PROJECT DATA INTELLIGENCE

Harshavardhan Chittaluru

Problem Statement:

Many individuals face difficulties in performing data analysis because they lack the expertise to write complex code or find the necessary Python modules to work with their datasets. This manual process often becomes time-consuming and error-prone, causing people to spend more time on technical details than on extracting meaningful insights. The goal of this project is to solve this problem by creating a tool that automates the entire data analysis process, allowing users to focus on outcomes rather than coding. The tool will simplify data analysis, making it more accessible and efficient for users of all skill levels.

Abstract:

This project presents a cloud-based web application designed to automate the data analysis process, addressing the common challenges users face in manually writing code and searching for modules in Python. By automating tasks such as data ingestion, cleaning, visualization, preprocessing, and machine learning model selection, the application allows users to focus on interpreting results rather than managing technical complexities. With features like auto-generated IPython notebooks, Word documents, and a dynamic trial-and-error system, the application offers an intuitive interface that accommodates various stages of the data analysis pipeline. This includes flexible data instance management, enabling users to experiment without affecting the original data. By integrating advanced tools, including OpenAI for visualization insights and model evaluation metrics, this tool significantly enhances efficiency and accuracy in data processing. The proposed solution empowers users—especially those with limited coding expertise—to streamline their workflows, reduce errors, and concentrate on deriving actionable insights from their datasets.

Project Overview:

This project proposes the design and development of a web application that automates data analysis using a cloud environment and pipeline architecture. The application enables users to store, retrieve, and analyze data in a seamless and efficient manner, streamlining the entire data analysis process from data ingestion to model evaluation.

The key feature of the application is its ability to automatically generate Word documents and IPython notebooks in the backend as soon as data ingestion is complete, allowing users to focus on data analysis rather than manual data preparation. Additionally, the application provides a unique "trial-and-error" feature, which enables users to experiment with different data cleaning, visualization, and preprocessing techniques without affecting the original data.

The application consists of seven interconnected pages, each catering to a specific stage of the data analysis process. The pages are designed to operate independently, allowing users to switch between them at any time and work on different aspects of their project simultaneously. The pages include:

- Data Ingestion: Users can connect to a data source or ingest a dataset with a unique name into the cloud and adjust data types as needed.
- Data Cleaning: Users can clear irrelevant data, manage null or NaN values, remove duplicates, handle erroneous data, and standardize data.
- Data Visualization: Users can select columns to visualize data and receive recommended graph types. They can also select a graph type and match it to the required data type columns. The application integrates with Open AI to analyze the graph and provide valuable insights.
- Data Preprocessing: Based on visualizations, users can select preprocessing techniques for categorical and continuous variables, such as mode, one-hot encoding, mean, median, and mode.
- Train-Test Split: Users can specify the percentage or number of records for the train-test split and choose between straight or random row selection.
- Machine Learning Model Selection: Users can select a machine learning model for their data prediction and specify the target variable and columns for training. The application provides default hyperparameter values, which users can adjust as needed.
- Model Evaluation and Comparison: The application displays metrics such as accuracy, F1 score, precision, and recall, as well as metric graphs like PR graphs and area under the curve. Users can compare results from different executions and view metrics and data immediately.s

A key feature of the application is its data instance management system. After each stage, users can create a new virtual file in the cloud, and the application will ask them to either override the original instance of the data or create a new instance. This allows users to experiment with different data versions and track changes throughout the analysis process. Additionally, users can change their data instances at any stage and check the data to ensure that it is correct. If a user changes the data instance in one stage and moves to the next stage, the application will ask them if they want to use the current working instance or the previous instance. This data instance implementation is designed to support future features and provide a flexible and efficient data analysis workflow.

To ensure transparency and accountability, the application also includes a change log that tracks all changes made to the data instances, including the user who made the changes, the date and time of the changes, and a description of the changes. The change log also reports any errors that occur during the analysis process, providing a clear audit trail of all activities. This feature enables users to track the history of their data analysis and ensure that their results are reproducible and reliable.

The application's pipeline architecture enables users to revisit and refine their work at any stage, ensuring that they can achieve optimal results. The proposed web application has the potential to revolutionize the data analysis process by automating tedious tasks, reducing errors, and increasing efficiency. Its user-friendly interface and flexible architecture make it an ideal tool for data analysts, scientists, and machine learning practitioners.