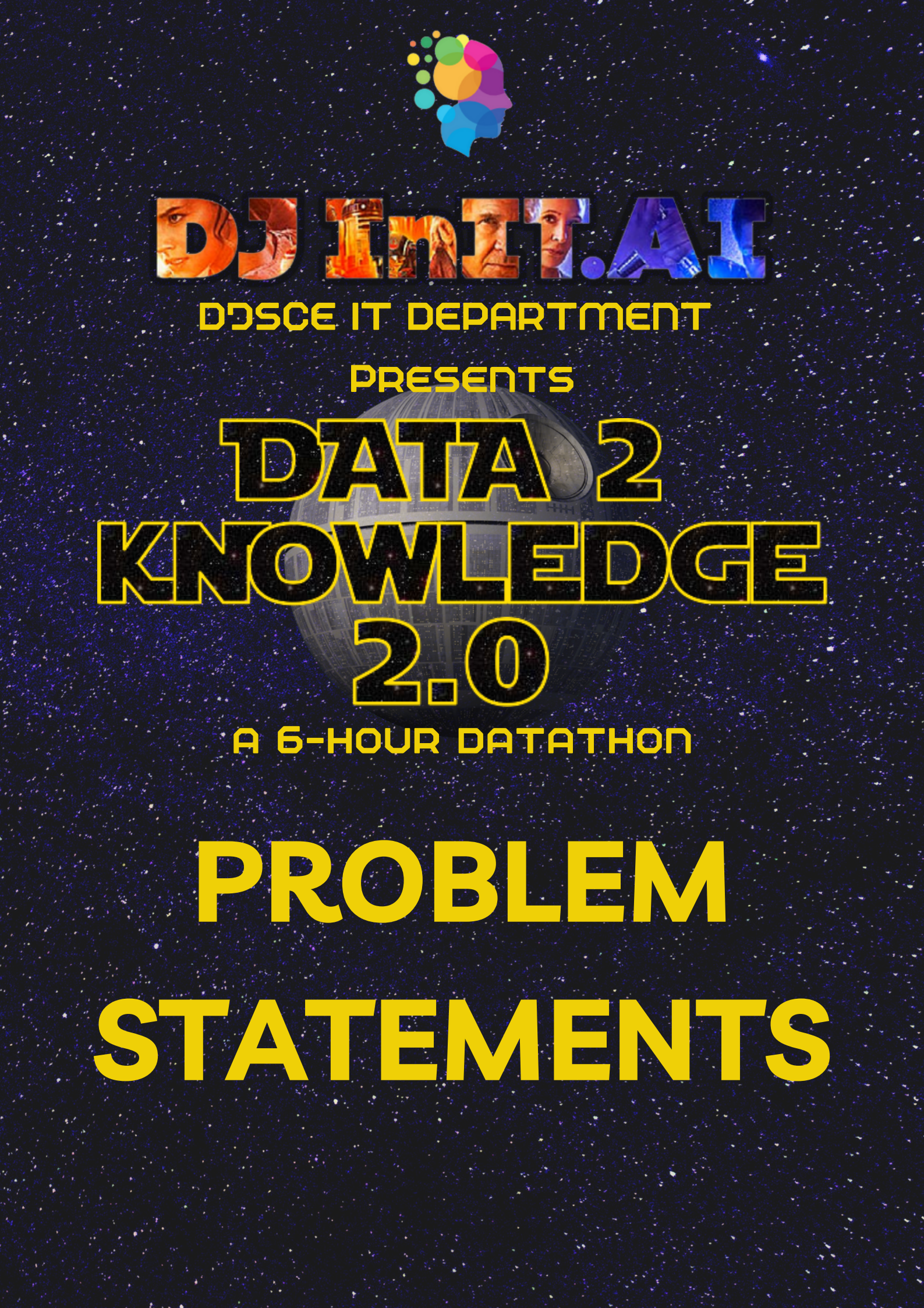

DJ InIT.AI
DJSCE IT DEPARTMENT
PRESENTS
DATA 2 KNOWLEDGE 2.0
A 6-HOUR DATATHON
PROBLEM STATEMENTS

DATA SCIENCE

POWERED BY

Dextra

delivering 10X transformation

Agritech: To Enhance Operational Efficiency through Data Report Insights using a Research Assistant.

Background :

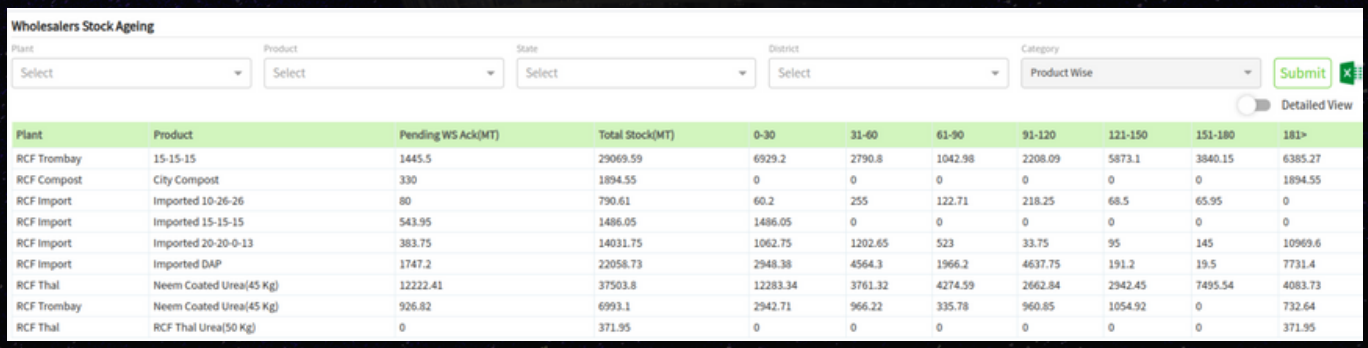
Organizations routinely generate diverse reports serving distinct purposes, often containing vast datasets. Extracting meaningful insights from these reports is paramount for informed decision-making in various business domains. The process of deriving insights from extensive datasets is not only laborious, error-prone, and time-consuming but also constrained by the creative and analytical capacities of the individuals interpreting the data. This limitation hinders the potential breadth and depth of actionable solutions derived from these insights.

A pivotal example in the market is from the agricultural industry, where the fertilizer companies distribute subsidies to farmers as a part of government initiatives through a complex supply chain. To ensure the successful execution of subsidy programs, it is imperative to manage stock efficiently and prevent the loss of subsidies. The current manual system relies on field officers to address issues like zero sales and aging stock at wholesalers and retailers end, ensuring subsidy retention.

Recognizing this, our organization has implemented a robust system that provides a detailed overview of stock aging at both retailer and wholesaler levels within the designated jurisdictions or districts. This comprehensive Aging Report serves as a strategic tool for key decision-makers responsible for overseeing multiple districts.

While Aging Report empowers users with consolidated information regarding the aging status of stock across their entire portfolio of districts, by leveraging this data, extracting valuable distribution insights remains a crucial yet challenging task for users.

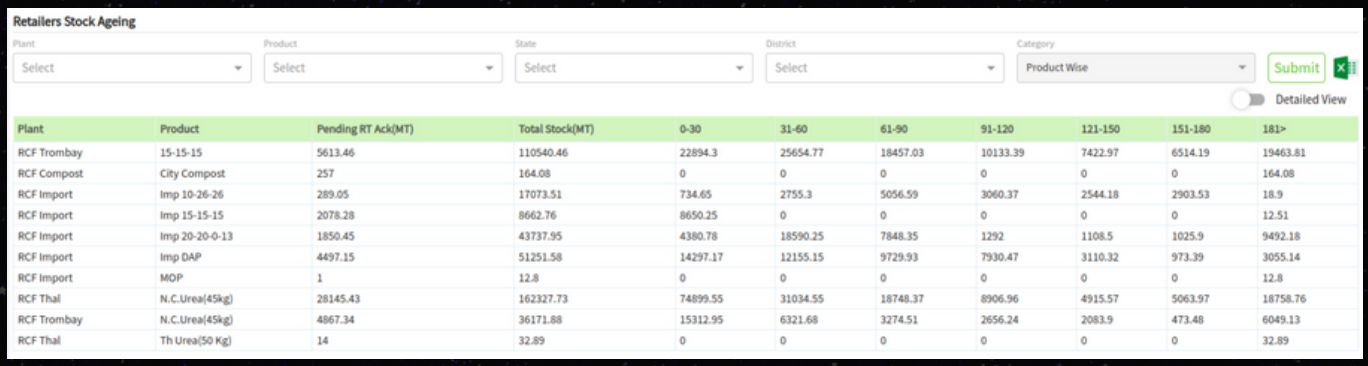
Sample of Aging Report :



Wholesalers Stock Ageing

Plant: Select | Product: Select | State: Select | District: Select | Category: Product Wise | Submit | Detailed View

Plant	Product	Pending WS Ack(MT)	Total Stock(MT)	0-30	31-60	61-90	91-120	121-150	151-180	181>
RCF Trombay	15-15-15	1445.5	29069.59	6929.2	2790.8	1042.98	2208.09	5873.1	3840.15	6385.27
RCF Compost	City Compost	330	1894.55	0	0	0	0	0	0	1894.55
RCF Import	Imported 10-26-26	80	790.61	60.2	255	122.71	218.25	68.5	65.95	0
RCF Import	Imported 15-15-15	543.95	1486.05	1486.05	0	0	0	0	0	0
RCF Import	Imported 20-20-0-13	383.75	14031.75	1062.75	1202.65	523	33.75	95	145	10969.6
RCF Import	Imported DAP	1747.2	22058.73	2948.38	4564.3	1966.2	4637.75	191.2	19.5	7731.4
RCF Thal	Neem Coated Urea(45 Kg)	12222.41	37503.8	12283.34	3761.32	4274.59	2662.84	2942.45	7495.54	4083.73
RCF Trombay	Neem Coated Urea(45 Kg)	926.82	6993.1	2942.71	966.22	335.78	960.85	1054.92	0	732.64
RCF Thal	RCF Thal Urea(50 Kg)	0	371.95	0	0	0	0	0	0	371.95



Retailers Stock Ageing

Plant: Select | Product: Select | State: Select | District: Select | Category: Product Wise | Submit | Detailed View

Plant	Product	Pending RT Ack(MT)	Total Stock(MT)	0-30	31-60	61-90	91-120	121-150	151-180	181>
RCF Trombay	15-15-15	5613.46	110540.46	22894.3	25654.77	18457.03	10133.39	7422.97	6514.19	19463.81
RCF Compost	City Compost	257	164.08	0	0	0	0	0	0	164.08
RCF Import	Imp 10-26-26	289.05	17073.51	734.65	2755.3	5056.59	3060.37	2544.18	2903.53	18.9
RCF Import	Imp 15-15-15	2078.28	8662.76	8650.25	0	0	0	0	0	12.51
RCF Import	Imp 20-20-0-13	1850.45	43737.95	4380.78	18590.25	7848.35	1292	1108.5	1025.9	9492.18
RCF Import	Imp DAP	4497.15	51251.58	14297.17	12155.15	9729.93	7930.47	3110.32	973.39	3055.14
RCF Import	MOP	1	12.8	0	0	0	0	0	0	12.8
RCF Thal	N.C.Urea(45kg)	28145.43	162327.73	74899.55	31034.55	18748.37	8906.96	4915.57	5063.97	18758.76
RCF Trombay	N.C.Urea(45kg)	4867.34	36171.88	15312.95	6321.68	3274.51	2656.24	2083.9	473.48	6049.13
RCF Thal	Th Urea(50 Kg)	14	32.89	0	0	0	0	0	0	32.89

Introduce a Research Assistant that will help the users gain valuable insights from any Report.

To enhance the utility of Data Reports, we propose the introduction of a Research Assistant – a digital tool designed to empower users with valuable insights, regardless of the report's nature. The model should be able to recognize what is the report about and what data the report holds.

In the example of Aging Report, Territory Managers will interact with the Research Assistant by asking questions related to the Aging Report. To guide their thought process and clarify the role of the Research Assistant, recommended or sample questions will be visible on the screen. This should streamline the user experience and help Territory Managers understand the potential of the Research Assistant in deriving actionable insights from the Aging Report.

Expected Solution :

1. The model should have the ability to identify the content and purpose of the report, recognizing the specific data it encompasses.

2. The solution is expected to grant the admin/backend team the capability to train the research assistant by imparting contextual understanding of the report.

3. Furthermore, it should facilitate the configuration of positive and negative outcomes while offering flexibility in setting validation parameters as required.

4. The assistant should be capable of supporting regional languages.

5. The Research Assistant should possess the capability to correlate information between two reports based on user queries.

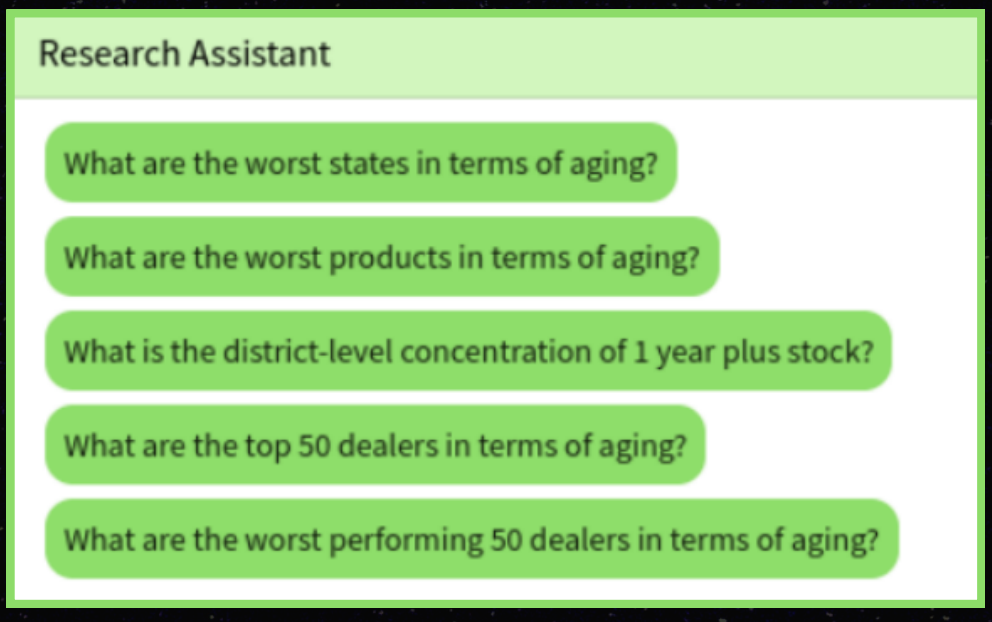
- For instance, if the aging of stock is elevated in a specific district, the Research Assistant should recommend an action to contact the Field Officer responsible for that district. The relevant district and Field Officer information, along with their performance metrics, might be stored in another report. In such scenarios, the Research Assistant is expected to provide insights derived from the correlation of data between these two reports

6. Users should be able to gain insights from the Report by asking the Research Assistant Questions.

- For eg. Territory Managers should be able to gain insights from the Aging report.
- Research Assistant should utilize the report information to identify specific trends prevailing in the markets where Territory Managers operate allowing them to make informed decisions based on localized insights.

7. Research Assistant should empower users to isolate outliers, pinpoint areas of concern, and recommend targeted actions to address any issues (for eg. related to aging stock).

Some of the examples of questions or insights that Territory Managers may require are mentioned below :



Evaluation Criteria:

1. Coverage of Questions: The solution will be evaluated based on its ability to address a variety of questions that the users may have regarding the Report Data.
For example : The questions that Territory Managers may have regarding stock aging.

a. Which are the Top 3 worst performing states in terms of aging?
b. Which district has the best stock aging?
c. Which state has the highest quantity of stock pending to be sold?

2. Correlation Accuracy: Assess the Research Assistant's precision in correlating data between two reports, ensuring accurate identification of patterns and relationships. Evaluate the ability of the Research Assistant to generate effective insights and actionable recommendations based on correlated data.

3. Contextual Training : Assess the solution's capability to facilitate the admin team in providing contextual understanding to train the Research Assistant effectively. Evaluate the solution's interface that enables configuration of positive and negative outcomes and to set validation parameters to enhance the training process.

Appendix :

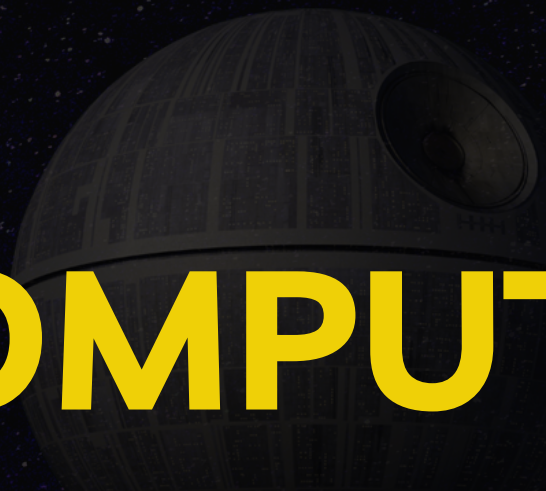
Term	Definition
Subsidy	A financial aid or support provided by the government to fertilizer companies for the purpose of reducing the cost of fertilizers for farmers. Subsidies aim to promote agricultural productivity and make fertilizers more accessible to farmers.
Aging Report	A comprehensive report that provides detailed information on the aging status of fertilizer stock at both retailer and wholesaler levels within specific jurisdictions or districts. The report includes data on the duration products have been in stock
Zero Sales	Occurs when no sales transactions for fertilizer products take place within a specific period. Addressing zero sales is a critical aspect of the subsidy management system as it helps prevent the loss of subsidies tied to product sales.

Term	Definition
Field Officers	Personnel assigned by fertilizer companies to engage with dealers (wholesalers and retailers). Field Officers play a key role in addressing zero sales, clearing aging stock, and ensuring effective communication within the distribution chain.
Retailer	An intermediary entity within the fertilizer distribution chain responsible for selling fertilizer products directly to end consumers, typically farmers. Retailers operate at the last stage of the distribution process, connecting directly with farmers for product transactions.
Wholesaler	An intermediary entity that purchases fertilizer products in bulk from manufacturers and sells them to retailers. Wholesalers play a crucial role in the distribution chain by facilitating the flow of products from manufacturers to retailers, ensuring availability at the retail level.

Term	Definition
Territory Manager	Territory managers oversee multiple sales teams of a company in a certain geographical area. They often work with different departments to help increase company sales and revenue through employee training, improved customer service tactics and impressive sales plans.

Datasets :

DataScience_PS_Dataset



COMPUTER VISION

POWERED BY

Dextra

delivering 10X transformation

Healthtech : Product Positioning Analysis through Image Recognition

Background :

In the competitive landscape of the pharmaceutical industry, the strategic positioning of products in retail outlets plays a pivotal role in influencing customer visibility and sales. To incentivize shop owners for optimal product placement, pharmaceutical companies offer perks and discounts based on contractual agreements. However, the manual inspection and analysis of product positions by Field Officers prove to be a time-consuming and intricate process.

Problem Description:

Currently, Field Officers from pharmaceutical companies conduct manual drives, visiting shop outlets to inspect and photograph the positioning of their company's products. The incentives and offers offered to shop owners are contingent upon the visibility and strategic placement of the pharmaceutical products. Higher incentives are granted for products that are immediately visible to customers, while obscured or poorly positioned products receive little to no incentives.

The manual inspection process involves analyzing images based on various parameters for example:

1. Visibility of the product from the shop front.
2. Placement of the product in corners or beside competitors.
3. Lighting conditions around the product.
4. Shelf placement of the product.
5. The angle from which the image was captured (e.g., shop front, sideways, lower angle).

Proposed Solution :

Introduce an Image Recognition System for Field Officers, streamlining and automating the product positioning analysis process. The application should empower Field Officers to capture images of the product setup within shops, and the Image Recognition System will analyze these images based on predefined parameters. The system should provide valuable insights to Field Officers and assist in deriving the incentive value for each product in every shop.

Expected Features of the Image Recognition System :

1. **Automated Analysis:** The application should automatically analyze product images, considering various parameters
   - Identification of boundary boxes for affiliated brands to easily recognise the products in the picture
   - Assessment of their proportional shelf occupancy.
   - Determination of their relative positions, considering factors like eye-level, visibility, lighting, bottom placement, and proximity to corners.

2. Criteria Definition for Admin: Enable administrators to define criteria for assessing the product position. This ensures flexibility and adaptability to changing business requirements.

3. GRID Configuration: Allow administrators to define the GRID in which product placement will be captured. Users should have the flexibility to define criteria in each block of the GRID, providing a customizable and granular approach to capturing product placement.

4. Image Capture Flexibility: Enable Users to capture placement images in real-time and facilitate the upload of product placement images for subsequent analysis.

5. Insight Generation: Provide actionable insights to Field Officers based on the analysis, helping them understand the effectiveness of the current product positioning.

6. Incentive Derivation: Assist Field Officers in deriving the incentive value for each product in each shop based on the analysis.

7. User-Friendly Interface: Ensure the application is intuitive and easy to use, enabling seamless integration into the workflow of Field Officers.

Evaluation Criteria :

1. Accuracy of Image Recognition: Evaluate the precision and accuracy of the Image Recognition System in identifying and analyzing key parameters.

2. GRID Configuration: Assess the level of customization offered in defining the GRID for capturing product placement, allowing users to set criteria independently for each block within the GRID.

3. Image Capture Flexibility: Evaluate the system's efficiency in facilitating real-time capture as well as upload images of product placements.

4. Criteria Definition for Admin : Assess the system's flexibility in allowing administrators to define and modify criteria for evaluating product position.

MLOPS

Problem Statement:

Develop an MLops Pipeline for Bias Detection and Mitigation in Heart Disease Diagnosis Models

Objective:

Create an end-to-end MLops pipeline that automates the identification and mitigation of potential biases in Heart Disease diagnosis models, aiming to enhance fairness and reliability across diverse demographic groups. The goal is to ensure equitable predictions and address disparities that may arise from biases in the training data or model architecture. Participants are encouraged to incorporate ensemble techniques to enhance model accuracy.

Background:

As machine learning advancements revolutionize healthcare diagnostics, it is crucial to guarantee that models are unbiased and provide fair predictions for various patient populations. In the context of Heart Disease diagnosis, biases can lead to accuracy discrepancies among different demographic groups. This Datathon challenges participants to build a comprehensive MLops pipeline to detect and mitigate biases in Heart Disease diagnosis models.

Deliverables:

1. Jupyter notebook or codebase containing the MLops pipeline.
2. Deployment of the model and pipeline.
3. Documentation explaining the approach, algorithms used, and the rationale behind bias detection and mitigation strategies, emphasizing the integration of ensemble techniques for model optimization.
4. Interface and visualization of findings, including reasoning behind the chosen approach.
5. Presentation outlining key findings, challenges faced, and potential improvements.

Participants are encouraged to collaborate and innovate, leveraging the latest advancements in MLops and fairness-aware machine learning techniques. The ultimate aim is to contribute to developing ethical and unbiased Heart Disease diagnosis models for improved healthcare outcomes.

Evaluation Criteria:

Participants will be evaluated based on the following criteria:

- Accuracy of the Heart Disease diagnosis model, considering ensemble techniques.
- Effectiveness in identifying biases across demographic groups.
- Transparency and interpretability of the MLops pipeline.
- Efficiency of bias mitigation strategies, particularly in an ensemble setting.
- Automation and monitoring capabilities for long-term model fairness.

Dataset:

MLOPS_PS_Dataset

NLP

Problem Statement:

Develop a Text-based Geolocation Extraction system to extract the geographical location or origin of given text content. The goal is to associate textual data, such as social media posts or news articles, with specific geographic locations, ultimately facilitating applications in content recommendation and regional linguistic analysis.

Objective:

Create a robust model that, given a piece of text, accurately extracts the likely geographical location or origin from where the text originated.

Background:

In an era where vast amounts of textual data are generated daily, associating this data with geographic locations has significant implications. The Text-based Geolocation Extraction system aims to enhance content recommendation and provide insights into regional linguistic variations. The extracted geolocation information can be visually represented on a map for a comprehensive understanding.

Deliverables:

1. **Model Implementation**: Develop a model capable of extracting geolocation information based on text input.
2. **Visualization:** Create a graphical representation of extracted geolocation information on a map.
3. **Documentation:** Provide clear and concise documentation detailing the model architecture, training process, and how to use the visualization.

Upon successful extraction of geolocation information, the locations will be visually represented on an interactive map. Each extracted location will be marked with a distinct marker or pin, allowing users to explore and analyze the distribution of extracted geolocations across the globe. This graphical representation aims to enhance the interpretability of the model's extractions and provide valuable insights into the geographical context of the analyzed text data.

Evaluation Criteria:

Participants will be evaluated based on the following criteria:

- **Model Accuracy:** The accuracy of the geolocation extraction model on a designated test dataset.
- **Visualization Quality:** The effectiveness and clarity of the graphical representation of geolocation information on the map.
- **User Interface:** The usability and user-friendliness of the visualization tool.
- **Documentation Quality:** The completeness and clarity of the documentation provided for the model and visualization tool.
- **Generalization:** The model's ability to generalize well to diverse textual data and accurately extract geolocation across different regions.

Dataset:

NLP_PS_Dataset

GENERATIVE AI

Problem Statement:

Content creators often struggle with optimizing their content for discoverability and engagement. Identifying relevant keywords and generating trending hashtags can be time-consuming and challenging. This problem statement aims to address these issues by developing a user-friendly tool that combines an accurate Keyword Extraction Module with a GenAI-Powered Hashtag Generator to streamline the content creation process.

Objective:

The main objective is to empower content creators with a tool that quickly extracts essential keywords from their content and generates trending hashtags through the use of Genetic Artificial Intelligence (GenAI). This tool aims to enhance content visibility on social media platforms by aligning keywords and hashtags with current trends and user preferences.

Background:

Content creators need efficient solutions for keyword extraction and hashtag generation to optimize their content for social media platforms. By integrating GenAI, the tool aims to provide dynamic and contextually relevant hashtags, ensuring that the content remains in tune with the latest trends.

Deliverables:

Keyword Extraction Module:

- Implementation of a robust algorithm for accurate keyword extraction from provided text.
- Prioritization of accuracy to ensure identified terms genuinely represent key concepts within the content.

GenAI-Powered Hashtag Generator:

- Integration of GenAI to analyze extracted keywords and generate trending, contextually relevant hashtags.
- Utilization of GenAI capabilities to adapt and align generated hashtags with current trends.

User-Friendly Interface:

- Design of an intuitive interface allowing content creators to input text easily.
- Options for customization, enabling users to adjust parameters and filter results based on their preferences.

Real-Time Feedback:

- Implementation of a real-time feedback mechanism to update extracted keywords and generated hashtags instantly as users input or modify content.

Scalability and Performance:

- Development of a tool capable of handling varying text lengths and complexities, ensuring scalability and optimal performance in different scenarios.

Evaluation Criteria:

Accuracy of Keyword Extraction:

- Evaluation of how accurately the tool identifies and extracts relevant keywords from diverse content.

Relevance of Hashtags:

- Assessment of GenAI's effectiveness in generating hashtags that align with current trends and are contextually relevant to extracted keywords.

User Interface Design:

- Evaluation of the interface's intuitiveness and user-friendliness for content creators.

Customization Options:

- Assessment of the adequacy and flexibility of customization options provided to users.

Real-Time Feedback Mechanism:

- Evaluation of the tool's responsiveness in providing instant updates as users input or modify content.

Scalability and Performance:

- Assessment of the tool's efficiency in handling varying text lengths and complexities for optimal performance in different content creation scenarios.

Overall Impact on Content Discoverability:

- Analysis of how effectively the tool enhances the discoverability and engagement of content on social media platforms.