Center for Open Access in Science ▪ https://www.centerprode.com/ojit.html
Open Journal for Information Technology, 2023, *6*(2), 143-156.
ISSN (Online) 2620-0627 ▪ https://doi.org/10.32591/coas.ojit.0602.05143m

A Hybrid Model for Detecting Insurance Fraud Using K-Means and Support Vector Machine Algorithms

Brian Ndirangu Muthura & Abraham Matheka

Kenyatta University, School of Engineering and Technology, Nairobi, KENYA

Received: 8 August 2023 ▪ Revised: 4 October 2023 ▪ Accepted: 15 November 2023

Abstract

Private stakeholders and governments across the globe are striving to improve the quality and access of healthcare services to citizens. The need to improve healthcare services, coupled with the increase in social awareness and improvement of people's living standards, has seen an increase in medical policyholders in the insurance industry. Even so, the healthcare sector is grappled with increased costs every other year, leading to revision of premiums and increased costs for the policyholders. One of the main factors contributing to the increased costs is fraudulent claims raised by the service providers and the policyholders, leading to unprecedented risks and losses for insurance firms. The insurance industry has set up fraud detection and mitigation systems to mitigate losses brought about by fraudulent claims, which come in two flavors: rule-based systems and expert claims analysis. With rule-based systems, conditions such as missing details, location of the claim vis a vis the location of the policyholder, among other rules, are evaluated by systems to assess the validity of the claims. On the other hand, insurance firms rely on the human intervention of experts using statistical analyses and artificial rules to detect fraudulent claims. The rule-based and expert analysis methods fail to detect patterns or anomalies in claims, which is central to efficient fraud detection. Data mining and machine learning techniques are being leveraged to detect fraud. This automation presents enormous opportunities for identifying hidden patterns for further analysis by insurance firms. This research aims to analyze a hybrid approach to detect medical insurance fraud using both K-Means (unsupervised) and Support Vector Machines (supervised) machine learning algorithms.

Keywords: fraud detection, machine learning, K-Means, support vector machines, hybrid algorithms.

1. Introduction

Insurance fraud is second to tax fraud in the frequency of occurrence (Association of Certified Fraud Examiners, 2019). The nature of the insurance business makes it susceptible to fraud. Insurance firms mainly manage risk for the policyholders by pooling and generating large cashflows through insurance premiums to pay loss claims. Insurance fraud occurs when the insured attempts to profit from the insurer while failing to comply with the policy's contractual terms and conditions, creating damage and losses for the insurer, and can occur at any stage of the policy term (Association of Certified Fraud Examiners, 2019). The losses span the long-term (comprised of life insurance) and short-term (comprised of motor and health insurance) insurance policies.

Correspondence: Brian Ndirangu Muthura, Kenyatta University, School of Engineering and Technology, Nairobi, KENYA.

The prevalence of insurance fraud is not localized in one country but spread globally. For instance, the Coalition Against Insurance Fraud estimates that over $80 billion is lost yearly due to insurance fraud in the United States of America. The Association of British Insurance recorded 107 000 fraudulent insurance claims in the United Kingdom in 2019 worth over £1.2 billion and depicted a 5% increase from 2018 (Association of Kenya Insurers, 2021). The national health fund in France (CNAM) estimated that $321.4 million was lost due to fraudulent schemes and claims. The inquiries revealed that health providers such as doctors and practitioners accounted for the highest percentage of fraudulent claims, 48%.

On the other hand, health institutions such as hospitals and clinics accounted for 31% of the fraudulent claims against 21% by the insured. In India, the estimated losses attributed to fraud amount to $6 billion annually, close to 8.5% of the total premiums remitted. The South African Insurance Crime Bureau estimated that out of $2.4 billion in insurance claims paid in 2019, $497.86 million could have been for fraudulent claims, which account for about 20% of the total claims raised in the year. In Kenya, the Insurance Fraud Investigation Unit identified 83 insurance fraud cases worth close to KES 386.34 million (Association of Kenya Insurers, 2021). The increase in insurance fraud, coupled with colossal sums of money involved, has led to an increase in the cost of insurance. Moreover, these figures are only estimates of claims deemed to be fraudulent and may not necessarily represent the precise magnitude of losses incurred.

Fraud in the insurance industry directly impacts a company's bottom line (Association of Kenya Insurers, 2021). Over 5% of a company's revenue is estimated to be lost to fraud yearly (ACFE, 2019). Through proper fraud detection mechanisms and validation of claims, insurance firms stand to benefit from increased profitability. Apart from the loss of revenue for the insurance firms, fraud schemes lead to the loss of the reputation of the insurers (Association of Certified Fraud Examiners, 2019). Insurance fraud is a global issue affecting the economy, state, community, and individuals.

Detecting and preventing fraud is a critical concern in the insurance industry (Matloob & Khan, 2019). While expert experience is critical in determining if a claim is fraudulent, the number of claims significantly raised surpasses the few experts tasked with analyzing these claims making it challenging to examine all insurance claims in real-time (Hanafy & Ming, 2021). Furthermore, differing experiences and perspectives from experts while dealing with the same claim cases contribute to decision bias. In the medical insurance industry, fraud detection has shifted from traditional domain expert analysis to rule-based systems (Gupta et al., 2021). The rule-based system contains sets of conditions that evaluate the validity of a claim which is an improvement from the domain expert analysis as the throughput of claims analyzed is much higher. However, there is an underlying need for even more efficiency in detecting insurance fraud in medical insurance (Matloob & Khan, 2019).

Researchers have proposed machine learning as a sophisticated technology that can be harnessed to assess claim patterns in medical insurance claims (Matloob & Khan, 2019). Machine learning can be applied to large datasets to discover unknown patterns and predict outcomes vital in fraud detection (Rawte & Anuradha, 2015). In medical insurance fraud detection, supervised learning is used to solve the classification problem into predefined labels (fraudulent and legitimate claims). In contrast, unsupervised machine learning algorithms address the clustering problem mainly through outlier detection (Rawte & Anuradha, 2015).

2. Problem statement

The implementation of robust fraud detection mechanisms is critical for medical insurance firms in the fight against fraudulent practices in the industry. A practical approach to fraud detection presents the opportunity of mitigating the losses attributed to fraudulent claims.

Subsequently, this presents an opportunity to reduce the cost of private medical insurance for the policyholders and increase profitability for the firms (Hanafy & Ming, 2021). Insurance firms have relied on domain expert analysis to detect fraud despite the benefits of a robust fraud detection technique. More recently, this approach has been automated by rule-based systems, which evaluate the validity of claims based on a set of rules as defined by domain experts. However, these attempts are ineffective in addressing fraud detection in the healthcare industry (Gupta et al., 2021).

Recent studies in fraud detection using machine learning and data mining techniques have focused on the efficacy of exclusively implementing supervised or unsupervised models. The supervised algorithms are mainly used to classify claims based on predefined labels, genuine and fraudulent claims. Similarly, unsupervised algorithms are used in clustering and outlier detections and do not need predefined labels. The combination of both supervised and unsupervised machine learning algorithms is complementary. Supervised algorithms learn from past fraudulent patterns, while unsupervised techniques target detecting new fraud patterns (Carcillo et al., 2021). The hybrid learning approach to fraud detection combines the advantages of supervised and unsupervised learning algorithms while reducing the inherent risk associated with either algorithm (Bauder et al., 2017). There is minimal research on using hybrid machine learning algorithms in fraud detection, more so with the combination of SVM and K-Means to analyze and solve classification and clustering problems, respectively.

3. Literature review

The chapter discusses the types of fraud in the medical insurance industry, the advancement of fraud detection techniques, and the use of supervised, unsupervised, and hybrid machine learning algorithms in fraud detection and prevention.

4. Types of medical insurance fraud

Healthcare fraud can be categorized based on the servicing pattern, that is, service-availing and service-providing patterns. The service-availing patterns are defined as fraudulent activities undertaken by the insured, while the service-providing patterns refer to the misrepresentation by the medical professionals (Matloob et al., 2020).

Healthcare fraud can also be categorized into service provider fraud, insurance subscriber fraud, insurance provider fraud, and conspiracy fraud. Service provider fraud may consist of charges incurred for medical services not performed, overbilling by the service provider, unbundling one medical procedure to multiple treatment stages, and billing each stage separately rather than consolidating, falsifying patients' diagnosis, and treatment history to validate superfluous medical procedures. Policy subscribers may commit fraud by filing claims for not receiving medical services, falsifying onboarding details to obtain a lower premium rate, and illegally claiming insurance benefits using another policyholder's coverage. On the other hand, insurance providers may commit fraud if they misrepresent the benefit offered for a particular scheme or product or make fake reimbursements to the service providers and policyholders. In conspiracy fraud, a combination of more than one of the tripartite parties is involved in getting undue benefits (Waghade & Karandikar, 2018).

5. Health care fraud detection methods

Fraud detection in medical insurance has evolved over the years. Traditional fraud detection techniques relied on rule-based methods. Claims would be evaluated for fraud based on

the rules outlined by domain experts (Zhou & Zhang, 2020). The efficacy of the rule-based evaluation method was constrained by the correctness of the rules (Zhou, He, Yang, Chen & Zhang, 2020). Traditional rule-based fraud detection methods relied on a few auditors to handle thousands of claims (Waghade & Karandikar, 2018). Only experienced auditors were able to uncover fraudulent claims. This approach was inefficient and time-consuming.

Electronic claim management systems have recently been implemented and integrated with healthcare systems (Kose, Gokturk & Kilic, 2015). Claim management systems are increasingly harnessed for auditing, review, and automatic claim processing. Electronic Claim Processing systems offer more efficiency and higher claim analysis throughput than traditional expert domain analysis (Ai, Lieberthal, Skyla & Wojciechowski, 2018).

The advancements in artificial intelligence, machine learning, and deep learning have resulted in new automated fraud detection methods with data mining and regression as the main approaches in medical insurance fraud detection (Joudaki et al., 2015).

6. Related works

Segal (2016) provides an introductory analysis of the capabilities of data mining techniques in pattern identification and matching in large data sets. Data mining can provide insights and trends in the underlying models. Moreover, data mining tools can detect anomalies and outliers by comparing them with known models and profiles.

In their study, Bauder et al. (2017) evaluate the use of data mining techniques in fraud detection and prevention by medical insurance firms. They posit that data mining in healthcare fraud detection involves structured and unstructured data. Structured data is a standardized data format that can be stored in tabular format in conventional databases. On the other hand, unstructured data is unformatted and unorganized data. Structured data is easy to analyze and model using data mining algorithms, while unstructured data requires additional steps such as parsing to analyze.

Zhou and Zhang (2020) explore in detail how data mining is used for fraud detection in the healthcare industry. They assert that fraud in medical treatment can be detected by analyzing abnormal data records and converting the fraud detection question into the classical outlier detection problem. The outlier detection method in data mining can further be divided into statistic-based, distance, clustering, and classification. The classification anomaly detection problem divides datasets into normal and abnormal types. In this outlier detection approach, the labelled data is applied for training and converts anomaly detection into a two-classification problem. Distance-based anomaly detection evaluates the range between the data sets (Zhou & Zhang, 2020). The local outlier factor (LOF) can be applied to each dataset to gauge the distance for each dataset. A large LOF value increases the likelihood of the dataset being an outlier. The statistical-based anomaly detection method assumes that outlier datasets do not conform to the model's distribution law of normal data. In the cluster-based approach, normal points often belong to clusters with multiple data points, while outliers belong to clusters with fewer or no data points (Zhou & Zhang, 2020).

Lawand and Kulkarni (2019) analyze insurance fraud prediction by solving the classification problem of the input space. Their research harnesses the Random Forest, decision tree, Naïve Bayesian Classification, and SVM algorithms to build a robust model with increased accuracy. It is evaluated using recall and precision metrics derived from the confusion matrix.

Hanafy & Ming (2021) Compare 13 machine learning methods in fraud detection in the insurance industry to show the impact of imbalanced datasets on the accuracy of the analysis. Resampling techniques such as Random Over Sampler, hybrid methods, and Random Under

Sampler are implemented to address the imbalanced datasets, thus enhancing the performance of the machine learning algorithms. They conclude that classifier algorithms cannot make accurate predictions with imbalanced datasets. SVM performs best using the Random Over Sampler, while C5.0 performs best using SMOTE and Random Under Sampler.

In their research, Naik and Laxminarayana (2017) state that in SVM, each data item is plotted as a point in n-dimensional space, where n denotes the number of unique features, with the value of each feature relating to the value of a coordinate. Classification is performed by finding the hyperplane that distinguishes the classes. Moreover, SVM is robust in outlier detection.

Rawte and Anuradha (2015) developed a hybrid machine learning algorithm using Evolving Clustering Model and Support Vector Machine. The Evolving Clustering Model is chosen since claim data is dynamic and constantly generated. At the same time, the support vector machine is used to solve the classification problems, thus detecting outliers and duplicate claims. The downside is that other forms of medical fraud are not detected.

In their paper, Naik and Laxminarayana (2017) note that the K-Means learning algorithm solves clustering problems by classifying a given input pace through several clusters (often denoted by k). The main concept is to define k centroids, each cluster having one centroid. These centroids should be placed carefully as different locations result in different outcomes. Data from the input dataset is associated with the nearest centroid. A loop is generated after every revision, and the k centroids may change their location until no further movements are possible.

Ogbuabor and Ugwoke (2018) compare the performance of K-Means and DBSCAN using Silhouette score values. The efficacy of the K-Means algorithm is evaluated using different distance metrics and different numbers of clusters. In contrast, the efficiency of the DBSCAN algorithm uses different distance metrics and the least number of points to form a cluster. The results indicate that K-Means and DBSCAN have solid inter-cluster separation and intra-cluster cohesion. Based on the research outcome, K-Means outperforms the DBSCAN algorithm in accuracy and execution time.

In their paper, Wakoli et al. (2014) apply the K-Means algorithm to medical claim records to cluster the claim type and the cost per claim. The Euclidean distance measure was used to flag suspicious claims that would be revalidated. Similarly, the research by Zhang et al. (2020) compares the fraud detection efficacy of clustering algorithms on known medical fraud records. From their research, traditional rule sorts had a 24% detection rate, while DBSCAN had a 33.0% accuracy. Similarly, K-Means, Isolation Forest, and Local Outlier Factor had a detection rate of 35.0%, 47.0%, and 45%, respectively.

7. Methodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was used for the research. Developed by a consortium of data mining agents through an initiative sponsored by the European Union, CRISP-DM depicted data mining as a six-phase cycle (Schröer, Kruse & Gómez, 2021). The methodology consisted of the following phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Ordering the phases in the CRISP-DM methodology is flexible (Schröer, Kruse & Gómez, 2021).

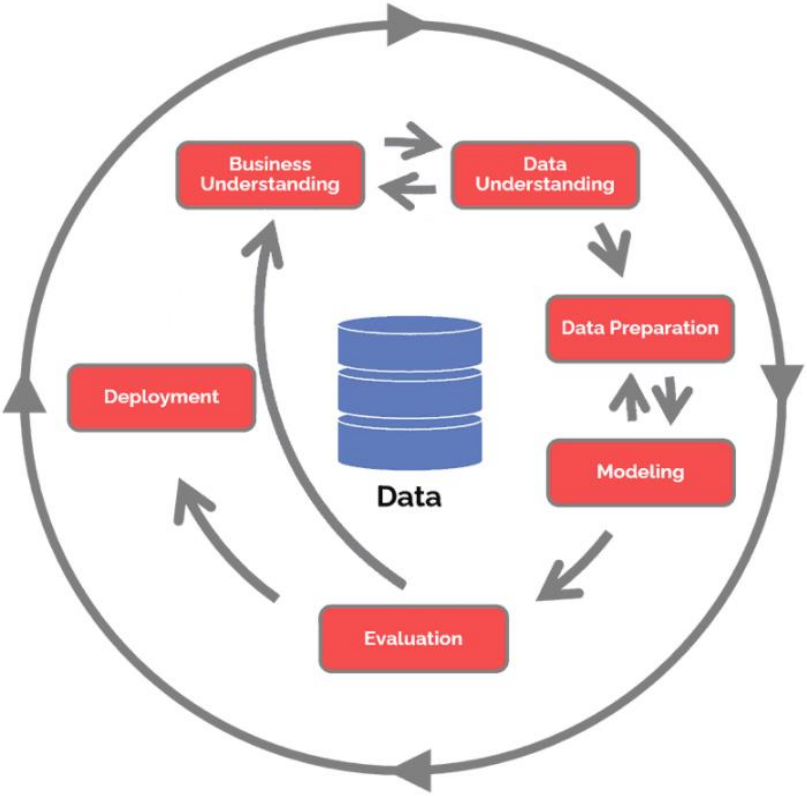

Figure 1. Phases of CRISP-DM (Schröer, Kruse & Gómez, 2021)

8. Business understanding

The business understanding phase of CRISP-DM involved a review of the journals and research papers on the various types of medical insurance fraud, how each fraud type is detected, and the current mitigation strategies for each fraud type assisting the research in gaining a more profound domain knowledge on the current controls, the evolution of fraud schemes, and which datasets can be mined for fraud detection in medical insurance. Moreover, this phase helped in selecting the appropriate technique to be used in answering the research questions.

9. Data understanding

The data understanding phase involved the collection of relevant medical claims datasets in tandem with the acquired domain knowledge. Subsequently, the research sought to familiarize with the claim datasets and ascertain the quality of datasets in developing the fraud detection model.

10. Data preparation

In this phase, the raw data provided was transformed into a standard acceptable format involving several activities, such as selecting relevant attributes and removing irrelevant attributes, handling null or missing values, and removing duplicate entries. The data cleansing step used a process called imputation which identified inaccurate and incomplete datasets,

substituted missing data with a placeholder, and noise reduction by removing data that did not relate to the research questions and objectives. Duplicated features that could be derived from existing feature sets or represented by another feature name were dropped. Identifying these features aided in identifying the strategy for feature engineering, feature relevance, and imputation strategies.

11. Modelling

The modelling phase attempted to solve the clustering and classification problems of the dataset involving implementing a hybrid machine learning approach where the K-Means algorithm was applied to the dataset for clustering similar features, and the Support Vector Machine (SVM) was harnessed for the classification of fraudulent and non-fraudulent claims. The merger of the two algorithms was achieved using a pipeline in Python.

12. Support vector machine model

The implementation phase sought to compare the performance of the hybrid fraud detection model vis the use of a sole supervised machine learning algorithm – SVM. These two models were also tuned as the last step of their iterations, and the performance metrics were recorded, resulting in four models: the lone SVM model, the tuned SVM model, the hybrid model, and the tuned hybrid model.

The transformed dataset was loaded with the SVM model with default hyperparameters for SVC, as shown in Table 1.

Table 1. Default Values for the SVC Model for Model 1

Parameter	Default Value
C	1
kernel	'rbf'
degree	3
gamma	'scale'
coef0	0
shrinking	TRUE
probability	FALSE
tol	1.00E-03
class_weight	None

The default hyperparameters were tuned using Python's Grid Search Cross Validation library to obtain the optimal parameters for the SVM classification algorithm. These optimized and non-optimized predictions were later used as the benchmark for evaluating the classification performance of the hybrid model.

13. Hybrid machine learning model

The first iteration of implementing the hybrid model involved using a pipeline with the K-Means and SVM. The pipeline workflow was designed to run the standardized dataset with the default K-Means for clustering and classifying the output using SVM. The clustering and

classification were performed using the default kernel hyperparameters of the K-Means (shown in Table 2) and SVM algorithms (shown in Table 1), respectively.

Table 2. Default Values for the K-Means algorithm for model 3

Hyperparameter	Default Value
n_clusters	8
init	'K-Means++'
n_init	10
max_iter	300
tol	1.00E-04
precompute_distances	'auto'
verbose	0
random_state	None
copy_x	TRUE
algorithm	auto'

The second iteration of the hybrid model applied the grid search library to exhaustively obtain the optimal parameters for the scaler, principal component analysis components, K-Means clusters, and the hyperparameter C in SVM. The parameter grid for the scaler parameter evaluated the Standard Scaler, Robust Scaler, and Quartile Transformer. The parameter grid for the PCA components ranged from 14 to 22 with an increment of 2. Similarly, the number of K-Means clusters ranged from 6 to 12 with an increment of 2.

14. Evaluation

After training the K-Means and SVM algorithms, the confusion matrix and the classification metrics were used to evaluate the efficiency and performance of the model on claims data on the insured. The model tested how many claims are categorized as false positives and false negatives (recall measure). Additionally, the model's performance was gauged by the percentage of correct classification of fraudulent claims (precision measure). The input space used a random resampling technique that rebalanced the imbalanced dataset's class distribution to improve the models' accuracy and reduce the false negatives and false positives. The performance evaluation metrics for the study included accuracy, precision, recall, and F1 score. These metrics were plotted on a Confusion Matrix to provide a detailed breakdown of the algorithm's true positive, true negative, false positive, and false negative predictions.

15. Results

The study summarized the performance of the four prototypes using a confusion matrix to evaluate the classification performance by categorizing the predicted and actual labels into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Classification accuracy, calculated by dividing the number of correct predictions by the total number of predictions, is a standard performance metric used to evaluate machine learning models by measuring the classified instances of true positives and true negatives (Altman & Krzywinski, 2017).

Table 3. Summarized accuracy of the models

Iteration	Model Description	Accuracy %
Model one	SVM with default hyperparameters	91.31%
Model two	SVM with optimal hyperparameters	97.05%
Model three	Hybrid model with default hyperparameters	68.08%
Model four	Hybrid model with tuned hyperparameters	97.49%

16. Confusion matrix

The study summarized the performance of the four prototypes using a confusion matrix to evaluate the classification performance by categorizing the predicted and actual labels into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

True Positives refer to the instances where the model predicted the positive class correctly with the actual label being positive. True Negatives occur when the model predicts a negative class correctly, thus matching with the negative label. False Positive (Type I error) occurs when the model incorrectly predicted a positive class, whereas the actual label is negative. False negative (Type II error) occurs when the model predicts a negative class, whereas the actual label is positive.

Table 4. Summary of the confusion matrix of the four models

Summary	TP	FP	FN	TN	Type I error rate	Type II error rate
Model one	1138	106	109	1122	4.28%	4.40%
Model two	1219	45	28	1183	1.82%	1.13%
Model three	784	328	463	900	13.25%	18.71%
Model four	1217	32	30	1196	1.29%	1.21%

The SVM model with default parameters noted a 4.28% Type I error and a 4.40% Type II error rate from the entire dataset. With the second iteration, both Type I and Type II error rates reduced to 1.82% and 1.13%, respectively, with an overall accuracy of 97.05% noted. The third prototype recorded increased Type I and Type II error rates at 13.25% and 18.75%, reducing the model's accuracy to 68%. The fourth model had the best accuracy rate of 97.45, mainly attributed to the lowest Type 1 error at 1.29%. The Type II error rate for the fourth model stood at 1.21%.

Further to the classification accuracy and confusion matrix, we examined other performance metrics such as precision, recall, and the F1 score to gauge the improvements of the four iterations of our model. Precision-measured the proportion of positively predicted cases against all predicted positive cases.

- Precision = TP / (TP+FP)

Recall measured the percentage of True Positives against all actual positive cases.

- Recall = TP / (TP +FN)

The F1 score evaluated the mean of precision and recall providing a balanced measure of the model's performance.

- F1 score = 2 * (precision * recall) / (precision + recall)

Table 5. Summary of the classification report on the algorithms

	Precision:	Recall:	F-Score:
Model one	91.48%	91.26%	91.37%
Model two	96.44%	97.75%	97.09%
Model three	70.33%	62.55%	66.21%
Model four	97.44%	97.59%	97.52%

17. Discussion

While comparing the performance of the four prototypes, the hybrid model with optimized hyperparameters performed better than the other three prototypes in the classification of fraudulent claim transactions. Prototype 4 recorded the highest accuracy, precision, and f-score among the four models. However, prototype 2 recorded the highest recall score at 97.75%. Bauder et al. (2017) posit that hybrid machine learning models have the potential to outperform a single algorithm due to their robust nature, adaptability, complementary strengths, and fusion of decisions.

The introduction of hyperparameter tuning in model 1 and model 3 improved the accuracy of the base models. Hyperparameter tuning is selecting optimal parameters for a machine learning model. Hyperparameters control the behavior and influence the model's performance to find the best combination of parameters that will lead to the best performance of the model on a specific dataset. The study adopted the grid search approach, which predefined values for each parameter. Model 2 and Model 4 exhaustively evaluated all possible combinations of values defined in the grid set and returned the best parameter values for selection for each grid entry. Bergstra and Bengio (2012) note that grid search is computationally expensive for large search spaces and grid entries. In our study, the grid search hyperparameter tuning from Model 1 to Model 2 runs for approximately 305 seconds, while that from Model 3 to Model 4 runs for 3740 seconds.

The iteration from model three to model four introduced Principal Component Analysis to the pipeline before the k-Means algorithm step to reduce the number of dimensions in the dataset. Dimensionality reduction is the feature reduction process while preserving as much relevant information as possible to mitigate overfitting, noise reduction and improve the computational efficiency of the model. While there is no predefined cutoff for the number of components to be used in the PCA, Abdi and Williams (2010) suggest that the number of components selected should explain a high percentage of the total variance in the dataset. The introduction of PCA to model 4 increased the overall rise in performance metrics.

18. Model verdict

Based on the benchmark results of the SVM and in comparison, with the hybrid models, we note that the models with tuned hyperparameters scored better than those with the default parameters. Model 4 has the best accuracy, precision, and F1 scores in this case. Model 2 came in second with the best overall recall but second in accuracy, precision, and F1 scores. Model 1 was ranked third, with all performance measures being the third best. Model 3 was ranked in the fourth position. The hybrid classification model that uses both K-Means and SVM recorded a slight improvement in the classification of fraudulent and genuine claims compared to the classification of a single SVM model.

19. Conclusion

The research proposed, evaluated, and ranked the performance of a hybrid machine learning model that consisted of clustering using K-Means before classification using SVM. Various studies indicated hybrid machine-learning models perform better than a single algorithm (Zang & Ma, 2020; Bauder et al., 2017; Abdallah et al., 2017). This research adds to the existing knowledge base and elicits that hyperparameter tuning is a crucial step for performance metrics to be improved in hybrid algorithms. In as much as hyperparameter tuning adds to the model's accuracy, there needs to be consideration of its impact on the speed of the model's performance, especially if multiple steps are in the pipeline. Each parameter set for hyperparameter tuning increases the computation time exponentially. Nonetheless, a balance needs to be sought between improving the accuracy of the model vis a vis the acceptable execution time of the model.

20. Recommendation

The study's findings can be extended to the existing fraud detection models in the insurance industry with added accuracy by using singular classification algorithms. The study answers the question of the performance of the hybrid model in fraud detection. The study recommends an integrated approach with the model's prediction capabilities and core applications to detect fraud in real-time, which can be achieved using Application Programmable Interfaces (APIs) to get the classification rating based on the dataset's features.

21. Future research

While the developed model recorded an accuracy rate of 97.49%, further research needs to be conducted on improving the computational and speed performance of tuning hyperparameters in a hybrid machine-learning model. The study adopted grid search cross validation which exhaustively fits the parameter set. The study can be validated against other medical insurance firms to revalidate the outputs and reinforce learning.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public commercial, or not-for-profit sectors.

The authors declare no competing interests.

References

Abdallah, A., Maarof, M., & Zainal, A. (2016). Fraud Detection System: A Survey. *Journal of Network and Computer Applications*, 90-113.

Abdi, H., & Williams, L. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433-459.

Ai, J., Lieberthal, R., Skyla, S., & Wojciechowski, R. (2018). Examining predictive modeling–based approaches to characterizing health care fraud. *Society of Actuaries*. https://www.soa.org/resources/research-reports/2018/healthcare-fraud.

Altman, N. S., & Krzywinski, M. (2017). Points of significance: Classification evaluation. *Nature Methods, 14*(8), 755-756.

Association of Certified Fraud Examiners (2019). *Insurance Fraud Handbook*. Association of Certified Fraud Examiners, Inc.

Association of Kenya Insurers (2020). *2020 Insurance Industry Report*. Nairobi: Association of Kenya Insurers.

Association of Kenya Insurers (2021). *Information Paper on Insurance Fraud*. Nairobi: Association of Kenya Insurers.

Bauder, R., Khoshgoftaar, T., & Seliya, N. (2017). A Survey on the state of healthcare upcoding fraud analysis and detection. *Health Services & Outcomes Research*, 31-55.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 281-305.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Obleb, F., & Bontempi, G. (2021, May). Combining unsupervised and supervised learning in credit card fraud. *Business Analytics Emerging Trends and Challenges*, *557*, 317-331.

Gupta, R. Y., Mudigonda, S. S., & Baruah, P. K. (2021, March). A comparative study of using various machine learning and deep learning-based fraud detection models for universal health coverage. *International Journal of Engineering Trends and Technology*, 96-102.

Hanafy, M., & Ming, R. (2021). Using machine learning models to compare various resampling methods in predicting insurance fraud. *Journal of Theoretical and Applied Information Technology, 99*(12), 2819-2833.

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015). Using data mining to detect health care fraud and abuse: A review of literature. *Global Journal of Health Science*, 194-202.

Kose, I., Gokturk, M., & Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing Journal, 36*, 283-299. https://doi.org/10.1016/j.asoc.2015.07.018

Lawand, S., & Kulkarni, U. (2019). Survey on fraud prediction for an application using data mining. *International Journal of Emerging Technologies and Innovative Research, 6*(6), 209-212. http://doi.one/10.1729/Journal.22988

Matloob, I., & Khan, S. (2019). A framework for fraud detection in government supported national healthcare programs. *Electronics, Computers and Artificial Intelligence, ECAI 2019*. Romania.

Matloob, I., Khan, S., ur Rahman, H., & Hussain, F. (2020). Medical health benefits management system for real-time notification of fraud using historical medical records. *Applied Sciences*, *10*(15). https://doi.org/10.3390/app10155144

Naik, J., & Laxminarayana, A. (2017). Designing hybrid model for fraud detection in insurance. In *National Conference on Advances in Computational Biology, Communication, and Data Analytics*, 24-30.

Ogbuabor, G., & Ugwoke, F. (2018). Clustering algorithm for a healthcare dataset using silhouette score value. *International Journal of Computer Science & Information Technology*, *10*(2), 27-37.

Rawte, V., & Anuradha, G. (2015). Fraud detection in health insurance using data mining techniques. In *2015 International Conference on Communication, Information & Computer Technology (ICCICT)*.

Schröer, C., Kruse, F., & Gómez, J. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 526-534.

Segal, S. Y. (2016). Accounting frauds – Review of advanced technologies to detect and. *Economics and Business Review*, 45-64.

Waghade, S. S., & Karandikar, A. (2018). A comprehensive study of healthcare fraud detection based on machine learning. *Nagpur: International Journal of Applied Engineering Research*.

Retrieved from https://www.ripublication.com/ijaer18/ijaerv13n6_140.pdf.

Wakoli, L., Orto, A., & Mageto, S. (2014). Application of the K-means clustering algorithm in medical claims fraud / abuse algorithm in medical claims fraud / abuse detection. *International Journal of Application or Innovation in Engineering & Management*, *3*(7), 142-151.

Zhang, C., Xiao, X., & Wu, C. (2020). Medical fraud and abuse detection system based on machine learning. *International Journal of Environmental Research and Public Health*, *17*(7265), 1-11.

Zhang, Y., & Ma, S. (2020). *Ensemble machine learning: Methods and applications*. Springer.

Zhou, S., & Zhang, R. (2020). A novel method for mining abnormal expenses in social medical insurance. *International IoT, Electronics, and Mechatronics Conference, Proceedings*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/IEMTRONICS51293.2020.9216354

Zhou, S., He, J., Yang, H., Chen, D., & Zhang, R. (2020). Big data-driven abnormal behavior detection in healthcare based on association rules. *IEEE Access*, 129002–129011. https://doi.org/10.1109/ACCESS.2020.3009006


C O A S