BARCELONA TECHNOLOGY SCHOOL
MASTER IN BIG DATA AND A.I. SOLUTIONS

FINAL PROJECT

CEO Anastasia Krivenkovskaya

CDO Daniel Espinoza

CTO Rogelio Martinez

CSO Juan Neuenschwander

COO Tehreem Malik

MENTORS: Ana Guasch, Sergio Gago

Table of Contents

Executive Summary

The agricultural commodities trading industry operates within a dynamic and complex global marketplace characterized by price volatility, supply chain intricacies, and evolving consumer demands. In response to these challenges, technological innovations have emerged as pivotal tools to enhance market analysis, forecasting, and decision-making processes. AGIA, an advanced chatbot solution developed for market operators in the agricultural commodities sector, integrates artificial intelligence (AI) to provide actionable insights, predictive analytics, and real-time data.

AGIA's foundation lies in its collaboration with Abastores, a leading online platform for agricultural commodities exchange. Through its integration with Abastores, AGIA demonstrates its ability to deliver unparalleled market intelligence, enabling users to make informed decisions that drive profitability and strategic growth.

This paper explores AGIA's technological underpinnings, operational methodologies, and practical applications through case studies and empirical evidence. It highlights AGIA's role in revolutionizing commodities trading by enhancing decision support systems with AI-driven insights.

Furthermore, this study contributes to the discourse on AI applications in commodities trading, emphasizing AGIA's impact on industry standards and its potential to reshape future innovations. By examining the implications of AI integration in market analysis and decision-making processes, this paper underscores AGIA's significance in fostering operational efficiency and strategic resilience for market operators.

In conclusion, AGIA represents a transformative solution at the nexus of technology and commodities trading, poised to redefine industry practices and empower market operators with advanced capabilities for navigating the complexities of the global agricultural commodities market.

Introduction

The introduction serves as an entry point to your academic paper, setting the context and laying out the objectives of the study. It provides a comprehensive overview of AGIA, its significance in the agricultural commodities trading industry, and the rationale behind its development.

Who, What, Why, and How

Who: AGIA is designed for market operators in the agricultural commodities sector, including farmers, traders, storekeepers, and institutional investors, who face challenges in accessing timely and actionable market insights.

What: AGIA is an AI-driven chatbot solution that provides real-time insights, historical price data, and predictive analytics to facilitate informed decision-making and superior trading outcomes.

Why: The agricultural commodities market is complex and influenced by numerous factors such as environmental conditions, oil prices, and currency fluctuations. Traditional market analysis tools often fall short in providing timely and actionable insights. AGIA addresses this gap by leveraging advanced AI technologies to deliver comprehensive market intelligence, enabling market operators to make informed decisions.

How: AGIA uses AI algorithms to analyze historical data, environmental indicators, oil prices, and currency exchange rates. It provides real-time analytics and forecasting capabilities, seamlessly integrating with existing platforms like Abastores to empower market operators.

Mission and Vision

Mission: To create a more transparent, efficient, and sustainable agricultural market, driving economic growth and resilience for all stakeholders.

Vision: To empower the agricultural sector with pioneering AI analytics, providing actionable insights and enabling informed decision-making for superior trading outcomes.

Business Model

This section delves into AGIA's business model, exploring its value proposition and detailing the components of its Business Model Canvas.

Value Proposition

AGIA offers an innovative AI-driven chatbot solution tailored specifically for the agricultural sector. Unlike traditional market analysis tools, AGIA's chatbot provides market operators with instant access to detailed and customized market insights. Through its seamless integration with platforms like Abastores, AGIA offers:

- **Real-Time Agricultural Insights:** Immediate access to the latest market data, enabling quick and informed decision-making.
- **Historical Price Data:** Comprehensive historical data analysis to identify trends and inform future strategies.
- **Advanced Predictive Analytics:** AI-powered forecasts that predict market movements, helping users stay ahead of the curve.
- **Customizable Reports:** Tailored insights and reports that meet the specific needs of different market operators.
- **User-Friendly Interface:** An intuitive and accessible chatbot interface that simplifies complex data analysis, making it easy for users to interpret and act upon.

Market and Competitors

To better understand our market, we conducted extensive market research. Our analysis revealed a plethora of companies offering data relevant to agricultural trading. However, these companies primarily provide dashboards featuring publicly accessible general data, which lack in-depth analysis, predictive capabilities, and customization. This approach leaves a significant gap in the market. AGIA aims to fill this gap by offering a sophisticated AI-driven chatbot solution. This chatbot not only provides detailed and customized market insights but also delivers real-time analytics and predictive analytics, setting us apart from the traditional dashboard-based solutions.

Business Model Canvas

The Business Model Canvas for AGIA outlines the essential elements that form the foundation of our business strategy. This model ensures a clear understanding of how AGIA delivers value, engages with customers, and sustains its operations.

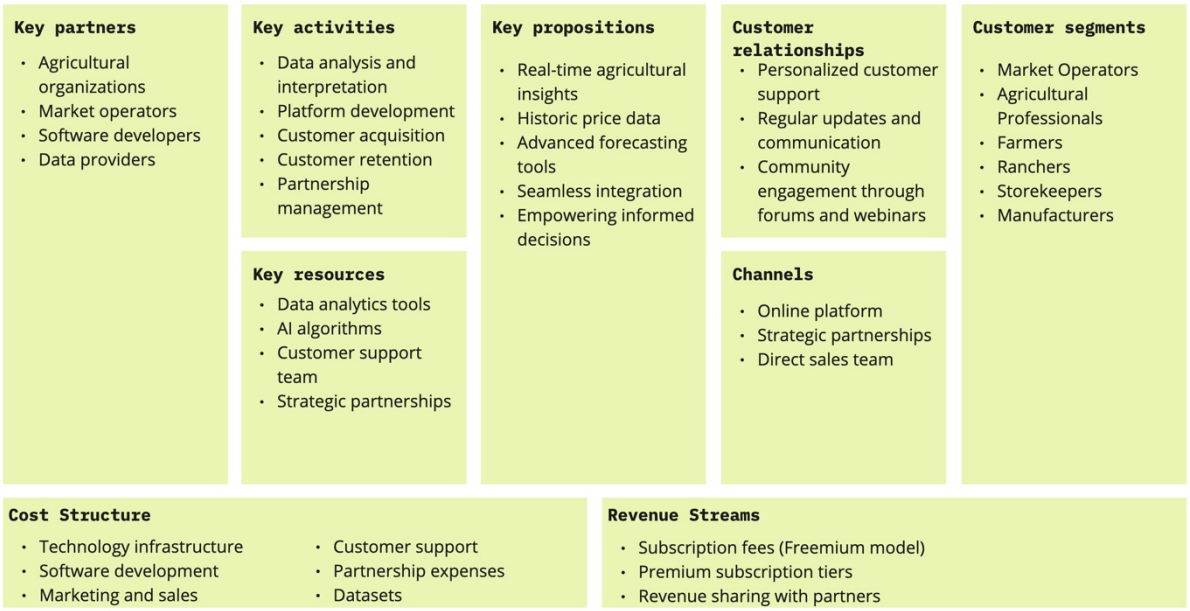


Image 1. AGIA's Business Canvas

Key Partners: AGIA's key partners include agricultural organizations, market operators, software developers, and data providers. These partnerships are critical for accessing valuable industry data and insights, enhancing our technological capabilities, and expanding our market reach. By collaborating with these partners, AGIA can continuously improve its AI-driven solutions and provide more accurate and relevant market insights.

Key Activities: The primary activities of AGIA revolve around data analysis and interpretation, platform development, customer acquisition, customer retention, and partnership management. Data analysis and interpretation are central to AGIA's value proposition, enabling the generation of real-time agricultural insights and advanced predictive analytics. Platform development ensures that AGIA remains user-friendly and technologically advanced. Customer acquisition and retention strategies are crucial for building and maintaining a robust user base. Partnership management involves nurturing relationships with key stakeholders to leverage their strengths and resources for mutual benefit.

Key Resources: AGIA relies on several key resources, including data analytics tools, AI algorithms, a customer support team, and strategic partnerships. Data analytics tools and AI algorithms are essential for processing vast amounts of data and generating actionable insights. The customer support team provides personalized assistance to users, ensuring high levels of satisfaction and engagement. Strategic partnerships enhance AGIA's capabilities and market presence, allowing for continuous growth and improvement.

Value Propositions: AGIA offers a unique value proposition by providing real-time agricultural insights, historical price data, advanced forecasting tools, customizable reports, and a user-friendly interface. These features enable market operators to make informed decisions, stay ahead of market trends, and optimize their trading strategies. The integration of AI-driven analytics ensures that AGIA's insights are accurate, relevant, and actionable, setting it apart from traditional market analysis tools.

Customer Relationships: AGIA aims to build strong and personalized relationships with its customers. This involves offering personalized customer support, regular updates and

communication, and community engagement through forums and webinars. Personalized customer support ensures that users receive tailored assistance based on their specific needs and preferences. Regular updates keep users informed about the latest market trends and insights. Community engagement fosters a sense of belonging and encourages knowledge sharing among users.

Channels: The primary channels through which AGIA engages with its customers include its online platform, strategic partnerships, and a direct sales team. The online platform serves as the main interface for users to access AGIA's features and insights. Strategic partnerships with agricultural organizations and market operators help expand AGIA's reach and enhance its offerings. The direct sales team is responsible for promoting AGIA's solutions and acquiring new users.

Customer Segments: AGIA targets a diverse range of customer segments, including market operators, agricultural professionals, farmers, ranchers, storekeepers, and manufacturers. Each of these segments has specific needs and preferences that AGIA addresses through its tailored insights and analytics. By understanding the unique challenges and opportunities faced by each segment, AGIA can provide more relevant and valuable solutions.

Cost Structure: AGIA's cost structure includes expenses related to technology infrastructure, software development, marketing and sales, customer support, partnership management, and data acquisition. Technology infrastructure and software development are critical for maintaining and enhancing the platform. Marketing and sales efforts are necessary to promote AGIA and attract new users. Customer support ensures high levels of user satisfaction and retention. Partnership management involves coordinating with key partners to leverage their resources and expertise. Data acquisition is essential for ensuring the accuracy and relevance of AGIA's insights.

Revenue Streams: AGIA generates revenue through subscription fees, including a freemium model and premium subscription tiers, as well as revenue sharing with partners. The freemium model offers basic features for free, encouraging users to try AGIA and upgrade to

premium tiers for additional features and insights. Premium subscription tiers provide advanced analytics and customized reports, catering to users with more specific needs. Revenue sharing with partners involves collaborating with strategic stakeholders to create mutually beneficial financial arrangements.

Financial Forecast

To provide a comprehensive financial forecast for AGIA, we will consider the potential revenue generated from a partnership with Abastores. Abastores has a user base of over 4,000 farmers, presenting a significant market opportunity for AGIA's subscription-based services. The following forecast assumes different adoption rates for each subscription tier and projects the potential revenue accordingly.

Assumptions

1. Adoption Rates:
    - 60% of users will opt for the Basic Subscription.
    - 30% of users will opt for the Professional Subscription.
    - 10% of users will opt for the Premium Subscription.

2. Subscription Costs:
    - Basic Subscription: €10 per month
    - Professional Subscription: €50 per month
    - Premium Subscription: €100 per month

3. User Base: 4,000 farmers

Financial Forecast Calculation

Basic Subscription:

- Adoption Rate: 60% of 4,000 farmers = 2,400 farmers
- Monthly Revenue: 2,400 farmers * €10 = €24,000
- Annual Revenue: €24,000 * 12 = €288,000

Professional Subscription:

- Adoption Rate: 30% of 4,000 farmers = 1,200 farmers
- Monthly Revenue: 1,200 farmers * €50 = €60,000
- Annual Revenue: €60,000 * 12 = €720,000

Premium Subscription:

- Adoption Rate: 10% of 4,000 farmers = 400 farmers
- Monthly Revenue: 400 farmers * €100 = €40,000
- Annual Revenue: €40,000 * 12 = €480,000

Total Annual Revenue:

- Basic Subscription: €288,000
- Professional Subscription: €720,000
- Premium Subscription: €480,000

Total Forecasted Annual Revenue: €288,000 + €720,000 + €480,000 = €1,488,000

Based on the projected adoption rates and subscription costs, partnering with Abastores and reaching their 4,000 farmers could generate an annual revenue of approximately €1,488,000 for AGIA.

Additional Considerations

- **Marketing and Customer Support Costs:** Achieving these adoption rates will likely require targeted marketing efforts and robust customer support to ensure high levels of user engagement and satisfaction.
- **Platform Development and Maintenance:** Continuous investment in platform development and maintenance is essential to provide a seamless user experience and integrate advanced features.
- **Partnership Revenue Sharing:** If a revenue-sharing agreement is established with Abastores, a portion of the generated revenue will be allocated to them, which should be factored into the overall financial planning.

Prototype Design & Development

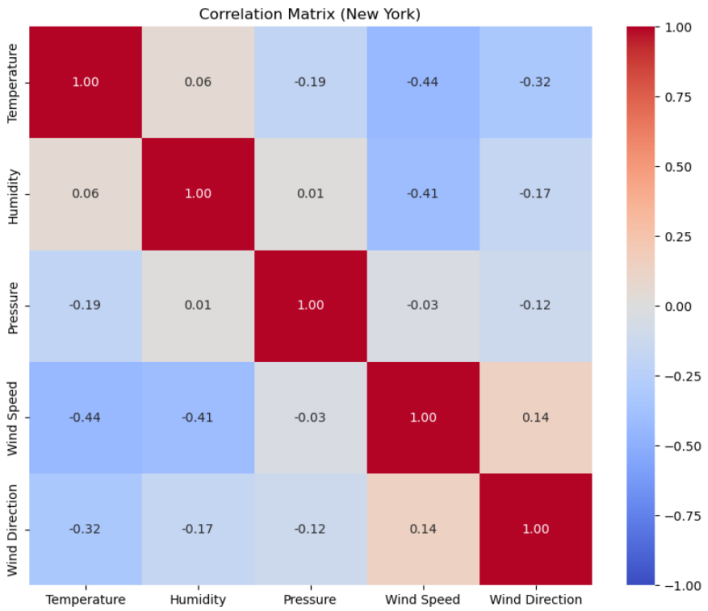
Data Exploration

Weather vs Wheat prices

One of our premises during this project was to predict the price of crops based on climate in regions of Spain. We decided to approach different weather data sources such as ClimateAPI or AccuWeather, however many of these sources turned out to be unreliable in location and data frequency, so we chose to go with Open weather, which gave us precision and more granularity than other sources, furthermore it had recent data entries, at the time were need to the suspicion of seasonality.

Combining this data with the Abastores data required sampling to work with the same granularity in the dates, in addition, we had to clean the data to be able to work efficiently.

As a result, we found a very low correlation with most of the proposed variables, with the highest correlation being 10% for wind speed, and a negative correlation of 36% for regional temperature.



Currency Exchange Rates vs. Abastores Prices

One of our premises during this project was to analyze the impact of currency exchange rates on the prices of goods in Abastores. We initially considered various data sources for exchange rates, such as XE and OANDA, but many of these sources turned out to be unreliable in terms of data frequency and granularity. Therefore, we chose to use Open Exchange Rates, which provided us with precise and granular data, including recent entries that were crucial for detecting potential seasonal patterns.Combining this data with the Abastores pricing data required careful sampling to ensure both datasets had the same date granularity. Additionally, we had to clean the data to remove inconsistencies and ensure efficient analysis.

As a result, we found a very low correlation between most of the proposed variables. The highest correlation was 12% for the exchange rate of USD to EUR, and a negative correlation of 28% for the exchange rate of GBP to EUR. This indicates that while there is some relationship between exchange rates and Abastores prices, it is not strong enough to make definitive predictions based solely on exchange rate data.

System Architecture

Our project is designed with a dual infrastructure approach to have both local testing and production deployment environments. This section outlines the distinct architectures implemented for these two environments. By having separate configurations, we ensure a streamlined development process while maintaining robust, scalable, and secure operations in production.

Local Environment

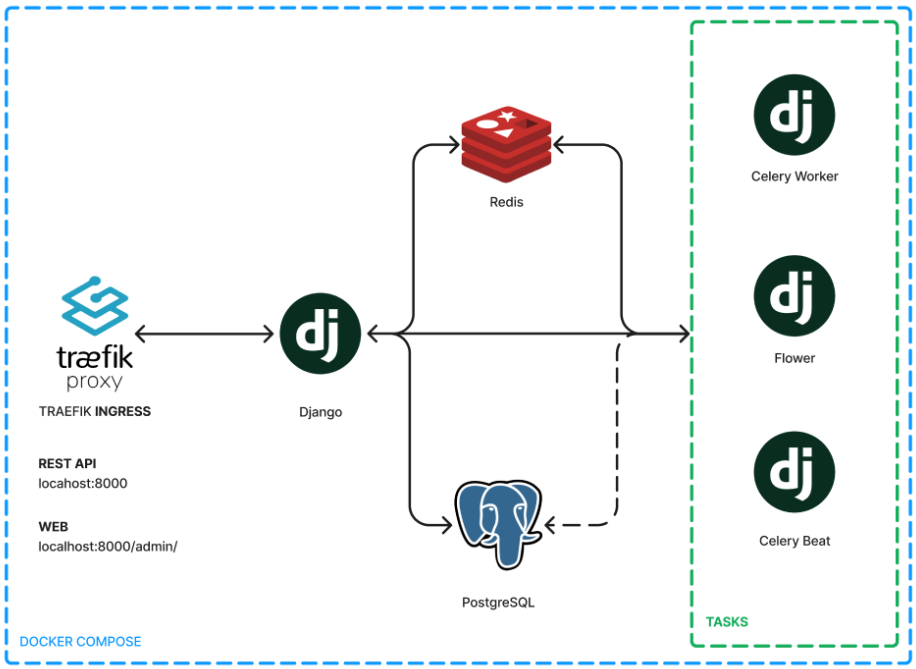


Image 2. AGIA's local environment schema

This system architecture, managed through Docker Compose, comprises several integrated components to ensure a robust and scalable web application. At the forefront, the Traefik Proxy serves as a traffic manager, directing incoming HTTP requests to the appropriate service within the Docker Compose network. The core of the system is the Django application, which handles web requests, processes data, and interacts with a PostgreSQL database for secure data storage and retrieval. Redis, an in-memory data structure store, is employed for quick-access data caching and as a message broker for Celery, facilitating efficient task queuing and execution. The Celery system includes Celery Workers for executing background tasks, Celery Beat for scheduling periodic tasks, and Flower for real-time monitoring of task execution and system performance.

This setup enables Django to offload long-running tasks to Celery, enhancing user experience by maintaining responsiveness. Overall, this architecture ensures scalability, efficient traffic routing, and centralized monitoring, making it a well-organized and maintainable system capable of handling various web application requirements effectively.

Production environment

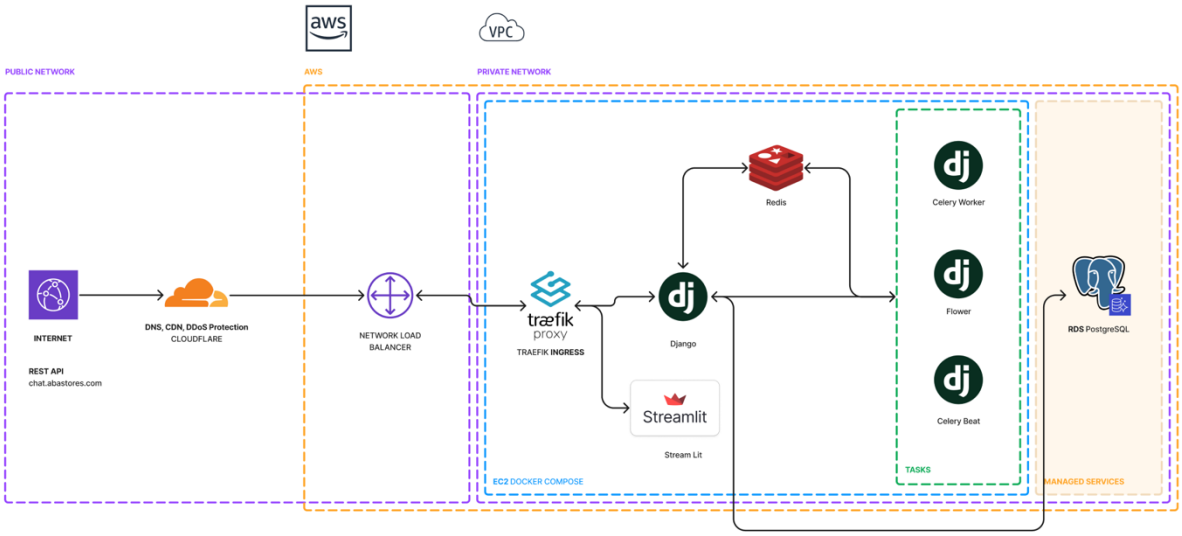
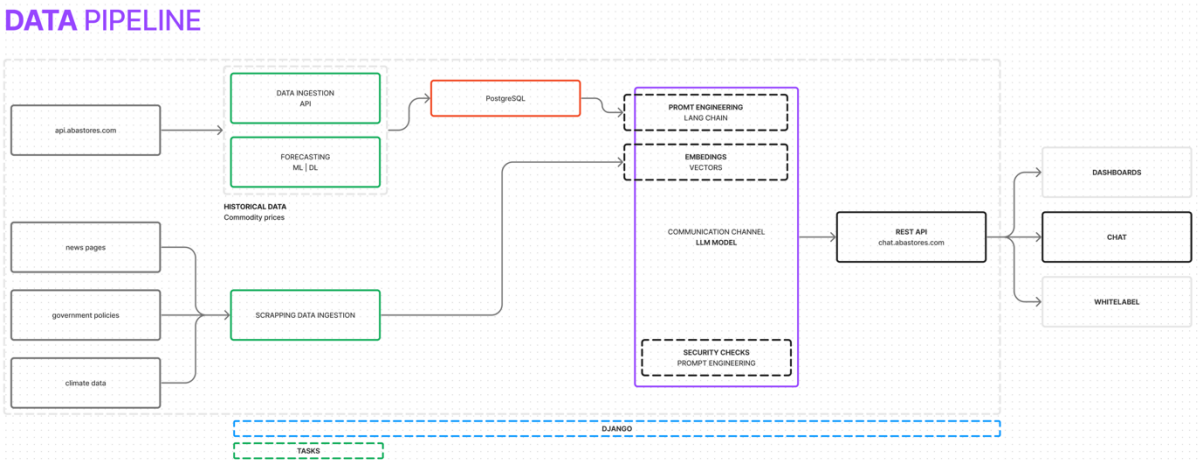


Image 3. AGIA's Product environment schema

The production architecture presents a robust web application infrastructure hosted on AWS, incorporating multiple services for optimal performance and scalability. External traffic enters through Cloudflare, which provides DNS resolution, CDN capabilities, and DDoS protection, enhancing security and speed. The traffic is then directed to the AWS Network Load Balancer, which distributes it to the Traefik Proxy within the private network. Traefik manages and routes requests to either the Django application or Streamlit. Django, as the core web framework, processes request, interacts with the managed PostgreSQL database (RDS PostgreSQL) for secure data storage, and uses Redis for caching and task management. Redis also serves as a message broker for Celery, which handles background tasks. Celery's components include Celery Worker for task execution,

Celery Beat for scheduling periodic tasks, and Flower for real-time monitoring of task status and performance. This architecture not only ensures scalability and efficient traffic management but also provides robust data handling, enhanced security, and interactive data visualization, making it a comprehensive solution for modern web application requirements.

Data pipeline



Data Sources

- **api.abastores.com:** Abastores has an API with data about regions of interest and prices of commodities. ~20 regions from more than 50 data sources, including ~30 different provinces and thousands of data prices.
- **News Pages:** It was identified that news might be a good source of data. Some news is relevant enough to change the prices in a very significant way.
- **Government Policies:** Collect information on government policies to measure the possible impact on agricultural practices, market prices, and trade conditions.
- **Climate Data:** Gather to forecast productivity and potential disruptions.

Data Ingestion and Processing

- **API Data Ingestion:** Real-time data from api.abastores.com is ingested through the API pipeline, this ingestion is scheduled to be run at least once a day. Later this can be updated to reflect more accurate estimations.
- **Scraping Data Ingestion:** News, government policy updates, and climate data are ingested using web scraping.

- **Forecasting:** Using machine learning (ML) models, we process historical data, such as commodity prices, to generate forecasts. This will allow the AI has a more insightful data.

Data Storage

- **PostgreSQL Database:** All ingested and processed data is stored in a PostgreSQL database, serving as a centralized repository. This can be used as a structured database as well as a repository for vectorized data.

Advanced Data Processing

- **Prompt Engineering and Embeddings:** Utilizing language models (like LangChain), we generate prompts and embeddings that help the AI understand and process user queries. These vectors are necessary for the contextual understanding necessary for the user quering.
- **Communication Channel (LLM Model):** Our large language model (LLM) is the backbone of the AI chat system, facilitating natural language interaction with users. Enhanced with security checks and additional prompt engineering, this model ensures reliable and secure communication.

Output and User Interaction

- **REST API:** The processed data and AI-generated insights will be displayed through a REST API (chat.abastores.com). This API serves as the interface for various user applications.
- **User Interfaces:** The REST API connects to multiple user interfaces, including:
    - **Chat Interface:** Allowing users to interact directly with the AI chat system for personalized advice and information. This is an option chosen by the client.

    - **Dashboards:** Providing visual representations of data trends and forecasts to help users understand market conditions on a visual representation.
- **White-Label Solutions:** Enabling integration of our system with other platforms, expanding the reach and utility of our AI chat service.

Infrastructure and Workflow Management

- **Django Framework:** The entire pipeline is managed using Django, a robust and scalable web framework that ensures smooth operation and efficient task management.
- **Task Organization:** Specific tasks and operations within the pipeline are organized and executed by Celery and Django, ensuring reliable and timely data processing. This helps us to keep one consistent environment easy to maintain.

Machine Learning

In this section of the project, we focus on an investigative study aimed at examining the feasibility of applying machine learning techniques to address the following hypothesis: "prediction of agricultural commodity prices." The purpose of this segment is to explore and evaluate various machine learning models and methods to determine their effectiveness in predicting the prices of these commodities, considering the inherent variability and complexity of agricultural data. Through this analysis, we aim to identify patterns and relationships that can enhance the accuracy of our predictions, thus providing a solid foundation for informed decision-making in the agricultural sector.

Abastores Price Data

In this phase of the project, we focus on data ingestion, which is carried out by making calls to the API provided by Abastores. This API returns data on the prices of grain

commodities. The data ingestion process involves fetching, processing, and storing this price data in a structured format suitable for analysis. By utilizing the Abastores API, we ensure that our dataset is up-to-date and comprehensive, capturing the latest market trends and price fluctuations. This data is critical for training and testing our machine learning models, enabling us to accurately predict future commodity prices in the agricultural sector.

```
base_url = "https://api.abastores.com/api/v2/marketdata/dataprice-rag/"

# Inicializar una lista para almacenar los datos de todas las páginas
all_data = []

# Iterar sobre las 36 páginas
for page_number in range(1, 38):
    # Construir el URL completo para la página actual
    url = f"{base_url}?page={page_number}&page_size=1000"

    # Enviar una solicitud GET al endpoint
    response = requests.get(url)

    # Verificar si la solicitud fue exitosa (código de estado 200)
    if response.status_code == 200:
        # Obtener los datos de la respuesta
        data = response.json()

        # Extraer valores de la columna "results" y crear columnas separadas
        df = pd.json_normalize(data['results'])

        # Agregar el DataFrame actual a la lista
        all_data.append(df)
    else:
        print(f"Failed to retrieve data for page {page_number}. Status code:", response.status_code)

# Concatenar todos los DataFrames en un solo DataFrame
final_df_1 = pd.concat(all_data, ignore_index=True)
```

Below is an example dataframe obtained from the API call to Abastores, showcasing the price data for various grain commodities. As can be seen, there are null values present, the dates are not well-sampled or standardized, and there are numerous columns that will not be used in our analysis.

[4]:

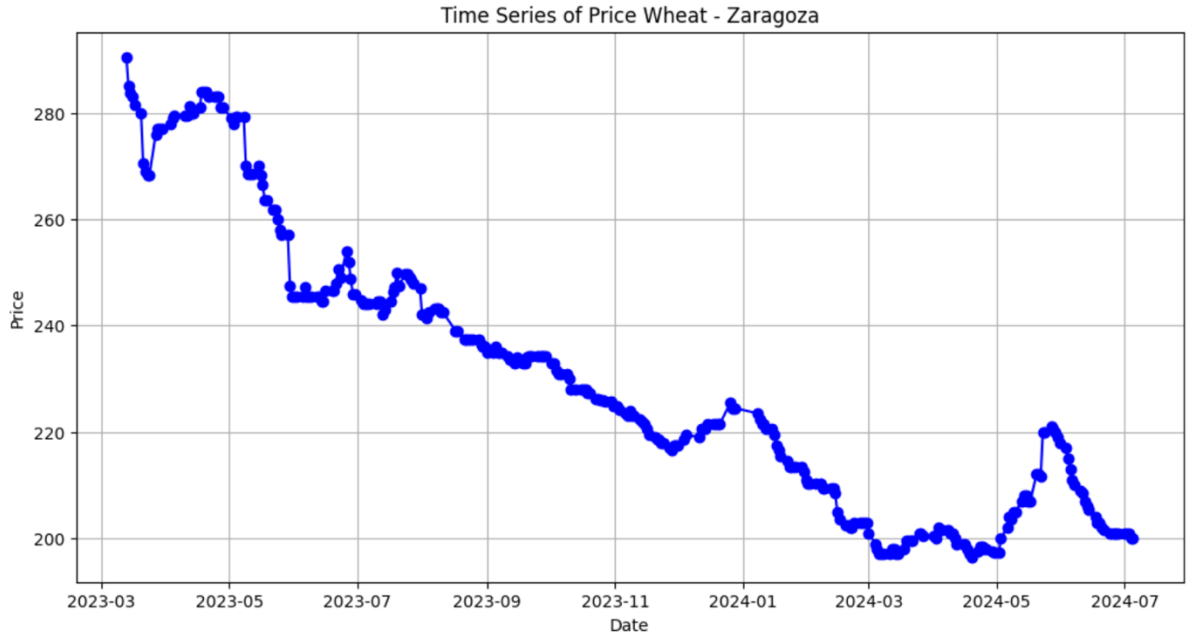
	id	date	price	quantity	region	product.id	product.meta_product.id	product.meta_product.name	product.meta_product.i
0	183785	2024-07-05T19:45:06.543632+02:00	228.25	NaN	NaN	21949	284	Trigo	a-wh
1	183786	2024-07-05T19:45:06.543632+02:00	215.00	NaN	NaN	21950	4	Maíz	a-c
2	183724	2024-07-05T14:10:00+02:00	200.00	NaN	NaN	1	284	Trigo	a-wh
3	183725	2024-07-05T14:10:00+02:00	214.00	NaN	NaN	1	284	Trigo	a-wh
4	183726	2024-07-05T14:10:00+02:00	210.00	NaN	NaN	1	284	Trigo	a-wh
...	...	...	...	...	...	...	...	...	
36995	109653	2002-10-04T02:00:00+02:00	112.69	0.0	16.0	554	13	Triticale	a-tritic
36996	109654	2002-10-04T02:00:00+02:00	108.18	0.0	16.0	547	6	Centeno	a-
36997	97689	2002-09-27T02:00:00+02:00	114.19	0.0	16.0	21872	5	Cebada	a-ba
36998	97690	2002-09-27T02:00:00+02:00	114.19	0.0	16.0	21783	5	Cebada	a-ba
36999	97691	2002-09-27T02:00:00+02:00	126.21	0.0	16.0	19088	8	Avena	a-

37000 rows × 19 columns

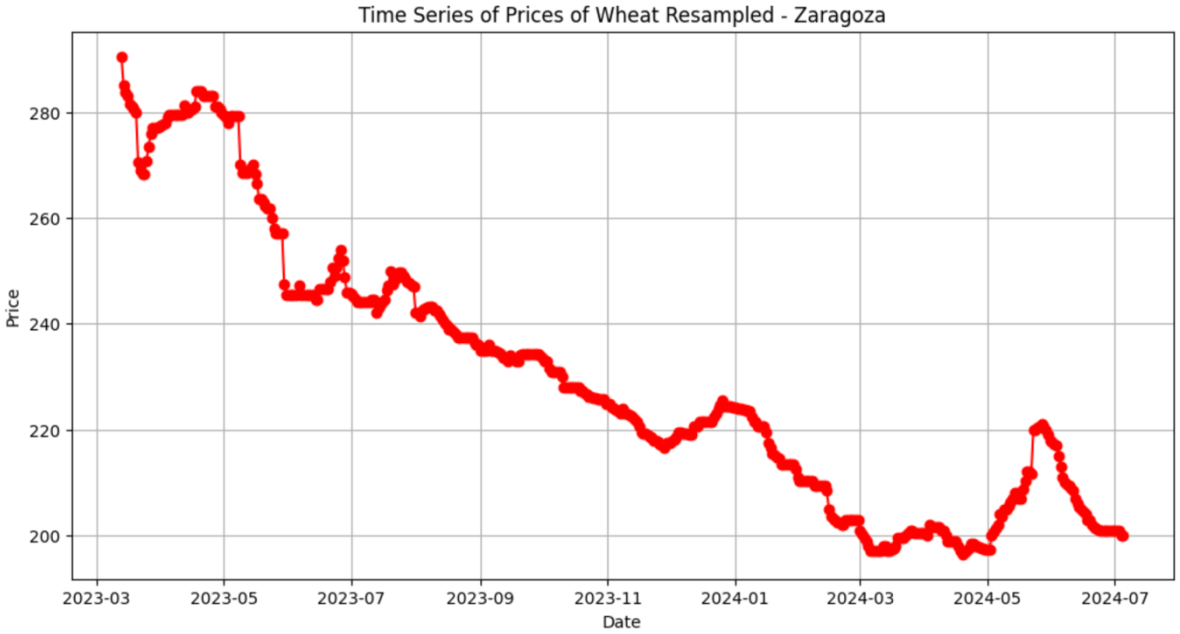
[5]:
```
final_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37000 entries, 0 to 36999
Data columns (total 19 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   id                         37000 non-null  int64
 1   date                       37000 non-null  object
 2   price                      37000 non-null  float64
 3   quantity                   26982 non-null  float64
 4   region                     26948 non-null  float64
 5   product.id                 37000 non-null  int64
 6   product.meta_product.id    37000 non-null  int64
 7   product.meta_product.name  37000 non-null  object
 8   product.meta_product.icon  35478 non-null  object
 9   product.family.id          37000 non-null  int64
 10  product.family.name        37000 non-null  object
 11  product.variety.id         37000 non-null  int64
 12  product.variety.name       37000 non-null  object
 13  data_source.id             37000 non-null  int64
 14  data_source.name           37000 non-null  object
 15  data_source.link           35917 non-null  object
 16  data_source.kind           37000 non-null  object
 17  province.id                37000 non-null  int64
 18  province.name              37000 non-null  object
dtypes: float64(3), int64(7), object(9)
memory usage: 5.4+ MB
```

We have chosen the price of wheat from Zaragoza for our univariate data analysis to determine if our hypothesis can be reproducible on one of the time series in the database; if successful, this solution can be scaled to the other time series.



To address the missing data and ensure a consistent time series, we will perform resampling to complete the dataset. By setting 'formatted_date' as the index, we can resample the data to have one sample per day, filling in missing days with the previous value using linear interpolation. This method allows us to maintain continuity in our time series and prepare the data for further analysis. Below is the code used for resampling and the resulting graph:



Weather Data

In this section, we will incorporate weather data obtained by geographical coordinates for the Zaragoza region. By analyzing this climate data in conjunction with the grain commodity prices, we aim to uncover potential relationships and influences of weather patterns on commodity prices. This integrative approach allows us to enhance our predictive models by considering external environmental factors that may impact agricultural market trends.

[179]:

	dt_iso	dt	timezone	lat	lon	temp	visibility	dew_point	feels_like	temp_min	...	weather_icon_02d	v
0	2010-01-01 00:00:00+00:00	1.262345e+09	3600.0	41.78681	-0.123294	7.591667	NaN	0.423333	4.404583	6.681250	...	0.000000	
1	2010-01-02 00:00:00+00:00	1.262432e+09	3600.0	41.78681	-0.123294	5.936250	NaN	1.635000	3.308750	5.188750	...	0.083333	
2	2010-01-03 00:00:00+00:00	1.262518e+09	3600.0	41.78681	-0.123294	5.539167	NaN	3.432500	4.485417	5.045417	...	0.000000	
3	2010-01-04 00:00:00+00:00	1.262605e+09	3600.0	41.78681	-0.123294	6.588333	NaN	4.985833	5.347500	5.953750	...	0.000000	
4	2010-01-05 00:00:00+00:00	1.262691e+09	3600.0	41.78681	-0.123294	7.656250	NaN	4.506250	6.057500	7.168750	...	0.000000	

5 rows × 51 columns

In this phase, we conduct thorough data exploration and cleaning, addressing any missing values to ensure data integrity. Additionally, we perform resampling of the 'date' column to align the temporal spacing with the previously processed price dataset. This step is crucial for maintaining consistency across datasets, enabling accurate and meaningful comparisons and analyses.

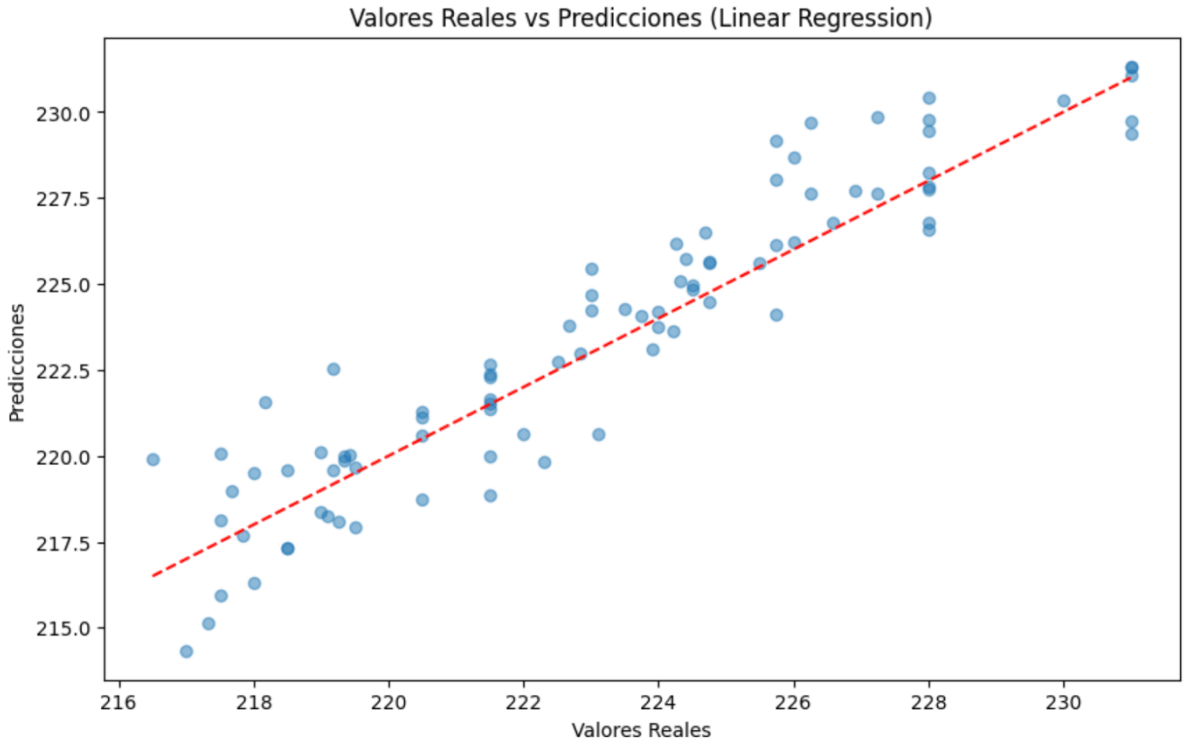
Benchmark Model

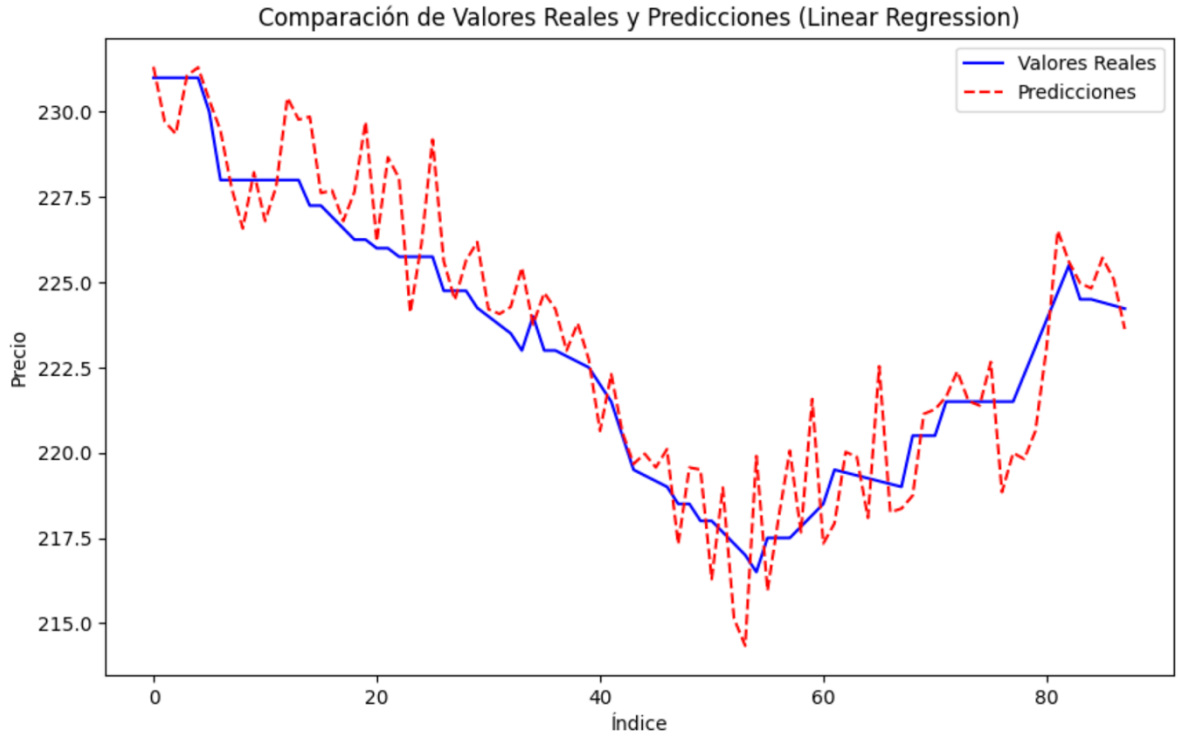
In this study, we aim to establish a benchmark model as the foundational reference for evaluating more sophisticated models aligned with our hypothesis. Given the temporal nature of the data, our initial approach involves creating lag features. These features, known as lagged variables, involve shifting each column's values backwards in time by a specified number of periods. For instance, a lag of 1 represents the immediate previous value, while a lag of 2 represents the value two periods ago, and so forth. This process allows us to incorporate historical price data as predictors, which can provide valuable insights into how past trends influence current prices. In this scenario, the prediction of the next value depends solely on the historical values within the same time series. No external variables influence the prediction process. Therefore, this approach is considered the simplest model, employing only one

variable (historical prices) for prediction. This basic model serves as a foundational step in our analysis, helping to establish a baseline performance against which more complex models, incorporating additional variables and features, will be evaluated.

[249]:
```python
# Función para crear características retardadas
def crear_caracteristicas_retardadas(df, lags):
    for col in df.columns:
        for lag in range(1, lags + 1):
            df[f'{col}_lag_{lag}'] = df[col].shift(lag)
    return df

# Crear características retardadas
lags = 3
df = crear_caracteristicas_retardadas(merged_df, lags)

# Eliminar filas con valores NaN debido a los lags
df.dropna(inplace=True)

# Definir características (X) y objetivo (y)
X = df.drop(columns=['price'])
y = df['price']
```

These were the results using a linear regressor with 3 lags of the time series of wheat prices for Zaragoza.





Mean Squared Error (Linear Regression): 2.2382096593478455
R^2 Score (Linear Regression): 0.8513807302871783

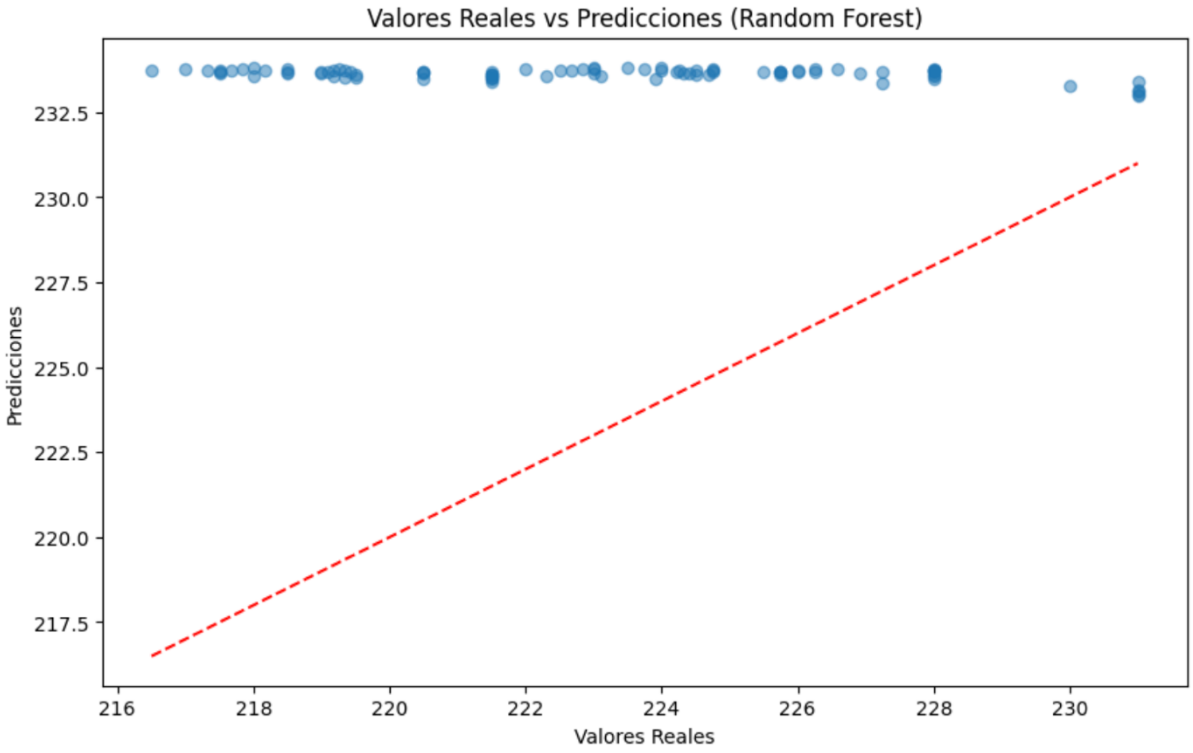
Price Prediction Using Weather Data

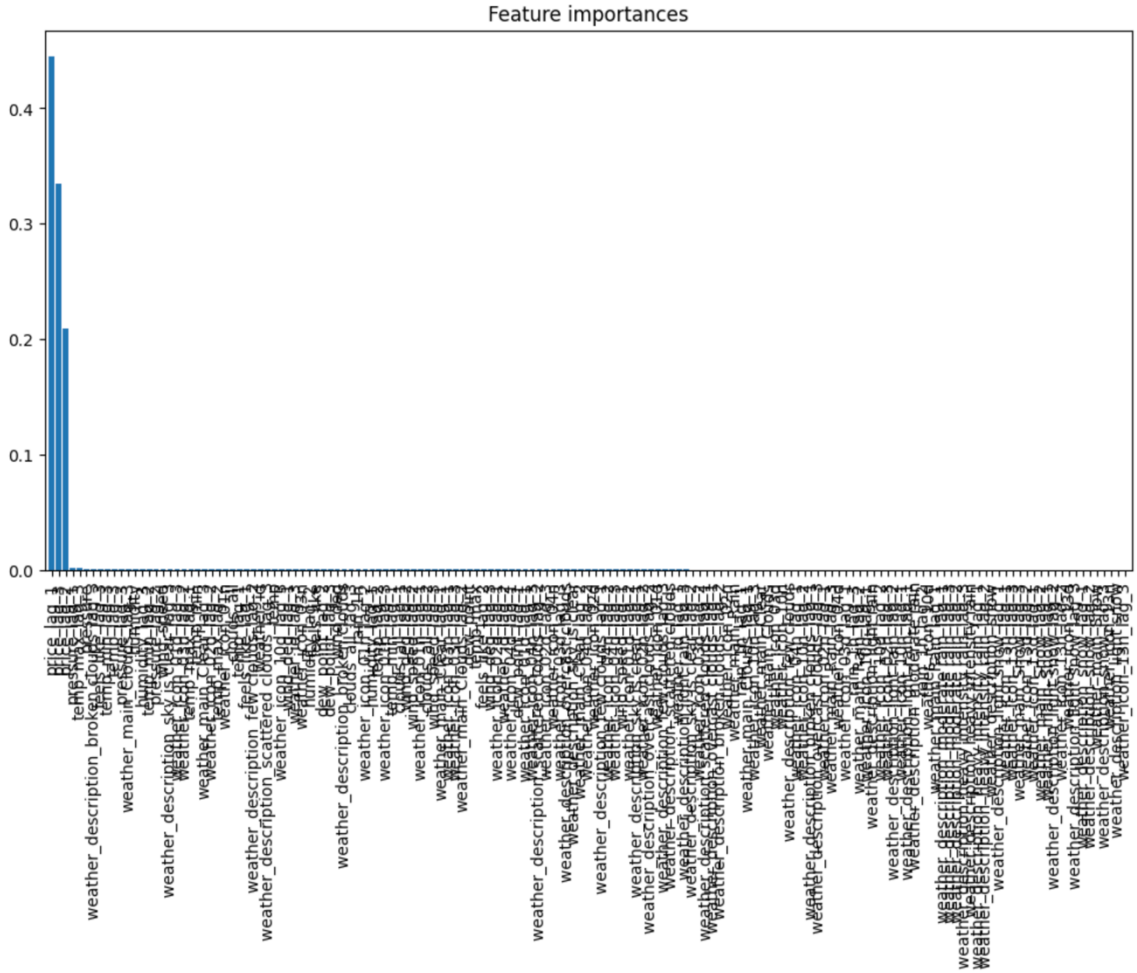
In order to better understand the factors that influence wheat prices in Zaragoza, we will perform a feature importance analysis on our dataset of wheat prices, along with their corresponding lagged values and weather data. This analysis will allow us to determine the relative importance of each variable in predicting wheat prices. By using a machine learning algorithm, such as a random forest or gradient boosting, we can train a model to predict wheat prices based on the available features and then calculate the importance of each feature in making those predictions. The results of our feature importance analysis are as follows: [insert results]. These results will provide insight into which variables have the greatest impact on wheat prices in Zaragoza and can help guide future data collection and modeling efforts.

This are the results for the first 20 columns , as we can see the importance of the variables to predict the target variable (price) is too low.

Feature ranking:
1. Feature price_lag_1 (0.4452732850929768)
2. Feature price_lag_3 (0.33401606558308533)
3. Feature price_lag_2 (0.20866628799421988)
4. Feature pressure_lag_1 (0.001238133323170466)
5. Feature temp_max_lag_3 (0.0010865772934177899)
6. Feature pressure (0.0004864019420421963 3)
7. Feature weather_description_broken clouds_lag_3 (0.00041889482468478536)
8. Feature feels_like_lag_3 (0.0004095339651271535)
9. Feature temp_min_lag_3 (0.00040570626100822033)
10. Feature temp_lag_3 (0.0003498331028258173)
11. Feature pressure_lag_3 (0.00034092119582295903)
12. Feature weather_main_Clouds_lag_3 (0.00023145581628593827)
13. Feature humidity (0.00022536432342895282)
14. Feature humidity_lag_3 (0.00019654858880071457)
15. Feature temp_min_lag_2 (0.00019408202174908652)
16. Feature pressure_lag_2 (0.00018426059992710696)
17. Feature wind_speed (0.0001820829109429744)
18. Feature weather_description_sky is clear_lag_3 (0.00018118009423936942)
19. Feature weather_icon_01d_lag_3 (0.00016775087157621754)
20. Feature weather_icon_03n_lag_2 (0.0001615813762131888)

We can see the whole comparison in the distribution of the importance of the variables in the next plot, as we can see the only related variables are the lags with the price we want to predict.


Valores Reales vs Predicciones (Random Forest)
Predicciones
Valores Reales
232.5
230.0
227.5
225.0
222.5
220.0
217.5
216
218
220
222
224
226
228
230


Feature importances
0.4
0.3
0.2
0.1
0.0

As part of our investigation into predicting wheat prices in Zaragoza, we conducted a series of experiments using different machine learning models to evaluate their performance in predicting wheat prices. We trained three models, specifically Support Vector Regression (SVR), Gradient Boosting Regressor, and Random Forest Regressor, using a dataset that included both lagged wheat prices and weather data. One of the goals of this analysis was to test the hypothesis that weather variables, which have been shown to have little importance in predicting wheat prices in Zaragoza, may not significantly improve the performance of the models.

Our results showed that, as expected, the inclusion of weather data did not significantly improve the performance of the models, it significantly decreased their performance. This suggests that weather data may not be a key driver of wheat prices in Zaragoza

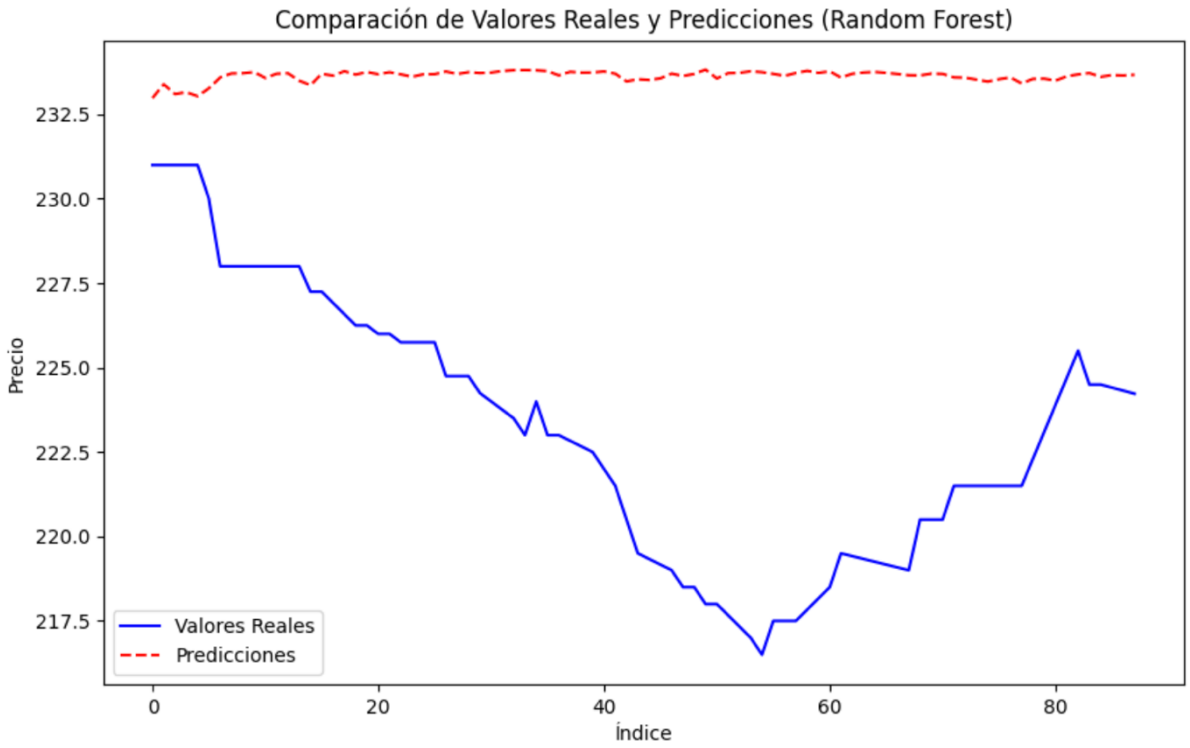
Random Forest Regressor

In the realm of forecasting, employing the Random Forest Regressor model provides a robust tool for predicting values such as prices. However, it is crucial to assess the accuracy of these predictions using standard metrics such as Mean Squared Error (MSE) and Coefficient of Determination (R^2).

In our analysis, we observed that the model yielded an MSE of 126.65, indicating a significant discrepancy between predicted and actual price values. Additionally, the R^2 Score of -7.41 suggests that the model is not effectively capturing data variations, demonstrating poor fit and limited ability to explain observed variability. These findings underscore the need to review and possibly enhance model configuration or consider alternative methods to improve the accuracy of our price predictions.

Mean Squared Error: 126.65237809161296

R^2 Score: -7.409839471809004

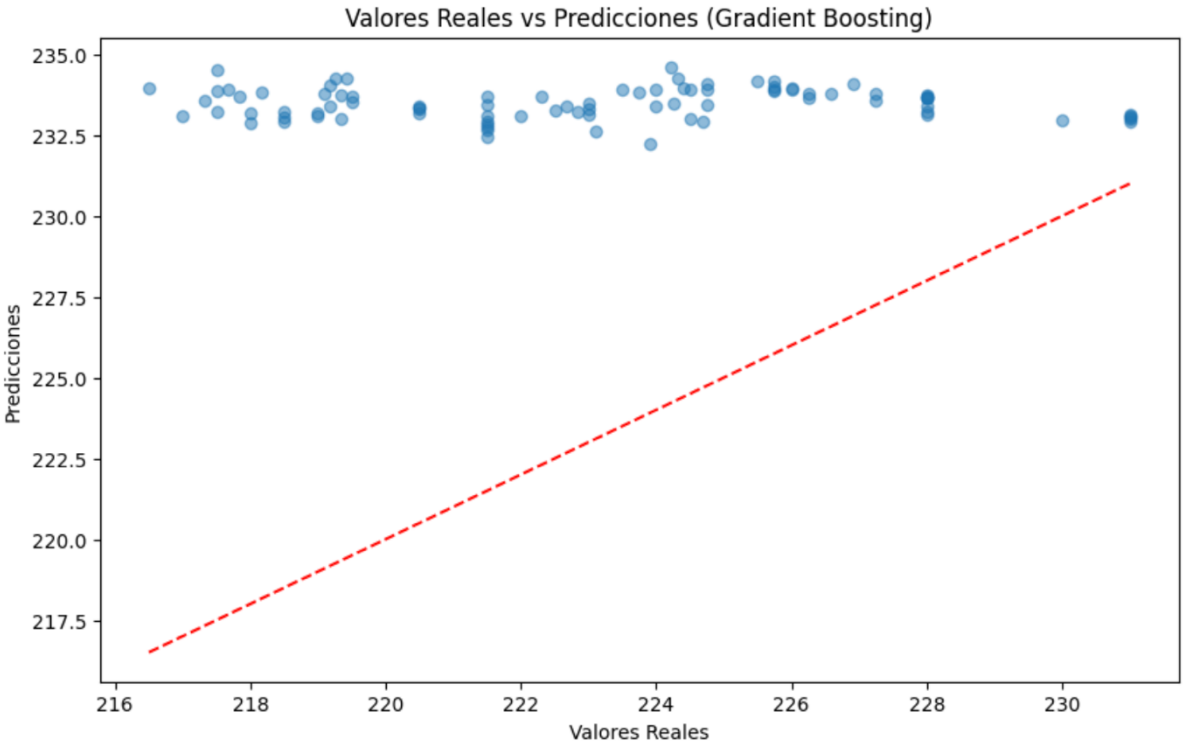


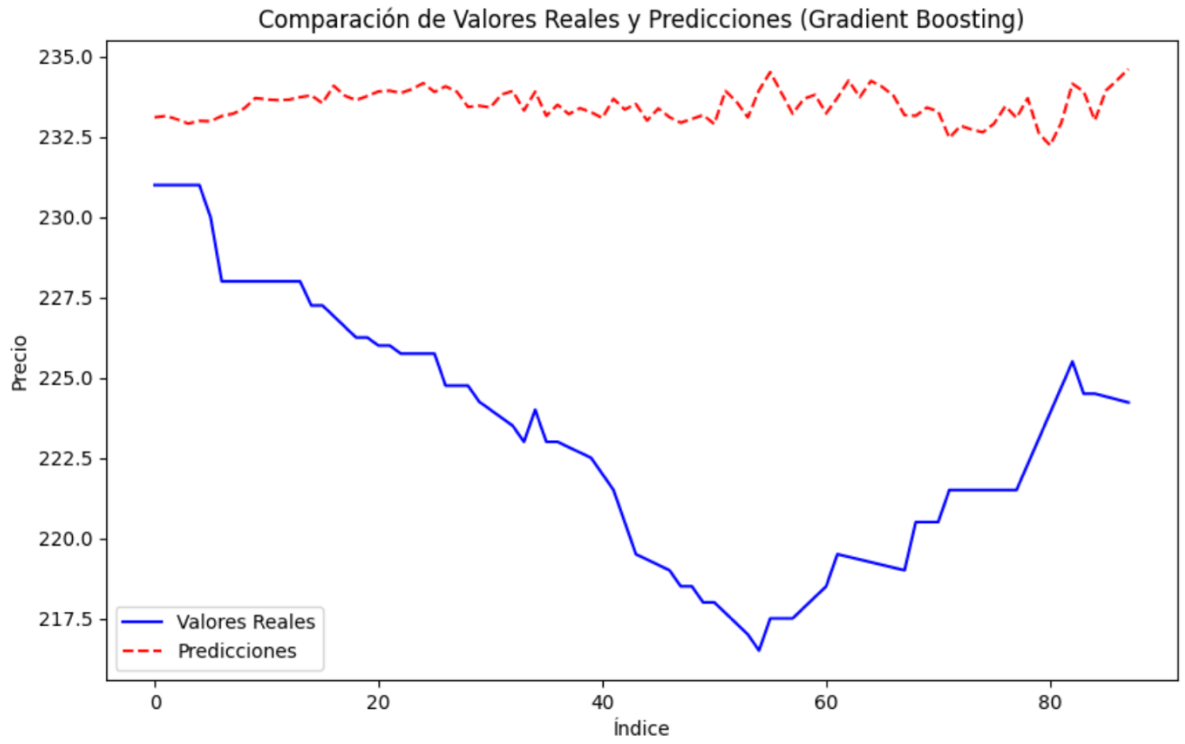
Gradient Boosting Regressor

In the domain of forecasting, employing the Gradient Boosting Regressor model provides a robust framework for predicting values such as prices. However, it is essential to evaluate the accuracy of these predictions using standard metrics such as Mean Squared Error (MSE) and Coefficient of Determination (R^2). Our analysis revealed that the model produced an MSE of 123.22, indicating a notable discrepancy between predicted and actual price values. Furthermore, the R^2 Score of -7.18 suggests that the model struggles to effectively capture data variations, showing inadequate fit and limited ability to explain observed variability. These results emphasize the importance of refining model parameters or exploring alternative methodologies to enhance the precision of our price predictions.

Gradient Boosting Mean Squared Error: 123.22196895021915

Gradient Boosting R^2 Score: -7.182057012162803


Valores Reales vs Predicciones (Gradient Boosting)
Predicciones
Valores Reales
235.0
232.5
230.0
227.5
225.0
222.5
220.0
217.5
216
218
220
222
224
226
228
230

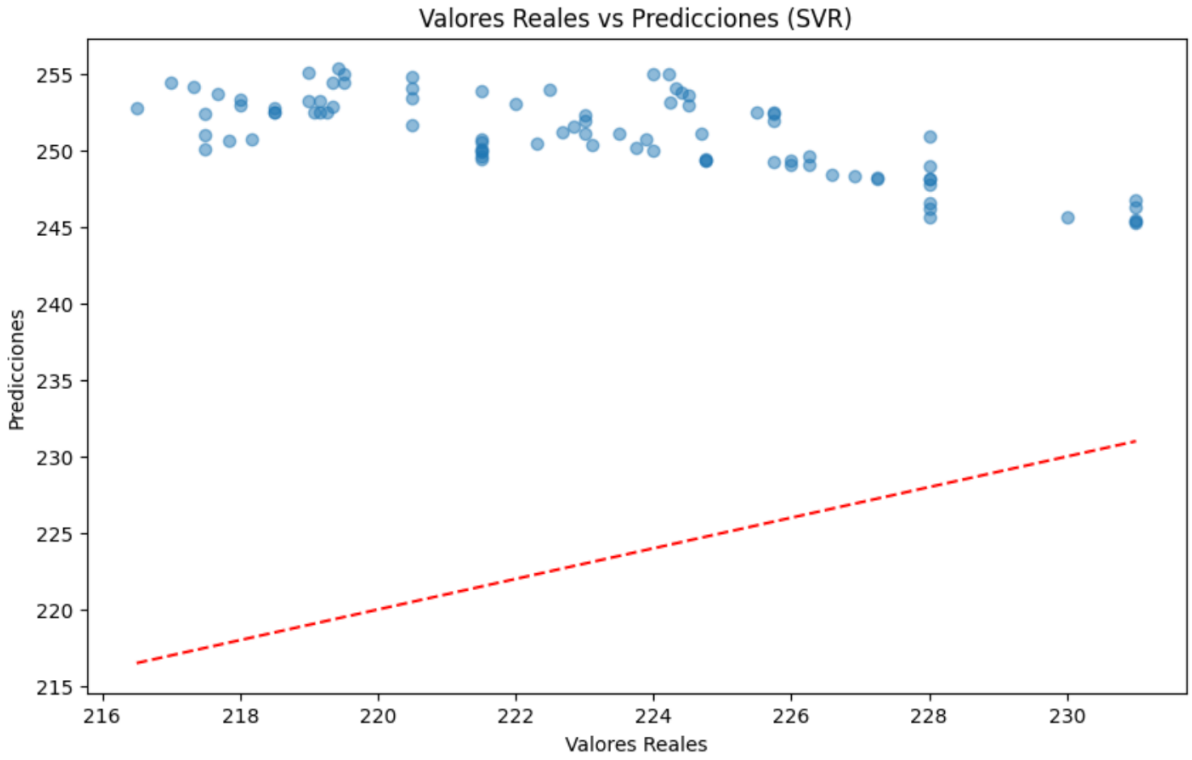

Comparación de Valores Reales y Predicciones (Gradient Boosting)
Precio
Índice
Valores Reales
Predicciones
235.0
232.5
230.0
227.5
225.0
222.5
220.0
217.5
0
20
40
60
80

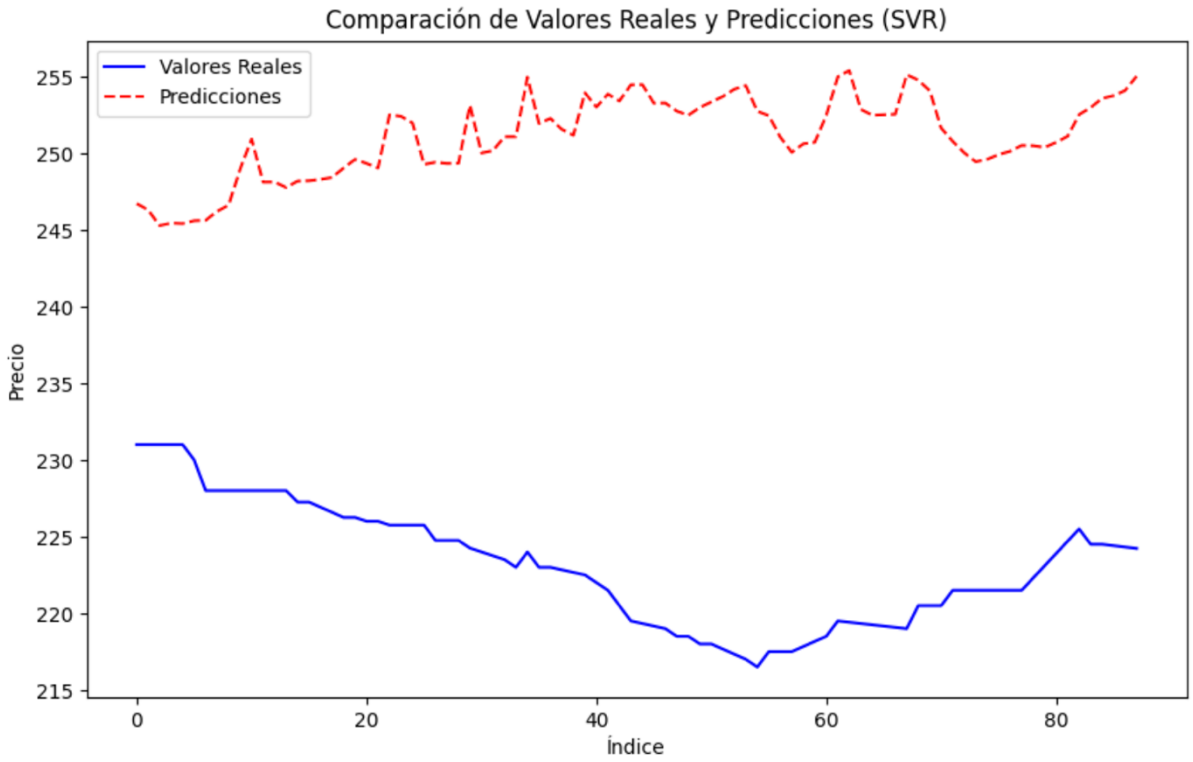
Support Vector Regressor

In the domain of forecasting, employing the Support Vector Regressor (SVR) provides a sophisticated approach for predicting values such as prices. Evaluating the accuracy of these predictions using standard metrics such as Mean Squared Error (MSE) and Coefficient of Determination (R^2) is crucial. Upon analysis, our SVR model yielded an MSE of 818.55 and an R^2 Score of -53.35.

These metrics indicate a substantial discrepancy between predicted and actual price values, suggesting that the model struggles significantly to capture the underlying data patterns. The negative R^2 Score highlights a poor fit and an inability to explain the observed variability in the data. These results emphasize the necessity of re-evaluating model parameters or exploring alternative methodologies to enhance the precision of our price predictions.

SVR Mean Squared Error: 818.5517700970266
SVR R^2 Score: -53.35262321645235




Comparación de Valores Reales y Predicciones (SVR)
Valores Reales
Predicciones
Precio
Índice
255
250
245
240
235
230
225
220
215
0
20
40
60
80

Technical Specifications

HARDWARE REQUIREMENTS

Category	Requirement	Details
Server	CPU	Minimum 4 cores (recommended 8 cores)
	Memory RAM	Minimum 2 GB (recommended 4 GB)
	Storage	3 GB SSD (recommended 20 GB SSD)

Notes: The requirements may vary based on the amount of data processed and user concurrency.

SOFTWARE REQUIREMENTS

Category	Requirement	Details
Operating System	Linux	Ubuntu 18.04 or later
Programming Language	Python	Python 3.10 or later
Development Environment	IDE	Visual Studio Code, PyCharm or Jupyter Notebook
Libraries and Frameworks	NLP	GPT-3.5 (Open AI)
	Web Frameworks	Django, DRF
Databases	SQL	PostgreSQL
DevOps Tools	Containers	Docker
	Version Control	Git, GitHub

Design and User Interface

The "Abastores Assistant" features a clean and modern design within a sleek Streamlit interface. The layout is intuitive, with a minimalist aesthetic that ensures ease of navigation. Users are greeted with a welcoming message, followed by a clear and concise chat interface. The design employs a light color scheme with soft accents, creating a professional yet approachable feel. The interface is designed to be user-friendly, with straightforward prompts and response areas, ensuring that users can effortlessly interact with the assistant. Overall, the design emphasizes simplicity and functionality, providing a seamless user experience.

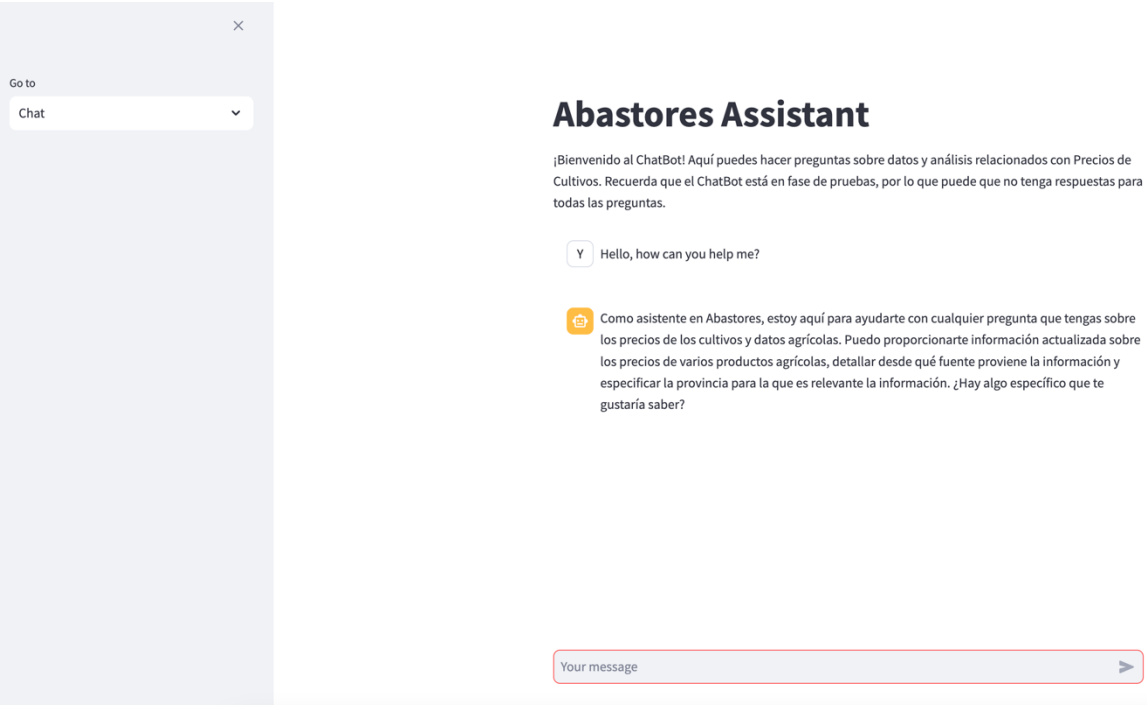


Image 4. Abastores Assistant UI

Testing and Quality Assurance

With the purpose of testing the LLM responses, we decided to prepare a set of 100 questions to be solved. These questions were designed with the objective of detecting failed responses, hallucinations, and jailbreak cases. Ensuring the reliability and accuracy of the LLM is crucial for its effectiveness and user trust.

Questions <Sample>:

- Explain the difference between machine learning and deep learning.
- Tell me a story about a robot that discovered its passion for painting.
- What are the ethical implications of artificial intelligence in healthcare?
- How does blockchain technology work, and what are its primary uses?
- Discuss the impact of climate change on ocean biodiversity.
- Imagine a future where humans can communicate with animals. Describe a day in such a world.
- What steps can individuals take to reduce their carbon footprint?
- Explain the concept of quantum computing and its potential impacts on society.
- Describe a historical event from the perspective of someone who lived through it.
- How can technology be used to improve educational access in remote areas?
- Create a dialogue between two historical figures from different eras.
- What are the challenges of space exploration, and how can humanity overcome them?
- Explain the process of photosynthesis in detail.
- Describe the cultural significance of tea in various countries.
- How do social media platforms impact human psychology and social dynamics?
- Discuss the importance of biodiversity and ecosystems for human survival.
- Imagine a world where renewable energy fully powers cities. Describe the transition.
- What are the pros and cons of telecommuting for both employees and employers?
- Explain the role of antibiotics and the issue of antibiotic resistance.

- Tell a story about a time traveler who accidentally changed history.
- What measures can governments take to protect digital privacy and security?
- Describe the life cycle of a star from birth to death.
- How do cultural differences affect communication in international business?
- What are the psychological effects of long-term isolation, and how can they be mitigated?
- Explain the principle of supply and demand in economics.
- Tell me about a fictional city where all transport is eco-friendly.
- Discuss the role of art and music in human societies throughout history.
- What strategies can be used to combat misinformation online?
- Describe the process of evolution and natural selection.
- Imagine a society where AI assists in making government decisions. Discuss the benefits and drawbacks.

Impact Test

This test consisted of checking whether the LLM could respond with coherence and reason. Normal questions that the LLM should be able to answer were included in this test. The primary focus was on evaluating the LLM's ability to provide accurate, relevant, and contextually appropriate responses to a variety of queries. This helps in understanding the LLM's general performance and its capability to handle standard interactions.

Fuzzled Test

This test involved checking whether the LLM would be able to differentiate normal questions from those that should not be answered. It allowed us to assess the LLM's knowledge limits and its ability to recognize and appropriately respond to ambiguous or nonsensical queries. By testing the LLM's boundaries, we can ensure that it provides meaningful responses and avoids confusion when faced with unclear or unusual inputs.

Adversary Test

This test focused on determining whether the LLM could differentiate normal questions from harmful ones. The goal was to evaluate the LLM's security measures and its ability to identify and mitigate potentially dangerous or malicious queries. Ensuring that the LLM does not respond to harmful inputs is critical for maintaining the safety and integrity of the system. This test also helps in assessing the robustness of the LLM's safeguards against adversarial attacks and exploitation attempts.

Comprehensive Evaluation

The combination of these tests provides a comprehensive evaluation of the LLM's performance, reliability, and security. By systematically analyzing the responses across various scenarios, we can identify areas for improvement and enhance the overall quality of the LLM. Continuous testing and refinement are essential for keeping the LLM up-to-date and capable of handling diverse and evolving user interactions.

Scalability and Prospects

Roadmap

This roadmap outlines the key phases and milestones necessary to successfully develop, launch, and expand the AGIA platform. By following this plan, AGIA aims to revolutionize the agricultural commodities trading sector through innovative AI-driven insights and strategic partnerships.

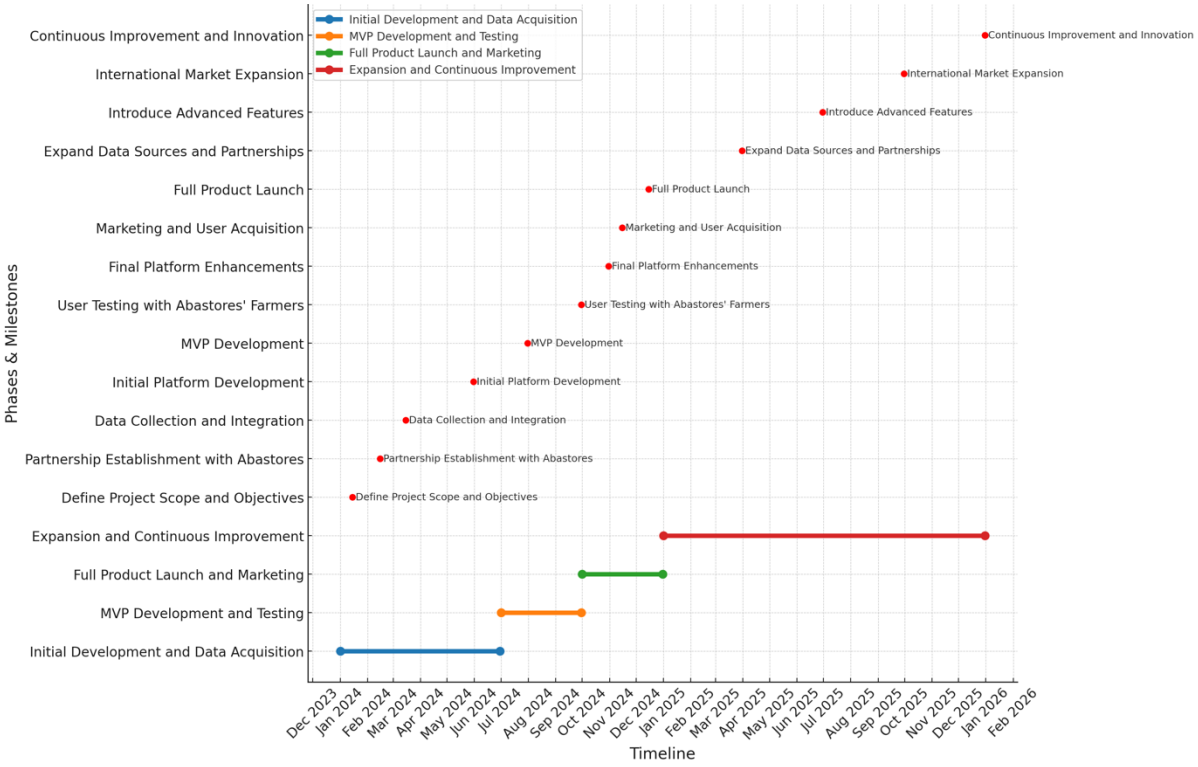


Image 5. AGIA's Roadmap

Learnings

The development and execution of the AGIA project have provided numerous valuable insights and learnings across various dimensions, including technical, operational, and strategic aspects. Here are some of the key learnings:

Technical Learnings

1. **Integration of AI and Data Analytics:**
    - Developing advanced AI algorithms for predictive analytics and real-time insights highlighted the importance of robust data processing and machine learning models.
    - The integration of diverse data sources, such as historical prices, environmental indicators, and market trends, is critical for generating accurate and actionable insights.

2. **Platform Development:**
    - Building a user-friendly chatbot interface demonstrated the need for intuitive design and seamless user experience.
    - Ensuring scalability and robustness in the platform required a strong focus on backend infrastructure and continuous testing.
3. **Data Quality and Management:**
    - The project underscored the importance of high-quality data for reliable analytics. Cleaning and preprocessing data are essential steps in this process.
    - Efficient data management practices, including data storage, retrieval, and real-time processing, are vital for maintaining platform performance.

Operational Learnings

1. **Effective Partnership Management:**
    - Collaborating with Abastores highlighted the significance of clear communication and well-defined roles and responsibilities in partnerships.
    - Establishing mutually beneficial agreements and aligning on common goals are key to successful long-term collaborations.
2. **User Engagement and Feedback:**
    - Conducting user testing and gathering feedback from Abastores' farmers provided valuable insights into user needs and preferences.
    - Iterative development, based on user feedback, is crucial for refining the product and ensuring it meets market demands.
3. **Project Management and Execution:**
    - Adhering to a structured project roadmap helped in organizing tasks, setting clear milestones, and tracking progress effectively.
    - Flexibility and adaptability in project plans are necessary to accommodate unforeseen challenges and changes in scope.

Strategic Learnings

1. **Market Research and Competitive Analysis:**
    - Conducting thorough market research revealed gaps in existing solutions and helped in identifying unique value propositions for AGIA.
    - Understanding the competitive landscape enabled us to differentiate AGIA by focusing on predictive analytics and customized insights, rather than just data dashboards.
2. **Business Model Development:**
    - Developing a tiered subscription model provided insights into pricing strategies and how different segments of users value various features.
    - Exploring revenue-sharing models with partners like Abastores highlighted the importance of aligning financial incentives for all stakeholders.
3. **Scalability and Expansion:**
    - Planning for international market expansion and incorporating advanced features into future iterations emphasized the need for scalable solutions.

Continuous innovation and improvement are essential for maintaining a competitive edge and addressing evolving market needs.

Conclusion

The AGIA project signifies a major advancement in the agricultural commodities trading sector, integrating cutting-edge AI-driven analytics and real-time data insights into a user-friendly chatbot platform. Throughout this report, we have demonstrated AGIA's potential to transform how market operators, farmers, and other stakeholders navigate the complexities of agricultural markets.

AGIA offers a unique value proposition by providing real-time agricultural insights, historical price data, advanced predictive analytics, and customizable reports through an intuitive chatbot interface. Our strategic partnership with Abastores has been instrumental in validating the concept, accessing critical data, and refining the product based on real-world feedback. The tiered subscription model and revenue-sharing agreements position AGIA for sustainable growth, catering to a diverse range of user segments from small farmers to large agribusinesses.

By leveraging sophisticated AI algorithms and integrating diverse data sources, AGIA delivers accurate and actionable market insights that empower users to make informed decisions. The platform’s scalable infrastructure, microservices architecture, and user-centric design ensure reliable performance and ease of use, accommodating future growth and feature enhancements. Personalized customer support, regular updates, and community engagement initiatives foster strong relationships with users, driving retention and satisfaction.

AGIA's strategic vision includes market expansion into new geographic regions and diversification into other commodities sectors. Ongoing feature development and enhancements, driven by user feedback and market trends, will keep AGIA at the forefront of agricultural market analytics. By focusing on scalable solutions and leveraging strategic partnerships, AGIA aims to achieve long-term success and profitability while contributing to the efficiency and transparency of agricultural markets.

Challenges such as ensuring high-quality data integration, balancing advanced technical capabilities with intuitive user experiences, and maintaining operational efficiency as AGIA scales will require continuous attention and refinement. However, the journey of AGIA, from conceptualization to execution, highlights the importance of innovation, collaboration, and strategic planning in creating impactful solutions that drive industry transformation.

As we look to the future, AGIA is committed to continuous improvement and expansion, striving to deliver unparalleled value to its users and making significant contributions to the agricultural commodities market. The insights and learnings from this project will guide our efforts in ensuring AGIA remains a pioneering force in the industry, fostering growth, efficiency, and sustainability for years to come.