Data Analysis & Visualization

APACHE BEAM

AMOD-5410H-A: Big Data

By Harshal Panchal & Kavya Gandhi

TRENT UNIVERSITY

LEVERAGING APACHE BEAM'S POTENTIAL: A PROFOUND ANALYSIS OF FOREST FIRE DATA

WHAT IS THE PROBLEM ?

Amid escalating global concerns over the ecological impact of forest fires, a pressing issue emerges. Forest fires pose a significant threat, necessitating urgent attention to understand and mitigate their effects. The focus is particularly on the Algerian region, grappling with distinct challenges and contributing to the urgency of finding effective solutions.

HOW CAN WE SOLVE IT ?

Our project takes a proactive stance, anchored in the formidable capabilities of the Apache Beam framework for data processing and analysis. By directing our exploration towards the "Algerian Forest Fires Dataset," a rich repository of vital information, we aim to uncover hidden patterns, identify trends, and highlight anomalies. Leveraging Apache Beam's prowess, our strategy is to shed light on the intricate factors influencing forest fires in Algeria. This comprehensive understanding serves as the foundation for devising more effective mitigation strategies, addressing the core problem at its roots.

used TECHNOLOGIES

Our project, implemented in Python, utilizes a powerful tech stack that seamlessly integrates key data science libraries. Python's readability and flexibility, coupled with libraries such as NumPy for numerical computing, Pandas for data manipulation, and Matplotlib/Seaborn for visualization, form the backbone of our data processing capabilities. Google Colab serves as our collaborative IDE, ensuring accessibility and ease of sharing. At the heart of our data processing engine is Apache Beam, a versatile framework for batch and stream processing. This unified approach, combined with Python and data science libraries, reflects our commitment to efficient data transformations and advanced analytics.

Apache Beam

Google Colab

Python

project

METHODOLOGY

Our project employs a robust tech stack, integrating key data science libraries for enhanced capabilities. Python's readability, coupled with NumPy for numerical computing, Pandas for data manipulation, and Matplotlib/Seaborn for visualization, forms the core of our data processing. Google Colab facilitates collaborative coding, ensuring accessibility. Apache Beam, at the heart of our processing engine, unifies batch and stream processing, reflecting our dedication to efficient data transformations and advanced analytics.

01

Integration and Collaboration:

Merged the project seamlessly into Google Colab for collaborative coding and documentation. Collaborated on implementation using Python and data science libraries.

02

Apache Beam Implementation:

Incorporated Apache Beam for unified batch and stream processing, enhancing data processing capabilities for large datasets.

03

Analysis & Statistical Calculation

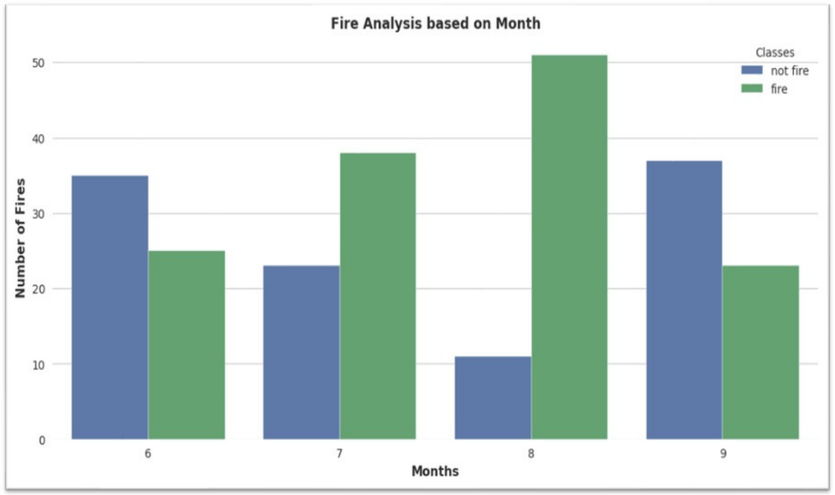
Leveraged Apache Beam for streamlined presentation of analysis results, showcasing key metrics such as average temperature, relative humidity, and total burned area and performed statisical analysis

04

Implementation of Machine Models

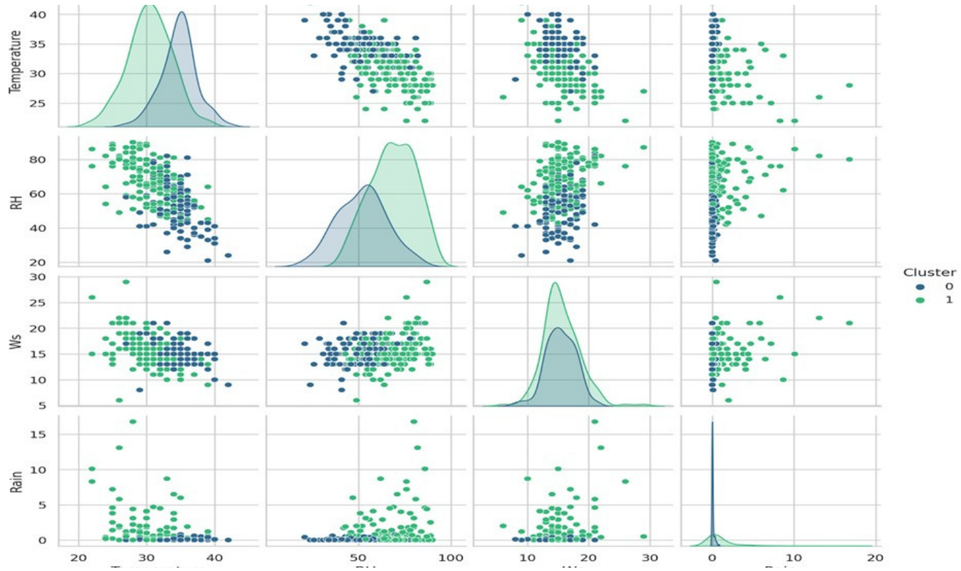
Embarked on the implementation of machine learning models to elevate data analysis. Leveraged Python's Scikit-Learn library to deploy predictive algorithms, enhancing the project's capability to derive valuable insights within the Algerian forest fire dataset.

RESULTS



Apache Beam's analysis highlights forest fire seasonality. The graph indicates August and September as peak months, emphasizing heightened fire occurrences during this period.

This visualization shows how the data points are grouped into clusters based on their similarities in weather conditions.



32.152 °C

Average Temperature

62.041%

Relative Humidity

18,915.70 HA*

Total Burned Area

*HA= HECTRE UNIT

Conclusion & RECOMMENDATIONS

- Apache Beam proved powerful for multidimensional datasets, with considerations for learning curves and scalability.
- The custom Apache Beam transform ensured data integrity, laying the groundwork for subsequent analyses.
- Month-wise analysis identified peak fire months i.e **August & September** aiding targeted preventive measures and resource allocation. how certain level of relative humidity and wind speed can affect the probability of forest fire.
- Preliminary exploration of machine learning integration, particularly K-Means clustering, showcased promising potential for identifying high-risk regions and enhancing predictive analysis.
- Python's data visualization tools created clear and intuitive representations of temperature, humidity, and fire occurrences, making complex data accessible to a broader audience.