2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) | 979-8-3503-6841-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSSAS64001.2024.10760834

Predicting Sham Claims in Vehicle Insurance using Machine Learning Techniques

Hema Ambiha. A
Department of Computer Science
Karpagam Academy of Higher Education
Coimbatore, India
hemaambiha.aravindakshan@kahedu.edu.in
https://orcid.org/0009-0000-7495-5107

V. Subithra
Department of Computer Applications
Karpagam Academy of Higher Education
Coimbatore, India
subithravaratharaj123@gmail.com

***Abstract*—The insurance industry is experiencing rapid growth and as a result, faces the challenge of managing vast amounts of data. One of the foremost concerns within this sector is the issue of fraudulent claims, which refers to deceptive attempts made for personal financial gain. This deceitful practice affects both insurance companies and their honest policyholders due to the prevalence of such claims. Traditionally, domain expertise and conventional methods have been employed to detect fraudulent activities in insurance claims. However, more recently, data mining techniques have emerged as valuable tools in the field of insurance analysis. These techniques enable the calculation of insurance premium amounts and the prediction of both fraudulent claims and accident occurrences. The premium amount is determined based on individual customers' confidential and financial information. Implementing this research significantly reduces human effort and financial losses in the examination of automobile insurance claims. Effective classifiers play a crucial role in determining whether a claim is fraudulent or not. Predicting the occurrence of accidents is a fundamental step in identifying fraudulent claims. Decision tree classifiers, such as Simple CART, SVM, and LMT, available in the WEKA tool, are employed to analyze accidents and claims. The execution of these classifiers is compared in terms of accuracy, with Random Forest demonstrating superior accuracy in predicting accident occurrences and yielding the best results among the classifiers. Furthermore, the SVM algorithm outperforms other classifiers in predicting fraudulent claims.**

Keywords—Fradulent claims, Insurance, Classifiers,CART, SVM, LMT

I. Introduction

The insurance database is vast, and in practice and theory, combating fraudulent claims is a formidable task. Typically, fewer than 3% of automobile insurance claims officially precede with suspicions of fraud, but elements of suspicion are found in a significant portion, ranging from 21% to 36% of such claims. When fraud goes unnoticed, insurance companies are forced to increase premium amounts to compensate for the losses, which can lead to a decline in their competitiveness in the insurance industry. This, in turn, burdens all insured clients with higher premium payments. Moreover, the time required to pay legitimate insurance claim is extended due to manual investigation methods. Detecting fraud is essential because it undermines the fundamental principle of trust, which is crucial in the insurance sector. Expert assessment and auditing are vital for the detection of automobile insurance fraud, but these methods can be costly and inefficient when exposing fraudulent claims manually. To mitigate financial losses for both policyholders and insurers, it is imperative to predict and detect fraud in claims before settlement. Detecting fraudulent claims is a crucial responsibility for insurance companies, and it revolves around assessing the likelihood of an accident occurring. To predict the likelihood of an accident, classification techniques are employed. The application of data mining techniques significantly accelerates the process of determining accident occurrence. When there is a high probability of an accident happening, the claim is deemed legitimate and proceeds to subsequent stages, including claim amount determination and premium percentage calculation. The premium percentage is computed based on the client's personal information and their neighborhood details.

The proposed work involves implementing an efficient classification system using both the decision tree and Bayesian classification algorithms. The Accident Occurrence dataset contains attributes that represent circumstances leading to accidents and those that do not. This dataset serves as the training set for classification. Classification algorithms are then applied to construct classifiers. From the resulting classifiers, the one that provides the most accurate results is selected for predicting the likelihood of 'accident occurrence.' If and only if the chance of an accident is high or very high, the claims are scrutinized for potential fraud. The Premium dataset is independent of the accident occurrence outcome. By applying an effective classification method to the Premium dataset, the percentage of the premium amount to be paid by individual customers is determined. The research's objective is to reduce the time needed to predict fraudulent claims. This work enables efficient classifiers to make precise predictions regarding accident occurrence, claims, and premium datasets.

II Literature Review

In this paper, a comprehensive review of various methodologies presented in prior research is conducted. The literature survey offers insight into different research works presented by various authors that align with and support the proposed work.

In 2007, the Automobile Fraud Claim Screening System was introduced [1], employing a cost-sensitive screening process with data mining methodologies. The vehicle insurance premium profiles were gathered during policy issuance, claims processing, and damage evaluation. Notably, the study didn't consider fraud details related to injuries or medical treatments. In 2008 [2], a proposal for using Bayesian classification in fraud detection for automobile insurance surfaced, evaluating the model's performance through metrics, a confusion matrix, and ROC curves.

In 2013, a proposal was made for predicting severity and duration of road traffic accidents [3]. This model, based on a 2010 accident dataset from China, forecasts indicators like fatalities, injuries, and accident damage. The paper recommends effective techniques to reduce accidents and enhance road safety, utilizing the integrity of the Ordered Probit method for prediction. Additionally, a comprehensive 2010 survey focused on data mining-based fraud detection [4], covering diverse fields such as terrorist detection, cost-effective crime detection, interruption, and spam detection. The survey explored various supervised and unsupervised algorithms employed in these contexts.

In 2013, the study "Predicting Severity and Duration of Road Traffic Accidents" [3] introduced a model forecasting three accident severity indicators based on a 2010 Chinese dataset. The paper provides recommendations for reducing accidents and improving road safety, employing the goodness-of-fit of the Ordered Probit model for predictions. In 2015, an application of One Class Support Vector Machine (OCSVM) based undersampling for churn prediction and insurance fraud detection was introduced [5]. The use of OCSVM-based undersampling enhanced classifier performance, simplifying the fraud detection system while yielding significant results.

In 2015, a proposal for An Identification Algorithm and model construction for automobile scam based on data mining was presented [6]. The paper employed an outlier detection method, utilizing nearest neighbors with pruning rules to identify automobile insurance fraud. Additionally, association rules were applied to extract patterns related to auto insurance fraud. The optimization of the algorithm's efficiency was achieved by pruning the dataset and reducing data spaces, minimizing unnecessary distance computations.

In 2016, A study and application of the Simple CART model in mining automobile insurance fraud were introduced [7]. This research highlighted the effectiveness of the Simple CART algorithm in reducing the number of variables when applied to large datasets with numerous relevant explanatory variables. The conclusion was that Simple CART is well-suited for handling substantial and unbalanced datasets. Another research project in 2016 [8] focused on analyze the vehicle insurance data using different classification methods for vehicle client behavior analysis. Emphasizing the importance of preprocessing methods like key element collection and correct corresponding methodologies, this work contributed to enhancing a classifier's effectiveness in distinguishing between regular and non-regular customers of an insurance company.

The specific research gap address is the lack of adaptive and accurate fraud detection systems capable of handling evolving fraudulent tactics. The research gap lies in the inability of existing fraud detection systems to effectively adapt to new and emerging fraud techniques. This framework outlines the limitations of current methods, such as their reliance on static rules and historical data that may not capture new fraud patterns. A combination of effectivw machine learning algorithms like SVM, CART and LMT and real-time data analysis is proposed to address these limitations.

III. Research Methodology

Classifying and predicting in datasets such as **Accident Occurrence, Premium,** and **Insurance Claim** can be challenging due to their high dimensionality, which results in increased storage space and processing time requirements. The issue of feature selection offers a solution to mitigate these challenges. It involves selecting a discriminative subset of features that are crucial for fraud prediction and can enhance dataset classification. This research work involves an analysis of several decision tree algorithms and the LMT algorithm to assess their accuracy in these tasks. Insurance companies typically maintain strict confidentiality regarding their client and claim details. In this research, a synthetic dataset is generated, as freely available datasets for analyzing insurance and premium information are not readily accessible. However, the accident dataset is obtained from the data.gov.in website. Drawing from case studies and field research on automobile insurance fraud, the attributes and nature of the synthetic datasets, specifically Premium and Insurance Claim are designed. The Accident Occurrence dataset contains information about accidents that occurred in 2012. By constructing a classifier using this dataset, it becomes feasible to determine whether an accident genuinely occurred. To predict suspicious claims and calculate premium amounts, classification rules need to be derived by mining these three datasets. The calculation of the premium amount percentage is contingent on factors like the insurer, policy type, customer's coverage choices, and company-specific rules. Fraud detection in a claim is only conducted if the accident genuinely occurred. Predicting accident occurrences serves as the fundamental screening process to identify fraudulent claims[9,10,11]. Currently, there is no established strategy for managing the percentage of the premium amount within insurance companies. This research proposes an efficient method for computing premium percentages using data mining techniques.

Phase I: In this phase, Dataset Generation takes place, and both the Insurance Claim and Premium Datasets are synthetically generated.

Phase II: Data preprocessing is the focus of Phase II, where the Insurance Claim and Premium Datasets are prepared for analysis. Data Generalization techniques are applied to ensure the datasets are in a consistent and unified format.

Phase III: Classifier Construction is the primary task in this phase. Various algorithms such as SVM, Simple CART and LMT are utilized to construct classifiers.

Phase IV: The final phase, Testing of Constructed Classifiers, involves performing classification on test data, and the outcomes are thoroughly analyzed.This workflow outlines the key steps involved in the proposed methodology for handling insurance-related datasets and fraud detection.

A. Dataset Generation

In this step, synthetic datasets are created to facilitate the prediction of fraudulent claims and the calculation of premium percentages for individuals. The Insurance Claim and Premium datasets are synthetically generated, drawing insights from various case studies. The Accident Occurrence dataset, on the other hand, was obtained from the **'www.data.gov.in'** website and contains information related to accident occurrences. This dataset comprises data that characterizes the circumstances that both lead to and do not lead to accidents. Claims in insurance generally fall into two categories. The first type pertains to claim amounts related to vehicle theft, while the second type concerns claim amounts associated with accidents.

B. Dataset Preprocessing

The raw accident data obtained from www.data.gov.in may contain unwanted and duplicate values. Additionally, the dataset retrieved from this source could have missing values or noisy data, which can result in incorrect classification of the data. To address missing values, various techniques can be employed, such as filling them with frequent values or random values. In this research work, missing values are filled with the more frequently occurred value of the attribute. Furthermore, the attributes gathered from different case studies for the Premium and Insurance Datasets require fine-tuning to achieve better results. This fine-tuning process likely involves refining and optimizing the attributes to improve the overall performance of the datasets in subsequent analyses and classifications.

C. Data Generalization

Data generalization proves valuable in managing Insurance Claim and Premium datasets. In the vehicle insurance claim dataset, key elements are consistently maintain in binary(dual) format, while the Premium dataset attributes are in nominal format, impacting data integrity. To address this, a transformation to a unified format is necessary. Generalization plays a key role by replacing basic model or basic data with deep level methods. For instance, concepts like youth, middle, old, and senior can be employed to map low-level numerical attributes such as age. This facilitates a more unified and cohesive representation of the data.

D. Dimensionality Reduction.

Dimensionality reduction is implemented on the Accident Occurrence Dataset, originally obtained from www.data.gov.in, comprising 15 attributes. Some attributes are considered irrelevant for predicting accidents, and their inclusion can lead to an inefficient model and increased processing time. To address this, the Gain Ratio estimatation method is employed to choose the most major attributes. **Attribute like 'Easting,' 'Northing,' 'Number of Vehicles,' and 'AccidentTime**' are identified as the least significant and subsequently removed. After dimensionality reduction, the data is stored in the Attribute-Relation File Format (ARFF) for further data mining and analysis. This process results in a more streamlined dataset, retaining only essential attributes for accurate and efficient prediction of accident occurrences.

E. Classifier Construction

Classification algorithms utilize features, attributes, or predictor variables to make predictions, with the "Class Variable" being the target variable. This variable, in either binary or categorical format, is what classification algorithms seek to predict. These techniques prove beneficial when dealing with categorical class labels, such as predicting student performance as 'good,' 'bad,' or 'fair,' or assessing cancer risk as high/medium/low. In the literature, SVM and Simple CART stand out as most supported decision tree based classified algorithms known for their accuracy. Additionally, Naive Bayes, a probabilistic based algorithm, is widely used for constructing classifiers and is employed in various research works[10]. These algorithms play a pivotal role in achieving accurate predictions and classifications based on the provided features and class variables.

B. Simple Cart Algorithm

CART **(Classification and Regression Trees)** differs from ID3 and C4.5 algorithms in that it forms binary trees, with each internal node having precisely two outgoing edges. Unlike ID3 and C4.5, which generate decision trees with variable branches per node, CART is distinct in its ability to perform regression analysis using regression trees. This analysis predicts a dependent variable based on predictor variables over a specified time period. The CART decision tree is a binary recursive partitioning method that handles both continuous and nominal attributes as targets and predictors. It employs the Gini index for splitting during tree growth, extending to a maximum size before being pruned back to the root through cost-complexity pruning. CART aims to produce a series of nested pruned trees, offering a range of candidate optimal trees. Additionally, the CART mechanism provides features like automatic class balancing, handling of missing values, support for cost-sensitive learning, dynamic feature construction, and probability tree estimation.

C. Logistic Model Tree Induction

In 2003, Niels Landwehr introduced **Logistic Model Tree induction**, combining linear logistic regression and tree induction concepts. Linear logistic regression is suitable for noisy or linearly structured data, while tree induction is preferred for highly non-linear datasets. The "Logistic Model Trees" merge these approaches, featuring a decision tree with logistic regression models at the leaves. This combination addresses classification tasks, providing both a final model with a single tree and class probability estimates. Notably, Logistic Model Trees (LMT) excel in probability estimates compared to other state-of-the-art learning schemes. The tree grows by initiating a logistic model at the root using the LogitBoost Algorithm, determining the number of iterations through fivefold cross-validation. Data is split into train and test sets for five iterations, and the LogitBoost algorithm runs, generating the logistic regression model at the root. Subsequent splits at the child nodes, constructed based on the data at the root, utilize LogitBoost algorithm runs to produce probability estimates.

D. Support Vector Machines

The **Support Vector Machine** (SVM) classifier, developed by Vapnik and others in 1992, originates from statistical learning theory. This versatile algorithm supports classification, regression, and various learning tasks. SVM, primarily a classifier, constructs hyperplanes in a multidimensional space to segregate instances with different class labels. Handling both regression and classification tasks, SVM accommodates multiple continuous and categorical variables. Operating on the concept of decision planes, SVM establishes decision boundaries to differentiate between objects with distinct class memberships. In its standard form, SVM takes input data, predicts the class for each input, and functions as a non-probabilistic binary linear classifier.

E. Classification over the Test Data

Once the classifier is built, its performance is assessed using a test dataset. The classifier's accuracy is determined using a formula (Eq. 1) that considers the number of correctly classified test records.

$$Classification\ Accuracy = \frac{No.\ of\ correctly\ classified\ test\ records}{Total\ no.of\ test\ records} \quad (1)$$

F. Classification Metrics

The chosen Non-Dominated feature subset undergoes validation through classification performance assessment. This validation includes the computation of crucial classification metrics like precision, recall, and accuracy both before and after the feature selection process. A rise in classification accuracy with the selected features provides confirmation of the viability of the Non-Dominated feature subset. Metrics such as precision, recall, and accuracy are determined using values such as True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

TP /True Positive - positive predicted as positive

TN/True Negative – negative predicted as positive

FP/False Positive – negative predicted as negative

FN/False Negative – positive predicted as negative

G. Precision or Positive Predicted Value

Precision, a metric assessing the proportion of retrieved instances that are relevant, calculates the number of correctly predicted positive instances divided by the total instances predicted accurately. It is expressed using the formula (Eq. 2), which quantifies precision as the ratio of True Positives (correctly predicted positive instances) to the sum of True Positives and False Positives (instances predicted as positive but are not).

$$\text{Precision} = \frac{\text{TP}}{\text{FP+TP}} \quad (2)$$

H. Recall or Sensitivity

Recall, or sensitivity, evaluates the proportion of relevant instances successfully retrieved. It quantifies the number of correctly predicted positive instances divided by the sum of correctly predicted positive instances and incorrectly predicted negative instances. The formula for recall (Eq. 3) reflects this calculation, providing insight into how well an algorithm captures all relevant positive instances.

$$\text{Recall} = \frac{\text{TP}}{\text{FN+TP}} \quad (3)$$

I. Accuracy

Accuracy serves as a metric measuring the overall correctness of predictions made by an algorithm. It is determined by dividing the number of correctly predicted positive and negative instances by the total number of instances in the dataset. The accuracy formula (Eq. 4) encapsulates this calculation.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{FP+FN+TP+TN}} \quad (4)$$

IV. Results and Discussions

A. Dataset Used

The Accident Occurrence dataset utilized in this research is obtained from www.data.gov.in, providing information about accidents that occurred in the year 2012. Inaddition, two other datasets, Insurance Claim and Premium, are synthetically generated based on specific case studies.

B. Experimental Results and Discussion

The classifiers, namely Simple CART, LMT, and SVM, are implemented on the Insurance Claim, Premium, and Accident Occurrence datasets using WEKA 3.8. The results obtained from these classifiers are documented across three test options: 50:50, 66:34, and 10CV-10 Cross Fold Validation, providing a comprehensive evaluation of the classifier performance[12,13,14].

C. Comparison between SVM, LMT and Simple CART Classifiers

(i) Accident Occurrence Dataset:

Table 4.1 summarizes the classifier performance on the unprocessed Accident Occurrence Dataset under three test options follows in the table (66:34,10CV & 50:50)

TABLE 4.1: PRECISION, ACCURACY, RECALL GAINED FOR ACCIDENT ATTAINED DATASET IN THE ABSENCE EARLY COMPILATION RESULT

Classifier	Validation	Precision Level	Accuracy Level (%)	Recall Level
Simple CART	66:34	0.94	94.5	0.94
	10CV	0.94	94.7	0.946
	50:50	0.92	91.2	0.92
SVM	66:34	0.93	93.5	0.94
	10CV	0.94	94.6	0.95
	50:50	0.93	93.1	0.93
LMT	66:34	0.93	93	0.93
	10CV	0.91	91.4	0.91
	50:50	0.92	93.9	0.94

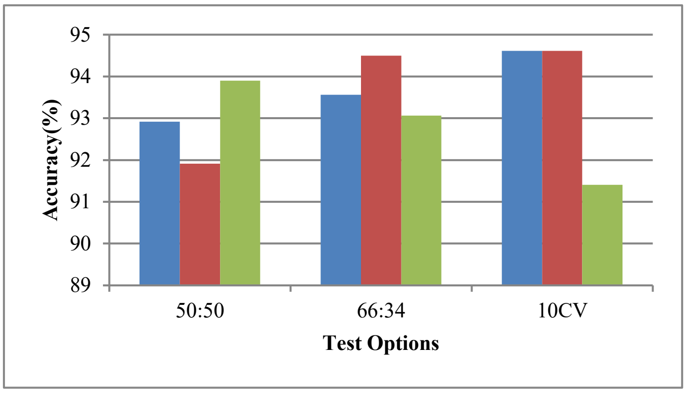


Fig. 4.1 Comparison between SVM, Simple CART and LMT under three test options over Accident Occurrence Dataset with preprocessing.

The analysis of Table 4.1 and Fig. 4.1 reveals that LMT demonstrates higher accuracy in the 66:34 and 10CV test options.

However, its accuracy drops in the 50:50 test option. Among the test options, 10CV consistently provides better overall performance compared to 50:50 and 66:34

Table 4.2 presents the performance of the Simple CART, LMT, and SVM classifiers on the Accident Occurrence Dataset after preprocessing. The preprocessing involved the use of the Gain Ratio attribute estimate the method to choose significant key elements, resulting in the removal of the "Type of Vehicle" attribute.

TABLE 4.2: PRECISION, ACCURACY, RECALL OBTAINED FOR ACCIDENT OCCURRENCE DATASET WITH PREPROCESSING

Classifier	Validation	Precision Level	Accuracy Level (%)	Recall Level
Simple CART	66:34	0.93	92.60	0.93
	10CV	0.90	90.70	0.91
	50:50	0.95	95.30	0.95
SVM	66:34	0.88	89.60	0.90
	10CV	0.90	90.40	0.90
	50:50	0.92	93.10	0.93
LMT	66:34	0.90	91.00	0.91
	10CV	0.88	88.40	0.89
	50:50	0.89	89.30	0.89

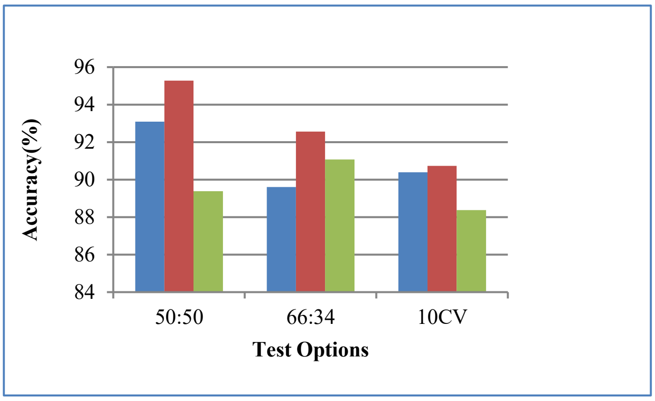


Fig. 4.2 Comparison between SVM, Simple CART and LMT under three test options over Accident Occurrence Dataset with preprocessing

From the analysis of Table 4.2 and Fig. 4.2, it is evident that the Simple CART algorithm performs better than the other two algorithms. Additionally, the 50:50 test option yields better results compared to 10CV and 66:34. It is noteworthy that the performance of the algorithms slightly decreases with preprocessing, indicating that the dataset is already adequately preprocessed and does not require further adjustments.

(ii)Insurance Claim Dataset:

Table 4.3 displays the performance of the SVM, LMT, and Simple CART classifiers on the Insurance Claim dataset three test options follows in the table (66:34,10CV & 50:50). Based on the data in Table 4.3 and Fig. 4.3, it is evident that the 10CV test option consistently yields the best results in terms of classifier accuracy, outperforming the 50:50 and 66:34 options. Among the three classifiers, the Simple CART algorithm stands out, achieving a classifier accuracy of 99.41% under the 66:34 test option, which is notably higher (0.62% and 0.22%)

than the accuracy obtained under the 50:50 and 10CV test options, respectively

TABLE 4.3: PRECISION, ACCURACY AND RECALL ATTAINED FOR INSURANCE ASSERT DATASET IN THE ABSENCE EARLY COMPILATION

Classifier	Validation	Precision Level	Accuracy Level (%)	Recall Level
Simple CART	66:34	0.95	99.40	0.99
	10CV	0.99	99.10	0.99
	50:50	0.99	98.80	0.99
SVM	66:34	0.97	97.00	0.97
	10CV	0.99	99.40	0.99
	50:50	0.98	97.50	0.98
LMT	66:34	0.98	98.20	0.98
	10CV	0.98	97.80	0.98
	50:50	0.94	97.50	0.97

Simple CART performs well in the 50:50 and 66:34 test options for the Insurance Claim dataset, outperforming other classifiers. However, its accuracy decreases in the 10CV test option. SVM, despite lower accuracy in 50:50 and 66:34, performs better in the 10CV test option.

Simple CART performs exceptionally well in the 66:34 and 50:50 test options, achieving 99.41% classifier accuracy, surpassing the 10CV option by 0.62% and 0.22%. Data preprocessing involved the use of the Gain Ratio key element estimating assessment method to select notable key elements in the insurance asset dataset[15,16,17].

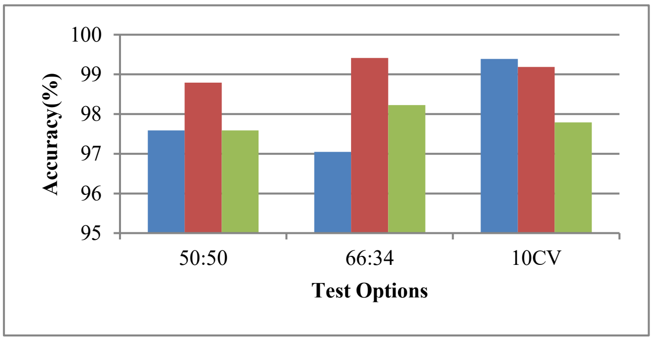


Fig. 4.3 Comparison between SVM, Simple CART and LMT under three test options over Insurance Claim Dataset without preprocessing.

The Key elements with gain values less than 0.01 are deemed less significant and have been excluded. Specifically, the attributes "vehicle insurance assert occurrence time is off extreme time,""unemployed," and "Continuously changing address/phone_no" exhibited gain values below 0.01 and were consequently removed. The classifier accuracy results for Simple CART, LMT, and SVM classifiers were evaluated comes under these three test options (50:50, 66:34, 10CV) on the Insurance Assert Dataset after preprocessing, as outlined in Table 4.4.

TABLE 4.4: PRECISION, ACCURACY AND RECALL ATTAINED FOR INSURANCE ASSERT DATASET ESCORTED BY EARLY COMPILATION

Classifier	Validation	Precision Level	Accuracy Level (%)	Recall Level
Simple CART	66:34	0.99	99.40	0.99
	10CV	0.99	99.80	0.99
	50:50	0.99	99.60	0.99
SVM	66:34	0.99	99.40	0.99
	10CV	0.99	99.40	0.99
	50:50	0.98	98.40	0.98
LMT	66:34	0.96	95.29	0.96
	10CV	0.98	98.20	0.98
	50:50	0.96	95.90	0.96

The Table 4.4 and Fig. 4.4, it is evident that the Simple CART algorithm is better than the other two algorithms. Additionally, the 10CV test option yields better results than 50:50 and 66:34. Notably, the accuracy of LMT decreases with preprocessing. In the 50:50 and 66:34 test options, LMT' accuracy is reduced by 1.61% and 2.94%, respectively, while it increases by 0.4% in the 10CV option.

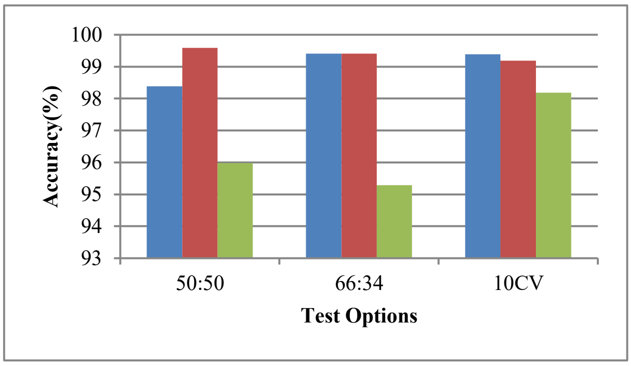


Fig. 4.4 Comparison between SVM, Simple CART and LMT under three test options over Insurance Claim Dataset with preprocessing.

(iii) Premium Dataset:

The performance of the Simple CART, LMT and SVM classifiers beneath three test options(50:50 / 66:34 & 10CV) over Premium Dataset in shown in Table 4.5.

TABLE 4.5: PRECISION, ACCURACY AND RECALL ATTAINED FOR INSURANCE ASSERT DATASET IN THE ABSENCE EARLY COMPILATION

Classifier	Validation	Precision Level	Accuracy Level (%)	Recall Level
Simple CART	66:34	0.74	72.90	0.72
	10CV	0.75	73.34	0.73
	50:50	0.62	68.20	0.68
SVM	66:34	0.67	68.20	0.68
	10CV	0.67	65.73	0.66
	50:50	0.63	65.00	0.65
LMT	66:34	0.83	78.90	0.78
	10CV	0.82	80.36	0.80
	50:50	0.78	77.10	0.77

Table 4.5 and Fig. 4.5 indicate that the 10CV test option provides the best results in terms of classifier accuracy. Among the three classifiers, LMT outperforms the others. Under the 10CV option, LMT achieves 80.37% classifier accuracy, which is 3.19% higher than the 50:50 and 2.18% higher than the 66:34 options.

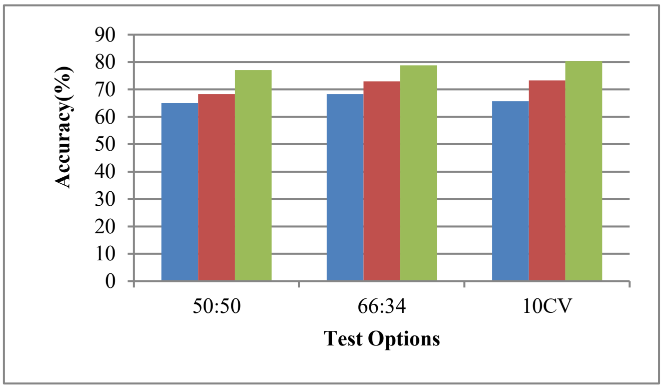


Fig. 4.5 Comparison between SVM, Simple CART and LMT under three test options over Premium dataset in the absence early compilation

TABLE 4.6: PRECISION, ACCURACY, RECALL OBTAINED FOR PREMIUM DATASET ESCORTED BY EARLY COMPILATION

Classifier	Validation	Precision Level	Accuracy Level (%)	Recall Level
Simple CART	66:34	0.66	67.60	0.68
	10CV	0.65	67.30	0.67
	50:50	0.64	65.90	0.66
SVM	66:34	0.82	77.60	0.78
	10CV	0.74	75.10	0.75
	50:50	0.77	75.90	0.76
LMT	66:34	0.80	77.00	0.78
	10CV	0.80	79.10	0.79
	50:50	0.81	79.10	0.79

To improve classifier accuracy, data preprocessing involves removing less important attributes and retaining the most valuable ones. The Gain Ratio Attribute evaluation method, implemented in Weka 3.8, is utilized in this research to rank and select the most relevant attributes for classification[18,19]. Attributes with a gain value less than 0.01 are deemed insignificant and consequently eliminated.

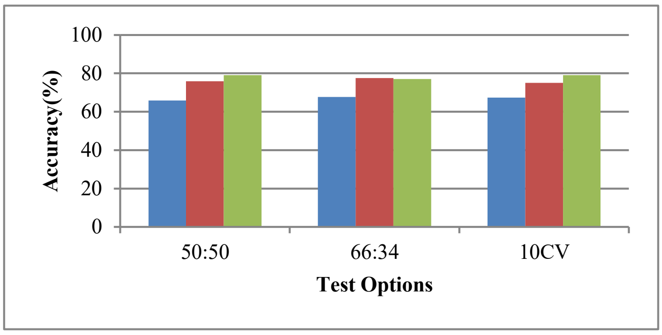
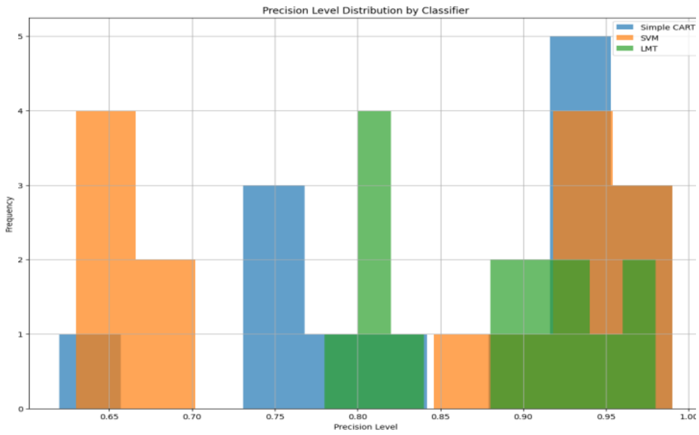


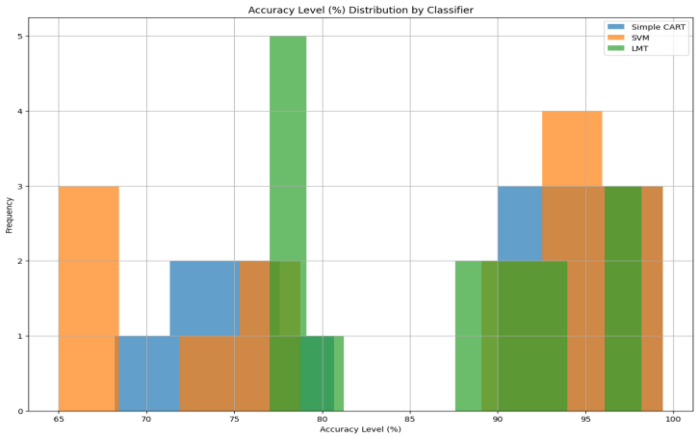
Fig. 4.6 Comparison between SVM, Simple CART and LMT under three test options over Premium Dataset escorted by early compilation.

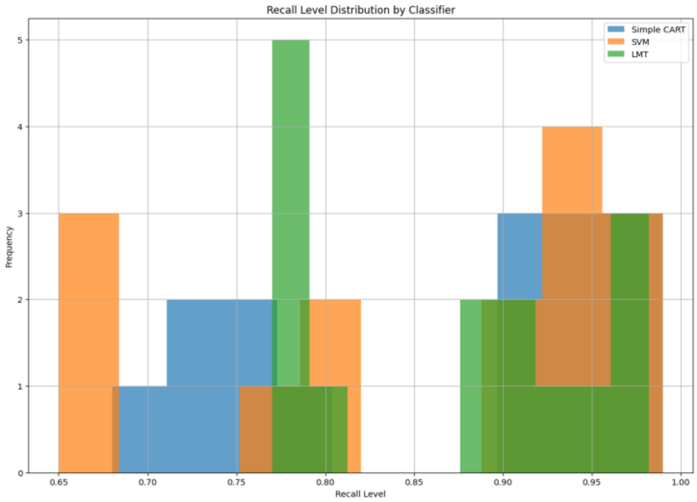
In the Premium dataset, the attributes 'accident history,' 'age,' and 'distance traveled per day' are the least ranked and have been removed accordingly.

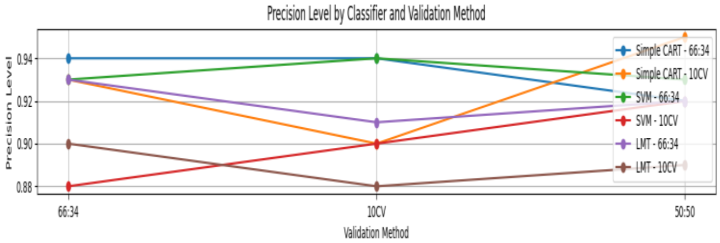
The overall performance of the three methodologies on the Insurance assert dataset, premium dataset, and Accident Occurrence dataset appears to remain relatively consistent, whether or not preprocessing is applied. This observation suggests that the Gain Ratio evaluation method may not be essential for further enhancing the performance of these datasets.

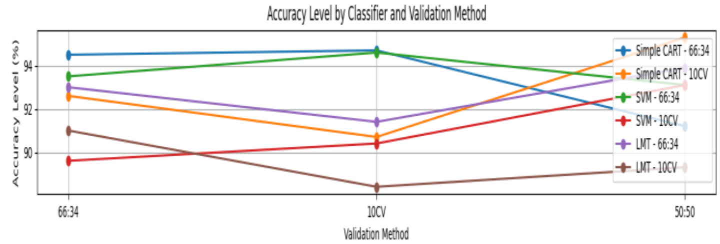
Upon analyzing the three datasets, it is evident that Simple CART performs the best in both the Insurance Claim and Accident Occurrence datasets. However, LMT outperforms the other classifiers on the Premium dataset, whether preprocessing is applied or not. In terms of test options, 10CV generally yields better results across all datasets, except for the Accident Occurrence Dataset without preprocessing[20,21,22].

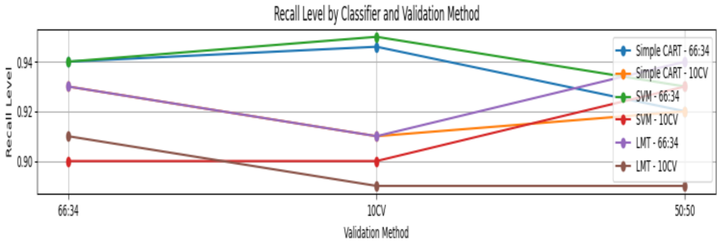












IV. Conclusion and Future Enhancement

This study aims to streamline the prediction of fraudulent claims in automobile insurance, focusing on reducing time and costs. It centers on predicting premium amounts for all clients, with fraudulence in claims predicted only when accurately forecasting accident occurrences. Employing three classifiers—SVM, Simple CART, and LMT—the research addresses both insurance fraud and premium predictions. A limitation is the constrained use of training data due to the confidentiality of client and claim details. The study proposes economically feasible solutions for clients and insurers. Notably, Simple CART outperforms other algorithms in Insurance Claim and Accident Occurrence datasets, while LMT excels in the Premium dataset. Future work will take away into the relationships between Insurance Claims, Premium datasets, and Accident Occurrence, aiming to optimize classification algorithms for improved results with real datasets.

References

[1] Mercedes Ayuso et.al “Strategies for Detecting Fraudulent Claims in the Automobile Insurance Industry”, European Journal of Operational Research 176(2007) 565-583

[2] Rekha Bhowmik, ”Data mining Techniques in Fraud Detection”, journal of Digital Forensics, Security and Law, vol. 3(2).

[3] Fang Zong,HuiyongZhang,HongguoXu,Xiumei Zhu, and Lu Wang, Predicting Severity and Duration of Road Traffic Accident ,Problems in Mathematical Engineering,2013

[4] Vincent Lee, Kate Smith, Clifton Phua and Ross Gayler, “A Comprehensive Survey of Data Mining based Fraud Detection Research,Cornell University Library, 2010

[5] Vadlamani Ravi, G.GaneshSundarkumar, and V.Siddeshwar, ”One Class Support Vector Machine based Undersampling: Application to Churn Prediction and Insurance Fraud Detection”, IEEE Int. Conf. on Computational Intelligence and Computing Research, 2015

[6] YaqiLi andChunyan ” The Identification Algorithm and The Identification Algorithm and Model Construction of Automobile Fraud based on Data mining”, Fifth Int. Conf. on Instrumentation and Measurement, Computer, Communication and Control, 2015

[7] YaqiLi, ChunYan, WeiLiu, and MaozhenLi, ”Research& Application of Simple CART Model in Mining Automobile Insurance Fraud”, 12th Int. Conf. on Natural Computation, Fuzzy Systems and Knowledge Discovery, 2016

[8] D.Saidur Rahman, KaziZawadArefin, SaqifMasud, Shahida Sultana, and RashedurM.Rahman,” Analyzing Life Insurance Data with Different Classification Techniques for Customer’s Behaviour A nalysis”, Advanced Topics in Intelligent Information and Database Systems, Studies in Computional Intelligence 710, Springer Int. Publishing

[9] Richard A.et.al ,” A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis”, IEEE transactions on knowledge and data engineering, vol 16,no 5, May 2004

[10] K.UlagaPriya and S.Pushpa ,” A Survey on Fraud Analytics using Predictive Model in Insurance Claims”, Int. Journal of Pure and Applied Mathematics, volume 114 no 7 2017, 755-767

[11] SureshYaram, “Machine Learning Algorithms for Document Clustering and Fraud Detection”, IEEE Int. Conf. on Data Science and Engineering (ICDSE), 2016

[12] Madhar Taamneh,SharafAlkhedar, Shalah Alkhedar,”Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates”, Journal of Transportation Safety and Security, Volume 9, 2017

[13] Apeksha V.Sakhare, prajaktas.kasbe, “A review on road accident data analysis using data mining techniques”, Int. Conf. on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017

[14] Sossi Alaoui, Safae, Yousef Farhaoui, and Brahim Aksasse. "A comparative study of the four well-known classification algorithms in data mining." In Advanced Information Technology, Services and Systems: Proceedings of the International Conference on Advanced Information Technology, Services and Systems (AIT2S-17) Held on April 14/15, 2017 in Tangier, pp. 362-373. Springer International Publishing, 2018.

[15] Dr. B. Lavanya and B. Divya, Predictive Analytics on Accident Data Using Rule Based and Discriminative Classifiers, Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 3 (2017) pp. 461-469.

[16] Fletcher Lu, J EfrimBoritz and Dominic Covey, ”Adaptive Fraud Detection using Benford’s Law”, Canadian AI,LNAI 4013,PP.347-358, Springer-Verlagberlin Heidelberg, 2006

[17] El BachirBelhadji, Georges Dionne, and FaouziTarkhani, “A Model for the Detection of Insurance Fraud”, The Geneva rs on Risk and Insurance vol 25. No.4 517-538, October 2000

[18] Muthukaruppan et.al (2011), ‘A Novel Hybrid Approach to Machine Learning’, Study Documents, Department of Computer Science and Engineering, University of Washington. vol. 2, no. 5, pp. 114-340

[19] Andri Irfan, Ronal AI Rasyid and Susanty Handayani,” Data mining applied for accident prediction model in Indonesia toll road”, AIP Conf. Proceedings 1977, 060001 2018

[20] V M Ramachandiran, et.al, Prediction of Road Accidents Severity using various algorithms, Int. Journal of Pure and Applied Mathematics, Volume 119 No. 12 2018, 16663-16669

[21] Dr. D. Ganeshkumar et.al “A data mining approach on various classifiers in email spam filtering” in Int. Journal of . Res. Appl. Sci. Eng. Technol 3 (1), 8-14

[22] Fancello Gion Franco, StefenoSoddu,PaolaFadda“An accident prediction model for urban road networks”,Journal of Transportation Safety and Security, Volume 10,2018