NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

GOVWISE PROCUREMENT VOCABULARY (GPV)

An alternative to the Common Procurement Vocabulary (CPV)

Ana Madalena Pinheiro Santos

Internship Report

presented as partial requirement for obtaining the Master Degree Program in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

GOVWISE PROCUREMENT VOCABULARY (GPV)

By

Ana Madalena Pinheiro Santos

Internship report presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence and Knowledge Management

Supervisor: Professor Doutor Flávio Luís Portas Pinheiro

November 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Frankfurt, 30/11/2022

ACKNOWLEDGEMENTS

With my sincere gratefulness to my advisor, Professor Doutor Flávio L. Pinheiro, for his constant guidance and availability on the orientation of this internship report. To both Francisco Figueiredo and João Silva, for all the support, knowledge sharing and inspiration throughout my internship at Govwise. To my family and friends, for their constant encouragement and comprehension.

ABSTRACT

In recent years, the world has witnessed emerging legislation on open data. Some of the main goals include stimulating economic growth with the re-use of the data, addressing societal challenges, enhancing evidence-based policymaking, and increasing efficiency in the public administrations, fostering the development of new technologies, such as AI, along with the enhanced participation of the citizens in political decisions and its transparency (European Commission, Open Data, 2021).

Govwise is an Advanced Analytics Platform developed to provide a wide range of data analytics to governmental organizations, via a SaaS model. The goal is to make use of the open data policies, by producing valuable information, tackling the challenges that such a data deluge arises. These challenges constitute the scope of the internship here reported, the whole process is described, starting with the data sources and the respective ETL process, on a more high-level structure, until the production of the analysis and dashboards, that constitute the product of Govwise. The focus will, however, be on the classification model developed to address a major necessity of the company to cluster the portuguese public procurement contracts. These are initially classified with a CPV code (common procurement vocabulary code), which does not satisfy the needs of Govwise. Therefore, the end goal of the model is to generate an alternative classification for contracts and tenders of the portuguese public procurement, the GPV (Govwise procurement vocabulary).

KEYWORDS

Open Data; ETL; Classification model; NLP; AI; Public Procurement; CPV

Sustainable Development Goals (SGD):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ETL	Extract, transform, load
AI	Artificial Intelligence
IT	Information technologies
NLP	Natural Processing Language
EU	European Union
ICT	Information and Communication Technologies
CPV	Common Procurement Vocabulary
GPV	GovWise Procurement Vocabulary
KDD	Knowledge Discovery in Databases
SEMMA	Sample, Explore, Modify, Model, Assess
VAT	Value added tax
CCP	Código Contratos Públicos
OGP	Open Government Partnership
AMA	Agência para a Modernização Administrativa
SVM	Support Vector Machine
OCDS	Open Contracting Data Standard
NIF	Número de identificação fiscal (Equivalent to the value added tax number, for portuguese authorities)
BERT	Bidirectional Encoder Representations from Transformers

1. INTRODUCTION

With an ever-increasing amount of data published within the EU, both public entities and corporations are struggling to generate value through open data. The push for open data will only increase the amount of data available, and therefore its total value. The lack of data-science professionals, combined with the decentralization of public entities and their data, are some of the main challenges open data users still face. Many times, data is published in distinct locations and inconsistent formats (national websites, regional databases, unstructured documents, semi-structured messages, and structured records) and is accessible in numerous ways (on a webpage, from file transfer or programmatically), making its extraction and transformation to produce valuable insights complex and time demanding (*The challenges posed by officially published open data, 2019).* In September 2011 Open Government Partnership (OGP), a multilateral initiative, was launched by 8 countries that aims to develop concrete commitments from governments in order to promote transparency, foster public participation, fight corruption and strengthen participative democracy through the use of new technologies (AMA, 2018). In May 2018, with the purpose of implementing the Portuguese participation on OGP, the Agência para a Modernização Administrativa (AMA), Portuguese agency for administrative modernization started developing an open data platform containing public procurement information. The data published in this platform follows the Open Contracting Data Standard (OCDS) international format, developed by the OCP aiming to ensure the transparency and quality of the e-procurement systems in each step of the public procurement purchases (Portal Base, 2022). This first approach towards an harmonization of the public procurement data in Portugal has proven itself to be very fruitful, however users still face some challenges, namely when it comes to produce added value on the published data. The missing and inaccurate information inserted on the platform by the users constitute the main triggers of the difficulties on analysing the data afterwards. Govwise software delivers a possible approach to tackle these challenges, which constitutes its value proposition and the scope of this internship report. In 2021, the contracts published in Portal Base represented 6.1% of the Portuguese PIB ("Contratação Pública Em Portugal 2021," 2022), this number has been tendentially increasing over the years, thus making it an increasingly more valuable data source.

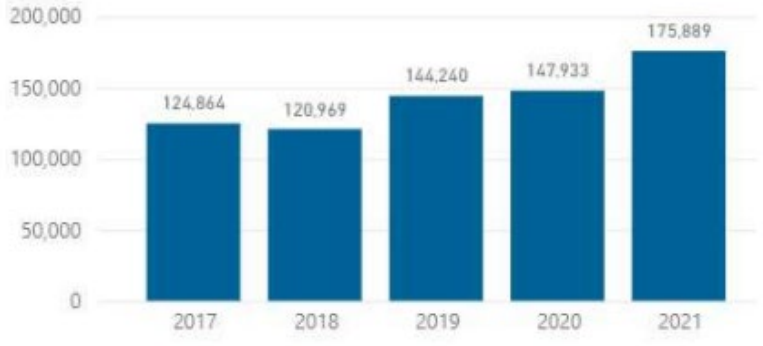

Figure 1 - Number of published contracts in Portal Base over the last 5 years

The more data is made available, the bigger is the challenge to produce added value, particularly considering the difficulties stated before, such as the lack of data professionals, as well as the human error when publishing the data. To produce valuable insights on the public procurement data, for this project particular case, aggregations are needed, which are made based on certain classes. One possible way to aggregate public procurement contracts is through the Common Procurement Vocabulary (CPV) code. The CPV is a single classification system for public procurement aimed at standardizing the references used by contracting authorities and entities to describe the subject of procurement contracts. At this stage, the problem arises due to many reasons related with the CPV classification, as identified, as well, in the *Final report| Revision of CPV*, *2017*:

- The incorrect use of the CPV from the publishers, when it comes to choosing the right match for a certain service or product, reducing largely the efficiency and transparency of the public procurement, compared to what was expected.
- The limited coverage of the CPV to all the phases of the e-procurement, it only applies to the publication and identification of tender notices, meaning that its limited scope will leave unclassified a lot of information in other phases of the procurement.
- At last, with particular emphasis for this report, the absence of a mapping of the CPV to other existing classifications. In order to allow interoperability in electronic data handling, this mapping arises as essential. It would promote the harmonization of several sources of open data, allowing increasingly more accurate analysis on public procurement.

The present internship report intends to tackle the stated problems, with an almost literal approach on the third one, given that by developing a solution to map the CPV to existing classifications, it also contributes for the solution of the two problems stated before. By developing a unique classification structure, the Govwise Procurement Vocabulary, it is then possible to aggregate public procurement contracts in an insightful way, thus generate relevant tender notifications for the clients, as well as provide a broad market analysis on Portuguese public procurement. Throughout the internship at Govwise, I am involved in a wide range of activities, all the way from the extraction of the data to its transformation and integration on the Datawarehouse, as well as in the production of dashboards and analysis, that constitute the main product of Govwise. GOVWISE is an Advanced Analytics Platform developed to provide a wide range of data analytics to governmental organizations, through a SaaS model. The start-up was founded in 2018, leaning on the new European Data Strategy and the EU Open Data Directive, positioning itself within the New Data Economy sector, with the purpose of compiling and systematically analysing the currently dispersed and untreated public data available, related to public procurement and spending, and employing state-of-the-art algorithms to transform scattered data into economical and societal value. As such, starting from e-tendering and e-invoicing data, the platform is built for the decision-makers, who can capitalize on the KPIs, and Insights produced by the company's algorithms. The goal is to display only information that matters through *data stories*, in a logical, clean, easy to understand format. The present internship report will focus on a very specific part of the ETL process, the implementation of a classification model with Natural Language Processing techniques embedded, to reclassify the gathered data into pre-defined classes. The classification model is being developed in such a way that allows Govwise to produce meaningful insights on Portuguese Public Procurement contracts and tenders, clustering the data according to classifications that bring added value, when compared to the CPV code that the data is already provided with, given the challenges stated before. The model constitutes a supervised multi-class algorithm applied to contracts and tenders celebrated/published by Portuguese entities, that will predict a final GPV classification for each one of the records. Each record includes text, numerical and categorical features that will be used to train the model, after being submitted to pre-processing techniques. The end goal is to get a solid, trustworthy classification model that guarantees the accuracy of the analysis produced on the classified data. This model will then be integrated on a pipeline in order to automatically reclassify the new data collected on real time, on a daily basis.

The present document will initially deepen some context on the CPV historical origins and the problems it entails considering the goal of this project, as well as its differences to the developed GPV, and the added value of the latest. Afterwards, related works will be discussed along with relevant literature for the project. On the following chapter the collection of the data, the pre-processing techniques applied to its features and a descriptive analysis is made. Still in the same chapter the methodology and workflow of the model are described. After describing the model development, a benchmark analysis is presented on the proposed models, as well as a performance evaluation for each. At this stage it is also elaborated a discussion on the results obtained. Finally, on the last chapter the conclusions are drowned, along with the limitations of the project and suggestions for future works.

2. RESEARCH CONTEXT

For the first phase of the project, it is crucial that the project objectives are made clear from a business perspective, afterwards, converted into a data mining problem definition to then develop a first plan to achieve the intended business goals (Shearer C., 2000).

THE COMMON PROCUREMENT VOCABULARY CODE

The CPV code was developed to standardise the terms the terms used by contracting authorities and entities to describe the subject of contracts, through a single classification system for public procurement within the EU market. It consists of a main vocabulary and a supplementary vocabulary. The main vocabulary defines the subject of a contract whereas the supplementary vocabulary may be used to add further qualitative information on the subject of the contract. The main vocabulary is based on a tree structure comprising codes of up to 9 digits (an 8 digit code plus a check digit) associated with a wording that describes the type of works, supplies or services forming the subject of the contract. In total, there are today 9,454 codes. In the supplementary vocabulary, the items are made up of an alphanumeric code with a corresponding wording allowing further details to be added regarding the specific nature or destination of the works/supplies/services to be purchased. In total, there are today 903 supplementary vocabulary items. It is meant to reduce the risk of error during translation, since it is available in all the EU's official languages. The first CPV coding system was developed in 1993, forming an 8-digit code, along with a supplementary code list. Since this first version was issued, the CPV structure has undergone several revisions, namely 1998, 2001 and between 2004-2007, the most recent one was being in 2017. Some updates were made to the codes, such as addition of new codes and removal of other ones, but the structure still followed nowadays is essentially the one published in 2008 and can be found at https://simap.ted.europa.eu/web/simap/cpv. The current structure of the CPV consists of a Main Vocabulary and a Supplementary Vocabulary, both available in 22 official EU languages. Find bellow an example a CPV code:

Division	35000000-4	Other transport equipment
Group	35100000-5	Ships and boats
	...	...
Class	35110000-8	Ships
Category	35112000-2	Ships and similar vessels for the transport of persons or goods
Sub-category	35112100-3	Cruise ships, ferry boats and the like, primarily designed for the transport of persons
	35112110-6	Ferry boats
	35112180-7	Cruise or excursion boats n.e.c.

Figure 2 - Example of CPV Code

In order to extract valuable insights and produce meaningful analytic from the public procurement data GovWise is working with, aggregating the data according to distinct classes was deemed as a critical task. The records being extracted from the open data sources, the public procurement contracts, are already provided with a classification, the Common Procurement Vocabulary Code. On a first approach, it was soon realized that the analytics being developed considering the CPV code were inaccurate and unreliable in many cases. For some records it was misplaced, for others it was either too general or too specific. The Cosinex Report from 2017 "*Revision of CPV"* states that "In around 10% of cases the code applied did not describe the work/supply/service procured; in some 8%, the code applied was too general, and in about 4%, the code was too specific". The Rambøll Study analysis mentioned in the same report states that "In 23% of the analysed cases the CPV was incorrectly used. The incorrect use of the CPV is a relevant drawback of the CPV in the view of bidders. The extent of incorrect use is most notable for works because around 28% of the notices tested were carrying an incorrect code.". Given all these reasons, the business objective is identified, and it consists of developing a new, reliable, accurate classification for the contracts.

GOVWISE PROCUREMENT VOCABULARY (GPV)

The Govwise Procurement Vocabulary is a classification system developed to tackle the problems related with the CPV, meeting simultaneously the requirements of Govwise. The development of this classification model is an ongoing project, considering the company's needs, always subjected to further developments and improvements. The work hereby reported focuses initially on the first level of the model, the GPV0, that classifies the records in one of the 12 categories of the first level of the tree and will then proceed through the most relevant classifications at the moment this report was written, namely the "Escritório, TI, consultoria e Processos" and "Saúde e acção social" branches. Therefore, the goal for the classification model developed, at this point, is to classify each contract within this branch, as accurate as possible.

Find bellow the 'as is' GPV tree today that illustrate the levels and branches mentioned before. This classification system is continuously a work in progress, open to changes as the company's product evolves.

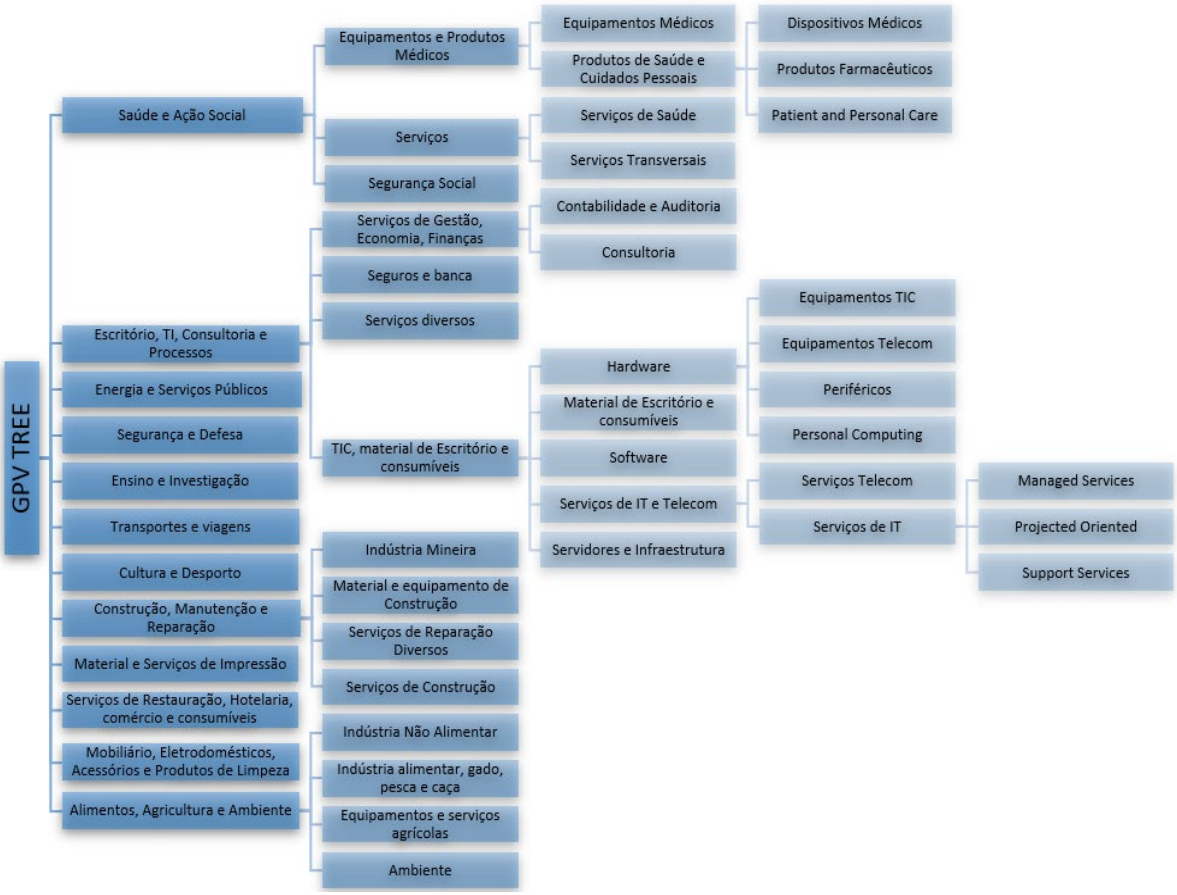

Figure 3 - Govwise Procurement Vocabulary Tree

The GPV tree is divided in 5 levels, starting with the GPV0, the level that covers all the main categories of the Portuguese public procurement, evolving to the GPV1, GPV2, GPV3, GPV4 that are respectively more granular levels of the one immediately before. The classification of the collected data into such classes allows Govwise to generate personalized notifications for each costumer according to their business areas, namely when a new public tender is released, thus giving them an advantage on the bidding process. Moreover, through the proposed classification it is then possible to develop statistical information regarding the public procurement in Portugal, for example, who are the biggest buyers, suppliers, per class of service/ product, which are the most purchased services/ products over the years within each category. This classification model allows, therefore, the filtering of the data accordingly to the needs and interest of each user. This filtering has been conveyed as a key need from the users, since the published data, so far, is published in a very noisy, unstructured way, blocking the possibility of getting relevant overview and insights on Portuguese public procurement. The accuracy of this machine learning model is expected to surpass the accuracy of the CPV attributed manually, that it is subjected to human error, which occurs rather often as concluded in previous analysis of the data.

RELATED WORK

Even though text classification has been widely explored in literature, works on the CPV specifically targeted for the Portuguese language are difficult to find. Kaan Görgün proposed a multilingual model, Multilingual-cpv-sector-classifier (2021), that takes public procurement descriptions written in 104 languages and classifies them into 45 classes from the CPV code descriptions. The F-score obtained was 0.686, and 0.631 for Portuguese language. Görgün made use of bert-base-multilingual-cased on tenders from the Economic Daily Public Procurement Data, being the output of the model the exact same structure of the current CPV code. The work from Navas-Loro et al. Multi-label Text Classification for Public Procurement in Spanish (2022) presents a very close approximation to the goal of this project, a classifier that uses textual description of the contracting process to then assign CPVs from the more generic 45 categories. This work classifies only data written in Spanish, and, similarly to Görgün (2021) aims to predict the CPV code as the output of the model. In their work, Navas-Loro et al. (2022) resorted to classifiers such as Multinomial NB (Naïve-Bayes), SVM (Support Vector Machine), KNN (K-Nearest Neighbour), Decision Tree, Random Forest and AdaBoost, achieving a maximum of 0.69 as F-score, when using the whole dataset to both train and test the data. In the research from Kayte and Schneider-Kamp A Mixed Neural Network and Support Vector Machine Model for Tender Creation in the European Union TED Database (2022), a new method is proposed to generate text automatically and subsequently classify tenders Tender Electronic Daily (TED). Their work is divided in two parts, the text generation, that through a few words generates a whole sentence, and the CPV code prediction. After generating the title, the model is meant to predict the most suitable CPV code. A Long Short-Term Memory (LSTM) model was implemented, achieving an accuracy of 97% for the text generation and 95% for the code classification, using SVM. Comparably to Görgün (2021) and Navas-Loro et al. (2022), the aim of this work was to predict the CPV code of the tender. All the input data was in English.

3. DATA AND METHODOLOGY

MODEL WORKFLOW

The model described in this report constitutes a classification problem, since the outputs are categorical. A classification algorithm assigns to new inputted data a pre-determined class, according to its features. The classification task can be divided in a binary classification, when there are two possible outcomes, or, in a multi-label classification, when there are more than two possible outcomes (Kesavaraj G., Sukumaran S., 2013). GPV constitutes, therefore, a multi-label classification problem. The main steps followed to develop a classification model can be summarized as (Sen P. C. et al., 2020):

1. Collect and clean the dataset or data pre-processing.
2. Make the classifier model initialized.
3. Split the dataset using cross-validation and feed the classifier model with training data. Python-based scikit-learn package has inbuilt methods named fit-transform (X, Y)/fit(X, Y) that map the input data member set X and corresponding label set Y to prepare the classifier model.
4. Predict the label for a new observation data. There is also a method predict (X) that returns the mapped label Y for the input instance X.
5. Evaluate error rate of the classifier model on the test dataset

The sourced data for mining comes in a raw state, meaning that it can many times include noisy data, irrelevant attributes, missing data, outliers. Therefore, data needs to be **pre-processed** before applying the classifier (Beniwal, S., 2021). Some pre-processing steps consist of data integration, data cleaning, discretization, attribute selection. The performance of the classification model is highly affected by the features that are selected to be included. Characteristics of this features, namely, their redundancy, correlation, and irrelevance must be assessed in order to mitigate the risk of hinder the model's performance, given that, many times, data mining algorithms that contain large amounts of features or attributes do not perform well. One of the most common approaches to tackle this issue is to resort to **feature selection techniques**, to transform or select a subset of the features (Ghotra B. et al., 2017). This approach will both improve the model performance, as well as prevent overfitting and provide faster and more cost-effective models (Beniwal, S., 2021). The final goal is to disclose the first optimal feature subset.

A simplified version of the steps followed to achieve the goals of this project is displayed next. The problem was identified, the raw data was collected and pre-processed, afterwards split between training and test data, the model was developed, and the classifier applied, the model was evaluated and finally the final model was chosen.

Figure 4 - Model Workflow

DATA SOURCE

The developed classification model targets both public procurement tenders and contracts established between public Portuguese entities and private parties. The EU directive on open data and the re-use of public sector information (directive (EU) 2019/1024, article 20) states that public sector bodies collect, produce, reproduce and disseminate documents to fulfil their public tasks. Use of such documents for other reasons constitutes re-use. Sequentially, the Portuguese directive launched on the 2nd of September n.º 284/2019 contemplates that public information regarding public procurement contracts shall be made available on Portal Base, namely regarding contracts and tenders that are contemplated on the CCP (Código Contratos Públicos). All the tenders released contemplated on the described categories must be published on Portal Base, containing mandatorily a few features, such as, the purchasing entity and some related relevant information (name, VAT), the procedure followed by the entity to announce bidding, the value of the tender, the description of the tender, and the CPV(s). On a later stage, once the tender results in a formal contract, the information is updated, now including the tenderer, or supplier, and the updates on the value of the contract, if applicable. The features enumerated previously are then used to build the classification model, accordingly to the information provided by the publishers.

DATA UNDERSTANDING AND PRE-PROCESSING

The data in use for the development of the classification model is stored in GovWise's data base. The sources are composed by open source portals made available for the public in general to consult. Each record represents a contract signed between a Portuguese public entity and a private entity. The collected data includes the main features of the contract, namely its signing date, the entities that celebrated the contracted, its value, among others. These are key attributes for the end goal classification of the developed model. Most attributes are extracted directly from the data source, others already went through a data pre-processing on SQL Server to generate new information with added value.

Data da publicação	05-10-2022
Tipos de contrato	Aquisição de bens móveis
Nº do acordo quadro	Não aplicável.
Descrição do acordo quadro	Não aplicável.
Tipo de procedimento	Ajuste Direto Regime Geral
Descrição	Aquisição de reagentes e consumíveis para laboratórios do IPP _2022"
Fundamentação	Artigo 20.º, n.º 1, alínea d) do Código dos Contratos Públicos
Fundamentação para recurso ao ajuste direto (se aplicável)	ausência de recursos próprios
Entidades adjudicantes	Instituto Politécnico de Portalegre (600028348)
Entidades adjudicatárias	Laborspirit, Lda (507485149)
Objeto do contrato	Aquisição de reagentes e consumíveis para laboratórios do IPP _2022"
Procedimento centralizado	-
CPVs	44423000-1
Data do contrato	08-09-2022
Preço contratual	12.343,62 €
Prazo de execução	60 dias
Local de execução	Portugal, Portalegre, Portalegre

Figure 5 - Example of a public procurement contract and some relevant features

The data selected to build the classification model consists of 458 092 records, where each record represents a public procurement contract, celebrated between a Portuguese public entity and a private company, between the years of 2019 and 2021. The original dataset is composed by 458 092 rows and 43 columns. Out of these 42 columns, 10 have missing values, some of them reaching 99% of missing values and will therefore be disregarded as features for the model. Additionally, some of the columns represent redundant information, e.g. “parsedPrice” and “initialContractualPrice”, where the first one is already a transformed version of the second one, the “€” sign and the decimal separator were removed to facilitate the use of the data. On a first “naked eye” analysis, the most relevant features for the classification are the object description (the contract summary), the supplier, the buyer, the respective VAT numbers, the CPVs and the contract types. There are no missing values for all of these features, and all of them are attributed with the correct data type. The columns that constitute the original dataset are the following:

Table 1 - The original dataset is constituted by the following columns:

Column name	Missings (%)
id	0 %
contractingProcedureUrl	78.77 %
publicationDate	0 %
endOfContractType	88.79 %
totalEffectivePrice	88.89 %
frameworkAgreementProcedureId	0 %
frameworkAgreementProcedureDescription	0 %
contractFundamentationType	0 %
contractingProcedureType	0 %
contractTypes	0 %
executionDeadline	0 %
cpvs	0 %
executionPlace	0 %
nonWrittenContractJustificationTypes	64.10 %
initialContractualPrice	0 %
objectBriefDescription	0 %
contractStatus	99.95 %
signingDate	0 %
contracted	0 %
contracting	0 %
cocontratantes	0 %
cpvCount	0 %
cpvFirst	0 %
cpv3est	0 %
cpv3estdesc	0 %
elegibleRenovation	0 %
renovationDate	0 %
country	0 %
municipality	12.18 %
location	18.10 %
parsedPrice	0 %
contractedCount	0 %
contractedFirstNif	1.52 %
contractingCount	0 %
contractingFirstNif	2.23 %
munguess	56.93 %
store_date	0 %
executionDeadlineDays	0 %
F_LeadProcessed	0 %
F_RenewalProcessed	0 %
N_NumberContestants	3.66 %
E_ORIGINAD	88.85%

Many of the features above went already through pre-processing techniques, after the raw collection from the source. This pre-processing runs on the pipeline every time new data is scrapped from the source, keeping the same data base structure. The columns that have no missing values are predominantly the ones that are original from the data source. Some of the columns added in a way that they keep only the first value of the column, already in order to facilitate future data analysis, and consequently might also be more relevant for the development of the model, some examples are the CPV First, that keeps only the first CPV code from each specific contract, in cases that the contract is attributed with more than one CPV code. Another example is the contracting first NIF, that keep only the NIF of the contracting entity that appears first. This could, indeed, generate bias on the analysis of the data, but the majority of the records (99%) have attributed both only one CPV (Figure 7) and one contracting NIF (Figure 8). Therefore, for the sake of simplicity, these features, along with the contracted NIF, will keep only the first record. The features with more than 50% of missing values will not be used to train the model, given both that they have too many missing values and additional those specific features do not bring added value compared to the other. Some of them are already a variance of other existing features, namely the 'Munguess', that derived from the 'Municipality'. The columns 'Contracting Procedure URL', 'End of Contract Type', 'Total Effective Price', 'Contract Status', 'Non Written Contract Justification Types', 'E Origin Ad' and 'Munguess' will therefore not be included in the training of the model.

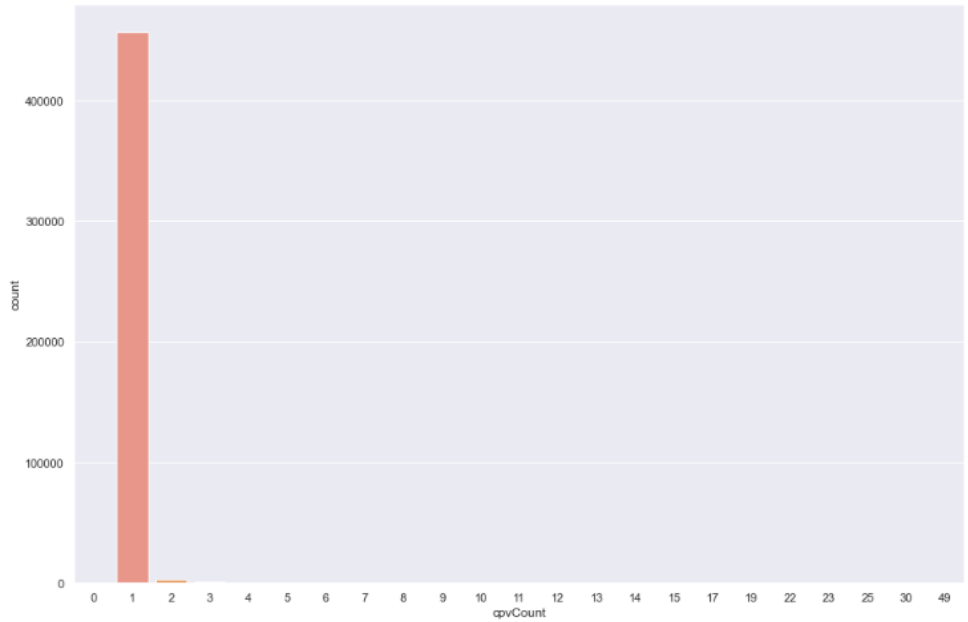

Figure 6 - Frequency of records with 1 or more CPV code

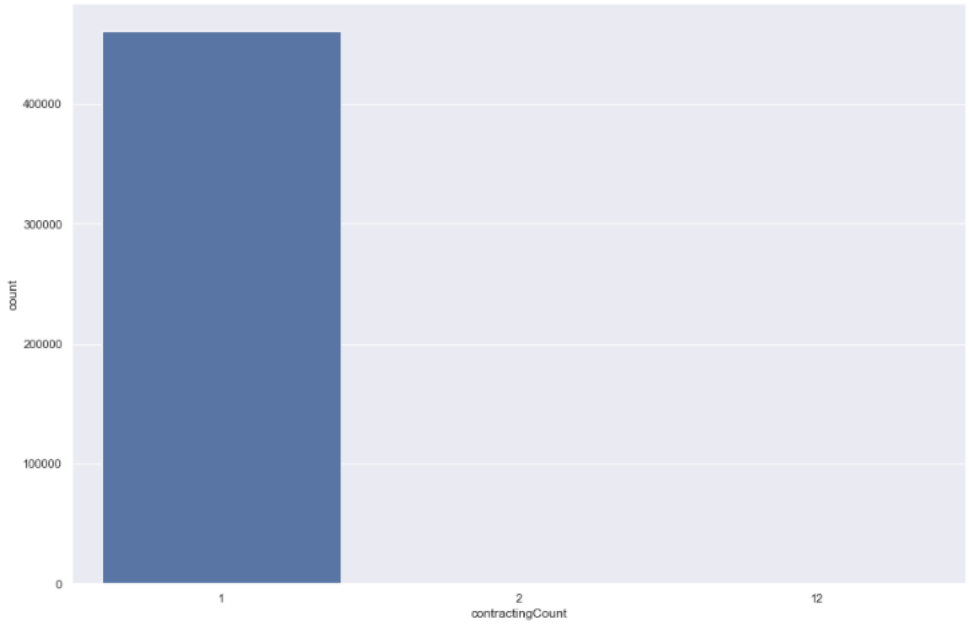

Figure 7 - Frequency of records with 1 or more CPV code

FEATURE SELECTION

Features highly correlated will not be kept to train the model, given that they don't add any value and increase the complexity and computational effort. The correlation of features was therefore tested. **Pearson** - The Pearson correlation evaluates the linear relationship between two continuous variables.

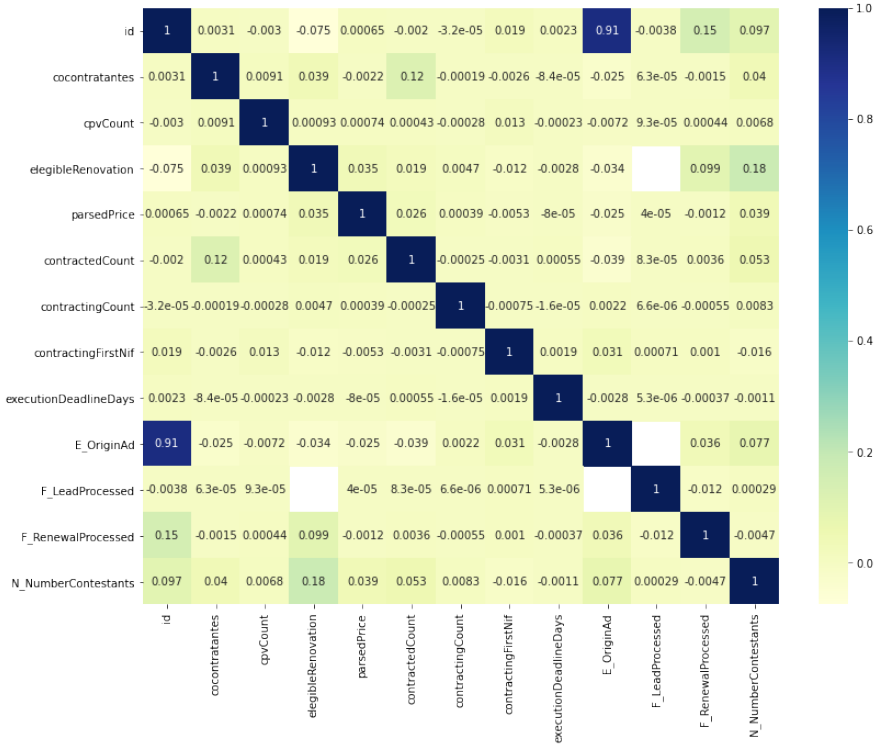

Figure 8 - Pearson Correlation

After the correlation analysis, features such as E_OriginAd, as concluded already before due to the missing values, or id and elegibleRenovation or executionDeadlineDays should not be both kept for training the model. According to the feature selection techniques applied, five features were kept to train the model, namely the Feature A, the CPV, the Feature B, the Feature C and the Feature D, given their importance, uniqueness, lack of correlation and relevance to the model, and the original ID of each record was set as the index. To the selected features, data pre-processing techniques were applied. The real names of the features are not disclosed to ensure the data privacy of the company is respected.

Table 2 - Selected Features to train the model

SELECTED FEATURES	DATA TYPE
ID (AS INDEX)	Int64
FEATURE A	Object
CPVFIRST	Object
FEATURE B	Object
FEATURE C	Int64
FEATURE D	Object

A) Common Procurement Value Code (CPV)

Starting with the CPV, three new columns were developed, the CPV1, CPV2 and CPV3, where the CPV1 keeps the first two digits of the attributed CPV, the CPV2 the first 3 and the CPV3 the first four. These three CPV levels still exhibit rather high accuracy values respectively, in contrast to the fourth level on, where the accuracy drops and the CPV tends to get too specific and present more errors. The division on the three levels of the CPV was made given that the output classification is also divided in more than one level. The distribution of the CPV1 on the dataset is rather asymmetric as it is possible to analyse on figure 11 below.

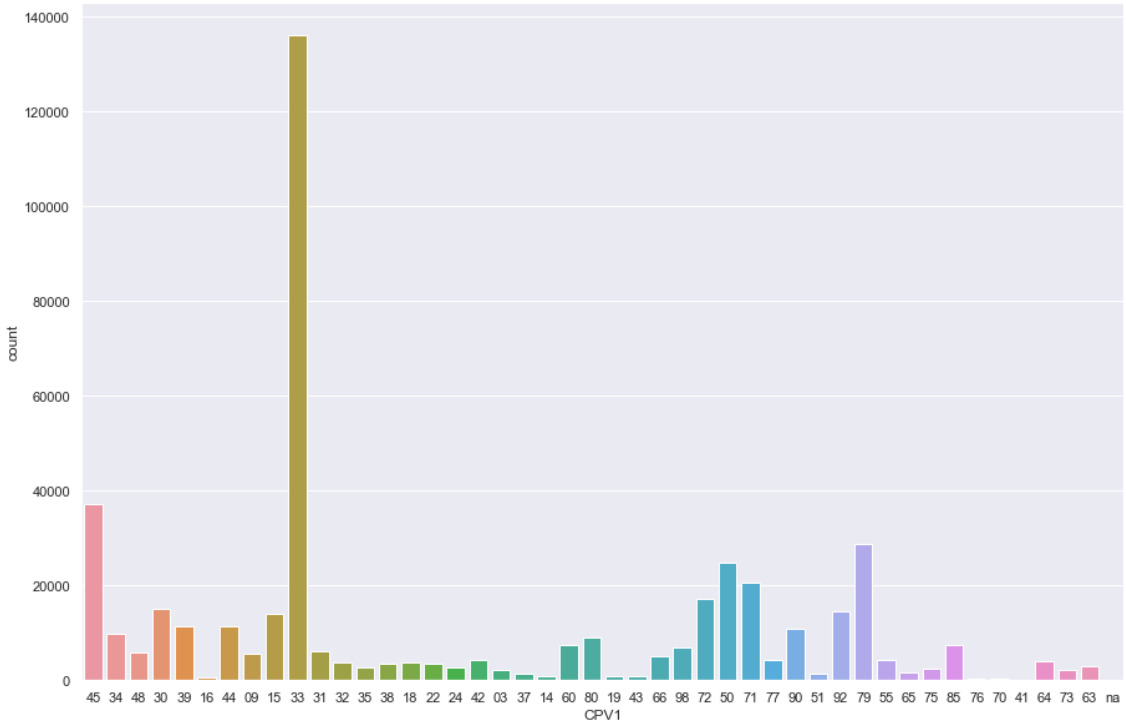

Figure 9 - Distribution of CPV level 1

As it is possible to observe on figure 11, a very considerable number of contracts established in the period of the observed data set (2019-2021) falls under the CPV 33, at the level 1, which stands for medical equipment, pharmaceuticals, and personal care products. The explanation is the fact that during this time frame, namely 2020 and 2021 the world, Portugal not being excluded, was going through the covid-19 pandemic, and public institutions acquired therefore massively amounts of health-related products. Even though this affects the distribution of this feature, there are still considerable amounts of records associated with the remaining CPVs. The second most common CPV in the dataset is the number 45, which is associated with Construction Work. Unlike the health products, construction work tends to be, steadily, one of the most used CPV both among different countries as well as over time.

In a premature attempt to ensure the best accuracy of the model, it was decided to input manual corrections to this specific feature. To the developed model, 1900 records were manually corrected, that had been detected to have the incorrect CPV code attributed. The manual corrections were a commercial decision and are still to be evaluated, considering the results of this intervention on the data, given that is a very time-consuming task. Finally, all records that don't include the CPV field are excluded for the model, there was no case for the current dataset, but there might be as the gathered data increases.

B) Feature A

Feature A is a text feature. Therefore, in order to get the best performance of the model, Natural Language Processing techniques were applied to this feature. It is important to reinforce that the entire data set is in Portuguese. A function was applied to the model to remove the punctuation as well as to lower all the capital letters. Additionally, stop words were removed, specifically Portuguese stop words such as 'a' and 'o', that don't bring any added value to the model, and this way the processing time is reduced and no unnecessary space on the database is occupied. Stemming was also applied, to normalize the word by truncating it to its stem word, but it resulted in the reduction of the accuracy of the model, the more complex form of the words was actually necessary to perform certain distinctions, therefore it wasn't kept to train the model. The output of this processing can be seen in the example bellow:

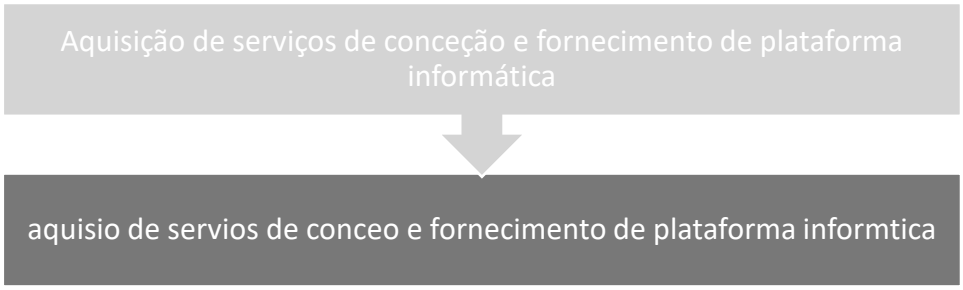

Figure 10 - Example of text feature after pre-processing

C) Feature B and Feature C

All features were transformed in 'object' data type, including the Feature C that was previously of the type integer. To these two features, given that they are more likely to have missing values, a univariate imputer was included in the model 'simple imputer', meant to fill in the missing values with rather simple strategies using descriptive statistics such as mean, median or most frequent.

The current model was developed having already in sight that it would be incorporated in a pipeline. As more data is saved on the data base, in real time, the model is supposed to classify each new record within the GPV classification. The data set is split using train_test_split, with the test size of 20%. Through the process of the pipeline, there are pre-processing steps inputted on the data, ending with the final estimator. For the five selected features to train the model, on each iteration, records containing missing values are dropped. The data frame containing the training features was converted to a NumPy array, making the processing tasks more efficient, given that it works well in large data sets, and tend to be faster comparing to working with data frames. Given that all the features are categorical, they were encoded using One hot encoder, since many scikit-learn estimators can only be fed by encoded categorical features. For the text feature, tf-idfTransformer was used in order to scale down the impact of tokens that occur very frequently, and that are therefore less informative than the ones that occur less frequently. Still for the text feature, Count Vectorizer was used for text pre-processing, namely, to remove stop words, remove the accents and to ignore terms that have a very low frequency (< 4). As mentioned before, for all the features, except for the text feature, Simple Imputer to replace missing values, using simple descriptive statistic (mean, median, most frequent), along each column, or using a constant value. Lastly, the estimator is applied, and the model is evaluated using accuracy, precision and recall. The function returns the records and each predicted GPV category for each one.

ALGORITHM SELECTION AND EVALUATION

The selection of the best machine learning algorithm is one of the most crucial tasks of the development of every machine learning model. This step will deeply affect the accuracy of the results, hence the performance of the model. The choice must consider the specifications of the model, and as it is stated by the “no free lunch” theorem (Wolpert & Macready, 1997), there is no one size fits all when it comes to select the right classifier. It is, although, possible to estimate a suitable selection of machine learning algorithms. This selection is dependent on the algorithm’s application and evaluation, given that it has been proved that no algorithm is universally more suitable on all the datasets, given their specific features and characteristics (Ali R. et al., 2016). In a classification model, a valid mapping function is performed based on the training dataset, to then predict the class label for the new data inputs. The classifier algorithm learns from the training set, particularly from its attributes or features, and later assign the new data inputs to a particular class (Sen P. C. et al., 2020). Many times, the best performing supervised learning models represent ensembles of base-level classifiers. The disadvantages are that this approach requires a lot of space to store all the classifiers, a long time to execute them, especially when it comes to large test sets and when the computational power is limited (Buciluǎ C. et al., 2006).

For the evaluation of classification models, fault is the basic concept to consider. If the predicted class by the model differs from the actual class in the test cases, then there are errors in the classification. One of the most popular evaluation metrics of the performance of classification models is the confusion matrix. It is represented as a table and can assume four primary values that can be gotten directly from examining the confusion matrix: true positive, true negative, false positive and false negative (Bisong, E., 2019). From these four primary values, there are three other metrics that can provide more information on the performance of the model: accuracy, precision and recall.

- **Accuracy**: It represents the ratio between the number of correctly classified examples and the total number of classified examples. Applying to the confusion matrix, it is the ration of the sum of true positive, TP, and true negative, TN, to the total classified examples. Some disadvantages of accuracy as an evaluation metric are that it neglects the differences between types of errors, and it is very dependent on the distribution of each class in the dataset (Novakovic J. et al. 2017). The **classification error rate**, inversely to the accuracy measures the ration of incorrect predictions over the total number of instances evaluated.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Precision**: Precision is the ratio of true positive, TP, to the sum of true positive, TP, and false positive, FP. In other words, precision measures the fraction of results that are correctly predicted as positive over all the results that the algorithm predicts as positive (Bisong, E., 2019).

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

- **Recall**: Recall is the ratio of true positive, TP, to the sum of true positive, TP, and false negative, FN. In other words, recall retrieves the fraction of results that are correctly predicted as positive over all the results that are positive. The sum TP + FN is also known as condition positive (Bisong, E., 2019).

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

Many of the classification metrics are defined for binary classification by default, in order to extend them to multiclass, several averaging techniques are used. **F-Score** represents the harmonic mean

between recall and precision values, to get a high F1 both false positives and false negatives must be low (Hossin, M. & Sulaiman, M.N., 2015). The advantages of accuracy is that it is easy to compute, it applies to multi-class and multi-label problems and easy to understand. Disadvantages are that it produces less discriminable values, which leads to less discriminating power to accuracy in selecting and determining the optimal the optimal classifier (Hossin, M. & Sulaiman, M.N., 2015).

Accuracy has been proven not to be such an appropriate performance measure for imbalanced classification problems, which is the case, due to the fact that discrepant number of occurrences between classes will weight more on the classes with the majority of the examples and way less in the classes with the minority of occurrences. For this reason, precision, recall and F-Score were also used to evaluate the performance of the model. To each of these metrics the weighted average was applied, accounting for the class imbalance, meaning that for each class the mean of each metric is calculated and afterwards is weighted given the number of actual occurrences of the class in the target dataset.

BENCHMARKING OF THE MODEL

Decision Tree

Decision trees are formed by the root node, branches and leaf nodes. On each internal node, an attribute is tested, the decision rule come in each branch, and class label, as a result, is on each leaf node. The root node is the parent of all nodes, and it is the topmost node in the tree. (Patel H. & Prajapati P., 2018) The decision tree deals with the problem step by step, data streaming with functioning a logic in each step, leading to the prediction of label on unlabelled data. It is statistical technique used both for regression and classification (P. Sen et al., 2020).

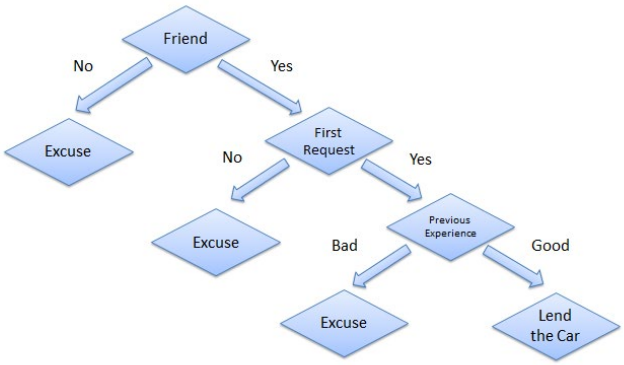

Figure 11 - Example of a decision tree

Decision trees work with satisfactory accuracy and relatively fast. The algorithm splits the set of data items into two or more homogeneous sets based on most significant attribute to make as distinct groups as possible (P. Sen et al., 2020). To split up the data into different groups, various techniques like information gain, chi-square, Gini, entropy, etc. are used. **Entropy** is considered to have a good discriminatory power for classification. The advantages of the decision tree are that it is a simple method, easy to understand and visualize, fast, requires less data pre-processing, and can deal with both categorical and numerical data. The disadvantages are that sometimes this algorithm may lead to a complex tree structure not being generalized enough, besides being a rather unstable model (P. Sen et al., 2020). Additionally, decision trees get computationally expensive as the data set grows, and once a mistake is done at a higher level, any sub-tree will present wrong results (Narayanan, U. et al., 2017). When applied, this method didn't involve additional pre-processing of the data. The default parameters of the sklearn library were used when applying the Decision Tree classifier. The obtained results were:

Table 3 - Performance Metrics results for the Decision Tree Model

	GPV0	*GPV1*	*GPV2*	*GPV3*	*GPV 4*	*Weighted Average*
Total Nº of Contracts	461 512	351 428	202 824	171 458	19 061	
Accuracy	0,9262	0,9082	0,9604	0,8153	0,8415	**0,9096**
Error Rate	0,0738	0,0918	0,0396	0,1847	0,1585	**0,0904**
Precision	0,9182	0,9075	0,9604	0,8198	0,8397	**0,9070**
F-Score	0,9262	0,9082	0,9588	0,8153	0,8415	**0,9093**

Logistic Regression

Logistic regression is a supervised machine learning algorithm, developed for learning classification problems. It maps a function from the features of the dataset to the targets to predict the probability that a new example belongs to one of the target classes (Bisong, E., 2019). When it comes to multinomial classification, the logistic regression setup (i.e., the cost function and optimization procedure) is structurally similar to logistic regression, being the only difference the output with more than 2 classes. In order to build a classification model with k classes, the multinomial logistic model is formally defined as (Bisong, E., 2019):

$$\hat{y}(k) = \theta_0^k + \theta_1^k x_1 + \theta_2^k x_2 + \cdots + \theta_n^k x_n$$

The preceding model takes into consideration the parameters for the k different classes. The softmax function, used to compute the probability that an instance belongs to one of the K classes when K > 2, is formally written as (Bisong, E., 2019):

$$p(k) = \sigma(\hat{y}(k))_i = \frac{e^{\hat{y}(k)_i}}{\sum_{j=1}^{k} e^{\hat{y}(k)_j}}$$

Adaptative Boosting

The idea behind boosting is to combine the output of multiple simple classifiers 'weak learners' into a powerful composed classifier, the 'strong learner'. It is an ensemble method meant to improve the performance of any learning algorithm. **AdaBoost** algorithm is considered to be one of the most accurate algorithms for multi-label classification (Al-Salemi, B. et al., 2016). It works by iteratively building a committee of weak hypotheses of decision stumps. In each learning round, all the features are examined, but only one feature is used to build a new weak hypothesis. This mechanism entails a high degree of computational complexity, namely as the dataset size increases. There are, although, a few ways to accelerate AdaBoost according to Al-Salemi, B. et al. (2016):

- Dimension-reduction-based acceleration: as in any other supervised machine learning classification algorithm, this includes pre-processing techniques on the original data, namely feature selection methods to reduce the number of features to be modelled.
- Representation-based acceleration: targeted for text classification models, Al-Salemi et al. (2016) proposed LDA-AdaBoost.MH, in which a Latent Dirichlet Allocation (LDA) topics model is used to extract latent topics as features to represent texts. Experimental results proved that representing texts using a small number of topics significantly accelerates AdaBoost.MH learning and performance.
- Weak-Learning-based acceleration: involves changing the number of pivot terms used to build the weak hypotheses, changing the base learner and reducing the search space of pivot terms in each boosting round.

Al-Salemi, B. et al. presented an improved version of AdaBoost, called the Rank-and-filter-based algorithm, **RFBoost**, that retains the simplicity and generality of AdaBoost with a new method of accelerating the weak learning based on reducing the search space of pivot terms in each boosting round. RFBoost filters a fixed number of feature terms in each boosting round to build a new weak hypothesis. The features are firstly ranked and then in each iteration, only the top k features among which the pivot term is most likely located to build a new weak hypothesis. RFBoost was not applied

in this model, but it is definitely to be kept in mind for future improvements, given that this classifier entails a high degree of computational complexity, namely as the dataset size increases, which would lead to the need of applying other strategies in order to keep the good performance of the model as more data is collected. The results obtained when recurring to the AdaBoost classifier were:

Table 4 - Performance metrics results for the AdaBoost Model

	GPV0	*GPV1*	*GPV2*	*GPV3*	*GPV 4*	*Weighted Average*
Total Nº of Contracts	461 106	351 957	200 124	171 232	18 226	
Accuracy	0,8883	0,9030	0,9602	0,8033	0,7915	**0,8910**
Error Rate	0,1116	0,0970	0,0398	0,1967	0,2085	**0,1090**
Precision	0,8752	0,9060	0,9530	0,8067	0,7906	**0,8369**
F-Score	0,8677	0,9003	0,9517	0,7928	0,7915	**0,8737**

Support Vector Machine

Support vector machine is an advanced supervised algorithm that can deal with both regression and classification tasks, although is considered better for classification. It can handle multiple continuous and categorical instances. The dataset records are represented, each having "n" number of features plotted as points in a n-dimensional space segregated into classes by clear margin widest possible, the hyperplane. Into that same n-dimensional space, data items are then mapped to get the prediction of the category they belong to, based on the side of hyperplane they fall (P. Sen et al., 2020). Known for their robustness, good generalization ability, and unique global optimum solutions, SVMs are probably the most popular machine learning approach for supervised learning, yet their principle is very simple. SVMs require that all the training data is stored in memory during the training phase, while the parameters of the SVM are learned. Once the model parameters are identified, SVM depends only on a subset of these training instances, called support vectors, for future prediction (Awad M. & Khanna R., 2015). Support vectors define the margins of the hyperplanes. Support vectors are found after an optimization step involving an objective function regularized by an error term and a constraint, using Lagrangian relaxation. The complexity of the classification task with SVM depends on the number of support vectors rather than the dimensionality of the input space. The number of support vectors that are ultimately retained from the original dataset is data dependent and varies, based on the data complexity, which is captured by the data dimensionality and class separability (Awad M. & Khanna R., 2015). Some advantages of SVM are that it shows a noticeable hike in performance where the "n" of the n-dimensional space is greater than the total size of sample set. Therefore, when dealing with high-dimensional data, it represents a good choice. If the hyperplane is well built, it shows high performance, being also memory efficient (P. Sen et al., 2020). The disadvantages are that the training

time is rather high when compared to other algorithms, so when dealing with very large datasets, the prediction task becomes very slow (P. Sen et al., 2020). It has also been showed that the performance of the algorithm is quite sensitive to noisy data (Narayanan, U. et al., 2017). Also, the drastic change in the position of the hyperplane due to a single additional point shows that the classifier is susceptible to high variability and can overfit the training data (Bisong, E., 2019). The results when applying the SVM as classifier were:

Table 5 - Performance metrics results for the SVM Model

	GPV0	*GPV1*	*GPV2*	*GPV3*	*GPV 4*	*Weighted Average*
Total Nº of Contracts	460777	351639	202821	171697	18672	
Accuracy	0.9336	0.9291	0.9681	0.8759	0.8798	**0.9290**
Error Rate	0.0664	0.0709	0.0319	0.1240	0.1202	**0.0710**
Precision	0.9390	0.9252	0.9672	0.8559	0.8803	**0.8968**
F-Score	0.9336	0.9387	0.9681	0.8522	0.8798	**0.9163**

Instance-based Learning Algorithm

Instanced-Based Learning (IBL) algorithm classifies or estimates new examples by comparing them to the ones already seen and in memory (Martin, 1995). This kind of algorithms are particularly useful for a problem that needs to be locally optimized. For huge datasets, the Instanced-based algorithms usually fail to generalize and are computational expensive. The Instanced-Based classifiers are also called as Lazy Learners because the important instances are determined every time the classification occurs, and a new local model is created. The computational work is mainly performed during the classification phase rather than in the learning phase (Figueiredo L., 2020). It becomes intuitive to realise that in Instanced-Based model the separation between training and testing phase is not as clear as it is in other algorithms. Nevertheless, the fundamental principle of classification algorithms remains the same, and IBL assume that similar instances must have similar classifications (Aha et al., 1991).

For a data record *t* to be classified, its *k* nearest neighbours are retrieved, and this forms a neighbourhood of *t*. Majority voting among the data records in the neighbourhood is usually used to decide the classification for *t* with or without consideration of distance-based weighting. However, to apply kNN we need to choose an appropriate value for *k*, and the success of classification is very much dependent on this value. The major drawbacks with respect to kNN are its low efficiency - being a lazy learning method prohibits it in many applications such as dynamic web mining for a large repository, and its dependency on the selection of a “good value” for *k* (Guo G. et al., 2003). The results obtained when applying KNN as classifier were:

Table 6 - Performance metrics results for the KNN Model

	GPV0	*GPV1*	*GPV2*	*GPV3*	*GPV 4*	*Weighted Average*
Total Nº of Contracts	461 106	317 085	208 283	171 999	19 021	
Accuracy	0,9198	0,8905	0,9418	0,8006	0,8530	**0,8973**
Error Rate	0,0802	0,1095	0,0582	0,1994	0,1470	**0,1027**
Precision	0,9177	0,8880	0,94362	0,8037	0,8492	**0,8966**
F-Score	0,9198	0,8905	0,9418	0,8006	0,8530	**0,8973**

Statistics-Based Algorithm

Statistic-base algorithms make use of distributive statistics to generalize the problem, look into the distribution structure to perform the predicting task. Naïve Bayes arises as one of the most popular statistics-based algorithms. The Naïve Bayes classifier produces the probabilities for every case, and then predicts the highest probability outcome (P. Sen et al., 2020). Naïve Bayesian Classifier has a very good accuracy in classification for large set of data. The problem with this method is that it take the entire attribute independently thus if the attributes are independent then only Naïve Bayesian classifier gives it full accuracy. Naive Bayesian classifier is a very strong statistical classifier when it comes to accuracy. But as the name suggests it is naive and takes the presumption that all attributes are independent of each other (Narayanan U., et al. 2017). These classifiers are capable of handling an arbitrary number of independent continuous and categorical variables efficiently. Let us consider a set of variables, X = {x1, x2, x3,…, xt}; it is required to find out the posterior probability for the event Cj from the sample space set C = {c1, c2, c3,…, ct}. Simply, the predictor is X and C is the set of categorical levels present in the dependent variable. Applying Bayes' rule:

$$P (Cj|x1, x2, x3,..., xt) \cdot P (x1, x2, x3,..., xt |Cj) P (Cj)$$

Where P(Cj |x1, x2, x3,…, xt) is the posterior probability that is the probability of the event X belonging to Cj is indicated. In Naive Bayes, there is an assumption that the conditional probabilities of the independent variables have statistical independence. Using Bayes' rule, a new case X is labeled with a class level Cj that accomplishes the highest posterior probability. The assumption that the predictor variables are independent of each other is not always accurate. This assumption makes the classification process simpler, as it allows the class conditional densities P (xd |Cj) to be calculated for each variable separately and thus a multidimensional task is reduced to some one-dimensional tasks. More precisely, it converts a high-dimensional density estimation task to a one-dimensional kernel density estimation. Classification task remains unaffected as this assumption does not greatly affect

the posterior probabilities, mainly in regions located closely around the decision boundaries (P. Sen et al., 2020).

Python and Libraries

The presented model was entirely developed in Python. The elected programming language was due to its popularity as well as its numerous available resources, namely when it comes to the all the libraries for Data Cleaning and Classification tasks related. Python is known for its high capacity of data processing, being especially useful for the development of machine learning models, and in this case, classification models. In the current project several python libraries were used, displayed in the following table:

Table 7 - Python libraries used in the model

LIBRARY	DESCRIPTION
PANDAS	Data analysis and manipulation
NUMPY	Working with arrays
MATPLOTLIB	Graphical Display
SKLEARN	Machine learning modelling
SEABORN	Data visualization
STATSMODELS	Estimating statistical models and performing statistical tests

4. RESULTS AND DISCUSSION

The main goal of this project was to develop a supervised classification model accurate enough to reclassify new data in real time, based on historical data already collected between 2019 and 2021. The model classifies the data inputs into pre-defined classes, according to the business needs of the company. At the time the model was developed, the focus was on the IT and health sector, given the nature of the target clients' business activity. The aim is to keep on expanding the model on the other GPV0 sectors, such as construction, culture, sports, etc, so the new contracts are classified to the most refined level possible on the GPV1, GPV2, GPV3 and GPV4 branches, while ensuring its good performance. Afterwards are displayed a couple of examples of classified records by the model with the best performance:

	Feature A	CPV1	CPV2	CPV3	Feature B	Feature C	Feature D	GPV0	GPV1	GPV2	GPV3	GPV4
4663		72	720	7200				Escritorio, TI, consultoria e processos	TIC, material de escritorio e consumiveis	Servicos IT e telecom	Servicos IT	Project Oriented
5160		72	720	7200				Escritorio, TI, consultoria e processos	TIC, material de escritorio e consumiveis	Servicos IT e telecom	Servicos IT	Project Oriented

Figure 12 - Examples of the classification output

Benchmarking of the models

The model that obtained the best performance is not described in the last chapter, and won't be fully disclosed, namely regarding the classifier in use, due to the data privacy rules of the company. The performance results of this model were the following:

Table 8 - Performance metrics results of the best performing model

	GPV0	*GPV1*	*GPV2*	*GPV3*	*GPV 4*	*Weighted Average*
Total Nº of Contracts	461108	351403	203799	171715	19008	
Accuracy	0.9445	0.9288	0.9652	0.8362	0.8712	**0.9269**
Error Rate	0.0555	0.0712	0.0348	0.1638	0.1288	**0.0731**
Precision	0.9417	0.9227	0.9642	0.8389	0.8690	**0.9242**
F-Score	0.9419	0.9257	0.9635	0.8362	0.8683	**0.9247**

A total of 461108 records were classified with this model, being 461 108 attributed to the first level (GPV0), 351 403 to the second level (GPV1), 203 799 to third level (GPV2), 171 715 to the fourth level (GPV 3) and 19 008 to the fifth level (GPV4). It makes sense that the number of classified records reduces as the level gets more specific, since some of the initial sectors on GPV 0 are still not ramified into more refined classes on the following branches.

To get an overview of the performance of all the models, and in order to allow their comparison, find next a table with the condensed results for each evaluation metric used, additionally find in appendix A the quantitative results of each model in each GPV level.

Table 9 - Comparison of the performance metrics results of all the models

	Weighted Accuracy	*Weighted Class. Error Rate*	*Weighted Precision*	*Weighted F-Score*	*Number of Records Classified*
AdaBoost	0,8910	0,1090	0,8369	0,8623	461 106
SVM	0,9290	0,0710	0,8968	0,9163	460 777
KNN	0,8973	0,1027	0,8966	0,8933	461 106
Decision Tree	0,9096	0,0904	0,9070	0,9081	461 512
Final Model	0,9269	0,0731	0,9242	0,9247	461 108

The model with the best accuracy was the SVM, but the Final Model outperformed both in Precision and F-Score. Given that the input data is rather unbalanced, the f-Score gains prevalence over the accuracy as a performance metric. Afterwards are displayed a visual distribution of the classifications of the Final Model in each GPV level.

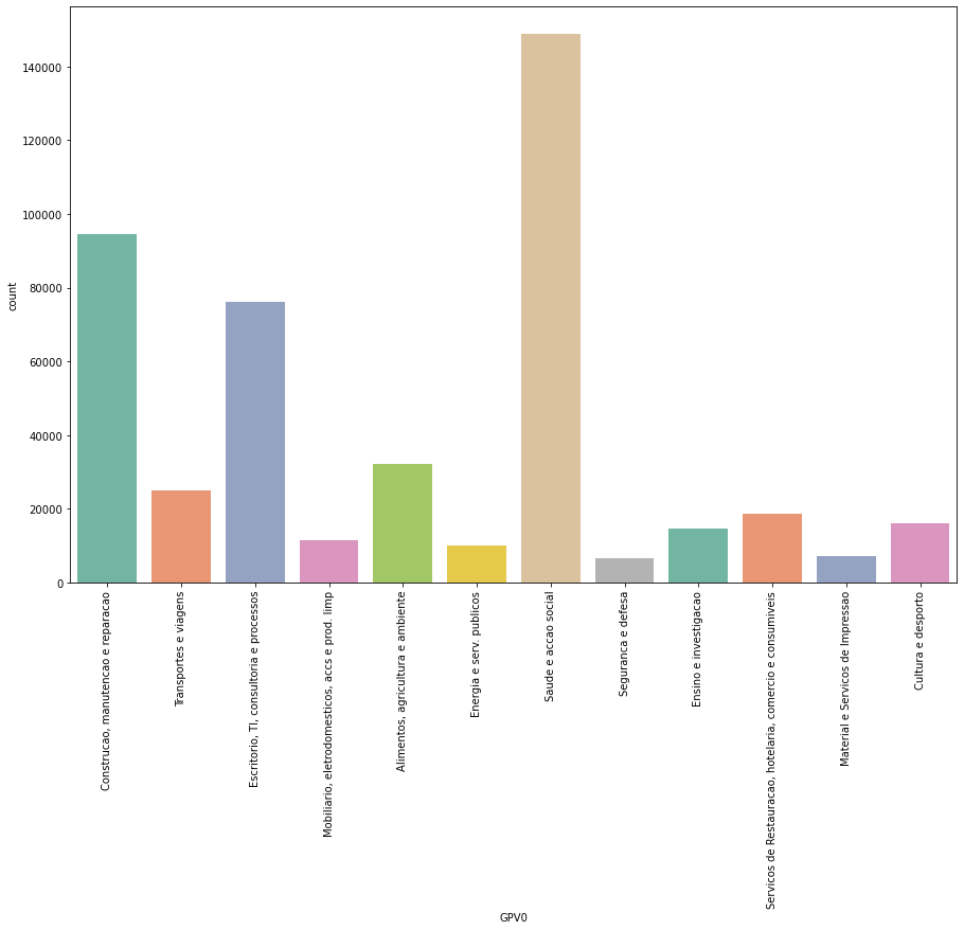

Figure 13 - GPV 0 Results Final Model

As initially analysed on the input data exploration section, a big part of the contracts is classified within the health sector (Saude e accao social), and the secondly within the construction sector (construcao, manutencao e reparacao). This might be an indicator that the model is performing well.

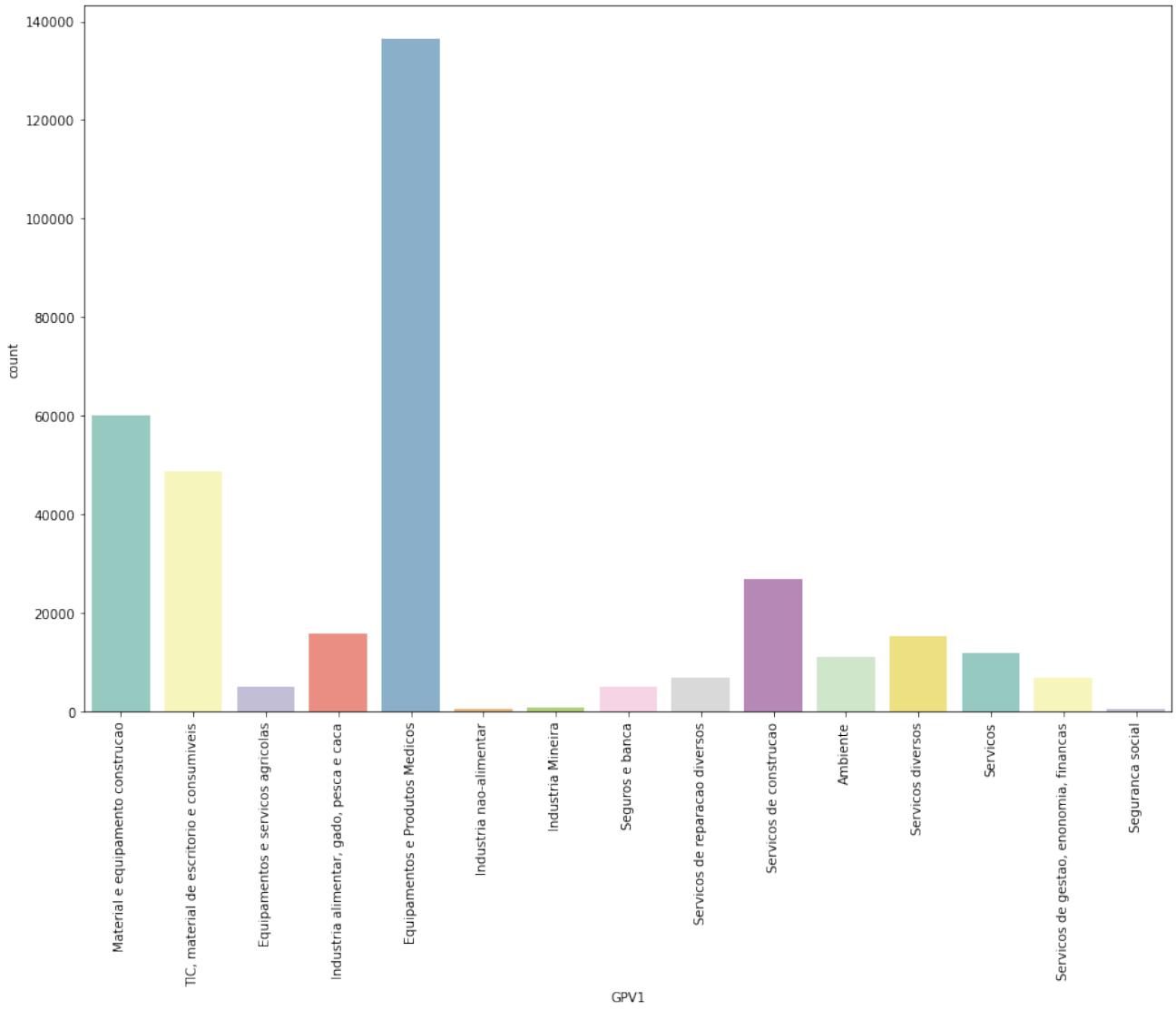

Figure 14 - GPV 1 Results Final Model

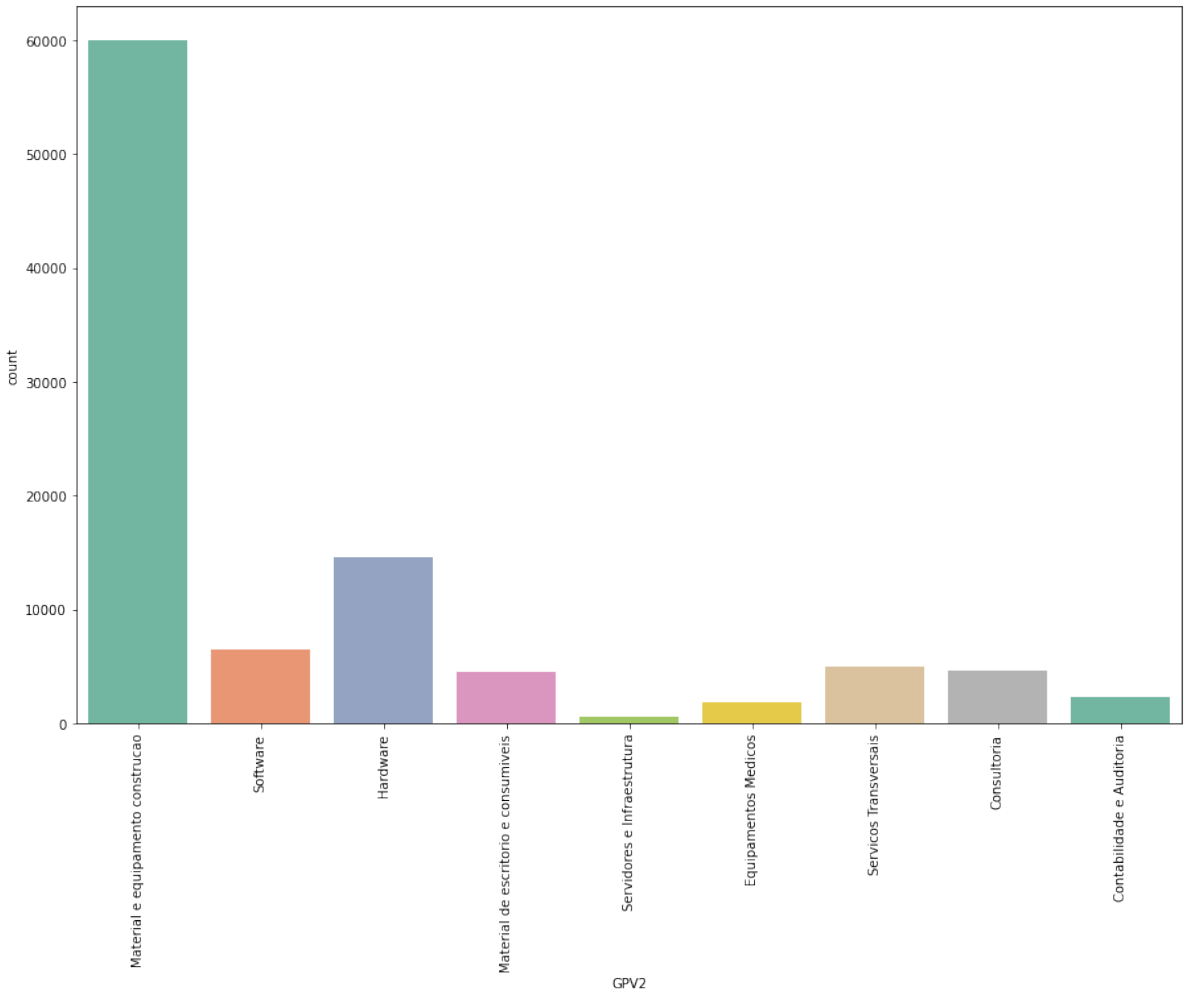

Figure 15 - GPV 2 Results Final Model

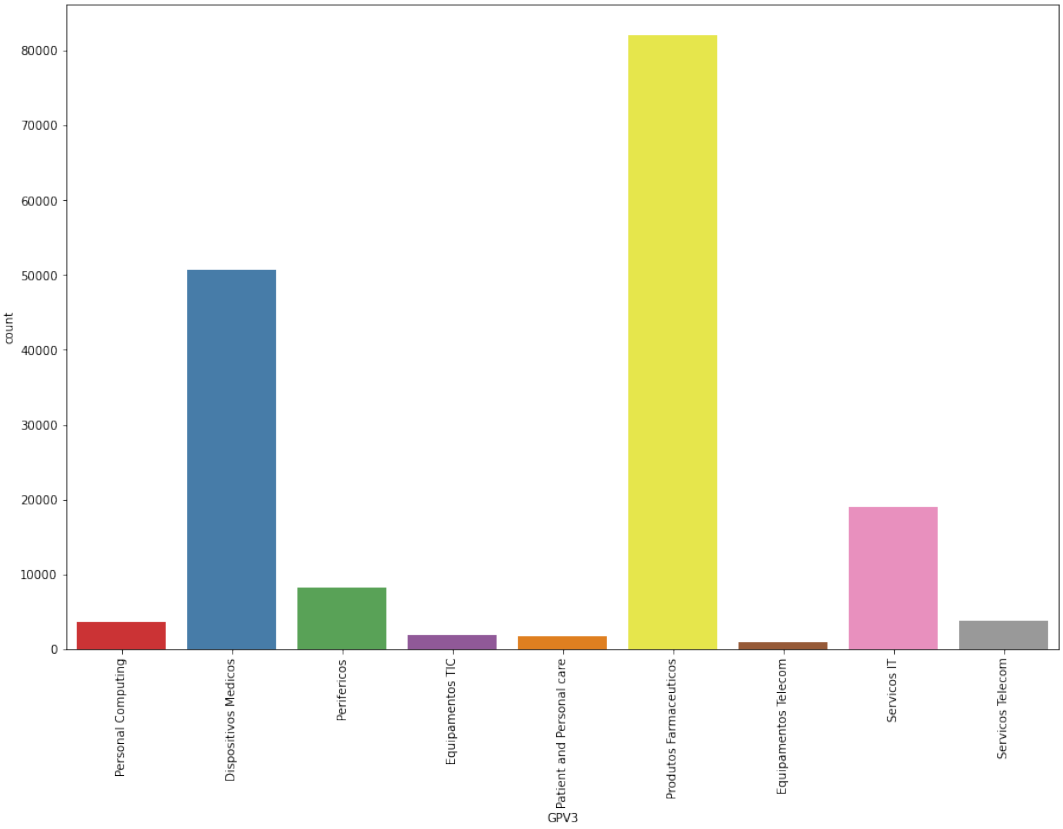

Figure 16 - GPV 3 Results Final Model

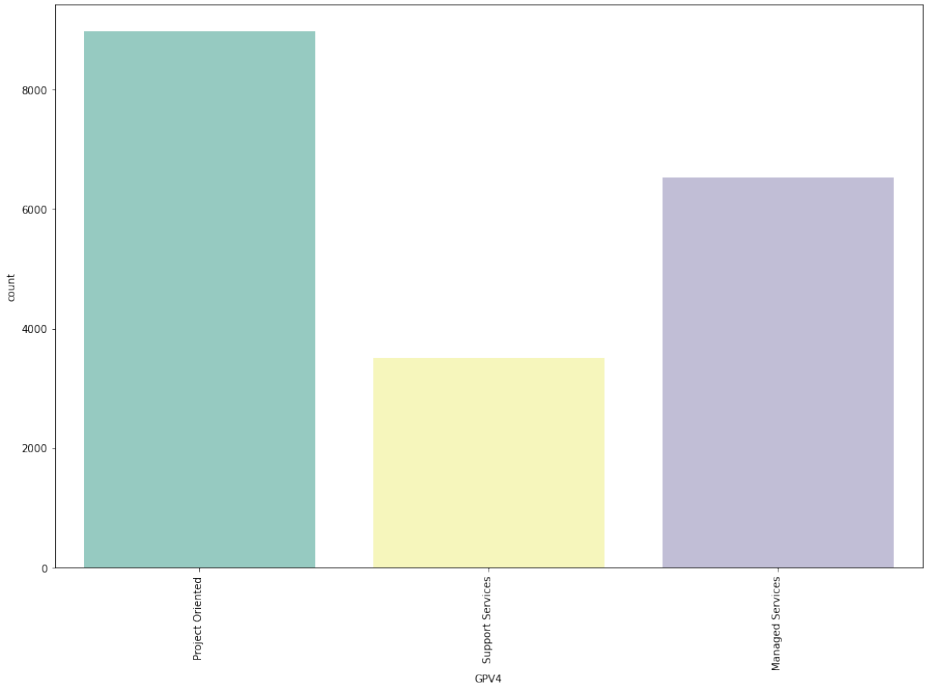

Figure 17 - GPV 4 Results Final Model

The results displayed above do not guarantee, although, the real accuracy of the model. For this reason, random sample tests were conducted to ensure that the model was indeed performing well. Over the first tests, there were very few examples spotted wrongly classified. These tests are although meant to be performed as new data is classified, and new outputs generated, and the wrong classifications that are spotted must be corrected directly on the data base, that will then keep on being used to train and improve the model, thus becoming more accurate. Bellow an example of a few classified records are displayed, just as displayed to the end user (client), that can also report if mistakes are found on the classification.

Buyer	Supplier	NIF Buyer	NIF Supplier	GPV (1) Category	GPV (2) Category	Original CPVs	Contract Description	Contract Value €	Contract Date	Contract Type	Link
Agrupamento de Escolas João de Deus Faro	Faro Doce Pastelaria Lda	600085350	505671034	Industria alimentar, gado, pesca e caca	Industria alimentar, gado, pesca e caca	15810000-9, Produtos de panificação, produtos frescos de pastelaria e bolos	Aquisição de Produtos de Pastelaria e bolos	€20,000	2019-12-31	Consulta Prévia	Open contract
Agrupamento de Escolas Frei João de Vila do Conde	Sunol Compal Marcas SA	600075516	505042037	Industria alimentar, gado, pesca e caca	Industria alimentar, gado, pesca e caca	15894700-8, Produtos de mercearia fina	Bufetes 2020	€4,932	2019-12-31	Consulta Prévia	Open contract
Freguesia de Alcochete	Ana Teresa Pereira Martins	506925447	218325070	Servicos de Restauracao, hotelaria, comercio e consumiveis	Servicos de Restauracao, hotelaria, comercio e consumiveis	98000000-3, Outros serviços comunitários, sociais e pessoais	Prestação de serviços de inserção social à população da freguesia	€15,600	2019-12-31	Consulta Prévia	Open contract

Figure 18 - Examples of the classified results by GPV (user view)

Considering the display of the classification on Figure 19, the end user can always report in case of misclassification. Although, so far, after the implementation of the model on the data, these are very rare events.

5. CONCLUSIONS AND FUTURE WORKS

The initial aim of the project was to create a unique classification system targeted specifically for the contracts and tenders of the Portuguese public procurement. The need for such system surged from the lack of accuracy of the already given CPV code on these records, as well as the fact that this code can be many times either too specific, or too general, and being vulnerable to human error, making it difficult to analyse the collected data in useful, insightful ways.

Even though the original dataset was unbalanced, the results of the classification model are rather satisfactory. The Final Model outperformed the other studied models in terms of precision and F-Score, being therefore the chosen one to be implement on the pipeline to classify new inputted data. After the implementation of the model on new data, human analysis tests were performed that reinforced the good accuracy of the model. There are still classification errors, that are meant to be tackled as more data is collected and used to train and improve the model, as well as more human corrections are made to the already wrongly classified data. Overall, these errors mostly occur on the more specific levels of the model, namely **GPV 3** and **GPV 4**. This might be explained by the fact that there is significantly less records at these levels, therefore the accuracy would improve as more data is inputted to train the model. Additionally, for these specific levels a different approach, namely data pre-processing, or even a different classifier could as well improve the performance of the model. This project is an ongoing work in progress, and there is, certainly, a lot of room for improvement.

The main limitations of the project were the time constraints, given that it is a very important part of the product, it should therefore be done as fast as possible with the best accuracy possible. It was although developed considering that it is a work in process and improvements are to be made in the long run. Additionally, the quality of the original data presented a challenge since many of the features have missing values and even incorrect values that at this point haven't been detected but that surely decrease the performance of the model. On the other hand, it was this human error that was intended to tackle with the current model in the first place.

For future developments, one clear next step includes completing the GPV tree has much as possible, considering that so far the records that don't fall under 'Saúde e ação social' or 'Escritório, TI, consultoria e processos' are classified in rather general categories. The pre-processing of the data is also subject to enhancements, being an important step for these classification problems.

Moreover, other classification methodologies could be explored, that could potentially improve the performance of the model, namely **BERT** (Bidirectional Encoder Representations from Transformers). BERT has been proven to perform very well when it comes to NLP tasks, namely text

classification tasks. Additionally, it can process high amounts of data at a comparably fast pace and low computational effort, and lastly, it performs well in several language, which would be a great advantage for Govwise's future expansion to other markets that make use of different languages on their public procurement. However, the use of Active Learning with BERT based models for multi class text classification has not been studied extensively (Prabhu S. et. al, 2021). The results Prabhu et al. (2021) obtained on the multi-class text classification work using BERT were rather satisfactory and is definitely a methodology to keep in mind for future experiments on the reported classification model. Overall, the results have proven to be very positive, both from a business perspective as well as from a starting point to develop a classification system more efficient and fruitful when it comes to the creation of an effective decision support system and to generating value on the increasing public data.

PRACTICAL APPLICATION

The ultimate goal of the project is to display the collected data in a way that provides valuable insights to the end user through user-friendly dashboards and data visualization tools. For this purpose, the classification of the data into the GPV classes was necessary. A few examples of the practical application of the model are displayed next.

I. Total value in Euros (€) per GPV category (filtered on **Industria alimentar, gado, pesca e caça**; **Cultura e desporto**; **Equipamentos e serviços agrícolas**; **Contabilidade e auditoria**; **Construção, Manutenção e reparação**, so the visualization of the chart is clearer), over the year of 2019 in public procurement contracts in Portugal. The values were intentionally hidden, to safeguard the data privacy of the company.

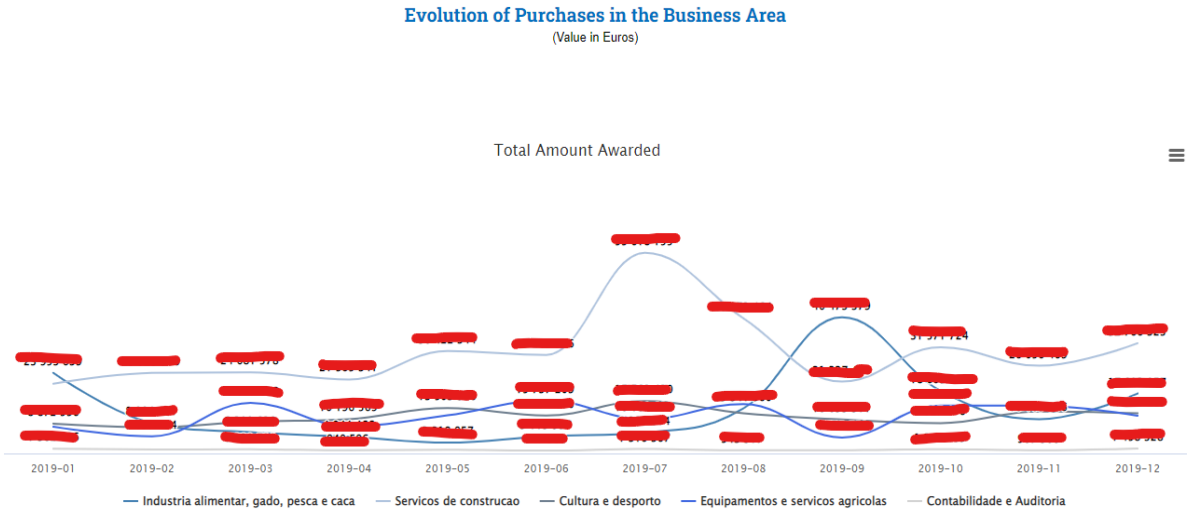

Figure 19 - Practical application of the model I

II. Total number of contracts per GPV category (for each filtered category), over the year of 2019 in public procurement contracts in Portugal.

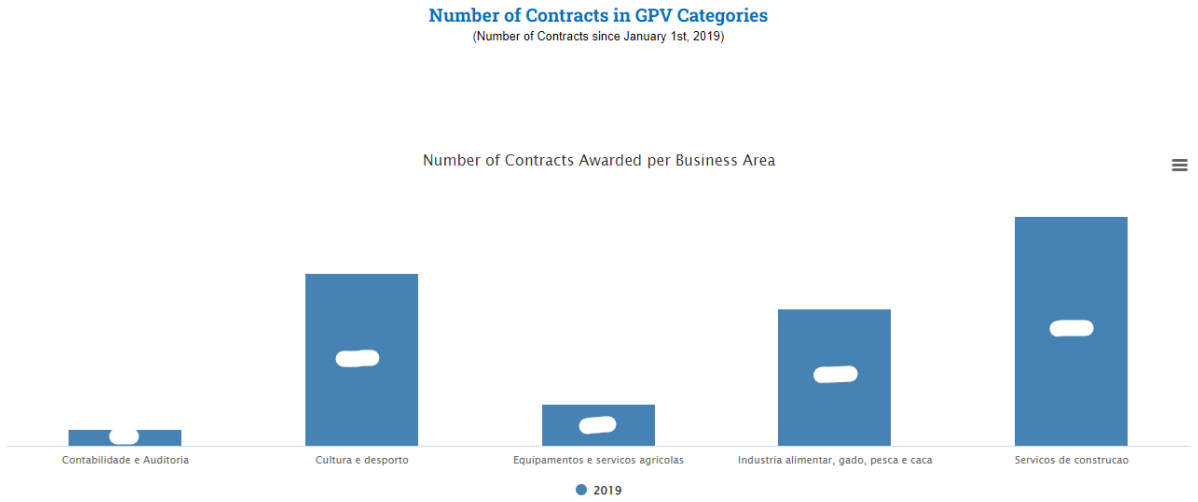

Figure 20 - Practical application of the model II

III. Main buyers and main providers per GPV category (for each filtered category), over the year of 2019 in public procurement contracts in Portugal, in total value (€) of contracts.

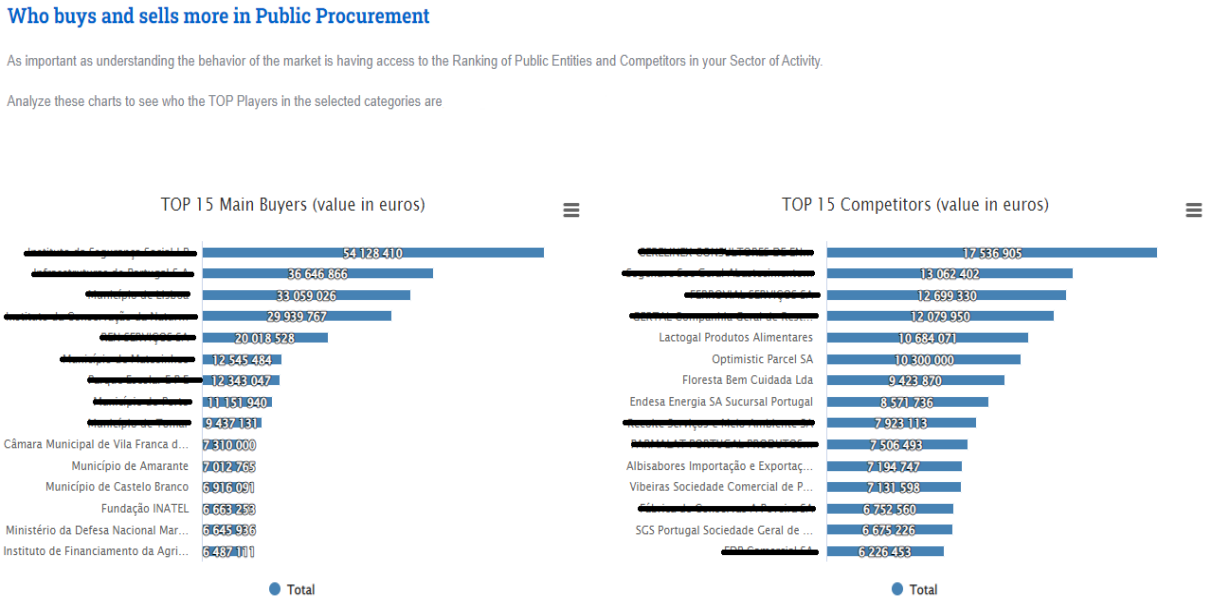

Figure 21 - Practical application of the model III

IV. Applying now the model on real time data, it is possible to analyse on which GPV categories are there open tenders that the suppliers can bid.

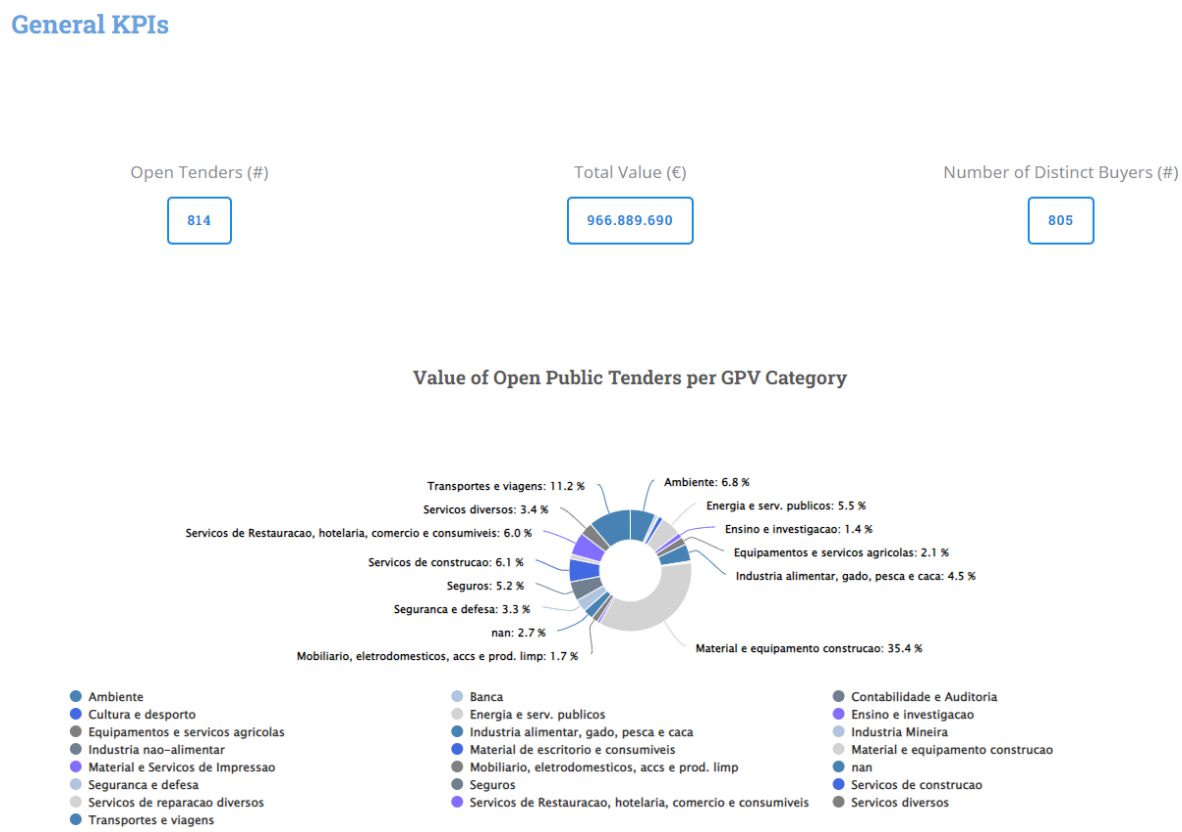

Figure 22 - Practical application of the model IV

BIBLIOGRAPHICAL REFERENCES

Agência para a Modernização Administrativa, IP. (n.d.). OpenGov. Open Government Partnership Portugal. Retrieved November 16, 2022, from https://ogp.eportugal.gov.pt/-/reforco-da-transparencia-na-contratacao-publica

Instituto dos Mercados Públicos, do Imobiliário e da Construção. (n.d.). Plataforma de dados abertos da AMA, Dados.Gov, os dados dos contratos públicos no portal BASE. Plataforma de dados abertos da ama, Dados.gov, Os Dados dos Contratos Públicos no portal base. Retrieved November 16, 2022, from https://www.base.gov.pt/Base4/pt/noticias/2019/plataforma-de-dados-abertos-da-ama-dados-gov-os-dados-dos-contratos-publicos-no-portal-base/

Patel, H. H., & Prajapati, P. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering, 6(10), 74–78. https://doi.org/10.26438/ijcse/v6i10.7478

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2888, 986–996. https://doi.org/10.1007/978-3-540-39964-3_62

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. Machine Learning, 6(1),37–66. https://doi.org/10.1007/bf00153759

Martin, B. (1995). Instance-based Learning: Nearest Neighbour with Generalisation. Department of Computer Science, University of Waikato.

Berends, J., Carrara, W. & Radu, C. (2020). Analytical Report 9: The Economic Benefits of Open Data. In European Data Portal (Publications Office of the European Union, 2020)

European commission (2017). Final report | Revision of CPV "Consultancy Services for Common Procurement Vocabulary expert group"

McDowell, C. (2021, October 5). The challenges posed by officially published open data. Open Access Government. Retrieved November 16, 2022, from https://www.openaccessgovernment.org/published-open-data/67170/

Gurin, J. (2014). Big Data and Open Data: How Open Will the Future Be? I/S: A Journal of Law and Policy for the Information Society, 10(3), 691–704. Retrieved from http://www.opendatanow.com/2013/09/back-to-school-with-open-data/.

Shukla, S., Kukade, V., & Mujawar, S. (2015). Big Data: Concept, Handling and Challenges: An Overview. International Journal of Computer Applications, 114(11), 6–9. https://doi.org/10.5120/20020-1537

Alvarez-Rodríguez, J. M., Labra-Gayo, J. E., Rodriguez-Gonzalez, A., & De Pablos, P. O. (2014). Empowering the access to public procurement opportunities by means of linking controlled vocabularies. A case study of Product Scheme Classifications in the European e-Procurement

sector. Computers in Human Behavior, 30, 674–688. https://doi.org/10.1016/j.chb.2013.07.046

Kayte, S., & Schneider-Kamp, P. (2019). A mixed neural network and support vector machine model for tender creation in the European union TED database. In IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (Vol. 3, pp. 139–145). SciTePress. https://doi.org/10.5220/0008362701390145

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. In Advances in Intelligent Systems and Computing (Vol. 937, pp. 99–111). Springer Verlag. https://doi.org/10.1007/978-981-13-7403-6_11

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37–53.

Mitchell, T. (1997). Machine Learning. New York: McGraw Hill. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR (p. 2).

Shearer, C., Watson, H. J., Grecich, D. G., Moss, L., Adelman, S., Hammer, K., & Herdlein, S. a. (2000). The CRISP-DM model: The New Blueprint for Data Mining. Journal of Data Warehousing.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., … Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Azevedo, A., & Santos, M. F. (2008). KDD, semma and CRISP-DM: A parallel overview. In MCCSIS'08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008 (pp. 182–185).

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), 217–222. Retrieved from http://www.ijisr.issr-journals.org/

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining (pp. 1–34). MIT Press.

Ali, R., Lee, S., & Chung, T. C. (2017). Accurate multi-criteria decision making methodology for recommending machine learning algorithm. Expert Systems with Applications, 71, 257–278. https://doi.org/10.1016/j.eswa.2016.11.034

Buciluǎ C., Caruana R., & Niculescu-Mizil A. (2006). Model compression. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '06.

Ghotra, B., McIntosh, S., & Hassan, A. E. (2017). A large-scale study of the impact of feature selection techniques on defect classification models. In IEEE International Working Conference on

Mining Software Repositories (pp. 146–157). IEEE Computer Society. https://doi.org/10.1109/MSR.2017.18

Beniwal, S., & Arora, J. (2012). Classification and Feature Selection Techniques in Data Mining. International Journal of Engineering Research & Technology (IJERT), 1(6), 1–6.

Narayanan, U., Unnikrishnan, A., Paul, V., & Joseph, S. (2018). A survey on various supervised classification algorithms. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017 (pp. 2118–2124). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICECDS.2017.8389824

Awad, M., & Khanna, R. (2015). Support Vector Machines for Classification. In Efficient Learning Machines (pp. 39–66). Apress. https://doi.org/10.1007/978-1-4302-5990-9_3

Bisong, E. (2019). Logistic Regression. In Building Machine Learning and Deep Learning Models on Google Cloud Platform (pp. 231–241). Apress. https://doi.org/10.1007/978-1-4842-4470- 8_19

Novaković, J. Dj., Veljović, A., Ilić, S. S., Papić, Ž., & Milica, T. (2017). Evaluation of Classification Models in Machine Learning. Theory and Applications of Mathematics & Computer Science (Vol. 7, pp. 39 – 46). Retrieved from https://uav.ro/applications/se/journal/index.php/TAMCS/article/view/158

Prabhu S., Mohamed M. & Misra H. (2021). Multi-class Text Classification using BERT-based Active Learning.https://doi.org/10.48550/arXiv.2104.14289

Silva, P. F. P. (2017). As políticas de Open Data em Portugal: análise da sua implementação e impacto [Master's Thesis, Universidade de Coimbra.].

Figuiredo, L. M. (2020). An Overview of the main Machine Learning Models. [Master's Thesis, Universidade Nova de Lisboa].

Contratação Pública em Portugal 2021. (2022). In *Instituto Dos Mercados Públicos, Do Imobiliário E Da Construção, I.P.* Direção Financeira, de Estudos e de Estratégia Instituto dos Mercados Públicos, do Imobiliário e da Construção, I.P. https://www.base.gov.pt/Base4/media/burjhsro/relat%C3%B3rio-anual-da-contrata%C3%A7%C3%A3o-p%C3%BAblica-2021.pdf

Görgün, M. K. (2021). *Multilingual-cpv-sector-classifier*. Huggingface. https://huggingface.co/MKaan/multilingual-cpv-sector-classifier

Navas-Loro, M., Garijo, D., & Corcho, O. (2022). *Multi-label Text Classification for Public Procurement in Spanish*. Universidad Politécnica de Madrid.

EU Directive:

Directive (EU) 2019/1024. *On open data and the re-use of public sector information.* European parliament and council on Official Journal of the European Union

APPENDIX A - Number of contracts in each GPV level per model

For GPV1 to GPV4 codes were used to ease the reading of the table. These codes are described in appendix B.

GPV 0	GPV 1	GPV 2	GPV 3	GPV 4	Total AdaBoost	Total KNN	Total SVM	Total Decision Tree	Total Final Model
Saúde e ação social					**146072**	**148443**	**148184**	**147689**	**148766**
	EPM				136280	136280	136280	136454	136281
		EM			1693	1696	1676	2415	1903
		PSCP			134587	134584	134604	134039	134378
			PF		90346	84602	82220	86134	81803
			DM		42588	48253	50713	46022	50895
			PPC		1653	1729	1671	1883	1680
	SSo				649	649	649	650	649
	S				9143	11514	11255	10585	11836
		Ssa			6790	6790	6790	6799	6790
		ST			2353	4724	4465	3786	5046
Escritório, TI, consultoria e processos					**75562**	**76045**	**75562**	**76271**	**76191**
	SB				5044	5065	5066	5043	5065
	SD				15817	15463	15537	15443	15444
	SGEF				6655	6673	6782	6881	6853
		CA			2345	2286	2275	2424	2301
		C			4310	4387	4507	4457	4552
	TMEC				48046	15681	48504	48904	48829
		MEC			4415	10962	4425	4491	4506
		SW			6478	4946	6501	6458	6467
		SI			508	493	485	536	519
		SITel			22027	22825	22485	22865	22810
			ST		3801	3804	3813	3804	3802
			SIT		18226	19021	18672	19061	19008
				MS	5727	6635	6329	6484	6634
				PO	9967	8863	9053	8854	8855
				Sse	2532	3523	3290	3723	3519
		H			14618	14590	14608	14554	14527
			P		8666	7804	8511	8083	8184
			Etel		798	880	796	838	874
			ETIC		1982	1844	1920	1887	1860
			PC		3172	4062	3381	3746	3609
Construção, manutenção e reparação					**98238**	**93678**	**95484**	**95370**	**94364**
	SC				26923	26477	27521	28184	26710
	MECo				59939	59934	59939	60001	59938
	SRD				10481	6367	7129	6288	6820
	IM				895	900	895	897	896
Alimentos, agricultura e ambiente					**32085**	**32082**	**32082**	**32098**	**32082**
	IAGPC				15689	15681	15688	15688	15680
	A				10965	10962	10962	10970	10962
	ESA				4946	4946	4946	4953	4946
	INA				485	493	486	487	494
Ensino e investigação					**14748**	**14742**	**14749**	**14765**	**14748**
Cultura e desporto					**15966**	**15966**	**15966**	**15969**	**15966**
Transportes e viagens					**24569**	**25960**	**24790**	**25176**	**24966**
Material e Serviços de Impressão					**7145**	**7349**	**7204**	**7282**	**7234**
Energia e serviços Públicos					**10201**	**10201**	**10201**	**10209**	**10201**
Segurança e defesa					**6467**	**6520**	**6439**	**6618**	**6473**
Mobiliário, eletrodomésticos, acessórios e produtos limpeza					**11437**	**11437**	**11437**	**11438**	**11437**
Serviços de Restauração, hotelaria, comércio e consumíveis					**18616**	**18683**	**18679**	**18627**	**18680**
Total					**461106**	**461106**	**460777**	**461512**	**461108**

APPENDIX B - Guide to the codes included in the results table in appendix A