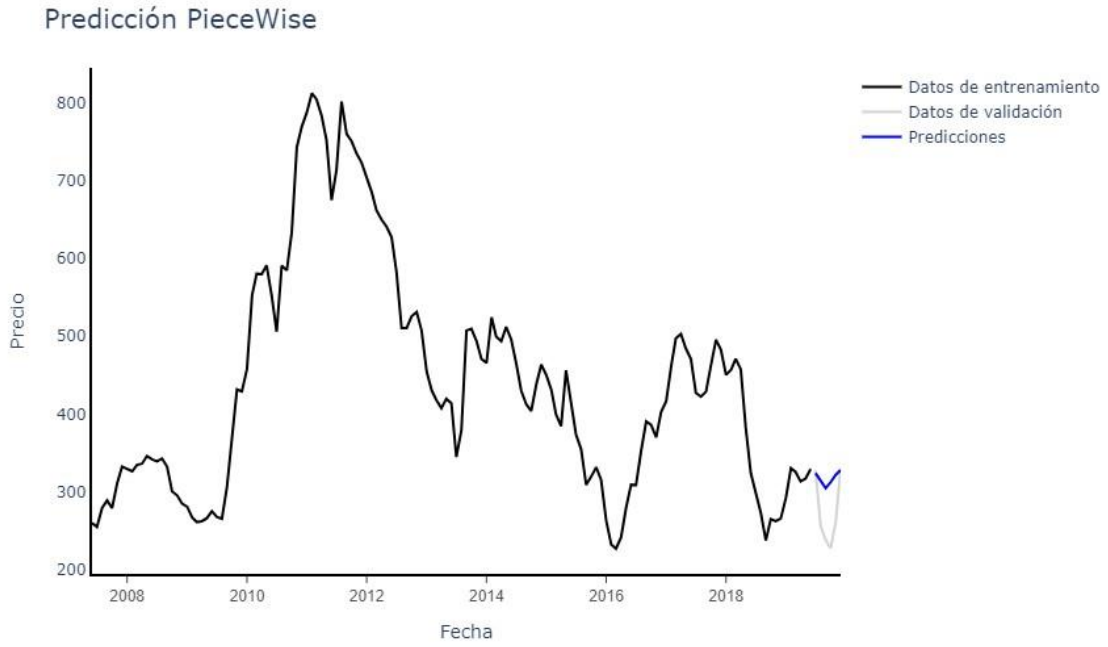
Una vez entrenado el modelo, se predijo a partir de los datos de validación y se obtuvieron las siguientes métricas:

RMSE = 55,74

MAPE = 18,77 %

Al igual que en los modelos anteriores, a continuación, se puede visualizar el rendimiento del mismo:

Figura 20. Rendimiento PieceWise



5.2.4. Splines

Los modelos Splines son una técnica utilizada en estadística y análisis de regresión para ajustar curvas suaves a conjuntos de datos, particularmente útiles cuando la relación entre las variables no es lineal. Los splines son funciones compuestas de piezas suaves que se combinan para formar una curva suave en todo el rango de los datos.

La característica principal de los splines es que son piezas de polinomios de bajo grado (generalmente polinomios de grado uno o dos) unidos en puntos llamados "nodos". Estos polinomios tienen la propiedad de ser continuos y suaves en los nodos, lo que significa que tienen la misma pendiente y curvatura en los puntos de unión.

Hay varios tipos de splines, pueden ser lineales o no lineales, dependiendo del grado de los polinomios utilizados. Los splines lineales se ajustan bien a datos con cambios graduales, mientras que los splines no lineales pueden adaptarse mejor a datos con cambios más abruptos.

Para este problema se entrenó un modelo de spline cúbico, donde se utilizaron polinomios cúbicos en cada intervalo entre nodos. Estos splines cúbicos son altamente flexibles y pueden adaptarse bien a una amplia variedad de formas de datos. Al igual que en los modelos anteriores se entrenó con la función lm() de R. El tiempo de ejecución del entrenamiento de este modelo fue de 1 segundo.

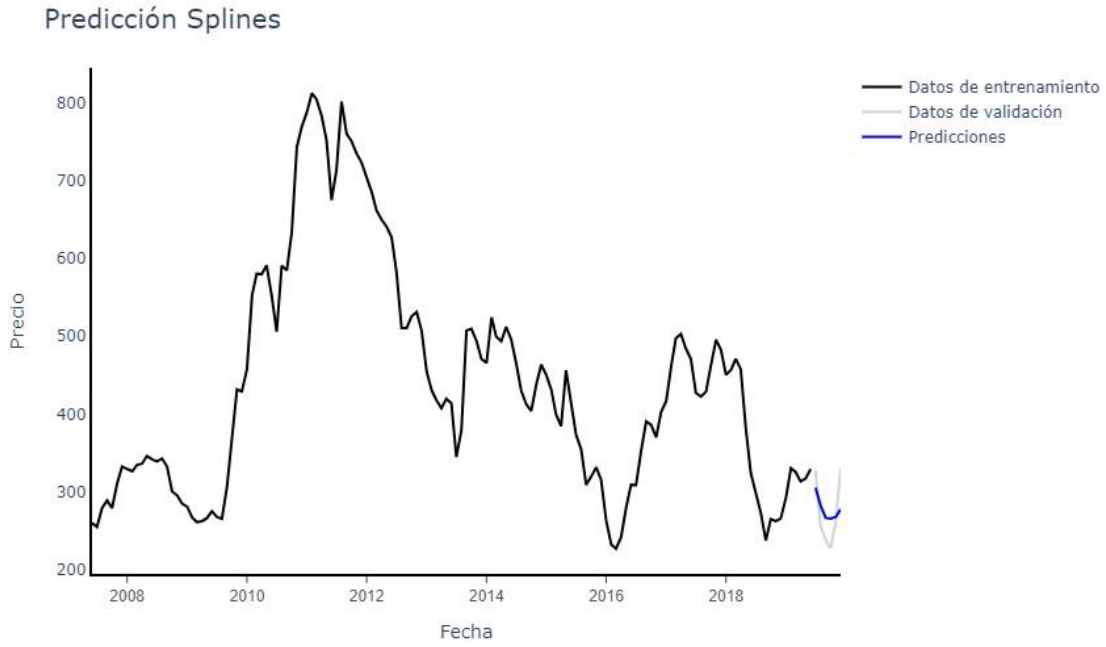
Una vez entrenado el mismo, se obtuvieron las siguientes métricos con las predicciones basadas en los datos de validación:

RMSE = 32,21

MAPE = 10,71 %

Por último, se puede visualizar el rendimiento del modelo a continuación:

Figura 21. Rendimiento Splines



5.2.5. Prophet

Prophet es una herramienta de código abierto desarrollada por Facebook para el análisis y pronóstico de series temporales. Se diseñó específicamente para abordar desafíos comunes en la predicción de series temporales, como los datos faltantes, las tendencias estacionales y las vacaciones irregulares. A diferencia de otros métodos más generales, Prophet está diseñado para ser fácil de usar y requiere menos ajuste de parámetros por parte del usuario.

Algunas de las características clave de Prophet incluyen:

- Modelo aditivo: Prophet modela la tendencia como una función no lineal de tiempo, con componentes adicionales para capturar los efectos estacionales y los días festivos.
- Manejo de datos faltantes: Prophet puede manejar de manera efectiva los datos faltantes y los cambios en la frecuencia de los datos, lo que lo hace útil para conjuntos de datos reales que pueden ser incompletos o tener irregularidades.
- Componente estacional flexible: Prophet puede modelar tanto las tendencias estacionales como las no estacionales, lo que lo hace adecuado para series temporales con patrones estacionales complejos.
- Capacidad de pronóstico a corto y largo plazo: Prophet puede generar pronósticos precisos tanto a corto plazo como a largo plazo, lo que lo hace útil para una amplia variedad de aplicaciones de predicción.

Para el entrenamiento de este modelo, se utilizó la librería “prophet” versión 1.0 [21] de R. En este modelo se definió la fecha como la variable relacionada al tiempo y el precio como la variable dependiente. A partir de allí se agregaron las variables adicionales como regresores al modelo. El tiempo de ejecución del entrenamiento de este modelo fue de 4 segundo.

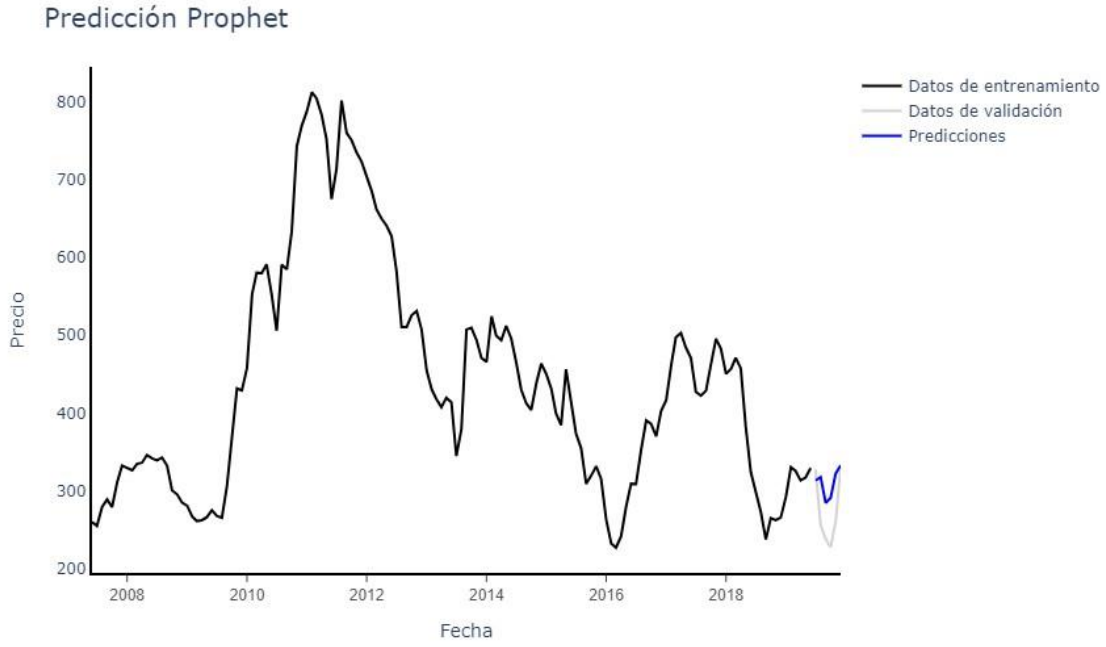
Las métricas obtenidas en validación por este modelo fueron:

RMSE = 48,28

MAPE = 16,71 %

Además, se puede visualizar el rendimiento del mismo a continuación:

Figura 22. Rendimiento Prophet



5.2.6. XGBoost

Al igual que en el problema 1, se utilizó el modelo de XGBoost para la predicción del precio. Además, se utilizó la librería "xgboost" versión 1.7.5.1 [15] de R, definiendo los siguientes parámetros:

- Booster: "gbtree", lo que define el método de árboles de decisión como base del modelo.
- Objective: "reg:squarederror", indicándole al modelo que se encuentra frente a un problema de regresión con el objetico de minimizar el error cuadrático medio.

Nuevamente, se procedió a la iteración algorítmica para la optimización de los hiperparámetros. Los intervalos posibles en este caso fueron los siguientes:

- nround: 5 a 50
- max_depth: 1 a 6
- eta (λ): 0,0025 a 0,1
- gamma: 0 a 1
- colsample_bytree: 0,6 a 1,0
- min_child_weight: 1 a 10
- subsample: 0,75 a 1,0

Una vez ejecutado el algoritmo, se obtuvo la siguiente tabla con los modelos que mejor rendimiento tuvieron en los datos de validación, juntos con aquellos hiperparámetros óptimos que se ajustan mejor a estos datos.

Tabla 5. Rendimiento de modelos XGBoost en problema 2

iteracion	nround	max_ depth	eta	gamma	colsample _bytree	subsa mple	min_child _weight	perf_tr	perf_vd
58	39	3	0,07351	0,93984	0,71855	0,89442	6,83916	23,04	25,97
56	37	2	0,06613	0,72315	0,75251	0,99162	5,41119	49,77	30,88
64	50	2	0,05304	0,63411	0,78355	0,89943	1,53028	44,07	31,57
80	40	3	0,07028	0,89809	0,7006	0,85956	7,11464	37,96	31,82
61	46	3	0,06878	0,77658	0,75402	0,95093	7,9324	30,57	31,97

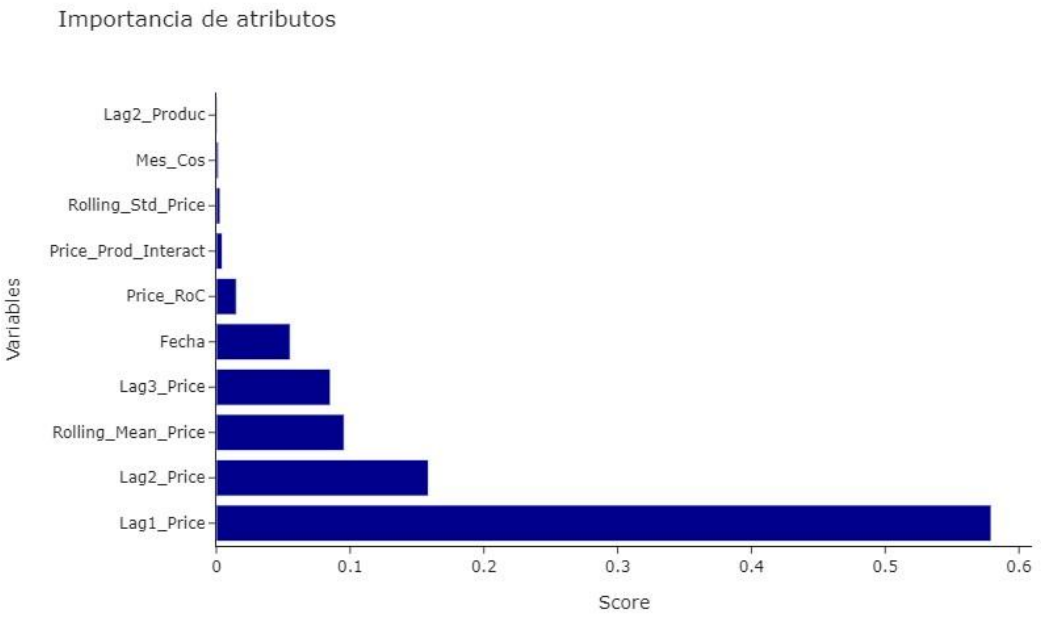
Como se puede observar, el modelo entrenado en la iteración número 58 del algoritmo es aquel que tiene un menor RMSE (última columna de la Tabla 4) en los datos de validación. Un punto importante a tener en cuenta es que, en la Tabla 4, se puede visualizar que el RMSE de entrenamiento es similar al de validación. Esto da un indicio de que no existe sobreajuste del modelo. El tiempo de ejecución del algoritmo utilizado para entrenar estos modelos y encontrar aquel que contenga los hiperparámetros óptimos fue de 3 minutos.

Además, se calculó la segunda métrica a tener en cuenta para este trabajo:

MAPE = 6,70 %

Luego, a partir del modelo seleccionado, se analizaron aquellas variables que son aportan un mayor poder predictivo y por ende son más importantes para el mismo en el siguiente gráfico:

Figura 23. Importancia de variables de XGBoost en problema 2

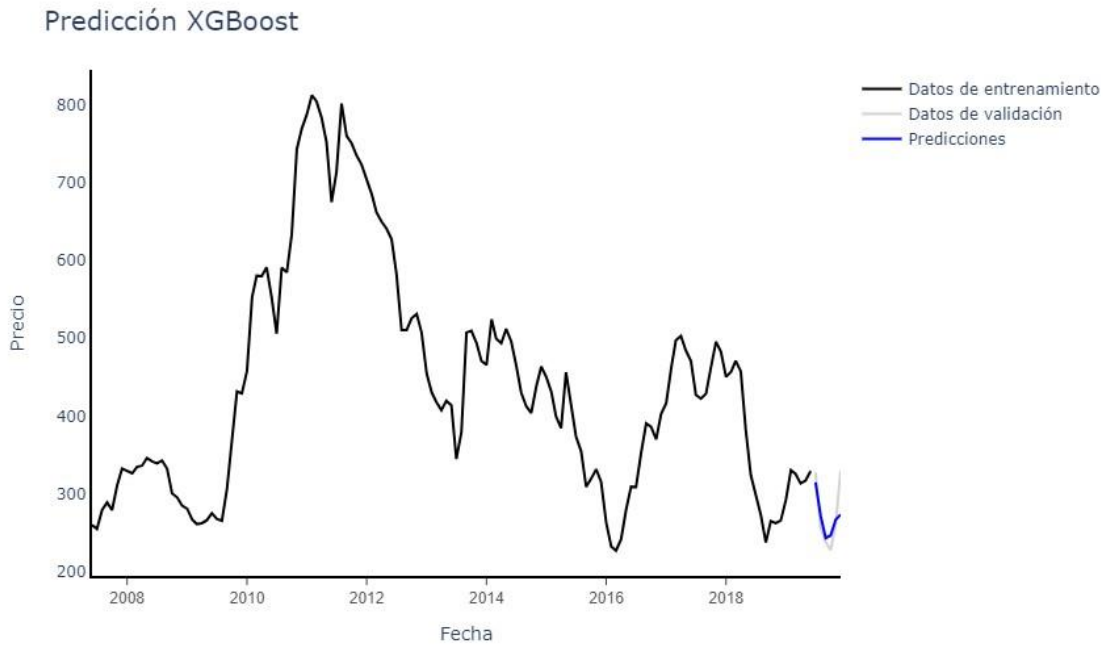


El histograma anterior muestra las variables más importantes del modelo de XGBoost que mejor rendimiento tuvo. Se puede observar que la variable que más aporta a la predicción del modelo es el precio del periodo anterior, lo cual resulta razonable ya que

tenía un valor de 0.98 en el análisis de autocorrelación mencionado con anterioridad. Además, si bien la producción no es una variable muy importante, la producción perteneciente a dos periodos anteriores aparece en noveno lugar realizando un aporte a la predicción del modelo. Esto puede ser un indicio de que el azúcar que se produce no tiene un impacto directo en el mercado, es decir que por ejemplo el precio de agosto recién se verá afectado por el resultante de la producción de mayo.

Por último, visualizamos el rendimiento del modelo de XGBoost a partir del siguiente gráfico:

Figura 24. Rendimiento XGBoost



5.2.7. Comparación de rendimiento de modelos problema 2

Al igual que en el problema anterior, luego del entrenamiento de los modelos, es necesario compararlos a través de todas las métricas definidas al inicio de este trabajo. A continuación, se puede observar la tabla 5 con esta comparativa:

Tabla 6. Comparativa resultados modelos de predicción del precio. Mejores corridas.

Método	Librería	Tiempo de Ejecución	RMSE Test	MAPE Test
Regresión Lineal	fucnión lm() de R	1 seg	45,00	16,25%
Regresion Exponencial	fucnión lm() de R	1 seg	104,31	32,71%
PieceWise	fucnión lm() de R	1 seg	55,74	18,77%
Splines	fucnión lm() de R	1 seg	32,21	10,71%
Prophet	prophet	4 seg.	48,28	16,71%
XGBoost	xgboost	3 min.	25,97	6,70%

Al igual que en el problema de predicción de la producción, el modelo de XGBoost es aquel que tiene menor RMSE y MAPE pero que cuenta con un tiempo de ejecución bastante elevado en comparación con los otros modelos. Se puede observar que todos los

modelos entrenados con la función lm() cuentan con un tiempo de ejecución extremadamente bajo, siendo el modelo Splines el que menor RMSE y MAPE tiene. Por último, el modelo entrenado con la librería de Facebook, Prophet, tiene un tiempo de ejecución bajo y valores de RMSE y MAPE intermedios, pero sin superar a Splines.

6. Elección de los modelos y ejecución en datos no observados

6.1. Elección de modelos

Luego de realizado el entrenamiento de todos los modelos estadísticos elegidos para ambos problemas relacionados con este trabajo, es necesario seleccionar aquellos que mejor desempeño presentaron a través de las métricas elegidas para su medición.

Como ya se describió en las secciones 4.2.4 y 5.2.7 no existe un solo modelo que sea mejor para todas las métricas.

Para el problema 1 de predicción de la producción, el modelo de XGBoost cuenta con menor RMSE y MAPE, pero el modelo de redes neuronales tiene un tiempo de ejecución considerablemente menor. Sin embargo, dado que el objetivo final de este trabajo se basa en la predicción del precio de los próximos 6 meses, al momento de aplicar este modelo en la vida real para obtener la información que será de utilidad para tomar decisiones estratégicas, el tiempo de ejecución de 50 minutos del modelo de XGBoost no se considera elevado.

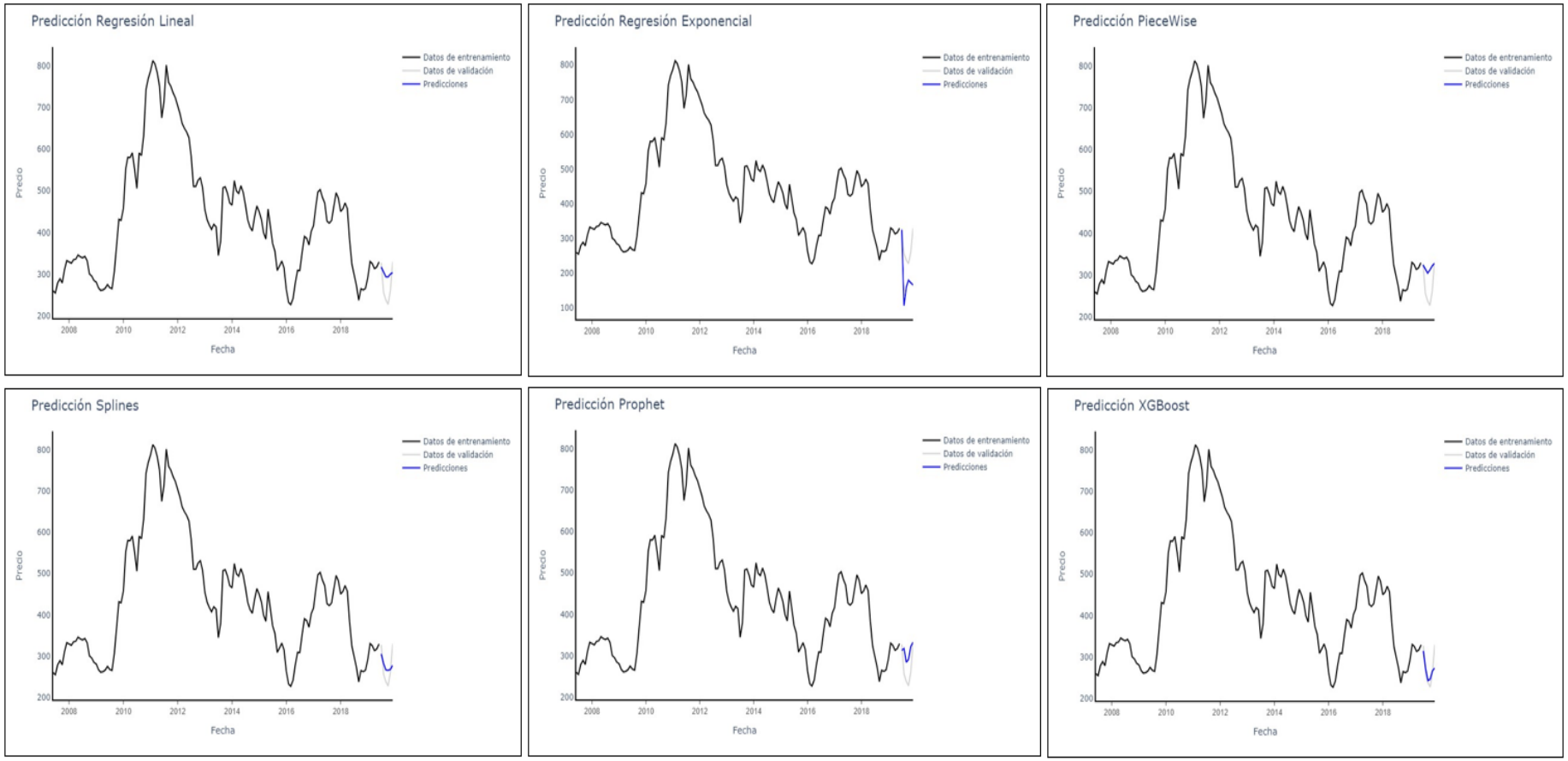
A partir de esto, se considera al modelo de XGBoost como aquel que mejor se ajusta al problema de predicción de azúcar en Argentina.

Por otro lado, para el problema 2 de predicción del precio, XGBoost muestra el mejor rendimiento en términos de precisión, con el RMSE más bajo y el MAPE más bajo, aunque requiere un tiempo de ejecución más largo en comparación con los otros modelos. Splines se destaca como el segundo mejor modelo en términos de precisión, con un RMSE competitivo y un MAPE bajo, junto con el menor tiempo de ejecución al igual que los otros modelos basados en la utilización de la función lm(). Los modelos de regresión lineal, PieceWise y Prophet tienen un desempeño intermedio en términos de precisión y tiempo de ejecución. Y, por último, la regresión exponencial muestra el peor desempeño en términos de precisión, con el RMSE más alto y el MAPE más alto.

Otra forma de comparar el rendimiento de los modelos, con un enfoque visual, es a través gráficos como se muestra en la figura 25. A partir de estos, si bien se puede observar que todos los modelos parecen haber captado el movimiento que tendrá en el precio, es decir una tendencia bajista con un posterior cambio de tendencia alcista, es evidente que las predicciones del modelo de XGBoost se asemejan más a los datos de validación.

Al igual que en el problema de predicción de la producción, el tiempo de ejecución del modelo en la práctica no se considera relevantes. Es por esto que, a partir de la descripción anterior se puede identificar que el modelo de XGBoost es aquel que mejor se ajusta al problema de predicción del precio de azúcar.

Figura 25. Comparación visual rendimiento modelos problema 2



6.2. Ejecución en datos no observados

Durante todo el proceso de entrenamiento de los modelos estadísticos, tanto para la predicción de la producción como para la predicción del precio, se dejó de lado un conjunto de datos pertenecientes al año 2020, al cual se denominó datos de testo.

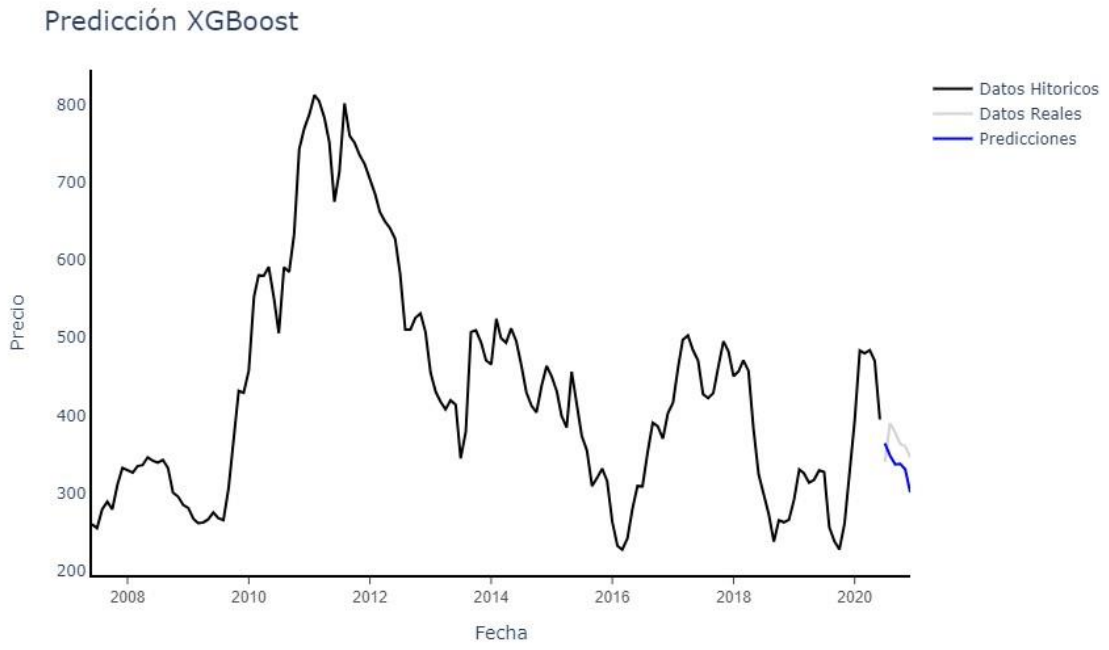
Cuando se construye un modelo predictivo, es fundamental verificar su capacidad para generalizar y realizar predicciones precisas en situaciones del mundo real más allá de los datos utilizados para entrenar y validar el modelo. Los datos de testeo proporcionan una forma de realizar esta evaluación final del modelo, permitiendo una evaluación imparcial y objetiva de su rendimiento.

Es por esto que, luego de haber identificado aquellos modelos estadísticos que mejor rendimiento tuvieron para los problemas planteados en este trabajo, se procedió a realizar predicciones con los datos de testo y evaluar su rendimiento.

En primer lugar, se predijo la producción de azúcar que tendrían los ingenios durante el año 2020. Al sumar estas predicciones, según el modelo de XGBoost, la producción nacional sería de 1.975.110 toneladas de azúcar. Ese año, la producción real fue de 1.781.328 toneladas de azúcar, lo que lleva a un error del modelo de 193.783 toneladas, equivalente a un 10,88%.

Luego, se incorporó dentro de los datos de testeo relacionados al precio la predicción de la producción anterior (1.975.110), y se predijo el precio que tendrá el azúcar en los meses de julio, agosto, septiembre, octubre, noviembre y diciembre del 2020. En el siguiente gráfico se puede visualizar el rendimiento final de este trabajo:

Figura 26. Rendimiento final



Por último, se identificaron las siguientes métricas:

RMSE = 35.55

MAPE = 9,50%

Si bien al observar la figura 26 anterior, se puede visualizar que el modelo no capto exactamente el movimiento el cual tuvo el precio en la realidad, los valores estimados no se alejan demasiado. Además, a partir de las métricas de rendimiento calculadas en los datos de testeo, se puede validar la robustez, estabilidad y la fiabilidad en diferentes contextos y escenarios de los modelos seleccionados anteriormente.

7. Conclusiones

Con el objetivo de brindar información útil para la toma de decisiones estratégicas para empresas dedicadas a la producción de azúcar, al comenzar este trabajo se planteó la interrogante de si era posible predecir el precio del azúcar en Argentina.

Luego de un análisis del contexto económico y político del sector en el país, y basándose en la ley de oferta y demanda, se arribó a la hipótesis de que el precio podría estar determinado por la cantidad de azúcar circulante en el país y por ende por la producción nacional total. Hipótesis la cual se validó en primera instancia con un análisis preliminar sobre los datos obtenidos.

A partir de allí surgiendo dos problemas. Por un lado, la predicción de la producción de azúcar. Y, por otro lado, la predicción del precio de la azúcar propiamente dicha.

Debido a los avances en materia estadística de las últimas décadas, impulsado por las facilidades de cómputo que ofrece la tecnología moderna, se decidió buscar una solución a ambos problemas a través de métodos de aprendizaje supervisado. Logrando, luego de numerosos análisis, obtener dos modelos que, combinados, tienen un gran rendimiento en la predicción del precio del azúcar.

Sin embargo, un punto importante a tener en cuenta es que, luego de analizar aquel modelo que proporciona el valor que tendrá el precio en el futuro, se pudo observar que el mismo no utiliza a la producción nacional como un factor determinante a la hora de realizar su predicción. Esto, no permite validar desde una perspectiva estadística la hipótesis desde la cual se parte al iniciar este trabajo.

Dicho esto, y como se mencionó anteriormente, la finalidad de este trabajo es brindar información útil para la toma de decisiones estratégicas. Objetivo el cual se puede obtener con la implementación de aquellos modelos que predicen el valor futuro del precio.

La combinación de modelos de XGBoost con un MAPE de 9,50% en la predicción del precio, proporciona información que brinda una ventaja competitiva a aquellos que la posean. Para un ingenio, poder conocer el movimiento que tendrá el precio lo puede llevar a tomar mejores decisiones que lo llevaran a aumentar sus beneficios, como ser el comprar azúcar hoy sabiendo que el precio de esta subirá en el futuro para luego venderla y obtener una ganancia de esta operación.

7.1. Puntos de mejora sobre el trabajo realizado

Luego de haber finalizado el trabajo, se pueden mencionar varias cuestiones que podrían mejorar el resultado obtenido del mismo.

El primer punto está relacionado con la inclusión de datos adicionales relacionados con el total de azúcar circulante en el mercado interno. Factor clave de la hipótesis inicial basada en la ley de oferta y demanda.

Incluir datos relacionados a la exportación podría ser relevante. De esta manera se podría tratar de cuantificar el stock con el que cuentan los ingenios y por ende el stock total existente en el país. Si a este stock, se le suma la producción del año, se obtendría un valor más preciso sobre el total circulante en el mercado interno.

Por otro lado, un análisis sobre la producción de alcohol y la cantidad de caña de azúcar que destina cada ingenio a la producción de este podría sentar las bases para cuantificar

el porcentaje en cantidad de caña de azúcar destinada a la producción de azúcar de cada ingenio.

Por último, en diciembre el 2023, en Argentina fue electo un presidente de un partido político distinto al cual se encontraba al momento de iniciar la realización de este trabajo. Es de esperarse que un presidente de ideales políticos distintos realice cambios en las condiciones de mercado del país. De esta forma, en diciembre del 2023, a través del Decreto de Necesidad y Urgencia que aprueba las Bases para la Reconstrucción de la Economía Argentina [22], deroga el artículo 1° de la Ley N° 25.715 [7]. Esta modificación, impacta en las condiciones del mercado interno, planteando la posibilidad de la entrada de productores extranjeros.

De esta forma, incluir datos relacionados al precio del azúcar en el mercado internacional puede ser una variable importante al momento de predecir el precio que se comercializa en el mercado interno.

Referencias

[1] Ley N° 27.640, MARCO REGULATORIO DE BIOCOMBUSTIBLES (3 de agosto de 2021) https://www.boletinoficial.gob.ar/detalleAviso/primera/247667/20210804

[2] Centro azucarero argentino. (s.f.). https://centroazucarero.com.ar/

[3] Presidencia de la Nación, Ministerio de Hacienda, Secretaría de Política Económica, Subsecretaría de Programación Microeconómica. (junio 2018). Informes de cadena de valor https://www.argentina.gob.ar/sites/default/files/sspmicro_cadenas_de_valor_azucar.pdf

[4] Secretaría de Alimentos y Bioeconomía del Ministerio de Agricultura, Ganadería y Pesca de Argentina. (s.f.). Azúcar. Recuperado de https://alimentosargentinos.magyp.gob.ar/contenido/revista/html/33/33_01_Azucar.htm

[5] Decreto 797/92, DERECHO ADICIONAL A LAS IMPORTACIONES (19 de mayo de 1992) https://www.argentina.gob.ar/normativa/nacional/decreto-797-1992-16804/actualizacion

[6] Ley N° 22.415, CODIGO ADUANERO (2 de marzo 1981) https://cancilleria.gob.ar/userfiles/ut/diaju-codigo-aduanero-ley_22.415.pdf

[7] Ley N° 25.715, AZUCAR (7 de abril de 2003) https://www.argentina.gob.ar/normativa/nacional/ley-25715-83889/texto

[8] Peralta González, J. A. (2021). Modelos de pronósticos del precio del azúcar en México (Doctoral dissertation, Universidad Autónoma Chapingo).

[9] Silva RF, Barreira BL, Cugnasca CE. Prediction of Corn and Sugar Prices Using Machine Learning, Econometrics, and Ensemble Models. Engineering Proceedings. 2021; 9(1):31. https://doi.org/10.3390/engproc2021009031

[10] Servicio Meteorológico Nacional. (2022, diciembre 14). Datos obtenidos vía correo electrónico. [cim@smn.gob.ar].

[11] Cactu Tucumán. (s.f.). Recuperado de https://www.cactutucuman.com/

[12] Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. New York: Springer. pp. 1

[13] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

[14] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

[15] XGBoost. (s.f.). Introduction to Boosted Trees. https://xgboost.readthedocs.io/en/latest/tutorials/model.html

[16] Ministerio de Agricultura, Ganadería y Pesca. (2015). Guía técnica del cañero [PDF]. Recuperado de https://www.argentina.gob.ar/sites/default/files/guia_tecnica_del_canero.pdf

[17] Scikit-learn. (s.f.). Support Vector Machines (SVM) - sklearn.svm.SVR. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html#sklearn.svm.SVR

[18] Keras. (s.f.). Keras Layers API. https://keras.io/api/layers/

[19] TensorFlow. (s.f.). Guía inicial de TensorFlow 2.0 para expertos. https://www.tensorflow.org/tutorials/quickstart/advanced?hl=es-419

[20] Banco de la Nación Argentina. (s.f.). Recuperado de www.bna.com.ar

[21] Facebook. (s.f.). Prophet: Forecasting at Scale. https://facebook.github.io/prophet/

[22] Decreto de Necesidad y Urgencia que aprueba las Bases para la Reconstrucción de la Economía Argentina Art. 146 DECRETO NACIONAL 70/2023. 20/12/2023. Vigente, de alcance general (B.O 21/12/2023) https://www.boletinoficial.gob.ar/detalleAviso/primera/301122/20231221