INTELLICLEANSE

Capstone Project Phase - I

Vaishnavi R P

SRN : PES1PG23CA159

Dr. Lekha A

Associate Professor

Department of CA

Abstract

Data cleaning ensures data accuracy and readiness for analysis by addressing errors, inconsistencies, and inaccuracies. This online tool simplifies cleaning files like ‘.csv‘ with features like duplicate removal, missing value imputation, outlier detection, and data profiling. Its user-friendly interface helps scientists, analysts, and businesses transform raw data into actionable insights.

Introduction

- **Problem Scenario:** Handling inconsistent, redundant, and incomplete datasets hampers effective data analysis and decision-making.
- **Proposed Solution:** Develop a robust web application, Intellicleanse, that streamlines data cleansing using automated tools and machine learning algorithms.
- **Purpose:** To enhance data quality by addressing issues such as missing values, outliers, and duplicates, enabling users to make accurate data-driven decisions.
- **Scope:** The application caters to data analysts, researchers, and enterprises by providing a comprehensive suite of data preprocessing tools, ensuring flexibility and scalability.

Related Work

- [1] L. Ahuja, B. Singh, and R. Simonv, "Data Cleaning: Paving a Way for Accurate and Clean Data""
  - **Year:** 2024
  - **Summary:** The paper ”Data Cleaning: Paving a Way for Accurate and Clean Data” discusses techniques to address issues like missing values and duplicates, highlighting automated tools for efficient cleaning. It emphasizes the importance of clean data for better decision-making and predictive accuracy.

Related Work

- [2] Peng Li , Xi Rao , Jennifer Blase , Yue Zhang , Xu Chu , Ce Zhang, "CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks"
  - **Year:** 2021
  - **Summary:** CleanML systematically investigates the impact of data cleaning on ML classification tasks. It evaluates how cleaning different types of errors in real-world datasets affects ML performance using diverse algorithms and cleaning methods.

Related Work

- [3] Nilu Singh, "Data Cleaning Methods"
  - **Year:** 2023
  - **Summary:** The paper highlights key techniques for improving data quality, including handling missing values, duplicates, and outliers. It emphasizes using tools like isnull(), fillna(), and dropna() for processing and underscores the importance of cleaning data to ensure accurate analysis.

Related Work

- [4] P. V. S. V. Narayana and V. M. G. S. Pandit, "Data Cleaning Approaches in Data Mining "
  - **Year:** 2018
  - **Summary:** This study reviews key data cleaning techniques in data mining, focusing on handling noise, missing values, and inconsistencies. It provides insights into algorithms that ensure data quality for improved mining outcomes and discusses real-world applications highlighting the importance of robust cleaning methods.

Related Work

- [5] Huang Shan, E. Gubin, "Data Cleaning for Data Analysis"
  - **Year:** 2019
  - **Summary:** This paper emphasizes the importance of data cleaning as a crucial preprocessing step in data analysis. It highlights common data issues such as missing values, outliers, and inconsistencies, and proposes methods to address these issues for improved data mining results.

Related Work

- [6] Virender Kumar, Cherry Khosla, "Data Cleaning : A Thorough Analysis and Survey on Unstructured Data"
  - **Year:** 2018
  - **Summary:** This paper addresses the challenges of cleaning unstructured data, a less-explored domain compared to structured datasets. It surveys methodologies and tools tailored to handle diverse formats like text, multimedia, and semi-structured data.

Related Work

- [7] Henning Koehler, Sebastian Link, "Possibilistic Data Cleaning"
  - **Year:** 2022
  - **Summary:** This paper explores a possibilistic approach to data cleaning, focusing on handling uncertainty and ambiguity in data. It introduces techniques based on fuzzy logic to clean datasets with incomplete or vague information.

Related Work

- [8] Xi Wang, Chen Wang, "Time Series Data Cleaning: A Survey"
  - **Year:** 2019
  - **Summary:** This paper surveys techniques for cleaning time-series data, emphasizing methods to address unique challenges such as temporal dependencies, irregular intervals, and seasonality.

Related Work

- [9]Otmane Azeroual , Gunter Saake, Mohammad Abuosba, "Data Quality Measures and Data Cleansing for Research Information Systems"
  - **Year:** 2018
  - **Summary:** The paper addresses the importance of data quality in research information systems (RIS) and explores strategies for data cleansing to ensure accurate and reliable data. It emphasizes error detection and corrective actions to improve decision-making and system trustworthiness. These measures aim to enhance the quality of research data across systems.

Related Work

- [10] Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon, "A Review on Data Cleansing Methods for Big Data"
  - **Year:** 2019
  - **Summary:** This paper reviews the data cleansing process and its challenges, specifically in the context of big data. It explores methods for identifying, detecting, and correcting data errors to enhance data quality, which is critical for accurate decision-making in organizations. The authors also discuss the limitations of traditional approaches and propose advanced techniques designed for the scale and complexity of big data.

Hardware and Software Requirements

- **Hardware Requirements:**
  - **Processor:** Intel Core i5 (9th Gen) / AMD Ryzen 5 or higher.
  - **RAM:** 8 GB (16 GB recommended).
  - **Storage:** 256 GB SSD (minimum).
  - **OS:** Windows 10/11, macOS Catalina+, or Linux (Ubuntu 20.04+).
  - **Network:** High-speed internet.
- **Software Requirements:**
  - **Frontend:** React.js, JavaScript, D3.js, Chart.js/Plotly.
  - **Backend:** Python, Node.js, Express.js, TensorFlow/Scikit-learn, NumPy, Pandas.
  - **Database:** MySQL, MongoDB.
  - **Server:** AWS.

Intellicleanse

Functional Requirements

- **Data Upload, Preview, and Profiling:**
  Users can upload `.csv` or `.xlsx` files and view an instant preview with a summary report showing column names, data types, missing values, duplicates, and key statistics.
- **Redundancy and Consistency Cleaning:**
  Detect and remove duplicate rows or columns, and handle missing values with configurable options like mean, median, custom values, or flagging for review.
- **Outlier Detection and Management with Visualizations:**
  Identify statistical outliers and visualize them using histograms, scatter plots, and box plots to allow intuitive management (e.g., capping, flagging, or removal).

Functional Requirements

- **Data Standardization and Validation:**

  Normalize data formats (e.g., dates, numerical values, text case) and allow users to define and apply validation rules to ensure schema consistency and compliance.

- **Automated Data Transformation Pipelines:**

  Create reusable workflows for repetitive cleaning tasks like data formatting, redundancy removal, or missing value imputation.

- **Collaborative Cleaning with Version Control:**

  Enable multiple users to work collaboratively on datasets with tracked changes, including the ability to save versions and revert to earlier stages.

Functional Requirements

- **Export Options with Cleaning Logs:**

  Download cleaned data in formats like `.csv` or `.xlsx` along with a detailed log of the changes made during the cleaning process.

- **Enhanced Cleaning Insights:**

  Extend profiling functionality with statistical insights, such as data distributions, correlations, and warnings for inconsistent patterns, to guide the cleaning process.

Intellicleanse

Non-Functional Requirements

- **Performance:**

  The system should handle 100 concurrent users with fast page load times under typical internet conditions.
- **Security:**

  User data must be encrypted, and all communications should occur over HTTPS.
- **Scalability:**

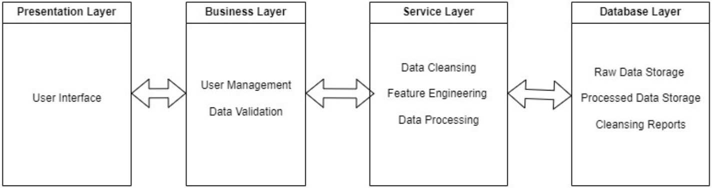
  The system must scale horizontally to handle increasing user demand.
- **Availability:**

  The system should have minimal downtime with failover mechanisms for continuous service.
- **Usability:**

  The application should be responsive and user-friendly across all devices.

Intellicleanse

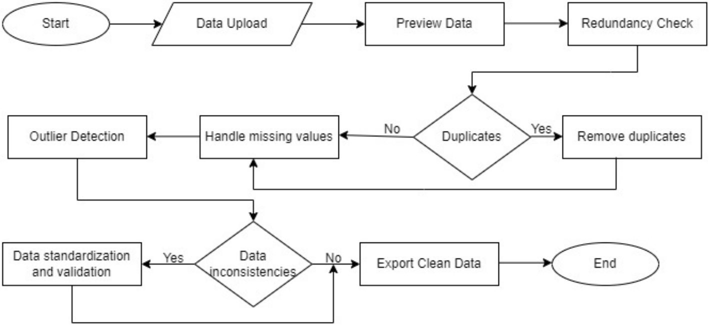
Architecture Diagram



ARCHITECTURE

Intellicleanse

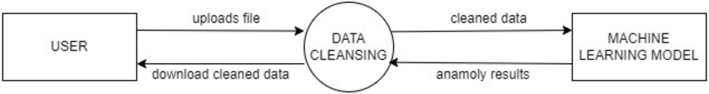
Process Flow Diagram



PROCESS FLOW

Intellicleanse

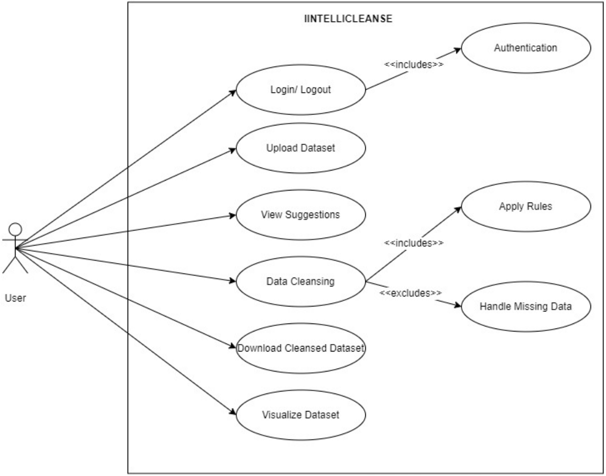
Context Flow Diagram



CONTEXT FLOW

Intellicleanse

Use Case Diagram



USE CASE

Application in Real World

- **Data Quality Enhancement:** Used by businesses to clean and preprocess large datasets, ensuring accuracy and consistency, particularly in industries like e-commerce, healthcare, and finance.
- **Data-driven Decision Making:**Helps organizations detect and eliminate data anomalies (outliers, redundancies) and standardize data for more reliable reporting and analysis.
- **Collaboration Efficiency:** Enables teams to collaboratively clean and refine data, improving the speed and quality of data analysis for actionable insights.

Reference

- D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT Press, 2001.
- I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Morgan Kaufmann, 2016.
- M. J. Pazzani, *Data Mining: Machine Learning Perspectives*, AAAI Press, 2000.
- P. D. Turney, *Types of Data Preprocessing*, 1999.
- D. J. Salgado, A. L. A. Coelho, and M. R. Silva, *Data Cleansing: Problems and Techniques in Data Science*, Springer, 2020.
- W. H. Inmon, *Building the Data Warehouse*, 4th ed., Wiley, 2005.
- P. V. S. V. Narayana and V. M. G. S. Pandit, *Data Cleaning Approaches in Data Mining*, Springer, 2018.

Reference

- H. V. Jagadish, A. A. Imran, and M. A. U. Zaman, *Data Quality and Cleaning in Data Science*, Springer, 2020.
- A. M. MacQueen, *Efficient Methods for Data Preprocessing and Cleaning*, 2015.
- R. Agerri and A. R. M. A. Yousaf, *Data Cleaning Techniques for Data Mining: A Review*, International Journal of Computer Applications, 2015.
- M. M. M. A. Mahmud, *Data Quality and Cleaning Challenges in Machine Learning Models*, 2017.
- S. C. G. J. H. P. J. Geurts, *Data Quality and the Impact of Data Cleansing on Predictive Performance*, Journal of Machine Learning Research, 2019.
- B. L. Masnadi-Shirazi, *Outlier Detection and Data Preprocessing in Data Science*, Springer, 2014.
- J. L. Ambroise and R. N. Dubes, *Data Cleaning and Normalization: A Comprehensive Survey*, Journal of Data Science, 2002.

Thank You

Vaishnavi R P

SRN : PES1PG23CA159

Department of Computer Applications