**COLLABORATIVE PROJECT WITH INTEL**

**PROJECT TITLE : KNOWLEDGE REPRESENTATION AND INSIGHT GENERATIONON STRUCTURED DATASET**

**TEAM NAME :** **CODECATCHER**

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

**Department of CSE-AI&ML**

[**rajinitv@gmail.com**](mailto:rajinitv@gmail.com)**,** [**rajinikanthtv@sreenidhi.edu.in**](mailto:rajinikanthtv@sreenidhi.edu.in)

**Ph. No: 9849414375**

**INTEL MENTOR : ARCHANA VAIDHEESWARAN**

**TEAM MEMBERS :**

**PREETHI** Roll no: 22311A6623 CSE-AIML **Team lead**

**(22311A6623@aiml.sreenidhi.edu.in)**

**VIRAJITHA** Roll no: 22311A6619 CSE-AIML Team member

**(22311A6617@aiml.sreenidhi.edu.in)**

**MANIKESH** Roll no: 22311A6691 CSE-AIML Team member

**(22311A6691@aiml.sreenidhi.edu.in)**

**VISHAL** Roll no: 22311A6690 CSE-AIML Team member

**(22311A6697@aiml.sreenidhi.edu.in)**

**HARINI** Roll no: 22311A6667 CSE-AIML Team member

**(22311A6692@aiml.sreenidhi.edu.in)**

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

**Institute Name** **:** **Sreenidhi Institute of Science and Technology**

**Yamnampet, Ghatkesar**

**Hyderabad – 501301**

**Date of Submission: 10-07-2024**

**INTRODUCTION**

* 1. **ABSTRACT**

In today’s data-driven world, organizations are inundated with vast amount of structured data. This data is rich in information but often remains underutilized due to the complexity of extracting meaningful insights. The challenge lies in effective representation of this Knowledge and generation on actionable insights that can aid in strategic decision making. Traditional methods of data analysis often fall short in handling the dynamic and multifaceted (many aspects) nature of datasets. Hence, there is a pressing need for an advanced AI-based solution that can seamlessly analyze, process and interpret structured data to provide valuable insights.

**KEY WORDS:** structured data, Data processing, Data analysis, Meaningful Insights , Data visualization, Knowledge Representation.

* 1. **OBJECTIVE**

Analyzing and processing structured data, represent knowledge effectively, generate meaningful insights to aid decision-making.

The primary objective of this project is to develop an AI-based solution capable of analyzing and processing structured datasets to generate meaningful insights. This solution aims to bridge the gap between raw data and actionable knowledge by leveraging the advanced algorithms and techniques. We have chosen a student dataset from Kaggle which is a csv file. We found the student performance by effectively representing knowledge via data visualization and graph-based methods, and employing sophisticated pattern identification techniques, the project seeks to enhance decision-making processes. Finally, the goal is to create a robust system that empowers users to gain deeper understanding and insightful information from their data, thereby driving more informed and strategic decisions.

**DATASET DESCRIPTION**

* 1. **DATASET SOURCE**

We have taken the dataset from kaggle. Size of the dataset is 395x33. Number of rows is 395 rows and number of columns is 33. Each column in the dataset is a feature. The file is in the csv format.

Data Set Link: https://www.kaggle.com/datasets/devansodariya/student-performance-data

Names of the columns (Features):

* School
* Sex
* Age
* Address
* Famsize
* Pstatus
* Medu
* Fedu
* Mjob
* Fjob
* Reason
* Guardian
* Traveltime
* Studying
* Failures
* Schoolsup
* Famsup
* Paid
* Activities
* Nursery
* Higher
* Internet

And so on.

* 1. **KEY FEATURES AND**

**ATTRIBUTES**

* **Age**
* **Sex**
* **Studytime**
* **Failures**
* **G1, G2, G3 (grades)**
* **Absences**

These key attributes provide insights into academic performance, factors affecting it, and potential areas for intervention.

**2.2.1 PREPROCESSING USED**

To prepare the dataset for insight generation, the following preprocessing steps are used-

* Handling Missing Values- ‘df.isnull()’ checks for any missing values in the dataset and handles them appropriately, either by filling them with suitable values or by removing the rows/columns with missing data.
* Duplicate removal- Checks for duplicate values and removes any duplicate rows to ensure data quality.

By performing these preprocessing steps, the dataset will be in better shape for generating meaningful insights and facilitating effective knowledge representation.

**METHODOLOGY**

* 1. **METHODS USED**

***3.1.1. Knowledge Representation***

1. ***Data Visualization***

Used visualization techniques to represent data in graphical formats, making complex datasets more comprehensible and accessible. Techniques such as bar charts, scatter plots, heat maps, histograms and more advanced visualizations like interactive dashboards are used to present the data insights effectively.

***3.1.2. Pattern Identification***

1. ***Clustering***

Algorithms like K-means are used to group similar data points together. This technique helps in identifying inherent structures and segmentations within the dataset, which can be crucial for targeted analysis and insight generation.

The cluster labels are calculated and displayed, which can help identify patterns in the data, such as the assignment of each data point to a cluster.

These pattern identification techniques can help identify various patterns and relationships in the data, such as:

Correlations between variables

Clusters of similar data points

Outliers or anomalous data points

Trends or patterns in the data

Relationships between variables

Central tendency and variability of each column

Skewness, outliers, and multimodality of each column

* 1. **TOOLS USED**

***3.2.1. Python-*** Python is the primary programming language used for data analysis and ML due to its extensive libraries and frameworks. Libraries used are-

* 1. **Pandas**

Pandas is a powerful python library for data manipulation and analysis and it offers data structures like Data Frames and Series for efficient handling of structured data.

* 1. **NumPy**

NumPy is a fundamental package for scientific computing in Python, providing support for arrays, matrices, and a wide range of mathematical functions to operate on these data structures.

* 1. **Scikit-Learn**

Scikit-learn is a widely-used machine learning library in Python, offering simple and efficient tools for data mining, data analysis, and machine learning, including classification, regression, clustering, and more.

* + 1. ***ydata\_profiling***

The code uses the ydata\_profiling library, a library that generates profile reports for pandas DataFrames. ydata\_profiling is used to generate a comprehensive report on the uploaded CSV file.

* + 1. ***Matplotlib***

The code uses the Matplotlib library, a popular data visualization library. Matplotlib is used to create scatter plots and other visualizations to represent the clustering results***.***

* + 1. ***Streamlit***

Streamlit app is used for the web interfacing of our model. Streamlit is a library which serves best for creating interactive web applications. Functions or say methods, used in our solution are set-page\_config, title, markdown, expander, code, sidebar, write, header, DataFrame, altair\_chart, warning.

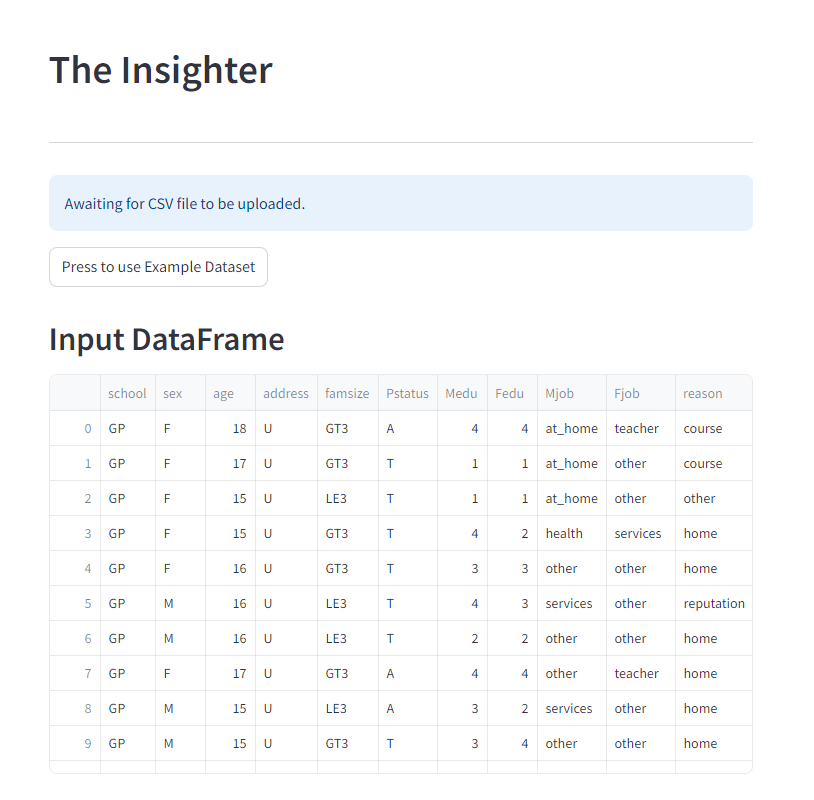
By using the above methods and tools, the project aims to create a comprehensive system that can effectively represent knowledge and generate insightful facilitate better understanding and utilization of data, ultimately aiding in more informed and strategic decision-making.

**RESULT AND DISCUSSION**

**Summary of Data Analysis:**

**Data Upload and Profiling**

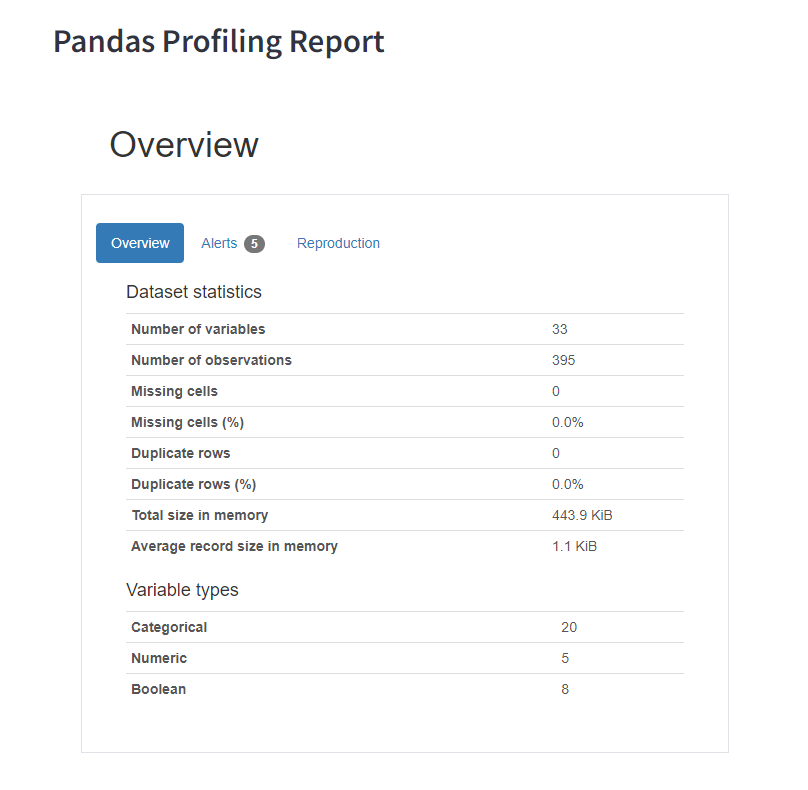
The user uploads a CSV file, which is then loaded into a Pandas DataFrame.



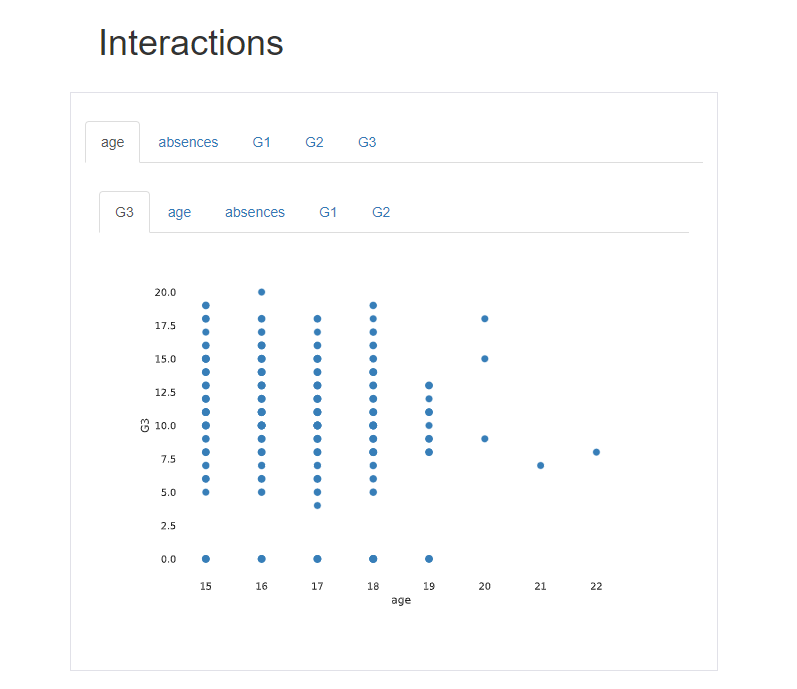
The Pandas Profiling Report is generated, which provides an overview of the dataset,

including:

Summary statistics (mean, median, mode, etc.) for each column.



Histograms and density plots for each column.



Correlation matrices and heatmaps.

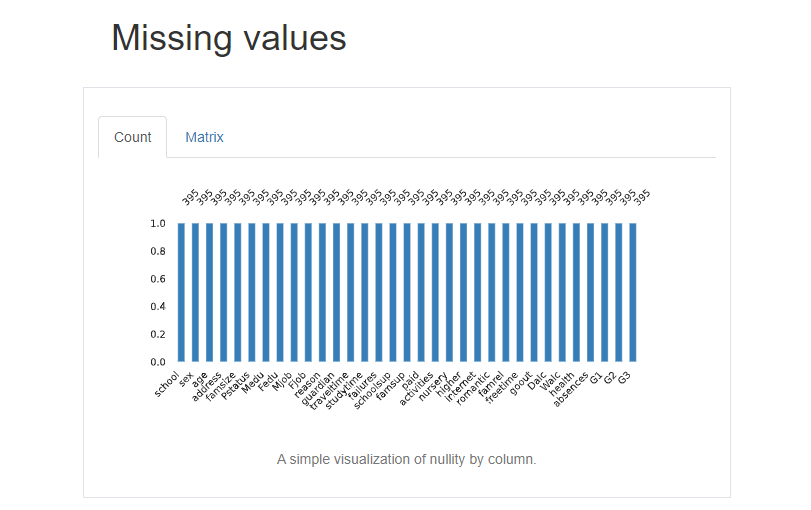
Missing value analysis.

**Handling Missing Values**

The code checks for missing values in the dataset and prompts the user to choose an action:

Delete rows with missing values.

Fill missing values with zeroes.



**K-Means Clustering**

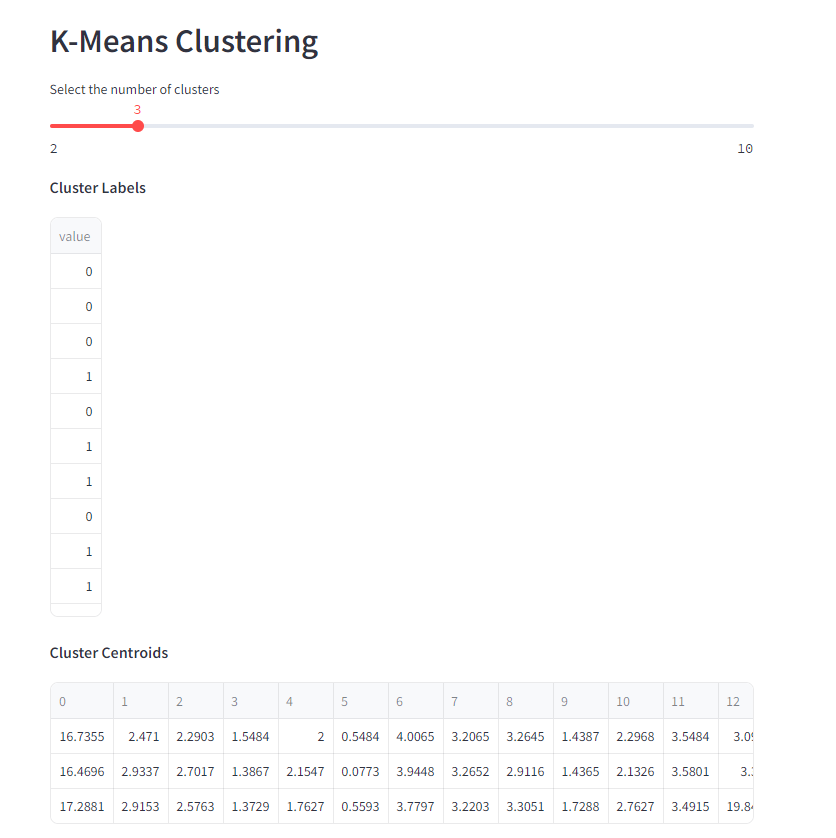
The user selects the number of clusters (k) using a slider.

The K-Means clustering algorithm is applied to the numerical columns of the dataset.

The clustering results are displayed, including:

Cluster labels for each data point.

Cluster centroids (mean values for each feature in each cluster).



A scatter plot is generated to visualize the clustering results, with each data point colored according to its cluster label.

Overall, this code provides a basic data analysis pipeline that includes data profiling, handling missing values, and K-Means clustering. The user can interact with the code through the Streamlit interface, selecting the number of clusters and choosing how to handle missing values**.**

**Key Insights**:

* The dataset is clean and complete.
* The features show moderate correlation, which is useful for clustering.
* The K-Means algorithm identified three distinct clusters.
* The clusters suggest specific groupings or categories.
* The visualizations and profile report provide a thorough understanding of the dataset and clustering results.
* **Conclusion**: The analysis revealed several interesting insights and patterns in the dataset. The clustering results provide a useful framework for understanding the data structure and can guide further investigation and decision-making.

APP LINK: [**The Insighter**](file:///C:\Users\tumma\OneDrive\Desktop\The%20Insighter.lnk)

GITHUB LINK: <https://github.com/dpreethi12/The-Insighter/tree/main>

**CONCLUSION**

**5.1 FINDINGS:**

The code serves as a powerful tool for:

Exploratory Data Analysis: Users can upload their data, generate a detailed summary report, and gain insights into the dataset’s structure and relationships.

Clustering Analysis: Perform K-Means clustering to uncover patterns, groups, and outliers within the data.

Visualization: The scatter plot provides an intuitive way to understand the clustering results and the distribution of data points.

**5.2 FUTURE WORK:**

Interactive Chatbot for User Queries:

To enhance the user experience, an interactive chatbot can be integrated into the app. This chatbot will be available to clarify user queries, provide guidance on data analysis, explain profiling reports, and assist with interpreting clustering results. By offering real-time support, the chatbot ensures users can navigate the tool effortlessly and maximize its potential.

History Saving:

To further improve the user experience, a history-saving feature will be added. This feature will allow users to save and revisit their previous analyses, including uploaded datasets, generated profiling reports, clustering parameters, and results. By maintaining a history of their work, users can track their progress, compare different analyses, and continue their exploration seamlessly without losing any previous insights.

REFERENCES

[1] Research on data cleaning in big data environment

Fengwei Zou

2022 International Conference on Cloud Computing, Big Data and Internet of Things (3CBIT)

[2] The Significance of using Data Extraction Methods for an Effective Big Data Mining Process

Manish Sharma;

Richa Gupta

2023 2nd International Conference for Innovation in Technology (INOCON)

[3] A Data Discovery and Visualization Tool for Visual Analytics of Time Series in Digital Agriculture

Jasmin K. Dhaliwal;

Megan E. Galbraith;

Carson K. Leung;

Da Tan

2023 27th International Conference Information Visualisation (IV)

[4] Getting Started with Streamlit for Data Science: Create and deploy Streamlit web applications from scratch in Python

Tyler Richards

[5] Foundation Data Space Models: Bridging the Artificial Intelligence and Data Ecosystems (Vision Paper)

Edward Curry;

Tarek Zaarour;

Yang Yang;

Mohan Timilsina;

2023 IEEE International Conference on Big Data (BigData)

[6] [NumPy](https://numpy.org/)