P.E.S. University
Dept. of Computer Applications Session:
Nov 2024 – Feb 2025

3rd Semester MCA
Capstone Project Phase-1
Synopsis

Project Title: Intellicleanse - An AI-Driven Data Cleansing Solution with Machine Learning

Abstract:

The quality of data is essential for producing trustworthy insights and promoting well-informed decision-making in today's data-driven society. The process of finding and fixing mistakes, inconsistencies, and inaccuracies in datasets to make sure they are correct, comprehensive, and prepared for analysis is known as data cleaning. This online application provides customers with a complete solution for efficiently cleaning `.csv` files. The tool streamlines the data cleaning process while improving data quality and usability with features including duplicate elimination, missing value imputation, outlier detection, and data profiling. Its user-friendly interface enables scientists, data analysts, and companies to convert unprocessed data into insights that can be put to use.

Functionalities:

1. 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.

2. 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.

3. 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).

4. 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.

5. Automated Data Transformation Pipelines
Create reusable workflows for repetitive cleaning tasks like data formatting, redundancy removal, or missing value imputation.

6. 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.

P.E.S. University
Dept. of Computer Applications Session:
Nov 2024 – Feb 2025

7. 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.

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

9. Dataset Balance Check
The system will analyse uploaded datasets to determine if they are balanced across categories. If an imbalance is detected, users will receive recommendations on handling unbalanced data through resampling techniques such as oversampling or under sampling.

Software Specifications:

Frontend	HTML CSS JavaScript
Backend	Python TensorFlow/Scikit-learn NumPy Pandas
Database	MySQL
Server	AWS

Submitted by:

SRN	Name	Student signature with date
PES1PG23CA159	Vaishnavi R Pachhapur	

Guide Name and Designation	Guide Signature with date
Dr. Lekha A	