**COLLABORATIVE PROJECT WITH INTEL**

**PROJECT TITLE :** Knowledge Representation and Insights Generation from Structured Datasets

**TEAM NAME :** TEAM\_ALPHABET

**TEAM MENTOR** : Dr T V RAJINI KANTH, Professor & Head,

Department of CSE-AI&ML

[**rajinikanthtv@sreenidhi.edu.in**](mailto:rajinikanthtv@sreenidhi.edu.in)

[**rajinitv@gmail.com**](mailto:rajinitv@gmail.com)**,** Ph. No: 9849414375

**TEAM MEMBERS:**

K. Sri Vani 21311A6627 CSM Team lead 21311A6627@sreenidhi.edu.in

Vishnu Vardhan21311A6607 CSM Team member 21311A6607@sreenidhi.edu.in

Deva Harsha 21311A6641 CSM Team member 21311A6641@sreenidhi.edu.in

Koushik Reddy 22311A66D1 CSM Team member 22311A66D1@sreenidhi.edu.in

Laxmi Chandana 22311A6609 CSM Team member 22311A6609@sreenidhi.edu.in

**INSTITUTE NAME:** Department of CSE-AI&ML

Sreenidhi Institute of Science and Technology

Yamnampet, Ghatkesar

Hyderabad – 501301

**DATE OF SUBMISSION:** 10 -07- 2024

**ABSTRACT**The abundance of structured data from diverse sources such as transactional databases, customer records, financial reports, and sensor data presents a significant challenge in deriving meaningful insights and actionable knowledge. Traditional data analysis methods often fall short in effectively representing the embedded knowledge, leading to underutilization of valuable information and missed opportunities for strategic decision-making. To address these challenges, advanced methodologies and tools are required to enhance data integration, pattern recognition, and overall analytical capability. This article compares the accuracy of a few major frameworks like: LangChain, pandas.ai and several other ai models. The study focuses on the application of these techniques in extracting actionable insights from structured datasets, aiming to improve decision-making processes and optimize data utilization. By exploring these advanced methodologies, the paper seeks to provide a comprehensive understanding of their potential to transform structured data into strategic assets.

**KEYWORDS:** Insight Generation, Patterns, Structured Datasets, Knowledge, Data Analysis

**INTRODUCTION**

In the current era, characterized by an explosion of data generation, organizations find themselves inundated with an unprecedented volume of structured data. This data originates from a multitude of sources, such as transactional databases, customer records, financial reports, and sensor data. Each source provides valuable information that can be harnessed to drive business decisions, improve operational efficiency, and gain competitive advantage. However, the true potential of this data remains largely untapped due to several challenges. The integration of data from heterogeneous sources, the complexity of data formats, and the identification of meaningful patterns and trends require sophisticated analytical tools and methods. Traditional data analysis techniques are often inadequate, resulting in the underutilization of data and missed opportunities. This paper explores these challenges and discusses advanced methodologies to enhance the extraction of actionable insights from structured datasets.

**MOTIVATION**

The motivation for studying the extraction of meaningful insights from structured data is to understand the existing data integration and analysis techniques, and to evaluate the performance of various advanced methodologies under practical constraints. Some of the challenges that structured data analysis faces include:

* Developing solutions for integrating heterogeneous data sources to ensure consistency and coherence in the combined dataset.
* Balancing the trade-off between the complexity of analytical models and their efficiency, as organizations require real-time analysis with high accuracy.
* Gaining hands-on experience with advanced data analysis algorithms and tools to enhance their practical implementation.
* Contributing to the field of data science by innovating new methodologies and tools that can overcome the limitations of traditional data analysis.
* Maximizing the strategic value of structured data by developing systems that can automatically extract and summarize knowledge, thereby reducing the reliance on human intervention and expertise.

The Intel Industrial Training initiative Unnati Program helps the students in getting the flavour of the Industrial View of the work planning, interaction, and guidance of the Intel Team and friendly competition with other college students.

**DATA SOURCES**

Studied and analysed the **Iris Dataset** for understanding the Knowledge Representation procedure from [UCI Repository](https://archive.ics.uci.edu/).

We have visualized the dataset accordingly with different algorithms and visualizations to create a model that represents the best for all the use cases and purposes of the user.

**PYTHON LIBRARIES USED IN THE PROGRAMS**

**Pandas**: Pandas is a fast, powerful, flexible, and easy-to-use open-source data analysis and manipulation library built on top of the Python programming language. It provides data structures like Data Frames and Series to efficiently handle large datasets and perform complex data analysis tasks.

**Streamlit**: Streamlit is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. Streamlit transforms data scripts into interactive web applications, making it simple to visualize data and model outputs.

**OS**: The OS module in Python provides functions for interacting with the operating system. It allows for the manipulation of files and directories, accessing environment variables, and performing other operating system tasks.

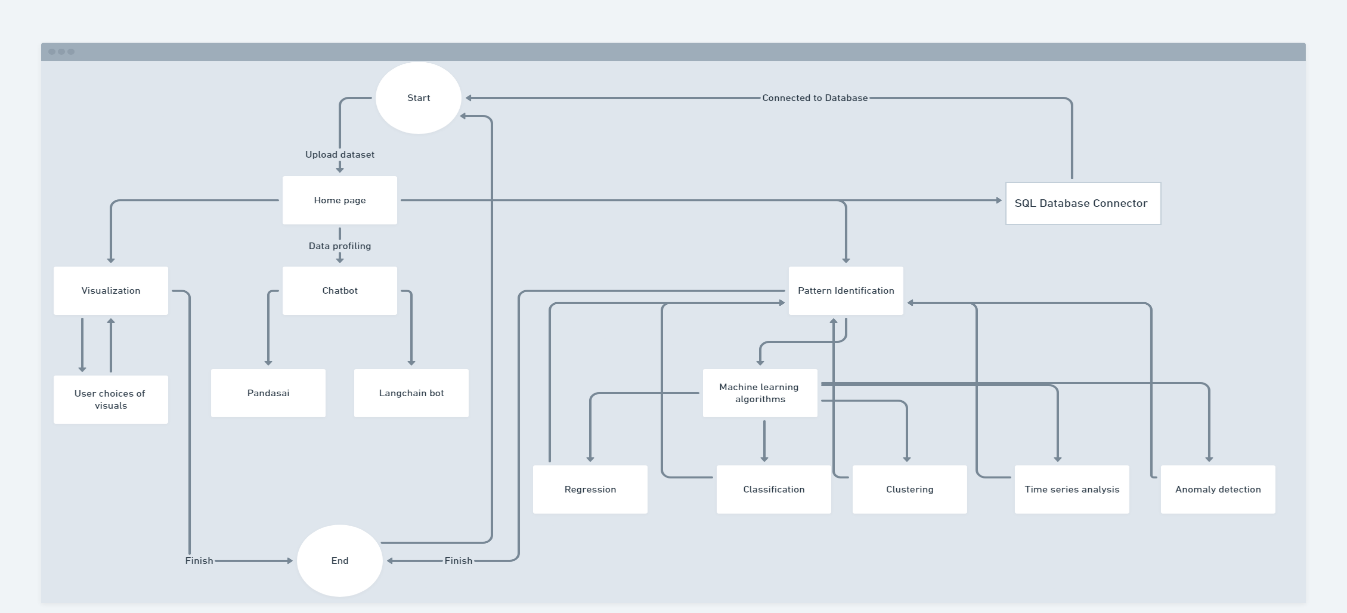
**Sklearn**: Sklearn (Scikit-learn) is a machine learning library for Python. It features various classification, regression, and clustering algorithms, including support vector machines, random forests, gradient boosting, k-means, and DBSCAN. It is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

**PandasAI**: PandasAI is an extension of the Pandas library, which leverages artificial intelligence to enhance data manipulation and analysis capabilities. It integrates machine learning algorithms directly into the Pandas workflow, providing intelligent data processing features.

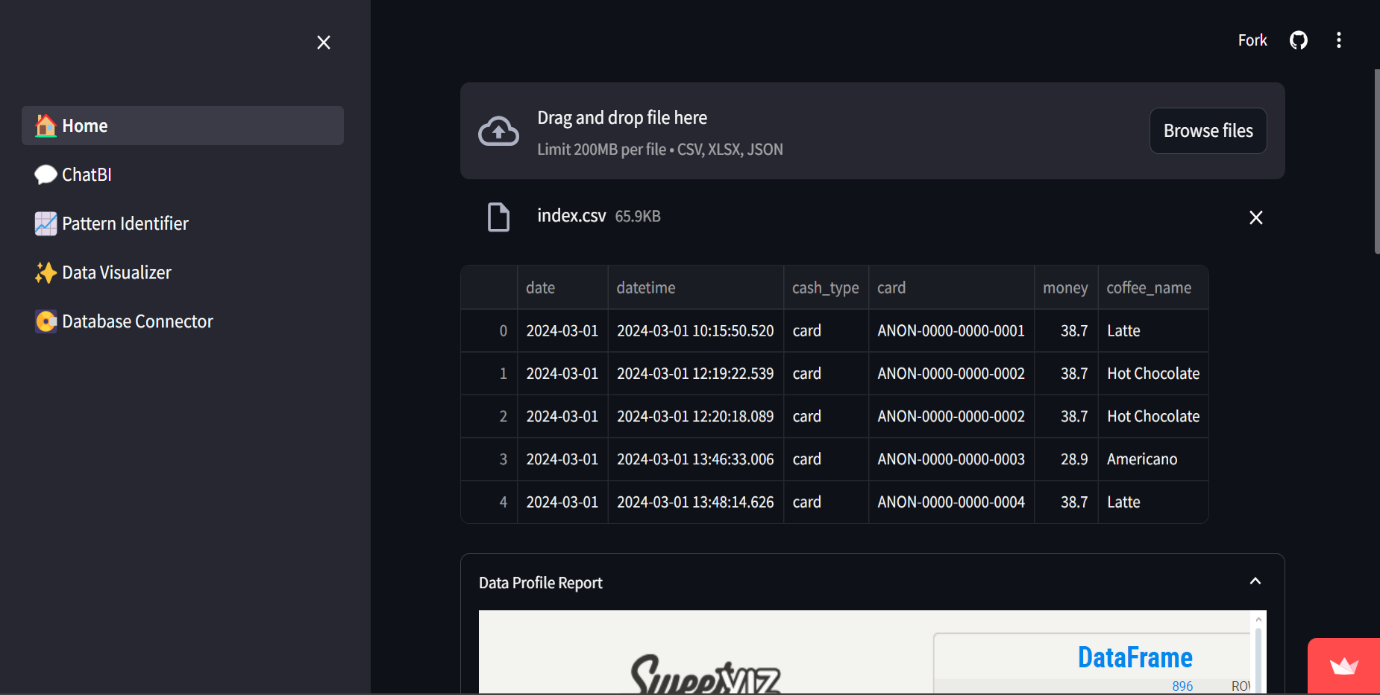
**LangChain**: LangChain is a library for building applications with language models. It provides a framework to create agents and tools that can interact with language models, enabling the development of sophisticated natural language processing applications.

**Gemini**: Gemini is a Python library designed for advanced data analysis and machine learning. It offers tools for data preprocessing, feature engineering, and model training, providing a comprehensive solution for developing and deploying machine learning models.

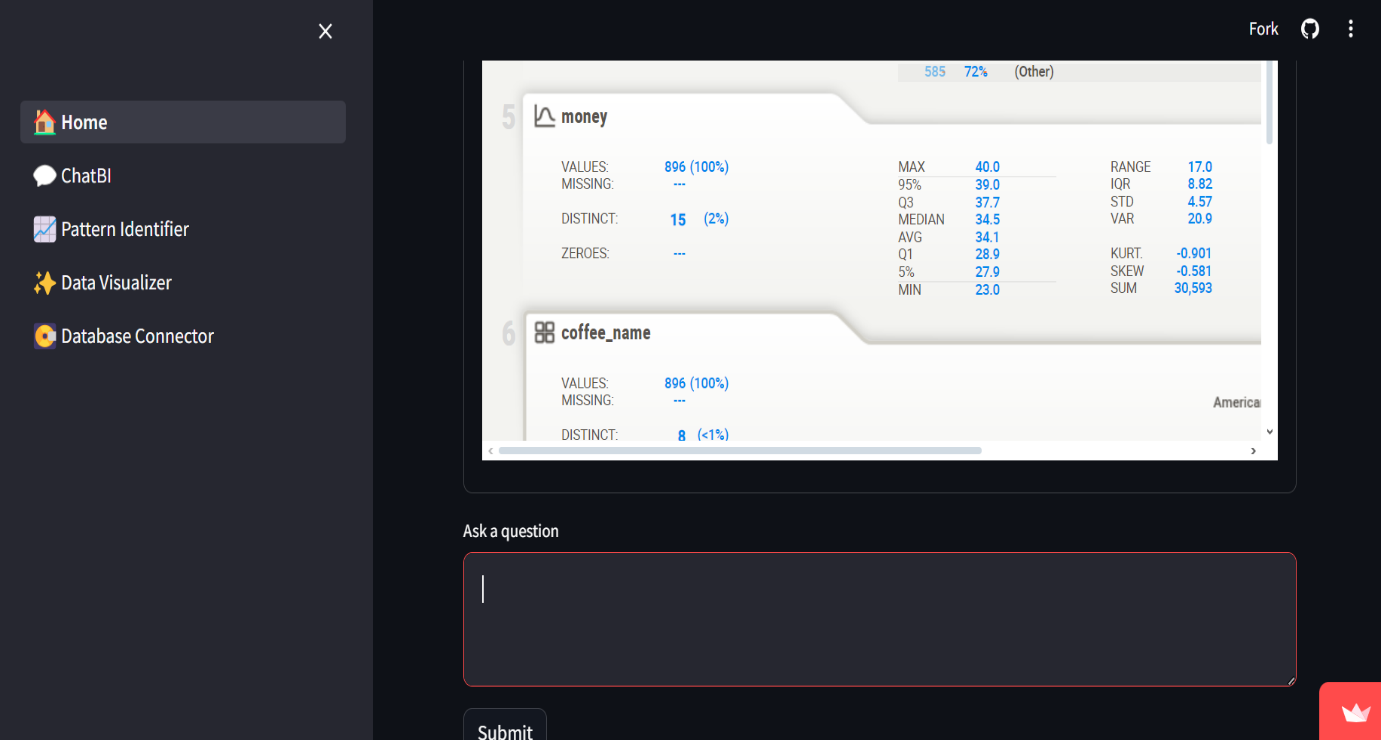
**ARCHITECTURE**

****

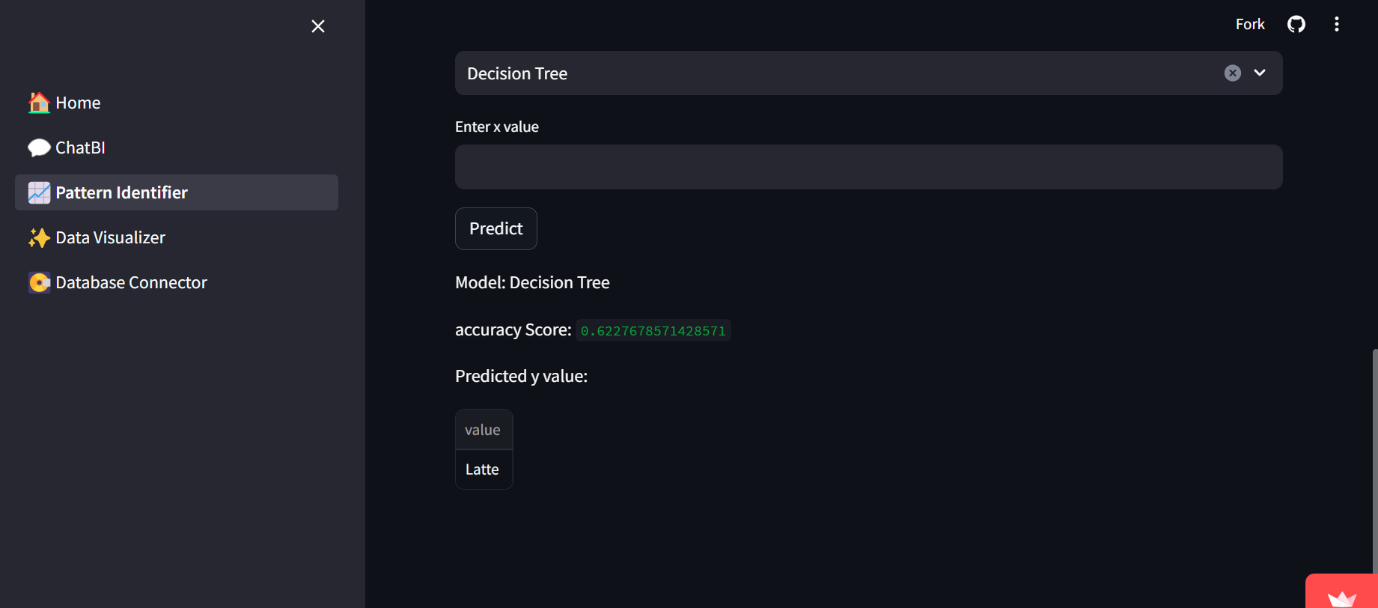
**RESULTS**

****

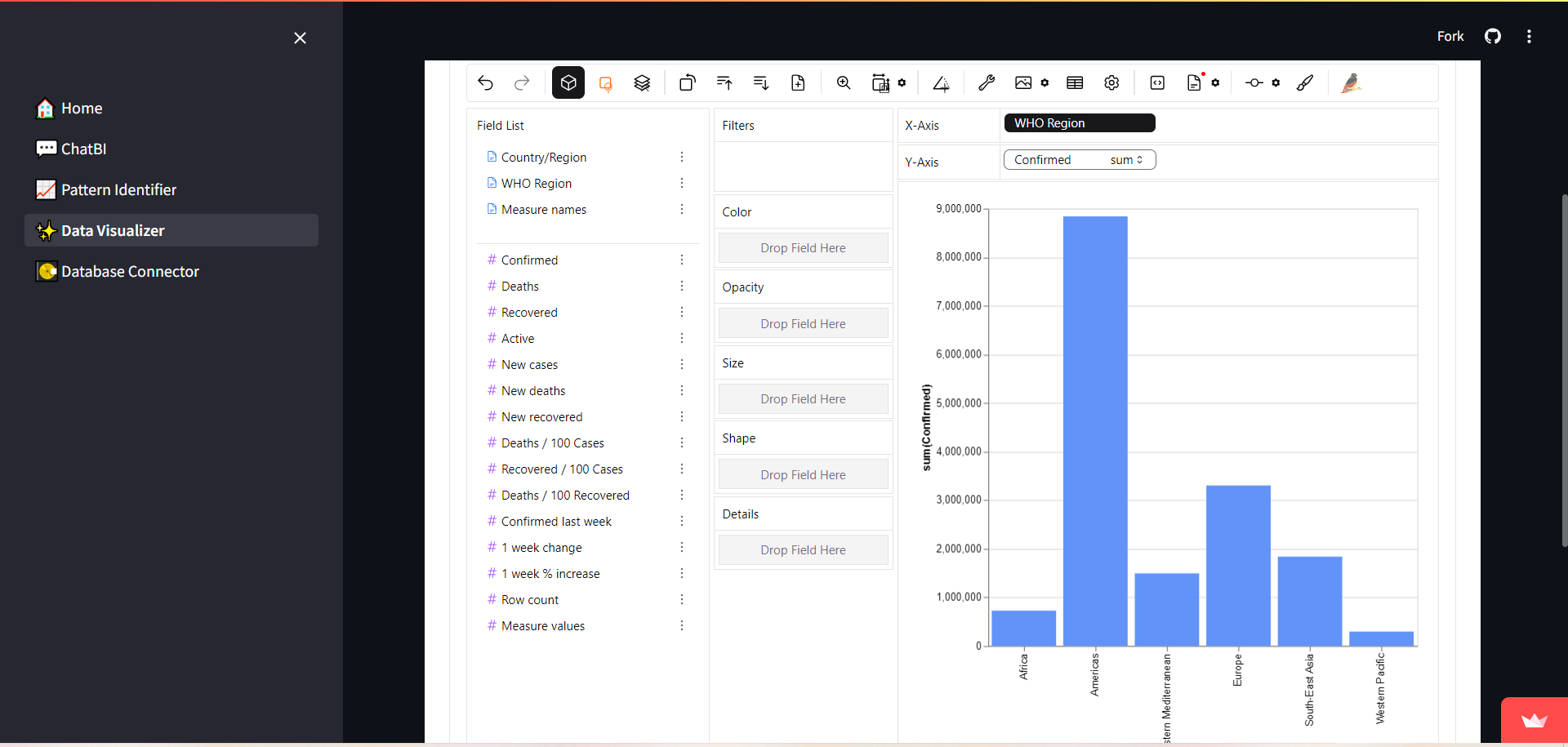
**Fig-1: Data Profile Report**

****

**Fig – 2: ChatBI**

****

**Fig – 3: Pattern Identifier**

****

**Fig – 4: Data Visualizer**

**CONCLUSION**

Analysed IDD Lite dataset, and developed a program to derive actionable insights from structured data and generate insights from data sets. Learned how to work on large-scale projects, and use pre-trained models to implement the solution. Successfully implemented LangChain and PandasAI to deliver insights and visualize the datasets.

**FUTURE SCOPE**

It can be developed further to seamlessly integrate data from various sources, identify subtle patterns and trends that traditional methods might overlook, enabling real-time analysis and decision-making and make it easier for stakeholders to understand and act upon the insights.

**REFERENCES**

 **"Heterogeneity in Data Integration"**

* Link: <https://ieeexplore.ieee.org/document/5472720>

 **"Data Mining and Knowledge Discovery: Challenges and Realities"**

* Link: <https://link.springer.com/article/10.1023/A:1009769707641>

 **"Advances in Data Analysis and Classification: Challenges, Approaches, and Applications"**

* Link: <https://link.springer.com/article/10.1007/s11634-017-0293-7>

 **"Data-Driven Decision Making: The Role of Analytics in Business"**

* Link: <https://www.sciencedirect.com/science/article/pii/S1877050918303090>

 **"Challenges in Real-Time Analytics: A Survey"**

* Link: <https://ieeexplore.ieee.org/document/7870073>

**SOURCE CODE AND CONFIGURATION FILES**

**Source code for the project is attached to the GitHub Link. (**[**GitHub Link**](https://github.com/vishan01/vishayamitra)**)**

**App Link (**[**Vishayamitra**](https://vishayamitra.streamlit.app)**)**

**(Please make sure that the repo is cloned before implementation.)**