

Master Thesis:

Impact of data quality on the prediction success of auto-classification tools

By

Iuliana Vulpe

Student ID: 477418

EXECUTIVE MASTER IN CUSTOMS AND SUPPLY CHAIN COMPLIANCE

MCSCC 2020-2023

Supervisor: Albert Veenstra


Co-reader: Morteza Pourakbar

Preface

The choice of the thesis subject resulted on one hand, from the numerous classification questions and inconsistencies I encountered in my daily work, and on the other hand, from my curiosity in new technologies. I conducted my desk and experimental research with the objective to have a better understanding about the nuances and limitations of an auto-classification process and to confirm or rebut some preconceived notions.

This research would have not been possible without the constant support of my supervisor at RSM, Albert Veenstra. I would like to thank him for sharing his sharp insights and helping me unlock some situations.

I would also like to take the opportunity to thank the co-reader for his thesis, Morteza Pourakbar, for his feedback and advices, as well as the members of the faculty who, by sharing their visions, enabled me to grow and expand my knowledge.

 28.11.23

Executive summary

Nowadays, companies with fast-paced logistics processes and involved in international trade must have streamlined processes to efficiently comply with international trade regulations and customs requirements. This is particularly true when companies must adapt, in a timely manner, to external regulation changes or internal events, such as the launching of a new product line. Product classification, also known as the Harmonized System (HS) of goods at World Customs Organization (WCO) level or as the Combined Nomenclature at EU level, is one of the domains impacted by the above-mentioned changes. This thesis aims at better understanding one specific aspect of the products classification: the impact of data quality on the auto-classification tools prediction scores.

After a familiarization with the theoretical concepts of product classification, two products data sets are analyzed. The first step of the analysis consists in confirming that the auto-classification tool chosen beforehand (based on criteria such as availability and ease of use) is appropriate for the classification of goods with the data sets at our disposal. This is achieved by submitting the unaltered (except formatting) data set to the auto-classification tool and compare the actual classification with the predicted classification. The tool is considered appropriate if its accuracy is above 90% (a subjective threshold) when used with both data sets.

In a second step, the data sets are manipulated in order to introduce errors in the products features (simulating human errors) or to remove products features (simulating an omission). These manipulated datasets are then submitted again to the auto-classification tool and classification indicators such as the accuracy score, the F_1 -score and the confusion matrix are extracted.

The analysis of these classification indicators leads to two main conclusions for the data sets at hand. First, some attributes carrying important commercial information might have no, or a limited, role in the auto-classification process, while the information driving the classification might be concentrated in a few others attributes (the “main material” and the “article sub-type” have been identified based on the data sets used in this research). Second, based on the results of the experiment, it is difficult to determine whether a type of error is worse than another, *i.e.* to determine whether the error of type “incorrect value maintained” would impact more significantly the performance of the auto-classification tool than the errors of type “no value maintained” or vice-versa.

The research of this thesis acts as a proof of concept by describing a methodology that can be easily reused (opensource software, code shared in appendix, easy/ready to use) on different data sets, or adapted to other classification algorithms.

Table of Contents

Preface	ii
Executive summary.....	iii
List of tables	vi
List of figures	viii
List of abbreviations.....	xi
1. Introduction.....	1
1.1 Problem statement.....	2
1.2 Research questions.....	3
2. Review of research literature	4
2.1 Classification codes.....	4
2.1.1 Harmonized System	4
2.1.2 Extensions of the Harmonized System.....	5
2.2 Classification rules.....	8
2.2.1 General Rules for the Interpretation	8
2.2.2 Tariff binding information	11
2.3 Classification expertise	12
2.3.1 Knowledge requirements.....	13
2.3.2 Customs view on auto-classification tools	14
2.4 Classification complexities	15
2.4.1 Nature of product	15
2.4.2 National interpretations.....	17
2.4.3 Master data availability in IT systems.....	19
3. State of the art of classification	21
3.1 K-nearest neighbors	22
3.2 Support Vector Classifier	23
3.3 Random forest.....	25
3.4 Classifier selection	27
4. Research approach.....	28
4.1 Random forest classifier	28
4.1.1 Step-by-step process.....	28
4.1.2 Model evaluation.....	28

4.2	Data Sets	30
4.2.1	Data set “chapter 71”	30
4.2.1	Data set “chapter 62”	32
4.3	Methodology.....	37
4.3.1	Step 1: training of the classifier	37
4.3.2	Step 2: testing of the classifier on initial data set.....	37
4.3.3	Step 3: testing of the classifier on data set with missing value.....	38
4.3.4	Step 4: testing of the classifier on data set with incorrect value	38
5.	Research results and analysis	38
5.1	Data set “chapter 71”	39
5.1.1	Initial data set.....	39
5.1.2	Data set with missing values	41
5.1.3	Data set with incorrect values.....	45
5.2	Data set “chapter 62”	50
5.2.1	Initial data set.....	50
5.2.2	Data set with missing value.....	54
5.2.3	Data set with incorrect value	56
6.	Conclusions.....	58
6.1	Conclusion research questions.....	58
6.2	Contribution for research and practice	60
6.3	Limitations and future research.....	60
7.	Bibliography.....	62
8.	Appendices	69
8.1	Python Code.....	69
8.2	Confusion matrices for data set “chapter 71”	72
8.3	Confusion matrices for data set “chapter 62”	80
8.4	Prediction metrics for data set “chapter 62”	87

List of tables

Table 1: Headings of chapter 71	31
Table 2: CN of interest from chapter 71	31
Table 3: Taxonomy of the attributes charactering the first batch (chapter 71)	32
Table 4: Relationship between the obsolete and current CN codes	33
Table 5: CN headings chapter 62	34
Table 6: CN of interest from chapter 62:	36
Table 7: Taxonomy of the attributes charactering the second batch (chapter 62).....	37
Table 8: Classifier's performance metrics for initial data set ch. 71.....	39
Table 9: Classifier's performance metrics for data set ch. 71 (without "center stone" feature)	42
Table 10: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring" features)	42
Table 11: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted" features).....	42
Table 12: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones" features).....	43
Table 13: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender" features).....	43
Table 14: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category" features).....	43
Table 15: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type" features)	44
Table 16: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type", "article type" features).....	44
Table 17: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type", "article type", "CITES" features)	44
Table 18: Classifier's performance metrics for data set ch. 71 (without "main material" feature)	45
Table 19: Classifier's performance metrics for data set ch.71 (introduced error: main material "silver" replaced by "gold")	46

Table 20: Classifier's performance metrics for data set ch.71 (introduced error: main material "silver" replaced by "steel")	47
Table 21: Classifier's performance metrics for data set ch.71 (introduced error: main material "silver" replaced by "pearl")	47
Table 22: Classifier's performance metrics for data set ch.71 (introduced error: main material "domestic calf" replaced by "pearl")	47
Table 23: Classifier's performance metrics for data set ch.71 (introduced error: main material "platinum" replaced by "silk")	48
Table 24: Classifier's performance metrics for data set ch.71 (introduced error: article type "jewellery" replaced by "accessories")	48
Table 25: Accuracy of classification for the initial data set ch.62 and reduced data sets scenarios	54
Table 26: Accuracy of classification for the initial data set ch.62 and data sets with errors scenarios.	57
Table 27: Prediction metrics for initial data set ch. 62	87
Table 28: Prediction metrics for data set ch. 62 (without "gender" feature)	88
Table 29: Prediction metrics for data set ch. 62 (without "material category" feature).....	89
Table 30: Prediction metrics for data set ch. 62 (without "article sub-type" feature).....	90
Table 31: Prediction metrics for data set ch. 62 (without "article type" feature).....	91
Table 32: Predictions metrics for data set ch. 62 (without "CITES" feature)	92
Table 33: Prediction metrics for data set ch. 62 (without "main material" feature)	93
Table 34: Prediction metrics for data set ch. 62 (introduced error: main material "wool" replaced by "polyester")	94
Table 35: Prediction metrics for data set ch. 62 (introduced error: main material "wool" replaced by "cotton").....	95
Table 36: Prediction metrics for data set ch. 62 (introduced error: main material "cotton" replaced by "silk").....	96

List of figures

- Figure 1: Structure of the HS code - Example of soja beans5
- Figure 2: HS comparison functionality of UNI-PASS website – Example of soja beans HS6
- Figure 3: HS code 90 29 10 and its two TARIC subdivisions 90 29 00 10 and 90 29 00 907
- Figure 4: TARIC codes 90 29 00 10 and 90 29 00 90 with associated duty rates and measures7
- Figure 5: Example of additional codes and export authorization measures.....7
- Figure 6: The Process Flow Chart for Classification of Goods9
- Figure 7: Scikit-learn machine learning diagram21
- Figure 8: One-hot Encoder principle.....22
- Figure 9: Illustration of K-Nearest Neighbors Algorithm Logic.....22
- Figure 10: Illustration of margin and support vectors of SVC24
- Figure 11: Data transformation to high-dimensional space24
- Figure 12: illustration of a decision tree example25
- Figure 13: Illustration of random forest classification27
- Figure 14: Confusion matrix example29
- Figure 15: Confusion matrix for initial data set ch. 7141
- Figure 16: Confusion matrix for data set ch.71 (introduced error: article type " jewellery" replaced by "accessories")49
- Figure 17: Confusion matrix for initial data set ch. 6251
- Figure 18: Confusion matrix for initial data set ch. 62, zoom on quadrant 1 (upper-left).....52
- Figure 19: Confusion matrix for initial data set ch. 62, zoom on quadrant 2 (upper-right).....52
- Figure 20: Confusion matrix for initial data set ch. 62, zoom on quadrant 3 (bottom-right)53
- Figure 21: Confusion matrix for initial data set ch. 62, zoom on quadrant 4 (bottom-left)54
- Figure 22: Confusion matrix for data set ch. 62 (without "article sub-type" feature).....55
- Figure 23: Confusion matrix for data set ch. 62 (without "main material" feature)56
- Figure 24: Confusion matrix for data set ch. 71 (without "center stone" feature)72

Figure 25: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring" features)	72
Figure 26: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted" features)	73
Figure 27: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones" features)	73
Figure 28: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender" features)	74
Figure 29: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category" features)	74
Figure 30: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type" features)	75
Figure 31: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type", "article type" features)	75
Figure 32: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type", "article type", "CITES" features)	76
Figure 33: Confusion matrix for data set ch. 71 (without "main material" feature)	76
Figure 34: Confusion matrix for data set ch.71 (introduced error: main material "silver" replaced by "gold")	77
Figure 35: Confusion matrix for data set ch.71 (introduced error: main material "silver" replaced by "steel")	77
Figure 36: Confusion matrix for data set ch.71 (introduced error: main material "silver" replaced by "pearl")	78
Figure 37: Confusion matrix for data set ch.71 (introduced error: main material "domestic calf" replaced by "pearl")	78
Figure 38: Confusion matrix for data set ch.71 (introduced error: main material "platinum" replaced by "silk")	79
Figure 39: Confusion matrix for data set ch. 62 (without "gender" feature)	80
Figure 40: Confusion matrix for data set ch. 62 (without "material category" feature)	81
Figure 41: Confusion matrix for data set ch. 62 (without "article type" feature)	82
Figure 42: Confusion matrix for data set ch. 62 (without "CITES" feature)	83

Figure 43: Confusion matrix for data set ch. 62 (introduced error: main material "wool" replaced by "polyester")84

Figure 44: Confusion matrix for data set ch. 62 (introduced error: main material "wool" replaced by "cotton").....85

Figure 45: Confusion matrix for data set ch. 62 (introduced error: main material "cotton" replaced by "silk").....86

List of abbreviations

AEO	Authorized Economic Operator
AI	Artificial intelligence
CBP	Customs and Border Protection
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora
CN	Combined Nomenclature
FN	False negative
FP	False positive
GIR	General Rules for the Interpretation of the Harmonized System
HS	Harmonized System
IS	Information System
IT	Information Technology
KNN	K-nearest neighbors
PCA	Principal Component Analysis
SGD	Stochastic Gradient Descent
SOP	Standard Operating Procedure
SVC	Support Vector Classifier
SVM	Support Vector Machine
TARIC	Integrated Tariff of the European Communities (TARif Intégré Communautaire)
TP	True positive
WCO	World Customs Organization

1. Introduction

All companies dealing with international trade of goods must classify the goods they import or export. Product classification can be a tedious process. For companies focusing on a few specialized raw materials (e.g.: rubber material), a few classification codes cover the full range of products. In such case, a fully manual classification process is acceptable. However, for companies operating in consumer goods industry where the range of applicable classification codes is much wider and is spread across multiple classification chapters, the use of an (semi-)automated classification solution can be an undeniable asset. For the latter type of companies, the classification process represents a time-consuming activity. This is especially true for sectors with seasonal collections, such as fashion industry, where a whole new goods collection must be classified within a short timeframe every season (every three months). Any delays in the classification process could result in import/export delays and by consequence in sales and profit loss.

Due to the redundant and time-consuming aspects of the classification, auto-classification tools have been developed to speed-up the process and alleviate the workload of customs and trade departments. Different types of auto-classification tools strategies exist: decision trees, automatic deduction of a target classification code based on a source classification (manually assigned), machine learning / artificial intelligence.

Given the complexity of these tools and the numerous interpretations of the “automatization” term, some stakeholders might be disappointed by the results provided by these tools [1]. The example of auto-classification tools using the customs description of the goods, highlights a first type of limitation: there is no gold standard labelled data set which could be used as a reference for a supervised algorithm. Each company defines these descriptions internally based on in-house knowledge which can vary from a company to another [2]. Next to the master data availability and quantity, the complexity of the nomenclature is another obstacle to a full automation. As explained in section 2.2 the usage of a good prevails over the good’s composition sometimes and vice-versa. Such complex rules are a hurdle to the automation. Also, the classification notes add a supplementary layer of complexity which is difficult to automate. The above few examples explain why the auto-classification tools might not meet the expectations of parties expecting a fully automated classification of their goods.

The above limitations should not undermine the interest for auto-classification tools. One should see it as a support tool instead of as an out-sourcing possibility for the classification process. A company with a mature customs department combining on the one hand, the understanding of this support point of view, and on the other hand, the product and nomenclature knowledge, is well equipped to

benefit from the advantages of an auto-classification tool while being able to recognize and avoid its pitfalls [1].

Another misconception regarding the auto-classification tools shared by immature customs departments is the (wrong) idea that these tools are plug and play tools that would allow to classify the goods overnight without any upstream preparation effort. On the contrary, a well-informed customs manager will anticipate the obvious prerequisites: the data quality and availability. According to [3], the data quality is a “*measure of the extent to which a database accurately represents the essential properties of the intended application*”. In this thesis the data of interest are the product attributes (or features, characteristics) and the intended application is the product auto-classification. In this context, the data quality can be interpreted as the accuracy of the products characteristics recorded in an information system (IS). In order to evaluate the impact of data quality on the auto-classification prediction success, this research alters the data quality of the studied data sets by simulating data errors, uses an auto-classification algorithm to classify the instances of the altered data set and compares the predicted classifications against the actual classification. For each altered data set, the higher the amount of discrepancies between the predicted and the actual classifications, the bigger the impact of data accuracy on the prediction success of the auto-classification algorithm.

1.1 Problem statement

In order to auto-classify products with an auto-classification tool, the tool must be fed with input data about the products, *i.e.* products features or characteristics. The product characteristics could be: material(s), usage, production methods, weights, etc. Regardless the auto-classification strategy, these characteristics are necessary as an input. The more classification digits (chapter, heading, subheading, ...) the tool has to predict, the more information about the product is needed. This “information about the product” is reflected in the product master data. The maintenance of such master data in IT systems is subject to human errors (incorrect data maintained) and to omissions (missing or incomplete data). Following the “garbage in, garbage out” notion, inaccurate inputs (product features) will not produce quality results (classification).

Assuming that the data quality of the initial data set is accurate, *i.e.* the product attributes are accurate and sufficient to determine the correct product classification, the data quality can be decreased by removing features or altering the features values. Hence, among the extended list of possible data quality problems [4], this thesis will address the “missing value” problem and the “incorrect value” problem as the data set used for this research is extracted from an IS where dynamic checks are implemented to prevent syntax violation, misspelling error and imprecise values problems.

1.2 Research questions

This thesis aims to identify the impact of data quality on the prediction success of auto-classification tools. In other words, it aims to answer the question *“what role does data quality play in the success of classification tool?”*.

To achieve this goal, the below elements will be addressed:

- First, several examples of classification systems (or nomenclatures) will be presented and their importance will be explained, the main legal rules driving the product classification process will be identified and their inherent complexities will be discussed. We will investigate how the nature of the products impacts the classification decision.
- Second, we will focus on the product attributes and classification criteria in order to determine the required granularity to reach satisfying classification results. We will also discuss the challenges of maintaining systematically such attributes in IT systems.
- Third, we will look at different mistakes in master data maintenance of product attributes and we will categorize them.
- Finally, we will analyze how the different categories of mistakes in data maintenance impact an auto-classification tool prediction success.

The first two sub-questions, *i.e.* *“what are the legal rules and inherent complexities driving the product classification?”* and *“what is the required granularity of product attributes to correctly determine the product classification?”* will be mainly answered by literature. The last two sub-questions, *“what are the possible categories of mistakes in data maintenance”* and *“what is their impact on the prediction success of an auto-classification tool”* will be answered by an experiment which will aim to identify the existence (or lack) of link between the certain types of mistakes (“small mistakes” vs. “big mistakes” vs. “missing information”) in the product master data and the correctness of the classification. This link (or absence of link) will have to be characterized per product category or classification chapter (as the classification of some product categories classified under a certain chapter might be more or less affected by master data errors than the classification of products falling under other chapters). In other words, the outcome that will be found for a certain experiment linked to some classification chapters might not be relevant for the classification of products falling under other chapters.

2. Review of research literature

The classification of goods is required by governments for three main reasons. The first one being the determination of customs duties and taxes. The second reason is for the determination of non-tax related measures such as export controls (and corresponding licenses), quotas or anti-dumping measures. Finally, the classification is used for the determination of trade statistics. An incorrect classification can lead to fines for companies, stricter and longer controls by customs authorities or loss of certifications which impact in turn business profitability [5]. These are the main incentives for companies to invest in customs compliance. To illustrate the consequences of an inappropriate classification, the Toyota recent case can be taken as an example: in September 2022 the Thailand Supreme Court ruling stated that the local Toyota's unit had to pay \$272 million in extra import duties because the goods imported during 2010 and 2012 were not subject to a reduced import rate under the Japan-Thailand Economic Partnership Agreement, *i.e.* they were wrongly classified as "car parts" while should have been treated as "complete knock-down kits" [6].

2.1 Classification codes

This section aims to provide an overview of the classification codes: we will describe the of HS concept [7] and provide a few examples of its application across the world such as in Korea, in Taiwan and in Europe. For the later, further details will be provided regarding the Combined Nomenclature (CN), the Integrated Tariff of the European Communities (TARIC) codes, additional digits and measures.

2.1.1 Harmonized System

The "Harmonized Commodity Description and Coding System", also referred to as the "Harmonized System" (HS) has been developed by the World Customs Organization (WCO) as of 1983 based on the "International Convention on the Harmonized Commodity Description and Coding System" and adopted as of 1988 [8]. Its development resulted from the desire to facilitate the international trade by using a standardized and common system for the classification of goods [9]. The HS nomenclature is used by more than 200 countries and economies worldwide [10]. The current version of the HS nomenclature has been enforced in January 2022. There is generally an update to the HS nomenclature every five years. These updates aim to reflect the trade changes and the newly developed technologies.

The HS is organized in 21 sections and each section is divided in chapters represented by the first two digits of the code. Each chapter is further divided into headings (3rd and 4th digits) and subheadings (5th and 6th digits) which is the lowest level of the HS. Figure 1 illustrates the HS code for soja beans: under the section II "*Vegetable products*" of the HS nomenclature the chapter 12 refers to "*Oil seeds and*

oleaginous fruits; miscellaneous grains, seeds and fruit; industrial or medicinal plants; straw and fodder”, the heading 01 refers to “Soya beans, whether or not broken” and the subheading 10 to “Seed”.

HS Code:	<u>12</u>	<u>01</u>	<u>10</u>
	Chapter	Heading	Subheading

Figure 1: Structure of the HS code - Example of soya beans
Source: [11]

Each country is free to supplement the six-digits HS code with further digits. All or part of these additional digits can also be defined at a union level as it is the case in the European Union with the Combined Nomenclature (CN) applied for exports from the EU and the Integrated Tariff of the European Communities (TARIC) applied for imports into the EU [12]. These additional digits can also be at association level as it is the case for the Association of Southeast Asian Nations (ASEAN) [13] or the Economic and Monetary community of Central Africa (CEMAC) [14]. A few examples of classification nomenclatures will be presented in the next sections.

2.1.2 Extensions of the Harmonized System

In this section we are looking at the way the HS codes are further extended at national, union or association level. We will first look at the South Korean nomenclature system. Via this use case, we will highlight the differences of nomenclature definitions between geographies. Then we will look at the European nomenclatures, the CN and the TARIC, which are even further extended with measures and additional codes.

2.1.2.1 Harmonized Schedule of Korea

The South Korean nomenclature system referred to as Harmonized Schedule of Korea (HSK) is composed of 10 digits [14], the first six being the HS defined by the WCO, the last four being the national digits. The HSK can be found on the *UNI-PASS* website which is the electronic customs clearance system developed by the Korean Customs Services (KCS). *UNI-PASS* provides a functionality to compare simultaneously up to three different nomenclatures [15]. In the example illustrated in Figure 2 we are looking again at the example of soya beans HS (12 01 10). We notice that the European Union does not have further subdivisions and reaches the 10 digits TARIC code by adding four zeros. For the same HS code, South Korea also extends it with four additional digits while Taiwan with five digits. Moreover both countries have subdivisions below the 12 01 10 code to differentiate either the soja seeds meant for bean sprouts from the other types of soja seeds (in the case of Korea), or the soja seeds below a certain weight from the others (in the case of Taiwan).

This example can be generalized to the other countries applying the HS defined by the WCO. First, the number of additional digits varies from a country to another. Some geographies might only extend the HS with only two digits while other geographies might extend it with four or five. Second, there is no one-to-one relationship between the local nomenclatures. A group of goods classified under the same code in a country, might be classified with two (or more) different codes in another country which deems necessary to reach a more detailed level of classification for certain goods. This element should be kept in mind in the context of auto-classification as the desired level of classification details will steer the decision of the number of product attributes to be used in the auto-classification process.

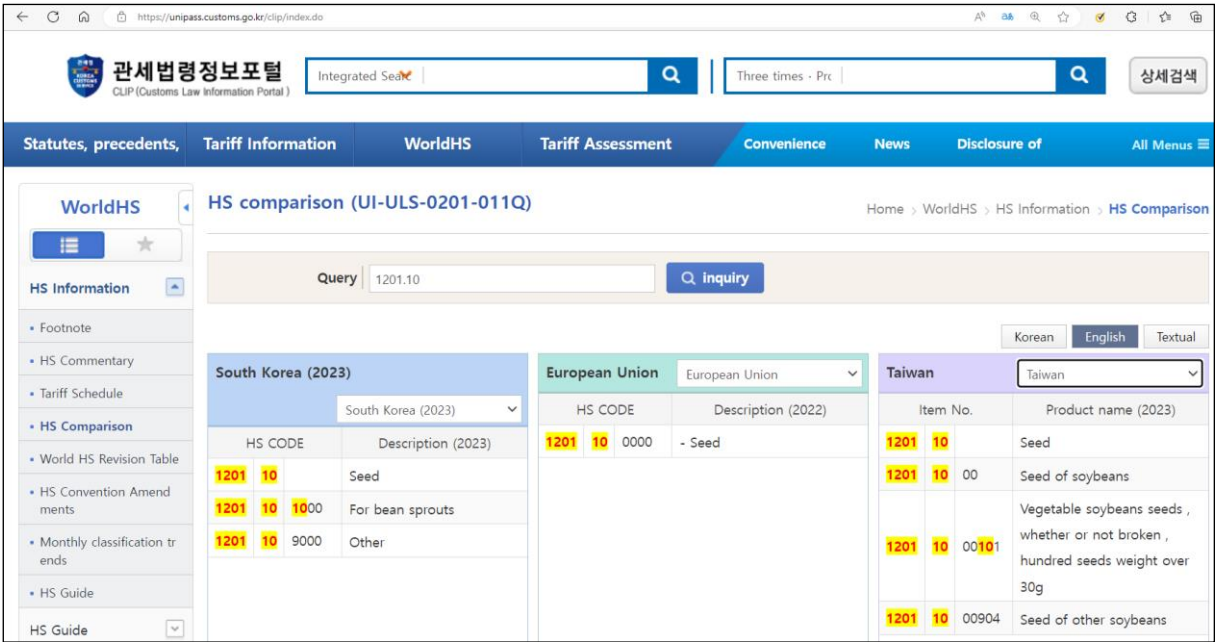


Figure 2: HS comparison functionality of UNI-PASS website – Example of soja beans HS
Source: [15]

2.1.2.2 Combined Nomenclature and the Integrated Tariff of the European Communities

In the EU, the Combined Nomenclature (CN) and the Integrated Tariff of the European Communities (TARIC) are further developments of the HS. On the one hand, the CN adds two digits to the HS code and is used for exports out of EU and for intra EU-trade statistics [16]. On the other hand, the TARIC is used for imports into the EU and extends the CN with two more digits, leading to a ten digits code. These last two digits allow a further subdivision of goods classification and hence of the tariff rates applied at importation [12]. Figure 3 and Figure 4 illustrate the subdivision of HS code 90 29 10 into two TARIC codes with their respective duty rates (0% or 1.9%) and (some of) their tariff measures (or conditions) such as the airworthiness tariff suspension reducing the duties rate to 0% upon presentation of a C119 certificate.

Moreover, as depicted by Figure 5, a code - CN or TARIC - can also be relevant for additional codes consisting of a 4 digits number and providing additional information about the traded goods such as whether they falls under a regulation (cfr. example of additional code 4099 in Figure 5). Similarly, a code can be relevant for non-tariff measures such as export authorization allowing the export upon the presentation of a Y935 document (cfr. example of Figure 5).

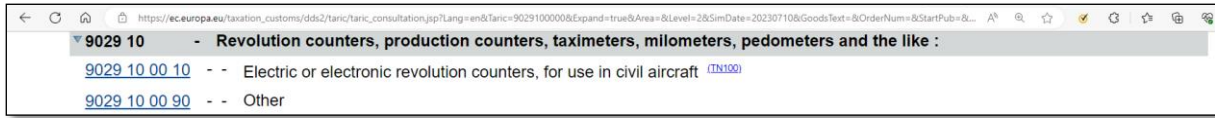


Figure 3: HS code 90 29 10 and its two TARIC subdivisions 90 29 00 10 and 90 29 00 90
Source: [17]

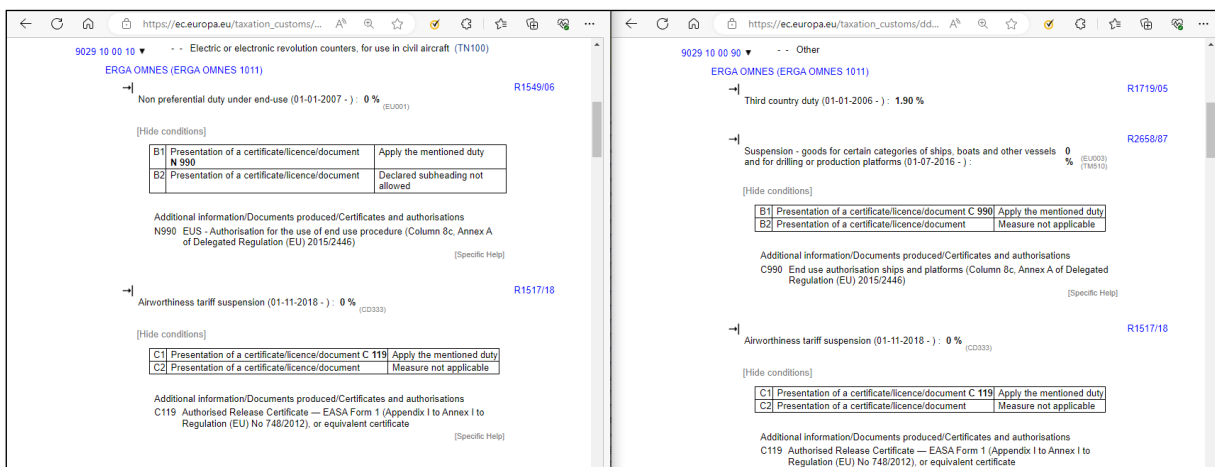


Figure 4: TARIC codes 90 29 00 10 and 90 29 00 90 with associated duty rates and measures
Source: [17]



Figure 5: Example of additional codes and export authorization measures
Source: [17]

The sensitivity of the information driven by the classification codes (duty rates, tariff measures, non-tariff measures and additional codes) demonstrates the importance of a correct classification. It also

justifies the fines, the loss of licenses or certifications, the stricter and longer controls negligent companies might incur.

2.2 Classification rules

This section provides an overview of the classification process because a high-level familiarization is necessary to understand the complexities and pitfalls an auto-classification tool can face. We will first explain the classification rules applicable at WCO level and then look in more details at the binding tariff information (BTI), the EU variation of the advance ruling for classification, as an additional tool to ensure the correct and uniform application of the HS classification.

2.2.1 General Rules for the Interpretation

The classification process is governed by the General Rules for the Interpretation of the HS (GIRs) developed by the WCO [18]. The GIRs constitute the single set of legal principles governing the classification and aim to ensure a certain uniformity in the product classification across the world. They consist of six rules (GIR 1 to GIR 6) to be applied sequentially until the classification is found (the first five rules govern the classification at heading level, the sixth rule governs the classification at subheading level). They provide guidance on how to classify a product based on its primary use, its composition or other characteristics. The diagram on Figure 6 illustrates the logic behind these rules.

To complement the diagram from Figure 6, the meaning of each GIR rule is explained below in details. The first rule, the GIR 1, is *“The titles of Sections, Chapters and sub-Chapters are provided for ease of reference only; for legal purposes, classification shall be determined according to the terms of the headings and any relative Section or Chapter Notes and, provided such headings or Notes do not otherwise require, according to the following provisions:”* (cfr. GIRs 2 to 6) [18]. This first rule prescribes to initiate the classification process of a product at the 4-digits level based on the terms (the wording) of headings and corresponding section or chapter notes. A particular attention should be paid to only compare headings between them (not with a subheadings). Also, this rule gives equal status to the heading terms and notes [19]. In case these terms and notes allow to uniquely classify a product, then the GIR 1 solves the heading classification. In the other cases, the subsequent rules must be considered.

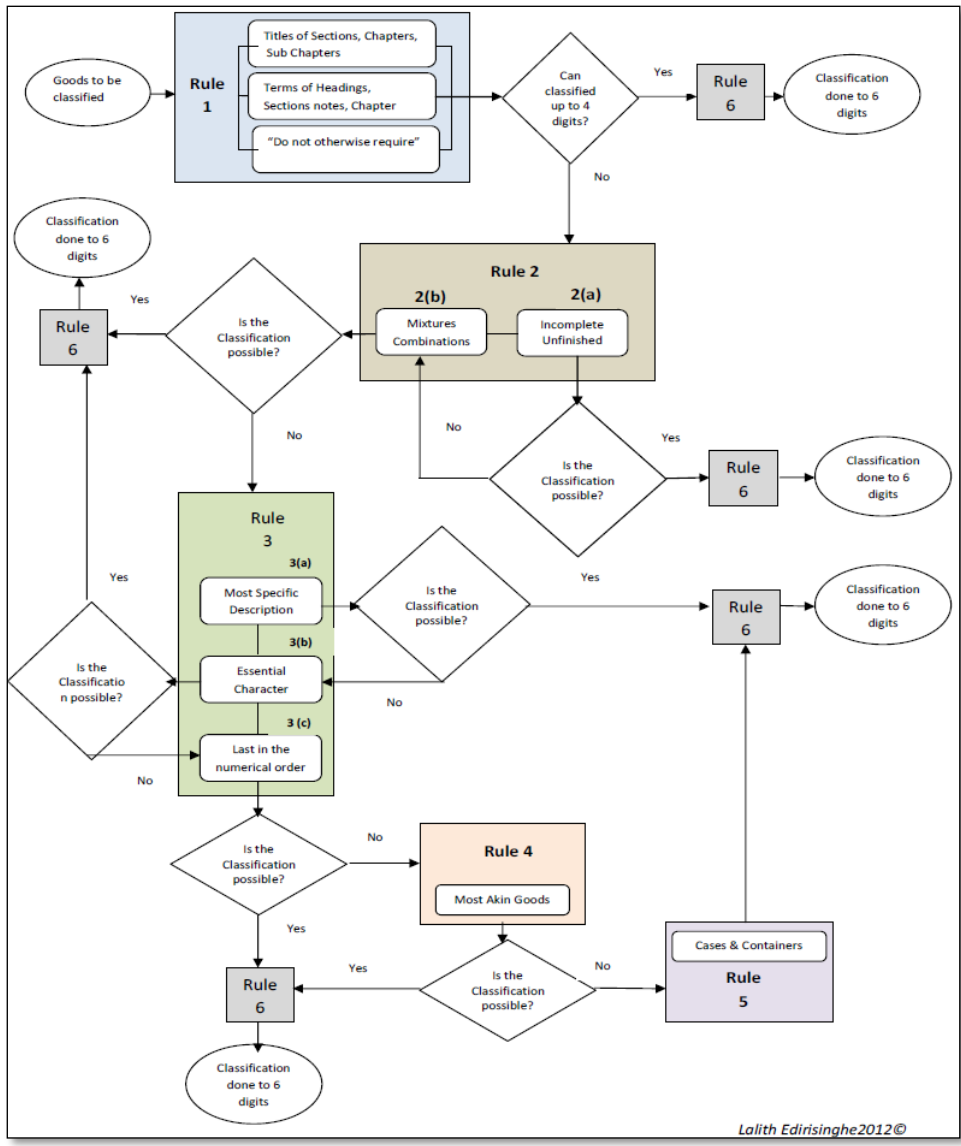


Figure 6: The Process Flow Chart for Classification of Goods
 Source: [20]

The second rule consists of two parts. The first part, GIR 2 (a), indicates “(a) Any reference in a heading to an article shall be taken to include a reference to that article incomplete or unfinished, provided that, as presented, the incomplete or unfinished article has the essential character of the complete or finished article. It shall also be taken to include a reference to that article complete or finished (or falling to be classified as complete or finished by virtue of this Rule), presented unassembled or disassembled.”. The GIR 2 (a) covers incomplete or unfinished (whether assembled or disassembled) goods by extending the scope of articles referred to in headings to the incomplete or unfinished articles presenting the essential character of the finished or complete article. Moreover, the term “as presented” refers to the state the goods were presented at the border. The second part, GIR 2 (b) indicates “(b) Any reference in a heading to a material or substance shall be taken to include a reference to mixtures or combinations of that material or substance with other materials or substances. Any

reference to goods of a given material or substance shall be taken to include a reference to goods consisting wholly or partly of such material or substance. The classification of goods consisting of more than one material or substance shall be according to the principles of Rule 3.”. The GIR 2 (b) covers mixtures and combination of materials by extending the scope of articles referred to in headings to the articles consisting of mixtures or of combinations of materials and leads to GIR 3. As an observation, it should be noted that the GIR 2 leads to a classification only in combination with GIR 1 or with GIR 1 and GIR 3.

The GIR 3 consists of three parts (*cfr.* [18] for full text) and starts with *“When by application of Rule 2 (b) or for any other reason, goods are, prima facie, classifiable under two or more headings, classification shall be effected as follows : (a) ... (b)... (c)...”*. The GIR 3 (a) indicates that the most specific heading should prevail in the classification process. Moreover, when each of the headings refer to only one of the materials or to articles included in a set, these headings should be equally considered regardless of how specific is their description. The GIR 3 (b) states that when *“mixtures, composite goods consisting of different materials or made up of different components, and goods put up in sets for retail sale”* cannot be classified by reference to GIR 3 (a), then their classification should be driven by the material or the component giving them their essential character. The GIR 3 (c) covers the classification of articles that cannot be classified by reference to GIR 3 (a) or (b). In such case, the article should be *“classified under the heading which occurs last in numerical order among those which equally merit consideration.”* [19].

The GIR 4 is self-explanatory and indicates that *“Goods which cannot be classified in accordance with the above Rules shall be classified under the heading appropriate to the goods to which they are most akin.”*

The GIR 5 consists of two parts. The first part, GIR 5 (a), indicates that cases or containers specially shaped to fit a (or a set of) article(s) should be classified with such article when sold together, *i.e.* the container must be presented with the article but the article does not have to be in the container [19]. The container should be *“suitable for long-term use”* (similar to the durability of the article) and be of *“a kind normally sold therewith”*. Additionally, this rule is not applicable *“to containers which give the whole its essential character”*. The GIR 5 (b) states that the packing materials or containers should be classified with the articles if they contain them at the moment when these are presented at the border (*“presented with the goods therein”*). Also, the rule involves that the packing material or containers *“are of a kind normally used for packing such goods”*. It should be noted that, in case the packing materials or containers are *“suitable for repetitive use”*, the GIR 5 (b) is not binding and each country can decide to apply it or not [19].

Finally, the GIR 6 can be applied to determine the subheading classification by using the GIRs 1 to 5 while considering the terms and notes of the subheadings (instead of the headings, given the “*mutatis mutandis*”). This process should be iterated twice: once for the fifth-digit subheadings and once for the sixth-digit subheadings as “*only subheadings at the same level are comparable*”. During this process, the section and chapter notes still apply unless mentioned otherwise.

2.2.2 Tariff binding information

Several legally binding instruments exist to ensure a uniform application of the classification: the GIRs, the legal notes (section, chapter, subheading notes for the HS), the additional notes (for the CN) [21], the Commission Implementing Regulations – classification of certain goods on the CN [22], the Court of Justice of the European Union (CJEU) judgments, and the advance rulings on classification also referred to as binding tariff information (BTI) in the EU. In this section we will focus on the latter.

The determination of the relevant classification code can be difficult, therefore economic operators have the possibility to apply for a BTI which is a written decision on classification issued by a national customs authority. The holder of a BTI benefits from a legal certainty regarding the way customs will consider the traded goods [23]. This legal certainty is particularly important when it comes to customs duties: an incorrect classification might lead to higher customs duties or, on the contrary, being exposed to retroactive customs duties payments due to an incorrect classification with a lower duty rate. Additionally to the financial aspect, the BTI also gives legal certainty with regards to the applicable measures, *i.e.* indicating which authorization or license is required or any other non-tariff obligation.

The BTI comes with legal certainty but also with obligations as it is binding on all EU member states customs authorities and on the holder. It is binding on all EU states customs authorities because, although it is issued at national level, all EU customs authorities have to accept the classification stated on the BTI. It is binding towards the holder because it must use the classification code resulting from the BTI decision, regardless whether it is advantageous or not in terms of customs duties payment of measures requirements, as such decision is not a recommendation but an enforcement. Hence, the economic operator should weigh the risk of a disadvantageous decision from customs against the benefit of a legal certainty before applying for a BTI [24].

The BTI application can be submitted electronically on the *Taxation and Customs Union* website towards the customs authorities of the country where the economic operator is established or where it intends to import or export the goods [25]. A BTI is valid for 3 years unless it is annulled, revoked or ceases to be valid. The annulment can be motivated by an inaccurate or incomplete information used in the application process. The revocation of a BTI can occur in case it is no longer compatible with the interpretation on the CN or HS (which could result from notes amendments, CJEU judgements or WCO

classification decisions or opinions), in case a Commission decision requests the revocation, or in case of an administrative error [26]. Finally, a BTI validity can cease in case it no longer complies with the legislation (which could be due to a change in the CN or HS or a Commission measure) [26]. Further legal details regarding the BTIs are given in Articles 22 to 37 of the Customs union ode (UCC), Articles 11 to 32 of the UCC Delegated Act, Articles 8 to 23 of the UCC Implementing act.

All valid BTIs are available in the European BTI (EBTI) database accessible on the *Taxation and Customs Union* website. This database is made public for four reasons: transparency, consultancy, uniformity and predictability [27]. The transparency with regards to the information - the description of the goods along with the classification justification leading to the decision - demonstrates that the decisions are based on objective data and according to the GIRs. These ready-made decisions and their justifications can serve as a consultancy by providing classification officers with reasoned confirmations or objections regarding their own classification reasoning. This results in a better uniformity of identical goods classification and in a higher predictability for the classification officers with regards to the customs decisions.

The BTI application process has also a few shortcomings. For example, customs authorities should in principle issue a decision within the 120 days following the application acceptance. In some circumstances, this time limit can be extended. These lengthy waiting periods can negatively impact businesses. Moreover, given the inherent feature of the BTIs decision (being made at national level), situations where two member states issue a different classification for similar goods can occur [23]. Such situations are referred to as divergent BTIs. National customs authorities should prevent, or at least reduce, the occurrence of divergent BTIs by consulting routinely and rigorously the EBTI database [28] as required by Article 17 of the UCC IA: *“The customs authority competent to take a decision shall, for the purposes of ensuring that a BTI decision which it intends to issue is consistent with BTI decisions that have already been issued, consult the electronic system referred to in Article 21 and keep a record of such consultations.”*. Despite this UCC obligation, such situations did occur in the past before this obligation came into force with the UCC (cfr. example in section 2.4.2), and can still happen [29], [30] for example due to the limited information available in the database (lack of details in the goods description or missing article pictures) but also due to non-compliance with the Article 17 [30].

2.3 Classification expertise

In this section we will discuss the requirements in terms of classification expertise: we will look at the profiles that are entitled to perform product classification in a company. Additionally, we will investigate the legitimate question of what is the legal ground for the use of auto-classification tools

and understand how they articulate with the work performed by classification officers and customs administrations.

2.3.1 Knowledge requirements

The classification of goods can be an in-house activity, meaning that a company manages the classification by its own, or it can be an outsourced activity performed by an external party such as a customs broker. In both cases, the company is liable for the classification as it is expected to put in place controls ensuring the quality of the external party work. Often the customs and trade departments have dedicated resources to the classification process. These resources have most of the time one to two years of learning or training - which is an indication of how complex the process is. In some cases, these resources must be “declared” to customs authorities as trusted and knowledgeable parties. Trainings or certificates serve to demonstrate their knowledge in terms of product characteristics, nomenclature, GIRs and other legal instruments. This is the case for Authorized Economic Operator (AEO) applicants in EU as stated by PART 2.II.2 of the AEO Guidelines “... *it is crucial that the staff is aware of the importance of non-fiscal requirements, the correct classification of goods and keeping the master data up to date. Regular training or self-study of the developing legislation is mandatory for businesses dealing with above mentioned goods*” and by the Annex 1a to the AEO Guidelines, the AEO Self-Assessment questionnaire [31]. In the questionnaire section 1.3 “*Information and Statistics on customs matters*” the applicant has to answer below questions regarding tariff classification:

- a) How, and by whom, is the tariff classification of goods decided?*
- b) What quality assurance measures do you take to ensure that tariff classifications are correct (e.g. checks, plausibility checks, internal working instructions, regular training)?*
- c) Do you keep notes on these quality assurance measures?*
- d) Do you regularly monitor the effectiveness of your quality assurance measures?*
- e) What resources do you use for tariff classification (e.g. database of standing data on goods)?*

Further details on how to answer these questions are given in Annex 1b, Explanatory notes for AEO-Self-Assessment Questionnaire. First, for question a), the name and position of the staff members responsible for the classification must be provided to customs authorities. Questions b) and d) explicitly mention that quality assurance measures are expected and that such measures can take multiple forms such as trainings, work instructions or operating procedures, regular checks and monitoring ensuring that the classification is done correctly and according to the instructions. Additionally, according to the notes for question c), the quality assurance measures are expected to be documented and made available as evidence to customs auditors. Moreover, for question d), the

applicant is expected to review the classification on a regular basis and keep a record of this activity. The record should indicate how, by whom and how often the review is performed [31]. Finally, as part of answer to e) question, the applicant should be able to present the list of the resources used, including potential BTIs and any information used to classify the goods. The above questions clearly demonstrates the level of control expected from a AEO applicant.

2.3.2 Customs view on auto-classification tools

When it comes to the use of auto-classification tools, there is no legal mentions or contraindications. This is explained by the fact that in case of incorrect classification of goods, the liability stays with the declarant and, if applicable, with his representative. They will bear (not the IT solution or its supplier) the consequences in terms of duty payment, authorization revocation or suspension, criminal sanctions or fines [1]. This liability leads to the practical consideration that auto-classification tools are to be used as support tool for classification officers, not as a replacement of these roles.

The use of auto-classification as support tool seems to be largely accepted and even promoted at WCO level via the BACUDA program [32] and its online “*AI HS code Recommendation Platform*” (a 6-digits HS code prediction using commercial description of goods) [33]. This WCO BACUDA program launched in 2019 aims to raise awareness and build capacity in data analytics among the WCO members [34]. Based on the input of experts from academia and research institutes, methodologies are developed to be deployed among customs administrations. More specifically the HS code recommendation algorithm is expected to be used by customs administrations during declaration processes in order to better assess risk and prevent misclassification fraud by providing field customs officials with HS code options based on the customs descriptions [35]. This free-access tool, among other, can be used as well by traders [36].

Some national and customs union authorities also provide expert search classification systems deployed by 3CE Technologies (owned by Alavara) [37]. We can cite: the European Commission Combined Nomenclature Search Engine [38], the U.S. Consensus Bureau Schedule B Search Engine [37], the Canada Tariff Finder [38] and the Federal German Government *Warenverzeichnis Online* search engine [39], [40].

These expert systems are not simple keyword engines that try to match the worlds entered by a user with the ones from the nomenclature, and hence, to be effectively useful, require the user to have a nomenclature knowledge. These expert classification systems have been built to address the main HS classification challenges: matching the commonly used commercial descriptions of goods with the nomenclature terminology; handling complex items such as sets, kits and parts; considering the GIRs and HS legal notes. Additionally they detect underspecified search (insufficient information provided

by a user) and prompts the user to answer additional question regarding the product [37]. Also, when relevant, these expert search systems retrieve the HS legal notes. However, it should be noted that although the search engine displays the CN explanatory notes under the proposed HS code, it does not consider these CN explanatory notes in the HS determination process. This is because the technology proposed by 3CE focuses on the HS regulation and does not take into account the EU regulations [1]. This specificity can lead to incorrect classification at (sub)heading level according to the EU regulations [1]. Finally, it should be noted that the claim of handling complex items such as sets does not necessarily mean that engine manages to classify the products but it rather triggers an error messages stating that *“The item you are classifying is considered a complex item (or set) which normally requires each component to be classified separately. Alternatively, you may request a binding classification ruling for your complex item (or set) from Customs in the country of import.”*. This message was, for example, displayed when *“sets of pens”* was entered in the “description of your product” field.

2.4 Classification complexities

Categorizing and describing the classification complexities is the first step to better understanding the challenges and limitations of the auto-classification tools. In this section we are looking first at the nature of the product and, via some examples, we will draw conclusions on their inherent classification complexity. We will also investigate the interpretation of codes per country and highlight how these national interpretations lead to classification discrepancies already at the 6 digits level. Finally, the master data availability in IT system will be discussed and the importance of the initial effort to setup (and maintain) the necessary attributes in the system will be stressed.

2.4.1 Nature of product

When the classification of a product cannot be derived solely from the wording of the nomenclature, additional classification criteria must be considered. In this section we will illustrate some of these cases. In the first case, the material of the product competes against its function in the classification process. In the second case, the intended use of the products is crucial to determine the classification. The third example illustrates two configurations of sets classifications. And the last example highlights the importance of technical knowledge and engineering criteria to enable the classification of certain products.

The first particular case is the rain gauges made of glass for which two heading are competing: 90 15 as a “meteorological” instrument and 70 20 for “other articles of glass”. Choosing between these two heading consists in answering what is the essential character of the product: is it its function or its material [39]? At first, the U.S. customs considered the material as the essential characteristic (*cfr.* customs rulings NY N296613, NY K81163, NY K80012, NY H88046 and NY G81419). However, ruling

H308673 from 2020 revoked the previous ones based on the HS Explanatory Note 37.0 to the heading 9015 [40]. This explanatory note states that “(V) METEOROLOGICAL INSTRUMENTS [...] The group does, however, include the following: [...] (8) Rain gauges and indicators, for measuring rainfall in a particular place. The simplest type consists of a funnel of known diameter fixed to a receptacle to collect the rain which is then measured in a calibrated tube.”. Although the HS Explanatory Notes are a non-binding instrument, the U.S. customs considered that they are generally indicative of the proper interpretation and hence considered that the essential character of the rain gauge is its function.

The second example illustrates the intended use criteria which is necessary to classify goods correctly under heading 84 32 “Agricultural, horticultural or forestry machinery for soil preparation or cultivation; lawn or sports-ground rollers”. More specifically, spreaders and distributors can only be classified under this header if they are respectively intended to spread manure (HS code 84 32 41) or to distribute fertilizers (HS code 84 32 42).

The third example consists of a ballpoint pen and a fountain pen sold together. According to the nomenclature, these articles should be classified under a single HS code 96 08 50 “Sets of articles from two or more of the foregoing subheadings” as the foregoing subheadings cover 96 08 10 “Ballpoint pens” and 96 08 30 “Fountain pens, stylograph pens and other pens”. However, when the articles are not covered by a heading, the classification becomes more complex and the applicability of GIR 3 b) must be checked and could lead to the determination of the item representing the essential character of the whole. Such determination can be based on different factors (for example the nature of the material, the weight, the value, or the significance for the function of the whole) [1].

Finally, the last example of complexity is illustrated by HS code 39 01 20 10 corresponding to chapter 39 “PLASTICS AND ARTICLES THEREOF”, heading 01 “Polymers of ethylene, in primary forms:”, subheading 20 “Polyethylene having a specific gravity of 0,94 or more:” and CN subheading 10 “Polyethylene in one of the forms mentioned in note 6(b) to this chapter, of a specific gravity of 0,958 or more at 23 °C, containing:

- 50 mg/kg or less of aluminium,
 - 2 mg/kg or less of calcium,
 - 2 mg/kg or less of chromium,
 - 2 mg/kg or less of iron,
 - 2 mg/kg or less of nickel,
 - 2 mg/kg or less of titanium and
 - 8 mg/kg or less of vanadium,
- for the manufacture of chlorosulphonated polyethylene”.

The wording of this CN code indicates that the classification of some products requires a keen knowledge of the engineering or manufacturing process.

2.4.2 National interpretations

All countries member of the WCO use the standardized HS framework and therefore, for a certain product, the first six-digits of each national classification should in principle always be identical with the first six-digits of other countries. In practice, several contradictions exist not only between countries (not related by a customs union or association) but also between EU member states.

The differences between countries (not related by a customs union or association) can be explained by a further understanding of the assumption that the classification code is the same up to the sixth digit. This assumption is only valid *“for the structure and the wording of the HS”*, while *“the interpretation of the scope of the (sub)headings and legal notes is still subject to the competent jurisdictions”* [1].

The reasons of the existing contradictions within EU member states are explained in section 2.2.2. A few examples of contradictory decisions made by different members of the WCO will be presented in this section.

An example of a product part of everyday life is a soap dispenser which has been classified differently by at least four countries. In 2022, Germany classified a soap dispenser under HS code 84 79 89 via the BTI DEBTI6384/22-1 [41]. This classification corresponds to chapter 84 *“NUCLEAR REACTORS, BOILERS, MACHINERY AND MECHANICAL APPLIANCES; PARTS THEREOF”*, heading 79 *“Machines and mechanical appliances having individual functions, not specified or included elsewhere in this chapter”* and subheading 89 *“Other”*.

A similar product, has been classified by the U.S. customs under code HS 84 24 89 via the customs ruling HQ H305296 [42] in 2020. Although the chapter is the same as the one used in Germany, the heading 24 is different and refers to *“Mechanical appliances (whether or not hand-operated) for projecting, dispersing or spraying liquids or powders; fire extinguishers, whether or not charged; spray guns and similar appliances; steam or sandblasting machines and similar jet projecting machines”* and the subheading 89 to *“Other”*. This ruling revoked or modified three previous rulings (NY N249630, NY N299353 and NY N298787) of the U.S. Customs and Border Protection (CBP) classifying similar soap dispenser with subheading 20 *“Spray guns and similar appliances”*. The reason of this opinion change lies in the definition of the *“spray”* and *“pump”* terms.

Taiwan classified a soap dispenser under HS code 84 13 20 via ruling 104AA0113 [43] in 2015. While the chapter is identical to the one used by Germany and the U.S., the heading 13 *“Pumps for liquids,*

whether or not fitted with a measuring device; liquid elevators” and the subheading 20 “Handpumps, other than those of subheading 8413 11 or 8413 19” are different.

A completely different reasoning has been applied by India when classifying a plastic mechanical liquid dispenser under HS code 39 24 90 via advance ruling GUJ/GAAR/R/34/2020 [44]. The chapter 39 refers to *“PLASTICS AND ARTICLES THEREOF”*, heading 24 to *“Tableware, kitchenware, other household articles and hygienic or toilet articles, of plastics”* and subheading 90 to *“Other”*. The reasoning of not using subheadings of code 84 24 is motivated by the fact that the items in question are not part of the description of the subheading.

The above soap dispenser example illustrates the classification complexities leading not only to international disagreements but also national reconsiderations, revocations and amendments of previous rulings.

A second example of classification disagreement for a product part of everyday life is the step stool which was classified under HS code 94 03 70 via the BTI DEBTI27378/22-1 by Germany in 2022 [45]. The German customs considered that this product was falling under chapter 94 *“FURNITURE; BEDDING, MATTRESSES, MATTRESS SUPPORTS, CUSHIONS AND SIMILAR STUFFED FURNISHINGS; LUMINAIRES AND LIGHTING FITTINGS, NOT ELSEWHERE SPECIFIED OR INCLUDED; ILLUMINATED SIGNS, ILLUMINATED NAMEPLATES AND THE LIKE; PREFABRICATED BUILDINGS”*, heading 03 *“Other furniture and parts thereof”* and subheading 70 *“Furniture of plastics”*. Whereas for a similar step stool, the U.S. ruled differently the same year by considering that the product is to be classified under HS code 39 24 90 [46] (*cf.* above for HS description). This divergent opinion lies in the interpretation of the product function: for the U.S. customs, the step stool is *“designed to elevate a standing person in order to reach something or perform a task at a greater height”* and therefore is considered as a *“household articles of plastics”*.

Divergent opinions on classification within the EU member states did occur on a regular basis before the UCC entered into force (*cf.* section 2.2.2). One example is the steering wheel cover which was classified under the HS code 87 08 94 *“Part of the vehicles”* via the BTI DE9149/15-1 by German customs in 2015 because it was intended to be attached to the steering wheel [47]. A similar product, was classified under HS code 42 05 00 *“Other articles of leather or of composition leather”* via the BTI PLPL-WIT-2014-01186 by Polish customs in 2014 [47].

Despite the Article 17 of the UCC IA (*cf.* section 2.2.2), divergent BTI still occur as mentioned by the Customs Code Committee Tariff and Statistical Nomenclature in their committee meeting minutes [30]. To prevent situations with divergent BTIs, member states must check the EBTIs database to identified previously issued BTI for similar kind of products by another member state before issuing

their own BTI (*cf.* section 2.2.2). Moreover, when divergent BTIs are discussed by the Committee, the member states are asked to review their BTIs and revoke them if they are not in accordance with the Committee's conclusion [30].

2.4.3 Master data availability in IT systems

As highlighted in section 2.4.1, the knowledge of product characteristics is essential for the classification process and often the commercial description is not sufficient. Such knowledge can be obtained by observing the product or reading its description (to determine the shape, the size, ...), by consulting technical drawings, by relying on supplier or manufacturers notes (to retrieve the composition, the intended use, ...) or, in case of in-house manufacturing, by contacting the engineering department (to determine more complex engineering criteria such as density or gravity, ...). These investigation activities can be performed on a case-by-case basis by a classification officer depending on the product at hand. However, when using an autclassification tool, such ad-hoc investigations and knowledge enrichments is not possible as all classification possibilities – within the range of the goods traded by a company - must be anticipated and specified from the start to define the minimal required scope of master data (*i.e.* the recording of the products characteristics in an IT system) to enable the differentiation between the possible classification codes.

This initial setup of the master data is crucial and represents a significative effort [1]. We will describe in the next paragraphs the different steps of such initial setup with the assumption that a company will continue trading the same type of products. In case, at some point in time, the company extends the scope of traded goods, some of the steps might need to be carried out again.

First, the range of products in scope, *i.e.* traded by a company, must be determined. Then, the corresponding classifications must be analyzed in order to derive the classification criteria from the nomenclature. Based on the previous sections, we can list here the most common criteria: the physical attributes (shape, size, weigh, ...), the compositions (main material, components or ingredients, quantity per component or ingredient, ...), the engineering characteristics (density, ...) and the intended use or function.

It should be noted that listing all the criteria, or attributes, that are relevant for classification is not sufficient. For each type of product, the necessary and sufficient criteria should be listed and their maintenance should either be enforced by the IT system or part of a standard operating procedure (SOP). Such maintenance enforcement or checks should be part of the article workflow creation or validation. This way, when a product - for example a leather bag - undergoes the article workflow checks, the IT system will request the user (or the user will know based on the SOP that he has) to maintain the main material (in this case "leather") but not the density (which does not play a role in

the classification of such type of articles). The definition of the necessary and sufficient criteria is a complex exercise and can only be performed by employees who have both the knowledge of the articles and of the nomenclature.

While systems checks and SOPs can help to consistently record the product attributes, the initial maintenance of these attributes remains a challenge for several reasons. The first reason is the large amount of data to be entered at once. This is particularly relevant for companies operating in industries such as consumer goods where trading a few thousands different goods is common. The second reason is related to the data quality. To ensure good data quality, the data entry should be carried out once again by employees who have a good knowledge of the goods and their properties [1]. Finally, the initial setup of the master data requires a cross-departments collaboration to ensure, for example, that the procurement department is responsive and contacts suppliers for additional product specification, that the production department allocates resources to answer questions on products attributes, and that the customs or legal department applies for BTI when necessary.

The effort of the initial setup should be seen in parallel with the benefits it would bring. The accuracy of the master data maintained as part of the initial setup determines the data quality that would feed the auto-classification tool. As demonstrated in chapter 5, the data accuracy of some attributes can have a drastic impact on the prediction success of the classifier. If the decision to use an auto-classification tool is taken, the master data maintenance should be performed methodically and some IT checks could be implemented to reduce the risk of human error. Such checks are, for example, the use of a predefined list of possible values for each attribute (instead of allowing the maintenance of free text which increases the risks of typos) or, as mentioned above, the enforcement of some attributes (to avoid missing values).

The final decision of investing in such initial effort should take into account the efficiency gain and the spare time an auto-classifier would bring to employees. Companies trading high volumes of seasonal consumer goods would typically experience a quick return on investment, whereas companies trading a few types of raw materials covering a limited range of the nomenclature might prefer to continue with an ad-hoc manual process.

3. State of the art of classification

In order to answer the research question on the impact of data quality on the prediction success of auto-classification tools, an auto-classification tool must be chosen to carry out the experiments. This chapter describes the selection of the classification tool.

For the purpose of this research, several classification algorithm have been considered. For the sake of simplicity, an open-source platform was preferred. Python with its *Scikit-learn* widely-used machine learning library offers multiples options. The three main classification algorithms are described below. For reference, the *Scikit-learn* library provides decision tree (*cf.* Figure 7) to help one select the right algorithm depending on the type and amount of data available. Although the algorithms are well known by academics, the indication related to the minimum amount of data necessary is an important information. The path made of black arrows illustrates the selection of the algorithms for this research. As described in section 4.2, the data sets at our disposal contain respectively 111.967 and 12.297 instances. The goal is to predict a category based on labeled data. Hence, a supervised learning algorithm (represented by the pink “classification” bubble in Figure 7) is required. As at least one of the data set contains less than 100k samples (the purpose being to apply the same algorithm to both data sets), the SGD classifier and the kernel approximation are discarded. The Naive Bayes methods are discarded due to the “naive” assumption of conditional independence between every pair of features, assumption which is probably not true for our data set (*cf.* Table 3 and Table 7 for the list of features). The three remaining algorithms are SVC (or SVM), K-Neighbors classifier and Ensemble Classifiers (the random forest classifier being is part of this group).

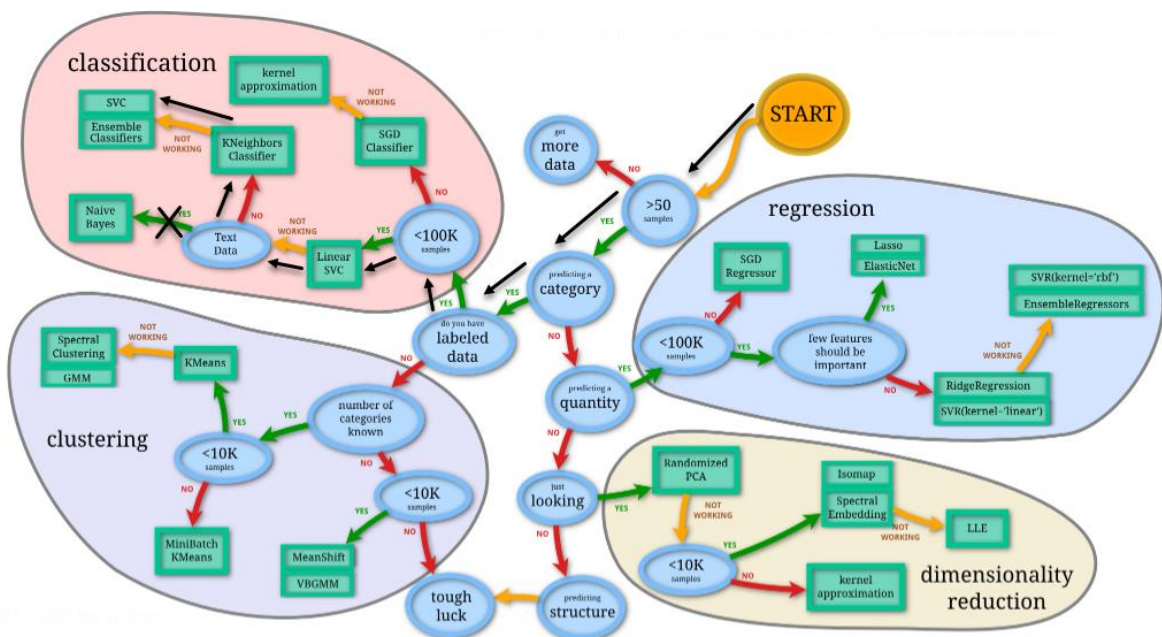


Figure 7: Scikit-learn machine learning diagram
Source: [48]

An important consideration is the fact that all three algorithms do not accept text data as input. As a consequence, the data sets containing text data (which is the case in this research) require an extra step of data preprocessing before being used as input to a machine learning model. The encoding from text categorical variables towards numeric categorical variables can be done via the Python One-Hot-Encoder functionality. This functionality has the advantage of keeping the same size for train and test data [49]. The Figure 8 illustrates how one-hot encoding works on a categorical feature.”

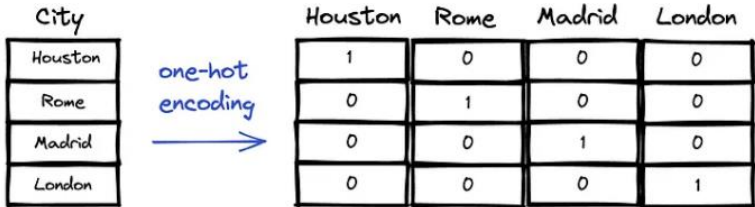


Figure 8: One-hot Encoder principle
Source: [50]

3.1 K-nearest neighbors

The K-nearest neighbors (KNN) classifier is an instance-based learning algorithm, *i.e.* it does not build a general model based on the training dataset [51]. Instead it simply memorizes the training data set and classifies a new data point based on how its neighbors are classified in a multi-dimensional space (each dimension representing a feature of the instance).

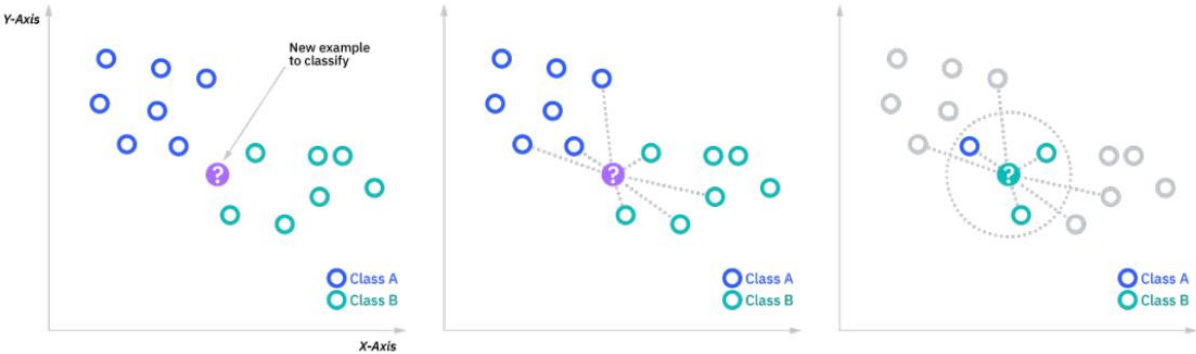


Figure 9: Illustration of K-Nearest Neighbors Algorithm Logic
Source: [52]

The logic is illustrated in Figure 9 for a two-dimensional space. On the left hand-side plot, the training data set is represented by blue and green circles in the two-dimensional space. The blue and green colors represent the two possible classification classes A (blue circles) and B (green circles). When a new instance (purple dot) has to be classified, the algorithm first places it in the two-dimensional space. Then a computational step takes place to determine the distance between the new instance and any point from the training data set (see middle plot from Figure 9). Finally, the K-nearest neighbors (here 3-nearest, see the right-hand side plot) of the training data set are selected before

proceeding to a simple majority vote of the nearest neighbors. As the majority of the neighbors (two out of the three) are part of class B, the new instance is classified as being a member of class B.

Different distance metrics (Manhattan distance, Minkowski distance, Hamming distance, ...) are possible [52], [53]. The most common one being the Euclidean distance and is defined as the distance between two vectors:

$$d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2}$$

Another choice to be made is the value of “K”, the number of nearest neighbors to consider for the majority vote. The optimal “K” value is data-dependent, for example for data sets with outliers, a large K-value would likely better perform. One structured way to determine the K-value is the use of cross-validation consisting in isolating a small part of the training data, referred to as validation set, and using it to assess the results of the algorithm with K-value being successively 1, 2, ..., n. The K-value that results in the best performance on the validation set is a good value for that data set.

A variation of the KNN classifier is a radius-based classifier which follows the same logic except for the determination of the neighbors involved in the majority vote. In such case, by defining a radius instead of the K-value, the algorithm will consider all training instances located within a circle with radius r centered of the new instance to be classified. This alternative is better suited for cases where the data is not uniformly sampled [51].

The main advantage of the KNN classifier is its simple and intuitive logic. However, the computational power required for the classification of each new instance (calculation of the distance from that new instance to all instances of the training data set) is a major inherent drawback.

3.2 Support Vector Classifier

The SVC segregates the classes of a data set by finding an optimal hyperplane in an iterative way. The optimal hyperplane is the one that maximizes the margin, *i.e.* the perpendicular distance between the hyperplane and the closest vectors from each class (referred to as support vectors, *cfr.* Figure 10) [54]. As different hyperplanes can divide the training data set into classes, the algorithm generates hyperplanes in an iterative way and selects the one maximizing the margin. This maximization of the margin acts as a reinforcement so that the new instances are classified with more confidence [55].

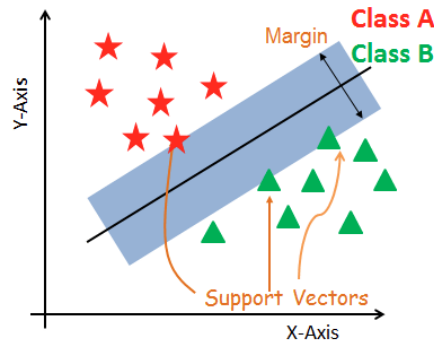


Figure 10: Illustration of margin and support vectors of SVC
Source: [54]

In case the segregation of the classes cannot be achieved with a hyperplane (a linear kernel function), *i.e.* the data are not linearly separable, the SVC can be used with non-linear functions such as polynomial, radial basis function (RBF), sigmoid, ... [56]. This kernel function is used to transform the data and the input space into a higher dimensional space where the segregation of classes via a hyperplane is possible. This approach is referred to as a “kernel trick” and is illustrated in Figure 11. The left-hand side plot displays the initial data that cannot be segregated via a hyperplane in a 2-dimensional space. The right-hand side plot displays the same data after a transformation via a kernel function into a 3-dimensional space (a Z-axis is introduced) where the segregation of the classes via a hyperplane is possible. The determination of the suited kernel function is achieved by trial and error.

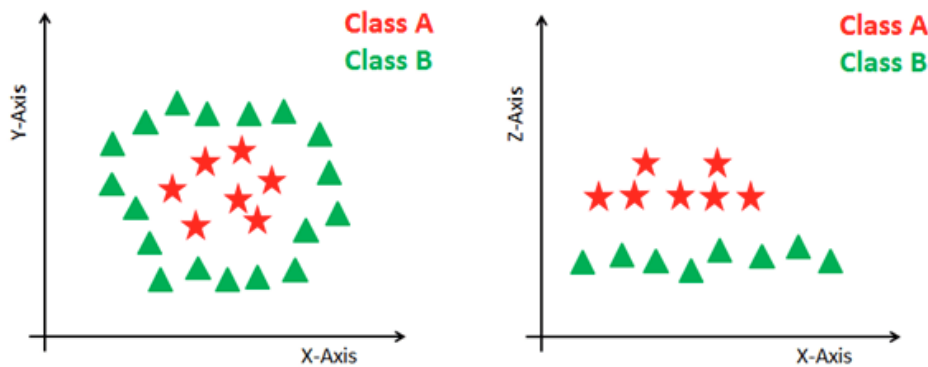


Figure 11: Data transformation to high-dimensional space
Source: [56]

The SVC is well suited for data sets with clear separation (no overlap) between classes. It is efficient with data set of a high dimension (especially when the number of dimension is higher than the number or training instances). In terms of memory, it is more efficient than the NKK because the SVC uses only a sub-set of training instances, the support vectors, for the classification phase. However, due to the high training time of SVC algorithms, they are not recommended for large data sets as the training time increases with the size of the data set. Additionally, for the data set where the number of features is significantly higher than the number of training instances there is an over-fitting risk if the chosen kernel function is not appropriate [57], [58].

3.3 Random forest

The Random forest classifier relies on the decision tree principle which is a series of “if-then-else rules” determined based on the training data. Figure 12 is an example of decision built on a data set having as features the salary, the commute time and free coffee availability, and as classes the acceptance or the decline of the job offer. The model classification model is a decision tree that will determine whether a candidate will accept or decline an offer based on the above mentioned features. The order of the features is important: the “root node” at the top of the tree, *i.e.* the first feature we split on, should be the most informative [59]. The “decision nodes” are the nodes at the origin of further splits, while the “leaf node” refers to the end of a branch (without further splits). The longest path from the root node to a leaf node is referred to as the “depth” of the decision tree.

The decision trees have the advantages of being intuitive and simple to interpret. However, they can result into over-complex trees (over-fitting) failing to generalize the data. Also, they are unstable as small variations in the training data set can result into fundamentally different decision tree model. Moreover, the determination of the optimal decision tree is complex. As a consequence, the decision trees determined in practice are often only local optimum, not global optimum [60].

These unsuitable behaviors can be mitigated by using ensemble methods, *i.e.* by combining several base estimators, in this case decision trees, in order to achieve a better robustness and generalizability compared to a single decision tree [61].

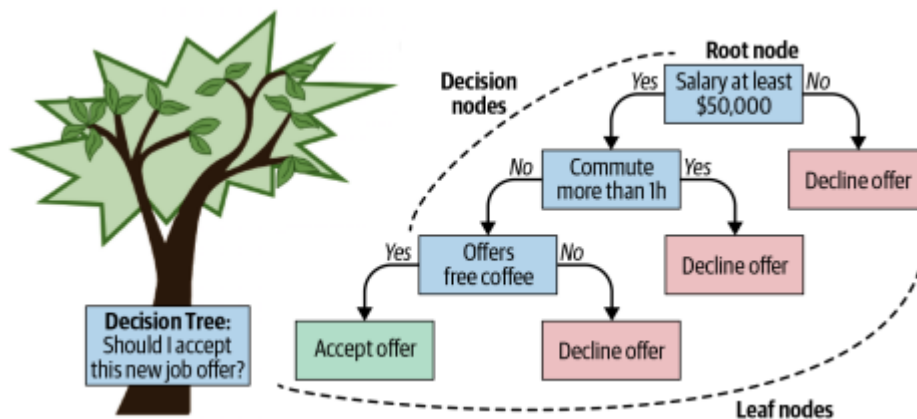


Figure 12: illustration of a decision tree example
Source: [59]

Within ensemble methods two sub-categories of methods are found. The first sub-category encompasses the boosting methods which aggregates several weak base models to create a strong one. The base models are built sequentially, each base model in the sequence is trained with the same data set but where the instances wrongly classified by the previous model are given a more important weight. This way, each base model is focusing its effort on the instances the previous model in the

iteration cannot handle [62]. The second sub-category covers the averaging methods which averages (or uses the majority vote mechanism on) the prediction of individually built base models. This approach considers that the majority vote prediction is probably closer to the true classification than the most of each individual prediction. The random forest classifier is part of this averaging method sub-category.

The above explains the “forest” part of the classifier name (ensemble of decision trees). The “random” part of the name is due to the randomness of the random forest classifier which is characterized by two hyperparameters (*i.e.* parameters that are not determined during the training, but control the structure of the model or the learning process): the bootstrapping and the random feature selection by columns [59]. The bootstrapping consists in sampling with replacement the training data set in order for each decision tree to be trained on a different sub-sample. The objective of this mechanism is to obtain slightly different decision tree models whose predictions will be aggregated at a later stage. This mechanism implies that the decision tree models are correlated, *i.e.* are trained to identify similar patterns in a data set. The random feature selection by columns will mitigate this situation by adding an additional layer of randomness. It consists in considering at each split of the decision tree only a subset of the features (columns). The resulting decision tree models are trained to identify different patterns in a data set. The combination of these models into the random forest ensemble generates a robust classifier with decoupled predictions errors [59], [61].

Figure 13 illustrates the random forest classification principle: the new instance to be classified is passed through each decision tree of the forest (each tree being characterized by the bootstrapping and the random feature selection); each decision tree predicts a class; the majority voting mechanism (or an average) determines the final class. One should note that the Python scikit-learn implementation used in this research determines the final class by averaging the probabilistic prediction of the decision trees (instead of using the majority vote) [61].

The key advantage of random forests is the reduced risk of overfitting induced by the averaging of uncorrelated decision trees. In terms of challenges, depending on the number and complexity of decision trees, this can result in a time-consuming process requiring memory resources to compute the and store the data for each decision tree [63].

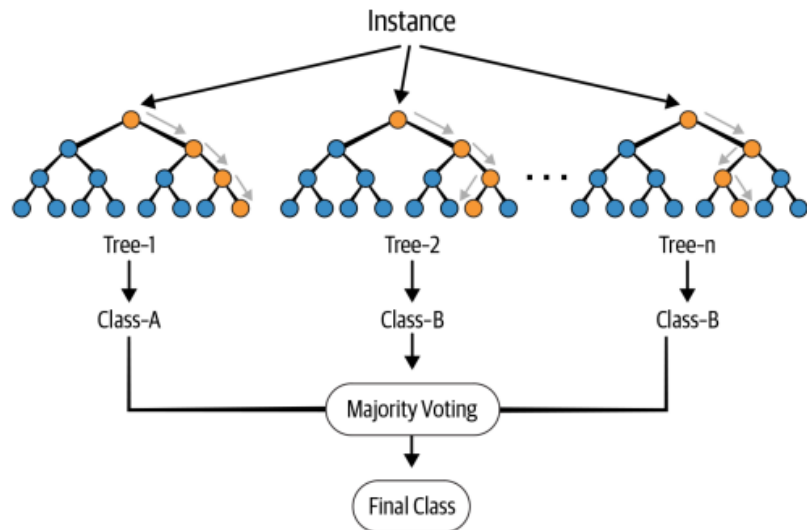


Figure 13: Illustration of random forest classification
Source: [59]

3.4 Classifier selection

Based on the advantages and disadvantages of the classification algorithms described in the above sections, the random forest classifier has been chosen to carry out the experiment of this research. The random forest algorithm is a good combination of an intuitive approach with limited risk of overfitting and reasonable computational power for the processing of the data sets at hand.

4. Research approach

This chapter consists of the different elements of the experimental research conducted in order to answer the question of this master thesis, *i.e.* the identification of the impact of product attributes errors on the auto-classification prediction success. First, the chosen classification algorithm is characterized. Second, the raw data sets and the formatting methods are described. Finally, the methodology used to perform the experiment is detailed.

4.1 Random forest classifier

The principle of the random forest classifier is explained in section 3.3. This section will focus on the practical setup of the classifier as well as on the model evaluation.

4.1.1 Step-by-step process

This section describes step by step how the random forest classifier is used. The first step when doing classification consists in splitting the data set into two sub-sets: one will be used to train the classifier (it will contain a vector X_{train} representing the product features, and y_{train} representing the class or target classification code), one will be used to test the classifier (similarly made of X_{test} and y_{test}). Several parameters allow to influence the split. One of them ensures that the sub-sets are representative of the population. This feature is useful when the initial data set is unbalanced (one or more classes are overrepresented as is the case for one of the batches described in section 4.2).

The second step is actually the creation of a classification model by maintaining some parameters such as the number of trees in the forest (50 in this research) or the number of features to consider when looking for the best split (square root value of 50 in this research), the randomness of the bootstrapping (0.33 in this research).

The third step consists in training the model built above with the train data set determined in the first step. No additional parameters are used at this step.

Finally the fourth step corresponds to the actual testing, the trained model takes as input the features (X_{test}) and predicts the $y_{\text{predicted}}$ classifications.

4.1.2 Model evaluation

This section explains the most common evaluation metrics for a classifier. The objective is to understand which information these metrics reflect in order to proceed with the analysis in chapter 5.

4.1.2.1 Accuracy score

By comparing the $y_{\text{predicted}}$ (determined by the trained model) with the y_{test} (actual labels), it is possible to evaluate the model's performance. The first evaluation score is the accuracy which indicates the percent of correct predictions. In case of imbalanced data set, the accuracy score is not a good measure of performance because it does not provide an insight per class [64]. The model might never predict some classes but due to the overrepresented classes (which are predicted), the accuracy score might still be high.

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total number of predictions}}$$

4.1.2.2 Confusion matrix

The confusion matrix provides a deeper insight as we can look at each class and identify where the model did some mistakes, *i.e.* when it got confused. The rows of the matrix represent the true labels (actual classification codes) and the columns represent the predicted labels (predicted classification codes). In case a model manages to predict perfectly each label, the confusion matrix will have the shape of a diagonal matrix where each element on the diagonal represent the actual and predicted labels ($y_{\text{test}} = y_{\text{predicted}}$), and the rest of the elements, above and below the diagonal, would be filled in with zeros. For the sake of visualization, the confusion matrix can be printed as a heat-map where the color scale reflects how many instances have been classified with a certain classification code.

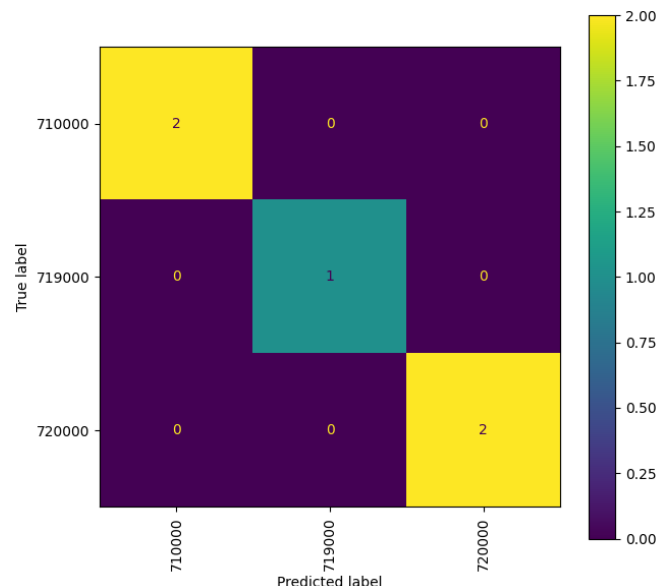


Figure 14: Confusion matrix example

4.1.2.3 Classification report

The classification report provides even more insights into the model performance and the types of errors via additional metrics such as the precision, the recall and the F_1 -score. The precision is the division of the true positives over the sum of true positives and false positives:

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

The recall is the division of the true positives over the sum of true positives and false negatives:

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

In general, a trade-off must be made between the precision and recall as is it difficult to train an algorithm to optimize both metrics simultaneously. In the product classification context, there is no preference for one metric over the other. A false positive would have a similar impact as a false negative. This can be nuanced for the labels corresponding to a higher duty rate. In such case, in terms of compliance and risk, it would be more acceptable to have a false positive (leading to unnecessary duties payment due to the prediction of a code with higher duties than the actual one) than to have a false negative (risk of fines or AEO certification revocation by customs authorities).

The F_1 -score is combining both precision and recall:

$$F_1 = \frac{2}{Recall^{-1} + Precision^{-1}} = \frac{2TP}{2TP + FP + FN}$$

Despite the fact that this metric is less intuitive, it is considered to be a good performance measure as it indicates whether a classifier is good at identifying the member of a class instead of being biased and assigning all instances to a large class.

4.2 Data Sets

For this research two data sets will be used: the first one containing products classified under chapter 71, the second one containing products classified under chapter 62. Each data set is composed of instances of articles features (or attributes) along with their corresponding CN classification. The number of attributes available for each data set is different (*cfr.* Table 3 and Table 7). As the attributes of the products are categorial text variables, the data sets must first undergo an One-Hot-Encoder preprocessing as described in chapter 3 .

4.2.1 Data set “chapter 71”

The first data set contains 111.967 instances of products classified under chapter 71 *“Natural or cultured pearls, precious or semi-precious stones, precious metals, metals clad with precious metal, and articles thereof; imitation jewellery; coin”*. Only the headings 7113, 7116 and 7117 are represented

(*cfr.* entries highlighted in grey in Table 1 for chapter and headings descriptions) by the 5 different CN codes (*cfr.* entries highlighted in grey in Table 2). A short analysis of these two tables, provides already a first insight: the five classification codes cover a relatively small range of the chapter 71, and are hence alike in terms of characteristics. For a non-professional classifier, a first look at the descriptions does not allow to identify the feature that could segregate the different classes.

Chapter 71: Natural or cultured pearls, precious or semi-precious stones, precious metals, metals clad with precious metal, and articles thereof; imitation jewellery; coin	
7101	Pearls, natural or cultured, whether or not worked or graded but not strung, mounted or set; pearls, natural or cultured, temporarily strung for convenience of transport
7102	Diamonds, whether or not worked, but not mounted or set
7103	Precious stones (other than diamonds) and semi-precious stones, whether or not worked or graded but not strung, mounted or set; ungraded precious stones (other than diamonds) and semi-precious stones, temporarily strung for convenience of transport
7104	Synthetic or reconstructed precious or semi-precious stones, whether or not worked or graded but not strung, mounted or set; ungraded synthetic or reconstructed precious or semi-precious stones, temporarily strung for convenience of transport
7105	Dust and powder of natural or synthetic precious or semi-precious stones
7106	Silver (including silver plated with gold or platinum), unwrought or in semi-manufactured forms, or in powder form
7107 00	Base metals clad with silver, not further worked than semi-manufactured
7108	Gold (including gold plated with platinum), unwrought or in semi-manufactured forms, or in powder form
7109 00	Base metals or silver, clad with gold, not further worked than semi-manufactured
7110	Platinum, unwrought or in semi-manufactured forms, or in powder form
7111 00	Base metals, silver or gold, clad with platinum, not further worked than semi-manufactured
7112	Waste and scrap of precious metal or of metal clad with precious metal; other waste and scrap containing precious metal or precious-metal compounds, of a kind used principally for the recovery of precious metal other than goods of heading 8549
7113	Articles of jewellery and parts thereof, of precious metal or of metal clad with precious metal
7114	Articles of goldsmiths' or silversmiths' wares and parts thereof, of precious metal or of metal clad with precious metal
7115	Other articles of precious metal or of metal clad with precious metal
7116	Articles of natural or cultured pearls, precious or semi-precious stones (natural, synthetic or reconstructed)
7117	Imitation jewellery
7118	Coin

Table 1: Headings of chapter 71
Source: [65]

Chapter 71: Natural or cultured pearls, precious or semi-precious stones, precious metals, metals clad with precious metal, and articles thereof; imitation jewellery; coin	
7113	Articles of jewellery and parts thereof, of precious metal or of metal clad with precious metal:
	- Of precious metal whether or not plated or clad with precious metal:
7113 11 00	- - Of silver, whether or not plated or clad with other precious metal
7113 19 00	- - Of other precious metal, whether or not plated or clad with precious metal
7113 20 00	- Of base metal clad with precious metal
...	...
7116	Articles of natural or cultured pearls, precious or semi-precious stones (natural, synthetic or reconstructed):
7116 10 00	- Of natural or cultured pearls
7116 20	- Of precious or semi-precious stones (natural, synthetic or reconstructed):
7116 20 11	- - Necklaces, bracelets and other articles made wholly of natural precious or semi-precious stones, simply strung without fasteners or other accessories
7116 20 80	- - Other
7117	Imitation jewellery:
	- Of base metal, whether or not plated with precious metal:
7117 11 00	- - Cuff links and studs
7117 19 00	- - Other
7117 90 00	- Other

Table 2: CN of interest from chapter 71
Source: [65]

This data set is composed of 10 product attributes. The maintenance of some of these attributes, such as the main material is mandatory, meaning that in the source IS no products can be created without the maintenance of these mandatory attributes (*cf.* Table 3). Other attributes, such as CITES, are optional and should be maintained only if applicable. Due to the dynamics checks implemented in the source IS, each attribute has a predefined number of possible values.

Product Attribute	Possible attribute values	Number of attribute values	Mandatory attribute
Main Material	diamond, rock crystal, mother of pearl (pinctada maxima), pearl (hyriopsis cumingii), gold, silver, platinum, palladium, titanium, steel, aluminium, brass, copper, zamak, cotton, cashmere wool (goat), silk, polyamid, polyester, polypropylen, nylon, wood (common walnut), leather (calf), horn (cow), leather (lamb), leather (cow), horn (water buffalo), ceramic, polymethyl methacrylate.	29	Yes
CITES ¹	pinctada maxima, ara hybrid (feather), meleagrina margaritifera, pteria margaritifera, pinctada fucata martensii, pinctada margaritifera, corallium elatius (branch), corallium konjoi (branch), corallium japonicum (branch), corallium secundum (branch), bubalus bubalis (horn), pinctada spp, hyriopsis cumingii, hyriopsis schlegeli, pteria penguin, pteria sterna, pinctada radiata, peafowl (feather), pinctada fucata, corallium rubrum (branch), tanygnathus megalorynchos (feather), phoeniconaias minor (feather), eudocimus ruber (feather), gallus gallus (feather), cyanocitta cristata (feather), phasianus colchicus (feather), [blank].	26	No
Article type	jewellery, leather goods, accessories, finished other, boutique accessories jewellery.	5	Yes
Article sub-type	ring, bracelet, necklace, earrings, clip, cufflinks, lapel, pendant with necklace, tie bar, bangle, dress stud, belt, belts, bracelet, key ring, usb key, money clip, cufflinks, charms, sunglasses jewellery, tie bar, bangle, brooch, hair accessories, accessories other, sales set, cordon.	27	Yes
Material category	metal, leather, textile, [blank].	3	No
Targeted gender	men's, unisex, women's, [blank].	3	No
Precious stones	no, yes.	2	Yes
Diamond set or mounted	no, yes.	2	Yes
Engagement ring	no, yes.	2	Yes
Central stone	no, yes.	2	Yes

Table 3: Taxonomy of the attributes charactering the first batch (chapter 71)

4.2.1 Data set “chapter 62”

The second data set of 12297 instances contains products classified under chapter 62 “ARTICLES OF APPAREL AND CLOTHING ACCESSORIES, NOT KNITTED OR CROCHETED”. The headings 6202, 6203, 6204, 6206, 6209, 6211, 6212, 6213, 6214, 6215, 6216 and 6217 are represented by 70 different CN codes (*cf.* entries highlighted in grey in Table 5 and Table 6 for the headings and CN descriptions).

An initial comparison between the list of CN represented in the data set and the current nomenclature (*cf.* Table 6), highlights the fact that some CN codes are not part of the current nomenclature anymore.

¹ CITES stand for “Convention on International Trade in Endangered Species of Wild Fauna and Flora” which is an international agreement between governments aiming to ensure that international trade in specimens of wild animals and plants does not threaten the survival of the species [60]. This attribute is maintained in case a product contain a part of a species whose trade is regulated.

One can assume that the missing CN codes are actually obsolete codes (used in former seasonal collections) but were never updated in the IS because these products were not sold anymore. This assumption is supported by the older versions of the ANNEX I to Council Regulation (EEC) No 2658/87.

An analysis of the current and older version of the nomenclature shows a one-to-one relationship between the old and new CN codes. This relationship is highlighted in Table 4. Given the fact that all the corresponding current CN codes are also part of the data set (the data set is mixing both old and new CN codes), the decision is taken to relabel the old values by the new ones instead of filtering out the impacted samples. This approach prevent the reduction of the data set size. After this relabeling, the data set counts 62 unique CN codes instead of 70 previously.

Obsolete CN code	Corresponding current CN code	Number of samples impacted by the replacement
62021100	62022000	157
62021210	62023010	83
62021290	62023090	5
62021310	62024010	47
62021390	62024090	6
62021900	62029000	25
62029200	62023090	5
62061010	62061000	5

Table 4: Relationship between the obsolete and current CN codes

A further analysis of the below Table 5 and Table 6 leads to the following observation: the data set covers a large range of the chapter 62. Even for a non-professional classifier, a first look at the CN descriptions gives an indication regarding the features paying a preponderant role in the segregation of the different classes: the main material (wool, cotton, ...) as well as the sub-category (suits, overcoats, ...) seems to be particularly important.

Chapter 62: ARTICLES OF APPAREL AND CLOTHING ACCESSORIES, NOT KNITTED OR CROCHETED	
6201	Men's or boys' overcoats, car-coats, capes, cloaks, anoraks (including ski-jackets), wind-cheaters, wind-jackets and similar articles, other than those of heading 6203
6202	Women's or girls' overcoats, car-coats, capes, cloaks, anoraks (including ski-jackets), wind-cheaters, wind-jackets and similar articles, other than those of heading 6204
6203	Men's or boys' suits, ensembles, jackets, blazers, trousers, bib and brace overalls, breeches and shorts (other than swimwear)
6204	Women's or girls' suits, ensembles, jackets, blazers, dresses, skirts, divided skirts, trousers, bib and brace overalls, breeches and shorts (other than swimwear)
6205	Men's or boys' shirts
6206	Women's or girls' blouses, shirts and shirt-blouses
6207	Men's or boys' singlets and other vests, underpants, briefs, nightshirts, pyjamas, bathrobes, dressing gowns and similar articles

6208	Women's or girls' singlets and other vests, slips, petticoats, briefs, panties, nightdresses, pyjamas, négligés, bathrobes, dressing gowns and similar articles
6209	Babies' garments and clothing accessories
6210	Garments, made up of fabrics of heading 5602 , 5603 , 5903 , 5906 or 5907
6211	Tracksuits, ski suits and swimwear; other garments
6212	Brassières, girdles, corsets, braces, suspenders, garters and similar articles and parts thereof, whether or not knitted or crocheted
6213	Handkerchiefs
6214	Shawls, scarves, mufflers, mantillas, veils and the like
6215	Ties, bow ties and cravats
6216 00	Gloves, mittens and mitts
6217	Other made-up clothing accessories; parts of garments or of clothing accessories, other than those of heading 6212

Table 5: CN headings chapter 62

Source: [65]

Chapter 62: ARTICLES OF APPAREL AND CLOTHING ACCESSORIES, NOT KNITTED OR CROCHETED	
6202	Women's or girls' overcoats, car-coats, capes, cloaks, anoraks (including ski-jackets), wind-cheaters, wind-jackets and similar articles, other than those of heading 6204 :
6202 20 00	- Of wool or fine animal hair
6202 30	- Of cotton:
6202 30 10	-- Of a weight, per garment, not exceeding 1 kg
6202 30 90	-- Of a weight, per garment, exceeding 1 kg
6202 40	- Of man-made fibres:
6202 40 10	-- Of a weight, per garment, not exceeding 1 kg
6202 40 90	-- Of a weight, per garment, exceeding 1 kg
6202 90 00	- Of other textile materials
6203	Men's or boys' suits, ensembles, jackets, blazers, trousers, bib and brace overalls, breeches and shorts (other than swimwear):
	- Trousers, bib and brace overalls, breeches and shorts:
6203 42	-- Of cotton:
	--- Trousers and breeches:
6203 42 11	---- Industrial and occupational
	---- Other:
6203 42 31	----- Of denim
6203 42 33	----- Of cut corduroy
6203 42 35	----- Other
...	...
6203 49	-- Of other textile materials:
	--- Of artificial fibres:
	---- Trousers and breeches:
6203 49 11	----- Industrial and occupational
6203 49 19	----- Other
	---- Bib and brace overalls:
6203 49 31	----- Industrial and occupational
6203 49 39	----- Other
6203 49 50	----- Other
6203 49 90	--- Of other textile materials
6204	Women's or girls' suits, ensembles, jackets, blazers, dresses, skirts, divided skirts, trousers, bib and brace overalls, breeches and shorts (other than swimwear):
...	...
	- Jackets and blazers:
6204 31 00	-- Of wool or fine animal hair
6204 32	-- Of cotton:
6204 32 10	--- Industrial and occupational
6204 32 90	--- Other
6204 33	-- Of synthetic fibres:
6204 33 10	--- Industrial and occupational
6204 33 90	--- Other
6204 39	-- Of other textile materials:
	--- Of artificial fibres:
6204 39 11	---- Industrial and occupational
6204 39 19	---- Other

6204 39 90	--- Of other textile materials
	- Dresses:
6204 41 00	-- Of wool or fine animal hair
6204 42 00	-- Of cotton
6204 43 00	-- Of synthetic fibres
6204 44 00	-- Of artificial fibres
6204 49	-- Of other textile materials:
6204 49 10	--- Of silk or silk waste
6204 49 90	--- Of other textile materials
	- Skirts and divided skirts:
6204 51 00	-- Of wool or fine animal hair
6204 52 00	-- Of cotton
6204 53 00	-- Of synthetic fibres
6204 59	-- Of other textile materials:
6204 59 10	--- Of artificial fibres
6204 59 90	--- Of other textile materials
	- Trousers, bib and brace overalls, breeches and shorts:
6204 61	-- Of wool or fine animal hair:
6204 61 10	--- Trousers and breeches
6204 61 85	--- Other
6204 62	-- Of cotton:
	--- Trousers and breeches:
6204 62 11	---- Industrial and occupational
	---- Other:
6204 62 31	----- Of denim
6204 62 33	----- Of cut corduroy
6204 62 39	----- Other
	--- Bib and brace overalls:
6204 62 51	---- Industrial and occupational
6204 62 59	---- Other
6204 62 90	---- Other
6204 63	-- Of synthetic fibres:
	--- Trousers and breeches:
6204 63 11	---- Industrial and occupational
6204 63 18	---- Other
	--- Bib and brace overalls:
6204 63 31	---- Industrial and occupational
6204 63 39	---- Other
6204 63 90	--- Other
6204 69	-- Of other textile materials:
	--- Of artificial fibres:
	---- Trousers and breeches:
6204 69 11	----- Industrial and occupational
6204 69 18	----- Other
	---- Bib and brace overalls:
6204 69 31	----- Industrial and occupational
6204 69 39	----- Other
6204 69 50	---- Other
6204 69 90	--- Of other textile materials
...	...
6206	Women's or girls' blouses, shirts and shirt-blouses:
6206 10 00	- Of silk or silk waste
6206 20 00	- Of wool or fine animal hair
6206 30 00	- Of cotton
6206 40 00	- Of man-made fibres
6206 90	- Of other textile materials:
6206 90 10	-- Of flax or ramie
6206 90 90	-- Of other textile materials
...	...
6209	Babies' garments and clothing accessories:
6209 20 00	- Of cotton
6209 30 00	- Of synthetic fibres
6209 90	- Of other textile materials:
6209 90 10	-- Of wool or fine animal hair
6209 90 90	-- Of other textile materials
...	...
6211	Tracksuits, ski suits and swimwear; other garments:
	- Swimwear:
6211 11 00	-- Men's or boys'

6211 12 00	-- Women's or girls'
6211 20 00	- Ski suits
	- Other garments, men's or boys':
6211 32	-- Of cotton:
6211 32 10	--- Industrial and occupational clothing
	--- Tracksuits with lining:
6211 32 31	---- With an outer shell of a single identical fabric
	---- Other:
6211 32 41	----- Upper parts
6211 32 42	----- Lower parts
6211 32 90	--- Other
6211 33	-- Of man-made fibres:
6211 33 10	--- Industrial and occupational clothing
	--- Tracksuits with lining:
6211 33 31	---- With an outer shell of a single identical fabric
	---- Other:
6211 33 41	----- Upper parts
6211 33 42	----- Lower parts
6211 33 90	--- Other
6211 39 00	-- Of other textile materials
	- Other garments, women's or girls':
6211 42	-- Of cotton:
6211 42 10	--- Aprons, overalls, smock-overalls and other industrial and occupational clothing (whether or not also suitable for domestic use)
	--- Tracksuits with lining:
6211 42 31	---- With an outer shell of a single identical fabric
	---- Other:
6211 42 41	----- Upper parts
6211 42 42	----- Lower parts
6211 42 90	--- Other
6211 43	-- Of man-made fibres:
6211 43 10	--- Aprons, overalls, smock-overalls and other industrial and occupational clothing (whether or not also suitable for domestic use)
	--- Tracksuits with lining:
6211 43 31	---- With an outer shell of a single identical fabric
	---- Other:
6211 43 41	----- Upper parts
6211 43 42	----- Lower parts
6211 43 90	--- Other
6211 49 00	-- Of other textile materials
6212	Brassières, girdles, corsets, braces, suspenders, garters and similar articles and parts thereof, whether or not knitted or crocheted:
6212 10	- Brassières:
6212 10 10	-- In a set made up for retail sale containing a brassière and a pair of briefs
6212 10 90	-- Other
6212 20 00	- Girdles and panty girdles
6212 30 00	- Corselettes
6212 90 00	- Other
6213	Handkerchiefs:
6213 20 00	- Of cotton
6213 90 00	- Of other textile materials
6214	Shawls, scarves, mufflers, mantillas, veils and the like:
6214 10 00	- Of silk or silk waste
6214 20 00	- Of wool or fine animal hair
6214 30 00	- Of synthetic fibres
6214 40 00	- Of artificial fibres
6214 90 00	- Of other textile materials
6215	Ties, bow ties and cravats:
6215 10 00	- Of silk or silk waste
6215 20 00	- Of man-made fibres
6215 90 00	- Of other textile materials
6216 00 00	Gloves, mittens and mitts
6217	Other made-up clothing accessories; parts of garments or of clothing accessories, other than those of heading 6212 :
6217 10 00	- Accessories
6217 90 00	- Parts

Table 6: CN of interest from chapter 62:

Source: [65]

This data set is composed of 6 product attributes. Again, the maintenance of some of these attributes, such as the main material is mandatory in the source IS, other attributes, such as CITES, are optional (*cfr.* Table 3).

Product Attribute	Possible attribute values	Number of attribute values	Mandatory attribute
Main Material	cotton, linen, hemp, ramie, wool (sheep), virgin wool (sheep), wool (alpaca), wool (bactrian camel), cashmere wool (goat), silk, viscose, modal, cupro, acetate fibre, polyamide, polyester, polyurethane, nylon, rayon, lyocell, mulberry silk, leather (ostrich), leather (calf), leather (lamb), leather (cow), fleece (lamb), merino wool (sheep), mohair wool (goat), mohair wool (young goat).	29	Yes
CITES	pinctada maxima, bubalus bubalis (horn), camelus bactrianus (wool), trochus niloticus, vicugna pacos (fur), turbo sarmaticus, [blank].	6	No
Article type	clothing accesories, leather goods, apparel / clothing articles, accessories.	4	Yes
Article sub-type	scarves, gloves, ties, handkerchief, cummerbund, bow tie, pareo, clothing accessories other, belt, belts, trousers, coats, shirts, skirts, jackets, jumpsuits, waistcoats, top, dresses, shorts, bodysuit, bathrobe, braces, towel.	24	Yes
Material category	denim, woven, knitwear, leather, jersey, textile.	6	Yes
Targeted gender	children's, men's, women's,	3	Yes

Table 7: Taxonomy of the attributes charactering the second batch (chapter 62)

4.3 Methodology

This section describes the experiments and the conditions under which they have been conducted. Similar types of experiment are conducted successively on both data sets, *i.e.* for the one covering the chapter 71, then for the one covering the chapter 62.

4.3.1 Step 1: training of the classifier

The first part of the experiment consists in training the random forest classifier with the training data as explained in section 4.1.1. This training data is a part of the complete data set available for this research as described respectively in sections 4.2.1 and 4.2.1, *i.e.* without any manipulation of the features except for the One-Hot-Encoder formatting.

4.3.2 Step 2: testing of the classifier on initial data set

The second part of the experiment consists in testing the classifier on the test data (*i.e.* the remaining instances of the unaltered data set that were not used in the training step). The resulting performance metrics serves as a reference for the subsequent experiments where the product features are manipulated.

4.3.3 Step 3: testing of the classifier on data set with missing value

The third part of the experiments consists in introducing errors of type “missing data” (omission of one or several features) before running the random forest classifier. A feature is selected and all its values is cleared out, except a few samples so each possible value is represented once. This representation is needed because the algorithm expects to see in the test data at least once the values he encountered during the training (*cfr.* section 4.1.1 regarding the representation of the initial population in both train and test data). This means for example that when testing the omission of feature “material category” on the data set “chapter 62”, the data set column corresponding to the feature “material category” will be empty except for 6 samples as this feature has 6 possible values (*cfr.* Table 7). This type of manipulations is carried on one feature at a time (to evaluate the impact that particular feature has on the prediction performance), but also by cumulating the features that are cleared in order to determine how many features can be disregarded before a significant drop in performance.

4.3.4 Step 4: testing of the classifier on data set with incorrect value

The fourth and last part of the experiment consists in introducing errors of type “incorrect data” before running the random forest classifier. A feature is selected and one of its values is systematically replaced by another one belonging to the same feature. For example for the feature “main material” and instances of “silver” are being replaced by “gold” (the rest of the data remaining unaltered). Please note the importance for both replaced and replacing values to be part of the same feature (due to the same reason of “representation of the initial population” explained above).

A final note regarding the errors introduced: as the random forest algorithm is used with categorial features, all errors are alike and there is no such thing as “a small” or “a big” error. This can be understood by realizing that the replacement of “silver” by “gold” is not numerously worse or better than replacing “silver” by “platinum”. Although both replacements are equally wrong, they could lead to a different impact on the prediction performance.

5. Research results and analysis

The analysis will consists first in comparing, for each data set, the classification results of the second step (with initial data) with the classification results of the third and fourth step (with modified data). The ultimate goal being the determination of the impact of errors on the classification performance. For the sake of clarity, the experiments are first applied to the data set “chapter 71” as the range on CN codes is smaller and will enable a better visualization of the different steps of the experiment and analysis.

5.1 Data set “chapter 71”

This section will present the outcome of the steps two to four of the experiment for the data set covering the chapter 71 of the nomenclature.

5.1.1 Initial data set

The classification results via the random forest algorithm for the initial data set show a very high accuracy (almost 1, *cfr.* Table 8) reflecting the fact that only 35 test instances out the 36.950 total of test instances are wrongly classified. Knowing the pitfall of the accuracy score (it gives a wrong impression of a high rate of successful predictions when in presence of overrepresented classes), the confusion matrix (*cfr.* Figure 15) helps to identify that in this case we are dealing with a highly imbalanced data set: the class with label 71131900 is overrepresented with 36.075 instances out of the 36.950 total of test instances.

	precision	recall	F_1-score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.98	0.98	0.98
71179000	0.83	0.93	0.88
Accuracy	0.9991		

Table 8: Classifier's performance metrics for initial data set ch. 71.

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

The precision and recall, and the F_1 -score per class are also high (1 or near to 1), except for:

- Class with label 71162080 which is never predicted. In this case the F_1 -score is artificially set to 0 (as it cannot be calculated). The reason why this class is never predicted by the classifier (while test samples corresponding to this CN code do exist) is the lack of features. This statement is inferred based on two elements: the analysis of the CN codes from subheading “Of precious or semi-precious stones (natural, synthetic or reconstructed)” (*cfr.* Table 2), and the additional articles knowledge. After investigation, it turns out that the samples classified under this CN code contain semi-precious stones (such as tiger’s eye or onyx – information found based on product pictures). As the “semi-precious stone” is not part of the product features (covering only the precious stones indicator), the classifier is not able to segregate these samples from the ones corresponding to heading 7113 (jewellery of precious metal) and 7117 (imitation jewellery).

- Class 71179000 which has a lower precision of 0.83 meaning that the classifier has the tendency to assign to this class some samples that actually belong to a different class generating this way false positives. This 71179000 class is wrongly assigned to some instances belonging to classes:
 - o 71162080 due to the missing “semi-precious stone” information reason as explained in the previous paragraph. In total only 9 classification predictions are impacted by this type of error.
 - o 71171900. This type of prediction error, mixing two CN codes from the 7117 heading (imitation jewellery), is less intuitive to understand. Based on the CN descriptions from Table 2, the two types of imitation jewellery differ from each other by the main material (whether or not of base metal). However, as the classifier is considering each feature equally important, a combination of other features (for example a combination of “article sub-type”, “targeted gender”, presence of “CITES” elements features) can play a more important role in the classification process compared to the main material feature. In total 13 classification predictions are impacted by this type of error. In order to avoid such prediction error types, the training data set should display more variability in term of feature combinations. For example, if all training instances with “necklace” (as “article sub-type”), “women’s” (as “targeted gender”), “wood” (as “CITES”) are classified under 71171900 (implying that the “main material” is a metal), the classifier will not be able to classify a similar instance only differing by the “main material” (being different from a metal) under the correct 71179000 class.

To conclude on the classifier model’s performance, the above analysis suggests that the classifier model can reach high prediction rates. Additionally, it provides the following insights about the data set: first, the data set is highly imbalanced and hence the high number of training samples (which is an indicator of good training) is less relevant as it is overrepresenting one class; second, some key features (such as the “semi-precious stone”) to enable a correct classification are missing in the data set; third, classifications 71171900 and 71179000 differing only by one attribute are not represented with sufficient variations of features in the training data set.

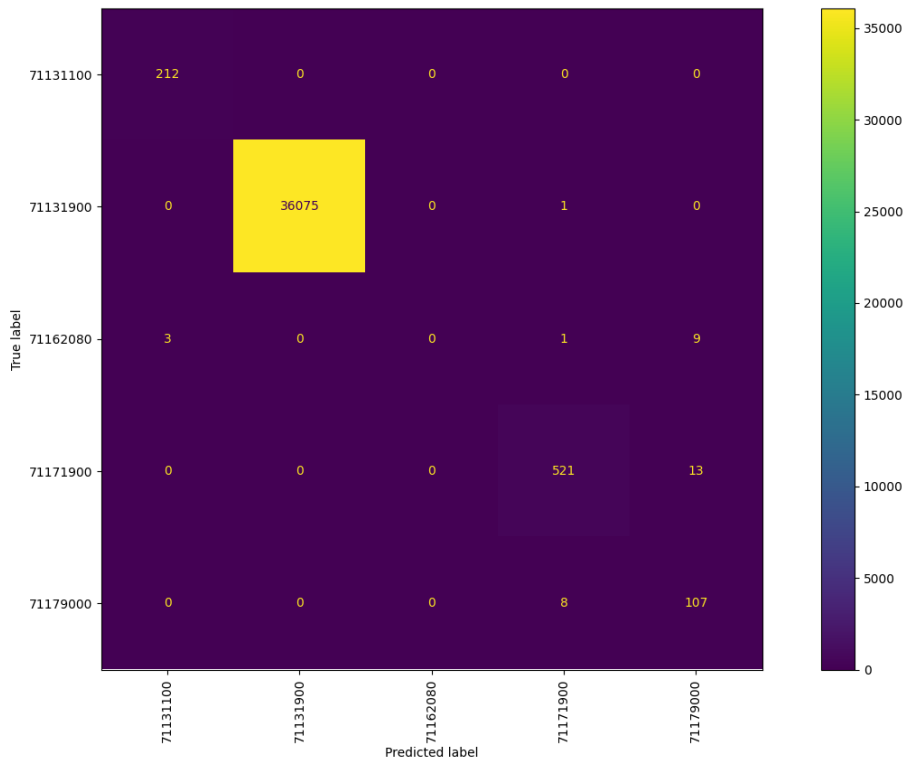


Figure 15: Confusion matrix for initial data set ch. 71

5.1.2 Data set with missing values

After the evaluation of the classifier model, this section aims to look at the performance of the same classifier model on an altered data set. The full data set “chapter 71” has been progressively reduced by removing one feature at a time (starting with the feature at the bottom of Table 3). Below Table 9 to Table 16 show the prediction metrics after each iteration of the data set reduction. For the sake of simplicity, the confusion matrices have not been added as part of this chapter (but can be found in Appendices 8.2) as there is little variation between the classification runs (except for the last run).

5.1.2.1 Cumulative reduction of the data set

After running successive classifications on a data set further reduced at each iteration, one more feature is removed at each run, we can notice that the accuracy score remains higher than 0.99 across the different classification runs. And the precision, recall and F_1 -score remain unchanged for the four first iterations of the data set reduction (*cfr.* Table 9 to Table 12). This suggests that the first four attributes (“center stone”, “engagement ring”, “diamond mounted”, “stones”) do not play a role in the classification process. Hence, omitting the maintenance of such article attributes in an IS has no impact on the auto-classification results via the random forest classifier model used in this research.

However, the F_1 -score start to decrease when the fifth attribute “targeted gender” is removed (*cfr.* Table 13). The impact is limited to classification labels 71171900 and 71179000 which were already

identified as being difficult to differentiate. The decrease in the precision for label 71179000 and the decrease of the recall for 71171900 label indicate that the omission of the “targeted gender” attributed accentuated the initial tendency of the classifier (*cfr.* section 5.1.1) to wrongly classify instances from 71171900 as being part of 71179000 label.

	precision	recall	F_1-score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.98	0.98	0.98
71179000	0.83	0.93	0.88
Accuracy	0.9991		

Table 9: Classifier's performance metrics for data set ch. 71 (without "center stone" feature)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

	precision	recall	F_1-score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.98	0.98	0.98
71179000	0.83	0.93	0.88
Accuracy	0.9991		

Table 10: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring" features)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

	precision	recall	F_1-score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.98	0.98	0.98
71179000	0.83	0.93	0.88
Accuracy	0.9991		

Table 11: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted" features)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

	precision	recall	F_1-score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.98	0.98	0.98
71179000	0.83	0.93	0.88
Accuracy	0.9991		

Table 12: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones" features)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

	precision	recall	F_1-score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.98	0.97	0.97
71179000	0.8	0.94	0.86
Accuracy	0.9989		

Table 13: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender" features)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

The F_1 -score continues to gradually decrease at each iteration of the data set reduction (cumulative removal of "material category" and "article sub-type" features, *cfr.* Table 14 and Table 15). This behavior is again only observed on labels 71171900 and 71179000, while the performance for labels 71131100 and 71131900 is not impacted. This remain true until the 8th feature, the "article type", is cumulatively removed from the test data set. Then, an impact (apparition of false negatives) is observed on label 71131900. Indeed, as per the confusion matrix (*cfr.* Figure 31 in Appendices), instances belonging to label 71131900 are wrongly classified under label 71171900. It is a first indicator that the "article type" is an attribute that plays a role in the classification process.

	precision	recall	F_1-score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.98	0.95	0.97
71179000	0.76	0.93	0.84
Accuracy	0.9987		

Table 14: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category" features)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining

both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

	precision	recall	F_1 -score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.97	0.9	0.94
71179000	0.62	0.88	0.72
Accuracy	0.9978		

Table 15: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type" features)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

	precision	recall	F_1 -score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.97	0.9	0.93
71179000	0.62	0.88	0.73
Accuracy	0.9977		

Table 16: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type", "article type" features)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

When looking at the impact of the cumulative removal of the "CITES" feature, there is no impact observed on the prediction metrics (precision, recall and F_1 -score are identical to the previous run, and the accuracy is still higher than 0.9978). The confusion matrix from Figure 32 only shows one instance being classified differently compared to the previous run (Figure 31).

	precision	recall	F_1 -score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.97	0.9	0.93
71179000	0.62	0.89	0.73
Accuracy	0.9978		

Table 17: Classifier's performance metrics for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type", "article type", "CITES" features)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

5.1.2.2 Isolated reduction of the data set

The isolated reduction of the test data consist in removing one single feature (instead of the cumulative data reduction performed until now): the “main material”. When the main material feature is left out of the test data set (while all other attributes are kept), a major drop is observed (cfr. Table 18):

- In precision (from 0.83 to 0.44) for class 71179000 reflecting an important number of false positives (108 instances from actual class 71171900 being wrongly classified under 71171900). This reinforces the observation made earlier indicating that the classifier needs the material information (whether the imitation jewellery is of base metal or not) to correctly segregate these two classes.
- In recall (from 1 to 0.11) for class 71131100 reflecting an important number of false negatives (188 instances from actual class 71131100 being wrongly classified under 71131900). This important impact can be explained by looking at the corresponding CN description in Table 2. These descriptions indicate that the precious metal (whether it is silver or not) is the attribute enabling the segregation of the two classes. By removing the main material, the precious metal (whether it is silver or not) is lost and the classifier is not able to differentiate the instances of each class.

So far, the main material is the attribute whose omission is impacting the most the auto-classification prediction. The above observations suggest that the main material is one of the driver attributes for the classification process.

	precision	recall	F₁-score
71131100	0.96	0.11	0.19
71131900	0.99	1	1
71162080	0	0	0
71171900	0.94	0.8	0.86
71179000	0.44	0.81	0.57
Accuracy	0.9996		

Table 18: Classifier's performance metrics for data set ch. 71 (without "main material" feature)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F₁-score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

5.1.3 Data set with incorrect values

After looking at the performance of the classifier model on data sets where the maintenance of some features has been omitted, we are looking now at the performance of the same classifier model on data sets where a feature's value is replaced systematically by another value. The full data set “chapter 71” has been successively altered (one alteration at a time) to simulate the impact of a mistake made

during the maintenance of feature's value in an IS. The errors introduced in the initial data set are not an exhaustive list. They are limited to the alteration of two features that have been identified in section 5.1.2 as having an impact on the prediction results: the main material and the article type.

5.1.3.1 Incorrect main material

Given the high amount of instances characterized by "silver" as main material, this "silver" value has been selected as a good candidate to showcase the impact of an incorrect value. The "silver" value has been successively replaced by: "gold" (a precious metal), "steel" (a non-precious metal) and "pearl" (an organic gemstone). The corresponding performance metrics are shown respectively in Table 19, Table 20 and Table 21.

By replacing the "silver" value by "gold" value in the main material feature, a drastic decreased of the recall value from 1 to 0.02 is observed for class 71131100 (*cfr.* Table 19). The analysis of the corresponding confusion matrix (Figure 34) indicates that 207 instances from class 71131100 have been incorrectly classified as being a member of class 71131900. This drop in the prediction success results was expected and is explained by the fact that the two CN codes differ only by their composition (whether made of silver or of other precious metals, *cfr.* Table 2). The complementary precision metric of class 71131900 is only slightly impacted (decrease from 1 to 0.99) as the number of 207 false positives is significantly lower compared to the total amount of instanced part of this class (more than 36.000) reflecting again the bias induced by a overrepresented class.

	precision	recall	F_1-score
71131100	1	0.02	0.04
71131900	0.99	1	1
71162080	0	0	0
71171900	0.98	0.98	0.98
71179000	0.83	0.93	0.88
Accuracy	0.9934		

Table 19: Classifier's performance metrics for data set ch.71 (introduced error: main material "silver" replaced by "gold")
The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

The next two replacements of "silver" value by "steel" then by "pearl" show a similar impact in terms of predictions (*cfr.* Table 20 and Table 21). These results indicate that the errors introduced by the replacement of "silver" by any other material (another precious metal, a non-precious metal or an organic gemstone) do not depend on the replacing value: as soon as the replacing value is different from "silver", the same type of errors are noticed (*cfr.* to Figure 34, Figure 35 and Figure 36 for the visualization).

	precision	recall	F₁-score
71131100	1	0.05	0.09
71131900	0.99	1	1
71162080	0	0	0
71171900	0.98	0.98	0.98
71179000	0.83	0.93	0.88
Accuracy	0.9936		

Table 20: Classifier's performance metrics for data set ch.71 (introduced error: main material "silver" replaced by "steel")
The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F₁-score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

	precision	recall	F₁-score
71131100	1	0.02	0.04
71131900	0.99	1	1
71162080	0	0	0
71171900	0.98	0.98	0.98
71179000	0.83	0.93	0.88
Accuracy	0.9934		

Table 21: Classifier's performance metrics for data set ch.71 (introduced error: main material "silver" replaced by "pearl")
The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F₁-score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

The last two alterations of the main material are not linked to the "silver" value anymore in order to investigate whether other material values play a similar role in the classification. Table 22 and Table 23 display the prediction metrics respectively for a replacement of "domestic calf" value by "pearl", and of "platinum" value by "silk" (representing a change in 3.700 instances). The results suggest that these type of errors do impact the prediction success rate as the overall precision and recall values are decreased. However, the decrease is limited (precision and recall remain higher than 0.8 across all classes) compared to the errors linked to the "silver" value.

	precision	recall	F₁-score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.98	0.97	0.98
71179000	0.8	0.95	0.87
Accuracy	0.9990		

Table 22: Classifier's performance metrics for data set ch.71 (introduced error: main material "domestic calf" replaced by "pearl")
The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F₁-score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

	precision	recall	F₁-score
71131100	0.99	1	0.99
71131900	1	1	1
71162080	0	0	0
71171900	0.98	0.98	0.98
71179000	0.83	0.93	0.88
Accuracy	0.9990		

Table 23: Classifier's performance metrics for data set ch.71 (introduced error: main material "platinum" replaced by "silk")
The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F₁-score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

5.1.3.2 Incorrect article type

As the "article type" has been identified as a second feature impacting the classification, this section is looking at one particular case: the replacement of "jewellery" value by "accessories" under this feature. Table 24 demonstrates that this replacement has a catastrophic impact on the prediction of classes 71131900 (the extremely low recall value of 0.02 indicates a high rate of false negatives) and 71171900 (the low precision value of 0.01 indicates a high rate of false positives). By coupling this result with the CN code descriptions, we understand that the "article type" attribute plays a decisive role in the segregation of the "jewellery (made of precious metals)" CN codes (covering the "jewellery" articles type) from the "imitation jewellery" CN codes (covering "accessories" article type). These errors can be easily visualized in the confusion matrix (cfr. yellow color position on Figure 16 compared to Figure 15) where the majority of instances (35.467) belonging to class 71131900 have been wrongly classified under 71171900 class.

	precision	recall	F₁-score
71131100	1	0.48	0.65
71131900	1	0.02	0.03
71162080	0	0	0
71171900	0.01	0.98	0.03
71179000	0.83	0.93	0.88
Accuracy	0.0362		

Table 24: Classifier's performance metrics for data set ch.71 (introduced error: article type "jewellery" replaced by "accessories")

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F₁-score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

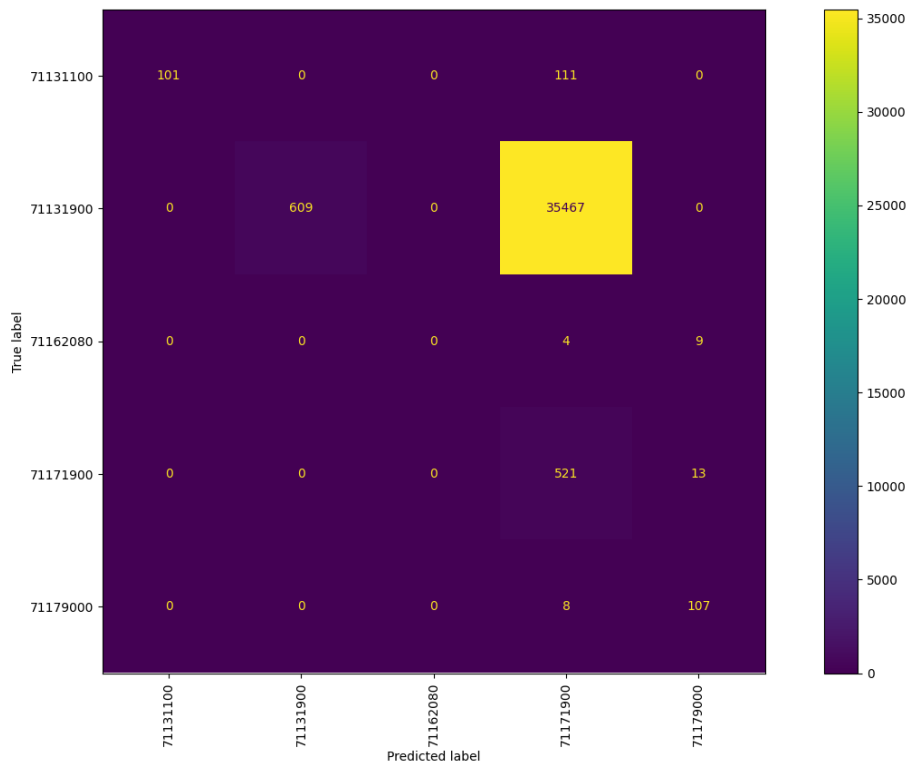


Figure 16: Confusion matrix for data set ch.71 (introduced error: article type "jewellery" replaced by "accessories")

5.2 Data set “chapter 62”

This section presents the outcome of the steps two to four of the experiment for the data set covering the chapter 62 of the nomenclature. Given the high amount of classes (62 CN codes) in this data set, the analysis will be focused on the visualization of the confusion matrices and accuracy instead of on the precision, recall, F_1 -score tables. Only the results corresponding to the most important outcomes are described in this chapter. For the sake of completeness, the remaining predictions metrics results can be found in Appendices 8.3 and 8.4.

5.2.1 Initial data set

The classification results for the initial data set show a high accuracy (higher than 0.91). However, it is lower than the accuracy of the data set “chapter 71”. This can be explained by the fact that the data set “chapter 62” is more balanced (*cfr.* lowest (purple) and highest (yellow) values in the confusion matrix in Figure 17). Although all classes are not evenly represented, the maximum difference in terms of instances between classes is lower than 325 (much lower than the 36.000 in data set “chapter 71”). As a consequence, the accuracy for “chapter 62” data set is less biased than in the previous case and can be considered as a meaningful indicator of the classifier’s performance.

For the sake of visualization, one can first focus on the confusion matrix. A first sight at the confusion matrix (*cfr.* Figure 17) indicates that most of the instances have been correctly classified (non-zero values concentrated on the diagonal). However, some classes have never been predicted (presence of a few “0” values on the diagonal) and some other classes are wrongly classified (some non-zero values next to the diagonal).

For a further analysis, Figure 18 to Figure 21 provide a zoom on each quadrant of the matrix. The upper-left quadrant (Figure 18) displays some classification errors. The two main errors, in terms of number of incorrect predictions, are:

- 132 instances belonging to class 62046239 are incorrectly classified as being part of class 62046231. This type of error is explained by the fact that on top of the “men’s trousers of cotton” (*cfr.* descriptions from Table 6) an additional information is required to differentiate these two CN codes: the type of cotton (denim or other). As this information is not part of the data set features, the segregation of these two classes is not possible.
- 20 instances belonging to class 62043990 are incorrectly classified as being part of class 62029000. The descriptions of these two CN codes are very similar and heading 6202 is defined as the complementary of heading 6204: “6202 - Women's or girls' overcoats, car-coats, capes, cloaks, anoraks (including ski-jackets), wind-cheaters, wind-jackets and similar articles, other

than those of heading 6204". Although, the "article sub-type" feature indicates whether the article is a "coat" or a "jacket", this information is not sufficient to differentiate the two classification possibilities.

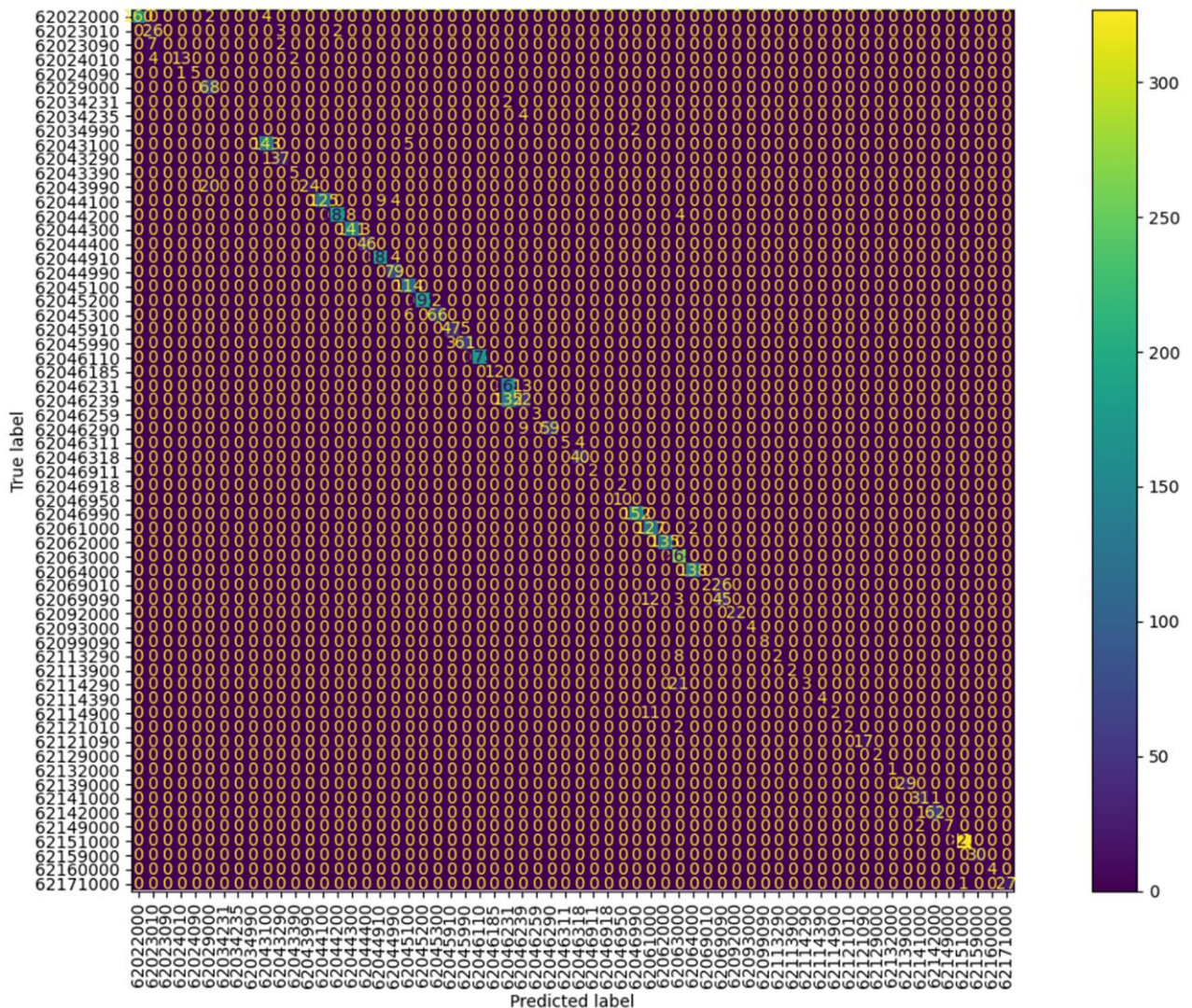


Figure 17: Confusion matrix for initial data set ch. 62

The upper-right quadrant (Figure 19) displays one classification error:

- 2 instances belonging to class 62034990 are incorrectly classified as being part of class 62046990. Again, the header descriptions indicate that both headers refer to the same type or clothes but for a different gender (6203 for women and 6204 for men). Although the "targeted gender" is part of the features and the error in the initial data set has been discarded (all instances of class 62034990 are characterized by "women's"), the classifier could not segregate the two classes. This analysis suggests that the combination of the other features have a preponderant importance compared to the "targeted gender" in the classifier model.

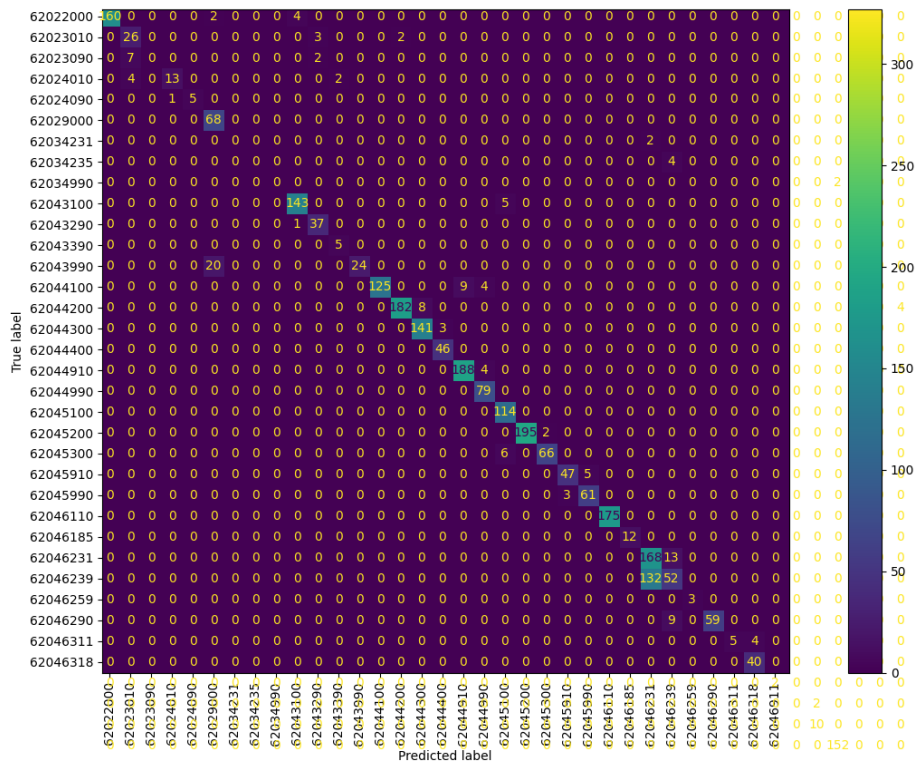


Figure 18: Confusion matrix for initial data set ch. 62, zoom on quadrant 1 (upper-left)

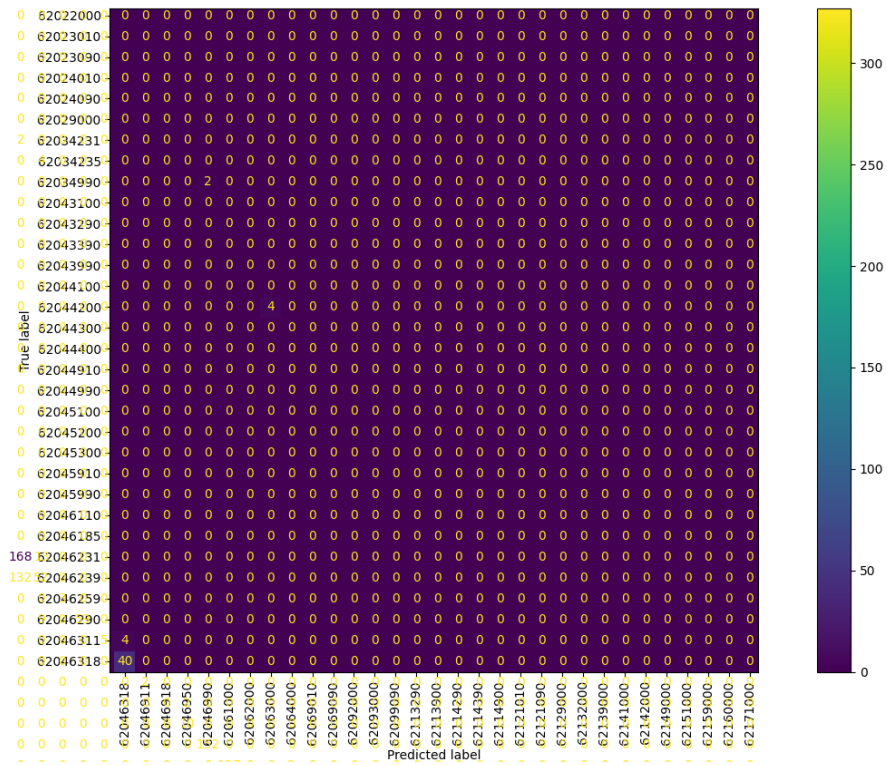


Figure 19: Confusion matrix for initial data set ch. 62, zoom on quadrant 2 (upper-right)

The bottom-right quadrant (Figure 20) displays some classification errors. The two main errors, in terms of number of incorrect predictions, are:

- 26 instances belonging to class 62069010 are incorrectly classified as being part of class 62069090. Both classes are very close to each other as their subheading refers to “Women's or girls' blouses, shirts and shirt-blouses: – Of other textile materials” and the last 2 digits of the CN are based on the type of “other textile materials” (“of flax or ramie” or “of other textile materials”) which is a granularity of information not covered by the available features.
- 21 instances belonging to class 62114290 are incorrectly classified as being part of class 62063000. This error is due to the fact that both share instances with identical values of features. An additional attribute would be required to enable the differentiation of these two classes. Identifying such attribute would also require a deep classification knowledge as the descriptions of the headings do not allow a straightforward decision: “6206 30 - Women's or girls' blouses, shirts and shirt-blouses of cotton” and “6211 42 - tracksuits, ski suits and swimwear; other garments; Other garments, women's or girls'; of cotton”.

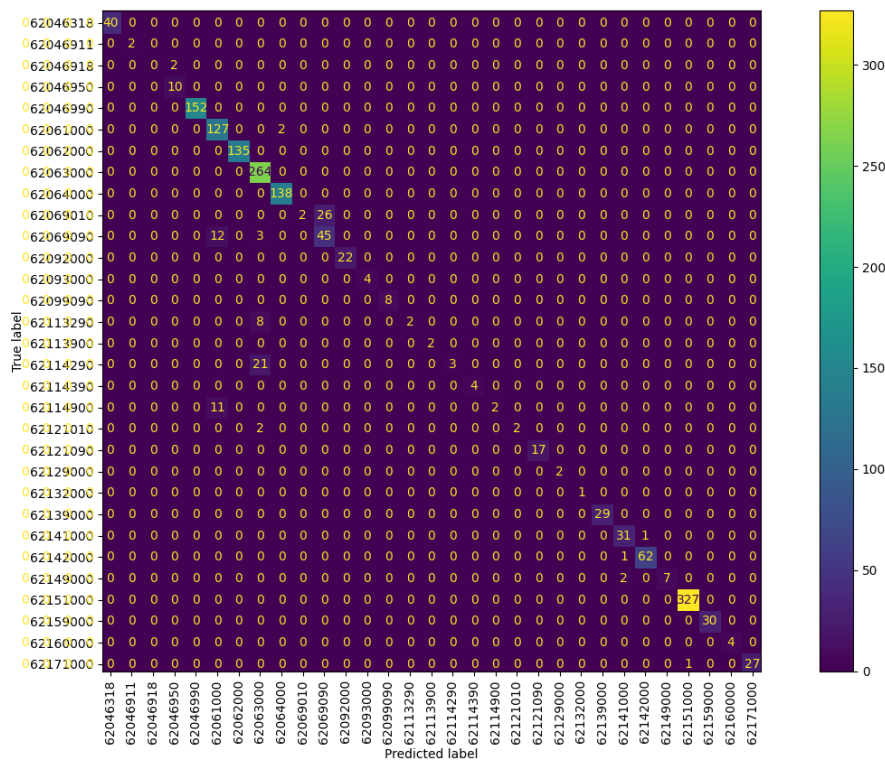


Figure 20: Confusion matrix for initial data set ch. 62, zoom on quadrant 3 (bottom-right)

The bottom-left quadrant (Figure 21) consists only of zero values indicating that there are no errors involving the combination of the “true labels” and “predicted labels” part of this quadrant of the matrix.

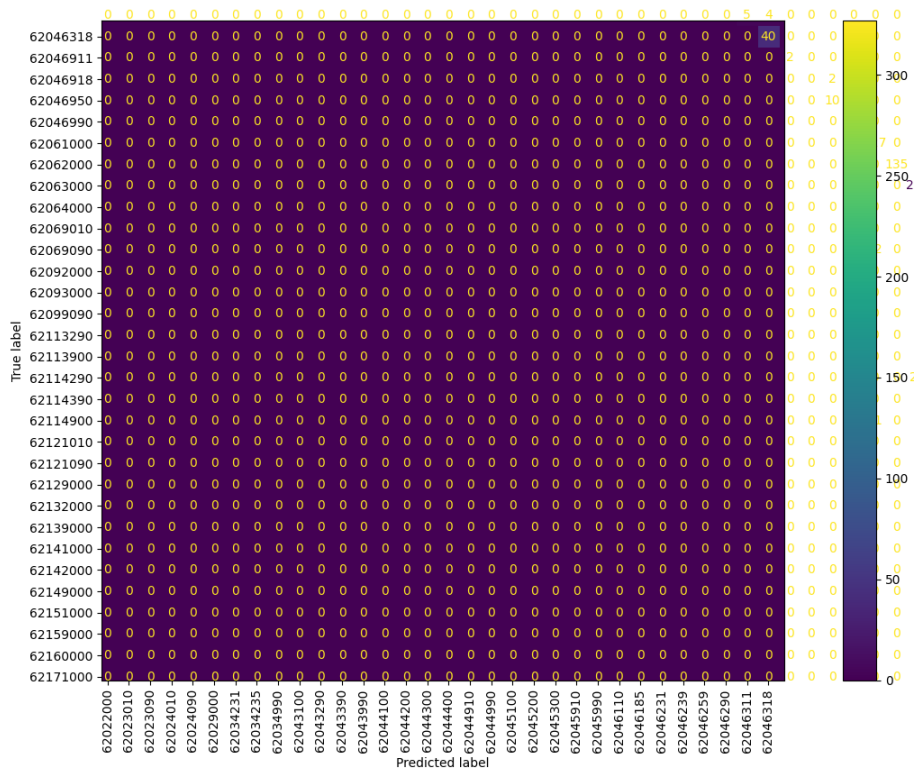


Figure 21: Confusion matrix for initial data set ch. 62, zoom on quadrant 4 (bottom-left)

To conclude on the classifier model’s performance, the above analysis indicates that the classifier model provides good predictions as mainly the diagonal of the confusion matrix is populated. Additionally, although a few classes are underrepresented and five of them are never predicted, this data set is much more balanced than the data set “chapter 71”. Hence, the accuracy metric (higher than 0.91) can be used to assess the classifier model. Moreover, the main error types in terms of occurrences are linked to the “main material” and “article sub-type” features suggesting that these features with their granularities are particularly important for the classification process.

5.2.2 Data set with missing value

For this data set, only isolated reductions of the test data, *i.e.* removal of one single feature at a time (no cumulative reduction), are carried out.

Scenario	Accuracy
Initial data set ch. 62	0.9116
Data set ch. 62 (without "gender" feature)	0.8465
Data set ch. 62 (without "material category" feature)	0.8862
Data set ch. 62 (without "article sub-type" feature)	0.2680
Data set ch. 62 (without "article type" feature)	0.9118
Data set ch. 62 (without "CITES" feature)	0.9017
Data set ch. 62 (without "main material" feature)	0.3321

Table 25: Accuracy of classification for the initial data set ch.62 and reduced data sets scenarios
The accuracy is a performance metric measured on a scale of 0 to 1, it represents the ratio of correct predictions to the total number of predictions. Further details on performance metrics can be found in section 4.1.2.

Table 25 provides an overview of the accuracy evolution for each classification run on a reduced data set. We observe that:

- The “article type” and “CITES” features have a very limited impact on the classification (accuracy is higher to 0.91 and very close to the accuracy for the initial data set). The “article type” feature limited impact can be explained by the fact that the information is too generic and does not provide sufficient granularity for the classification. The “CITES” feature does not seem to play a role at all based on the CN descriptions for this data set.
- The “material category” and “targeted gender” have some impact on the classification as without these features the accuracy is reduced: it drops from 0.91 to a value between 0.84 and 0.88. This was an expected result as the gender (*women’s, men’s*) is one of the key differentiator according to the CN descriptions.; while the “material category” is completing the information part of the “main material” feature.
- The “article sub-type” and the “main material” features have a major impact of the classification: the accuracy drop respectively to 0.26 and 0.33. Again, this was an expected result given the way the CN chapter 62 is structured: the materials and type of clothing articles are key decision factors for the classification. This confirms the findings from section 5.2.1.

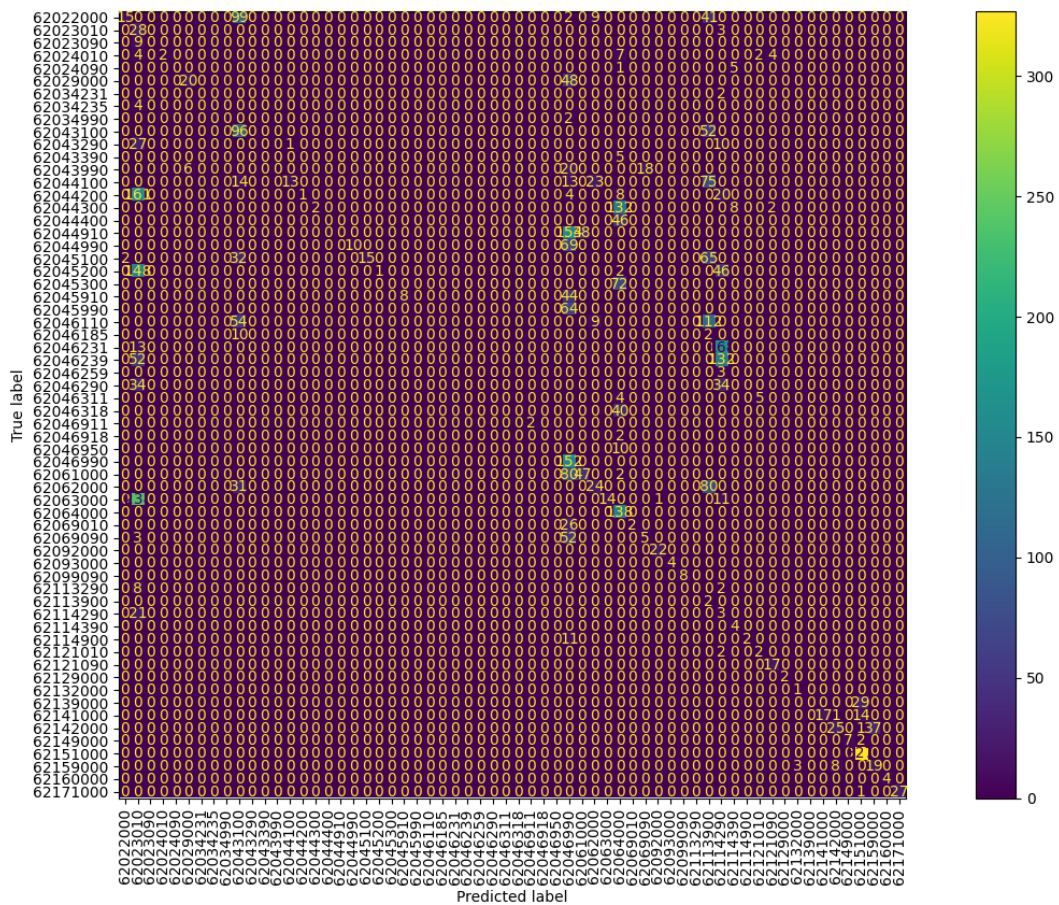


Figure 22: Confusion matrix for data set ch. 62 (without "article sub-type" feature)

A closer analysis of the confusion matrices for the last scenarios with missing “article sub-type” (Figure 22) and the “main material” (Figure 23) features indicates that the omission of these two features generates different types of error: the omission of the “article sub-type” results in a scattered matrix (Figure 22) where the predicted class is often incorrect at heading level, while the omission of the “main material” results into clusters around the diagonal (Figure 23) indicating that the errors are impacting the CN level. This observation is in line with the structure of the nomenclature for chapter 62.

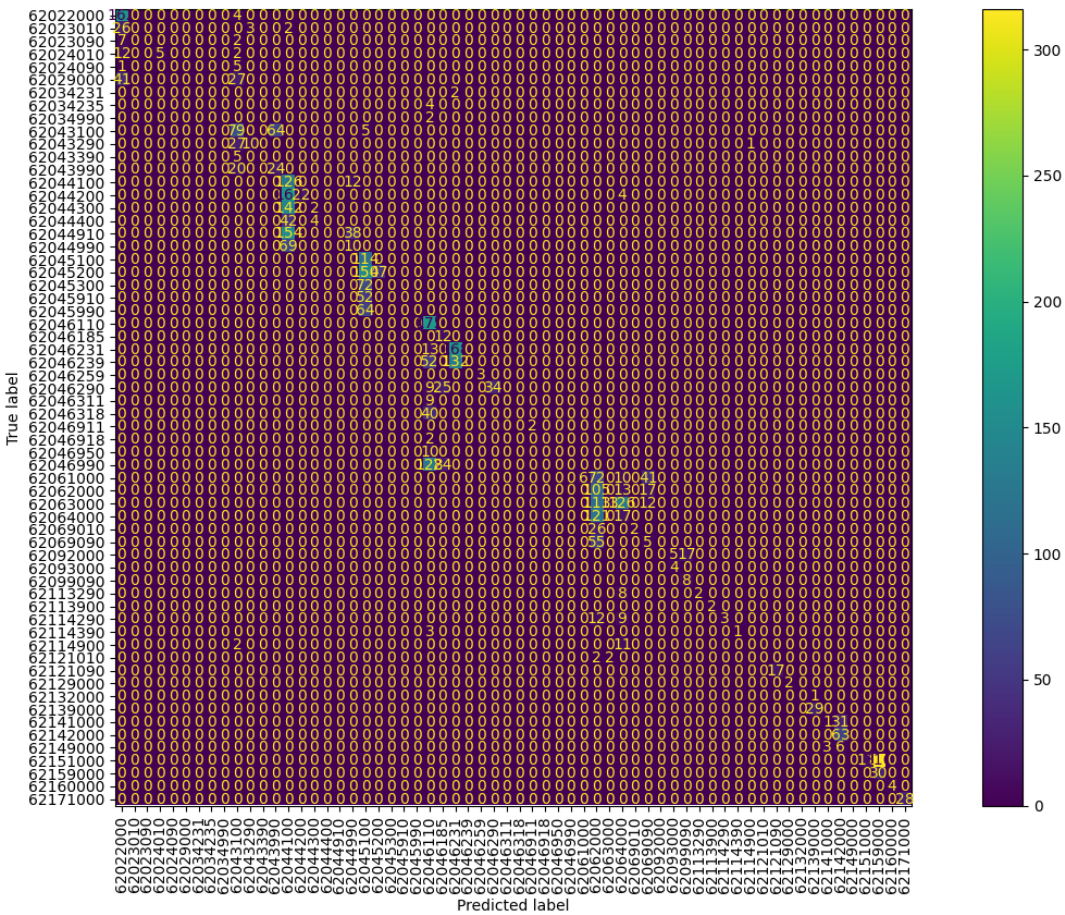


Figure 23: Confusion matrix for data set ch. 62 (without "main material" feature)

5.2.3 Data set with incorrect value

This section is focusing on the performance of the same classifier model on data sets where a feature’s value is replaced systematically by another value. The data set “chapter 62” has been altered (one alteration at a time) to simulate the impact of a mistake made in the “main material” feature (as suggested by the analysis of section 5.2.2).

Table 26 displays the accuracy for each error introduced. It indicates that a mistake replacing the “wool” value by “polyester” or “cotton” does impact the classification performance (accuracy drop to 0.85). The replacement of “cotton” by “silk” has a more significant impact as the accuracy drops lower

to 0.66. Alike the features omission, the incorrect values mainly results in errors at CN level (last two digits of the classification) as illustrated by the confusion matrix in Figure 43 , Figure 44 and Figure 45 where some predictions are located near to the diagonal.

Scenario	Accuracy
Initial data set ch. 62	0.9116
Data set ch. 62 (introduced error: main material "wool" replaced by "polyester")	0.8574
Data set ch. 62 (introduced error: main material "wool" replaced by "cotton")	0.8556
Data set ch. 62 (introduced error: main material "cotton" replaced by "silk")	0.6644

Table 26: Accuracy of classification for the initial data set ch.62 and data sets with errors scenarios
The accuracy is a performance metric measured on a scale of 0 to 1, it represents the ratio of correct predictions to the total number of predictions. Further details on performance metrics can be found in section 4.1.2.

6. Conclusions

6.1 Conclusion research questions

The identification of the impact of data quality on the prediction success of auto-classification tools is addressed in this thesis via four sub-questions. The first sub-question, *“what are the legal rules and inherent complexities driving the product classification?”*, is answered via two elements: the nomenclature and the classification rules. The explanation of the HS nomenclatures and its further developments at national level gives a first insight of the large range of possible codes and highlights the differences between countries purely from a nomenclature definition point of view. The example of soja bean classification at the 11th digit level in Taiwan illustrates the fact that the longer the classification code, the more attributes are usually required to classify a product, and the higher the effort to retrieve and maintain such data. Even if the analysis is limited to the HS or CN level, the terms (wordings) of the headings and sub-headings give an indication of which attributes are required and how specific they must be (*cf.* to the “density” or “gravity” attributes discussed in section 2.4). Next to the nomenclature definition, the explanation of classification rules - the GIRs - highlights the interlinks between the nomenclature, the legal notes, the national regulations, the advance rulings on classification and the court rulings. This high number on considerations to be taken into account to correctly classify products are main the drivers of the classification complexity. Additionally, despite the acknowledgement of these rules at WCO level, when there is room for interpretation each country is free to decide on its own the HS classification resulting in contradictory classifications for products part of everyday life.

The second sub-question, *“what is the required granularity of product attributes to correctly determine the product classification?”*, is answered by the combination on both nomenclature and product knowledge. Classification experts can provide, per type of product, an exhaustive list of criteria (attributes) necessary for the classification. However, the knowledge of the products traded by a company can help to reduce the number of attributes: for example if a women apparel brand is trading exclusively women clothes, there is no need to maintain the targeted gender for each product. It should be noted that the testing phase of an auto-classification tool as described in this thesis (for example by analysis the errors near to the diagonal in the confusion matrix) can also help to identify potentially missing attributes or granularity in the master data such as the “semi-precious stone” information discussed in section 5.1.1 or the “type of cotton (denim or other)” and “other textile materials” (“of flax or ramie” or “of other textile materials”) in section 5.1.1. Additionally, not all available attributes in an IT system are relevant for classification as it is the case for the “CITES” attribute discussed in section 5.1.2.1. Hence, the insights of a classification expert should be requested

to exclude unnecessary attributed an avoid unnecessary processing memory consumption during auto-classification.

The third sub-question, *“what are the possible categories of mistakes in data maintenance”* is addressed in sections 4.3.3 and 4.3.4. as part of the methodology development but also in section 6.3 as part of the limitations. Due to the choice of the classification algorithm, not all type of errors could be simulated in this research, namely the omission (missing value) and the replacement of a value by another value of the same feature (incorrect value) have been studied. Next to these, other types of errors exist such as the replacement of a value by another value belonging to another feature (another type of incorrect value), or the typos occurring during the maintenance of a value in a free text field (without additional checks or pre-defined values). While the occurrence of certain types of error can be reduced or prevented by IT checks (for example by pre-defining a list of possible values), some types of error such as the maintenance of a value by another one belonging to the same feature cannot be prevented.

Finally, the last sub-question, *“what is the impact of the possible categories of mistakes on the prediction success of an auto-classification tool”* is answered by the experiments whose results are detailed in sections 5.1.2, 5.1.3, 5.2.2 and 5.2.3. The first main conclusion drawn based on these results is that many attributes seeming to bring relevant information to the classification (such as the “center stone”, “engagement ring”, “diamond mounted” and “stones” for the data set of chapter 71, or “gender”, “article type” and “CITES” for chapter 62) are playing no, or a limited, role in the auto-classification process as their omission does not impact the prediction metrics. This means that if users miss to maintain these attributes, the auto-classifier will still correctly classify the goods. Moreover, the results suggest that the information driving the auto-classification is concentrated in a few main attributes: the “article sub-type” and “main material” for both data sets of chapter 71 and chapter 62. The second main conclusion is that it is difficult to label the types of errors in terms of impact: in one case the omission of an attribute results in worse prediction metrics than the maintenance of an incorrect value (*cfr.* the “main material” for chapter 62), in another case the maintenance of an incorrect value triggers a drastic drop of the prediction metrics compared to the omission of the same value (*cfr.* “article type” for the chapter 71). Is it hence not possible, based on the results of the data sets at stake, to state whether a certain type of error is worse than another.

To conclude, the master data errors - simulated in this research to emulate different degrees of data quality - do play a role in the prediction success of auto-classification tools when these errors are linked to attributes identified by the algorithm as carrying the information allowing the segregation between classification codes. Depending on the type of products and the chapter they are linked to, the data

quality of a same attribute might play a minor role in classification (and the prediction metrics remain close to 1), or drastically impact it (important drop of the prediction metrics). The determination of these attributes is of course influenced by the nomenclature but also by the data set used to train the classifier. Hence, a proper choice of the training data set is crucial.

6.2 Contribution for research and practice

The research conducted as part of this thesis resulted in three main axes. The first axis is the setup of a goods auto-classification methodology as a proof of concept. The methodology is explained step by step in order to allow the reproducibility of the experiment on different data sets or extend it to other classification algorithms. Additionally, the tools (opensource software and libraries) and parameters (for example the number of trees in the forest, the randomness of the bootstrapping) used to conduct the experiments are also detailed and made available (full code is shared in appendix) allowing any interested parties to reuse them and assess the impact of master data quality on the auto-classification prediction success for their own range of products.

The second axis is the contextualization of existing classification algorithms within the customs domain. Although the tools used are generic classification algorithms based on machine learning, they are used here in the context of goods classification. This context allows customs managers or classification officers - who typically are not machine learning experts - to demystify the technical concepts of auto-classification, understand the requirements of its setup and be aware of its limitations. This thesis contributes to translate the technical concept into accessible content for customs professionals and prepares them for a potential use of auto-classification as a support tool.

Finally, by considering on the one hand, the regulatory requirements of classification and on the other hand, the analysis of the experiment results, this thesis demonstrates the importance of the classification nomenclature knowledge but also of the products knowledge for the initial setup of an auto-classification tool. The contribution of employees with expert knowledge drives the relevance of classification criteria or attributes and the product master data quality which, in turn, impacts the prediction success of auto-classification tools.

6.3 Limitations and future research

The experiment of this research is focused on data sets covering partially CN codes from chapters 62 and 71. Hence, the research results only hold for these CN and cannot be extrapolated to other chapters or even to the full coverage of these chapters. The methodology developed in this research is a proof of concept and should be reproduced with data sets covering a larger range of the

nomenclature in order to draw further insights on master data quality impact on the classifier's performance.

Additionally, the data set "chapter 71" is heavily unbalanced as one of the CN codes is overrepresented compared to the others. In presence of such unbalanced data sets, the classification algorithm might determine an overfitted model resulting into a simplified vision of the data set. By using more balanced data sets covering the same range of CN codes, additional observations could be made.

It should be noted that, due to the algorithm choice, only categorical variables have been used to describe the goods in this research. As a next step, it would be interesting to study also quantitative variables such as size or weight.

In this research all available product attributes have been used to feed the autotool. To avoid high computational memory usage, a Principal Component Analysis (PCA) could be carried out to reduce the dimension of the data set by transforming the highly correlated variables into fewer uncorrelated variables referred to as principal components.

Due to the choice of the random forest algorithm, incorrect values errors could only be generated by replacing one value by another belonging to the same feature (for example, replacing "silver" by "gold" where both values are part of the "main material" feature). Another choice of classification algorithm might allow to study another type of error such as replacing the value of one feature (for example, main material "silver") by the value of another feature (for example, the article sub-type "ring"). Another type of error to be considered in future research would be the typos during the value maintenance in a free text field. These both errors would aim to simulate human maintenance error in the context data maintenance based on SOP.

Further research on the data quality impact on auto-classification prediction success could consider additional data sources on top of the product attributes maintained in an IT system by a company. Such additional data sources could be, in case of CN classification, the EBTI database or the more exhaustive CLASS (Classification Information System) database [66] which is a single access point to different types of classification information (conclusions of the Customs Code Committees, classification regulations, CJEU rulings, CN and CN explanatory notes and TARIC information). Finally, product images could also be processed to generate additional data (attributes) about the products to be classified.

7. Bibliography

- [1] M. Lux and C. Matt, "Classification of Goods: What are the Hurdles and Pitfalls in the Use of Automation or IT Support?," *Global Trade and Customs Journal*, vol. 16, no. 6, pp. 237-247, 2021.
- [2] H. Chen, B. van Rijnsoever, M. Molenhuis, D. van Dijk, Y.-h. Tan and B. Rukanova, "The use of machine learning to identify the correctness of HS Code for the customs import declarations," Piscataway, Oct 06, 2021.
- [3] M. L. Brodie, "Data quality in information systems," *Information & Management*, vol. 3, no. 6, pp. 245-258, 1980.
- [4] P. Oliveira, F. Rodrigues and R. H. Pedro , "A Formal Definition of Data Quality Problems," in *International Conference on Information Quality*, 2005.
- [5] L. Ding, Z. Fan and D. Chen, "Auto-Categorization of HS Code Using Background Net Approach," *Procedia Computer Science*, vol. 60, pp. 1462-1471, 2015.
- [6] "Reuters: Thailand's top court rules Toyota unit must pay \$272 mln in import duties," 15 September 2022. [Online]. Available: <https://www.reuters.com/business/autos-transportation/thailands-top-court-rules-toyota-unit-must-pay-272-mln-import-duties-2022-09-15/>.
- [7] C. Weerth, "Structure of customs tariffs worldwide and in the European Community," *Global Trade & Cust. J.*, vol. 3, p. 221, 2008.
- [8] "WCO: INTERNATIONAL CONVENTION ON THE HARMONIZED COMMODITY DESCRIPTION AND CODING SYSTEM," [Online]. Available: <https://www.wcotradetools.org/en/node/100042>. [Accessed 20 03 2023].
- [9] "WCO: Harmonized System," [Online]. Available: <https://www.wcotradetools.org/en/harmonized-system>.
- [10] "WCO: List of Contracting Parties to the HS Convention and countries using the HS," [Online]. Available: <https://www.wcoomd.org/en/topics/nomenclature/overview/list-of-contracting-parties-to-the-hs-convention-and-countries-using-the-hs.aspx>. [Accessed 20 03 2023].
- [11] E. MacCarthy, "Harmonized System Nomenclature for the Classification of Goods (HS Codes)," 25 August 2022. [Online]. Available: <https://www.westpandi.com/news-and-resources/news/august-2022/harmonized-system-nomenclature-for-the-classificat/>.
- [12] "Taxation and Customs Union: What is the Common Customs Tariff?," [Online]. Available: https://taxation-customs.ec.europa.eu/customs-4/calculation-customs-duties/customs-tariff_en#:~:text=The%20'Common%20Customs%20Tariff'%20,and%20where%20they%20come%20from.. [Accessed July 2023].

- [13] "Association of Southeast Asian Nations: Harmonisation of Tariff Nomenclature, Customs Valuation and Procedures," 3 October 2012. [Online]. Available: <https://asean.org/harmonisation-of-tariff-nomenclature-customs-valuation-and-procedures/>.
- [14] "WCO: CEMAC Tariff Committee gauges progress in implementation of HS 2022 amendments," 1 August 2022. [Online]. Available: <https://www.wcoomd.org/en/media/newsroom/2022/august/cemac-tariff-committee-gauges-progress-in-implementation-of-hs-2022-amendments.aspx>.
- [15] "UNI-PASS CLIP (Customs Law Information Portal) : World HS - HS Comparison," [Online]. Available: <https://unipass.customs.go.kr/clip/index.do>.
- [16] "Taxation and Customs Union: The Combined Nomenclature," [Online]. Available: https://taxation-customs.ec.europa.eu/customs-4/calculation-customs-duties/customs-tariff/combined-nomenclature_en. [Accessed April 2023].
- [17] "Taxation and Customs Union: TARIC Consultation," 2023. [Online]. Available: https://ec.europa.eu/taxation_customs/dds2/taric/taric_consultation.jsp?Lang=en.
- [18] "General Rules For The Interpretation Of The Harmonized System," 16 03 2023. [Online]. Available: https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/nomenclature/instruments-and-tools/hs-interpretation-general-rules/0001_2012e_gir.pdf?la=en.
- [19] G. Grooby, "World Customs organization - Knowledge Academy for Customs and Trade: HS - General rules for interpretation," 1 July 2021. [Online]. Available: https://na.eventscloud.com/file_uploads/409ffa6a9585ec9220bcc22cf70555c3_GaelGROOBY EN.pdf.
- [20] L. Edirisinghe, "Customs Harmonized System Coding; Simplifying the Classification process," *OPA Journal*, vol. 33, pp. 61-64, 2017.
- [21] "WCO: HS CLASSIFICATION HANDBOOK," November 2023. [Online]. Available: http://http://harmonizedsystem.wcoomdpublishings.org/pdfs/WCOOMD_MSH_EN.pdf.
- [22] D.-G. f. T. a. C. U. European Commission, "COMMISSION IMPLEMENTING REGULATION (EU) 2023/1427 of of 4 July 2023 concerning the classification of certain goods in the Combined Nomenclature," *Official Journal of the European Union*, no. 175, pp. 3-5, 2023.
- [23] M. Laszuk, "Rules of obtaining binding tariff information in the EU—analysis of selected problems," *World Customs Journal*, vol. 12, no. 1, pp. 81-90, 2018.
- [24] R. L. F. J. Ploum, "Decisions on binding tariff information (BTI decisions)," 2023. [Online]. Available: <https://ploum.nl/en/expertises-en/practice-areas/customs/Decisions-on-binding-tariff-information-BTI-decisions>.
- [25] "Taxation and Customs Union: Apply for a BTI decision," [Online]. Available: https://taxation-customs.ec.europa.eu/apply-bti-decision_en. [Accessed 2023].

- [26] "Taxation and Customs Union: Validity period of BTI decisions," [Online]. Available: https://taxation-customs.ec.europa.eu/validity-period-bti-decisions_en.
- [27] T. Kawazoe, "Advance Rulings on Tariff Classification: What? Why? Where?," 01 March 2023. [Online]. Available: https://nz.linkedin.com/posts/taichikawazoe_advance-rulings-on-tariff-classification-activity-7039400383315685376-avaJ.
- [28] D.-G. f. T. a. C. U. European Commission, "Administrative Guidance on the Binding Tariff Information Process," 21 December 2018. [Online]. Available: https://taxation-customs.ec.europa.eu/system/files/2023-03/bti_guidance_en.pdf.
- [29] S.-C. Chen, "World Customs Journal:," *In the name of legal certainty? Comparison of advance ruling systems for tariff classification in the European Union, China and Taiwan*, vol. 10, no. 2, pp. 47-64, 2016.
- [30] "European commission: Comitology Register: USTOMS CODE COMMITTEE TARIFF AND STATISTICAL NOMENCLATURE SECTION MINUTES OF THE 220TH MEETING OF THE CUSTOMS CODE COMMITTEE (TEXTILES AND MECHANICAL / MISCELLANEOUS SUB-SECTION)," 29 June 2021. [Online]. Available: <https://ec.europa.eu/transparency/comitology-register/screen/documents/074555/1/consult?lang=en>.
- [31] D.-G. f. T. a. C. U. European Commission, "AUTHORISED ECONOMIC OPERATORS GUIDELINES," 11 March 2016. [Online]. Available: https://taxation-customs.ec.europa.eu/system/files/2017-03/aeo_guidelines_en.pdf.
- [32] WCO, "WCO: How Artificial Intelligence (AI) can help Customs in automating HS Classification," 27 April 2022. [Online]. Available: <https://www.wcoomd.org/en/media/newsroom/2022/april/how-ai-can-help-customs-in-automating-hs-classification.aspx>.
- [33] "BACUDA project: AI HS Code Recommendation Platform," [Online]. Available: <http://49.50.165.5:19090/>. [Accessed 2023].
- [34] "WCO BACUDA Project," [Online]. Available: <https://bacuda.wcoomd.org/>. [Accessed 2023].
- [35] WCO, "BACUDA: Capacity Building Framework for Data Analytics," [Online]. Available: https://bacuda5.files.wordpress.com/2023/04/bacuda_english.pdf. [Accessed 2023].
- [36] WCO, "WCO: WCO BACUDA experts develop a neural network model to assist classification of goods in HS," 03 March 2022. [Online]. Available: <https://www.wcoomd.org/en/media/newsroom/2022/march/wco-bacuda-experts-develop-a-neural-network-model-to-assist-classification-of-goods-in-hs.aspx>.
- [37] R. -. D. o. B. D. & G. T. a. A. Rotchin, "WCO Mag: What impact is technology having on efforts to improve HS classification efficiency and accuracy?," 24 February 2022. [Online]. Available: <https://mag.wcoomd.org/magazine/wco-news-97-issue-1-2022/impact-technology-hs-classification-efficiency-and-accuracy/>.
- [38] Eurostat, "Eurostat: Combined Nomenclature 2023 Search Engine," [Online]. Available: <https://eurostat.prod.3ceonline.com/##current-question-pos>. [Accessed 2023].

- [39] T. Kawazoe, "Classifying HS, based on Material or Function?," 03 09 2020. [Online]. Available: <https://www.customslegaloffice.com/global/classify-hs-based-on-material-or-function/>.
- [40] "U.S. Customs and Border Protection - CUSTOMS RULINGS ONLINE SEARCH SYSTEM (CROSS): H308673," 10 August 2020. [Online]. Available: <https://rulings.cbp.gov/ruling/H308673>.
- [41] "Taxation and Customs Union: BTI Details," 12 July 2022. [Online]. Available: https://ec.europa.eu/taxation_customs/dds2/ehti/ehti_details.jsp?Lang=en&selectedReference=&reference=DEBTI6384/22-1&refcountry=&valstartdate=&valstartdateto=&valenddate=&valenddateto=&suppldate=22/07/2019&nomenc=&nomencto=&keywordsearch=&excludekeywordse.
- [42] "U.S Customs and Border Protection - CUSTOMS RULINGS ONLINE SEARCH SYSTEM (CROSS): H305296: Revocation of NY N249630 and NY N299353; Modification of NY N298787; Classification of pump dispensers from China," 21 January 2020. [Online]. Available: <https://rulings.cbp.gov/ruling/H305296>.
- [43] "CPT - Customs Clearance services: Certificate-free enquiry service / Tax rate inquiry / GC431," 05 March 2015. [Online]. Available: https://portal.sw.nat.gov.tw/APGQ/LoginFree?request_locale=zh_TW&breadCrumbs=JTdCJTlyYnJlYWRDcnVtYnMlMjllM0EINUIIN0IIMjUyYU1JTlYJTNBJTlyJU01JTg1JTJhEJU04JUFEJTg5JU02JTIGJUE1JU04JUE5JUEyJU02JTIDJThEJU01JTJhCJTk5JTlyJTJDJTlydXJsJTlyJTNBJTlyJTlyJTdEJTJDJTdCJT.
- [44] "Taxguru: Goods and Services Tax - Judiciary - HSN Code & GST rate for Plastic Mechanical Liquid Dispenser," 03 July 2020. [Online]. Available: <https://taxguru.in/goods-and-service-tax/hsn-code-gst-rate-plastic-mechanical-liquid-dispenser.html>.
- [45] "Taxation and Customs Union: BTI Details," 16 September 2022. [Online]. Available: https://ec.europa.eu/taxation_customs/dds2/ehti/ehti_details.jsp?Lang=en&selectedReference=DEBTI27378&reference=DEBTI27378/22-1&refcountry=&valstartdate=&valstartdateto=&valenddate=&valenddateto=&suppldate=&nomenc=&nomencto=&keywordsearch=&excludekeywords.
- [46] "U.S Customs and Border Protection - CUSTOMS RULINGS ONLINE SEARCH SYSTEM (CROSS): H305296: Revocation of NY N249630 and NY N299353; Modification of NY H305377; Tariff classification of one-step step stools," 4 October 2022. [Online]. Available: <https://rulings.cbp.gov/search?term=H305377&collection=ALL&sortBy=RELEVANCE&pageSize=30&page=1>.
- [47] T. Kawazoe, "How "Staring wheel cover" is classified under HS code," 21 February 2020. [Online]. Available: <https://www.customslegaloffice.com/global/how-staring-wheel-cover-is-classified-under-hs-code/>.
- [48] "Choosing the right estimator," 18 03 2023. [Online]. Available: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html#.

- [49] "OneHotEncoder," 18 03 2023. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html#sklearn-preprocessing-onehotencoder>.
- [50] S. Yildirim, "Scikit-learn 1.1 Comes with an Improved OneHotEncoder," 22 October 2022. [Online]. Available: <https://towardsdatascience.com/scikit-learn-1-1-comes-with-an-improved-onehotencoder-5a1f939da190>.
- [51] "Scikit-learn - 1.6. Nearest Neighbors," [Online]. Available: <https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification>. [Accessed 12 03 2023].
- [52] "IBM - What is the k-nearest neighbors algorithm?," [Online]. Available: <https://www.ibm.com/topics/knn>. [Accessed 04 03 2023].
- [53] "Towards Data Science: A Simple Introduction to K-Nearest Neighbors Algorithm," 8 June 2019. [Online]. Available: <https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e>. [Accessed 13 01 2023].
- [54] "DataCamp: Support Vector Machines with Scikit-learn Tutorial," 12 2019. [Online]. Available: <https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python>. [Accessed 13 01 2023].
- [55] "Medium: Support Vector Machines(S.V.M) — Hyperplane and Margins," 25 09 2020. [Online]. Available: <https://medium.com/@apurvjain37/support-vector-machines-s-v-m-hyperplane-and-margins-ee2f083381b4#:~:text=The%20hyperplane%20will%20be%20generated,hyperplanes%20that%20could%20be%20chosen..> [Accessed 13 01 2023].
- [56] "IBM: How SVM Works," 17 08 2021. [Online]. Available: <https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works>. [Accessed 13 01 2023].
- [57] "Scikit-learn - 1.4. Support Vector Machines," [Online]. Available: <https://scikit-learn.org/stable/modules/svm.html#classification>.
- [58] "Le Data Scientist: Support Vector Machines (SVM) en python," [Online]. Available: <https://ledatascientist.com/support-vector-machines-svm-en-python/>. [Accessed 13 01 2023].
- [59] J. S. Damji, B. Wenig, T. Das and D. Lee, Learning Spark: Lightning-Fast Data Analytics, Sebastopol: O'Reilly Media,, 2020.
- [60] "Scikit-learn - 1.10. Decision Trees," [Online]. Available: <https://scikit-learn.org/stable/modules/tree.html>. [Accessed 13 01 2023].
- [61] "Scikit-learn - 1.11. Ensemble methods," [Online]. Available: <https://scikit-learn.org/stable/modules/ensemble.html>. [Accessed 13 01 2023].
- [62] "Towards Data Science: Ensemble methods: bagging, boosting and stacking," 23 04 2019. [Online]. Available: <https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205>. [Accessed 13 01 2023].

- [63] "IBM - What is random forest?," [Online]. Available: <https://www.ibm.com/topics/random-forest>. [Accessed 13 01 2023].
- [64] J. Kreiger, "Evaluating a Random Forest model," 13 01 2020. [Online]. Available: <https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56>. [Accessed 16 02 2023].
- [65] "Official Journal of the European Union - COMMISSION IMPLEMENTING REGULATION (EU) 2022/1998 of 20 September 2022 amending Annex I to Council Regulation (EEC) No 2658/87 on the tariff and statistical nomenclature and on the Common Customs Tariff," 20 09 2022. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R1998&from=EN#d1e36-3-1>. [Accessed 31 03 2023].
- [66] "Taxation and Customs Union: CLASS Consultation," 2023. [Online]. Available: <https://webgate.ec.europa.eu/class-public-ui-web/#/search>.
- [67] C. Weerth, "Basic Principles of Customs Classifications under the Harmonized System," *Global trade and customs journal*, vol. 3, pp. 61-67, February 2008.
- [68] R. Rotchin, "What impact is technology having on efforts to improve HS classification efficiency and accuracy? – WCO," *WCO News*, vol. 97, pp. 53-55, 2022.
- [69] *World Customs Organization*.
- [70] A. Naydenov, *File:TARIC code structure.JPG*, 2008.
- [71] G. Li and N. Li, "Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network," *Electronic commerce research*, vol. 19, pp. 779-800, 2019.
- [72] H.-K. Chan, H. Zhang, F. Yang and G. Fischer, "Improve customs systems to monitor global wildlife trade," *Science (American Association for the Advancement of Science)*, vol. 348, pp. 291-292, April 2015.
- [73] "What is CITES?," [Online]. Available: [https://cites.org/eng/disc/what.php#:~:text=CITES%20\(the%20Convention%20on%20International,the%20survival%20of%20the%20species..](https://cites.org/eng/disc/what.php#:~:text=CITES%20(the%20Convention%20on%20International,the%20survival%20of%20the%20species..) [Accessed 30 03 2023].
- [74] "Official Journal of the European Communities - COUNCIL REGULATION (EEC) No 2658/87 of 23 July 1987 on the tariff and statistical nomenclature and on the Common Customs Tariff," 23 07 1987. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31987R2658&from=FR#d1e32-7-1>. [Accessed 20 02 2023].
- [75] H. Chen, M. Molenhuis, Y.-H. Tan and B. Rukanova, "The use of machine learning to identify the correctness of HS Code for the customs import declarations," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, 2021.
- [76] K. Orr, "Data quality and systems theory," *Communications of the ACM*, vol. 41, no. 2, pp. 66-71, 1998.

- [77] "Real Python: The k-Nearest Neighbors (kNN) Algorithm in Python," [Online]. Available: <https://realpython.com/knn-python/>. [Accessed 13 01 2023].
- [78] P. Chakraborty, "Medium: 10 Classification Methods From Scikit Learn We Should Know," 6 01 2021. [Online]. Available: <https://cprosenjit.medium.com/10-classification-methods-from-scikit-learn-we-should-know-40c03ab8b077>. [Accessed 13 01 2023].
- [79] "KSC: Overview of HS Code," [Online]. Available: <https://www.customs.go.kr/engportal/cm/cntnts/cntntsView.do?mi=7311&cntntsId=2333>.
- [80] "Customs clearance services: Waiver enquiry service - Tax rate inquiry (GC431) - Comprehensive inquiry of pre-examination tax cases," 05 March 2015. [Online]. Available: https://portal.sw.nat.gov.tw/APGQ/LoginFree?request_locale=zh_TW&breadCrumbs=JTdCJTlyYnJIYWRDcnVtYnMIMjllM0EINUIIN0IIMjUyYW1JTIyJTlJyJTU1JTg1JTJEJUU4JUFJEJg5JUJ2JTIGJUE1JUJ4JUE5JUEyJUJ2JTIDJThEJUU1JTJCJTk5JTlyJTJDJTIydXJsJTlyJTlJyJTdEJTJDJTdCJT.
- [81] U. S. Census Bureau, "United States Census Bureau Schedule B Search Engine," [Online]. Available: <https://uscensus.prod.3ceonline.com/#!#current-question-pos>. [Accessed 2023].
- [82] "Canada Tariff Finder," [Online]. Available: <https://www.tariffinder.ca/en/>. [Accessed 2023].
- [83] "DE STATIS: Warenverzeichnis Suchmaschine," [Online]. Available: <https://destatis.3ce.com/>. [Accessed 2023].
- [84] D. STATIS, "DE STATIS: Warenverzeichnis Suchmaschine - Online-Suchmaschine für Warennummern," [Online]. Available: <https://www.destatis.de/DE/Methoden/Klassifikationen/Aussenhandel/warenverzeichnis-suchmaschine.html?nn=205976>. [Accessed 2023].

8. Appendices

8.1 Python Code

```
# Code used

import pandas
import os
import numpy
import tkinter
from tkinter import filedialog

from sklearn.feature_extraction import DictVectorizer

def c_open_file():
    from sklearn import svm, datasets
    from sklearn.model_selection import train_test_split
    from sklearn.metrics import ConfusionMatrixDisplay

    #####Import data
    rep = filedialog.askopenfilename()
    print(rep)
    #file = pandas.read_csv(rep, header=0) #explicitely indicates that
the header is the first line
    print("Hello")
    file = pandas.read_csv(rep, low_memory=False) #low_memory=False
argument to False, you're basically telling Pandas not to be efficient,
and process the whole file, all at once
    print(file)
    print("Hello1")
    #print("file.head", file.head(10)) #first 10 of the list
    #print("file.tail", file.tail(10)) #last 10of the list

    X = file.drop(["HS"], axis = 1)
    print("X", X)
    print("typ", type(X))
    from sklearn.preprocessing import OneHotEncoder

    # create an encoder and fit the dataframe #OneHoEncoder is better
than Dummies because keeps the same size for train and test data - see
https://albertum.medium.com/preprocessing-onehotencoder-vs-pandas-get-dummies-3de1f3d77dcc
    enc = OneHotEncoder(sparse_output=False).fit(X) #sparse_output is
important for the format
    print("enc", enc)

    encoded = enc.transform(X)
    print("encoded", encoded)

    # convert it to a dataframe
    X_encoded_df = pandas.DataFrame(encoded, columns =
enc.get_feature_names_out()) #transforms & gives names to columns
    print("X_encoded_df", X_encoded_df)

    head = X_encoded_df.head() #print by default first 5 lines
    print("X_encoded_df(head)", head)

    y= file["HS"]
    print("y", y)
```



```

# On divise en base d'apprentissage et de test :
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X_encoded_df, y,
test_size=0.33, random_state=44, stratify=y)
#I set a random_state; this ensures that if I have to
rerun my code, I'll get the exact same train-test split, so my results
won't change.
#stratify=y. This tells train_test_split to make sure
that the training and test datasets contain examples of each class in the
same proportions as in the original dataset.
#https://medium.com/analytics-vidhya/evaluating-a-random-
forest-model-9d165595ad56

# Puis on cale un modèle d'apprentissage :
from sklearn.ensemble import GradientBoostingClassifier,
RandomForestClassifier
rf_model = RandomForestClassifier(n_estimators=50,
max_features='sqrt', random_state=44)
rf_model.fit(X_train, y_train) #train the model

print("X_train")
print(X_train)
print("y_train")
print(y_train)

y_predictions = rf_model.predict(X_test)
print("X_test")
print(X_test)
print("y_test")
print(y_test)

print(y_predictions)

###Evaluate precision
import matplotlib.pyplot as plt
import seaborn
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report

# View count of each class
y_counts = y.value_counts() #==> list of y with the number of
occurrence ==> classes must be balances to do modeling
#can do this pretty easily with some tools from Imbalanced-Learn.
print("y_counts", y_counts)

#EVALUATION
#https://medium.com/analytics-vidhya/evaluating-a-random-forest-
model-9d165595ad56https://medium.com/analytics-vidhya/evaluating-a-
random-forest-model-9d165595ad56
#for our first evaluation of the model's performance: an accuracy
score.
#This score measures how many labels the model got right out of the
total number of predictions.
# View accuracy score
accuracy = accuracy_score(y_test, y_predictions)
print("Accuracy score:", accuracy) #But remember that accuracy is not
a great measure of classifier performance when the classes are imbalanced

#A confusion matrix is a way to express how many of a classifier's
predictions were correct, and when incorrect, where the classifier got

```

confused (hence the name!). In the confusion matrices below, the rows represent the true labels and the columns represent predicted labels. Values on the diagonal represent the number (or percent, in a normalized confusion matrix) of times where the predicted label matches the true label.

```
# View confusion matrix for test data and predictions
conf_matrix = confusion_matrix(y_test, y_predictions)
print("conf_matrix:", conf_matrix)

###Draw confusion matrix
# Get and reshape confusion matrix data
matrix = confusion_matrix(y_test, y_predictions,
labels=rf_model.classes_) #link with labels to be printed
matrix = matrix.astype('float') / matrix.sum(axis=1)[:,
numpy.newaxis]

from sklearn.metrics import ConfusionMatrixDisplay
color = 'white'
displmatrix = ConfusionMatrixDisplay(confusion_matrix = conf_matrix,
display_labels=rf_model.classes_)
displmatrix.plot()
plt.xticks(rotation = 90)
plt.show()

#Accuracy report
from sklearn.metrics import classification_report
# View the classification report for test data and predictions
print(classification_report(y_test, y_predictions))
print(classification_report(y_test, y_predictions,
labels=numpy.unique(y_predictions)))

c_open_file()
```

8.2 Confusion matrices for data set “chapter 71”

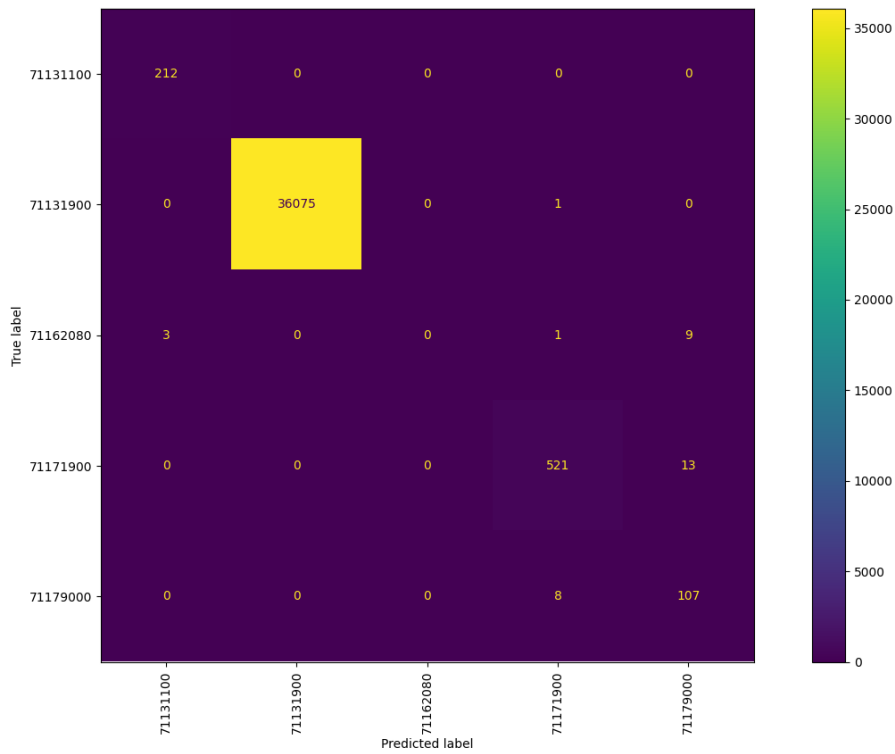


Figure 24: Confusion matrix for data set ch. 71 (without "center stone" feature)

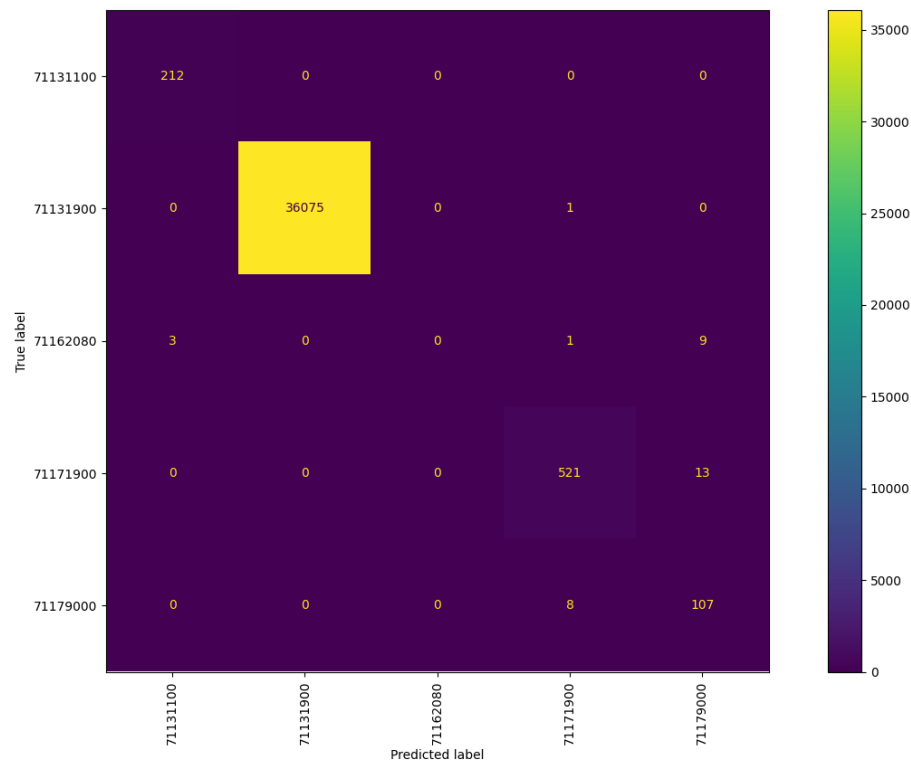


Figure 25: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring" features)

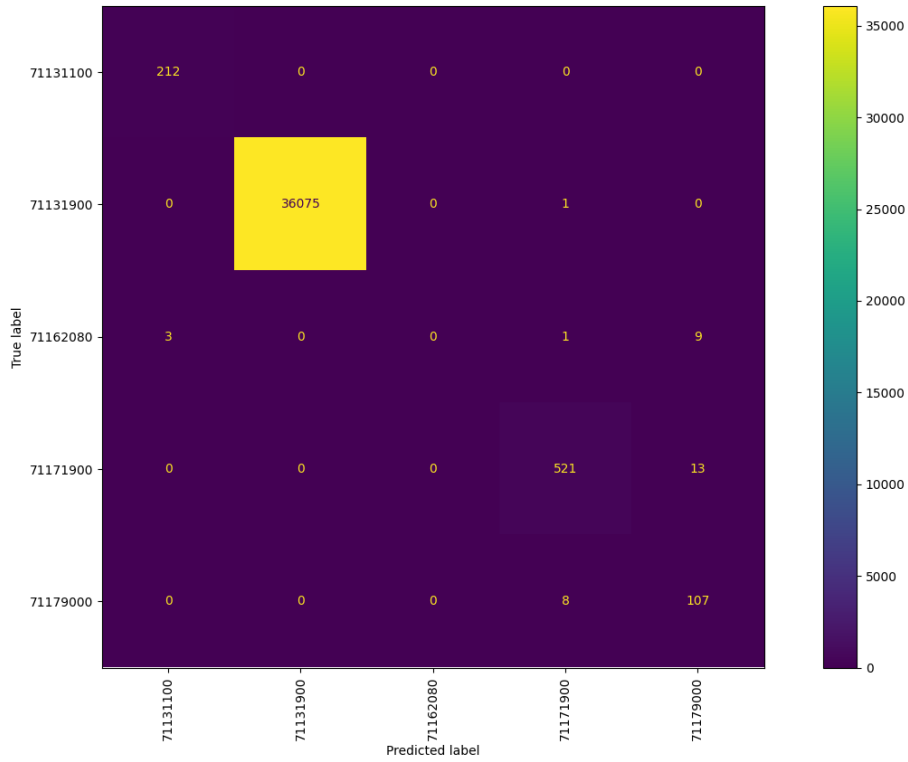


Figure 26: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted" features)

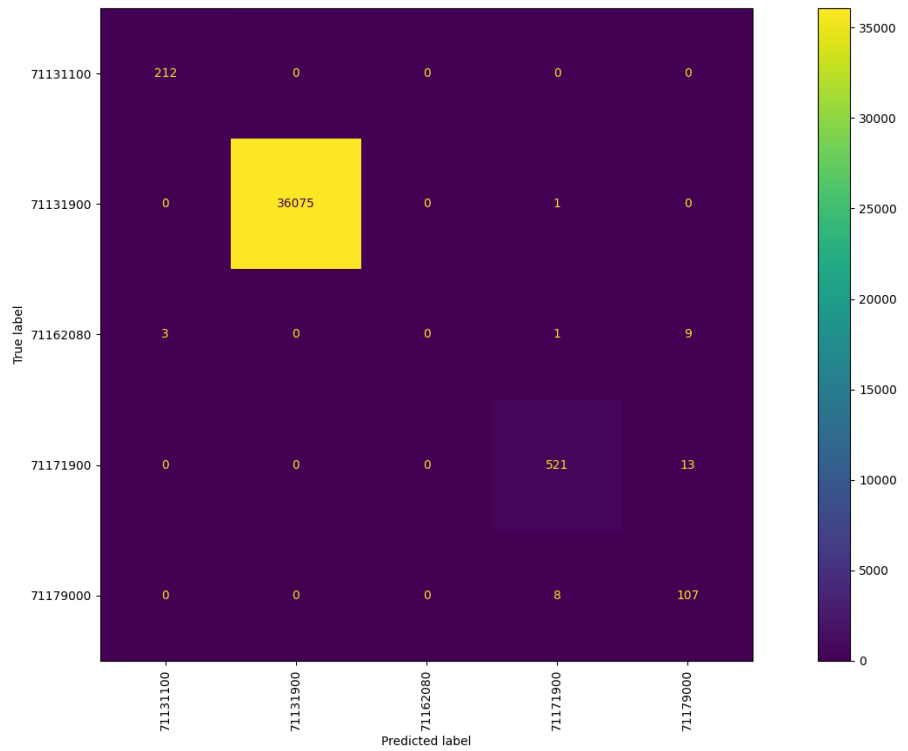


Figure 27: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones" features)

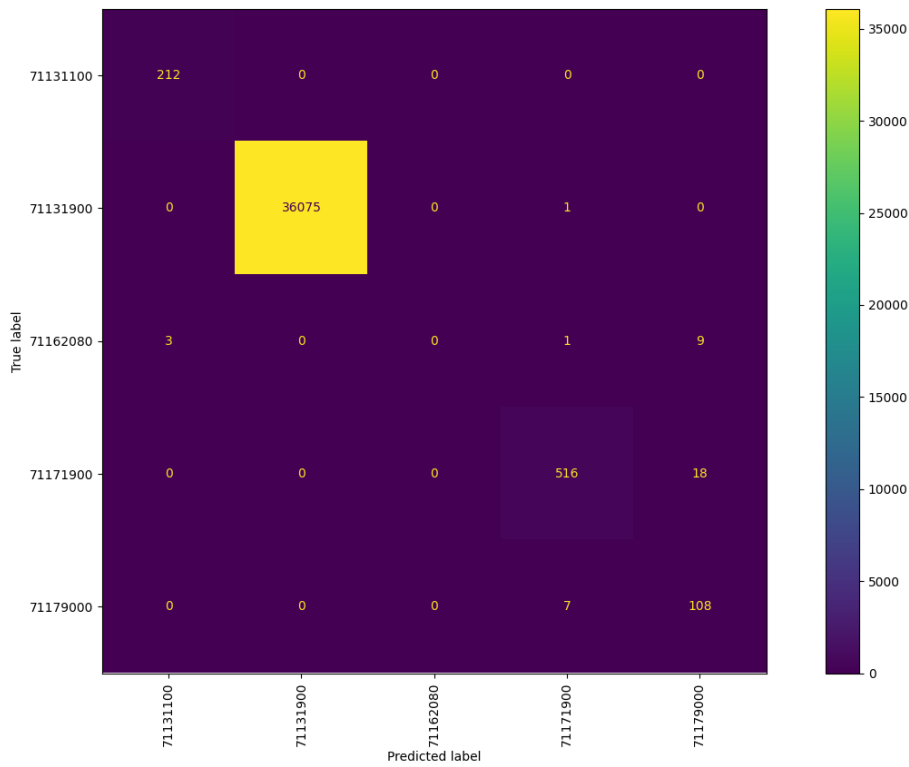


Figure 28: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender" features)

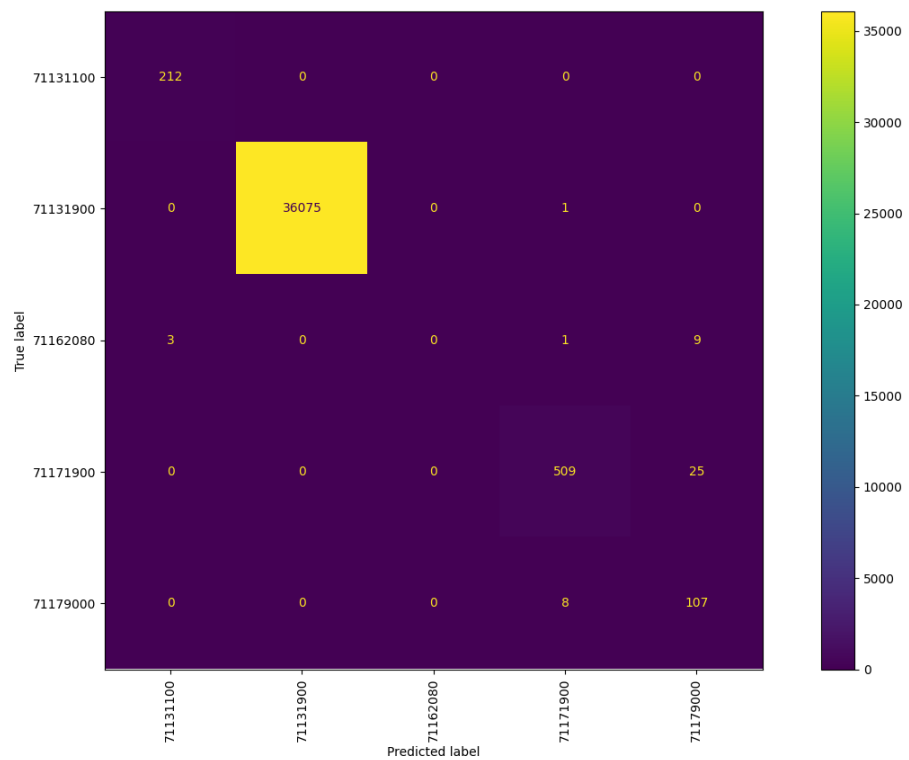


Figure 29: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category" features)

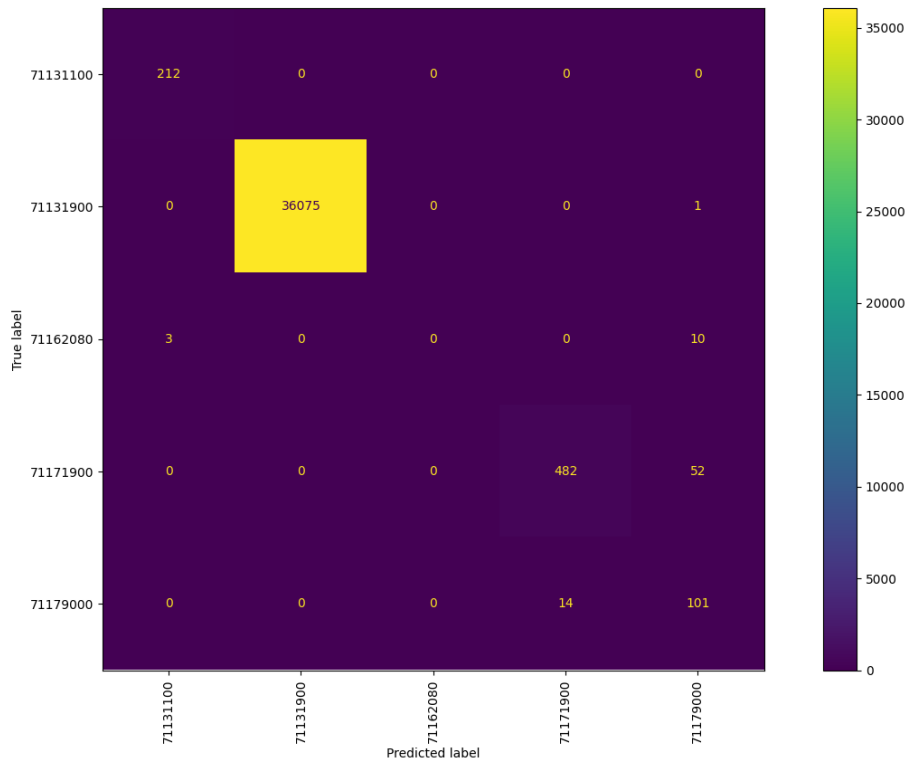


Figure 30: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type" features)

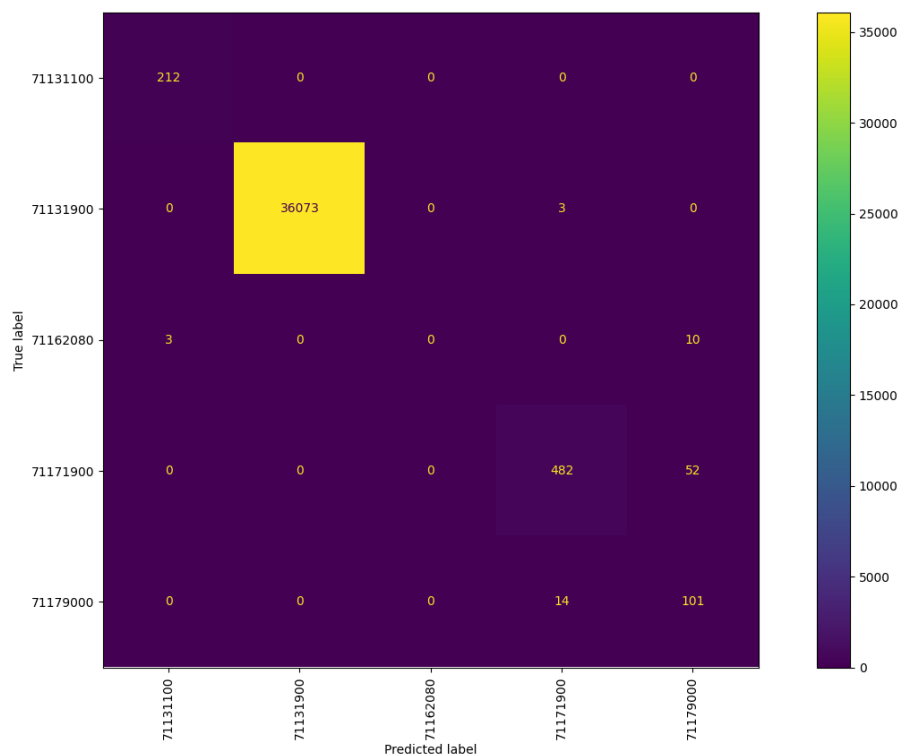


Figure 31: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type", "article type" features)

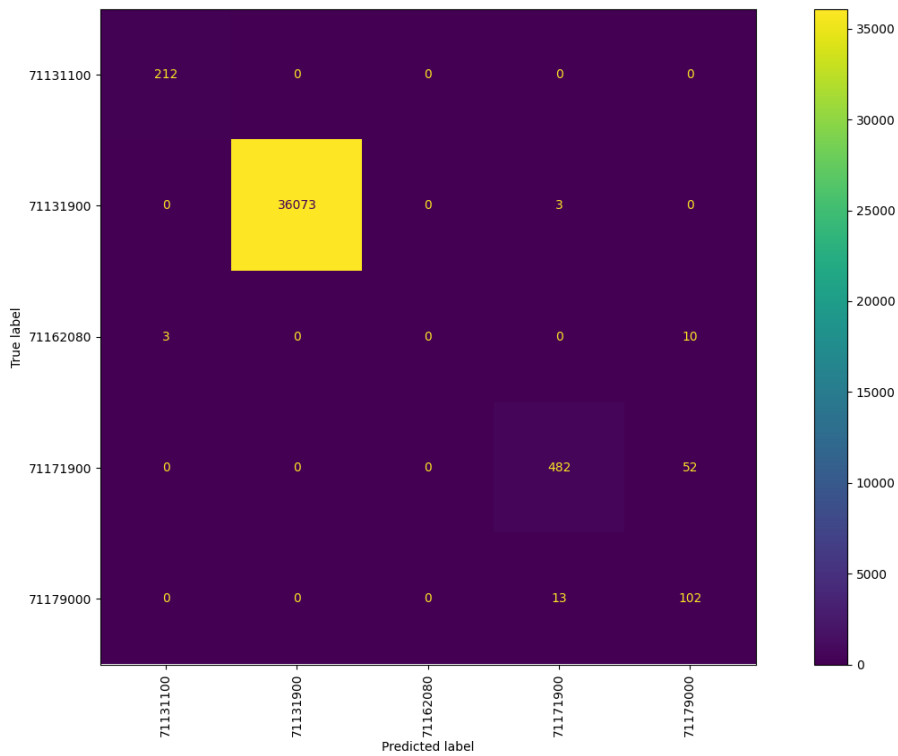


Figure 32: Confusion matrix for data set ch. 71 (without "center stone", "engagement ring", "diamond mounted", "stones", "gender", "material category", "article sub-type", "article type", "CITES" features)

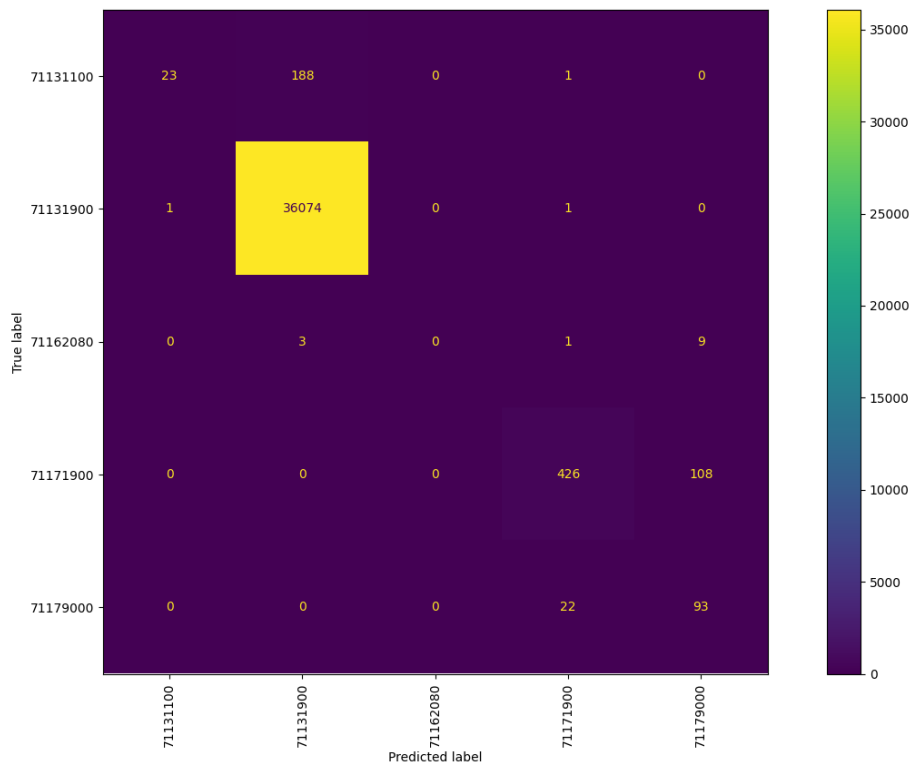


Figure 33: Confusion matrix for data set ch. 71 (without "main material" feature)

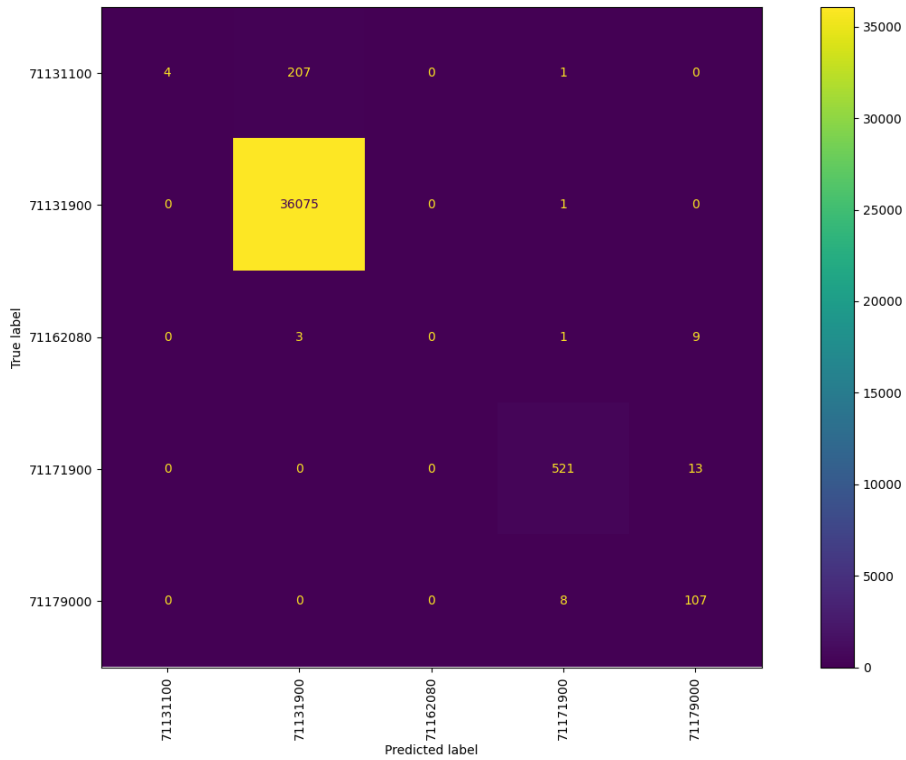


Figure 34: Confusion matrix for data set ch.71 (introduced error: main material "silver" replaced by "gold")

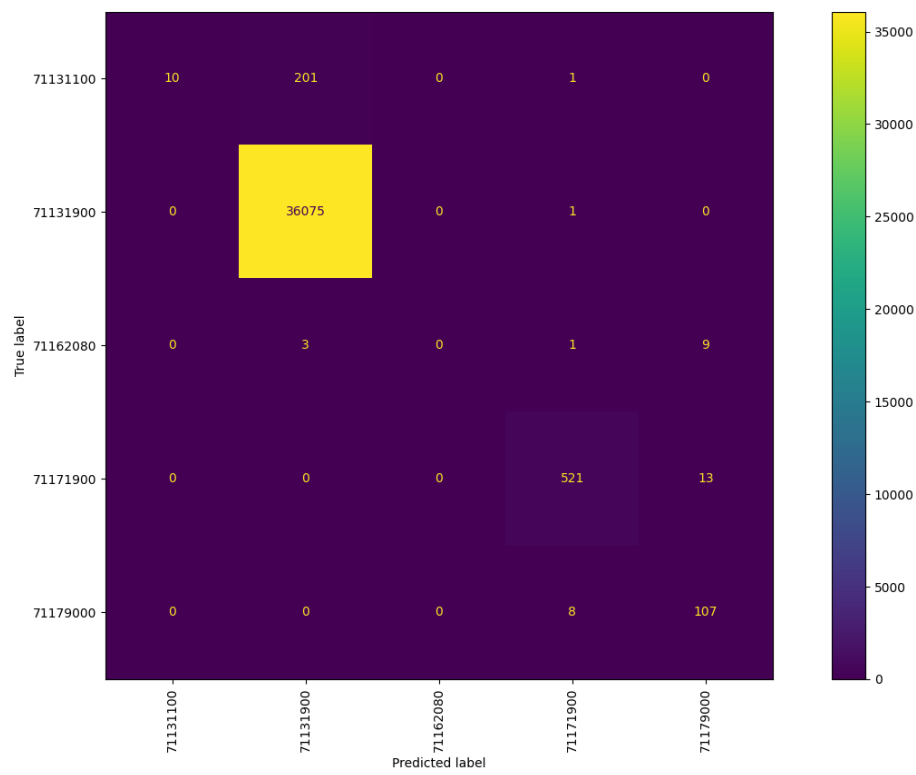


Figure 35: Confusion matrix for data set ch.71 (introduced error: main material "silver" replaced by "steel")

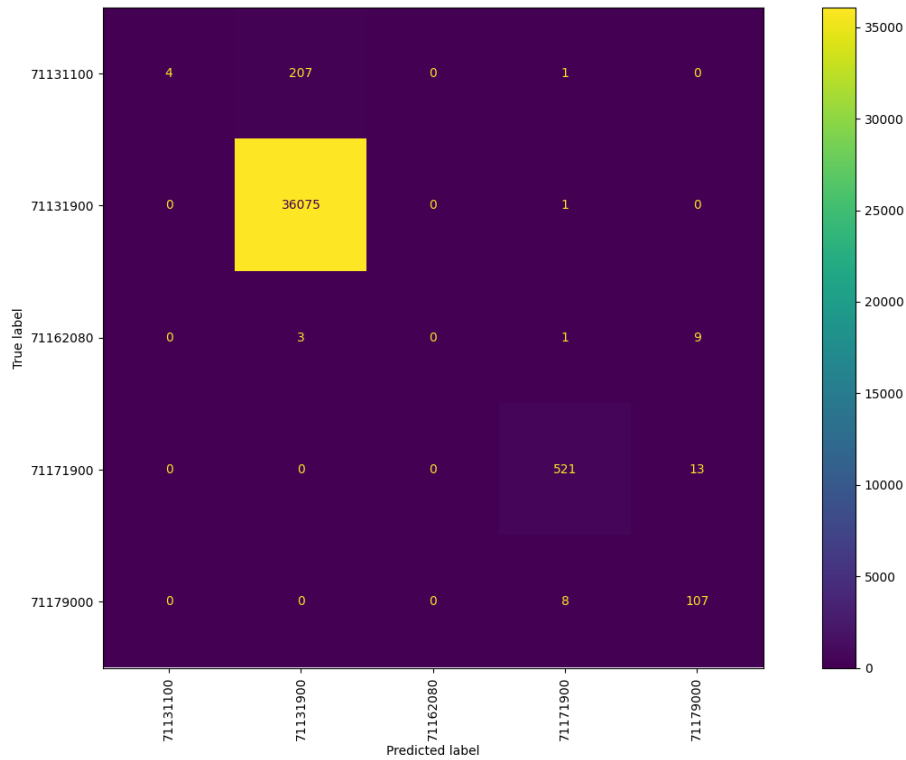


Figure 36: Confusion matrix for data set ch.71 (introduced error: main material "silver" replaced by "pearl")

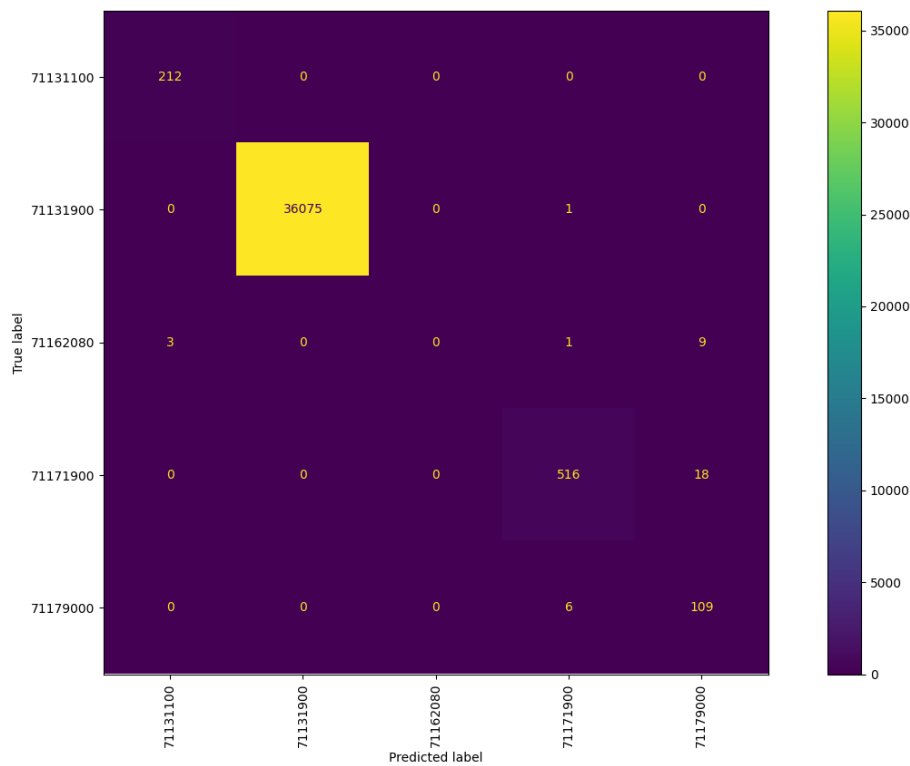


Figure 37: Confusion matrix for data set ch.71 (introduced error: main material "domestic calf" replaced by "pearl")

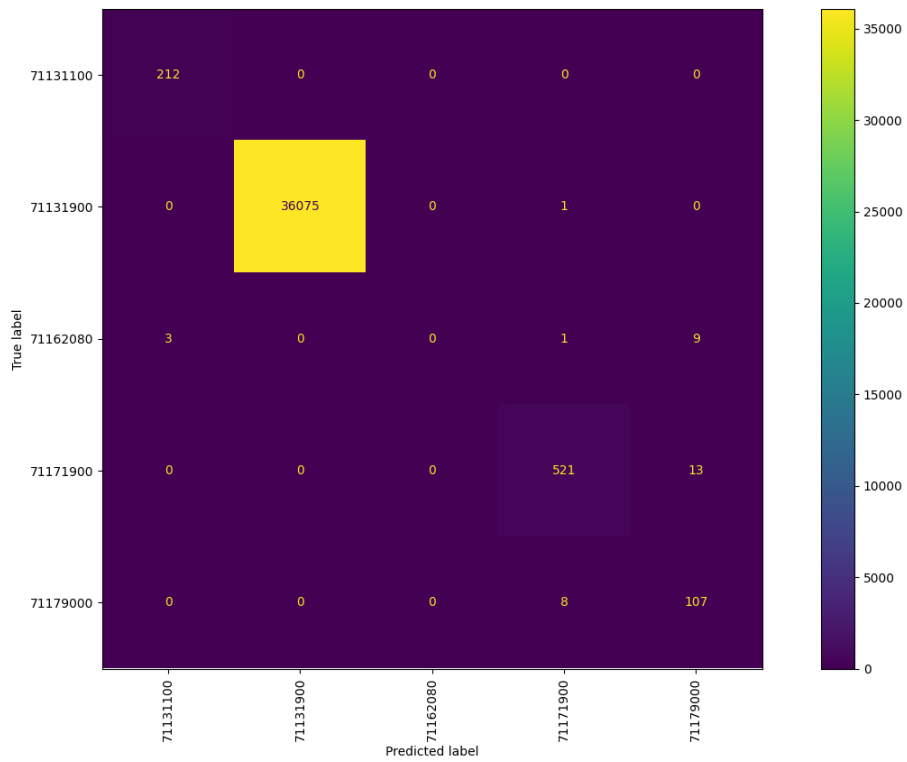


Figure 38: Confusion matrix for data set ch.71 (introduced error: main material "platinum" replaced by "silk")

8.3 Confusion matrices for data set “chapter 62”

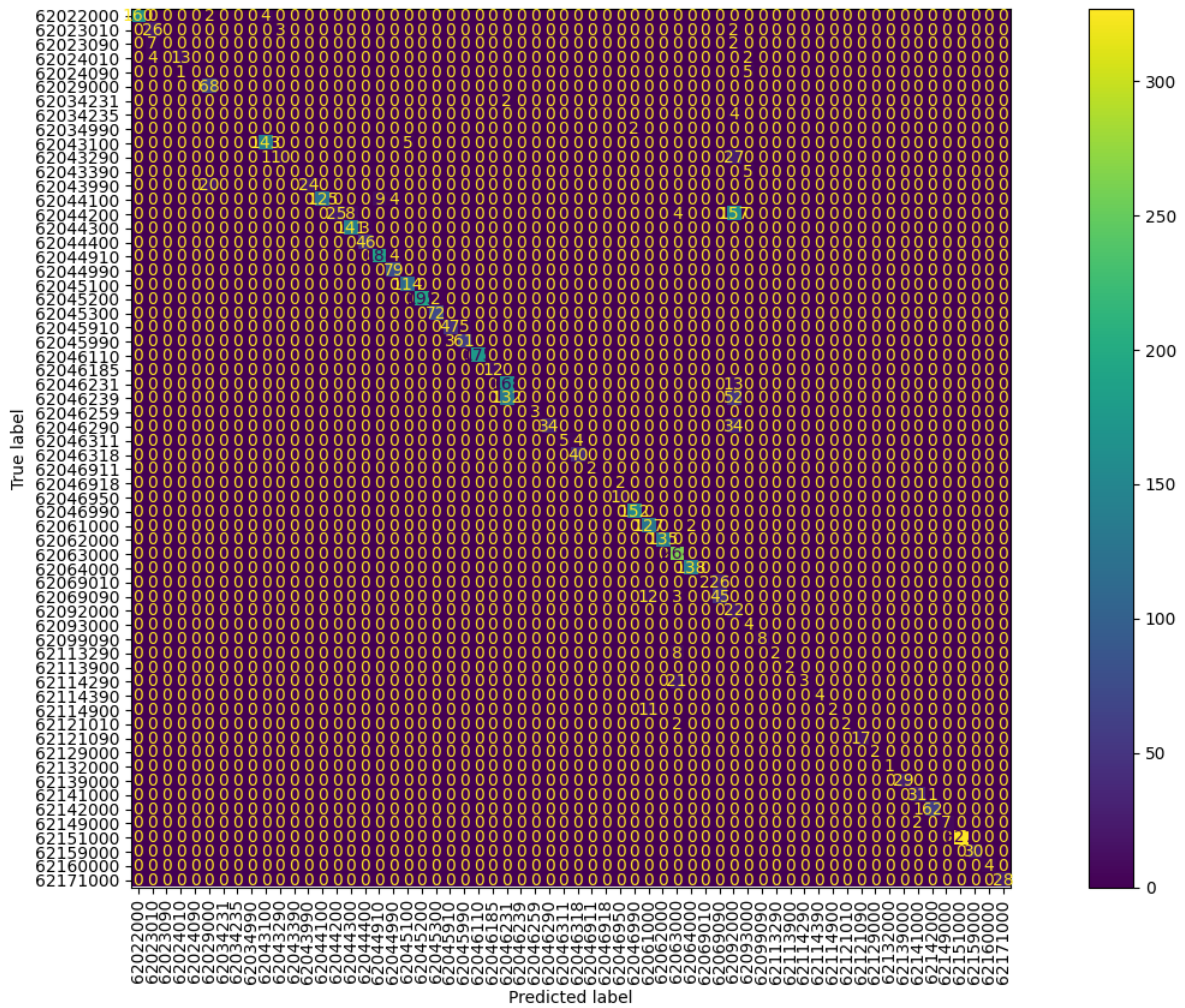


Figure 39: Confusion matrix for data set ch. 62 (without "gender" feature)

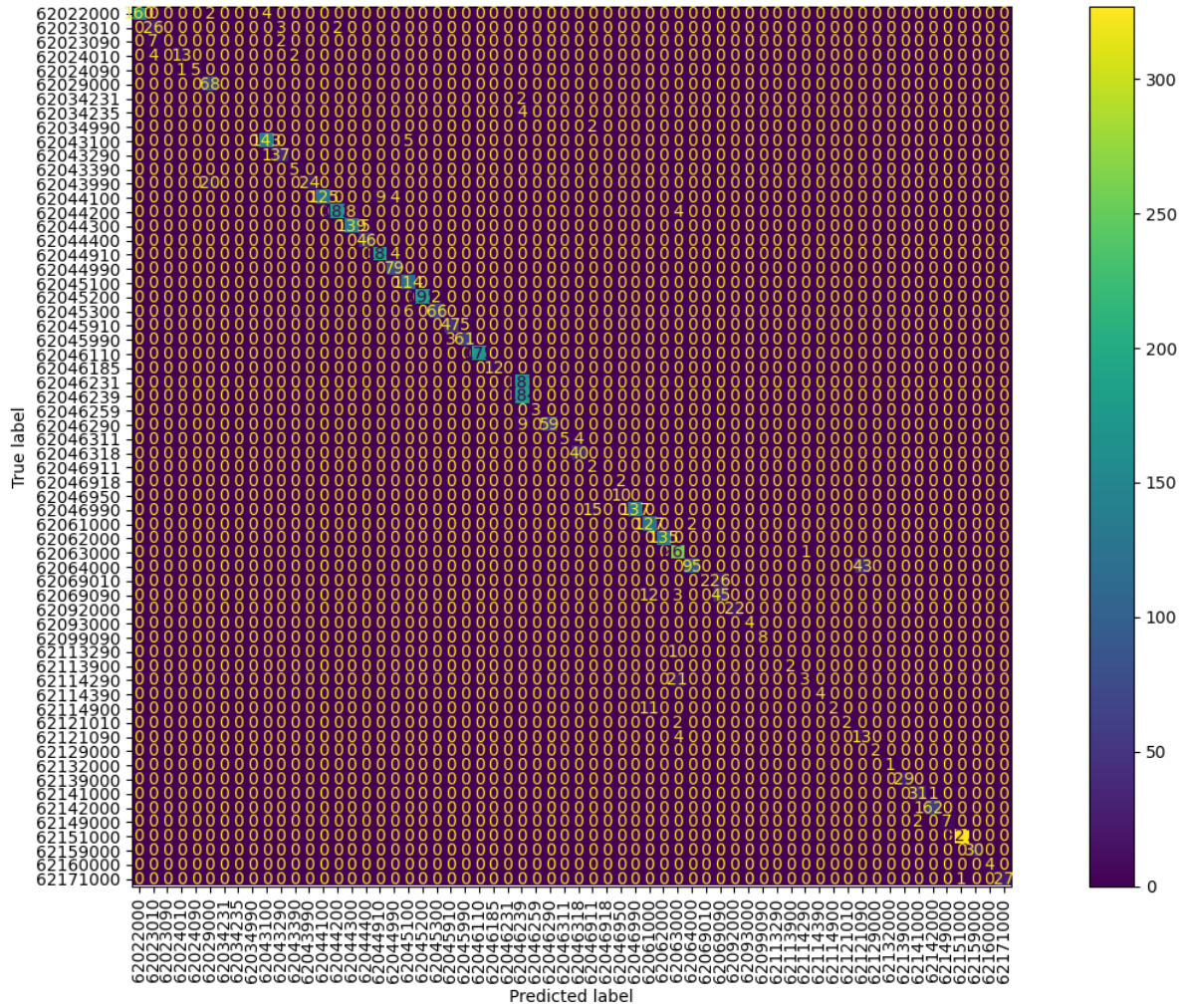


Figure 40: Confusion matrix for data set ch. 62 (without "material category" feature)

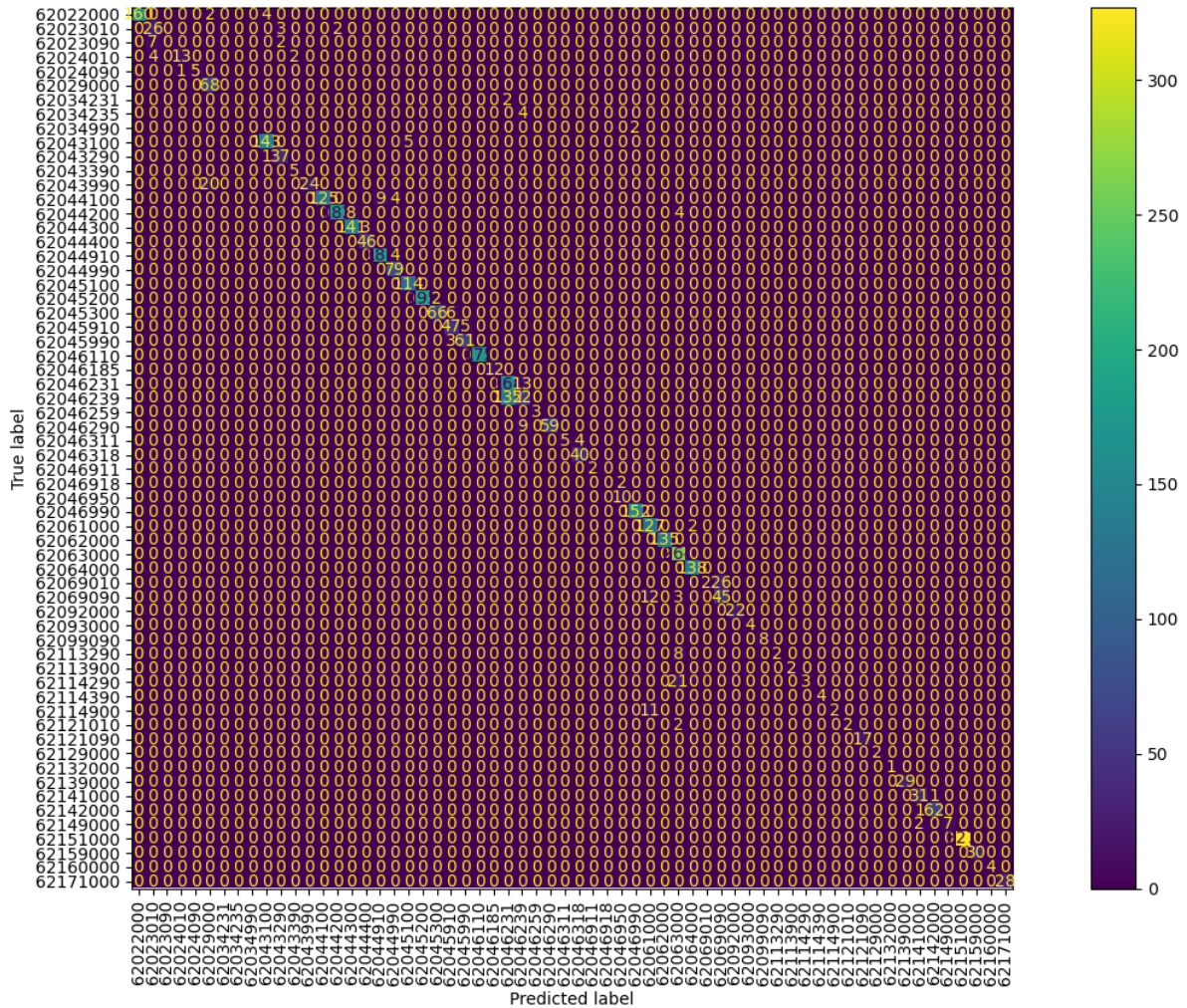


Figure 41: Confusion matrix for data set ch. 62 (without "article type" feature)

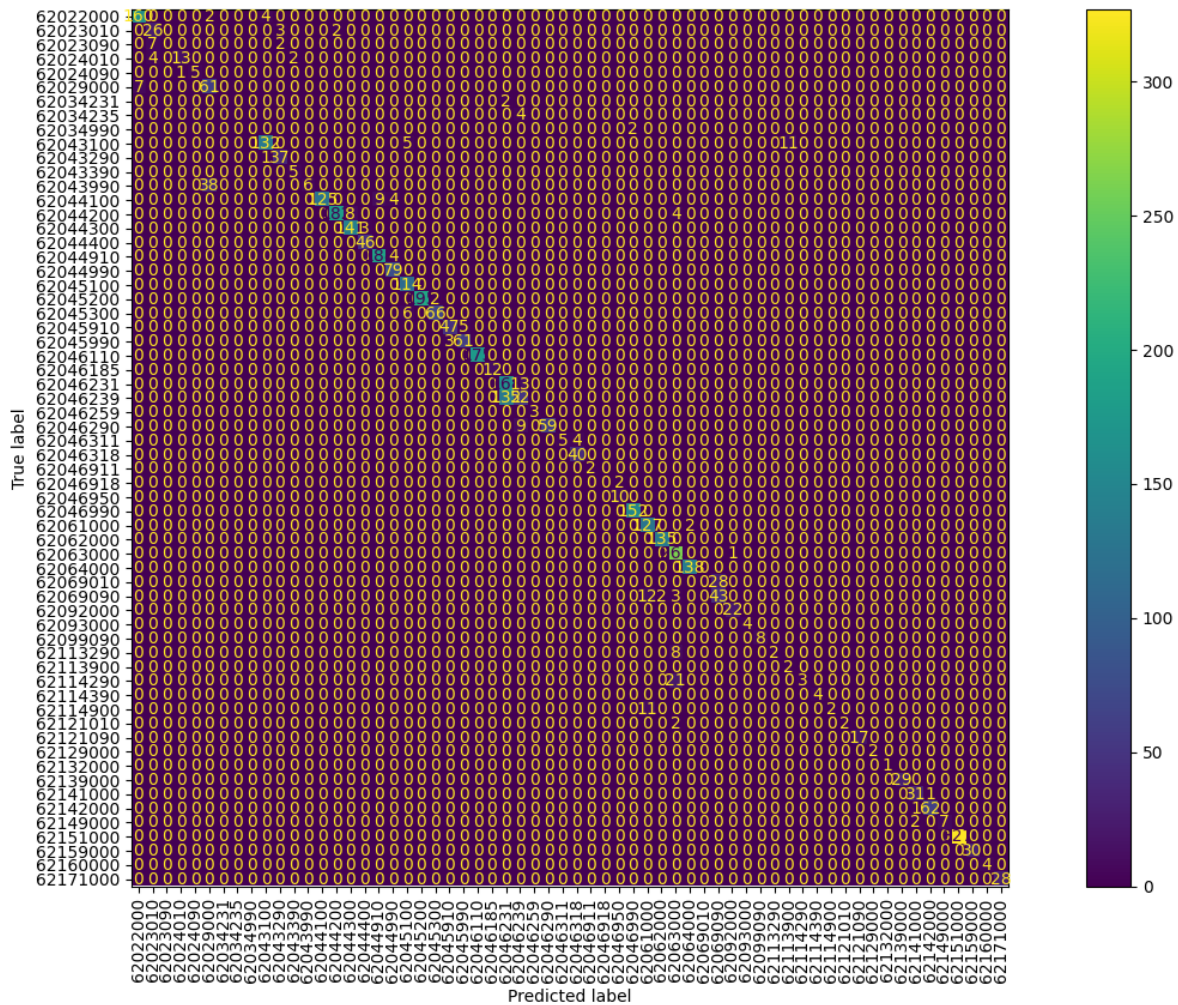


Figure 42: Confusion matrix for data set ch. 62 (without "CITES" feature)

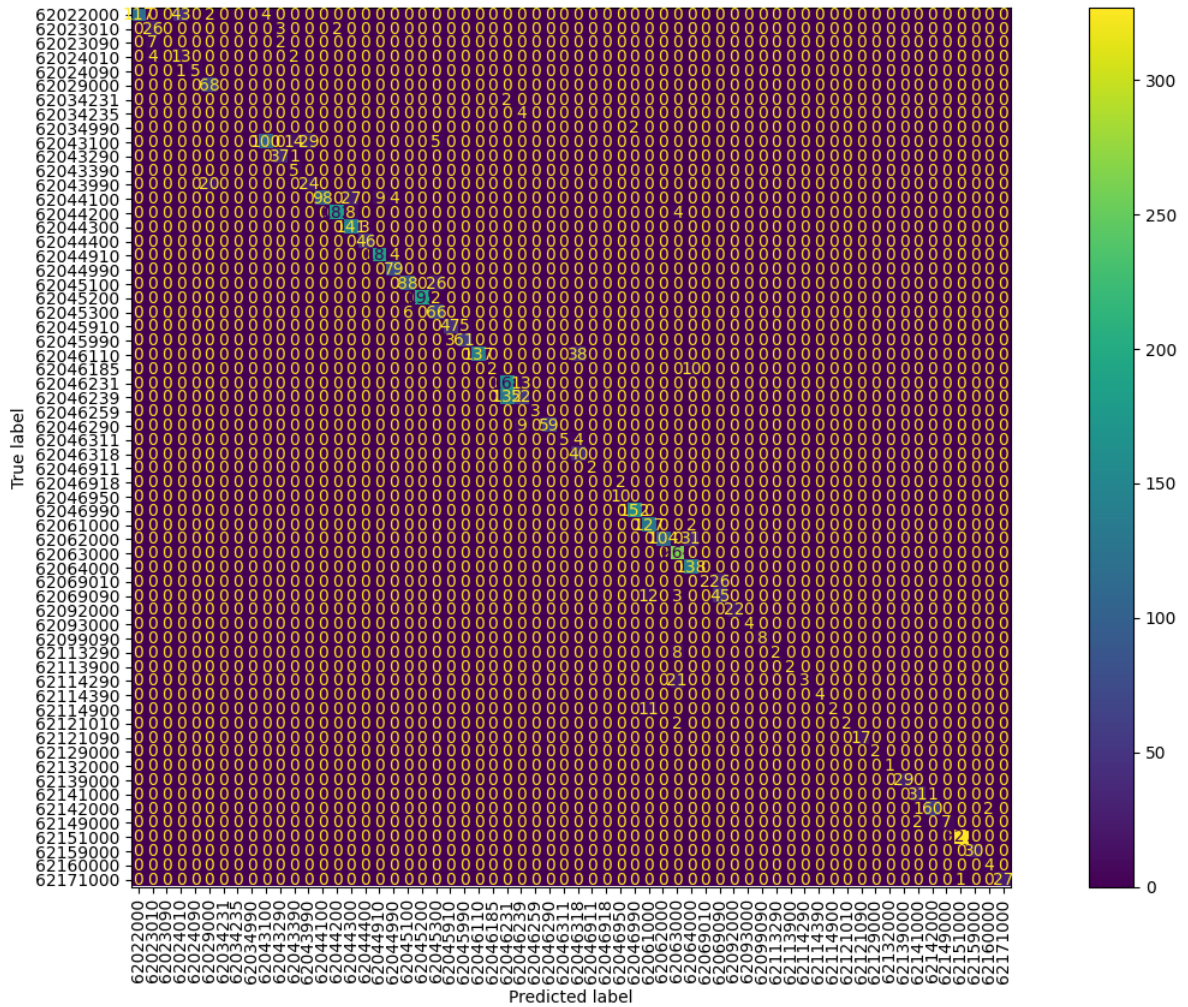
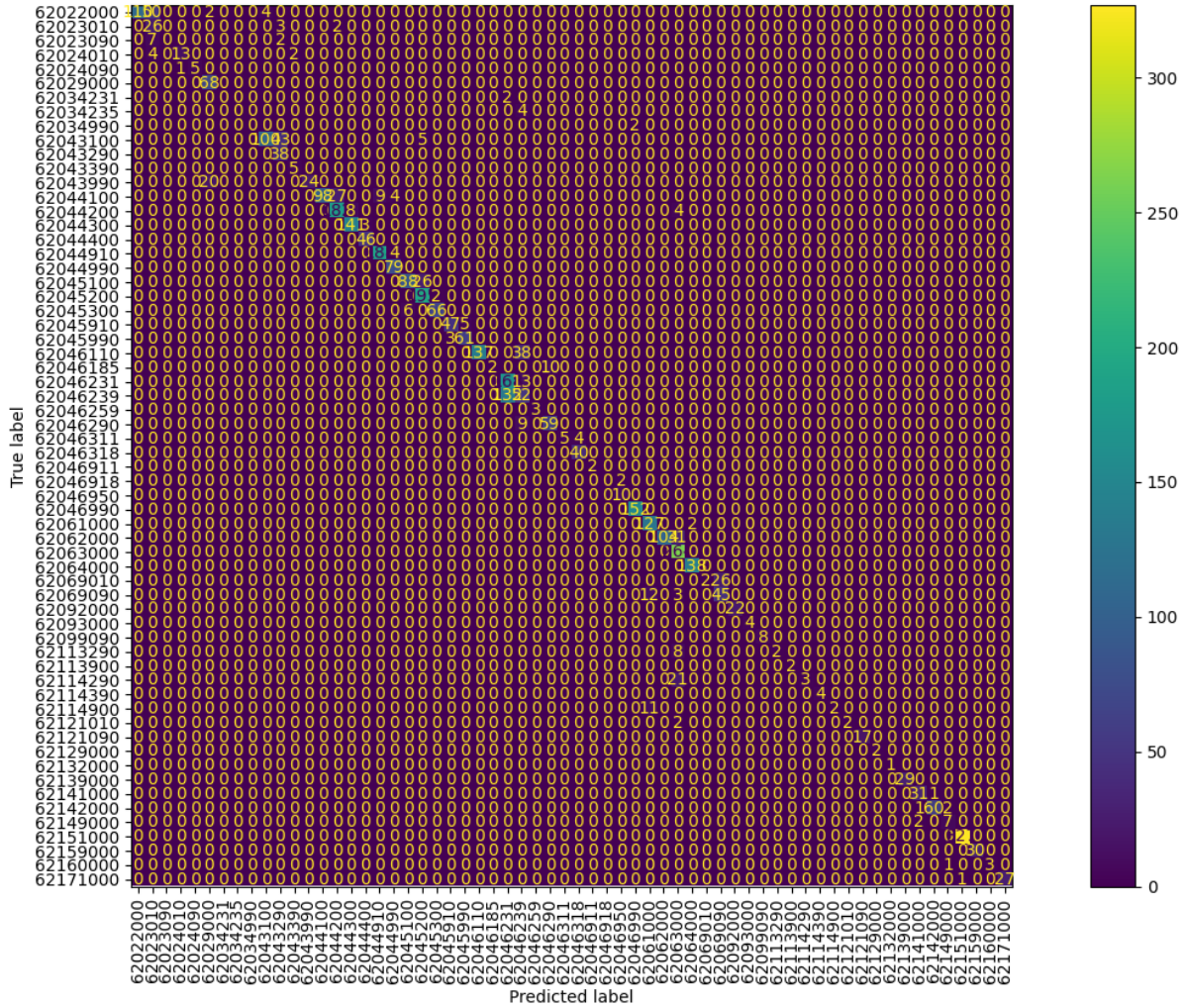


Figure 43: Confusion matrix for data set ch. 62 (introduced error: main material "wool" replaced by "polyester")



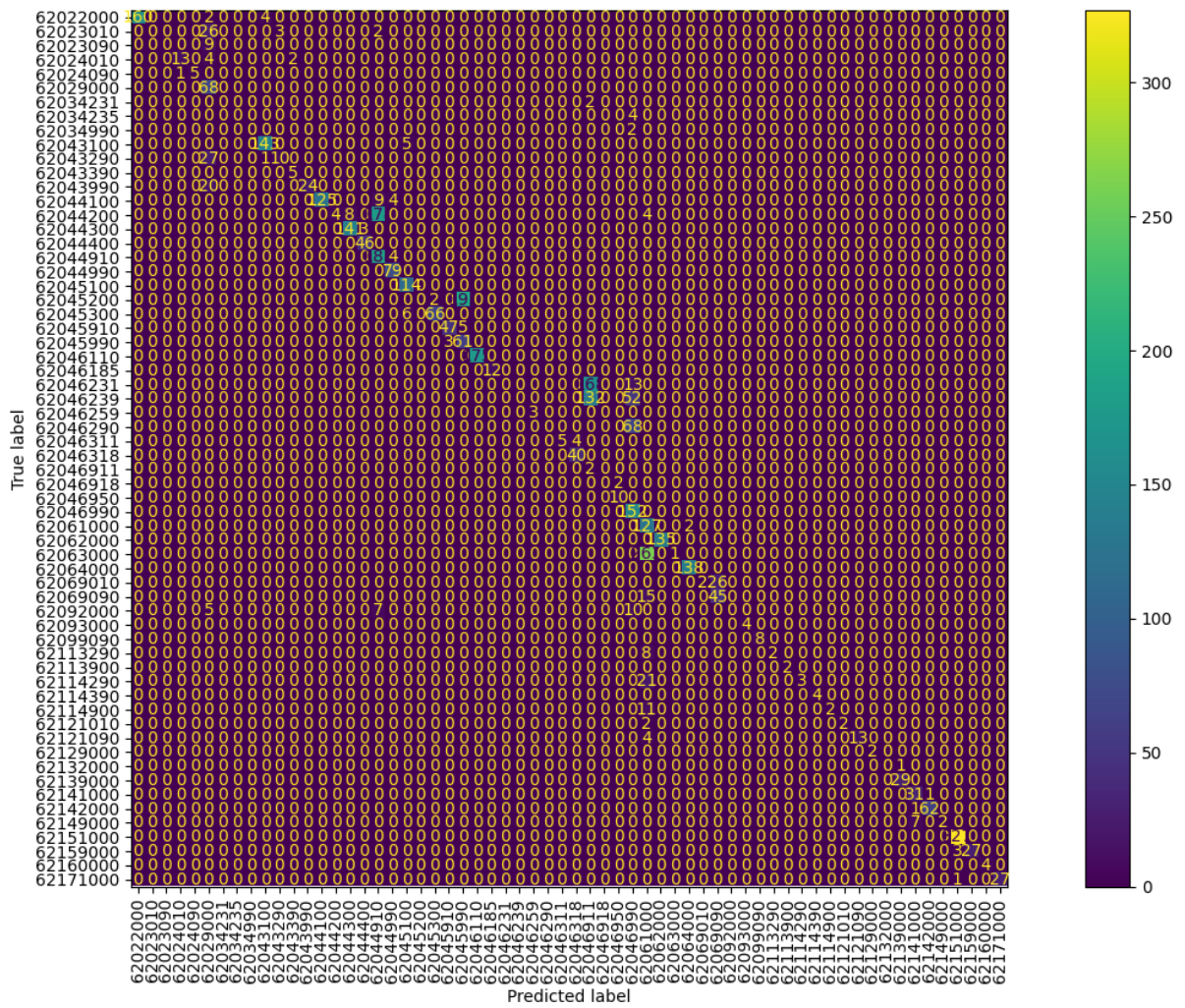


Figure 45: Confusion matrix for data set ch. 62 (introduced error: main material "cotton" replaced by "silk")

8.4 Prediction metrics for data set “chapter 62”

classification	precision	recall	F ₁ -score
62022000	1	0.96	0.98
62023010	0.7	0.84	0.76
62023090	0	0	0
62024010	0.93	0.68	0.79
62024090	1	0.83	0.91
62029000	0.76	1	0.86
62034231	0	0	0
62034235	0	0	0
62034990	0	0	0
62043100	0.97	0.97	0.97
62043290	0.88	0.97	0.93
62043390	0.71	1	0.83
62043990	1	0.55	0.71
62044100	1	0.91	0.95
62044200	0.99	0.94	0.96
62044300	0.95	0.98	0.96
62044400	0.94	1	0.97
62044910	0.95	0.98	0.97
62044990	0.91	1	0.95
62045100	0.91	1	0.95
62045200	1	0.99	0.99
62045300	0.97	0.92	0.94
62045910	0.94	0.9	0.92
62045990	0.92	0.95	0.94
62046110	1	1	1
62046185	1	1	1
62046231	0.56	0.93	0.7
62046239	0.67	0.28	0.4
62046259	1	1	1
62046290	1	0.87	0.93
62046311	1	0.56	0.71
62046318	0.91	1	0.95
62046911	1	1	1
62046918	0	0	0
62046950	0.83	1	0.91
62046990	0.99	1	0.99
62061000	0.85	0.98	0.91
62062000	1	1	1
62063000	0.87	1	0.93
62064000	0.99	1	0.99
62069010	1	0.07	0.13
62069090	0.63	0.75	0.69
62092000	1	1	1
62093000	1	1	1
62099090	1	1	1
62113290	1	0.2	0.33
62113900	1	1	1
62114290	1	0.12	0.22
62114390	1	1	1
62114900	1	0.15	0.27
62121010	1	0.5	0.67
62121090	1	1	1
62129000	1	1	1
62132000	1	1	1
62139000	1	1	1
62141000	0.91	0.97	0.94
62142000	0.98	0.98	0.98
62149000	1	0.78	0.88
62151000	1	1	1
62159000	1	1	1
62160000	1	1	1
62171000	1	0.96	0.98
Accuracy		0.9116	

Table 27: Prediction metrics for initial data set ch. 62

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F₁-score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

classification	precision	recall	F_1 -score
62022000	1	0.96	0.98
62023010	0.7	0.84	0.76
62023090	0	0	0
62024010	0.93	0.68	0.79
62024090	0	0	0
62029000	0.76	1	0.86
62034231	0	0	0
62034235	0	0	0
62034990	0	0	0
62043100	0.97	0.97	0.97
62043290	0.77	0.26	0.39
62043390	0	0	0
62043990	1	0.55	0.71
62044100	1	0.91	0.95
62044200	1	0.13	0.23
62044300	0.95	0.98	0.96
62044400	0.94	1	0.97
62044910	0.95	0.98	0.97
62044990	0.91	1	0.95
62045100	0.96	1	0.98
62045200	1	0.99	0.99
62045300	0.97	1	0.99
62045910	0.94	0.9	0.92
62045990	0.92	0.95	0.94
62046110	1	1	1
62046185	1	1	1
62046231	0.56	0.93	0.7
62046239	0	0	0
62046259	1	1	1
62046290	1	0.5	0.67
62046311	1	0.56	0.71
62046318	0.91	1	0.95
62046911	1	1	1
62046918	0	0	0
62046950	0.83	1	0.91
62046990	0.99	1	0.99
62061000	0.85	0.98	0.91
62062000	1	1	1
62063000	0.87	1	0.93
62064000	0.99	1	0.99
62069010	1	0.07	0.13
62069090	0.63	0.75	0.69
62092000	0.07	1	0.13
62093000	0.25	1	0.4
62099090	1	1	1
62113290	1	0.2	0.33
62113900	1	1	1
62114290	1	0.12	0.22
62114390	1	1	1
62114900	1	0.15	0.27
62121010	1	0.5	0.67
62121090	1	1	1
62129000	1	1	1
62132000	1	1	1
62139000	1	1	1
62141000	0.91	0.97	0.94
62142000	0.98	0.98	0.98
62149000	1	0.78	0.88
62151000	1	1	1
62159000	1	1	1
62160000	1	1	1
62171000	1	1	1
Accuracy		0.8465	

Table 28: Prediction metrics for data set ch. 62 (without "gender" feature)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

classification	precision	recall	F_1 -score
62022000	1	0.96	0.98
62023010	0.7	0.84	0.76
62023090	0	0	0
62024010	0.93	0.68	0.79
62024090	1	0.83	0.91
62029000	0.76	1	0.86
62034231	0	0	0
62034235	0	0	0
62034990	0	0	0
62043100	0.97	0.97	0.97
62043290	0.88	0.97	0.93
62043390	0.71	1	0.83
62043990	1	0.55	0.71
62044100	1	0.91	0.95
62044200	0.99	0.94	0.96
62044300	0.95	0.97	0.96
62044400	0.9	1	0.95
62044910	0.95	0.98	0.97
62044990	0.91	1	0.95
62045100	0.91	1	0.95
62045200	1	0.99	0.99
62045300	0.97	0.92	0.94
62045910	0.94	0.9	0.92
62045990	0.92	0.95	0.94
62046110	1	1	1
62046185	1	1	1
62046231	0	0	0
62046239	0.48	1	0.65
62046259	1	1	1
62046290	1	0.87	0.93
62046311	1	0.56	0.71
62046318	0.91	1	0.95
62046911	0.11	1	0.19
62046918	0	0	0
62046950	0.83	1	0.91
62046990	1	0.9	0.95
62061000	0.85	0.98	0.91
62062000	1	1	1
62063000	0.86	1	0.92
62064000	0.98	0.69	0.81
62069010	1	0.07	0.13
62069090	0.63	0.75	0.69
62092000	1	1	1
62093000	1	1	1
62099090	1	1	1
62113290	0	0	0
62113900	1	1	1
62114290	0.75	0.12	0.21
62114390	1	1	1
62114900	1	0.15	0.27
62121010	1	0.5	0.67
62121090	0.23	0.76	0.36
62129000	1	1	1
62132000	1	1	1
62139000	1	1	1
62141000	0.91	0.97	0.94
62142000	0.98	0.98	0.98
62149000	1	0.78	0.88
62151000	1	1	1
62159000	1	1	1
62160000	1	1	1
62171000	1	0.96	0.98
Accuracy		0.8862	

Table 29: Prediction metrics for data set ch. 62 (without "material category" feature)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

classification	precision	recall	F_1 -score
62022000	0.88	0.09	0.16
62023010	0.04	0.9	0.07
62023090	0	0	0
62024010	1	0.11	0.19
62024090	0	0	0
62029000	0.77	0.29	0.43
62034231	0	0	0
62034235	0	0	0
62034990	0	0	0
62043100	0.29	0.65	0.4
62043290	0	0	0
62043390	0	0	0
62043990	0	0	0
62044100	0.93	0.09	0.17
62044200	1	0.01	0.01
62044300	1	0.01	0.03
62044400	0	0	0
62044910	0	0	0
62044990	1	0.13	0.22
62045100	1	0.13	0.23
62045200	1	0.01	0.01
62045300	0	0	0
62045910	1	0.15	0.27
62045990	0	0	0
62046110	0	0	0
62046185	0	0	0
62046231	0	0	0
62046239	0	0	0
62046259	0	0	0
62046290	0	0	0
62046311	0	0	0
62046318	0	0	0
62046911	1	1	1
62046918	0	0	0
62046950	0	0	0
62046990	0.21	1	0.34
62061000	0.55	0.36	0.44
62062000	0.37	0.18	0.24
62063000	1	0.05	0.1
62064000	0.29	1	0.45
62069010	1	0.07	0.13
62069090	0.22	0.08	0.12
62092000	0.96	1	0.98
62093000	1	1	1
62099090	1	1	1
62113290	0	0	0
62113900	0	1	0.01
62114290	0.01	0.12	0.01
62114390	0.24	1	0.38
62114900	1	0.15	0.27
62121010	0.22	0.5	0.31
62121090	0.74	1	0.85
62129000	1	1	1
62132000	0.25	1	0.4
62139000	0	0	0
62141000	1	0.53	0.69
62142000	0.74	0.4	0.52
62149000	1	0.78	0.88
62151000	0.87	1	0.93
62159000	0.34	0.63	0.44
62160000	1	1	1
62171000	1	0.96	0.98
Accuracy		0.2680	

Table 30: Prediction metrics for data set ch. 62 (without "article sub-type" feature)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

classification	precision	recall	F_1 -score
62022000	1	0.96	0.98
62023010	0.7	0.84	0.76
62023090	0	0	0
62024010	0.93	0.68	0.79
62024090	1	0.83	0.91
62029000	0.76	1	0.86
62034231	0	0	0
62034235	0	0	0
62034990	0	0	0
62043100	0.97	0.97	0.97
62043290	0.88	0.97	0.93
62043390	0.71	1	0.83
62043990	1	0.55	0.71
62044100	1	0.91	0.95
62044200	0.99	0.94	0.96
62044300	0.95	0.98	0.96
62044400	0.94	1	0.97
62044910	0.95	0.98	0.97
62044990	0.91	1	0.95
62045100	0.96	1	0.98
62045200	1	0.99	0.99
62045300	0.97	0.92	0.94
62045910	0.84	0.9	0.87
62045990	0.92	0.95	0.94
62046110	1	1	1
62046185	1	1	1
62046231	0.56	0.93	0.7
62046239	0.67	0.28	0.4
62046259	1	1	1
62046290	1	0.87	0.93
62046311	1	0.56	0.71
62046318	0.91	1	0.95
62046911	1	1	1
62046918	0	0	0
62046950	0.83	1	0.91
62046990	0.99	1	0.99
62061000	0.85	0.98	0.91
62062000	1	1	1
62063000	0.87	1	0.93
62064000	0.99	1	0.99
62069010	1	0.07	0.13
62069090	0.63	0.75	0.69
62092000	1	1	1
62093000	1	1	1
62099090	1	1	1
62113290	1	0.2	0.33
62113900	1	1	1
62114290	1	0.12	0.22
62114390	1	1	1
62114900	1	0.15	0.27
62121010	1	0.5	0.67
62121090	1	1	1
62129000	1	1	1
62132000	1	1	1
62139000	1	1	1
62141000	0.91	0.97	0.94
62142000	0.98	0.98	0.98
62149000	1	0.78	0.88
62151000	1	1	1
62159000	1	1	1
62160000	1	1	1
62171000	1	1	1
Accuracy		0.9118	

Table 31: Prediction metrics for data set ch. 62 (without "article type" feature)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

classification	precision	recall	F_1 -score
62022000	0.96	0.96	0.96
62023010	0.7	0.84	0.76
62023090	0	0	0
62024010	0.93	0.68	0.79
62024090	1	0.83	0.91
62029000	0.6	0.9	0.72
62034231	0	0	0
62034235	0	0	0
62034990	0	0	0
62043100	0.96	0.89	0.93
62043290	0.88	0.97	0.93
62043390	0.71	1	0.83
62043990	1	0.14	0.24
62044100	1	0.91	0.95
62044200	0.99	0.94	0.96
62044300	0.95	0.98	0.96
62044400	0.94	1	0.97
62044910	0.95	0.98	0.97
62044990	0.91	1	0.95
62045100	0.91	1	0.95
62045200	1	0.99	0.99
62045300	0.97	0.92	0.94
62045910	0.94	0.9	0.92
62045990	0.92	0.95	0.94
62046110	1	1	1
62046185	1	1	1
62046231	0.56	0.93	0.7
62046239	0.67	0.28	0.4
62046259	1	1	1
62046290	1	0.87	0.93
62046311	1	0.56	0.71
62046318	0.91	1	0.95
62046911	1	1	1
62046918	0	0	0
62046950	0.83	1	0.91
62046990	0.99	1	0.99
62061000	0.85	0.98	0.91
62062000	0.99	1	0.99
62063000	0.87	1	0.93
62064000	0.99	1	0.99
62069010	0	0	0
62069090	0.61	0.72	0.66
62092000	0.96	1	0.98
62093000	1	1	1
62099090	1	1	1
62113290	1	0.2	0.33
62113900	0.15	1	0.27
62114290	1	0.12	0.22
62114390	1	1	1
62114900	1	0.15	0.27
62121010	1	0.5	0.67
62121090	1	1	1
62129000	1	1	1
62132000	1	1	1
62139000	1	1	1
62141000	0.91	0.97	0.94
62142000	0.98	0.98	0.98
62149000	1	0.78	0.88
62151000	1	1	1
62159000	1	1	1
62160000	1	1	1
62171000	1	1	1
Accuracy		0.9017	

Table 32: Predictions metrics for data set ch. 62 (without "CITES" feature)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

classification	Precision	recall	F_1 -score
62022000	0.65	0.98	0.78
62023010	0	0	0
62023090	0	0	0
62024010	1	0.26	0.42
62024090	0	0	0
62029000	0	0	0
62034231	0	0	0
62034235	0	0	0
62034990	0	0	0
62043100	0.46	0.53	0.49
62043290	0.77	0.26	0.39
62043390	0	0	0
62043990	0.27	0.55	0.36
62044100	0.18	0.91	0.3
62044200	1	0.11	0.2
62044300	0.33	0.01	0.03
62044400	0	0	0
62044910	0	0	0
62044990	0.17	0.13	0.14
62045100	0.25	1	0.4
62045200	1	0.24	0.39
62045300	0	0	0
62045910	0	0	0
62045990	0	0	0
62046110	0.39	1	0.56
62046185	0.2	1	0.33
62046231	0.56	0.93	0.7
62046239	0	0	0
62046259	1	1	1
62046290	1	0.5	0.67
62046311	0	0	0
62046318	0	0	0
62046911	1	1	1
62046918	0	0	0
62046950	0	0	0
62046990	0	0	0
62061000	1	0.05	0.09
62062000	0.21	0.78	0.33
62063000	0.87	0.05	0.09
62064000	0.09	0.12	0.1
62069010	1	0.07	0.13
62069090	0.07	0.08	0.07
62092000	0	0	0
62093000	0.44	1	0.62
62099090	0.32	1	0.48
62113290	1	0.2	0.33
62113900	1	1	1
62114290	1	0.12	0.22
62114390	1	0.25	0.4
62114900	0	0	0
62121010	0	0	0
62121090	1	1	1
62129000	1	1	1
62132000	0	0	0
62139000	0.97	1	0.98
62141000	0.25	0.03	0.06
62142000	0.63	1	0.77
62149000	0	0	0
62151000	1	0.03	0.07
62159000	0.09	1	0.16
62160000	1	1	1
62171000	1	1	1
Accuracy		0.3321	

Table 33: Prediction metrics for data set ch. 62 (without "main material" feature)

The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

classification	precision	recall	F_1 -score
62022000	1	0.7	0.83
62023010	0.7	0.84	0.76
62023090	0	0	0
62024010	0.23	0.68	0.34
62024090	1	0.83	0.91
62029000	0.76	1	0.86
62034231	0	0	0
62034235	0	0	0
62034990	0	0	0
62043100	0.96	0.68	0.79
62043290	0.88	0.97	0.93
62043390	0.23	1	0.37
62043990	0.45	0.55	0.49
62044100	1	0.71	0.83
62044200	0.99	0.94	0.96
62044300	0.8	0.98	0.88
62044400	0.94	1	0.97
62044910	0.95	0.98	0.97
62044990	0.91	1	0.95
62045100	0.94	0.77	0.85
62045200	1	0.99	0.99
62045300	0.67	0.92	0.77
62045910	0.94	0.9	0.92
62045990	0.92	0.95	0.94
62046110	1	0.78	0.88
62046185	1	0.17	0.29
62046231	0.56	0.93	0.7
62046239	0.67	0.28	0.4
62046259	1	1	1
62046290	1	0.87	0.93
62046311	1	0.56	0.71
62046318	0.49	1	0.66
62046911	1	1	1
62046918	0	0	0
62046950	0.83	1	0.91
62046990	0.99	1	0.99
62061000	0.85	0.98	0.91
62062000	1	0.77	0.87
62063000	0.87	1	0.93
62064000	0.76	1	0.87
62069010	1	0.07	0.13
62069090	0.63	0.75	0.69
62092000	1	1	1
62093000	1	1	1
62099090	1	1	1
62113290	1	0.2	0.33
62113900	1	1	1
62114290	1	0.12	0.22
62114390	1	1	1
62114900	1	0.15	0.27
62121010	1	0.5	0.67
62121090	1	1	1
62129000	1	1	1
62132000	1	1	1
62139000	1	1	1
62141000	0.91	0.97	0.94
62142000	0.98	0.95	0.97
62149000	1	0.78	0.88
62151000	1	1	1
62159000	1	1	1
62160000	0.67	1	0.8
62171000	1	0.96	0.98
Accuracy		0.8574	

Table 34: Prediction metrics for data set ch. 62 (introduced error: main material "wool" replaced by "polyester")
The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

classification	precision	recall	F_1 -score
62022000	1	0.66	0.8
62023010	0.3	0.84	0.44
62023090	0	0	0
62024010	0.93	0.68	0.79
62024090	1	0.83	0.91
62029000	0.76	1	0.86
62034231	0	0	0
62034235	0	0	0
62034990	0	0	0
62043100	0.96	0.68	0.79
62043290	0.44	1	0.61
62043390	0.71	1	0.83
62043990	1	0.55	0.71
62044100	1	0.71	0.83
62044200	0.86	0.94	0.9
62044300	0.95	0.98	0.96
62044400	0.94	1	0.97
62044910	0.95	0.98	0.97
62044990	0.91	1	0.95
62045100	0.94	0.77	0.85
62045200	0.86	0.99	0.92
62045300	0.97	0.92	0.94
62045910	0.94	0.9	0.92
62045990	0.92	0.95	0.94
62046110	1	0.78	0.88
62046185	1	0.17	0.29
62046231	0.56	0.93	0.7
62046239	0.45	0.28	0.35
62046259	1	1	1
62046290	0.86	0.87	0.86
62046311	1	0.56	0.71
62046318	0.91	1	0.95
62046911	1	1	1
62046918	0	0	0
62046950	0.83	1	0.91
62046990	0.99	1	0.99
62061000	0.85	0.98	0.91
62062000	1	0.77	0.87
62063000	0.79	1	0.88
62064000	0.99	1	0.99
62069010	1	0.07	0.13
62069090	0.63	0.75	0.69
62092000	1	1	1
62093000	1	1	1
62099090	1	1	1
62113290	1	0.2	0.33
62113900	1	1	1
62114290	1	0.12	0.22
62114390	1	1	1
62114900	1	0.15	0.27
62121010	1	0.5	0.67
62121090	1	1	1
62129000	1	1	1
62132000	1	1	1
62139000	1	1	1
62141000	0.91	0.97	0.94
62142000	0.98	0.95	0.97
62149000	0.7	0.78	0.74
62151000	1	1	1
62159000	1	1	1
62160000	1	0.75	0.86
62171000	1	0.96	0.98
Accuracy		0.8556	

Table 35: Prediction metrics for data set ch. 62 (introduced error: main material "wool" replaced by "cotton")
The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.

classification	precision	recall	F_1 -score
62022000	1	0.96	0.98
62023010	0	0	0
62023090	0	0	0
62024010	0.93	0.68	0.79
62024090	1	0.83	0.91
62029000	0.42	1	0.59
62034231	0	0	0
62034235	0	0	0
62034990	0	0	0
62043100	0.97	0.97	0.97
62043290	0.77	0.26	0.39
62043390	0.71	1	0.83
62043990	1	0.55	0.71
62044100	1	0.91	0.95
62044200	1	0.02	0.04
62044300	0.95	0.98	0.96
62044400	0.94	1	0.97
62044910	0.49	0.98	0.65
62044990	0.91	1	0.95
62045100	0.91	1	0.95
62045200	0	0	0
62045300	0.97	0.92	0.94
62045910	0.94	0.9	0.92
62045990	0.23	0.95	0.38
62046110	1	1	1
62046185	1	1	1
62046231	0	0	0
62046239	0	0	0
62046259	1	1	1
62046290	0	0	0
62046311	1	0.56	0.71
62046318	0.91	1	0.95
62046911	0.01	1	0.01
62046918	0	0	0
62046950	0.83	1	0.91
62046990	0.5	1	0.67
62061000	0.28	0.98	0.43
62062000	1	1	1
62063000	1	0	0.01
62064000	0.99	1	0.99
62069010	1	0.07	0.13
62069090	0.63	0.75	0.69
62092000	0	0	0
62093000	1	1	1
62099090	1	1	1
62113290	1	0.2	0.33
62113900	1	1	1
62114290	1	0.12	0.22
62114390	1	1	1
62114900	1	0.15	0.27
62121010	1	0.5	0.67
62121090	1	0.76	0.87
62129000	1	1	1
62132000	0	0	0
62139000	0.97	1	0.98
62141000	0.79	0.97	0.87
62142000	0.98	0.98	0.98
62149000	1	0.22	0.36
62151000	0.99	1	0.99
62159000	1	0.9	0.95
62160000	1	1	1
62171000	1	0.96	0.98
Accuracy		0.6644	

Table 36: Prediction metrics for data set ch. 62 (introduced error: main material "cotton" replaced by "silk")
The four performance metrics used in this table are measured on a scale of 0 to 1. The accuracy is the ratio of correct predictions to the total number of predictions. The precision is the ratio of true positives to the sum of true positives and false positives. The recall is the ratio of true positives to the sum of true positives and false negatives. The F_1 -score is combining both precision and recall and is calculated as follows: $F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$. Further details on performance metrics can be found in section 4.1.2.