Enhancement of Backward Feature Elimination as a Pre-Processing Method for K Nearest Neighbor Algorithm Applied to Insurance Fraud Detection

Ma. Pauline Yvana B. Amores[1], Charisse Nicole P. Aberin[2], Vivien A. Agustin[3], Herminiño C. Lagunzad[4], Richard C. Regala[5], Raymund M. Dioses[6]

[1] *Student, College of Engineering - Pamantasan ng Lungsod ng Maynila, Philippines.* *mpybamores2019@plm.edu.ph*

[2] *Student, College of Engineering - Pamantasan ng Lungsod ng Maynila, Philippines.* *cnpaberin2019@plm.edu.ph*

[3] *Thesis Adviser, College of Engineering - Pamantasan ng Lungsod ng Maynila, Philippines.* *vaagustin@plm.edu.ph*

[4] *Thesis Adviser, College of Engineering - Pamantasan ng Lungsod ng Maynila, Philippines.* *hclagunzad@plm.edu.ph*

[5] *Thesis Adviser, College of Engineering - Pamantasan ng Lungsod ng Maynila, Philippines.* *rcregala@plm.edu.ph*

[6] *Thesis Adviser, College of Engineering - Pamantasan ng Lungsod ng Maynila, Philippines.* *rmdioses@plm.edu.ph*

Abstract: This study presents novel enhancements to the Backward Feature Elimination (BFE) method for improved insurance fraud detection using the K-Nearest Neighbor (KNN) algorithm. The research addresses issues inherent in the baseline BFE process, such as the over-reliance on p-values, the potential for misleading results, and suboptimal feature selection leading to overfitting. To address these, the study integrates confidence intervals and feature importance into the BFE process, establishing a more robust and reliable criterion for feature selection. Moreover, feature engineering techniques are introduced during preprocessing to enhance model performance. The modified BFE method demonstrates superior performance over the baseline model regarding the recall, precision, and F1 score. Stratified K-Fold Cross-Validation, ROC-AUC Score, and Coefficient of Variation (CV) confirm the consistency and robustness of the enhanced model across varying data subsets. These innovations offer a comprehensive and reliable solution to feature selection in the BFE method, applied to the KNN model for effective insurance fraud detection. The study mitigates the issues related to p-value dependence and boosts model performance, paving the way for more accurate and robust fraud detection systems.

Keywords: Backward Feature Elimination, Confidence Interval, Fraud Detection, K-Nearest Neighbor Algorithm.

1. INTRODUCTION

Insurance fraud incurs substantial financial losses for the industry each year. The Philippines Insurers and Reinsurers Association (2018) reported car insurance scams in the Philippines costing over a billion pesos annually [1]. Furthermore, the Coalition Against Insurance Fraud (2022) estimated that fraud costs businesses and consumers $308.6 billion annually [2]. Consequently, data analytics and machine learning algorithms have emerged as practical

tools for combating insurance fraud, with the K-nearest neighbors (KNN) algorithm being widely employed due to its ability to identify patterns indicative of fraudulent activities.

However, the effectiveness of the KNN algorithm can be influenced by the number and relevance of the features used to make predictions. Identifying the most pertinent features while eliminating the least significant ones is vital in enhancing the algorithm's performance. Feature selection is a crucial concept in machine learning, enabling the construction of models with essential and relevant features that significantly impact their predictive accuracy.

Backward Feature Elimination (BFE) is a popular method for feature selection, involving the iterative removal of the least significant features from the dataset [3]. This technique provides insights into the importance of each feature, improving the overall performance and efficiency of machine learning models [4], yet still poses some constraints.

The traditional BFE method relies solely on p-values to assess the statistical significance of features, which has limitations and may lead to inaccurate results [5]. In this study, the researchers propose an enhanced version of BFE for the KNN algorithm in detecting fraudulent claims, integrating confidence intervals as a more reliable measure of statistical significance [6]. Moreover, the proponents introduce the consideration of feature importance that aligns with the nature of the study, alongside p-values, as another criterion for feature selection.

The objective of this research is to evaluate the performance of the enhanced BFE method, incorporating confidence intervals and feature importance, compared to the baseline BFE method that relies only on p-values. The findings of this study have the potential to revolutionize feature selection for the KNN algorithm in insurance fraud detection, offering practical implications for the industry's fraud detection approaches.

2. RELATED WORKS

Several studies and works have emphasized the importance of data preprocessing in machine learning for fraud detection tasks [7]. Preprocessing techniques, including handling missing values, noise reduction, and resolving inconsistencies, enhance data quality and improve the accuracy and reliability of machine learning models [7].

Combating fraudulent claims is a challenge for insurance firms, and using digital advances is essential in this regard [8]. With the use of machine learning algorithms, insurers are now able to process massive amounts of data, spot hidden patterns, and detect fraudulent trends in claim processing and customer background checks [8]. This helps insurers save a lot of money by reducing the number of fraudulent claims.

The KNN algorithm has gained attention for its simplicity and effectiveness in various problem domains, including fraud detection [9]. KNN is an effective option for spotting fake patterns because it relies on feature proximity and similarity. Additionally, KNN's performance can be improved by using feature selection techniques to lower the dimensionality of highly dimensional data [10].

BFE has been presented as a technique for minimizing features. Feature selection plays a crucial role in optimizing machine learning models [11]. It streamlines models, enhances performance, and pinpoints the most crucial aspects by iteratively deleting the least important or relevant features [11].

However, several academics have emphasized the drawbacks of merely using p-values to understand trial outcomes [5], [12], [13], [14]. Various researchers stress the utility of confidence intervals because they offer a range of tenable values and transmit in-depth details about the nature, magnitude, and clinical significance of an effect [5], [12], [13], [14]. Confidence intervals offer valuable insights beyond binary significance, facilitating a more meaningful interpretation of research results.

3. PROPOSED METHOD

To address the limitations of Backward Feature Elimination Method in its disregard of effect size and confidence intervals, the unreliability and sensitivity of p-values, and the potential for suboptimal feature selection and overfitting. These limitations highlight the need for enhancements in the feature selection process to address these issues and improve the effectiveness of the BFE method as a pre-processing method for KNN algorithm. The BFE method is enhanced in two ways; considering important features before feature selection and applying confidence intervals with p-values to T ensure a more reliable and robust feature selection process. The proposed work is shown in Fig. 1.

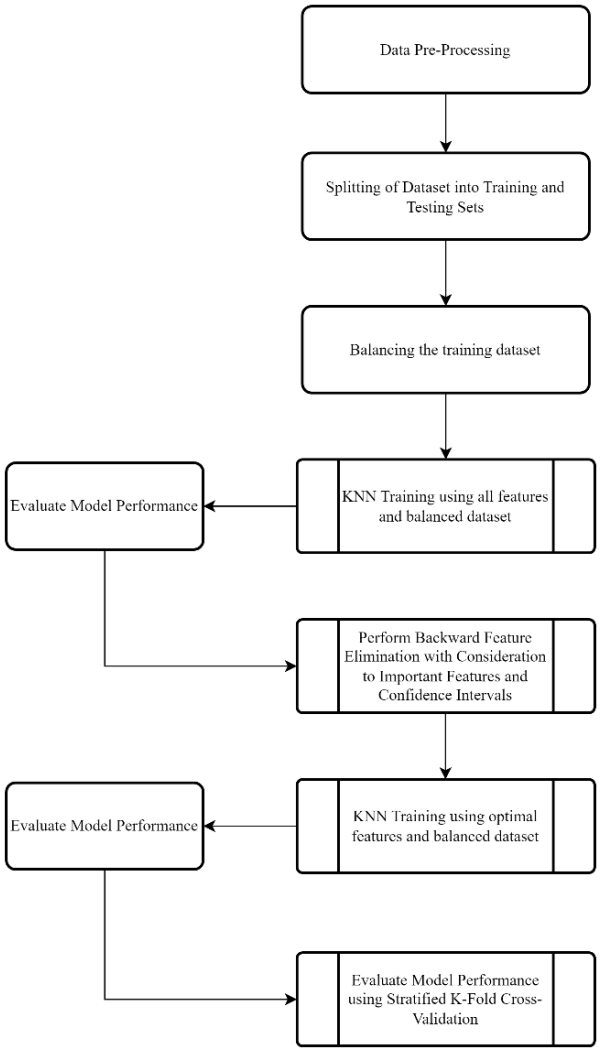

Fig. 1: Proposed Architecture

A. Data Set Information

This study utilized the "Auto Insurance Claims Data" dataset, initially uploaded by Shah (2018) on the Kaggle website. The dataset contains 40 columns, each representing different features related to auto insurance claims. The dataset provided the necessary information to evaluate the effectiveness of the enhanced Backward Feature Elimination (BFE) method in conjunction with the K-Nearest Neighbors (KNN) algorithm for insurance fraud detection. It provided a realistic set of features, enabling a thorough evaluation of the proposed enhancements and their impact on the model's overall performance.

B. Attribute Information

The attributes of the dataset underwent pre-processing methods to reduce complexity and dimensionality. These steps included cleaning the data, handling missing values, converting categorical variables into numerical ones, normalizing numerical values, and reducing dimensionality. From 40 initial features, the dataset was reduced to 29 features, which retained the most meaningful information while minimizing computational complexity and redundancy. This refined dataset offers a solid foundation for predictive modeling, enhancing the ability to discern the factors most impactful in predicting insurance fraud.

C. Baseline Backward Feature Elimination (BFE)

Backward Feature Elimination (BFE) is a well-known machine learning technique used to enhance models' performance by selecting the most relevant features.

Effective feature selection can considerably improve the accuracy of predictive models such as the K-nearest neighbors (KNN) algorithm in the context of insurance fraud detection. As shown in Fig. 2, BFE involves training a model iteratively with all available features, removing one feature at a time, and comparing the model's performance. This procedure is repeated until no additional features can be eliminated without significantly degrading model performance [15].

Step-1: Firstly, we need to select a significance level to stay in the model. (SL=0.05)

Step-2: Fit the complete model with all possible predictors/independent variables.

Step-3: Choose the predictor which has the highest P-value, such that.

a. If P-value >SL, go to step 4.

b. Else Finish, and Our model is ready.

Step-4: Remove that predictor.

Step-5: Rebuild and fit the model with the remaining variables.

Fig. 2: Pseudocode of Baseline BFE method

D. Enhanced Backward Feature Elimination (BFE)

Considering the discussion from earlier in the introduction, the current conventional approach to BFE has several drawbacks that could reduce the efficacy of the model. As shown in Fig. 3, the researchers devised a revised pseudocode to address these issues.

Step-1: Initialized the list of important features with a feature deemed important. (e.g., 'incident_severity')

Step-2: Create a copy of the important features list (e.g., 'best_features') and features that are not on the important features list (e.g., 'other_features')

Step-3: Set the p-value threshold (e.g., 0.05) and confidence level (e.g., 0.95)

Step-4: Initialize a variable for previous number of features with a number larger than the important features list.

Step-5: Perform the feature elimination loop while other_feature list has remaining features:

a. Add a constant term to the predictor set by concatenating best_features and other_features with the training data.

b. Fit an Ordinary Least Squares (OLS) model using the training labels and the predictor set.

c. Retrieve the p-values and confidence intervals of the features from the OLS model.

d. Identify the feature with the maximum p-value and its confidence interval does not include 0.

1. If yes, remove the feature from other_features.

2. If not, move the feature from other_features to best_features.

e. Update previous number of features with the current number of other_features.

f. Break the loop if there is no change in the number of other_features.

g. Break the loop if all the remaining features have p-values below the threshold.

Step-6: If there are selected features (e.g., best_features) available:

a. Train the KNN algorithm using the optimal feature set (e.g., best_features) and the resampled training data.

b. Print the selected features.

Step-7: If no feature satisfies the p-value threshold:

a. Set the first feature from the original feature list as the default feature.

b. Train the KNN algorithm using the default feature and the resampled training data.

Fig. 3: Pseudocode of Enhanced BFE method

The enhanced version of Backward Feature Elimination (BFE) method integrates several key statistical concepts and techniques to improve the selection of important features in a predictive model. It involves computing a score for each input feature in the model, which represents the "importance" or effect of the feature on the model's predictions.

Additionally, features with high correlation are more linearly dependent and tend to have similar effects on the dependent variable. Confidence intervals and p-values are also integrated into the BFE method, providing statistical evidence to guide feature selection.

Finally, p-values are used to test the statistical significance of each feature. This process is repeated until no more features can be removed, resulting in a final set of important features.

4. RESULTS AND DISCUSSION

The performance and effectiveness of the proposed modification to the BFE method for KNN algorithm applied to insurance fraud detection were analyzed using several performance evaluation metrics and validations.

A. Performance Evaluation Metrics

To know if the enhanced model significantly improved, various evaluation metrics such as average accuracy, average precision, average recall, average F1-score, and ROC-AUC score were used compared to the baseline model.

Table I: Comparison of Performance Evaluation Metrics of Enhanced and Baseline Algorithm

	Performance Evaluation Metrics				
	Average Accuracy	**Average Precision**	**Average Recall**	**Average F1-score**	**ROC-AUC Score**
KNN with Baseline BFE	0.604	0.557	0.563	0.560	0.479
KNN with Enhanced BFE	0.568	0.843	0.754	0.796	0.753

The outcomes of the feature selection process utilizing the Backward Feature Elimination (BFE) method were compared to a baseline method, enhanced by the inclusion of confidence intervals and feature importance alongside p-values. This comparison aimed to assess the enhanced method's effectiveness in improving the model's performance for insurance fraud detection.

As seen, Table I presents the evaluation metrics used to assess the performance. The baseline approach had an average accuracy of 0.604, meaning 60.4% of the occurrences were identified correctly. However, the average accuracy decreased slightly to 0.568. While the decrease in accuracy may seem concerning, it is crucial to consider the impact on other evaluation metrics.

Including confidence intervals and feature importance alongside p-values in the feature selection process led to notable improvements in several evaluation metrics. The average precision, which measures the proportion of correctly predicted positive instances out of all instances predicted as positive, significantly increased from 0.557 in the baseline approach to 0.843 in the enhanced approach. This improvement indicates that the enhanced feature selection method was able to identify better and classify instances related to insurance fraud.

Similar improvements were seen in the average recall, which increased from 0.563 in the baseline to 0.754 in the enhanced approach. The average recall reflects the proportion of adequately predicted positive cases out of all positive instances. This improvement shows that a higher proportion of positive incidents were captured by the enhanced feature selection method, essential for effectively identifying insurance fraud.

A balanced criterion known as the average F1 score, which considers both precision and recall, increased considerably from 0.560 in the baseline to 0.796 in the enhanced approach. This improvement suggests that the enhanced feature selection method achieved a better balance between correctly predicting positive instances and minimizing false positives.

Additionally, the ROC-AUC score, which evaluates how well the model differentiates between positive and negative instances, increased noticeably from 0.479 in the baseline to 0.753 in the enhanced method. The ROC-AUC results

suggest that the model's capacity to distinguish between fraudulent and legitimate insurance claims was improved by the enhanced feature selection method.

B. Performance Validation

The researchers garner the results of performance validation of each model using Stratified K-Fold Cross-Validation to further identify the overall execution of the enhanced model compared to the baseline model.

Table II: Comparison of Baseline and Enhanced Algorithm with Stratified K-Fold Cross Validation

Performance Validation Metrics	KNN with Baseline BFE	Standard Deviation	KNN with Enhanced BFE	Standard Deviation
Accuracy	0.558	0.030	0.807	0.027
Precision	0.557	0.027	0.844	0.033
Recall	0.563	0.059	0.754	0.038
F1-score	0.559	0.041	0.796	0.030

Table II presents a comparison of the Baseline and Enhanced algorithms using K-nearest neighbors (KNN) with Stratified K-Fold Cross Validation. The results clearly demonstrate the significant improvements achieved through the application of the enhanced Backward Feature Elimination (BFE) method.

In terms of accuracy, the Baseline algorithm achieves an accuracy of 0.558, with a standard deviation of 0.030. However, the Enhanced algorithm shows a substantial improvement with an accuracy of 0.807 and a lower standard deviation of 0.027. This indicates that the Enhanced algorithm consistently performs better and achieves a higher level of correct classification.

Similarly, the precision of the Baseline algorithm is 0.557 with a standard deviation of 0.027. In contrast, the Enhanced algorithm demonstrates a remarkable increase in precision, reaching 0.844, while maintaining a lower standard deviation of 0.033. This improvement suggests that the Enhanced algorithm excels in correctly classifying positive instances, reducing false positives significantly.

Regarding recall, the Baseline algorithm achieves a recall of 0.563 with a standard deviation of 0.059. Once again, the Enhanced algorithm outperforms the Baseline, achieving a recall of 0.754, and showcasing a lower standard deviation of 0.038. This implies that the Enhanced algorithm consistently identifies a higher proportion of actual positive instances, reducing false negatives.

Finally, the F1-score of the Baseline algorithm is 0.559, with a standard deviation of 0.041. On the other hand, the Enhanced algorithm demonstrates a noteworthy improvement, achieving an F1-score of 0.796, with a lower standard deviation of 0.030. This signifies that the Enhanced algorithm strikes a better balance between precision and recall, resulting in a more robust and reliable performance.

The results and discussion clearly support the efficacy of the proposed enhancements to the BFE method applied to the KNN algorithm for insurance fraud detection. The Enhanced algorithm consistently outperforms the Baseline algorithm in all metrics, including accuracy, precision, recall, and F1-score. These findings validate the enhancements made to the feature selection process, improving the model's ability to detect insurance fraud effectively.

The observed improvements in performance and reduced standard deviations indicate that the Enhanced algorithm is not only more accurate but also more stable and reliable across different data subsets. This enhanced stability enhances the algorithm's potential for generalization to new and unseen data, making it a valuable tool for real-world insurance fraud detection applications.

Table III: Coefficient of Variation for Baseline and Enhanced Algorithm

Performance Validation Metrics	KNN with Baseline BFE	KNN with Enhanced BFE
Accuracy	0.0543	0.0339
Precision	0.0491	0.0386
Recall	0.1040	0.0504
F1-score	0.0731	0.0381

The findings presented in Table III suggest that the Enhanced model shows a higher level of consistency in its performance metrics, including accuracy, precision, recall, and F1 score, compared to the Baseline model. This observation is supported by the calculated coefficients of variation (CV) for each metric. The Enhanced model exhibits lower CV percentages for all metrics, indicating a more stable and reliable performance.

Analyzing the results for each metric, we find that the Enhanced model's accuracy has less variation (3.39%) compared to the Baseline model's accuracy (5.43%). This suggests that the Enhanced model consistently achieves a higher level of correct classification.

Similarly, the Enhanced model's precision demonstrates lower variation (3.86%) compared to the Baseline model's precision (4.91%). This implies that the Enhanced model consistently performs better in correctly classifying positive instances and reducing false positives.

In terms of recall, the Enhanced model exhibits lower variation (5.04%) compared to the Baseline model (10.40%). This indicates that the Enhanced model consistently identifies a higher proportion of actual positive instances, resulting in fewer false negatives.

Furthermore, the Enhanced model's F1 score shows less variation (3.81%) compared to the Baseline model's F1 score (7.31%). This implies that the Enhanced model maintains a more consistent balance between precision and recall.

Overall, the findings suggest that the Enhanced model outperforms the Baseline model in terms of both mean performance and stability across different data subsets. The lower coefficients of variation for all metrics indicate less variability around the mean, highlighting the Enhanced model's reliability and robustness when applied to novel data. The balanced approach to precision and recall in the Enhanced model indicates an improved compromise between false positives and false negatives.

In summary, the Enhanced model's superior performance, stability, and consistency make it a promising choice for insurance fraud detection, providing more dependable results and enhancing the potential for accurate predictions in real-world scenarios.

5. CONCLUSION AND RECOMMENDATION

In conclusion, the proposed enhancements to the Backward Feature Elimination (BFE) method applied to the K-nearest neighbors (KNN) algorithm for insurance fraud detection have successfully addressed the limitations of the baseline model. By incorporating confidence intervals and considering feature importance, the enhanced method provides a more reliable and robust feature selection process, resulting in improved model performance and the ability to detect insurance fraud more effectively.

Based on the positive outcomes of this study, it is recommended that future research in the field of insurance fraud detection continues to build upon these enhancements. Researchers should explore alternative feature scaling techniques, such as standardization, normalization, or robust scaling, to further refine the preprocessing stage and improve the KNN algorithm's performance and accuracy in fraud detection.

Additionally, the use of hyperparameter tuning should be employed to optimize the KNN algorithm's parameters. This will ensure that the algorithm is configured in the most effective way, leading to enhanced fraud detection capabilities.

Furthermore, it is recommended to extend the application of the proposed enhancements to other classification algorithms, including random forest, naive Bayes, support vector machines, and others. This broader investigation will provide a comprehensive understanding of the performance and generalizability of the enhancements across different classifiers.

By pursuing these recommendations, future research can contribute to the advancement of fraud detection systems in the insurance industry. The accuracy and efficiency of these systems will be improved, leading to more effective identification and prevention of fraudulent activities. Ultimately, these efforts will help protect insurance companies from financial losses and ensure fair and reliable services for policyholders.

ACKNOWLEDGEMENT

Our sincere appreciation goes to everyone who has helped us throughout this research project. We are incredibly grateful to have been able to work with such a skilled and dedicated group of individuals and receive their invaluable assistance and support.

6. REFERENCES

[1] E. Reyes and Author: erwin reyes Erwin has a combined experience of more than 15 years in the car insurance industry in the Philippines and Australia. Loves cars and enjoys to sourcing out great deals for its clients, "Insurance fraud in the Philippines that you should be aware of," iChoose.ph, https://ichoose.ph/blogs/insurance-fraud-philippines-aware/ (accessed May 23, 2023). J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.

[2] Coalition Against Insurance Fraud, "Fraud stats," InsuranceFraud.org, https://insurancefraud.org/fraud-stats/ (accessed May 23, 2023).

[3] P. Pedemkar, "Backward elimination: How to apply backward elimination?," EDUCBA, https://www.educba.com/backward-elimination/ (accessed May 23, 2023).

[4] M. Shah, "Machine learning: Feature selection with backward elimination," Medium, https://medium.com/@mayankshah1607/machine-learning-feature-selection-with-backward-elimination-955894654026 (accessed May 23, 2023). M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] S. C. Karpen, "P value problems," American journal of pharmaceutical education, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5738950/#:~:text=Indeed,%20a%202016%20joint%20statement,or%20range%20of%20an%20effect (accessed May 23, 2023).

[6] M. J. Gardner and D. G. Altman, "Confidence intervals rather than P values: Estimation rather than hypothesis testing," British medical journal (Clinical research ed.), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1339793/ (accessed May 23, 2023). Punith, "Insurance claims‐fraud detection using machine learning," Medium, https://medium.com/geekculture/insurance-claims-fraud-detection-using-machine-learning-78f04913097 (accessed May 23, 2023).

[7] Punith, "Insurance claims‐fraud detection using machine learning," Medium, https://medium.com/geekculture/insurance-claims-fraud-detection-using-machine-learning-78f04913097 (accessed May 24, 2023).

[8] "Insurance fraud detection using machine learning: What you should know: News," managementevents.com, https://managementevents.com/news/insurance-fraud-detection-using-machine-learning-what-you-should-know/ (accessed May 23, 2023).

[9] Mahima, "Features of KNN algorithm," Edureka Community, https://www.edureka.co/community/46176/features-of-knn-algorithm (accessed May 23, 2023).

[10] J. Brownlee, "K-nearest neighbors for Machine Learning," MachineLearningMastery.com, https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning (accessed May 23, 2023).

[11] "Backward Elimination in Machine Learning - Javatpoint," www.javatpoint.com, https://www.javatpoint.com/backward-elimination-in-machine-learning (accessed May 23, 2023).

[12] N. Pandis, "Confidence Intervals rather than P values," Redirecting, https://doi.org/10.1016/j.ajodo.2012.11.012 (accessed May 23, 2023).

[13] S. K. Das, Published online October 2019 in IJEAST (http://www.ijeast.com ..., https://www.ijeast.com/papers/278-282,Tesma406,IJEAST.pdf (accessed May 22, 2023).

[14] H. T. Davies and I. K. Crombie, "What are confidence intervals and p-values?," Bandolier - Evidence based thinking about health care, https://www.bandolier.org.uk/painres/download/whatis/What_are_Conf_Inter.pdf (accessed May 23, 2023).

[15] H. Singh, "Backward feature elimination and its implementation," Analytics Vidhya, https://www.analyticsvidhya.com/blog/2021/04/backward-feature-elimination-and-its-implementation/ (accessed May 23, 2023).

DOI: https://doi.org/10.15379/ijmst.v10i1.2685