

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.

InsightNation serves as a data analytics engine 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**

Customer churn, also known as attrition, represents the rate at which customers discontinue using a service or cease purchasing products over a defined period. This metric is critical for businesses, particularly those reliant on subscription models, where retaining customers is essential to maintaining revenue streams. Churn plays a pivotal role in customer lifetime value (CLV) calculations, helping businesses forecast potential profits from ongoing customer relationships. In competitive sectors such as Software as a Service (SaaS), the availability of numerous alternatives makes it vital for companies to understand and mitigate churn.

The telecommunications sector, especially in markets like India, faces similar challenges. Telecom operators experience constant pressure to keep customers satisfied due to intense competition and the presence of new entrants offering comparable services at lower costs. Customer dissatisfaction with service quality, pricing, or alternatives can lead to higher churn rates. Predicting which customers are likely to churn and implementing strategies to retain them is crucial for maintaining profitability and market position.

Customer churn prediction is a data-driven approach that combines historical customer data with **machine learning (ML)** techniques to create predictive models. These models can forecast which customers are most likely to discontinue their service, allowing companies to take corrective actions to reduce churn. Predictive churn analysis involves the use of **Exploratory Data Analysis (EDA)** to uncover underlying patterns in customer behaviour, identifying factors that influence their decision to stay or leave.

By utilizing **machine learning algorithms** such as **Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbors**, and more advanced techniques like **XGBoost** and **Support Vector Machines (SVM)**, businesses can develop robust churn prediction models. These models are trained on historical data, learning from past customer behaviour to make future predictions with a high degree of accuracy. Each algorithm has its strengths and weaknesses, with some excelling at handling complex, non-linear relationships within the data, while others may be better suited for smaller datasets or simpler patterns.

By implementing effective churn prediction models, telecom companies can improve customer retention strategies, reduce churn rates, and ultimately enhance their profitability. This analysis also highlights the importance of developing data-driven approaches tailored to the Indian telecom industry, where customer loyalty plays a significant role in determining market success.

# **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**

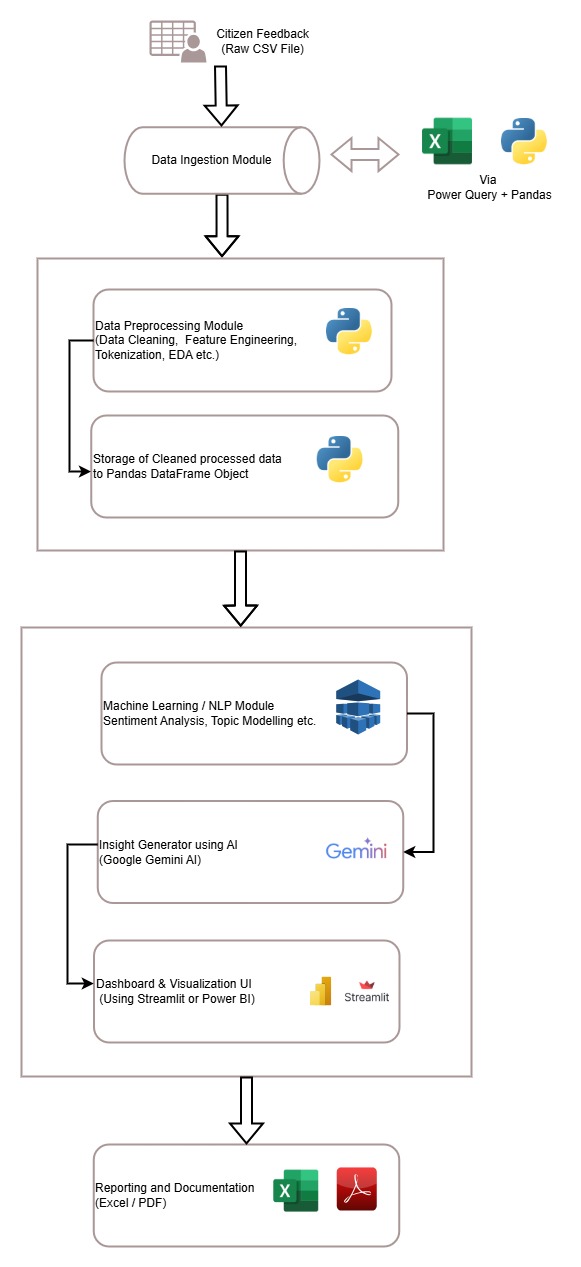
The models will be evaluated based on key metrics such as accuracy, precision, recall, and AUC-ROC scores. These metrics will help identify which model provides the best predictive accuracy and generalizes well to new data.

# **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)