MARRI LAXMAN REDDY GROUP OF INSTITUTIONS

MLR INSTITUTE OF TECHNOLOGY

(UGC AUTONOMOUS)

Laxman Reddy Avenue, Dundigal, Hyderabad - 500 043, Telangana , India

DEPARTMENT OF DATA SCIENCE & CYBER SECURITY

II CSD-B Mini Project

A.Y. 2023-2024

Date: 04-04-2024

Domain / Areas	Machine Learning, Data Science	
Title of the Project	Python-based Data Deduplication Tool for CSV Files	
Team Leader Name	1. K. Purna Vamsi	22R21A6787
Team Members Name with Roll No	2. M. Abhay Reddy	22R21A67A2
	3. S. Sai Nath	22R21A67B9
Guide Name	Y. Anjali	
Guide Signature		

ABSTRACT:

This project focuses on developing a Python-based tool for efficient data deduplication from CSV files, addressing the prevalent challenge of duplicate records. Existing solutions often lack flexibility and scalability, hindering effective data preprocessing. Therefore, our tool aims to provide a customizable and user-friendly solution, empowering users to easily identify and remove duplicate entries based on specified criteria. Leveraging Python's versatility, the tool offers scalability and efficiency in deduplication processes, catering to datasets of varying sizes. By streamlining the deduplication process and optimizing storage resources, the project aims to enhance overall data quality, improve analysis efficiency, and facilitate more accurate insights and informed decision-making.

Software/Hardware Needs	Python, CSV Module, IDE

Signature of the Incharge **Signature of the HOD(CS&DS)**