INTELLICLEANSE

Capstone Project Phase - I (First Presentation)

Vaishnavi R P

SRN : PES1PG23CA159

Dr. Lekha A

Associate Professor

Department of CA

Abstract

- **Overview:**
  - Large, varied, and noisy datasets are difficult for businesses in the financial, medical, and e-commerce sectors to manage. These datasets frequently contain errors and inconsistencies that impede analysis and decision-making.
- **Objective:**
  - This project aims to develop an automated data cleaning and preprocessing system that leverages artificial intelligence and data transformation pipelines to improve the quality of data, ensuring it is reliable for decision-making and analysis.

Abstract

- **Methodology:**
  - The system uses AI techniques like machine learning and statistical methods to automatically clean and preprocess data, handling missing values, outliers, and inconsistencies.
- **Expected Outcomes:**
  - A scalable, automated solution is expected to streamline data cleaning, reduce manual effort, minimize human error, and improve data usability, particularly in large dataset-dependent domains.

Intellicleanse

Problem Scenario

- **Context:**
  - Organizations increasingly rely on large datasets for analytics and decision-making, spanning domains like finance, healthcare, and e-commerce.
  - Example: ”A financial institution analyzing transaction data to detect fraud or a healthcare provider evaluating patient data for medical diagnosis.”
- **Specific Problem:**
  - Raw data often contains inconsistencies, errors, and missing values, compromising its quality and hindering effective analysis.
- **Implications:**
  - These issues lead to flawed analysis, unreliable insights, and wasted resources, impacting organizational decision-making and performance.

Intellicleanse

Problem Domain

- **Definition:**
  - The project operates within the domain of Data Engineering, focusing on the design, building, and maintenance of systems that manage and process large volumes of data for business decision-making.
- **Scope:**
  - This project specifically addresses the challenges of handling diverse datasets through tools for data preprocessing, transformation, and integration of dual databases—MySQL for structured data and MongoDB for semi-structured data.
- **Relevance:**
  - Data Engineering is crucial for businesses and organizations to derive meaningful insights from large datasets. This project ensures that raw data is processed into a structured and accessible format, enabling data-driven decision-making.

Motivation

- **Real-World Problem:**
  - Unclean data is a widespread issue that hampers the accuracy of insights and decision-making across various industries.
  - The inability to efficiently clean and process large datasets can lead to significant time delays, wasted resources, and flawed business strategies.
- **Academic and Research Relevance:**
  - This project enhances data science by automating data cleaning with machine learning, improving data quality for real-world applications.
  - It introduces innovative techniques for managing large, diverse datasets, advancing both academic research and data-driven decision-making.

Motivation

- **Personal Interest:**
  - Passion for improving data quality and its impact on decision-making in real-world applications.
  - Curiosity about leveraging machine learning and automation to solve common data-related challenges.

- **Potential Impact:**
  - Enhanced data quality will lead to more reliable analyses, supporting better decision-making and improving efficiency in various sectors like finance, healthcare, and e-commerce.
  - Automating data cleaning will save time, reduce errors, and enable businesses to focus on insights rather than data preparation.

Intellicleanse
Functionalities

Functionality	Description
Data Upload and Preview	Users can upload structured and semi- structured files and view a summary of the dataset, including column names, data types, and sample rows.
Data Profiling and Summary Report	Automatically generate a report highlighting missing values, duplicate entries, data types, and key statistics for quick insights.
Interactive Visualizations	Visualize data distributions and outliers through histograms, scatter plots, and box plots.

Intellicleanse
Functionalities

Functionality	Description
Missing Value Imputation	Provide options to handle missing values, such as filling with averages, medians, or custom values, or flagging them for further review.
Outlier Detection and Management	Identify statistical outliers and offer actions like removal, capping, or flagging for anomalies.
Data Standardization	Normalize and format data by standardizing dates, case sensitivity, and numerical formatting.

Intellicleanse
Functionalities

Functionality	Description
Customizable Data Cleaning Rules	Allow users to define and apply custom rules for data validation and cleaning.
Duplicate Removal	Detect and remove duplicate rows or columns to reduce redundancy.
Data Transformation Pipelines	Create reusable workflows to automate common cleaning tasks.

Intellicleanse
Functionalities

Functionality	Description
Collaborative Cleaning	Multiple users can work collaboratively, review, and edit datasets.
Version Control and Undo	Save data versions at different stages and allow users to revert changes if needed.
Export Options	Download cleaned data in various formats (e.g., .csv, .xlsx) as per user choice, along with a detailed cleaning log.

Note: These functionalities are subject to refinement as the project progresses.

Intellicleanse

Tools & Technologies

Category	Tools/Technologies Used
Programming Languages	Python, JavaScript(ES6+)
Frameworks and Libraries	TensorFlow/Scikit-learn, NumPy, Pandas, React.js, Node.js, Express.js, D3.js, Chart.js/Plotly
Software Tools	Visual Studio Code, Git, Docker, MySQL, MongoDB
Hardware/Infrastructure	AWS, Servers, NVIDIA

Note: All tools are chosen based on project requirements and scalability.

Intellicleanse
Application in the Real World

- **Problem Addressed:**
  - This project addresses the challenge of inefficient, error-prone data cleaning processes by automating the identification and correction of data issues in large datasets, ensuring higher data quality for better decision-making.
- **Target Audience:**
  - The target audience includes businesses, researchers, data scientists, and organizations handling large datasets that need accurate and reliable data for analysis.

Application in the Real World

- **Use Cases:**
  - This project can be applied in e-commerce, healthcare, finance, and marketing sectors to automate data cleaning tasks, ensuring accurate analysis, improved decision-making, and better predictive modelling.
- **Potential Impact:**
  - By automating data cleaning, this project can significantly reduce manual effort, improve data quality, and lead to more accurate insights, ultimately optimizing decision-making and operational efficiency.

Intellicleanse
Project Work Status

- **Short-Term Goals (1-2 weeks):**
  - Complete detailed literature review.
  - Finalize tools, libraries, or frameworks for implementation.
- **Mid-Term Goals (3-4 weeks):**
  - Develop a prototype or preliminary solution.
  - Test basic functionalities.

Intellicleanse
Project Work Status

- **Initial Work Done :**

Task 1 : Conducted background research on data cleaning techniques and identified common issues such as missing values, outliers, and data duplication.

Task 2 : Defined the project scope and determined the necessary functionalities for the web application, such as user authentication, data preprocessing, and outlier detection.

Task 3 : Developed a conceptual framework for the system architecture, including the integration of MySQL and MongoDB for data storage and management.

Thank You

Vaishnavi R P

SRN : PES1PG23CA159

Department of Computer Applications