Advance Data Mining Applications in Life Insurance

Dr. Tapaswini Nayak
CSE Department
MITS
Rayagada, Odisha.
nayak_roma@yahoo.co.in

Dr. Gouri Kumar Panda
CSE Department
MITS
Rayagada, Odisha.
drgkpmail@gmail.com

Dahnanjay Kumar P
CSE Department
MITS
Rayagada, Odisha.
balaramnaik24557@gmail.com

Abstract-In today's fast-evolving insurance landscape, companies struggle to optimize customer acquisition, retention, and risk management amidst increasing data complexity. This smart insurance analytics framework addresses these challenges by deploying an integrated, data-driven system that enhances decision-making and operational efficiency. Leveraging advanced predictive modeling, the framework employs survival analysis to predict policy lapses, uplift modeling to boost customer retention, and graph-based algorithms for real-time fraud detection. A built-in explainability module ensures transparency, while fairness metrics and continuous operational monitoring maintain regulatory compliance. In cases of detected anomalies, such as potential fraud or customer churn risk, the system triggers automated alerts to stakeholders, prompting timely interventions. If initial actions do not yield desired outcomes, periodic reminders and adaptive strategies are deployed to re-engage customers or refine risk assessments. By incorporating causal inference and temporal dynamics, this framework minimizes inefficiencies, reduces financial losses, and ensures ethical alignment, providing a scalable and practical solution for life insurance analytics.

Keywords: Life insurance analytics, Predictive modeling, Survival analysis, Explainability, Data-driven decision-making, Fraud detection

I. INTRODUCTION

The insurance sector has long been recognized as one of the most data-intensive industries in the financial services domain. Every interaction—whether through customer applications, premium payments, claims processing, or policy renewals—generates valuable information. In recent years, the explosion of digital channels, online applications, and customer relationship management systems has further amplified the volume, velocity, and variety of insurance data. This evolving data-rich environment provides unprecedented opportunities for insurers to make better-informed decisions, optimize their operations, and enhance customer experience. However, realizing these opportunities requires more than the traditional statistical tools once used for actuarial modeling or risk assessment. It calls for the strategic use of data mining—a set of techniques for uncovering meaningful patterns, hidden relationships, and predictive insights from complex datasets.

Early research in this area has shown that data mining can play a vital role in life insurance by supporting tasks such as customer acquisition, retention, product design, fraud detection, and underwriting. For instance, clustering techniques help segment customers into homogeneous groups, association rule mining can reveal policy bundling opportunities, and classification algorithms can link customer demographics with likely policy selections. These insights are directly tied to insurers' strategic goals: expanding the customer base, improving retention rates, and reducing losses through effective risk management. The research paper by Sharma and Panchal (2012) is an example of such work, where classical algorithms such as k-means, Apriori, and k-nearest neighbors were introduced with illustrative examples to demonstrate their application in insurance decision-making.

While such contributions are valuable for providing conceptual clarity and pedagogical demonstrations, they leave significant questions unanswered. Most existing research in this domain is either descriptive in nature, showcasing techniques through toy examples, or limited to high-level discussions of potential benefits. Rarely do such studies test methods on large-scale, real-world insurance datasets or evaluate their performance against business-relevant metrics. In addition, the focus on classical algorithms, while useful for foundational understanding, overlooks the evolution of modern machine learning and data mining methods that have become standard in other industries such as banking, retail, and healthcare. This creates an opportunity for new research that not only builds on earlier ideas but also fills these critical gaps.

One of the key limitations of prior work is its lack of attention to temporal dynamics in insurance data. Policies are not static contracts; they evolve over time through premium payments, renewals, customer service interactions, and potential lapses or surrenders. Understanding when a customer is likely to lapse or surrender a policy is as important as knowing whether they belong to a particular customer segment. Techniques such as survival analysis, time-to-event modeling, and recurrent neural networks have shown great promise in modeling similar longitudinal data in healthcare and subscription-based services. Their application to life insurance

represents a natural extension that has not yet been fully explored in the literature.

Another underexplored area is customer retention and uplift modeling. While association rule mining can highlight correlations between purchased products, it does not tell us which interventions—such as offering a premium discount, bundling an additional product, or improving service—will actually increase the likelihood of customer retention. Uplift modeling, which focuses on estimating the incremental effect of an action on an outcome, offers a more powerful approach. By modeling counterfactual outcomes, insurers can allocate resources more effectively, targeting interventions only to customers who are likely to respond positively.

Fraud detection in insurance is another critical challenge where data mining can play a transformative role. The paper reviewed mentions fraud as a potential application but does not provide concrete methodologies or evidence. In reality, fraudulent activity in insurance often involves collusion among multiple parties, including agents, providers, and claimants, which makes it difficult to detect using simple classification methods. Graph-based anomaly detection, network analysis, and semi-supervised learning have been shown to outperform traditional rule-based systems in fraud detection for other industries. Applying such methods to insurance claims can significantly reduce financial losses while improving efficiency in claims processing.

Equally important is the issue of interpretability, fairness, and governance in insurance analytics. Because insurance decisions—such as pricing, underwriting, and claims approval—directly affect individuals' financial well-being, regulatory bodies demand transparency and fairness in model outcomes. Yet much of the existing research has focused purely on accuracy and pattern discovery, neglecting the equally important question of whether models can be explained and trusted. Modern interpretability techniques, such as SHAP values and counterfactual explanations, provide opportunities to design models that are both accurate and transparent. Furthermore, fairness-aware algorithms can help prevent biases against protected groups, an increasingly pressing concern in regulated industries. These considerations are almost entirely absent from earlier data mining literature in insurance, highlighting a significant gap where academic research can add real-world value.

Operational challenges also remain largely unexplored. Implementing a data mining solution is not just about developing an accurate model; it also requires robust pipelines for data preparation, model monitoring, and retraining to handle concept drift. Insurance data is particularly prone to drift due to changes in regulatory frameworks, macroeconomic conditions, and customer preferences. Without continuous monitoring and recalibration, even the best-performing models can degrade over time, leading to suboptimal or unfair decisions. Addressing these issues calls for an integration of machine learning operations (MLOps) principles into the design of insurance analytics systems.

This research paper seeks to advance the field by building on the foundational work of earlier studies while explicitly addressing these gaps. We propose an integrated framework for applying data mining in life insurance that combines predictive, causal, and interpretable methods. Specifically, our study emphasizes (1) survival analysis for modeling policy lapse and surrender, (2) uplift modeling for targeted customer retention, (3) graph-based techniques for fraud detection, (4) explainable AI methods for interpretability and fairness, and (5) operational considerations such as drift detection and monitoring. Unlike earlier works that remain conceptual, our research will ground these methods in empirical experimentation on realistic datasets and evaluate them against business-relevant metrics such as retention lift, fraud detection accuracy, and customer lifetime value.

By moving beyond descriptive insights to actionable, validated, and ethically aligned decision support, this study aims to demonstrate how modern data mining techniques can transform life insurance practices. The ultimate goal is not only to showcase technical novelty but also to provide practical, scalable solutions that insurers can adopt in real-world settings. In doing so, this research will contribute both to academic knowledge and to industry practice, offering a more comprehensive roadmap for data-driven decision-making in life insurance.

II. LITERATURE REVIEW

This literature review synthesizes two primary pieces of material: (1) the uploaded 2012 IJDKP paper that surveys classical data-mining techniques for life insurance, and (2) the more recent InsurTech excerpt (citing Kasy, 2018; Kaur, Sharma & Mittal, 2018). Together, they help place the uploaded paper in the broader landscape of insurance analytics, surface thematic findings, and highlight methodological and applied gaps that motivate a student research project. The IJDKP paper is primarily pedagogical—it demonstrates how clustering, association rules, classification, and correlation can be mapped onto life insurance tasks such as customer acquisition, policy retention, product design, and risk assessment. While valuable as a conceptual bridge between data mining and insurance practice, it does not extend to empirical validation, real-world deployment, or advanced modeling approaches.

Building on that foundation, contemporary InsurTech literature illustrates how the industry has evolved to embrace AI and machine learning more fully. Insurance, one of the world's oldest industries with origins in early marine insurance, has always relied on data to guide decision-making. With today's explosion of customer data, researchers and practitioners highlight opportunities for sharper policy enrollment, risk selection, claims processing, fraud detection,

and marketing (Khan et al., 2014; Knighton et al., 2020; Kose et al., 2015; Kraus et al., 2020). Modern AI enables faster underwriting, policy personalization, automated mobile claims, advanced fraud detection, workforce insights, and data-driven marketing strategies (Larson & Sinclair, 2021; Maehashi & Shintani, 2020; Nian et al., 2016). Compared with the classical methods showcased in the IJDKP paper, these advances push the field toward strategies that deliver tangible, measurable benefits for insurers, thereby setting the stage for research that moves beyond descriptive analysis into predictive, causal, and operational solutions.

Modern AI/ML trends in insurance (themes from recent literature)

Recent literature and the user-supplied paragraph highlight key AI/ML trends transforming the insurance industry. AI accelerates underwriting and pricing by leveraging predictive models trained on rich datasets, enabling sharper risk selection and personalized policy design tailored to individual needs (Larson & Sinclair, 2021; Maehashi & Shintani, 2020; McGlade & Scott-Hayward, 2019; Mita et al., 2021). Claims processing is streamlined through mobile-first intake and automated triage, while ML-driven anomaly detection and graph-based methods enhance fraud detection, addressing its significant economic impact (Nian et al., 2016). Customer retention and marketing are evolving with ML tools like uplift modeling and causal inference, moving beyond descriptive analytics to measure intervention impacts and estimate customer lifetime value. Workforce analytics adapt similar behavioral signals to boost broker and employee performance (Ozbayoglu et al., 2020; Sengupta et al., 2020). Meanwhile, growing emphasis on interpretability, fairness, privacy (via federated learning and synthetic data), and production monitoring ensures AI systems are accountable and compliant with regulatory demands.

S No.	Author	Year	Application	Techniques used
1.	Bartl et al.	2020	Claim Prediction in Export Credit Finance	Decision Tree, Random Forest, Neural Networks (NN) & Probabilistic Neural Networks (PNN) for prediction and Accuracy, Cohen's K & R-square for assessment
2.	Baudry et al.	2019	Claim Reserving	Non-parametric ML model used for prediction, chain ladder method used for assessment w.r.t. bias and variance of estimates.
3.			Insurance Claim Analysis	Naïve Bayes, Naïve Bayes Updatable, Multi-Layer Perceptron, J48, Random Tree, LMT, Random Forest used for prediction with Recall, Precision, F-Measure, MCC, PRC & ROC Area for assessment
4.			Risk Prediction in Life Insurance	Univariate & Bivariate analysis for data visualization, PCA & Correlation-based feature selection for dimensional reduction and multiple linear regression, multilayer perceptron, REPTree & Random Tree for prediction
5.	Dehghanpour et al.	2018	Portfolio Insurance Strategy	Adaptive Neuro-Fuzzy Inference Systems (ANFIS) for prediction with the Markowitz portfolio optimization model for determining optimal portfolio weights
6.	Kang et al.	2018	Aggregate Auto-Insurance	Feature Selection techniques to

			Data Analysis	classify the dataset into homogenous risk groups
7	Panigrahi et al.	2018	Auto-Insurance Fraud Detection	Univariate, L1 based & Tree-based feature selection and Decision Tree, Naïve Bayes, KNN & Random Forest classification algorithms
8	Patil et al.	2018	Survey on ML techniques used for Fraud Detection	Supervised, Unsupervised and Hybrid-Bagging, Boosting, Stacking & Ensemble Learners
9	Quan et al.	2018	Predictive Analytics of Claims	Multivariate Decision Trees compared using R2, Gini, ME, MPE, MSE, MAE and MAPE
10	Wang et al.	2018	Auto-Insurance Fraud Detection	Deep Learning model with Latent Dirichlet Allocation (LDA) based text analytics
11		2018	Role of Data Mining in the Insurance Industry	Classification, Clustering, Regression, Association, Summarization
12	Rao et al.	2013	Factors influencing Claims in General Insurance, India	Regression analysis
13	Guelman	2012	Insurance Lost Cost Modeling	Gradient Boosting (GB) compared to the Linear Model approach
14	Salcedo-Sanz et al.	2004	Insolvency Prediction	Simulated Annealing (SA) and Walsh Analysis for feature selection using SVM as underlying classifier
15	Viaene et al.	2002	Insurance Claim Fraud Detection	Logistic Regression, C4.5 Decision Tree, K-Nearest Neighbor, Bayesian Multilayer Neural Network, Naïve Bayes and SVM for classification and PCC, AUROC and ROC curves for assessment

Table 1. Research work in the Insurance Industry.

2.1 RESEARCH GAP

Evidence Gap:One of the most significant shortcomings of the paper is the lack of empirical evidence to support its proposed applications. While it introduces popular data mining techniques such as clustering, association rules, classification, and correlation, the work relies entirely on illustrative examples instead of real-world insurance datasets. Without the use of authentic customer or policyholder records, there is no way to measure the accuracy, efficiency, or business impact of these methods in practice. Furthermore, the absence of baselines, benchmarks, and standardized evaluation protocols leaves the claims unverified and makes it difficult to compare the suggested approaches with existing industry practices. This evidence gap limits the credibility and practical applicability of the findings.

Temporal & Survival Modeling Gap:Insurance data is inherently temporal, with policies evolving over years and customer behaviors shifting across life stages. However, the paper does not explore methods that account for time-to-event processes such as policy lapse, surrender, or claim occurrence. Techniques like survival analysis, hazard models, or time-series forecasting could reveal critical insights into customer lifecycle risks, seasonal claim patterns, or cohort-specific trends. By ignoring the temporal dimension, the analysis treats customer behavior as static, missing the opportunity to predict

when key events will happen—a crucial capability for proactive risk management and customer retention in insurance.

Customer Lifetime Value (CLV) & Uplift Gap:Although the paper mentions customer acquisition and retention, it frames retention primarily through simple association rules (e.g., cross-sold products). This overlooks more advanced approaches that estimate customer lifetime value (CLV) or incremental uplift. CLV models allow insurers to prioritize high-value customers, while uplift modeling identifies which customers are most likely to change their behavior if given a targeted offer. Without these methods, the proposed strategies remain descriptive rather than prescriptive, failing to optimize for long-term profitability and personalized retention strategies.

Fraud Detection Gap:Fraud detection is highlighted as an area where data mining can add value, but the paper does not provide concrete methodologies or empirical support. Insurance fraud is complex, often involving hidden relationships between claimants, agents, and providers. Techniques like anomaly detection, graph-based analysis, and semi-supervised learning are well-suited for identifying suspicious patterns in such networks. By not addressing these approaches, the paper reduces fraud detection to a vague assertion rather than a demonstrable application, leaving a critical research and operational gap unfilled.

Causality & Experimentation Gap:The approaches discussed in the paper focus primarily on correlations and associations, but insurance companies also need to understand causality—what actually drives customer decisions. For example, offering a discount may correlate with higher renewals, but without causal analysis it is unclear whether the discount itself caused the retention. The absence of counterfactual reasoning, A/B testing frameworks, and causal inference methods means that the recommendations risk being misinterpreted and may lead to ineffective or even counterproductive strategies. This gap highlights the need for research that distinguishes between correlation and causation to guide actionable business interventions.

MLOps Gap:Even if the proposed techniques were applied, the paper does not address how these models would be maintained, scaled, or integrated into real-world operations. Insurance companies deal with dynamic environments where customer preferences, economic conditions, and regulatory requirements evolve over time. Without guidance on data pipelines, feature engineering, drift detection, or model monitoring, there is a risk that deployed models will quickly degrade in performance. Incorporating MLOps practices ensures that data mining systems remain reliable and trustworthy, a vital step missing in the paper's discussion.

Governance Gap:Insurance is a highly regulated industry where fairness, interpretability, and privacy are non-negotiable. However, the paper does not consider governance aspects such as explainable AI (e.g., SHAP values), fairness audits to detect discriminatory pricing, or privacy-preserving learning techniques like federated learning. Ignoring these dimensions raises ethical and legal risks, as opaque models could lead to biased decisions or privacy violations. Addressing governance concerns is crucial not only for regulatory compliance but also for building trust with customers and stakeholders, making this a significant research gap.

Data Design Gap:Finally, the paper simplifies the data requirements to a few demographic and product-related attributes, neglecting the richness of modern insurance datasets. In practice, data sources may include agent behavior, communication channels, call notes, payment histories, geographic indicators, and even household-level relationships across multiple policies. Incorporating such diverse features—structured, unstructured, and relational—can greatly enhance predictive power and actionable insights. The omission of realistic data design considerations leaves the proposed models underdeveloped and limits their real-world effectiveness.

2.2 CLAIM ANALYSIS IN INSURANCE SECTOR

Claim analysis plays a critical role in the insurance sector, where nearly 80% of premium revenue is spent on claims (Pappas & Woodside, 2021). Traditionally, statistics and actuarial science have formed the foundation of insurance risk assessment. However, with the exponential growth of data, predictive analytics—which combines data mining, predictive modeling, and machine learning techniques such as classification, regression, clustering, and outlier detection—has gained prominence (Pal et al., 2012; Yang et al., 2021; Palanisamy & Thirunavukarasu, 2019). By analyzing claim data, insurers can identify relationships among variables and build models to forecast outcomes, thereby improving decision-making (Petropoulos et al., 2020). Beyond structured datasets, unstructured data offers significant potential, providing insights that can help segment beneficiaries, estimate expected payouts, and optimize claim processing (Pramanik et al., 2020; Pourhabibi et al., 2020; Waring et al., 2020). Key performance indicators (KPIs) for claims management include cycle time, customer satisfaction, fraud detection, recovery, and handling costs (Richter & Khoshgoftaar, 2018). Despite the abundance of available data, insurers typically leverage only 10–15% of it (Ringshausen et al., 2021; Saggi & Jain, 2018). Applying machine learning can significantly expand this utilization, enabling automation of routine processes, reduction of claim cycle times, enhanced fraud detection, improved recovery, and greater customer satisfaction.

III. PROPOSED COMPUTATIONAL METHODOLOGY

In this study, two different insurance claim datasets are analyzed using supervised machine learning (ML) classification algorithms. Supervised learning is applied when the dataset consists of input variables (features) and a

corresponding output variable (target). Within supervised learning, classification algorithms are used when the target variable is categorical (e.g., claim approved vs. claim rejected), while regression algorithms are used when the target variable is continuous (e.g., predicting premium or claim amount). Since the target variable in both datasets is categorical, classification algorithms are most appropriate for this analysis.

The outcomes vary across the two datasets because of differences in their feature sets and data characteristics, which influence model performance. To ensure a systematic approach, the analysis follows a structured ML workflow based on Yufeng Guo's 7 Steps of Machine Learning. This framework guides the process from data collection and preparation to model training, evaluation, and deployment.

Beyond supervised learning, the insurance industry also benefits from unsupervised learning techniques. In these cases, there is no target variable, and algorithms such as clustering and association are used. Clustering helps identify groups of customers with similar attributes or detect patterns of attrition, while association rule mining uncovers hidden relationships that support cross-selling or product bundling strategies. Together, supervised and unsupervised methods provide complementary insights for insurance claim analysis and customer management.
Fig. 2 shows the framework used to perform claim analysis.

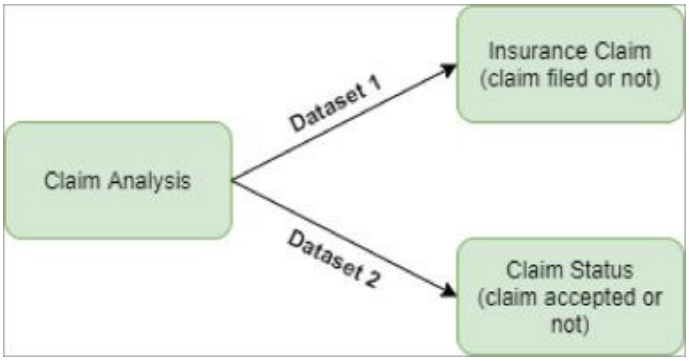

Fig. 2.1. Outcomes of claim analysis done in this analysis.

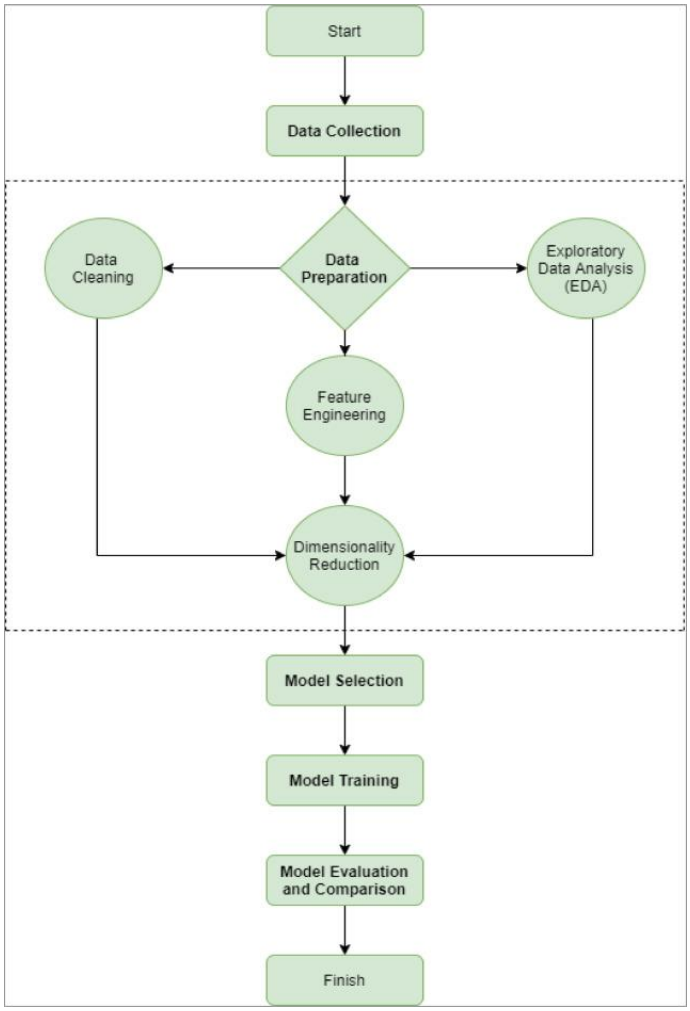

Fig. 2.2. ML Framework used for Claim Analysis.

3.1 DATA COLLECTION

The machine learning process begins with data collection, which can be achieved through various sources and techniques. One of the most common and effective methods is searching for and sharing data. This involves sourcing data from the web or accessing it from centralized repositories like a 'data lake,' with sharing often facilitated through online platforms. Another approach is data augmentation, where existing datasets are enriched with external data to enhance diversity, rather than gathering entirely new data. This technique is particularly valuable in deep learning for training large artificial neural networks (ANNs). Additionally, data can be obtained through crowdsourcing or by generating synthetic datasets to meet specific needs. The datasets used in this analysis were sourced from Kaggle.com and GitHub.com, with further details provided in the "Case Studies" section.

3.2 DATA PREPARATION

Getting data ready for machine learning is all about shaping raw information into something algorithms can work

with effectively. This step, often called data preparation, plays a huge role in how well a model performs. It involves a few key tasks: cleaning up the data by fixing errors, filling in missing pieces, or removing inconsistencies; digging into exploratory data analysis (EDA) to spot trends, patterns, or outliers that might influence the model; scaling data through normalization to ensure all values are on a similar range for fair comparison; and trimming down complexity with dimensionality reduction to focus on the most relevant features without losing critical insights. Each of these steps helps ensure the data is reliable and tailored to the algorithm's needs, ultimately boosting the model's accuracy and efficiency. Skipping or rushing this process can lead to skewed results, making thorough preparation a cornerstone of successful machine learning projects.

3.2.1 DATA CLEANING

Data cleaning, often referred to as data pre-processing, is the crucial first step after collecting data for machine learning. It focuses on identifying and correcting inaccurate, incomplete, corrupted, or irrelevant records to ensure a high-quality dataset. A widely used method is variable-by-variable cleaning, where each feature is examined individually to eliminate issues like illegal or misspelled values. This involves checking that values fall within acceptable ranges (e.g., no extreme outliers beyond minimum or maximum thresholds), ensuring variance and standard deviation stay within reasonable limits, and correcting any typographical errors in the data. Depending on the severity of the issue, problematic values are either removed or adjusted. When dealing with missing data, the approach depends on the extent of the gaps: features with many missing values may be dropped entirely, while those with fewer gaps might have missing values replaced with a placeholder (treating the absence as a distinct value) or imputed using techniques like mean substitution, which is a popular method for filling in missing data. Thorough data cleaning sets a solid foundation for reliable model performance by ensuring the dataset is accurate and consistent.

3.2.2 EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is a vital step in understanding a dataset before applying machine learning models. By using various visualization techniques, such as graphs and charts, EDA reveals key characteristics of the data, including hidden relationships and patterns among features that might not be apparent from simply reviewing raw data tables (e.g., Table 3). These visualizations help uncover trends, outliers, and correlations, providing critical insights that guide model development and improve overall analysis.

3.2.3 FEATURE ENGINEERING

Feature engineering is a critical phase in preparing data for model training and evaluation. During this stage, new features are crafted by leveraging insights from exploratory data analysis (EDA) and domain expertise to enhance model performance. This process is often one of the most challenging and time-intensive aspects of data preparation, with studies indicating that data scientists dedicate roughly 80% of their efforts to this phase. New features are derived through various calculations, such as ratios, mathematical transformations, or domain-specific statistical or scientific formulas, to create more impactful variables.

Feature engineering can be performed manually by statisticians or through automated techniques like feature encoding for categorical data. Contrary to the common belief that feature engineering is only useful for linear regression or text classification tasks, it has proven highly effective for a range of models, including support vector machines, random forests, neural networks, and gradient boosting algorithms. Encoding is essential because machine learning models rely on mathematical frameworks, which typically cannot directly handle distinctions between categorical and continuous data. Encoding techniques are divided into two main approaches: nominal and ordinal. Nominal encoding is used when the order of categories is unimportant, while ordinal encoding is applied when the order matters.

Beyond encoding, other feature engineering techniques, such as normalization, are commonly used. Normalization scales all values in a dataset to a standardized range, typically between 0 and 1, to ensure consistency and improve model performance.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

While normalization can enhance a model's numerical stability and performance, it's not universally beneficial and may negatively impact certain models if applied indiscriminately.

3.2.4 DIMENTIONAL REDUCTION

Feature engineering is a critical phase in preparing data for model training and evaluation. During this stage, new features are crafted by leveraging insights from exploratory data analysis (EDA) and domain expertise to enhance model performance. This process is often one of the most challenging and time-intensive aspects of data preparation, with studies indicating that data scientists dedicate roughly 80% of their efforts to this phase. New features are derived through various calculations, such as ratios, mathematical transformations, or domain-specific statistical or scientific formulas, to create more impactful variables.

Feature engineering can be performed manually by statisticians or through automated techniques like feature encoding for categorical data. Contrary to the common belief

that feature engineering is only useful for linear regression or text classification tasks, it has proven highly effective for a range of models, including support vector machines, random forests, neural networks, and gradient boosting algorithms. Encoding is essential because machine learning models rely on mathematical frameworks, which typically cannot directly handle distinctions between categorical and continuous data. Encoding techniques are divided into two main approaches: nominal and ordinal. Nominal encoding is used when the order of categories is unimportant, while ordinal encoding is applied when the order matters.

Beyond encoding, other feature engineering techniques, such as normalization, are commonly used. Normalization scales all values in a dataset to a standardized range, typically between 0 and 1, to ensure consistency and improve model performance.

3.2.4.1 FEATURE SELECTION

Feature engineering is a critical phase in preparing data for model training and evaluation. During this stage, new features are crafted by leveraging insights from exploratory data analysis (EDA) and domain expertise to enhance model performance. This process is often one of the most challenging and time-intensive aspects of data preparation, with studies indicating that data scientists dedicate roughly 80% of their efforts to this phase. New features are derived through various calculations, such as ratios, mathematical transformations, or domain-specific statistical or scientific formulas, to create more impactful variables.

Feature engineering can be performed manually by statisticians or through automated techniques like feature encoding for categorical data. Contrary to the common belief that feature engineering is only useful for linear regression or text classification tasks, it has proven highly effective for a range of models, including support vector machines, random forests, neural networks, and gradient boosting algorithms. Encoding is essential because machine learning models rely on mathematical frameworks, which typically cannot directly handle distinctions between categorical and continuous data. Encoding techniques are divided into two main approaches: nominal and ordinal. Nominal encoding is used when the order of categories is unimportant, while ordinal encoding is applied when the order matters.

Beyond encoding, other feature engineering techniques, such as normalization, are commonly used. Normalization scales all values in a dataset to a standardized range, typically between 0 and 1, to ensure consistency and improve model performance.

3.2.4.1.1 Chi-Square Test

This refers to a statistical filter method that assesses the correlation between features by analyzing their frequency distributions. In this approach, feature selection relies on the inherent characteristics of the features themselves and is independent of any machine learning algorithm.

3.2.4.1.2 Recursive Feature Elimination (RFE)

This describes a wrapper method for feature selection known as Recursive Feature Elimination (RFE). The term "wrapper" reflects how this method integrates a classifier within the feature selection process. In RFE, features are iteratively eliminated from the dataset based on weights assigned by an external estimator, typically the classifier, which evaluates feature importance based on performance. As a greedy algorithm, RFE aims to identify the optimal subset of features that maximizes model performance.

3.2.4.1.3 Tree-Based Feature Selection

This refers to an embedded method for feature selection, where an algorithm's built-in mechanism evaluates and ranks features based on their importance. Embedded methods leverage algorithms with inherent feature selection capabilities to generate a prioritized set of features during the modeling process.

3.3 MODEL SELECTION

Once data preparation is complete and the dataset is split into training and testing sets, the next step is to select appropriate models for training. Given that this is a classification task, suitable classifiers must be chosen. Both datasets in this analysis involve binary classification. The classifiers employed in this study include Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Mixed Naïve Bayes, and K-Nearest Neighbors.

3.4 MODEL TRAINING

After selecting the models, the training data is first used to train the chosen models with all available features. Subsequently, the classification algorithms are applied exclusively to the features identified through feature selection techniques.

3.5 MODEL EVALUTION AND CONPARISON

To determine the most effective model and feature selection technique for the two datasets, a comparative analysis is conducted between the models and the feature selection methods applied. Model performance is evaluated using four equally weighted metrics: Precision, Recall, F1 Score, and Accuracy, ensuring a balanced assessment without prioritizing any single metric.

IV. EXPERIMENTATION RESULTS

The analysis comprises two case studies: one focused on the health insurance sector and the other on the travel insurance sector. As illustrated in Fig. 1, the outcomes of these case studies differ, with the first reflecting the beneficiary's perspective and the second representing the insurance company's perspective.

4.1CASE STUDY 1: HEALTH INSURANCE

For this case study, the analysis is built on a dataset from Kaggle.com. It includes 1,338 entries, each with nine associated attributes. Eight of these are input features used for prediction, and the ninth is the output variable we aim to predict. Table 2 outlines the specifics of each variable.

S. No.	Column Heading	Description
1.	Age	Age of the beneficiary
2.	Sex	Gender of the beneficiary (female = 0, male = 1)
3.	BMI	Body Mass Index of the beneficiary, i.e. the ratio of weight to height (kg / m^2), ideally 18.5 to 25
4.	Steps	Average walking steps per day of the beneficiary
5.	Children	Number of children or dependents of the beneficiary
6.	Smoker	Smoking status of the beneficiary (non-smoker = 0, smoker = 1)
7.	Region	The place of residence of the beneficiary in the US (northeast = 0, northwest = 1, southeast = 2, southwest = 3)
8.	Charges	Individual medical costs billed by health insurance
9.	Insurance Claim	Whether the beneficiary files a claim or not (Yes = 1, No = 0)

Table 2. Description of Health Insurance Dataset.

Statistics	Age	BMI	Steps	Children	Charges
Min	18	15.96	3000	0	1121.87
Max	64	53.13	10010	5	63770.42
Mean	39.2	30.66	5328.62	1	13270.42

Table 3. Statistics of features of the Health Insurance Dataset.

The initial step in the data preparation process involves checking the dataset for missing values. In this case, the dataset is complete, with no missing values detected. Following this, various statistical measures of the features are examined to gain a comprehensive understanding of the data's characteristics and distributions.

To facilitate exploratory data analysis (EDA), one-hot encoding is applied to the 'region' feature. This process transforms the categorical 'region' variable into four new binary features: 'NorthEast USA', 'NorthWest USA', 'SouthEast USA', and 'SouthWest USA', while the original 'region' feature is removed from the dataset. This encoding enables a clearer understanding of the distribution of policyholders across different regions in relation to other features.

Analysis of the data, as depicted in Fig. 3, reveals distinct patterns in claim behavior based on BMI and age. Policyholders with a BMI between 14 and 24 are the least likely to file a claim, while those with a BMI between 24 and 29 show a neutral tendency. In contrast, policyholders with a BMI exceeding 29 are the most likely to file claims. Additionally, the age group of 29–39 contains the highest number of policyholders who have not filed a claim.

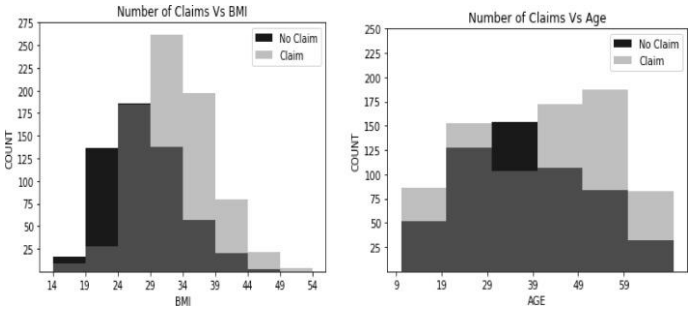

Fig. 3. Graphical Representation of relationship b/w Number of Claims and the BMI of the beneficiary & Number of Claims and the Age of the beneficiary.

Based on Fig. 4, it is evident that most smokers tend to file claims, whereas non-smokers exhibit a neutral stance toward filing claims. The policyholder's sex has no significant impact on claim behavior, as both male and female smokers are highly likely to file a claim.

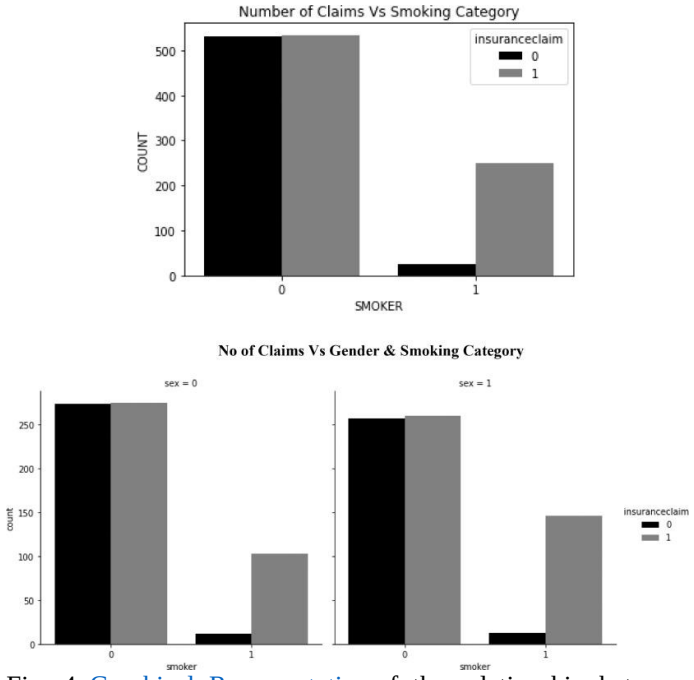

Fig. 4. Graphical Representation of the relationship between Number of Claims and the smoking category of the policyholders, Number of Claims and the smoking category when the policyholders are female & Number of Claims and the smoking category when the policyholders are male.

Analysis of Fig. 5 reveals distinct patterns in claim behavior among policyholders. Those with charges below $9,999 are the least likely to file a claim, while those with charges exceeding $39,999 are highly likely to file one. Additionally, policyholders averaging 3,000–6,000 steps per day file claims most frequently, whereas those exceeding 6,000 steps per day are the least likely to file. Policyholders with no children also tend to file claims the most.

Regarding regional differences, policyholders in NorthWest USA display a neutral tendency toward filing claims. In contrast, policyholders in SouthEast USA are more likely to file claims compared to other regions.

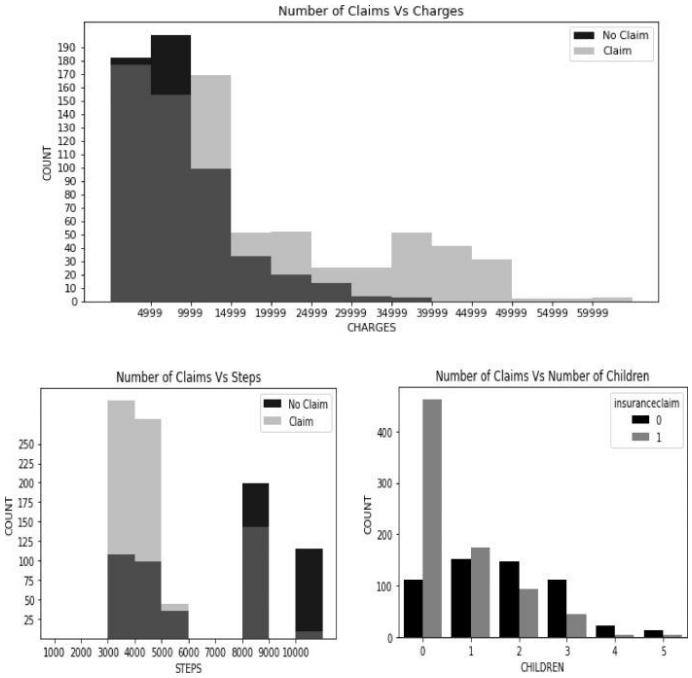

Fig 5. Graphical representation of the relationship between Number of Claims by the policyholder and the Charges billed by the health insurance, Number of Claims and the average walking steps of the policyholder per day & Number of Claims and the number of children of the policyholder.

Data Preparation:The dataset is prepared for modeling by selecting a final set of features: 'age', 'sex', 'BMI', 'steps', 'children', 'smoker', 'NorthEast USA', 'NorthWest USA', 'SouthEast USA', 'SouthWest USA', 'charges', and the target variable 'insurance claim'. Initial evaluation of classifier performance, as shown in Table 6 and Fig. 8, indicates that Logistic Regression, Random Forest, and Decision Tree Classifiers perform best. Among these, the Decision Tree Classifier outperforms the others, achieving higher values in three out of four performance metrics, with no single metric prioritized over others. To further investigate the impact of feature selection, the dataset is retrained using eight classifiers, applying various feature selection techniques to assess changes in performance and identify the most effective method.

Impact:Further analysis from Table 7 and Fig. 8 confirms that Logistic Regression, Random Forest, and Decision Tree Classifiers remain top performers after applying feature selection. Random Forest emerges as the best, with all four performance metrics surpassing those of other classifiers. Using the Chi-Squared Test, four features are eliminated, resulting in a reduced dataset with seven features: 'age', 'BMI', 'steps', 'children', 'smoker', 'NorthWest USA', and 'SouthEast USA'. Notably, the SVM Classifier's performance declines, while Gaussian and Mixed NB Classifiers show improvement, and the KNN Classifier's performance improves significantly. Subsequent evaluation in Table 8 and Fig. 8 highlights Logistic Regression, Random Forest, Decision Tree, Gaussian NB, and Mixed NB as top performers, with Random

Forest again leading. Logistic Regression, Gaussian NB, and Mixed NB exhibit equivalent performance. Compared to the Chi-Squared Test results, SVM, Gaussian NB, Bernoulli NB, and Mixed NB show improved performance, while others decline. The Chi-Squared Test, being a statistical method independent of machine learning models, identifies seven features as optimal, which is used as a benchmark for comparing feature selection techniques. Using Recursive Feature Elimination (RFE), the seven most important features are identified as 'NorthEast USA', 'age', 'BMI', 'charges', 'children', 'smoker', and 'steps'. In contrast, the Tree-Based Feature Importance Method selects 'children', 'steps', 'BMI', 'charges', 'age', 'smoker', and 'sex' as the top seven features, with Decision Tree outperforming others, though most classifiers show reduced performance compared to RFE, except for Decision Tree.

Conclusion:By comparing the performance of the eight classifiers with and without feature selection, it is evident that the Decision Tree Classifier performs best without feature selection, while Random Forest excels with feature selection. RFE proves to be the most effective feature selection method, as most classifiers achieve their best performance with it, except for KNN, which performs best with Chi-Squared Test features, and Decision Tree, which excels with the Tree-Based method. Logistic Regression performs best without feature selection, while Gaussian NB and Mixed NB consistently yield identical results, suggesting that continuous variables are more critical than categorical ones for this dataset. Consequently, the optimal feature set is determined to be 'NorthEast USA', 'age', 'BMI', 'charges', 'children', 'smoker', and 'steps'.

4.2 CASE STUDY 2: TRAVEL INSURANCE

The dataset for this case study, sourced from Kaggle.com, comprises 62,288 rows and 12 columns, including 11 features and 1 target variable. Table 4 provides a detailed description of these features and the target variable to enhance understanding of the dataset's structure and content.

S.No.	Column Heading	Description
1.	Agency	Name of the insurance agency
2.	Agency Type	Type of agency: travel or airlines
3.	Distribution Channel	Distribution channel of the insurance agency: online or offline
4.	Product Name	Name of the insurance policies (products)
5.	Duration	Duration of travel of the policyholder
6.	Destination	Destination of travel
7.	Net Sales	The total amount of sales of the insurance policies
8..	Commission (in value)	The commission received to the mediator (agent)
9.	Gender	Gender of the policyholder
10.	Age	Age of the policyholder
11.	ID	ID of the policyholder
12.	Claim	Claim status of the insurance policy: accepted or denied

Once the data is collected, the next step is data preparation. First, the data is checked for any missing values. There are 39,575 missing values for Gender i.e. 63.54 %. Hence, the column is dropped from the dataset before proceeding for further evaluation. Also, the ID column is dropped as it has no significance in predicting claim acceptance. Next, different statistics of the features are observed.

According to the statistics observed as in Table 5, the maximum age is 118 which is not a suitable age to travel for anybody, also most of the insurance companies do not provide insurance to people above 85 years of age, hence considering 100–118 as an outlier category which comprises of 1.44% of the total policyholders, it is replaced by 99 hence keeping 99 as the maximum age of a policyholder. The minimum negative value of Net Sales is justified as net sales are calculated as the difference between the value for which the insurance was sold and the expenses incurred, or the claim amount paid by the insurance company to the policyholder/beneficiary. So the net sale may be negative if the claim amount paid or even if the claim amount is not paid it can be negative when the claim is rejected and the expenses incurred for doing the investigation is more than the actual policy amount paid by the policyholder. The minimum value of duration is -2 is which not possible under any circumstances and the maximum value is 4881. Even if the unit for the duration is considered to be in days then also this value is not possible as travel insurance policies can be applied for a maximum duration of 1–2 years in the case of an Annual Plan. Considering a maximum duration of 731 days i.e. one year and one leap year, the values above 731 and below 1 are imputed as they constitute only 0.026% of the whole dataset. Values above 731 are imputed as 731 and values below 1 are imputed by the median value of duration.

Statistics	**Age**	**Commission**	**Net Sales**	**Duration**
Min	0	0	-389	-2
Max	118	262.76	682	4881
Mean	39.67	12.83	50.71	60.96

Table 5. Statistics of the features of the Travel Insurance Dataset.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.8137255	0.96078	0.95098	0.52941	0.80392	0.73529	0.80392	0.5490196
Recall	0.8556701	0.95146	0.9898	0.7013	0.73874	0.68182	0.73874	0.6021505
F1 Score	0.8341709	0.9561	0.97	0.60335	0.76995	0.70755	0.76995	0.574359
Accuracy	0.8768657	0.96642	0.97761	0.73507	0.81716	0.76866	0.81716	0.6902985

Table 6. Classification of Health Insurance dataset without Feature Selection.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.773913	0.99107	0.97321	0.5625	0.875	0.76786	0.875	0.723214
Recall	0.872549	0.97368	0.9646	0.62376	0.75385	0.66667	0.75385	0.771429
F1 Score	0.820276	0.9823	0.96889	0.59155	0.80992	0.71369	0.80992	0.746544
Accuracy	0.854478	0.98507	0.97388	0.67537	0.82836	0.74254	0.82836	0.794776

Table 7. Classification of Health Insurance dataset with Chi-Squared Test.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.7714286	0.99048	0.96639	0.54622	0.80672	0.79832	0.80672	0.5714286
Recall	0.8350515	0.97196	0.95833	0.80247	0.83478	0.73077	0.83478	0.7234043
F1 Score	0.8019802	0.98113	0.96234	0.65	0.82051	0.76305	0.82051	0.6384977
Accuracy	0.8507463	0.98507	0.96642	0.73881	0.84328	0.77985	0.84328	0.7126866

Table 8. Classification of Health Insurance dataset with Recursive Feature Elimination.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.754237	0.9322	0.97458	0.55932	0.80508	0.69492	0.80508	0.542373
Recall	0.89899	0.99099	1	0.74157	0.79832	0.76636	0.79832	0.752941
F1 Score	0.820276	0.9607	0.98712	0.63768	0.80169	0.72889	0.80169	0.630542
Accuracy	0.854478	0.96642	0.98881	0.72015	0.82463	0.77239	0.82463	0.720149

Table 9. Classification of Health Insurance dataset with Tree Based Feature Importance.

By observing Fig. 6, it can be inferred that Net Sales and Acceptance % are directly proportional to each other and only 3 agencies have a good amount of sales. Also, Agency 'C2B' despite having high net sales and acceptance % provide a low commission to the mediator or maybe most of their sales are direct and there is no mediator in between.

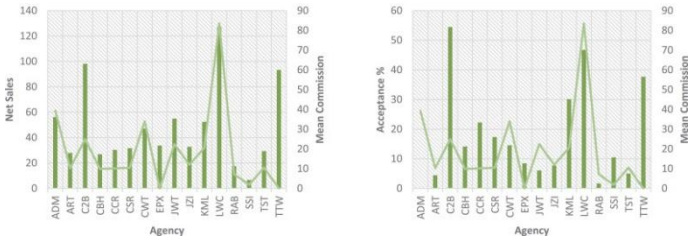

Fig. 6. Graphical representation of the relationship between Agency and Net Sales of policies & Agency and Acceptance % of claims along with the mean commission.

From Fig. 7, it can be inferred that products having high Commission Value have high Net Sales as well as high Acceptance %. Apart from these findings, it is also observed that most of the Travel Agency claims are denied and despite having a low count of claims under Airlines Agency, around 40% of the claims are accepted. Before moving onto the next comparison, the age is divided into three groups: Child (less than 21 years old), Adult (21–50 years old) and Senior (above 50 years old).

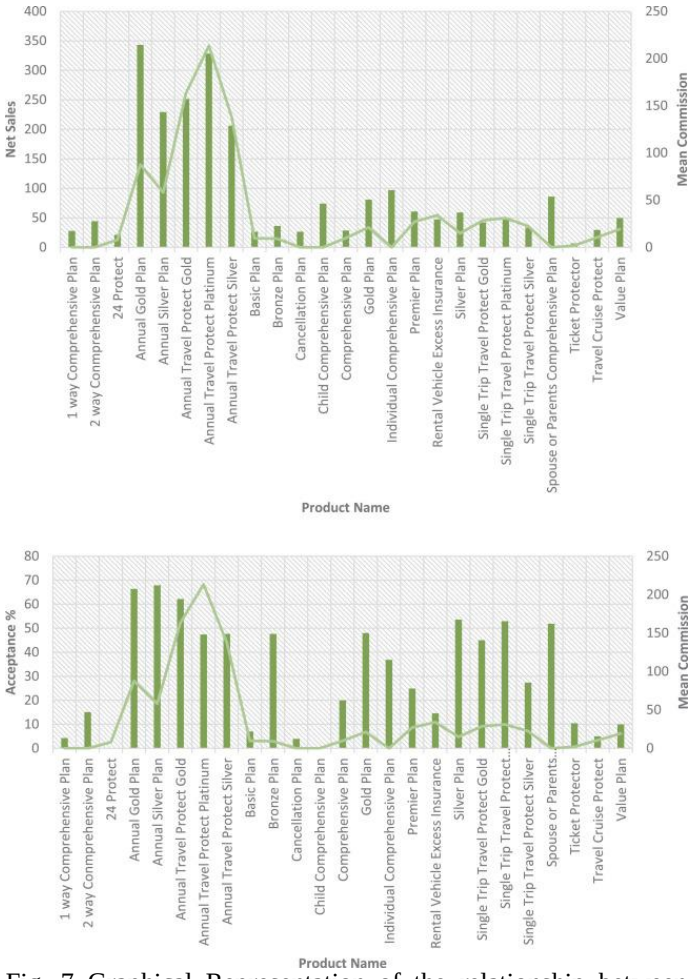

Fig. 7. Graphical Representation of the relationship between Product Name and Net Sales along with Mean Commission &

between Product Name and Acceptance % along with mean commission.

Fig. 8. Graphical Representation of the performance of the aforementioned classifiers without feature selection, with Chi-Squared Test, with RFE (using LogisticRegression Classifier) and with Tree-Based Feature Importance (using ExtraTreesClassifier).

The next step after EDA is Feature Engineering. In this stage, label, dummy, and frequency encoding are performed to deal with the categorical columns. Frequency Encoding is done for Destination, Agency, and Product Name columns. Dummy Encoding is done for Agency Type and Distribution Channel columns & Label Encoding is performed for Destination Category. Destination Category is decided based on Destination Risk. Destination Risk is calculated based on the count of claims. If the value is more than 0.3 i.e. more than 30% of the policyholders (travellers) have claimed the destination is marked as 'High Risk'. Similarly, if the value of risk is between 0.2 and 0.3, then the destination is marked as 'Moderate Risk' and if the value of risk is between 0 and 0.2, then the destination is marked as 'Low Risk'.

The final set of features selected before proceeding to feature selection are 'Age', 'Commission (in value)', 'Duration', 'Net Sales', 'Dest_freq_encoding', 'Agency_freq_encoding', Product_Name_freq_encoding', 'Destination Category (labels)', 'Agency Type_Travel Agency', 'Distribution Channel_Online' and the target variable 'Claim'.

After feature engineering, the next step is to do dimensionality reduction. In both the datasets, as previously discussed there is no need for feature extraction. In this section, first of all, both the datasets are trained and evaluated using the aforementioned eight classifiers without performing feature selection. Then feature selection is performed using three different methods, namely filter, wrapper and embedded methods & then the models are evaluated using four performance metrics. The feature selection techniques used under these methods in this analysis are the Chi-Squared Test, Recursive Feature Elimination using Logistic Regression Classifier & Tree-Based Feature Importance using ExtraTrees Classifier.

From Table 10 and Fig. 9, it can be inferred that Random Forest, Decision Tree and KNN show the best performance. Random Forest is best among all the classifiers as all four of its performance metrics have a higher value than the rest of the classifiers.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.91919	0.98727	0.97604	0.97323	0.83818	0.86154	0.83818	0.939743
Recall	0.799163	1	1	0.83515	0.90457	0.8981	0.90457	0.999147
F1 Score	0.854985	0.99359	0.98787	0.89892	0.87011	0.87944	0.87011	0.968535
Accuracy	0.750361	0.98981	0.98082	0.82477	0.79965	0.81088	0.79965	0.951116

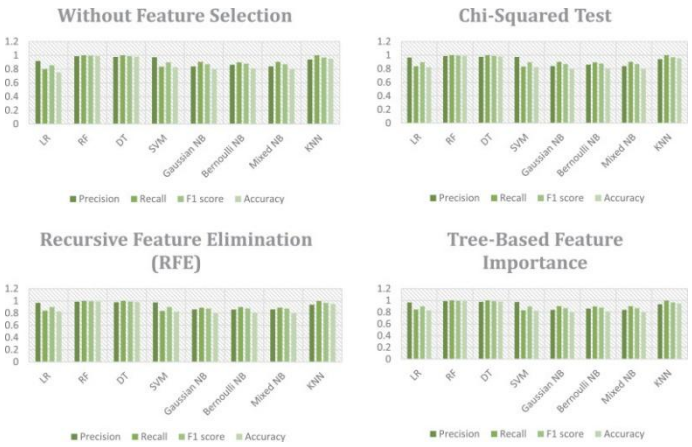

Fig. 9. Graphical Representation of the performance of the aforementioned classifiers without Feature Selection, with Chi-Squared Test, with RFE (using LogisticRegression Classifier), with Tree-Based Feature Importance (using ExtraTreesClassifier).

Now, once again the dataset is trained using the same eight classifiers but with feature selection techniques to observe the changes in the result as well as evaluate the best feature selection technique for all the classifiers.

From Table 11 and Fig. 9, it can be inferred that Random Forest, Decision Tree and KNN classifiers show the best performance. Random Forest is once again the best among all the classifiers. Using Chi-Squared Test, only one feature is discarded leaving the dataset with nine features: 'Age', 'Commission (in value)', 'Duration', 'Net Sales', 'Dest_freq_encoding', 'Agency_freq_encoding', 'Product_Name_freq_encoding', 'Destination Category (labels)' and 'Agency Type_Travel Agency'. The performance of Logistic Regression, Decision Tree and KNN classifiers has increased and the performance of the remaining classifiers has decreased as compared to modelling without feature selection.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.965773	0.98706	0.97662	0.97461	0.83895	0.86033	0.83895	0.942103
Recall	0.837788	1	1	0.83266	0.90204	0.89452	0.90204	1
F1 Score	0.89724	0.99349	0.98817	0.89806	0.86935	0.87709	0.86935	0.970189
Accuracy	0.823086	0.98965	0.9813	0.82301	0.79828	0.80711	0.79828	0.953684

Table 11. Classification of Travel Insurance dataset with Chi-Squared Test.

From Table 12 and Fig. 9, it can be inferred that Random Forest, Decision Tree and KNN classifiers show the best performance. Random Forest is the best among all the classifiers used for the dataset. Except for Decision Tree and KNN, the rest all the classifiers have increased performance as compared to the results of the Chi-Squared Test. Also, the performance of the Bernoulli NB classifier has decreased as compared to its performance without any feature selection technique. As Chi-Squared Test is a statistical test and has no involvement of any ML model, the number of features considered important by this test i.e. nine is considered as the number of comparing all the feature selection techniques. The nine most important features identified by RFE are: 'Age', 'Agency Type_Travel Agency', 'Commission (in value)', 'Dest_freq_encoding', 'Destination Category (labels)', 'Distribution Channel_Online', 'Duration', 'Net Sales' and 'Product_Name_freq_encoding'.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.966322	0.98751	0.97601	0.97401	0.85927	0.85767	0.85927	0.938331
Recall	0.839476	1	1	0.83533	0.88913	0.89825	0.88913	0.999574
F1 Score	0.898444	0.99371	0.98786	0.89936	0.87395	0.87749	0.87395	0.967985
Accuracy	0.825574	0.98997	0.98074	0.82493	0.80093	0.80767	0.80093	0.950153

Table 12. Classification of Travel Insurance dataset with Recursive Feature Elimination.

From Table 13 and Fig. 9, it can be inferred that Random Forest, Decision Tree and KNN classifiers show the best performance. Random is once again the best among all the classifiers. The performance of Logistic Regression, Random Forest, Decision Tree and Bernoulli NB classifiers has increased, rest four classifiers' performance has decreased compared to RFE. The nine best features selected by the Tree-Based Feature Importance method are: 'Duration', 'Age', 'Net Sales', 'Dest_freq_encoding', 'Commission (in value)', 'Agency_freq_encoding', 'Destination Category (labels)', 'Product_Name_freq_encoding' and 'Agency Type_Travel Agency'. These are the same features as selected by Chi-Squared Test. The performance of the models using both feature selection methods is also almost the same.

Metrics Used	Classification Algorithms Used							
	Logistic Regression	Random Forest	Decision Tree	Support Vector Machine	Gaussian Naïve Bayes	Bernoulli Naïve Bayes	Mixed Naïve Bayes	K-Nearest Neighbors
Precision	0.965455	0.98794	0.97659	0.97408	0.83886	0.86277	0.83886	0.936095
Recall	0.844541	1	1	0.83306	0.90348	0.89626	0.90348	0.999569
F1 Score	0.90096	0.99394	0.98816	0.89807	0.86997	0.8792	0.86997	0.966792
Accuracy	0.829347	0.99037	0.9813	0.82333	0.79965	0.81056	0.79965	0.948788

Table 13. Classification of Travel Insurance dataset with Tree Based Feature Importance.

By observing the performance of all the eight classifiers with and without feature selection it can be concluded that Random Forest is the best classifier both with and without feature selection. Chi-Squared and Tree-Based Feature Importance methods are the best feature selection techniques for the dataset as four models perform their best with features selected from these two. The slight difference between the results of these two techniques is neglected as both have selected the same set of features and any ML model trains itself continuously & both the techniques are not applied parallelly. Gaussian NB and Mixed NB produce the same results in all the cases; hence it can be concluded that continuous variables are more important than categorical variables for the dataset. Bernoulli NB classifier performs best without any feature selection. Therefore, the best set of features for the dataset is: 'Duration', 'Age', 'Net Sales', 'Dest_freq_encoding', 'Commission (in value)', 'Agency_freq_encoding', 'Destination Category (labels)', 'Product_Name_freq_encoding' and 'Agency Type_Travel Agency'.

V. CONCLUSION

The blend of traditional data mining and modern InsurTech innovations highlights both the potential and hurdles of turning insurance into a fully data-driven field. A 2012 IJDKP study showed how classic methods like clustering, association rules, classification, and correlation could aid tasks such as attracting customers, keeping them engaged, and designing products. However, these approaches, while insightful, were often more academic than practical, lacking real-world testing. Today's InsurTech research showcases how AI and machine learning have expanded possibilities, speeding up risk assessments, tailoring policies, automating claims through mobile platforms, and improving fraud detection. Together, these efforts trace the shift in insurance analytics from basic descriptions to predictive, cause-focused, and practical tools that deliver clear business benefits.

Yet, challenges persist. A 2016 PwC report noted that only 28% of major insurers collaborate with InsurTech startups, and just 14% engage with incubators or venture partnerships. This disconnect hampers the growth of new InsurTech firms, which often focus heavily on policy sales rather than risk evaluation, claims handling, or regulatory

compliance. Without stronger ties and smarter resource allocation, innovation risks staying isolated rather than scaling up. Additionally, highly imbalanced data—where rare events like fraud or policy dropouts are underrepresented—makes modeling tricky. Solutions like advanced resampling, clustering dominant groups, or using techniques such as SMOTE, XGBoost, or AdaBoost can improve predictions on these uneven datasets.

Looking ahead, the focus must shift from theoretical examples to practical, evidence-based, and ethical solutions. Promising paths include uplift modeling and causal inference to measure the real impact of retention efforts, cross-selling, or policy tweaks. Survival analysis and time-to-event modeling can predict when policies might lapse or claims arise, adding a time-based perspective missing in earlier methods. Graph-based and semi-supervised approaches strengthen fraud detection by spotting unusual patterns across customers, agents, and providers that older rule-based systems might overlook. Optimizing customer lifetime value (CLV) also shifts retention from short-term wins to long-term profitability, aligning marketing, risk assessment, and service efforts.

Operational and ethical factors are just as critical. Insurers need robust MLOps systems for data quality, performance monitoring, and model updates to ensure lasting results. With regulatory scrutiny growing, tools like SHAP values help explain decisions, ensure fairness, and build trust with customers and regulators. Ethical concerns around bias, privacy, and fair access call for innovations like federated learning and fairness-focused algorithms. Future efforts should also tap into diverse data sources—agent notes, call transcripts, or digital behaviors—to gain deeper insights into customers and employees.

In short, while the 2012 IJDKP study laid a solid groundwork for data mining in insurance, and InsurTech shows the power of AI-driven tools, there's still work to do to connect theory with practice. Future efforts should prioritize tested, time-sensitive, cause-driven, and transparent models that are practical and ethical. Tackling structural issues—like limited partnerships, imbalanced data, and resource constraints—will be key to ensuring these advancements are sustainable and scalable. By pursuing these paths, researchers and industry professionals can help transform insurance into a predictive, adaptable, and customer-focused field that delivers real value and resilience.

REFERENCES

[1] Alex Berson and Stephen J. Smith, "Data Warehousing, Data Mining, And OLAP",MC Graow–Hill,1997.

[2] Bigus and Joseph P, "Data Mining With Neural Networks", MC Graw–Hill, New York 1996.

[3] Christopher J. Matheus, Gregory Piatetshy–Shapiro and Dwight Mcneill", Selecting and Reportingwhat is Interesting The Kefir Application to Health Care Data", Advances in Knowledge Discoveryand Data Mining, AAA1 Press/The MIT Press, 1996.

[4] Dasrathy B. V., Ed, "Nearest Neighbor Norms: NN Pattern Classification Techniques",IEEE,Computer Society Press, Calif. 1990.

[5] David Cheung, Vincent T., Ada W. Fu and Yongjian Fv, "Efficient Mining of Association Rules inDistributed Databases", IEEE, 1996.

[6] Seema Rawat, Aakankshu Rawat, Deepak Kumar, A. Sai Sabitha, Application of machine learning and data visualization techniques for decision support in the insurance sector, doi.org/10.1016/j.jjimei.2021.100012

[7] Aswani et al., 2020 - R. Aswani, S.P. Ghrera, S. Chandra, A.K. Kar, A hybrid evolutionary approach for identifying spam websites for search engine marketing, doi.org/10.1007/s12065-020-00461-1

[8] Bacry et al., 2020 - E. Bacry, S. Gaïffas, F. Leroy, M. Morel, D.P. Nguyen, Y. Sebiat, D. Sun, SCALPEL3: A scalable open-source library for healthcare claims databases, doi.org/10.1016/j.ijmedinf.2020.104203

[9] Waring et al., 2020 - J. Waring, C. Lindvall, R. Umeton, Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, doi.org/10.1016/j.artmed.2020.101822

[10] Ringshausen et al., 2021 - F.C. Ringshausen, R. Ewen, J. Multmeier, B. Monga, M. Obradovic, R. van der Laan, R. Diel, Predictive modeling of nontuberculous mycobacterial pulmonary disease epidemiology using German health claims data, doi.org/10.1016/j.ijid.2021.01.003

[11] Pramanik et al., 2020 - M.I. Pramanik, R.Y.K. Lau, M.A.K. Azad, M.S. Hossain, M.K.H. Chowdhury, B.K Karmaker, Healthcare informatics and analytics in big data, doi.org/10.1016/j.eswa.2020.113388

[12] R. Burbidge et al. - Drug design by machine learning: Support vector machines for pharmaceutical data analysis, Computers & Chemistry(2001)

[13] Palanisamy and Thirunavukarasu, 2019 - V. Palanisamy, R. Thirunavukarasu, Implications of big data analytics in developing healthcare frameworks – a review, doc.org/10.1016/j.jksuci.2017.12.007

[14] Ozbayoglu et al., 2020 - A.M. Ozbayoglu, M.U. Gudelek, O.B. Sezer, Deep learning for financial applications: A survey, doc.org/10.1016/j.asoc.2020.106384

[15] McGlade and Scott-Hayward, 2019 - D. McGlade, S. Scott-Hayward, ML-based cyber incident detection for electronic medical record (EMR) systems, doc.org/10.1016/j.smhl.2018.05.001

[16] Mita et al., 2021 - Yosuke Mita, Ryo Inose, Ryota Goto, Yoshiki Kusama, Ryuji Koizumi, Daisuke Yamasaki, Masahiro Ishikane, Masaki Tanabe, Norio Ohmagari, Yuichi Muraki, An alternative index for evaluating AMU and anti-methicillin-resistant Staphylococcus aureus agent use: A study based on the National Database of Health Insurance Claims and Specific Health Checkups data of Japan, doc.org/10.1016/j.jiac.2021.02.009