TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Serkan Ahmet Koch 194189IVGM

MODERNIZATION OF COMMODITY CLASSIFICATION PRACTICES IN TRADE FOR EU CUSTOMS WITH MACHINE LEARNING WEB SOLUTIONS

Master thesis

Supervisor: Eric Blake Jackson
PhD candidate
Co-supervisor: Markko Liutkevičius
PhD candidate

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Serkan Ahmet Koch 194189IVGM

MASINÕPPEL TUGINEVATE VEEBILAHENDUSTEGA EUROOPA TOLLINDUSES KAUBA KLASSIFITSEERIMISPRAKTIKATE KAASAJASTAMINE

Magistritöö

Juhendaja: Eric Blake Jackson
Doktorant

Kaasjuhendaja: Markko Liutkevičius
Doktorant

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Serkan Ahmet Koch

10/5/2020

Abstract

Proper classification of trade commodities both internationally and in the EU is an increasingly demanded capability by private and public based organizations. International organizations such as the WTO and WCO have fortunately come up with internationally renowned standards that, if utilized appropriately, can very effectively benefit the process of identifying trade goods. The task of identifying/ classifying trade items according to standards or nomenclatures issued by the mentioned governing bodies is usually still carried out manually in many places across the globe. Europe is not an exception, and EU member states that are not well acquainted with digital advancements fall in this category. At the same time, the European Commission is creating new regulatory requirements for classification of e-commerce goods and customs declarations that accompany them, the swift adoption of which is expected and rather required by the EU. Reasons for the emerging legislation include significant loses in VAT revenue, combatting fraudulent e-commerce practices involving tax evasion and the lack of proper classification mechanisms. Novel research in the field of machine learning, emerges as an encouraging factor for the research & development of automated systems that can help nations of the EU cope with the growing demand for digitalization when it comes to e-commerce and customs related procedures. Thereby, this thesis will focus on a set of concerns including the emerging regulation, international and European product classification standards and the adoption of machine learning for automation of product classification procedures through the creation of software solution.

This thesis is written in English and is 82 pages long, including 7 chapters, 29 figures and 7 tables.

Annotatsioon

Kaubanduskaupade nõuetekohane liigitamine nii rahvusvaheliselt kui ka ELis on erasektori ja avalikel organisatsioonidel üha nõudlikum suutlikkus. Rahvusvahelised organisatsioonid, nagu WTO ja WCO, on õnneks välja töötanud rahvusvaheliselt tunnustatud standardid, mis, kui neid kasutatakse asjakohaselt, võivad väga tõhusalt kasu saada kaubanduskaupade identifitseerimise protsessist. Kaubandusüksuste identifitseerimise/klassifitseerimise ülesanne vastavalt nimetatud juhtorganite poolt välja antud standarditele või nomenklatuuridele toimub tavaliselt käsitsi paljudes kohtades üle maailma. Euroopa ei ole erand ja ELi liikmesriigid, kes ei ole digitaalsete edusammudega hästi kursis, kuuluvad sellesse kategooriasse. Samal ajal loob Euroopa Komisjon uusi regulatiivseid nõudeid e-kaubanduse kaupade ja nendega kaasas olevate tollideklaratsioonide klassifitseerimiseks, mille kiiret vastuvõtmist EL eeldab ja pigem nõuab. Esilekerkivate õigusaktide põhjused hõlmavad käibemaksutulu märkimisväärset kaotamist, pettusega seotud e-kaubandustavade vastu võitlemist, mis hõlmavad maksudest kõrvalehoidumist ja nõuetekohaste klassifitseerimismehhanismide puudumist. Uudsed uuringud masinaõppe valdkonnas on julgustavaks teguriks automatiseeritud süsteemide teadus - ja arendustegevusele, mis võib aidata ELi riikidel toime tulla kasvava nõudlusega digitaliseerimise järele e-kaubanduse ja tolliga seotud menetluste puhul. Seega keskendub see väitekiri paljudele muredele, sealhulgas esilekerkivale määrusele, rahvusvahelistele ja Euroopa toodete klassifitseerimise standarditele ning masinõppe vastuvõtmisele toote klassifitseerimise protseduuride automatiseerimiseks tarkvaralahenduse loomise kaudu.

Lõputöö on kirjutatud Inglisee keeles ning sisaldab teksti 82 leheküljel, 7 peatükki, 29 joonist, 7 tabelit.

List of abbreviations and terms

EU	European Union
ML	Machine Learning
ID3	Iterative Dichotomiser 3
IS	Information System
SE	Software Engineering
VAT	Value-added Tax
HS	Harmonized System
HCDCS	Harmonized Commodity Description and Coding System
WTO	World Trade Organization
WCO	World Customs Organization
UI	User Interface
JSON	JavaScript Object Notation
CN	Combined Nomenclature
CCT	Common Customs Tariff
UCC	Union Customs Code
API	Application Programming Interface
MVC	Model View Controller
OOP	Object Oriented Programming
IG	Information Gain
REST	Representational State Transfer
SDLC	Software Development Life-Cycle
NCMM	Nomenclature Classification Mining Model
FEDS	Framework for Evaluation in Design Science Research
UX	User Experience

UNCTAD	United Nations Conference on Trade and Development
ITC	International Trade Centre
NLP	Natural Language Processing

Table of contents

List of figures

List of tables

1 Introduction

Over the past decade, the field of commerce conducted through websites and platforms or otherwise more widely known today as e-commerce, has faced an exponential growth in capital and has righteously become the new most preferred standing ground for means of conducting commerce worldwide. E-commerce tech giants like Amazon and Alibaba have managed to captivate most of the global customer audiences while smaller up and coming e-commerce companies and start-ups have also seemingly managed integrating into the e-commerce environment and continue to do so in terms of quickly targeting a given audience with the use of abundant information system tools and software that can support an accelerated utilization approach. Seamless to say, "E-commerce plays an important role in today's business environment, and that role will continue to grow each year" (Reinsch, 2005). Today's world of e-commerce is quite far ahead of what its landscape used to be ten or twenty years back in time per say, while the information communication capacities that accompany e-commerce professionals allow them to leverage their audience reach and handle immense loads of transaction throughput if all necessary tools for that are justifiably targeted and prioritised. This is to say, that "the e-commerce of today typically requires advanced ICT infrastructures to work" (Falk & Hogsten, 2015).

Just as in the more "archaic" times of e-commerce back when the world faced the dotcom bubble, not much has changed today and as Holsapple and Singh (2000) have stated, there are two main approaches when it comes to developing an electronic presence of a business with those referring to Informational and Transactional Approach. In the past, consumers would rather accommodate more compliably with the informational aspect of conducting e-commerce with emphasis on product research and communication capabilities rather than initiating in unphased straightforward transactions and product or service purchase through the online presence of a business. Non-surprisingly, consumer mentality today has shifted to the dramatic extent where consumer's non hesitantly do initiate in online purchases with an increased degree of comfort, trust, and reliability. Due to these ever-increasing transaction requests and demands, e-commerce systems are more than often required to possess the functional capabilities of sharp product categorization, identification, and attribution.

Despite the endless possibilities and ease of incorporating e-commerce and making it an extension of a business with reliance on all existing automating tools and software solutions, the field of e-commerce still exhibits flaws and inconsistencies that can manifest on both ends of the spectrum, potentially affecting buyers, sellers and even the hosting solution, in that case the transaction facilitating platform. E-commerce related fraud has been under the scope of the European Union for quite enough time now and has motivated the commission towards the creation and application of new regulatory means and legislations that can stabilize the situation and if not eliminate, then at least significantly decrease the number of monetary resources being lost annually as a result of e-commerce fraud. It is emphasised that, “the amount of current losses from customs fraud, even by means of an estimate, is unknown, as the customs gap (i.e. the difference between the duties collected and duties that are legally due) – contrary to the VAT gap – is currently not measured.” (European Parliament Policy Department D for Budgetary Affairs, 2019). Despite figures not being certainly depicted, some non-official sources have stated that tax revenue losses from e-commerce, under the form of VAT approximate between five to seven billion euros annually. These numbers are of course arguable, but they do awake concerns for the magnitude and significance of financial losses the EU has been facing as a result of unregulated and non-standardized e-commerce procedures.

It becomes evidently observable that fraudulent e-commerce tactics are in fact, systematically carried out with the intent of avoiding appropriate taxation on goods ought to be imported in the union. According to EU research, the most prominent tactics for customs related fraud are “mis-declaration of tariff classification; mis-declaration of value (undervaluation); mis-declaration of origin (preferential or non-preferential)” along with customs related VAT frauds such as “customs procedure 42 abuse; undervaluation and non-declaration, and export VAT fraud” (European Parliament Policy Department D for Budgetary Affairs, 2019). Needless to say, the most well-known tech giants of today which happen to facilitate most of the world’s e-commerce do not necessarily care for the micro-consequences emerging as a result of fraudulent procedures that can be incorporated in e-commerce. Consequently, there emerges an obvious understanding for the need of readjusting and re-modifying the currently existing tools and most importantly, information technology frameworks the EU possess for fixating, monitoring, and evaluating e-commerce transactions, consignments, and specifications of goods.

1.1 Research Motivation

As the world witness's exponential advancement leaps in the fields of digitalization and cross-border commerce, the European Commission, as the legislative and partially executive body that it is, has implemented a regulation of crucial importance in the beginning of 2021. The intended regulation aims to invoke member states towards amending and improving practical approaches of how the union and its member states handle e-commerce and customs related operations. Some of the addressed operations such as consignment commodity classification along with duty and tax calculation have been justified to be vital factors in the EU's newly proposed strategy to battle e-commerce related fraud. This exact matter, in congruence with the EU's aspirations to prioritize AI accompanied by the ever-growing capabilities of ML are the main drivers behind the expressed interest in conducting this research.

Furthermore, the author's experience in the fields of web development and software engineering, practical involvement in an e-commerce associated start-up project and the desire to acquaint oneself with the personal grasp and expertise in implementing ML solutions make up for a well-rounded thesis disposition.

The main audiences that benefit from this research are IS developers and engineers, who may be interested in exploring the possibilities that ML has to offer for tackling cross-border commerce related issues within the boundaries of the EU, tied to the CN and HS for commodity description and classification. Other parties that may benefit as well, include member state customs authorities and professionals working in institutions, agencies, or private firms looking for ML solutions for commodity classification problems.

1.2 Problem Formulation

The author's involvement in the Archimedes (https://archimedes.ee/en/archimedes-foundation/) research and development project 'e-Commerce EU VAT and Duty Declaration Digitalization 2021' serves as the foundational foreground for framing the research problem that contemplates the involved research questions. During the period of involvement, efforts were made towards the utilization of underlying tax classification capabilities with the use of HS codes for an e-commerce marketplace.

Furthermore, the involvement in this project bestows an affluent inspiration for the shaping of the dispositions made by this thesis, which associates with the project's aims to utilize automated issuance of customs declarations for commerce consignments bound to importation for Europe. This is a project focusing on the use case of tailor-made system developments for the acceleration of logistic operations and VAT compliance compatible with the new regulatory propositions of the already mentioned VAT directive for the case of Estonia. During the Archimedes Project as stressed by Liutkevičius, Pappel, Butt, and Pappel (2020), five group interviews were conducted with the Estonian customs authority as a mean of gathering qualitative data needed for justifying the need for automatization of customs declarations.

As a result, Liutkevičius et al. (2020), go on to conclude that the results of the research imply for the coping with numerous data requirements for declaring goods, which has been deemed achievable through the incremental implementation of novel logistics and customs-oriented systems. Additionally, the need for establishing harmonization and data uniformity throughout the EU for systems similar to the aforementioned ones, further promotes European classification standards such as the Combined Nomenclature.

1.3 Research Questions

The purpose of this section is to establish an elaborative overview of the methodology that is to assist in finding the most plausible and appropriate approaches for tackling the posed research questions.

Europe's VAT regulation on low-value goods, the threshold abolishment for custom duties and the obligation imposed by the EU to establish information systems for declaring, processing, and disseminating customs declarations by customs authorities of member states are already existing factors contributing to the reality to be abided by. The process of creating customs declarations involves an activity referred to as Harmonized System Code assignment for individual entities/ goods attributed to respective commodity groups. The process of allocating and assigning HS codes to e-commerce goods shipped to Europe from both EU and non-EU countries have stemmed various concerns due to the fact that the EU has been losing considerable amounts of revenue under the form of VAT taxes and Customs Duties. These loses occur annually as a result of mischievous and fraudulent e-commerce practices along with the incapability to correctly classify commerce commodities and goods with the use of proper software tools. Falsifying declared values of shipments, routing shipments through third countries as a means of disguise, falsely describing merchandise and splitting shipments into multiple smaller shipments all contribute negatively to the effectivity and legitimacy of properly assigning Harmonized System Codes to shipment associated goods.

Machine learning algorithms for identifying Harmonized System codes have already become an important topic with respect to the HS and have shown potential for tackling most of the problems that custom authorities face when processing e-commerce shipments and consignments. Despite the potential of machine learning in that regard, ML algorithms have only been implemented for HS code recognition with relatively low success rate and with negligence for VAT tax and duty calculation functionalities. Therefore, this thesis focuses on the development of a web interface (client-side) with a back end (server-side) which is to utilize machine learning algorithms for relatively accurate HS code identification along with VAT calculation based on member-state and Customs Duty calculation based on HS commodity groups and their respective tariff rates. The goal of this research is to determine how web-applications that make use of machine learning algorithms can aid EU customs authorities and independently operating retailers with marketplace presences in tackling tax and duty loses resulting from lacking classification utilities in information systems. This is to be accomplished by means of software engineering, data mining and machine learning features utilized in a homogenous information system environment with the use web technologies. The efficiency, effectiveness, and ease of utilizing all underlying ML algorithms are to be cohesively analysed and evaluated in accordance with nomenclature standards such as the Harmonized System and Combined Nomenclature.

This thesis aims to answer three main research questions, two of which can be considered as meta questions. To answer those, it is first expected to answer all underlying sub-questions that may comprise any of the aforementioned.

RQ1: How can the specialized application of renowned commodity nomenclatures & ML web technologies in information systems contribute towards more competent classification practices on trade goods throughout the EU?

RQ2: Can the findings/ answers to RQ1 support the proposal of a web technology deployment model for the adoption of commodity classification mining in web applications?

RQ3: How to evaluate the utility of the HSCODESYS?

Table 1. Research Sub-Questions Table

RQx.x	**Sub-Question**	**Hypothesis**
RQ1.1	Can ML algorithms in web applications assure nomenclature code identification and tax assumption of goods subject to cross-border trade with satisfying accuracy?	ML algorithms can very accurately (70%) classify trade commodities based on pre-programmed constraint logic.
RQ1.2	What is a feasible and plausible machine learning approach for mining commodity classification nomenclatures in web server architectures?	The HS and CN nomenclatures can be mined with the use of scripted Decision Tree Algorithms.
RQ3.1	What is the utility representation of the HSCODESYS?	Utility Tree Figure 22
RQ3.2	What is the most feasible DSR evaluation method suitable for measuring the utility of the HSCODESYS?	Hybrid in-situ DSR Simulation methodology

1.4 Research Methodology

The purpose that this chapter is to provide an overview of the methodology that will best serve the answering of the posed research questions by justifiably clarifying the reasons for using any specifically proposed research methods.

In order to establish a clear understanding of the value that the Harmonized System Nomenclature serves in e-commerce and how it can contribute towards the efforts made in tackling cross-border associated fraud within the boundaries of the EU, three main components will be emphasised for the basis of this research. The first component refers to the conceptual understanding of new regulations and directives proposed by the European Union for battling fraud and tax evasion in e-commerce. Subsequently, the presented overview of any new legislation is to serve as basis for highlighting the importance of accurate product classification for logistic operators and e-commerce retailers with the use of ML, HSCDCS and CN. Lastly, based on the comprehensive elaboration that is to be made, it is to be proceeded with the proposition of a web application/ interface that is envisioned to utilize algorithms for ML outcomes by using the HS and CN datasets for classifying/ allocating product descriptions and their applied taxes and duties.

The first component of this research will provide the needed foundation for identifying all ongoing changes and readjustments proposed by the EU for battling e-commerce tax revenue losses. This is done by analysing and evaluating the importance of recent legislative documents, directives and reports rolled out by the European Commission. The first part of this paper is intended to contribute towards the understanding of all regulatory reforms taking place so that tools, frameworks, and software components needed for the second part of this paper are justifiably identified.

The second component of this research focuses on comprehensively describing the structure, importance and modern-day use cases of the Harmonized Commodity Description and Coding System Nomenclature and its perspectives in Machine Learning, all in the spectrum of e-commerce and commerce in general. This is to be achieved by collecting secondary data about application use cases of ML for data model training and mining capabilities over the HS dataset. The secondary data is to be mainly collected from other studies and research-based papers highlighting already implemented and tested application approaches and results revolving around the use of ML algorithms for commodity code classification.

The third and last component of this research will for the most part, revolve around the development of a web application that makes use of machine learning functionality for scanning/mining/allocating HS code records from the Harmonized System dataset, based on which tax and duty calculation can be further functionally estimated, again with the use of ML. This task is to be accomplished by abiding to design science principles and by following design cycles that will ensure the production of the two expected artefacts, one of them being the application software, and the second being the architectural framework behind it. The web application is then expected to undergo series of various success rate statistical tests that would support the proposed hypothesis of the first sub-question while also serve as a comparable example to the existing secondary data. The analysis of the validation and test results of the secondary ML use cases for HS code classification, with the one this paper aims to propose, will be the foundational approach for validating the produced artefact/s while also providing context for answering sub-questions 2 and 3.

1.5 Design Science Approach

The purpose of this section is to provide an elaborative overview of the methodology that is to be used for accomplishing the underlying tasks associated with some of the research questions along with the proposition of a design science research framework that is to guide the development of design artefacts and processes.

Information System development is a rather complex process which non-surprisingly is tied to science related factors that can influence its life cycle and outcomes. Hevner, March, Park, and Ram (2004) argue that the sole purpose of implementing any kind of information system within an organization is to achieve greater improvements in effectiveness and efficiency of the organization. Respectively, the process of creating information systems is one that involves engineering envisioning through a wide and varying range of lenses where design is one of the prominent predetermining factors of development. Subsequently, in software engineering, as a contributing factor to IS development, “practical problems are always problems in the design, construction or maintenance of software systems—the software engineering domain.” Wieringa (2014).

It seems as if design is the steppingstone in building or creating any product or service in the domains of IS and SE. This is to also be the case with the expected outcomes that this research intends on producing, which is why the most suitable and promising research methodology for achieving the estimated outcomes is the design-science methodology/ paradigm. It is stressed by Simon (1996), that the roots of the design science methodology lie in engineering and sciences associated with the artificial where the paradigm itself is focused on solving problems.

It is further elaborated by Denning (1997) and Tsichritzis (1998) that the methodology aims to establish a blueprint of practical approaches, tools and fundamentals that would allow the creation, use and analysis of information systems to be carried out. The ideas that this methodology intends on materializing can be perceived as artefacts, supported by technological tangibility, which as stated by Sjoberg, Dyba, and Jorgensen (2007), can manifest under the forms of "process models, methods, techniques, tools and languages".

Furthermore, design science, as opposed to behavioural science, and "as the other side of the IS research cycle, creates and evaluates IT artefacts intended to solve identified organizational problems" (Hevner et al., 2004). Hevner et al. (2004) further stress, that design within the boundaries of the IS research paradigm can be perceived both as a process consisting of a set of activities and a product, or otherwise referred to as the design artefact, which emerge from applying a sequential set of expert activities. Respectively, design artefacts are always subject to continuous evaluation throughout the course of the research, which serves as feedback for better comprehension of the existing problem as well as a motive for the continuous improvement of the artefact's quality and the design process itself.

During the creative process that design science makes use of, evolvement of the design artefact and/or the design process are very much expected. Additionally, March and Smith (1995), have identified two design processes and four design artefacts which could be produced during any design-science research in the field of information systems. The two processes refer to build and evaluate while the artefacts refer to constructs, models, methods, and instantiations. It becomes clear that, in order to utilize the design science methodology fruitfully, an artefact is first expected to be built, after which it is evaluated and validated. For the purpose of producing the particular to this research design artefacts, a framework for information system research is to be abided by.

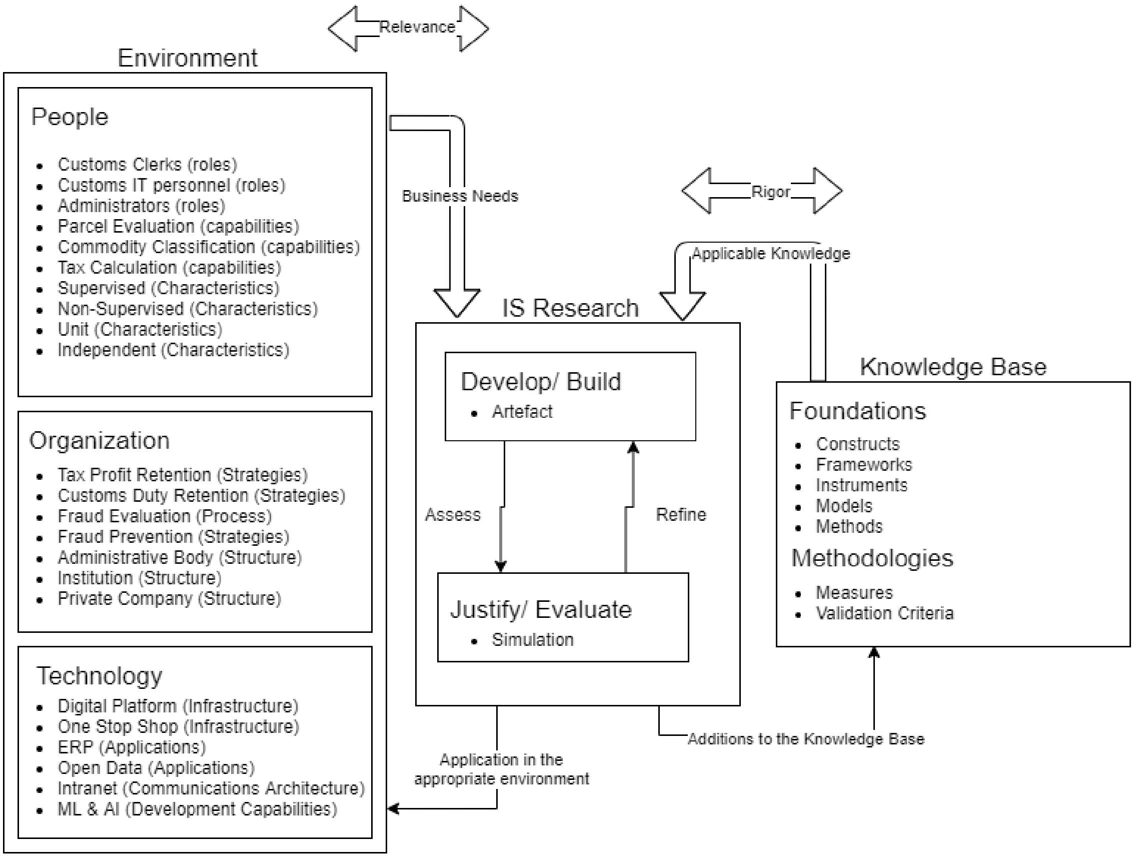

Relevance
Environment
People
Customs Clerks (roles)
Customs IT personnel (roles)
Administrators (roles)
Parcel Evaluation (capabilities)
Commodity Classification (capabilities)
Tax Calculation (capabilities)
Supervised (Characteristics)
Non-Supervised (Characteristics)
Unit (Characteristics)
Independent (Characteristics)
Organization
Tax Profit Retention (Strategies)
Customs Duty Retention (Strategies)
Fraud Evaluation (Process)
Fraud Prevention (Strategies)
Administrative Body (Structure)
Institution (Structure)
Private Company (Structure)
Technology
Digital Platform (Infrastructure)
One Stop Shop (Infrastructure)
ERP (Applications)
Open Data (Applications)
Intranet (Communications Architecture)
ML & AI (Development Capabilities)
Business Needs
Rigor
Applicable Knowledge
IS Research
Develop/ Build
Artefact
Assess
Refine
Justify/ Evaluate
Simulation
Knowledge Base
Foundations
Constructs
Frameworks
Instruments
Models
Methods
Methodologies
Measures
Validation Criteria
Additions to the Knowledge Base
Application in the appropriate environment

Figure 1 - Information System Research Framework for producing the intended by this thesis artefact. Example reproduced in resemblance with the one presented by Hevner et al., (2004).

2 Background & Theoretical Framework

2.1 State of the Art Regulatory Landscape Review

The existing background information on how the European Union and its member states in particular can battle e-commerce fraud more persistently revolve around matters of increased digitalization and re-engineering of existing information system tools and software put in place to accommodate and support the outlets facilitating digital commerce transactions. By using tailor made bridging portals and online platforms for declaring, structuring, and disseminating information on VAT and customs declarations of goods, the burdens of non-monitored and dismissed declarations will be slightly decreased, effecting positively the overall flow of declarations information crucial for identifying potential fraud.

According to Chapter 2, Section1, Article 6 of the new EU Parliament Regulation No. 952/2013, addressing the proposed obligations and rights of persons involved in the processes of information provisioning related to customs declarations, it is stated that “all exchanges of information, such as declarations, applications or decisions, between customs authorities and between economic operators and customs authorities, and the storage of such information, as required under the customs legislation, shall be made under electronic data-processing techniques.” (REGULATION No 952/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, 2013).

This proposition constitutes, that the EU and more precisely, it’s member states should be well equipped with all necessary electronic data-processing tools, where each and every member state is responsible for establishing and developing or compliantly modifying any existing tools in accordance with the regulation. Furthermore, this development of new electronic means of customs information dissemination also requires of the member states to ensure proper reporting obligations which will contribute towards the EU’s overall efforts of monitoring and evaluation of micro as well as macro progress of the electronic systems under development.

Continuously, as referred to Article 278a of Regulation (EU) No 952/2013, it is stressed that “In accordance with paragraph 4 of this Article, Member States are to provide the Commission, twice per year, with an updated table on their own progress in developing and deploying the electronic systems.” (Council of the European Union, 2019). Such collective efforts increase the likelihood of efficiently reaching milestones in that regard as well as the consistency of parallel member state endeavours towards the implementation of the necessary electronic systems. Research on the topic of machine learning and product classification incorporation into e-commerce shows that “categorizing products into a hierarchical taxonomy has become a central part of the organizational efforts of eCommerce companies” (Krishnan & Amarthaluri, 2019).

Machine learning algorithms can result in accurate calculations of price and tax on products based on metadata attributes of both goods and customers. From the research conducted by Li, Kok, and Tan (2018) on two major real-world data sets, it is noted that “state-of-the-art machine translation (MT) models surpass previous classification approaches”. In addition, as the European Commission Help Desk (2020) points out, “all products are classified under a tariff code that carries information on duty rates and other levies on imports and exports, any applicable protective measures (e.g. anti-dumping), external trade statistics, import and export formalities and other non-tariff requirements". Much of the on-going tax fraud and lost VAT and Customs Duty revenue happens because of missing product classification software and functionality of existing e-commerce marketplaces.

This is just a fraction of the standing ground supporting the need for inputting metadata of goods, such as tariff codes into machine learning algorithms that can calculate taxes and duty. It is thought that although these electronic systems will be designed and implemented for member states customs authorities exclusively, they can also be compatibly developed for retail online marketplaces as well. This is the case with Eurora Solutions, a private start-up company from Estonia which has shown great effort into developing what is now known as, perhaps, one of Europe's first cross-functional business to government and vice-versa customs declarations and product classification software solution. Among the vital for customs authorities' services that the Eurora Intelligent Cross Border Compliance Platform provides, are "HS Code Allocation, Duty & Tax Calculation, Electronic Declarations, Label & Tracking" and more ("Solution", 2020).

Another pioneer in this field, hailing from Estonia as well is the Cybernetica ICT company that specializes in the field of secure systems and technologies. The company is known for its mission-critical projects and developments, one of which happens to be the Customs System, also known as "e-Customs" product suite which makes use of the in-house developed Customs Engine for covering major business processes tied to customs documents such as "customs declarations, manifests, TIR carnets, transit declarations, anticipated export records, exit summary declarations and arrival notifications" ("Business Domains", 2020). Furthermore, it has been firmly stated by the company that it's utilized Customs Engine contains the "full implementation of the European Union's common customs system: New Computerized Transit System (NTCS), Export Control System (ECS), and Import Control System (ICS)" ("Business Domains", 2020).

In a sense, Estonian based ICT companies seem to have swiftly proceeded with implementing what the European Union and its commission have proposed in terms of the new VAT & Customs regulation for member states. It is such examples, that can pose and be perceived as the ultimate blueprint for implementing these newly proposed computerized product classification and automated declaration issuing systems. EU member states have much to learn from the achievements and developments of the Estonian ICT sector. Despite the remarkable advancements in cross-border compliance services and functionalities, there is room for improvements of the product classification capabilities the above-mentioned examples possess. This statement is more precisely aimed at addressing how the currently utilized Harmonized System Code Allocation algorithms work, and how they can be comprehensively integrated with frameworks like the Integrated Tariff also known as TARIC, for acquiring information on trade policies and tariff measures for products in the EU but also other international tariff frameworks for achieving global compliance, compatibility, and competitiveness.

2.2 Comparative Analysis of related work

This section aims to support the comparative analysis of existing research efforts and elaborate upon documented applications of machine learning for the purpose of automated commodity classification by customs authorities. Referential overview of research conducted in this area by institutional or private organizations is conducive for the validation of propositions made by this thesis.

The need of adopting machine learning solutions for the purpose of automating the classification of tradeable products in commerce has exponentially risen due to the boom of e-commerce along with the internationally favorable logistics circumstances. The stage of introducing novel solution developments for tackling misclassification of goods and revolving fraudulent practices require the adoption of ML. One of the reasons for this as Spichakova and Haav (2020) stress is the "terminological and the semantic gap between product descriptions in the HS nomenclature and goods descriptions in trade (i.e. commercial terms)".

This observation has led to the awareness and demand for ML solutions that make use of NLP and sentence encoding along with convolutional image classifying approaches. Relative works which have been proposed, use "text-image adaptive convolutional neural network to effectively utilize website information and facilitate the customs classification process" (Li & Li, 2019). Another study conducted by Altaheri and Shaalan (2020), implies the incorporation of ML for HS code prediction by making use of user input describing commodities utilized by a learning model.

In the example of using user input to predict certain HS codes, Altaheri and Shaalan (2020) go on to further elaborate the concept of tokenization used in NLP to mathematically determine the weighted importance of each word comprising a commodity description. Ding, Fan, and Chen (2015) also mention the potentiality of text categorization based on sets of keywords comprising a vector space model that can facilitate incremental learning using graph-based background nets.

In his works for developing an automated tool for HS codes of e-commerce products, Van der Hejde (2019), acknowledges the use case for artificial intelligence technologies such as NLP and information retrieval. Van der Hejde (2019) goes on to further emphasize that a taxonomical categorization process involves multiple algorithm derived responses consecutively pertained from attribute deduced decision making. As it can be noted by the referred examples, there is a tendency of deploying NLP methods for classification of trade goods in accordance with the Harmonized Nomenclature.

On the other hand, machine learning utilization for automation of customs processes across-the board is an increasingly advancing practice because of its recently noticed potential along with increasing demand throughout numerous customs related procedures. One such procedure refers to the customs activity associated with the creation, processing, and dissemination of customs declarations. As elaborated by Liutkevičius et al. (2020), the latest EU VAT directive constitutes the abolishment of de minimis thresholds adherent to each EU member state, which as a result assumes the exponential increase in customs declarations ought to be handled on a national level.

This sought-after demand that is expected to occur can be congruently justified by the observation made in a study conducted by Basalisco, Wahl, and Okholm (2016), which underlines that an approximate of 70% e-commerce bound consignments pass in between public postal channels, resulting in unpaid VAT on 65% of consignments, leading to a loss of 1 billion EUR in VAT revenue. Conclusively, the emergence of fraudulent activities such as those described in the introductory chapter of this thesis presently exist thus posing the need of developing countering IS mechanisms, which assuredly may make use of ML technologies.

When it comes to witnessing quick adoption and compliance with emerging legislation such as the EU VAT directive, examples like Sweden, as emphasized by Rozbroj (2020), truly depict how early consideration for transition embracement can bring together Fintech companies in the sphere of logistics, state owned postal-companies and marketplaces for the purpose of automation and regulation compliance. The Swedish efforts and results of cross-sector cooperation, as Rozbroj (2020) further argues, are seen as a positive step towards regulation compliance and a notable achievement for the reduction of the national VAT gap.

2.3 Validation of proposed problem/ solution

Similarly, to the referred examples in section 2.2 of this chapter, the involving solution of the Archimedes project made use of a machine learning application that uses marketplace specified HS codes of e-commerce products for automated calculation of VAT and Duty taxes upon checkout. The fundamental aim of the Archimedes backed R&D project was to bridge gaps between different entities (logistics providers, customs authorities, marketplaces) involved in the process of comprising and transmitting customs declarations.

Unlike the already mentioned examples of section 2.2 which capitalize of ML methods such as neural networks and NLP, the solution to the problem for addressing the research questions at hand will prioritize the finding and proposition of a method capacious for taxonomy classification based on a two-dimensional input approach. Van der Hejde (2019), goes on to emphasize that the “process of deriving high-quality information from texts” also known as data mining, is already used by some commerce companies. Spichakova and Haav (2020) additionally argue that the taxonomical kernel of HS codes can be used as an integral apprehension to plain text.

Complementary to the expressed remarks on the aspect of taxonomy employment, product classification, as argued by Gupta, Karnick, Bansal, and Jhala (2016), can be perceived as the result of taxonomical path prediction for products assigned to pre-declared taxonomy-based hierarchies associated with textual data. The inherent taxonomical nature of the HS and CN systems present an untapped potential for the utilization of a data mining solution that can exploit the analogous comprising parts of each code construct for the mining of emanate nomenclature data imperative to customs declarations.

The development of such a solution is to be inspired by techniques such as the "Hierarchical Product Classification (HPC) framework for the purpose of classifying products using hierarchical product taxonomy" (Van der Hejde, 2019). Another ingenuity relative to the proposition at hand is the "Cross-Industry Process for Data Mining methodology which comprises six phases, namely business understanding, data understanding, data preparation, building prediction model, performance evaluation and model deployment." (Altaheri & Shaalan, 2020).

To the best of knowledge preceding the argumentations of this thesis, no research has been published that covers the classification of 8-digit combined nomenclature codes based on the 2-digit taxonomical levels of ordinary 6-digit harmonized system codes. Therefore, this will stand as the argumentation supporting the preference of using the HS as the uniform standard it is proclaimed to be for developing a web tool that utilizes supervised machine learning for mining of CN codes.

2.4 Union Customs Code Legislation

The union customs code legislation, is yet another piece of EU regulation, proposed on 9th of October 2013, which lays down all the specifications, requirements, and guidelines for the establishment, of what is envisioned to be Europe's modern framework for customs and trade. Essentially, the UCC is to serve as a key element in Europe's efforts to modernize its customs IT infrastructure and operations since "It provides a comprehensive framework for customs rules and procedures in the EU customs territory adapted to modern trade realities and modern communication tools" (European Commission, n.d.). Continuously, as seen throughout the REGULATION (EU) No 952/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 9 October Laying down the Union Customs Code (2013), some of these operations include the proper measurement and application of import and export duties, CCT and tariff classification of goods, production of customs declarations and many more. The whole set of proposals made in the UCC regulation are made with the intent of leading Europe towards a more paperless, automated and to some extent uniform customs union.

As it is further stressed by the Taxation and Customs Union of the European Commission (n.d.), in order to achieve the above-mentioned objective, “the Commission and the Member States have to upgrade some existing electronic systems and introduce some new ones for the completion of all customs formalities”. Other objectives of the UCC include the establishment of an overall simpler, more serviceable, and more efficient environment that is fully electronic while also enhancing the competitiveness of European business, safeguarding the financial and economic interests of the EU, and protecting the flow of goods moving in and out of the union.

2.5 Supervised Learning and Classification with Decision Trees

While examining potential approaches that can be implemented for Harmonized System code classification functionalities in web applications, it becomes obvious that two very prominent approaches present themselves as very promising for that matter. These refer to the uses of neural networks and data structure-based algorithms. Despite the fact that application cases of neural networks have been mostly applied in server-side scripting and programming environments such as Python and Golang, JavaScript frameworks for direct neural network training in client-side environments have also gained considerable traction.

Neural networks have shown great success in application scenarios such as image classification, object detection, speech recognition, face detection and many more recognition aimed capabilities based on pre-trained models. Despite all notable achievements of neural networks, the classification capabilities that they exhibit are still less avail for non-encoded stand-alone textual and numeric data when compared to decision tree learning algorithms. That is why the focus of this research will revolve around identifying and reaffirming an implementation scenario of decision tree classification on the HS system and the newly proposed by the EU CN with the use of a live example.

When mentioning classification, it should be emphasized that, essentially it is the “processing of finding a set of models (or functions) which describe and distinguish data classes or objects” (Vijendra, Parashar & Vasudeva, 2011). Furthermore, any model that is to be derived from classification procedures, is the product of analysis carried out on various sets of training data. Decision trees can be truly diverse in the inferences they produce, as Vijendra et al. (2011) further emphasize, a “derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks”. Classification, decision, and regression tree algorithms have presented themselves as integral principles of data mining, while persistently presenting equally important properties that can be tremendously productive in machine learning.

Decision trees, or more specifically speaking, classification and regression trees are a family of machine learning algorithms that are most notably used in software engineering for building classification models that resemble the form of a tree structure. Before dissecting and analysing what CART’s are and how they work, it is important to note that the most concise definition of classification refers to “the task of assigning objects to one of several predefined categories“ (Howbert, 2012). Throughout time, classification algorithms have shown great potential for tackling data classification problems in a varying range of technically associated domains that deal with data, especially big data, and it has been pointed out that “one of the most intuitive tools for data classification is the decision tree” (Aggarwal, 2014). As further emphasised by Tan, Steinbach, and Kumar (2015), a more detailed definition of classification refers to “the task of learning a target function f that maps each attribute set x to one of the predefined class labels y” where the target function serves as the classification model.

To further support the pointed-out definition, it would be of matter to stress that “The input data for a classification task is a collection of records. Each record, also known as an instance or example, is characterized by a tuple (x, y) where x is the attribute set and y is a special attribute, designated as the class label (also known as category or target attribute)” (Tan et al., 2015).

The definition on how decision trees work on the other hand, as the most famous member of the classification algorithm family, is very identical to that of classification itself. Aggarwal (2014) argues that decision trees essentially partition data in a hierarchical way, by binding leaf nodes to various class instances with the use of conditional split criterions on attributes with the aim to hierarchically distinct splits as univariate (former) and multivariate (latter). In more simpler terms, in order to initiate any classification task, there must be a cohesive dataset at hand, which allows for the specification of features and classes in order to comprise a model, which can then be subjugated to classification algorithms. Some of the most well-known classification techniques refer to "Decision Trees, Rule-Based methods, Logistic Regression, Discriminant Analysis, k-Nearest Neighbour, Naïve Bayes, Neural Networks, Support vector machines, Bayesian belief networks", etc. (Howbert, 2012).

This research will mainly focus on the implications and utility of decision tree techniques and algorithms for classification purposes. Decision trees have a hierarchical structure that makes use of root nodes, internal nodes, and leaf nodes where "each internal node corresponds to a partitioning decision, and each leaf node is mapped to a class label prediction" (Tan et al., 2015). Based on this interpretation, as further emphasised by Tan et al. (2015), the process of classifying data items involves iterative traversal cycles that begin at the root node of the specified tree which make use of internal node programming based on splitting rules that govern the process of reaching leaf nodes. This iterative traversal cycle is usually carried out by classification algorithms that carry out the task of inducing nodes and reaching the classification outcome that is based on the pre-defined decision rules. Such algorithms refer to "Hunt's algorithm (one of the earliest), CART, ID3, C4.5, SLIQ, SPRINT" (Howbert, 2012).

The above-mentioned classification process and associated algorithms refer to the concept of decision tree induction. The way that decision tree induction works is that it makes use of the three types of nodes that a decision tree consists off, referring to "a root node that has no incoming edges and zero or more outgoing edges; internal nodes, each of which has exactly one incoming edge and two or more outgoing edges; leaf or terminal nodes, each of which has exactly one incoming edge and no outgoing edges" (Tan et al., 2015). Additionally, as further emphasised by Tan et al. (2015), in ordinary decision tree structures "each leaf node is assigned a class label" where "non-terminal nodes which include the root and other internal nodes, contain attribute test conditions to separate records that have different characteristics". The hierarchical structure of the prospective nomenclatures (HS and CN) that this thesis aims to examine along with the proven and documented utility cases of such algorithms serve as the standing ground for considering decision trees as the governing approach, with a due diligent emphasis on the ID3 algorithm.

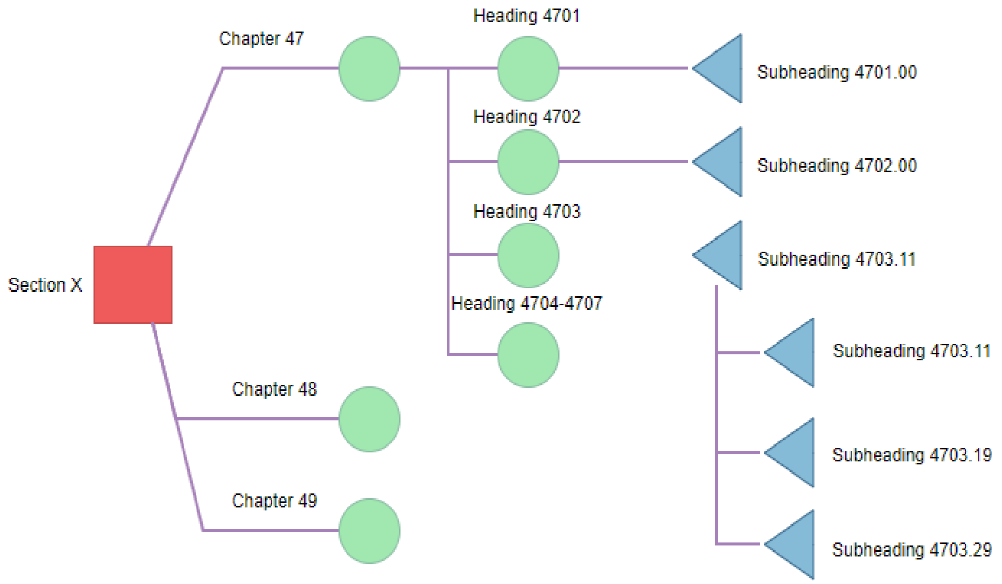


Figure 2 - The Harmonized System Code Construct perceived in a Tree-Like Semantic Hierarchy (Example given for Section 10 (X))

2.6 Harmonized System Nomenclature

It would be a negligent approach to address the need for using tariff metadata and machine learning algorithms for classifying tariff estimates on goods traded across borders without mentioning the Harmonized System Nomenclature and its implications on cross border trade. According to Ding et al. (2015), "The Harmonized System of tariff nomenclature created by the Brussels-based World Customs Organization is widely applied to standardize traded products with Code, Description, Unit of Quantity, and Duty for Classification, to cope with the rapidly increasing international merchandise trade".

Additionally, as the renowned World Customs Organization elaborates, the HS nomenclature is the resulting outcome of the Harmonized System Convention, the objectives of which is stated to be "by harmonizing the description, classification and coding of goods in international trade, to facilitate international trade and the collection, comparison and analysis of statistics" (WCO, n.d.). In essence the Harmonized System is a huge metadata set consisting of approximately six thousand product classification codes that are structured in a hierarchical order with respective groups of adhering products being nested within their respective Section, Chapter, Heading and Subheading. More specifically speaking, stand-alone HS codes are comprised of 6 digits which are internationally recognizable by all nation state adoptees of the nomenclature.

The first two digits of an ordinary HS code represent the Chapter to which the product falls in, followed by a Heading, represented with another two subsequent digits which adhere to a more detailed and specific description of the good. Headings are then followed by another two subsequently positioned digits which are referred to as Subheadings and provide further specific detail for the given good. HS codes do not end there, since “In addition to the internationally standardized 6-digit code, each country is able to further add more digits to extend the code to 8, 10 or 12-digit for its own tariff and statistics purpose.” (Ding et al., 2015).

The quintessentially sub categorical nature of the nomenclature itself, makes it a perfect prospect for applying machine learning operations and algorithms. Understanding how Harmonized System codes work and how valuable they can be for the development and training of algorithms for identifying commerce related goods is vital for logistic operators and customs authorities. This can only be reaffirmed by the observation made by Ding et al. (2015), which underlines that approximately 200 countries worldwide have adopted the nomenclature which adheres to 98% of international trade. National customs authorities and private e-commerce businesses have already turned towards utilization approaches of the nomenclature with some achieving success and others not, while on a general basis, “studies show that, about 30% of declaration submission uses wrong HS Code.” (Ding et al., 2015).

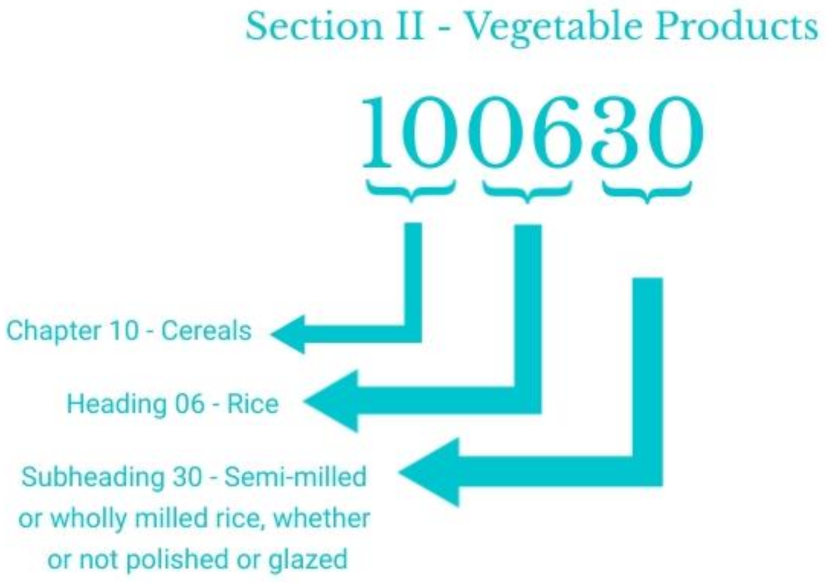

Section II - Vegetable Products
100630
Chapter 10 - Cereals
Heading 06 - Rice
Subheading 30 - Semi-milled or wholly milled rice, whether or not polished or glazed

Figure 3 - Example of HS code structuring based on Chapter, Heading & Subheading

Achieving highly accurate pre-trained HS classification networks remains a real challenge even today. That is why, for the sake of this research, any involving efforts aimed at developing a solution that makes use of the HS code system or EU's very own CN will only make use of the renowned 6-digit HS codes which are internationally standardized along with 8-digit CN codes which are EU standardized. This approach serves the purpose of establishing a solution that focuses solely on the utilization of standardized nomenclature codes for the sake of interoperability, ease of migration and scalability. After all, the above stated standardized formats of the nomenclatures to be used for this research can be easily accessed online while also being more easily interpretable and implementable by IT professionals in the field.

2.7 Combined Nomenclature

The Combined Nomenclature is a tool that has been proposed by the European Union for utility in various areas, some of which concern satisfying Common Customs Tariff requirements as well as use in internal and external trade statistics. A regulatory piece of the European Council that stresses the extended and specialized use of the CN was implemented on the 21st of September 2020 and is expected to be put in force starting the 1st of January 2021. One of the most major implications of this regulation, as pointed out by the Council of the European Union (2020), is to propose the CN, and establish it as the uniform goods nomenclature of the EU and its member states. Essentially, CN is expected to function as Europe's universal standard/ tool for classifying goods subject to external (international) and internal (EU) trade while also acting as the main supportive feature of the CCT.

In continuity, "The term 'tariff classification of goods' is defined in Article 57 of the Union Customs Code (UCC)" (European Commission, n.d.). As mentioned by Taxation and Customs Union of the European Commission (n.d.), the CN is nothing more than a further developed version of the HS intended for specific use by the union, which utilizes more deeply nested hierarchical subdivisions of commodities for more accurate product descriptions and duty rates classification. The classification capabilities that the Combined Nomenclature would allow for do not revolve only around matters of customs duty rate estimation but also "non-tariff measures" such as import/export licencing, restriction identification, origin certificate issuance, product liability, etc. (European Commission, n.d.).

Lastly, it is important to note that, according to the Taxation and Customs Union of the European Commission (n.d.), the CN is one of two elements of the EU classification system, with the other one referring to TARIC, which "provides information on all trade policy and tariff measures that apply to specific goods in the EU". An integrated tariff code makes use of CN codes (8 digits), while adding two additional digits that could be referred to as a TARIC (10 digits) subheading. TARIC is already integrated within the taxation and unions consultancy platform that serves as the EU's comprehensive and centralized information system for tariff and commodity related data. The integrated tariff will not be of great relevance for the aims of this research, since it is an enclosed remote system, that does not allow for external software integrations and does not provide an API of its own to the vast public. It is rather ways of utilizing the CN and the HS for ML capabilities in web apps/ IS's that will contribute the most for achieving the proposed objectives of this research.

Table 2 - CN Subheading structure example (European Commission, n.d.)

Chapter in the Harmonized System (HS)	2 digits	e.g., 'Chapter 18 - Cocoa and Cocoa Preparations'
HS Heading	4 digits	e.g., '1806 - Chocolate and other food preparations containing cocoa'
HS Subheading	6 digits	e.g., '1806 10 - Cocoa powder, containing added sugar or sweetening matter'
CN Subheading	8 digits	e.g., '1806 10 15 - Containing no sucrose or containing less than 5 % by weight of sucrose (including invert sugar expressed as sucrose) or isoglucose expressed as sucrose'

2.8 EU VAT Tariff Profiles

Table 3 - VAT Rates of European Union Member States (excluding GB as of 1st January 2021) ("2021 VAT Rates in Europe", 2021)

Country	Super-reduced VAT Rate (%)	Reduced VAT Rate (%)	Parking VAT Rate (%)	Standard VAT Rate (%)
Austria (AT)	–	10 / 13	13	20
Belgium (BE)	–	6 / 12	12	21
Bulgaria (BG)	–	9	–	20
Croatia (HR)	–	5 / 13	–	25
Cyprus (CY)	–	5 / 9	–	19
Czech Republic (CZ)	–	10 / 15	–	21
Denmark (DK)	–	–	–	25
Estonia (EE)	–	9	–	20
Finland (FI)	–	10 / 14	–	24
France (FR)	2.1	5.5 / 10	–	20
Germany (DE)	–	7	–	19
Greece (GR)	–	6 / 13	–	24
Hungary (HU)	–	5 / 18	–	27
Ireland (IE)*	4.8	9 / 13.5	13.5	21
Italy (IT)	4	5 / 10	–	22
Latvia (LV)	–	5 / 12	–	21
Lithuania (LT)	–	5 / 9	–	21
Luxembourg (LU)	3	8	14	17
Malta (MT)	–	5 / 7	–	18
Netherlands (NL)	–	9	–	21
Poland (PL)	–	5 / 8	–	23
Portugal (PT)	–	6 / 13	13	23
Romania (RO)	–	5 / 9	–	19
Slovakia (SK)	–	10	–	20
Slovenia (SI)	–	5 / 9.5	–	22
Spain (ES)	4	10	–	21
Sweden (SE)	–	6 / 12	–	25
United Kingdom (GB)	–	5	–	20

2.9 EU Customs Duty Tariff Profiles

Table 4 - The 2020 EU Tariff Profile Ranges for HS commodity groups (WTO, 2020)

Product Groups	AVG Duty Rate (in %)	
	Final Bound Duties	MFN Applied Duties
Animal Products	17.2	16.3
Dairy Products	42.2	37.5
Fruit, vegetables, plants	12.2	10.9
Coffee, tea	6.0	5.9
Cereals & preparations	18.0	13.9
Oilseeds, fats & oils	5.7	5.3
Sugars and confectionery	26.9	24.5
Beverages and tobacco	19.8	19.3
Cotton	0.0	0.0
Other agricultural products	4.6	3.1
Fish and fish products	11.4	11.6
Minerals & metals	1.9	2.0
Petroleum	3.1	2.5
Chemicals	4.5	4.5
Wood, paper, etc.	0.9	0.9
Textiles	6.6	6.5
Clothing	11.5	11.5
Leather, footwear, etc.	4.2	4.1
Non-electrical machinery	1.7	1.8
Electrical machinery	2.4	2.3
Transport equipment	4.1	4.7
Manufactures, n.e.s.	2.4	2.2

The table above resembles customs duty tariff profile ranges for all the commodity groups that make up the Harmonized System Nomenclature of the WTO/ WCO. The table consists of 22 product groups, which could be perceived as descriptive bindings of customs duty tariff rates assigned to specific ranges of HS Chapters. The example below depicts how such binding duty ranges can be specified.

Table 5 - Product group definition and corresponding HS Heading & Subheading ranges for Non-electrical and electrical machinery (WTO, 2020)

Product Group	MTN	Harmonized System Nomenclature 2017
Non-electrical machinery	7	7321-22, Ch. 84 (except 846721-29), 850860, 852842, 852852, 852862, 8608, 8709
Electrical machinery	8	846721-29, Ch. 85 (except 850860, 852842, 852852, 852862, 8519-8523 but including 852352)

3 Web Application (HSCODESYS) for classification of trade goods

This chapter is devoted to comprehensively introducing, describing, specifying, and defining all boundaries, components, functionalities, and underlying mechanisms of the artefact that are ought to be produced for the purposes of this research. Furthermore, the section will focus on defining proper requirements of the artefact in terms of functionality, design, and architecture along with proper reasoning and emphasis on why the particular technologies (algorithms, frameworks, web technologies) ought to be used have been chosen.

3.1 Artefact Description

The artefact that this research aims to produce in order to support the theories and propositions of this thesis is a web application that encompasses a wide range of classification-oriented machine learning functionalities along with a simple yet concise design and an easily integrable infrastructure that can be supported by any modern browser. The artefact is expected to make use of various commodity group datasets issued by governing bodies like the European Union and the WCO/ WTO with the intent of executing supervised machine learning algorithms in order to classify commodities under certain categories and assign certain attributes of importance for customs clearance like CN and HS codes.

3.2 The SDLC of HSCODESYS

The software development lifecycle that stands as most suitable for the development of the aforementioned artefact is the waterfall model. As mentioned by Van Casteren (2017), the waterfall software development sequence was first introduced by Dr. Winston Royce and was later on affirmed as a legitimate SDLC by Bell and Thayer in 1976. The waterfall model is considered to be one of the first SDLC models and finds its application even today. This is a traditional development model that is most suitable for projects with concretely predefined requirements bound to none or minimal future change with a life cycle that is rather hierarchical and sequential. The utilization of this model for the development of the HSCODESYS is appropriate due to the straight forwardness that it provides along with the fact that the artefact's components are concretely specified.

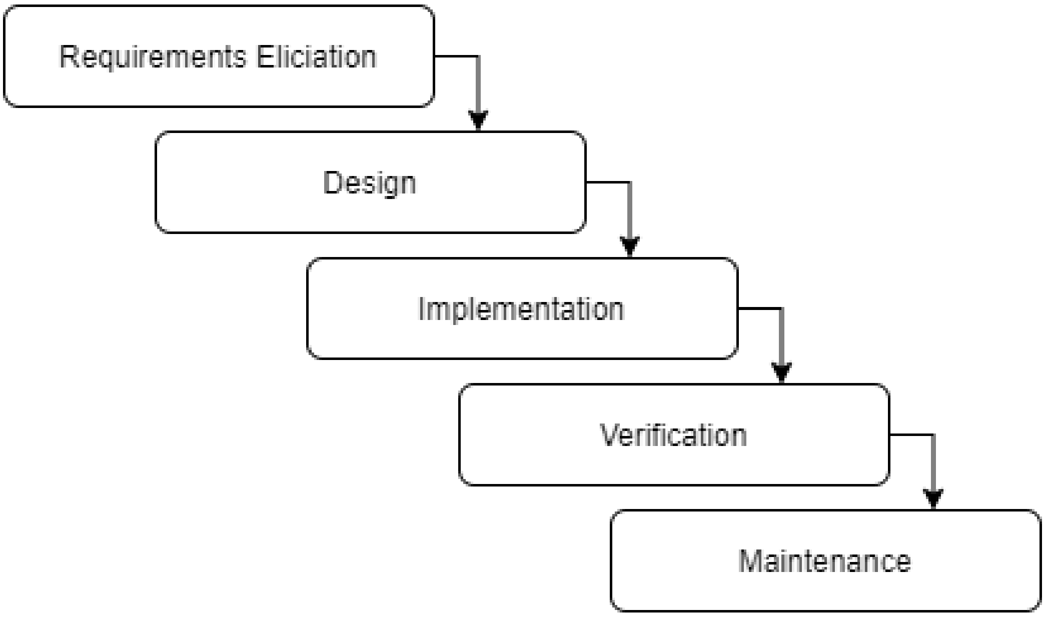


Figure 4 - The Waterfall SDLC

3.2.1 Design Requirements & Constraints

The artefact's design is envisioned as one that is clear, concise, and unsaturated in terms of graphical media features and extensive styling. For the task of establishing a proper and cohesive design, no extensive tools, libraries, or frameworks are expected to be used apart from CSS3 (Cascading Style Sheets). A set of two main colours is to be utilized as the prominent colour scheme of the artefact, in congruence with the less colours, less saturation principle. Additionally, UI elements are expected to be modest and stylized just enough as to not distract users from the artefact's main purpose and result in an overshadowed functionality. Overall, the artefact's UI is envisioned to resemble some aspects of ERP (Enterprise Resource Planning) interfaces, but not all. Percentage representation and visualization is expected to be done due diligently with well adhering graphical components.

3.3 Artefact Architecture

3.3.1 Intended domain and used technologies.

The intended domain of the tool mainly refers to desktop browsers that support JavaScript. The tool itself is a web application constructed on top of a simple full-stack solution. The stack components refer to the Server Side, Middleware and Client Side, along with the used datasets which can be regarded as external files of greater importance for the functionality of the app. Starting off with the server side, as the most important part of the application, is to be built by using NodeJS which is a back-end JavaScript run-time environment for executing JS code outside of browsers. The ID3 algorithm that the server side makes use of is a NodeJS library that supports the creation of decision trees and can be installed with the use of NPM (Node Package Manager). Therefore, the mentioned back-end decision tree library is expected to take JSON files as input, based on which respective traversal functions with the use of the ID3 formula can be carried out.

3.3.2 High-Level Architectural Diagram

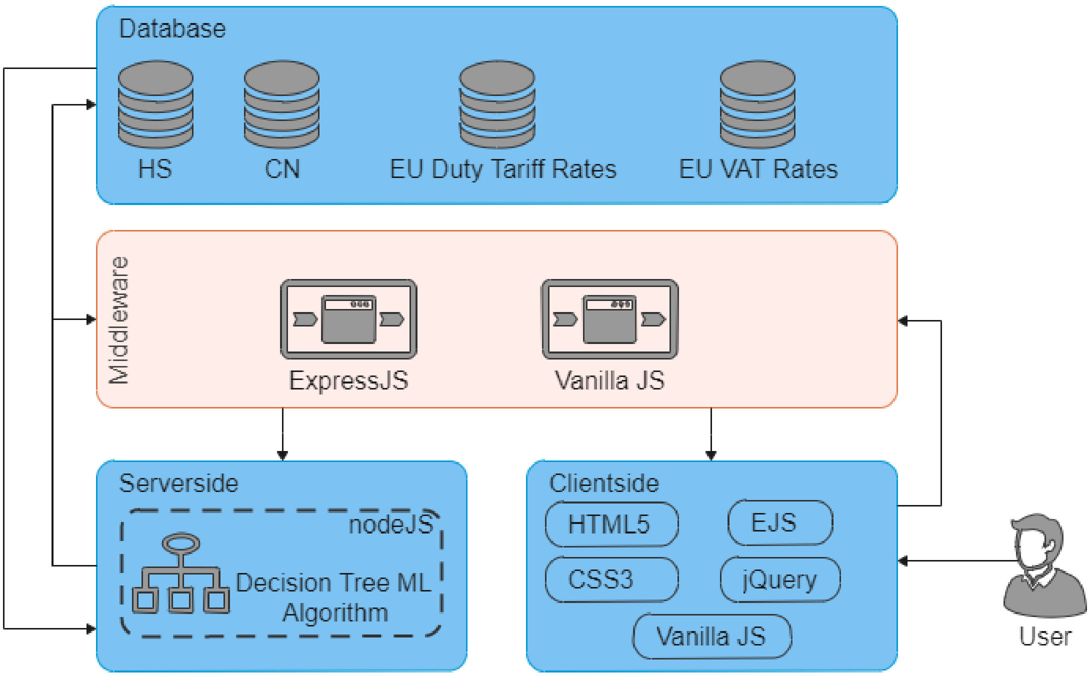


Figure 5 - High-Level Architectural Diagram that is to be used for constructing and developing the proposed artefact.

3.4 Artefact Components

This section aims to elaborate upon each of the components that are to be constructed/ engineered in order for the artefact to be able to perform the envisioned functionalities and tasks. The below components make up the stack infrastructure that the application is intended to operate on.

3.4.1 Server Side

Starting off with the most important component for the implementation and development of the proposed artefact, the server side/ back-end is envisioned to serve as the backbone of the artefact's operational infrastructure. It is important to distinct this particular component with the argument that the "back-end is the far side of the web page or screen functionality; to set the analogy, it could be imagined as the brain or the engine of a website or app" (Yevtushenko & Yalanska, 2021). That also happens to be the case with the development scenario that this research's artefact entails.

The back end is to be implemented with the use of NodeJS, which refers to a server-side platform based on JavaScript and Google Chrome’s V8 engine. The specialty and strength of NodeJS has to do with the fact that it is the industry’s most preferred technology when it comes to server-side programming. It supports a vast variety of JavaScript libraries and frameworks, which in fact, happen to exist solely because of the JS technology itself. This mutually supportive ecosystem allows for the development of truly versatile apps. In addition to Node, the artefact’s back end is envisioned to utilize ExpressJS, which is a NodeJS framework that allows for the functional programming of MVC architectures, identical to ones exhibited in OOP.

In addition to the above-mentioned framework, the server-side also makes use of a node component which respectively refers to the Decision Tree package (utilizing ID3 – Iterative Dichotomiser 3). The decision tree package simply resembles an ID3 implementation case that can be utilized in the NodeJS scripting environment of any web application’s server-side. Such deployable functionality would allow the proposed artefact to possess the capability of creating decision trees using the aforementioned formula for various classification purposes, in this case emphasising HS and CN attribute allocation based on a fixed set of supervised features.

1. The ID3 (Iterative Dichotomiser 3) Decision Tree Algorithm

The ID3 decision tree algorithm, as mentioned by Rizvi (2010), was invented by Ross Quinlan in the year 1968 at the University of Washington. Similarly, to many other decision tree algorithms, the ID3 uses pre-defined attributes, often referred to as features, which are used to calculate numeric metrics like entropy and information gain of nodes. This respective process is executed on each node followed by a splitting of the associated node into two separate nodes. Continuously the metric calculation and splitting process is repeated on all subsequently generated nodes until the simplest micro-decision tree with lowest entropy and highest information gain, as compared to its parent node is generated.

As Brutus, Daniely, and Maslach (2019) point out, “This splitting operation is based on a splitting criterion, which promotes reduction of the training error”. In addition to this statement, it has to be emphasised that “after representing data through these extracted features, the learning algorithm will utilize the label information as well as the data itself to learn a map function f (or a classifier) from features to labels, such as f (features) → labels" (Aggarwal, 2014). In support, to the above-mentioned statement about entropy and IG, it can be further added that "each step during tree construction, we choose the split that causes the largest decrease in impurity, which is the difference between the impurity of data reaching node m and the total entropy of data reaching its branches after the split" (Alpaydin, 2009).

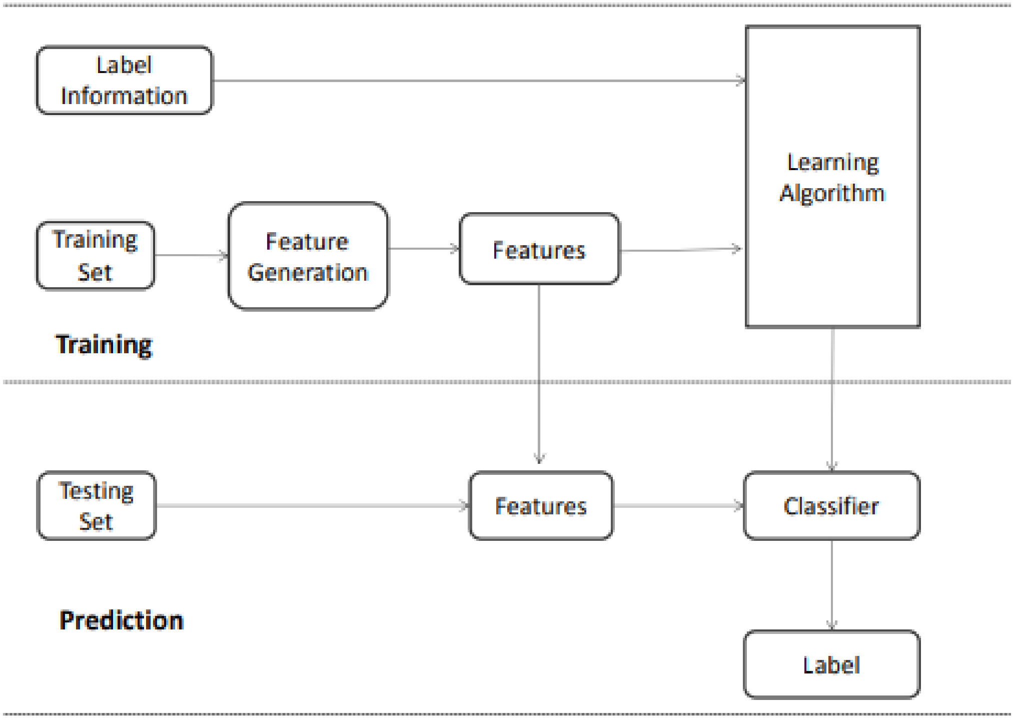
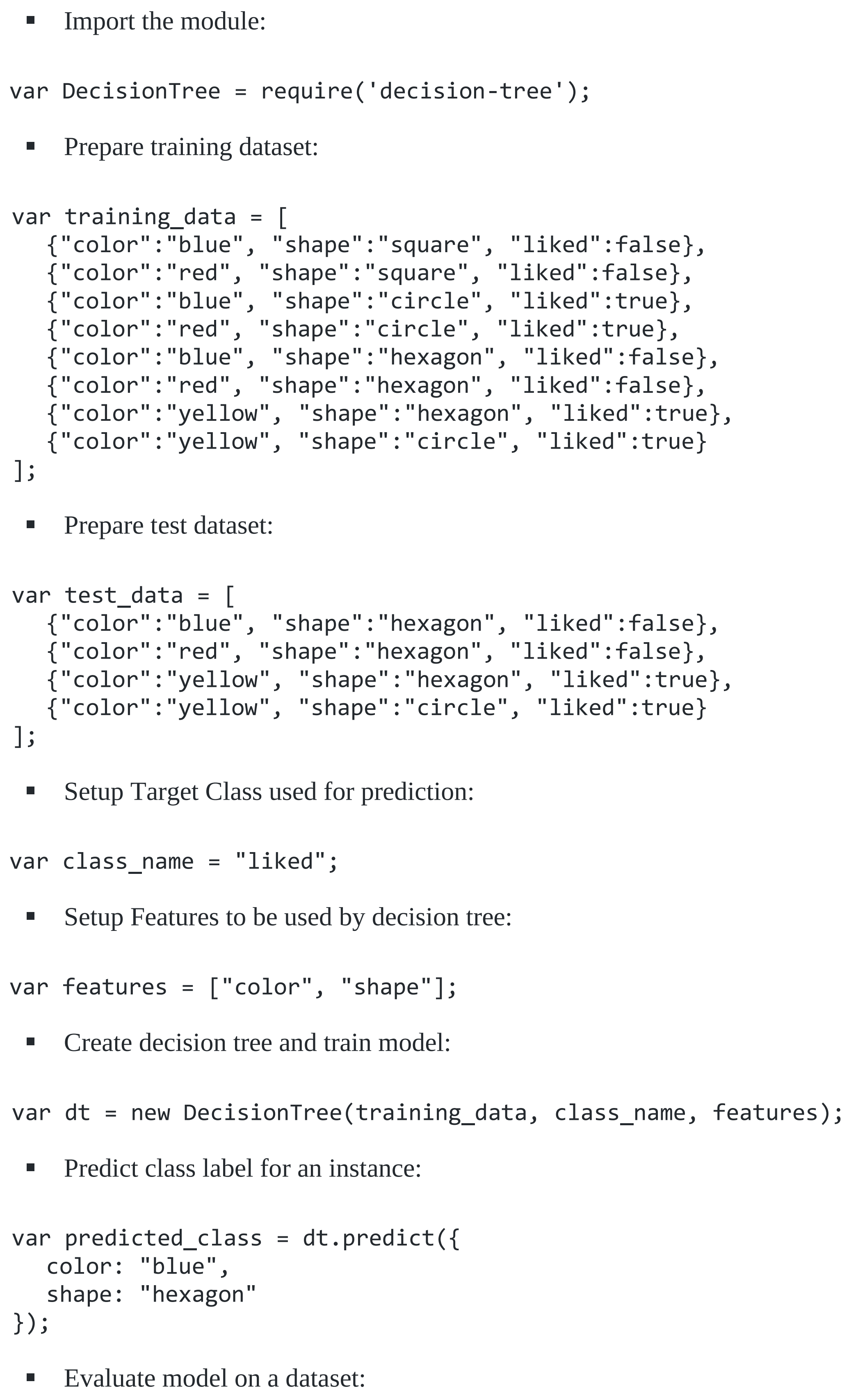


Figure 6 - A general process of data classification (Aggarwal, 2014)

The longevity and affirmed utility of the ID3 algorithm across many sectors has stapled it as one of the most important algorithms that have contributed for the further development of ML solutions for decision trees. Another argument concerning ID3 underlines that "the popularity of this algorithm stems from its simplicity, interpretability and good generalization performance" (Brutus et al., 2019).

Fortunately, the algorithm itself has already been implemented for some of the most well-known programming languages including JavaScript. That is why the NodeJS package that provides implementable functionality of the ID3 for web applications stands as the perfect fit for developing the required underlying functionalities of the artefact. The software module that is ought to be used for generating the required decision trees adhering to the HS and CN datasets needed for the development of the artefact, will make use of a training dataset, testing dataset, target class and a feature set in order to execute a prediction function, the accuracy of which can be evaluated with the testing intended dataset. The modules capability for traversing through large JSON files deem it as flexible and capable of perfectly handling the prospectus datasets referring to the CN and HS nomenclatures.



- Import the module:

```
var DecisionTree = require('decision-tree');
```

- Prepare training dataset:

```
var training_data = [
  {"color":"blue", "shape":"square", "liked":false},
  {"color":"red", "shape":"square", "liked":false},
  {"color":"blue", "shape":"circle", "liked":true},
  {"color":"red", "shape":"circle", "liked":true},
  {"color":"blue", "shape":"hexagon", "liked":false},
  {"color":"red", "shape":"hexagon", "liked":false},
  {"color":"yellow", "shape":"hexagon", "liked":true},
  {"color":"yellow", "shape":"circle", "liked":true}
];
```

- Prepare test dataset:

```
var test_data = [
  {"color":"blue", "shape":"hexagon", "liked":false},
  {"color":"red", "shape":"hexagon", "liked":false},
  {"color":"yellow", "shape":"hexagon", "liked":true},
  {"color":"yellow", "shape":"circle", "liked":true}
];
```

- Setup Target Class used for prediction:

```
var class_name = "liked";
```

- Setup Features to be used by decision tree:

```
var features = ["color", "shape"];
```

- Create decision tree and train model:

```
var dt = new DecisionTree(training_data, class_name, features);
```

- Predict class label for an instance:

```
var predicted_class = dt.predict({
  color: "blue",
  shape: "hexagon"
});
```

- Evaluate model on a dataset:

Figure 7 - An example of how to set up a decision tree with the ID3 algorithm in NodeJS. (Decision Tree for NodeJS, n.d.).

3.4.2 Client Side

The client-side component of the artefact makes use of the traditional web technology stack (universal code) which makes use of HTML5, CSS3 and JavaScript. The use of HTML5 is intended for building the UI structure and visible frame of the app along with the use of EJS which is a templating framework for HTML that allows for the inline embedding of plain JavaScript. The app’s frame and structure that is to be developed is then expected to be styled with the use of the latest Cascading Style Sheets technology, without the use of any external styling frameworks, allowing for light, easily receptive and modifiable source code. Further along the specified technologies, the use of Plain/ Vanilla JavaScript is essential for the responsiveness and functionality of UI components utilized on the front-end of the application. JavaScript serves as the universal scripting language of the web and is almost seen as a necessity when developing web applications, due to its cross-compatibility among browsers and the ease of migration in terms of code base.

In addition to the above-mentioned scripting solution, the open-source lightweight jQuery library for JavaScript is also to be used for further support of JS functions and event handling with the use of less source code. In compliment to the above specified technologies, the client-side portion of the artefact also makes use of additional functionality such as local storage for the preservation of user selected or specified inputs and options. The previously elaborated EJS utility operates in coherence and association with the middleware technology (ExpressJS), which in essence serves as an API allowing for the transmission of GET and POST request data, easily displayable to the end-user with inline scripting. This same task of forwarding server-side data to the client-side layer with EJS is accomplished with the use of template literals and parameter interpolation.

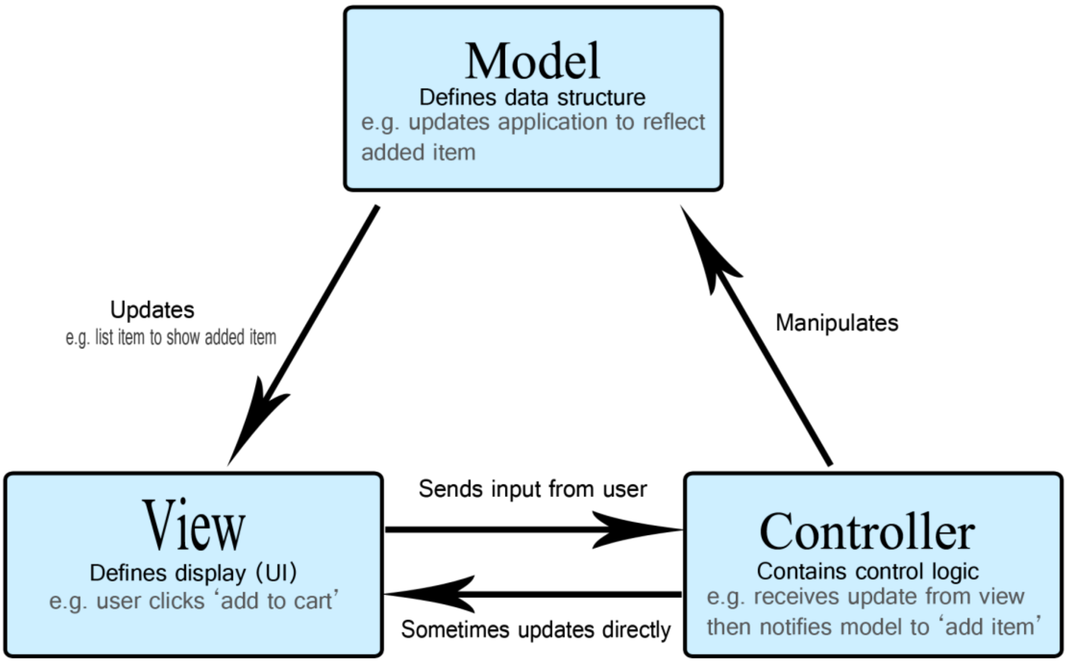
3.4.3 Middleware

The middleware portion of the artefact serves the purpose of back-and-forth data dissemination and user request processing based on HTTP methods and REST operations. The ExpressJS framework for NodeJS which makes up for most of the middleware’s source code, works with the so called “Express Routes” which are a well-known feature of the framework used to programmatically create and expose APIs for server-client communication and vice-versa. It would be negligent not to emphasise what the underlying architecture and mechanism of the REST model aims to deliver, since it is of utmost importance for the envisioned utility of the artefact. As Elkstein (n.d.) points out, the abbreviation stands for “Representational State Transfer” which “relies on a stateless, client-server, cacheable communications protocol -- and in virtually all cases, the HTTP protocol is used”. Moreover, REST “is an architecture style for designing networked applications. The idea is that, rather than using complex mechanisms such as CORBA, RPC or SOAP to connect to machines, simple HTTP is used to make calls between machines” (Elkstein, n.d.).

1. Model View Controller (MVC) Architecture

The MVC architecture model is another mechanism that assures proper communication handling between essential components of the artefact. MVC (Model-View-Controller) is a “pattern in software design commonly used to implement user interfaces, data, and controlling logic” (MDN Web Docs Glossary, n.d.). By establishing a data structure model, in that case with the use of NodeJS for server functions, the view with which the user is presented with can accept inputs and respectively forward them to the controller (ExpressJS). The controller can update the view/ UI based on its pre-programmed logic, and more than certainly it does manipulate the back-end data structure functionality based on the input provided with the purpose of coherently updating the UI (in this particular case with the use of EJS).

This very same architecture is very representative of how the used web technologies for building the artefact interact with each other. When the user provides any input through the presented UI, he feeds the control logic (in this case the middleware), which acts as an intermediary data disseminator to both the UI and server (depending on request and sequence). The same input then feeds methods in the server logic after being sent by the middleware, serving them as function parameters for respective methods, the successful response of which are forwarded directly to the UI. This very mechanism “emphasizes a separation between the software’s business logic and display. This "separation of concerns" provides for a better division of labor and improved maintenance.” (MDN Web Docs Glossary, n.d.).



Model
Defines data structure
e.g. updates application to reflect added item
Updates
e.g. list item to show added item
Manipulates
View
Defines display (UI)
e.g. user clicks 'add to cart'
Sends input from user
Sometimes updates directly
Controller
Contains control logic
e.g. receives update from view then notifies model to 'add item'

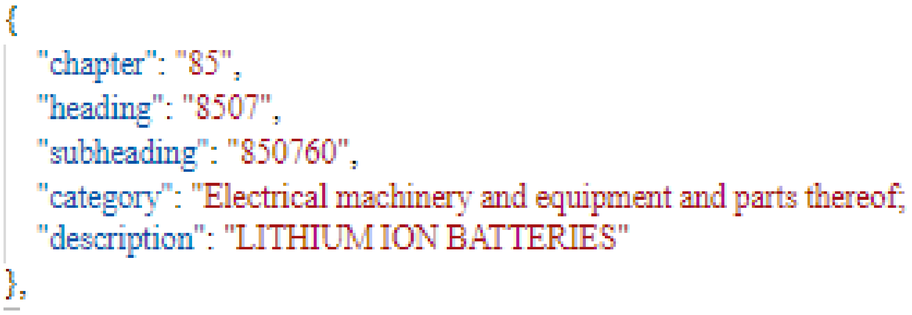
Figure 8 - The MVC Architecture (MDN Web Docs Glossary, n.d.)

2. JavaScript Object Notation (JSON) Datasets for web-based machine learning

JSON, which stands for JavaScript Object Notation, "is a lightweight semi-structured data format based on the data types of programming language JavaScript." (Lv et al., 2019). Furthermore, Lv et al. (2019) mention that, "It is a popular data exchange format over the World Wide Web and becomes a dominant standard format for sending API (Application Programming Interface) requests and responses in the past few years". The fact that a simple object notation format for the internet like JSON has become so influential for the likes of API & web development further affirm the potential that it bears. JSON formatting of data can be justifiably characterised as categorical and hierarchical, which does not make it a perfect prospect for statistics while the notation does happen to be "strictly more expressive than the vector representation" of data (Harris, 2015).

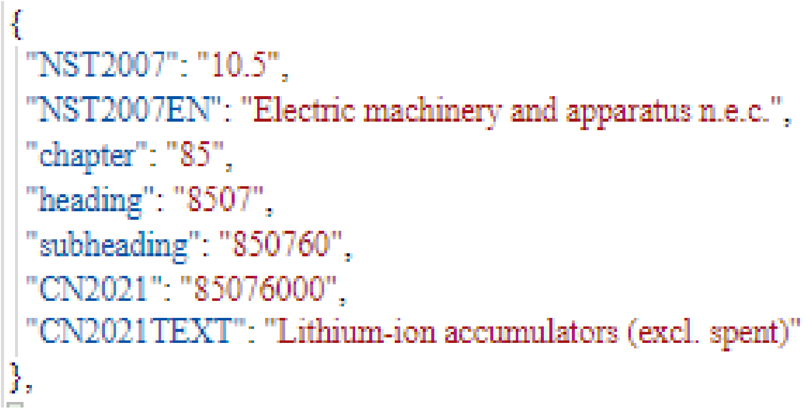
When comparing JSON with other notation formats such as XML, it has to be noted that a JSON document rather adheres to "a set of "key-value" pairs, in which the "value" itself can be a JSON document, which allows arbitrary levels of nesting, so it is more flexible to use and more difficult to process accordingly" (Lv, Yan, & He, 2019). The rich in data attribution and feature specification datasets for the CN and HS, makes JSON the perfect choice for facilitating and employing the above-mentioned nomenclatures in a web application. Not to mention that, just like the CSV format, JSON has become one of the most preferred formats for the likes of open data hosted by numerous institutions around the globe.

This statement can be easily backed-up by the fact that “JSON data model and schema are not only foundations for other data management technologies, such as data indexing, data querying, data searching, data mapping, data integrating, and data mining, but also has important theoretical significance and application prospects to provide theoretical basis and technical means for other related research, such as data integration, data conversion and other semi-structured and unstructured data queries” (Lv et al., 2019). The NodeJS ID3 algorithm implementation that the proposed artefact makes use of has been designed for full compatibility with JSON files and the only drawback of using JSON in this scenario is that “Although the large-scale data represented by JSON provides data resources for data analysis and data mining, which enables us to gain unprecedented insight into data, the cost of processing and querying large-scale JSON data is often very high” (Lv et al., 2019). Nevertheless, the format does remain the best possible prospect for achieving the intended classification and mining functionalities of the envisioned web application.



```
{
  "chapter": "85",
  "heading": "8507",
  "subheading": "850760",
  "category": "Electrical machinery and equipment and parts thereof;
  "description": "LITHIUM ION BATTERIES"
},
```

Figure 9 - The HS nomenclature specification of Lithium Batteries as a JSON object



```
{
 "NST2007": "10.5",
 "NST2007EN": "Electric machinery and apparatus n.e.c.",
 "chapter": "85",
 "heading": "8507",
 "subheading": "850760",
 "CN2021": "85076000",
 "CN2021TEXT": "Lithium-ion accumulators (excl. spent)"
},
```

Figure 10 - The CN nomenclature specification of Lithium Batteries as a JSON object

3.5 Mock-Up Design

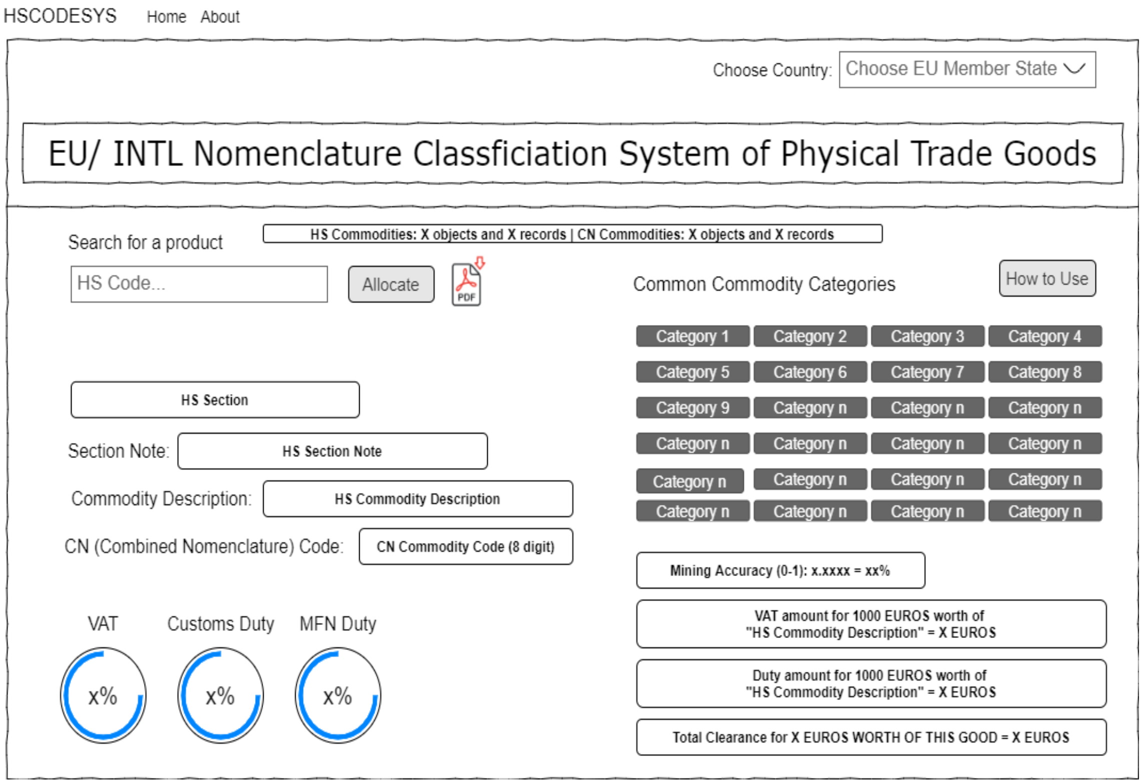

HSCODESYS
Home
About
Choose Country:
Choose EU Member State
EU/ INTL Nomenclature Classficiation System of Physical Trade Goods
Search for a product
HS Commodities: X objects and X records | CN Commodities: X objects and X records
HS Code...
Allocate
PDF
Common Commodity Categories
How to Use
Category 1
Category 2
Category 3
Category 4
Category 5
Category 6
Category 7
Category 8
Category 9
Category n
HS Section
Section Note:
HS Section Note
Commodity Description:
HS Commodity Description
CN (Combined Nomenclature) Code:
CN Commodity Code (8 digit)
Mining Accuracy (0-1): x.xxxx = xx%
VAT amount for 1000 EUROS worth of "HS Commodity Description" = X EUROS
Duty amount for 1000 EUROS worth of "HS Commodity Description" = X EUROS
Total Clearance for X EUROS WORTH OF THIS GOOD = X EUROS
VAT
Customs Duty
MFN Duty
x%
x%
x%

Figure 11 - The envisioned wireframe design for the artefact

4 HSCODESYS and the NCMM (Commodity Classification Mining Model)

The following chapter aims to comprehensively elaborate and analyse the deployed functional components of the produced software by closely examining its ML utility with further detail. The findings of this elaborative process are then to be used for the support and proposal of NCMM (Nomenclature Classification Mining Model) for comprehensive and effective allocation, classification, and mining of cross-border commodities by private or public bodies in web applications.

4.1 Introduction

This chapter is to serve the purpose of answering SQ1 and SQ2 of RQ1 by investigating the deployment specifications and outcomes of the decision tree algorithm (ID3) that has been implemented for the intended classification capabilities of the artefact (HSCODESYS).

RQ1 - How can the specialized application of renowned commodity nomenclatures & ML web technologies in information systems contribute towards more competent classification practices on trade goods throughout the EU?

> **RQ1.1 - Can ML algorithms in web applications assure nomenclature code identification and tax assumption of goods subject to cross-border trade with satisfying accuracy?**
>
> **RQ1.2 - What is a feasible and plausible machine learning approach for mining commodity classification nomenclatures in web server architectures?**

RQ1.1 is to be answered collectively through justifiably concluded artefact development outcomes emphasised in-detail throughout Chapter 4, Section 5.3 of Chapter 5, and Chapter 6 as a result of abiding to the requirements elicitation preceding the developments. Consequently, RQ1.2 is additionally answered by Sub-Section 4.2.2 and Section 4.3, with empirically backed mathematical foundations and a web application architecture justified through a documented as well as evaluated simulation. Based on the underlying theoretical premise that encourages the development of such a solution, the involved creation and presentations of the HSCODESYS stand as the supportive foreground for answering RQ1 in Chapter 7 of this thesis.

The perception that the use of ML approaches associated with decision trees might bear potential when it comes to traversing through big sets of structured or unstructured data for the purpose of classification stands as very justifiable. This is due to the fact that the utilized ID3 algorithm for the HSCODESYS makes use of underlying variables with respective formulas that govern the outcome, accuracy, and success of the allocation process. As so, in continuation, the following sub-points of this chapter will elaborate on underlying models and conceptual representations of the component's utility procedural behaviour within the ecosystem of the artefact.

4.2 The HSCODESYS Component Composition

As already specified above, the main component of the artefact refers to the ID3 Decision Tree Algorithm for data mining, deployed with the use of a NodeJS library. As such, the functional outcomes that can be derived from the use of the deployed algorithm are dependent on mathematical formulas and statistical metrics along with various web technologies associated with the artefact. These mathematical features and visually important elements are to be elaborated accordingly in the chapter sections that follow.

4.2.1 Common Commodity Categories (CCC)

One of the most important visual components of the HSCODESYS artefact is the UI element representing the CCC (Common Commodity Categories). It is a basic representation of commodities subject to e-commerce or general commerce which can be involved in procedures such as customs clearance and processing. Conclusively, it has to be stressed that the options provided are based on product categories present on websites of international e-commerce companies, and the choice of using them for display is solely based on their frequent appearance and importance throughout the e-commerce sphere.

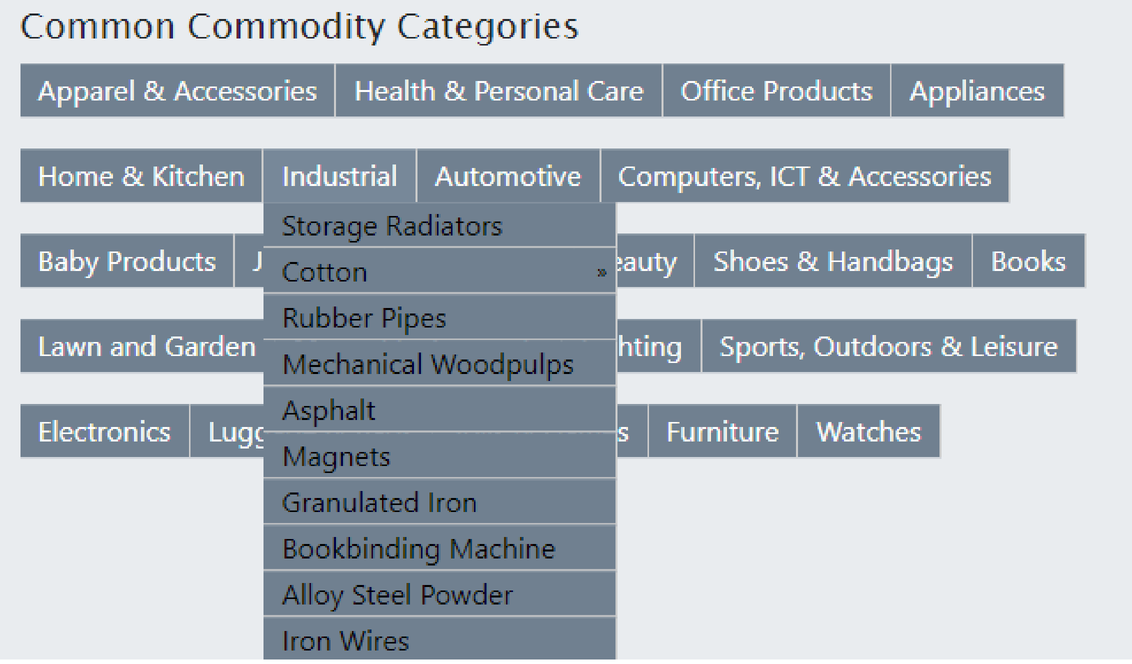


Figure 12 - Common Commodity Categories representing commerce goods that can be classified in accordance with the CN and HS nomenclatures

Essentially, all comprising categories are grouped into a visual element that binds a wide range of cross-border trade goods with their respective Harmonized System codes (6 digits). In order to trigger the decision tree functionality of the artefact, the user is expected to click on a dropdown option that corresponds to a given trade good. Trade good options that can be clicked for information retrieval by the user are represented by labels which are predefined in the server-side of the application with the use of programming logic.

When a given option is clicked, the user is simply presented with the corresponding HS code of the depicted good, as a result of executing an underlying function of the artefact's server side. This click-response mechanism is the preceding step to initiating the mentioned decision tree functionalities of the artefact, which are dependent on a coherent HS code. Upon the successful retrieval of a correct HS code for a specifically clicked option, the user can then proceed with triggering the DT information mining functionality of the artefact which in the given scenario, makes use of 4 datasets, three of which are traversed by using the ID3 algorithm.

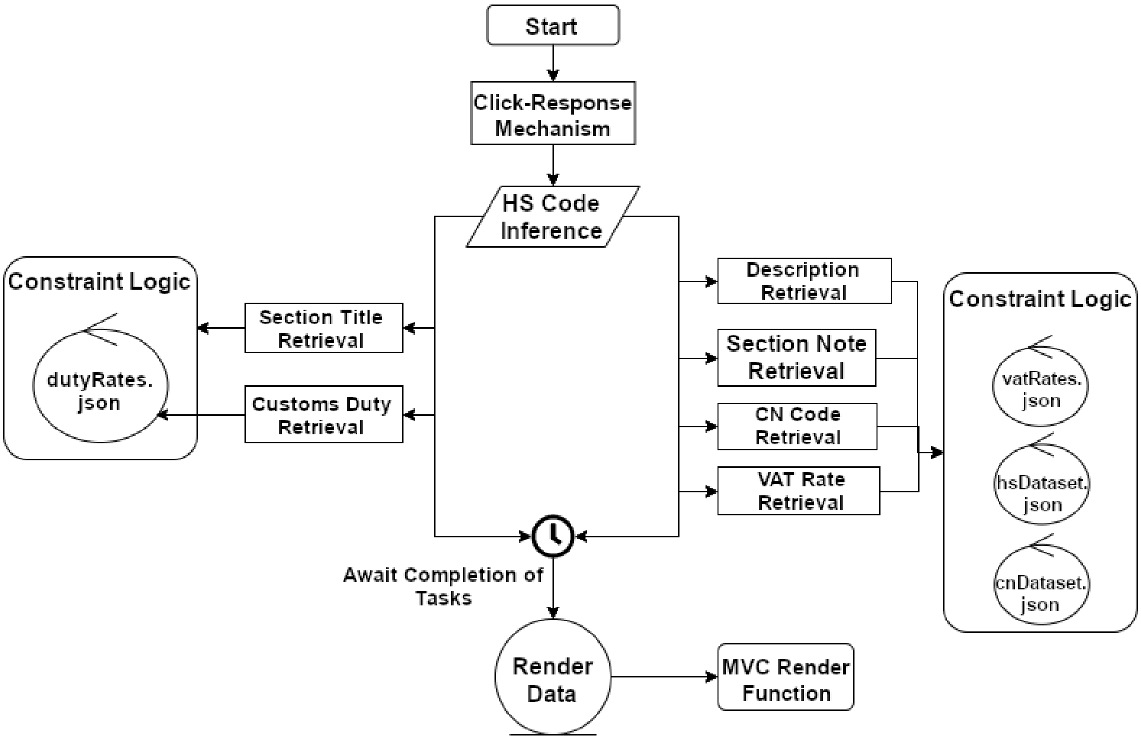


Figure 13 - Operational Flow Diagram for the HSCODESYS

As depicted in Figure 12, the operational chain of tasks that get executed whenever the user initiates in allocating information for a given HS code are specified. It has to be noted that the execution order of each task is irrelevant for the overall outcome of the operation. This has to do with the entity object that gets produced as a result of the multifaceted data traversal process which deems it ready for rendering only upon the completion of all associated retrieval tasks. The exact nature and details of how exactly the pointed-out retrieval tasks are carried out is to be elaborated in the following section.

4.2.2 ID3 Role & Measurement Metrics

The purpose of the ID3 algorithm within the context of the mining model to be proposed, is to mine data from the functionality associated datasets of the HSCODESYS artefact with the sole intent of rendering valuable commodity associated information back to the client-side (browser). This mostly descriptive information can later be used for important customs/ logistics operations associated with the evaluation of trade goods and their respective tariffs. Additionally, the commodities that make up the involved datasets and are ought to be mined by the algorithm can be perceived as ontological information entities, the allocation of which is crucial for the envisioned functionality. The capability for mining data from JSON datasets such as the HS or CN, is established with the use of server-side library for machine learning, also known as the decision-tree library for the server-side scripting language NodeJS. As already elaborated throughout Chapter 3 and more specifically Section 3.4 adhering to the architecture of the artefact, the ID3 algorithm is one of the most influential and important decision tree algorithms due to its ease of use and well renowned flexibility.

The particular deployment/ instantiation of the ID3 algorithm within the architecture of the artefact can be made under a single function which takes user input as a parameter. The user input that is needed for the mining function to succeed is the actual HS code inference produced from the Common Commodity Categories functionality, as shown in the previous section. The code that then gets assigned to the text-field for further allocation of information is split in three parts that get assigned to three different variables. These variables resemble the three double digit sets of an HS code, which adhere to Chapter, Heading and Subheading, as depicted in Figure 3 of this paper. In order for the ID3 to initiate any mining activity on the used datasets, the above stated set of features needs to be provided, again as mentioned in the sub-point of Section 3.4.1.

This set of features is pre-declared throughout the source code of the ID3 function, and it is the most important factor when it comes to carrying out successful allocation/ traversal. The purpose of the allocation function that makes use of the ID3 algorithm fundamentally, in the context of the artefact, is to dissect an HS code, and use its sub-parts as algorithm features for finding the correspondent or closest in similarity of sub-parts CN code. For this purpose, the function is required to traverse the whole CN dataset which consists of approximately 50,000 records.

Such magnitude of data records can be easily regarded as big data, thus classifying the referred task as a process otherwise referable as big data mining. That also happens to be the case with other data mining associated functionalities of the artefact such as HS information mining and VAT tax estimation, with the only difference being the relatively smaller size of the involved datasets. As already mentioned in throughout Section 3.4.1, the statistical properties by which the ID3 algorithm reaches the most likely label associated with the provided features, refer to entropy and information gain.

1. **Entropy**

Entropy is a formula-based criterion in decision trees, especially important for the fundamental functionality of the ID3. From a non-technical perspective, the implications of entropy on information, generally imply that “the greater the information in a message, the lower its randomness, or ‘noisiness’ hence the smaller its entropy” (Burgin, 2003). Another elaboration on entropy refers to it as “the average amount of information needed to determine the outcome of a random variable” deeming it as the actual “uncertainty of the outcome” (Host, 2019). From the above pointed out statements, it becomes evident that entropy is in fact a very important factor for data mining activity in general, and as Chauhan (2020) further adds, “the higher the entropy, the harder it is to draw any conclusions from the information”.

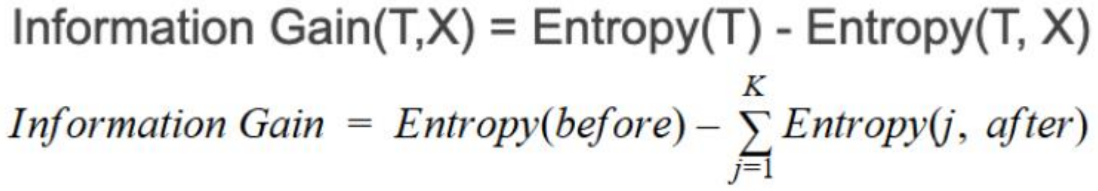
$$\mathrm{H}(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Figure 14 - “Entropy **H(*S*)** is a measure of the amount of uncertainty in the (data) set ***S*** (i.e. entropy characterized the (data) set ***S***)” (Wikipedia, n.d.)

The way ID3 measures entropy in the context of the produced artefact is in accordance with the features provide, which as already mentioned in the previous section, are the actual sub-parts that comprise an HS code. Upon execution, the respective mining function reaches out to each branch of the dataset and calculates the entropy of the given branch. If a branch happens to have an entropy value of 0, it is recognized by the algorithm as a leaf that could be the potentially corresponding classifier to the initially provided features. If the ID3 algorithm happens to calculate a value above 0 on a given branch, this implies the non-correspondence to the features being used, resulting in a further splitting of the branch as to decrease the encountered entropy. This splitting process re-iterates itself until a leaf node with satisfying entropy is being allocated.

2. Information Gain

Information gain is another formula-based criterion that is utilized in the ID3 algorithm and is used for the evaluation of inference that results from already calculated entropy values. Contrary to how entropy works, information gain is a formula that measures the congruence between allocation variables (features) and actual dataset values involved in the mining process. As Chauhan (2020) describes it, “Information gain or IG is a statistical property that measures how well a given attribute separates the training examples according to their target classification”.

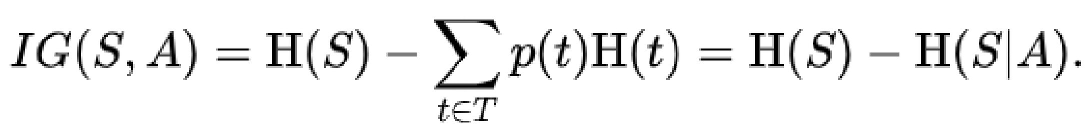


$$\text{Information Gain(T,X) = Entropy(T) - Entropy(T, X)}$$

$$Information\ Gain\ =\ Entropy(before) - \sum_{j=1}^{K} Entropy(j,\ after)$$

Figure 15 - Abstract mathematical formula representation for calculating Information Gain (Chauhan, 2020)

To understand information gain, it must be noted, in accordance with what Brutus et al. (2019) argue, that the recursive nature of how ID3 decision trees are produced, involves recursive leaf splitting based on variables that have the highest information gain. By comparing the utilized dataset’s before and after state regarding decision tree inferences and modifications such as splitting of branch nodes and entropy calculations, the IG property concludes any observable variances in actual values. If the entropy values of the newly produced decision tree resulting from branch splitting in accordance with provided values is less than the entropy values of the non-modified initial tree, then the inference is regarded as positive. That is due to the fact that “Information gain is a decrease in entropy” and as such it is computed based on “the difference between entropy before split and average entropy after split of the dataset based on given attribute values” (Chauhan, 2020).



$$IG(S,A) = \mathrm{H}(S) - \sum_{t\in T} p(t)\mathrm{H}(t) = \mathrm{H}(S) - \mathrm{H}(S|A).$$

Figure 16 - “Information gain ***IG(A)*** is the measure of the difference in entropy from before to after the set ***S*** is split on an attribute ***A***. In other words, how much uncertainty in ***S*** was reduced after splitting set ***S*** on attribute ***A***.” (Wikipedia, n.d.)

3. ID3 Conclusion

As emphasised in the sub-sections of the 4.2.2 Section of Chapter 4 addressing the ID3 algorithm and its criterion associated metrics and formulas, it becomes clear that the algorithm works in a recursive manner and its outcome and success rate depend on statistical properties that are produced throughout the execution process of the algorithm. The sequence of steps by which the ID3 algorithm propagates itself in a successive manner, as Chauhan (2020) addresses, consist of five consecutive phases referring to the initial one as the perception of the dataset by the algorithm as the root node.

Further speaking, each consecutive iteration that follows intends on producing entropy and information gain values that are associated with the unused attributes of the particular dataset. Upon calculation of the mentioned values, by which specific attributes are represented, the algorithm selects those that have their information gain outweigh their entropy. Consequently, the attribute that has been selected is used to further split the dataset into a subset upon which the already mentioned metrics are expected to be yet again calculated, leading to an ongoing recursive process that considers only attributes that have not been previously selected or subject to any metric calculation.

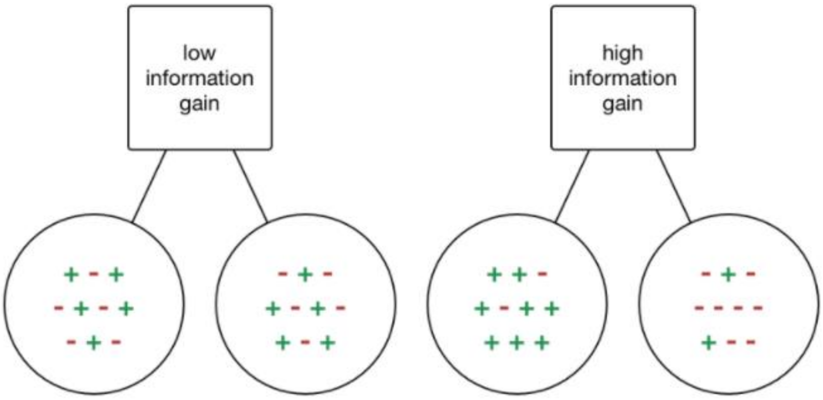


Figure 17 - Information gain representational schema. Low IG is represented by greater levels of similarity between completely split decision trees and their initial state counterparts whereas high IG is represented by greater variance of the associated values. (Chauhan, 2020) Note: the green pluses and red minuses represent positive/ negative entropy values.

4.3 The NCMM (Nomenclature Classification Mining Model)

This chapter serves the purpose of introducing the second artefact that this research aims to propose and evaluate as a direct result of the design and engineering process involved in building the HSCODESYS artefact. The findings and justifications that are concluded from answering the sub-questions of RQ1 are envisioned to support the proposition of the artefact, in this case a model, which is to be foundationally backed-up by the conclusions drawn from Chapters 3 and 4.2 of this paper. The model that is ought to be proposed in the subsequent sections of this chapter is expected to undergo respective evaluation based on which the value of the artefact can be justified. By doing so, the findings of this chapter along with the utility measurements and simulation results inferred from Chapter 5 and Chapter 6 are then to be used for answering RQ2.

4.3.1 Introduction

The NCMM (Nomenclature Classification Mining Model) is the underlying infrastructural foundation that the HSCODESYS makes use of for the fundamental facilitation and provision of DT functionalities for mining trade good nomenclatures. As such, it serves the purpose of establishing an architectural framework by the use of which, machine learning functionalities for querying commodity classifications from well-known and renowned nomenclature datasets such as the HS (Harmonized System) and CN (Combined Nomenclature) can be made possible. The NCMM bears the best deployment outcomes for web applications that make use of the NodeJS server-side scripting language while not being strictly limited to it, also allowing for deployment in other scripting web environment that make use of the MVC (Model View Controller) architecture.

4.3.2 The Model and Features

In its essence, the NCMM is a model that aims to resemble an architectural/ infrastructure-based blueprint for web applications by addressing all three layers of the ordinary web stack (client-side, middleware, server-side). The fundamental purpose of this model is for it to be used as a deployment framework for the embodiment of supervised learning capabilities for trade goods classification in server-side applications. This is to say that, by following this model, IS engineers and designers can develop web applications similar in their underlying functionality to that of the HSCODESYS. Simply put, the model consists of three main layers, each of which has its own components, features and attributes that characterize its role in the given architectural framework schema. The expected conceptual outcome upon successful deployment of the proposed model refers to the idea of decentralized classification.

4.3.3 NCMM Model Schema

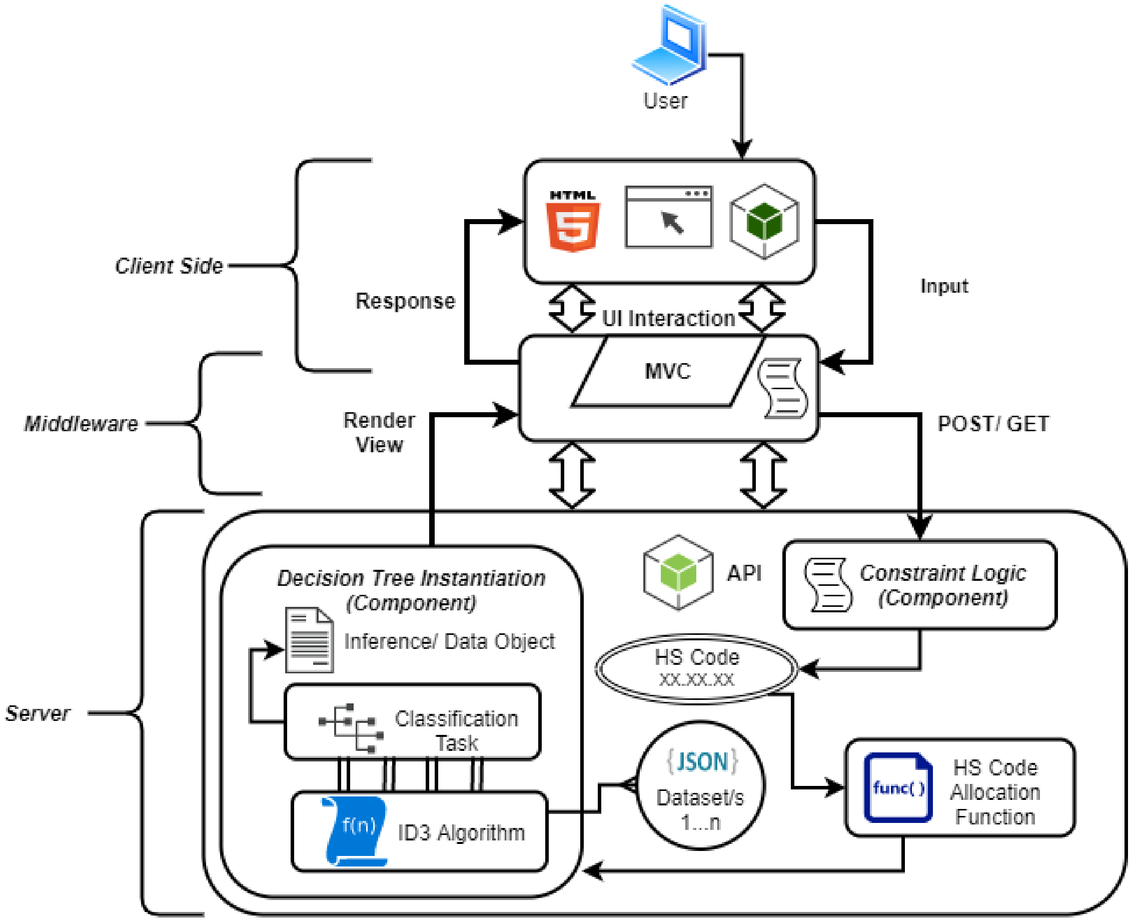


Figure 18 - NCMM Model Schema (Low-level architectural framework diagram representation for the HSCODESYS)

The schema above resembles the model schema representative of the NCMM. As already mentioned, this model makes use of three layers that communicate with each other on various request instantiations. These types of requests refer to GET and POST, which happen to be two of the most majorly used methods when building and consuming REST (Representational State Transfer) API's. Non-exceptionally, the HSCODESYS app has an underlying API built with NodeJS that deploys POST methods to modify parts of the underlying model that can be rendered on the view with the use of the controller. The deployed communication mechanic between server and user happens with the use of a MVC, as described in the first point of Section 3.4.3 of Chapter 3.

Nevertheless, the most important layer of the schema is the server layer where the depicted classification process takes place. The classification processes of the model can be referred to as decentralized due to its capability of instantiating the ID3 algorithm on more than one JSON dataset in a simultaneous yet single-threaded manner. This statement can be additionally used to further support the argument that the NCMM model preaches the deployment and mining of multiple datasets that may or may not have common data attributes. Every instantiation of a POST/ GET request, as displayed on Figure 18, is handled by the MVC layer through UI interaction and further processed by the deployed server components of the model. Respectively, the input that is provided in each request is used to trigger specifically bound functions by making use of the aforementioned ML technology. The first step of the chained classification process utilizes the already emphasised CCC for producing inference based on user interaction through the provided dropdown menus of the app. The type of inference to be produced, depends solely on the deployed constraint logic behind the already mentioned dropdown UI element.

Furthermore, inference produced by the facilitated interaction between UI and browser by the use of the CCC is used as input for the Decision Tree component, important for triggering and accomplishing the multifaceted classification task of the model. It is in this phase of the instantiation process, which utilizes a formula-based method, where the already produced inferential HS code by the constraint logic of the server, is used as input for executing the DT component task itself. The component makes use of the already extensively elaborated ID3 algorithm, the purpose of which is to execute traversal tasks on bound datasets (CN, HS, VAT) to allocate and further present the user with accurate classification labels of goods, based on inferential HS codes.

It is important to note that, the time complexity of this traversal classification task is influenced by the datasets at hand. Subsequently, the time needed for presenting whatever data to the user is heavily dependent on the volume of all involved datasets and their records. Lastly, when the DT algorithm completes its execution by having traversed through all data records of the involved datasets (in this case JSON objects), the response that gets produced as a result of this process can be referred to as a data object that gets rendered on the client-side with the use of the deployed MVC layer.

5 DSR Methods and Utility Evaluation

The purpose of this section is to emphasise the utility of the initially proposed artefact (HSCODESYS) of this thesis by adhering to a utility tree approach that addresses every valuable aspect of the given artefact. The framing and elaboration of the supposed utility tree is to be accompanied with an emphasis on existing DSR evaluation methods along with the determination of the most appropriate method for evaluating each specified utility feature.

5.1 Introduction

Throughout the course of this chapter, every utility associated feature of the artefact is to be examined and elaborated so as to provide affirmation for justifiably answering RQ3 and its sub-questions. Every identified utility feature of the artefact is to be evaluated according to the most plausible DSR method, based on the foundations provided by design science research.

RQ3: How to evaluate the utility of the HSCODESYS?

- **RQ3.1: What is the utility representation of the HSCODESYS?**
- **RQ3.2: What is the most feasible DSR evaluation method suitable for measuring the utility of the HSCODESYS?**

The answer to RQ3.1 can be seen on Figure 22, and throughout the whole of Section 5.3. Utility representation simply refers to the visually comprehendible depiction of utility features presently embodied in the HSCODESYS artefact, as displayed, through the use of a utility tree. Subsequently, RQ 3.2 is answered throughout Section 5.2 as well as Sub-Section 6.1.2 of Section 6.1, Chapter 6. More specifically, Figure 21 adheres to an accurate depiction of what the most suitable DSR evaluation method/ path for the artefact of this thesis is.

5.2 DSR Evaluation Methods

Despite design science presenting itself as a rather novel research methodology associated with the design and development of information systems, it has shown importance and legitimacy in the IS field (Gregor and Hevner, 2013). Through the application of this methodology, researchers intend on producing artefacts that can be perceived as tangible research outcomes under a conceivable material form within the digital field. In order to support the underlying abstract but knowledge-based theories that IS research is usually accompanied by, artefacts ought to be produced in the process have to be due diligently evaluated. This can be additionally supported by the argument Winter, Zhao, and Aier (2010) make, which suggests that solution inventions like artefacts as a result of DSR are perceived as the core activity while being surrounded by theoretical establishments and proper evaluation.

Respectively, proper evaluation is of utmost importance in design science for it measures the knowledge produced from applying the associated design theory principles and the values that can be derived as a direct result of their implementation. Furthermore, evaluation within the DSR context can be referred to as "the process of determining how well the artifact performs" (March & Smith, 1995). There exist various methods that have been comprehensively documented throughout design science literature as viable for measuring the performance and effectiveness derived values of artefacts.

There are two majorly known and distinctive approaches when it comes to evaluating artefacts and those refer to naturalistic and artificial. As the term suggests, the naturalistic approach is associated with the positioning and testing of an artefact in real-world circumstances with real people and operational environments. As additionally underlined by Venable, Pries-Heje, and Baskerville (2012), naturalistic evaluation usually touches on the "performance of a solution technology in its real environment".

Artificial evaluation on the other hand, again as the term might suggest, is based around means of simulation and virtual testing which might include "laboratory experiments, field experiments, simulations, criteria-based analysis, theoretical arguments, and mathematical proofs" (Venable et al., 2012). With the rather short emphasis on the above-mentioned evaluation approaches, the question of how to evaluate in DSR is rather clear with the two possible options being specified. The subsequent question that would then follow respectively poses the question of when to evaluate a given artefact.

Design science literature usually points out two well renowned options that address the timing for conducting evaluation on DSR derived artefacts, with these referring to ex-ante and ex-post. Additionally, as Venable et al. (2012) further emphasise, ex ante is the evaluation of an artefact before it has been constructed while ex post is the post-production evaluation of an artefact. This perception of how-to evaluation DSR derived artefacts can also be supported by the statement that “ex post evaluation is evaluation of an instantiated artifact (i.e. an instantiation) and ex ante evaluation is evaluation of an uninstantiated artifact, such as a design or model” (Venable et al., 2012). There are various inferentially and statistically associated factors that are used for selecting and carrying out each respective evaluation approach.

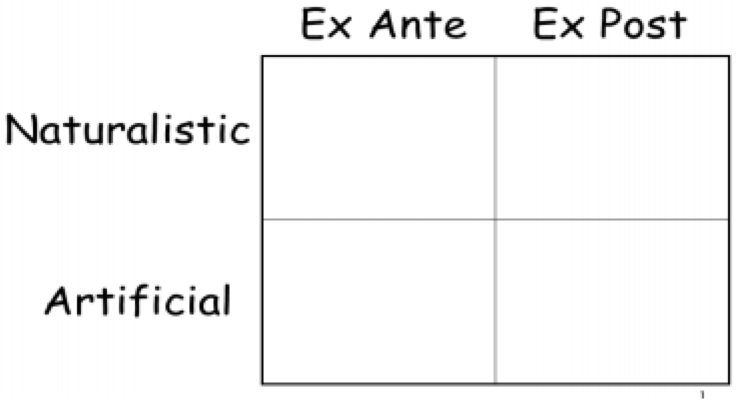
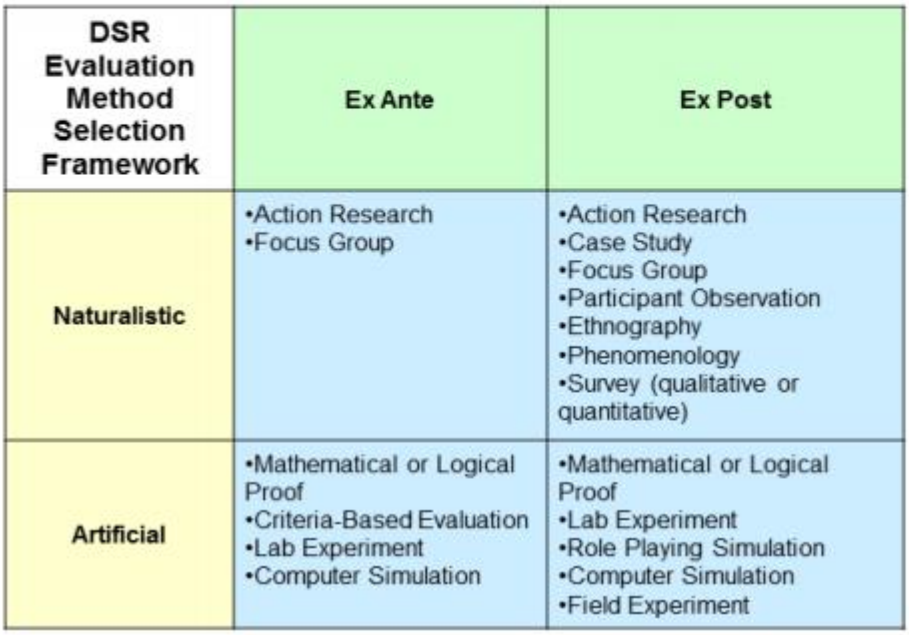


Figure 19 - DSR Evaluation Strategy Framework (Pries-Heje, Baskerville & Venable, 2008)

As a result of choosing a specific evaluation approach, the core concerning question which arises regarding the course of a DSR evaluation, is the specification of actual research methods ought to be used. Following the context of the artificial approach, the methodological tools that rather exist as depicted in Figure 20 by Venable et al. (2012), refer to field experiments, simulations, mathematical proofs, etc.

Consequently, based on the preceding elaborations on evaluation methods, the conclusion that can be drawn is one that supports the artificial evaluation of the HSCODESYS artefact in an ex-post manner. The evaluation of the artefact stands as appropriate due to its construction in a short period of time, in congruence with the time schedule of this research which does not allow for a naturalistic evaluation approach. The artefact, in this case a prototype of type instantiation, is most suitable for ex post evaluation since all functionality and design associated with the utility evaluation have to be implemented in advance.



DSR Evaluation Method Selection Framework	Ex Ante	Ex Post
Naturalistic	•Action Research •Focus Group	•Action Research •Case Study •Focus Group •Participant Observation •Ethnography •Phenomenology •Survey (qualitative or quantitative)
Artificial	•Mathematical or Logical Proof •Criteria-Based Evaluation •Lab Experiment •Computer Simulation	•Mathematical or Logical Proof •Lab Experiment •Role Playing Simulation •Computer Simulation •Field Experiment

Figure 20 - DSR Evaluation Method Selection Framework (Venable et al., 2012)

Therefore, the knowledge goal that underlies the scientific enquiry of this specific evaluation approach needs to be emphasised as well. As Herwix and Rosenkranz (2018) address, DSR presents a duality aspect of its own when it comes to producing design as well as science associated knowledge. It is further stressed that, “whereas design is practical, generally creative and hard to structure process concerned with the construction of useful artifacts, science is a rigorous, systematic and highly structured endeavour concerned with the discovery of new knowledge about the nature of things – DSR is difficult to grasp because it aims to integrate the two in a mutually supporting whole” (Herwix & Rosenkranz, 2018).

As a result, deciding on the best adherent inquiry, inherent to the depicted DSR method, can be done based on the inquiry genres for design science knowledge table established by Baskerville, Kaul and Storey (2015), and further modified by Akoka, Comyn-Wattiau, Prat, and Storey (2017). Using the supposed table for specifying the enquiry genre based on knowledge scope and goals, the uttermost knowledge production episode for the depicted DSR can be chosen. Conclusively, idiographic design stands out as the best fit for the intended knowledge production process associated with the creation of the HSCODESYS artefact. The nature of this design approach adheres to “Knowledge necessary for the research-and-development of individual product” where the “knowledge role of the artifact is one of materializing or embodying this knowledge” (Baskerville et al., 2015).

Table 6 - Design Science Knowledge Table for Genre Enquiry (Baskerville et al., 2015)

		Knowledge Goals	
		Design	Science
Knowledge Scope	Nomothetic	Nomothetic Design	Nomothetic Science
	Idiographic	Idiographic Design	Idiographic Science

As to conclude, the question of why to evaluate which might also translate as the purpose of the specified artefact, also needs to be emphasised. Venable, Pries-Heje, and Baskerville (2016) stress that, according to the FEDS strategy, functional purposes behind a given DSR evaluation serve as one of two specified dimensions of the strategy and could be either formative or summative. Additionally, the second dimension of the FEDS strategy refers to “paradigm of the evaluation (artificial or naturalistic)” (Venable et al., 2016).

Furthermore, Sein, Henfridsson, Purao, Rossi, Lindgren (2011), argue that when a given artefact is in its alpha version, the evaluation that takes place addresses formative purposes while evaluation on beta versions is aimed towards summative purposes. This statement can additionally be supported by the observation made by Gregor and Hevner (2013), which states that whenever considerable efforts are being put towards the development of a given artefact with already conducted formative testing, summative testing should not be as elaborative when the artefact happens to be developed by a third person.

In the case of this thesis, where the artefact is not built by a third person and extensive formative testing has not been prioritised due to the current beta version, what is left is a comprehensive summative approach. Lastly, by selecting the most suitable DSR evaluation method based on the already mentioned framework and strategy, the main aim remains the justification of worth by the use of "final summative tests in case studies or experiments, expert review, simulations, statistics on usage data for implemented systems, and evidence of impact in the field" (Gregor & Hevner, 2013).

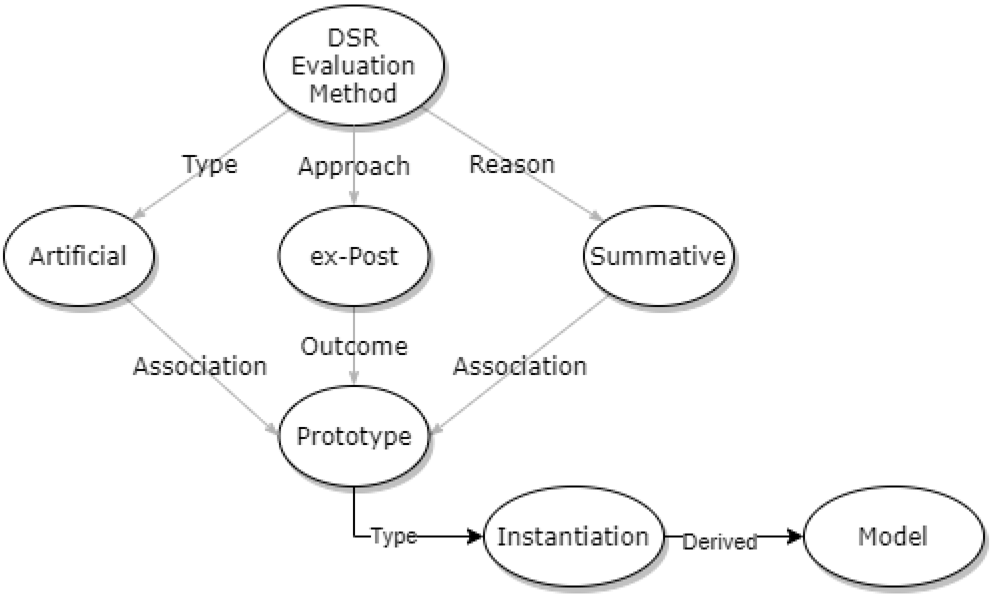


Figure 21 - The DSR Evaluation path best suitable for the artefact of this research. In congruence with the DSR Evaluation Framework proposed by Venable et al., (2012).

To summarize the findings of this chapter and as the above figure depicts, the DSR evaluation method that this research aims to undergo refers to that of type artificial, along with the specified approach being ex-post and the evaluative reasoning referring to summative. As Figure 22 further displays, the actual type of artefact/s ought to be evaluated are the HSCODESYS (Instantiation) and NCMM (Model). The main objective of the chosen evaluation approach is to justify and approve of the asserted utility of the initially proposed artefact, a web application, based on which the deriving and underlying model, an architectural framework, can be congruently justified.

5.3 HSCODESYS Utility Tree

The purpose of this section is to establish a utility tree adhering to features embodied in the HSCODESYS, the branches and sub-branches of which are to be evaluated accordingly as to produce a summative outcome/ conclusion that will serve for justifying the utility traits and features of the artefact. As such, all testing required by this section is to be conducted on a beta version of the artefact deployed for online access through the Heroku platform, which support facilitation and integration of NodeJS and NPM for web applications.

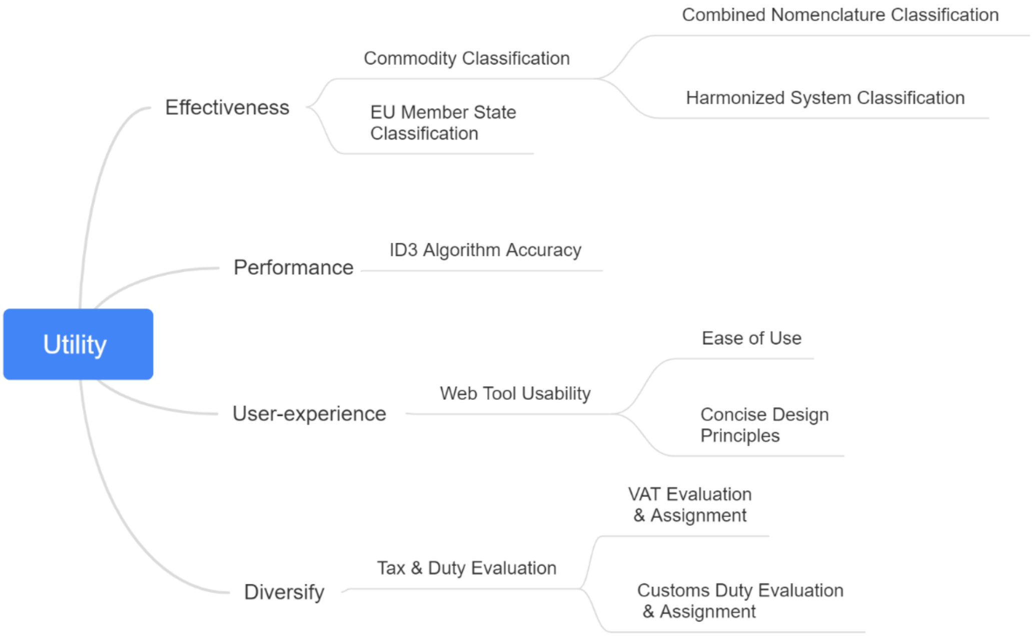

Utility
Effectiveness
Commodity Classification
Combined Nomenclature Classification
Harmonized System Classification
EU Member State Classification
Performance
ID3 Algorithm Accuracy
User-experience
Web Tool Usability
Ease of Use
Concise Design Principles
Diversify
Tax & Duty Evaluation
VAT Evaluation & Assignment
Customs Duty Evaluation & Assignment

Figure 22 - Utility Tree for the HSCODESYS Artefact (Addressing Utility associated features)

Effectiveness in terms of utility can be resembled as one of the most valuable utility features of the depicted tree in Figure 22 due to its association and responsibility with guarantying the main classification capabilities and functionalities of the web application. The effectiveness branch of the utility tree, as shown on the already mentioned figure, is further split into two sub-branches which refer to commodity classification and EU member state classification.

As perhaps stressed in the requirements elicitation table of this paper, the capabilities of the artefact for facilitating commodity classification functionality is among the main development priorities of this research and its assurance is of utmost importance. As already elaborated, the classification methods that the artefact makes use of are associated with internationally used trade good nomenclatures, the usage of which under the form of a dataset, is crucial for satisfying the effectiveness utility branch. EU member state classification on the other hand, happens to be the second branch associated with effectiveness, and as such, its implementation is vital for evaluating VAT tax estimates on classified trade goods, which happens to be another objective associated with the diversification of utility for the associated artefact.

5.3.1 Commodity Classification

As perhaps already pointed out, the commodity classification utility of the HSCODESYS can be split into two respective parts. Regarding the first part, also depicted as Harmonized System classification, it has to be noted that it makes use of a unique HS code intended for producing classification data. This classification data, in the context of the web tool, adheres to the taxonomy associated labels used for hierarchical structuring of HS codes and their representation by "series of 4-digit headings, most of which are further subdivided into 5- and 6- digit subheadings" (WCO, n.d.).

The Combined Nomenclature classification functionality of the proposed web tool on the other hand, also makes use of HS codes for the triggering of mining activity, although for different purposes. The aim of mining CN codes based on HS code fragments revolves around matters of mathematically reaching the most adhering CN code, based on the functional instantiation of the ID3 algorithm.

1. **Combined Nomenclature Classification**

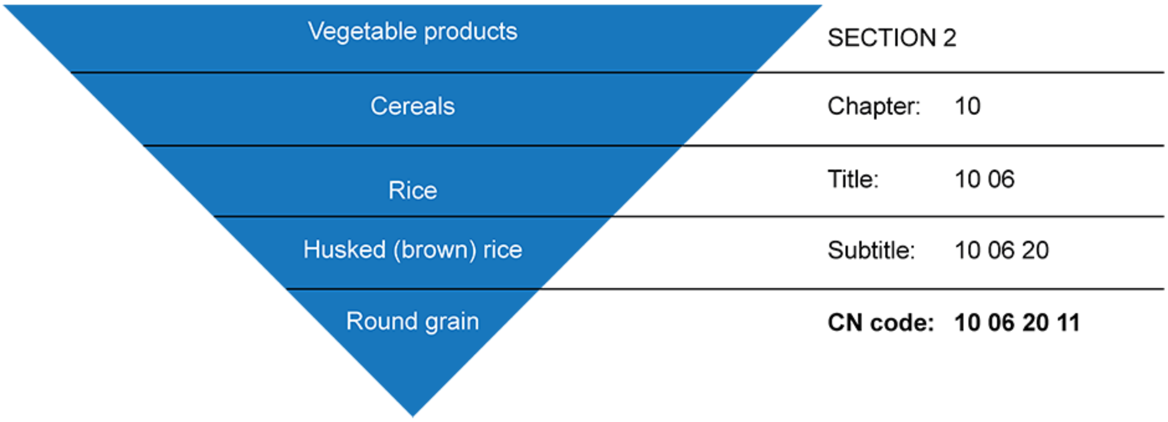


Figure 23 - Taxonomical Structure of a Combined Nomenclature Code (Statistics Denmark, 2021)

As perhaps shown on Figure 23, the hierarchical structure of a CN code resembles that of an HS code. The fact of the matter is that a combined nomenclature code is essentially an establishment which is based on the pre-defined taxonomical structure of the Harmonized System, and as such can be perceived as an extension of it. As perhaps already mentioned, the CN stands as one of two major constructs used by the EU for classification of goods and is additionally used for adhering to CCT (Common Customs Tariff) rules.

In order to classify a CN code from a dataset of thousands of records, the provision of accurately adhering parameters needed by the underlying ID3 algorithm is very important and decisive to the final result. The web application that this thesis aims to develop and propose, makes use of the 2020 version of the CN issued by Eurostat. More precisely, the dataset itself, which is of format type JSON, consists of 34500 records that make up 8625 JSON objects. By making use of designated HS codes and their double-digit sub-parts that respectively comprise CN codes, the intended mining functionality can be achieved.

For this given reason, the HSCODESYS web tool utilizes preservation of code inferences derived from UI interaction with the use of client-side scripting. Upon completion of the ID3 traversal process, the underlying function which facilitates the algorithm itself can then compare the derived class based on a pre-defined testing data set which produces an accuracy metric. This exact metric is also associated with the evaluation of the ID3 instantiation performance-based accuracy.

2. Harmonized System Classification

As already emphasised throughout various sections and chapters of this thesis, the Harmonized System nomenclature stands as the most valuable piece of data construct utilized throughout the architectural boundaries of the web application. This is due to its international scope and renowned use for cross-border purposes globally, backed-up by the statement that it “contributes to the harmonization of Customs and trade procedures, and non-documentary trade data interchange” (WCO, n.d.). The Harmonized System nomenclature dataset is in the epicentre of shared and common classification functionalities that the artefact embodies.

Moreover, going back to the importance of this nomenclature and its implications for ML utility, it is argued by the WCO (n.d.) that, the adherent collection of tax revenues is dependent on proper classification means, thus reaffirming that accurate classification by using the HS is utmost needed for safeguarding trade regulation, trade statistics and policy data. Conclusively, “the more accurate the application of the HS, the more it serves the needs of its users” by prioritizing “proper application of the HS for day-to-day Customs work, policy development and trade” (WCO, n.d.).

In technical terms, by addressing the deployed algorithm of this paper’s artefact and its derived functionality it can be argued that the comprising parts of a given HS code are the crucially needed inputs for carrying out classification tasks based on instantiations of the ID3. Among some of the ML functionalities dependent on HS code inferential induction include section note mining, commodity description mining, combined nomenclature mining along with ordinary satisfaction of constraint logic deployed by the server of the artefact responsible for customs duty tariff percentage estimation. Needless to say, most of the utility features that the HSCODESYS intends on providing can be perceived possible as a result of integrating the HS nomenclature.

Ultimately, the mere consideration of deploying HS code mining is because of the capabilities it presents for classifying sections and chapters of commodities. By being able to locate a commodity and all of its associated attributes with adherent descriptions and labelling, core activities revolving around estimation of tariff rates on trade goods can be carried out cohesively, as it follows to be addressed throughout this chapter.

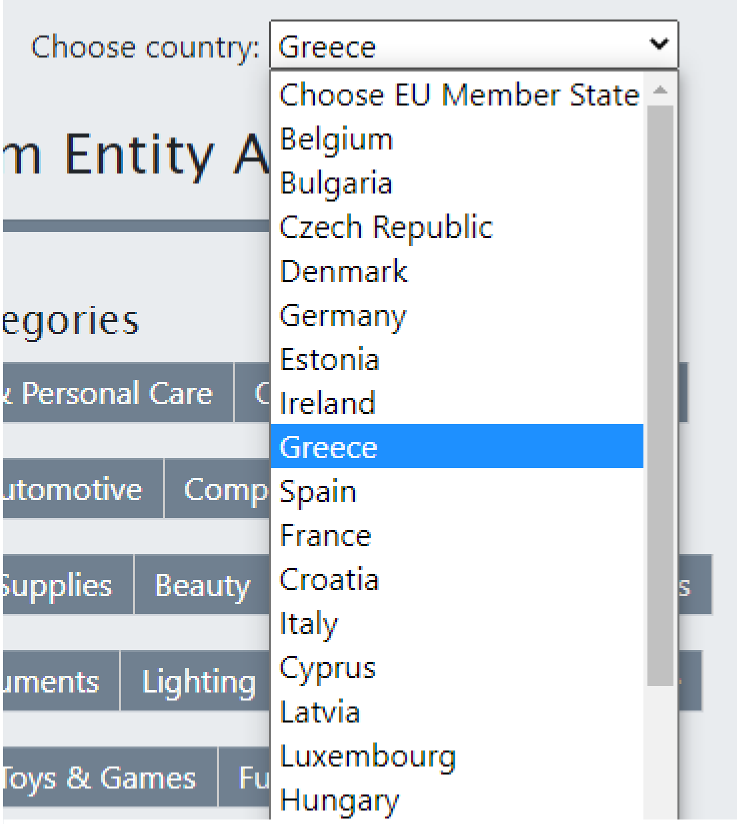
Deploying such functionality can only enrich and pave ways for further scaling and development of classification operations that mostly rely on the Harmonized System nomenclature as a fundamental basis. This on its own, is a noticeable feature of the produced artefact as it shows how certain classification functionalities can be further built on top of already established functional constructs.

5.3.2 EU Member State Classification

The EU member state classification functionality of the artefact is directly associated with the VAT estimation feature that is also presently deployed. By providing the user with a UI element that allows for selecting the EU member state for which the product might be envisioned for importation, further calculation of applicable VAT for the given good can be established. Respectively, this is made possible with the use of a dropdown HTML element which binds each member state option to its VAT rate, as laid down and specified by the European Commission’s (2020) VAT rates declaration.

Upon having chosen a member state option, the user can then proceed with allocating trade good data from the UI categories element (CCC), which in term will result in retrieval and rendering of all associated classification labels for the given option along with a VAT rate estimation presented under the form of a percentage. The programmatic decision of choosing and assigning an intended VAT rate for a specific country option clicked by the user makes use of an ID3 decision tree instantiation. As perhaps already mentioned, this is achieved by declaring a server-side function that makes use of a JSON file facilitating the VAT rates data associated with EU member states.

This server based functional feature of the web application is to be more comprehensively emphasised in the following diversification sub-section of this chapter constituting tax & duty evaluation. Consequently, as to finalize this elaboration, it has to be stressed that the utility that this feature bears is derived from its implication and connection with tax evaluation on the server level and the depicted sense of choice (within the EU) conveyed through the UI element itself.



Choose country: Greece

Figure 24 - The HSCODESYS EU Member states dropdown with underlyingly bound VAT rates

5.3.3 ID3 Algorithm Accuracy

The ID3 algorithm instantiation which is present throughout the architectural foundation of the artefact, makes use of the emphasised mathematical metrics and formulas in Chapter 4.2.2. of this paper for concluding accuracy of inference produced. As perhaps already mentioned, these metrics for the most part, refer to entropy and information gain. The way in which the Node JS based algorithm instantiation of the artefact produces the so called “accuracy” variable, is by means of decision tree comparison.

Prior to the completion of allocation/ traversal queries on the deployed HS and CN datasets, the algorithm itself constructs a decision tree based on the features and target class provided by the user. Right before completing the task in terms of script execution, the functional constraint logic of the deployed server-side script then initiates a comparison of the already constructed DT with one that is based on a testing data sample. The two decision trees are then compared as to produce the accuracy variable which essentially, is an evaluation of the initially created decision tree based on the testing dataset. Formally speaking, a testing dataset has to be comprised of JSON objects which are required to be also present in the mining dataset (HS and CN) used by the algorithm for creating the first DT instance, responsible for producing all taxonomy and classification-based labels ought to be rendered to the user.

As to conclude, the general accuracy of the ID3 algorithm that is deployed by the artefact can be perceived as appeaseable, as inferred from conducting functionality and performance associated unit tests with NodeJS testing framework also known as Mocha. Associated testing results show that, wherever instantiated, the algorithm performs faultlessly with a maximum success rate when it comes to appropriately retrieving all expected response data as declared throughout the pre-defined constraint logic. Additionally, a survey conducted on ordinary users of the HSCODESYS web app with varying amounts of knowledge and understanding on customs procedures shows that the artefact is perceived to perform well and accurately.

5.3.4 Web Tool Usability

This section will for the most part, focus on the usability characteristics and attributes of the artefact that are associated with how the design principles that have been implemented can benefit the user when it comes to understanding and comprehending the deployed UI adaptively for quick and well perceived usage. The reasoning behind using certain design elements and features for constructing the artefact will be explored, so as to elaborate the design rules that the web application abides by as well as the direct value that can be derived from those for presenting an overall easy-to-use interface.

1. **Ease of Use**

As already emphasised, the web tool refers to a Node JS application that can be hosted online and accessed through any browser. Its use of ordinary styling with CSS and the existence of a few graphical elements present on the UI stand out as coherently responsive. This observation can be concluded from recorded survey responses of participants who have interacted with the artefact and witnessed its functionality. The sample size that underlies the survey is comprised of 24 individuals who take active part in the digital transformation field both professionally and academically.

The usability portion associated with the proposed artefact, has a lot to do with its perceived UX characteristics and their utility, which are associated with design and styling attributes.

2. Concise Design Principles

One of the utility features that the artefact proclaims is the embodiment of concise design principles which are mostly associated with how UI and UX serve the end user of the web tool. This utility associated feature mostly has to do with how the design features of the application itself can influence the way in which the user uses the intended artefact and its functionalities in the best way possible.

By having applied simple and concise design features when it comes to style, graphical imagery and animations that might be used, the main aim is to engage the user in a non-distracting way. By doing so the user can proceed towards carrying out the main and most important features of the artefact which, as already emphasised, are mostly classification oriented.

Some of the actual principles that have been applied refer to simplicity, visual hierarchy, grid-based layout, loading time, responsiveness, and consistent colour scheme. By making use of the user feedback survey already mentioned in the previous section of this chapter, the expressed efforts and principles used for achieving a concise design can be further supported.

5.3.5 Tax & Duty Evaluation

This particular feature of the web tool can be divided in two parts. The first part adheres to the VAT calculation and assignment on commodities of the HS and CN, while the second part associates with how duties can be estimated while also abiding to the international nomenclature rules. As perhaps already mentioned in section 5.3.2 of this Chapter as well as Chapter 2.8, the EU VAT tariff profiles are the single most definitive and well-founded data needed for the development of the intended constraint logic behind VAT estimation.

Oppositely, customs duty estimation is based on data from the world tariff profiles publication issued by the WTO, ITC and UNCTAD. The World Trade Organization

(WTO) is said to be the “only global international organization dealing with the rules of trade between nations” and as such, its “main function is to ensure that trade flows as smoothly, predictably and freely as possible” (WTO, 2020). The International Trade Centre (ITC) on the other hand, “is the joint agency of the World Trade Organization and the United Nations” and its main focus revolves around the integration of enterprises and their competitiveness in the international trade and commerce landscape (WTO, 2020). Lastly, the United Nations Conference on Trade and Development (UNCTAD), as emphasised by the WTO (2020), is responsible for the promotion of developmental advancements and the integration of developing countries into the international economy that exists.

Conclusively, it is the produced data by the above-mentioned international organizations and agencies, that stands as most suitable, practically applicable, and reliable information source for establishing the given utility purpose. In terms of specificality, it is the 2020 EU Tariff profile of the mentioned publication which is of most value and importance for the development of the customs duty tax evaluation feature that the HSCODESYS makes use of.

1. VAT Evaluation & Assignment

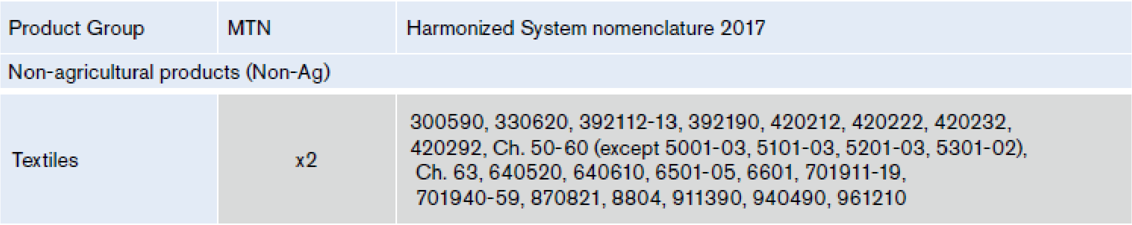
As already elaborated in section 5.3.2 of this chapter, the development and implementation of the VAT evaluation and assignment utility of the application is directly tied with the previously established functionality associated with the EU member state’s dropdown element of the UI. By making use of an underlying JSON file that contains all data needed for initiating and completing this procedure, the constraint logic responsible for this functionality which happens to make use of the ID3 algorithm can be carried out in accordance. The moment a user clicks on a given dropdown option corresponding to an EU member state, the input attributes required for the intended functionality to be executed get passed on to variables which are later on used to initiate an ID3 traversal.

VAT estimation works on the basis of providing a country code and country name, which are preserved whenever a user makes his EU country choice. By using the just mentioned data attributes as features, by which the algorithm intends on finding the most coherent class (VAT rate of an EU member state), the mining process can be successfully initiated. This functionality in itself is not a complicated one and does not revolve around the traversal of a huge dataset. The value of this utility is associated with the diversification of data ought to be presented to the end user, and as such resembles the applicable VAT rate associated with a given classified trade good based on chosen EU member state for importation.

Conclusively, the utility-based value of this feature is mostly derived from using the ID3 algorithm which assures accurate and efficient classification of VAT rates based on country codes and names of EU member states. In addition, as expected for further elaboration in the following and perhaps final section of this chapter, both of the diversification utilities associated with tax and duty evaluation and their contextual value are to be reaffirmed by considering results of the undertaken user acceptance survey.

2. Customs Duty Evaluation & Assignment

Specialized identification and evaluation of tariff numbers such as those present in the HS and CN is crucial when it comes to complying with internationally established regulation. When stressing the international regulatory importance of tariff classification based on uniform standards, it is mentioned that commodities which might share common characteristics "must have the same uniform and binding code number, regardless of whether it is imported into Mexico, Germany or South Africa" (Wegner, 2018). Wegner (2018) goes on to further emphasise that, classification is equally important for private companies as much as it is for public customs authorities, and customs duty on products can vary alike due to the tremendous magnitude and variety of goods bound to international trade.

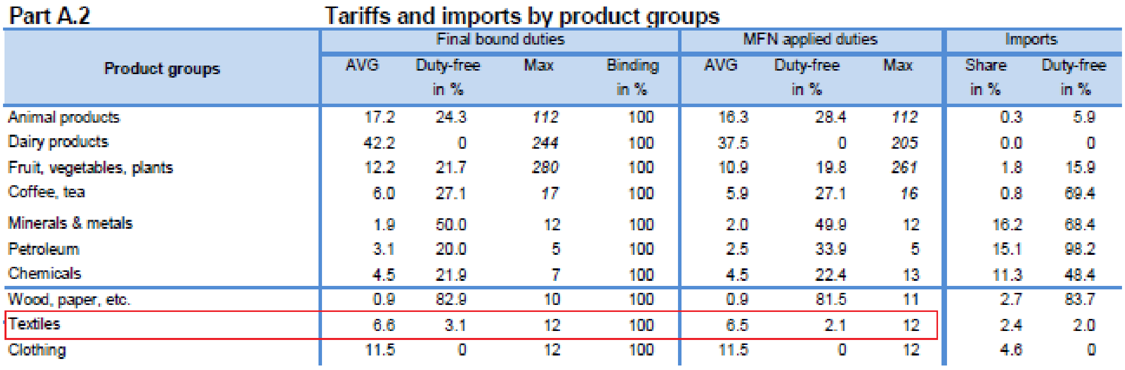


Product Group	MTN	Harmonized System nomenclature 2017
Non-agricultural products (Non-Ag)		
Textiles	x2	300590, 330620, 392112-13, 392190, 420212, 420222, 420232, 420292, Ch. 50-60 (except 5001-03, 5101-03, 5201-03, 5301-02), Ch. 63, 640520, 640610, 6501-05, 6601, 701911-19, 701940-59, 870821, 8804, 911390, 940490, 961210

Figure 25 - An example from the world tariff profiles as proposed by the WTO (n.d.), representing the textile commodity group and the Harmonized System chapters, headings and subheading that are bound to it. Customs duty tariffs for textile-based products can only apply to the above specified HS codes and bear the same implications for all nations that have adopted the nomenclature.

In the case of the European Union, where the CN has been established and deployed as the EU standard for classifying trade goods, the fact that "the six-digit number generated according to the above model is extended by two more digits" supports the statement that a CN number is a mere extension of an HS code for a given good (Wegner, 2018). Subsequently, this observation does underline the renowned nature of the HS as a whole as well as positioning it in the core of activities carried out by international customs associated organizations. Proper handling of the HS is in the forefront of tariff associated activities, and as such it can be regarded as the single most important tariff classification standard.

Fortunately, the WTO along with ITC and UNCTAD, as perhaps previously mentioned, publish annual documentations of the world tariff profiles which happen to have statistical implications. Particularly important for the establishment of the associated customs duty evaluation utility for the given artefact of this paper (HSCODESYS), is the most lately published tariff profile of the EU itself. It is further emphasized that "the statistics related to applied tariffs and imports are calculated using data which are based on the HS nomenclature adopted by the country for the reference year" (WTO, 2020). Additionally, it must be pointed out that the 2020 publication adhering to the world tariff profiles makes use of the 2017 Harmonized System nomenclature binding, which is to be used as the practical implementation example for the HSCODESYS.



Part A.2 Tariffs and imports by product groups

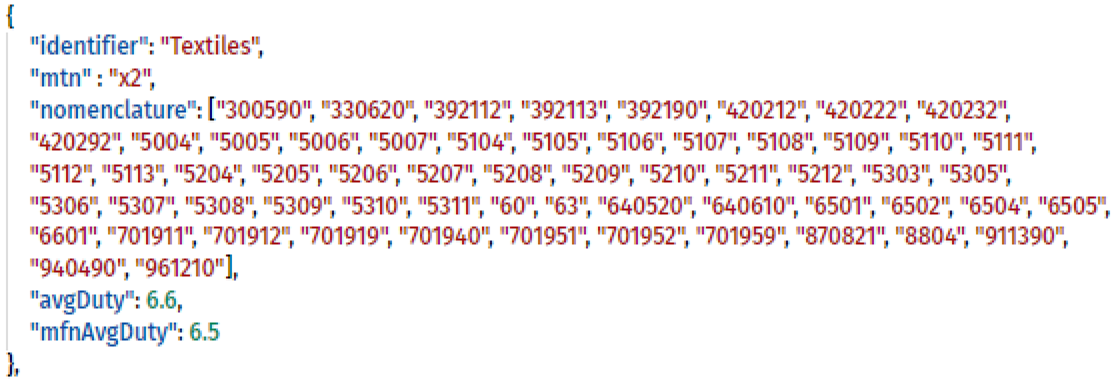
Product groups	Final bound duties				MFN applied duties			Imports	
	AVG	Duty-free in %	Max	Binding in %	AVG	Duty-free in %	Max	Share in %	Duty-free in %
Animal products	17.2	24.3	112	100	16.3	28.4	112	0.3	5.9
Dairy products	42.2	0	244	100	37.5	0	205	0.0	0
Fruit, vegetables, plants	12.2	21.7	280	100	10.9	19.8	261	1.8	15.9
Coffee, tea	6.0	27.1	17	100	5.9	27.1	16	0.8	69.4
Minerals & metals	1.9	50.0	12	100	2.0	49.9	12	16.2	68.4
Petroleum	3.1	20.0	5	100	2.5	33.9	5	15.1	98.2
Chemicals	4.5	21.9	7	100	4.5	22.4	13	11.3	48.4
Wood, paper, etc.	0.9	82.9	10	100	0.9	81.5	11	2.7	83.7
Textiles	6.6	3.1	12	100	6.5	2.1	12	2.4	2.0
Clothing	11.5	0	12	100	11.5	0	12	4.6	0

Figure 26 - A table representing final bound duty tariffs that apply to respective product groups subject to importation in the European Union (WTO, 2020).

By replicating the data structure depicted by the given tariff profile under a programmable data format, the underlying constraint logic of the artefact that would support such a classification functionality can be utilized. More specifically, the EU tariff profile for the year 2020 is replicated under the form of a JSON file which holds the specified data structure of tariffs associated with how customs duty rates on all described chapters, headings and subheadings of the Harmonized System can be classified accordingly.

Conclusively, the way in which the JSON structure representing the given tariff profile is iterated through refers to a recursive functional loop whenever an HS code is being mined through the web application. Respectively, this function-based activity serves the purpose of satisfying the deployed constraint logic of the server script for finding the adhering average duty rate of a given commodity.

Subsequently, by replicating all necessary tables such as “tariffs and imports by product groups” for the European Union as well as “Definition of groups used in part A.2” under the form of a JSON file, the envisioned functionality for customs duty evaluation can be respectively implemented for programmatic usage (WTO, n.d.).



```
{
  "identifier": "Textiles",
  "mtn" : "x2",
  "nomenclature": ["300590", "330620", "392112", "392113", "392190", "420212", "420222", "420232",
  "420292", "5004", "5005", "5006", "5007", "5104", "5105", "5106", "5107", "5108", "5109", "5110", "5111",
  "5112", "5113", "5204", "5205", "5206", "5207", "5208", "5209", "5210", "5211", "5212", "5303", "5305",
  "5306", "5307", "5308", "5309", "5310", "5311", "60", "63", "640520", "640610", "6501", "6502", "6504", "6505",
  "6601", "701911", "701912", "701919", "701940", "701951", "701952", "701959", "870821", "8804", "911390",
  "940490", "961210"],
  "avgDuty": 6.6,
  "mfnAvgDuty": 6.5
},
```

Figure 27 - A JSON object, representing the textile commodity group example and its binding HS code ranges as replicated from the world tariff profiles documentation (Figure 25).

6 Evaluation

This chapter aims to emphasise the evaluation approach that is to be undertaken for validating the proposed utility features of the HSCODESYS implied through its implementation and deployment. For this purpose, the validation perspective that follows to be pursued, is one that entails the consideration for satisfactory validation of all depicted utility features in Figure 22. As a result, the validation techniques and methods used for the evaluation objective might be of varying types, leading to the establishment of a mixed evaluation approach.

6.1 Utility Validation

The artefact's utility evaluation as deliberately envisioned and perhaps already mentioned, is to be carried out with the use of non-identical validation techniques that imply validation of specific aspects and matters associated with the constructed prototype such as performance, design, and user acceptance. By addressing more than a single utility concerning principle, a multifaceted validation approach can be established contributing for an overall more encompassing, diversified, and pervading validation that can be trusted. This chapter section will focus on addressing each validation angle comprising the depicted approach as also shown on Figure 28.

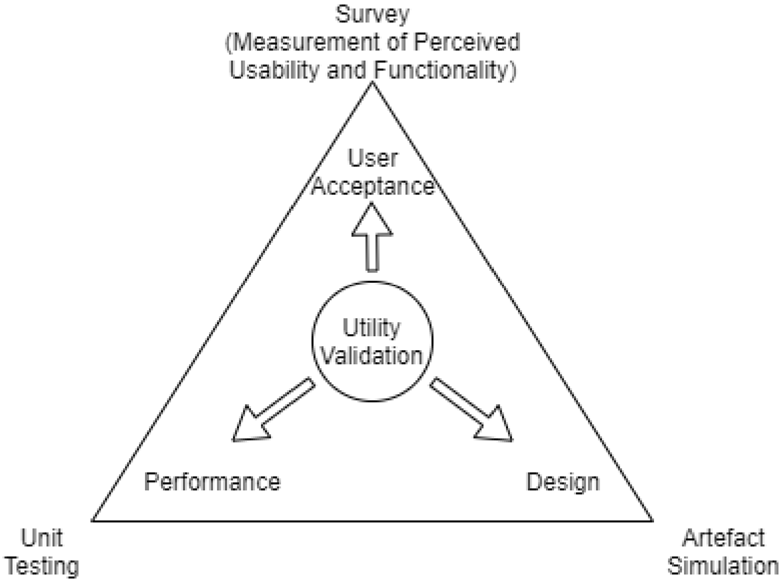


Figure 28 - A triangulated utility validation approach incorporating three evaluation aspects.

6.1.1 Survey Results

The purpose of drafting and making use of a survey, particular to this thesis, is one that addressed the accompanying needs derived from creating the proposed artefact at first hand, which refers to utility validation and overall evaluation. Respectively, this section is to address all matters revolving around the involved survey including emphasis of tools used for creation, reasoning behind incorporated questions and targeted audience.

The creation process that this survey entails, is one that makes use of a specific platform needed for the initial creation of the survey as well as deployment, hosting and accessibility by the intended target audience. Respectively, the platform identified as suitable and appropriate for this task is Typeform, which as emphasised on Wikipedia (n.d.), is a privately held SaaS (software as a service) company which "specializes in online form building and online surveys". Typeform stands as the platform of choice, due to its distinctiveness and reliability along with its engaging and easy to use interface.

Continuously, the survey that has been drafted through the use of Typeform, is one that is comprised of 10 survey questions. Half of the intended questions have been created so as to associate and address important factors such as features, capabilities, characteristics, and utilities specific to the developed by this thesis solution. The other half of involved questions aims to capture participant opinion and perception on important classification matters in cross-border trade, so as to frame participant understandings, correlated to the objectives of this thesis. The nature of the involved questions and the responses that they aim to capture, entail the conclusion for this survey to be for the most part qualitative, with a single quantitative question.

In addition, it is important to elaborate upon the target audience for which this survey is intended. The involved participants who have taken part in this survey mostly refer to individuals who are part of the digital transformation field in information technology, with varying specialties and backgrounds. Among the involved individuals, some of them can be distinguishably identified as software engineers/ developers, data analysts, web designers, academia researchers and postgraduate students.

Before addressing the summarized responses of the survey, it is perhaps important to stress the significance of the survey yet again in accordance to assessing the produced artefact and affirming its utility features along with perceived usefulness among the involved audience. The purpose of HSCODESYS specific questions has to do with the measurement of perceived usability and functionality, as constituted on Figure 28, where the user acceptance factor can be concluded from analyzing HSCODESYS specific survey questions, while non-specific to the artefact questions deal with levels of participant familiarity and awareness regarding the broader research matter.

The summarized results that are about to be elaborated are concluded through the successful completion of the intended survey by the involving 24 participants. It is perhaps important, to start off the survey summary review by addressing the results of involved questions that may not be explicitly related to the artefact utilities. These questions deal with objective assumptions and predispositions that might be already existent or not, such as the use of electronic data processing techniques by EU based customs organizations for commodity classification, proper taxation, and the use of internationally renowned nomenclatures.

The first question which deals with capturing the participants perception and understanding of matters that drive the conduction of this research, refers to if whether electronic data processing techniques are being deployed and used through EU customs authorities. Consequently, 12 participants (50%) have answered with “To a lesser extent”, 11 (45.8%) have chosen the option of “To a greater extent”, 1 (4.2%) has answered with “Most definitely” and 0 have chosen the “Not at all” option. The conclusion that can be drawn from these results implies that the sample can be divided to two groups, each one of which believes such processing techniques are being deployed and used, regardless of rate and magnitude.

The second state of the art question asks from participants, if misclassification of trade goods by customs authorities can lead to wrongfully taxed goods, to which 21 (87.5%) have answered with “Yes”, while 3 (12.5%) have chosen the “No” option. Another one of the non-related to the HSCODESYS functionalities questions asks of participants if they “find it useful to be able to classify trade goods through electronic data processing means?”, to which 14 (58.3%) of participants have answered with “Very Useful”, 9 (37.5%) having chosen the “Somewhat useful” option, 1 (4.2%) standing for “Not that useful”, and 0 for the option of “Not useful at all”.

Lastly, the final question that deals with correct classification of trade goods based on international standards, asks of participants to point out based on a scale of 0-10 how important do they think it is “to classify trade goods based on internationally renowned and adopted standards, nomenclatures and protocols?”.

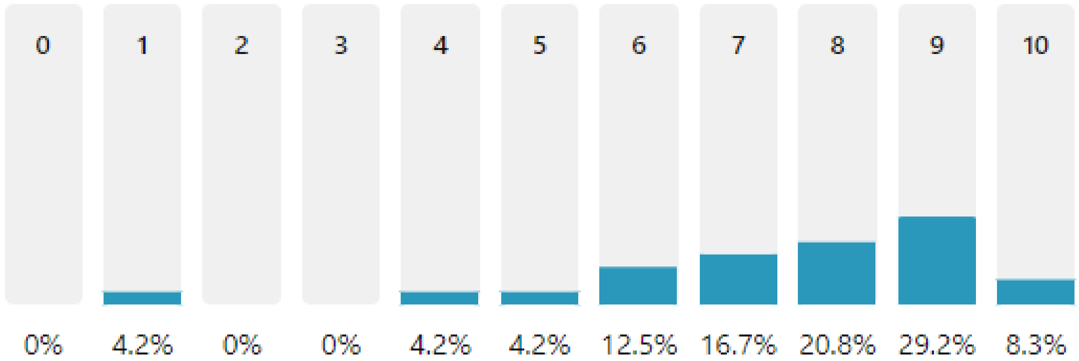


Figure 29 - 0 to 10 scale capturing survey participant perception on the importance of international trade nomenclatures.

As shown on Figure 29, most of the survey participants (7 respondents for 9/10, 5 respondents for 8/10, 4 respondents for 7/10, 3 respondents for 6/10 and 2 respondents for 10/10) do believe that abiding to internationally adopted nomenclatures is important in the sphere of cross-border commodity trading with only 3 respondents (option 1/10, 4/10 and 5/10 respectively) expressing disagreement. This survey question is of particular significance for affirming the need to use renowned standards such as the HS and CN, which subsequently encourages their adoption in various solutions while particularly upholding their utility and importance throughout the development context of the HSCODESYS artefact.

When posed with the questions of "How do you find the performance and accuracy of the traversal process given the magnitude of records used?", 12 (50%) participants reported "Good", 9 (37.5%) reported "Very Good", 3 (12.5%) having chosen "Medium" and 0 for "Bad", given the already mentioned sample size. Subsequently, when presented with the question of "Does the HSCODESYS provide an assured sense of classifying trade goods?", 11 (45.8%) of the same sample answered with "Somewhat", 10 (41.7%) have reported "Agree" and 3 (12.5%) having chosen the "I would like to see more functionality" answer.

Nevertheless, the survey results that can be concluded show a positive perception of the web tool with existing room for greater advancements and deployment of functionality. Any compromise of functionality and inability to conclude an alpha version of the artefact are due to the complex and multifaceted research approach along with the limited period of time present at hand for carrying out this research and the time costly process of designing, engineering, and implementing the presented solution. Nonetheless, the fundamental grounds for algorithm instantiation and application in web software environments established and proposed by this thesis can further serve as essential basis for conducting research of greater magnitude and scale in the same direction.

In terms of affirming the delivery of envisioned design and ease of use results, when presented with the questions of "Is the HSCODESYS web application easy to use?", 19 (79.2%) of the involved participants answered with "Yes", while 5 (20.8%) have answered with "No". By observing the answers to this survey question, the perceived ease of use for the artefact can be sufficed as satisfactory and positive, with the majority of respondent feedback supporting this claim.

Subsequently, when posed with the questions of “How do you find the design and styling of the HSCODESYS?”, 16 (66.7%) participants responded with “Concise and just right”, 5 (20.8%) responded with “Complex”, and 3 (12.5%) responded with “Too simple”. The way design gets perceived can very much be a subjective matter, which is why the gathering of supportive means like the ones produced from the used feedback survey can serve the affirmation of a statement, in this case, the perceived conciseness of the artefact. Additionally, when asked if the degree of sufficiency and descriptiveness of data displayed on the artefact’s UI is seen as satisfactory, 18 (78.3%) respondents have chosen “Yes”, while 5 (21.7%) have decided to choose the “No” option.

In terms of validating the usefulness of the deployed VAT and Customs Duty estimation functionality present in the artefact, by means of participative testing carried out on the simulated artefact, survey results underline that 20 (83.3%) participants out of 24 responded with “It is a useful feature”, 3 (12.5%) chose the option of “It needs further development”, and 1 (4.2%) decided that “It is redundant”.

To briefly conclude the results produced by this survey, it can be said that the expected feedback needed for the constructive and appropriate evaluation of the artefact’s user acceptance concern has been provided accordingly. The overall qualitative analysis associated with evaluating utilities embodied by the artefact and their user approved value further contribute for the proper carriage of the triangulated utility validation approach specified throughout Section 6.1 and Figure 28. By referring to the information that has been presented here, a confidential elaboration on the permissibility of deliverables produced by the used DSR methodology can be further supported.

6.1.2 Artefact Simulation

Before emphasising how simulation of the proposed and developed by this thesis artefact, can be used for evaluating the perceivable design that has been established for it, a distinctive elaboration on the matter of simulation in the context of complex design artefacts is to be laid out. Nguyen, Eiring, and Poo (2018), argue that “Design science artifacts can be highly sophisticated due to the substantial complexity of their situated organizational structures, user characteristics, and technological infrastructure.”, which seemingly happens to be the case for the involved artefact of this study. It is then further expressed that, “Evaluation of complex design science artifacts is challenging; because they may involve long-term experimental procedures, impose uncontrollable impacts to users or environments, entail multiple role structures and diverse contexts, or require human participation in true settings” (Nguyen et al., 2018).

Since the HSCODESYS entails a diverse plethora of functional capabilities that revolve around providing certain utilities, it can be regarded as the initial version of a system that aims to exhibit complex characteristics and features. The sophisticated architectural infrastructure that underlies the web application along with its diverse set of deployed design elements suggest the requirement and use of a simulation approach that resembles real situation-based environments for artefacts that are already developed and implemented, respectively. That is the reason why, the in-situ simulation approach as mentioned and elaborated by Nguyen et al. (2018), stands out as a confidential and visibly as well as evidentially provable method for the accompanying validation strategy that is ought to be carried out.

For the sake of conducting the mentioned simulation, the artefact, in this case HSCODESYS, in accordance with the remarks expressed by Aggarwal et al. (2010), intends on making use of human subjects, as well as computer-based agents, making up a hybrid validation scenario. This process aims to deliberately amplify the user experience that is derivable through direct interactive means exhibited by the UI at hand. As elaborated by Nguyen et al. (2018), artefacts resulting from design science research such as the one presented here, prioritize situated utility, therefore calling for “simulation in real situations, called as in-situ simulation”.

It is then additionally comprehended that this simulation type may “enable ex-post consideration of performance measures, safety assessment, modelling natural or human systems, examining effects of alternative design suggestions and courses of actions when a naturalistic experimental design is not practicable.”, with the last underlined observation being specifically adhering to the case of this paper (Nguyen et al., 2018). Consequently, Nguyen et al. (2018), proceed with further pointing out examples of when in-situ simulation is most suitable, with the most corresponding ones being that “the evaluation of design artifacts requires experimental procedures that run a long period of time” as well as the one entailing that “the realization of design artifacts requires human participation and expertise, which cannot be done with computer-based simulation techniques to replicate some properties of the true settings.”.

The methodological steps to be taken for conducting the simulation will be those described in the paper published by Nguyen et al. (2018), which refer to “planning, simulation, data collection, analysis and reporting”. By following this methodology, a streamlined process consisting of consecutive steps can properly assure the successful production of simulation-based outcomes needed by the utility validation of the proposed artefact. The consecutive steps that make up the methodological simulation approach consist of the above pointed out steps, with planning being resembled by the requirements elicitation phase of the HSCODESYS and the drafting of the intended development life cycle along with design requirements and constraints.

Subsequently, simulation becomes possible by the portrayal of various testing method on the developed beta version of the artefact with those being unit testing, user surveying, design requirements and constraint verification. As follows, the third step of this chained methodological approach is data collection, which is for the most part backed up by the survey and unit tests results derived from the previously mentioned simulation phase. Lastly, the analysis and reporting steps, can be justifiably perceived as the validated utility traits of the artefact and their evidentially proven value as a result of survey and test data drawn from the DSR evaluation method at hand.

Nguyen et al. (2018), further argue that there exist several simulation types, with the most renowned ones referring to those that are either computer, or human-based. Computer-based simulating, as pointed out by Prat, Comnyn-Wattiau and Akoka (2015), stands as the most preferred technique, documented in over 30 percent of DSR related papers. Human-based simulation, as elaborated by Ammenwerth et al. (2012) and Dieckmann, Clemmensen, Sørensen, Kunstek, and Hellebek (2016), opposite of the computer-based simulation type, constitutes the “human-in-the-loop” factor, which resembles user interactions, deemed irreplaceable by “agent-based automators”. Nevertheless, the mutual sense of agreement between DSR academicians on how design science needs to be approached, with simulation accompanying phases such as “identifying problems, design, development, evaluation, and knowledge abstraction” stands as valid and evidentially renowned. (Nguyen et al., 2018).

Furthermore, Nguyen et al. (2018) point out, that in order for the simulation step of the in-situ process to bear best results, it has to underline clearly defined goals along with a well perceivable level of correspondence and precision presented to test prospects that may interact with the artefact, so as to provoke cognitive perceptiveness during the stated simulation. This can be regarded as a crucial step when acquainting participants with the design and functionality of the artefact, for the sake of harnessing proper provision of inputs that may be needed, smoother interaction experience along with valid and rigorous simulation results.

Lastly, the testable hypotheses associated with the artefact that this thesis aims to evaluate, are for the most part defined and described in Chapter 5.3. addressing the utility tree for the HSCODESYS. As perhaps elaborated, such hypotheses can be regarded as "elements that can be validated to determine whether the design artifacts met the meta-requirements, developed in the design phase of the DSR lifecycle" (Nguyen et al., 2018). To conclude, this section stresses the validation significance of utilizing the in-situ simulation approach throughout the depicted DSR life cycle that the involved artefact is to be evaluated by.

By following the pointed-out steps of the simulation methodology at hand, the consecutive processes of drafting proper requirements and design constraints aim to establish all designated utility features. Subsequently, the tasks of collecting data, analysing, and reporting become presupposed, in accordance with the type of simulation as mentioned previously, which could be computer/ human-based, or both (hybrid). This particular disposition also supports the utilization of the triangulated utility validation approach as depicted on Figure 28, by making use of a hybrid simulation environment where computer-based agents as well as human participants all contribute for the overall validation course of the artefact.

6.1.3 Unit Testing

When it comes down to affirming software related functionalities of applications and programs in the sphere of web development, notable and well accepted types of testing are evidently utilized more often than not. Web application testing is presented as a "software testing technique exclusively adopted to test the applications that are hosted on web in which the application interfaces and other functionalities are tested" (Jaleel, 2019). As emphasised on the Guru99 website (n.d.), there exist different types of unit test levels applicable to software with the most notable ones referring to unit, integration, system, and acceptance testing.

Table 7 - Table of Unit Tests performed on the HSCODESYS web tool.

Function Name	Function Description	Test Instances	Test Input	Test Output	Status
getVAT()	Return VAT rate based on country code and name	10	Country Code Country Name	VAT Rate	(10/10) Successful
getDescription()	Returns lowest-level description of an HS Code	15	HS Code	HS Subheading Description	(15/15) Successful
getSectionNote()	Return the section note of an HS Code	15	HS Code	HS Heading Description	(15/15) Successful
getCN()	Returns most identical and adhering CN code based on HS code provided	15	HS Code	Combined Nomenclature Code	(15/15) Successful

Since this section aims to address the performance validation of the depicted application prototype, the underlying functional effectiveness and successful execution of source code components/ units (functions) is of utmost regard. Such an elaboration would constitute the utilization of unit testing, or otherwise referred to as component testing. This has been described by Sreeraman (n.d.), in congruence with the ISTQB – International Software Testing Qualification Board specification, as a software testing approach concerned with individual pieces of code described as units, the results of which can bear potential for validating software performance and functionality. Consequently, it is also mentioned that "unit testing has been believed to be one of the pillars of code quality over a long period of time" (Gren & Antinyan, 2019). Based on the expressed observations and statements, unit testing is to be utilized with the use of the already mentioned in Chapter 5 Mocha testing framework for NodeJS as to assure the functionality of underlying components responsible for utility features exhibited by the proposed artefact.

As displayed on Table 7, four major units responsible for deliberate functionality have been tested accordingly. The results of the conducted tests show that, each functionality performs as intended based on specified input/ output variables used within the tested constraint logic. Respectively, these script-based methods that have been tested underlie the artefact’s utility features as shown on Figure 22. The conclusion that each one of the utility associated functions performs as expected and is accompanied by realized simulation and documentation concretize the fulfilment of the initially foreseen and probable utility capacities.

7 Conclusion & Future Work

The purpose of this chapter is to conclude this thesis, by establishing an overview of this research which emphasises all major milestones and objectives posed along the way. Section 7.1 of this chapter focuses on providing a comprehensive but summarized recapitulation of this thesis while subsequently being followed by Section 7.2 of this paper, which aims to present a run-through reference of answers provided for each one of the research questions as laid out in Chapter 1. Respectively, Section 7.3 of this Chapter presents the limitations that revolve around the involving research work while stressing observations and remarks that can be culminated as a result. Lastly, Section 7.4 bears a highlight of any future work that can arise or evolve from the premise and foregrounds established by this thesis.

7.1 Conclusion

This research deals with the inspection and analysis of a software intended machine learning algorithm (ID3) use case suitable for classification of trade goods in the EU, based on internationally applied taxonomical standards in information systems. The Harmonized System and Combined Nomenclature are two very good examples of globally adopted classification blueprints responsible for labelling, tariffs, category-based listings, and stand as a core component needed by the arguments this thesis proposes. The experimentation and probing of such an algorithm suitable for the specified tasks, is made possible by engaging in a software development lifecycle. The SDLC incorporated in this research intends on producing a working solution, in this case an artefact named HSCODESYS, that is capable of utilizing the ID3 algorithm for successful classification of trade commodities in accordance with the HS and CN.

A literature overview of the need for proper classification of trade goods subject to the internal or external borders of the EU is provided with the intent of stressing tax and duty associated revenue losses currently occurring, as well as fraudulent practices that might be evermore present in the ever-growing e-commerce sphere. The importance of abiding to the novel legislation of the European Commission concerning the Taxation and Customs Union is also elaborated, with a crucial emphasis on the importance of the establishment and deployment of electronic data processing techniques for customs-associated institutional bodies throughout the EU.

The creation process of the artefact revolving around its crafting and envisioning is a process substantially influenced and governed by the Design Science Research (DSR) methodology which intends on supporting the development process behind artefacts such as the HSCODESYS, by using justifiable and empirical means of research motives such as design theory and knowledge delivery through scientific rigour. Upon successful realization of the artefact, its utility, as perhaps specified throughout this thesis, is to be evaluated in accordance with renowned DSR evaluation method deemed most suitable along with supportive validation procedures such as testing, simulation, and surveying.

The works of this thesis are then summarized with an overview, addressing the results obtained from the successful conduction of the documented DSR evaluation method (Figure 21) and the triangulated utility validation (Figure 28), which will serve the overall value approval of the HSCODESYS. Lastly, this elaboration is to ensure the adhering congruence of the envisioned and expected utility with that which was documented and assessed after the associated development or realization has occurred.

7.2 Research Questions

This thesis does not make use of a distilled meta question but rather a set of three distinct main questions that may or may not have respective sub-questions associated with them. Comprehensive answers to each research question can be found within the contents of sections responsible for delivering the associated answers and justifications. This section rather focuses on the brief reference of the aforementioned question answers by establishing finalized conclusive statements.

7.2.1 RQ-1: How can the specialized application of renowned commodity nomenclatures & ML web technologies in information systems contribute towards more competent classification practices on trade goods throughout the EU?

In order to comprehend the summarized answer for this research question, the cross-reference of answers provided to the related sub-questions in Chapter 4.1 should be noted. Competent classification practices revolving around the identification of certain commodity codes such as Harmonized System and Combined Nomenclature codes, as seen throughout this thesis, require a multi-layered solution that makes use of multi-faceted technology deployment. By planning, designing, and developing a working solution example such as the HSCODESYS, the competency level and effectiveness of embodied functionalities proclaiming to benefit the overall process of conducting classification on commodities can be evaluated correspondingly.

By implementing specialized and tailor-made solutions like the one subjected by this thesis, unresolved issues and inconveniences pertaining throughout the classification practices of respective organizations can be elaborated so as to envision and realize working solutions. By observing the concluded state of the artefact along with validation and evaluation results, it can be argued that the intertwining of open data (HS, CN) and ML technologies do contribute for the establishment of competent and effective classification practices in information systems. Specialized application of correlated technologies, such as the ones used by the HSCODESYS, present the capability of overcoming marginal classification with non-diversified data.

Lastly, the significance of using renowned commodity classification nomenclatures such as the HS and CN for engagement in diversified classification activities can be observed and justified through the usage of the HSCODESYS artefact. Utility matters such as VAT and Customs tax evaluation are not possible without the abidance to and utilization of the mentioned nomenclature standards. The preference and option of using decision tree machine learning algorithms like the ID3 on the other hand, present possibilities for establishing well performing and effective data processing capabilities for web-based information systems.

7.2.2 RQ-2: Can the findings/ answers to RQ1 support the proposal of a web technology deployment model for the adoption of commodity classification mining in web applications?

The answer to RQ1 is based on the proven and validated use case of commodity classification nomenclatures under the form of open data formats (JSON, CSV) and the use of distributable server scripting package named decision-tree adhering to a scripted implementation of the ID3 algorithm for usage in server applications. The illations and findings implied by the answer to RQ1, serve as justifiable and research backed reasoning valuable for the underpinning foundation to be used for answering RQ2.

The simulated outcome along with the involved engineering cycle used for realizing the HSCODESYS artefact are the basis for proposing its underlying architectural framework as a web technology deployment model usable by similar use cases. As a result, the NCMM (Nomenclature Classification Mining Model) as depicted on Figure 18 in Chapter 4.3.3 of this thesis, can serve the purpose of presenting itself as a blueprint suitable for adoption by any information system striving to establish classification utilities and functionalities similar to those of the HSCODESYS.

7.2.3 RQ-3: How to evaluate the utility of the HSCODESYS?

In order to comprehend the summarized answer for this research question, the cross-reference of answers provided to the related sub-questions in Chapter 5.1 should be noted. The artefact of this thesis is one that is created according to DSR, which on its own suggest the due diligent and vigorous evaluation of proclaimed artefact utilities so as to concretely affirm any value or utility ought to be delivered. This thesis makes use of a chapter devoted to representing any embodied or suggestable utilities that can be evaluated by DSR evaluation means.

For this purpose, a utility tree, as shown on Figure 22, is constructed for visually conceivable and representational means. Furthermore, Chapter 5 along with Chapter 6 comprehensively stress the matter of evaluating utilities by abiding to the most suitable DSR evaluation method for the artefact at hand, as depicted on Figure 21. Lastly, Figure 28 of this thesis displays the second evaluation layer or perhaps method, of a triangulated type (making use of three evaluation techniques), important for further validating the outcomes posed by the artefact production.

To conclude, utilities presented by the HSCODESYS can be evaluated by the DSR evaluation method and triangulated validation approach used by this thesis. The referred methods focus on simulating and testing an artificial prototype after its creation (ex-post) for summative means such as instantiation testing (ID3 algorithm) and model deriving (NCMM).

7.3 Limitations

Limitations of the concluded research results presented by this thesis are not an exception, and those of most significance for the envisioned outcomes of this thesis will be elaborated. One of the more affecting limitations is that associated with the lack of available stakeholders, needed for gathering feedback inputs from individuals working in customs authorities. The involvement of such stakeholders would have assured the pragmatic authority-based attestation of the web tool created for this research.

Despite the persistent effort of reaching out and contacting such prospectus employees, the response and reach out can be considered of insignificant value resulting to no feedback or discussion whatsoever. This matter is most likely associated with the Covid-19 pandemic and all social complications revolving it at present, which as a result encouraged the more extensive involvement in mixed evaluation techniques. The feedback deficiency described above respectively led to validating the outcomes of this thesis, with greater emphasis on production of valuable quantitative data resulting from artificial simulation means. The usage of survey feedback from digital governance and transformation practicing individuals intends on filling the resulting gap partially but not completely, which is why additional methods such as testing, and simulation are also used.

Lastly, the performance of the instantiated ID3 algorithm for the working solution presented by this thesis, can be comparatively bounded depending on the size of datasets being used. By observing the performance results obtained from the extensive testing and simulation of the HSCODESYS artefact, it can be noted that deployable datasets of large magnitudes with more than fifty thousand records can result in delayed traversal and querying processes. Owing to this limitation, performance boundaries can be tackled with the extensive trimming of involved datasets with the consideration of preserving the most valuable and use-case necessary records of a dataset ought to be used.

7.4 Future Work

The established and presented works of this thesis are subject to involving concerns and matters that can be addressed with further developments and extensions to the currently present results. The multifaceted undertake of varying functional capabilities displayed by the artefact of this thesis present a wide spectrum of possible additions and improvement of existing utilities.

Such an addition could be the integration of TARIC (Integrated Tariff of the European Union) for the creation of more encompassing classification capabilities of the web application in terms of EU standards and regulatory requirements. Additionally, the works of this research might provoke interest towards the greater appendance of ML capacities by using novel AI technologies such as recurrent neural networks or convolutional neural networks.

The presented web tool solution of this thesis also holds potential for the extension of functionalities in terms of customs associated operations with an emphasis on the creation of customs declarations for consignments adhering to the used nomenclature standards. The ability of drafting and producing customs declarations on the go for commerce consignments of commodities classified by the HSCODESYS stands as a captivating capability, which seems to attract more demand throughout the spectrum of information system potentialities for customs and logistics operations.

References

Falk, M., & Hagsten, E. (2015). E-commerce trends and impacts across Europe. International Journal Of Production Economics, 170, 357-369. doi: 10.1016/j.ijpe.2015.10.003

Reinsch, R. (2005). E-commerce: managing the legal risks. Managerial Law, 47(1/2), 168-196. doi: 10.1108/03090550510771377

Holsapple, C., & Singh, M. (2000). Toward a unified view of electronic commerce, electronic business, and collaborative commerce: a knowledge management approach. Knowledge And Process Management, 7(3), 151-164. doi: 10.1002/1099-1441(200007/09)7:3<151::aid-kpm83>3.0.co;2-u

European Parliament Policy Department D for Budgetary Affairs. (2019). Protection of EU financial interest on customs and VAT: Cooperation of national tax and customs authorities to prevent fraud

REGULATION (EU) No 952/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 9 October 2013 Laying down the Union Customs Code. (2013). OJ 02013R0952

Council of the European Union (2019). Commission Implementing Decision (EU) 2019/2151 of 13 December 2019 establishing the work programme relating to the development and deployment of the electronic systems provided for in the Union Customs Code. OJ L325/168

Business Domains. (2020). Retrieved from https://cyber.ee/competences/business-domains/#customs-systems

Solution. (2020). Retrieved from https://eurora.com/solution

Krishnan, A., & Amarthaluri, A. (2019). Large Scale Product Categorization using Structured and Unstructured Attributes. ArXiv, abs/1903.04254.

Ding, L., Fan, Z., & Chen, D. (2015). Auto-Categorization of HS Code Using Background Net Approach. Procedia Computer Science, 60, 1462-1471. doi: 10.1016/j.procs.2015.08.224

Vijendra, S., Parashar, H. and Vasudeva, N. (2011). A New Method for Classification of Datasets for Data Mining. 3rd International Conference on Machine Learning and Computing (ICMLC 2011), [online] Available at: <https://arxiv.org/ftp/arxiv/papers/1612/1612.00151.pdf>

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75. doi: 10.2307/25148625

Wieringa, R. (2014). Design Science Methodology for Information Systems and Software Engineering. doi: 10.1007/978-3-662-43839-8

Council of the European Union. (2020). Commission Implementing Regulation (EU) 2020/1577 of 21 September 2020. Official Journal of the European Union. [online] Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2020:361:FULL&from=EN>

European Commission. (n.d.). Classification of goods - Taxation and Customs Union. Retrieved from https://ec.europa.eu/taxation_customs/business/calculation-customs-duties/what-is-common-customs-tariff/classification-goods_en.

European Commission. (n.d.). Union Customs Code - Taxation and Customs Union. Retrieved from https://ec.europa.eu/taxation_customs/business/union-customs-code_en.

European Commission. (n.d.). The Combined Nomenclature - Taxation and Customs Union. Retrieved from https://ec.europa.eu/taxation_customs/business/calculation-customs-duties/what-is-common-customs-tariff/combined-nomenclature_en.

2021 VAT Rates in Europe. (2021). Retrieved from https://taxfoundation.org/value-added-tax-2021-vat-rates-in-europe/.

Aggarwal, C.C. (Ed.). (2014). Data Classification Algorithms and Applications. 1st ed. New York: Chapman and Hall/CRC., pp.1-114. https://doi.org/10.1201/b17320

Howbert, J. (2012). Classification - Basic Concepts, Decision Trees, and Model Evaluation. Introduction to Machine Learning [Adobe Acrobat slides]. Retrieved from http://courses.washington.edu/css490/2012.Winter/lecture_slides/04_classification_basics.pdf.

Tan, P., Steinbach, M. & Kumar, V. (2015). Introduction to Data Mining. 1st ed. Dorling Kindersley: Pearson.

WTO. (2020). World Tariff Profiles 2020. Retrieved from https://www.wto.org/english/res_e/publications_e/world_tariff_profiles20_e.htm

Yevtushenko, A., & Yalanska, M. (2021). Back-End Development. Foundation of Digital Product. Retrieved from https://blog.tubikstudio.com/back-end-development-foundation-of-digital-product/.

MDN Web Docs Glossary. (n.d). MVC: Definitions of Web-related terms. Retrieved from https://developer.mozilla.org/en-US/docs/Glossary/MVC

Rizvi, A. (2010). ID3 Algorithm. [Adobe Acrobat slides]. Retrieved from http://athena.ecs.csus.edu/~mei/177/ID3_Algorithm.pdf.

Alpaydin, E. (2009). Introduction to Machine Learning (2nd ed.). Cambridge: MIT Press.

Wikipedia. (n.d.). ID3 Algorithm. Retrieved from https://en.wikipedia.org/wiki/ID3_algorithm.

Decision Tree for NodeJS. (n.d.). Retrieved from https://www.npmjs.com/package/decision-tree.

Brutus, A., Daniely, A., & Maslach, E. (2019). On the Optimality of Trees Generated by ID3. doi: arXiv:1907.05444

Elkstein, M. (n.d.). Learn REST: A Tutorial. Retrieved from http://rest.elkstein.org/.

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A Comprehensive Framework for Evaluation in Design Science Research. Lecture Notes In Computer Science, 423-438. doi: 10.1007/978-3-642-29863-9_31

Harris, N. (2015). Machine Learning over JSON. Retrieved from https://www.naftaliharris.com/blog/machine-learning-json/

Lv, T., Yan, P., & He, W. (2019). On Massive JSON Data Model and Schema. Journal Of Physics: Conference Series, 1302, 022031. doi: 10.1088/1742-6596/1302/2/022031

Winter, R., Zhao, J., & Aier, S. (2010). Global Perspectives on Design Science Research. Lecture Notes In Computer Science. doi: 10.1007/978-3-642-13335-0

Herwix, A., & Rosenkranz, C. (2018). Making Sense of Design Science in Information Systems Research: Insights from a Systematic Literature Review. Designing For A Digital And Globalized World, 51-66. doi: 10.1007/978-3-319-91800-6_4

Chauhan, N. (2020). Decision Tree Algorithm, Explained - KDnuggets. Retrieved 8 March 2021, from https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html#:~:text=Classification%20is%20a%20two%2Dstep,prediction%20step%2C%20in%20machine%20learning.&text=In%20the%20prediction%20step%2C%20the,algorithms%20to%20understand%20and%20interpret.

Van Casteren, W. (2017). The Waterfall Model and the Agile Methodologies : A comparison by project characteristics. doi: 10.13140/RG.2.2.36825.72805

Li, M., Kok, S., & Tan, L. (2018). Don't Classify, Translate: Multi-Level E-Commerce Product Categorization Via Machine Translation. ArXiv, abs/1812.05774.

Burgin, M. (2003). Information Theory: a Multifaceted Model of Information. Entropy. 5(2):146-160. https://doi.org/10.3390/e5020146

Host, S. (2019)."Information Measures" in Information and Communication Theory, IEEE, pp.37-68, doi: 10.1002/9781119433828.ch3. Retrieved from https://ieeexplore.ieee.org/xpl/ebooks/bookPdfWithBanner.jsp?fileName=8699087.pdf&bkn=8698969&pdfType=chapter.

Gregor, S., & Hevner, A. (2013). Positioning and Presenting Design Science Research for Maximum Impact. MIS Quarterly, 37(2), 337-355. doi: 10.25300/misq/2013/37.2.01

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a Framework for Evaluation in Design Science Research. European Journal Of Information Systems, 25(1), 77-89. doi: 10.1057/ejis.2014.36

Pries-Heje, J., Baskerville, R., Venable, J.R. (2008) Strategies for Design Science Research Evaluation. In: Proceedigns of the 16th European Conference on Information Systems (ECIS 2008), Galway, Ireland

Statistics Denmark. (2021). Retrieved 6 April 2021, from https://www.dst.dk/en/Indberet/oplysningssider/intrastat/udenrigsoekonomi#

European Commission. (2020). VAT Tariff Profiles. Retrieved 7 April 2021, from https://ec.europa.eu/taxation_customs/sites/taxation/files/resources/documents/taxation/vat/how_vat_works/rates/vat_rates_en.pdf

WCO. (n.d.). The Harmonized System: A Universal Language for International Trade. Retrieved 13 April 2021, from http://www.wcoomd.org/-/media/wco/public/global/pdf/topics/nomenclature/activities-and-programmes/30-years-hs/hs-compendium.pdf

Jaleel, H. (2019). Testing Web Applications. International Journal Of Computer Science And Engineering, 6(12), 1-9. doi: 10.14445/23488387/ijcse-v6i12p101

Gren, L., & Antinyan, V., (2019). On the Relation Between Unit Testing and Code Quality. Chalmers University of Technology and the University of Gothenburg. 412-92. doi: arXiv:1904.04748v1

Wegner, T. (2018). Customs Tariff Codes - What you need to know » O&W Rechtsanwälte. Retrieved 22 April 2021, from https://www.owlaw.com/german-customs-law/9746-customs-tariff-codes-what-you-need-to-know/

Nguyen, H., Eiring, Ø., & Poo, D. (2018). In-Situ Simulation in Design Science Research: Evaluation of Complex Design Artifacts. International Conference On Information Systems San Francisco. Retrieved from https://www.researchgate.net/publication/328019662_In-Situ_Simulation_in_Design_Science_Research_Evaluation_of_Complex_Design_Artifacts

Wikipedia. (n.d.). Typeform (service). Retrieved 29 April 2021, from https://en.wikipedia.org/wiki/Typeform_(service)

Spichakova, M., & Haav, H. (2020). Using Machine Learning for Automated Assessment of Misclassification of Goods for Fraud Detection. Communications In Computer And Information Science, 144-158. doi: 10.1007/978-3-030-57672-1_12

Liutkevičius, M., Pappel, K., Butt, S., & Pappel, I. (2020). Automatization of Cross-Border Customs Declaration: Potential and Challenges. Lecture Notes In Computer Science, 96-109. doi: 10.1007/978-3-030-57599-1_8

Rozbroj, R., (2020). The Upcoming EU 2021 VAT E-commerce Package From Consumer Perspective. M.Sc. Thesis. Tallinn University of Technology. Available at: https://www.researchgate.net/publication/344058449_The_Upcoming_EU_2021_VAT_E-_Commerce_Package_From_Consumer_Perspective (Accessed: 6 May 2021)

Van der Hejde, J. (2019). Automated classification tool for e-commerce products and their HS code. M.Sc. Thesis. Eindhoven University of Technology. Available at: https://pure.tue.nl/ws/portalfiles/portal/145366508/Master_Thesis_Jorg_van_der_Heijde.pdf (Accessed: 24 March 2021)

Altaheri, F., & Shaalan, K. (2020). Exploring Machine Learning Models to Predict Harmonized System Code. Information Systems, 291-303. doi: 10.1007/978-3-030-44322-1_22

Li, G., & Li, N. (2019). Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network. Electronic Commerce Research, 19(4), 779-800. doi: 10.1007/s10660-019-09334-x

Simon, H. A. (1996). The Sciences of the Artificial (3rd ed.), MIT Press, Cambridge, MA

Denning, P. J. (1997). A New Social Contract for Research. Communications of the ACM (40:2), pp. 132-134.

Tsichritzis, D. (1998) “The Dynamics of Innovation” in Beyond Calculation: The Next Fifty Years of Computing, P. J. Denning and R. M. Metcalfe (eds.), Copernicus Books, New York, pp. 259-265.

Sjoberg, D., Dyba, T., & Jorgensen, M. (2007). The Future of Empirical Methods in Software Engineering Research. Future Of Software Engineering (FOSE '07). doi: 10.1109/fose.2007.30

Basalisco, D.B., Wahl, J., Okholm, D.H. (2016). e-Commerce imports into Europe: VAT and customs treatment. Copenhagen Economics

Gupta, V., Karnick, H., Bansal, A., & Jhala, P. (2016). Product Classification in E-Commerce using Distributional Semantics. Retrieved from http://arxiv.org/abs/1606.06083.

March, S., & Smith, G. (1995). Design and natural science research on information technology. Decision Support Systems, 15(4), 251-266. doi: 10.1016/0167-9236(94)00041-2

Akoka, J., Comyn-Wattiau, I., Prat, N., & Storey, V.C. (2017). Evaluating Knowledge Types in Design Science Research: An Integrated Framework. International Conference on Design Science Research in Information Systems, pp. 201-217. Springer.

Baskerville, R., Kaul, M., & Storey, V.C. (2015). Genres of Inquiry in Design-Science Research: Justification and Evaluation of Knowledge Production. MIS Quarterly 39, 541-564.

Sein, M.K., Henfridsson, O., Purao, S., Rossi, M., Lindgren, R. (2011). Action Design Research. MIS Quarterly, vol. 35, pp. 37-56.

Aggarwal, R., Mytton, O. T., Derbrew, M., Hananel, D., Heydenburg, M., Issenberg, B., MacAulay, C., Mancini, M. E., Morimoto, T., Soper, N., Ziv, A., and Reznick, R. (2010). Training and Simulation for Patient Safety. Quality and Safety in Health Care (19: Suppl 2), pp. i34–i43.

Prat, N., Comyn-Wattiau, I., and Akoka, J. (2015). A Taxonomy of Evaluation Methods for Information Systems Artifacts. Journal of Management Information Systems (32:3), pp. 229–267.

Ammenwerth, E., Hackl, W. O., Binzer, K., Christoffersen, T. E. H., Jensen, S., Lawton, K., Skjoet, P., and Nohr, C. (2012). Simulation Studies for the Evaluation of Health Information Technologies: Experiences and Results. Health Information Management Journal (41:2), pp. 14–21.

Dieckmann, P., Clemmensen, M. H., Sørensen, T. K., Kunstek, P., and Hellebek, A. (2016). Identifying Facilitators and Barriers for Patient Safety in a Medicine Label Design System Using Patient Simulation and Interviews. Journal of Patient Safety (12:4), pp. 210–222.

Sreeraman, S. (n.d.). ISTQB - Unit Testing & Integration Testing. Retrieved 18 April 2021, from https://www.getsoftwareservice.com/unit-testing-integrated-testing/#:~:text=Unit%20or%20Component%20or%20Module,called%20Component%20or%20Module%20testing.

Guru99. (n.d.). Unit Testing Tutorial. Retrieved 16 April 2021, from https://www.guru99.com/unit-testing guide.html#:~:text=UNIT%20TESTING%20is%20a%20type,an%20application%20by%20the%20developers.

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Serkan Ahmet Koch

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis MODERNIZATION OF COMMODITY CLASSIFICATION PRACTICES IN TRADE FOR EU CUSTOMS WITH MACHINE LEARNING WEB SOLUTIONS, supervised by Eric Blake Jackson
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

10.05.2021

[1] The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – Non-Functional Requirements Elicitation Table

NFR1	Scalability	It is required of the artefact to be scalable in terms of dataset consumption, as it is expected to take more than one JSON dataset as input. This is a crucial requirement that aims to ensure diverse and compatible data streams for all intended ML tasks.
NFR2	Usability	The artefact is expected to be easily usable in terms of UI. Establishing a simple and concise UI that provides all necessary functionality and design principles for navigating through a web application in a cohesive manner is utmost needed.
NFR3	Compatibility	The artefact is expected to be compatible with all modern browsers that utilize JavaScript.
NFR4	Portability	The artefact has to be constructed in a way that would allow for future portable capabilities that can support the migration and portability of the web application from one machine to another.
NFR5	Reliability	The server side of the artefact is expected to operate fairly effectively with no major disruptions or miss communications between server and client. The artefact can be either hosted online or used offline (locally).
NFR6	Manageability	The artefact's media sources (styling, images, etc.) as well as all source codes and script files are expected to be organized in a way that allows for coherent management of the sources. Significant portions of the source code can be commented out for better navigation.
NFR7	Security	No major security requirements revolve around the artefact's functionality and operability. No sensitive user data is expected to be provided by the user, and all data that the app makes use of is pre-defined and concretized.
NFR8	Performance	The artefact is expected to perform fairly well in terms of data retrieval from the server. Loading time for data request should not surpass the 10 second mark, and most less significant data requests should happen under 3-5 seconds.
NFR9	Regulatory	The artefact's designation and domain of operation should be specifically aimed at customs commodity classification in

		accordance with EU law and international conventions like the Harmonized System Convention by the WCO.

Appendix 3 – Functional Requirements Elicitation Table

FR1	Commodity Search	An ordinary input field that accepts HS codes as user input.
FR2	Allocate Button	A button that triggers allocation functions upon click.
FR3	JSON Inputting	A functionality that ensures proper allocation and traversal of JSON records and files.
FR4	Section Allocation	Allocating the corresponding Harmonized System section of a given HS code
FR5	Input string concatenation	Underlying functionality that concatenates HS codes to respective fractions (Chapter, Heading, Subheading, etc.).
FR6	Description Allocation	Allocating commodity descriptions for respective Harmonized System entities based on provided Heading and Subheading (w. ID3).
FR7	Section Note Allocation	Allocation binding section notes (categories) for respective HS codes based on provided Heading and Subheading (w. ID3)
FR8	CN Code Allocation	The functionality of allocating respectively binding Combined Nomenclature codes for corresponding Harmonized System Codes, based on Chapter, Heading and Subheading (w. ID3)
FR9	VAT calculation	Calculating the VAT tax for a given commodity based on chosen EU member state
FR10	EU Member States Dropdown List	An ordinary dropdown that presents the user with all EU Member States to pick from
FR11	Dropdown Option Caching	Caching the chosen by the user EU member state for following sessions with the use of local Storage
FR12	Template Literals Fetching	Fetching POST request responses from the server to the user with the use of EJS (Embedded JavaScript) Template Literals
FR13	Input Field Dynamic Placeholder	Upon clicking any commodity from the provided dropdowns, the user is to be presented with a dynamic

		placeholder positioned inside the input field that signals the beginning of the HS binding
FR14	Hiding HTML Divisions	Upon page loading, it is expected that div (division) tags associated with HS queries are to be hidden from the user
FR15	Displaying HTML Divisions	Upon successful HS queries and respective data allocation, all associated div (division) tags are to be made visible to the user
FR16	Search Loader	Upon clicking the "Allocate" button, the user is to be presented with an intuitive loader for the duration of the allocation
FR17	Null Value Identification & Notification	Identifying null values provided by the user and notifying of inconsistent HS codes
FR18	Customs Duty Allocation (incl. MFN Duty)	The functionality of allocating custom duty tariff rate for a specific commodity (based on provided HS and specified attribute ranges)
FR19	Smart Commodity Dropdowns	Ordinary dropdowns based on an underlying pre-trained neural net that binds each commodity to its respective HS code.
FR20	Nested Commodity Dropdowns	Ordinary dropdowns with the additional feature of having nested child dropdowns.
FR21	Decision Tree Functionality	The implementation of a decision tree functionality based on an existing decision tree classification formula/ algorithm
FR22	PDF Download Button	A button that allows for the download of the index page in PDF format with all of the retrieved and fetched information related to a commodity
FR23	Mining Accuracy Estimation	Upon the initiation of a traversal task, a mining accuracy attribute gets calculated and assigned for fetching to the UI

Appendix 4 – User Acceptance Survey Questions

Question No.	Survey Questions
1	Do you believe electronic processing techniques are deployed and used throughout customs authorities of EU member states?
2	Do you think misclassification of trade goods by customs authorities could lead to wrongfully taxed goods?
3	Do you find it useful to be able to classify trade goods through electronic data processing means?
4	Does the HSCODESYS provide an assured sense of classifying trade goods?
5	Do you find the capability of estimating VAT and Duty for cross-border products on the go useful?
6	How important do you think it is to correctly classify trade goods based on internationally renowned and adopted standards, nomenclatures, and protocols?
7	How do you find the design and styling of the HSCODESYS?
8	Is the HSCODESYS web application easy to use?
9	How do you find the performance and accuracy of the traversal process given the magnitude of records used?
10	Is the information displayed by the UI sufficient and descriptive enough?