**AutoInsurance Fraud Detection using Machine Learning**

**<Student Name>**

**<Student email>**

**University of North Texas**

**Professor: <Professor Name>**

**Introduction**

According to the *Federal Bureau of Investigation*, non-health insurance costs an estimated loss of $40 Billion US dollars per year, or about $400 to $700 per U.S family in annually increased premiums. Such numbers of currency represent money that actually goes to fraud from the pockets of innocent people, who seek safety and financial help at time of crisis.

Automobile insurance fraud is a frequently occurring fraudulent activity; according to a *study by Versisk,* auto insurers take a loss of at least $29 Billion US dollars per year! This total is a direct result of fraudulent activities that involve unrecognized drivers ($10.3 Billion), underestimated mileage ($5.4 Billion), violations/accidents ($3.4 Billion), and false garaging to lower premiums ($2.9 Billion).

All of these add up to the direct increase in premiums for other policy holders. This problem is usually addressed with a combination of various methods and techniques, in order to determine any fraudulent activity that might have taken place, such as:

* *Claims Investigation*: The insurance companies usually have teams of investigators who review and investigate the claim on ground. These teams look for inconsistencies within a filed claim, by interviewing witnesses, examining the vehicles and site at which the incident had occurred, additionally reviewing the police reports and performing extended research on their own.
* *Past Claims and Information Sharing*: Insurance companies often share information about known or suspected fraudsters identified by law enforcement agencies, such as the National Insurance Crime Bureau’s (NCIB) datasets in the United States.
* *Auditing and Review*: Insuring companies also conduct regular audits and additional reviews of claims, and underwriting processes to identify potential frauds and risks.

These methods and techniques are both expensive to conduct and time consuming which requires cooperation between various parties involved. This makes it less than ideal in order to potentially detect an insurance fraud sooner rather than later. Using Machine Learning Algorithms, I intend to train a model that can identify inconsistencies for a filed claim, when this is paired with the above activities the overall time that might be taken to detect and report a fraud will be reduced significantly.

The proposed idea is to have a Machine Learning model: Decision Tree Learning Algorithm, which undergoes supervised learning over the dataset of previously filed claims, and has a target class of *fraud\_detected*. This simple solution can be used by insurance companies to work on a statistical estimate and proceed with the necessary activities to verify this hypothesis of fraud, which is a result of my model. Rather treating every case the same way which would be a waste of both time and resources.

**Literature Review**

**2.1 Use of Optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection**

This is a research paper, published by Authors: *Sharmila Subudhi* and *Suvasini Panigrahi*, which aims to provide a hybrid approach to detect fraud in automobile insurance claims using a combination of Genetic Algorithm based Fuzzy C-means clustering and supervised classifier models. They first took a portion of the insurance data as a test set, leaving the rest as a training set. They applied the FCM clustering to the training set to create groups of similar data. Then, they used these groups to classify the test set data into three categories: genuine, malicious, and suspicious.

They discarded the genuine and malicious data and focused on the suspicious cases. These cases were analyzed using four different classifiers: Decision Tree, Support Vector Machine, Group Method of Data Handling, and Multi-Layer Perceptron. They used a technique called 10-fold cross-validation to train and validate their models.

By conducting experiments on a real-world car insurance dataset, they demonstrated that their proposed method effectively detects automobile insurance fraud.

**2.2 Detecting insurance claims fraud using machine learning techniques**

This research paper, is a good reference to the work that I intend to do, this paper authored by *Riya Roy*, and *Thomas George K.* aims to enhance the detection process by utilizing machine learning techniques. Machine learning allows computers to learn from data and make predictions or decisions without being explicitly programmed to do so. This approach can potentially improve the accuracy and efficiency of detecting fraudulent claims*.*

The authors compare the performance of various machine learning techniques such as Decision Tree and Random Forest by calculating a confusion matrix, which is a table used to describe the performance of a classification model on a dataset for which the true values are known. The confusion matrix helps determine important evaluation metrics such as accuracy, precision, and recall. The findings may contribute to improving the detection process and reducing the financial impact of fraudulent claims on the insurance industry.

However, I intend to determine hypermeters that are best suited for Decision Tree and other learning algorithms, also draw contrast among these to identify the best possible solution in order to classify the label “fraud”.

**2.3 A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification**

This paper authored by: *Yaqi Li, Chun Yan, Wei Liu,* and *Maozhen Li*  presents a new and improved method for detecting car insurance fraud by combining three techniques: Random Forest, Principle Component Analysis (PCA), and Potential Nearest Neighbor. The goal is to make fraud detection more accurate and reduce errors. The method is tested against other similar techniques and shows better results. The authors also claim that it is also applied to real car insurance fraud cases to find fraud patterns.

The authors use PCA to change the data at each step when deciding how to split it, which makes the trees in the Random Forest more diverse and improves accuracy. The relationship between Random Forest and Adaptive Nearest Neighbors is studied, and a new voting method based on Potential Nearest Neighbor is proposed to replace the traditional majority vote.

They tried to test this new method by using 12 different data sets from various fields. And then compared these results to those generated by other methods like Oblique Decision Tree Ensemble, Rotation Forest, and basic Random Forest. The authors, conclude that the proposed new method is less prone to errors, and that it is a newer approach for detecting automobile insurance fraud, without losing sight of important aspects of the case.

**References**

Federal Bureau of Investigation. Insurance fraud. Retrieved from <https://www.fbi.gov/stats-services/publications/insurance-fraud>

Insurance Information Institute. (2022, August 1). Background on: Insurance fraud. Retrieved from <https://www.iii.org/article/background-on-insurance-fraud>

Verisk Insurance Solutions. (2017). The challenge of auto insurance premium leakage: Mispriced risk is a $29 billion annual problem for the industry. Retrieved from <https://www.verisk.com/siteassets/media/campaigns/gated/underwriting/verisk-the-challenge-of-auto-insurance-premium-leakage.pdf?__FormGuid=8c509869-699d-4698-9ac3-9ada3d271c97&__FormLanguage=en-US&__FormSubmissionId=65299f67-8c20-408c-97bd-221fc1cee1bf>

Subudhi, S., & Panigrahi, S. (2017). Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University - Computer and Information Sciences*. <https://doi.org/10.1016/j.jksuci.2017.09.010>

Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques*. International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. Kollam, India. <https://doi.org/10.1109/ICCPCT.2017.8074258>

Li, Y., Yan, C., Liu, W., & Li, M. (2018). A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. *Applied Soft Computing*, 70, 1000-1009. <https://doi.org/10.1016/j.asoc.2017.07.027>