MSME (Micro, Small and Medium Enterprises) Data Analysis in Telangana Using PySpark: A Big Data Analytical Study

Abstract

This study presents a comprehensive analysis of Micro, Small, and Medium Enterprises (MSMEs) in Telangana using PySpark, a powerful Big Data framework. The research leverages large-scale MSME registration datasets to identify patterns, trends, and industrial distributions across the state. By applying data preprocessing, aggregation, and classification methods, this study provides insights into the structure and evolution of MSME activity, supporting evidence-based policy formulation and economic planning.

Introduction

Micro, Small, and Medium Enterprises (MSMEs) represent one of the most dynamic sectors in Telangana's economy. They contribute significantly to employment generation, innovation, and regional economic development. Understanding the geographical, sectoral, and temporal distribution of MSMEs is crucial for designing targeted industrial policies. This study employs PySpark, a distributed computing framework, to perform large-scale analysis of MSME registration data and to extract key insights on the state's entrepreneurial ecosystem.

Methodology

The methodology followed in this study integrates data engineering, big data analytics, and exploratory data analysis techniques. The MSME dataset was obtained from official government records containing enterprise names, registration dates, sectors, and districts. The steps include:

1. Data Cleaning: Removal of missing and inconsistent records.
2. Data Transformation: Standardization of column names and categorical encoding.
3. Data Analysis: Aggregation of enterprises by district, sector, and registration year.
4. Classification: Categorization into micro, small, and medium sectors using keyword-based tagging.
5. Validation: Ensuring consistency and accuracy in classification results.

Results

The analysis revealed a strong concentration of MSME registrations in industrially developed district such as Rangareddy. Manufacturing and service-based enterprises constituted the majority of registrations. Temporal analysis indicated consistent growth in MSME registrations over recent years, reflecting Telangana's improving business environment. Additionally, micro-enterprises accounted for a dominant share, highlighting the need for further support to facilitate their transition into small and medium segments.

Discussion

The findings emphasize the concentration of MSME activity around urban industrial clusters, particularly in and around Hyderabad. While this indicates economic dynamism, it also reflects uneven industrial development, as rural districts show relatively lower MSME density. The predominance of micro-enterprises suggests limited scalability and access to capital. Targeted interventions in financing, infrastructure, and technology adoption could enhance the growth potential of smaller enterprises.

Limitations

The study's limitations stem primarily from the scope of the dataset. Certain enterprise attributes such as turnover, employment size, and sectoral sub-classifications were not available. Additionally, since the dataset represents registrations, it may not fully reflect currently active enterprises. Future analyses can incorporate dynamic datasets and integrate GIS-based visualization for enhanced spatial accuracy.

Data Completeness

The dataset demonstrated a high degree of completeness regarding registration details, including enterprise names, district locations, and classification types. However, some records lacked standardized naming conventions or contained typographical errors, which required preprocessing. Post-cleaning validation confirmed that the usable portion of the data retained over 95% of the original entries, ensuring representativeness.

Data Accuracy

Data accuracy was ensured through a multi-step validation process involving duplicate removal, normalization, and format verification. Enterprise names were cross-referenced using keyword-based classification to ensure correct categorization across sectors. While minor inconsistencies were observed in regional spellings, these did not materially affect aggregate analytical outcomes.

Conclusion

The analysis underscores the critical role of MSMEs in Telangana's economic fabric. The PySpark-based framework successfully demonstrated the efficiency of big data analytics in managing and interpreting large-scale industrial datasets. Findings indicate an expanding MSME ecosystem, dominated by micro-level enterprises, with potential for diversification and technological upgrading.

Future Work and Recommendations

Future studies can integrate additional data dimensions such as employment size, capital investment, and financial performance to enrich the analysis. Implementing predictive models for MSME growth forecasting can also help identify emerging industrial clusters. It is recommended that policymakers enhance data reporting mechanisms, promote digital transformation among MSMEs, and provide region-specific support schemes to ensure inclusive economic growth.

References

[1] Ministry of Micro, Small and Medium Enterprises, Government of India, Annual Report 2024.
[2] Telangana State Industrial Infrastructure Corporation (TSIIC) Reports, 2024.
[3] Apache Software Foundation, PySpark Documentation.
[4] World Bank MSME Finance Gap Report, 2023.