An Abstract on

“DATA DETOX: A SMART AND VISUAL SYSTEM FOR ANOMALY DETECTION AND INTERACTIVE CLEANING”

Submitted in partial fulfillment of the requirement for the
award of the degree of

Bachelor of Technology in
Artificial Intelligence & Data Science

By

H. KEERTHI	**(22701A3039)**
K. DINESH REDDY	**(22701A3019)**
A. MAHAMMAD SOHAIL	**(22701A3047)**
A. KHADEERULLA	**(22701A3042)**

Under the esteemed guidance of

Mrs. N. Swathi, MTech
Assistant Professor
Department of Artificial Intelligence & Data Science

Department of Artificial Intelligence & Data Science
Annamacharya Institute of Technology and Sciences
(Autonomous)
(Affiliated to J.N.T.University, Anantapur)
New Boyanapalli, Rajampet – 516126, Annamayya (Dist.), A.P

2025-2026

SUPERVISOR **PROJECT COORDINATOR** **HEAD OF THE DEPARTMENT**

ABSTRACT

Smart Data Cleaning Framework with Interactive Visualization and Outlier Discovery Abstract Data is one of the most precious means in moment's digital frugality, driving invention, decision- timber, and robotization across diligence. still, raw data is frequently agonized by issues similar as noise, missing values, outliers, inconsistencies and redundancies. These defects reduce the delicacy of machine literacy models, increase storehouse and computational costs and laggardly down data processing channels. In line with the principle of Garbage In, Garbage Out (GIGO), icing data quality before it enters analytics or AI systems is critical. To address these challenges, this design introduces an Automatic Data Cleaning Framework(AutoCleanAI) that combines anomaly discovery, profiling and channel integration to deliver cleaner and further dependable datasets. The system leverages insulation timber, a machine learning algorithm designed for unsupervised anomaly discovery, to identify unusual records or outliers that do n't conform to normal patterns within the dataset. By automatically segregating these irregularities, the frame prevents them from turning downstream analytics and prophetic models. In addition, the result incorporates automated profiling through Exploratory Data Analysis (EDA) reports which give precious perceptivity into data characteristics, including missing values, data distributions, correlations and indistinguishable entries. This not only pets up the preprocessing stage but also assists data scientists in making informed opinions with minimum homemade trouble. The integration of these ways within a data channel ensures that data is constantly covered, gutted and prepared for analysis in real- time. As a result, associations benefit from reduced storehouse costs, faster query prosecution, bettered model delicacy and a significant reduction in downstream crimes.

Keywords:

Automatic Data Cleaning, Anomaly Detection, Isolation Forest, Outlier Detection, Exploratory Data Analysis (EDA), Data Profiling, Data Pipeline Optimization and Reliable Machine Learning.

Objective:

The main aim of this project is to design and implement an automated and user-friendly system for data cleaning and anomaly detection. The application focuses on addressing common data quality issues such as missing values, duplicate entries, inconsistencies, and outliers. By combining machine learning techniques (Isolation Forest) with data profiling tools (EDA reports) and an interactive Streamlit interface, the system simplifies the data preprocessing stage. This ensures the availability of cleaner and more reliable datasets, which in turn improves the accuracy of downstream machine learning models and enhances decision-making processes.

Methodology:

The project workflow allows users to upload datasets into a Streamlit-based platform, where several automated operations are carried out.

- **Data Profiling:** Detailed profiling reports are generated using ydata_profiling, offering insights into distributions, correlations, missing values, and duplicate records.
- **Data Cleaning:** The framework applies automated techniques to handle missing values, remove duplicates, and correct inconsistencies within the dataset.
- **Anomaly Detection:** The Isolation Forest algorithm is employed to identify outliers and unusual data points that could negatively impact analysis.
- **Visualization:** Tools such as Seaborn, Matplotlib, and Plotly are used to create interactive visualizations, helping users better understand anomalies and verify cleaning results.

By integrating these components, the system provides a reliable and efficient solution that reduces manual effort and speeds up data preparation.

Key Findings:

The outcomes of this project highlight that automated data cleaning considerably decreases preprocessing time while improving the overall quality of datasets. The use of the Isolation Forest model proved effective in detecting anomalies and outliers, preventing unreliable records from influencing further analysis. Automated profiling enabled quick assessment of dataset health, while the Streamlit interface made the system easy to use for both technical and non-technical users. Overall, the framework produced cleaner datasets, minimized human errors, enhanced model accuracy, and improved confidence in analytics pipelines. These results demonstrate the system’s potential as a practical and scalable solution for real-world, data-driven applications.