The Revenues Authorities are the bank of confidential information from taxpayer and businesses in a digital format. They usually exchange that information with other governmental bodies. In East African Community (EAC), Revenues Authorities (RA) have put a lot of work into developing the updated systems with standards enabling the automatic exchange of information under multilateral or bilateral agreements. This paper show that Data Mining is helping the RA to Fight against Tax Fraud. It investigates and shows that RA are undertaking to monitor data of their information Technology systems and countering potential threats involving fraudulent communications aiming to steal identity and fraudulently claim tax refunds. It was found that 72% of revenues authorities already have the ability of data mining to identify tax fraud. An estimated 80% of RA will not attend the annual target if they don’t well manage data and follow their income. It was evident that data mining has impact on existing and potential data economy. Ultimately, this study finds KNN as the best model to be used in RA to detect and predict fraud.

**Keywords** Data mining, analytics, advanced analytics, predictive analytics, Revenues Authorities

**DATA GOVERNANCE IN EAST AFRICAN, COMMUNITY**

African countries have started with digital growth, at the continental and regional levels, there are identifiable institutions that create or make data governance related laws, regulations or policies in Africa. This permit the deployment of data-driven technologies to transform most aspects of the daily lives and work into quantifiable data that can be tracked, monitored, analysed and monetised has become such a phenomenon that the term “datafication” has been coined to describe it. The data governance ecosystem in Africa involves institutions that can ensure the availability, usability, integrity, security and quality of data shaped by functional regulations, contextual ethical principles and technical infrastructure.
DATA MINING IN REVENUES AUTHORITIES - USING ANALYTICS TO ENHANCE TAX COMPLIANCE IN EAST AFRICAN COMMUNITY

Nadine NIBIGIRA-University of Burundi, Pr Vincent HAVYARIMANA-Ecole Normale Superieure

DATA PREPROCESSING

The data used covers a period of one year. It was found that ARs in EAC have already started using data analysis techniques. Clustering is performed using k-clusters based on similarities between clusters and a simple K-mean algorithm with K=2 is used. Based on the values of the attributes grouped in the clusters, it was suggested to make two groups of which group constitutes taxpayers who do not engage in tax cheating (Fraud) and group 2 of taxpayers who engage in tax cheating (Zero Fraud).

The database comprises data of six RA of EAC from 2205 taxpayers, with 1.565. 089 lines (entries) and 12 columns (features). From the database, we train 1803 taxpayers and data are labelled as “Zero Fraud” or “Fraud”, so that and we use the supervised learning.

<table>
<thead>
<tr>
<th>Label</th>
<th>Number of Taxpayers</th>
<th>Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Fraud</td>
<td>1051</td>
<td>504.125</td>
</tr>
<tr>
<td>Fraud</td>
<td>752</td>
<td>180.383</td>
</tr>
</tbody>
</table>

Table1: Labelled dataset

In this data set fraudulent taxes are rare compared to normal taxes.

SUMMARY OF THE RESULT

Looking at the precision metric, the three algorithms used SVM, K-nearest neighbours (KNN) and Neural Network (NN ) work very similarly and the precision was at 95.32%. Therefore, the recall was at 89.22% for SVM, 72.18% for KNN and 82.02% for NN. And the accuracy was 98.03% for SVM, 99.22% for KNN and 99.11% for NN. We know accuracy can be misleading in the case of fraud detection. With highly imbalanced fraud data.

To decide which final model is best, we have considered how bad it is not to catch fraudsters, versus how many false positives the fraud analytics team can deal with. Ultimately, this study finds KNN as the best model to be used in RA to detect and predict fraud.

ROC Score:

0.9922598952689644670147

Classification Report:

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1051</td>
</tr>
<tr>
<td>1</td>
<td>0.97</td>
<td>0.88</td>
<td>0.88</td>
<td>752</td>
</tr>
</tbody>
</table>

| accuracy     | 0.99      | 0.72   | 0.94     | 1803    |
| macro avg    | 0.98      | 0.93   | 0.94     | 1083    |
| weighted avg | 0.99      | 0.93   | 0.99     | 1803    |

Confusion Matrix:

```
[[1800  19]
 [51  95]]
```

The model predicts 1803 cases of fraud, of which 650 are actual frauds and a very high accuracy score was found. Recall is therefore not as good as precision.
CONCLUSION

Tax compliance here refers to taxpayers fulfilling their registration, filing, reporting, and payment obligations correctly and on time. Intentionally failing to file personal income tax returns or filing an income tax return for natural or legal persons and deliberately understating the amount of income earned during the tax year comes up in cases of fraud. It should be noted however that KNN works very well in the when of the RA of the EAC region.

DISCUSSION

The study was conduct to data mining techniques to detect and predict tax fraud practice by taxpayers in different RA of EAC. As a result of fraud, tax revenues are compromises public investment. The detection of tax fraud has become one of the priorities in the EAC region. This is why the latter is one of the main priorities of the regional tax authorities who are required to develop profitable strategies to solve this problem.

Much of the recent work in tax evasion detection relies on supervised machine learning techniques that leverage labelled or audit-assisted data. Unfortunately, auditing tax returns is a slow and expensive process, so access to tagged historical information is extremely limited.

The fraud detection and prevention through machine learning is a collection of artificial intelligence (AI) algorithms trained with the historical data to suggest risk rules. It can then implement the rules to block or allow certain user actions, such as suspicious logins, identity theft, or fraudulent transactions. Detecting fraud in RA is essential and continuing.

The study suggests KNN algorithm which used to develop a model for predicting the annual revenue collection for RA and their performance has been compared for evaluation so as to get the best performer. According to the results there are high similarities between predicted actual data for both SVR, KNN and NN.
First of all, I would like to thank Almighty God for giving me strength, peace of mind and good health. This study is the result of the symbiotic framework where inspiration found in many directions. I would like to thank everyone from far or near who has contributed, directly or indirectly, to this work. I believe I could not have found better ingredients for it. I hope the results will serve our common goal of improving the fight against tax evasion in the EAC.

[1] STRENGTHENING DATA GOVERNANCE IN AFRICA, Project Inception Report, Centre for the Study of the economies of Africa. Sone Osakwe (Research Fellow at CSEA) and Adedeji Adeniran (Director of Research at CSEA)
[2] Data governance in EAST AFRICAN COMMUNITY
[4] Credit Card Fraud Detection - Machine Learning methods, Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla Faculty of Technical Sciences University of Novi Sad Novi Sad, Serbia