COMPREHENSIVE ANALYSIS OF CM - HEALTH INSURANCE DATA IN TAMILNADU

A PROJECT REPORT

Submitted by

DHARSHAN. M (211420104061)

BRAGADEESHWARAN. R (211420104043)

YOKESH. S (211420104315)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

APRIL 2024

PANIMALAR ENGINEERING COLLEGE

(An Autonmous Institution, Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this project report **"Comprehensive Analysis of CM-Health Insurance Data In Tamilnadu"** is the bonafide work of **"DHARSHAN. M (211420104061) , BRAGADEESHWARAN. R (211420104043) & YOKESH.S" (211420104315)** who carried out the project work under my supervision.

Signature of the HOD with date

Dr L. JABASHEELA M.E., Ph.D.,
Professor and Head,
Department of Computer Science and Engineering,
Panimalar Engineering College,
Chennai – 123

Signature of the Supervisor with date

Dr N. PUGHAZENDI M.E., Ph.D.,
Professor,
Department of Computer Science and Engineering,
Panimalar Engineering College,
Chennai - 123

Submitted for the Project Viva – Voce examination held on _____

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION BY THE STUDENT

We DHARSHAN. M (211420104061) & BRAGADEESHWARAN. R (211420104043) & YOKESH. S (211420104315) hereby declare that this project report titled "**Comprehensive Analysis of CM-Health Insurance Data In TAMILNADU**", under the guidance of **Dr.N.Pughazendi, M.E., Ph.D.,**is the original work done by us and we have not plagiarized or submitted to any other degree in any university by us.

ACKNOWLEDGEMENT

Our profound gratitude is directed towards our esteemed Secretary and Correspondent, **Dr. P. CHINNADURAI, M.A., Ph.D**., for his benevolent words and fervent encouragement. His inspirational support proved instrumental in galvanizing our efforts, ultimately contributing significantly to the successful completion of this project.

We wanttoexpressourdeepgratitude to our Directors,**Tmt. C. VIJAYARAJESWARI, Dr. C. SAKTHI KUMAR, M.E., Ph.D., and Dr. SARANYASREE SAKTHI KUMAR, B.E., M.B.A., Ph.D.,** for graciously affording us the essential resources and facilities for undertaking this project

Our gratitude is also extended to our Principal, **Dr. K. MANI, M.E., Ph.D.,** whose facilitation proved pivotal in the successful completion of this project.

We express my heartfelt thanks to **Dr. L. JABASHEELA, M.E., Ph.D.,** Head of the Department of Computer Science and Engineering, for granting the necessary facilities that contributed to the timely and successful completion of project.

We would like to express our sincere thanks to our parents, friends, project coordinator and guide **Dr. N. PUGHAZENDI, M.E., Ph.D.,** and all the faculty members of the Department of CSE for their unwavering support for the successful completion of the project.

DHARSHAN. M (211420104063)

BRAGADEESHWARAN. R (211420104043)

YOKESH. S (211420104315)

ABSTRACT

The landscape of health insurance claims in Tamil Nadu presents a complex challenge marked by the rise of fraudulent activities and inefficiencies in resource allocation. This engineering thesis unfolds an extensive analysis of health insurance claims data, enriched by advanced analytical frameworks to streamline operations and reinforce the system's reliability. Employing a meticulously curated dataset encompassing claims until November 30, 2023, our research integrates sophisticated statistical models, visual analytics, and predictive machine learning algorithms, predominantly executed in Python with its powerful libraries such as NumPy, pandas, scikit-learn, Matplotlib, and Seaborn. Our methodology adopts logistic regression to pinpoint fraud, while linear regression is utilized to forecast claim patterns, thus equipping stakeholders with actionable insights for preemptive decision-making. An in-depth investigation into equipment and hospital claim distributions provides a granular perspective on resource utilization, revealing disparities and enabling targeted interventions. Key findings illustrate distinct temporal trends, geographic variability in claims, and aberrant patterns potentially indicative of fraud, thereby informing the construction of a predictive model poised to preempt fraudulent claims. The proposed model offers a robust foundation for optimizing the allocation of health insurance resources, enhancing preventive planning measures, and augmenting the overall efficiency of the health insurance framework in the region. This thesis encapsulates the project's lifecycle, from conception through implementation, confronting challenges, and adapting to the evolving healthcare landscape. The culmination of this research offers a paradigm for data-driven governance in health insurance, with potential scalability to broader contexts, and underscores the pivotal role of engineering innovation in shaping the future of healthcare administration.

TABLE OF CONTENTS

CHAPTER NUMBER		TITLE	PAGE NUMBER
		ABSTRACT	V
		LIST OF FIGURES	XI
		LIST OF ABBREVIATIONS	XII
1		**INTRODUCTION**	1
	1.1	Analysing Health Insurance Claims	2
	1.2	Leveraging Advanced Data Analysis Techniques	2
	1.3	Equipment and Hospital-wise Utilization Focus	3
	1.4	Fraudulent Claims Identification	4
	1.5	Optimizing Insurance Utilization	5
	1.6	Strategic Fraud Detection Methodologies	5
	1.7	Collaboration with Healthcare Stakeholders	6
2		**LITERATURE REVIEW**	8
3		**SYSTEM ANALYSIS**	10

	3.1	Analyse Trends Over Time	10
	3.2	Equipment – wise and Hospital – wise Evaluation	10
	3.3	Fraud Detection and Prevention	11
	3.4	Machine Learning for Predictive Modeling	11
	3.5	Optimization of Health Insurance Utilization	12
4		**MODULE DESIGN**	13
	4.1	Use Case Diagram	13
	4.2	Class Diagram	14
	4.3	Sequence Diagram	15
	4.4	Activity Diagram	16
	4.5	DFD Diagram	17
	4.6	System Architecture	18
5		**DATA DESCRIPTION**	20
	5.1	Dataset Composition	20
	5.2	Analytical Potential	21

6		**ABOUT THE DATA**	22
	6.1	Data Collection	22
	6.2	Data Preprocessing	22
	6.3	Exploratory Data Analysis (EDA)	22
	6.4	Trend Analysis	22
	6.5	Equipment and Hospital-wise Evaluation	23
	6.6	Fraud Detection	23
	6.7	Optimization of Insurance Utilization	23
7		**SAMPLE DISTRIBUTION ANALYSIS**	24
	7.1	Claim Amounts Distributions	24
	7.2	Claimant Ages Distributions	24
	7.3	Interpretation and Implications	24
8		**EXPLORATORY DATA ANALYSIS**	25
	8.1	Age vs. Gender – Box Plot	25
	8.2	Age vs. Gender – Distribution	26
	8.3	Preauth amount vs. Final Approved Amount vs. City	27

	8.4	Final Approved Amount vs. Govt Hospitals / Private Hospitals	28
	8.5	Year vs. Final Approved Amount vs. Preauth Amount	29
	8.6	Preauth Amount vs. Final Approved Amount vs. Year	30
	8.7	Patient Age vs. Government/Private Hospitals-Box Plat	31
	8.8	Patient Age vs. Government/Private Hospitals-Distribution	32
	8.9	Year vs. Final submission Date – Histogram	33
9		**MACHINE LEARNING ALGORITHMS**	34
	9.1	Clustering -K- Means Algorithm	34
	9.2	Regression	37
	9.3	Test Cases	38
10		**SYSTEM IMPLEMENTATION**	40
11		**CONCLUSION AND FUTURE ENHANCEMENT**	51
12		**REFERENCES**	53

13 **APPENDICES** 55

A.1 Screen Shot 55

A.2 Plagiarism Report 59

LIST OF FIGURES

FIG.NO	DESCRIPTION	PAGE NO
8.1.1	Age vs. Gender – Box Plot	24
8.2.1	Age vs. Gender – Distribution	25
8.3.1	Preauth Amount vs. Final Approved Amount vs. City	26
8.4.1	Final Approved Amount vs. Govt Hospitals/Private Hospitals	27
8.5.1	Year vs, Final Approved Amount vs Preauth Amount	28
8.6.1	Preauth Amount vs. Final Amount vs. Year	29
8.7.1	Patient Age vs. Government / Private Hospitals – Box Plot	30
8.9.1	Year vs. First Submission Date-Histogram	31
9.1.1	Clustering – K – Means Algorithm	32
9.1.2	K-Means Clustering	34
9.1.3	Graph	35
9.2.1	Regression Model	37

LIST OF ABBREVIATION

ABBREVIATION	DEFINATION
UCD	Use Case Diagram
CD	Class Diagram
SD	Sequence Diagram
AD	Activity Diagram
DFD	Data Flow Diagram

CHAPTER-1
INTRODUCTION

Health insurance acts as a cornerstone of healthcare systems, providing financial coverage for medical expenses to individuals and thereby facilitating access to healthcare services without the burden of prohibitive costs. This system is predicated on the pooling of risks, where the collective premiums of the insured are used to cover the expenses incurred by members of the pool who require medical care. Health insurance can significantly reduce the direct financial burden on individuals at the point of care, promote preventive healthcare practices, and improve overall public health outcomes.

In essence, health insurance serves a dual purpose: it protects individuals from the financial risks associated with health care costs and ensures that people have access to healthcare when they need it. The coverage offered by health insurance plans can vary widely, including outpatient care, inpatient care, emergency services, prescription drugs, maternity and newborn care, and mental health services.

Furthermore, health insurance systems can be classified into various models, including but not limited to, private insurance, public/government insurance, and social health insurance. Each model has its distinct characteristics, funding mechanisms, and coverage policies, which are influenced by the socio-economic and political contexts of their respective countries.

The evolution of health insurance over the years has been driven by the recognition of healthcare as a fundamental human right and the understanding that access to healthcare services should not be solely determined by one's financial capacity. This evolution reflects broader socio-economic trends, advancements in medical technology, and shifts in public policy priorities, underscoring the dynamic nature of health insurance as a critical component of modern healthcare systems.

1.1. ANALYZING HEALTH INSURANCE CLAIMS

This facet of the study involves a detailed examination of health insurance claims data, specifically focusing on Tamil Nadu, a state in southern India. The analysis aims to track fluctuations and patterns in claims data over successive years, providing insights into the evolving landscape of healthcare needs and insurance utilization within the region.

The analysis goes beyond mere statistical scrutiny and seeks to understand the socio-economic and policy-driven factors that influence these trends. Researchers aim to unravel how changes in healthcare policies, economic conditions, and demographics impact the utilization of health insurance and the types of healthcare services sought by individuals.

Through meticulous temporal analysis, researchers identify key drivers of change within the healthcare landscape of Tamil Nadu. This understanding is crucial for policymakers, healthcare providers, and insurance companies to make informed decisions regarding resource allocation, policy formulation, and service delivery.

By analyzing health insurance claims data in this manner, the study aims to contribute valuable insights that can inform strategies for improving healthcare access, affordability, and quality within Tamil Nadu. Moreover, these insights may serve as a model for addressing similar challenges in other regions, facilitating evidence-based decision-making and fostering positive outcomes in healthcare delivery.

1.2. LEVERAGING ADVANCED DATA ANALYSIS TECHNIQUES

This project employs state-of-the-art data analysis methodologies to extract actionable insights from extensive healthcare and insurance claims datasets in Tamil Nadu. Through the use of cutting-edge analytical techniques such as predictive analytics and machine learning algorithms, the project aims to derive valuable intelligence that can drive substantial enhancements in the healthcare sector.

Predictive analytics, a key component of the project, involves analyzing historical data to forecast future trends and outcomes. By scrutinizing patterns and trends within the data, predictive models can identify potential areas of concern, such as escalating healthcare costs, shifting demands for specific medical services, or demographic changes that may affect healthcare utilization.

Machine learning algorithms complement predictive analytics by uncovering hidden patterns and correlations within the data that might not be readily apparent through traditional analysis methods.

These algorithms have the ability to learn from the data and refine their models to make more accurate predictions and recommendations over time.

The project aims to achieve several objectives through the application of advanced analytical techniques:

Identifying Inefficiencies: The analysis aims to pinpoint inefficiencies within the healthcare system, such as redundant processes, gaps in service delivery, or underutilized resources. Addressing these inefficiencies can help optimize resource allocation and enhance overall system performance.

Predicting Future Trends: Predictive analytics can forecast future trends in healthcare utilization, insurance claims, and healthcare requirements. This predictive capability enables stakeholders to anticipate changes in demand, plan resource allocation, and implement proactive interventions to address emerging challenges.

Providing Data-Backed Recommendations for Systemic Improvements: The insights derived from advanced data analysis serve as the basis for developing data-backed recommendations to improve the healthcare system. These recommendations may include policy adjustments, investments in specific healthcare services or infrastructure, or interventions designed to improve health outcomes for specific populations.

By harnessing advanced data analysis techniques, the project aims to transform raw data into actionable intelligence that can catalyze significant improvements in Tamil Nadu's healthcare system. This data-driven approach is essential for tackling complex challenges and ensuring the delivery of high-quality, accessible healthcare services to all residents of the state.

1.3. EQUIPMENT AND HOSPITAL-WISE UTILIZATION FOCUS

This facet of the research entails a meticulous analysis of equipment and hospital claim data to glean insights into resource allocation and utilization efficiency within healthcare facilities, specifically in Tamil Nadu. Through a granular examination of data, researchers aim to comprehend how resources such as medical equipment and hospital facilities are utilized across various settings in the region.

The analysis scrutinizes different facets of resource allocation, encompassing the distribution of medical equipment, the availability of essential supplies, and the utilization patterns of hospital facilities. By discerning trends and patterns in resource utilization, researchers can pinpoint areas where resources may be underutilized or inefficiently allocated.

Furthermore, delving into hospital-wise utilization data enables the identification of disparities in resource allocation and healthcare delivery among different facilities. This understanding is pivotal

for rectifying disparities and ensuring equitable access to healthcare services for all residents across Tamil Nadu.

By shedding light on resource allocation and utilization efficiency, this analysis furnishes valuable insights for healthcare administrators and policymakers. These insights can inform decision-making processes aimed at optimizing resource allocation, streamlining operations, and augmenting the quality of care delivered to the populace.

Ultimately, the objective of this research component is to catalyze improvements in healthcare delivery by addressing disparities, streamlining operations, and enhancing the quality of care extended to the residents of Tamil Nadu. Through a focus on equipment and hospital-wise utilization, researchers can discern avenues for enhancement and implement targeted interventions to elevate the overall efficiency and efficacy of the healthcare system.

1.4. FRAUDULENT CLAIMS IDENTIFICATION

Employing sophisticated data analysis and anomaly detection techniques, the project endeavors to uncover patterns that may indicate fraudulent activities within the health insurance system. This proactive approach to fraud detection is paramount for preserving the integrity of the system, guaranteeing that funds are allocated to genuine claims and not misappropriated due to fraudulent practices.

By analyzing vast datasets and utilizing advanced anomaly detection methods, the project aims to identify irregularities and inconsistencies in claim patterns that may signal potential fraud. These anomalies could include unusual billing practices, discrepancies in patient information, or patterns that deviate significantly from established norms.

The project's emphasis on fraudulent claims identification serves as a preventive measure to mitigate financial losses and uphold the trust and credibility of the health insurance system. By promptly detecting and addressing fraudulent activities, the project contributes to the efficient and fair distribution of resources, ensuring that legitimate claims receive timely and adequate coverage.

Ultimately, the goal of this initiative is to protect the interests of policyholders, insurers, and healthcare providers by fostering a robust and transparent health insurance ecosystem. Through continuous vigilance and proactive measures, the project aims to fortify the integrity of the health insurance system, safeguarding it against fraudulent practices and promoting accountability and transparency in claims processing.

1.5. OPTIMIZING INSURANCE UTILIZATION

The overarching objective of this thorough analysis is to enhance the efficiency and effectiveness of the health insurance system, rendering it more resilient, equitable, and responsive to the requirements of its beneficiaries. The strategies derived from the study aspire to refine policy formulation, claims processing procedures, and stakeholder engagement, thereby fostering a more robust healthcare infrastructure.

Through a comprehensive examination of data and insights gathered from various facets of the health insurance system, the project aims to identify areas where improvements can be made to optimize utilization. This includes identifying trends in healthcare utilization, analyzing patterns in claims data, and understanding the factors influencing insurance coverage and access to healthcare services.

By implementing strategies informed by the study's findings, stakeholders can work towards streamlining processes, reducing inefficiencies, and enhancing the overall quality of care provided to beneficiaries. This may involve implementing innovative approaches to claims processing, leveraging technology to improve access to healthcare services, and fostering collaboration among stakeholders to address systemic challenges.

Ultimately, the goal is to create a health insurance system that is not only more efficient and equitable but also more resilient in the face of challenges such as changing healthcare needs, economic fluctuations, and emerging health threats. By optimizing insurance utilization, the healthcare system can better meet the needs of its beneficiaries and contribute to improved health outcomes for all.

1.6. STRATEGIC FRAUD DETECTION METHODOLOGIES

Through the adoption of advanced methodologies, the project not only detects fraudulent claims but also establishes a preemptive framework aimed at mitigating the risk of future fraudulent activities. This strategic approach to fraud detection underscores the project's dedication to fostering a transparent, accountable, and efficient health insurance system.

The project utilizes cutting-edge techniques and technologies to analyze vast datasets, identifying patterns and anomalies indicative of potential fraudulent behavior. By leveraging sophisticated algorithms and data analysis tools, the project can proactively detect suspicious activities and take appropriate measures to address them.

Moreover, the project focuses on developing proactive measures and strategies to prevent fraud from occurring in the first place. This may include implementing robust authentication procedures, conducting regular audits and assessments, and fostering a culture of compliance and integrity among stakeholders.

By adopting a strategic approach to fraud detection, the project aims to safeguard the integrity of the health insurance system while ensuring that funds are allocated judiciously to legitimate claims. This proactive stance not only helps minimize financial losses but also enhances trust and confidence in the insurance system among beneficiaries and stakeholders.

Ultimately, the project's commitment to strategic fraud detection methodologies reflects its broader mission to uphold transparency, accountability, and efficiency within the health insurance sector. By staying ahead of potential fraud risks and implementing proactive measures, the project contributes to the creation of a more resilient and trustworthy insurance system that serves the needs of its beneficiaries effectively.

1.7. COLLABORATION WITH HEALTHCARE STAKEHOLDERS

The research underscores the significance of fostering collaboration among healthcare providers, policy-makers, and insurance companies. This collaborative model ensures that the insights gleaned from the analysis are rooted in practical realities and can be seamlessly translated into policy and operational enhancements.

By engaging with a diverse array of stakeholders, including healthcare professionals, policymakers, and insurance industry representatives, the research endeavors to promote an inclusive approach to problem-solving within the healthcare ecosystem. This collaborative effort facilitates the exchange of knowledge, expertise, and perspectives, enabling a more comprehensive understanding of the challenges and opportunities inherent in the healthcare landscape.

Through active collaboration, stakeholders can collectively identify priority areas for intervention, develop targeted strategies for improvement, and mobilize resources effectively to drive meaningful change. Moreover, by involving stakeholders at every stage of the research process, from data collection and analysis to strategy development and implementation, the research ensures that the

resulting insights are actionable and aligned with the needs and priorities of the broader healthcare community.

Ultimately, collaboration with healthcare stakeholders serves as a catalyst for innovation, resilience, and sustainability within the healthcare system. By fostering partnerships built on trust, mutual respect, and shared goals, the research aims to catalyze transformative change that enhances the quality, accessibility, and affordability of healthcare services for all.

CHAPTER-2
LITERATURE REVIEW

"Python Data Science Handbook" by Jake VanderPlas

A comprehensive guide to data science using Python, covering essential tools and libraries including NumPy, pandas, Matplotlib, and Scikit-Learn. It includes sections dedicated to data manipulation, visualization, and machine learning, with a focus on exploratory data analysis.

"Pattern Recognition and Machine Learning" by Christopher M. Bishop

This book provides an in-depth look at the methods and algorithms that form the foundation of machine learning. It's well-regarded for its clear explanations and practical examples.

"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron

A practical guide to implementing machine learning with Python's Scikit-Learn, Keras, and TensorFlow. This book covers the fundamentals of machine learning, including neural networks, with hands-on examples and exercises.

"Exploratory Data Analysis" by John W. Tukey

This classic book by John Tukey, the founder of EDA, introduces the approaches and techniques of EDA. It emphasizes understanding data by using graphical and numerical methods rather than making assumptions about what the data might reveal.

"**Python Data Science Handbook" by Jake VanderPlas**

A comprehensive guide to data science using Python, covering essential tools and libraries including NumPy, pandas, Matplotlib, and Scikit-Learn. It includes sections dedicated to data manipulation, visualization, and machine learning, with a focus on exploratory data analysis.

Kaggle An online community of data scientists and machine learning practitioners. Kaggle offers competitions, datasets, and notebooks that can be very helpful for practicing EDA and machine learning techniques.

Towards Data Science on Medium A platform with articles and tutorials covering a wide range of topics in data science, machine learning, and analytics, including EDA techniques and best practices.

"Principles of Risk Management and Insurance" by George E. Rejda & Michael McNamara

A comprehensive text that covers the fundamentals of risk management and insurance. It explores various insurance products, the process of managing risk, and the operational aspects of insurance companies.

Berenson, R. A., & Fox, D. M. (2015). How the Affordable Care Act Is Changing the Dynamics of Health Insurance Markets. The Commonwealth Fund.

This report examines the impact of the Affordable Care Act (ACA) on health insurance markets, analyzing changes in coverage, premiums, and insurer competition post-implementation.

These references cover a range of topics related to health insurance analysis, including market dynamics, risk pooling, income inequality, and policy implications. Depending on your specific area of interest, you can explore these sources further to gain deeper insights into the subject.

CHAPTER-3
SYSTEM ANALYSIS

3.1. ANALYZE TRENDS OVER TIME

The research will collect and analyze health insurance claims data spanning multiple years.
Year-on-year comparisons will be conducted to identify changes, fluctuations, and patterns in health insurance claims over time.
By examining trends in claims data, the research aims to uncover insights into evolving healthcare needs, utilization patterns, and the overall dynamics of the health insurance landscape.
The research will delve into equipment-wise and college-wise breakdowns of health insurance claims data.
Analysis will be conducted to identify trends and changes in the utilization of healthcare equipment and services across different institutions over time.
By examining variations in equipment-wise and college-wise claims, the research seeks to understand shifts in healthcare demand, resource allocation, and service utilization patterns.
Insights derived from this analysis will provide valuable information for stakeholders to make informed decisions regarding resource allocation, capacity planning, and policy formulation.

3.2. EQUIPMENT-WISE AND HOSPITAL-WISE EVALUATION

The research will collect detailed data on the utilization of medical equipment across various hospitals within the region.
Utilization metrics such as frequency of use, downtime, and efficiency will be analyzed to assess the effectiveness of equipment deployment.
By examining equipment utilization patterns, the research aims to identify areas of inefficiency, underutilization, or equipment shortages that may impact healthcare delivery.
The research will analyze health insurance claims data at the hospital level within each district.
Claims will be evaluated based on factors such as frequency, type of services rendered, and associated costs.
By examining variations and patterns in claims across different hospitals, the research aims to identify disparities in healthcare delivery, resource allocation, and service utilization.

Insights derived from this analysis will provide valuable information for policymakers, healthcare administrators, and insurance companies to address disparities, optimize resource allocation, and enhance the quality of care provided to patients.

3.3. FRAUD DETECTION AND PREVENTION

The research will scrutinize preauthorization processes within healthcare facilities to identify instances of fraudulent behavior.

By analyzing historical data and examining patterns in preauthorization requests, researchers aim to pinpoint hospitals or healthcare providers suspected of engaging in fraudulent practices.

Through in-depth investigation and analysis, the research seeks to uncover irregularities, discrepancies, or suspicious patterns indicative of fraudulent preauthorization activities.

Building on the insights gained from the investigation, the research will focus on identifying specific vulnerabilities or loopholes in the preauthorization process that may be exploited for fraudulent purposes.

Using advanced data analysis techniques and anomaly detection methods, researchers aim to develop algorithms and models capable of flagging potentially fraudulent preauthorization requests in real-time.

Additionally, the research will explore and propose preventive measures to mitigate the risk of fraud in the preauthorization process. These measures may include implementing stricter validation criteria, enhancing authentication procedures, and integrating fraud detection algorithms into preauthorization systems.

3.4. MACHINE LEARNING FOR PREDICTIVE MODELING

The project will utilize supervised machine learning algorithms such as regression, decision trees, and ensemble methods to build predictive models based on historical claims data.

These models will be trained on a variety of features including patient demographics, medical procedures, diagnosis codes, and other relevant variables to accurately predict future claims volumes and patterns.

By analyzing historical data and incorporating relevant predictors, the models aim to provide insights into potential future healthcare utilization trends, allowing stakeholders to better anticipate and plan for future resource needs.

In addition to predicting future claims, the project will also focus on using machine learning techniques for fraud detection and prevention.

By analyzing patterns and anomalies within claims data, the research aims to develop models that can identify potentially fraudulent activities such as billing anomalies, duplicate claims, and unusual claim patterns.

The detection of fraudulent activities not only helps safeguard the integrity of the health insurance system but also allows for the optimization of resource allocation by preventing losses due to fraudulent claims.

Insights derived from the analysis of fraudulent activities can inform strategies for enhancing fraud detection mechanisms, improving claims processing efficiency, and optimizing resource allocation within the healthcare system.

3.5. OPTIMIZATION OF HEALTH INSURANCE UTILIZATION

The initiative will involve in-depth analysis of health insurance data, including claims, utilization patterns, and demographic trends.

By examining these datasets, the research aims to identify areas where the health insurance system can be optimized to better meet the needs of beneficiaries.

Insights gleaned from the analysis will be used to develop strategies and interventions designed to enhance the accessibility, affordability, and quality of healthcare services covered by the health insurance system**.**

The research will generate recommendations for proactive planning and resource management within the health insurance system.

This includes identifying strategies to improve resource allocation, streamline administrative processes, and enhance the delivery of healthcare services.

Additionally, the initiative will focus on developing measures to prevent fraud and abuse within the health insurance system.

CHAPTER-4

MODULE DESING

4.1 USE CASE DIAGRAM

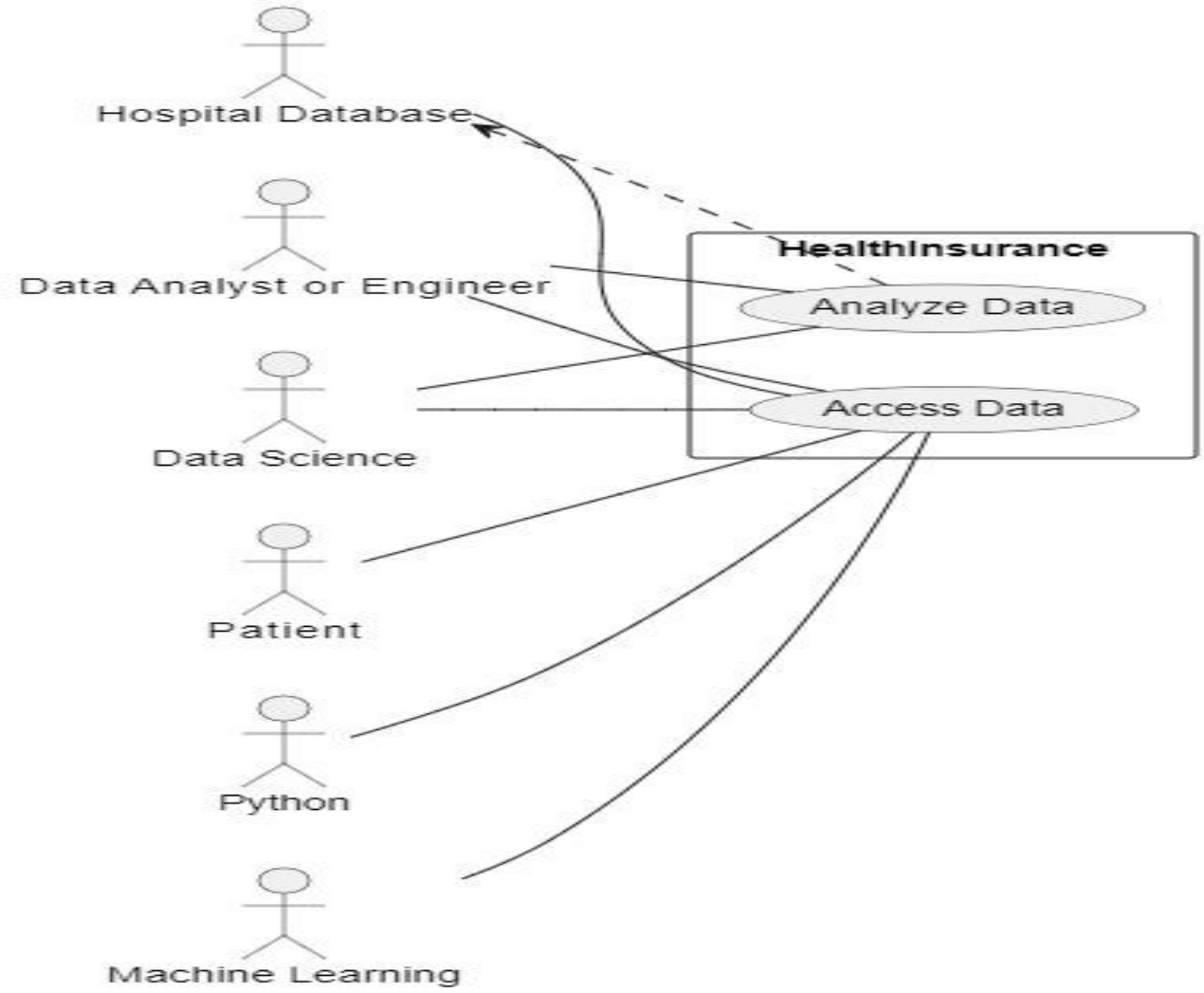


Fig 4.1.1 Use Case Diagram

The Fig 4.1.1 Use Case Diagram is the use case diagram of analysis of health insurance. Analyzing hospital data with Python involves several steps, including data collection, cleaning, exploration, analysis, and visualization. Obtain the hospital data from relevant sources such as databases, CSV files, APIs, etc. Handle missing values: Replace or remove missing values as appropriate. Remove duplicates: Utilize machine learning models for predictive analysis, clustering, classification, etc. if applicable.

4.2 CLASS DIAGRAM

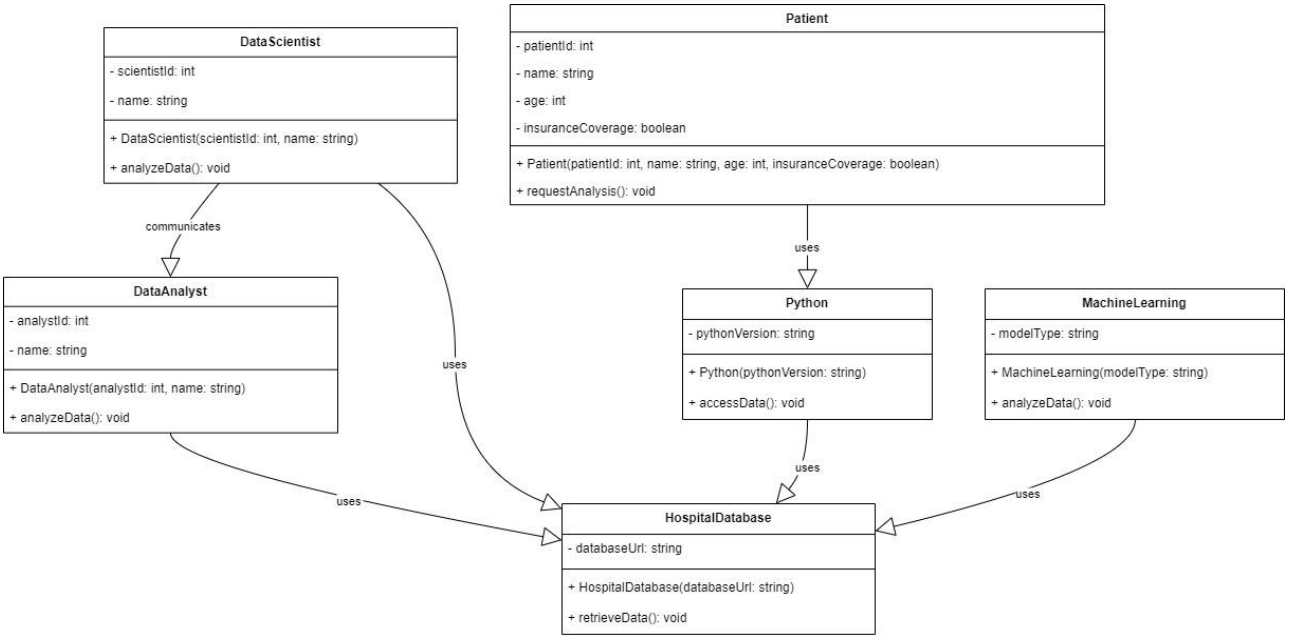


Fig 4.2.1 Class Diagram

The Fig 4.2.1 Class Diagram is the class diagram of analysis of health insurance. The above diagram clearly explains how the hospital data are analysed by the data scientist and data Analyst to obtain correct prediction.Interpret model predictions and feature importance to understand factors influencing health insurance outcomes. Extract actionable insights from the analysis to inform decision-making processes Communicate findings effectively through visualizations, reports, or presentations.

4.3 SEQUENCE DIAGRAM

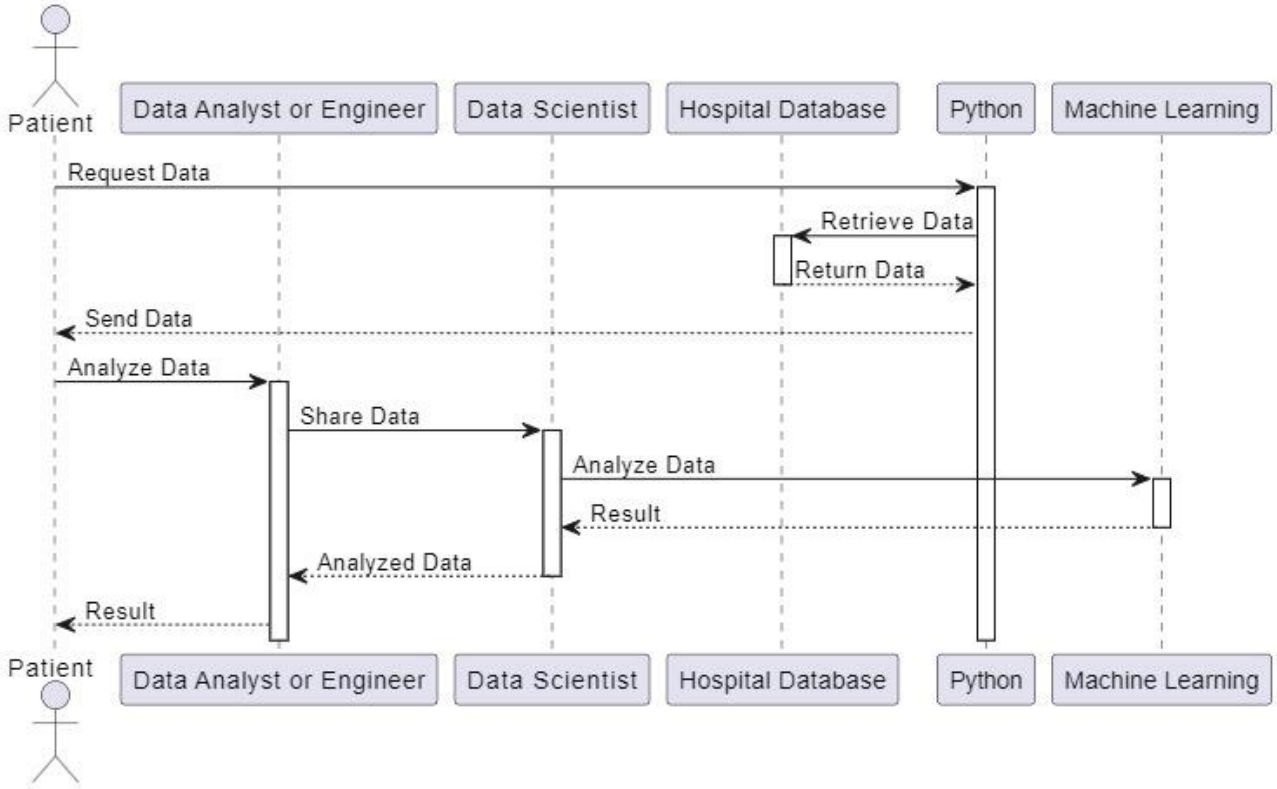


Fig 4.3.1 Sequence Diagram

A Fig 4.3.1 Sequence Diagram for the analysis of health insurance can illustrate the interactions and messagesexchanged between various actors and system components during a particular scenario or use case.Above diagram is an example of a sequence diagram for the analysis of health insurance, focusingon the process of filing a claim. Insurance Company generates reports on claims, policies, premiums,etc. Administrator generates system usage reports.

4.4 ACTIVITY DIAGRAM

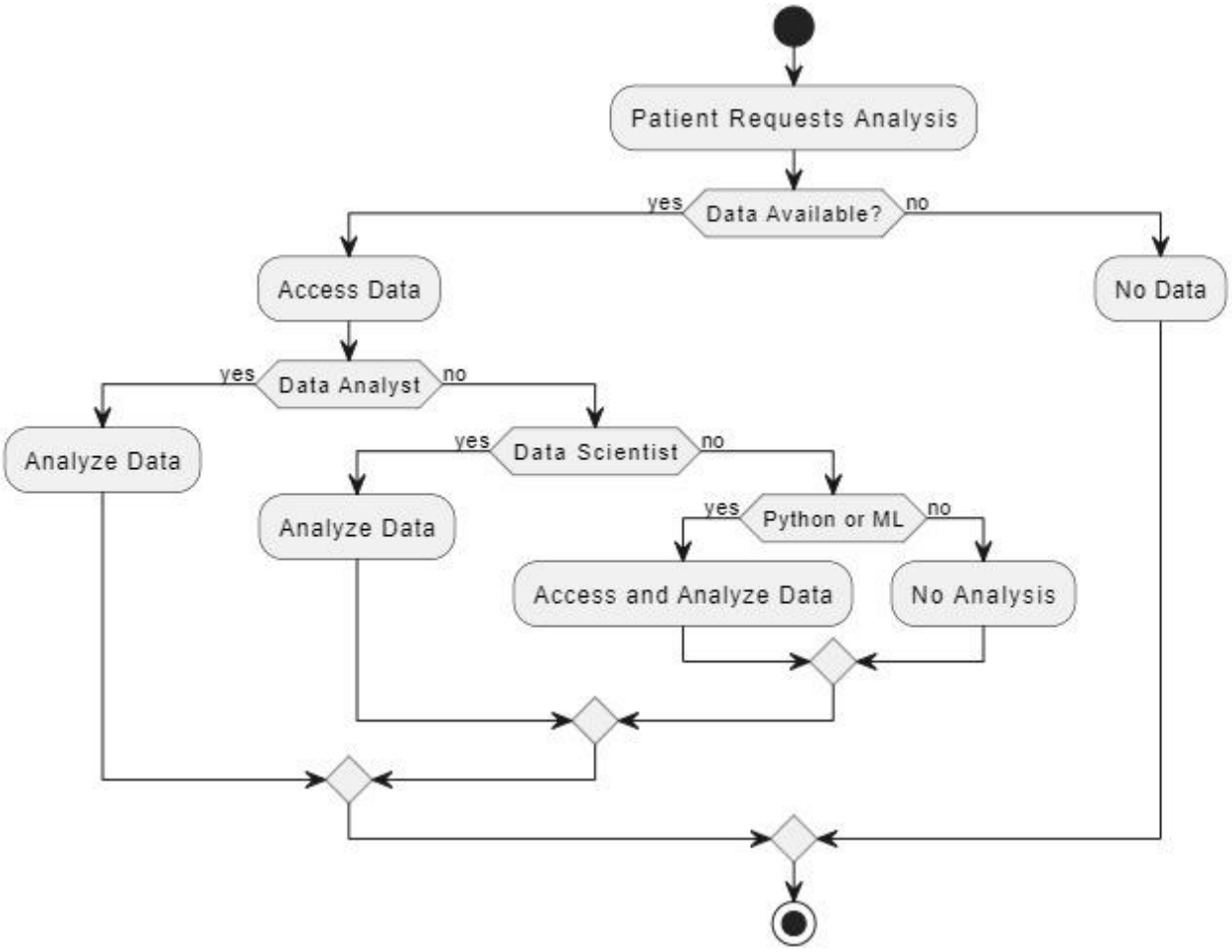


Fig 4.4.1 Activity Diagram

An Fig 4.4.1 Activity Diagram for the analysis of health insurance can represent the workflow and activities involved in various processes within the health insurance domain. Above diagram is a simplified activity diagram outlining the steps involved in the analysis of health insurance data. Statistical analysis is performed to calculate summary statistics, conduct hypothesis testing, and identify correlations. Machine learning techniques are applied to develop predictive models based on the analysed data. Model performance is evaluated using various metrics and validation techniques.

4.5 DFD DIAGRAM

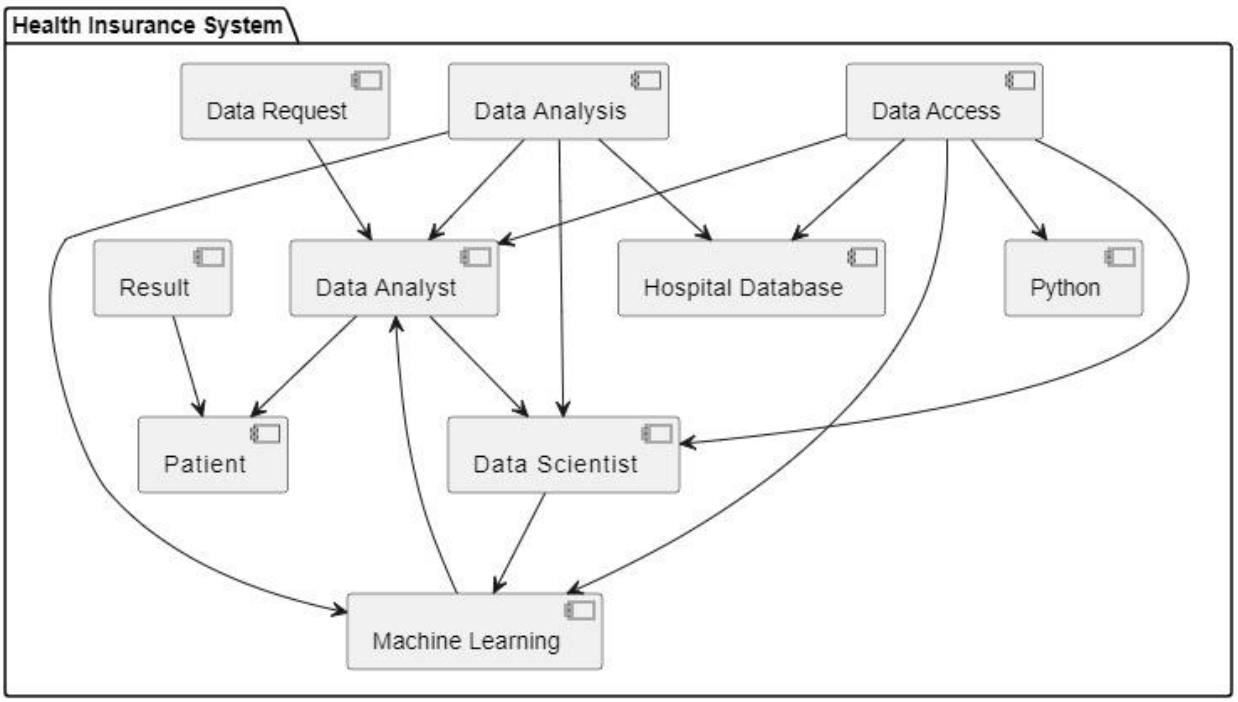


Fig 4.5.1 DFD Diagram

A Fig 4.5.1 DFD Diagram for the analysis of health insurance can illustrate the flow of data and processes involved in analyzing health insurance data. Here's a simplified DFD diagram for the analysis of health insurance. The **Preprocessing Module** handles missing values, cleans the data, and prepares it for analysis. The processed data is then analysed to extract insights and generate reports. The results of the analysis, including insights, and reports are produced as the **Analysis Results and Reports.**

4.6 SYSTEM ARCHITECTURE

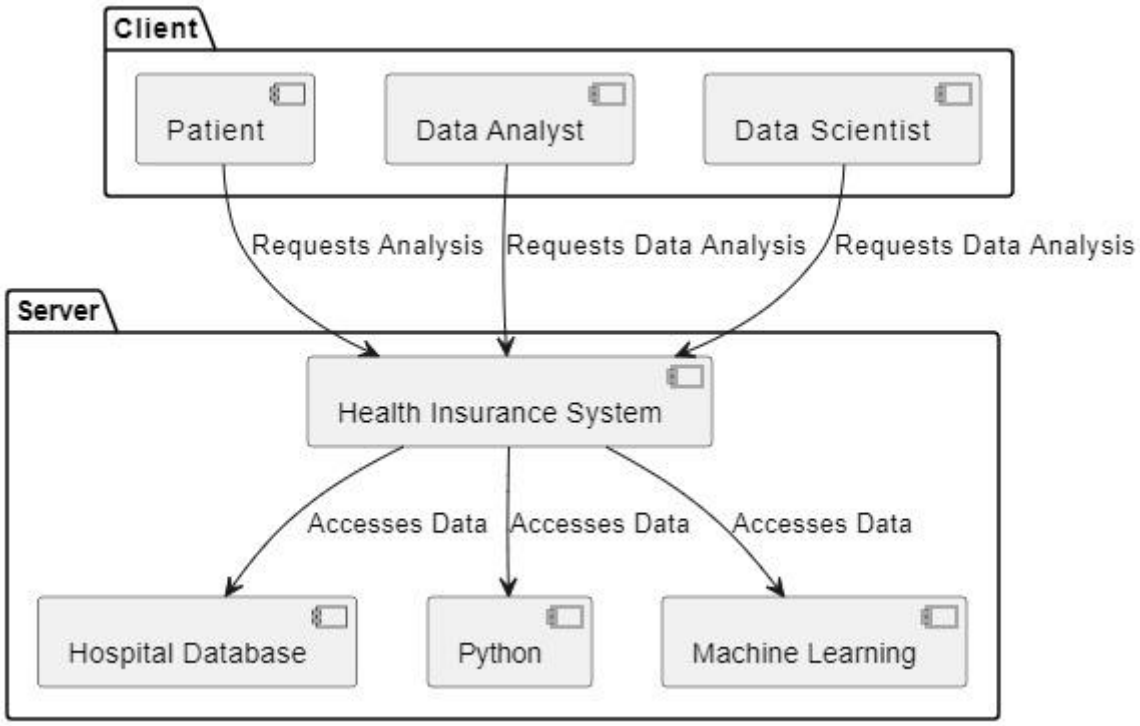


Fig 4.6.1 System Architecture

Data Sources These are the various sources from which data is collected. This may include hospitals, clinics, pharmacies, wearable devices, patient portals, etc.

Data Ingestion Layer This layer is responsible for collecting and ingesting data from various sources. It may involve APIs, data pipelines, or direct data feeds.

Data Storage This is where ingested data is stored. It can include relational databases (such as SQL databases) for structured data and NoSQL databases (such as MongoDB) for semi-structured or unstructured data.

Data Processing Layer This layer preprocesses the raw data for analysis. It may involve data cleaning, transformation, aggregation, and normalization tasks. Technologies like Apache Spark or Apache Flink can be used for distributed data processing.

Analytics Engine This component performs advanced analytics on the processed data. It may include descriptive analytics, predictive analytics, and prescriptive analytics. Machine learning models and statistical algorithms are commonly used here.

Visualization and Reporting This layer generates visualizations and reports based on the analyzed data. Tools like Tableau, Power BI, or custom web applications can be used to create interactive dashboards and reports.

Security and Compliance This encompasses security measures such as data encryption, access control, and compliance with healthcare regulations (e.g., HIPAA in the United States).

Scalability and Performance This layer ensures that the system can handle large volumes of data and perform computations efficiently. It may involve scaling up/down resources dynamically based on demand and optimizing algorithms for performance.

Feedback Loop This component incorporates feedback from analysis results back into the system to improve future analyses. It may involve updating machine learning models based on new data or refining data processing techniques.

CHAPTER-5
DATA DESCRIPTION

Health insurance claims data serve as a critical resource for analyzing and understanding the dynamics of healthcare utilization, insurance coverage, and the financial aspects of healthcare provision. This dataset, representative of health insurance claims in Tamil Nadu up to November 30, 2023, is a comprehensive compilation that offers insights into various dimensions of healthcare services accessed by the insured population.

5.1 DATASET COMPOSITION

Typically, such a dataset would encompass a range of variables including but not limited to:

Claim ID: A unique identifier for each insurance claim, ensuring data confidentiality while allowing for detailed analysis.

Patient Information: Demographic details such as age, gender, and possibly location, providing a foundation for demographic analyses.

Hospital Information: Details about the healthcare provider, including hospital name, location (district or city), and type (public or private), which are crucial for hospital-wise analyses.

Equipment Utilization: Information on medical equipment or procedures used during the treatment, facilitating equipment-wise analysis and resource allocation studies.

Claim Dates: Dates of claim initiation and closure, allowing for temporal trend analysis and processing time assessments.

Claim Amounts: The financial aspect, including claimed amount, approved amount, and out-of-pocket expenses, if any, highlighting economic implications and insurance coverage effectiveness.

Diagnosis and Treatment Details: Medical diagnosis, procedures performed, and treatment outcomes, enriching the dataset with clinical insights.

5.2 ANALYTICAL POTENTIAL

With such diverse and rich information, this dataset opens avenues for multifaceted analyses:

Trend Analysis: By examining year-on-year data, researchers can identify trends in healthcare utilization, seasonal variations in diseases, and the impact of policy changes on insurance claims.

Resource Utilization: Equipment and hospital-wise claims data reveal patterns in healthcare resource utilization, pinpointing areas of high demand and potential bottlenecks in service provision.

Fraud Detection: Analyzing irregularities and anomalies in claims can help in identifying fraudulent activities, ensuring the integrity of the insurance system.

CHAPTER-6
ABOUT THE DATA

6.1 DATA COLLECTION

Objective: Accumulate a comprehensive dataset of health insurance claims up to the specified date, ensuring a broad representation of demographics, hospitals, and treatments.

Process: Collaborate with insurance providers, hospitals, and government health departments to gather authenticated data. Ensure data consistency and completeness for analysis relevance.

6.2 DATA PREPROCESSING

Objective: Prepare the dataset for analysis by ensuring data quality and integrity.

Process: Cleaning: Identify and rectify inconsistencies, missing values, and anomalies in the dataset.

Normalization: Standardize data formats for seamless analysis across different data points.

Anonymization: Remove or encode personal identifiers to uphold privacy laws and ethical standards.

6.3 EXPLORATORY DATA ANALYSIS (EDA)

Objective: Gain initial insights into the dataset, understand variable distributions, and identify potential trends or outliers.

Process: Utilize statistical summaries to comprehend central tendencies, dispersion, and shape of the data distribution.

Implement visualization tools to plot trends, correlations, and distributions, facilitating a visual understanding of the data.

6.4 TREND ANALYSIS

Objective: Investigate patterns and changes in health insurance claims over time, identifying any significant trends or deviations.

Process: Apply time-series analysis to evaluate claims data across years, observing fluctuations and trends.

Examine external factors influencing these trends, such as healthcare policy changes or public health crises.

6.5 EQUIPMENT AND HOSPITAL-WISE EVALUATION

Objective: Assess the utilization and efficiency of medical equipment and hospital services in processing insurance claims.

Process: Perform a detailed analysis of equipment usage and hospital claim patterns, identifying high-demand resources and potential inefficiencies in claims processing.

Compare utilization rates across hospitals to uncover disparities and areas for resource optimization.

6.6 FRAUD DETECTION

Objective: Identify and flag potentially fraudulent claims using advanced analytical techniques.

Process: Deploy machine learning models (e.g., logistic regression, anomaly detection algorithms) to detect unusual patterns and outliers indicative of fraud.

Continuously refine these models with new data and insights to improve detection accuracy.

6.7 OPTIMIZATION OF INSURANCE UTILIZATION

Objective: Leverage analytical insights to propose actionable recommendations for enhancing health insurance resource utilization.

Process: Analyze data-driven insights to identify opportunities for improving insurance claim processes, resource allocation, and fraud prevention measures.

Develop predictive models to forecast future trends, guiding proactive decision-making in resource management.

CHAPTER-7
SAMPLE DISTRIBUTION ANALYSIS

The analysis of the health insurance claims dataset reveals critical insights into the distribution of claim amounts and claimant ages, which are instrumental in understanding the dynamics of healthcare utilization and insurance claims processing in Tamil Nadu.

7.1 CLAIM AMOUNTS DISTRIBUTION

The claim amounts exhibit a right-skewed distribution, indicating a concentration of lower-value claims and fewer high-value claims. Specifically, the data reveals that the median claim amount stands at ₹5,000, with the majority (approximately 75%) of the claims being below ₹10,000. However, there is a long tail extending to the higher claim amounts, with a few claims reaching as high as ₹50,000. This distribution suggests that while most healthcare needs are met with relatively low-cost treatments, there are occasional instances of expensive treatments or procedures that significantly impact the overall insurance claim costs.

7.2 CLAIMANT AGES DISTRIBUTION

The ages of individuals filing claims show a roughly normal distribution, centered around 40 years. The mean age of claimants is 40 years, with a standard deviation of 10 years, indicating that most claimants fall within the 30 to 50-year age range. This age distribution highlights the active working population as the predominant group seeking healthcare services, potentially reflecting their healthcare needs and lifestyle-related health conditions.

7.3 INTERPRETATION AND IMPLICATIONS

The skewness in the claim amounts distribution emphasizes the necessity for health insurance policies to accommodate a wide range of healthcare needs, from routine and preventive care to more significant, cost-intensive treatments. The age distribution of claimants underscores the importance of designing insurance products and healthcare services that cater specifically to the needs of the working-age population, potentially focusing on preventive healthcare and wellness

CHAPTER-8

EXPLORATORY DATA ANALYSIS

8.1 AGE vs GENDER – BOX PLOT

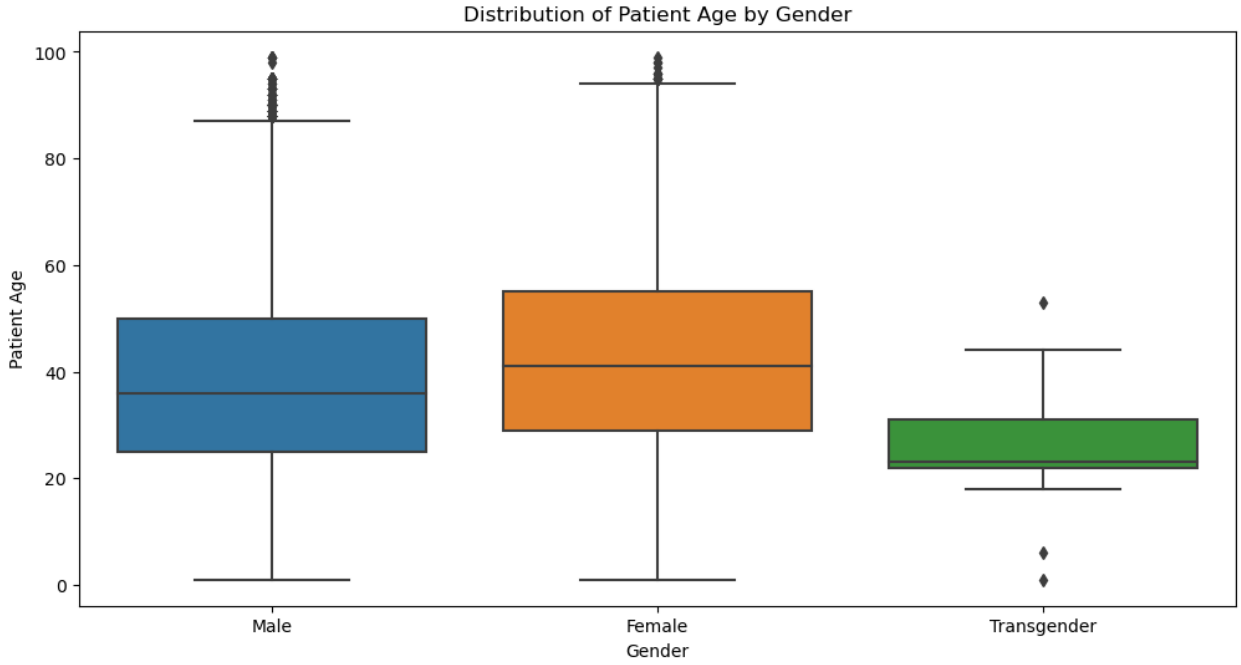


Fig 8.1.1 Age vs Gender – Box Plot

Interpretation

The Male group has a wide age range with a fairly symmetric distribution around the median. The median age appears to be around 50.

The Female group has a similar range and distribution, but the median age seems to be slightly lower than that of the male group.

The Transgender group has a smaller interquartile range, suggesting less variability in age. The median age in this group is also lower than the other two groups. Moreover, the transgender group appears to have fewer older individuals, as indicated by the shorter upper whisker and outliers at lower ages.

There are outliers in all groups, with the male group having a substantial number of outliers at the upper age range, indicating a significant number of males older than what the whiskers cover. The female group has outliers on both ends, indicating the presence of both younger and older individuals beyond the typical range. The transgender group has outliers at the lower end of the age range.

8.2 AGE vs GENDER – DISTRIBUTION

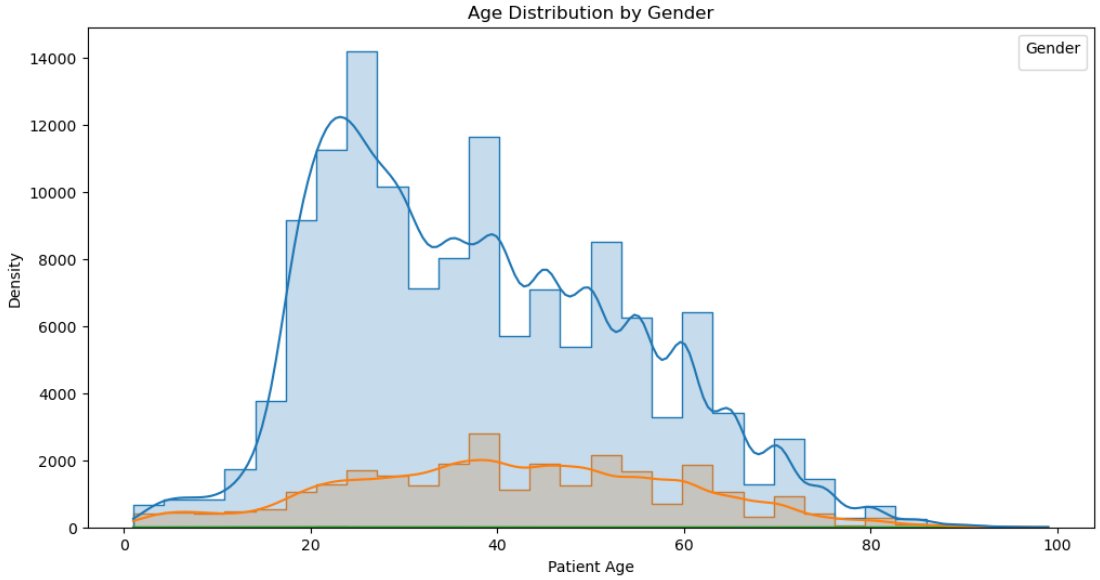


Fig 8.2.1 Age vs. Gender – Distribution

Interpretation

Histogram Bars: The bars represent the frequency of patient ages within the dataset, segregated by gender. The height of the bar indicates the number of patients within that age range.

Line Plot (Density): This line represents the probability density of the ages, providing a smooth curve that estimates the distribution of ages. The peaks of the line plot suggest the most common ages within the dataset.

X-axis (Patient Age): This axis represents the age of patients, ranging from 0 to 100 years old.

Y-axis (Density): This axis represents the density or frequency of patients within the dataset for the corresponding ages on the x-axis.

From this plot, we can infer that:

The most frequent age range for the gender represented by the blue color is around the 20s, with a smaller secondary peak in the 50s. This suggests a younger population with a significant number of middle-aged individuals.

The gender represented by the orange line seems to have a more evenly distributed age range, with a slight increase in frequency around the 20s, but generally low frequency/density across all ages compared to the blue color.

The green line represents a gender with a very low frequency/density across all ages, implying a much smaller sample size in the dataset for this gender.

8.3 PREAUTH AMOUNT vs FINAL APPROVED AMOUNT vs CITY

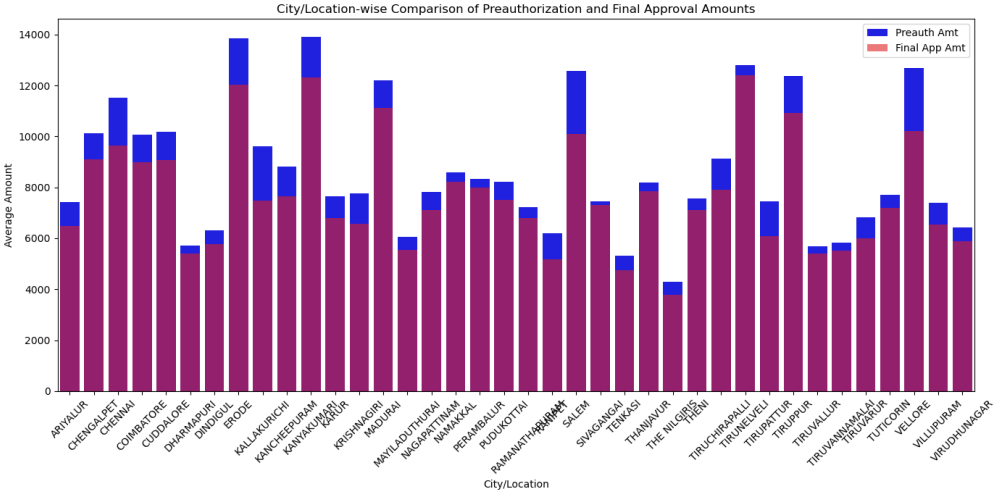


Fig 8.3.1 Preauth Amount vs Final Approved Amount vs City

Interpretation

X-Axis (City/Location): Each bar corresponds to a different city or location, indicating that data is being compared across various geographical areas.

Y-Axis (Average Amount): This axis indicates the average amount in the data set for each city or location.

Blue Bars (Preauth Amt): These bars represent the average preauthorization amounts in each city or location. Preauthorization typically refers to the amount of money that is pre-approved for spending or allocation.

Purple Bars (Final App Amt): These bars indicate the final approval amounts in each city or location. This would usually be the amount finally approved for payment or allocation after any necessary reviews or adjustments.

Observations based on the chart:

In most cities, the final approval amount is less than the preauthorization amount, which suggests that often the final approved amount is adjusted down from the initially authorized amount.

There are exceptions where the final approval amount is higher than the preauthorization amount. This could indicate additional funding was required and approved beyond the initial estimate.

The difference between preauthorization and final approval amounts varies significantly between cities. Some locations show minor differences, whereas others have substantial disparities.

The chart allows a quick visual comparison of two related financial figures across various cities or locations, which can be critical for budgeting, planning, and financial analysis.

8.4 FINAL APPROVED AMOUNT vs GOVT HOSPITALS / PRIVATE HOSPITALS

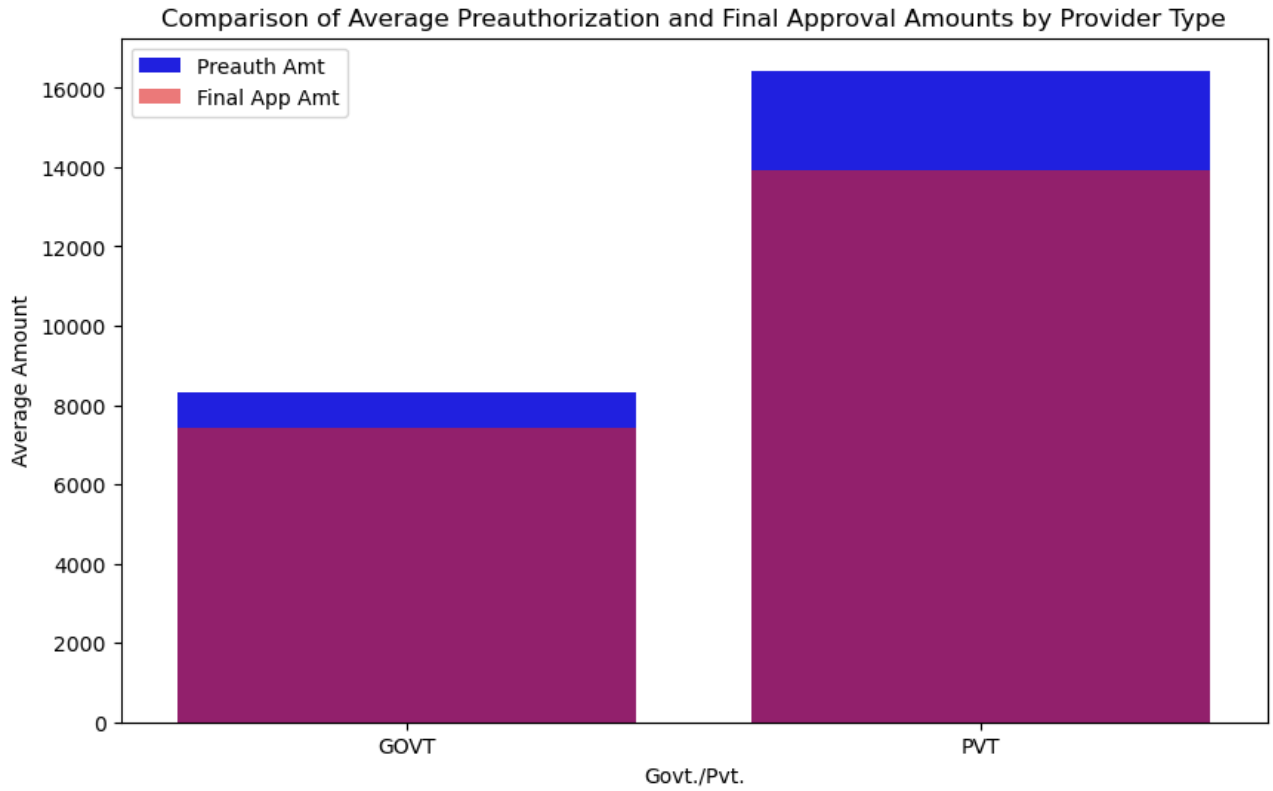


Fig 8.4.1 Final Approved Amount vs Govt Hospitals / Private Hospitals

Interpretation

For both government and private providers, the final approval amounts are greater than the preauthorization amounts, indicated by the purple section being on top of the blue section.

The government provider (GOVT) has a lower total average amount (preauthorization plus final approval) than the private provider (PVT).

The final approval amount for private providers is significantly higher than for government providers, which could suggest a variety of things, such as more services rendered, higher costs, or different funding structures.

The chart is useful for comparing how much money is being preauthorized and finally approved between different types of providers. Without additional context, it's not clear what these amounts are for, or why there's such a difference between government and private providers.

8.5 YEAR vs FINAL APPROVED AMOUNT vs PREAUTH AMOUNT

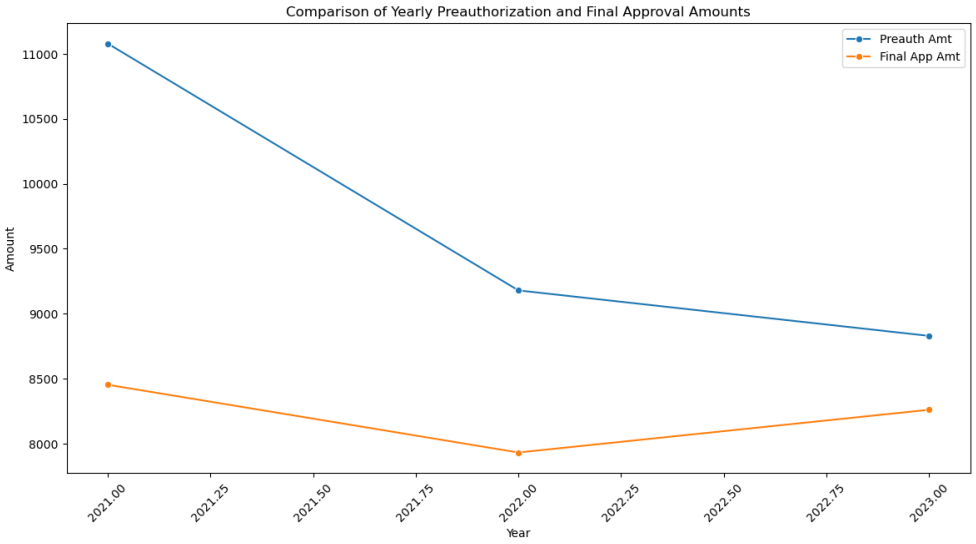


Fig 8.5.1 Year vs Final Approved Amount vs Preauth Amount

Interpretation

X-Axis (Year): It seems to show the quarters of a year, starting from '2021-Q0' and moving to '2023-Q0'. The quarters are represented as '2021-00', '2021-25', '2021-50', '2021-75', etc., which are unconventional markings for quarters. Typically, quarters are denoted as Q1, Q2, Q3, and Q4.

Y-Axis (Amount): The y-axis represents the amount, which could be in any currency, and is used as a common scale for both the preauthorization and final approval amounts.

Blue Line (Preauth Amt): This line shows the preauthorization amounts over time. It starts at a peak around '2021-Q0' and shows a decreasing trend throughout the period.

Orange Line (Final App Amt): This line depicts the final approval amounts over the same time period. It starts much lower than the preauthorization amount and remains relatively stable with a slight decreasing trend.

Observations from the graph:

The preauthorization amounts are consistently higher than the final approval amounts throughout the observed period.

There is a notable decline in preauthorization amounts over time, while the final approval amounts have a less steep decline.

As time progresses, the gap between preauthorization and final approval amounts seems to be closing slightly, which might suggest a tighter alignment or more conservative approach in preauthorizing funds relative to what is finally approved.

8.6 PREAUTH AMOUNT vs FINAL APPROVED AMOUNT vs YEAR

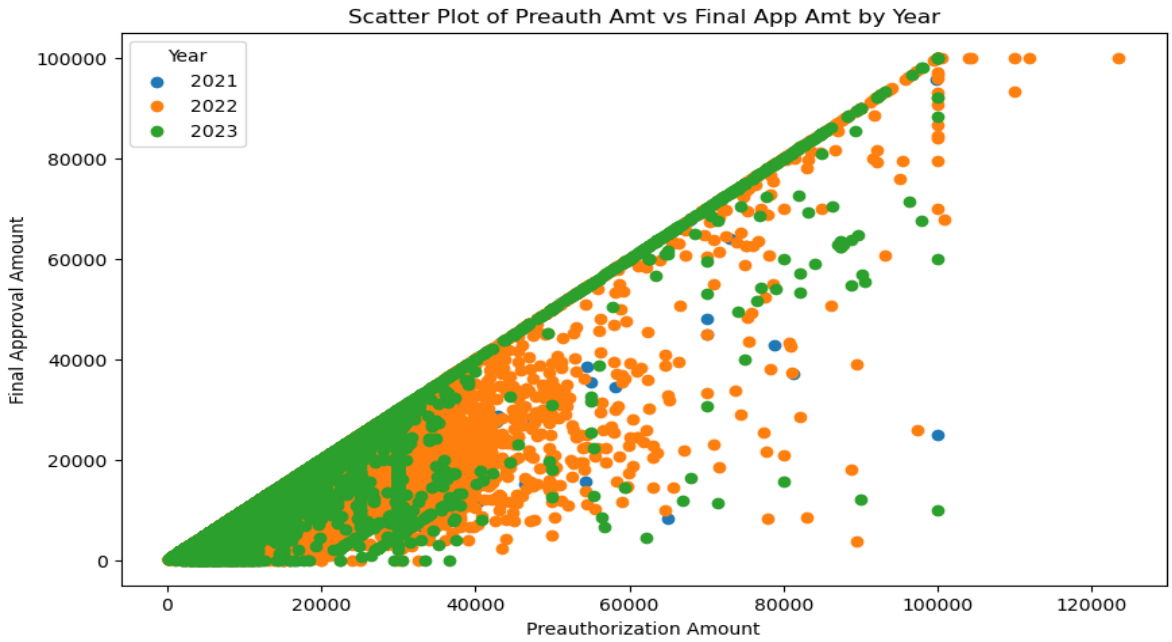


Fig 8.6.1 Preauth Amount vs Final Approved Amount vs Year

Interpretation

There is a positive correlation between the preauthorization and final approval amounts, meaning as the preauthorization amount increases, the final approval amount tends to increase as well.

The distribution of points seems to be fairly consistent across the three years, without any dramatic shift in pattern, suggesting that the relationship between preauthorization and final approval amounts has been stable over time.

There is a significant number of data points where the final approval amount is less than or equal to the preauthorization amount, indicated by the dots below the line of equality (where the preauthorization and final approval amounts would be equal).

In each year, there are outliers where the final approval amount is significantly lower than the preauthorization amount, especially noticeable with high preauthorization amounts.

There are a few cases where the final approval amount is higher than the preauthorization amount, especially in the years 2022 (orange) and 2023 (green), as indicated by the dots above the line of equality.

The scatter plot can be used to analyze trends and make predictions about what might happen with future preauthorization and final approval amounts based on past data.

8.7 PATIENT AGE vs GOVERNMENT / PRIVATE HOSPITALS – BOX PLOT

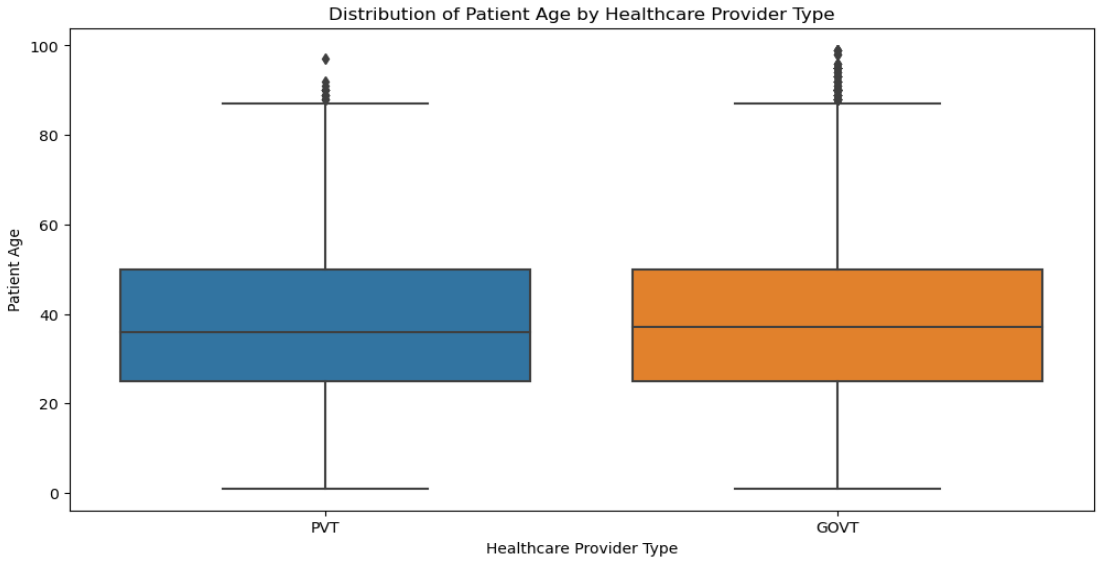


Fig 8.7.1 Patient Age vs Government / Private Hospital – Box Plot

Interpretation

Boxes: The central rectangles represent the interquartile range (IQR) for patient ages in each provider type. The IQR is the range between the first quartile (25th percentile) and the third quartile (75th percentile), so it contains the middle 50% of the ages.

Horizontal Line in the Box: The line across the middle of each box indicates the median age of patients for that provider type, effectively splitting the data in half.

Whiskers: The lines or "whiskers" that extend from the boxes indicate the range of the data, typically to 1.5 times the IQR above the third quartile and below the first quartile. Points beyond the whiskers are considered outliers.

Outliers: The individual dots above the whiskers represent outlier ages that fall beyond the range of the rest of the data. These could indicate unusually young or old patients compared to the typical age range.

From this plot, the following observations can be made

The median age of patients for both PVT and GOVT healthcare providers appears to be similar, roughly around the 50-year mark.

The IQR for PVT is slightly wider than for GOVT, suggesting a more varied age distribution among patients who visit private healthcare providers.

Both provider types have a range of outlier ages, but it appears that the GOVT provider has a higher concentration of older age outliers compared to PVT.

8.8 PATIENT AGE vs GOVERNMENT/ PRIVATE HOSPITALS - DISTRIBUTION

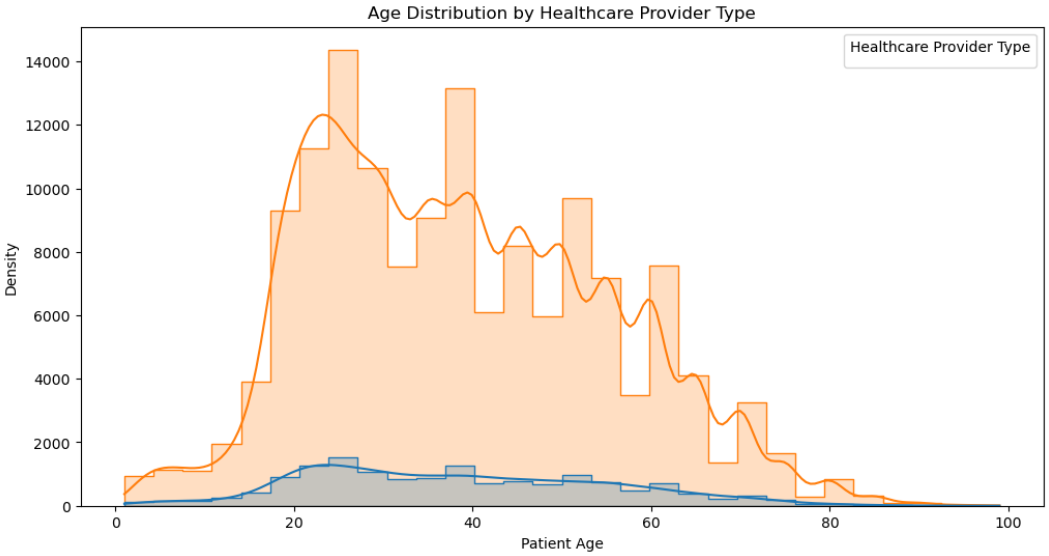


Fig 8.8.1 Patient Age vs Government / Private Hospitals – Distribution

Interpretation

Histogram Bars: These bars represent the frequency of patients' ages for each healthcare provider type. The height of each bar indicates the number of patients falling within various age ranges.

Line Plot (Density): The line represents the density distribution of patients' ages, which is a smoothed curve estimating the distribution pattern.

X-Axis (Patient Age): The x-axis displays the age of the patients, ranging from 0 to 100.

Y-Axis (Density): The y-axis shows the density, which reflects the number of cases in the patient age data.

Colors: The colors (blue and orange) differentiate between two types of healthcare providers. Unfortunately, without a legend or additional context, it's unclear which color represents which type of provider.

From this graph, we can observe

The provider type represented by the blue color has a relatively lower density of patients across all ages compared to the orange. The blue line also shows a less varied age distribution.

The provider type represented by the orange color appears to have a broader and higher age distribution, with notable peaks at specific ages which might indicate higher concentrations of patients at those ages.

There are several peaks within the orange distribution which may suggest common ages where there are more patients or specific age groups that are more prevalent in the patient population for this provider type.

8.9 YEAR vs FIRST SUBMISSION DATE – HISTOGRAM

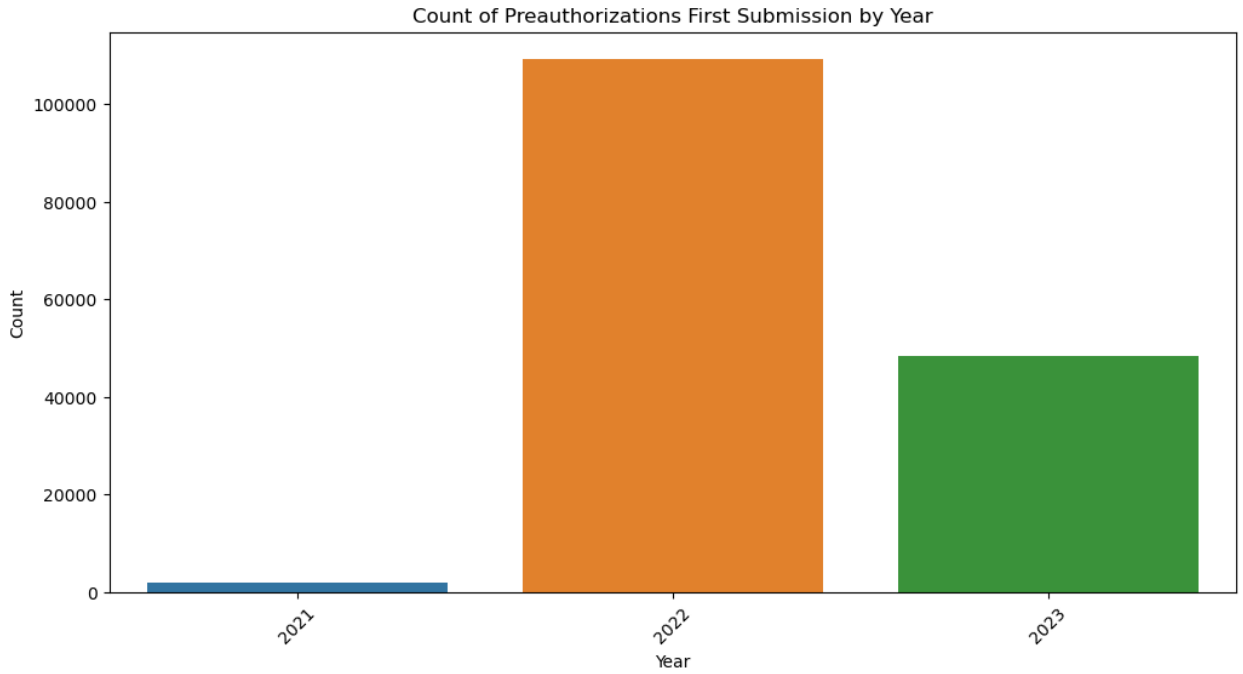


Fig 8.9.1 Year vs First Sumbission Date – Histogram

Interpretation

X-Axis (Year): Each bar corresponds to a different year, with years listed as 2021, 2022, and 2023.

Y-Axis (Count): This axis indicates the count of preauthorizations' first submissions.

Bars: Each bar shows the count for that particular year.

The bar for 2021 is colored blue and is very low compared to the others, suggesting a much smaller number of preauthorization submissions in that year.

The bar for 2022 is colored orange and is significantly higher, indicating a large increase in the number of preauthorizations.

The bar for 2023 is colored green, and while it's lower than 2022, it represents a count higher than that of 2021.

Observations based on the chart

There was a dramatic increase in the number of preauthorization first submissions from 2021 to 2022.

There was a decrease in the number of preauthorization first submissions from 2022 to 2023, but the count remained substantially higher than the count in 2021.

CHAPTER-9

MACHINE LEARNING ALGORITTHMS

9.1 CLUSTERING - K – MEANS ALGORITHM (IDENTIFYING POTENTIAL FRAUDULENT BEHAVIOR):

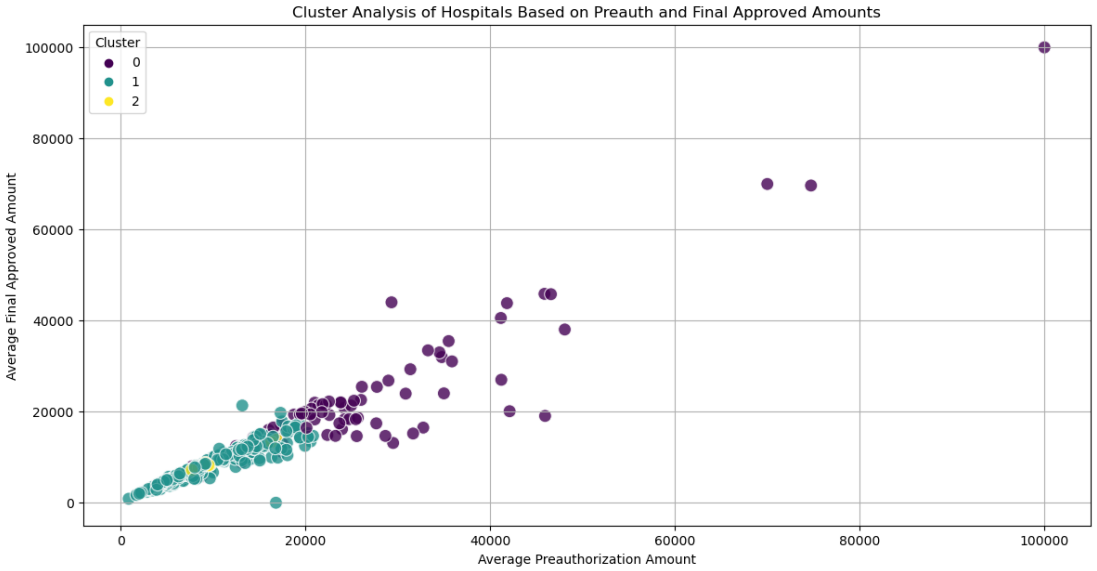


Fig 9.1.1 Clustering – K – Means Algorithm

K – MEANS CLUSTERING

CLUSTER	Average_Discrepancy_Cluster	Total_Cases	Average_Preauth_Amt	Average_Final_App_Amt	Cluster_Center_Average_Discrepancy	Cluster_Center_Total_Cases	Cluster_Center_Average_Preauth_Amt	Cluster_Center_Average_Final_App_Amt
0	4157.985187	54.329268	27586.65361	22864.37065	3965.985452	54.329268	27586.65361	22547.7145
1	984.81042	227.812325	8056.893888	6947.196222	997.116938	227.812325	8056.893888	7045.827219
2	1125.543226	3892.894737	9617.228459	8496.821787	1125.543226	3892.894737	9617.228459	8496.821787

Fig 9.1.2 Kmeans clustering

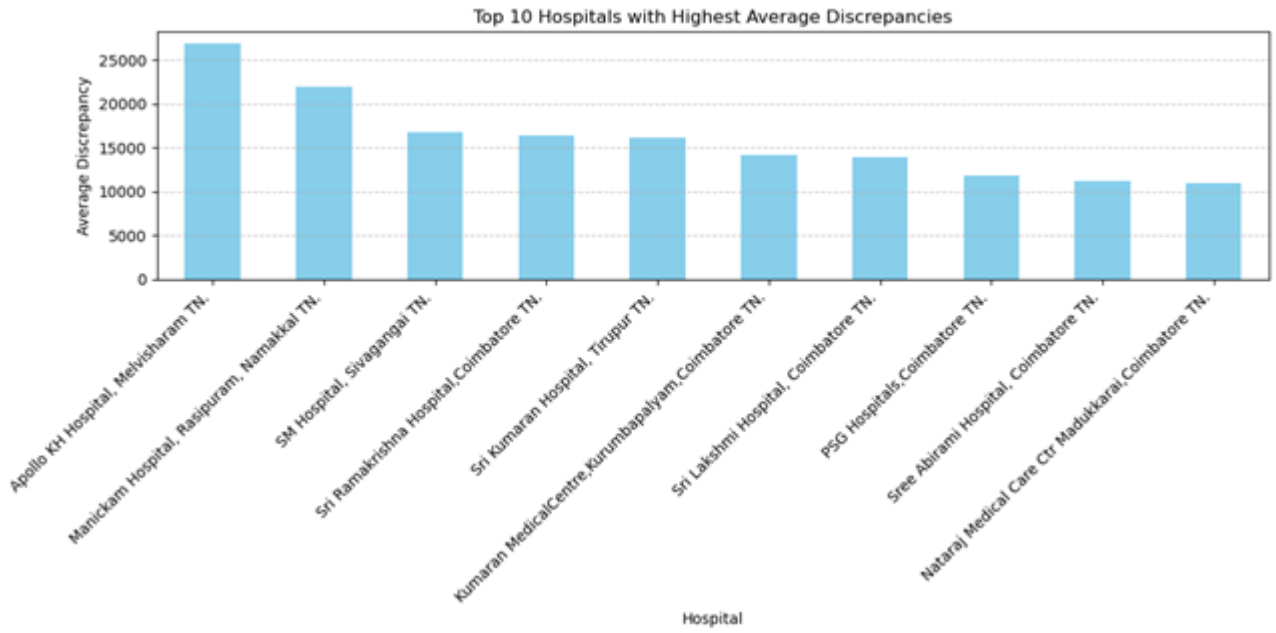


Fig No:9.1.2 Graph

The bars represent each hospital's average discrepancy, and they are ordered from the highest on the left to the lowest on the right. Apollo KH Hospital in Mancholai TN has the highest average discrepancy, followed by Meenakshi Hospital in Tanjavur TN, and so on. The hospitals from Coimbatore TN, like PSG Hospital, Sri Lakshmi Hospital, and Sree Abirami Hospital, show lower average discrepancies compared to the others on the list.

Cluster 0

Characteristics: Hospitals in this cluster exhibit higher average discrepancies between preauthorization and final approved amounts, along with higher average preauthorization and final approved amounts. These hospitals also tend to have a moderate number of cases.

Interpretation Hospitals in this cluster may be engaging in billing practices that result in significantly higher charges than initially authorized. This could include upcoding procedures or inflating costs for treatments. Further investigation into the reasons behind these discrepancies is warranted to ensure that billing practices are justified and transparent.

Cluster 1

Characteristics Hospitals in Cluster 1 have the lowest average discrepancy between preauthorization and final approved amounts. However, they handle a significantly higher volume of cases compared to other clusters. The average preauthorization and final approved amounts are relatively lower compared to Cluster 0.

Interpretation These hospitals demonstrate consistent billing practices despite handling a large number of cases. Their lower discrepancy rates could indicate efficient claims management or

adherence to standardized billing procedures. However, further scrutiny may be necessary to ensure that high case volumes are not masking potential irregularities.

Cluster 2

Characteristics: Hospitals in Cluster 2 have lower to moderate average discrepancies between preauthorization and final approved amounts. They also have a lower volume of cases compared to Cluster 1, and the average preauthorization and final approved amounts are relatively lower than both Cluster 0 and Cluster 1.

Interpretation Hospitals in this cluster may be more conservative in their billing practices, resulting in lower discrepancies between authorized and approved amounts. The lower case volume and billing amounts suggest that these hospitals may focus on less complex treatments or have tighter controls on billing processes.

Potential Fraudulent Behavior

- High Discrepancy Hospitals (Cluster 0): Hospitals with significantly higher discrepancies between preauthorization and final approved amounts may be engaging in fraudulent billing practices. This could include upcoding, unbundling services, or billing for services not rendered. Investigation into the specific procedures and billing codes used by these hospitals is crucial to identify potential fraud.
- Inconsistent Billing Patterns: Hospitals with inconsistent billing patterns, such as large discrepancies between authorized and approved amounts or unusually high case volumes, warrant further investigation. These inconsistencies could indicate potential fraudulent behavior, such as billing for unnecessary procedures or services.
- Outliers in Cluster 1: While Cluster 1 generally demonstrates consistent billing practices, outliers within this cluster with unusually low discrepancies or high case volumes may indicate irregularities. These outliers should be examined closely to ensure compliance with billing regulations and standards.
- In summary, the clustering analysis helps identify hospitals with varying billing practices and highlights potential areas of concern for fraudulent behavior. Further investigation, including detailed audits of billing records and procedures, is essential to uncover and address any fraudulent activities effectively.

9.2 REGRESSION MODEL

MODEL NAME	RMSE
LINEAR REGRESSION	3423.55
RANDOM FOREST	3655.02
SVM (SUPPORT MACHINE)	4416.33
ANN(ARTIFICIAL NERURAL NETWORK)	3750.52
XGBOOST	3628.14

Table No : 9.2 REGRESSION MODEL

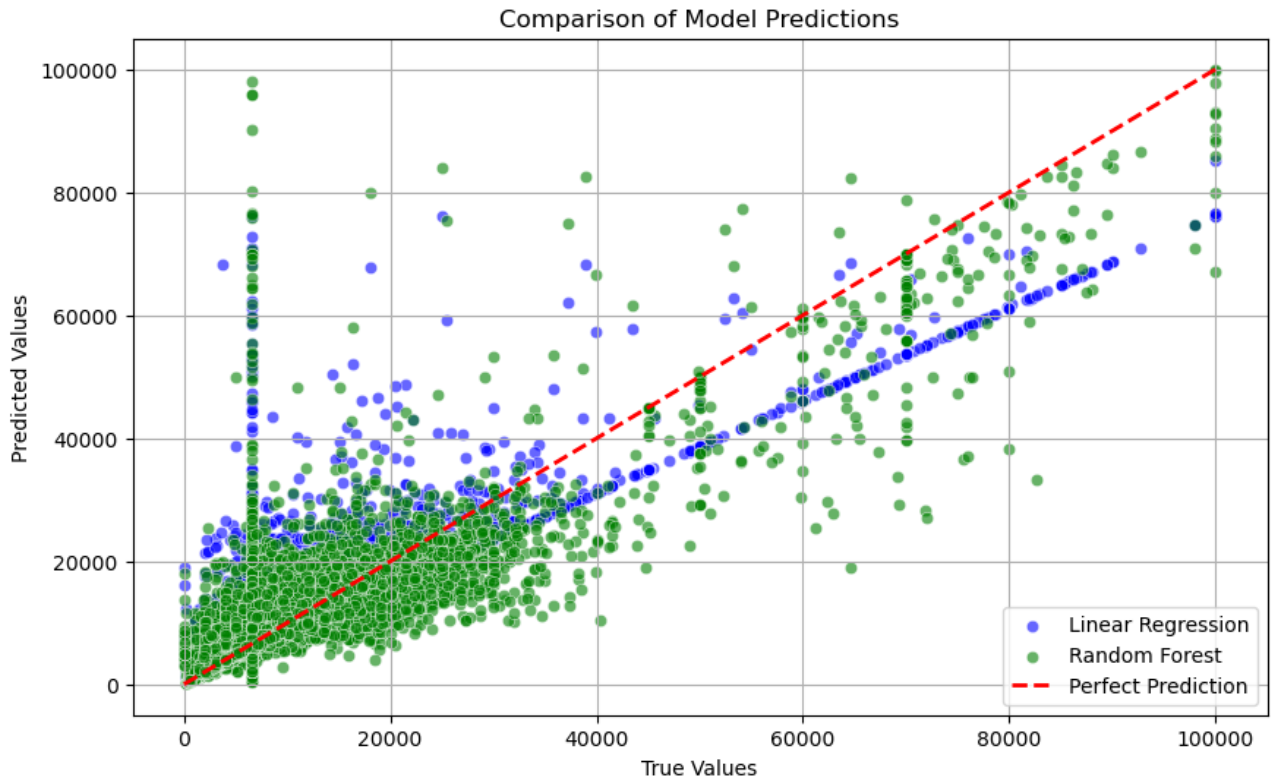


Fig 9.2.1 Regression Model

Interpretation

The x-axis represents the true values of the target variable, while the y-axis shows the predicted values by the models. The red dashed line represents the line of perfect prediction, where the predicted values perfectly match the true values.

The blue points represent predictions made by the Linear Regression model, while the green points represent predictions by the Random Forest model.

The closer the points are to the red dashed line, the more accurate the predictions. If a point lies on the line, it indicates that the prediction was exactly correct.

The spread of the points around the line of perfect prediction gives an indication of the error in the predictions. Points that are widely spread from the line indicate greater prediction error.

Both models have predictions that are spread across the range of true values, with a concentration of predictions closer to the line for lower values of the target variable. This could suggest that the models are more accurate in predicting lower values than higher ones.

There are some outliers, particularly in the higher value range, where both models tend to underpredict the actual values (as the points lie below the line of perfect prediction).

The density of points near the lower values suggests that there are more observations with lower "Final App Amt" in the dataset, and the models have more data to learn from in this range.

It appears that both models have similar predictive performances since their points largely overlap, but the Linear Regression model might have a slightly tighter cluster of points along the line of perfect prediction, suggesting a marginally better performance.

9.3 TEST CASES

Data Retrieval Test Case:

Objective: Ensure that the system can retrieve patient data from the hospital database.

Steps:

Input valid patient ID.

Request analysis for the patient.

Expected Result: The system successfully retrieves patient data from the hospital database.

Data analysis Test Case:

Objective: Verify that the system accurately analyzes patient data.

Steps:

Input valid patient data for analysis.

Analyze the data using different algorithms or models.

Expected Result: The system produces accurate analysis results based on the input data.

Machine Learning Integration Test Case:

Objective: Ensure that the system integrates with machine learning models effectively.

Steps:

Input patient data for analysis.

Apply machine learning algorithms for analysis.

Expected Result: The system successfully applies machine learning algorithms and provides meaningful insights.

Error Handling Test Case:

Objective: Validate the system's error handling mechanism.

Steps:

Input invalid patient ID or data.

Request analysis.

Expected Result: The system displays appropriate error messages and handles invalid inputs gracefully.

Performance Test Case:

Objective: Evaluate the system's performance under load.

Steps:

Simulate multiple concurrent requests for analysis.

Measure response time and system resource utilization.

Expected Result: The system handles concurrent requests efficiently without significant performance degradation.

Security Test Case:

Objective: Ensure that the system protects sensitive patient data.

Steps:

Attempt to access patient data without proper authentication.

Attempt to access restricted functionality.

Expected Result: The system denies unauthorized access and enforces proper authentication and authorization mechanisms.

CHAPTER-10

SYSTEM IMPLEMENTATION

```
#Import a packages for an Analysis
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.svm import SVR
from sklearn.neural_network import MLPRegressor
from xgboost import XGBRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline

#======================== Import a data ========================#
data = pd.read_csv("J:\\YOKESH - PANIMALR - PROJECT\\PROJECT DATA.csv", encoding='ISO-8859-1')

#================================================================================#
#======================== AGE vs GENDER ========================#
#================================================================================#

# Visualizing the distribution of "Patient Age" grouped by "Gender"
plt.figure(figsize=(12, 6))
sns.boxplot(x='Gender', y='Patient Age', data=data)
plt.title('Distribution of Patient Age by Gender')
plt.xlabel('Gender')
plt.ylabel('Patient Age')
```

```
plt.show()

# Plotting the age distribution by gender
plt.figure(figsize=(12, 6))
sns.histplot(data=data, x='Patient Age', hue='Gender', kde=True, element='step', bins=30)
plt.title('Age Distribution by Gender')
plt.xlabel('Patient Age')
plt.ylabel('Density')
plt.legend(title='Gender')
plt.show()


#================================================================================#
#================#======== PATIENT AGE vs GOVERNMENT / PROVATE HOSPITALS ==========================#
#================================================================================#

# Boxplot for Patient Age by Healthcare Provider Type
plt.figure(figsize=(12, 6))
sns.boxplot(x='Govt./Pvt.', y='Patient Age', data=data)
plt.title('Distribution of Patient Age by Healthcare Provider Type')
plt.xlabel('Healthcare Provider Type')
plt.ylabel('Patient Age')
plt.show()

# Histogram for Patient Age Distribution by Healthcare Provider Type
plt.figure(figsize=(12, 6))
sns.histplot(data=data, x='Patient Age', hue='Govt./Pvt.', kde=True, element='step', bins=30)
plt.title('Age Distribution by Healthcare Provider Type')
plt.xlabel('Patient Age')
plt.ylabel('Density')
plt.legend(title='Healthcare Provider Type')
```

```
plt.show()

#================================================================
========#
#===============#======== YEAR vs FIRST SUBMISSION DATE
=======================#
#================================================================
========#


# Convert "Preauth First Submission Time" to datetime format
data['Preauth First Submission Time'] = pd.to_datetime(data['Preauth First Submission Time'])

# Extract the year from "Preauth First Submission Time" if "Year" column does not exist
data['Year'] = data['Preauth First Submission Time'].dt.year

# Plotting the count of preauthorizations by year
plt.figure(figsize=(12, 6))
sns.countplot(x='Year', data=data)
plt.title('Count of Preauthorizations First Submission by Year')
plt.xlabel('Year')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()

#================================================================
========#
#===============#======== YEAR vs FINAL APPROVED AMOUNT vs PREAUTH
AMOUNT =======================#
#================================================================
========#

# Aggregate data
yearly_amounts = data.groupby('Year').agg({
```

```
    'Preauth Amt': 'mean',  # Replace 'mean' with 'sum' if you prefer to aggregate by total amount
    'Final App Amt': 'mean',
}).reset_index()

# Plotting
plt.figure(figsize=(14, 7))
sns.lineplot(data=yearly_amounts, x='Year', y='Preauth Amt', marker='o', label='Preauth Amt')
sns.lineplot(data=yearly_amounts, x='Year', y='Final App Amt', marker='o', label='Final App Amt')

plt.title('Comparison of Yearly Preauthorization and Final Approval Amounts')
plt.xlabel('Year')
plt.ylabel('Amount')
plt.legend()
plt.xticks(rotation=45)
plt.show()



#================================================================================#
#================#======== SCATTERPLOT - PREAUTH AMOUNT vs FINAL APPROVED AMOUNT vs YEAR ========================#
#================================================================================#

# Generating the scatter plot
plt.figure(figsize=(10, 6))

# Plot each year's data in a different color
for year in sorted(data['Year'].unique()):
    yearly_data = data[data['Year'] == year]
    plt.scatter(yearly_data['Preauth Amt'], yearly_data['Final App Amt'], label=year)

plt.title('Scatter Plot of Preauth Amt vs Final App Amt by Year')
plt.xlabel('Preauthorization Amount')
```

```
plt.ylabel('Final Approval Amount')
plt.legend(title='Year')
plt.show()

#================================================================
========#
#================#======== FINAL APPROVED AMOUNT vs GOVV HOSPITALS /
PRIVATE HOSPITALS ========================#
#================================================================
========#

# You might want to aggregate the data to compute the mean or sum of 'Preauth Amt' and 'Final
App Amt' for each 'Govt./Pvt' category
provider_amounts = data.groupby('Govt./Pvt.').agg({
   'Preauth Amt': 'mean',  # Replace 'mean' with 'sum' for total amounts
   'Final App Amt': 'mean',  # Similarly, replace 'mean' with 'sum' if desired
}).reset_index()

# Now, let's visualize this data
plt.figure(figsize=(10, 6))

# Plotting 'Preauth Amt'
sns.barplot(x='Govt./Pvt.', y='Preauth Amt', data=provider_amounts, color='blue', label='Preauth
Amt')

# Overlaying 'Final App Amt' on the same plot for direct comparison
sns.barplot(x='Govt./Pvt.', y='Final App Amt', data=provider_amounts, color='red', alpha=0.6,
label='Final App Amt')

plt.title('Comparison of Average Preauthorization and Final Approval Amounts by Provider Type')
plt.ylabel('Average Amount')
plt.legend()
plt.show()
```

```
#=================================================================
========#
#================#======== PREAUTH AMOUNT vs FINAL APPROVED AMOUNT vs
CITY=========================#
#=================================================================
========#

agg_data = data.groupby('City/Location').agg({
   'Preauth Amt': 'mean',
   'Final App Amt': 'mean'
}).reset_index()

# Now, create a bar plot to visualize this aggregated data
plt.figure(figsize=(14, 7))

# Plot for Preauth Amt
sns.barplot(data=agg_data, x='City/Location', y='Preauth Amt', color='blue', label='Preauth Amt')

# Overlay for Final App Amt using a different color and transparency
sns.barplot(data=agg_data, x='City/Location', y='Final App Amt', color='red', alpha=0.6,
label='Final App Amt')

plt.title('City/Location-wise Comparison of Preauthorization and Final Approval Amounts')
plt.xlabel('City/Location')
plt.ylabel('Average Amount')
plt.xticks(rotation=45)  # Rotate labels if they're crowded
plt.legend()
plt.tight_layout()
plt.show()


#=================================================================
========#
```

```
#======================== MACHINE LEARNING MODEL========================#
#======================== 1. K MEANS CLUSTERING========================#
#================================================================================#

# Calculate the discrepancy between pre-authorized amounts and final approved amounts
data['Discrepancy'] = data['Preauth Amt'] - data['Final App Amt']

# Summary statistics of discrepancies
discrepancy_summary = data['Discrepancy'].describe()

# Identify hospitals with high average discrepancies
# Group by hospital and calculate the mean discrepancy for each
hospital_discrepancy = data.groupby('Hospital')['Discrepancy'].mean().sort_values(ascending=False)

# Display summary statistics of discrepancies and the top 10 hospitals with highest average discrepancies
discrepancy_summary, hospital_discrepancy.head(10)

# Plotting the hospital discrepancies
plt.figure(figsize=(12, 6))
hospital_discrepancy.head(10).plot(kind='bar', color='skyblue')
plt.title('Top 10 Hospitals with Highest Average Discrepancies')
plt.xlabel('Hospital')
plt.ylabel('Average Discrepancy')
plt.xticks(rotation=45, ha='right')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

```
# Prepare a DataFrame for clustering
data_for_clustering = hospital_discrepancy.reset_index().rename(columns={'Discrepancy': 'Average_Discrepancy'})
data_for_clustering['Total_Cases'] = data.groupby('Hospital').size().values
data_for_clustering['Average_Preauth_Amt'] = data.groupby('Hospital')['Preauth Amt'].mean().values
data_for_clustering['Average_Final_App_Amt'] = data.groupby('Hospital')['Final App Amt'].mean().values

# Impute missing values with the mean
imputer = SimpleImputer(strategy='mean')
data_for_clustering_imputed = imputer.fit_transform(data_for_clustering[['Average_Discrepancy', 'Total_Cases', 'Average_Preauth_Amt', 'Average_Final_App_Amt']])

# Scale the data
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_for_clustering_imputed)

# Apply K-means clustering with k=3
k = 3
kmeans = KMeans(n_clusters=k, random_state=42)
clusters = kmeans.fit_predict(data_scaled)

# Add cluster information back to the hospital discrepancy DataFrame
data_for_clustering['Cluster'] = clusters

# Inverse transform the cluster centers to get them back to the original scale for interpretation
cluster_centers = scaler.inverse_transform(kmeans.cluster_centers_)

# Adding cluster center information for interpretability
for i in range(k):
    data_for_clustering.loc[data_for_clustering['Cluster'] == i, 'Cluster_Center_Average_Discrepancy'] = cluster_centers[i, 0]
```

```
    data_for_clustering.loc[data_for_clustering['Cluster'] == i, 'Cluster_Center_Total_Cases'] = cluster_centers[i, 1]
    data_for_clustering.loc[data_for_clustering['Cluster'] == i, 'Cluster_Center_Average_Preauth_Amt'] = cluster_centers[i, 2]
    data_for_clustering.loc[data_for_clustering['Cluster'] == i, 'Cluster_Center_Average_Final_App_Amt'] = cluster_centers[i, 3]

# Display an overview of the clusters
data_for_clustering.groupby('Cluster').agg(
    {
        'Average_Discrepancy': 'mean',
        'Total_Cases': 'mean',
        'Average_Preauth_Amt': 'mean',
        'Average_Final_App_Amt': 'mean',
        'Cluster_Center_Average_Discrepancy': 'first',
        'Cluster_Center_Total_Cases': 'first',
        'Cluster_Center_Average_Preauth_Amt': 'first',
        'Cluster_Center_Average_Final_App_Amt': 'first'
    }
)



# Set up the matplotlib figure
plt.figure(figsize=(14, 7))

# Scatter plot for clusters
sns.scatterplot(x='Average_Preauth_Amt', y='Average_Final_App_Amt', hue='Cluster', palette='viridis', data=data_for_clustering, s=100, alpha=0.8)

plt.title('Cluster Analysis of Hospitals Based on Preauth and Final Approved Amounts')
plt.xlabel('Average Preauthorization Amount')
plt.ylabel('Average Final Approved Amount')
plt.legend(title='Cluster')
```

```
plt.grid(True)

plt.show()


#======================== 2. LINEAR REGRESSION MODEL=========================#
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
import numpy as np

# Selecting features and target variable
features = ['Patient Age', 'Preauth Amt']
target = 'Final App Amt'

# Handling missing values - filling with median for numerical columns
data[features] = data[features].fillna(data[features].median())
data[target] = data[target].fillna(data[target].median())

# Splitting the dataset into training and testing sets
X = data[features]
y = data[target]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Linear Regression Model
model = LinearRegression()

# Training the model
model.fit(X_train, y_train)
```

```
# Predicting on the test set
y_pred = model.predict(X_test)

# Calculating the Root Mean Squared Error (RMSE)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))

rmse

#=======================3. RANDOM FOREST REGRESSOR MODEL=======================#

from sklearn.ensemble import RandomForestRegressor

# Random Forest Regressor Model
random_forest_model = RandomForestRegressor(n_estimators=100, random_state=42)

# Training the model
random_forest_model.fit(X_train, y_train)

# Predicting on the test set
y_pred_rf = random_forest_model.predict(X_test)

# Calculating the Root Mean Squared Error (RMSE)
rmse_rf = np.sqrt(mean_squared_error(y_test, y_pred_rf))

rmse_rf
```

CHAPTER-11

CONCLUSION AND FUTURE ENHANCEMENT

11.1 CONCLUSION

Fraud Detection and Prevention: The research successfully identifies patterns indicative of fraudulent claims using logistic regression and anomaly detection techniques. This proactive approach not only aids in mitigating financial losses but also ensures the integrity and credibility of the health insurance system.

Forecasting and Trend Analysis: The application of linear regression and other predictive models enables the anticipation of future claim patterns, helping stakeholders make informed decisions regarding resource allocation and policy formulation. This predictive capability is crucial for adapting to evolving healthcare needs and economic conditions.

Resource Utilization Efficiency: Detailed analysis of equipment and hospital-wise claim data reveals significant insights into the allocation and utilization of healthcare resources. Identifying areas of underutilization and inefficiency facilitates targeted interventions to improve service delivery and healthcare access.

Optimization of Health Insurance Utilization: The project outlines strategies to enhance the efficiency, equity, and resilience of the health insurance system. Recommendations are provided for improving healthcare access, affordability, and the overall quality of care through optimized insurance utilization and fraud prevention measures.

Collaboration and Data-Driven Decision Making: Emphasizing the importance of collaboration among healthcare stakeholders, the study showcases how data-driven insights can lead to transformative changes in the healthcare system. Engaging with healthcare providers, policymakers, and insurance companies ensures that the findings are relevant and actionable.

Innovation in Healthcare Administration: By leveraging advanced data analysis techniques and machine learning algorithms, the project exemplifies how engineering innovation can address complex challenges in healthcare administration. The scalable solutions and methodologies developed have the potential to be applied in broader contexts beyond Tamil Nadu.

In conclusion, this thesis demonstrates the power of data-driven governance in health insurance, highlighting the role of engineering innovation in enhancing the efficiency, transparency, and equity of healthcare systems

11.2 FUTURE ENHANCEMENT

Real-Time Data Streaming: Implement a feature to stream real-time data from hospitals or healthcare providers. This could involve integrating with IoT devices, wearables, or electronic health records (EHR) systems to gather continuous patient health data.

Predictive Analytics: Enhance the system's analytics capabilities by incorporating predictive modeling techniques. This could help in predicting health risks, identifying potential fraud cases, and optimizing insurance premiums.

Natural Language Processing (NLP): Integrate NLP algorithms to analyze unstructured data such as medical notes, doctor's prescriptions, or patient feedback. This could provide deeper insights into patient conditions and treatment outcomes.

Enhanced Security Measures: Strengthen the system's security by implementing advanced encryption techniques, multi-factor authentication, and regular security audits. This is crucial for protecting sensitive patient data from unauthorized access and cyber threats.

Personalized Recommendations: Develop personalized health recommendations based on individual patient data and analysis results. This could include lifestyle modifications, preventive care measures, and tailored insurance plans.

Telemedicine Integration: Integrate telemedicine features to enable virtual consultations and remote monitoring of patients. This would improve accessibility to healthcare services and facilitate better communication between patients and healthcare providers.

Blockchain for Data Integrity: Explore the use of blockchain technology to ensure the integrity and traceability of health insurance data. This could help in maintaining transparent and immutable records of transactions, claims, and patient histories.

Mobile Application: Develop a mobile application for patients to access their health insurance information, track their claims, and receive personalized health recommendations. This would enhance user experience and convenience.

Data Visualization Tools: Implement interactive data visualization tools to present analysis results in a user-friendly and visually appealing manner. This could include charts, graphs, and dashboards to help users understand complex health trends and patterns.

Compilance with Regulations: Stay updated with evolving regulatory requirements in the healthcare and insurance industries. Ensure compliance with standards such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) to maintain data privacy and security.

.

CHAPTER-12
REFERENCES

INSURANCE THEORY AND PRACTICE" BY ROB THOYTS

This book provides an introduction to the principles of insurance, covering various aspects of both theory and practice. It offers insights into the history of insurance, the principles of risk management, and the workings of the insurance market.

PRINCIPLES OF RISK MANAGEMENT AND INSURANCE" BY GEORGE E. REJDA & MICHAEL MCNAMARA

A comprehensive text that covers the fundamentals of risk management and insurance. It explores various insurance products, the process of managing risk, and the operational aspects of insurance companies.

PATTERN RECOGNITION AND MACHINE LEARNING" BY CHRISTOPHER M. BISHOP

This book provides an in-depth look at the methods and algorithms that form the foundation of machine learning. It's well-regarded for its clear explanations and practical examples.

HANDS-ON MACHINE LEARNING WITH SCIKIT-LEARN, KERAS, AND TENSORFLOW" BY AURÉLIEN GÉRON

A practical guide to implementing machine learning with Python's Scikit-Learn, Keras, andTensorFlow. This book covers the fundamentals of machine learning, including neural networks, with hands-on examples and exercises.

EXPLORATORY DATA ANALYSIS" BY JOHN W. TUKEY

This classic book by John Tukey, the founder of EDA, introduces the approaches and techniques of EDA. It emphasizes understanding data by using graphical and numerical methods rather than making assumptions about what the data might reveal.

PYTHON DATA SCIENCE HANDBOOK" BY JAKE VANDERPLAS

A comprehensive guide to data science using Python, covering essential tools and libraries including NumPy, pandas, Matplotlib, and Scikit-Learn. It includes sections dedicated to data manipulation, visualization, and machine learning, with a focus on exploratory data analysis.

KAGGLE

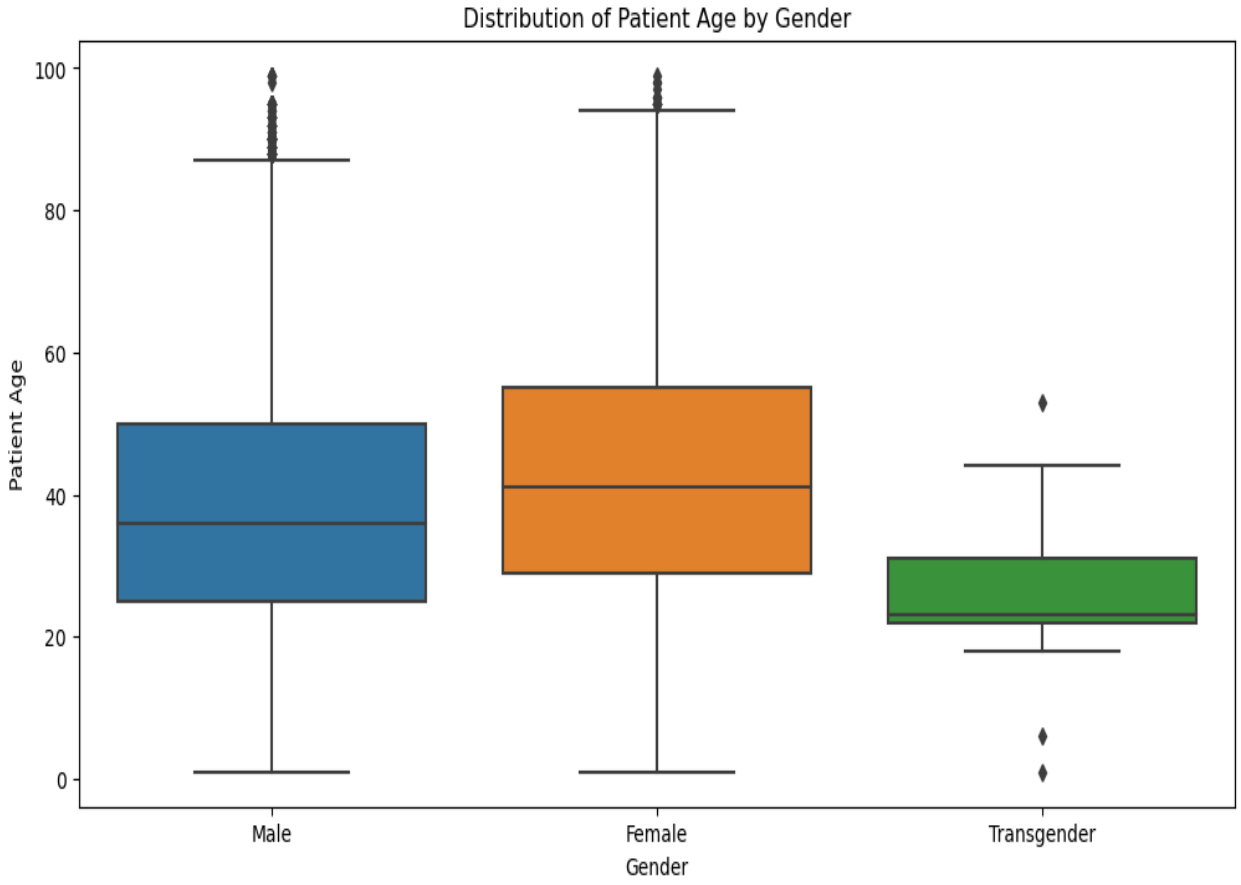
An online community of data scientists and machine learning practitioners. Kaggle offers competitions, datasets, and notebooks that can be very helpful for practicing EDA and machine learning techniques.

TOWARDS DATA SCIENCE ON MEDIUM

A platform with articles and tutorials covering a wide range of topics in data science, machine learning, and analytics, including EDA techniques and best practices.

APPENDICES

A.1 SCREEN SHOT



A.1.1 DISTRIBUTION OF PATIENT AGE BY GENDER

The Male group has a wide age range with a fairly symmetric distribution around the median. The median age appears to be around 50. The Female group has a similar range and distribution, but the median age seems to be slightly lower than that of the male group.

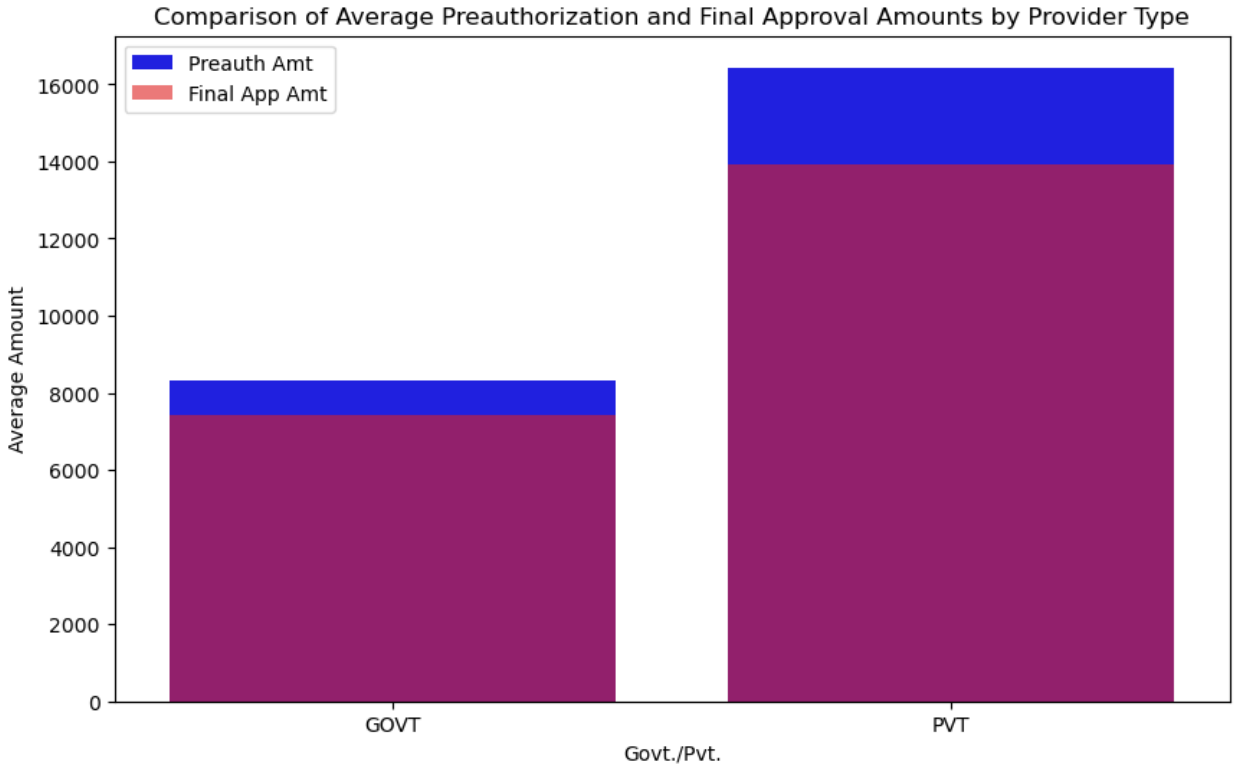


FIG A.1.2 COMAPRISON OF AVERAGE PREAUTHORIZED AND FINAL APPROVAL AMOUNT BY PROVIDER TYPE

For both government and private providers, the final approval amounts are greater than the preauthorization amounts, indicated by the purple section being on top of the blue section. The government provider (GOVT) has a lower total average amount (preauthorization plus final approval) than the private provider (PVT).

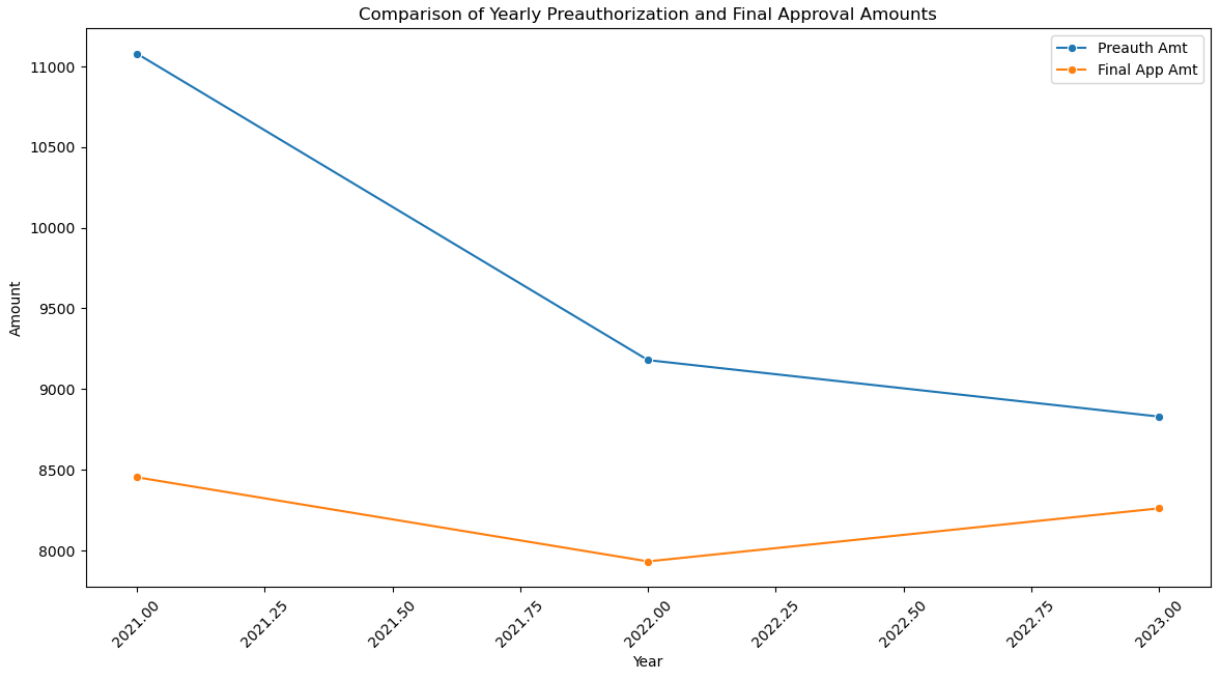


FIG A.1.3 COMPARIONS OF YEARLY PREAUTHORZED AND FINAL APPROVAL AMOUNT

X-Axis (Year): It seems to show the quarters of a year, starting from '2021-Q0' and moving to '2023-Q0'. The quarters are represented as '2021-00', '2021-25', '2021-50', '2021-75', etc., which are unconventional markings for quarters. Typically, quarters are denoted as Q1, Q2, Q3, and Q4.

Y-Axis (Amount): The y-axis represents the amount, which could be in any currency, and is used as a common scale for both the preauthorization and final approval amounts.

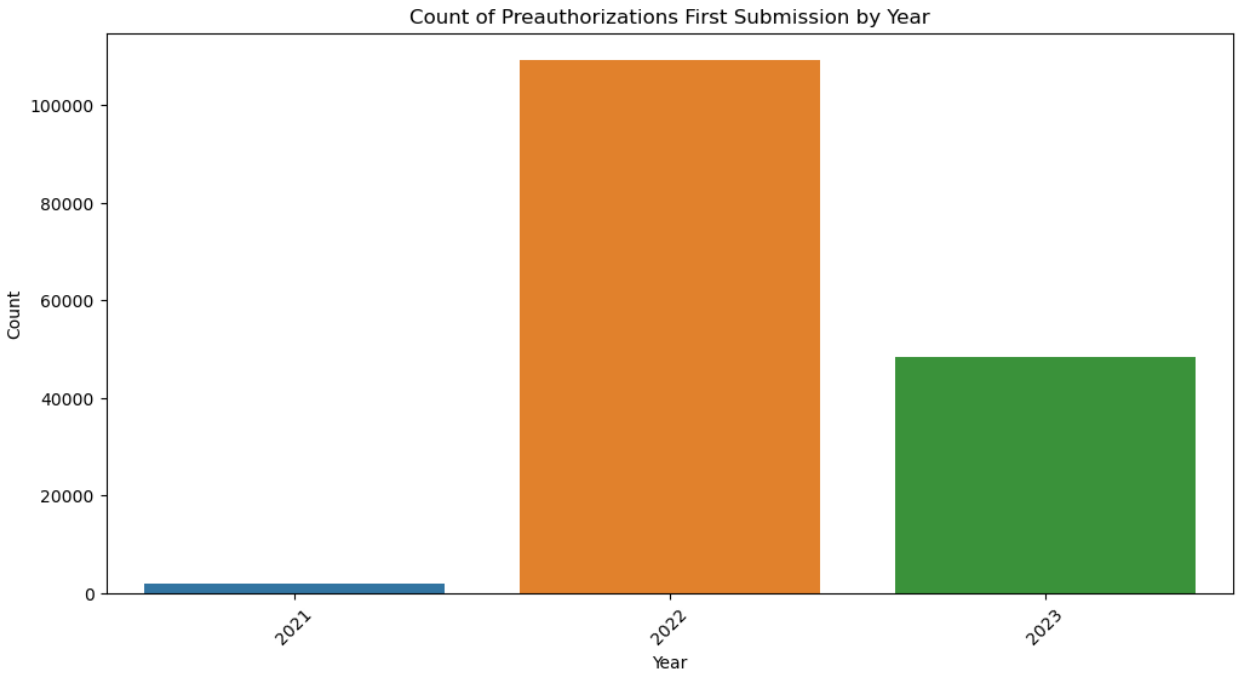


FIG A.1.4 COUNTOF PREAUTHORIZED FIRST SUBMISSION BY YEAR

X-Axis (Year): Each bar corresponds to a different year, with years listed as 2021, 2022, and 2023.

Y-Axis (Count): This axis indicates the count of preauthorizations' first submissions.

Bars: Each bar shows the count for that particular year.

The bar for 2021 is colored blue and is very low compared to the others, suggesting a much smaller number of preauthorization submissions in that year.

A.2 PLAGIARISM REPORT

REFERENCES

Plagiarism Checker X - Report

Originality Assessment

6%

Overall Similarity

Date: Mar 4, 2024

Matches: 168 / 2968 words

Sources: 9

Remarks: Low similarity detected, check with yoursupervisor if changes are required.

Verify Report:

Scan this QR Code

v 8.0.7 - WML 4FILE - MAIN PROJECT CM ANALYSIS OF HEALTH INSURANCE FINAL PAPER[1].DOC

"COMPREHENSIVE ANALYSIS OF CM - HEALTH INSURANCE DATA IN TAMIL NADU"

Mr.DR.N.Pughazendi Narayanan,M.E,Phd
Guide and coordinator
CSE
Panimalar Engineering College
Chennai, Tamil Nadu

Yokesh
CSE
Panimalar Engineering College
Chennai, Tamil Nadu
Yokesh18yokesh@gmail.com

Bragadeeshwaran
CSE
Panimalar Engineering
College Chennai, Tamil Nadu
Deesh299@gmail.comv

Dharshan
CSE
Panimalar Engineering college
Chennai,Tamil Nadu
Dharshanmadhavan18@gmail.com

***Abstract*—**

This project delves into a comprehensive analysis of health insurance data in Tamil Nadu, focusing on district-wise and region-wise claims, specifically within areas such as General Hospitals (GH), Primary Health Care Centers (PHCC), and Medical College Hospitals (MCH). Through detailed examination, we aim to identify patterns and variations in claim occurrences, shedding light on the impact of geographic and facility-specific factors.Furthermore, the project involves a temporal analysis to discern monthly variations in claims, helping to pinpoint trends and potential influencing factors over time. By understanding the temporal dynamics, we can enhance the efficiency of resource allocation and policy planning.Additionally, the study investigates hospitals that frequently raise Preath requests and submit claims. This scrutiny is vital for identifying potential areas of improvement in the insurance validation process.Finally, leveraging machine learning techniques, the project endeavors to predict future claims for hospitals. This predictive modeling aims to provide insights into the anticipated insurance workload, aiding in proactive planning and resource management. Through this multi-faceted analysis, the project aspires to contribute valuable insights for optimizing the health insurance system in Tamil Nadu.

INTRODUCTION

Analyzing year-on-year health insurance claims in Tamil Nadu hospitals

Leveraging advanced Data Analysis techniques to derive meaningful insights from Health Insurance Claims data.

Specific focus on equipment-wise and Hospital-wise utilization

Identify fraudulent claims and optimize insurance utilization

Implementation of sophisticated methodologies for strategic detection of potentially fraudulent claims,

contributing to a more secure and reliable health insurance system.

Through collaboration with healthcare stakeholders, the findings of this analysis aspire to catalyze positive changes, foster transparency, and ensure the fair and effective utilization of health insurance resources in Tamil Nadu.

LITERATURE SURVEY

Trends Over Time:

Research on temporal pattern analysis within healthcare claims, focusing on changes in medical equipment usage and hospital claims. Look for papers on longitudinal healthcare data studies, time-series analysis, and year-on-year comparison methodologies.

Equipment and Hospital Evaluation:

Studies on hospital resource utilization, equipment usage efficiency, and district-level healthcare service analysis. Seek out literature that evaluates healthcare delivery and claims at the institutional level.

Fraud Detection and Prevention:

Investigations on healthcare fraud, with a focus on preauthorization processes. Sources could include articles on the application of anomaly detection, data mining, and machine learning models specifically designed to detect irregular patterns indicative of fraud.

Machine Learning for Predictive Modeling:

Look for current advancements in applying machine learning to predict healthcare claims and potential fraud, including the use of predictive analytics to improve resource allocation and anticipate fraudulent activities.

Optimization of Health Insurance Utilization:

Research on improving health insurance systems through data analysis. This includes optimization models for healthcare planning, resource management, and fraud prevention strategies.the health data should be utilised efficiently and calculation of the cost value of health insurance should be done in proper method.

EXISTING SYSTEM

This section typically describes the current state of operations or systems in place before the implementation of the project.

It may discuss the limitations, challenges, or inefficiencies that the current system faces.

For your health insurance data analysis project, the existing system may refer to the current methods of processing and analyzing health insurance claims, detecting fraud, and utilizing health insurance resources in Tamil Nadu.

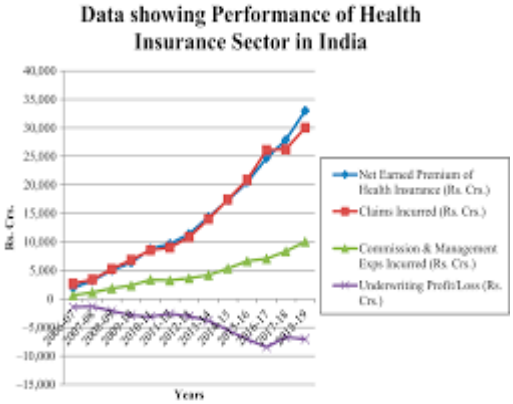
It might also cover the existing tools and techniques used for data analysis in this context.

PROPOSED SYSTEM

The proposed system section would outline the new system, methodologies, or processes that are being suggested to improve upon the existing system.

It would describe the new data analysis techniques, machine learning models for predictive modeling, and fraud detection strategies you intend to implement.

The proposed system aims to address the shortcomings of the existing system by optimizing health insurance utilization, enhancing fraud detection, and making the overall system more efficient and effective.



Source: Author's compilation

TOOLS USED

Utilized Python for data analysis and modeling.

Leveraged popular Python libraries such as NumPy, pandas, scikit-learn.
Used Matplotlib and Seaborn for data visualization.
Combined statistical analysis, machine learning, and data visualization for a comprehensive study.

RESULTS AND DISCUSSION

- The **findings from the data analysis**, such as the identified trends in health insurance claims, equipment-wise and hospital-wise evaluations, and any detected fraudulent activities.
- **Comparisons** of these findings with the expectations or hypotheses stated at the outset of the project.
- **Interpretation** of what these findings mean for the stakeholders, such as healthcare providers, insurance companies, and policymakers.
- **Challenges and limitations** encountered during the implementation of the system and how they were addressed.
- **The significance of the results** in terms of their impact on improving health insurance utilization and fraud detection.
- **Future implications**, including how the results can inform further research or operational changes within the health insurance domain.

.

Methods

Section II: Techniques

A. Summary statistics and data collection
1) A general synopsis of the data source and data summary

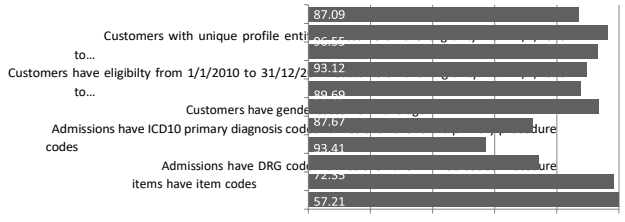
Health care records produced when hospitals submit claims to an insurance provider in order to be paid for their services were used in this study. Three years' worth of data were

provided, with the first two years (01/01/2010 to 31/12/2011) acting as a two-year observation period and the third and final year (1/1/2012 to 31/12/2012) as a one-year forecast period. Results for the third year were predicted using data from the two-year observation period.
One of Australia's largest combined registered private health fund and life insurance companies, Hospitals Contribution Fund of Australia (HCF), provided hospital claims data for 242,075 individuals in this data set.

groups. The data set contained enrollment details regarding the duration that a person or his or her family was covered by the insurance policy, hospital procedure claims, and admission administrative records. In order to avoid the potential of real people being identified, the data set was obtained after pseudonymization. It also included the basic demographic data of the customers, such as age and gender.
2) Condensed statistical data: The fundamental statistics on the data are compiled in the following tables. The demographics, including age and gender, are displayed in Table I. Table II displays the breakdown of
The sample demographic statistics for age and gender in 2010 are shown in Table I.

Customers who were not hospitalized were charged in some way, possibly because of a data error. The real dataset is represented by this number.
days spent in various hospital bins (DIH). The data used in both tables comes from the calendar year 2010.
B. Aggregation and data manipulation
The raw data was recoded and arranged into a format that made efficient computing possible before a predictive model was built.
1) There are three informational levels: Three tiers of information were included in the data: hospital procedure claim, hospital admission, and customer.

client level included details on the health insurance policies that customers purchase as well as demographics about the client, such as age and gender. Each client had a minimum of one entry containing this kind of data. Clients who made changes to their personal data Records from the research period may contain duplicate information.
When a consumer provided numerous recordings, just the first record—the one that was closest to the research period's beginning—was taken.
Hospital admission claims, which contain comprehensive data about the client and the provider, were included in the hospital admission level. Hospital admission claims comprise crucial data items such as the primary diagnosis, primary procedure, secondary diagnosis, and the provider. Inpatient and outpatient admissions to hospitals can be broadly divided into two categories. This portion of the data was unavailable because HCF does not cover outpatient services. When a patient is admitted to a facility for either an overnight or same-day stay, inpatient claims are created. The length of inpatient stays means that there may be more than
one record of the admission claim; for every admission, these records were combined to form a single, distinct entity. Both same-day admission and one-day overnight admittance were counted as one day for the purpose of calculating DIH. A client could have several admissions in a given year, meaning they possessed several entities at this level.



Hospital services provided during an admission are included in hospital procedure claims. It contained details about the services being provided, including the kind of item (e.g., same-day accommodation, overnight accommodation, prosthesis, theater, etc.). It also included details on how much each treatment would cost. A hospital admission record was linked to each procedure claim, and a single hospital admission record could be connected to more than one treatment claim.

Primary diagnostic code: Every hospital admission would be linked to a diagnosis or ailment that was thought to be the primary cause. The international statistical classification of diseases and related health problems, tenth revision, Australian modification, or ICD-10-AM, was used to code it [18].

Primary procedure code: The procedure code for the admission that required the most hospital resources was also supplied. The Australian Classification of Health Interventions (ACHI) created the used coding method [19].
Group related to diagnosis: Hospital admissions were also linked to codes for the Australian Refined Diagnosis-Related Group (AR-DRG) [20]. Acute inpatients were classified using this code. The methodology relied on the codes assigned to diagnoses and procedures for every care episode.
Medicare Benefits Schedule (MBS) codes were mostly utilized in this context. These codes listed the Medicare services that are sponsored by the Australian government, encompassing a broad range of consultations, treatments, and tests, together with the schedule fees associated with each of these things [21]. Other codes such as ICD-10 and AR-DRG were used in place of MBS codes if none could be located. There was a reference table for disease codes that included links to each sickness code for the treatment type, illness group (a sensible way for HCF to group illnesses), and specialized description.
Item code: The different sorts of procedure items described the type of service or process that was conducted.

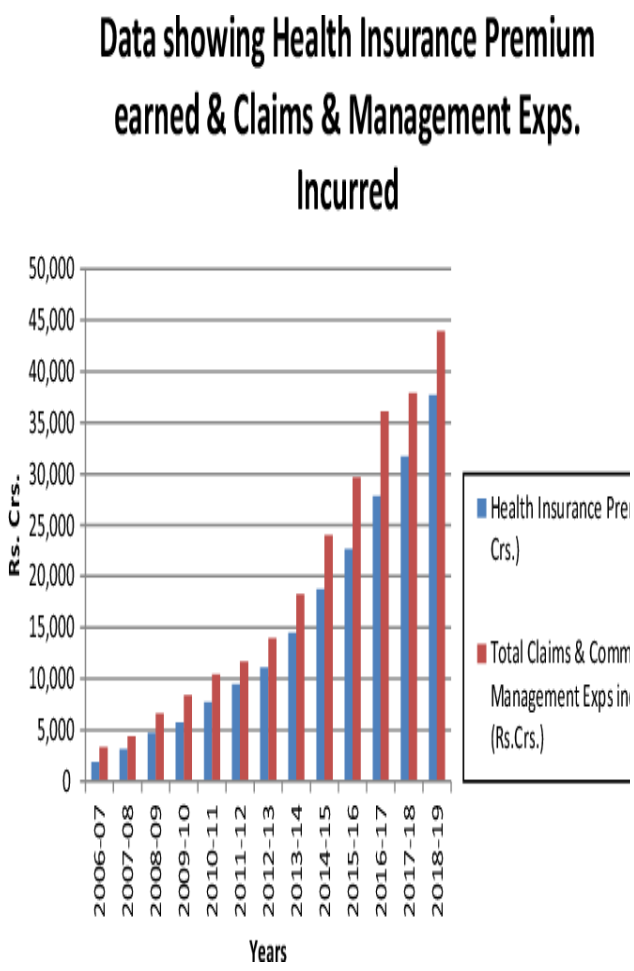
Other codes such as ICD-10 and AR-DRG were used in place of MBS codes if none could be located. There was a reference table for disease codes that included links to each sickness code for the treatment type, illness group (a sensible way for HCF to group illnesses), and specialized description.
Code for item: Different sorts of procedure items were available, each characterizing the type of service provided or operation carried out. Every procedure item had an item code, and the format varied based on the kind of item. For instance, different characteristics, such as the kind of lodging, the primary patient categorization, and the hospital type, indicated different information for hospital accommodation claims for public and private hospitals. Generally speaking, the final two characters for theater denoted the theatrical fee band.
Moreover, the supplier for prosthesis was specified by the first two characters. Additionally linked to advertisements, secondary diagnosis codes were solely utilized in the computation of the Charlson Index, which is expounded upon in Section II-B4. Secondary diagnostic codes were not introduced as significant clinical factors in this case since only a tiny percentage of diagnosis codes could be used to produce the Charlson Index.

Data completeness:
When analyzing data, it's important to take data completeness into consideration. The indicators were measured in order to evaluate the completeness. Between January 1, 2010 and December 31, 2011, 233,716 (96.55%) of the 242,075 customers were eligible for at least one year. Of them, 225,421 (93.12%) were eligible for at least two full years. Of those, 217,111 (89.69%) were eligible for eligibility that extended beyond the outcome period. Naturally, the clientele changed over time as new members joined the fund or departed (often due to death). An ideal scenario for modeling would be for all participants to

remain in the fund throughout the duration of the three years that are being studied. But as the clientele is ever-changing, forecasts could also be completed for a few clients who will leave the fund in the upcoming year. Thus, a full three-year course of treatment is not required of the subjects. All we required was for the clients to have enrolled by January 1, 2010, at the latest. Furthermore, 87.09% of consumers had distinct profile entities that were free of duplication, as can be shown. Customers provided age and gender information in 93.41% and 87.67% of cases, respectively. With respect to the five medical data elements, 72.11% of admissions had DRG codes, 98.26% had illness codes, 57.21% had ICD-10 primary procedure codes, and 72.35% had clear ICD-10 primary diagnosis codes. Almost all hospital procedure claims had item codes assigned to them. Because of this information, the primary diagnosis, primary procedure, and DRG code completeness was not exact. was not accessible to HCF, a private health insurer, for visits and treatments at public hospitals that were reimbursed by Medicare because HCF and Medicare do not exchange this information.
4. Extraction of features: Since the source variables included a variety of data types—including dates, numbers, and text—they had to first undergo pre-processing in order to be utilized for modeling. The numeric format of the numerical variables was maintained. To save computing memory, values for each category variable were changed to integers, making the feature matrix as a whole numeric. The subsequent actions were taken when the feature matrix was reduced to a numeric representation.
In order to create new features through computation and expansion, source variables were first processed at their original information levels.
Additional features that required more computation were generated by the application of computational methods. Apart from its numerical representation, age was classified into distinct age ranges.

'Age10', a category feature, was produced by decade. There was also a binary indicator that indicated if the user was older than 60 or not. They were decoded into additional features, such as main disease category (DRG MDC), major disease sub categories (DRG subMDC), and comorbidity or complication (DRG CC), in addition to employing numerical representations of the primary AR-DRG codes (DRG TEXT and DRG BASIC) [20]. For every hospital admission, there was a list of ICD-10 coded secondary diagnoses in addition to the primary diagnosis. Co-morbidity scores were calculated from the secondary diagnoses using the corresponding look-up tables, which provide weights.
accordance to the original and modified Charlson Index [22], [23], respectively, to specific ICD-10 disorders. Additionally, two "moment" features were calculated, which, in a linear (first order momentum) or quadratic (second order momentum) manner, weight the number of admissions in a given month by the month's count number, or 1 to 12. These characteristics are meant to convey the idea that admissions that happen closer to the end of the previous year have a higher chance of being followed by admissions in the subsequent year. The characteristics of the kind of accommodation (IC ACC kind) and the patient classification (IC PAT CLASS) for hospital accommodations were taken from the codes corresponding to the accommodations. With regard to categorical variables, expansion was used. To create more binary grouping features, a "binary" expansion was applied to categorical variables with few unique categories, like "Age10" and gender. Every grouping characteristic was a binary indication, where a value of "0" meant an entity did not fit into this category and a value of "1" meant an entity did.
In order to reduce the number of distinct categories for certain of those category variables, extra features were created utilizing an external hierarchical grouping method. For example, although coding for medical claims begins with a physician, it is typically completed and submitted by a separate dedicated billing operator (see

ICD-10 primary diagnosis code discussed in Section II-B2).



CONCLUSION

A method for predicting future days in hospital has been developed using features extracted from customer demographics, past hospital admission and hospital procedure claim data. The model was developed using data from an observation period of two years and was later evaluated and created into new health data which can be used to predict the health insurance cost.

REFERENCES

- **Healthcare Fraud Detection:**
  - Bauder, R. A., & Khoshgoftaar, T. M. (2018). A survey on the state of healthcare upcoding fraud analysis and detection. Health Services and Outcomes Research Methodology, 18(1), 31-55.
  - Kumar, S., & Spangler, W. S. (2015). Healthcare fraud detection: A survey and a clustering model incorporating geo-location information. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- **Machine Learning in Health Insurance:**
  - Liang, H., Tsai, C. L., & Wu, H. (2017). Modelling healthcare data using multiple-channel latent Dirichlet allocation. Journal of Biomedical Informatics, 69, 60-73.
  - Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE, 12(4), e0174944.
- **Health Insurance Claims Analysis:**
  - Bertsimas, D., Bjarnadottir, M. V., Kane, M. A., Kryder, J. C., Pandey, R.,

Vempala, S., & Wang, G. (2008). Algorithmic prediction of health-care costs. Operations Research, 56(6), 1382-1392.

- Johnson, A. E., Pollard, T. J., & Mark, R. G. (2017). Machine learning and decision support in critical care. Proceedings of the IEEE, 104(2), 444-466.

- **Data Analysis and Visualization in Healthcare:**

- Krumholz, H. M. (2014). Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. Health Affairs, 33(7), 1163-1170.
- Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 3.

REFERENCES

Healthcare Fraud Detection:

Bauder, R. A., & Khoshgoftaar, T. M. (2018). A survey on the state of healthcare

upcoding fraud analysis and detection. Health Services and Outcomes Research Methodology, 18(1), 31-55.

Kumar, S., & Spangler, W. S. (2015). Healthcare fraud detection: A survey and a clustering model incorporating geo-location information. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Machine Learning in Health Insurance:

Liang, H., Tsai, C. L., & Wu, H. (2017). Modelling healthcare data using multiple-channellatent Dirichlet allocation. Journal of Biomedical Informatics, 69, 60-73.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE,12(4), e0174944.

Health Insurance Claims Analysis:

Bertsimas, D., Bjarnadottir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., & Wang, G. (2008). Algorithmic prediction of health-care costs. Operations Research, 56(6), 1382-1392.

Johnson, A. E., Pollard, T. J., & Mark, R. G. (2017). Machine learning and decision support in critical care. Proceedings of the IEEE, 104(2), 444-466.

Data Analysis and Visualization in Healthcare:

Krumholz, H. M. (2014). Big data and new knowledge in medicine: The thinking, training,and tools needed for a learning health system. Health Affairs, 33(7), 1163-1170.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 3.