4th International Engineering Conference (IEC 2022)
Federal University of Technology, Minna, Nigeria

MACHINE LEARNING MODELS FOR RISK MANAGEMENT IN NIGERIAN CUSTOMS: AN INVESTIGATIVE PERFORMANCE ANALYSIS

*Aisha M. K. N[1], Alhassan, J. K[1], Aliyu, H. O[1], & Abdullahi, I. M[2]
[1]Department of Computer Science, Federal university of Technology, Minna.
[2]Department of Computer Engineering, Federal university of Technology, Minna.
*Corresponding author email:eesha006@yahoo.com +2348063954956

ABSTRACT
Customs administrations utilize risk analysis to identify which people, products, and modes of transportation should be scrutinized and to what extent. Risk analysis and risk assessment are analytical techniques for determining which risks are the most significant and should be treated first or have corrective action performed first. Several ML models were investigated to determine the suitable model for custom data. This is necessary due to the unavailability of such research work. The Machine Learning (ML) models considered are; Support Vector Machine (SVM), Decision Tree (DT) classifier, K-Nearest Neighbor (KNN), Ensemble and Discriminant analysis classifiers. The dataset was collected and pre-processed. The Models were trained and tested using 70% of data for training and 30% for testing. The result shows that the ensemble models produce the highest accuracy of 66.6% for Boosted Trees classifier when compared with the other models. The medium and coarse tree produced an accuracy of 66.1%. This shows that the tree-based algorithms performs averagely better than others and recommended for further exploration.

Keywords: *Artificial Intelligence; Consignment Risk; Customs Management; Machine Learning*

1 INTRODUCTION

The emergence of big data provides firms with the chance to better understand their customers, develop revenue-generating initiatives, and create new business models. However, only a small percentage of data acquired by businesses is evaluated. This situation opens a gap that could deprive established businesses additional revenue and put their long-term survival in jeopardy if new market entrants use it. Data is analyzed by intelligence-driven organizations (Kavoya, 2020). Customs is one of such organizations that generates huge amount of data that requires analysis for better revenue generation. The activities of Customs all over the world are vital due to their ability to generate huge revenue and boost the economies of their countries. In Nigeria for instance, the Nigerian Customs Service (NCS) generated about 2.3 trillion Naira in 2021, an amount well above the estimated target of 1.679 trillion Naira (Vanguard, Dec. 20, 2021). Despite these successes, several challenges still hinder the agency from reaching its full potential. These challenges include; man power shortage, slow digitization processes and the unrelenting efforts at smuggling and corruption.

The amended Kyoto Convention recommends limiting intrusive customs inspections, and this is also a proposal being addressed as part of the World Trade Organization's (WTO) trade facilitation negotiations (Laporte, 2011). To limit such inspections, more modern administrations use computerized data interchange and risk analysis to intervene at all stages of the customs chain, focusing their resources on a posteriori control. Developing country customs administrations are lagging behind in this regard. As a result, risk analysis appears to be a top concern for developing countries as they modernize their customs systems. Because of the huge volume of export, import, and transit transactions, many Customs administrations utilize risk analysis to identify which people, products, and modes of transportation should be scrutinized and to what extent. Risk analysis and risk assessment are analytical techniques for determining which risks are the most significant and should be treated first or have corrective action performed first (Bezabeh, 2019). Risk management is the systematic use of management systems and practices to provide Customs with the information they need to address movements or consignments that pose a risk. Risk management's major goal is to determine if a shipment requires physical inspection, documented checks, or immediate release. Some of the automated methods used include statistical scoring and rule-based methods which fails as the data volume increases (Regmi & Timalsina, 2018). Data mining has been widely used for risk management.

Data mining is a flexible technique that have grown rapidly over decades which is being used by corporate organizations to extract data and other valuable details, patterns whence large data sets is concerned. Moreso, data mining techniques has seen wide adoption and application in various domains with the sole aim of facilitating daily activities and ease human burdens Zhang et al., (2022). The use machine learning and data mining techniques for risk management have only been given little attention. Hence, in order to explore the advantages of data mining and machine learning techniques for effective risk management. This paper presents an investigative analysis of several machine learning models for risk management in Nigerian Customs with a view of finding the most suitable model.

2 DATA, METHODOLOGY AND EVALUATION

The methodology adopted to achieve the aim of this paper is shown in Figure. Dataset were collected and preprocessed, followed by data partitioning and ML model designs. Each of the model is trained, tested and performance evaluated.

2.1 Data

The dataset used in this paper is the trade record from the single good declaration of the Nigerian Custom service (NCS) which can be obtained from NCS website. The dataset comprised of over 6 million records collected up to 2019. The attributes of the data described the nature of goods to be imported and include information such as: the importers name, the method of importation which can either be by road, sea, or air, the declarants name, the item number, quantity, net and gross mass, price, value, tax, invoice and description. Finally, the datasets contain the category of the record being high risk or low risk item. Figure 1 shows a snap shot some samples of the dataset and their attributes without the importers and declarants details.

CTY_ORIGIN	MODE_OF_TRANSPORT	HS_DESC	ITM_NUM	QTY	GROSS_MASS	NET_MASS	ITM_PRICE	STATISTICAL_VALUE	TOTAL_TAX	INVOI
MANY	Road transport	35	0.2	1	2000	2000	404240	404240	744843	1734
Italy	Road transport	20	1	1	6250	6250	100417	100417	452499	1346
Germany	Sea transport	5	0.75	100	4000	4000	552510	623809	400330	5264
United Kingdom	Road transport	35	1	1	1637	1637	663084	663084	297379	663
United States	Road transport	35	1	1	14042	14042	1145764	1142855	512544	1145
China	Road transport	10	0.33333333	450	10000	10000	2934000	3391378	900051	9258
China	Road transport	5	0.66666667	250	5000	5000	2144450	2569970	0	6681
Saudi Arabia	Road transport	5	1	30	20000	20000	13652880	16138921	2000175	18060
Japan	Road transport	10	0.25	64	2000	2000	1304000	1458850	438881	3145
China	Road transport	20	0.42857143	322	3000	3000	3010808	3734842	3694063	10324
United Arab Emirates	Air transport	20	0.5	26	1000	1000	97800	134007	90029	228

Figure 1: A snapshot of the data sample

To test the machine learning models, about 5000 samples were selected that cut across low and high-risk samples.

2.2 Methodology

The dataset was preprocessed by removing unwanted attributes manually and converting the textual attributes to numerical values. This is followed by data normalization and partitioning. The data was partitioned into training and testing data in the ratio of 70:30 respectively. Five popular ML models were selected to be investigated. The essence is to determine the best model for determining the risk of consignments. The ML models considered are; Support Vector Machine (SVM), Decision Tree (DT) classifier, K-Nearest Neighbor (KNN), Ensemble and Discriminant analysis classifiers.

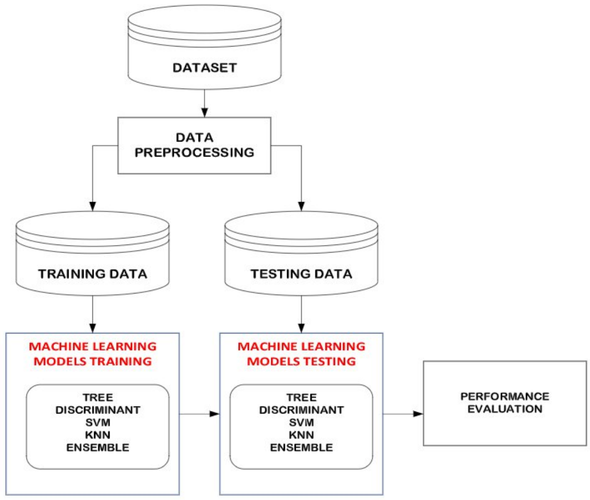

Figure 2: Model Investigation Methodology

For each model, different types were investigated, such as; Fine, Medium and Coarse Tree models, Linear and Quadratic discriminant models, Linear, Quadratic, Cubic, Fine gaussian, Medium gaussian and Coarse gaussian SVM models; Fine, Medium, Coarse, Cosine, Cubic and Weighted KNN, and finally, Boosted, Bagged and RUSBoosted Ensemble Tree models.

Experiments were performed for each ML model using the partitioned dataset to train and test each model. The performance of the models was evaluated using accuracy, True Positive Rate (TPR) and False Negative Rate (FNR).

These metrics were calculated from the True Positive (TP), True Negative (TN), False positive (FP) and False Negative (FN) values obtained from the confusion matrices. Accuracy shows the percentage of samples correctly classified for risky and non-risky samples. The TPR shows the ability of the model to detect risky consignment, while the FNR shows the number of risky samples wrongly classified as not risky. To evaluate the performance of the proposed model, the following metrics will be used.

i. Accuracy:

The number of consignment correctly classified divided by the total number of classified consignments.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \qquad (3.1)$$

ii. True Positive Rate:

The proportion of positive classifications that are truly positive.

$$TPR = \frac{TP}{TP + FP} \times 100\% \quad (3.2)$$

iii. False Positive Rate:
The proportion of actual Positives that are correctly classified.

$$Recal = \frac{FP}{FP + TN} \times 100\% \quad (3.3)$$

3 RESULTS AND DISCUSSION

The number TP, TN, FP and FN obtained for each model type are shown in Table 4. From Table 4, Fine KNN produced the highest TP of 600 while Linear SVM produced the least with 265 resulting in the lowest and highest FN of 400 and 735 respectively. Cubic SVM has the highest FP of 441 compared to Linear SVM with just 32 FP.

Table 4: Confusion matrix

Models		TP	TN	FP	FN
Tree	Fine Tree	530	766	234	470
	Medium Tree	500	822	178	500
	Coarse Tree	504	818	182	496
Discriminant	Linear Discriminant	468	779	221	532
	Quadratic Discriminant	280	946	54	720
SVM	Linear SVM	265	968	32	735
	Quadratic SVM	504	731	269	496
	Cubic SVM	435	559	441	565
	Fine Gaussian SVM	537	715	285	463
	Medium Gaussian SVM	449	831	169	551
	Coarse Gaussian SVM	266	967	33	734
KNN	Fine KNN	600	616	384	400
	Medium KNN	514	766	234	486
	Coarse KNN	530	720	280	470
	Cosine KNN	512	768	232	488
	Cubic KNN	501	782	218	499
	Weighted KNN	582	659	341	418
ENSEMBLE	Boosted Trees	481	852	148	519
	Bagged Trees	569	703	297	431
	RUSBoosted Trees	500	822	178	500

From the TP, TN, FP, and FN values, the accuracy, TPR and FPR were calculated and shown in Figure 3 and Figure 4. Figure 3 is the bar chart of the accuracy of all the investated models. The result shows that the ensemble models produced the highest accuracy of 66.6% for Boosted Trees and 66.1% for RusBoosted Trees. Cubic SVM produced the lowest accuracy of 49.7%. Figure 4 shows the TPR and FNR of all the models. The Fine KNN model produces the highest TPR (60%) than any other while the Linear SVM model produces the lowest TPR of 26.5%.

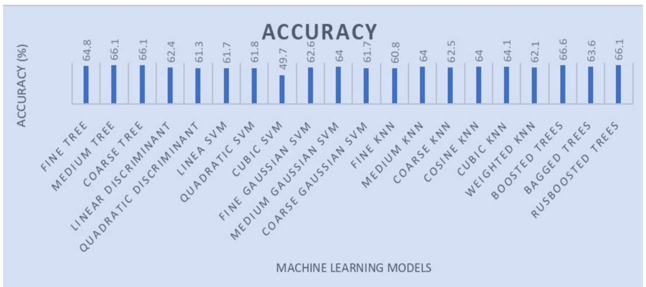

Figure 3: Models accuracy Bar chart

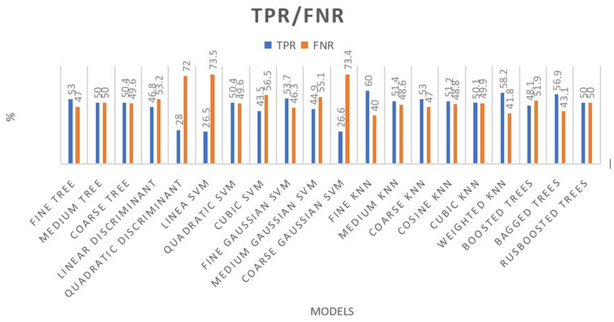

Figure 4: TPR and FPR Bar chart

4 CONCLUSION

Several machine learning models were investigated to determine the suitability of the models for detection of consignment risks in Customs. Datasets were collected, preprocessed and partitioned into training and testing. The models were trained and tested using the datasets. The performance each model was evaluated using the accuracy and TPR/FNR measure. The ensemble models performed better in terms of accuracy for both risks and non-risk consignments. The KNN model produces highest positive detection rate while SVM produces lowest positive detection rate. The results indicate that ensemble and KNN models can be recommended for adoption and further investigation.

FUTURE RESEARCH DIRECTION

To improve the performance of the models, it can be recommended that more pre-processing of the data be carried out using other pre-processing techniques and data balancing approaches. Also, other data mining approach such as feature selection using appropriate metaheuristic algorithms like Pastoralist optimization Algorithm (POA) be carried out to improve the performance of the models. Finally, other performance measures can be checked such as; precision, recall and F1-score.

ACKNOWLEDGEMENTS

I give thanks to The Almighty for making this work possible. I also acknowledge the effort of my infadigable supervisor, Prof. J. K. Alhassan for taking his time out of his tight schedules to criticize my work constructively and also offering assistance by giving me a guide to follow. My acknowledgment is also extended to Dr Abdullahi and Dr Aliyu who are co-authors of this work.

REFERENCE

Bhero, E., & Hoffman, A. (2017). Improving customs infraction-detection and risk management with GPS data. *2017 IEEE AFRICON: Science, Technology and Innovation for Africa, AFRICON 2017*, 1534–1540. https://doi.org/10.1109/AFRCON.2017.8095710

Hammadi, L., Ouahman, A. A., De Cursi, E. S., & Ibourk, A. (2018). An approach based on FMECA methodology for a decision support tool for managing risk in customs supply chain: A case study. *International Journal of Manufacturing Technology and Management*, *32*(2), 102–123. https://doi.org/10.1504/IJMTM.2018.090060

Hartama, D., Perdana Windarto, A., & Wanto, A. (2019). The Application of Data Mining in Determining Patterns of Interest of High School Graduates. *Journal of Physics: Conference Series*, *1339*(1). https://doi.org/10.1088/1742-6596/1339/1/012042

Kavoya, J. (2020). Machine learning for intelligence-driven Customs management. *African Tax and Customs Review*, *9535*, 50–58. https://www.researchgate.net/publication/341462392

Laporte, B. (2001). *Risk management systems : using data mining in developing countries ' customs administrations*. *5*(1), 17–28.

Regmi, R. H., & Timalsina, A. K. (2018). Risk Management in customs using Deep Neural Network. *Proceedings on 2018 IEEE 3rd International Conference on Computing, Communication and Security, ICCCS 2018*, *October 2018*, 133–137. https://doi.org/10.1109/CCCS.2018.8586834

Vanguard. (2021, December 20). *Nigeria Customs Service revenue hits N2.3trn – Official*. https://www.vanguardngr.com/2021/12/nigeria-customs-service-revenue-hits-n2-3trn-official/

Wuletaw, A. (2021), Developing automatic fraud detection model for customs transaction using deep learning technique. Dspace Repository. https://ir.bdu.edu.et/handle/123456789/13225

Yaqin, W., & Yuming, S. (2010). Classification model based on association rules in customs risk management application. *Proceedings - 2010 International Conference on Intelligent System Design and Engineering Application, ISDEA 2010*, *1*, 436–439.

https://doi.org/10.1109/ISDEA.2010.276

Zhang, L., Liu, K., Ilham, I., & Fan, J. (2022). Application of Data Mining Technology Based on Data Center. *Journal of Physics: Conference Series*, *2146*(1).https://doi.org/10.1088/1742-6596/2146/1/012017