**DataCronyx: AutoEDA & AutoTrainer**

**Project Report submitted in the partial fulfilment**

**Of**

Master of Technology

In

Artificial Intelligence

by

**Shardul Gore (R016)**

Under the supervision of

**Name of Faculty Mentor**

(Designation, Name of the department, MPSTME)

# SVKM’s NMIMS University

(Deemed-to-be University)



**MUKESH PATEL SCHOOL OF TECHNOLOGY MANAGEMENT & ENGINEERING (MPSTME)**

**Vile Parle (W), Mumbai-56**

**(2025-26)**

# CERTIFICATE

****

This is to certify that the project entitled DataCronyx: AutoEDA & AutoTrainer, has been done by **Mr. Shardul Gore** under my guidance and supervision & has been submitted in partial fulfilment of the degree of Master of Technology in Artificial Intelligence of MPSTME, SVKM’s NMIMS (Deemed-to-be University), Mumbai, India.

Project mentor (name and Signature) Examiner ( name and Signature) (Intenal Guide)

**Date**

**Place: Mumbai**

**(HoD)** (name and Signature)

# ACKNOWLEDGEMENT

Student Names

|  |  |  |
| --- | --- | --- |
| **NAME** | **ROLL NO.** | **SAP ID** |
| Shardul Gore | R016 | 70282400003 |
|  |  |  |
|  |  |  |
|  |  |  |

# ABSTRACT

This project presents DataCronyx, an intelligent, automated system designed to simplify and enhance the process of exploratory data analysis (EDA) and machine learning model development for diverse tabular datasets. The work addresses the common challenges faced by analysts and decision-makers in deriving insights and building predictive models efficiently, especially when technical expertise is limited.

DataCronyx offers a dual-mode functionality that significantly contributes to data analytics workflows. The first mode empowers users with a custom, user-friendly interface where they can manually guide the data cleaning, feature selection, and model training processes. This interactive environment enables fine control and customization suited to specific business needs. The second mode leverages AI agents to autonomously perform end-to-end data analysis and model training, reducing human intervention and accelerating the time from raw data to meaningful insights.

A key contribution of this work is the integration of automated EDA with high-quality visualizations and comprehensive data quality assessments, allowing for rapid and clear understanding of data patterns and anomalies. The system utilizes well-established machine learning libraries to support multiple modeling algorithms, providing robust model evaluation and comparison tools. Another novel aspect is the incorporation of large language model (LLM) technology—specifically Google Gemini API—to translate statistical findings and model outcomes into coherent, actionable narratives. This bridges the gap between complex analytics and accessible interpretation, enhancing decision-making for both technical and non-technical stakeholders.

Through the development and testing of DataCronyx, this project demonstrates how combining customizable workflows with intelligent automation can improve the efficiency, accuracy, and interpretability of data analysis pipelines. The system’s adaptability to various data sources, including file uploads and direct SQL database connections, further extends its practical applicability. Overall, DataCronyx advances the state of automated data science by delivering a scalable, explainable, and user-centric platform that empowers organizations to derive deeper insights and build predictive capabilities without extensive coding or domain expertise.

**Table of Contents**

**Topics Page**

**List of Figures** **i**

**List of Tables** **xvi**

**Abbreviations** **xix**

**Chapter 1 Introduction**

**1.1 Background of the project**

**1.2 Motivation and Scope**

**1.3 Problem Statement**

**1.4 Salient Contribution**

**Chapter 2 Literature survey 2.1 Introduction**

**2.2 Exhaustive Literature survey**

**Chapter 3 Methodology and Implementation**

**3.1**

**3.2**

**Chapter 4 Result and Analysis**

**4.1**

**4.2**

**Chapter 5 Advantages, Limitations and Applications**

* 1. **Advantages**
  2. **Limitations**
  3. **Applications**

**Conclusion and Future Scope**

**References**

**Appendix A: Soft Code Flowcharts Appendix B: Data Sheets Appendix C: List of Components**

**Appendix D: List of Paper Presented and Published.**

List of the figures

|  |  |  |
| --- | --- | --- |
| **Fig No** | **Name of the figure** | **Page No** |
|  |  |  |
|  |  |  |

List of tables

|  |  |  |
| --- | --- | --- |
| **Table No** | **Name of the Table** | **Page No** |
|  |  |  |
|  |  |  |

Abbreviations

|  |  |  |  |
| --- | --- | --- | --- |
| **Table No** |  | **Name of the Table** | **Page No** |
|  |  |  |  |
|  |  |  |  |

**Note: Each table should be included on separate page.**

**Chapter 1 Introduction**

* 1. **Background of the project topic**

In the digital age, organizations and individuals generate and collect vast volumes of data from a variety of sources, including business transactions, sensors, social media, and online platforms. The effective use of this data through analytics and predictive modeling has become central to innovation, competitiveness, and informed decision-making across industries. Central to this process are two foundational practices: Exploratory Data Analysis (EDA) and machine learning (ML). EDA serves as the bridge between raw data and valuable insights, enabling users to investigate the structure, patterns, anomalies, and relationships within datasets. Once the quality and key features of the data are understood, machine learning techniques further unlock its predictive potential, informing forecasts, classifications, optimizations, and recommendations.

Despite remarkable advances in the field, the practical adoption of data analytics remains hindered for many users. Typical analytics workflows are fragmented, involving data cleaning, visualization, transformation, and modeling steps, each often requiring different tools and programming skills. Furthermore, there is a steep learning curve associated with understanding statistical outputs and the implications of model performance. This creates barriers for non-experts while leaving data scientists encumbered by repetitive, manual tasks that detract from higher-value creative work. The vision to lower these barriers and create a more seamless, interpretable, and efficient path from raw data to decision-ready insights is what motivates this project.

* 1. **Motivation and scope**

Motivated by the need to democratize access to advanced analytics, the present work focuses on developing DataCronyx, a comprehensive and intelligent data analytics platform. The motivation is twofold: First, to enable users—regardless of their technical background—to conduct thorough and insightful EDA coupled with robust machine learning, and second, to automate and accelerate these workflows using the latest advancements in artificial intelligence (AI) agents.

The scope of this report includes the design, implementation, and evaluation of a platform that can handle tabular data from a wide array of sources, including file uploads (CSV, Excel, JSON) and direct SQL databases. DataCronyx is envisioned as an easy-to-use web interface, incorporating both a traditional, interactive mode—where users retain full control over parameter selection—and an automated mode, where AI agents seamlessly process and analyze data with minimal input. This dual approach ensures that both novice and expert users can benefit from the platform, choosing either a guided experience or leveraging built-in automation for efficiency.

* 1. **Problem statement**

The field of data analytics continually faces several persistent challenges:

* Analysts often spend a disproportionate amount of time on routine data wrangling, quality checks, and plotting, rather than on generating actionable insights.
* Manual feature selection, preprocessing, and model evaluation are both error-prone and time-intensive, particularly for large or complex datasets.
* While automated solutions exist, many focus exclusively on either EDA or model automation, lacking an integrated, holistic approach or flexibility for user intervention.
* The presentation of findings is frequently in technical or statistical format, alienating business users or stakeholders who may lack domain expertise but are responsible for making key decisions.
* There is a growing demand for systems that combine robust automation with explainability, flexibility, and accessibility, narrowing the gap between data and impactful action.
  1. **Salient contribution**

This project brings forth several innovative contributions to the domain of automated analytics and AI-assisted decision support:

* A dual-mode data analysis workflow: DataCronyx allows users to select between a fully custom, user-driven EDA/ML flow, and an AI agent-driven automatic pipeline that manages data exploration, feature engineering, and model training end-to-end.
* Flexible data ingestion and preparation: The system facilitates both conventional file uploads and live connections to SQL databases, broadening its applicability and ensuring compatibility with real-world business environments.
* Comprehensive, automated EDA and reporting: With automated generation of descriptive statistics, high-resolution visualizations, and data quality diagnostics, users can rapidly assess and understand their datasets regardless of scale or complexity.
* Integrated machine learning with user choice: Multiple models are available, with side-by-side evaluation and leaderboard presentation; the best models can be exported for deployment or further analysis.
* Natural language, AI-powered explanation: By leveraging advanced large language model (LLM) technology (Google Gemini API), the platform summarizes results and model findings in clear, accessible language, closing the gap between technical analytics and business application.
* Accessible user interface: Intuitive design principles ensure DataCronyx is approachable for non-technical users, while also offering depth and customization for power users and data professionals.

**Chapter 2 Literature survey**

* 1. **Introduction to overall topic**

Exploratory Data Analysis (EDA) is a foundational step in data science that involves summarizing the main characteristics of datasets, detecting anomalies, and uncovering underlying patterns to inform subsequent modeling and decision-making. Traditionally, EDA relied heavily on manual inspection, statistical summaries, and visualization techniques, which can be time-consuming and require significant expertise—especially when dealing with large, complex, or heterogeneous data sources. With the explosive growth of data across domains and the increasing demand for rapid and robust analysis, automation in EDA has become a critical area of research.

In recent years, Artificial Intelligence (AI) and Large Language Models (LLMs) have shown promise in transforming the EDA landscape by automating data preprocessing, feature engineering, and insight generation while bridging the gap between technical complexity and user accessibility. Automated Machine Learning (AutoML) frameworks complement this by reducing the burden of model selection and hyperparameter tuning, enabling faster predictive modeling. However, real-world EDA presents challenges such as complex database schemas, unclear user intent, lack of cross-domain generalization, and the need for interpretable and personalized insights.

This literature review consolidates existing approaches to automated exploratory data analysis, text-to-SQL query generation, AI-driven visualization recommendations, and natural language powered data interpretation, drawing on insights from recent peer-reviewed publications. It sets the foundation for developing end-to-end systems that empower users to conduct thorough, explainable, and scalable data analysis using both custom interactive workflows and AI agent-driven automation.

* 1. **Exhaustive literature survey**

Exploratory Data Analysis (EDA) is fundamental to understanding the structure, trends, and anomalies of a dataset, forming the basis for more advanced modeling and informed decision-making. Traditionally, EDA has relied on manual processes involving descriptive statistics and visualizations, which, while effective for small datasets, are cumbersome and inadequate for the scale and complexity of modern data [2]. Manual EDA is further limited by the need for significant domain expertise, the risk of human error, and difficulties in scaling to big data environments.

In response to these challenges, research has explored the use of Artificial Intelligence (AI) and Machine Learning (ML) to automate the EDA process [2]. AI-based EDA solutions harness machine learning, deep learning, and natural language processing (NLP) to automate tasks such as data cleaning, transformation, feature extraction, anomaly detection, and the generation of interactive, context-aware visual summaries [2]. By eliminating manual intervention and accelerating analytical workflows, such systems facilitate more efficient, scalable, and accurate exploration of vast datasets, thus bridging the gap between raw information and actionable insights.

Yet, the literature acknowledges significant limitations in both traditional and AI-based EDA systems. Many early AI-driven EDA frameworks were constrained to structured, numerical data and specific formats like CSV, lacking effective handling of high-volume, high-variety, or unstructured datasets [2]. Furthermore, these tools often provided only basic visualizations or static charts with minimal support for real-time interaction, adaptive insights, or user-centered explanation. The growing adoption of Natural Language Processing in EDA systems has allowed users to interrogate data in plain language, but the quality and accuracy of responses—particularly in ambiguous or domain-specific scenarios—remain open areas for improvement.

Recent research has introduced new paradigms to overcome these barriers. Zhu et al. present TiInsight, an end-to-end, LLM-augmented cross-domain EDA solution designed to automate the entire process: from schema summarization and intent clarification, to text-to-SQL translation and adaptive visualization [1]. TiInsight utilizes hierarchical data context generation and map-reduce frameworks to address real-world challenges like complex, enterprise-scale database schemas and vague user queries, outperforming several benchmark systems in execution accuracy and user satisfaction. Their research underlines the important shift towards context-aware, explainable, and highly automated EDA workflows. However, the study also recognizes the continued difficulties in generalizing across data domains, managing context window limitations, and offering universal, user-friendly interfaces.

Wu et al. further extend the automation of EDA through Attribute Frequency Statistical Feature Ratio (AFSFR), enabling iterative data focusing, dynamic field selection, and feedback loops for user-inferred cognition [3]. Their AutoEDA-Segment system incorporates clustering, parallel coordinates-based filtering, and field type identification, resulting in more personalized, adaptive, and interpretable exploration. The effectiveness of this method is validated through case studies in meteorology and healthcare, demonstrating improved analyst efficiency and depth of insight [3].

Despite these advancements, the literature agrees on outstanding research gaps. First, most platforms lack seamless integration between user-driven and fully automated workflows, limiting flexibility for varying user expertise and use cases [1][2][3]. Second, there remains insufficient support for diverse, cross-domain data ingestion, including unstructured and multi-relational datasets [1][2]. Third, natural language interfaces still struggle with intent ambiguity and do not offer robust multi-step reasoning or advanced visualization recommendations in all scenarios [1]. Fourth, issues of explainability, bias mitigation, data privacy, and high computational costs persist as significant challenges for production-grade EDA systems [2].

In summary, while AI-driven EDA systems have made notable strides by automating repetitive analysis and improving accessibility, current approaches do not sufficiently unify customizable, user-guided EDA with fully autonomous AI-driven pipelines in one extensible platform [1][2][3]. There is also a persistent need for improved natural language interaction, cross-domain data handling, deeper integration of advanced visualization techniques, and transparent, explainable results. Hence, the problem addressed in this project is the design and implementation of a scalable, dual-mode EDA platform—DataCronyx—that empowers both user-driven and AI-agent-driven workflows, accommodates varied data sources, and leverages LLM technology for highly interpretable and accessible insights.

**Chapter 3 Methodology and Implementation**

* 1. **Block diagram**
  2. **Software description, flowchart / algorithm**

This chapter can comprise of actual implementation photos and their description.

**Chapter 4 Results and Analysis**

This shall include a thorough evaluation and investigation carried out. It should also bring out your contributions from the study. The discussion shall logically lead to inferences and conclusions as well as scope for possible further future work.

**Note:**

**Include the IEEE or any other standards that you have adhered to test the validity of the results.**

**Link for IEEE standards**

https://[www.ieee.org/content/ieee-org/en/standards/index.html/](http://www.ieee.org/content/ieee-org/en/standards/index.html/)

**Chapter 5**

**Advantages, Limitations and Applications**

**Chapter 6 Conclusion and Future Scope**

* A brief report of the work carried out, conclusions derived from logical analysis presented in the Results and Discussions chapter.
* Scope for future work should be stated lucidly in this chapter.

**References**

**[1] Zhu, J. P., Niu, B., Cai, P., Ni, Z., Wan, J., Xu, K., ... & Liu, Q. (2024). Towards Automated Cross-domain Exploratory Data Analysis through Large Language Models. arXiv preprint arXiv:2412.07214.**

**[2] Prof. Jyoti Gaikwad, Aniket Manohare, Shweta Munde, Anwar Shaikh, and Diksha Subhedar, “AI-Based Exploratory Data Analysis”, Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol, vol. 11, no. 2, pp. 3876–3884, Apr. 2025, doi: 10.32628/CSEIT25112860.**

**[3] Wu, T., Wang, S., & Peng, X. (2024, October). AutoEDA: Iterative Data Focusing and Exploratory Analysis Based on Attribute Frequency. In 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 4113-4118). IEEE.**

**Appendix A: Soft Code Flowcharts**

**Appendix B: Data Sheets**

**Appendix C: List of Components**

**Appendix D: List of Paper Presented and Published**

* List of papers on the topic of the report published by the candidates.
* This may also be included in the contents.
* The candidates may also include reprints of his/her publications after the literature citation.