MTU - JBMS

Malatya Turgut Özal Üniversitesi İşletme ve Yönetim Bilimleri Dergisi
Malatya Turgut Özal University Journal of Business and Management Sciences

ISSN: 2717-7890

***Cilt/Volume**:5, **Sayı/Issue:**1, **Yıl/Year:**2024, **Sayfa/Page**:39-56, **ISSN:** 2717-7890*

RISKIFIED FRAUD DETECTION USING MACHINE LEARNING: INSURANCE CLAIMS

Makine Öğrenmesi Kullanarak Riskli Dolandırıcılık Tespiti: Sigorta Talepleri

Hakan KAYA*

**Doktorant*, hkkaya5@std.okan.edu.tr, ORCID: 0000-0002-0812-4839*

Research Article / Araştırma Makalesi

Received Date:
09.02.2024
Acceptance Date:
02.04.2024

Keywords:
Fraud Detection, Insurance, Machine Learning Algorithms

JEL Codes:
C53, C80, G22

Similarity Rate:
Ithenticate: %8

ABSTRACT

In the insurance industry, fraud presents a significant and widely recognized challenge. With fraudulent claims posing a substantial financial burden on insurers, it's crucial to distinguish between legitimate and false claims. Given the impracticality of manually scrutinizing every claim due to the associated time and cost, employing advanced technology becomes imperative. This article delves into utilizing predictive models powered by machine learning algorithms to analyze claim data. For the study, a dataset was prepared from the damage records of a private insurance company. Eleven predictive models (Ada Boost, Cat Boost, Decision Tree, Extremely Randomized Tree, Gradient Boosting, KNN, LightGBM, Random Forest, Stochastic Gradient Boosting (SGB), Support Vector Classification (SVC), and Voting Classifiers) are applied for developing a fraud detection mechanism. Algorithms will be compared in terms of score the algorithm that gives the best values will be determined. GridSearchCV, Confusion Matrix and Classification Report methods (Accuracy, Precision, Recall, and F1-Score) of the used to calculate and display all metrics. As a result of this study, the Random Forest and Decision Tree Classifiers outperformed the other models with have the highest classification accuracy of 75.6%. The findings of this study are beneficial for fraud detection and the underlying framework holds a functionality for real-time problem-solving in the insurance sector.

Geliş Tarihi:
09.02.2024
Kabul Tarihi:
02.04.2024

Anahtar Kelimeler:
Dolandırıcılık Tespiti, Sigorta, Makine Öğrenimi Algoritmaları

JEL Kodları:
C53, C80, G22

Benzerlik Oranı:
Ithenticate: %8

ÖZ

Sigorta sektöründe dolandırıcılık önemli ve yaygın olarak kabul edilen bir sorundur. Sahte iddiaların sigortacılara önemli bir mali yük getirdiği göz önüne alındığında, meşru ve sahte iddialar arasında ayrım yapmak çok önemlidir. İlgili zaman ve maliyet nedeniyle her iddiayı manuel olarak incelemenin pratik olmadığı göz önüne alındığında, gelişmiş teknolojinin kullanılması zorunlu hale gelmektedir. Bu çalışmanın amacını, sigorta endüstrisinde dolandırıcılığı tespit etmek için makine öğrenimi algoritmalarıyla tahmin modellerinin kullanıldığı bir çerçeve oluşturmaktır. Çalışma için özel bir sigorta şirketinin hasar kayıtlarından bir veriseti hazırlanmıştır. Dolandırıcılık tespiti mekanizması geliştirmek için on bir tahmin modeli (Ada Boost, Cat Boost, Decision Tree, Extremely Randomized Tree, Gradient Boosting, KNN, LightGBM, Random Forest, Stochastic Gradient Boosting (SGB), Support Vector Classification (SVC) ve Voting Classifiers) uygulanmaktadır. Algoritmalar doğruluk değeri açısından karşılaştırılacak, en iyi değerleri veren algoritma belirlenecektir. Tüm metrikleri hesaplamak ve görüntülemek için GridSearchCV, Karmaşıklık Matrisi ve Sınıflandırma Raporu yöntemleri (Doğruluk, Kesinlik, Geri Çağırma ve F1-Puanı) kullanılmıştır. Bu çalışmanın sonucunda, Random Forest ve Decision Tree algoritmaları %75,6 ile en yüksek sınıflandırma doğruluğuna sahip olarak diğer modellerden daha iyi performans göstermiştir. Bu çalışmanın bulguları, sigorta sektöründe dolandırıcılık tespiti için faydalı ve temel çerçeve, sigorta sektöründe gerçek zamanlı problem çözme için bir işlevselliğe sahiptir.

.

Citation / Atıf: Kaya, H. (2024). Riskified Fraud Detection Using Machine Learning: Insurance Claims. *Malatya Turgut Ozal University Journal of Business and Management Sciences,* 5(1), 39-56.

1. INTRODUCTION

The successful implementation of data science algorithms hinges on the specific problem or task at hand and the characteristics of the available data. These algorithms demonstrate versatile applications across various domains and industries, presenting solutions for a wide range of challenges.

Vehicle insurance fraud involves collaborating to submit misleading or exaggerated claims regarding property damage or personal injuries following an accident. This can encompass staged accidents, fictitious passengers, and overstated injury claims.

Based on the provided information, it's clear that insurance companies have faced numerous challenges concerning fraud and abuse methods within the auto branch. The notification data from January 2023 to 2024 sheds light on the prevalence of these issues. Reasons for SISBIS notification: Additional research requirement (14.46%), waivers obtained from insured persons (37.92%), determination of false statements (25.67%), fraud cases resolved by courts (1.04%), driver information cases (0.28%), and others (20.62%). For abuses methods: Issuance of policies after damages (16.65%), driver change due to insufficient license (26.99%), damage applications with false documents (21.35%), malinger attempts by insured person (15.88%), determination of fictitious damage (9.91%), and suspect-based damage applications (9.22)(https://siseb.sbm.org.tr).

Insurance companies encounter a considerable obstacle with claim leakage, leading to significant financial setbacks. Fraudulent claims pose a costly problem, with the potential to incur billions of dollars in annual costs for the industry. In response, cutting-edge machine learning methods are being suggested to accurately identify fraudulent activities within the insurance sector. This proposed approach is assessed using Python and dataset samples obtained from insurance agencies.

The study's objective is to research in the application of machine learning algorithms demonstrates an innovative approach to tackling fraud detection, addressing a critical issue of fraudulent claims detection. In this pioneering research paper, the answers to these there questions which are, How do machine learning algorithms help in distinguishing between legitimate and false insurance claims? What are the key findings of the study regarding the performance of different predictive models in fraud detection? How can the insights from this study be applied to improve fraud detection mechanisms in insurance companies? Given the limited of research, particularly pertaining to damages within the insurance sector for experts and academic researchers in this field, indicating the potential impact of the research on future studies and industry practices.

This article includes an Exploratory Data Analysis and Predictive Machine Learning Models for detecting fraud. The paper is structured as follows: Section 2 covers the literature related to the work. In Section 3, the research design, methods, data collection and analysis procedures, as well as the utilization of machine learning models are described. The outcomes of the research, along with the evaluation criteria for assessing model performance, are discussed in Section 4, and Section 5 presents the paper's conclusion.

2. LITERATURE REVIEW

Upon reviewing the literature, it seems to revolve around the common theme of utilizing machine learning and data analysis techniques for fraud detection in the auto insurance industry.

Hybrid Approaches: Studies like Subudhi and Panigrahi (2020) and Sathya and Balakumar (2022) employ hybrid approaches combining different techniques like Genetic Algorithms, Fuzzy C-Means clustering, and blockchain technology. Strengths lie in their potential to capture complex patterns and improve accuracy by leveraging the strengths of multiple methods. However, weaknesses may include increased complexity, computational overhead, and potential challenges in integrating disparate techniques seamlessly. Evaluation and Comparison Studies: Hanafy and Ming (2021, 2022), Geren (2020), and Nordin et al. (2024) focus on evaluating and comparing various machine learning algorithms for fraud detection. Strengths include providing insights into the relative performance of different methods, aiding practitioners in selecting appropriate algorithms. Weaknesses may involve variability in datasets, making direct comparisons challenging, and potential bias in algorithm selection or evaluation metrics. Application of Reinforcement Learning: Choi et al. (2021) explore the application of reinforcement learning techniques like DQN and DDQN. Strengths include the ability to learn optimal strategies through interactions with the environment. However, weaknesses may include the need for extensive computational resources, complex model tuning, and potential challenges in defining the reward function accurately. Literature Reviews: Ali et al. (2022) conduct a systematic literature review, summarizing existing research on ML-based fraud detection. Strengths include providing a comprehensive overview of the field, identifying trends, and highlighting gaps for future research. Weaknesses may include potential bias in study selection or synthesis and the reliance on existing literature, which may not capture the latest developments. Application of Machine Learning in Insurance Sector: Jones and Sah (2023) study the broader application of machine learning in the insurance sector. Strengths include exploring various aspects beyond fraud detection, such as risk assessment and customer segmentation, offering holistic insights. However, weaknesses may include less focus specifically on fraud detection and potential challenges in generalizing findings to specific fraud detection contexts. Automation of Evaluation Process: Kalra, Singh, and Kumar (2022) propose an automated system for evaluating insurance claims using machine learning techniques. Strengths include improving efficiency and consistency in claim processing. Weaknesses may include potential biases in algorithmic decision-making, lack of interpretability in automated decisions, and challenges in handling complex or novel fraud scenarios. Enhancing Classification Performance: Itri et al. (2020) introduce a novel oversampling technique, TH-SMOTE, to improve classifier performance. Strengths include addressing imbalanced datasets common in fraud detection, potentially leading to better model performance. Weaknesses may include sensitivity to parameter tuning and limited generalizability across different datasets or fraud scenarios.

3. METHODOLOGY

The application developed in the study is written in the Python programming language. Real damage data obtained from a private insurance company is utilized. Visual Studio Code was used during the development of the application. The compared algorithms are Ada Boost, Cat Boost, Decision Tree, Extremely Randomized Tree, Gradient Boosting, KNN, LightGBM, Random Forest, Stochastic Gradient Boosting (SGB), Support Vector Classification (SVC) and Voting Classifiers. In the application, accuracy rates will be measured based on the detection of fraudulent damage data. A prediction will be made for each algorithm regarding fraudulent records, and the results will be compared.

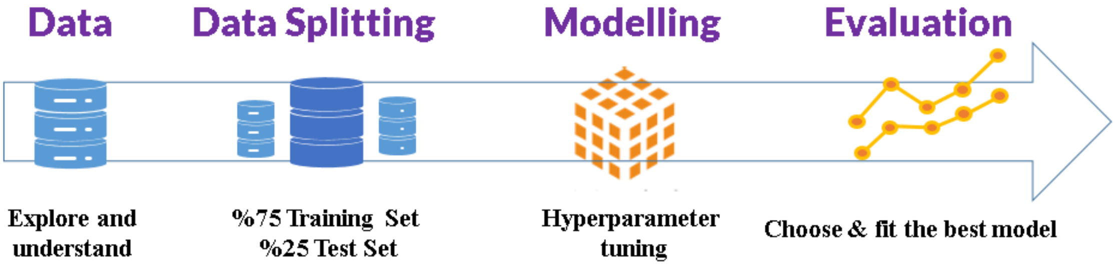

Figure 1. ML Based Fraud Detection Framework

In the model training process, a grid of parameter values is created for both the Decision Tree classification model and the AdaBoost classification model. For the Decision Tree model, the grid of parameter values includes: min_samples_split and min_samples_leaf with values ranging from 2 to 10. max_depth with values of 3, 5, 7, or 10. ‘criterion’ which searches through 'gini' or 'entropy' to find the ideal criterion. The Decision Tree model is used to predict the value of a target variable by learning simple decision rules inferred from the data features. For the AdaBoost model, the grid of parameter values includes: n_estimators with values of 50, 70, 90, 120, 180, or 200. ‘learning_rate’ with values of 0.001, 0.01, 0.1, 1, 10. ‘algorithm’ which searches through 'SAMME' to find the ideal algorithm. The AdaBoost algorithm involves a trade-off between learning rate and the number of estimators. The process commences with the fitting of a regressor on the original dataset. Subsequently, additional replicas of the regressor are fitted on the same dataset, with adjustments made to the instance weights based on the error of the current prediction.

This approach allows for the creation of models that predict the target variable using decision rules and iterative adjustments to sample weights and model weights, ultimately optimizing the performance of the models.

In Figure 2 shows, 753 frauds for "No" and and 247 frauds as "Yes" were reported.

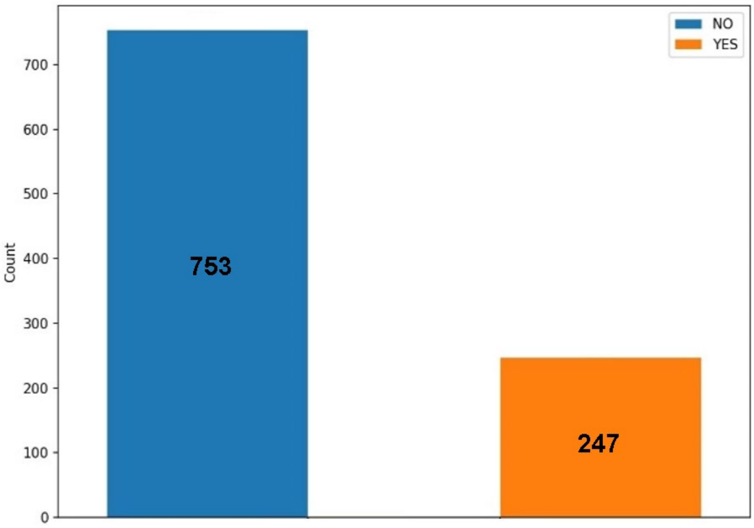

Figure 2. Number of Frauds Reported

3.1. DATA SET AND MODELS

By using 75% of the data for training and 25% for testing, out of 1000 records randomly selected the training set contains 750 and the test set contains 250.

When conducting model training and testing, a normal train/validation/test split involves the model training on a specific randomly selected portion of the data, validating on a separate set of data, and finally testing on a holdout dataset.

In contrast, cross-validation works by splitting the dataset into random groups, holding one group out as the test set, and training the model on the remaining groups. This process is repeated for each group being held as the test group, and then the average of the models is used for the resulting model.

Implementing grid search allows running the model over a grid of hyperparameters to identify the optimal result. It involves keeping the holdout test data consistent between models but using cross-validation and grid search for parameter tuning on the training data to evaluate the resulting outputs. This approach offers a more comprehensive assessment of the model's performance and allows for better parameter tuning.

GridSearchCV is a technique involves searching for the optimal hyperparameters for a given model. It allows you to define a grid of parameters that will be searched using cross-validation. This helps in fine-tuning the model for better performance. Confusion Matrix is a quality measurement of predictions. With data from the confusion matrix interprets the results by looking at the classification report (Liashchynskyi and Liashchynskyi, 2019:3). The classification report returns the metrics relevant to evaluating classification model:

Accuracy represents the rate of correct predictions. It is calculated by dividing the number of test observations with correctly predicted labels by the total number of test observations. The accuracy score ranges between 0 and 1, and scores approaching 1 are considered indicative of a successful model (Naseer, 2022:12961).

$$\text{Accuracy} = (\text{TP}+\text{TN}) / (\text{TP}+\text{TN}+\text{FN}+\text{FP}) \qquad (1)$$

Precision is calculated as the ratio of True Positive predictions to the total predicted positives (Sokolova, Japkowicz, and Szpakowicz, 2006:1016).

$$\text{Precision} = \text{TP} / [\text{TP} + \text{FP}] \qquad (2)$$

Recall is the ratio of correctly predicted observations within a class to all the observations that are actually correct within that class (Muneer and Fati, 2020: 196761).

$$\text{Recall (Sensitivity)} = \text{TP} / [\text{TP} + \text{FN}] \qquad (3)$$

F1-Score is the harmonic mean of the Precision and Recall values, making it a good indicator of overall performance (Sokolova, Japkowicz, and Szpakowicz, 2006:1016).

$$\text{F1-Score} = 2 / \{[1 / \text{Precision}] + [1 / \text{Recall}]\} \qquad (4)$$

The F1-Score reaches 1 only when both precision and recall are 1, indicating perfect precision and recall. A high F1-Score is achieved when both precision and recall are high, making it a robust metric for evaluating model performance.

Moreover, the F1-Score's use of the harmonic mean of precision and recall positions it as a superior measure compared to accuracy, particularly in the presence of class imbalance. This feature enables the F1-Score to offer a more equitable evaluation of a model's ability to capture both Positive and Negative instances.

If the accuracy is high but the recall or precision is low, it's important to consider the possibility of an imbalance.

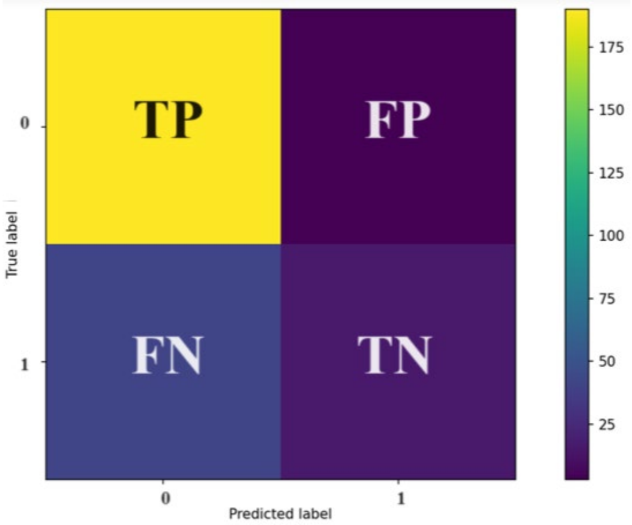

Figure 3. Confusion Matrix

True Positive (TP) refers to the instances in a binary classification task where the model correctly predicts the positive class. In other words, TP represents the number of positive cases that were correctly identified as positive by the model. You guessed positively and it is correct.

False Positive (FP) refers to the instances in a binary classification task where the model incorrectly predicts the positive class. In other words, FP represents the number of negative cases that were falsely identified as positive by the model. You guessed positive and it is wrong.

True Negative (TN) denotes the instances in a binary classification task where the model correctly predicts the negative class. In essence, TN represents the number of negative cases that were accurately identified as negative by the model. You guessed the negative and it's true.

False Negative (FN) refers to the instances in a binary classification task where the model incorrectly predicts the negative class. In other words, FN represents the number of positive cases that were falsely identified as negative by the model. You guessed negative and it is wrong.

In the evaluation of the confusion matrix, the TN and TP values provide the numbers of our correct predictions.

3.1.1. Ada Boost

The AdaBoostClassifier is an influential ensemble learning technique that aggregates the predictions of multiple weak classifiers to create a robust classifier. It operates by assigning greater weights to misclassified data points in each iteration, enabling subsequent weak learners to prioritize these previously misclassified samples. This iterative approach culminates in the creation of a strong classifier through the amalgamation of the weighted sum of individual weak classifiers (Freund and Schapire, 1996:150).

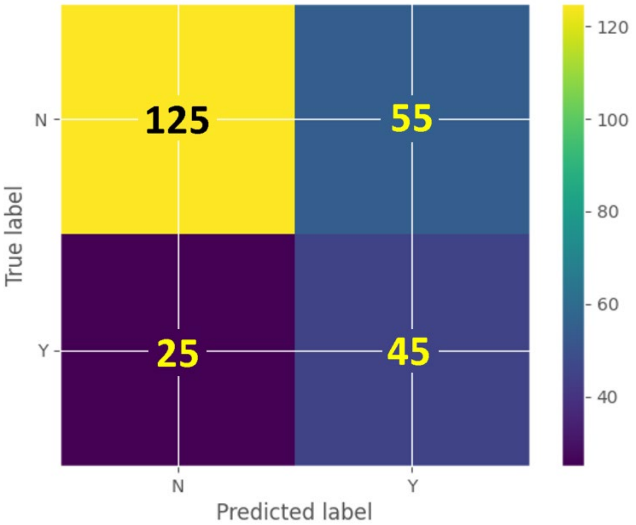

Figure 4. Confusion Matrix of Ada Boost

3.1.2. Cat Boost

The CatBoostClassifier is a machine learning algorithm that excels in effectively handling categorical features within the framework of gradient boosting. It combines the principles of gradient boosting and randomization to enhance prediction accuracy. CatBoost distinguishes itself by effectively managing categorical variables, including those with a large number of levels, through techniques such as ordered boosting. Additionally, its innovative algorithm for processing categorical features notably accelerates both training and inference processes (Prokhorenkova et al., 2018:31).

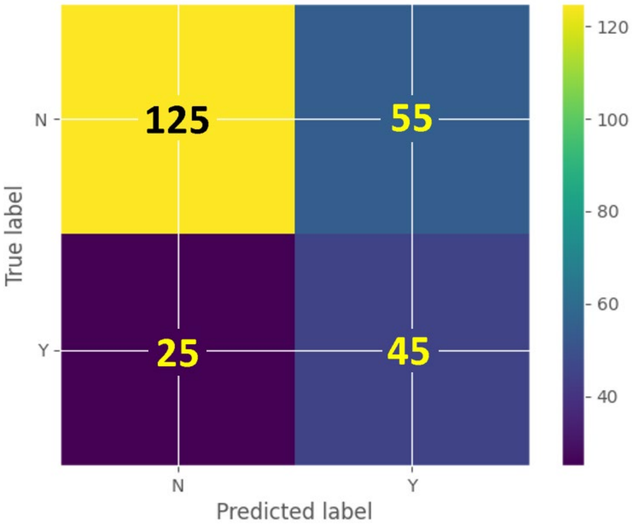

Figure 5. Confusion Matrix of Cat Boost

3.1.3. Decision Tree

Decision trees are built using an algorithmic approach that splits the dataset based on various conditions. This recursive process divides the data into subsets, with each node representing tests on input features and each leaf node representing predicted class labels (Charbuty & Abdulazeez, 2021:21).

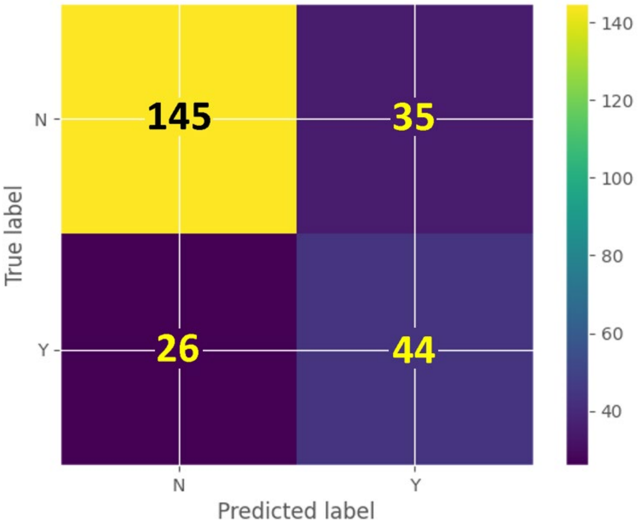

Figure 6. Confusion Matrix of Decision Tree

3.1.4. Extremely Randomized Tree

The ExtraTreesClassifier, an extension of the Random Forest algorithm, employs additional randomness in the tree building process to create a diverse set of decision trees, ultimately reducing variance and enhancing generalization performance. This is achieved by randomly selecting subsets of training data and features at each node split, as well as using random thresholds for feature-based data splits. These randomization techniques effectively mitigate overfitting, enhance diversity among trees, and bolster the classifier's robustness against noise and outliers in the dataset (Goetz et al., 2014:8).

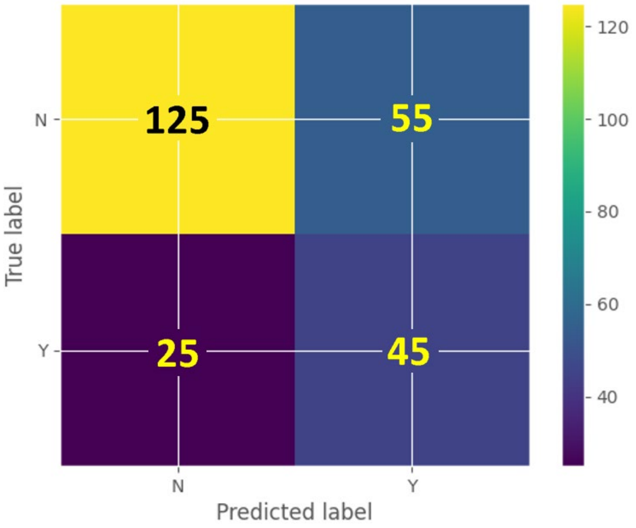

Figure 7. Confusion Matrix of Extremely Randomized Tree

3.1.5. Gradient Boosting

The GradientBoostingClassifier is an influential ensemble learning method that sequentially builds decision trees, with each tree focusing on rectifying the errors of its predecessors. This iterative approach minimizes prediction errors by optimizing a differentiable loss function, resulting in the development of a highly accurate predictive model (Chakrabarty et al., 2019:657).

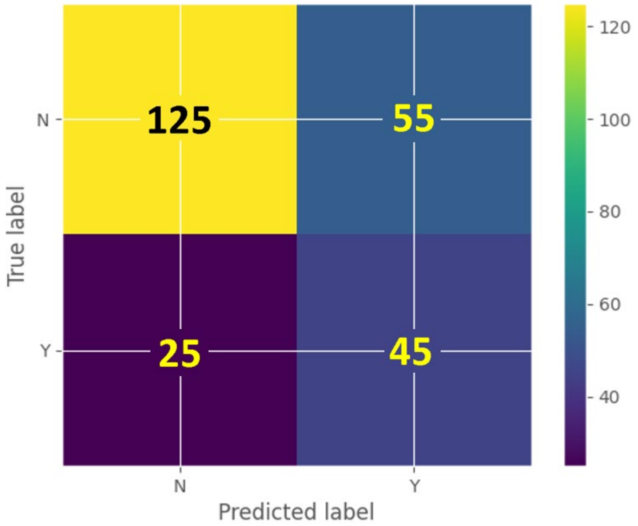

Figure 8. Confusion Matrix of Gradient Boosting

3.1.6. KNN

The K-Nearest Neighbors (KNN) algorithm is a foundational non-parametric method employed for classification and regression tasks. It functions by identifying the ‘K’ closest data points (neighbors) to a query instance using a selected distance metric, and subsequently making predictions based on the majority class (for classification) or the average (for regression) of these neighboring points (Kramer, 2013:14).

To find the ideal value for ‘K’ in KNN classification, the analysis of algorithm errors is pivotal. By plotting the graph of ‘K’ values and the corresponding metric for the test set, we can determine the ‘K’ that minimizes loss. In Figure 7 shows the F1-Score is highest when ‘K’ equals 12. Subsequently, plotting the F1-Score values against ‘K’ values and retraining the classifier using 12 neighbors is recommended to maximize the F1-Score. Figure 8 is retrains our classifier with 12 neighbors.

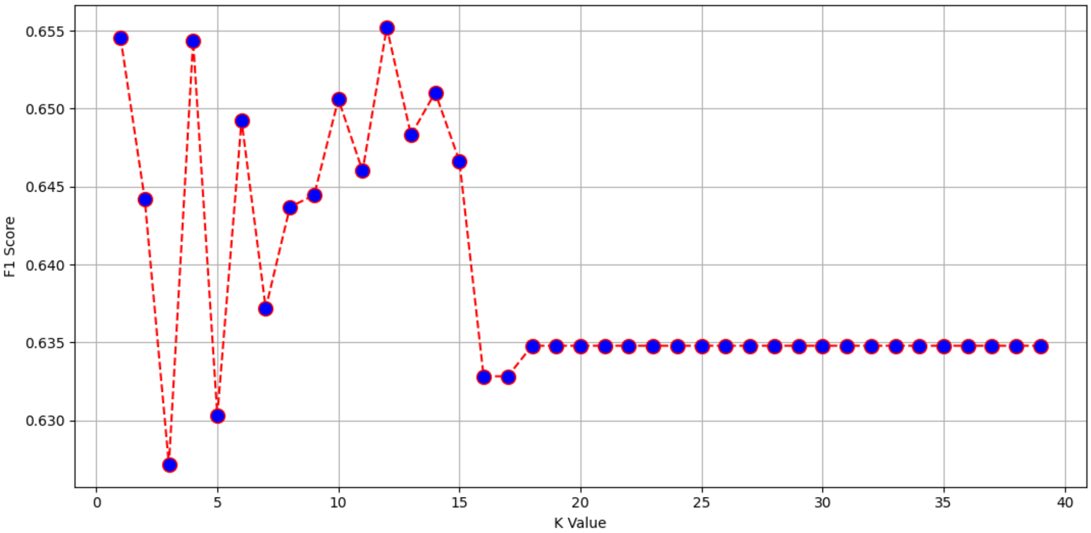

Figure 9. Discovering the Optimal 'K' for KNN Classification

It's important to note that the weighted F1-Score for a class decreases when it becomes more unreliable.

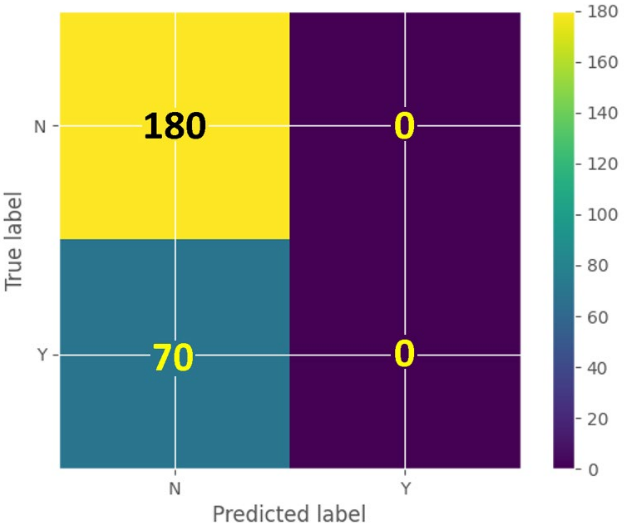

Figure 10. Confusion Matrix of KNN

3.1.7. LightGBM

LightGBM is a widely used gradient boosting framework utilizing tree-based learning algorithms, valued for its efficiency and speed in machine learning and data analysis. It leverages the gradient-based one-side sampling technique and the exclusive feature Bundling (EFB) algorithm to handle large datasets more efficiently than traditional gradient boosting engines. LightGBM is specifically designed to optimize accuracy, speed, and resource usage, rendering it well-suited for large-scale and distributed machine learning tasks (Ke et al., 2017:3148).

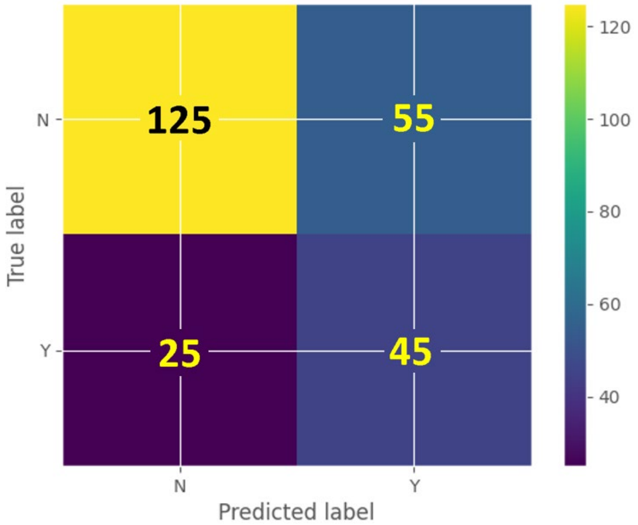

Figure 11. Confusion Matrix of LightGBM

3.1.8. Random Forest

Random Forest is a powerful ensemble learning method used for classification, regression, and various machine learning tasks. It constructs multiple decision trees during training and then determines the mode of the classes (for classification) or mean prediction (for regression) from these individual trees. The algorithm introduces randomness in two key ways: firstly, by randomly selecting a subset of features when splitting nodes, and secondly, by building multiple trees with bootstrapped samples of the data. This deliberate randomness helps decorrelate the trees, leading to the creation of a more resilient and precise model (Au, 2018:1743).

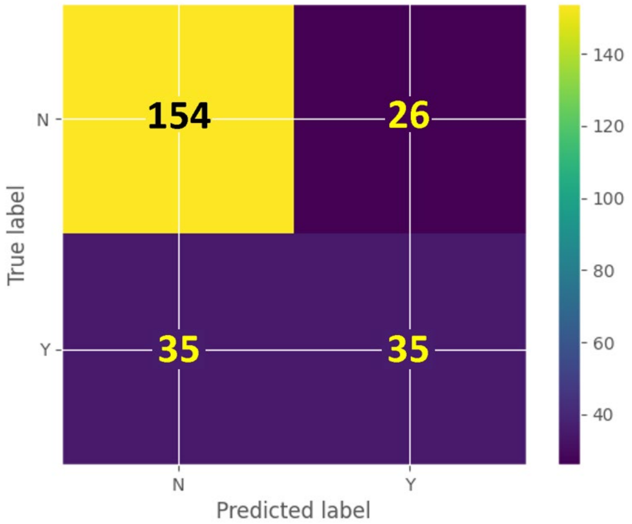

Figure 12. Confusion Matrix of Random Forest

3.1.9. Stochastic Gradient Boosting (SGB)

Stochastic Gradient Boosting (SGB) is an effective machine learning framework that extends the gradient boosting method by introducing stochasticity. This is achieved through subsampling of training data and features, enhancing generalization and reducing overfitting. By sequentially building trees to correct errors, SGB improves the robustness and performance of the gradient boosting algorithm. It is known for its ability to handle diverse loss functions and has demonstrated provable generalization guarantees (Ustimenko and Prokhorenkova, 2021:10491).

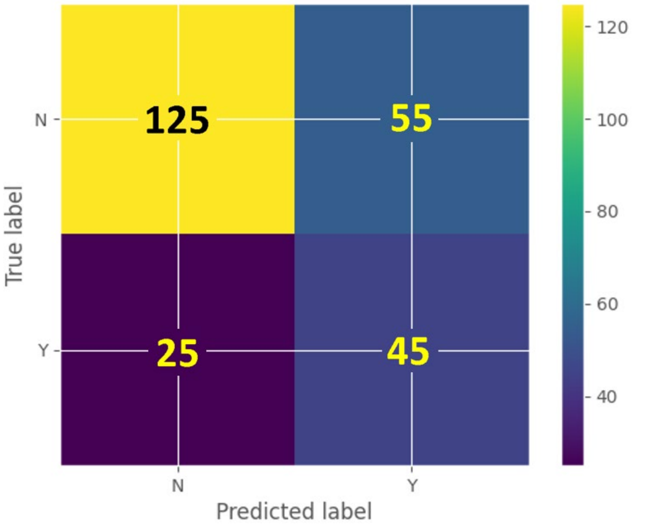

Figure 13. Confusion Matrix of SGB

3.1.10. Support Vector Classification (SVC)

Support Vector Classification (SVC) is a robust supervised learning algorithm tailored for classification tasks. It excels in identifying the optimal hyperplane within high-dimensional spaces to distinctly separate classes in input data. Key features of SVC encompass its adeptness at addressing linear and non-linear classification challenges, its resilience against overfitting via the margin maximization principle, and its proficiency in managing high-dimensional datasets (Liu, 2022:48).

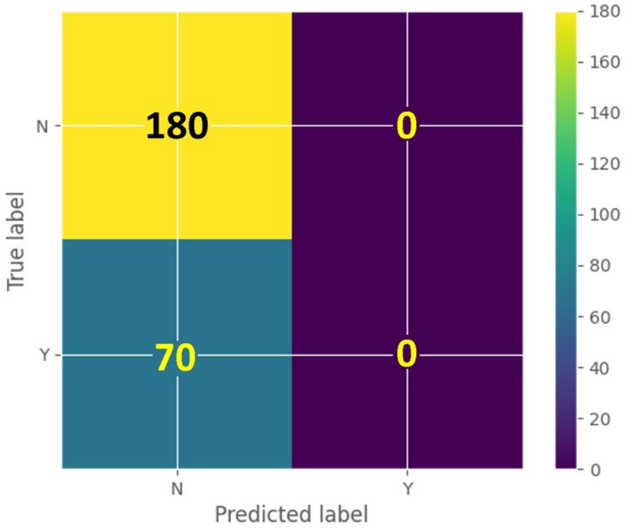

Figure 14. Confusion Matrix of SVC

3.1.11. Voting Classifier

The Voting Classifier is an adaptable ensemble learning method that integrates predictions from diverse individual models, which may encompass different algorithms like Support Vector Machines (SVM), Decision Trees, and Random Forests. It functions by consolidating predictions from each base model and determining the class label through majority voting for classification tasks or averaging for regression tasks (Bandi et al., 2023:522).

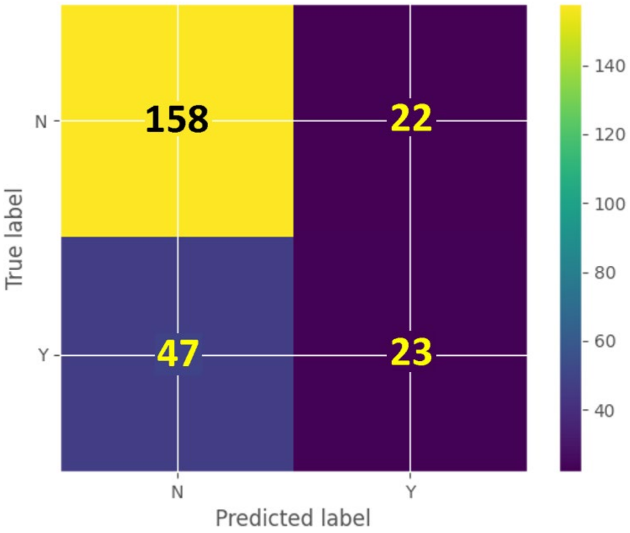

Figure 15. Confusion Matrix of Voting Classifier

3.1.12. GridSearchCV

GridSearchCV is a robust technique for hyperparameter tuning in machine learning. It aims to find the hyperparameters of a given algorithm that yield the best performance on a validation set.

By exhaustively searching through a specified parameter grid and systematically testing different configurations, GridSearchCV helps fine-tune model parameters, ultimately enhancing predictive accuracy (Sarang, 2023:103). Decision Tree and Ada Boost Classifiers were used in the article.

4. ANALYSIS AND EMPIRICAL FINDINGS

In our research on utilizing machine learning algorithms to detect fraudulent insurance claims in the insurance sector, we utilized a dataset comprising one thousand records. We compared the test and processing accuracies of the algorithms and conducted an assessment of their performance. The accuracy rates of all algorithms are detailed in Table 1.

Table 1. Algorithms Test Results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Training Accuracy (%)	Testing Accuracy (%)
Decision Tree	75.6	77.0	76.0	76.0	83.46	75.6
Random Forest	75.6	75.0	76.0	75.05	97.33	75.6
Extra Trees	72.8	71.0	73.0	69.36	100	72.8
Voting Classifier	72.4	70.0	72.0	70.2	93.86	72.4
SVC	72.0	52.0	72.0	60.27	85.6	72.0
KNN	72.0	52.0	72.0	60.27	77.20	72.0
LGBM	71.6	68.0	72.0	69.36	100	71.6
Ada Boost	68.0	73.0	68.0	69.36	83.46	68.0
Cat Boost	67.2	69.0	67.0	69.36	90.93	67.2
SGB	38.0	60.0	38.0	69.36	95.06	38.0
Gradient Boost	36.8	60.0	37.0	69.36	94.26	36.80

Conclusion from models, they are evident that the highest accuracy rate was achieved with the Random Forest and Decision Tree Classifiers.

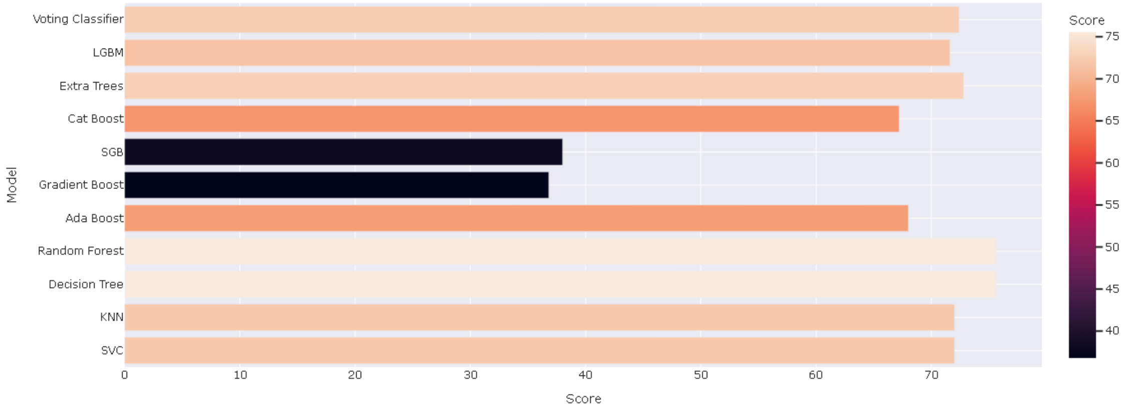

Figure 16. Models Comparison of Accuracy

Decision Tree Classifier with GridSearchCV the accuracy score of 75.6%. Here our model predicts 145 TP cases out of 180 positive cases and 44 TN cases out of 70 cases. It predicts 35 FP cases out of 180 positive cases and 26 FN cases out of 70 cases. It gives the F1-Score of 76.0%.

Random Forest Classifier with the accuracy score of 75.6%. Here our model predicts 154 TP cases out of 180 positive cases and 35 TN cases out of 70 cases. It predicts 26 FP cases out of 180 positive cases and 35 FN cases out of 70 cases. It gives the F1-Score of 75.05%.

The Decision Tree model involved fitting 5 folds for each of 512 candidates, totaling 2560 fits, and the resulting parameters were {'criterion': 'gini', 'max_depth': 3, 'min_samples_leaf': 8, 'min_samples_split': 2}. Moving on to the Random Forest, after creating the grid, the GridSearchCV model ran with RandomForestClassifier() as the estimator parameter, performing 5 folds for each of 60 candidates, totaling 300 fits. Finally, the AdaBoost model utilized {'algorithm': 'SAMME', 'learning_rate': 0.001, 'n_estimators': 90}. These efforts are contributing to the development of robust models for predictive analysis.

5. CONCLUSION AND RECOMMENDATION

The article discussed the development of a model for detecting auto insurance fraud using machine learning techniques. The model aims to reduce losses for insurance companies by accurately identifying fraudulent claims. The challenge of fraud detection in machine learning is addressed, emphasizing the rarity of fraudulent cases compared to legitimate claims.

The project utilized eleven different classifiers, including Random Forest and Decision Tree, with thorough analysis and default parameter values. The results indicated that these models performed best with the dataset, demonstrating their suitability for the task. The study concludes that with appropriate machine learning algorithms and thorough analysis, it is feasible to predict fraudulent claims to a significant degree using the available data.

The use of cross-validation and grid search offers significant benefits in model development, enhancing the robustness of the resulting models. However, it's important to consider that each iteration of the model, up to ‘K’ times, requires running the full model, which can become computationally expensive as the dataset grows larger and as the value of 'K' increases.

By leveraging cross-validation and grid search, meaningful results were achieved compared to the original train/test split, with minimal tuning. Cross-validation plays a vital role in creating better fitting models by training and testing on all parts of the training dataset.

In specific reference to the Decision Tree model, the selected hyperparameters led to a notable improvement, resulting in an accuracy score of 75.6% on the validation set. This performance indicates that the chosen hyperparameters have effectively enhanced the model's predictive capability, making it well-suited for deployment in real-world scenarios.

Insurance fraud is described as encompassing various improper activities aimed at obtaining favorable outcomes from insurance companies, highlighting the importance of accurately detecting fraudulent cases. Machine learning techniques are emphasized for enhancing predictive accuracy while maintaining low FP rates.

The model's high accuracy in distinguishing between fraudulent and legitimate claims is highlighted, emphasizing the role of machine learning in swiftly detecting fraud to minimize costs. However, the study acknowledges limitations such as a small sample size and the need for more comprehensive state data to capture incident claims effectively.

Addressing fraud and abuse in insurance companies requires a multi-faceted approach that involves preventive measures, detection strategies, and collaborative efforts. Here are some key recommendations and solutions: Utilize advanced data analytics and AI to detect patterns and anomalies indicating fraudulent activities, enabling early identification and prevention of fraudulent claims. Invest in fraud detection software and platforms that use machine learning algorithms and predictive modeling to identify unusual patterns and potential fraud indicators, enhancing the ability to detect fraudulent activities. Strengthen verification processes for policy issuance, claims, and policyholder information changes through background checks, documentation validation, and identity verification to decrease the likelihood of fraudulent activities. Empower insurance professionals by providing comprehensive education and training on identifying, reporting, and preventing fraud and abuse, fostering vigilance and proactive measures. Foster collaboration with law enforcement agencies, regulatory bodies, and industry associations to share intelligence, best practices, and coordinate efforts in combating insurance fraud. Launch customer awareness programs to educate policyholders about the consequences of insurance fraud and the importance of reporting suspicious activities, aiming to deter potential fraudsters and promote honest behavior. Ensure strict adherence to regulatory compliance and standards, including regular audits and checks to maintain integrity within the insurance industry. Foster an ethical company culture that promotes transparency, integrity, and a zero-tolerance policy towards fraudulent activities. Encourage employees to report any suspicions or irregularities.

REFERENCES

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., & Saif, A. (2022). Financial fraud detection based on machine learning: a systematic literature review. *Applied Sciences, 12*(19), 9637.

Au, T. C. (2018). Random forests, decision trees, and categorical predictors: the" absent levels" problem. *The Journal of Machine Learning Research, 19*(1), pp. 1737-1766.

Bandi, R., Likhit, M. S. S., Reddy, S. R., Bodla, S. R., & Venkat, V. S. (2023). Voting Classifier-Based Crop Recommendation. *SN Computer Science, 4*(5), 516. https://doi.org/10.1007/s42979-023-01995-8

Chakrabarty, N., Kundu, T., Dandapat, S., Sarkar, A., & Kole, D. K. (2019). Flight arrival delay prediction using gradient boosting classifier. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2,* pp. 651-659. https://doi.org/10.1007/978-981-13-1498-8_57

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends, 2*(01), pp. 20-28. https://doi.org/10.38094/jastt20165

Choi, J. M., Kim, J. H., & Kim, S. J. (2021). Application of Reinforcement Learning in Detecting Fraudulent Insurance Claims. *International Journal of Computer Science & Network Security, 21*(9), pp. 125-131.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *icml,* Vol. 96, pp. 148-156.

Geren, Y. (2020). Makine Öğrenmesi ile Sigorta Hasarlarında Sahtecilik Tespiti. *Turkish Studies-Information Technologies and Applied Sciences, 15*(2), pp. 195-209.

Goetz, M., Weber, C., Bloecher, J., Stieltjes, B., Meinzer, H. P., & Maier-Hein, K. (2014). Extremely randomized trees based brain tumor segmentation. *Proceeding of BRATS challenge-MICCAI,* 14, pp. 6-11.

Hanafy, M. O. H. A. M. E. D., & Ming, R. (2021). Using machine learning models to compare various resampling methods in predicting insurance fraud. *J. Theor. Appl. Inf. Technol, 99*(12), pp. 2819-2833.

Hanafy, M., & Ming, R. (2022). Classification of the Insureds Using Integrated Machine Learning Algorithms: A Comparative Study. *Applied Artificial Intelligence, 36*(1), 2020489.

Insurance Fraud Information System (SISBIS) Statistics, (2024) https://siseb.sbm.org.tr/tr/istatistikler

Itri, B., Mohamed, Y., Omar, B., & Mohamed, Q. (2020). Empirical oversampling threshold strategy for machine learning performance optimisation in insurance fraud detection. *International Journal of Advanced Computer Science and Applications, 11*(10).

Jones, K. I., & Sah, S. (2023). The Implementation of Machine Learning In The Insurance Industry With Big Data Analytics. *International Journal of Data Informatics and Intelligent Computing, 2*(2), pp. 21-38. https://doi.org/10.59461/ijdiic.v2i2.47

Kalra, H., Singh, R., & Kumar, T. S. (2022). Fraud Claims Detection in Insurance Using Machine Learning. *Journal of Pharmaceutical Negative Results,* pp. 327-331.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems, 30 (NIPS 2017)*, pp. 3146-3154.

Kramer, O. (2013). K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors,* pp. 13-23. https://doi.org/10.1007/978-3-642-38652-7_2

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv preprint arXiv:1912.06059*.

Liu, T., Jin, L., Zhong, C., & Xue, F. (2022). Study of thermal sensation prediction model based on support vector classification (SVC) algorithm with data preprocessing. *Journal of Building Engineering, 48*, 103919. https://doi.org/10.1016/j.jobe.2021.103919.

Muneer, A., & Fati, S. M. (2020). Efficient and automated herbs classification approach based on shape and texture features using deep learning. *IEEE Access*, *8*, pp. 196747-196764.

Naseer, S., Fati, S. M., Muneer, A., & Ali, R. F. (2022). iAceS-Deep: Sequence-based identification of acetyl serine sites in proteins using PseAAC and deep neural representations. *IEEE Access*, *10*, pp. 12953-12965.

Nordin, S. Z. S., Wah, Y. B., Haur, N. K., Hashim, A., Rambeli, N., & Jalil, N. A. (2024). Predicting automobile insurance fraud using classical and machine learning models. *International Journal of Electrical and Computer Engineering (IJECE), 14*(1), pp. 911-921.

Pranavi, P. S., Sheethal, H. D., Kumar, S. S., Kariappa, S., & Swathi, B. H. (2020). Analysis of Vehicle Insurance Data to Detect Fraud using Machine Learning. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 8*(7), pp. 2033-2038.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems,* 31.

Sarang, P. (2023). Ensemble: Bagging and Boosting: Improving Decision Tree Performance by Ensemble Methods. In *Thinking Data Science: A Data Science Practitioner's Guide*, pp. 97-129. https://doi.org/10.1007/978-3-031-02363-7_5

Sathya, M., & Balakumar, B. (2022). Insurance Fraud Detection Using Novel Machine Learning Technique. *International Journal of Intelligent Systems and Applications in Engineering, 10*(3), pp. 374-381.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian* *joint conference on artificial intelligence*, pp. 1015-1021. Berlin, Heidelberg: Springer Berlin Heidelberg.

Subudhi, S., & Panigrahi, S. (2020). Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University-Computer and Information Sciences, 32*(5), pp. 568-575.

Ustimenko, A., & Prokhorenkova, L. (2021). SGLB: Stochastic gradient langevin boosting. In *International Conference on Machine Learning,* pp. 10487-10496.