

Exploring Machine Learning Models to Predict Harmonized System Code

استكشاف نماذج التعلم الآلي للتنبؤ بكود النظام المنسق

by

FATMA ALI MOHAMED ALI ALTAHERI

**Dissertation submitted in fulfilment
of the requirements for the degree of
MSc INFORMATICS**

at

The British University in Dubai

November 2019

DECLARATION

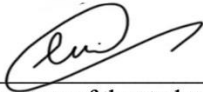
I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.



Signature of the student

COPYRIGHT AND INFORMATION TO USERS

The author whose copyright is declared on the title page of the work has granted to the British University in Dubai the right to lend his/her research work to users of its library and to make partial or single copies for educational and research use.

The author has also granted permission to the University to keep or make a digital copy for similar use and for the purpose of preservation of the work digitally.

Multiple copying of this work for scholarly purposes may be granted by either the author, the Registrar or the Dean only.

Copying for financial gain shall only be allowed with the author's express permission.

Any use of this work in whole or in part shall respect the moral rights of the author to be acknowledged and to reflect in good faith and without detriment the meaning of the content, and the original authorship.

Abstract

Globalization has shaped the way governments and government agencies operate; alongside how said phenomenon has consequently paved the way toward economic growth. With globalization, use of modern technology has also become a vital component in the public sector. Customs, for one, recognizes the importance of technology in ensuring efficiency of international trade. The Harmonized System (HS) Code is widely used across all customs departments because of the several benefits it yields for the government agency including a more convenient and easier approach for calculating fees and taxes. In that regard, it is the purpose of this study to explore ways to reduce the complexity, gaps and many other challenges in using HS Code in Dubai Customs, UAE using a case study approach and a machine learning-based HS Code Prediction model. This study uses six machine learning models based on the CRISP-DM framework. Initially, the study acquires the datasets from Dubai Customs and then analyses the data. Following this is the preparation of data for processing and the creation of the machine learning models. The results of the study indicate that machine learning models are effective tools in predicting HS Code for the user goods descriptions. In this study, six machine learning-based models have been implemented to determine the ability of detecting the HS Code based on the user's input description, where the highest achieved accuracy is 76.3% using linear support vector machine model.

Keywords: Harmonized System Code (HS Code); Customs, Revenue Protection; Machine Learning; Predictive Models; Trade.

ملخص

أدت العولمة على تشكيل طريقة حديثة لتعامل بين الجهات الحكومية إضافة إلى النمو الاقتصادي التي تشهده دول عديدة نتيجة إلى المنافع العديدة للعولمة علاوة إلى ذلك استنفادة القطاعات الخاصة بشكل كبير من العولمة من خلال لتكنولوجيا التي أصبحت وليدة للعولمة. ولعل العولمة أثرت بشكل ملحوظ في قطاع الجمارك مما أدى إلى تسهيل العمليات الجمركية والتجارة الدولية. من أهم الاستخدامات الدورية في قطاع الجمارك هي استخدام كود النظام المنسق المعروف ب (HS) وذلك للإيجابيات الهائلة على قطاع الدوائر الجمركية العالمية ومن أهمها تسهيل عملية حساب الرسوم الجمركية والضرائب. ومن هذا المنطلق قمت بدراسة تنص على استكشاف وتقليل التعقيدات والفجوات التابعة لهذا النظام وجعله أكثر سلاسة وسهولة لجمارك دبي التابعة لدولة الإمارات العربية المتحدة من خلال استخدام نهج دراسة الحالة ونموذج التنبؤ القائم على التعلم الآلي ستخدم هذه الدراسة ستة نماذج للتعلم الآلي بناءً على إطار عمل CRISP-DM. في البداية، بجمع الدراسة على مجموعات بيانات من جمارك دبي ثم تقوم بتحليل البيانات. ومن ثم إعداد البيانات للمعالجة وإنشاء نماذج التعلم الآلي. تشير نتائج الدراسة إلى أن نماذج التعلم الآلي هي أدوات فعالة في توقع كود النظام المنسق لوصف البضائع المستخدم، تم تنفيذ ستة نماذج معتمدة على التعلم الآلي لتحديد قدرة اكتشاف كود النظام المنسق استناداً إلى وصف مدخلات المستخدم، حيث بلغت أعلى دقة تم تحقيقها 76.3% باستخدام طراز جهاز ناقل الخطية linear support vector machine model.

Dedication

I dedicate my dissertation work to my parents, my sisters, my brothers, my whole family and friends. I thank them for their patience, prayers and their constant encouragement and support.

Acknowledgement

First and foremost, I have to thank my research supervisor, Prof. Khaled Shaalan and My colleague Mr. Omar AlQaryouti Without their assist and devoted involvement in every step throughout the process, this paper would have never been accomplished. I would like to thank you very much for your provision and support.

Most importantly, none of this could have occurred without my family. My parents, who offered their encouragement daily. Moreover, when faced some challenges I was and prepared to quit, you did not let me, and I am forever thankful. This dissertation stands as an evidence to your unqualified love and encouragement.

Table of Contents

Abstract.....	
1. Chapter One: Introduction.....	1
1.1. Background of the Study.....	3
1.2. Purpose and Objectives.....	5
1.3. Research Question.....	6
1.4. Problem Statement.....	6
1.5. Research Rationale.....	8
1.6. Thesis Structure.....	10
1.7. Summary.....	11
2. Chapter Two: Literature Review.....	12
2.1. Overview of the Harmonized System (HS) Code Classification.....	13
2.2. Benefits and Challenges of the use of the Harmonized System Code Classification16	
2.3. Theoretical Background on Machine Learning.....	20
2.4. The Concept of Learning in the Perspective of Machine Learning.....	21
2.5. Integration of Machine Learning to HS Code.....	24
2.6. Application of machine learning-based model.....	26
2.7. Summary.....	29
3. Chapter Three: Methodology.....	30
3.1. Business Understanding.....	31
3.2. Data Understanding.....	32
3.2.1. Data Analysis.....	32
3.2.2. Data Cleansing.....	33

3.2.3. Sample Processing.....	34
3.2.4. Tokenization	35
3.3. Experiments.....	36
3.3.1. Performance Measures	36
3.3.2. HS Code Prediction Models	38
4. Chapter Four: Discussion	44
5. Chapter Five: Conclusion and Future Prospects.....	45
References	47

List of Tables

Table 1: The confusion matrix attributes.....	37
Table 2: Performance Evaluation Adopted Machine Learning Models.....	42

List of Figures

Figure 1: Structure of the HS Codes based on HS 2017 (WCO, 2018)	14
Figure 2: Components of HS (WCO 2018).....	15
Figure 3: Machine Learning Techniques and Their Required Data (Mohammed, Khan and Bashier 2017).....	23
Figure 4: Architecture of Data Driven Smart Customs (Youyi, 2017)	26
Figure 5: Research Methodology	31
Figure 6: Machine Learning Prediction and Evaluation Approach.....	36

1. Chapter One: Introduction

The emergence of globalization has paved way to opening a number of opportunities towards achieving economic growth and prosperity. However, there are still certain challenges that pose as barriers to the effective implementation of processes within local governments. Customs is one of the most important government agencies, in the global perspective, that governs international trade and services process such as the declaration of goods and services, particularly in the aspect of trade and commerce. Generally, the verifying commodity descriptions ensures that goods and/or services are in compliance with government regulations to prevent improper or unlawful entry to the country of destination (Che, Xing and Zhang 2018). In this sense, Customs has the responsibility of ensuring that declared goods and/or services to be imported or exported are classified accordingly based on commodity descriptions.

Goods classification is one of the most important obligation of importer and exporter compliance. As such, it has become vital to customs departments worldwide to ensure that there is a delicate balance of reducing classification uncertainties and promoting effective classification systems. In response to this, several customs departments across the globe has integrated the use of modern technology by adopting the Harmonized System (HS) Code. According to Ding, Fan and Cheng (2015), the HS Code, also known as the Harmonized Commodity Description and Coding System, was developed by Brussels-based World Customs Organization (WCO) in order to cope with the rapidly increasing international trade worldwide. In line with this, Weerth (2008) explained that goods in a

customs tariff must be described fully within the customs declaration so they can be classified accordingly. For example, a wooden chair can be classified according to material condition or its function as furniture (Weerth 2008). This means that the use of HS Code has become important for customs departments because it promotes easier ways of calculating duties, fees and taxes, determining appropriate permits, licenses and certificates required and collecting trade statistics (Ding, Fan and Chen 2015).

Traditionally, declared goods are analyzed and inspected by inspectors using inspection images wherein results are used as grounds for decision making (Che, Xing and Zhang 2018). Using this approach, a number of weaknesses and concerns can be identified such as difficulty in identifying hazards in substances and ineffective classification of declared goods due to possible inefficiencies in the designation of harmonized system codes (Che, Xing and Zhang 2018; United Nations 2013). Aside from these weaknesses, there are also some challenges in the application of HS Code in relation to achieving satisfactory accuracy. These challenges include HS complexity, gaps in terminology and the evolving nature of the HS Code among others (Ding, Fan and Chen 2015).

This study aims to contribute to reducing HS complexity and gaps by developing a machine learning-based HS Code Prediction model. In order to provide a better outlook as to the impact of the adoption of the machine learning-based HS Code Prediction model, the case of Dubai Customs is used. This means that this study employs a case study approach focusing on the Customs Department in Dubai, UAE. It is expected that the integration of technological advancement such as Artificial Intelligence application to the

HS Code of Dubai Customs can contribute to addressing HS Code complexity and to enhancing its accuracy.

In carrying out the study, six machine learning models based on the CRISP-DM framework is used. Since the HS Code classification problem can be defined as a multilevel classification problem (Deng 2014), the datasets will be obtained from Dubai Customs. The datasets include the dataset with formal HS Codes and descriptions and the dataset with the customer provided descriptions and the corresponding HS Codes. The six machine learning models are integrated by performing 10-fold cross validation technique to assess the performance of machine learning models. The six machine learning models aim to use unseen data in order to measure the accuracy of learning behavior of machine learning model. The six machine learning models to be used in conducting this study include Naïve Bayes, K-Nearest Neighbor, Decision Tree, Random Forest, Linear Support Vector Machine and Adaboost. In carrying out this study, two settings are used. In the first setting, machine learning model predicts the entire HS Code correctly and in the second setting, the prediction ability of the header of the HS code is tested.

This study is based on the CRISP-DM framework which consists of six basic steps including business understanding, data understanding, data preparation, modeling, evaluation and deployment. Accordingly, this framework is mostly focused on collecting, describing and exploring the data (Roy, 2018). Since the integration of the machine learning models need large amounts of data, the CRISP-DM framework is considered as the most appropriate methodology for machine learning projects. According to Foroughi and Luksch (n.d.), the CRISP-DM framework describes the analytical process and is a beneficial tool because it is well-defined, structured and is a documented process of data

mining. Machine learning can be described as systems, tools or approaches that make computers recognize patterns from mass amount of correlated data to optimize operational efficiency and gain increased competitive advantage. As such, the use of the CRISP-DM framework can be viewed as the most viable basis for analyzing data to be used in the six machine learning models to establish a more efficient and accurate HS prediction code as a solution to forecasting problems.

1.1. Background of the Study

Dubai Customs is one of the earliest government agencies in the UAE, which operates a range of services and procedures to clients, including customs declaration and clearances, finance, customs registration and licensing, public auctions and others (Dubaicustoms.gov.ae). One of the most important services offered by Dubai Customs is customs declaration services, in which the agency process millions of declarations annually. By definition, customs declaration pertains to tan official document that lists and provides information about the goods that are to be exported or imported (Alsaedi, Burnap and Rana 2017). Goods are declared through the use of the Harmonized System (HS) Code, which is an important system for classifying traded goods. For example, The HS code 3818000 pertains to ‘chemical elements doped for use in electronics, in the form of discs, wafers, or similar forms.

The HS system can comprise around 5,300 product descriptions, in which correctly identifying the HS code for goods would be a challenging undertaking for customers and require the needed knowledge on how to apply the HS code. Any mistakes in the correct coding of the product for declaration would also cause some problems in the department (e.g. potential customer dissatisfaction, and revenue loss). Therefore, it would be the potential to create an HS code prediction model to address this issue and streamline the

efficiency of correctly coding goods to be declared. In developing the machine learning-based model towards its application to the HS code system of Dubai Customs, an important question can be answered: Can machine learning techniques be used in building a harmonized system code prediction model?

The use of machine learning techniques is already being adopted by several firms in various fields. In the study of Osoba and Davis (2017), machine learning methods, a subfield of artificial intelligence methods, rely on optimization procedures towards improving the processes of complex tasks and fostering deep and reinforcement learning through the use of data mining. As described by Murphy (2019, p. 6), “machine learning is a technical approach to AI that uses statistical algorithms to build (or learn) a prediction model by processing large amounts of multivariate data related to the phenomena of interest”. This suggests that machine learning can be used to develop a prediction model that can help predict outcome accurately from new sets of input data. This suggests that the use of machine learning in the implementation of the HS code can contribute to enhancing the accuracy of the classification system. In line with this, Mikuriya (2018) noted that having an accurate HS code means also having reliable information that is crucial to informing policy areas (i.e. economics, security, health and wellbeing, etc.) and to risk management. Thus, the application of the machine learning-based HS code prediction model can enable Dubai Customs to reliably predict outcomes (i.e. determining whether x-ray image of declared goods contains hazardous or illegal substances).

1.2. Purpose and Objectives

In the general perspective, the HS code is used to provide legal and logical structure comprising of a series of four-digit headings to enable countries and organizations to make further subdivisions in accordance to their needs. As such, the process of classification is

a complex task that requires the use of an effective and harmonized coding systems. Yet, despite the significant contributions of the HS Code system as a classification system at Dubai Customs, certain challenges still need to be addressed. Questions on HS still being fit for purpose in the contemporary are currently being asked. Thus, the main purpose of this study is to contribute to reducing HS complexity and gaps by developing a machine learning-based HS Code Prediction model. In particular, this research aims to address the following objectives:

- To analyze the effectiveness and accuracy of the HS Code System in relation to customs classification coding and strategic trade control implementation.
- To identify and explore the challenges of using the HS Code System in customs.
- To analyze the impact of the integration of a machine learning based HS Code Prediction model to the accuracy and effectiveness of the HS Code System at Dubai Customs.

1.3. Research Question

In line with the objectives of this research, this study aims to seek answers to the following research question:

- Can we use machine learning techniques to build a Harmonized System Code Prediction Model?

1.4. Problem Statement

As already mentioned, the Harmonized System (HS) Code is used by governments worldwide for the purpose of assessing duties and taxes as well as for the application of customs laws and regulations (Mikuriya 2018). In addition, the HS code system is also used by businesses to manage their trade regulatory obligations as well as to oversee their

supply chains (Mikuriya 2018). As such, the need for continuous revision of HS became imperative in order to ensure relevancy to the changing patterns of international trade. According to Mikuriya (2018), HS has already been amended six times since 2002 and the seventh amended is already underway in pursuit of producing terms and categorization that can serve international trade until its next revision cycle. Despite this, there are still some challenges in relation to meeting the high expectations of the international trading community. Issues such as simplifying the complexity of the HS structure, ensuring compliance, classification issues and concerns of whether HS can still be considered as a sufficient solution to achieving consistent classification (Mikuriya 2018; Grooby 2018). Therefore, the HS amendments pose considerable challenges wherein they may cause certain ambiguities in the transposition of HS codes in the new versions. Moreover, among the identified challenges in the implementation of HS include (United Nations Statistics Division (UNSD) 2017):

- HS is relatively complex thereby creating the need for extensive training of users with regards to its use to prevent the occurrence of major classification errors.
- Absence of stand-alone descriptors can lead to duplication of work because several nations and international organizations also develop similar descriptors.
- Inaccuracy in classification such that there are different factors at play such as differences in countries' circumstances and statistical priorities and use of groupings for certain commodities among others.
- Frequent revisions of HS can result in conflicts regarding discontinuation or merging of past and/or existing codes with new ones.

In Dubai, UAE, one of the oldest government departments is Dubai Customs. This department is strict in the dealing with traded commodities and imported goods. In line

with this, Dubai customs follows the unified tariff for GCC states that outlines unified description and classification codes to be used in categorizing traded goods/services. The appropriate classification of goods is the responsibility of the customs department so any wrong classification may cause revenue loss or cause customer dissatisfaction. This may include the duties underpayment or overpayment according to the provided classification. However, since there are millions of types of products and each product may contain hundreds or even thousands of subcategories, the classification of the HS code database of all items is a challenging task. Therefore, it is vital to explore machine learning approaches for faster HS code classification with less human error.

1.5. Research Rationale

Dubai customs holds one of the most important positions in the emirate of Dubai because of its significant role in maintaining the legality and security of international trade practices both at the regional and international levels. According to the Federal Customs Authority (2019), Dubai customs was able to continue its development and has already gained a good reputation at the global level. Part of its success can be attributed to the virtue of its advanced infrastructure and management facilities and services. In addition, Dubai customs also adopts the HS code in order to unify the classification of goods. The HS code system used by Dubai customs is an eight-digit code and requires all consignee to specify HS codes instead of just describing goods in the invoice prior to the generation of Bill of Entry (International Business Publication 2016).

Whilst there are several benefits in the use of HS code, there are also a number of challenges that need to be addressed. This study is focused on addressing gaps in the application of HS code at Dubai customs by developing a machine learning-based HS Code Prediction model. This study evolved out of the interest of the researcher in machine

learning, digitization and artificial intelligence. This interest developed due to the growing interest of scholars and researchers in understanding the positive benefits of machine learning in its integration with existing systems among different organizations worldwide but there are limited literatures available that explores the use of a model based on machine learning on classification systems such as the HS code. Therefore, the main rationale of this study is motivated from the interest of the researcher to contribute to empirical knowledge towards filling knowledge and awareness gaps with regard to the use of machine learning approaches in enhancing HS code at Dubai customs.

Interestingly, exploring machine learning approaches for promoting faster HS code classification with less human error is a subject that has received less attention in literature. According to Shalev-Shwartz and Ben-David (2014), there is a need for the adoption of machine learning rather than just programming computers or systems to carry out tasks to address problems of complexity and the need for adaptivity. This means that there are tasks that require machine learning programs in order to achieve satisfactory results. Despite this study focusing on the adoption of machine learning in enhancing the performance of HS code at Dubai customs, machine learning is an interdisciplinary field wherein it shares common threads with other mathematical fields, computer science and artificial intelligence.

To the best of the researcher's knowledge, this subject area focusing on Dubai customs is the first to be studied in the field of machine learning in the Arab region. Therefore, there is a great opportunity in the development of HS code prediction model towards streamlining the efficiency of accurate and relevant coding of declared and traded goods at Dubai customs. In addition, there is also a potential of making use of this model

as a framework towards enhancing the classification and coding practices in other customs organizations across the globe.

1.6. Thesis Structure

In undertaking this study, the paper is divided into six chapters in order to present a more concise flow of literature. In particular, this thesis is divided into the following chapters:

Chapter 1 is the introduction section. In this chapter, the subject of inquiry is introduced. The background of the study and the problem area are also discussed. In addition, this is the chapter where the objectives and questions of the study are presented. It is also in this chapter that the rationale for the study is explained.

Chapter 2 is the section of the literature review. In this chapter, several literatures about harmonized system code and machine learning are reviewed and analyzed. This chapter serves as the foundation of the collection of data wherein findings from the literature review are used as basis for justifying the case study and providing support to the answers to the research questions.

Chapter 3 is the methodology section. In this chapter, the different methods used in carrying out the research are explained in detail. It discusses the research design, research approach, sample and sampling method, data collection method, data analysis, ethical consideration and limitations of the research.

Chapter 4 is the results and findings section. It is in this chapter that the results of the study are presented and analyzed to generate the findings. The results are presented in visual format using graphs, charts or tables in order to promote a better understanding of information collected from large amount of data.

Chapter 5 is the discussion section. In this chapter, the findings from the study are discussed and applied to the case study. This is the chapter where an overview of the Dubai customs is presented as well as the application of the HS code prediction model to the HS code of Dubai customs is analyzed and justified.

Chapter 6 is the conclusion and future works section. In this chapter, key findings are discussed, and recommendations are suggested based on the findings. This is the final chapter of the thesis that also presents the generalizations grounded on the results of the study.

1.7. Summary

In this chapter, the topic about machine learning-based model for harmonized system code classification was introduced. The background of the study and the problem statement were also discussed in order to provide a better overview of what that this study is all about. To guide the research, the objectives and research questions were identified. Also, the motivation for and underlying principle of the research were explained in this chapter. Finally, the structure of this thesis was presented in order to provide a better overview of what to expect in the succeeding chapters.

2. Chapter Two: Literature Review

This chapter presents a review of related literatures related to the harmonized system (HS) Code, machine learning, machine learning-based prediction model and the application of HS code in government and private organizations. In addition, this chapter also reviews related literatures related to the benefits and challenges of the HS Code in the global perspective. As such, this chapter is aimed at addressing the following research objectives by conducting secondary research.

- To analyze the effectiveness and accuracy of the HS Code System in relation to customs classification coding and strategic trade control implementation.
- To identify and explore the challenges of using the HS Code System in customs.
- To analyze the impact of the integration of a machine learning based HS Code Prediction model to the accuracy and effectiveness of the HS Code System at Dubai Customs.

It is expected that this literature review can provide justifications and explanations about the subject area being discussed. This chapter serves as the foundation for formulating generalizations regarding machine learning-based model for harmonized system code classification. In line with this, this chapter is divided to the following subheadings:

- Overview of the Harmonized System Code Classification
- Benefits and Challenges of the use of the Harmonized System Code Classification
- Theoretical Background on Machine Learning
- Application of machine learning-based model

There may be some challenges in studying the application of machine learning based model in the perspective of HS code classification which is what this research is all about. To the best of the researcher's knowledge, there are limited studies that discussed customs HS code classification using a machine learning based model which makes it challenging to justify using secondary research. However, there are enough studies found in literature that attempted to explore the use of machine learning based model in different fields other than that of HS code classification. Therefore, section 2.4 mainly focuses on the application of machine learning based model in different fields as opposed to its use in HS code classification.

2.1. Overview of the Harmonized System (HS) Code Classification

The Harmonized System (HS) is a popular tool used by different organizations worldwide for the purpose of classifying goods and/or services. As described by United Nations International Trade Statistics (2017), HS is “an international nomenclature for the classification of products”. This means that the HS code allows participating countries to make use of standardized codes for classifying traded goods particularly for custom purposes. Thus, it can be noted that the harmonized system can be considered as the universal language in the aspect of international trade.

The HS code was introduced in 1988 and since then, has been adopted by several countries across the globe. It was developed by the World Customs Organization (WCO) as an instrument in pursuit of addressing one of the most fundamental needs of the government – to have the ability to classify and categorize goods and/or services being traded (WCO 2018). In addition, WCO (2018, p. 5) explained that HS “enables both decisions on immediate actions for specific goods (for example duty collections, restrictions or controls) and the use of the collated information to underpin economic and

trade related policies and planning”. This suggests that HS was developed in order to improve proper facilitation of international trade and regulation.

As a general description, the harmonized system is a structured nomenclature that consists of a series of 4-digit headings. Most of those headings are further subdivided in 5- or 6-digit subheadings in order to accommodate certain groupings of related products. Thus, HS represents a valuable instrument that can be used for different purposes whilst retaining a structure required for the purpose of tariff classification (WCO 2018). This means that the harmonized system was developed by the World Customs Organization as a core system to enable countries adopting it to make further subdivisions in accordance to their particular needs. Additionally, as illustrated in Figure 1 HS is also described as a multipurpose classification system (WCO 2018). This means that HS can be used for all types of transportable goods even if those goods may not be involved in international trade thereby reflecting flexibility and being multipurpose.

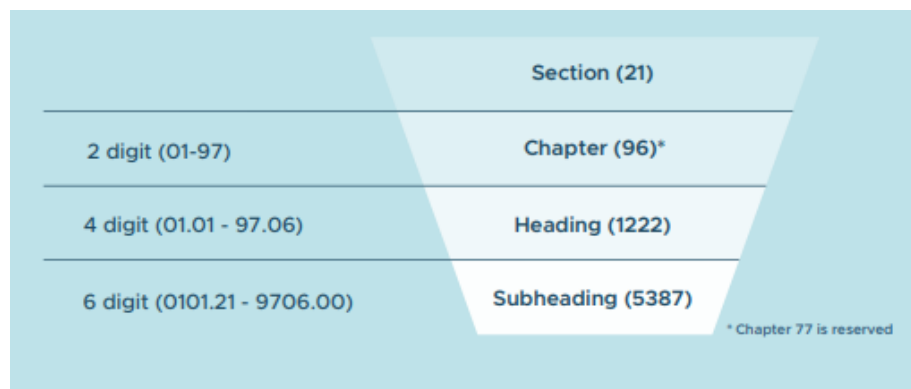


Figure 1: Structure of the HS Codes based on HS 2017 (WCO, 2018)

In a simpler description, harmonized system is described as an internationally standardized system of names and numbers or description and codes being used to classify traded goods. According to (Bernardino 2013), there are over two hundred (200) countries,

customs and economic unions that use HS which represents about 98 per cent of world trade. More so, Bernardino (2013) identified that the primary uses of HS include: (1) to determine custom tariffs or import duty; (2) to collect international trade statistics; (3) rules of origin; (4) to ascertain eligibility of a product under a free trade agreement; and (5) compliance with customer requirements.

The harmonized system is composed of the general interpretative rules (GIRs), sections, chapters, headings and subheadings as shown in Figure 2. The GIRs are the rules (consisting of 6 rules) which provides the principles of classification of goods under the HS. The sections are divided into 21 sections which are then divided into chapters. The HS has 99 chapters, chapter 77 of which, is being reserved for future use. The chapters are then divided into headings and are assigned with 4-digit codes. Finally, headings are divided into subheadings and are assigned with 6-digit codes (Bernardino 2013).

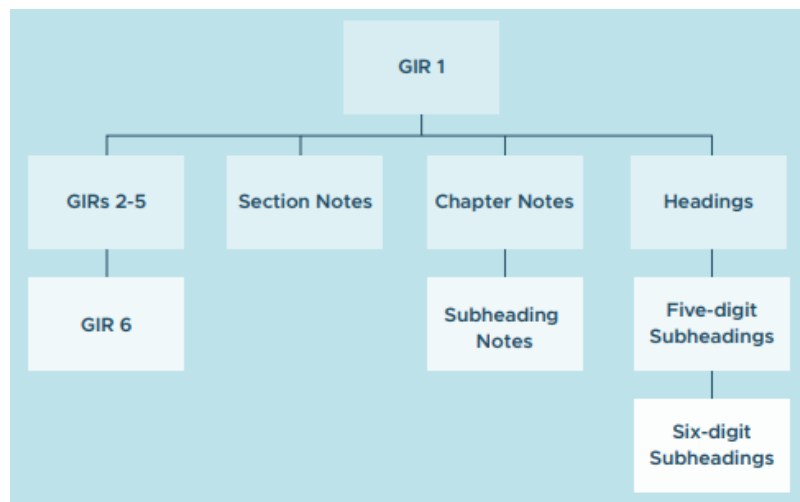


Figure 2: Components of HS (WCO 2018)

In the adoption of the HS, classification of codes of traded goods became plausible. Participating countries follow the fundamental rules relating to the classification of traded

goods in customs tariff that is in compliance with the HS in terms of description and coding of goods (Weerth 2008). As such, the use of the HS has become important in the process of classification of goods because developed and developing countries and several international organizations use it as a tool for implementing different trade policy instruments and international trade (Ulo & Ott 2006). Therefore, the adoption and implementation of HS has become an important part of strategic decision making and in the promotion of legal and unbiased international trade practices.

2.2. Benefits and Challenges of the use of the Harmonized System Code Classification

Whilst several countries worldwide have become accustomed to the adoption of the HS code, it is important to understand its impact on customs practices and operations. It has already been mentioned that the HS code is a complex system which means that it presents various benefits and challenges to governments and organizations implementing it. In the general perspective, the main benefit of the use of the harmonized system is allowing countries implementing it to classify and categorize traded goods using systematic and unified classification codes thereby promoting regulated trade that presents advantages to businesses and international trade (WCO 2018; Walden 1987). Despite this, the main challenge in the implementation of HS is essentially questions of classification relating to the development of possible current and future contingencies (WCO 2018).

The harmonized system is used as a basis for customs tariffs and collection of international trade statistics thereby making it known as the universal language of international trade. Under the HS convention, it is required that General Rules for Interpretation (GRIs) are used for HS application. This means that the HS application provides a unified and systematic nomenclature for the purpose of international trade

thereby enabling international comparability of trade statistics and facilitating detailed analysis of international trade (UNSD 2017). In addition, the application of the HS code contributes to ensuring greater ability for adopting countries to protect and monitor the value of their tariff concession (Yu 2008).

Furthermore, Walden (1987) asserted that the adoption of the HS provides the benefit of allowing countries to establish basis for advance clearance of traded goods. In particular, Walden (1987, p. 17) identified the following immediate benefits of the use of the harmonized system:

- HS provides more objective descriptions of goods with the use of characteristics that are easily measurable and observable by the inspectors.
- HS reduces chances of misclassification and/or poor description due to the agreement with import and export documents.
- HS improves information exchange between customs services.
- HS fosters more accurate commodity statistics that can enable trade negotiators to develop informed positions and refined economic predictions.

Finally, one of the most noted advantages of using the HS code is for countries to adopt a previously defined and structured system that contributes to reducing efforts required by trading parties in documentation of traded goods (Blue Whale 2019). This means that goods or commodities can move across customs in faster and smoother manner without having time constraints issues. As such, the movement of classified commodities flow efficiently. Also, the flexibility and reliability of the HS code makes it ideal for organizations, both private and public, to use for multiple purposes including trade policies,

custom tariffs, internal taxes and economic research, transport and trade statistics and analysis among others.

The use of the HS code has a number of benefits and advantages not just for customs but to other private, public and international organizations as well. Yet, along with these benefits, there are also certain challenges with the adoption and implementation of the HS code. As identified by Blue Whale (2019), among the disadvantages of the use of HS include:

- Due to the harmonization of standards of goods and commodities, the participating nations need to cope with the already set international standards by developed nations thereby enabling them to deal with their level of technology and advancements in manufacturing goods (Blue Whale 2019).
- According to Blue Whale (2019), the biggest challenge with the implementation of the HS code is related to the interpretation. The challenge is grounded on the subjectivity of the classification which means that there may be bias with the classification of goods based on the perspectives of the experts doing the classification. This means that there may be risk of commodities being wrongly classified that can lead to inaccurate taxes or penalties.
- There are organizations resorting to changing of codes to classify new commodities to prevent incorrect classification. The change of code, once approved, can be costly and can have legal and technical concerns relating to tax laws or delays at international borders (Blue Whale 2019).

Moreover, the United Nations Statistics Division (2017) noted that the HS is relatively complex and difficult to implement particularly in the absence of extensive

training. This means that in order to implement HS effectively, extensive training is needed in order to avoid significant classification errors which is costly. In addition, UNSD (2017) asserted that the definition of commodity groups in HS is not always satisfactory in relation to economic analysis which means that there is a need to develop different analytical classifications. Furthermore, UNSD (2017) also claimed that the frequent revisions of HS may result in discontinuation and/or merging of some codes that can cause breaks in time series that are important for analytical purposes. As highlighted by Yu (2008), the amendments to the HS also pose considerable challenges for its users because the codes and descriptions need to be transposed post amendment in the new version of HS nomenclature that fosters complexity of HS amendments.

Finally, implementation of HS can foster satisfactory accuracy in relation to classification, however, this is quite challenging to achieve. According to Ding, Fan and Chen (2015), the difficulties in the implementation of HS are mainly due to three factors including HS complexity, Gaps in terminology and the evolving nature of HS. Ding and colleagues explained that due to the structured multipurpose nomenclature of HS, there may be difficulties related to the proper classification of products. Additionally, HS needs to be revised or amended in a continuous manner in order to cope with the changes in the international trade market. This means that HS codes are being changed frequently thereby creating the need for having a classification system that is robust and adaptable to the continuous changes in the descriptions of goods (Ding, Fan and Chen 2015).

In order to promote effective adoption and implementation of HS, it is therefore important that countries are able to build on the strengths of HS and minimize weaknesses. UNSD (2017) suggested that it can be a good practice to make use of other product classifications as applicable depending on serving the purpose and needs of the

participating countries. Therefore, understanding the benefits and challenges of HS implementation can aid in promoting better decision making particularly in relation to classification decisions.

2.3. Theoretical Background on Machine Learning

In today's contemporary era, the concept of machine learning is of growing interest. It is currently being integrated with a number of far-reaching applications to systems with computational complexity. According to Shalev-Shwartz and Ben-David (2014, p. vii), "The term machine learning refers to the automated detection of meaningful patterns in data. In the past couple of decades, it has become a common tool in almost any task that requires information extraction from large data sets". This suggests that machine learning machine learning based technologies are currently popular integrated tools to foster automated learning.

Technological advancements have created numerous opportunities such as the creation of intelligent machines making it a possibility for machines to think and learn. Mohammed, Khan and Bashier (2017) asserted that the concept of machine learning was grounded on the discovery of Artificial Intelligence (AI) which started when scientists, David Rumelhart, Geoffrey Hinton and Ronald Williams, discovered a method that allowed networks to learn to discriminate between nonlinear separable classes. Based on this, an adapted computing system capable of learning was invented. As defined by KPMG International (2018, p. 44), machine learning is an application of artificial intelligence "...that explores and constructs algorithms that are able to learn and make predictions from data" wherein such algorithms are used to learn information directly from data without the need to rely on predetermined equation as a model. This means that there are a number of

machine learning models that can be used to identify and learn patterns and relations between data (labeled or unlabeled).

2.4. The Concept of Learning in the Perspective of Machine Learning

The concept of learning being used in machine learning can best be explained from the examples of naturally occurring animal learning. Shalev-Shwartz and Ben-David (2014) noted that there is learning mechanisms when animals use past experience towards acquiring expertise in achieving their goals. This means that the concept of prediction is at play wherein the animals can predict positive or negative effects when encountered in the future with the use of their past experience. Thus, Das and Behera (2017) noted that there may not be a significant difference in relation to how humans and animals learn with time and experience and how a machine can learn from experience.

In the perspective of machine learning, machines can be programmed to learn to do specific tasks. For example, a typical machine learning task is filtering spam e-mails wherein the machine can be programmed to memorize all previous emails labeled as spam by the human user and use this information to match old and new emails to detect spam emails (Shalev-Shwartz and Ben-David 2014). This method is known as learning by memorization. However, there are still some gaps with regard to the use of the learning by memorization approach such that machines should be able to learn without them being programmed explicitly (Mohammed, Khan and Bashier 2017).

Accordingly, learning should not be confined to humans alone because it is a phenomenon that can be exhibited by other living organisms such as animals, plants or amoeba among others. In addition, Simeone (2018) noted that machines can also learn with the help of the adoption of machine learning methodologies. In the traditional sense,

algorithms are developed and used in order to produce a trained machine that makes learning possible. In line with this, Das and Behera (2017) noted that the traditional approach of developing algorithms to promote learning can be enhanced through the use of machine learning.

In line with this, Mohammed, Khan and Bashier (2017) identified four general machine learning methods including supervised, unsupervised, semi-supervised and reinforcement learning methods. These machine learning methods can be used to achieve the primary goals of machine learning – “to enable machines to make predictions, perform clustering, extract association rules, or make decisions from a given dataset” (Mohammed, Khan and Bashier, 2017 p. xxi). Similarly, Simeone (2018) explained that the taxonomy of machine learning methods is represented by the three main classes of machine learning approaches known as supervised, unsupervised and reinforcement learning methods.

Supervised learning method encompasses the use of training set with the goal of learning a mapping between input and output spaces (Simeone 2018). In unsupervised learning, machines attempt to find a hidden structure from unlabeled data without any supervision (Mohammed, Khan and Bashier 2017). In relation to this, Mohammed and colleagues explained that in semi-supervised learning, the machine learns from a combination of labeled and unlabeled data towards generating an appropriate model for data classification. This is grounded on the theoretical analysis that machines can learn from a similar process of semi-supervised learning wherein individuals are supplied with unlabeled data by the environment and labeled data by the supervisor (Mohammed, Khan and Bashier 2017). Finally, reinforcement learning is situated between supervised and

unsupervised learning (Simeone 2017). This means that reinforcement learning applies to decision making problems wherein the learner interacts with an environment by means of taking sequential actions based on observations whilst receiving feedback about each of the selected action (Simeone 2017). Figure 3 illustrates the various machine learning techniques and its corresponding classifications.

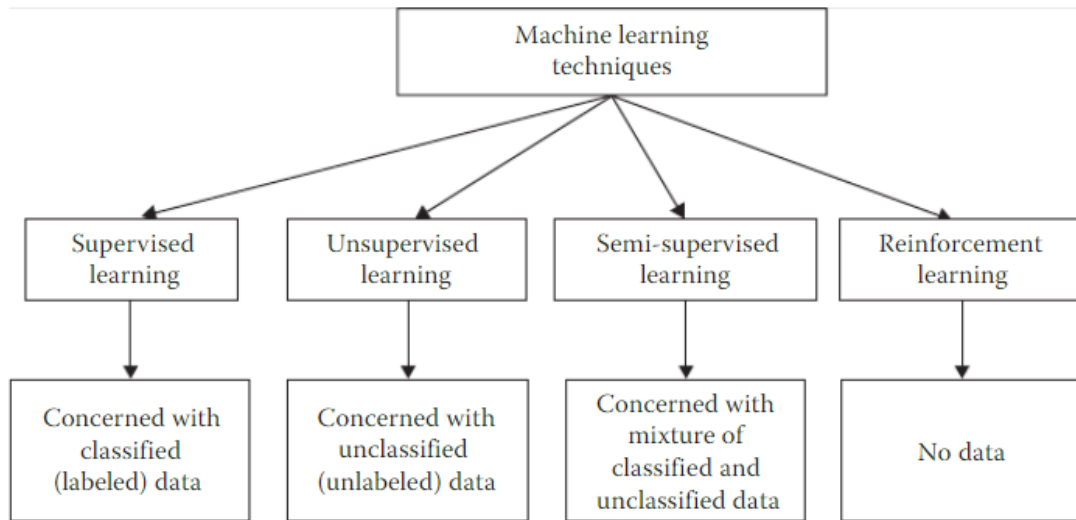


Figure 3: Machine Learning Techniques and Their Required Data (Mohammed, Khan and Bashier 2017)

Machine learning can be applied in various settings. As identified by Das and Behera (2017), among the most important real-life applications of machine learning include speech recognition, computer vision, bio surveillance, robotics or automation control and empirical science experiments. However, there are also some future sub-problems that must be considered including explaining human learning, programming languages that contains machine learning primitives and perception (Das and Behera 2017). Despite these concerns, machine learning in the overall perspective, can be used to provide software with flexibility and adaptability when necessary (Mohammed, Khan and Bashier 2017).

Therefore, it can be noted that machine learning algorithms can be applied to address real world challenges.

2.5. Integration of Machine Learning to HS Code

Machine learning methods are currently being integrated in various programs and systems in order to help organizations make better decisions particularly with regards to predictive analysis and pattern recognition. For example, machine learning methods are majorly being integrated in the security field such as to enhance facial recognition from using large amount of data from various sources which is difficult for humans to do manually (Mohammed, Khan and Bashier 2017). In addition, machine learning is also being integrated to various systems to promote automation and improve efficiency and accuracy. For example, machine learning techniques are being integrated in HS codes to foster accuracy, intelligence and automation (Zang, et al 2008). This suggests that multiple machine learning strategies integrated in various programs and systems can contribute to enhancing system performance, continuous learning and promoting more informed decision making.

In the perspective of customs organizations/ departments, the integration of machine learning has become vital in order for them to stay competitive in the international trade facilitation (Ding, Fan and Chen 2015). As explained by Ding and colleagues, different machine learning techniques are already being applied in capturing and learning long term and stable criteria for text categorization. For example, the use of keywords is one of the most common approaches of machine learning through the adoption of the vector space model (VSM) (Ding, Fan and Chen 2015). In addition, machine learning approaches are also being integrated in HS in pursuit of addressing HS code prediction problems (Luppés 2019).

Furthermore, machine learning strategies are also being integrated in HS in order to enhance automation process. According to KPMG International (2018), machine learning can be used to develop a knowledge base towards learning and developing a set of algorithms from large amount of data in order to make informed predictions. KPMG International (2018, p. 37) explained that “A combination of natural language processing and machine learning makes it possible to automate the capture, array and analysis of unstructured data and transform it into structured data that may be used in a tax application”. This suggests that the integration of machine learning in HS can contribute to process efficiency by means of improving quality, consistency and accuracy of code classification due to reduced likelihood of human errors. There are several machine learning tools and techniques that have been developed. Yet, machine learning can be viewed as the one that provides the technical basis for data mining (Witten, Frank & Hall 2011).

In order to promote effective application and integration of machine learning to HS, it is important to understand how it works. In the general perspective, machine learning can be viewed as the new technology for mining knowledge from data in order to learn and generate accurate predictive outcomes (Witten, Frank & Hall 2011). As such, machine learning can be viewed as being directly correlated with data mining. In line with this, machine learning techniques integrated with HS can also foster smart matching and classification through building data driven smart customs (Youyi 2017). As explained by Youyi (2012), a data driven smart customs can promote intelligent law enforcement, intelligent risk control and intelligent revenue collection wherein data collections can lead to intelligent applications (see Figure 4).

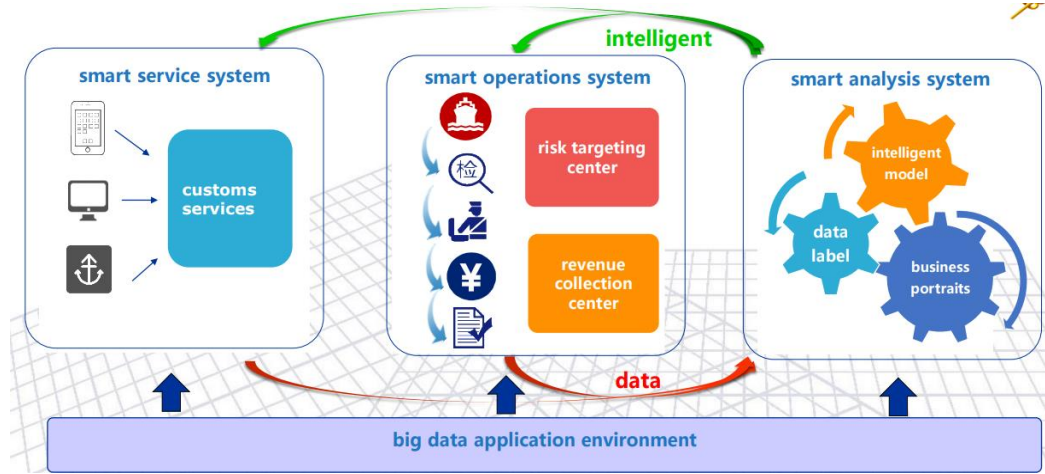


Figure 4: Architecture of Data Driven Smart Customs (Youyi, 2017)

Therefore, data mining is an important part of creating a machine learning framework that can contribute to promoting accurate and efficient classification and detection of tariff code as integrated in HS code systems. Collecting large amount of data set and generating useful information from collected data will not become a challenge due to the integration of machine learning in HS. As highlighted by Bramer, et al (2015), using machine learning based approaches and models can help enhance predictive analysis on streaming data and detect changes and inconsistencies in data to address HS code classification and description and business intelligence problems.

2.6. Application of machine learning-based model

In understanding the concept of machine learning, it can be noted that machine learning techniques can be applied effectively to HS code classification in order to enhance predictive analytics and reduce classification errors. According to Wu, et al (2019), a machine learning based model can be used as a framework to drive effective predictive control towards promoting improved system performance. Wu and colleagues explained that formulating a model predictive control (MPC) can help drive process performance to

desired set point. In line with this, Bishop (2013) asserted that the model based approach to machine learning can aid in creating a custom model that is tailored to specific applications used. Thus, Bishop noted that the model based machine learning can best be implemented using a model specification language.

The machine learning based model has been used in different fields such as in healthcare, security and education among others because of its ability to promote accurate prediction outcomes. For example, machine learning based model has been integrated in the medical field for the prediction of long term outcomes in medical related diseases such as acute stroke (Heo, et al 2019). Conversely, the machine learning based model uses big data to drive accuracy of predictive outcomes. As highlighted by Najafabadi, et al (2015), the adoption of big data analytics contributes to driving deep learning applications towards extracting complex patterns from large volume of data, semantic indexing, data tagging and simplifying multifaceted tasks.

Moreover, the machine learning based model is also being adopted to enhance categorization processes. Che, Xing and Zhang (2018) noted that deep learning methods as part of the machine learning based model, contributes to filtering out unreasonable predictions and justifying the resulted HS codes. Accordingly, integrating the machine learning based model in HS code classification can help address identical classification coding and classification inaccuracies. Moreover, Bishop (2013) specified that main goals of the integration of the machine learning based model approach to HS including fostering suitable inference and/or learning algorithms to generate more accurate predictive outcomes, can be designed in accordance to the requirements of the particular application, modifications can be integrated automatically and promoting transparency of functionality among others.

Furthermore, the machine learning based model can also be used in integration with the supervised learning strategy. According to Kaur, et al (2019), the machine learning based model can be used in enhancing predictive analytics towards evaluating nationwide happiness and subjective wellbeing. This suggests that the machine learning based model can be viewed as an effective predictive analytic model that provides better predictions and more accurate and consistent outcomes. Hence, a prediction model can be developed based on the machine learning based model to achieve appropriate predictors, data sources and outcomes.

The use of machine learning based models can aid in enhancing detection processes and code classification. In addition, the adoption of machine learning based model can be considered as being more cost effective as compared to the use of other specialized machine learning models (Nascimento, et al 2018). Nascimento and colleagues explained that the integration of the context aware approach to the machine learning based model can contribute to improving results by reducing bias in different machine learning tasks. Similarly, Choi, et al (2018) suggested that using machine learning based model on big data can contribute to the development of an effective prediction model. According to Choi and colleagues, learning based models can be developed by applying machine learning approaches such as decision trees, random forests, bagging and boosting. Therefore, it can be argued that using machine learning based model can contribute to promoting accurate and effective prediction outcomes that can be used to make better decisions.

2.7. Summary

There are a number of studies that explored the use of machine learning techniques in different fields such as in medicine (healthcare), big data, digital processes and others but there are very few that attempted to explain the integration of machine learning based models in harmonized system codes classification. This chapter therefore focused on reviewing literatures related to the harmonized system, machine learning and machine learning based models in the general perspectives. This chapter also explored the general context of harmonized system (HS) wherein it was noted that there are a number of benefits but also certain challenges in HS implementation. Thus, to address the challenges in HS implementation, a number of strategies have been adopted including the use of machine learning strategies. Based on literary findings, machine learning techniques are known to be effective in contributing to enhancing accuracy of predictive outcomes. In addition, it was also suggested that the development and use of machine learning based models can help improve predictive results, enhance detection processes and code classification and simplifying multifaceted tasks among others.

3. Chapter Three: Methodology

This study aims to investigate the possibility of implementing machine learning model to predict the HS code for the commodities based on the provided descriptions by the users. To perform this text mining task, we adopted the Cross-Industry Process for Data Mining methodology (CRISP-DM) (Shearer 2000). As illustrated in Figure X, this methodology consists of the following processes, namely business understanding, data understanding, data preparation, data modelling, performance evaluation and deployment. Business understanding refers mainly to the importance of the problem that is being address which is reducing the loss revenue by building machine learning model to correctly predict the HS Code. Data Understanding refers to the process of analyzing the available input information in order to produce an efficient machine learning model. Data preparation represents the preprocessing steps which are typically performed in order to remove any factor which may degrade the performance of the machine learning model. The machine learning modelling, evaluation and deployment involves the selection the appropriate machine learning techniques in order to investigate and evaluate these models and figure out the best machine learning model that produces the highest performance for deployment.

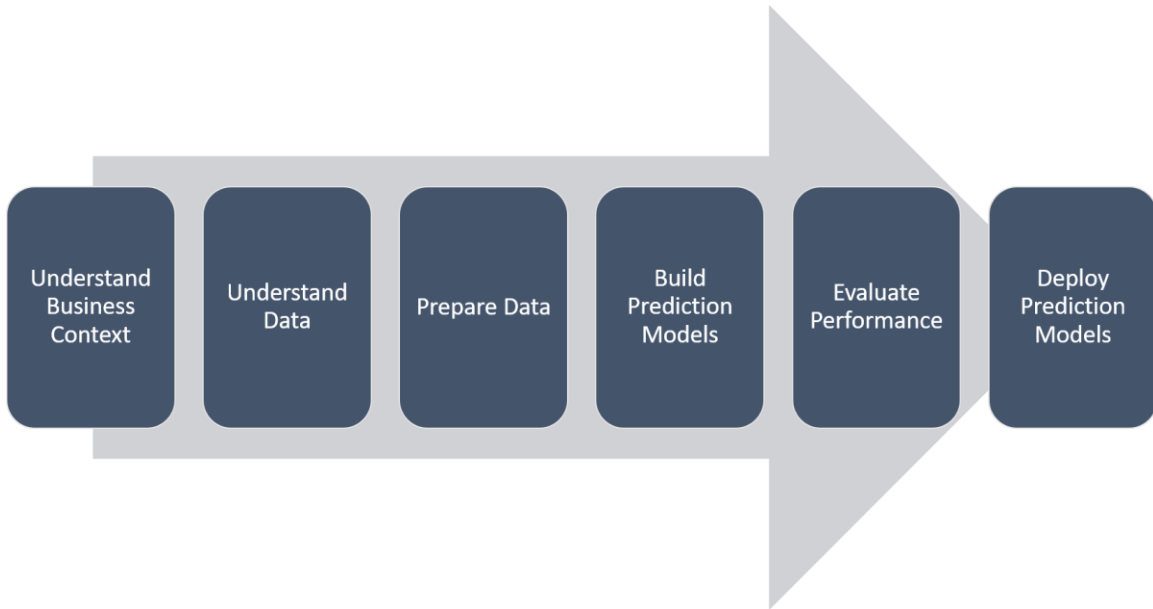


Figure 5: Research Methodology

This chapter starts by describing the source of the HS Codes dataset that is used to conduct the machine learning experiments. Furthermore, a detailed analysis on how the obtained dataset can be used to answer the driving research question. And this is followed by the data preparation tasks that includes data cleansing, sample processing and tokenization that will be used to build the inputs features to the machine learning models. This chapter also explains in details the machine learning models that are adopted and implemented throughout the HS Code prediction models.

3.1. Business Understanding

The importance of this work scope is to reduce the loss revenue caused by misclassifying the goods. This misclassification results in loss of duties. Such misclassification normally occurs to either the ability to the user not to correctly determine the required HS Code or in another situation the manipulation by the users to reduce the amount of duty that he/she is required to pay. In this work, the focus is on the first issue

since the second issue requires other authentication techniques to determine whether the user is intentionally misclassifying the goods.

3.2. Data Understanding

The data used in this work are provided by Dubai Customs through the Artificial Intelligence (AI) hackathon competition that was conducted on October 2019. This data consists of 22,346,194 records where each record has two attributes; the Harmonized System Code (HS Code) and the user inputs description. In this section, we will describe the processing techniques that we adopted in order to review, analyze and prepare the input data.

3.2.1. Data Analysis

During the analysis of the provided data, we have noticed several major factors which must be taken into consideration during the development of the machine learning based model. These factors are mainly related to the quality of the user's description and the number of expected classes (labels).

About the user's description, several issues have been noted. For instance, some of the descriptions do not contain any valuable keywords from the English dictionary. In another words, the spelling mistakes in these descriptions are too severe to the point that they could not be understood by the human mind. Also, we have noticed that the same description has been used several times to describe totally different items.

Regarding the expected number of classes, the provided data has almost 8,000 labels. And this underlines the problem associated with establishing any machine learning model to predict the label based on the provided description. Additionally, it has been noticed that they are severe variations in terms of the number of records that are associated with each

class. For instance, some of the classes have around 10 records while other classes may have up to 700 records. This factor must be addressed in order to avoid having any bias behavior in the machine learning model.

3.2.2. Data Cleansing

The above-mentioned factors have been taken into consideration during the cleaning of the data. And in this section, we will present the procedural steps that have been performed during the cleaning of the data. These steps can be summarized as follows: (1) remove duplications (2) remove punctuation and remove stop words (3) remove non-English words (4) remove numbers and (5) lemmatization.

Remove Duplications: Duplicated records might have negative impact on the overall performance of the machine learning models. Therefore, we have removed all duplicated records based on the users' description as a pre-processing step.

Remove Punctuation and Stop Words: Eventually, the punctuation and stop words have no impact in the training behavior of any machine learning model since they are common words that do not add any distinct features to the description. Therefore, the punctuation and stop words have been removed in order to speed up the training process.

Remove Non-English Words: Normally, the descriptions are provided by users who have basic understanding of the English language. And this may result in writing words with spelling mistakes which may confuse the behavior of the training because at the training stage, these non-English words will be interpreted with different words that have different meaning. Therefore, these non-English words were removed during the cleansing step.

Remove Numbers: In this domain, it has been noticed that numbers do not provide any valuable features to the goods. Thus, the number have been removed as well in order to speed up the training for the machine learning model.

Lemmatization: This process is used to return the word to its origin. Thus, applying lemmatization is expected to reduce the vocabulary size and improve the performance since words that have the same origin must be treated equally.

In addition to the above-mentioned cleansing steps, all text has been converted to lower case in order to reduce the vocabulary size since any machine learning model is case sensitive.

3.2.3. Sample Processing

Having classes with very big variation in terms of records is expected to create a significant impact on the training behavior. For instance, if we assume using the K-nearest neighbor to predict a binary class label where the first class has 100 records and the second class has only 10 records. In this case, the model may mis-classify the data towards labeling the reading by the 100-records label class. Accordingly, this sampling issue must be addressed to ensure consistency in terms of the classification behavior.

To address this problem, we have reported the minimum and maximum records for the 8,000 classes that are provided in the dataset. During this step, it has been noticed the gap between the minimum and the maximum reaches to 750 records. Thus, we have performed down sampling to achieve 100 records as a gap between the minimum and maximum class size.

In this work, due to computational limitation of the used machine, we have selected only 500,000 at random as the input data. Additionally, once all pre-processing steps are performed, 217,700 records are kept as the final data set input.

3.2.4. Tokenization

In Natural Language Processing (NLP), tokenization is an important step in order to determine weight (importance) of each word in the text. In general, two major techniques have been widely used to perform tokenization. These techniques are: (1) bag of words and (2) Term Frequency - Invert Document Frequency (TF-IDF). The bag of words is simply a counting method where each word is given a value that represents the frequency in which the word appears in the text. Accordingly, the use of this technique depends whether the number of appearances of each word in the text is considered as distinct features in each model.

However, the TF-IDF is considered as a more advanced technique since it uses the frequency of the word in order to determine the uniqueness of each word in the text. The TF-IDF starts by calculating the frequency of each word in the text ($Fr(w)$). This is calculated by dividing the number of times word (w) appears in the text over the total number of words. This denotes the term frequency in the calculation (TF). The inverse document frequency (IDF) is calculated as follows:

$$IDF(w) = \log\left(\frac{N}{Nw}\right)$$

Where N denotes to the total number of records (the users' HS Code descriptions), and Nw denotes the number of records that has the word w . This inverse is normally calculated to associate a weight value with each word where words with less appearance

have more weight. Eventually, the TF-IDF score for a word w can be represented as follows:

$$TF-IDF(w) = TF(w) \times IDF(w)$$

The equation shows that the TF-IDF aims to magnifies the weight of the words that appears in a smaller number of records.

3.3. Experiments

In this section, we present and discuss the machine learning models that are used in this work and the corresponding performance of each model. These experiments were performed using Python 3.5 with scikit learn library for model specifications. To evaluate the performance for each model, we have adopted the following evaluation metrics: precision, recall, F1-measure and accuracy. The machine learning prediction approach is illustrated in Figure 6.

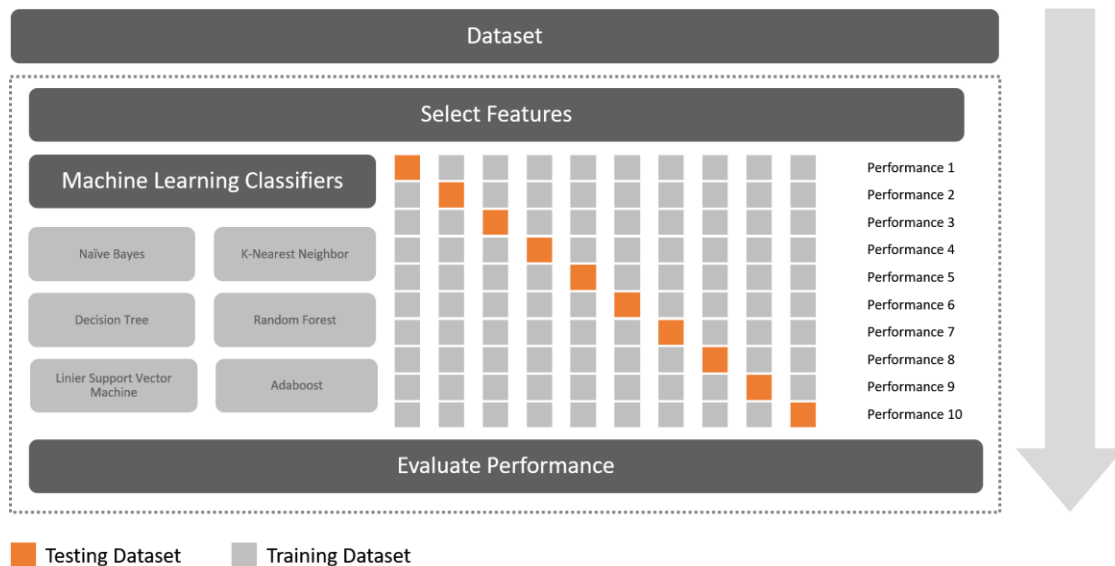


Figure 6: Machine Learning Prediction and Evaluation Approach

3.3.1. Performance Measures

For each HS Code label, precision refers to the number of correctly classified description that belongs to this label. For instance, the precision for HS Code (00000000)

refers to the percentage of correctly classified goods that belong to this label where the equation of calculating the precision is show below:

$$P = \frac{TP}{TP + FP}$$

Where, number of goods that are classified as (00000000) label is referred as True-Positive (TP). The False-Positive (FP) denotes the number of goods (descriptions) that have been misclassified as (00000000) label.

Recall (R) denotes the correctly classified labels. For instance, the recall for the class (00000000) in the above example refers to the correctly classified goods that belong to this label and is calculated as per the below equation:

$$R = \frac{TP}{TP + FN}$$

Where, false-negative (FN) denotes the number of goods (records) that belong the class (00000000) and are misclassified as other classes.

The **F1-score** is used a measure to determine the quality of the classification, where a higher F1-measure score highlights better classification quality. The F1-measure score is calculated according to the following equation:

$$F1 = 2 \times \frac{R \times P}{R + P}$$

The accuracy refers to the number correctly classified goods by the machine learning model and it is calculated as follows:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

Table 1 highlights the relationship between the TP, TN, FP and FN where this relationship is expressed for each class label.

Table 1: The confusion matrix attributes

Actual	Predicted	
	Yes	No

	Yes	<p style="text-align: center;">TP (true positive) Number of goods that are correctly classified</p>	<p style="text-align: center;">FP (false positive) Number of goods that are wrongly classified</p>
	No	<p style="text-align: center;">FN (false negative) Number of goods that do not belong to a particular class, but is classified as a member of this class</p>	<p style="text-align: center;">TN (true negative) Number of goods do not belong to a particular class and actually classified as it does not belong to this class</p>

3.3.2. HS Code Prediction Models

In the experiments that are presented in this work, we have performed 10-fold cross validation technique. This technique which is also denoted as rotation estimation is typically used to assess the performance of the machine learning model by generating different independent equal sets ($k=10$). In general, any machine learning model works by using the training set to learn the behavior of the data. Whereas, the test set is the unseen data that is used to measure the accuracy of the learning behavior of the machine learning model. In cross validation, the experiment is performed k times where in each round the data is divided into training and testing sets. In each round the selected testing set must be different where the performance is monitored over the k runs in order to better analyze the model performance.

The main ingredients of this work are to use the following machine learning models in order to evaluate and answer the proposed research question. These machine learning models are: Naïve Bayes, K-Nearest Neighbor, Decision Tree, Random Forest, Linear Support Vector Machine and Adaboost. In these machine learning models, we will perform

the prediction model on two experiment settings. The first setting, we will test whether the machine learning model is able to predict the entire HS Code Correctly. Whereas, in the second setting, we will test the ability of only predicting the header of the HS Code. The header denotes the first four digits of the HS Code which represent the chapter (2 digits) and section (2 digits). The chapter and section represent the commodity type (e.g. coffee, skin products)

The input data for the machine learning models are expressed as a tuple that consists of the following features: $\langle HS\ Code, Description\ Text \rangle$ where, once the tokenization is performed, the description text is converted to several features that represent the weights of the words in each description. HS Code represents the label that will be predicted.

In its core, the Naïve Bayes classifier uses the Bayes theorem to predict the probability of membership. The main idea in this classifier is to use conditional probability in which the probability is calculated between each input feature and the corresponding classes. In other words, if we assume that we have 5 features and 2 classes, the mechanism of this classifier works by determining the conditional probability of each feature and the 2 classes. This classifier is very efficient in text-based modelling where the input is expected to have distinct words that highlight the advantages of using the conditional probability concept. To test the performance of the Naïve Bayes classifier in predicting the HS Code, we ran the experiment where Table 2 shows the performance metrics. From the table, we can see that the accuracy of this classifier is relatively low (52.43%) and this highlights the fact that the input description consists of relatively common words that do not distinguish between the input record efficiently. Table 2 shows the performance measure of this classifier where the label of each class only represents the header. Recall that the header represents the first 4 digits of the HS Code. And this header classifies an

abstract level the goods type. From this table, we can see that the by considering the header of the label, the performance of the classifier has improved significantly, where the classifier achieves 66.21% accuracy. This is related to the fact that by aggregating the records and only perform the classification on the HS Code header, we reduce the pressure on the classifier since the number of labels have been reduced significantly.

K-Nearest Neighbor is a discriminated classifier in which the prediction works by taking into consideration the observation around the location of the new data to be predicted. To clarify this concept, assume that we are planning to predict the class label of an input record using 5-Nearest Neighbor. In this example, once the new record is represented in the solution space, the 5-Nearest Neighbor works by determining the class label of the 5 nearest point to the new record. Then, a voting mechanism is applied and the new record is assigned to the label with the major points. The key performance factor of this classifier is the number of neighbors that are considered during the search (k). in this work, we have noticed that the performance of this classifier stabilizes once the k reaches to 15 ($k=15$). Thus, in the experiment shows below, we have sat k to 15. Table 2 illustrates the performance measures for this classifier. From the table, we can see that the K-Nearest Neighbor classifier performs slightly better than the Naïve Bayes model. This improvement (around 5%) in performance is related to the cleansing procedures that are applied in this work since it contributes towards eliminating unnecessary features and therefore reduce the dimensionality of the solution space. In addition, Table 2 shows the performance for only the header of the HS Code. From this table, it is clear that the accuracy of this classifier has been improved (71.72% accuracy) due to the above-mentioned factors similar to Naïve Bayes.

Decision Tree classifier uses the input data to build a tree. This tree is eventually a conditional statement (if-else statement). In this mechanism, the node of the trees is selected based on the corresponding label. In other words, by starting from the root, the selection of the node will be based on the features and their impact on the class label. In each level, the features with the most impact on the class label will be selected as this particular level nodes. The behavior and performance of this classifier depends on the length of the tree and the number of labels. In our experiment it is not expected that this classifier to obtain significant classification behaviors due to the high number of labels. This is confirmed by the results which is shown in Table 2, where this classifier achieves 47.84% accuracy for the entire HS-code matching experiment.

Random Forest is an ensemble learning technique in which the classification is performed based on the outcome of several learning models. In this work, the sub-learning model represents decision trees. By combining several decision trees into a random forest, the boundary of prediction is expected to be more accurate. And this is was clearly illustrated by the achieved results (66.21%) where the accuracy has improved significantly compared to the previously discussed Decision Tree classifier as shown in Table 2.

Linear Support Vector Machine is an efficient well-known method that has been widely adopted to address several machine learning problems. Typically, a machine learning model tries to determine a hyperplane which divides the solution space according to the classes. For instance, assume that we have a binary problem, a hyperplane in this case is expected to divide the space such that one of the classes will be located to the right of the hyperplane and the other is located to the left of the hyperplane. In Linier Support Vector Machine, the optimization objective is to determine the location of this hyperplane such that the distance between the nearest point in each class and the hyperplane is

maximized. This distance is typically referred to as a margin and therefore, the problem addressed in this situation can be denoted as maximizing the total margin. The results shown in Table 2 indicates that the linear Support Vector Machine achieves significant results than other classifiers in all of the cases (75.4%), where in the header only case, the improvement is much significant compared to the entire HS Code experiment (84.58%). This is related to the fact that in the header only case, the number of classes will be significantly lower and this reduces the complexity of locating an efficient hyperplane to separate the classes. In other situation where the experiment is performed to predict the entire HS Code, the complexity of the solution space increases the pressure on the classifier to obtain an efficient hyperplane locality.

Table 2: Performance Evaluation Adopted Machine Learning Models

Machine Learning Model	Experiment Settings	Precision	Recall	F1-Measure	Accuracy
Naïve Bayes	HS Code Header	73.66%	55.66%	63.40%	66.21%
	Entire HS Code	59.67%	29.45%	39.44%	52.43%
K-Nearest Neighbor	HS Code Header	72.83%	57.75%	64.42%	71.72%
	Entire HS Code	55.30%	26.66%	35.97%	57.94%
Decision Tree	HS Code Header	70.21%	46.92%	56.25%	61.62%
	Entire HS Code	51.00%	20.72%	29.47%	47.84%
Random Forest	HS Code Header	96.35%	64.39%	77.19%	79.99%
	Entire HS Code	94.00%	38.19%	54.31%	66.21%
Linear Support Vector Machine	HS Code Header	96.35%	70.55%	81.46%	84.58%
	Entire HS Code	95.06%	51.41%	66.73%	75.40%
Adaboost	HS Code Header	73.66%	67.44%	70.41%	75.40%
	Entire HS Code	51.00%	25.10%	33.65%	57.02%

Adaboost is an ensemble learning technique where the main idea of this classifier is to determine the weight of each sub-classifier based on its behavior in the overall evaluation. For instance, assume that Adaboost consists of three round optimizations. In each round, the weight of the classification factors depends on the outcome of the previous round of evaluations. Thus, this classifier aims to optimize the classification factors weights in iterative manner based on the behavior of the previous round. Unexpectedly, this classifier performance was relatively low compared to other machine learning classifiers as shown in Table 2. This is related to the fact that in our problem the number of features (after tokenization) is high. Therefore, Adaboost, failed to obtain an efficient classification behavior during the experiments which is bounded by 24 hours running time.

4. Chapter Four: Discussion

This study aims to investigate the problem of predicting the HS Code for the goods based on the users' input description. This prediction aims to reduce the revenue loss caused by using wrong HS Codes in order to reduce amount of duty that the trader is expected to pay to the customs administration. From the results, it is clear that the machine learning models can be used to help in predicting the HS Code for the user goods descriptions. The relatively low classification performance can be contributed by different factors such that the quality of the description and in some situations the planning of providing falsified descriptions. Since any machine learning model will be able to predict the description quality is bad due to poor English skills or due to the intention of the user to provide such low-quality description. The work that has been carried out to answer the research question which deals with the ability of using machine learning techniques to build HS Code Prediction Model establishes the ability to perform such tasks using machine learning techniques. To answer the main research question of this dissertation, the dataset that is provided by Dubai Customs was used to predict the HS Code of the provided users' description using six machine learning models. The adopted machine learning models have achieved promising results specially the linier Support Vector Machine which achieved the highest accuracy of 76.3%.

5. Chapter Five: Conclusion and Future Prospects

This study underpins the challenges that confront Customs relating to the use of Harmonized System (HS) Code in their operations. Through the use of the case study approach and the development of a machine learning-based HS Code Prediction Model. Considering the benefits of using the HS code, there is subsequently a need to also address the issues it entails. The significance of this study is rooted on the interest of the research to contribute to empirical knowledge, especially in filling in gaps about the use of machine learning approaches in improving HS Code at Dubai Customs. As the results of the study reveal, the machine learning models are useful in predicting HS Code for the user goods descriptions. In addition, it is also noted that one of the six machine learning models used, the Linear Support Vector Machine, reveals a high accuracy of 76.3% in relevance with predicting HS Code using the dataset provided by the Dubai Customs.

However, this study is also not without limitations. Since a case study approach was adopted, the research is only limited to Dubai Customs as the machine learning models were only used on the dataset obtained from said government agency for predicting HS Code. On the other hand, information presented in the study can be used for future research, specifically on how gaps and other challenges in HS Code can be addressed among government agencies that use the modern technology. It can also be used to explore and identify strategies that will positively affect the efficiency of using HS Code in the public sector. The adoption of machine learning models as predictive frameworks and their respective effectiveness and/or efficiency can additionally be examined in future

Due to the complexity of the problem, as a future step, we will explore the applicability of developing a divide-and-concur approach to address this problem. Whereas, hierarchical prediction can be built starting from the HS Code header until identifying all of the HS Code subsections. Additionally, we are planning to explore the benefits of employing deep learning-based approaches to address this problem.

References

- Alsaedi, N., Burnap, P. and Rana, O. (2017). Can we predict a riot? disruptive event detection using twitter. *ACM Transactions on Internet Technology (TOIT)*, 17(2), 1–26,
- Bernardino, A. D. (2013). Seminar on Classification, GIRs and Import Duty Determination. Retrieved from http://www.philexport.ph/c/document_library/get_file?uuid=721b6c4e-96ab-4c53-b90f-ec4ecef91bb&groupId=127524
- Bishop, C. M. (2013). Model-Based Machine Learning. *Phil Trans R Soc A*, 371(1984), 1–17
- Blue Whale. (2019). Harmonized Commodity Description and Coding System or Harmonized System (HS) Code. Retrieved from <http://www.blue-whale.in/customs/harmonized-commodity-description-and-coding-system-or-harmonized-system-hs-code/>
- Che, J., Xing, Y. and Zhang, L. (2018). *A comprehensive solution for deep-learning based cargo inspection to discriminate goods in containers*. IEEE Xplore.
- Choi, C., Kim, J., Kim, J., Kim, D., Bae, Y. and Kim, H. S. (2018). Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data. *Advances in Meteorology*, 2018, 1-12
- Das, K. and Behera, N. (2017). A Survey on Machine Learning: Concept, Algorithms and Applications. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(2), 1301-1309

- Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning ERRATUM. *APSIPA Transactions on Signal and Information Processing*, 3.
- Ding, L., Fan, Z. and Chen, D. (2015). Auto-Categorization of HS Code Using Background Net Approach. *Procedia Computer Science*, 60, 1462-1471
- Federal Customs Authority. (2019). Dubai Customs. Retrieved from <https://www.fca.gov.ae/En/UAE-Customs/Pages/DubaiCustoms.aspx>
- Goodfellow, I., and Bengio, Y. and Courville, A. (2016). Deep Learning. MIT Press
- Grooby, G. (2018). Is the HS Still Fit For Purpose? In: *The Harmonized System from Every Angle*. World Customs Organization.
- Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S. and Heo, J. H. (2019). Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. *Stroke*, 50(5), 1263-1265
- International Business Publications. (2016). *United Arab Emirates Investment and Business Guide, Volume 1 Strategic and Practical Information*. Washington, DC: International Business Publications
- Kaur, M., Dhalaria, M., Sharma, P. K. and Park, J. H. (2019). Supervised Machine-Learning Predictive Analytics for National Quality of Life Scoring. *Applied Sciences*, 9(1613), 1-15
- KPMG International. (2018). Transforming The Tax Function Through Technology. Retrieved from <https://home.kpmg/content/dam/kpmg/au/pdf/2018/transforming-the-tax-function-au.pdf>

- Luppés, J. (2019). *Classifying Short Text for the Harmonized System with Convolutional Neural Networks*. Master Thesis, Radboud University
- Mikuriya, K. (2018). The Harmonized System, 30 years old and still going strong! In: *The Harmonized System from Every Angle*. World Customs Organization.
- Mohammed, M., Khan, M. B. and Bashier, E. B. M. (2017). *Machine Learning - Algorithms and Applications*. Boca Raton, FL: CRC Press
- Murphy, R. F. (2019). *Artificial Intelligence Applications to Support K-12 Teachers and Teaching*. RAND Corporation
- Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R. and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), 1-21
- Nascimento, N., Alencar, P., Lucena, C. and Cowan, D. (2018). *A Context-Aware Machine Learning- Based Approach*. Proceeding of CASCON '18 Proceedings of the 28th Annual International Conference on Computer Science and Software Engineering, Markham, Ontario, Canada — October 29 - 31, 2018
- Osoba, O. A. and Davis, P. K. (2017). *An Artificial Intelligence/Machine Learning Perspective on Social Simulation: New Data and New Challenges*. RAND Corporation
- Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

- Singh, A. and Sahu, R. (2004). *Decision support system for HS classification of commodities*. The 2004 IFIP International Conference on Decision Support Systems (DSS2004).
- Ulo, H. & Ott, K. (2006). Classification and coding: Approach of different international organizations. *Transport*, 21(3), 189-196
- United Nations. (2013). *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*, 5th Revised Edition. Geneva: United Nations
- United Nations International Trade Statistics. (2017). Harmonized Commodity Description and Coding System (HS). Retrieved from <https://unstats.un.org/unsd/tradekb/Knowledgebase/50018/Harmonized-Commodity-Description-and-Coding-Systems-HS>
- United Nations Statistics Division (UNSD). (2017). Benefits and Challenges Associated with the Use of the HS. Retrieved from <https://unstats.un.org/wiki/pages/viewpage.action?pageId=7405716>
- Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4), 13-22.
- Walden, E. (1987). AMS. In: Operation Alliance. Department of the Treasury US Customs Service
- Weerth, C. (2008). Basic Principles of Customs Classifications under the Harmonized System. *Global Trade and Customs Journal*, 3(2), 61-67
- Witten, I. H., Frank, E. and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd Edition. Burlington, MA: Elsevier

- World Customs Organization (WCO). (2018). *The Harmonized System*. Belgium: World Customs Organization.
- Wu, Z., Tran, A., Ren, Y. M., Barnes, C. S., Chen, S. and Christofides, P. D. (2019). Machine Learning-Based Model Predictive Control of Distributed Chemical Processes. *IFAC Papers Online*, 52(2), 120-127
- Youyi, W. (2017). Exploration of Data-Driven Intelligent Customs. Retrieved from https://www.eiseverywhere.com/file_uploads/c7e054aa02ad13907d6ad513ea57b8d9_session3-YouyiWu.pdf
- Yu, D. (2008). *The Harmonized System - Amendments and Their Impact on WTO Members' Schedules*. World Trade Organization (WTO).
- Zang, B., Li, Y., Xie, W., Chen, Z., Tsai, C. and Laing, C. (2008). An ontological engineering approach for automating inspection and quarantine at airports. *Journal of Computer and System Sciences*, 74, 196-210