MIS272 – Predictive Analytics

T2 2024

Assignment 1 – Individual
Student name: Thi Ngoc Hong Nguyen (Claire Nguyen)
Student number: 222052067

Executive summary (1 page)

DAX Compensation Lawyers are having trouble recognising false personal injury insurance claims. The company provides a data set of over 3000 claims to seek assistance in detecting fraudulent claims, as well as strategic advices to maintain financial health and operational efficiency.

Objectives/Anticipated Advantages:

- Enhance efficiency and accuracy of fraud detection
- Reduce financial losses by optimising claim investigation resource allocation
- Assist data-driven decision-making

Recommendations:

- The company should employ the best predictive model for claim screening and update it periodically with new data to assure accuracy: Cross-validated decision trees have the highest kappa (0.186) and accuracy (96.83%).
- The company should automate the first claim screening process to detect high-risk claims bassed on visualisation observations: claimants suffering from back issues or head injuries (Figure 3), claims with nature of injury is sprain/strain or confusion (Figure 4), claims related to lifting and striking objects as the cause of injury (Figure 5), and claimants with age range 31-35 (Figure 8). These aspects might highly involved in fraudulent claims.

Data exploration, pattern discovery, and preparation (2 pages)

Preparation:

- With over 3000 claims, a robust dataset must be maintained while minimising the loss of information. Therefore, our approach for replacing missing values guarantees that the dataset is as comprehensive as feasible
- We replace missing values with specific values, modes, or averages ensures dataset consistency:

- Categorical attributes:
+) Adjustor Notes, Claimant Marital Status, Nature of Injury, and Cause of Injury: replace with “Unknown”
+) Fraud Flag and Vehicle Flag: replace with mode value determined by the Aggregate operator.

- Numerical attributes (Claimant Age): replace with average age to maintain central trend without bias.

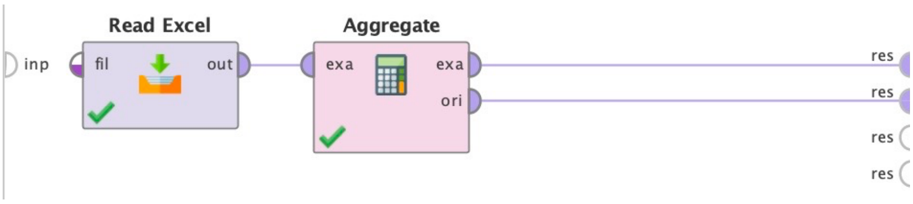
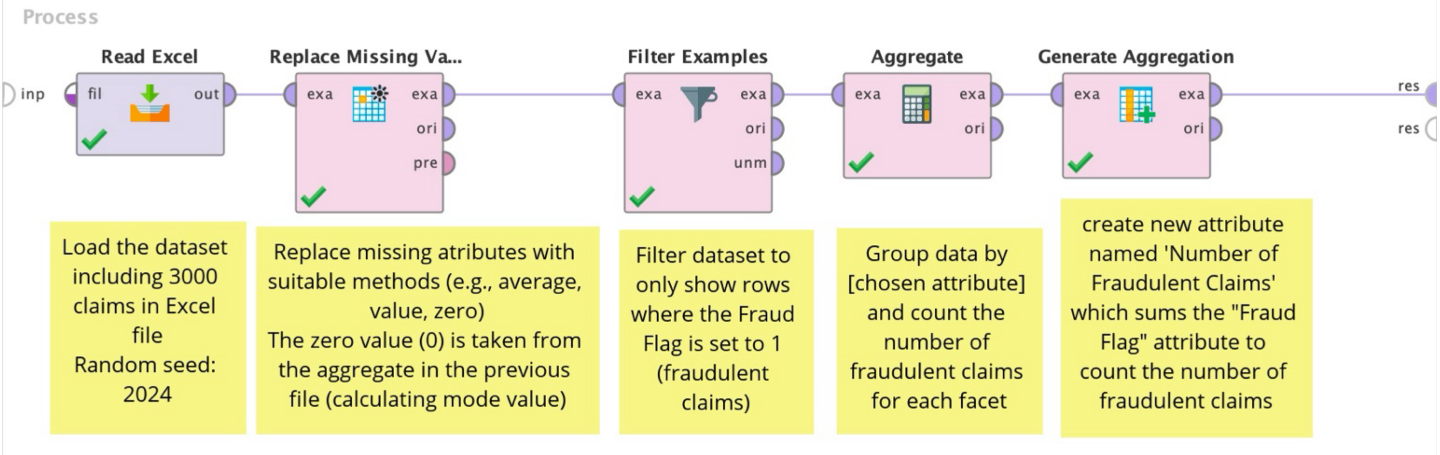


Figure 1: Preparation: Mode calculation



Load the dataset including 3000 claims in Excel file
Random seed: 2024

Replace missing atributes with suitable methods (e.g., average, value, zero)
The zero value (0) is taken from the aggregate in the previous file (calculating mode value)

Filter dataset to only show rows where the Fraud Flag is set to 1 (fraudulent claims)

Group data by [chosen attribute] and count the number of fraudulent claims for each facet

create new attribute named 'Number of Fraudulent Claims' which sums the "Fraud Flag" attribute to count the number of fraudulent claims

Figure 2: Process for visualisation

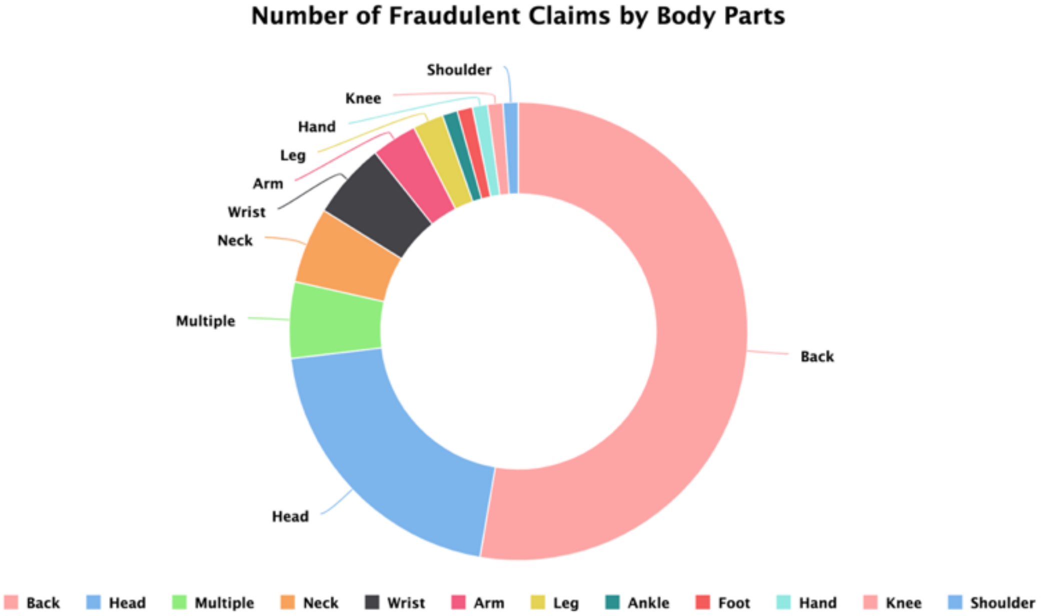
<table>
<tr><th>Visualisation</th><th>Interpretation</th></tr>
<tr><td>


Figure 3: Number of Fraudulent Claims by Body Parts
</td><td>Claimants suffering from back issues might have higher fraudulent claims, followed by head-problem claimants.</td></tr>
</table>

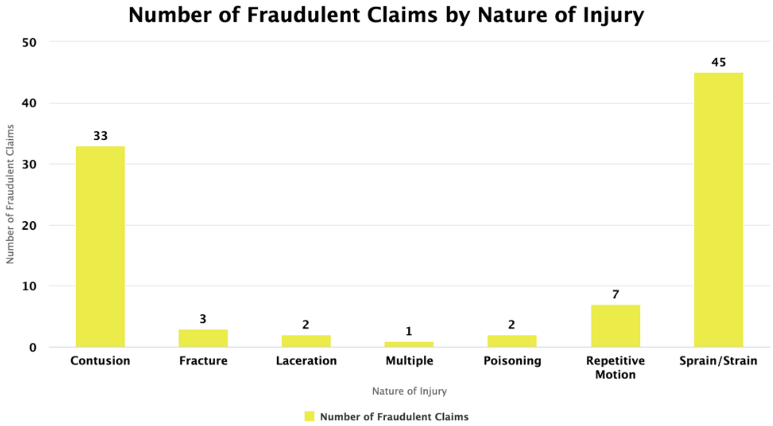


Figure 4: Number of Fraudulent Claims by Nature of Injury

Claimants with injuries such as sprains or strains most frequently file fraudulent claims, followed by confusion.

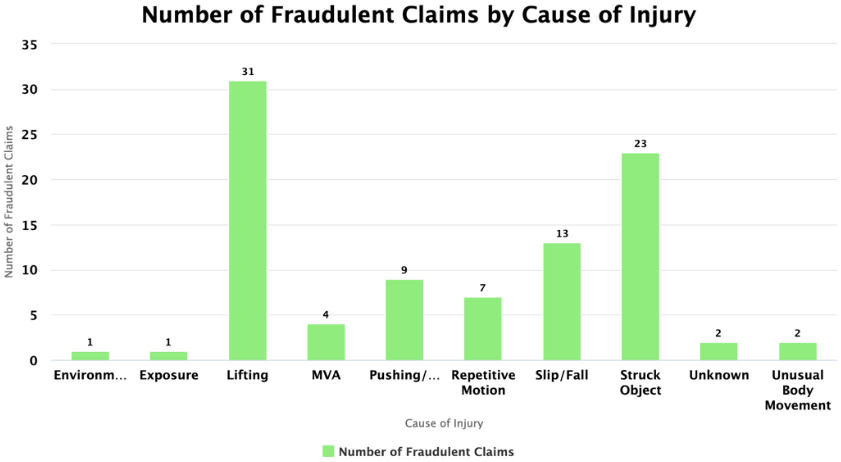
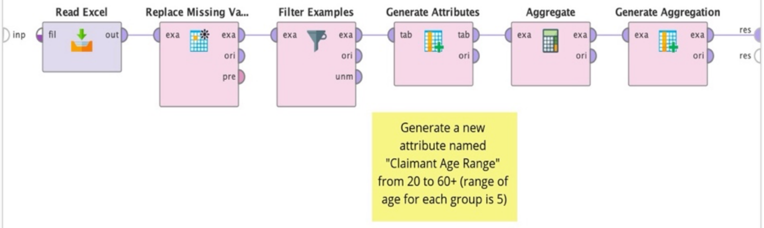


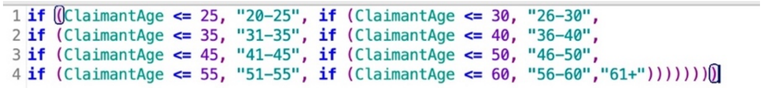
Figure 5: Number of Fraudulent Claims by Cause of Injury

Regarding the cause of injury, claimants who have lifted and struck objects due to health issues may have a higher likelihood of fraudulent claims.



Generate a new attribute named "Claimant Age Range" from 20 to 60+ (range of age for each group is 5)

Figure 6: Process



```
if (ClaimantAge <= 25, "20-25", if (ClaimantAge <= 30, "26-30",
if (ClaimantAge <= 35, "31-35", if (ClaimantAge <= 40, "36-40",
if (ClaimantAge <= 45, "41-45", if (ClaimantAge <= 50, "46-50",
if (ClaimantAge <= 55, "51-55", if (ClaimantAge <= 60, "56-60","61+"))))))))
```

Figure 7: Function expression for Age Range

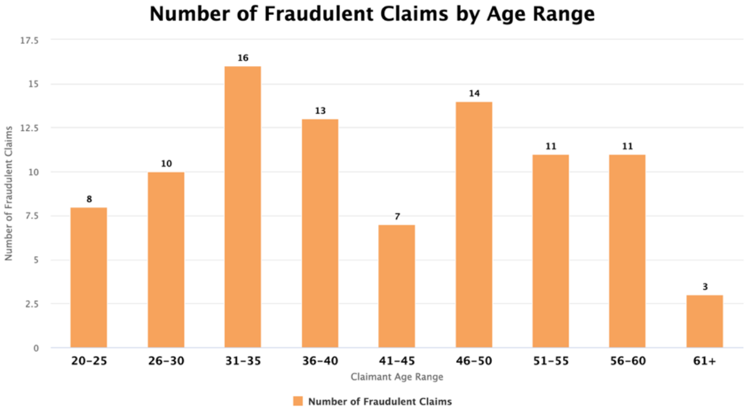


Figure 8: Number of Fraudulent Claims by Age range

We divide the claimant age into different age ranges (e.g., 20-25, 26-30, 31-35, etc.).

Age range 31-35 has the greatest number of fraudulent claims, followed by 46-50 and 36-40.

Predictive modelling (2 pages)

- **Data Preparation:**

- Filter Examples operator: filter through all missing attributes
- Numerical to Binominal: transform Fraud Flag (1=Yes 0=No) from numerical into binominal where “0” might represent “No” and “1” represents “Yes”
- Select Attributes: exlude Witness Present. After incorporating all attributes, the decision tree over-relied on the binary "Witness Present" attribute, resulting in an underperforming tree. Thus, removing this attribute encouraged the model to use more informative attributes and creating a deeper and more accurate decision tree.
- Set Role: sets “Fraud Flag” attribute as the label for the model to anticipate fraudulent claims.
- Split Data: splits data into two subsets: 70% data used for training and 30% for testing sets.

1. Decision Tree – Partial

Set maximum depth to 5 to avoid overfitting and ensure model generalization. This balance might prevents the model from becoming overly training data-specific.

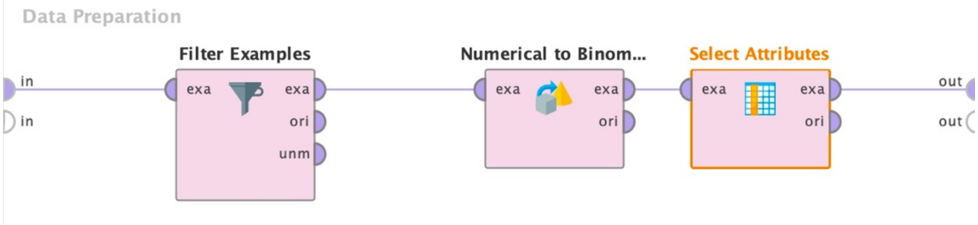


Figure 9: Operators inside Subprocess: Data Preparation (in all models)

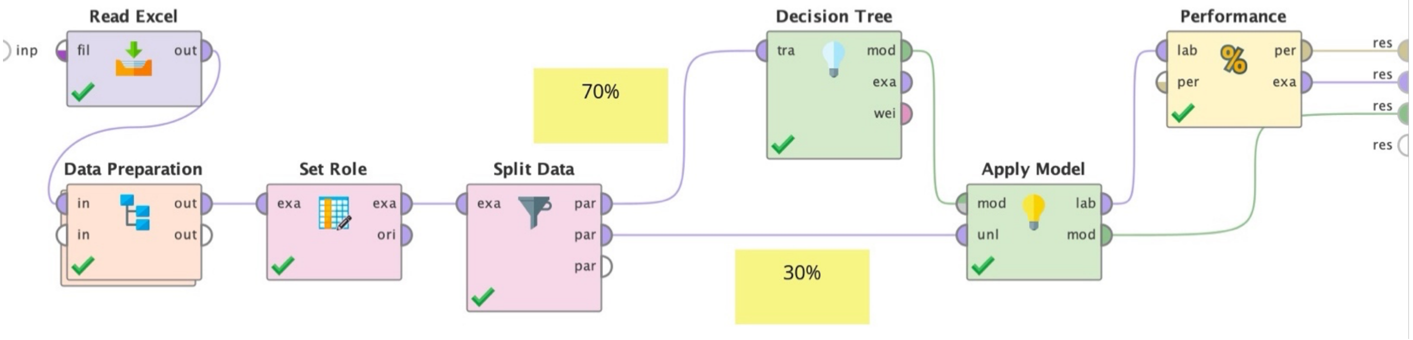


Figure 10: Decision Tree – Partial Model

2. k-NN Partial Model

Set k to 5 in k-NN to balance bias, variance, and capture patterns without being overly sensitive to noise.

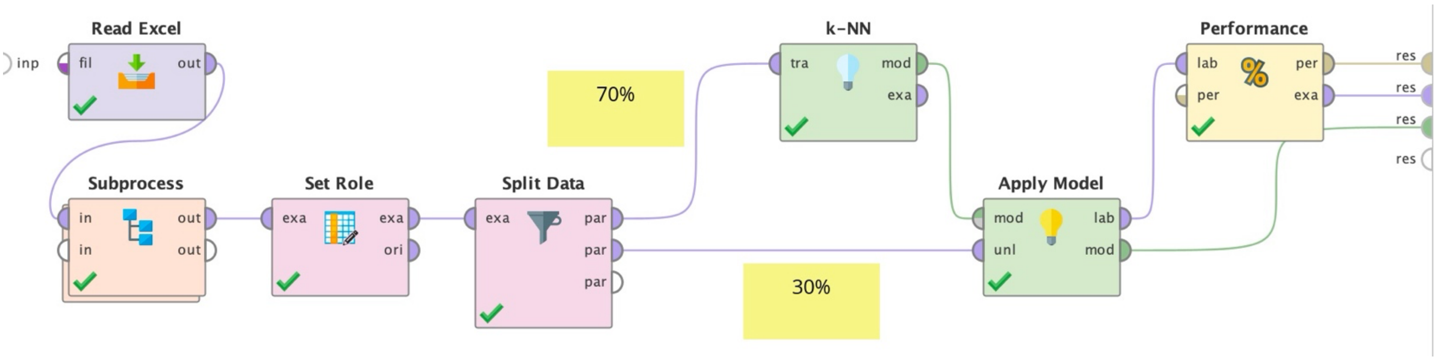


Figure 11: k-NN Partial Process

3. Decision Tree Cross Validation Model & k-NN Cross Validation Model

Both models use the same process but have different Cross Validation operators (Decision Tree and k-NN). The Performance operator computes accuracy and kappa to evaluate the model's fraud prediction ability. Cross Validation uses 10 folds to balance bias and variance and estimate the model's performance on unknown data, even with 3000 claims.

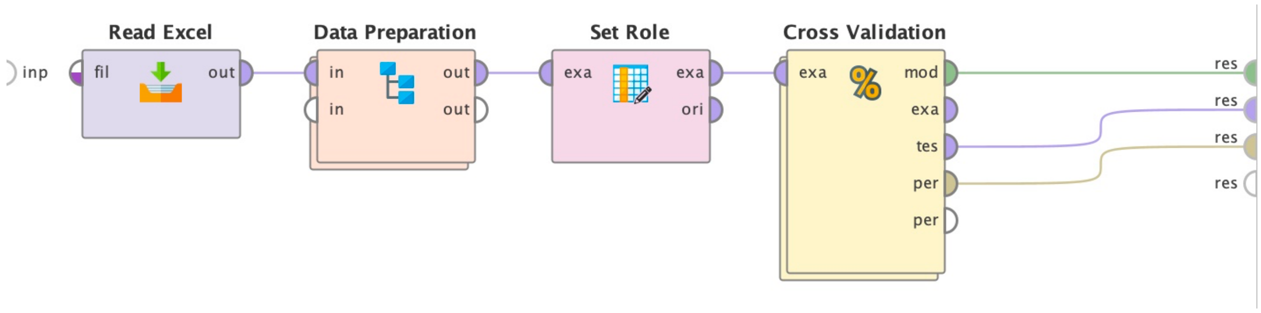


Figure 12: Decision Tree/k-NN – Cross Validation Process

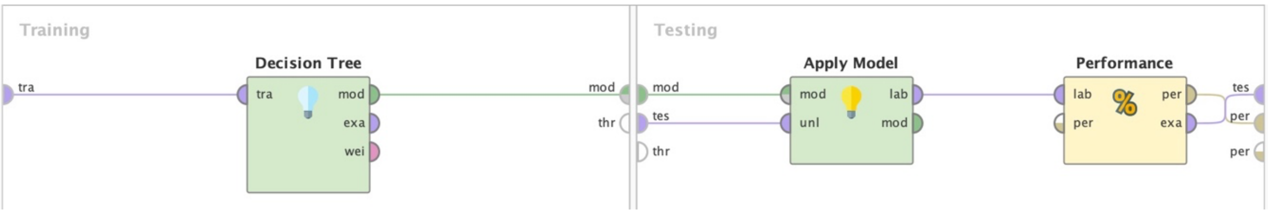


Figure 13: Operators inside Cross Validation (Decision Tree)

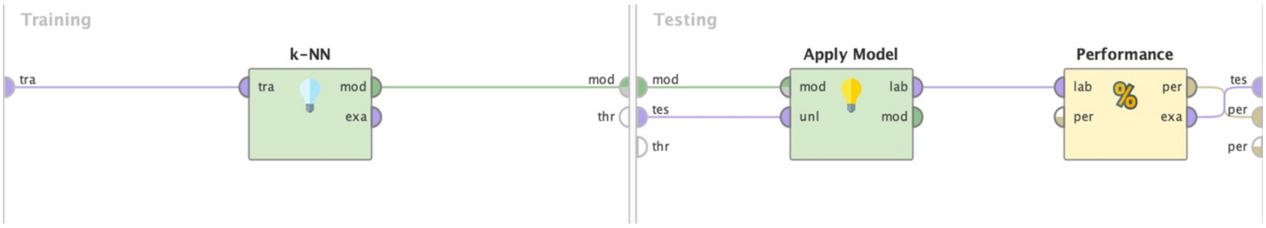


Figure 14: Operators inside Cross Validation (k-NN)

4. Ensembles with Decision Tree, k-NN, and Naïve Bayes Model

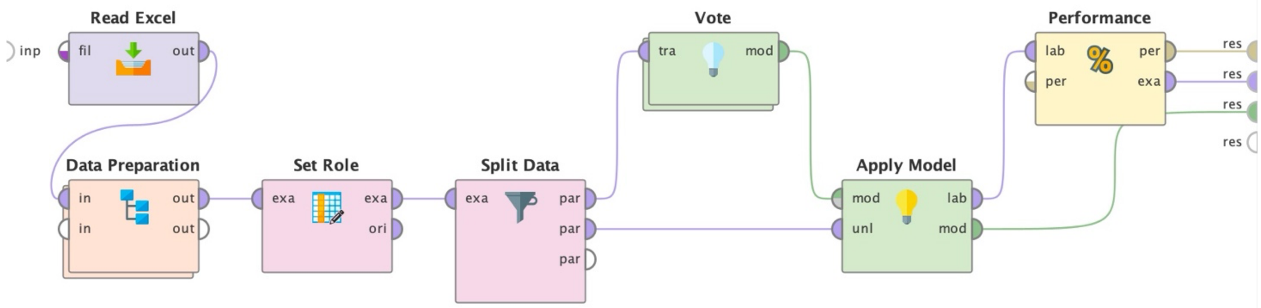


Figure 15: Ensembles with Decision Tree, k-NN, and Naïve Bayes process

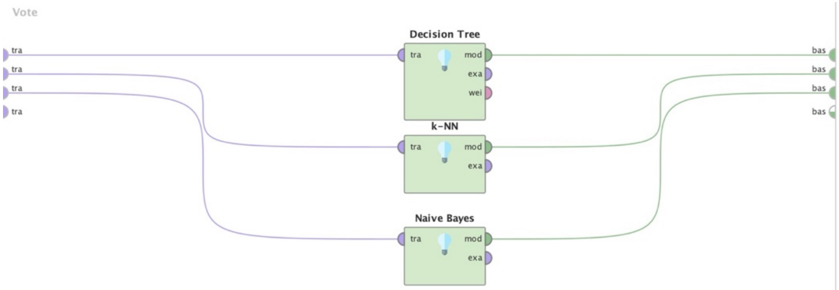


Figure 16: Process in the Vote operator

Model evaluation and improvement (1 page)

RESULT TABLE			
Predictive Modelling	**Parameter**	**Accuracy**	**Kappa**
Decision Tree - partial	Maximum depth = 5	96.47%	0.143
K-NN – partial	k = 5	96.91%	-0.002
Decision Tree – Cross Validation	Number of folds = 10 Maximum depth = 5	***96.83%***	***0.186***
K-NN – Cross Validation	Number of folds = 10 k = 5	96.76%	-0.004
Ensembles (K-NN, Decision Tree, Naive Bay)	Maximum depth = 5 K = 5	96.91%	0.166

Performance Improvements

Kappa increases from 0.143 to 0.186 shows cross-validation improves the decision tree model. The ensemble model also demonstrated improved performance by combining multiple models

Conclusion

When evaluating predictive models, although it is crucial to consider both accuracy and Kappa, the priority depends on the dataset. The dataset has fraudulent claims much fewer than non-fraudulent ones, which results in class imbalance. Besides, kappa can accurately distinguish between fraudulent and non-fraudulent claims; therefore, it becomes a more critical metric than accuracy for evaluating model performance.

Therefore, while ensemble and k-NN models obtained higher accuracy, Decision Tree with cross-validation gives the optimum balance of accuracy and higher kappa, it is the preferred model, indicating better performance in identifying fraudulent claims.