

Amity University Online, Noida, Uttar Pradesh, India

In partial fulfilment of the requirements for the award of the degree

**Masters of Business Administration – Data Science**

**Title:** InsightNation - Government Data Analytics Platform for Citizen Opinion and Public Service Enhancement

**Guide Det:**

**Name:** Vasanthi Chandran

**Designation:**

**Submitted By:**

**Student Name:** PRANOY CHAKRABORTY

**Enrolment No:** A9920123006194

**Course Name:** Dissertation (MSDS600)

**Date:**

ANNEXURE B

**DECLARATION**

I, **Pranoy Chakraborty**, a student pursuing **MBA, Semester 4 (Specialization: Data Science)** at **Amity University Online**, hereby declare that the project work entitled **“InsightNation – Government Data Analytics Platform for Citizen Opinion and Public Service Enhancement”** has been prepared by me during the academic year **2023-2025** under the guidance of **Ms. Vasanthi Chandran, Project Guide from Qollabb**. I assert that this project is a piece of original bona fide work done by me. It is the outcome of my own effort, and it has not been submitted to any other university for the award of any degree.

Name and signature of the student

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

PRANOY CHAKRABORTY

**PLAGARISM REPORT**

This is to certify that I, **Pranoy Chakraborty**, enrolled in the 4th semester of the degree program “Master of Business Administration”, and undertaking the course by the title “Dissertation (MSDS600)”, for the third semester in the academic session of July’ 2023, have submitted this report under strict compliance of the guidelines specified by Amity University by keeping the percentage of plagiarism below the permissible limits.

This plagiarism in this report has been checked using the tool “Dupli Checker” and it came out to be 100%.

**ACKNOWLEDGEMENT**

I would like to convey my profound gratitude to **Ms. Vasanthi Chandran,** my professor and supervisor, for her invaluable guidance, mentorship, and steadfast support throughout this project. Her expertise and encouragement have been instrumental in enhancing my understanding of customer churn dynamics and the application of data analytical techniques.

I am also indebted for her astute advice, assistance, and generous dissemination of knowledge. Her guidance and motivation have empowered me to engage in rigorous research, address complex data challenges independently, and navigate intricate machine learning methodologies with confidence. Additionally, her moral support has been a significant source of strength throughout this endeavour.

Finally, I extend my heartfelt appreciation to all individuals who have contributed directly or indirectly to this project. Your support and encouragement have been invaluable, and I am deeply appreciative of the collective effort that has facilitated this undertaking.

**ABSTRACT**

In the era of data-driven decision-making, the need for responsive governance and citizen-centric public service delivery has become more critical than ever. Traditional approaches to understanding public sentiment and service satisfaction often rely on slow, manual surveys or narrowly scoped feedback loops, which limit the scope and accuracy of actionable insights. As societies continue to urbanize and digitalize, there is a growing need for governments and civic agencies to adopt more scalable, intelligent, and adaptive methods for interpreting citizen feedback and improving services in real time. In response to this challenge, the current project introduces InsightNation – a robust, AI- and ML-powered analytics platform designed to bridge the gap between public opinion and smarter public service enhancement.

This dissertation project, ‘InsightNation – Government Data Analytics Platform for Citizen Opinion and Public Service Enhancement’ serves as a data analytics platform that ingests, processes, analyzes, and visualizes multi-dimensional feedback from citizens across a wide spectrum of public service categories such as sanitation, transportation, parks and recreation, library services, and safety. The system employs modern techniques in natural language processing (NLP), statistical analytics, and supervised machine learning to mine actionable insights from structured survey data. It transforms raw citizen input into meaningful dashboards, predictive models, and strategic recommendations for government stakeholders, municipal planners, and civic organizations. By offering real-time visibility into what citizens are experiencing and expecting, the platform seeks to assist decision-makers in identifying gaps, measuring satisfaction, and forecasting future needs.

The project's architecture is designed for extensibility and scalability, allowing for flexible growth and adaptation. The backend pipeline is powered by Python and pandas for data wrangling, scikit-learn and SpaCy for machine learning and NLP, and Matplotlib/Plotly for visualization. The frontend is developed using Streamlit, allowing users to interact with the system through a clean, intuitive dashboard that supports file uploads, dynamic charts, chatbot-style Q&A, and visual summaries of citizen sentiment. Data input primarily consists of cleaned and structured CSV survey data collected from diverse urban populations, comprising multiple demographic segments and service categories. The dataset used for this project includes over 5,000 citizen records, each with detailed service-level feedback and open-text suggestions.

One of the core innovations of the InsightNation platform lies in its ability to apply sentiment classification to open-ended citizen responses using advanced NLP pipelines. After pre-processing textual feedback with SpaCy (including tokenization, lemmatization, stopword removal, and named entity recognition), the platform uses machine learning models such as Logistic Regression and Support Vector Machines (SVM) to classify sentiments into positive, negative, or neutral categories. These classifications are further aggregated and visualized to identify trends by city, age group, gender, or service type. The system’s learning pipeline is designed to be extensible to other models, including BERT or LSTM-based architectures, to improve classification accuracy in future iterations.

In addition to traditional charts and model outputs, InsightNation integrates conversational AI through Google’s Gemini LLM (via the Gemini API), enabling natural language interaction with the analytics platform. Users can ask contextual questions about trends, seek strategy advice, or request summaries of findings in plain English. This feature empowers non-technical stakeholders, such as municipal leaders or citizen engagement officers, to access AI-generated insights without needing to understand the underlying data science models. Moreover, this conversational layer includes tools for SWOT analysis, business-like recommendations, and memory-based Q&A to simulate expert consultants.

To ensure robust usability and modular growth, the platform is divided into distinct functional phases: dataset upload and cleaning, exploratory data analysis (EDA), NLP preprocessing, ML modeling, data visualization, and AI-powered advisory modules. Each phase is linked to an intuitive tab in the Streamlit interface and is supported by Python scripts organized in a standardized folder structure, ensuring clean codebase management and future scalability.

From a project management standpoint, InsightNation was developed over 12 structured weeks, adhering to an agile methodology with iterative development, testing, and refinement. Weekly milestones covered problem identification, system architecture, model experimentation, UI/UX design, performance validation, and final integration. The deliverables include a fully functional Streamlit-based analytics platform, trained ML/NLP models, custom visualization assets, and a detailed project report documenting methodology, results, findings, and strategic implications.

The outcomes of this project demonstrate the power and necessity of integrating AI and citizen feedback to improve public services. Through machine learning and interactive dashboards, decision-makers can now pinpoint areas of concern, recognize regional disparities, and deploy targeted interventions with data backing. Furthermore, the use of NLP ensures that even qualitative suggestions—often ignored in traditional feedback pipelines—are now incorporated into performance reviews and planning strategies. Ultimately, this results in a more participatory governance model where citizens feel heard and empowered, and where governments respond faster and more precisely to evolving public needs.

In conclusion, InsightNation redefines how governments and civic agencies can listen to and act upon public opinion using modern data science tools. It lays the groundwork for scalable public service intelligence that goes beyond static survey reports, offering a continuous, AI-augmented decision-making loop. The successful implementation of this platform sets a strong precedent for replicating this model across regions, departments, and even entire nations.

Future Scaling and Expansion: Looking ahead, InsightNation can be scaled to integrate real-time feedback channels such as mobile apps or social media APIs, allowing for live citizen sentiment tracking. Additionally, advanced AI integrations such as GPT-based summarization, multilingual feedback parsing, and smart alert systems for anomaly detection can enhance the platform’s utility in larger, more complex public service ecosystems.

Keywords: Public Service Analytics, Citizen Feedback, Data Science, Machine Learning, Natural Language Processing (NLP), Sentiment Analysis, Streamlit Dashboard, AI-Powered Governance, Google Gemini API, Civic Engagement, Public Satisfaction, Predictive Analytics.

# **CHAPTER 1 – INTRODUCTION**

In today’s era of digital transformation and data-driven governance, citizen feedback is no longer a passive form of communication—it has emerged as a powerful instrument to enhance the efficiency, transparency, and responsiveness of public service delivery. Government agencies, municipal bodies, and public institutions are increasingly recognizing the value of listening to the voice of the citizen, both as a metric of satisfaction and as a compass for strategic improvements. Against this backdrop, the need for structured, intelligent, and scalable analytics platforms that can process, analyze, and derive actionable insights from public opinion has become more critical than ever.

The capstone project titled **“InsightNation – Government Data Analytics Platform for Citizen Opinion and Public Service Enhancement”** is conceived as a strategic and technological response to the increasingly urgent need for responsive, data-informed governance. As societies grow more urbanized and citizens demand higher standards of public service delivery, it becomes critical for governing bodies to not only listen to feedback but to systematize its collection, processing, and analysis. This project addresses that precise challenge by building a platform that empowers public administrators, planners, and policymakers with data-driven insights derived directly from citizens’ lived experiences and service interactions.

At the heart of the InsightNation platform lies the recognition that citizen feedback is not just an afterthought or a box to be checked—it is a powerful diagnostic tool that can inform resource allocation, uncover systemic inefficiencies, and highlight areas of improvement in real time. The project is therefore anchored in the belief that **data is a dialogue**, and that turning qualitative and quantitative feedback into structured, actionable intelligence is essential for making public service delivery more efficient, inclusive, and accountable.

The scope of the platform encompasses feedback from multiple touchpoints within urban public services—including **sanitation facilities, public parks, transport infrastructure, library access, and local governance mechanisms**. These areas were deliberately chosen because they represent core dimensions of urban livability and are commonly encountered by a wide spectrum of citizens across age groups, genders, and geographies. By analyzing feedback across these domains, the project ensures that insights generated are both **comprehensive and multi-faceted**, reflecting the complex realities of public life.

Technically, the project integrates **data engineering, statistical analysis, machine learning (ML), and natural language processing (NLP)** into a seamless pipeline that takes raw citizen feedback—often fragmented and unstructured—and transforms it into clean, interpretable formats. These are then analyzed to detect patterns, identify satisfaction gaps, and prioritize areas needing urgent attention. The platform also enables **demographic-level segmentation**, allowing public agencies to customize their interventions based on city, age group, or gender, thereby aligning service delivery with the actual needs of their constituents.

Importantly, this project is not simply about creating dashboards or performing one-off analytics. Rather, it demonstrates the viability of a **scalable, modular, and policy-aligned architecture** that can be deployed by municipalities or civil organizations seeking to embed analytics into their feedback loops. It is a **proof-of-concept** that public feedback analytics can go beyond surveys and summaries, evolving into an engine for civic intelligence and smarter decision-making. Ultimately, InsightNation aims to bridge the long-standing gap between **what people say and what governments do**—not through guesswork or assumptions, but through **data, analysis, and insight**.

Public institutions across the world have historically struggled with inefficiencies, bureaucratic bottlenecks, and outdated feedback mechanisms. Paper-based surveys, occasional community meetings, and static suggestion boxes are often inadequate in reflecting the dynamic needs and grievances of a digitally connected and increasingly aware populace. While some developed nations have adopted e-governance portals and smart feedback collection systems, many regions—including several urban and semi-urban areas in developing countries—still lack real-time, analytics-backed systems for tracking service performance. This project aims to fill that void by building an intelligent data analytics platform tailored to the nuances of public service interactions.

**Contextual Relevance and Problem Background**

The conventional model of public service feedback collection is often fragmented, delayed, and unstructured. Citizens may provide feedback in multiple formats—verbal complaints, online reviews, social media posts, or structured forms—but public agencies typically lack the infrastructure to integrate and analyze these inputs holistically. Moreover, the absence of sentiment classification, thematic grouping, and performance dashboards makes it difficult for public officials to prioritize actions or track improvement over time.

With the proliferation of smart cities and digital citizenship initiatives, the time is ripe to introduce AI-powered platforms that bring structure to the chaos of public opinion. A data analytics-driven feedback platform holds the promise of quantifying subjective experiences, identifying recurring pain points, and highlighting regional variations in service delivery quality. For example, if transport services are routinely flagged as unsafe by female commuters in a particular city zone, such signals can help civic authorities deploy gender-sensitive policy interventions more effectively. Similarly, poor satisfaction scores for library access or cleanliness issues in public parks can be traced and resolved proactively if detected early through systematic analytics.

In this context, the InsightNation platform presents a transformative approach to civic management. By combining structured survey data, Natural Language Processing (NLP) for free-text suggestions, and visual analytics through dashboards, the platform equips administrators with a 360-degree view of public sentiment. Not only does this reduce reliance on intuition and anecdotal evidence, but it also enables a culture of evidence-based governance.

**Justification of Selecting the topic:**

The selection of this project topic is grounded in its societal impact, analytical complexity, and technological relevance. As an MBA student specializing in Data Science, the intersection of civic engagement, AI-driven insights, and service optimization offers a rich, multi-dimensional problem space that aligns well with academic objectives and real-world applicability. Moreover, this topic provides an opportunity to explore various facets of data science—data wrangling, statistical analysis, machine learning, NLP, and dashboarding—within the context of a high-stakes, socially beneficial domain.

From a societal standpoint, public dissatisfaction with services—be it poor sanitation, unreliable transportation, or underutilized civic amenities—has tangible repercussions. It not only erodes trust in public institutions but also hampers the quality of life in communities. By building a platform that can intelligently harness and analyze citizen sentiment, this project contributes meaningfully to solving a chronic problem that affects millions of people, particularly in urbanizing regions.

Technologically, the project aligns with current industry trends such as smart governance, civic tech innovation, and AI for social good. Leading global cities are investing heavily in platforms that can automate public feedback processing using data science. For instance, New York City’s 311 service, London’s open data portal, and Singapore’s Smart Nation initiative all represent institutional efforts to embrace data-led decision-making. This project aspires to bring similar capabilities within the reach of local municipalities and citizen engagement programs in developing nations, using open-source tools and scalable machine learning models.

Furthermore, the topic offers an academic advantage in that it encapsulates multiple modules from the MBA Data Science curriculum. The project involves data preprocessing, feature engineering, supervised learning (for sentiment classification), and unsupervised techniques (for clustering or topic modeling, where applicable). It also includes dashboard development using Streamlit, making it suitable for real-time data visualization and executive-level decision support.

In selecting this topic, the goal was not just to complete a project for academic fulfillment but to prototype a potentially scalable solution that could be deployed in real-life urban or civic contexts. The modular design and open-source architecture ensure that the platform can be extended to integrate with mobile apps, voice-based feedback systems, or even multilingual NLP models in the future. The implications of this work stretch beyond technical execution—they touch upon policy innovation, public-private partnership models, and the democratization of data access in governance.

**Current Landscape and Research Gaps:**

In recent years, the idea of citizen-centric governance has gained significant momentum globally, driven by a growing recognition that public participation is essential for building responsive and accountable government systems. Governments at national, state, and municipal levels are increasingly turning to digital platforms to solicit feedback from citizens on a wide range of public services. Whether it's about sanitation facilities, public transport, urban green spaces, safety in neighbourhoods, or the efficiency of local libraries, the demand for real-time citizen insights is escalating. However, despite this growing interest, the actual implementation and utilization of comprehensive feedback analytics platforms remain largely limited and underdeveloped.

Currently, most digital governance initiatives focus primarily on front-end engagement—developing mobile applications, web portals, or survey systems that allow citizens to register complaints, give ratings, or submit suggestions. While this infrastructure is critical, it addresses only the initial stage of the data lifecycle. What follows—namely, backend analytics, intelligent processing, and actionable insight generation—is often either absent or implemented in a rudimentary manner. These systems are frequently restricted to basic summary statistics or manual review processes, limiting their scalability and impact.

Even where basic analytics exist, they are often narrow in scope. For example, sentiment analysis, if employed at all, tends to be binary or polarity-based (positive vs. negative), without considering the nuanced themes, domain-specific terminology, or contextual variations embedded in feedback. Many models also fail to consider how feedback might vary based on demographics like age group, gender, or geographic location, which are crucial for equity-focused public service delivery. Furthermore, despite the diversity of service domains covered—such as sanitation, safety, transport, libraries, and public parks—there is little effort to develop multi-domain feedback models that reflect the complexity of citizen experiences.

Additionally, multilingualism in countries like India poses a significant challenge to NLP-based public opinion analytics. Many local feedback platforms struggle to effectively process regional languages or dialects, leading to the exclusion of non-English speakers from data-driven decision-making processes. Even when translations are performed, semantic context and emotional tone are often lost, leading to erroneous interpretations. This highlights a critical technological gap in natural language understanding tailored to culturally and linguistically diverse populations.

Data privacy and governance concerns also limit the deployment of large-scale feedback platforms. Many local government bodies lack the infrastructure or policies necessary to ensure secure data handling, leading to public distrust in feedback collection mechanisms. As a result, adoption rates remain low, and data quality suffers due to limited participation or superficial responses. Moreover, the lack of standardization in data formats and storage protocols across different municipal systems makes inter-agency data integration a significant bottleneck.

From a research standpoint, the academic and industry literature on public feedback analytics remains relatively sparse when compared to other domains like e-commerce, healthcare, or financial services, where customer sentiment and behavioral data have long been mined for strategic advantage. Numerous studies have developed sophisticated NLP models to analyze product reviews, patient feedback, or investment sentiment. However, analogous efforts in the public service domain—especially those that cut across multiple services and citizen attributes—are few and far between.

Even within existing civic research, many studies tend to be issue-specific (e.g., transport planning or sanitation improvement), failing to adopt a holistic, cross-service analytical approach. Moreover, while there is considerable literature on service delivery models and governance metrics, the incorporation of real-time citizen feedback into these models remains mostly theoretical. There is a lack of operational platforms that bridge academic research with practical implementation in this area.

This project, InsightNation, is positioned to address these deficiencies by creating an integrated analytics platform that is capable of ingesting, preprocessing, analyzing, and visualizing citizen feedback across multiple service domains. It does not merely aim to summarize responses but to derive patterns, trends, and insights that are both actionable and policy-relevant. The project leverages modern data science methodologies—including machine learning (ML), natural language processing (NLP), and interactive visual analytics—to build a pipeline that simulates how government agencies might meaningfully interpret large-scale public feedback.

The dataset used in the project comprises over 4,000 records, collected through structured citizen surveys spanning a diverse set of demographic attributes and public service areas. Each record includes responses about sanitation cleanliness, safety perception, transport satisfaction, library usage, and park facilities, among others. It also captures metadata such as city, gender, age group, and service usage frequency, enabling multi-dimensional slicing of the feedback.

Such a dataset offers a rich substrate for not just exploratory data analysis (EDA) and descriptive statistics, but also for advanced modeling and hypothesis testing. For example, the project investigates whether certain cities are underperforming in specific domains (e.g., library services), whether satisfaction levels correlate with age or gender, or what themes emerge in open-ended suggestions for public services. These questions are not just academically intriguing; they are deeply consequential from a public policy and urban governance perspective.

By embedding these questions into the design of its analytics workflows, the InsightNation platform seeks to provide a proof-of-concept for how public agencies can move from passive data collection to active, insight-driven decision-making. This approach not only bridges the gap between data and action but also contributes to the emerging field of digital public service innovation, where feedback loops between citizens and institutions are both real-time and data-informed.

In summary, the current landscape of public feedback analytics is fragmented and underdeveloped, particularly in its backend intelligence and multi-domain modeling capabilities. The research gaps in this space—ranging from sentiment analysis in public services to demographic-driven satisfaction modeling—are vast but addressable. InsightNation’s goal is to contribute a tangible step forward in closing these gaps by demonstrating the potential of a citizen feedback analytics platform that is technologically robust, ethically grounded, and policy-aligned.

**Real-World Implications**

In practice, platforms like InsightNation can serve multiple stakeholders. For municipal governments, it offers a tool to diagnose service gaps and respond to citizen needs in real-time. For non-governmental organizations (NGOs), it can serve as an advocacy tool to highlight underserved communities or services. For researchers and data journalists, it offers a mine of structured public sentiment data that can inform articles, studies, and investigations.

At a time when trust in public institutions is fragile, and citizen expectations are rapidly evolving, a data-backed feedback loop can serve as the foundation for collaborative governance. By involving citizens not just as complainants but as co-creators of urban experiences, platforms like InsightNation shift the paradigm from reactive to proactive administration.

Additionally, this approach has significant potential for scalability. With minor adjustments, the core architecture can be repurposed for use in education (student feedback), healthcare (patient satisfaction), or even electoral systems (voter sentiment). In an age where data is the new oil, civic data—properly refined and utilized—can be the fuel that powers smarter cities and more empathetic governance.

# **CHAPTER 2 - OBJECTIVE OF THE STUDY**

The primary objective of this study is to develop a robust customer churn prediction model within the telecommunications sector by leveraging exploratory data analysis (EDA) and machine learning algorithms. Using the Telco Customer Churn Dataset, the study aims to analyze customer behaviour patterns, identifying the key factors that contribute to customer attrition. Logistic Regression will be employed as the baseline predictive model, and its performance will be compared with other machine learning algorithms such as Support Vector Classifier (SVC), Random Forest, Decision Tree, and Naive Bayes classifiers to determine the most accurate model for predicting churn.

In addition to the analysis using the Telco dataset, the study will extend its scope to examine churn trends within the Indian telecom sector using more recent data. The goal is to provide a comprehensive view of churn behaviour and to offer practical insights into customer retention strategies that can be adopted by telecom companies. Ultimately, the research aims to contribute to better decision-making processes in customer management, helping businesses reduce churn rates, minimize revenue loss, and enhance long-term customer loyalty.

# **3. LITERATURE REVIEW**

Customer churn, defined as the percentage of customers who discontinue using a company's products or services, poses a significant challenge for businesses, particularly within the telecommunications industry. Research has demonstrated that understanding the factors contributing to customer attrition is vital for developing effective retention strategies. In this context, various studies have explored the application of machine learning (ML) techniques and data analysis methods to predict and mitigate churn.

Web Chin-Ping Wei and I-Tang Chiu (2016) proposed a churn prediction technique utilizing the C4.5 decision tree algorithm on customer call data, emphasizing the importance of understanding customer behaviours to enhance retention efforts. Their approach highlighted that predictive models could identify customers likely to churn, enabling organizations to implement targeted interventions. Similarly, Yi-Fan Wang et al. (2018) introduced a recommender system that also employed decision tree algorithms to predict churn, analyzing over 60,000 transactions. Their findings underscored the effectiveness of decision trees in handling large datasets, providing actionable insights for customer management strategies.

Moreover, Jadhav and Pawar (2019) designed a decision support system that utilized backpropagation algorithms on customer billing data to forecast churn behaviour. Their study illustrated the potential of neural network approaches in achieving high accuracy in churn predictions, reinforcing the notion that advanced machine learning techniques can significantly enhance predictive capabilities in the telecommunications sector.

In a more comprehensive analysis, Kamalraj and Malathi (2020) explored the application of various data mining techniques to better understand churn prediction. They emphasized the utility of machine learning models within the context of Customer Relationship Management (CRM), advocating for their integration into retention strategies to mitigate customer attrition effectively. This perspective aligns well with the objectives of this project, as it seeks to leverage machine learning algorithms, including logistic regression and support vector classifiers, to analyze customer churn within the Indian telecom industry.

Research by Adwan et al. (2020) further supports the use of machine learning for churn prediction, showcasing a multi-layer perceptron neural network (MLPNN) model on actual customer data from a major Jordanian telecommunications firm. Their results indicated that MLPNN could successfully predict churn, reinforcing the efficacy of neural networks in this domain. Additionally, Farhad Shaikh’s study (2021) highlighted the combination of classification and clustering techniques to rank churn clients and identify underlying reasons for their attrition, thereby facilitating tailored retention strategies.

While the majority of existing literature emphasizes the application of various ML models for churn prediction, there remains a need to focus on business implications and customer retention strategies. The analysis of churn in the context of competitive markets is crucial, as demonstrated by Ismail et al. (2022), who noted the intense rivalry among telecommunications providers and the necessity of deploying robust predictive models to stay ahead. Their findings indicated that understanding churn dynamics could directly impact an organization's competitive positioning and profitability.

The datasets utilized in these studies vary, but many have recognized the relevance of using historical data to inform churn predictions. The Telco Customer Churn dataset, prepared by IBM, serves as a significant reference point, being six years old yet still pertinent due to its comprehensive nature. Complementing this, the inclusion of more recent data from the Indian telecom sector ensures that the analysis remains relevant in the current competitive landscape.

# **4. RESEARCH METHODOLOGY**

Research methodology involves a structured approach and strategy for carrying out research. It includes the various methods, techniques, and processes used to gather, assess, and interpret data with the goal of addressing research questions or testing hypotheses. A well-defined research methodology is essential for ensuring that the research process remains objective, valid, and dependable. This section includes Data Collection Approach, Data Source and Research methods.

## **4.1 DATA COLLECTION APPROACH**

The data for this project will be gathered from secondary sources, specifically from publicly available datasets. The main data source will be the Telco Customer Churn Dataset prepared by IBM. This dataset consists of customer information such as demographics, account details, and churn status. Additionally, I will utilize recent data from the Indian Telecom Sector to provide more localized insights.

## **4.2 SOURCES USED**

*Kaggle Database:* Kaggle is a platform that provides datasets and serves as a learning and competition space for data scientists and machine learning enthusiasts.

*Telco Customer Churn Dataset (Prepared by IBM):* The Telco Customer Churn Dataset, which is publicly available on Kaggle, has been curated by IBM to help with churn analysis for telecommunication industries. This dataset contains over 7,000 customer records, including various attributes related to customer demographics, service usage, account information, and whether the customer has churned. It’s an ideal dataset for training predictive models due to its clean, well-structured nature and the variety of customer behaviour variables it captures. This study will leverage this dataset to build models that can be applied to the Indian telecom market.

*Indian Telecom Sector Data:* This research will incorporate a more recent dataset from the Indian telecom sector from Kaggle Database, representing customer behaviour and churn patterns in India. This dataset is approximately one year old, offering a more localized and current understanding of churn in the Indian telecom industry. It will be integrated to provide a comparative analysis and highlight strategies specifically tailored for the Indian market.

## **4.3 RESEARCH METHODS**

The following steps outline the methodology for analyzing and predicting customer churn:

**Data Preprocessing:** Before applying machine learning models, the datasets will undergo several preprocessing steps, such as handling missing values, normalizing variables, and encoding categorical features. Feature engineering may also be performed to create new variables that could improve model performance.

**Exploratory Data Analysis (EDA):** A thorough exploratory data analysis will be conducted to understand patterns, correlations, and trends within the dataset. EDA will help uncover key factors that contribute to customer churn, which can inform both the model-building process and business strategy recommendations.

**Predictive Modeling:**  
Various machine learning models will be applied to predict customer churn. These include:

1. Logistic Regression
2. Support Vector Classifier (SVC)
3. Random Forest Classifier
4. Decision Tree Classifier
5. Naive Bayes Classifier

## **4.4 MODEL EVALUATION AND SELECTION**

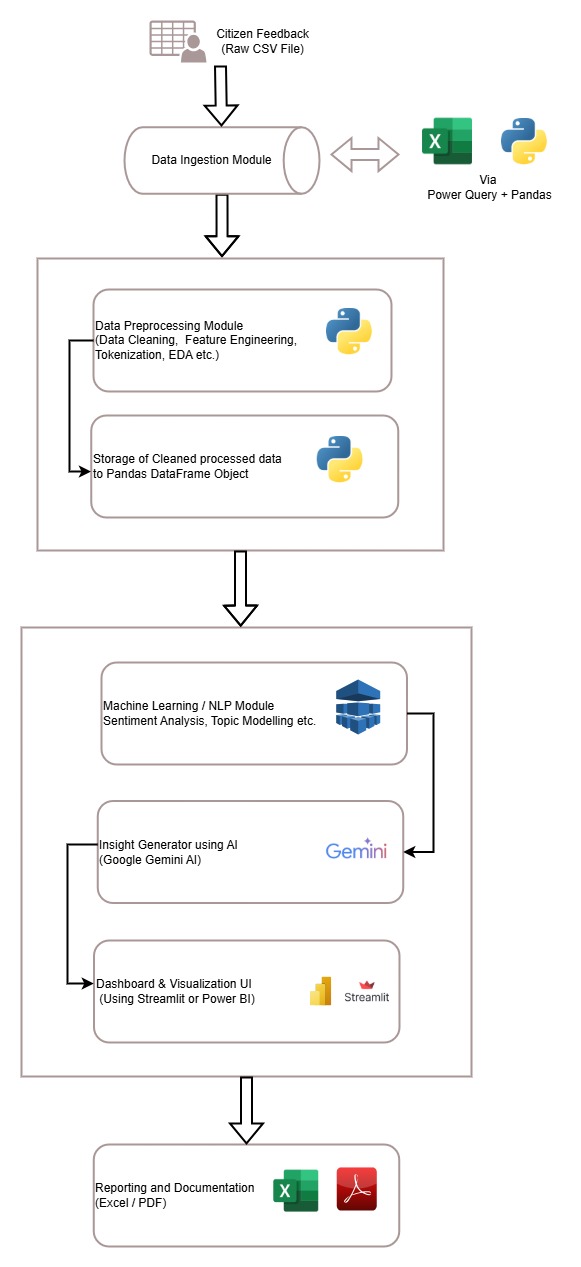
The models will be evaluated based on key metrics such as accuracy, precision, recall, and AUC-ROC scores. These metrics will help identify which model provides the best predictive accuracy and generalizes well to new data.

# **9. RESULT DISCUSSION**

In this study, five machine learning models were evaluated to predict customer churn within the Indian telecom sector: Logistic Regression, Support Vector Classifier (SVC), Random Forest, Decision Tree, and Naive Bayes. Initial data analysis showed low correlations among features, favouring models that excel with complex, non-linear patterns.

The Random Forest Classifier achieved the highest accuracy, benefiting from its ability to capture intricate customer patterns through multiple decision trees. SVC also performed well, demonstrating its strength in defining precise decision boundaries in low-correlation data. Logistic Regression offered a reliable baseline with moderate performance, while Decision Tree lagged behind Random Forest. Naive Bayes, which assumes feature independence, scored lowest due to the complex dependencies in the data.

Overall, Random Forest proved most effective for this telecom dataset, suggesting it as a strong choice for accurate churn predictions and strategic customer retention efforts.



# **10. CONCLUSION AND FUTURE SCOPE**

# **11. BIBLIOGRAPHY**

1. Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). **Predicting customer churn in telecom industry using multilayer perceptron neural networks: modeling and analysis**. *Life Sci. J.*, 11(3), 75-81. [View the Document](https://d1wqtxts1xzle7.cloudfront.net/77251766/011_22840life110314_75_81-libre.pdf?1640341187=&response-content-disposition=inline%3B+filename%3DPredicting_Customer_Churn_in_Telecom_Ind.pdf&Expires=1730321081&Signature=NHbQ6SlLj086xDFnUeB2A5qzj~pFcsPm~nH-OZj38iHrVLOZK3myV00UGGb6CfAj9f0IIeR2vkfsrP1aItQ63XnKvs~9-iXMbqk36-gxu5B149Xjo9k6Tk4volcL2wIa2n-iQ3VikOaqoZ~p4mW2JiWMsil34So3DeYUT1WWfdU2o5rtXIAqWogANkMogcP0wqz~g2Nc-nt6tyyUOOVRSwtj9RGEKGZ7VOS6K~HSey2Oo7urODaXjlJni5DJF~Sfxg1zFTAvrJxcgx41hW90ZfiNEx0urVTtXCcyZtJS17j5irdxQya5Z76xHZWu4C471W1Uwg5DLNFoMOekn9puwg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)
2. Ismail, M. R., Awang, M. K., Rahman, M. N. A., & Makhtar, M. (2015). **A multi-layer perceptron approach for customer churn prediction**. *International Journal of Multimedia and Ubiquitous Engineering*, 10(7), 213-222. [View the Document](https://gvpress.com/journals/IJMUE/vol10_no7/22.pdf)
3. Sharma, A., & Panigrahi, P. K. (2011). **A neural network-based approach for predicting customer churn in cellular network services**. *Int. J. Comput. Appl.*, 27(11), 26-31. [View the Document](https://www.ijcaonline.org/volume27/number11/pxc3874605.pdf)
4. Ahmed, A., & Linen, D. M. (2017). **A review and analysis of churn prediction methods for customer retention in telecom industries**. *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, 1-7. [View the Document](https://ieeexplore.ieee.org/abstract/document/8014605)
5. Babu, S., Ananthanarayanan, N. R., & Ramesh, V. (2016). **A study on efficiency of decision tree and multi layer perceptron to predict the customer churn in telecommunication using WEKA**. *Int. J. Comput. Appl.*, 140(4), 26-30. [View the Document](https://www.ijcaonline.org/research/volume140/number4/babu-2016-ijca-909274.pdf)
6. Kayaalp, F. (2017). **Review of customer churn analysis studies in telecommunications industry.** *Karaelmas Science & Engineering Journal*, 7(2). [View the Document](https://www.researchgate.net/profile/Fatih-Kayaalp/publication/321723500_Review_of_Customer_Churn_Analysis_Studies_in_Telecommunications_Industry/links/5a2e583945851552ae7f1b99/Review-of-Customer-Churn-Analysis-Studies-in-Telecommunications-Industry.pdf)
7. Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). **A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector.** *IEEE Access*, 7, 60134-60149. [View the Document](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8706988&utm_source=sciencedirect_contenthosting&getft_integrator=sciencedirect_contenthosting)
8. Jinde, V., & Amit, S. (2020). **Customer churn prediction system using machine learning.** International Journal of Advanced Science and Technology, 29(5), 7957-7964. [View the Document](https://link.springer.com/article/10.1007/S00607-021-00908-Y)
9. Yahaya, R., Abisoye, O. A., & Bashir, S. A. (2020). **An enhanced bank customers churn prediction model using a hybrid genetic algorithm and K-means filter and artificial neural network**. 2020 IEEE 2nd International Conference on Cyberspace (CYBER NIGERIA), IEEE, 52-58. [View the Document](https://ieeexplore.ieee.org/document/9428805)
10. Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). **A comparison of machine learning techniques for customer churn prediction**. Simulat. Model. Pract. Theor., 55, 1-9. [View the Document](https://www.sciencedirect.com/science/article/abs/pii/S1569190X15000386)