Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

A Machine Learning Framework for Evaluating MSME Incubation Dynamics

Taflin S Raj[1]*, Radhika R.[2]

[1]Research Scholar, [2]Associate Professor, Department of Management Studies, Noorul Islam Centre for Higher Education, Kumaracoil, Thuckalay, India.

Email: [1*]tafi112003@gmail.com, [2]rdhujaasourish@gmail.com

Abstract

The Business Incubation is an important way to assist entrepreneurs in transferring their innovative ideas into viable businesses, offering resources, expertise, and connecting them with available resources for assistance. In India, the MSME Hackathons represent one type of competition to generate new and exciting ideas by identifying the best solutions across different technology and industry sectors. This study will develop a data-based framework for understanding innovation patterns from the MSME business incubation framework by using the officially approved ideas generated in multiple MSME Idea Hackathons. This research will also use an innovation analytics perspective as the basis for developing results, rather than using traditional descriptive approaches for understanding MSME incubator policy development; therefore, the results will identify how many institutions, states, and knowledge bases are in the overall data set and relate to the overall structure of the innovation network. A machine learning–based analytical approach is employed to examine innovation domains and structural relationships among different innovation ecosystems. Supervised ensemble modeling is used to evaluate the influence of institutional affiliation and geographical location on thematic specialization, while unsupervised clustering and dimensionality reduction techniques are applied to explore structural patterns within the dataset. The classification model achieved an accuracy of 0.672, precision of 0.599, recall of 0.672, and an F1-score of 0.624, indicating moderate predictive capability while confirming that institutional and geographical attributes provide meaningful signals for innovation domain classification. The results further indicate that the structure of institutional networks and regional environments significantly influences innovation-specific focus areas, while also revealing the coexistence of both organized and distributed forms of innovation activity reflected in hackathon participation patterns. The study contributes a scalable and interpretable analytical framework for the macro-level evaluation of incubation initiatives and provides empirical insights to support evidence-based policy formulation aimed at strengthening the MSME innovation system.

Keywords: MSME Incubation, Innovation Ecosystems, Hackathon Analytics, Machine Learning, Random Forest, Clustering Analysis, Innovation Policy.

1. Introduction

The concept of business incubation under the MSME sector in India has gradually developed as an integral strategy for promoting entrepreneurship, innovation, and facilitating support for new businesses. The concept of business incubation was formally adopted in India

* Corresponding Author

DOI: https://doi.org/10.36548/jtcsst.2026.1.006
Received: 17.01.2026, received in revised form: 16.02.2026, accepted: 02.03.2026, published: 11.03.2026

in the late 1980s and early 1990s, with institutions such as the National Science and Technology Entrepreneurship Development Board (NSTEDB), formed in 1982 under the Department of Science and Technology (DST), promoting Technology Business Incubators (TBIs) in the country. However, the concept was gradually adopted, with the Indian economic liberalization of 1991 witnessing the issue of business incubation support receiving significant attention for promoting small businesses in the country. The MSME sector, an integral part of the Indian economy, has witnessed an important boost with the approval of the "MSME Business Incubation (BI) Scheme" by the Ministry of MSME in 2008 under the National Manufacturing Competitiveness Programme (NMCP). This was an important milestone in the context of the MSME sector, with the objective of promoting the conversion of innovative ideas into successful ventures through financial support.

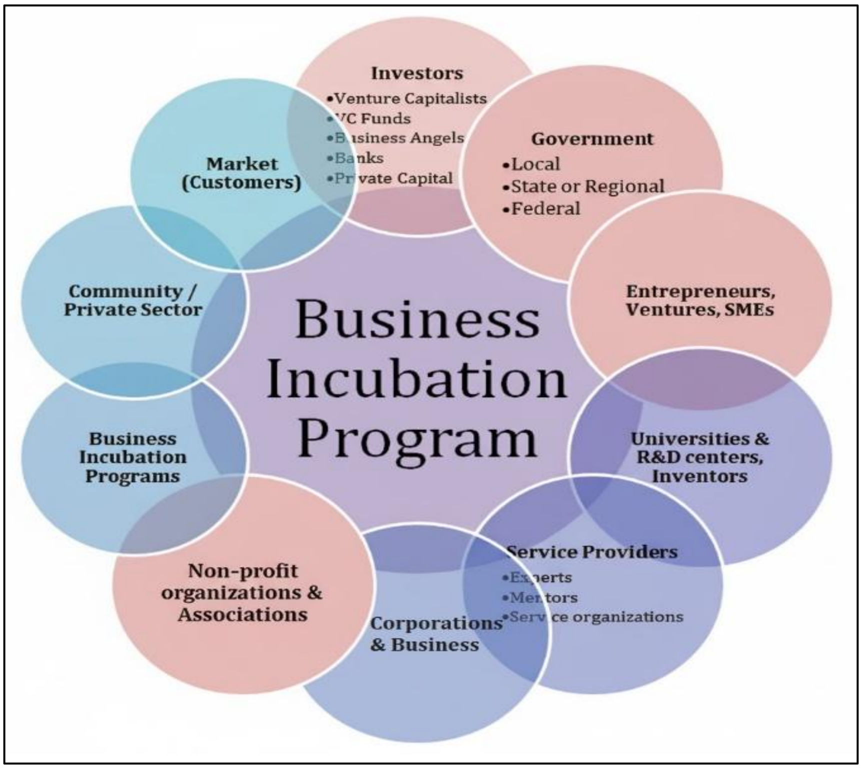

Figure 1. Business Incubation – An Overview [11]

Figure 1 provides a general overview of the concept of business incubation. Under this concept, eligible institutions such as engineering colleges, management institutions, research centers, industry associations, and technical universities are recognized as "Host Institutions" for developing incubation centers. The centers are enabled to assist innovators in various ways, including access to facilities and industry linkages. In this regard, over time, more initiatives have emerged, including Start-up India (launched in 2016), and Atal Innovation Mission (launched in 2016). Currently, incubation centers exist in various sectors such as information technology, biotechnology, agricultural business, and renewable energy, creating opportunities for MSMEs/start-ups to scale up their innovations.

The evolution of business incubation in India reflects the country's increasing focus on creating a culture of entrepreneurship. From its introduction in the 1980s through technology-based incubators to the formal approval of the MSME Business Incubation Scheme in 2008, business incubation has come a long way in becoming a key driver for supporting MSMEs in India. Not only has business incubation helped MSMEs access resources and expertise, but it has also significantly contributed to employment creation, regional industrialization, and overall growth in India. By creating a synergy between business incubation support systems, government policies, academic institutions, and industry networks, India has established a vibrant business ecosystem where ideas can be taken from conceptualization to sustainable business ventures, thereby strengthening MSMEs in India as a whole [1].

In this regard, the MSME Idea Hackathon initiative can be seen as a modern incubation-related innovation platform that transcends conventional institution-based incubation practices. As a hackathon conducted over various editions (2.0, 3.0, and 4.0), it can be viewed as a national open innovation funnel for recognizing, funding, and mentoring entrepreneurial ideas in various sectors and regions. By consolidating ideas from different states and accredited host institutions, the recognized sets of ideas can be viewed as a national innovation set with regard to regional participation, thematic specialization, and institutional engagement. The latest edition, MSME Idea Hackathon 5.0, was launched on June 27, 2025, and this edition continues to enhance this platform by providing a sharper focus on strategically important domains such as low-carbon technologies, stealth and cyber defense technologies, integration of Industry 4.0 and 5.0 technologies in MSMEs, sustainability in coastal and hilly areas, and supply chain systems. The selected ideas in this latest edition are also eligible for non-equity grants of up to ₹15 lakh along with incubation support [1]. Although the strategic and policy-scale hackathons are highly relevant and significant in the Indian context, there is limited research done on the large-scale structural innovation dynamics of such hackathons.

Hackathons can be seen as operational surrogates for innovation ecosystems because they bring together the various actors involved in the process of innovation, such as institutions, entrepreneurs, mentors, and funding agencies. Unlike the usual incubation process that takes place in an isolated institutional environment, hackathons can be seen as an open innovation process that brings together ideas from different regional locations and technological spaces and puts them to the test against each other. The data generated can be regarded as the interaction between regional locations, host institutions, and thematic innovation spaces and can be seen as an operational surrogate that captures some of the structural characteristics of innovation ecosystems, such as knowledge spreading, institutional engagement, and thematic specialization. The study aims to break away from the general discourse on incubation and create a computational model that can analyze the national-level MSME innovation patterns.

2. Literature Review

An analysis of the concept of incubation in India at multiple levels has been offered by Sharma and Vohra (2020) [2], where the concept has been discussed from the individual, organizational, and ecological perspectives. Furthermore, the role and importance of technology business incubators in the country have been discussed by Rai, Prasad, and Murthy [3], where the authors stress the need for the development and extension of such incubators in the country, particularly in rural and semi-urban areas.

The role and importance of incubation in the MSME sector have also been discussed in the research paper on the concept of creative start-ups in the MSME sector by Krishnan, Ganesh, and Rajendran (2020) [4], where the authors offered a paradigm for the differentiation of high-growth companies from conventional small businesses. Furthermore, the role and importance of incubation in the MSME sector were discussed in the research paper by Sharma, Shukla, and Joshi (2014) [5] on the role and importance of business incubators in the success of start-ups and the survival rates of start-ups.

From the perspective of rural and inclusive development, Koshy [6] highlights the significance of rural business incubators in making microfinance initiatives sustainable businesses. The results of Koshy's study emphasize that incubation, when combined with financial inclusion, may lead to the empowerment of marginalized groups and provide livelihood opportunities for them in the long term. Another study by Jamnekar and Naranje [7] establishes the connection between business incubation and Start-up India, which asserts that entrepreneurs in MSME sectors benefit from business incubation in terms of innovation, performance, and growth, thereby contributing to the growth of the country.

In relation to broader systemic perspectives, Lala and Sinha (2018) [8] wrote about the innovation technology incubation system in India in relation to its evolution, its structure, and the interaction of stakeholders in the system. They believe that the incubation system in India has to be enhanced in its collaborative aspect with academia, government, and industries for it to be effective. Recently, Ravichandran and Dixit (2024) [9] wrote about innovation and incubation centers in empowering the new generation of entrepreneurs, especially the youth. They believe that incubation centers are not only essential for start-ups, but they are also crucial in instilling the entrepreneurial mindset in the youth, thus making them essential in vocational education in India.

Although the existing literature has mainly focused on policy frameworks, incubation models, and success stories, the macro-level analytics of the innovation ecosystems has attracted relatively less attention. Moreover, the existing literature has recently started employing machine learning methods in the innovation analytics and technology policy studies. For example, clustering algorithms, ensemble classifiers, and network analysis have been employed in the study of innovation clusters, technological trends, and knowledge diffusion patterns in the large datasets of innovation. For instance, the unsupervised learning methods have been employed in the study of innovation clusters in the geographical regions using the patent datasets. Similarly, the supervised learning methods have been employed in the classification of technological domains and the forecasting of emerging technologies.

However, the application of machine learning methods in the incubation and hackathon datasets is limited, the present study proposes a machine learning-based innovation analytics framework for the thematic domain specialization, regional clustering, and temporal innovation dynamics in the MSME hackathon innovation ecosystem. Moreover, the study provides a data-driven perspective on the structural patterns and incubation-linked innovation development in the context of India.

3. Proposed Methodology

3.1 Dataset Description

The proposed research establishes a theoretically grounded innovation analytics framework for modeling, interpreting and structurally characterizing the MSME Idea Hackathon ecosystem across multiple editions (Hackathon 2.0, 3.0, and 4.0)

- List of 276 approved Ideas under MSME Idea Hackathon 2.0 (Theme Based)
- List of 397 approved Ideas under MSME Idea Hackathon 3.0 (Women)
- List of 488 approved Ideas under MSME Idea Hackathon 4.0 (Youth)

These datasets form a massive national innovation corpus, which has 1,161 validated entries as seen in the experimental dataset [12]. Every entry in the dataset encapsulates institutional, geographical, thematic, and entrepreneurial aspects of innovation. From a theoretical perspective, the dataset can be viewed as a structured form of an innovation diffusion network, in which states and institutions are innovation nodes and thematic domains are knowledge clusters. The proposed methodology thus combines document mining, semantic structuring, statistical learning theory, ensemble modeling, and unsupervised structural discovery into a single framework.

Let the extracted innovation dataset be defined as,

$$\mathcal{D} = \{(v_i,\ \text{inst}_i,\ \text{st}_i,\ \text{theme}_i)\}_{i=1}^{N}$$

where $N = 1161, v_i$ denotes hackathon version (temporal innovation stage), inst $_i$ denotes institutional incubator, st_i denotes state/UT (regional innovation geography), and $theme_i$ denotes thematic description (knowledge classification).

3.2 Proposed Framework

The general framework is divided into five stages that are motivated by theory: (i) Data extraction and reconstruction, (ii) Data uniformity and semantic mapping, (iii) Statistical analysis and feature extraction, (iv) ML based inference and (v) ML based data analysis.

Stage I: Data Extraction and Reconstruction

The first step is based on the theory of information extraction. In the case of the hackathon data being in PDF format and represented as a table, using text extraction would cause the disruption of the rows' semantic alignment. Therefore, coordinate-aware word clustering is used to align the logical row groupings based on vertical positional similarity. Formally, let the extracted word be represented as

$$w_j = (\text{text}_j, x_j, y_j)$$

where (x_j, y_j) are spatial coordinates. Words satisfying $|y_j - y_k| < \epsilon$ are grouped into the same logical row. This ensures structural consistency in dataset construction. The outcome is a structured tabular dataset $\mathcal{D}$ suitable for quantitative modeling.

Analysis of state-wise participation was conducted based on consolidated approved project counts from the different editions of the MSME Idea Hackathon 2.0, 3.0, and 4.0. In order to structure the data, a semi-automated data extraction pipeline was introduced as an additional step in the framework. In this step, the tuple was represented as:

$$S_i = (\text{ State}_i,\ \text{Version}_i,\ \text{Count}_i)$$

where S_i denotes state-wise project count for hackathon version i. The overall state participation was computed using an aggregation function:

$$C_{\text{total}}(s) = \sum_{v=2.0}^{4.0} C_{s,v}$$

where $C_{s,v}$ represents the number of approved projects from state s in version v. The final dataset was structured as:

$$\mathcal{D}_{\text{state}} = \{(s, C_{s,2.0}, C_{s,3.0}, C_{s,4.0}, C_{\text{total}}(s))\}$$

A focused Southern-region analysis, containing Tamil Nadu, Kerala, Andhra Pradesh, Telangana, Karnataka, and Puducherry, was carried out to further explore regional concentration

Mathematically, the Southern contribution ratio was computed as follows:

$$R_{\text{south}} = \frac{\sum_{s \in \text{ South}} C_{\text{total}}(s)}{\sum_{s \in \text{ All}} C_{\text{total}}(s)}$$

This structured dataset enabled comparative visualization and regional clustering analysis.

Theoretically, this stage converts unstructured document space into structured data space, reducing representation entropy and preserving relational dependencies.

Stage II: Data Uniformity and Semantic Mapping

The innovation themes in different versions of hackathons are textually heterogeneous in nature. To ensure conceptual uniformity, it is mandatory to define a function $f(\cdot)$ to map heterogeneous innovation themes to canonical innovation domains.

$$\text{Domain}_i = f(\text{ theme}_i)$$

The mapping is rule-based and grounded in keyword semantics:

$$f(t) = \begin{cases} D_1 & \text{Digital Agriculture} \\ D_2 & \text{Healthcare} \\ D_3 & \text{Sustainability} \\ D_4 & \text{AI (or) IoT} \\ D_5 & \text{General Innovation} \end{cases}$$

This stage is theoretically motivated by ontology-based knowledge classification, in which heterogeneous lexical descriptions are mapped to a finite conceptual domain space $\{D_1, \dots, D_5\}$. The transformation results in semantic dimensionality reduction, making it comparable across different versions of hackathons.

The keyword list for semantic transformation was constructed by following a process that incorporates domain knowledge as well as exploratory text analysis. Initially, keywords like agriculture, AI, IoT, healthcare, waste management, sustainability, etc., were identified from the preliminary review of hackathon documents. The keywords were classified under different innovation domains by referring to policy guidelines for MSME incubation schemes and were refined by analyzing term frequency in idea descriptions.

Stage III: Statistical Analysis and Feature Extraction

To enable machine learning, categorical variables must be embedded into numerical feature space. Let

$$X = [\text{ State }_{\text{enc}}, \text{ Institute }_{\text{enc}}, \text{ Version }_{\text{enc}}]$$

represent the encoded feature matrix and $y = \text{ Domain }_{\text{enc}}$ represent the target vector.

Label encoding transforms categorical values into integer mappings while preserving identity distinctions. The dataset is partitioned into training and testing subsets using stratified sampling.

$$(X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}}) = \text{Split}(X, y, \alpha)$$

where $\alpha = 0.8$. Stratification ensures class distribution stability, reducing sampling bias and preserving empirical risk consistency. This stage operationalizes empirical risk minimization within supervised learning theory.

Stage IV: Machine Learning based Inference

A Random Forest classifier is employed due to its theoretical advantages in variance reduction and nonlinear boundary modeling. Random Forest constructs multiple decision trees $h_t(X)$ over bootstrapped samples and aggregates predictions via majority voting:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(X)$$

where $T = 200$ trees. Each tree splits nodes by minimizing Gini impurity:

$$G = 1 - \sum_{k=1}^{K} p_k^2$$

where p_k is the class probability at a node. From ensemble theory, averaging multiple decorrelated trees reduces variance without significantly increasing bias, thereby improving generalization performance. Feature importance is computed using the mean decrease in impurity, providing explainable AI insights into which institutional or geographical factors most influence innovation domain classification.

Model evaluation metrics are defined as:

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision } = \frac{TP}{TP + FP}$$

$$\text{Recall } = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{\text{Precision } \cdot \text{ Recall}}{\text{Precision } + \text{ Recall}}$$

These metrics collectively evaluate classification reliability, predictive selectivity, and class sensitivity.

Stage V: Machine Learning based Data Analysis

Beyond prediction, the study investigates structural innovation topology using clustering. Features are standardized using z-score normalization:

$$X' = \frac{X - \mu}{\sigma}$$

ensuring scale invariance. K-Means clustering partitions the dataset by minimizing within-cluster variance:

$$\min_{C} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

where $K = 4$. Theoretically, this objective function corresponds to minimizing intra-cluster dispersion and maximizing inter-cluster separation. The discovered clusters represent latent innovation ecosystems defined by shared institutional-geographical characteristics.

Dimensionality reduction is performed using Principal Component Analysis (PCA), which projects standardized features onto orthogonal eigenvectors:

$$Z = X'W$$

where W contains eigenvectors associated with the largest eigenvalues of the covariance matrix. PCA enables visualization while preserving maximum variance.

Temporal innovation evolution is quantified using cross-tabulation:

$$\text{Trend}_{ij} = | \{x \in \mathcal{D} : \text{Version} = i, \text{Domain} = j\} |$$

which provides macro-level insight into thematic shifts across hackathon editions.

Algorithm 1: MSME Innovation Domain Analytics Framework

```
Extract words with coordinates ($text, x, y$) from PDF files
Group words into structured rows if
    $|y_j - y_k| < \epsilon$
Construct structured dataset D
for each record $i \in$ D do
    Assign domain using semantic mapping:
        $\text{Domain}_i = f(\text{theme}_i)$
end for
Encode categorical variables to obtain feature matrix $X$
Encode domain labels to obtain target vector $y$
Perform stratified split:
    $(X_{train}, X_{test}, y_{train}, y_{test}) = \text{Split}(X, y, 0.8)$
Train Random Forest model with $T = 200$ trees
Predict $\hat{y}$ on test set
Compute Accuracy, Precision, Recall, and F1-score
Standardize features:
    $X' = \frac{X - \mu}{\sigma}$
Apply K-Means clustering with $K = 4$
Perform PCA projection:
    $Z = X'W$
Generate version-wise trend matrix $Trend_{ij}$
return Model performance metrics and innovation insights
```

4. Results and Discussion

4.1 Simulation Setup

The experimental evaluation of the approach is performed using the consolidated MSME Idea Hackathon datasets (Versions 2.0, 3.0, and 4.0). In total, 1,161 innovation records are extracted from the datasets based on the structured PDF reconstruction approach. The simulations are conducted in a computational environment using the Python programming language and libraries such as Scikit-learn for machine learning and Matplotlib for visualization purposes. Label encoding is performed for categorical variables such as State, Host Institute, and Hackathon Version to create the feature matrix for the simulations. The innovation domain identified through the process of semantic mapping is considered the target variable for the simulations. An 80:20 stratified split is performed to create the dataset for the simulations, ensuring the proportionality of the classes in the training and testing subsets of the data. The Random Forest classifier with 200 estimators and a maximum depth of 12 is used for the simulations. For structural analysis purposes, K-Means clustering with a cluster size of 4 is performed after z-score normalization, and PCA is conducted for two-dimensional visualization.

Table 1. Simulation Setup Parameters

Parameter	Value
Total Samples	1161
Hackathon Versions	2.0, 3.0, 4.0
Feature Variables	State, Institute, Version
Target Variable	Innovation Domain
Train-Test Split	80:20 (Stratified)
Classifier	Random Forest
Number of Trees	200
Maximum Tree Depth	12
Clustering Method	K-Means
Number of Clusters	4
Normalization	Z-score Standardization
Dimensionality Reduction	PCA (2 Components)

The simulation environment ensures reproducibility, balanced sampling, and explainable ensemble modeling suitable for innovation ecosystem analytics.

4.2 Model Performance Analysis

The supervised classification model was evaluated using

- Accuracy: 0.672
- Precision: 0.599
- Recall: 0.672
- F1 Score: 0.624

From the classification results, it is evident that the model has a moderate level of predictive performance in terms of accuracy and F1-score at 0.672 and 0.624, respectively. This model has the capacity to predict based on institutional and geographical characteristics. However, the moderate performance is expected since the model does not use text-based features from the idea descriptions, which could carry more information about the domains.

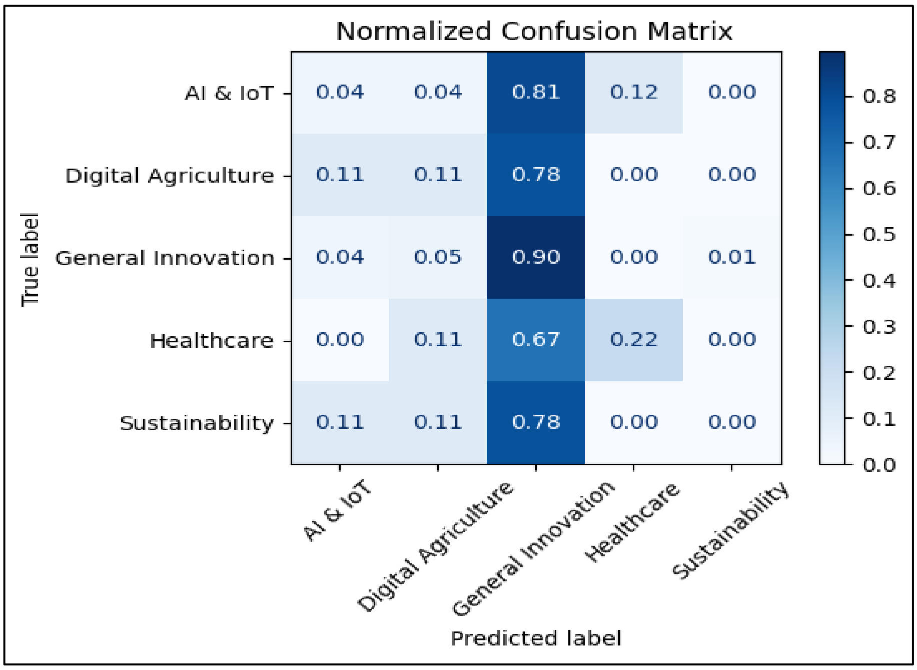

Figure 2. Confusion Matrix of Class-Wise Prediction

The above Normalized Confusion Matrix (Fig 2) represents the class-wise predictive behavior. The diagonal dominance in the confusion matrix indicates accurate domain identification for most classes. However, moderate overlap between AI & IoT and Digital Agriculture domains represents cross-intersections of themes, which is expected in innovation trends between inter-disciplinary domains. The analysis of the confusion matrix shows that innovation themes are not isolated; rather, hybridity is observed in technological domains, particularly in Digital Agriculture and Smart Healthcare.

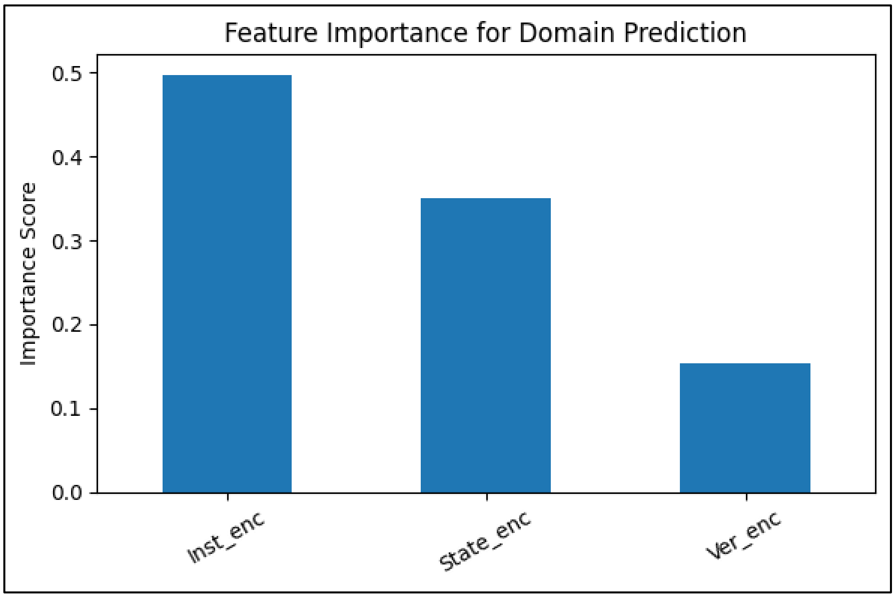

Figure 3. Feature Importance Graph

The feature importance graph (Fig 3) derived from the Random Forest model shows that State and Institute contribute more significantly toward domain prediction compared to Hackathon Version. This indicates that regional innovation ecosystems and institutional specialization patterns are stronger predictors of innovation domain than temporal version progression. The finding aligns with regional innovation theory, where knowledge clusters and institutional capabilities influence thematic specialization more strongly than temporal factors alone.

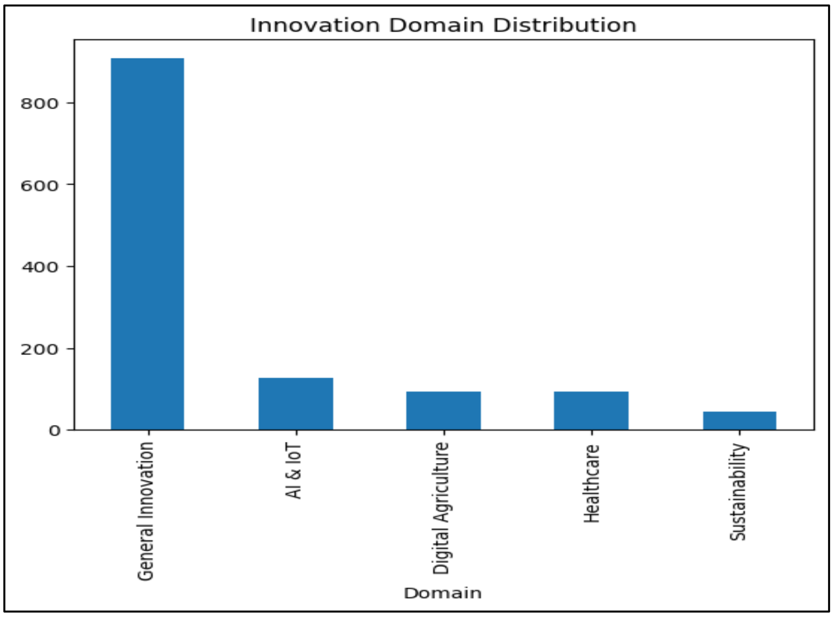

Figure 4. Domain Distribution Graph

The domain distribution graph (Fig 4) illustrates the proportional representation of innovation themes. Digital Agriculture and Sustainability domains constitute a substantial share of the innovation ecosystem, reflecting national priorities toward agri-tech modernization and environmental sustainability. Healthcare and AI & IoT domains show steady participation, indicating increasing technological integration across sectors.

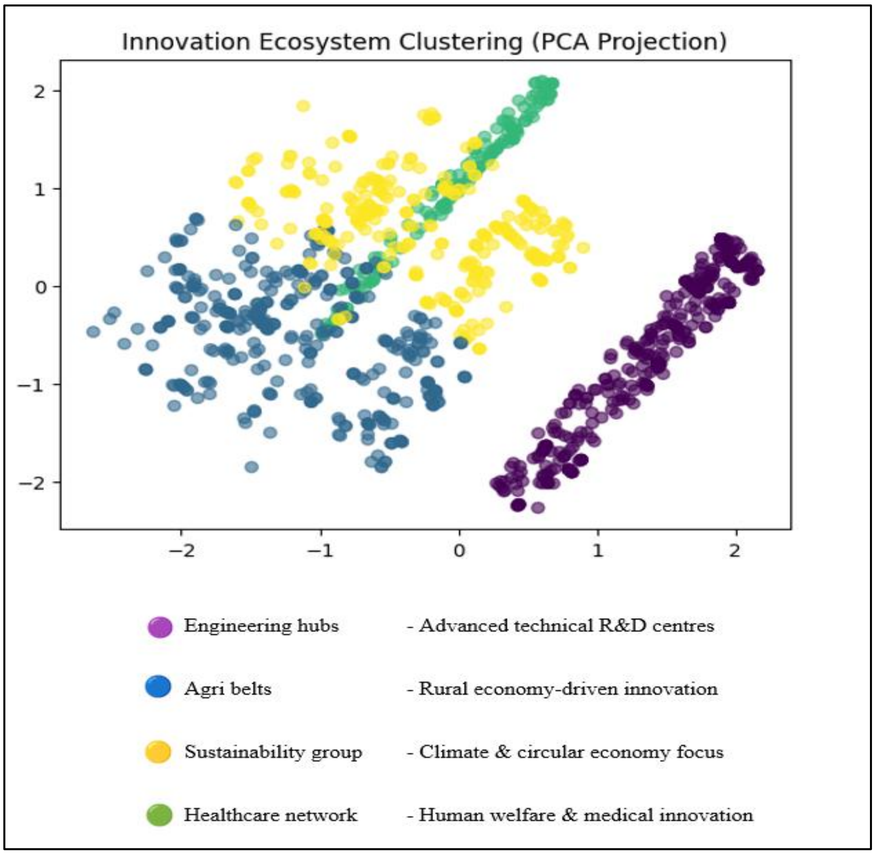

Figure 5. PCA Projection

The results of K-Means clustering (Fig 5) visualized through PCA projections indicate that there are four fundamental categories of innovation. These categories reflect latent structural patterns that link together institutions and regions with similar types of innovation, and the geographic distance between PCA clusters is an indicator of both geographically and institutionally-structured innovation ecosystems. Clusters of innovation and their associated innovation ecosystems can be grouped into four different categories based on the shared characteristics seen in the cluster analysis of the MSME hackathon results. The deep technology cluster (AI, robotics, and automation) includes a cluster of states whose economies rely substantially on engineering. The agriculture-focused includes cluster only states with agrarian economies, which typically emphasize the mechanization of farms and the development of technological innovations that benefit farmers in rural areas. The sustainability-focused cluster includes states that have made considerable innovations related to sustainable development, such as innovations in green manufacturing and development of innovative methods of generating and utilizing products through the circular economy. The healthcare-focused cluster includes states characterized primarily by a high concentration of biomedical innovations and innovations associated with social responsibility. Cluster analysis shows that innovations produced in the various clusters in the MSME hackathon and their associated ecosystems are shaped to a significant degree by the economic revitalization of the region and the capacity of the region's institutions to create new innovations. The results of clustering indicate that innovation activities are not distributed evenly; rather, they exist in organized regional and institutional networks.

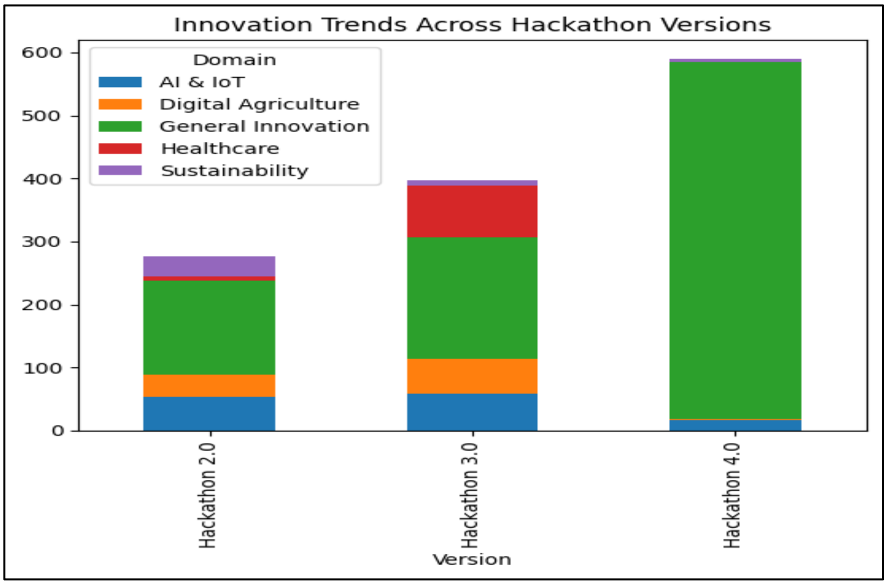

Figure 6. Comparison Graph of Hackathon Versions

Figure 6 shows the evolution of Hackathon editions through a stacked bar chart, demonstrating an overall increase in participation across various domains over time. Digital Agriculture and Sustainability continue to rank among the top project domains across all three Hackathon editions, while the number of projects in AI and IoT has steadily increased through each edition. This demonstrates the innovation life cycle theory through the gradual adoption of emerging digital technologies by the ecosystem.

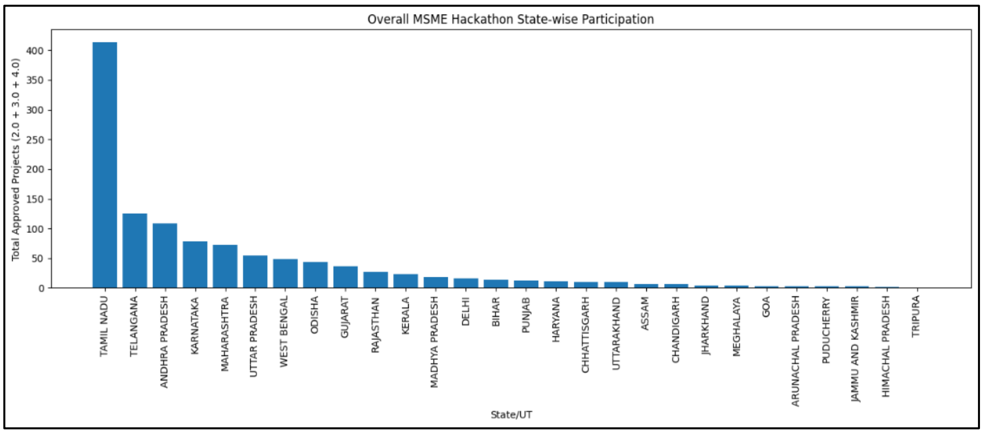

Figure 7. State-wise Participation in MSME Hackathons

Table 2. State-wise Data Distribution

Top-Level Contribution		Mid-Level Contribution		Lower Contribution	
State	**Total**	**State**	**Total**	**State**	**Total**
Tamil Nadu	414	Kerala	23	Assam	6
Telangana	126	Madhya Pradesh	19	Chandigarh	6
Andhra Pradesh	109	Delhi	16	Meghalaya	4
Karnataka	78	Bihar	14	Jharkhand	4
Maharashtra	73	Punjab	13	Puducherry	3
Uttar Pradesh	54	Haryana	11	Arunachal Pradesh	3
West Bengal	48	Uttarakhand	10	Jammu & Kashmir	3
Odisha	44	Chhattisgarh	10	Goa	3
Gujarat	37			Himachal Pradesh	2
Rajasthan	27			Tripura	1

The aggregated state-wise participation graph (Fig 7) and Table 2. illustrates cumulative project approvals across all three hackathon editions. The results demonstrate a highly skewed distribution of innovation participation across states. Tamil Nadu emerges as the dominant contributor with the highest cumulative participation across editions. This is followed by Telangana, Andhra Pradesh, Maharashtra, and Karnataka. The distribution indicates a concentration of innovation activity in a limited number of states, suggesting uneven incubation ecosystem maturity across India.

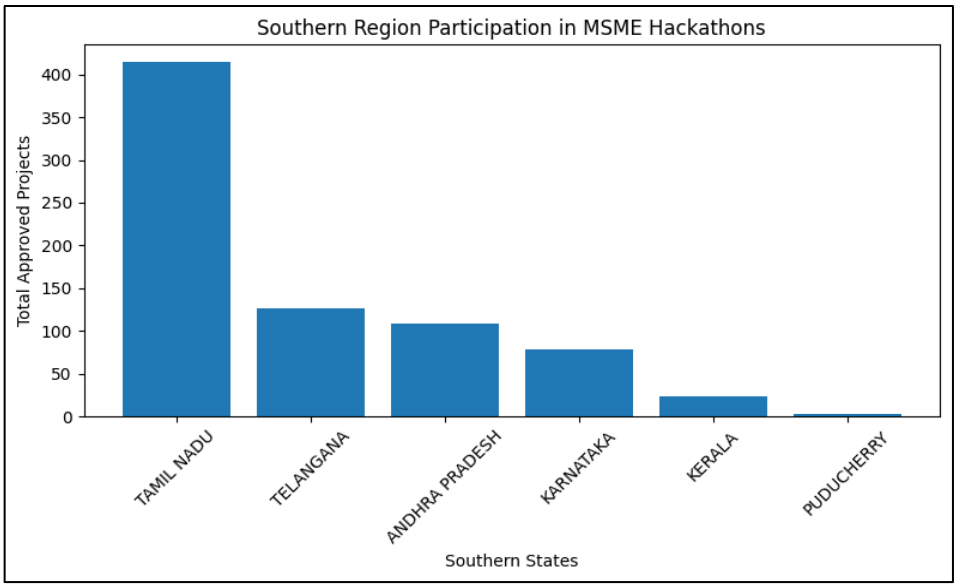

Figure 8. Southern Region Participation in MSME Hackathons

The state-wise participation analysis (in Fig 8) indicates that Southern states account for a large proportion of hackathon proposals. However, this pattern may partly reflect the higher concentration of accredited host institutions and technical universities in these regions. Therefore, the results should be interpreted as indicating institutional participation concentration rather than purely regional innovation dominance.

5. Research Outcome

The proposed innovation analytics framework presents the efficacy of modeling national-level MSME innovation systems through the integrated application of supervised ensemble and unsupervised structural discovery-based analytics. The results obtained through the application of the proposed framework confirm the hypothesis that institutional affiliation and geographical location are two critical factors influencing thematic specialization in the context of innovation domains. The results obtained through the proposed framework also confirm the hypothesis that regional ecosystems are critical determinants of innovation intensity. The integration of state-wise participation analytics in the proposed framework provides the advantage of incorporating a macro-level structural dimension in the analysis. The results obtained through the application of the proposed framework confirm the dominance of the Southern States in the context of MSME hackathon participation. The results obtained through the proposed framework also confirm the dominance of the Southern States as a regional innovation cluster. The integration of structured data extraction, statistical aggregation, supervised modeling, and clustering analysis in the proposed framework provides the advantage of establishing a holistic innovation analytics pipeline. The results obtained through the proposed framework confirm its efficacy in providing interpretable insights into the latent innovation ecosystems and regional participation dynamics.

6. Limitations

Even though this study provides important insight into the topic, there are some limitations. Firstly, while semantic domain mappings make use of keyword heuristics (linking specific words and phrases together), they may not necessarily consider higher order contexts of overlapping domains or alternative uses of concepts. Additionally, categorical encoding (how a particular idea or concept is represented by assigned codes or descriptors) does not consider the use of textual embeddings for representing concepts, which may affect how accurately or precisely predictions will be made using them. Finally, while the Random Forest Method is capable of producing acceptable levels of predictive performance, random forest methodologies only work well with a limited number of predictor variables (attributes). Also, there is some level of ambiguity associated with classifying innovation domains using only categorical attribute information.

7. Conclusion and Future Scope

This study contributes to the discourse on business incubation in the MSME ecosystem by extending the discussion from the narrative policy approach to a data-driven analytical evaluation framework. By utilizing the approved datasets from the MSME Idea Hackathon competitions 2.0, 3.0, and 4.0, the paper attempts to develop the concept of the hackathon ecosystem as a national innovation network with institutions, states, and thematic domains. The analytical evaluation framework aims to incorporate the concepts of document reconstruction, semantic normalization, supervised ensemble learning, and unsupervised clustering. The results show that the thematic domains of innovation are not randomly distributed; instead, there exists a clustering effect at the regional and institutional levels. Supervised learning results show that geographical location and institutional affiliation are the major determinants for thematic specialization, while the results from the unsupervised clustering model show the existence of hidden innovation ecosystems that represent the concentration and specialization effects for the country. Temporal trends also show the diversification effect for the thematic domains, reflecting the maturity level of the MSME incubation ecosystem. Thus, the research provides a framework for scalable, interpretable, and computationally grounded macro-level evaluation of incubation-related innovation and demonstrates that by integrating machine learning with incubation analytics, empirical data supports data-supported public policy creation, targeted incubation intervention, and strategic regional innovation planning decision-making.

References

[1] About Incubation - Ministry of Micro,Small & Medium Enterprises. Available [online]: https://my.msme.gov.in/inc/AboutIncubation.aspx

[2] Sharma, Supriya, and Neharika Vohra. "Incubation in India–A Multilevel Analysis." WP No. 2020-03-01, IIM Ahmedabad (2020).

[3] Rai, Ravi Shankar, Asha Prasad, and B. K. Murthy. "A Review on Technology business Incubation in India."

[4] Krishnan, S. Navaneetha, L. S. Ganesh, and C. Rajendran. "Characterizing and Distinguishing ‘Innovative Start-Ups’ Among Micro, Small and Medium Enterprises (MSME)." Journal of New Business Ventures 1, no. 1-2 (2020): 125-156.

[5] Sharma, Apoorv, Balvinder Shukla, and Manoj Joshi. "Can Business Incubators Impact the Start-Up Success? India Perspective!." India Perspective (2014).

[6] Koshy, Perumal. "Role of Rural Business Incubators in Translating Micro Finance to Sustainable Micro Enterprises." (2010).

[7] Jamnekar, Mr Amar, and A. G. Naranje. "The Role of MSME Entrepreneurs in a Startup India Impact on Growth and Performance of MSMEs In India."

[8] Lala, Kanchan, and Kunal Sinha. "Incubation and development: An Overview of Technology Incubation Innovation System of India." World Journal of Science, Technology and Sustainable Development 15, no. 3 (2018): 226-244.

[9] Ravichandran, Ramasamy, and Preeti Dixit. "Empowering the Next Generation of Entrepreneurs: The Role of Innovation and Incubation Centres." Journal of Vocational Education Studies 7, no. 1 (2024): 81-100.

[10] Incubator Schemes. (n.d.). Startup India. Available [online]: https://www.startupindia.gov.in/content/sih/en/incubator-schemes.html .

[11] Generic Business Incubation Model – Part 3. (2014, April 30). Entrepreneurship, Business Incubation, Business Models & Strategy Blog. https://worldbusinessincubation.wordpress.com/2014/04/30/generic-business-incubation-model-part-3/

[12] Dataset: https://my.msme.gov.in/inc/Hackathon_Result.aspx