The International Journal of Engineering and Science (IJES)
|| Volume || 13 || Issue || 1 || Pages || PP 01-15 || 2024 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

# Adaptive Customization Detection Model based on Knowledge Discovery

Rehab Khaled Mohamed[a]* Ayman E. Khedr[b]* Mona Nasr[c]
[a] *Information Systems Department, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt, rehabkhaledM.D@gmail.com*
[b] *Information Systems Department, Faculty of Computers and Information Technology, Future University in Egypt (FUE), Cairo, Egypt, ayman.khedr@fue.edu.eg*
[c] *Information Systems Department, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt, drmona_nasr@yahoo.com*

**----------*ABSTRACT*----------**
*Due to the fast growth of internet technology, there is a lot of text data online that can be used to classify text. Taxes on international trade have traditionally brought in a lot of money for most countries' governments. Goods that crossed national borders were easy to track down, and goods were held until taxes and duties were paid. This made it harder to avoid paying taxes, and duty rates were often clear, so most problems with valuation were avoided. Customs administrations handle an enormous volume of trade. Among its responsibilities are risk management and the discovery of abnormalities and illegal consignments in import declarations. These activities are crucial since import tariffs make up a significant share of total tax collection.*
*Even though the customs system says it is anti-corruption and anti-fraud, as shown by the cases above, malfeasance continues in the revenue collection system. This research aims to find out how corruption and fraud affect the effectiveness of the customs duties revenue collection system by improving the ability to process text data with unbalanced distributions in customs fraud.*
*As a result, an effective model for the classification task of customs fraud may be constructed. A set of simulations were run to assess the performance of the suggested strategies. The efficacy and practicality of the suggested approaches are validated by simulation results when compared to TF-IDF using state-of-the-art classification algorithms that result in greater overall accuracy and processing time. In comparison to our approach, the maximum accuracy achieved with the random forest classifier was 98.97%, the accuracy achieved with MLP was 96.83%, the accuracy achieved with stochastic gradient descent was 95.45% , and the accuracy achieved with SVM was 98.78%.*
***Keywords:*** *Customs Fraud, Text mining, Classification*

---

Date of Submission: 23-12-2023 Date of acceptance: 03-01-2024

---

## I. Introduction

Big Data technologies, regardless of what we call them: algorithms, artificial intelligence (AI), or machine learning (ML), are more than simple "tools. "They are a common gesture that may be used to travel, make a choice, or trade, but they can also be used to police, control, or punish [1].
We no longer rely on machines to help us create things; rather, we rely on them to help us select, in the sense of determining, forecasting, or anticipating. As a result of the ease with which machines can make decisions with more rapidity and precision than humans can, there is a concern that we are losing control of something. This concern is a counterpoint to the ease with which machines can make choices. Technologies are dictating our behavior, judging our decisions, and directing us based on the results of calculations, with the results being imposed on us as mathematical evidence. As a result, proof is becoming increasingly remote and obscure to non-specialists, and as a result, it is becoming less and less disputable[2].
International trade taxes have long been an important way for most governments to make money. Goods that crossed national borders could be easily identified; goods were kept until taxes and duties were paid, which made tax fraud harder; duty rates were often clear, which got rid of the question of how much something was worth. So, from a tax administration point of view, customs taxes were easier to collect than other types of taxes. From a broader fiscal management point of view, their relative security and predictability was a plus. Tariffs on international trade were also supported by economic reasons[3].
Export taxes were put in place so that foreign buyers would pay them instead of domestic sellers. This meant that citizens didn't have to pay the tax. Import tariffs were seen as a tool for industrialization because they kept local manufacturers from having to compete with imports. This made both company owners and employees

more likely to support local manufacturers. In industrial countries, taxes on foreign trade now make up a small part of total income. This is due to a number of factors that have happened in recent years. But in developing countries, trade taxes still bring in a big but steadily shrinking share of total tax revenue [4].

Several things can explain this pattern. Tax administration has become more complex, and it is now easier to tax all kinds of economic activity thanks to organized firms, simple accounting systems, computerized record keeping, and more taxpayer compliance. Income taxes, general sales taxes, excise taxes, and property taxes can all be collected more efficiently than they were in the past. People have also said that trade tariffs hurt growth and job creation from an economic point of view. If a country didn't have a strong monopoly, export taxes tended to make it less competitive and bring in less money from exports. The fact that taxes on rubber, tin, coffee, and cocoa hurt the economy and are now mostly gone is proof of this [5].

Most people agree that trade liberalization is good, and the systematic removal of tariffs has been a major success of the General Agreement on Tariffs and Trade (GATT), the World Trade Organization (WTO), and a number of bilateral and regional trade agreements. Even with these changes, international trade tax administration is an important government job that needs to be supported and strengthened, especially in countries that depend a lot on trade taxes. As late as the beginning of the 2000s, 28 percent of all money made in Africa came from customs taxes. In the Middle East, this number was 22%, in East Asia and the Pacific it was 15%, and in the Western Hemisphere it was 13%. (De Wulf and Sokol 2005, 23). At the same time, customs administrations are in charge of collecting the value-added taxes that are paid on imports. This task requires some of the same steps that are used to charge customs fees. [6].

Text mining is when a computer finds new information that was not known before by automatically pulling information from different written sources. Text mining is the process of getting information from unstructured or semi-structured sets of text data. Data mining looks for patterns in structured databases or XML files [7].

This material may include emails, websites, medical abstracts, news stories, and business reports. With the vast amount of text data now accessible via the Internet, text mining has become increasingly valuable. It has been implemented in numerous fields, including business, healthcare, government, education, manufacturing, history, sociology, research, and criminal investigation. Text mining exists at the intersection of numerous disciplines, including data mining, knowledge discovery, information retrieval, machine learning, and natural language processing (NLP). Additionally, it borrows numerous algorithms and techniques from these fields. Information extraction (IE) is another field that significantly contributes to the discovery of information in text mining. IE involves the extraction of specific, structured information and predefined relationships, whereas text mining entails the discovery of general, unexpected information and new relationships. IE can be of great assistance during the knowledge extraction phase [8].

## II. Objectives

The research aims to accomplish and establish, among other things, the following: Identify the sources of fraud and corruption in the customs system.

managerial objectives:

- to analyze the risk of substantial revenue misstatements due to fraud.
- to identify anti-corruption tactics and evaluate their implementation.
- to evaluate the effectiveness of existing anti-corruption programs in reducing the risk of fraud and corruption.
- to establish a connection between fraud and corruption and revenue.

technical objectives:

- Provide a machine learning-based customs fraud detection architecture (TF-IDF).
- Make classification analysis using machine learning models.

## III. Background

3. Recognizing fraud Examines and investigates in detail the consequences of fraud and corruption on the effectiveness of customs duty revenue collection, as well as the causes of fraud and corruption. The paper covers fraud risks and how they affect revenue recognition. Examining the effectiveness of the connection between fraud, corruption, and income The student learned about anti-corruption measures via textbooks, periodicals, and the internet. in addition to other important sources.

The customs procedure employs a "human-in-the-loop" inspection paradigm, in which physical inspections are performed with the assistance of a fraud detection model, and customs agents determine whether the claimed item is fraudulent or not. If the inspection uncovers fraud, the officers may take on more tasks, and the findings will be utilized to enhance the model for detecting fraud. Most of the time, the algorithm identifies the most suspect products to be inspected, which might be problematic for nations with shifting trade trends.

Though the fraud detection model should continue to capture recognized scams and preserve income, it should

also learn about novel methods of theft. This is known as the "exploration-exploitation problem" [9], in which a balance must be struck between picking illegal objects to generate money immediately and discovering new things to maintain long-term performance.

There are different ways to deal with finding fraud. We talk about the use of neural networks, Bayesian networks , expert systems, rule-based systems , and finding statistical outliers [10, 11, 12]. These methods can be put into two groups: those that are supervised and those that are not. In the supervised approaches, there is a training set of operations that are either marked as fraud or as normal. Some systems, like neural network systems, need labelled inputs to build the model that will be used to spot fraud. These operations are used as inputs to those systems.

These techniques are classified into two types: supervised and unsupervised. There is a training set of operations that are designated as either fraudulent or normal in the supervised techniques. These actions are fed into some systems, such as neural network systems, which require labelled inputs to build the model that will be used to identify fraud [13].

## IV. Related Work

Algorithms for detecting customs fraud Earlier efforts to detect customs fraud relied on rule-based and random selection algorithms [17]. While some customs offices have used machine learning [18], many offices in underdeveloped nations continue to rely on rule systems and expert knowledge [19]. Recent research used off-the-shelf algorithms, such as the ensembled SVM, to detect customs fraud [20]. The Dual Attentive Tree-aware Embedding (DATE) model, for example, uses gradient boosting and attentions to build transaction-level embeddings and deliver interpretable judgments (Kim et al. 2020). Some of the more recent models use idea drift to better describe shifting trade patterns over time [21].

Techniques for Domain Adaptation Domain adaptation tries to learn universal representations that don't depend on the domain. One example is when the source and target domains have the same latent distribution[22]. To extract discriminative features between classes, contrastive learning is employed [23], and the memory module is used to supplement target characteristics with incremental information [24]. Negative transfer is a long-standing issue in domain adaptation. It refers to atypical conditions in which the source domain data causes poorer learning performance in the target domain due to a huge disparity in data distributions [25]. To address this issue, regularization and adaptive source selection approaches have been developed [26]. Most domain adaptation strategies require that the source and target data be accessible continuously, which may be impractical for multi-national customs administrations.

### 4.1 economic impact of Customs fraud

The danger of major income understatement due to fraud: Revenue in yearly financial statements may be misrepresented fraudulently by deceit and change of accounting records, false representation or deliberate omission of facts, and purposeful misapplication of accounting rules pertaining to categorization, disclosure, and presentation. [28] Organizations should examine the fraud risks associated with revenue misrepresentation in order to implement the suggested actions [30] outlines. This component of the literature review seeks to investigate or establish the fraud risks that exist inside organizations and lead to revenue overstatement.

Fraud risks leading to misstatement of revenue: According to [29], side agreements increase the risk of revenue misstatement. They arise when the terms and conditions of sale are changed or revised outside the appropriately recognized sales process or reporting channels, so impacting revenue recognition. According to [30], common modifications include return privileges, unlimited payment terms, refunds, and exchanges. To recognize revenue prior to the completion of the transaction, sellers may give these terms and conditions in side letters, emails, or verbal agreements.

### 4.2 Customs fraud detection

Previous research in the field of customs fraud detection suggests two main approaches for potential solutions. The first body of work includes ways for analyzing fresh samples, ranging from the simple but intuitive random inspection [31] to more complicated active learning solutions using uncertainty [32] and diversity. The second group of research focuses on using already obtained data. This includes using heuristic approaches and standard machine learning algorithms.

Machine learning is "focused on generating predictions based on previously observed trends and patterns from data analytics" [31]. From this description, we can figure out how to get to what governments would ideally want to be able to do with strategic trade: predict if a certain transaction fits a pattern that suggests it might be a regulated strategic good.

It is not a novel notion to use big data analysis or machine learning to analyze international commerce data. Utilizing mirror trade information is one approach for detecting misclassification and tax evasion. By comparing

an exporting state's statement with information from the importing state, authorities may be able to discern patterns of undervaluation or overvaluation, therefore identifying transactions or commodities in which companies routinely dodge taxes or levies [33]. In other cases, authorities may be able to identify trends in new transactions that indicate strategic products or tax/duty evasion simply by using past transaction data [34]. Recent research by [35] applies distinct techniques to decision tree algorithms for risk identification and profiling at customs. This paper proposes a strategy that builds on previous attempts but focuses on identifying strategic goods rather than the risk of duty evasion.

The efficiency of fraud detection is improved by inspecting random things, even if it means sacrificing some known frauds [36]. This research aims to determine the exploration ratio for the understudied human-in-the-loop fraud detection challenge.

## V. Methodology

In this section, we describe a method for the Customer Behavior Mining Framework based on data mining techniques in a customs fraud case. This framework considers the customers' behavior patterns and predicts the way they may act in the future.

The TF-IDF features are used first, then the implementation portfolio analysis is used to do the preprocessing analysis, and the k-means algorithm is used to divide past customers into groups based on their sociodemographic characteristics.

Finally, classification analysis using machine learning models was evaluated. MLP, Random Forest, and Stochastic Gradient Descent (SGD) Algorithms

### 5.1 Materials and Methods

Customs fraud is any effort to fraudulently decrease the customs duty (or tariff or tax) levied on goods, as indicated in Table 1. This dataset was created by professionals in customs fraud. The dataset is divided into eight columns [البند, الصنف, القيمة قبل التحسين بالدولار, القيمة بعد التحسين بالدولار, معدل التغير, القيمة المقبولة بالجنيه, الرسوم الجمركية المحصلة بالجنيه, بلد المنشأ] , each of which contains 11368 records.

*Table 1: dataset of customs fraud*

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | البند | الصنف | القيمة قبل التحسين بالد | القيمة بعد التحسين بال | معدل التغير | القيمة المقبولة بالجنيه | الرسوم الجمركية المحصلة بالج | بلد المنشأ |
| 2 | 8443320010 | طابعات كمبيوتر ملونة | 159,571 | 283,603 | 78% | 4,550,040 | 0 | مصر |
| 3 | 8487900090 | أجزاء آلات أو أجهزة أخ | 663,210 | 1,432,399 | 116% | 23,952,840 | 2,147,344 | المانيا الاتحادية |
| 4 | 8302410090 | تركيبات ولوازم وأصناف | 13,931 | 37,810 | 171% | 631,127 | 0 | البرتغال |
| 5 | 7013999000 | مصنوعات أخر من زجا | 28,402 | 42,968 | 51% | 723,550 | 431,148 | الصين الشعبية |
| 6 | 8512300010 | أجهزة كهربائية للإشارة | 218 | 546 | 150% | 9,068 | 0 | المانيا الاتحادية |
| 7 | 8716801000 | تروليات لزوم حضانات | 2,000 | 3,000 | 50% | 48,016 | 0 | مصر |
| 8 | 8414800010 | مضاغط تصرف حتى 5 | 533 | 1,333 | 150% | 22,413 | 1,121 | اليابان |
| 9 | 8515800090 | آلات وأجهزة لحام أخر | 9,518 | 16,232 | 71% | 272,162 | 13,608 | الصين الشعبية |
| 10 | 9606100000 | أزرار كباسة (حابكة بالآ | 135,926 | 205,020 | 51% | 3,303,841 | 165,192 | مصر |
| 11 | 8412310000 | محركات وآلات محركة | 1,562 | 15,558 | 896% | 258,807 | 0 | المانيا الاتحادية |
| 12 | 8452290000 | آلات خياطة غير منزلية | 1,677,493 | 2,755,744 | 64% | 44,619,875 | 0 | مصر |
| 13 | 8544209010 | كابلات أخر متحدة المد | 456,648 | 734,097 | 61% | 12,076,804 | 1,033,370 | مصر |
| 14 | 8708999090 | أجزاء وقطع منفصلة أخ | 32,028 | 398,246 | 1143% | 6,616,086 | 660,414 | السوق الاوربية المشتركة |
| 15 | 8507200090 | مدخرات [جماعات ] ك | 42,986 | 68,764 | 60% | 1,414,187 | 192,579 | المانيا الاتحادية |
| 16 | 1211900050 | جذور سوس [ عرق سوس | 202,295 | 232,640 | 15% | 3,886,200 | 0 | سوريا |
| 17 | 8701100090 | جرارات صغيرة [موتوكلا | 2,805 | 11,791 | 320% | 197,632 | 9,882 | اليابان |
| 18 | 8419810000 | آلات وأجهزة ومعدات أ | 289 | 5,009 | 1633% | 83,682 | 0 | مملكةمتحدةبريطانياعظمى وش.ايرلندواسكا |
| 19 | 9108900000 | عدد حركة ساعات ،عدا | 2,520 | 3,150 | 25% | 50,877 | 5,088 | مصر |
| 20 | 9026100000 | أجهزة وأدوات لقياس أو | 2,483 | 3,792 | 53% | 63,046 | 0 | المانيا الاتحادية |
| 21 | 7228100000 | قضبان من صلب متحم | 40,153 | 81,633 | 103% | 1,311,888 | 65,594 | مصر |
| 22 | 8501403090 | محركات كهربائية لآلات | 38,550 | 64,550 | 67% | 1,037,593 | 20,752 | مصر |
| 23 | 8544491000 | موصلات كهربائية أخر، | 6,070 | 10,420 | 72% | 173,682 | 0 | دولة الامارات العربية |
| 24 | 9207100000 | أدوات موسيقية ذات مة | 24 | 277 | 1050% | 4,692 | 1,408 | اسبانيا |
| 25 | 8464200000 | آلات شحذ أو صقل الحـ | 237,080 | 522,380 | 120% | 8,393,345 | 0 | مصر |

We checked for customs fraud in the data using the TF-IDF features, then made a classification algorithm.

## 5.2 proposed approach

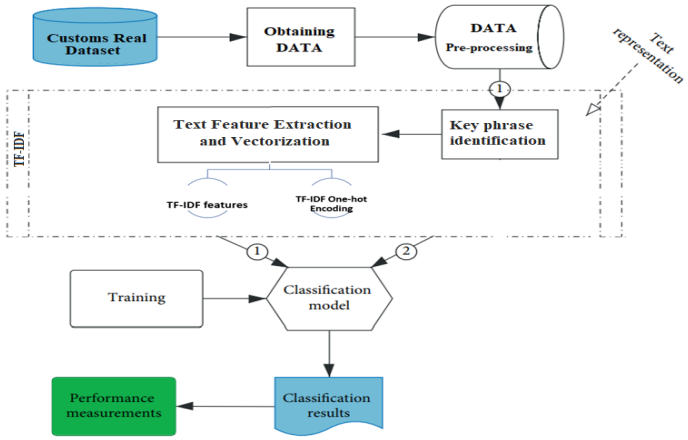


*Figure 1:Methodology for Custom fraud*

We go over our Proposed Approach procedure, which is seen in Fig. 1. First, we downloaded the customs fraud dataset. After that, the raw data was processed to eliminate the abnormalities. Second, several feature selection and extraction approaches were used on the processed data, and last, various machine learning classifiers were utilized to develop a detection model. Each step is thoroughly discussed.

1.Collect datasets or use benchmark datasets that contain data on customs fraud.
2.The dataset is pre-processed and cleared.
3.Make Key phrase identification with TF-IDF.
4.Make Text Feature Extraction and Vectorization
1.TF-IDF features
2.TF-IDF One-hot Encoding
5.Make a classification using the data features you chose.

## 5.3 Obtaining The Data

Egypt's customs fraud yielded the dataset containing real and fake customs fraud. The shape of the dataset was 11368 rows and 8 columns.

*Table 2: sample of dataset*

| | البند | الصنف | القيمة قبل التحسين بالدولار | القيمة بعد التحسين بالدولار | معدل التغير | القيمة المقبولة بالجنيه | الرسوم الجمركية المحصلة بالجنيه | بلد المنشأ |
|---|---|---|---|---|---|---|---|---|
| 0 | 8443320010 | طابعات كمبيوتر ملونة | 159571 | 283603 | 0.78 | 4550040 | 0 | مصر |
| 1 | 8487900090 | أجزاء آلات أو أجهزة أخر، غير محتوية علي موصلات... | 663210 | 1432399 | 1.16 | 23952840 | 2147344 | المانيا الاتحادية |
| 2 | 8302410090 | تركيبات ولوازم وأصناف مماثلة أخر ممايستعمل للم... | 13931 | 37810 | 1.71 | 631127 | 0 | البرتغال |
| 3 | 7013999000 | مصنوعات أخر من زجاج أخر من الأنواع المستعملة ل... | 28402 | 42968 | 0.51 | 723550 | 431148 | الصين الشعبية |
| 4 | 8512300010 | أجهزة كهربائية للإشارة الصوتية من الأنواع المس... | 218 | 546 | 1.50 | 9068 | 0 | المانيا الاتحادية |

## 5.4 Pre-processing

Data cleaning is the process of fixing or deleting inaccurate, corrupted, improperly formatted, duplicate, or incomplete data from a dataset. When combining data from many sources, there are numerous opportunities for data duplication and mislabeling. Even though the conclusions and algorithms seem accurate, you cannot rely on them if the data are incorrect. Since the methods vary from dataset to dataset, it is impossible to definitively state the correct steps for cleaning data. To create a template for your data cleaning process, do the following:

1-remove "null" from the data.
2-create labels in the data.
3-Determine the number of labels in the data.

## 5.5 Key phrase identification with TF-IDF

We will compare several scikit-learn detection models. To extract features, bag-of-words from the dataset and Term Frequency—Inverse Document Frequency (TF—IDF) are used. The lengthy article is easily divided into words by counting frequency [37].

TF-IDF scores are used to analyze key phrases in customs fraud. The words with the highest scores provide relevant information about customs fraud. However, many terms in the fraud domain may refer to the same concept. To identify that topic, TF-IDF modelling is used, and various topics are identified. This displays the top words in a specific topic, as well as the likelihood that they belong to that topic. Finally, the output will be used to highlight potentially fraudulent activity or transactions.
TF-IDF is a mix of term frequency (TF) and inverse document frequency (IDF) (IDF). The TF representation is one of the most straightforward TWSs since it utilises the document's original word frequency value. The term "frequency value" (TF) presupposes that a phrase with a higher frequency value is considered more significant than one with a lower frequency value. Only the frequency with which a common phrase appears in a customs scam is relevant. Due to a lack of collection frequency information, the TF's capacity to discern pertinent customs fraud papers from other irrelevant records was previously quite limited [38].
In order to address this issue, the inverse document frequency (IDF) was developed with collecting frequency in mind. This enhanced a term's capacity to differentiate across texts.
IDF stands for "document frequency," which refers to the number of papers in which a certain phrase occurs. It is recommended on the basis that a phrase that appears in fewer papers is more significant than one that appears in more instances of customs fraud[39].

### 5.6 Text Feature Extraction and Vectorization

**For predictive modelling in fake news analysis, we use specialised ways to prepare text data. To get text data ready for predictive modelling, sample text is tokenized by breaking it up into words. These words are then encoded as integers or floating-point values with vectorization so that features can be extracted. The Python scikit-learn libraries are used in the work at hand. CountVectorizer and TfidfVectorizer can turn text into word count vectors and word frequency vectors, respectively. [40].**

#### 5.6.1 TF-IDF features

Term Frequency-Inverse Document Frequency (TF-IDF) is an information retrieval and text mining weighting system. The TF-IDF value reflects the significance of a phrase inside a corpus document. Note that "document" refers to a pathology report, "corpus" to a group of reports, and "term" to a single word inside a report. The TF-IDF weight of a phrase t in document d is given by [41]:

$$\begin{aligned} TF(t,d) &= \frac{\text{Number of times } t \text{ appears in } d}{\text{Total number terms in } d}, \\ IDF(t) &= log\Big(\frac{\text{Total number of documents}}{\text{Number of documents with } t}\Big), \\ TF\text{-}IDF(t,d) &= TF(t,d) * IDF(t) \end{aligned} \quad (1)$$

To convert a pathology report into a feature vector, we went through the following steps:

- Create a vocabulary set that includes all the unique words from all the pre-processed training reports.
- Create a zero-length vector fd of equal length as the vocabulary.
- Set TF-IDF as the index for each word. (t, d).
- The final output fd is a very sparse feature vector.

#### 5.6.2 TF-IDF One-hot Encoding

This algorithm generates a vector with a length equal to the number of categories in your dataset, where a category is a single distinct word. Let's say we want a single hot encoding description for each of the three documents below, which are customs fraud reviews.
One popular encoding is a vector portrayal of words in "jargon.". Each word in the jargon is addressed by a vector of size "n," where "n" is the total number of words in the jargon. For example, if a jargon contains 25 words, the vector comparing each word will have a size of 25, and that too with twofold qualities like "0" and "1," and we use it in customs fraud [39].

## VI. Results

Following the collection and preparation of hundreds of fraud instances from different categories of customs fraud (customs fraud is any fraudulent effort to lower the customs duty (or tariff or tax) imposed on goods), table-3 displays the key to customs fraud from these dataset samples. These top customs fraud keys from the dataset are then utilized to identify items. Multiple goods may be chosen based on the specifications.

### 6.1 Pre-processing

```
Number of All rows: 11367
```

We discovered that there were 11367 rows in the data after counting them.

```
Number of null rows: 158
```

We found 158 rows of data to remove after counting the number of nulls in the data.

*Table3:sample of nulls data*

| | البند | الصنف | القيمة قبل التحسين بالدولار | القيمة بعد التحسين بالدولار | معدل التغير | القيمة المقبولة بالجنيه | الرسوم الجمركية المحصلة بالجنيه | بلد المنشأ |
|---|---|---|---|---|---|---|---|---|
| **123** | 8507300000 | مدخرات (جماعات) كهربائية من النيكل - كادميوم ... | 133067 | 160080 | 0.20 | 2571606 | 128580 | NaN |
| **183** | 8481901000 | أجزاء للاصناف الواردة فى البندين 10 10 81... 84 ، | 4705 | 5772 | 0.23 | 92889 | 37156 | NaN |
| **465** | 8515190000 | آلات وأجهزة أخر للحام بمعادن مالئة ( عيدان لحا... | 224 | 336 | 0.50 | 5435 | 272 | NaN |
| **541** | 8407349000 | محركات الاحتراق الداخلي ذات مكابس متناوبة للمر... | 318 | 2975 | 8.36 | 47703 | 4770 | NaN |
| **598** | 8541100000 | صمامات ثنائية ، عدا الصمامات الثنائية الحساسة ... | 420 | 1050 | 1.50 | 16871 | 0 | NaN |

In table-2, null data will be removed and made clear.

```
Number of Non Null rows: 11209
```

To create labels in the data, we select the column name [الصنف].

```
0                                    طابعات كمبيوتر ملونة
1     أجزاء آلات أو أجهزة أخر، غير محتوية على موصلات...
2     تركيبات ولوازم وأصناف مماثلة أخر ممايستعمل للم...
3     مصنوعات أخر من زجاج آخر من الأنواع المستعملة ل...
4     أجهزة كهربائية للإشارة الصوتية من الأنواع المس...
5                          تروليات لزوم حضانات التفريخ.
6     مضاغط تصرف حتى 15م مكعب فى الدقيقة وضغط حتى 10...
7     ألات وأجهزة لحام أخر بخلاف ماذكر أعلاه ألات وأ...
8                  أزرار كباسة (حابكة بالكبس ) ، وأجزاؤها .
9     محركات وآلات محركة ، تعمل بالهواء المضغوط [بني...
10                 آلات خياطة غير منزلية ، غير ذاتية الحركة .
```

To extract and count the number of labels in the data, we use a library called [nltk] to remove stop words like [.,] punctuation in Arabic, then we count the number of labels. A tokenizer divides a string by matching either the tokens or the separators between tokens using a regular expression. Then, we calculate the number of words in the data as well as the number of times each word is repeated, and we add spaces between each word.

أجزاء آلات أو أجهزة أخر غير محتوية علي موصلات أوعوازل أو وشائع كهربائية أو أدوات تماس أو غيرها من تركيبات كهربائية عدا رفاصات مراوح دافعة للسفن أوالقوارب وريشها غيرمذكورة إ
لخ
تركيبات ولوازم وأصناف مماثلة أخر مماياستعمل للمباني من معادن عادية فيماعدا الأبواب
مصنوعات أخر من زجاج آخر من الأنواع المستعملة للمائدة المطبخ التواليت للمكتب للتزيين الداخلي أو لاستعمالات مماثلة عدا الأصناف الداخلة في البند 10 70 أو البند الخ
39 85 أجهزة كهربائية للإشارة الصوتية من الأنواع المستعملة للسيارات عدا الأصناف الداخلة في البند
ترولیات لزوم حضانات التفريخ
مضاغط تصرف حتى 15م مكعب في الدقيقة وضغط حتى 10 جوى ومركبة أو متصلة بمحركات تعمل بالضغط حتى 92 كيلو وات
آلات وأجهزة لحام أخر بخلاف ماذكر أعلاه آلات وأجهزه كهربائية للرش الساخن للمعادن والخلائط المعدنية الخزفية سيرميت المصهورة
أزرار كباسة حابكة بالكبس وأجزاؤها
محركات وآلات محركة تعمل بالهواء المضغوط بنيوماتيك بحركة خطية مستقيمة اسطوانات
آلات خياطة غير منزلية غير ذاتية الحركة

*Figure 2:sample of labels*

We use our dictionary to remove prepositions, verbs, and some words that have no meaning and may not be used as indicative words.

*Table 4 :our dictionary*

| | الكلمات المستبعدة | حروف الجر | الالوان | اسماء الاشارة |
|---|---|---|---|---|
| 0 | ي | من | احمر | هذا |
| 1 | في | الى | برتقالى | هذه |
| 2 | كل | عن | اصفر | الذى |
| 3 | لم | على | اخضر | التى |
| 4 | لن | فى | ازرق | ذات |

The data after applying and removing looks like this:

أجزاء آلات أو أجهزة أخر غير محتوية علي موصلات أوعوازل أو وشائع كهربائية أو أدوات تماس أو غيرها من تركيبات كهربائية عدا رفاصات مراوح دافعة للسفن أوالقوارب وريشها غيرمذكورة' إلخ'

However, there are some words that must be removed again; we make them in our code, such as {ف , او , و } . Then, word by word, we create a dictionary of labels by dividing each sentence by total_word_length to extract words and measure word strength.

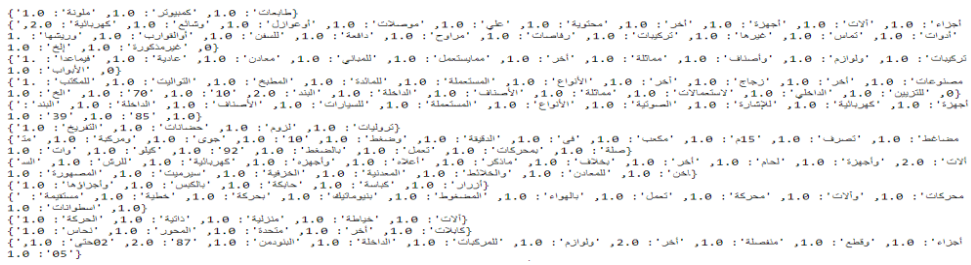

Figure 3:dictionary of labels

1.0: عوامل
1.0: عضوية
1.0: نشاط
1.0: سطحي
1.0: الصابون
1.0: أنيونية
1.0: سالبة
1.0: الشحنة
1.0: غيرمهيأة
1.0: للبيع
1.0: بالتجزئة

*Figure 4:sample of dictionary of labels arranged.*

After collecting word labels from sentences in the dictionary, we extract strong words from sentences and then save them in a column in the dataset.

*Table 3:sample of [ الكلمات الدلالية ]*

```
0: طابعات, كمبيوتر, ملونة
1: كهربائية, أجزاء, آلات, أجهزة, محتوية, موصلات, عوازل, شائع, أدوات, تماس
2: تركيبات, لوازم, أصناف, مماثلة, مماسستعمل, للمباني, معادن, عادية, الأبواب
3: البند, مصنوعات, زجاج, الأنواع, المستعملة, للمائدة, المطبخ, التواليت, للمكتب, للتزيين
4: أجهزة, كهربائية, للإشارة, الصوتية, الأنواع, المستعملة, للسيارات, الأصناف, الداخلة, البند
5: تروليات, لزوم, حضانات, التفريخ
6: مضاغط, تصرف, 15م, مكعب, الدقيقة, ضغط, 10, جوى, مركبة, متصلة
7: آلات, أجهزة, لحام, بخلاف, ماذكر, أعلاه, أجهزه, كهربائية, للرش, الساخن
8: أزرار, كباسة, حابكة, بالكبس, أجزاؤها
9: محركات, آلات, محركة, بالهواء, المضغوط, بنيوماتيك, بحركة, خطية, مستقيمة, اسطوانات
```

Save [الكلمات الدلالية ] in the dataset's table and use it as a label for each row of data.

*Table 4: dataset with [ الكلمات الدلالية ]*

| | البند | الصنف | القيمة قبل التحسين بالدولار | القيمة بعد التحسين بالدولار | معدل التغير | القيمة المقبولة بالجنيه | الرسوم الجمركية المحصلة بالجنيه | بلد المنشأ | الكلمات الدلالية |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 8443320010 | طابعات كمبيوتر ملونة | 159571 | 283603 | 0.78 | 4550040 | 0 | مصر | طابعات, كمبيوتر, ملونة |
| 1 | 8487900090 | أجزاء آلات أو أجهزة أخر، غير محتوية على موصلات... | 663210 | 1432399 | 1.16 | 23952840 | 2147344 | الدانيا الاتحادية | كهربائية, أجزاء, آلات, أجهزة, محتوية, ... ,موصلات |
| 2 | 8302410090 | تركيبات ولوازم وأصناف مماثلة أخر مماسستعمل للم... | 13931 | 37810 | 1.71 | 631127 | 0 | البرتغال | تركيبات, لوازم, أصناف, مماثلة, مماسستعمل, ...للمب |
| 3 | 7013999000 | مصنوعات أخر من زجاج أخر من الأنواع المستعملة ل... | 28402 | 42968 | 0.51 | 723550 | 431148 | الصين الشعبية | البند, مصنوعات, زجاج, الأنواع, المستعملة, ...للما |
| 4 | 8512300010 | أجهزه كهربائيه للإشاره الصوتيه من الأنواع المس... | 218 | 546 | 1.50 | 9068 | 0 | الدانيا الاتحادية | أجهزة, كهربائية, للإشارة, الصوتية, الأنواع, ...ال |
| 5 | 8716801000 | .تروليات لزوم حضانات التفريخ | 2000 | 3000 | 0.50 | 48016 | 0 | مصر | تروليات, لزوم, حضانات, التفريخ |
| 6 | 8414800010 | مضاغط تصرف حتى 15م مكعب فى الدقيقة وضغط حتى 10... | 533 | 1333 | 1.50 | 22413 | 1121 | اليابان | مضاغط, تصرف, 15م, مكعب, الدقيقة, ...,ضغط, 10, جوى |
| 7 | 8515800090 | آلات وأجهزه لحام أخر بخلاف ماذكر أعلاه آلات و... | 9518 | 16232 | 0.71 | 272162 | 13608 | الصين الشعبية | آلات, أجهزة, لحام, بخلاف, ماذكر, أعلاه, ...,أجهزه |
| 8 | 9606100000 | . أزرار كباسة (حابكة بالكبس ) ، وأجزاؤها | 135926 | 205020 | 0.51 | 3303841 | 165192 | مصر | أزرار, كباسة, حابكة, بالكبس, أجزاؤها |
| 9 | 8412310000 | محركات وآلات محركة ، تعمل بالهواء المضغوط إنبي... | 1562 | 15558 | 8.96 | 258807 | 0 | الدانيا الاتحادية | محركات, آلات, محركة, بالهواء, المضغوط, ...بنيومات |

## 6.2 Extract TFIDF Features

Customs data will be weighted using the TF-IDF approach, which can be seen in Table 7, based on the results of the preprocessing. One Hot Encoding TF-IDF word frequency or weighting
Using the CountVectorizer function, we turn a supplied text into a vector based on the frequency (count) of each word across the full text.
Then use the Tfidf Vectorizer function, which needs to perform both term frequency and inverse record frequency for you, whereas Tfidf Transformer requires you to use the Scikit-CountVectorizer class from Scikit-Learn to do term frequency.

Then compute Term Frequency and enter the results in column [val].

*Table 8: Term frequency of TF-IDF*

| | القيمة قبل التحسين بالدولار | القيمة بعد التحسين بالدولار | معدل التغير | القيمة المقبولة بالجنيه | الرسوم الجمركية المحصلة بالجنيه | 0 | 1 | 2 | 3 | 4 | ... | 5066 | 5067 | 5068 | 5069 | 5070 | 5071 | 5072 | 5073 | 5074 | val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 159571 | 283603 | 0.78 | 4550040 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.777284 |
| 1 | 663210 | 1432399 | 1.16 | 23952840 | 2147344 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.159797 |
| 2 | 13931 | 37810 | 1.71 | 631127 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.714091 |
| 3 | 28402 | 42968 | 0.51 | 723550 | 431148 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.512851 |
| 4 | 218 | 546 | 1.50 | 9068 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.504587 |
| 5 | 2000 | 3000 | 0.50 | 48016 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.500000 |
| 6 | 533 | 1333 | 1.50 | 22413 | 1121 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.500938 |
| 7 | 9518 | 16232 | 0.71 | 272162 | 13608 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.705400 |
| 8 | 135926 | 205020 | 0.51 | 3303841 | 165192 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.508321 |
| 9 | 1562 | 15558 | 8.96 | 258807 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 9.960307 |
| 10 | 1677493 | 2755744 | 0.64 | 44619875 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.642775 |
| 11 | 456648 | 734097 | 0.61 | 12076804 | 1033370 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.607577 |
| 12 | 32028 | 398246 | 11.43 | 6616086 | 660414 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 12.434307 |
| 13 | 42986 | 68764 | 0.60 | 1414187 | 192579 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.599684 |
| 14 | 202295 | 232640 | 0.15 | 3886200 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.150004 |

Following the completion of table edits

| | القيمة قبل التحسين بالدولار | القيمة بعد التحسين بالدولار | معدل التغير | القيمة المقبولة بالجنيه | الرسوم الجمركية المحصلة بالجنيه | 0 | 1 | 2 | 3 | 4 | ... | 6934 | 6935 | 6936 | 6937 | 6938 | 6939 | val | Mean | Sum | Count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 159571 | 283603 | 0.78 | 4550040 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.777284 | 1.0 | 3 | 3 |
| 1 | 663210 | 1432399 | 1.16 | 23952840 | 2147344 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.159797 | 1.1 | 11 | 10 |
| 2 | 13931 | 37810 | 1.71 | 631127 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.714091 | 1.0 | 9 | 9 |
| 3 | 28402 | 42968 | 0.51 | 723550 | 431148 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.512851 | 1.1 | 11 | 10 |
| 4 | 218 | 546 | 1.50 | 9068 | 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.504587 | 1.0 | 10 | 10 |

## 6.3 Classification

### 6.3.1 Random Forest Classification

*Table 9:Random Forest performance matrix*

| | |
|---|---|
| Accuracy | 98.97% |
| Sensitivity | 99.21% |
| Specificity | 98.77% |

Finally, Random forest Classification: 98.97%

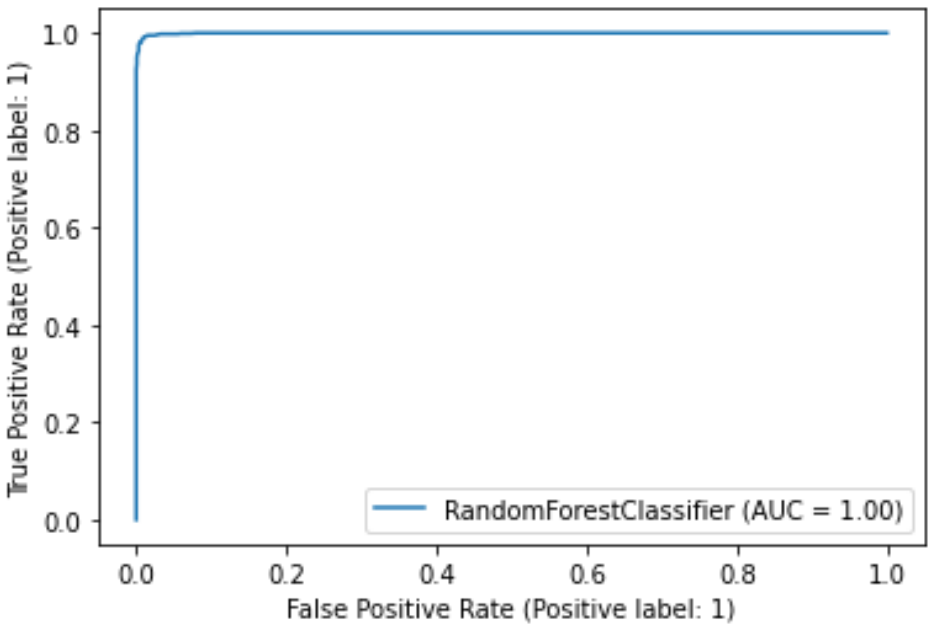


*Figure 5:Roc curve of Random Forest*

The true positive rate and false positive rate are shown in the confusion matrix in Figure 11, and the AUC is set to 1.00.

6.3.2 Multi-layer perceptron (MLP) Classification

*Table 10:MLP performance matrix*

| | |
|---|---|
| Accuracy | 96.83% |
| Sensitivity | 99.71% |
| Specificity | 94.44% |

Finally, multi-layer perceptron (MLP) Classification: 96.83%

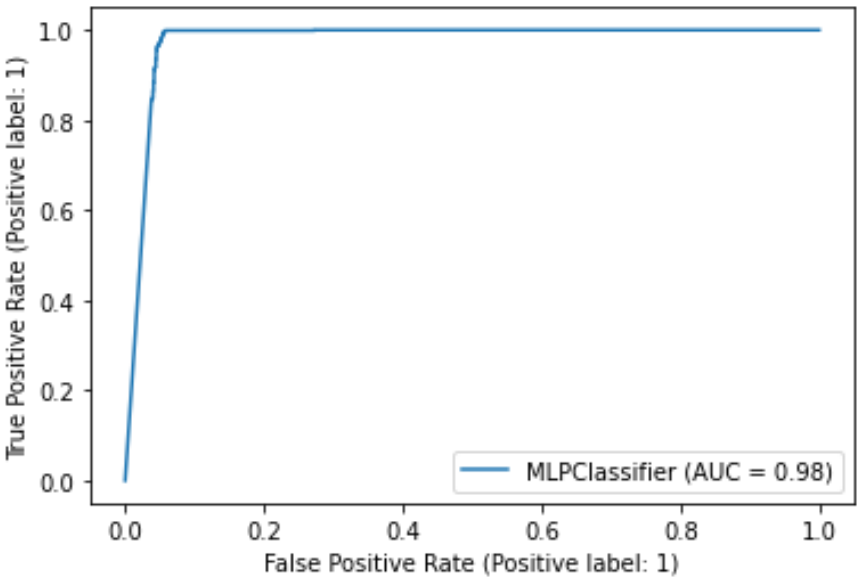


*Figure 6:Roc curve of MLP*

The true positive rate and false positive rate are shown in the confusion matrix in Figure 12, and the AUC is 0.98.

6.3.3 Stochastic gradient descent (SGD) Classification

*Table 11:Stochastic gradient descent performance matrix*

| | |
|---|---|
| Accuracy | 95.45% |
| Sensitivity | 94.6% |
| Specificity | 96.16% |

Finally, Stochastic gradient descent (SGD) Classification: 95.45%

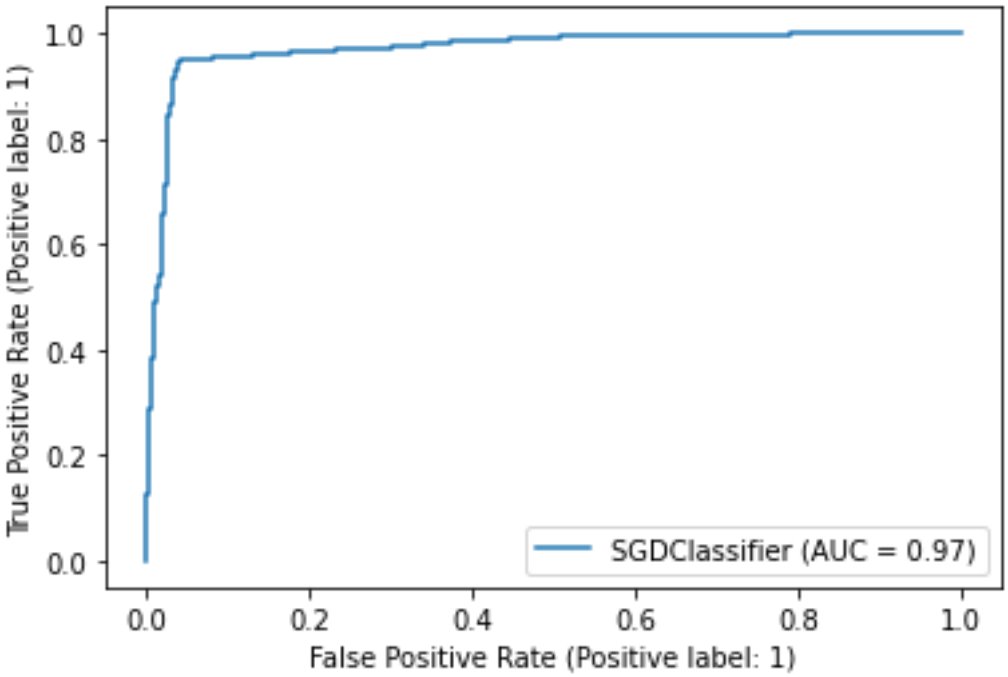


*Figure 7:Roc curve of Stochastic gradient descent*

The true positive rate and false positive rate are shown in the confusion matrix in Figure 13, and the AUC is 0.97.

6.3.4 Support vector machine (SVM) Classification

*Table 12 :SVM performance matrix*

| | |
|---|---|
| Accuracy | 98.78% |
| Sensitivity | 94.6% |
| Specificity | 75.7% |

Finally, SVM Classification: 98.78%%

6.4 **Discussions**

The algorithm would be trained on the data and assessed against a known-classification test set. Parameters or features might be adjusted based on the test. This model fits a strategic need. After testing, the technique may be iterated to construct models for a portfolio of strategic products based on a state-level risk or priority evaluation. Once built, these models may be used with fresh data. Pre-shipment data or shippers' export declarations are used to assess whether a transaction contains a strategic product. Multiple models take into consideration commodity features and anticipate which strategic goods may be involved. A broad methodology to determine whether a transaction involves a strategic good would blend commodities with different risks and priorities.

Our technique, which included Random Forest, stochastic gradient descent, and multi-layer perceptron as well as a large data framework, resulted in greater overall accuracy and processing time. In comparison to our approach, where the maximum accuracy achieved with the random forest classifier was 98.97%, the accuracy achieved with MLP was 96.83%, and the accuracy achieved with stochastic gradient descent was 95.45%, our study has certain limitations. To begin, we just utilized two machine learning models. Second, there are only two classifications in the dataset.

.

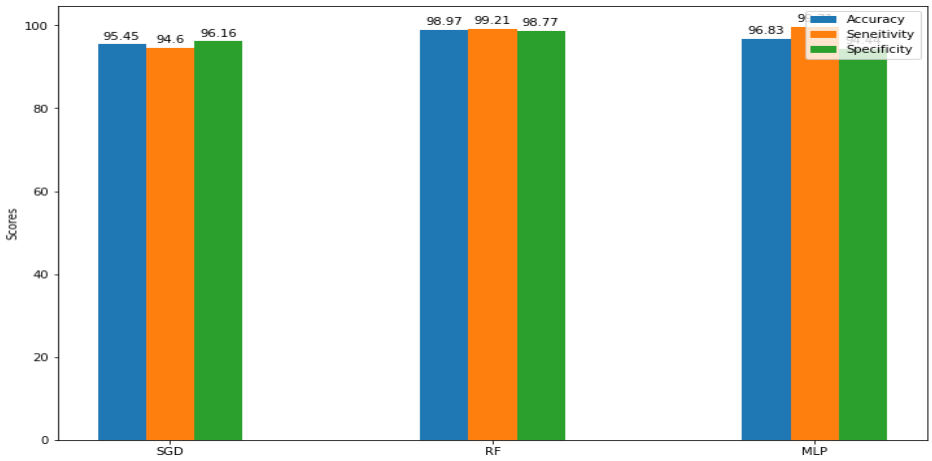


*Figure 8:comparison between our algorithms*

Classifying strategic goods transactions has several uses for nations. This technique would provide improved profiling of strategic goods transactions using real-world data. Training a random forest model on strategic good transactions and others may help identify transaction trends. Once trained and adjusted, these models might be applied to incoming transactions, improving risk profiling, documentation, or end-use checks. Modeling based on high-priority strategic items might improve resource allocation and justify inspections. This technique would also help nations assess strategic trade movements and pinpoint transshipment locations. The random forest may consider transaction origin and destination. A critical node in the model may help with enforcement targeting or outreach.

This strategy might also be used to increase customs efficiency via outreach. First, using current data to train algorithms, unlicensed strategic goods transactions might be identified. These organizations might get export control training and end-use verification. Because this technique is used to indicate how entities transport strategic goods, it could be used to improve how governments and entities categorize which products should be used in strategic product commerce. This might give a baseline for how state organizations operate and how customs officials can engage with them and international partners to better utilize the customs system identified. These organizations might get export control training and end-use verification. Because this technique is used to indicate how entities transport strategic goods, it could be used to improve how governments and entities categorize which products should be used in strategic product commerce. This might give a baseline for how state organizations operate and how customs officials can engage with them and international partners to better utilize the customs system. If this technique works, customs data will get completer and more reliable over time, boosting modelling. The exchange of transaction models for strategic goods with other trading partners may improve import detection and the sharing of best practices.

## VII. Conclusion

The use of machine learning in customs control has shown that it has a lot of promise to improve risk ratings, regulation, and marketing in foreign business. As more and more transaction data is collected, the models used to classify key goods can be improved and changed. Using data about the state as a whole, these models can find useful things based on things like location, trade partners, and industrial powers.

Machine learning techniques in customs control have had to deal with two main problems over the years: the need to move away from rule-based systems and the difficulty of balancing multiple goals at the same time. But a lot of work has been made in dealing with these problems, which has led to good applications.

The technique is a mixed solution that can be used with different types of artificial intelligence. Still, speed can be improved by looking into more creative preparation methods and making use of document-level traits and context. Also, RF, MLP, and SGD were chosen based on how well they fit the problem at hand, taking into account things like accuracy, scaling, and how easy they are to understand.

Overall, machine learning has a lot of promise to change how customs control works, and more study and new ideas will pave the way for even better and more effective ways to do things in the future.

**Declaration of interests**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


[1]. W. L. Collaborative, Customs fraud, 2020. URL: https://www.whistleblowerllc.com/ what-we-do/financial-fraud/customs-fraud/.

[2]. Y. Sahin, E. Duman, Detecting credit card fraud by decision trees and support vector machines, IMECS 2011 - International MultiConference of Engineers and Computer Scientists 2011 1 (2011) 442–447.

[3]. S. Y. Huang, Fraud detection model by using support vector machine techniques, 2013.

[4]. E. Kirkos, C. Spathis, Y. Manolopoulos, Data mining techniques for the detection of fraudulent financial statements, Expert systems with applications 32 (2007) 995–1003.

[5]. V. Čekanavičius, G. Murauskas, Statistika ir jos taikymai, Vilnius: teV 1 (2000).

[6]. J. Cohen, Statistical power analysis for the behavioral sciences, 2nd edn. á/l, 1988.

[7]. H.-Y. Kim, Statistical notes for clinical researchers: Chi-squared test and fisher's exact test, Restorative dentistry & endodontics 42 (2017) 152–155.

[8]. N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent Data Analysis (2002) 429–449.

[9]. G. E. A. P. A. Batista, R. Prati, M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explorations 6 (2004) 20–29.

[10]. ] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, J. Artif. Int. Res. 16 (2002) 321–357.

[11]. P. Gajjewar, Understanding fuzzy neural network using code and animation, 2018. URL: https://medium.com/@apbetahouse45/ understanding-fuzzy-neural-network-with-code-and-graphs-263d1091d773.

[12]. P. K. Simpson, Fuzzy min—max neural networks—part 1: Classification, IEEE Trans. On Neural Networks 3 (1992) 776–786.

[13]. A. Tharwat, Classification assessment methods, Applied Computing and Informatics (2018). URL: http://www.sciencedirect.com/science/article/pii/S2210832718301546. doi:https://doi.org/10.1016/j.aci.2018.08.003

[14]. J. J. Filho, "Artificial intelligence in the customs selection system through machine learning (SISAM)," Receita Federal do Brasil, 2015.

[15]. K. Mikuriya and T. Cantens, "If algorithms dream of customs, do customs officials dream of algorithms? a manifesto for data mobilization in customs," World Customs Journal, vol. 14, no. 2, 2021.

[16]. S. Kim, T.-D. Mai, T. N. D. Khanh, S. Han, S. Park, K. Singh, and M. Cha, "Take a chance: Managing the exploitation-exploration dilemma in customs fraud detection via online active learning," arXiv preprint arXiv:2010.14282, 2020.

[17]. J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 12, pp. 2346–2363, 2020.

[18]. T. Bayoumi, "Changing patterns of global trade," International Monetary Fund, 2011.

[19]. J.-Y. Audibert, R. Munos, and C. Szepesvari, "Exploration–exploitation ´ tradeoff using variance estimates in multi-armed bandits," Theoretical Computer Science, vol. 410, no. 19, pp. 1876–1902, 2009.

[20]. C. Han and R. Ireland, "Performance measurement of the KCS customs selectivity system," Risk Management, vol. 16, no. 8, pp. 25–43, 2014.

[21]. N. Houlsby, F. Huszar, Z. Ghahramani, and M. Lengyel, "Bayesian ´ active learning for classification and preference learning," 2011.

[22]. O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in ICLR, 2018.

[23]. J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," in ICLR, 2020.

[24]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in KDD, 2016, pp. 785–794.

[25]. J. Vanhoeyveld, D. Martens, and B. Peeters, "Customs fraud detection: Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue," Pattern Analysis and Applications, vol. 23, 2020.

[26]. S. Kim, Y.-C. Tsai, K. Sigh, Y. Choi, E. Ibok, C.-T. Li, and M. Cha, "DATE: Dual attentive tree-aware embedding for customs fraud detection," in KDD, 2020, pp. 2880–2890.

[27]. J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in Advances in Artificial Intelligence – SBIA 2004, A. L. C. Bazzan and S. Labidi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 286–295.

[28]. N. Lu, G. Zhang, and J. Lu, "Concept drift detection via competence models," Artificial Intelligence, vol. 209, pp. 11–28, 2014.

[29]. L. Du, Q. Song, and X. Jia, "Detecting concept drift: An information entropy based method using an adaptive sliding window," Intelligent Data Analysis, vol. 18, 03 2014.

[30]. C. Alippi and M. Roveri, "Just-in-time adaptive classifiers — Part I: Detecting nonstationary changes," Neural Networks, IEEE Transactions on, vol. 19, pp. 1145 – 1153, 08 2008.

[31]. C. Han and R. Ireland, "Performance measurement of the KCS customs selectivity system," Risk Management, vol. 16, no. 8, pp. 25–43, 2014.

[32]. N. Houlsby, F. Huszar, Z. Ghahramani, and M. Lengyel, "Bayesian ´ active learning for classification and preference learning," 2011.

[33]. O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in ICLR, 2018.

[34]. J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," in ICLR, 2020.

[35]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in KDD, 2016, pp. 785–794.
J. Vanhoeyveld, D. Martens, and B. Peeters, "Customs fraud detection: Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue," Pattern Analysis and Applications, vol. 23, 2020.

[36]. S. Kim, Y.-C. Tsai, K. Sigh, Y. Choi, E. Ibok, C.-T. Li, and M. Cha, "DATE: Dual attentive tree-aware embedding for customs fraud detection," in KDD, 2020, pp. 2880–2890.

[37]. H. Wang and Z. Abraham, "Concept drift detection for streaming data," in IJCNN, 2015.

[38]. P. Ball, J. Parker-Holder, A. Pacchiano, K. Choromanski, and S. Roberts, "Ready policy one: World building through active learning," in ICML, 2020, pp. 591–601.
[39]. A. Slivkins, "Introduction to multi-armed bandits," 2019.
[40]. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," Proceedings of IEEE 36th Annual Foundations of Computer Science, Aug 1998.
[41]. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," SIAM Journal on Computing, vol. 32, no. 1, p. 48–77, 2002.
[42]. R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, "POT: Python optimal transport," Journal of Machine Learning Research, vol. 22, no. 78, pp. 1–8, 2021.

Rehab Khaled Mohamed, et. al. "Adaptive Customization Detection Model based on Knowledge Discovery." *The International Journal of Engineering and Science (IJES)*, 13(1), (2024): pp. 01-15.