IMsured

HELPING YOU STAY PREPARED

SOMAIYA VIDYAVIHAR UNIVERSITY

IMsured – Helping you stay prepared

Submitted in Partial Fulfillment of Requirements

for the Degree of

Master of Science

(Statistics)

Under the Guidance of

Prof. Jyoti Mantri

By

Ms. Tanvi Akre : 31031823001

Ms. Vidhi Murdeshwar : 31031823008

SOMAIYA VIDYAVIHAR UNIVERSITY

S K Somaiya College

Vidyavihar, Mumbai - 400 077

2024-25

Student Details	
Name	Ms. Tanvi Akre Ms. Vidhi Murdeshwar
Seat No.	31031823001 31031823008
Department	Mathematics and Statistics
Degree	Master of Science in Statistics
Institution	S K Somaiya College
Academic Year	2023 - 2025
Teacher in Charge	Prof. Jyoti Mantri
Title of the Project	IMsured- Helping you stay prepared
Location	Vidyavihar
Duration	40 hours
Signature of the Guide/Faculty In-charge	
Signature of the Head of Department	

Certificate of Authentication

This is to certify that the project entitled ”IMsured- Helping you stay prepared” is a bonafide work of Tanvi Akre- 31031823001 and Vidhi Murdeshwar- 31031823008 submitted to the S K Somaiya College in partial fulfillment of the requirement for the award of the degree of M.Sc. in the subject of Statistics.

I considered that the thesis has reached the standards and fulfilling the requirements of the rules and regulations relating to the nature of the degree. The contents embodied in the thesis have not been submitted for the award of any other degree or diploma in this or any other university.

External Supervisor

Guide/Faculty In-charge

Head of Department

Date: _____

Place: _____

Declaration by the Student

I certify that

(a) The work contained in the thesis is original and has been done by myself under the supervision of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

(e) Whenever I have quoted written materials from other sources, due credit is given to the sources by citing them.

(f) From the plagiarism test, it is found that the similarity index of the whole thesis is within 25%, and a single paper is less than 10% as per the university guidelines.

Date: _____

Place: Mumbai

Student Signature

Tanvi Akre

31031823001

Vidhi Murdeshwar

31031823008

SOMAIYA VIDYAVIHAR UNIVERSITY
S K Somaiya College

Somaiya TRUST

Department of Mathematics and Statistics

CERTIFICATE

This is to certify that Ms. of M.Sc. Statistics, has satisfactorily completed the Project titled "IMsured- Helping you stay prepared" for the Partial fulfillment of the Degree by the Somaiya Vidyavihar University, during the Academic year 2023-25.

Signature of the Guide	Signature of the HOD
Signature of the Examiners	Signature of Director
Date of Examination:	College seal

ACKNOWLEDGEMENT

We would like to express our gratitude and sincere thanks to our Project Guide, for instilling confidence in us to carry out this study and extending valuable guidance and encouragement from time to time, without which it would not have been possible to undertake and complete this project.

We also wish to extend my appreciation to our Head – Academics, Mrs. Jyoti Mantri and the Program Coordinators, for their kind co-ordination and support.

Also, our special thanks to the people of Mumbai for providing us with their vital information, opinions, valuable time & support.

Chapter 1: Introduction & Literature Survey

INDEX

Sr. No.	Content details	Page No.
1	Title page	
2	Student details	1
3	Certificate of authentication	2
4	Declaration by the student	3
5	Department certificate	4
6	Examiner Approval sheet	4
7	Acknowledgement	5
8	Contents	6
9	List of Abbreviations	8
10	List of Figures	9
11	Abstract	10
Chapter 1	**Introduction & Literature Survey**	
	1.1 Introduction	11
	1.2 Background of Research	13
	1.3 Review of Literature	14
	1.4 Problem Statement of the Work	15

	1.5 Aims and Objectives of the Work	16
Chapter 2	**Research Methodology**	
	2.1 Introduction to Data Collection Methods	17
	2.2 Sampling Technique	17
	2.3 Data Processing & Encoding	18
	2.4 Graphical Representations	19
Chapter 3	**Exploratory Data Analysis (EDA)**	28
Chapter 4	**Statistical Analysis & Predictive Modeling**	
	4.1 K-Nearest Neighbors	32
	4.2 Chi-Square Tests	36
	4.3 Random Forest Model	42
	4.4 Clustering Analysis (Fuzzy C-Means)	48
	4.5 Multinomial Logistic Regression with L2 Regularization	55
Chapter 5	**Conclusion and Future Prospects**	62
	Software & Codes used	64
	Questionnaire	70
	References	76

LIST OF ABBREVIATIONS

Abbreviation	Full Form
IRDI	Insurance Regulatory and Development Authority of India
EDA	Exploratory Data Analysis
KNN	K-Nearest Neighbours
ML	Machine Learning
NCB	No-Claim Bonus
PCA	Principal Component Analysis
SMOTE	Synthetic Minority Over-sampling Technique
L2	Ridge Regularization
CGHS	Central Government Health Scheme
ESIS	Employees' State Insurance Scheme

LIST OF FIGURES

Figure No.	Description
Figure 2.1	Distribution of medical insurance policyholder.
Figure 2.2	Age distribution of respondents
Figure 2.3	Gender distribution of respondents
Figure 2.4	Distribution of respondents by Area type
Figure 2.5	Distribution of respondents by Income level
Figure 2.6	Distribution of respondents by the number of their dependents
Figure 2.7	Health Insurance Policy Types Distribution
Figure 2.8	Health Insurance Policy Premium Distribution
Figure 2.9	Health Insurance Policy Renewal Frequency
Figure 2.10	Distribution of Respondents by Policy Duration
Figure 2.11	Distribution of Respondents by Claim Status
Figure 2.12	Frequency of Claims Made Under Current Policy
Figure 2.13	Distribution of Respondents by Most Recent Claim Amount
Figure 2.14	Distribution of Respondents by Claim Status
Figure 2.15	Key Concerns in Health Insurance Policies
Figure 2.16	Respondents' Willingness to Opt for Government Healthcare Plans
Figure 3.1	Claim count distribution by area type
Figure 3.2	Average Claim Amount by Area Type
Figure 3.3	Correlation Heatmap of Factors Affecting Last Claim Amount
Figure 4.1	KNN Accuracy for Different k Values
Figure 4.4	Feature importance scores from Random Forest model showing key policy attributes influencing satisfaction.
Figure 4.5.1	Fuzzy Cluster Visualization with PCA
Figure 4.5.2	Cluster Distribution

ABSTRACT

This study investigates healthcare claim patterns among rural and urban populations under private health insurance policies. By analyzing demographic, socioeconomic, and health-related factors, the research identifies key differences in claim behaviors, expenditure trends, and high-risk groups. Statistical and machine learning techniques, including clustering and hypothesis testing, are employed to uncover disparities in claim frequency, treatment types, and insurance utilization. The findings provide actionable insights for insurers to design tailored policies that improve accessibility, affordability, and healthcare outcomes for diverse populations. This research contributes to data-driven decision-making in the health insurance sector, fostering more equitable coverage across geographic regions.

Chapter 1: Introduction & Literature Survey

1.1 Introduction

Health insurance is a critical component of financial security, ensuring that individuals and families can access medical care without the burden of excessive out-of-pocket expenses. In India, the health insurance sector has witnessed significant growth over the past few decades, driven by rising healthcare costs, increased awareness, and government initiatives promoting health coverage. However, disparities remain in how different populations particularly rural and urban communities—access and utilize health insurance.

India's healthcare landscape is characterized by a dual system comprising public and private healthcare providers. While government schemes like Ayushman Bharat (PM-JAY) aim to provide financial protection to economically vulnerable groups, private health insurance is becoming increasingly important in bridging gaps in coverage and offering comprehensive medical services. Despite its significance, private health insurance penetration remains uneven, with urban populations benefiting more from private policies due to better financial literacy, employment-linked insurance, and accessibility to a well-developed healthcare infrastructure. In contrast, rural populations face multiple barriers, including low awareness, affordability issues, and limited healthcare facilities, leading to underutilization of health insurance.

This study aims to analyze the healthcare claim patterns of rural and urban populations under private health insurance policies, focusing on key demographic, geographic, socioeconomic, and health-related factors influencing claims and expenditures. By understanding these patterns, insurers can develop tailored products that better meet the needs of diverse populations.

A significant factor contributing to the urban-rural disparity in health insurance utilization is the accessibility and availability of healthcare services. Urban areas are home to well-equipped hospitals, multi-specialty clinics, and advanced medical technologies, enabling policyholders to seek treatment more frequently and file more insurance claims. In contrast, rural regions often lack adequate healthcare infrastructure, resulting in fewer medical consultations and claims despite a high burden of illnesses. Moreover, the prevalence of diseases varies between rural and urban populations. Urban areas report a higher incidence of lifestyle-related diseases such as diabetes, hypertension, and cardiovascular disorders due to sedentary habits and dietary patterns, while rural areas face a greater burden of communicable diseases, malnutrition, and maternal health issues.

Another key aspect influencing insurance utilization is the ease of claim processing. Urban policyholders have better access to digital tools, cashless hospitalization, and efficient claim settlement processes, making insurance more convenient to use. Rural policyholders, however, often struggle with documentation, claim rejection due to inadequate paperwork, and delays in reimbursement, discouraging them from fully utilizing their health insurance benefits.

To bridge these gaps, insurers and policymakers need data-driven insights to design more inclusive health insurance policies that cater to the unique needs of both rural and urban populations. This study employs statistical and machine learning techniques, including clustering and hypothesis testing, to identify key differences in claim behaviors, expenditure trends, and high-risk groups. The findings will help insurers optimize their products and services, making private health insurance more accessible, affordable, and efficient across different geographic regions.

By shedding light on these disparities and their underlying causes, this research contributes to data-driven decision-making in the health insurance sector, fostering equitable healthcare coverage and financial protection for all segments of society. The ultimate goal is to support insurers, healthcare providers, and policymakers in developing targeted strategies that enhance insurance penetration and utilization, ensuring that healthcare remains a fundamental right rather than a privilege limited to certain sections of the population.

1.2 Background of Research in India

Early Developments

The concept of health insurance in India dates back to the early 20th century, with the introduction of the Employees' State Insurance Scheme (ESIS) in 1948, which provided coverage for industrial workers. The Central Government Health Scheme (CGHS) was launched in 1954 to offer healthcare benefits to government employees and their families. These schemes laid the foundation for structured health insurance in India.

Growth of the Private Health Insurance Sector

Until the 1990s, public healthcare schemes dominated the insurance landscape. However, with economic liberalization and the entry of private insurers, the health insurance market expanded rapidly. The Insurance Regulatory and Development Authority of India (IRDAI), established in 1999, played a crucial role in regulating and promoting private health insurance. Today, numerous private and public sector insurers offer a variety of health plans, catering to different income groups and needs.

Current Scenario and Challenges:

- Rural vs. Urban Disparities: While urban areas have greater access to private hospitals and better awareness about health insurance, rural areas struggle with limited network hospitals, lower literacy rates, and affordability concerns.
- Claim Settlement Issues: Many policyholders face claim rejections or delays due to a lack of documentation, complex procedures, or limited digital infrastructure in rural areas.
- Health Risks and Insurance Utilization: Urban populations tend to claim more frequently due to better healthcare access, whereas rural policyholders might only claim in emergencies due to a lack of preventive care services.
- Low Insurance Penetration: Despite government schemes like Ayushman Bharat, India’s overall health insurance penetration remains low compared to global standards, particularly in rural regions.

1.3 Review of Literature

1.3.1 Socio-Economic Factors in Insurance Claims

Several studies have highlighted how economic status, education levels, and access to services affect insurance claim behaviour.

- A study by Smith et al. (2020) found that urban policyholders tend to submit more frequent claims due to better awareness and accessibility to insurance services.
- Rural areas, on the other hand, showed higher rejection rates for claims due to insufficient documentation or lack of awareness about claim procedures (Brown & Patel, 2019).

1.3.2 Risk Assessment and Fraud Detection

Insurance fraud is a significant concern, with fraudulent claims leading to substantial financial losses.

- Machine learning models, such as logistic regression, decision trees, and neural networks, have been used to detect anomalies in claim data (Chen & Zhao, 2021).
- Studies suggest that urban areas may experience more fraudulent health and auto insurance claims, whereas rural areas may see higher fraud rates in agricultural and property insurance.

1.3.3 Predictive Modeling in Insurance

Regression and machine learning techniques are widely applied for predicting claim trends and risk factors.

- Linear regression and logistic regression models have been effective in identifying key predictors of claim amounts and frequencies (Williams et al., 2018).
- Random forests and deep learning approaches have demonstrated superior performance in classifying high-risk claims and detecting outliers.

While past research has explored these factors individually, there remains a research gap in comprehensive comparative studies focusing on both rural and urban claim patterns using advanced statistical techniques and machine learning models. This study aims to bridge this gap.

1.4 Problem Statement of the Work

Insurance claims processing and policyholder behavior differ significantly between rural and urban areas, influenced by demographic, behavioral, and economic factors. Insurers often struggle with identifying high-risk policyholders, optimizing claim settlement processes, and understanding customer satisfaction drivers. The lack of a data-driven approach to analyze these differences leads to inefficiencies in claim approval, policyholder dissatisfaction, and increased fraud risks.

Moreover, factors such as health behaviors (smoking, alcohol consumption, physical activity) and policy preferences play a crucial role in determining claim outcomes. However, the impact of these factors remains underexplored in the insurance sector. Additionally, policyholder switching behavior from private to government insurance schemes is influenced by several unknown variables, necessitating deeper analysis for better decision-making.

This research aims to address these gaps by employing statistical models, machine learning techniques, and clustering approaches to:

- Identify key factors affecting claim settlement rates across rural and urban populations.
- Classify high-risk policyholders using predictive modeling.
- Analyze policyholder satisfaction and switching behavior trends.
- Provide actionable insights for insurers to enhance risk assessment and policy optimization.

By leveraging advanced data analysis methods, this study will provide valuable insights for insurance companies, leading to more efficient claims processing, improved policyholder experience, and equitable insurance policies across different demographics.

1.5 Aims and Objectives of the Work

Aim:

The primary aim of this study is to analyze insurance claim patterns and policyholder behaviour in rural and urban areas using statistical and machine learning techniques. The research focuses on identifying key risk factors, satisfaction levels, claim settlement trends, and policyholder switching behaviours, ultimately providing data-driven insights for insurers to improve risk assessment and policy optimization.

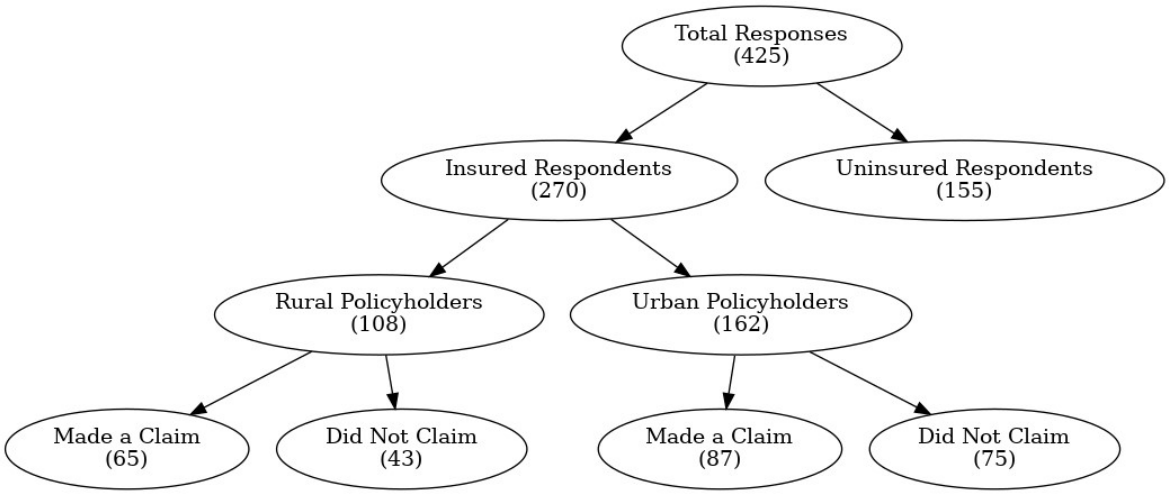
Objectives:

Chapter 2: Research Methodology

2.1 Introduction to Data Collection Methods

For primary data collection, a survey approach was employed, focusing exclusively on Maharashtra state. A structured research questionnaire was designed to gather information on respondents' health insurance claims. Initially, a pilot survey was conducted with a sample of 25 individuals, allowing us to refine the questionnaire based on feedback.

The final survey was administered to a total of 425 individuals through Google Forms and in-person interactions. The collected data was then entered into an Excel sheet, cleaned, and encoded for analysis. Of the 425 respondents, 270 individuals had active health insurance policies, comprising 108 from rural areas and 162 from urban areas. The remaining 155 respondents did not have any insurance coverage.



2.2 Sampling Technique

In the present study, Stratified Convenience Sampling was used to collect data from the defined population. This sampling method, a type of non-probability sampling, involves first dividing the population into distinct subgroups (strata) based on specific characteristics and then selecting participants based on accessibility and willingness to participate.

In this case, the population was stratified based on geographic location (rural/urban) and insurance status (with/without insurance) to ensure representation from each group. However, within each stratum, participants were selected through convenience sampling, meaning data was gathered from individuals who were readily available and willing to respond. No strict inclusion criteria were set before participant selection, and all individuals within the strata were invited to take part in the study.

2.3 Data Processing & Encoding

1. Data Cleaning

- Missing values were identified and handled through appropriate imputation techniques.
- Inconsistent and duplicate entries were removed to ensure data integrity.
- Standardized categorical variables were reviewed to maintain uniformity across datasets.

2. Encoding of Categorical Variables

Since machine learning models typically require numerical input, categorical variables were transformed into numerical representations using the following techniques:

- **Label Encoding**: Applied to ordinal variables such as satisfaction levels and policy duration.
- **One-Hot Encoding**: Used for nominal variables such as Policy name, Coverage option and Factors to improve.
- **Binary Encoding**: Implemented for binary variables like gender (Male/Female) and Chronic Illness (Yes/NO) .

3. Standardization & Normalization

- Numerical data, including policy premiums, claim amounts, and customer age, were standardized using **Z-score normalization** to ensure comparability across features.
- Some variables, such as claim frequency, were normalized to avoid the influence of extreme values.

4. Handling Imbalanced Data

To address any imbalance in the dataset**,** SMOTE (Synthetic Minority Over-Sampling Technique) was applied to improve classification performance, especially for claim approval predictions.

2.4 Graphical Representations

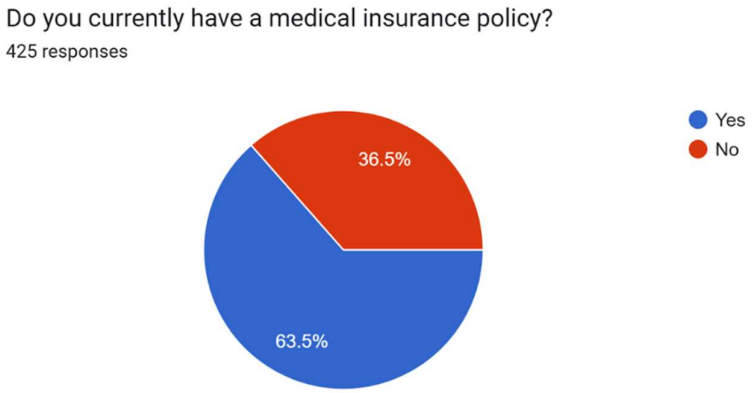


Figure (2.1) Distribution of medical insurance policyholder.

- **Out of 425 people there are 270 people who have active medical insurance policy and 155 who do not have medical Insurance policy.**

INSURED VS. UNINSURED POLICYHOLDERS

Age responses of the people:

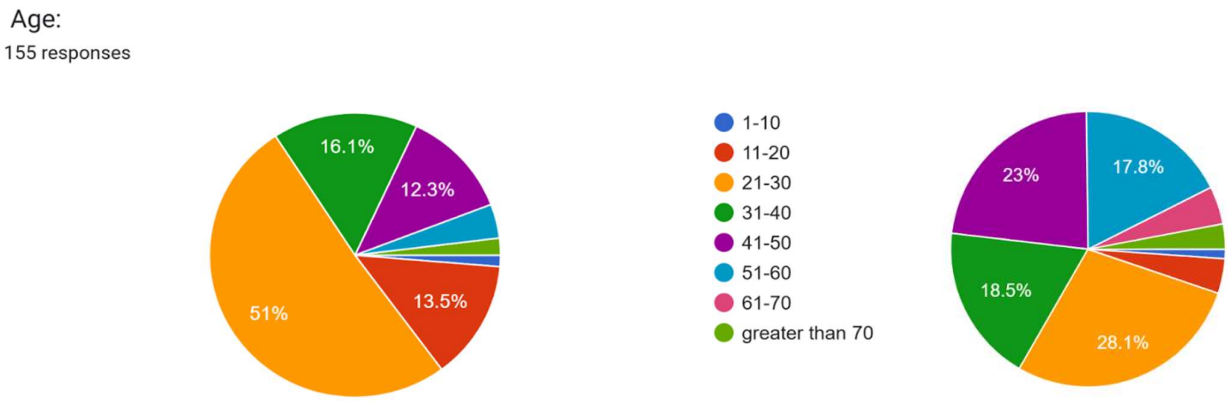


Figure (2.2) Age distribution of respondents

51% which means there are 79 people falling in the age group of 21-30 who do not have a medical insurance policy	28.1% which means there are 76 people falling in the age group of 21-30 who have a medical insurance policy

16.1% which means there are 25 people falling in the age group of 31-40 who do not have a medical insurance policy	23% which means there are 62 people falling in the age group of 41-50 who have a medical insurance policy
13.5% which means there are 21 people falling in the age group of 11-20 who do not have a medical insurance policy	18.5% which means there are 50 people falling in the age group of 31-40 who have a medical insurance policy
12.3% which means there are 19 people falling in the age group of 41-50 who do not have a medical insurance policy	17.8% which means there are 19 people falling in the age group of 41-50 who do not have a medical insurance policy

Gender responses of the people:

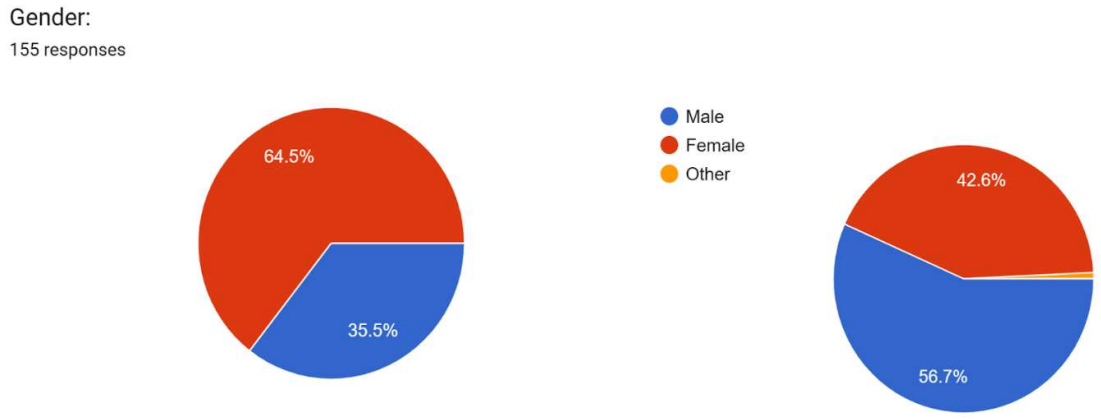


Figure (2.3) Gender distribution of respondents

There are 64.5% which accounts to 100 females who do not have a medical insurance policy.	There are 42.6% which accounts to 115 females who have a medical insurance policy.
There are 35.5% which accounts to 55 males who do not have a medical insurance policy.	There are 56.7% which accounts to 153 females who have a medical insurance policy.

Type of area responses of the people:

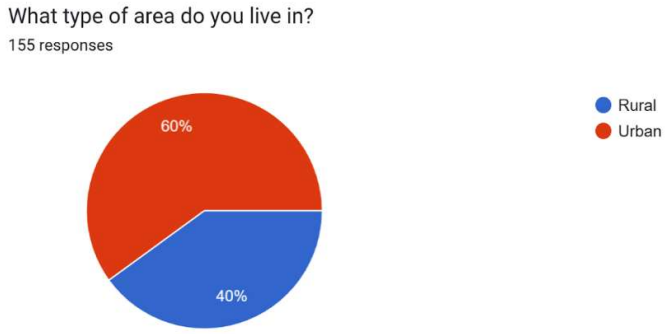


Figure (2.4) Distribution of respondents by Area type

57.4% meaning 89 individuals belonging to the rural population do not have a medical insurance	40% meaning 108 individuals belonging to the rural population have a medical insurance
42.6% meaning 66 individuals belonging to the urban population do not have a medical insurance	60% meaning 162 individuals belonging to the rural population have a medical insurance.

Income Level Responses:

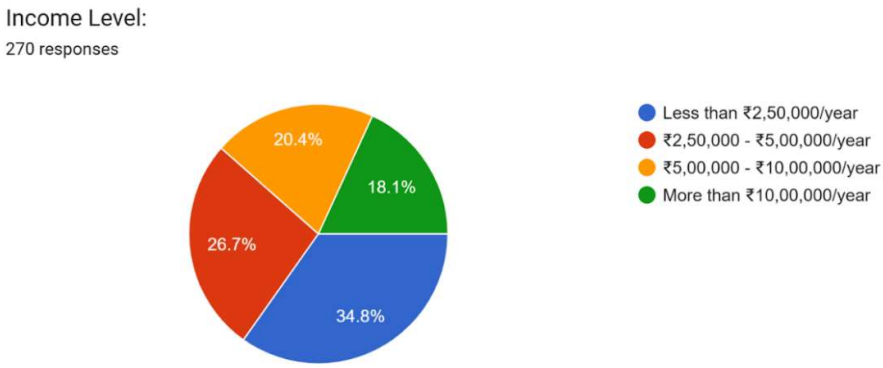


Figure (2.5) Distribution of respondents by Income level

- 34.8% of the respondents, equivalent to 94 individuals, have an income of less than ₹2,50,000 and hold a medical insurance policy where as 26.7% of the respondents, equivalent to 72 individuals, have an income between ₹2,50,000-₹5,00,000 and hold a medical insurance policy.

Number of Dependents:

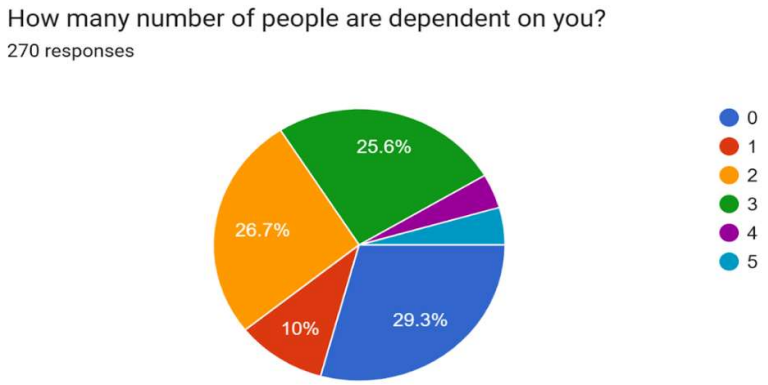


Figure (2.6) Distribution of respondents by the number of their dependents

- 26.7% of respondents (72 people) have two dependents, while 29.3% (79 people) have five dependents.

Type of Health Insurance Policy Held:

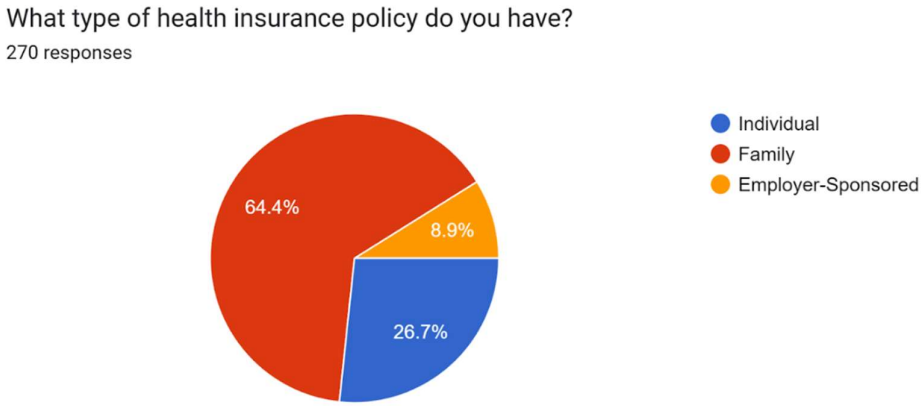


Figure (2.7) Health Insurance Policy Types Distribution

- 64.4% of the respondents (174 individuals) have a family insurance policy, 26.7% (72 individuals) have an individual insurance policy, and 8.9% (24 individuals) have an employer-sponsored insurance policy.

Policy Premium Amount:

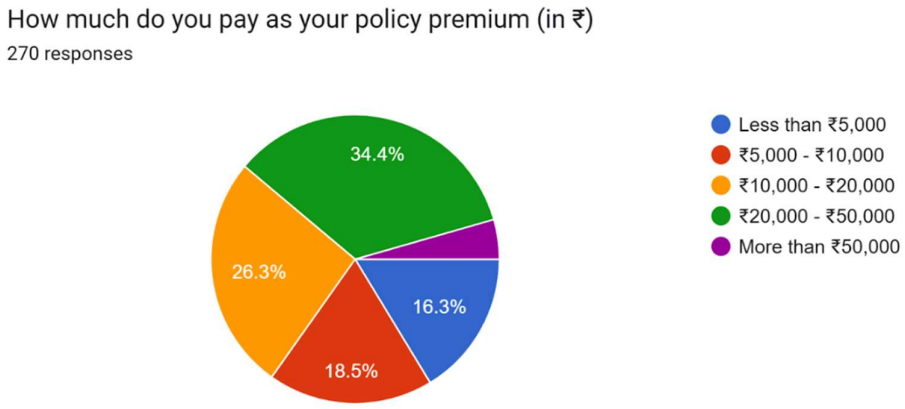


Figure (2.8) Health Insurance Policy Premium Distribution

- 34.4% of the respondents (93 individuals) pay a policy premium between ₹20,000 and ₹50,000, while 16.3% (44 individuals) pay less than ₹5,000.

Policy Renewal Frequency:

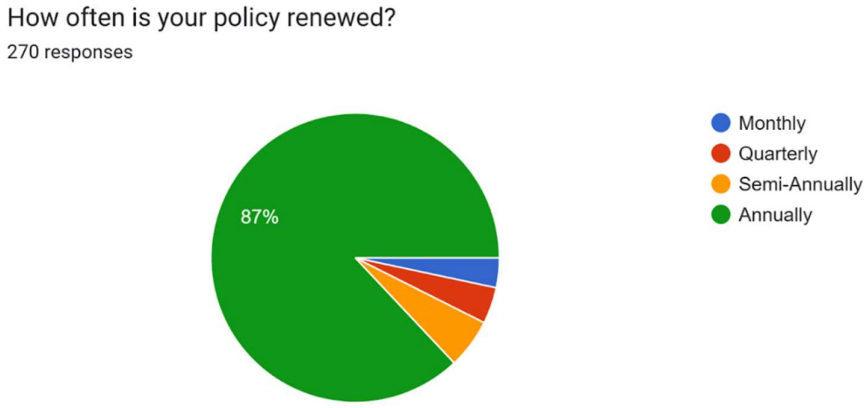


Figure (2.9) Health Insurance Policy Renewal Frequency

- 87% of the insured respondents, equivalent to 235 out of 270 individuals, renew their policy annually.

Policy Duration Responses:

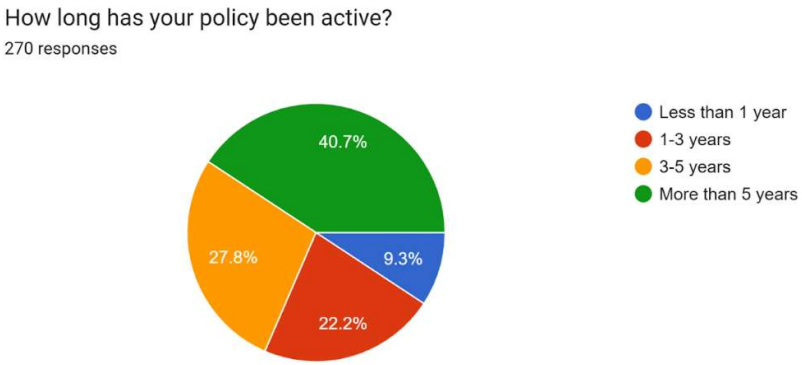


Figure (2.10) Distribution of Respondents by Policy Duration

- 40.7% of respondents (110 people) have had an active policy for more than five years, while 9.3% (25 people) have had it for less than one year.

Claim History Responses:

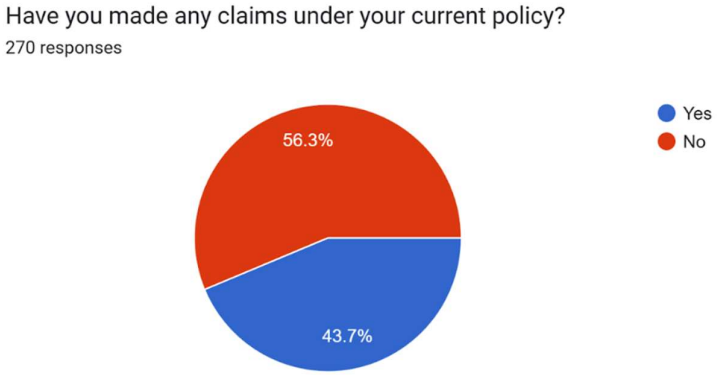


Figure (2.11) Distribution of Respondents by Claim Status

- 56.3% of the respondents (152 individuals) have made insurance claims, while 43.7% (118 individuals) have not.

Number of Claims Responses:

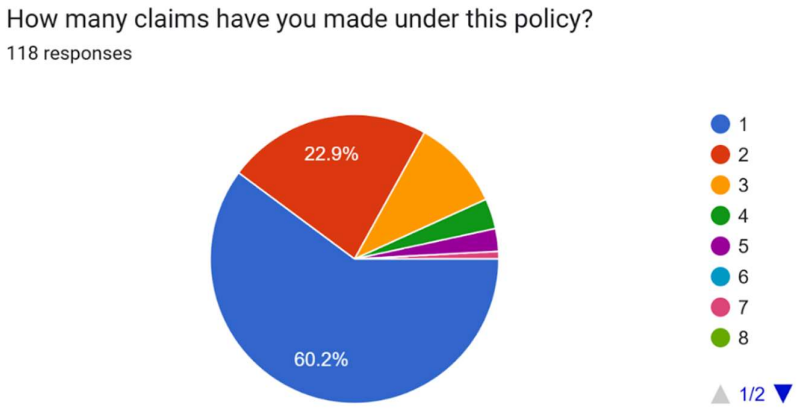


Figure (2.12) Frequency of Claims Made Under Current Policy

- 60.2% of the respondents (71 individuals) have made a single claim so far, while 22.9% (27 individuals) have made two claims.

Recent Claim Amount Responses:

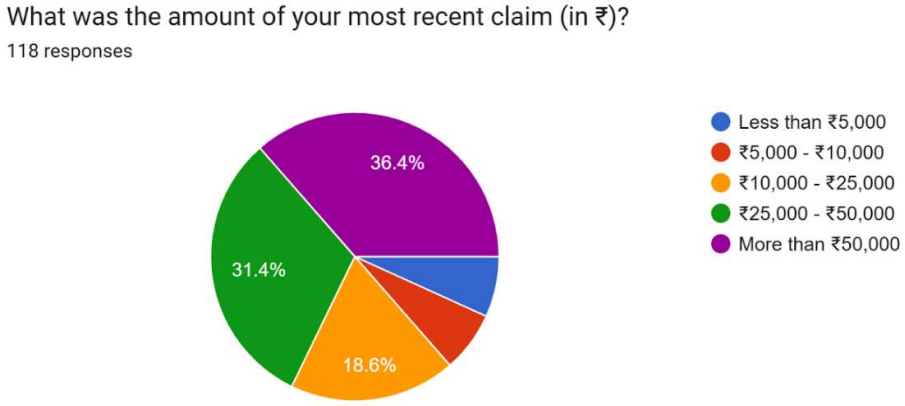


Figure (2.13) Distribution of Respondents by Most Recent Claim Amount

- 36.4% of respondents (43 people) have made a recent claim exceeding ₹50,000, while 31.45% (37 people) have claimed between ₹25,000 and ₹50,000.

Claim Status Responses:

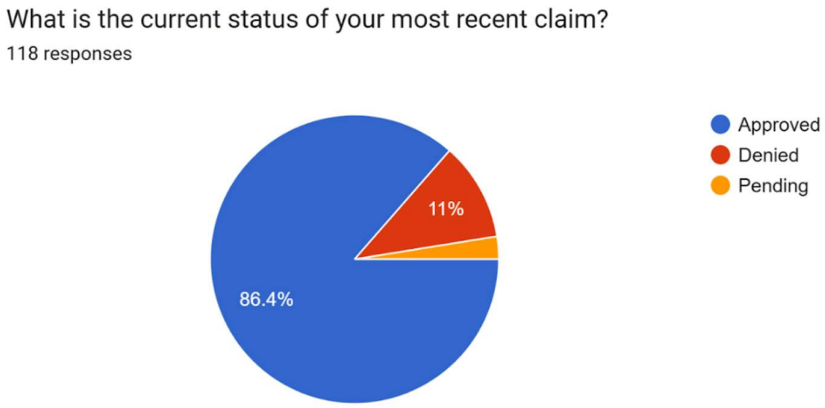


Figure (2.14) Distribution of Respondents by Most recent Claim Status

- 86.4% of the respondents (102 individuals) had their claims approved, 11% (13 individuals) had their claims denied, and 2.5% (3 individuals) have claims still pending.

CLAIMED VS. NOT CLAIMED RESPONSES

Factors for Improving Satisfaction Responses:

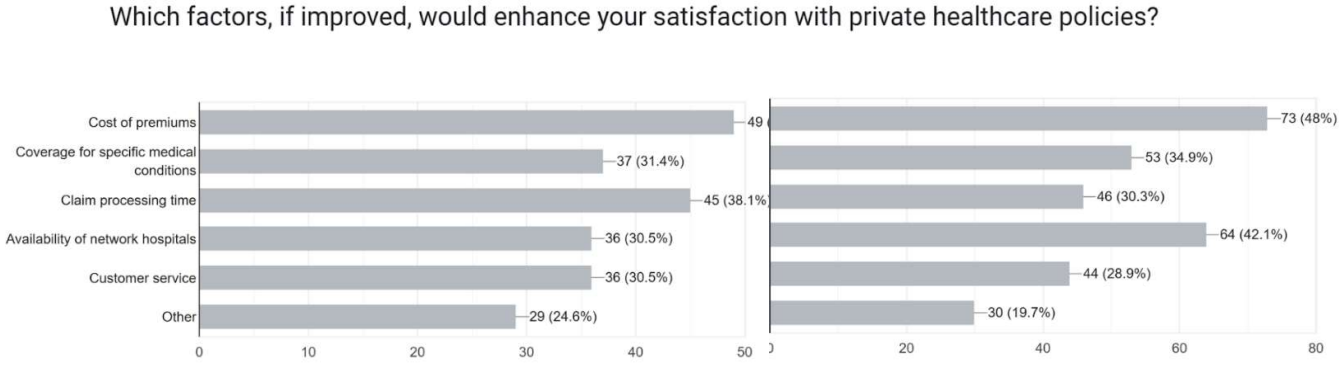


Figure (2.15) Key Concerns in Health Insurance Policies

- **41.5% of claimants and 48% of non-claimants believe lower premiums would improve their satisfaction.**
- **38.1% of claimants and 30.3% of non-claimants believe improving claim processing time would enhance their satisfaction.**

Willingness to Switch to Government Healthcare Policies – Responses:

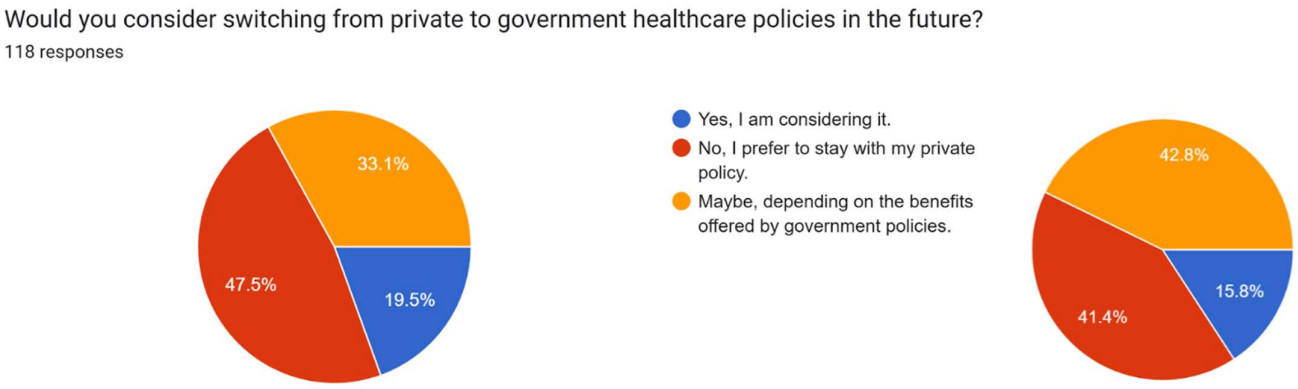


Figure (2.16) Respondents' Willingness to Opt for Government Healthcare Plans

47.5% of claimants prefer to stay with private insurance and are not considering switching to government plans.	41.4% of non-claimants prefer to stay with private insurance and are not considering switching to government plans.
19.5% of claimants are considering switching to government insurance.	15.8% of non-claimants are considering switching to government insurance.
33.1% of claimants might switch to government insurance, depending on the benefits offered.	42.8% of non-claimants might switch to government insurance, depending on the benefits offered.

Chapter 3: Exploratory Data Analysis (EDA)

To gain insights into the insurance claim patterns between urban and rural policyholders, we conducted an in-depth exploratory data analysis (EDA). The analysis focuses on claim count distribution, average claim amounts, and correlation analysis to identify key factors influencing high claim amounts.

♦ **Data Overview**

The dataset consists of 270 records and 81 features, covering:

- Demographics: Age, Gender, Marital Status, Income Level, Dependents, and Area Type (Urban/Rural).
- Insurance Details: Policy type, Premium, Coverage details, Claim count, Last Claim Amount, and Last Claim Status.
- Service-related factors: Customer service experience.

3.1 Claim Distribution Analysis

To understand the frequency of claims, we analyzed the Claim Count Distribution across area types.

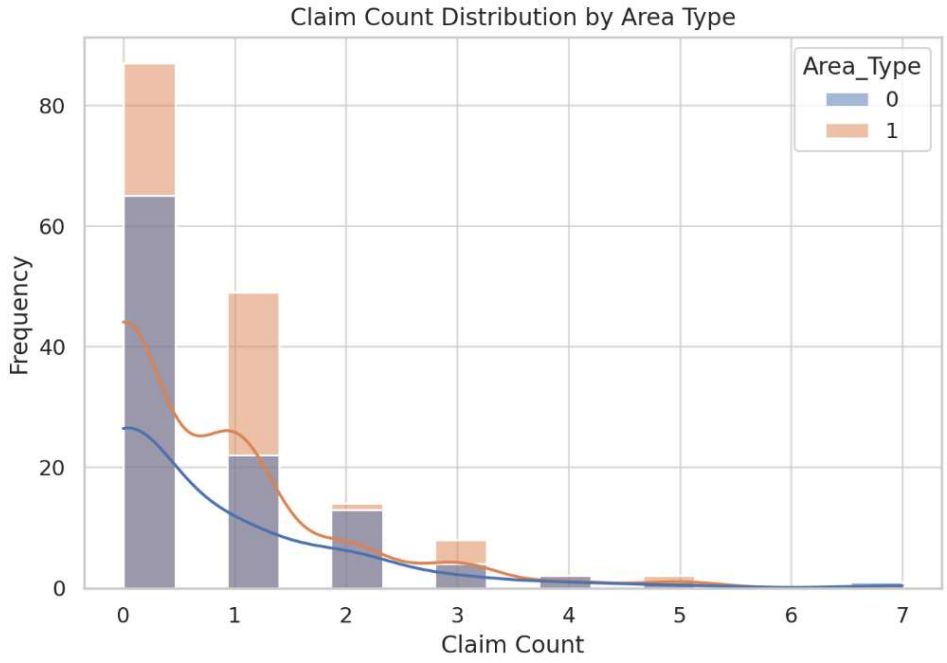


Figure (3.1) Claim count distribution by area type

- The majority of both urban and rural policyholders have either zero or one claim in their records.
- However, urban policyholders exhibit a higher frequency of multiple claims compared to rural policyholders. This could indicate increased claim activity in urban areas due to higher policy usage or more frequent risks.

3.2 Average Claim Amount by Area Type

Next, we examined the differences in average claim amounts between urban and rural policyholders.

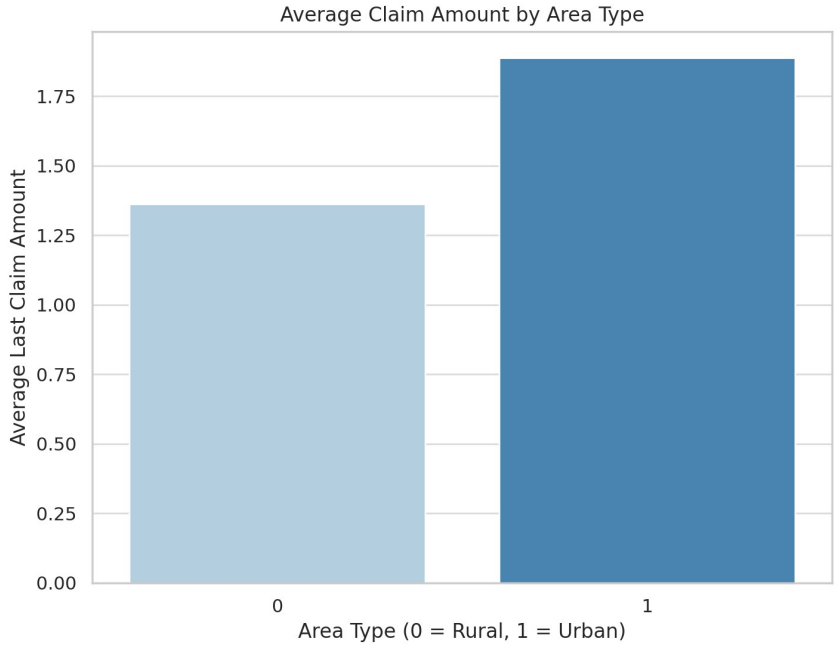


Figure (3.2) Average claim amount by area type

- The average claim amount is significantly higher in urban areas compared to rural areas.
- This suggests that urban claims might involve higher repair costs, increased policy coverage, or higher damage severity, making them more expensive.

3.3 Correlation Analysis: Key Factors Influencing Claim Amount

To identify the most significant factors influencing Last Claim Amount, we conducted a correlation analysis on numerical variables. The heatmap a below highlights the strongest correlations.

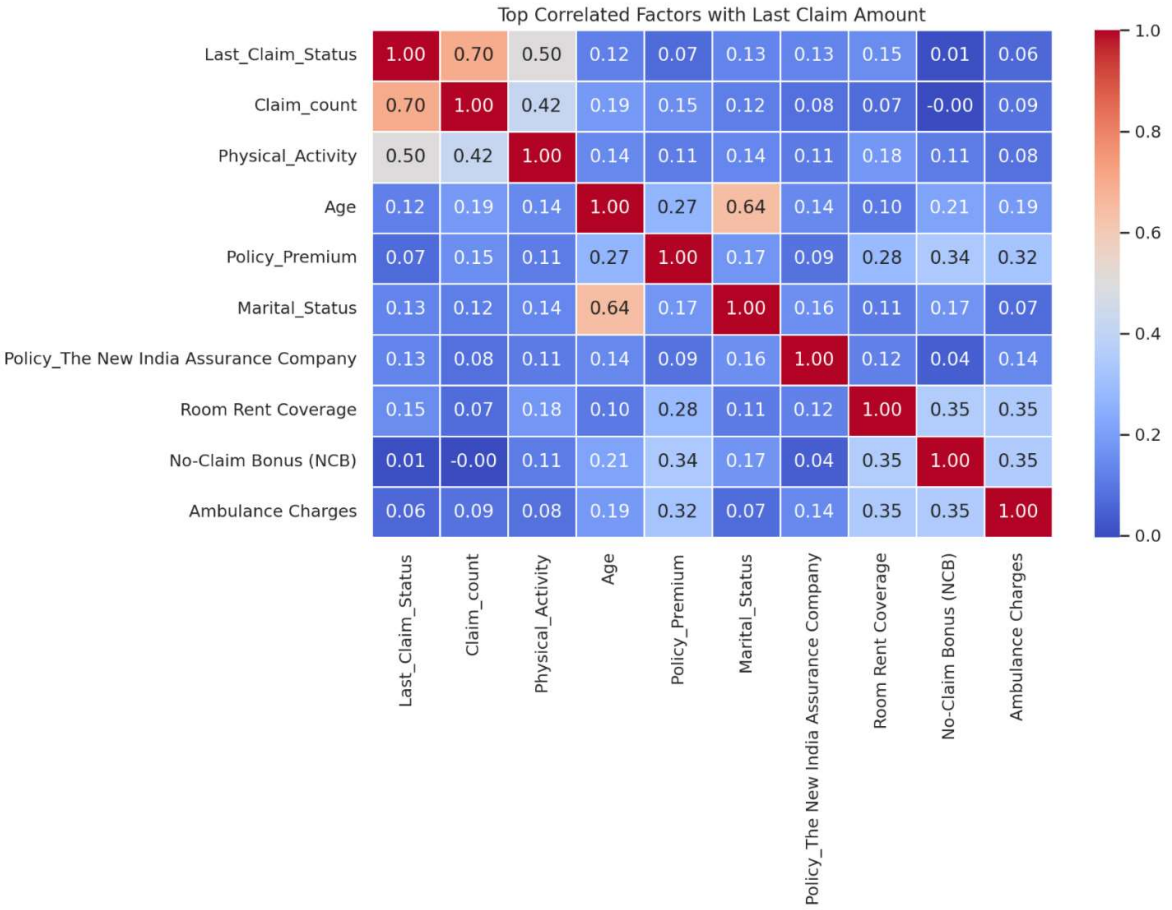


Figure (3.3) Correlation Heatmap of Factors Affecting Last Claim Amount

1. Claim Status (+0.80):
   - A strong positive correlation indicates that approved claims tend to have higher claim amounts.
2. Claim Count (+0.71):
   - More frequent claims are associated with higher cumulative claim amounts.
   - Policyholders with multiple claims tend to have higher individual claim values as well.
3. Physical Activity (+0.52):
   - A higher level of physical activity correlates with increased claim amounts, possibly due to medical expenses related to injuries.

4. Age (+0.25):

- Older policyholders tend to file higher claims, possibly due to increased health-related insurance usage.

5. Policy Premium (+0.22):

- Higher premiums are linked with higher claim amounts, suggesting comprehensive coverage leads to larger payouts.

6. Marital Status, Room Rent Coverage, and Ambulance Charges (+0.15 to +0.17):

- These factors have moderate correlations, indicating they contribute to claim amounts but are not primary drivers.

Implications for Insurers:

- Regular claimers and large claims should be monitored for possible fraud or premium changes.
- Premiums can be adjusted based on age and activity levels for fair pricing.
- Urban policyholders may have higher claims due to lifestyle risks, so insurance plans should be tailored accordingly.

Chapter 4: Statistical Analysis & Predictive Modeling

4.1 K-NEAREST NEIGHBORS

The k-nearest neighbors (KNN) algorithm is a straightforward, supervised machine-learning technique used for both classification and regression tasks. It is easy to implement and interpret; however, its performance significantly slows down as the dataset size increases.

KNN operates by calculating the distance between a given query point and all examples in the dataset. It then selects the K closest neighbors and determines the output based on them—either by majority voting for classification or by averaging the values for regression.

Choosing the optimal K value is crucial for model performance. This is typically done by testing multiple values and selecting the one that yields the best results.

OBJECTIVE 1

To predict whether an individual will or will not make an insurance claim based on their smoking status, alcohol consumption, and physical activity levels.

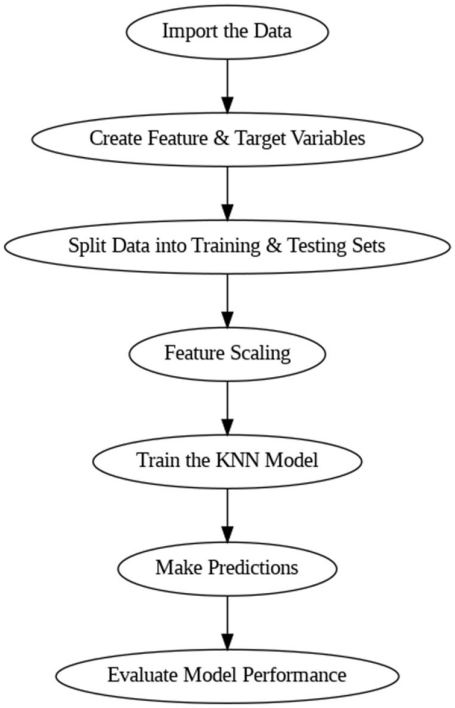
PROS

1. K-NN is pretty intuitive and simple.
2. K-NN has no assumptions.
3. It constantly evolves.
4. Very easy to implement for multi-class problem.
5. Can be used both for Classification and Regression.
6. One Hyper Parameter.
7. Variety of distance criteria to be chosen.

CONS

1. K-NN is slow algorithm.
2. Curse of Dimensionality.
3. K-NN needs homogeneous features.
4. Optimal number of neighbors.
5. Imbalanced data causes problems.
6. Outlier sensitivity.
7. Missing Value treatment.

STEPS TO IMPLEMENT K-NN ALGORITHM



DATA ANALYSIS & INTERPRETATION

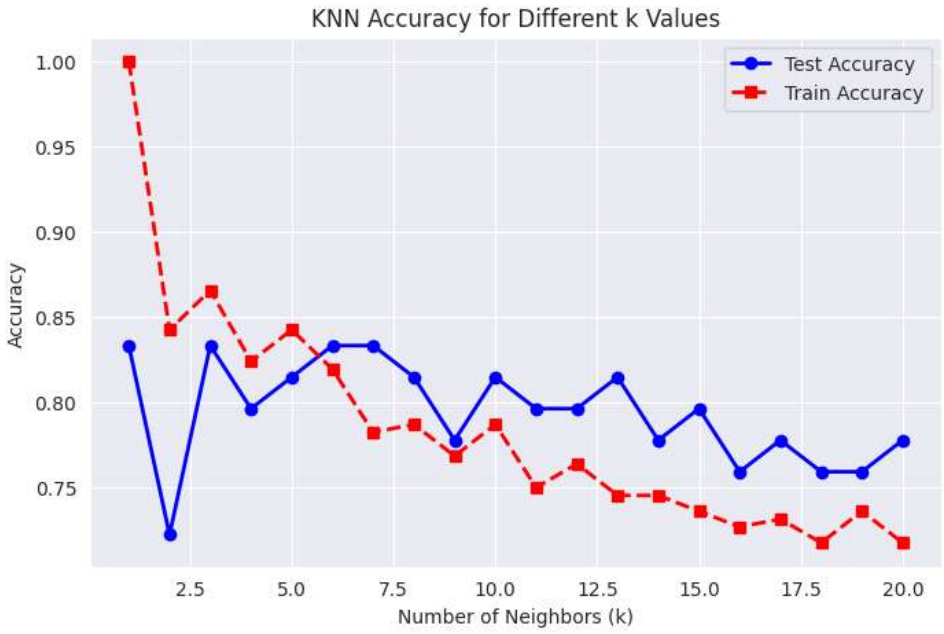
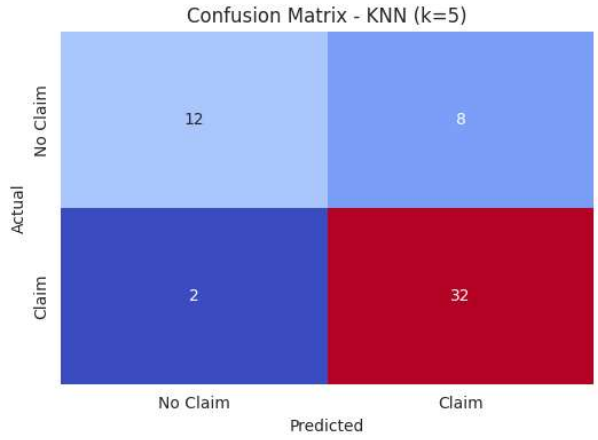


Figure (4.1) KNN Accuracy for Different k Values

Since k = 5-7 provides the best generalization, we select k = 5 for the KNN algorithm as the optimal choice.

1. Confusion matrix

Actual / Predicted	No Claim	Claim
No Claim	12	8
Claim	2	32



Using K=5

Accuracy= (TN+TP)/(TN+TP+FN+FP)

Accuracy score = 81.48%

2. Classification Report

Class	Precision	Recall	F1-Score
No Claim	0.86	0.60	0.71
Claim	0.80	0.94	0.86
Accuracy	-	-	0.81
Macro Avg	0.83	0.77	0.79
Weighted Avg	0.82	0.81	0.81

- Accuracy (81.48%) implies that, the model correctly predicts whether a claim is made in 81.48% of cases.
- Macro Avg (F1-score: 0.79) implies that, the model performs fairly well across both classes, considering balance.
- Weighted Avg (F1-score: 0.81), meaning this metric accounts for class imbalance, showing strong performance.
- NoClaim
  Precision (0.86): When the model predicts "No Claim," it is correct 86% of the time.

  Recall (0.60): The model correctly identifies 60% of actual "No Claim" cases.

- Claim
  Precision (0.80): When the model predicts "Claim," it is correct 80% of the time.
  Recall (0.94): The model successfully detects 94% of actual "Claim" cases.

Therefore, the model is better at detecting "Claim" cases than "No Claim" cases (higher recall for "Claim" at 94%).

4.2 CHI-SQUARE TESTS

Chi Square test of independence measures whether there is a relationship between two categorical variables. The Chi Square statistic is a non-parametric tool designed to analyze group differences when the dependent variable is measured at nominal level. It does not require equality of variances among the study groups or Homoscedasticity in the data.

Chi Square is robust with respect to the distribution of the data. Unlike most statistics, the Chi Square can provide information not only on the significance of any observed differences, but also provides detailed information on exactly which categories account for any differences found.

OBJECTIVE 2

To Compare claim behaviours between urban and rural policyholders.

PROS

1. It is easier to compute.
2. It identifies the difference between observed and expected values.
3. It does not assume anything about the data distribution.

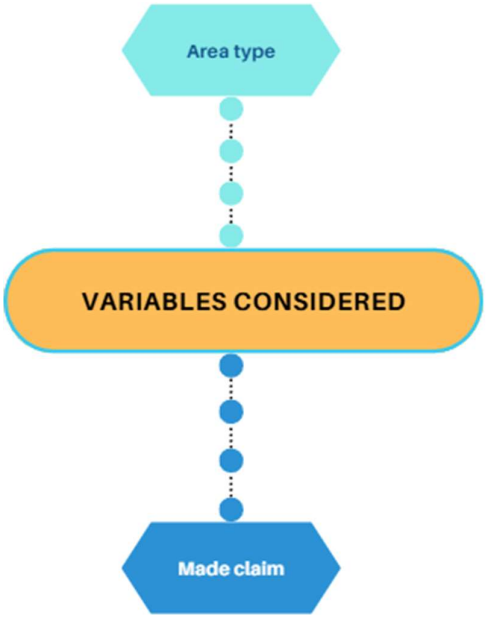
CONS

1. The number of observations should be more than 20.
2. It can't use percentages.
3. Data must be frequency data.
4. It is sensitive to small frequencies (below 5) which leads to erroneous conclusions.

ASSUMPTIONS

1. A sample with sufficiently large size is assumed.
2. The observations are always assumed to be independent of each other.
3. The categories are mutually exclusive i.e., each subject should fit in only one category.
4. It assumes that the data for the study is randomly picked from the population.

VARIABLES IDENTIFICATION



VARIABLES ENCODED

Area Type	Rural	0
	Urban	1
Made Claim	Yes	1
	No	0

Hypothesis :

H_0: Area type and Made claim are independent.

(There is no significant relationship between them.)

H_1: Area type and Made claim are dependent.

(There is a significant relationship between them.)

TEST STATISTIC :

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where, O_i= Observed frequency,

E_i= Expected frequency

D.F: (r-1) (c-1)

Where, r = Number of rows,

c = Number of columns

DECISION CRITERIA (For ɑ = 0. 05)

If p value is less than 0.05 then we reject null hypothesis & conclude that there is association between two variables, i.e., they are dependent on each other.

```
Chi-Square Test Results:
Chi-Square Statistic: 0.8587
Degrees of Freedom: 1
P-Value: 0.3541
```

CONCLUSION

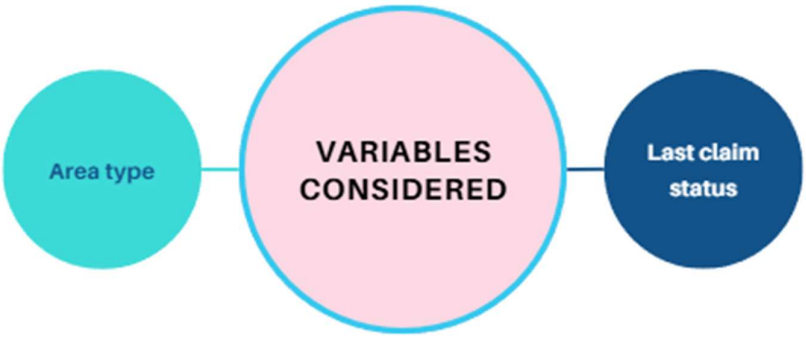
P-value = 0.3541 > 0.05 concluding that we fail to reject the null hypothesis.

This means there is no significant association between Area type and Last claim status.

OBJECTIVE 3

To test the association between claim settlement and area type.

VARIABLES IDENTIFICATION:



VARIABLES ENCODED

Area type	Rural	0
	Urban	1
Last claim status	Approved	0
	Denied	1
	Pending	2

STEPS TO IMPLEMENT FOR CHI-SQUARE TEST OF INDEPENDENCE

- Import Libraries
- Load Dataset
- Check Unique Values
- Create Contingency Table
- Perform Chi-Square Test
- Interpret Results

Hypothesis:

(H_0): There is no significant association between Area type (Rural/Urban) and Last claim Status

(H_1): There is a significant association between Area type (Rural/Urban) and Last claim Status

TEST STATISTIC

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where, O_i= Observed frequency

E_i= Expected frequency

D.F: (r-1) (c-1)

Where, r = Number of rows

c = Number of columns

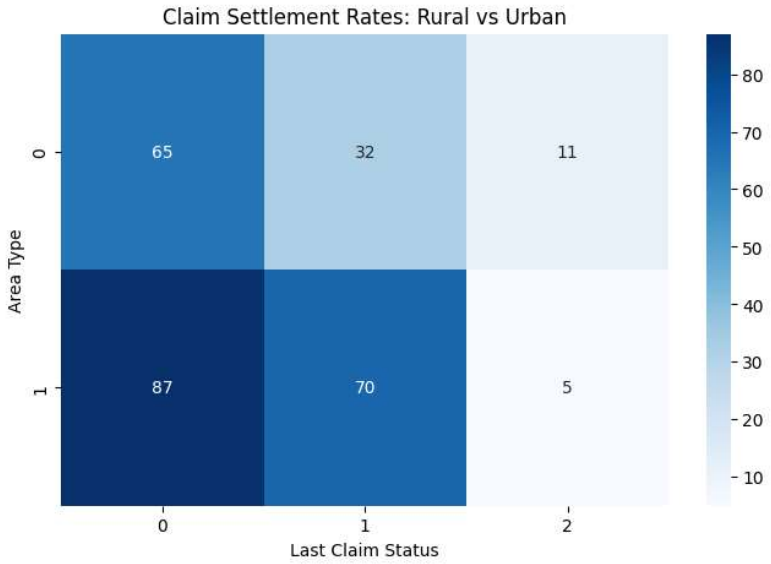
DECISION CRITERIA (For α = 0. 05)

If p value is less than 0.05 then we reject null hypothesis & conclude that there is association between two variables, i.e., they are dependent on each other

Chi-Square Statistic: 9.15736799105607

P-value: 0.010268400669748498

Degrees of Freedom: 2



Since p-value < 0.05, we reject the null hypothesis (H_0). And conclude that there is a significant association between Area type and Last claim status. And they are dependent of each other.

CONCLUSION

Location (Rural vs. Urban) significantly impacts claim settlement rates.

- Urban areas (1) have higher claim settlements (87 cases) compared to Rural (65 cases).
- Pending claims (status = 1) are also higher in Urban areas (70) compared to Rural (32).
- Rejections (status = 2) are slightly higher in Rural (11) compared to Urban (5).

4.3 RANDOM FOREST MODEL

Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their outputs to improve prediction accuracy and reduce overfitting. It is well-suited for handling complex, non-linear relationships and high-dimensional data.

OBJECTIVE 4

To assess policyholder satisfaction levels and determine key policy features influencing overall satisfaction.

PROS

1. Handles large datasets with higher accuracy than individual decision trees.
2. Reduces overfitting through bootstrapping and feature randomness.
3. Works well with both categorical and continuous predictor variables.
4. Provides feature importance, helping in understanding key drivers of satisfaction.
5. Can handle missing data by averaging predictions across trees.

CONS

1. Computationally intensive, requiring more memory and processing power.
2. Harder to interpret than logistic regression due to multiple decision trees.
3. Not ideal for very sparse datasets.
4. Model training can be slower compared to simpler algorithms.

ASSUMPTIONS

1. Observations should be independent (no repeated or related data points).
2. No need for data to follow a normal distribution or have a linear relationship.
3. Identifies which features (variables) are most important for prediction.
4. Uses techniques like bootstrapping (resampling data) and aggregation (combining results) for better accuracy.
5. Works best with a sufficiently large dataset for reliable predictions.

RANDOM FOREST MODEL

- Random Forest constructs multiple decision trees based on bootstrapped samples of the dataset and aggregates their predictions. Each tree is trained on a subset of features to increase diversity.

- **Prediction Formula:**

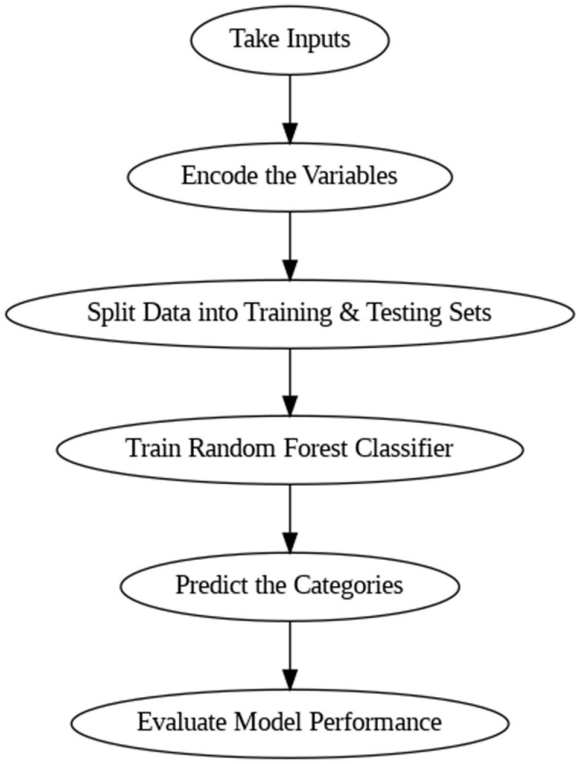
$$f(X)=\frac{1}{N}Tree_i(X)$$

where:
N is the number of decision trees in the forest.
$Tree_i(X)$represents the individual decision tree predictions.

VARIABLE IDENTIFICATION

Variable Name	Type
Overall Satisfaction	Categorical (Ordinal)
Room Rent Coverage	Binary
Personal Accident Coverage	Binary
No-Claim Bonus (NCB)	Binary
Hospitalization Expenses	Binary
Maternity , Newborn Coverage	Binary
Mental Health Treatment	Binary
Preventive Health Check-ups	Binary
Daycare Procedures	Binary
Domiciliary Treatment	Binary
Ambulance Charges	Binary
Cashless Treatment	Binary
Critical Illness Coverage	Binary

STEPS TO BE IMPLEMENTED IN LOGISTIC REGRESSION



DATA ANALYSIS & INTERPRETATION

1. Feature Importance Analysis

The most important features in driving satisfaction scores are:

1. Cashless Treatment – Customers highly value policies that provide seamless cashless hospitalizations.
2. No-Claim Bonus (NCB) – Policies rewarding customers for not making claims improve satisfaction.
3. Critical Illness Coverage – Policies covering major illnesses significantly impact customer trust.
4. Maternity & Newborn Coverage – Comprehensive maternity benefits improve policyholder sentiment.

5. Ambulance Charges & Hospitalization Expenses – These affect the convenience and affordability of policies.

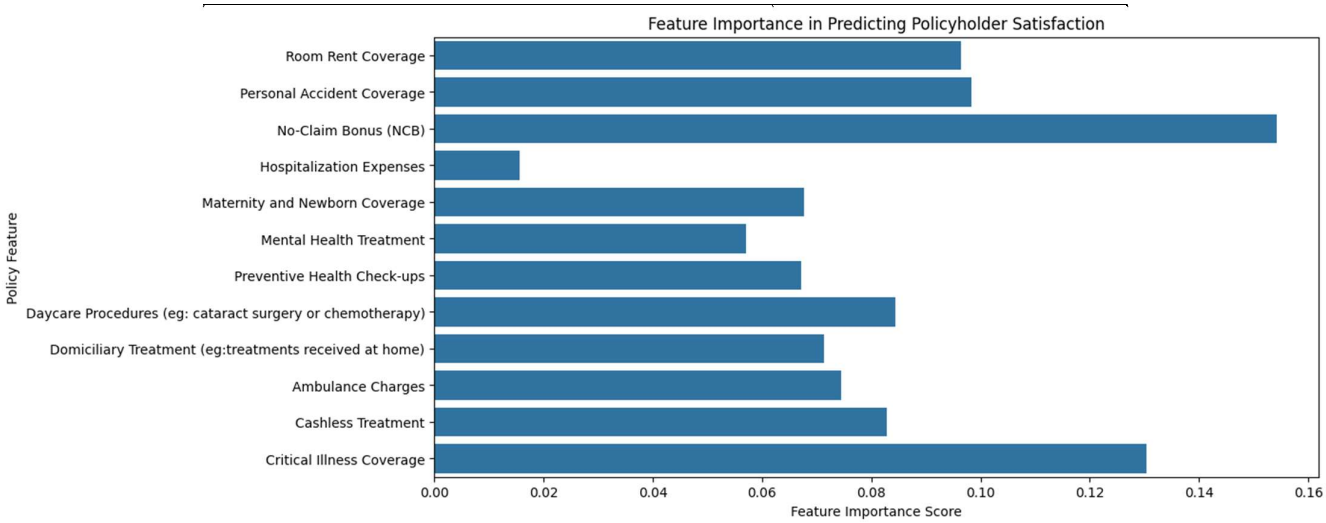


Figure (4.3) Feature importance scores from Random Forest model showing key policy attributes influencing satisfaction.

2. Model Performance Analysis

2.1. Confusion Matrix

Actual → Predicted	0	1	2	3
Class 0	15	1	0	0
Class 1	1	17	2	0
Class 2	0	2	8	0
Class 3	0	0	4	4

Accuracy=(TN+TP)/(TN+TP+FN+FP)

Accuracy=(TN+TP)/(TN+TP+FN+FP)

Accuracy = 81.48%

2.2. Classification Report

Class	Precision	Recall	F1-score
0	0.94	0.94	0.94
1	0.85	0.85	0.85
2	0.57	0.80	0.67
3	1.00	0.50	0.67

1) Class 0 (F1 = 0.94): Strong overall Performance

- Precision = 0.94: When the model predicts class 0, it is correct 94% of the time.
- Recall = 0.94: Out of all actual class 0 instances, 94% were correctly identified.

2) Class 1 (F1 = 0.85) → Good Performance

- Precision = 0.85: When predicting class 1, it is correct 85% of the time.
- Recall = 0.85: The model correctly detects 85% of actual class 1 cases.

3) Class 2&3 (F1 = 0.67) → Moderate Performance

The Random Forest model shows strong predictive performance, especially for major satisfaction classes (0 and 1), while some misclassifications occur for classes 2 and 3.

Interpretation

- Customers prioritize seamless medical services.

  Policies offering cashless treatments and critical illness coverage are highly valued.

- Rewarding loyalty increases satisfaction.

  Offering higher No-Claim Bonuses (NCB) encourages customers to stay insured.

- Maternity & newborn benefits are crucial**.**

  Family-centric policies should be strengthened for higher satisfaction.

- Cost-related factors impact policyholder decisions.

Ambulance charges and hospitalization expenses should be optimized for affordability.

Factors to focus in order to improve customer satisfaction :

1. Enhance cashless treatment networks – Partner with more hospitals to expand seamless claims.
2. Improve policy transparency – Clearly communicate coverage benefits to customers.
3. Incentivize long-term policyholders – Increase NCB and offer premium discounts.
4. Expand maternity benefits – Include additional support for new born care.
5. Adjust pricing for competitive advantage – Ensure policy affordability while maintaining profitability.

CONCLUSION:

The Random Forest model successfully predicts customer satisfaction with 81.48% accuracy, highlighting the importance of key policy features. By optimizing these factors, insurance companies can enhance customer experience and loyalty, ultimately improving retention rates.

4.5 CLUSTERING ANALYSIS

Clustering is an unsupervised machine-learning technique used to group similar data points together based on their features. It helps in identifying patterns, segmenting large datasets, and detecting anomalies without requiring labeled data. Clustering is widely used in fields like customer segmentation, image recognition, and anomaly detection.

OBJECTIVE 6

To identify distinct policyholder segments using clustering techniques based on claim patterns and demographics.

PROS

1. Finds Patterns – Helps discover hidden trends in data.
2. Simplifies Data – Groups similar data points to make analysis easier.
3. Better Targeting – Helps in customer segmentation for personalized marketing.
4. Detects Anomalies – Identifies unusual patterns like fraud.
5. Supports Decisions – Aids businesses in making data-driven choices.
6. Optimizes Resources – Helps allocate resources efficiently.
7. Improves Recommendations – Enables better product/service suggestions.

CONS

1. Difficult to Choose Clusters – Hard to decide the right number of groups.
2. Computationally Expensive – Large datasets require more processing power.
3. Algorithm Sensitivity – Different methods give different results.
4. Affected by Noise – Outliers can distort clusters.
5. Hard to Interpret – Requires domain expertise to understand results.
6. High Dimensionality Issues – Too many features can cause poor clustering.
7. Unstable Results – Outcomes may vary based on initialization and parameters.

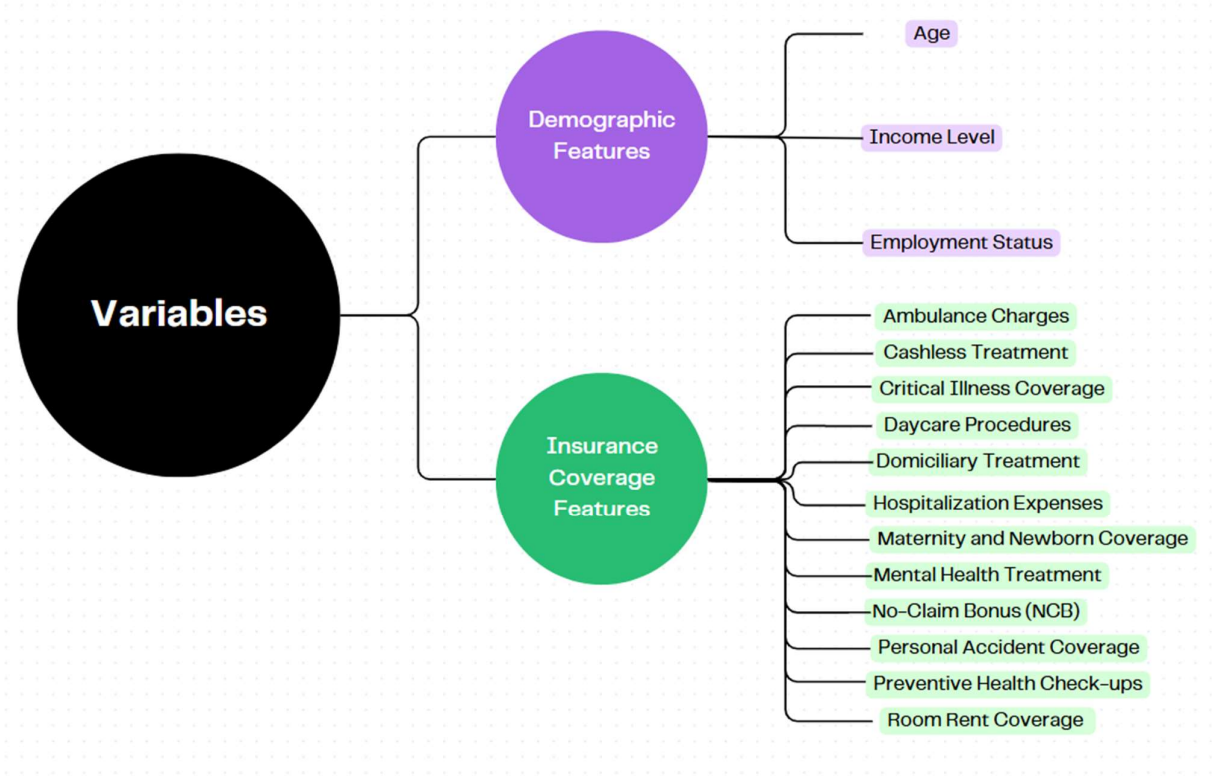
ASSUMPTIONS

1. Homogeneity Within Clusters – Policyholders within the same cluster share similar characteristics.
2. Heterogeneity Between Clusters – Clusters should be distinct from each other.
3. Relevant Feature Selection – Only meaningful features are used to avoid noise.
4. Standardization of Data – Features are scaled to ensure equal influence.
5. Sufficient Data Size – Enough data points are needed for stable clusters.
6. Independence of Features – Highly correlated features should be minimized.
7. Soft Clustering Validity – Fuzzy C-Means allows partial membership in multiple clusters.
8. Cluster Shape and Density – Clusters may have irregular shapes, so soft clustering helps in better segmentation.

CLUSTERING TECHNIQUE USED

This study employs **Fuzzy C-Means (FCM) clustering** to segment policyholders based on claim patterns and demographics. Unlike K-Means**,** which assigns each data point to a single cluster, FCM allows partial membership**,** meaning a data point can belong to multiple clusters with varying probabilities. This flexibility is ideal for overlapping customer segments**,** providing a more realistic classification. By capturing nuanced relationships, FCM enhances risk assessment, personalized policy recommendations, and informed decision-making in insurance underwriting.

FEATURE SELECTION



DATA STANDARDIZATION

Since different variables have different scales, StandardScaler from sklearn.preprocessing was used to transform the data. This ensures all features have a mean of 0 and a standard deviation of 1, preventing any single feature from dominating the clustering results.

DIMENSIONALITY REDUCTION USING PCA

- Applying Principal Component Analysis (PCA)

To enhance computational efficiency and remove redundant information, Principal Component Analysis (PCA) was applied to the dataset. This technique transforms the original features into a new set of uncorrelated variables (principal components) while retaining most of the data's variance. The number of components was chosen to retain 95% of the total variance.

- Results

After applying PCA, the dataset was reduced to 14 principal components from the original 15 features. The cumulative variance retained was **95%**, ensuring that minimal information was lost during dimensionality reduction. This transformation helped improve clustering performance by eliminating noise and redundant features while maintaining essential patterns in the data.

Selecting the Number of Clusters

An optimal number of clusters (k = 2) was chosen based on analysis and business relevance. This was adjusted based on different trials and evaluation metrics.

RESULTS AND VISUALIZATION

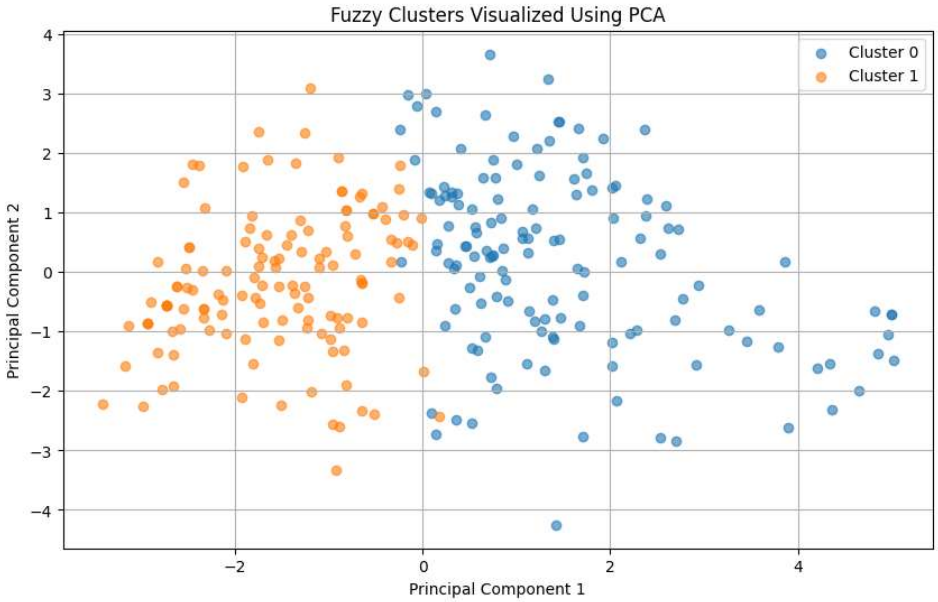


Figure (4.5.1) Fuzzy Cluster Visualization with PCA

Interpretation of the Fuzzy Clustering Visualization

- The two clusters (blue & orange) represent different groups identified by Fuzzy C-Means Clustering.
- The X and Y axes are Principal Components from PCA, reducing high-dimensional data for better visualization.
- Some points overlap, showing fuzzy clustering's soft assignment (data points can belong to both clusters with different probabilities).
- Cluster 0 (blue) likely represents individuals with higher insurance coverage, while Cluster 1 (orange) includes those with basic plans.

Cluster Distribution Analysis

Figure (4.5.2) Cluster Distribution

- The bar chart represents the distribution of data points across two clusters (0 and 1).
- Both clusters have a nearly equal number of data points, indicating a balanced segmentation.
- This suggests that no single cluster dominates, meaning the data is well-separated into two distinct groups.
- A more detailed analysis of cluster characteristics is needed to understand their differences (e.g., demographics, preferences, or behaviors).

CONCLUSION AND INSIGHTS

```
Cluster 0 Features:
 Age                                                  4.826087
Income_Level                                          2.492754
Employment_Status                                     2.905797
Ambulance Charges                                     0.702899
Cashless Treatment                                    0.869565
Critical Illness Coverage                             0.594203
Daycare Procedures (eg: cataract surgery or chemotherapy)    0.572464
Domiciliary Treatment (eg:treatments received at home)       0.340580
Hospitalization Expenses                              0.978261
Maternity and Newborn Coverage                        0.304348
Mental Health Treatment                               0.260870
No-Claim Bonus (NCB)                                  0.724638
Other.1                                               0.152174
Personal Accident Coverage                            0.557971
Preventive Health Check-ups                           0.608696
Room Rent Coverage                                    0.789855
Name: 0, dtype: float64
```

```
Cluster 1 Features:
 Age                                                  4.037879
Income_Level                                          1.931818
Employment_Status                                     2.409091
Ambulance Charges                                     0.128788
Cashless Treatment                                    0.416667
Critical Illness Coverage                             0.181818
Daycare Procedures (eg: cataract surgery or chemotherapy)    0.121212
Domiciliary Treatment (eg:treatments received at home)       0.030303
Hospitalization Expenses                              0.893939
Maternity and Newborn Coverage                        0.045455
Mental Health Treatment                               0.030303
No-Claim Bonus (NCB)                                  0.204545
Other.1                                               0.106061
Personal Accident Coverage                            0.227273
Preventive Health Check-ups                           0.128788
Room Rent Coverage                                    0.340909
Name: 1, dtype: float64
```

Cluster 0 Characteristics:	**Cluster 1 Characteristics:**
Lower Age (4.04)	Lower Age (4.04)
Higher Income Level (2.49)	Lower Income Level (1.93)
More Stable Employment (2.91)	Less Stable Employment (2.41)
Greater Utilization of Health Insurance Benefits:	Lower Usage of Health Insurance Benefits:
Higher Ambulance Charges (0.70)	Lower Ambulance Charges (0.13)
More Cashless Treatment (0.87)	Less Cashless Treatment (0.42)
Greater Personal Accident Coverage (0.56)	Fewer Preventive Health Check-ups (0.13)
Better Room Rent Coverage (0.79)	Less Room Rent Coverage (0.34)
Higher Critical Illness Coverage (0.59)	Lower Critical Illness Coverage (0.18)
More No-Claim Bonus (NCB) (0.72)	Lower No-Claim Bonus (NCB) (0.20)

Interpretation:

Cluster 0 consists of older, financially stable individuals with higher income and employment stability. They actively use health insurance benefits, including cashless treatments, ambulance services, and critical illness coverage.

Cluster 1 consists of younger individuals with lower income and employment stability. They use fewer insurance benefits, indicating they may opt for minimal coverage or have lower healthcare needs.

```
Top 10 Differentiating Features Between Clusters:
 Age                                                                0.788208
Ambulance Charges                                                   0.574111
Income_Level                                                        0.560935
No-Claim Bonus (NCB)                                                0.520092
Employment_Status                                                   0.496706
Preventive Health Check-ups                                         0.479908
Cashless Treatment                                                  0.452899
Daycare Procedures (eg: cataract surgery or chemotherapy)           0.451252
Room Rent Coverage                                                  0.448946
Critical Illness Coverage                                           0.412385
Name: 1, dtype: float64
```

The features that contribute most to separating the clusters are:

1. Age (+0.79 difference) → Older individuals in Cluster 0.
2. Ambulance Charges (+0.57 difference) → Higher in Cluster 0, indicating more usage of emergency services.
3. Income Level (+0.56 difference) → Higher in Cluster 0.
4. No-Claim Bonus (NCB) (+0.52 difference) → Cluster 0 receives more NCB, implying fewer claims or better coverage.
5. Employment Status (+0.50 difference) → More stable in Cluster 0.
6. Preventive Health Check-ups (+0.48 difference) → Cluster 0 takes better preventive care.
7. Cashless Treatment (+0.45 difference) → Cluster 0 utilizes this more.
8. Daycare Procedures (+0.45 difference) → More frequent in Cluster 0.
9. Room Rent Coverage (+0.45 difference) → Cluster 0 has better room rent benefits.
10. Critical Illness Coverage (+0.41 difference) → Cluster 0 has more comprehensive protection.

Interpretation:

1. Marketing & Insurance Strategy
   Cluster 0: Can be targeted for premium insurance plans with added benefits.

   Cluster 1: Needs affordable, entry-level insurance plans with flexible payment options.
2. Healthcare Planning
   Cluster 0 likely requires more frequent healthcare services. Cluster 1 may benefit from awareness programs to encourage preventive care.

MULTINOMIAL LOGISTIC REGRESSION WITH L2 (RIDGE) REGULARIZATION

Multinomial logistic regression is an extension of binary logistic regression used when the dependent variable has more than two categorical outcomes. In this case, the dependent variable, **"Switch To Government,"** represents different categories of switching behavior.

Since simple linear regression is not suitable for categorical dependent variables, multinomial logistic regression is applied to model the relationship between independent variables (predictors) and the multinomial outcome.

Traditional multinomial logistic regression can suffer from overfitting when predictors are highly correlated. L2 (Ridge) Regularization helps:

- Prevent overfitting by penalizing large coefficients.
- Improve model stability when predictors are correlated.
- Enhance generalization for better performance on unseen data

OBJECTIVE 5

To assess the impact of various factors last claim status, overall satisfaction, policy premium, area type, employment status, and dependents on the likelihood of switching from private to government health insurance.

PROS

1. Handles Multi-Class Problems
2. No Need for Feature Scaling
3. Probabilistic Interpretation
4. Handles Categorical and Continuous Features
5. Works Well with Linear Relationships

CONS

1. Assumes Independence of Irrelevant Alternatives (IIA)
2. Sensitive to Outliers
3. Computationally Expensive
4. Requires Large Sample Size
5. Struggles with Non-Linear Relationships

ASSUMPTIONS

1. No Multicollinearity
2. Linearity of Log-Odds
3. Independence of Observations
4. Sufficient Sample Size
5. Independence of Irrelevant Alternatives (IIA)

MULTINOMIAL LOGISTIC MODEL

The probability of switching to government is modelled as:

$$P(Y=j|X) = \frac{e^{(\beta 0j+\beta 1jX1+\beta 2jX2+\ldots+\beta njXn}}{1+\sum_{k=1}^{k-1} e^{(\beta 0k+\beta 1kX1+\beta 2kX2 \ \ldots+\beta nkX \ )}}$$

Where:

π(x): Probability of switching to government.

- **Y:** Dependent variable (Switching to government) yes = 1, No= 2, Maybe=3).
- **X_1, X_2, X_3, X4, X5, X6:** Independent variables (Last Claim Status, Overall Satisfaction, Policy Premium, Area Type, Employment Status, Dependents).
- **β_0:** Intercept.

  β_1, β_2, β_3,….. : Coefficients of the predictors.

Log-Odds (Logit Transformation)

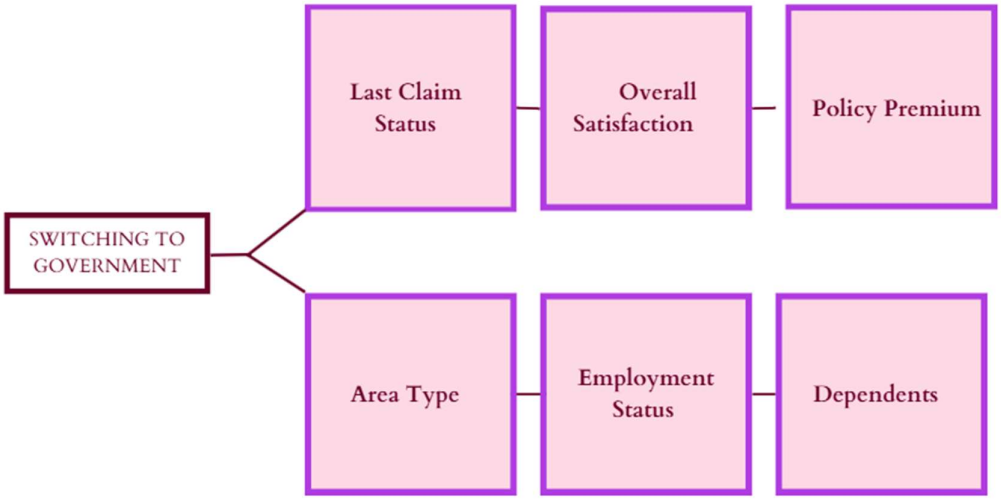
Since probabilities range from 0 to 1, taking the **logarithm of odds** transforms it into an **unbounded scale**.

$$\log(\frac{P(Y=1)}{P(Y=j)}) \ \beta 0j + \beta 1jX1 + \beta 2jX2+\ldots+\beta njXn$$

where:

- β0j is the intercept for category jjj.
- β1j, β2j,…,… are the coefficients for predictors.

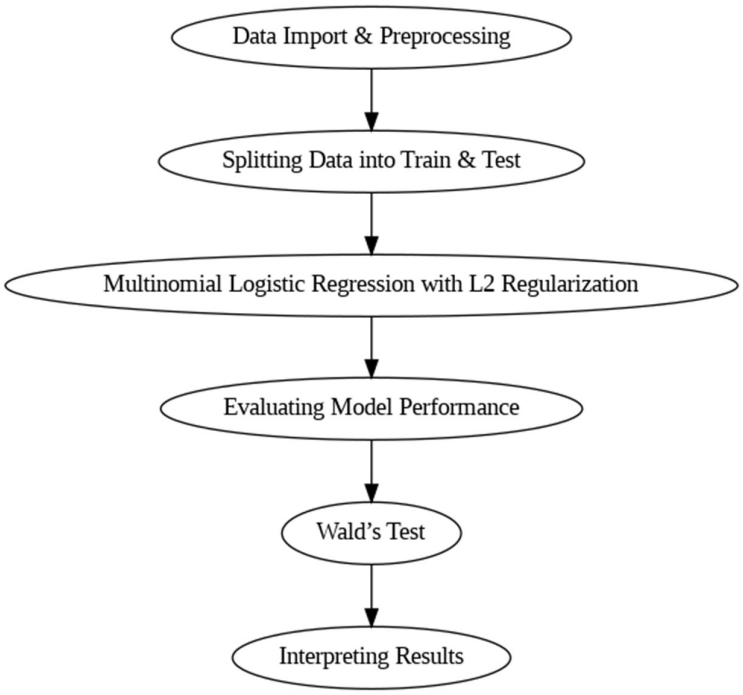
VARIABLE IDENTIFICATION



VARIABLES ENCODED

Switch To Government	1 = Yes, 2 = No, 3 = Maybe
Last Claim Status	0, 1, 2
Overall Satisfaction	1, 2, 3, 4, 5
Policy Premium	1, 2, 3, 4
Area Type	0 = Rural, 1 = Urban
Employment Status	1, 2, 3, 4, 5, 6
Dependents	0, 1, 2, 3, 4

STEPS TO BE IMPLEMENTED IN MULTINOMIAL LOGISTIC REGRESSION



Hypothesis

H0: There is **no significant relationship** between the predictor variables (**Last_Claim_Status, Overall_Satisfaction, Policy_Premium, Area_Type, Employment_Status, Dependents**) and the likelihood of switching to government insurance.

HA: At least one predictor variable has a **significant effect** on switching behavior.

CHECKING THE ASSUMPTIONS

Multicollinearity

	Feature	VIF
0	Last_Claim_Status	1.668632
1	Overall_Satisfaction	8.901672
2	Policy_Premium	7.820603
3	Area_Type	2.722411
4	Employment_Status	4.145344
5	Dependents	2.790424

Since, all the VIF's are less than 10, we can conclude that there is no multicollinearity present between the independent variables.

DATA ANALYSIS & INTERPRETATION:

Model Performance Summary

Metric	Score
Accuracy	64.81%
AUC Score	0.81

- **Accuracy**: **64.81%** – The model correctly classifies **64.81%** of the test data.
- **AUC Score**: **0.81** – The model has strong discriminatory power, meaning it effectively distinguishes between different classes of "Switch_To_Government."

Wald's Test

Feature	p-value
Last Claim Status	0.0005
Overall Satisfaction	0.0001
Policy Premium	0.027
Area Type	0.134
Dependents	0.008
Employment Status	0.095

INTERPRETATION:

Last Claim Status (p = 0.0005)
• Since p < 0.05, Last Claim Status significantly affects the dependent variable (insurance claim). This suggests that the status of an individual's previous claim is an important predictor of whether they will make a future claim.

Overall Satisfaction (p = 0.0001)
• Since p < 0.05, Overall Satisfaction has a significant impact on the likelihood of making a claim. This indicates that customer satisfaction levels may influence insurance claim behavior.

Policy Premium (p = 0.027)
• Since p < 0.05, Policy Premium significantly affects the dependent variable. This implies that the cost of the policy premium plays a role in determining whether an individual files an insurance claim.

Area Type (p = 0.134)
• Since p > 0.05, we fail to reject H_0, meaning Area Type does not significantly impact the likelihood of making a claim.

Dependents (p = 0.008)
• Since p < 0.05, Dependents significantly influence the dependent variable. This suggests that the number of dependents an individual has is an important factor in predicting insurance claims.

Employment Status (p = 0.095)
• Since p > 0.05, we fail to reject H_0, meaning Employment Status does not significantly impact the likelihood of making a claim.

Classification Report Breakdown

Class (Switch To Government)	**Precision**	**Recall**	**F1-score**	**Support**
1 (Low likelihood of switching)	0.50	0.22	0.31	9
2 (Moderate likelihood of switching)	0.64	0.80	0.71	20
3 (High likelihood of switching)	0.68	0.68	0.68	25

- Class 2 (Moderate Switching) has the highest recall (0.80), meaning the model is good at detecting moderate switchers.
- Class 3 (High Switching) has balanced precision and recall (~0.68), suggesting the model performs well in predicting high switchers.
- Class 1 (Low Switching) has low recall (0.22), meaning the model struggles to correctly classify people who are unlikely to switch.

Interpretation :

- Customer satisfaction is the strongest factor—insurers should focus on improving customer experience to reduce switching rates.
- Higher premiums increase churn, suggesting the need for competitive pricing strategies.
- Last claim status is important—insurers may need to improve claim handling to retain customers.
- Area type does not matter, so marketing strategies should be customer-focused rather than location-based.

Chapter 5: Conclusion and Future Prospects

KNN Algorithm

- The K-Nearest Neighbors (KNN) model achieves an accuracy of 81.48%, indicating strong predictive performance.
- It is particularly effective at identifying "Claim" cases, with a high recall of 94%

CHI SQUARE TEST

- The results (p = 0.3541) indicate no significant association between area type (urban vs. rural) and claim behavior.
- This suggests that policyholders from both areas make claims at similar rates.
- Since location does not impact claim likelihood, insurers should focus on other factors, such as demographics or policy features, to refine risk assessment strategies.

CHI SQUARE TEST

- The results indicate a significant association between area type and claim settlement status.
- Urban areas have higher claim approvals (87 vs. 65 in rural areas) and more pending claims (70 vs. 32), while rejections are slightly higher in rural areas.
- The Insurers must focus on streamlining claim processing in urban areas to reduce pending claims and improve outreach in rural areas to address higher rejection rates.

RANDOM FOREST

- The model predicts policyholder satisfaction with 81.48% accuracy, identifying cashless treatment, No-Claim Bonus (NCB), and critical illness coverage as key drivers.
- Policies with better maternity benefits and cost-efficient hospitalization coverage also enhance satisfaction.
- Insurers should expand cashless networks, improve transparency, offer higher NCB rewards, enhance maternity benefits, and optimize pricing to boost customer loyalty and retention.

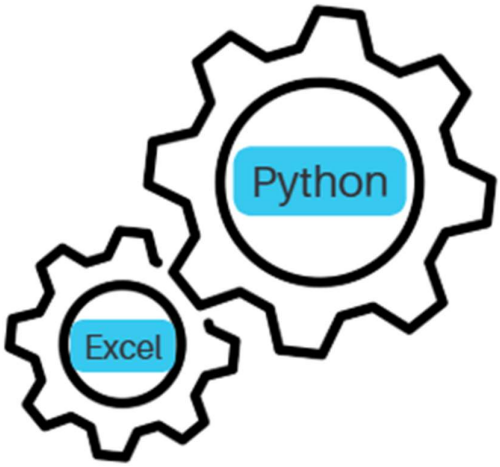
CLUSTER ANALYSIS

- Cluster 0 consists of older, financially stable individuals who actively use health insurance benefits, while Cluster 1 includes younger, lower-income individuals with minimal insurance usage.
- Key factors driving cluster separation include age, income, employment stability, and healthcare utilization.
- Insurers can offer premium plans with added benefits for Cluster 0 and affordable, flexible plans for Cluster 1, along with awareness programs to encourage preventive care.

MULTINOMIAL LOGISTIC REGRESSION

- The model with L2 regularization predicts switching to government insurance with 64.81% accuracy and an AUC of 0.81, indicating strong discriminatory power.
- Policyholders with previous claims and lower satisfaction are most likely to switch, while policy premium has a slight positive effect.
- Employment status shows a minor influence, whereas area type has no significant impact.
- The model performs well for moderate and high switchers but struggles to classify low switchers accurately.

SOFTWARES USED & CODES



K-NEAREST NEIGHBORS

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
data = pd.read_excel("dataset.xlsx")  # Update with actual path
X = data.iloc[:, :-1].values  # Features
y = data.iloc[:, -1].values  # Target (Made Claim)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"KNN Model Accuracy: {accuracy:.4f}")
```

```
print("Classification Report:\n", classification_report(y_test, y_pred, target_names=["No Claim", "Claim"]))

sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d', xticklabels=["No Claim", "Claim"], yticklabels=["No Claim", "Claim"])

plt.xlabel("Predicted")

plt.ylabel("Actual")

plt.title("Confusion Matrix - KNN (k=5)")

plt.show()

neighbors = np.arange(1, 21)

train_acc = []

test_acc = []

    knn = KNeighborsClassifier(n_neighbors=k)

    knn.fit(X_train, y_train)

    train_acc.append(knn.score(X_train, y_train))

    test_acc.append(knn.score(X_test, y_test))


plt.plot(neighbors, test_acc, marker='o', label="Test Accuracy")

plt.plot(neighbors, train_acc, marker='s', label="Train Accuracy", linestyle='dashed')

plt.xlabel("Number of Neighbors (k)")

plt.ylabel("Accuracy")

plt.title("KNN Accuracy for Different k Values")

plt.legend()

plt.show()
```

MULTINOMIAL LOGISTIC REGRESSION

```
import pandas as pd

import numpy as np

from sklearn.model_selection import train_test_split, cross_val_score

from sklearn.preprocessing import StandardScaler

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score, classification_report, roc_auc_score

X = df[["Last_Claim_Status", "Overall_Satisfaction", "Policy_Premium", "Area_Type",
"Employment_Status", "Dependents"]]

y = df["Switch_To_Government"]X_scaled = StandardScaler().fit_transform(X)X_train, X_test,
y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

model = LogisticRegression(penalty='l2', solver='liblinear', random_state=42)

model.fit(X_train, y_train)

y_pred = model.predict(X_test)

print(f"Accuracy: {accuracy_score(y_test, y_pred):.4f}")

print(f"AUC: {roc_auc_score(y_test, model.predict_proba(X_test), multi_class='ovr'):.4f}")

print(classification_report(y_test, y_pred))

cv_scores = cross_val_score(model, X_scaled, y, cv=5, scoring='accuracy')

print(f"CV Mean Accuracy: {np.mean(cv_scores):.4f} ± {np.std(cv_scores):.4f}"
```

CHI SQ FOR CLAIM SETTLEMENT

```
import pandas as pd

import scipy.stats as stats

import seaborn as sns

import matplotlib.pyplot as plt

contingency = pd.crosstab(df["Area_Type"], df["Last_Claim_Status"])

chi2, p, dof, expected = stats.chi2_contingency(contingency)

print(f"Chi-Square: {chi2:.4f}, P-value: {p:.4f}, DF: {dof}")

print("Expected Frequencies:\n", expected)

sns.heatmap(contingency, annot=True, fmt="d", cmap="Blues")

plt.title("Claim Settlement: Rural vs Urban")

plt.xlabel("Last Claim Status") plt.ylabel("Area Type") plt.show()
```

CHI SQ FOR CLAIM BEHAVIOUR

```
import pandas as pd
import scipy.stats as stats
df.rename(columns=lambda x: x.strip().replace(" ", "_").lower(), inplace=True)
contingency_table = pd.crosstab(df['area_type'], df['made_claim'])
chi2, p, dof, expected = stats.chi2_contingency(contingency_table)
print("Chi-Square Test Results:")
print(f"Chi-Square Statistic: {chi2:.4f}")
print(f"Degrees of Freedom: {dof}")
print(f"P-Value: {p:.4f}")
```

RANDOM FOREST

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Define target and predictors
target = "Overall_Satisfaction"
predictors = ["Room Rent Coverage", "Personal Accident Coverage", "No-Claim Bonus (NCB)",
    "Hospitalization Expenses", "Maternity and Newborn Coverage", "Mental Health Treatment",
    "Preventive Health Check-ups", "Daycare Procedures (eg: cataract surgery or chemotherapy)",
    "Domiciliary Treatment (eg:treatments received at home)", "Ambulance Charges",
    "Cashless Treatment", "Critical Illness Coverage"]
df[target] = df[target].astype("category").cat.codes

# Split data
X_train, X_test, y_train, y_test = train_test_split(df[predictors], df[target], test_size=0.2, random_state=42)

# Train model
rf = RandomForestClassifier(n_estimators=100, max_depth=10, random_state=42).fit(X_train, y_train)
```

```
# Predictions & Evaluation
y_pred = rf.predict(X_test)
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
print("\nModel Accuracy:", accuracy_score(y_test, y_pred))

# Hyperparameter Tuning
grid = GridSearchCV(RandomForestClassifier(random_state=42), {
    "n_estimators": [50, 100, 200], "max_depth": [5, 10, 20], "min_samples_split": [2, 5, 10]
}, cv=5, scoring="accuracy", n_jobs=-1).fit(X_train, y_train)
print("\nBest Parameters:", grid.best_params_)

# Feature Importance
plt.figure(figsize=(12, 6))
sns.barplot(x=rf.feature_importances_, y=predictors)
plt.xlabel("Feature Importance Score")
plt.ylabel("Policy Feature")
plt.title("Feature Importance in Predicting Policyholder Satisfaction")
plt.show()
```

CLUSTERING ANALYSIS

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import skfuzzy as fuzz

# Selected features for clustering
features = ['Age', 'Income_Level', 'Employment_Status', 'Ambulance Charges',
            'Cashless Treatment', 'Critical Illness Coverage', 'Daycare Procedures',
            'Domiciliary Treatment', 'Hospitalization Expenses', 'Maternity and Newborn
Coverage',
            'Mental Health Treatment', 'No-Claim Bonus (NCB)', 'Personal Accident Coverage',
            'Preventive Health Check-ups', 'Room Rent Coverage']

# Standardize features
X_scaled = StandardScaler().fit_transform(df[features])
```

```
# PCA for dimensionality reduction (95% variance retained)
X_pca = PCA(n_components=0.95, random_state=42).fit_transform(X_scaled)
print(f"Selected {X_pca.shape[1]} Principal Components")

# Fuzzy C-Means clustering with optimal clusters (adjust if needed)
cntr, u, _, _, _, _, _ = fuzz.cluster.cmeans(X_pca.T, c=2, m=2, error=0.005, maxiter=1000)
df['Cluster'] = np.argmax(u, axis=0)  # Assign cluster labels

# Save results
df.to_csv("/content/fuzzy_clustered_data.csv", index=False)
print("Fuzzy clustering complete! Data saved.")

# Visualization of clusters in PCA space
plt.figure(figsize=(10, 5))
for cluster in range(2):
    plt.scatter(X_pca[df['Cluster'] == cluster, 0], X_pca[df['Cluster'] == cluster, 1],
label=f"Cluster {cluster}", alpha=0.6)
plt.xlabel("PC1"), plt.ylabel("PC2"), plt.title("Fuzzy Clusters"), plt.legend(), plt.grid()
plt.show()

# Cluster distribution
sns.countplot(x=df['Cluster'], palette='viridis')
plt.xlabel("Cluster"), plt.ylabel("Count"), plt.title("Cluster Distribution"), plt.grid()
plt.show()
```

QUESTIONNAIRE

Do you currently have a medical insurance policy?

- Yes
- No

(If No)

1) Age:

- 1-10
- 11-20
- 21-30
- 31-40
- 41-50
- 51-60
- 61-70
- Greater than 70

2) Gender:

- Male
- Female
- Other

3) What type of area do you live in?

- Rural
- Urban

4) What is the primary reason for not having a medical insurance policy?

- Too expensive
- Lack of awareness
- Employer provides coverage
- Don't feel the need for it
- Other

(If Yes)

1) Age:

- 1-10
- 11-20

- 21-30
- 31-40
- 41-50
- 51-60
- 61-70
- Greater than 70

2) Gender:

- Male
- Female
- Other

3) Marital Status:

- Single
- Married
- Divorced
- Widowed

4) Education Level:

- High School
- Undergraduate
- Graduate
- Postgraduate
- Other

5) Income Level:

- Less than ₹2,50,000/year
- ₹2,50,000 - ₹5,00,000/year
- ₹5,00,000 - ₹10,00,000/year
- More than ₹10,00,000/year

6) How many number of people are dependent on you?

- 1
- 2
- 3
- 4
- 5

7) What type of area do you live in?

- Rural
- Urban

8) Employment Status:

- Student
- Employed
- Self-Employed
- Unemployed
- Retired
- Homemaker
- Other

9) Do you have any long-term illnesses or conditions (e.g., diabetes, hypertension, heart disease)?

- Yes
- No

10) How often do you engage in the following habits?

Habit	Never	Occasionally	Daily	Frequently
Smoking				
Alcohol				
Physical Activity				

11) What type of health insurance policy do you have?

- Individual
- Family
- Employer-Sponsored

12) How much do you pay as your policy premium (in ₹)?

- Less than ₹5,000
- ₹5,000 - ₹10,000
- ₹10,000 - ₹20,000
- ₹20,000 - ₹50,000
- More than ₹50,000

13) How often is your policy renewed?

- Monthly
- Quarterly
- Semi-Annually
- Annually

14) What is the name of your health insurance policy?

15) How long has your policy been active**?**

- Less than 1 year
- 1-3 years
- 3-5 years
- More than 5 years

16) Have you made any claims under your current policy?

- Yes
- No

(If Yes)

1) How many claims have you made under this policy?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

2) What was the amount of your most recent claim (in ₹)?

- Less than ₹5,000
- ₹5,000 - ₹10,000
- ₹10,000 - ₹25,000
- ₹25,000 - ₹50,000
- More than ₹50,000

3) What is the current status of your most recent claim?

- Approved
- Denied
- Pending

4) What type of coverage options does this policy hold?

- Hospitalization Expenses

- Daycare Procedures (eg: cataract surgery or chemotherapy)
- Ambulance Charges
- Domiciliary Treatment (eg: treatments received at home)
- Cashless Treatment
- Preventive Health Check-ups
- Critical Illness Coverage
- Maternity and Newborn Coverage
- Mental Health Treatment
- Room Rent Coverage
- Personal Accident Coverage
- No-Claim Bonus (NCB)
- Other

5) How satisfied are you with your current private healthcare policy?
(Please rate on a scale of 1 to 5, where 1 = Very Dissatisfied, 2= Dissatisfied, 3= Neutral, 4= Satisfied and 5 = Very Satisfied)

6) Which factors, if improved, would enhance your satisfaction with private healthcare policies?

- Cost of premiums
- Coverage for specific medical conditions
- Claim processing time
- Availability of network hospitals
- Customer service
- Other

7) Would you consider switching from private to government healthcare policies in the future?

- Yes, I am considering it.
- No, I prefer to stay with my private policy.
- Maybe, depending on the benefits offered by government policies.

(If No)

1) What type of coverage options does this policy hold?

- Hospitalization Expenses
- Daycare Procedures (eg: cataract surgery or chemotherapy)

- Ambulance Charges
- Domiciliary Treatment (eg: treatments received at home)
- Cashless Treatment
- Preventive Health Check-ups
- Critical Illness Coverage
- Maternity and Newborn Coverage
- Mental Health Treatment
- Room Rent Coverage
- Personal Accident Coverage
- No-Claim Bonus (NCB)
- Other

2) How satisfied are you with your current private healthcare policy?
(Please rate on a scale of 1 to 5, where 1 = Very Dissatisfied, 2= Dissatisfied, 3= Neutral, 4= Satisfied and 5 = Very Satisfied)

3) Which factors, if improved, would enhance your satisfaction with private healthcare policies?

- Cost of premiums
- Coverage for specific medical conditions
- Claim processing time
- Availability of network hospitals
- Customer service
- Other

4) Would you consider switching from private to government healthcare policies in the future?

- Yes, I am considering it.
- No, I prefer to stay with my private policy.
- Maybe, depending on the benefits offered by government policies.

REFERENCES

Research Papers & Articles

- https://www.cureusjournals.com/articles/3010-predictive-precision-unraveling-health-insurance-claim-patterns-with-logistic-regression-and-decision-trees?utm_source=chatgpt.com#!/

https://www.researchgate.net/publication/369436192_Prediction_for_Insurance_Premiums_Based_on_Random_Forest_and_Multiple_Linear_Regression/download

Industry Reports & Websites:

- **Insurance Industry Performance in India**
  ResearchGate Link
- **Health Insurance Trends**
  Forbes Advisor
- **Statistical Techniques in Insurance Analytics**
  Emerald Insight

Statistical & Machine Learning Methods:

- **Chi-Square Tests and Hypothesis Testing**
  Lardbucket.org
- **Random Forests and Decision Trees in Risk Estimation**
  ResearchGate
- **Fuzzy Clustering Algorithms for Insurance Data Segmentation**
  ResearchGate
- **Multivariate Statistical Analysis**
  Applied Multivariate Statistics – Johnson & Wichern
- **K-Nearest Neighbors Algorithm**
  GeeksForGeeks

SOMAIYA VIDYAVIHAR UNIVERSITY
S K Somaiya College

Somaiya TRUST

Progress report for the Month of Oct 2024 – Mar 2025			
Date of report :	Guide's signature	Internal mentor signature	Roll Number :
			31031823001 31031823008
Name of student	Tanvi Akre Vidhi Murdeshwar		
Leave/early off taken by the student during the reporting period	Number of days	With prior permission	Reason
Current status of project : October -2024 November -2024 December -2024 January -2025 February -2025 March -2025	1.Introduction and guidance for collecting and reading the research papers as directed by Prof. Jyoti Mantri 2. Shortlisted topic choices 3. Finalised project topic 1. Specified the Aim and Objectives of the project 2. Worked on framing a sample questionnaire 3. Discussed the techniques and methodology to be used. 4. Finalised the questionnaire 1. Designed google form 2. Conducted a pilot survey 3. Analysed the pilot survey's results 4. Accordingly made changes in the questionnaire 1. Circulated updated questionnaire and collected data. 1. Studied and analysed the data using different statistical techniques 1. Report generation 2. Final submission		
Planned work executed in time	YES	No, Reason	
Remarks by guide			