IS 3053/ ST 3070 GROUP PROJECT

ANALYSIS ON HOUSE INSURANCE DATASET

PREPARED BY

GROUP 03

Tishani Wijekoon - s16379
Lasandi Perera - s16388
Amidu Yohan - s16385
Thenuka Yatawara - s16383
Nihindu Omal - s16276

Abstract

This study analyzes a dataset on property insurance claims. During the analysis, key features associated with fraudulence were identified. By conducting exploratory data analysis and fitting predictive models using classification methods and Machine Learning techniques, factors that help determine fraudulence were discovered. Further, cluster analysis was applied to identify groups with higher fraud likelihood and their similar features were observed. The study offers insights for improved fraud detection and risk management.

Table of contents

1. **Objective………………………………………………………………………….…1**
2. **Dataset Description……………………………………………………………….....2**
3. **Pre-Processing…………………………………………………………………….....2**
4. **Descriptive Analysis…………………………………………………………….…..3**
    - **Pearson's Correlation between numerical predictors ..………………………….3**
    - **Spearman's Rank Correlation for Response with numerical variables……..….3**
    - **Pearson's Chi squared test for Response with categorical variables………..….3**
    - **Univariate Analysis………………………………………………………………...3**
    - **Bivariate Analysis…………………………………………………………………..4**
    - **Associations between predictors…………………………………………......5**
5. **Advanced Analysis ……………………………………………………………………..5**
    - **Classification Trees……………………………………………………...…5**
    - **Random Forest …………………………………………………………………6**
    - **Support Vector Machine (SVM) ………………………………………………7**
    - **k-Nearest Neighbours (kNN) ……………………………………………...7**
    - **Cluster Analysis ………………………………………………………………8**
6. **Conclusion……………………………………………………………………………10**

Objective of Study

Identify the features that distinguish fraudulent claims from legitimate ones through the application of descriptive statistics, data pre-processing, and advanced analytical techniques. The analysis aims to uncover key patterns, isolate clusters with a higher incidence of fraud, and derive actionable insights for fraud detection.

Dataset Description

The dataset, *insurance_claims.csv*, consists of 4415 records across 21 variables. It contains detailed information on insurance claims, including:

Variable No	Variable Name	Variable descriptions	Variable No	Variable Name	Variable descriptions
1	**claimid**	Claim ID	12	**gender**	Gender
2	**incident date**	Date of incident	13	**dob**	Date of birth
3	**claim type**	Type of claim	14	**edcat**	Level of education
4	**uninhabitable**	Property uninhabitable	15	**job_start_date**	Employment starting date
5	**claim amount**	Cost of claim in thousands	16	**retire**	Retired
6	**fraudulent**	Fraudulent claim	17	**income**	Household income in thousands
7	**policyid**	Policy ID	18	**marital**	Marital status
8	**policy date**	Date policy went into effect	19	**reside**	Number of people in household
9	**coverage**	Amount of coverage in thousands	20	**occupancy date**	Date of occupancy
10	**deductible**	Deductible	21	**primary residence**	Property is primary residence
11	**townsize**	Size of hometown			

Pre-processing

The pre-processing phase involved several key steps to ensure data quality and prepare the dataset for fraud analysis:

1. **Data Import & Cleaning** – The workspace was cleared, a random seed set, and the dataset imported. Duplicate rows and records with missing values were removed.

2. **Data Reduction & Type Conversion** – Unnecessary identifiers (claim/policy IDs) were dropped. Date columns were standardized to a datetime format for time-based calculations.

3. **Feature Engineering** – Age was computed from the incident date and date of birth, and policy/occupancy durations were calculated. Irrelevant date columns were removed.

4. **Dataset Balancing & Partitioning** – To address class imbalance, all fraudulent claims were retained, and 500 non-fraudulent claims were randomly selected. The dataset was shuffled and split into training (80%) and test (20%) sets.

Descriptive Analysis

Pearson's Correlation between numerical predictors

	income	Claim_amount	coverage	deductible	age	Policy_duration
income	1	0.57051	0.27046	0.24177	0.07235	0.07427
claim amount	0.57051	1	0.42337	0.53226	0.10899	0.15195
coverage	0.27046	0.42337	1	0.20418	0.05993	0.07259
deductible	0.24177	0.53226	0.20418	1	0.23565	0.24199
age	0.07235	0.10899	0.05993	0.23565	1	0.82866
Policy_duration	0.07427	0.15195	0.07259	0.24199	0.82866	1

Table 2: Pearson's Correlation Table

The correlation matrix highlights key relationships in insurance data. Age and policy duration (0.83) show the strongest association, while claim amount and deductible (0.57) also have a notable correlation. Moderate relationships exist between claim amount, coverage (0.42), and deductible (0.53), suggesting financial influences on claims. These insights aid in risk assessment and pricing strategies.

Spearman's Rank Correlation for Response with numerical variables

	income	claim_amount	coverage	deductible	age	policy_duration
Fraudulent Claims	-0.00429	-0.0223	-0.02394	0.03708	-0.06972	-0.09885

Table 3: Spearman's Rank Correlation Table

The numerical predictors show negligible correlation with fraudulence, suggesting weak associations. Other methods may be needed for better prediction.

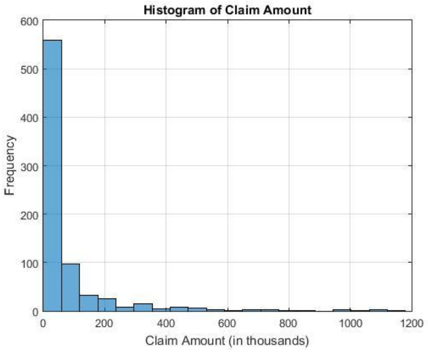
Pearson's Chi squared test for Response with categorical variables.

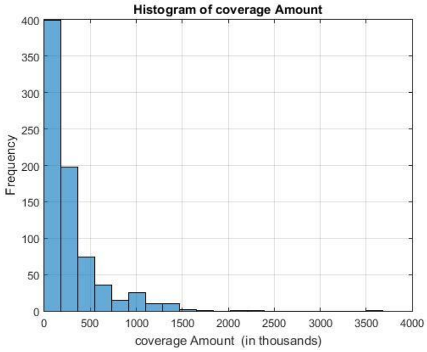
Categorical Variable	claim type	Uninhabitable	town size	gender	edcat	retire	marital	reside	primary residence
P Value	0.0607	0.0207	0.4363	0.8422	0.6076	0.0580	0.9606	0.0764	0.2177

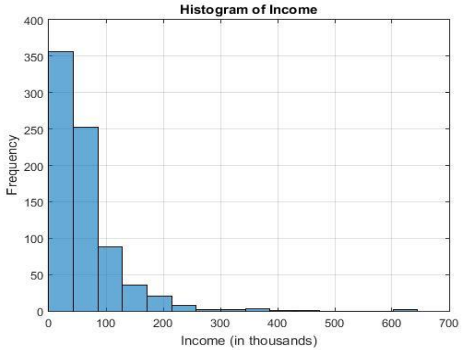
Table 4: Pearson's Chi squared test values Table

Pearson's chi-square test revealed that there was a significant relationship only between the 'fraudulent' response variable and the 'uninhabitable' predictor variable.

Univariate Analysis







The income, claim amount, and coverage distributions exhibit negative skewness.

Bivariate Analysis

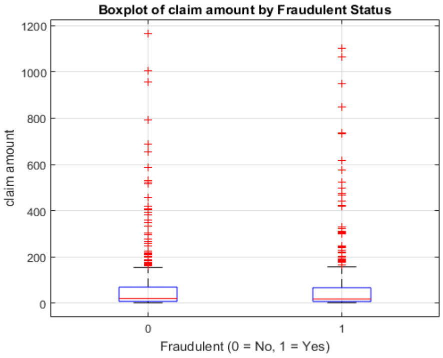


Figure 2: Boxplot of claim amount by Fraudulent Status

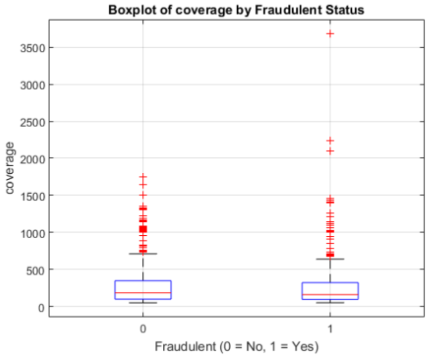


Figure 3: Boxplot of claim amount by Fraudulent Status.

Fraudulent Status with Claim Amount and Coverage

Figures 2 and 3 show that, there is no significant difference between the median of coverage for fraud and non-fraud status or between the median of claim amount for fraud status. This can be due to the presence of outlying observations.

Fraudulent Status with Policy Duration

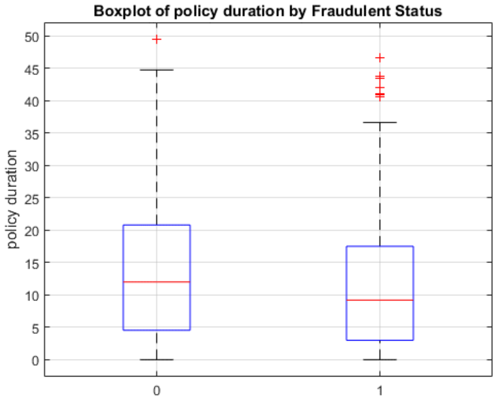


Figure 4: Boxplot of claim amount by Fraudulent Status.

A significant difference between policy duration and fraud status is clear in the figure with some outliers present. Mann-Whitney test was applied as the normality assumption was violated by the QQ plots. With a p-value of 0.006, result confirms there is a significant difference between fraudulent status and Policy Duration.

"Challenger et. Al(2011) found that fraudulent claims were more likely among policyholders with short-term or recently issued policies."

Fraudulent Status with Habitat Status

A key observation from the figure 5 is that there is a significant difference between the percentages of fraudulent claims among people from uninhabitable houses and habitable houses. Chi-square test also revealed that 'Uninhabitable' is a significant predictor.

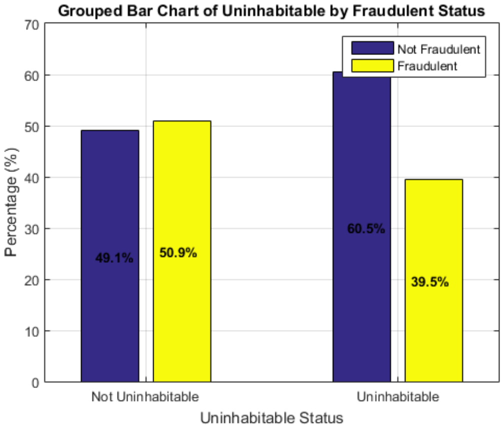


Figure 5: Grouped Bar graph of Habitat Status by Fraudulent Status

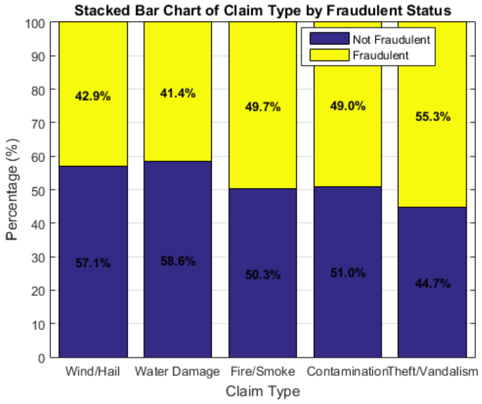


Figure 6: Grouped Bar graph of Claim Type by Fraudulent Status

Fraudulent Status with Claim Type

Figure 6 shows that Theft/Vandalism has the highest fraud percentage (55.3%) while natural disasters such as water damages and wind/hail have the lowest fraud percentages. Pearson's Chi-squared test also confirms that claim type is marginally significant.

Association between predictors

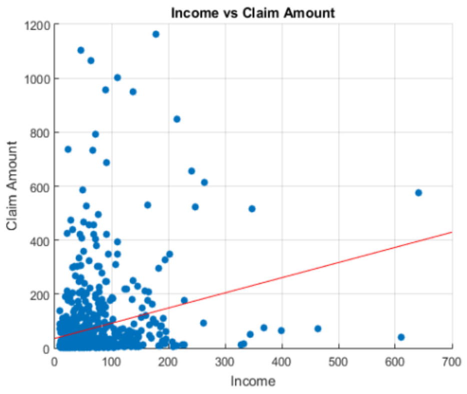


Figure 7: Scatterplot of income vs claim amount

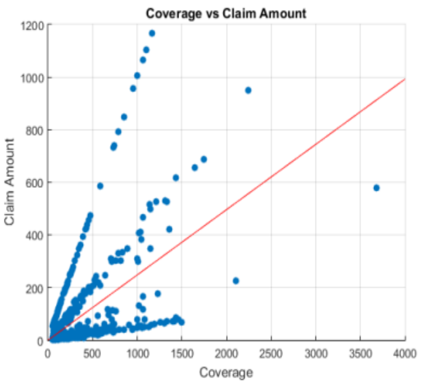


Figure 8: Scatterplot of claim amount vs coverage.

From the figures 7 and 8, it is clear that there is a positive correlation between Income and Coverage with Claim Amount. Figure also suggests the presence of clusters which will be further examined under Advanced Analysis.

Advanced Analysis

Classification Tree

Decision trees inherently perform a form of variable selection by identifying the most important features (predictors) that best split the data to maximize information gain or minimize impurity (e.g., Gini impurity or entropy). The features used at the top levels of the tree are generally the most significant in predicting the target variable.

To further refine the model and ensure it generalizes well to unseen data, tree pruning was applied. The pruning process helps balance the trade-off between model complexity and predictive accuracy by reducing the number of terminal nodes while maintaining performance.

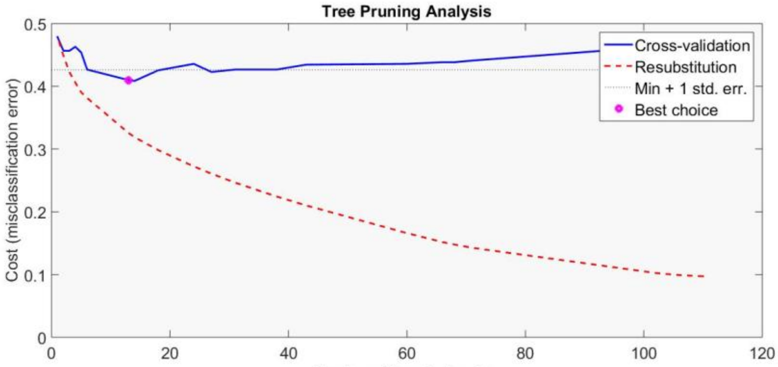
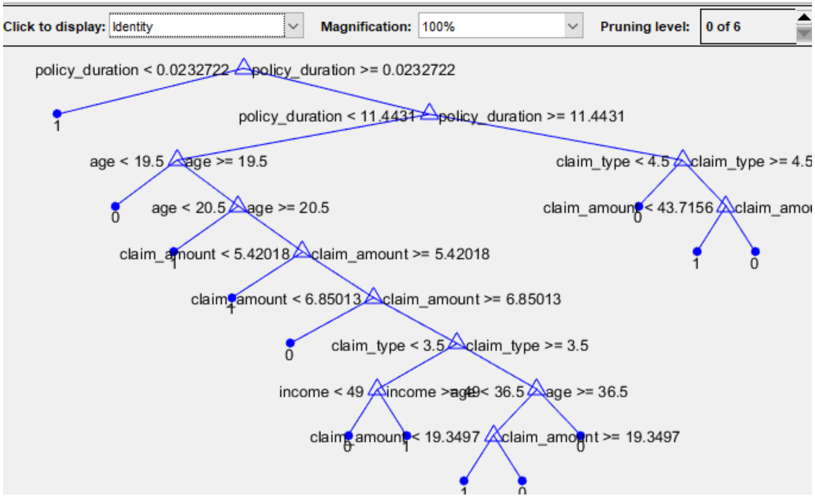


Figure 9: Pruning Analysis Tree



Click to display: Identity

Magnification: 100%

Pruning level: 0 of 6

Figure 10: Decision Tree

As shown by figure 9, cross-validation was used to evaluate the misclassification error across different levels of tree complexity. The graph illustrates the relationship between the number of terminal nodes and the error rate, highlighting how pruning can effectively reduce overfitting. The optimal tree size was determined using the "Min + 1 std. err." rule, which selects the simplest model within one standard error of the minimum error. This approach ensures that the pruned tree is both interpretable and robust, providing a reliable foundation for variable selection and predictive modeling. By leveraging pruning, the final decision tree model achieves a balance between simplicity and accuracy, making it a powerful tool for understanding the underlying structure of the data and identifying key predictors.

The final pruned decision tree, corresponding to level 15, is visualized in figure 10. The tree structure demonstrates how the model makes decisions based on key predictors such as policy duration, age, claim_type, claim amount, and income.

	Default Tree	**Optimal Tree**
Resubloss	0.097276	0.32555
Accuracy	47.94%	**56.06%**

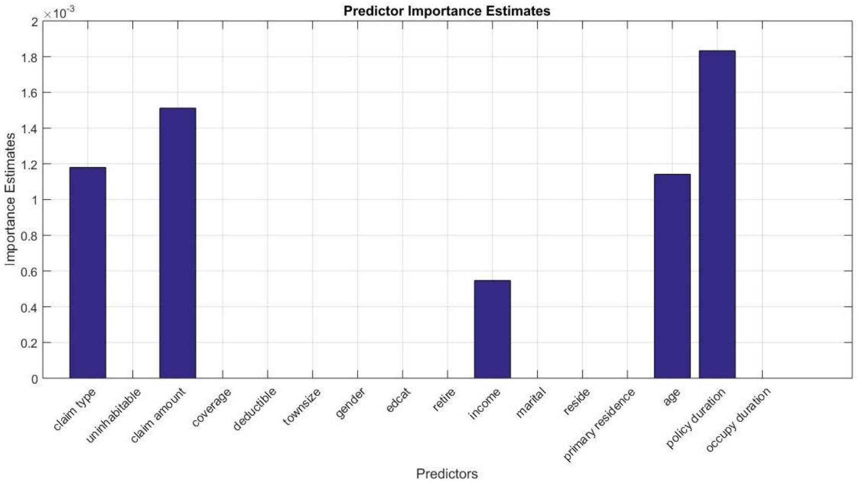


Figure 11: Bar Graph of Predictor Importance Estimates

Thus, by using the variable importance plot (figure 11) the important predictors were selected. In this dataset the most important variables are:

- Claim Type
- Claim Amount
- Inome
- Age
- Policy Duration

Random Forest

Random Forest Classification is an ensemble learning method that builds multiple decision trees during training and combines their outputs to make predictions. By utilizing randomly selected features and bootstrapped samples, it delivers robust and accurate results while minimizing overfitting. This technique is particularly effective for complex datasets, as it reduces variance and enhances generalization.

we employed Random Forest to achieve reliable and interpretable results. The model's inherent randomness in feature selection and data sampling ensures it captures diverse patterns, making it well-suited for our classification task.

Using **k-fold cross-validation**, we determined that **470 trees** provided the optimal balance between model complexity and predictive accuracy. This configuration achieved a **test accuracy of 58.33%**, with the following confusion matrix:

	Fraudulent claim	Non fraudulent claim
Fraudulent claim	66	32
Non fraudulent claim	48	46

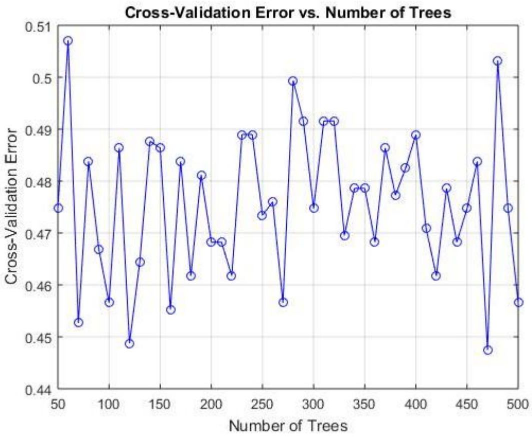


Figure 12: Line chart of cross validation Error over num of Trees

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that separates data points of different classes with the maximum margin. SVM is particularly effective for high-dimensional data and can handle both linear and non-linear decision boundaries using kernel functions (e.g., linear, RBF, polynomial). Its ability to generalize well and avoid overfitting makes it a popular choice for various machine learning applications.

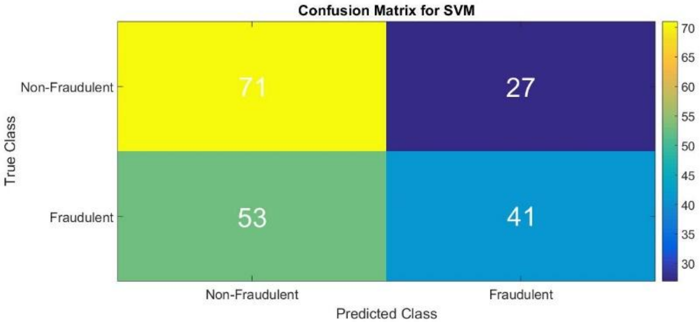


Figure 13: Confusion Matrix for SVM

The SVM model was trained and evaluated using a **5-fold cross-validation** approach to ensure robustness. The model achieved a **cross-validation accuracy of 57.20%**, indicating its ability to generalize well to unseen data during training. When tested on the independent test set, the model achieved a **test accuracy of 58.33%**, demonstrating consistent performance.

KNN Algorithm

k-Nearest Neighbours (kNN) is a simple algorithm used for classification and regression. It predicts the class or value of a test point by considering the k nearest training samples. The algorithm works by comparing distances between the test point and training data, making it effective for capturing local patterns. Its performance depends on the choice of k and the distance metric.

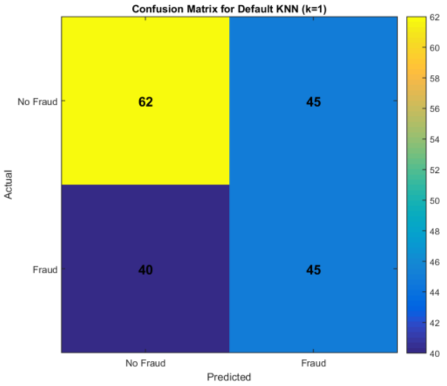


Figure 14: Confusion Matrix for Default KNN

Since kNN algorithm is sensitive to outliers, Box-Cox trasnformation was applied beforehand. The importance features identified from Classification tree were only considered. Categorical variables were handled with one-hot encoding.
The default kNN algorithm was applied with **k=1** and achieved a **test accuracy** of **55.73%.** Figure 14 shows the confusion matrix for the model.

Subsequently, **hyper parameter tuning** was performed with **10 fold cross validation** on the number of neighbours, k, ranging from 1 to 15. It was found that **optimal k = 15** with **cosine distance metric** yielded the best **test accuracy** of **60.42%.** Figure 15 shows the confusion matrix of the optimal model.

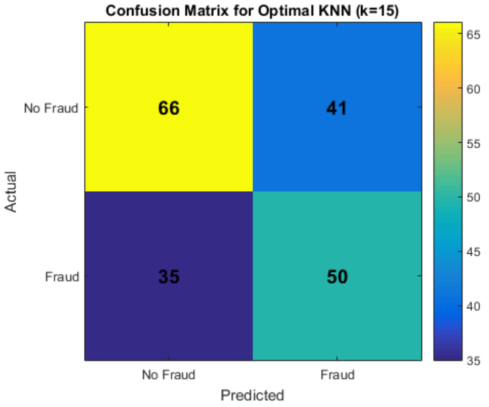


Figure 15: Confusion Matrix for Tuned KNN

Cluster Analysis

Dividing claims into similar groups can help find common patterns among them. This can also help identify groups that may have more fraudulent claims. To do this, we used cluster analysis to group the claims based on their similarities. The k-means clustering algorithm was selected for this study with the distance measure, city blocks. As shown in below figure it ended up with 2 clusters whose average silhouette value is maximum which is 0.4984. The clusters have been separated as the following figure of silhouette plot which shows how perfectly the clusters have been divided. There are some negative values, but considering other cluster numbers, this is
the optimal choice.

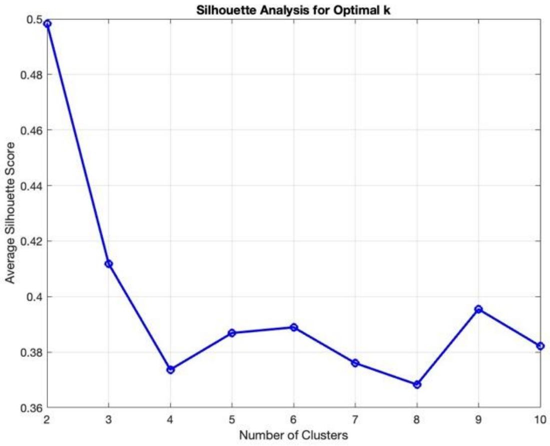


Figure 16: Silhouette Values

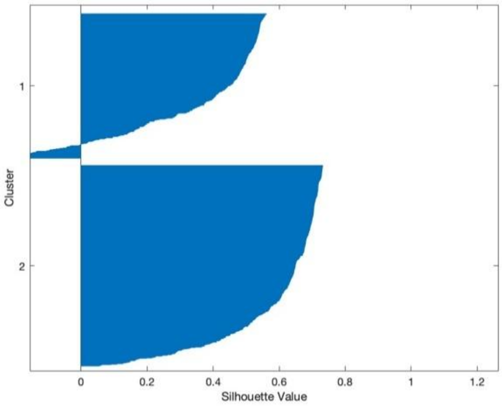


Figure 17: Silhouette Plot

By the clustering, from the training data set 448 observations were allocated to cluster 1 and 323 observations were allocated to cluster 2. Through the analysis it was identified that the cluster which has the highest chance of defaulting in the future is cluster 1.

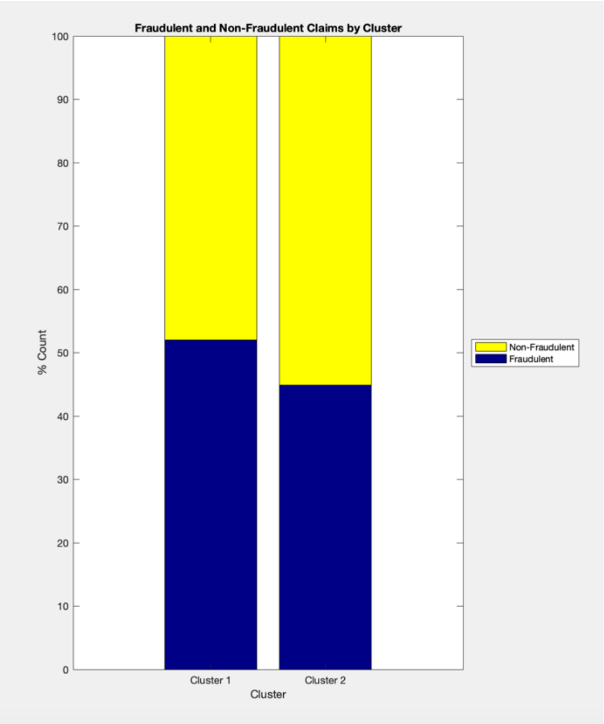


Figure 18: Fraudulent and Non-Fraudulent

After counting the fraudulent claims in each group, we found that Cluster 1 had the highest percentage of fraud. This suggests that this group may have more fraudulent claims. We then looked at the characteristics of the claims in Cluster 1 to find common patterns that could help identify fraud.

	Fraudulent	Non-fraudulent
Cluster 1	52.01%	47.99%
Cluster 2	44.89%	55.11%

More Findings on Important Variables

1. Claim Amount

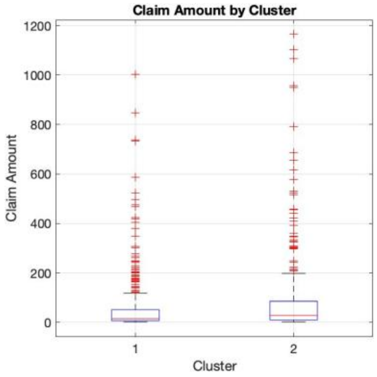


Figure 19: Box plot of claim amount

	Cluster 1	Cluster 2
Medians of Claim Amount	15,652	28,152

The boxplot shows Cluster 1 has more concentrated claim amounts (smaller box), while Cluster 2 has a wider distribution with more outliers at the higher end. This suggests Cluster 2 has higher-value claims that may need a closer look.

2.Income

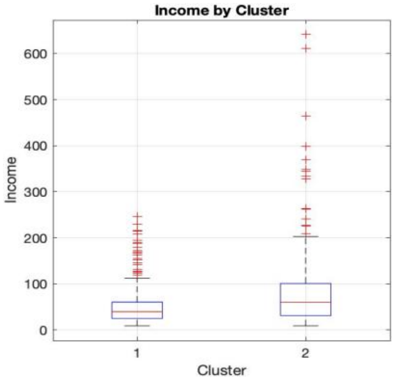


Figure 20: Box plot of Income

	Cluster 1	Cluster 2
Medians of Income	40	60

The income distribution in Cluster 1 is more compressed, while Cluster 2 shows greater variability and higher values. This indicates a potential relationship between income level and claim behaviour.

3. Age

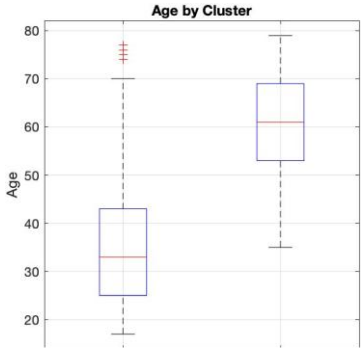


Figure 21: Box plot of Age

	Cluster 1	Cluster 2
Medians of Age	33	61

Cluster 1: Younger policyholders (median age 33)

Cluster 2: Older policyholders (median age 61)

The age difference between clusters is substantial, with Cluster 1 showing a clear concentration of younger individuals. This represents one of the most distinctive separating features between the clusters. \

4. Policy Duration

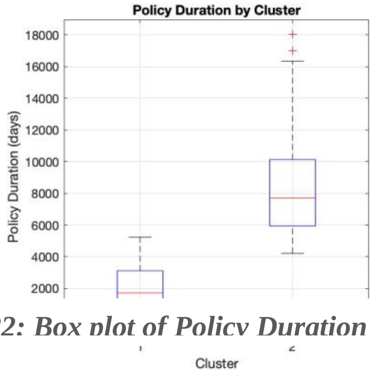
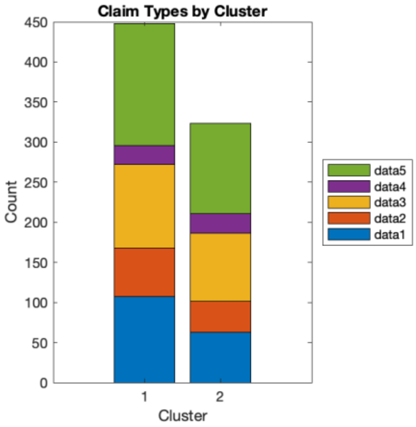


Figure 22: Box plot of Policy Duration

	Cluster 1	Cluster 2
Medians of Policy Duration	1714.5	7711

The dramatic difference suggests Cluster 2 represents long-term policyholders (over 21 years on average). Cluster 1 contains newer customers with less established relationships.

5. Claim Type



Cluster 1 has more total claims, the relative proportions of claim types are somewhat similar between clusters.
Type 5 (green) represents the largest proportion in both clusters.

Figure 23: Stacked bar charts for claim type

Conclusion

Descriptive analysis was conducted to examine the distribution of variables. Significant features associated with fraudulent claims were identified through correlation analysis and statistical tests. Classification trees were utilized to determine the most influential variables. Predictive models, including Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbours (kNN), were applied and compared for performance evaluation. Cluster analysis revealed two distinct groups, with one exhibiting a higher proportion of fraudulent claims. The findings highlight key risk factors that differentiate fraudulent and non-fraudulent claims. Model performance was assessed to determine the most effective approach for fraud detection. The results provide useful insights for improving fraud prevention in insurance.

Appendix

The MATLAB codes used in the project can be easily accessed through the following Google Drive Link:

https://drive.google.com/drive/folders/1UbtSnYiCvyER6UT5VwfbI5JBlFIDSX_g?usp=drive_link