FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

IA.SAE - Prevenção e Fiscalização Inteligente de Risco Alimentar

José João Coelho Dias

Mestrado Integrado em Bioengenharia

Supervisor: Alexandra Alves Oliveira

Co-Supervisor: Brígida Mónica Faria

June 28, 2020

IA.SAE - Prevenção e Fiscalização Inteligente de Risco Alimentar

José João Coelho Dias

Mestrado Integrado em Bioengenharia

June 28, 2020

Abstract

The food chain considers the exchange of food between several actors, beginning in the food harvest until reaching the final consumer. These exchanges must obey a set of rules so that the products do not endanger the public health or economic infrastructure.

This way, in 2005, the Portuguese government created the Autoridade de Segurança Alimentar e Económica in order to inspect and prevent compliance with legislation on economic activities, in the food and non-food sectors. The institution also assures the assessment and communication of risks in the food chain, communicating with similar entities, at an international level. Due to the high number of economic agents operating in the country, this institution is in need of an automatic and explicable system for inspection planning based on the estimated level of risk of the agent.

The system must focus on risk matrices that take into account the consumption volume, growth rate, service and other factors. For a better understanding of the subject in question, a set of state of art works was collected and analyzed in detail.

In the analysed collection of work, was possible to be acquainted to techniques that allowed risk identification. Hazard Analysis and Critical Control Point systems identify food security gaps in the food chain and Risk Based Systems identify the most likely risks to occur and their severity. Both systems can act in a preventive way by establishing measures to minimize these risks.

In the literature was reported the use of machine learning techniques to determine risk in different areas have been proved useful reaching good performances.The ensemble classifiers outperformed the simple machine learning classifiers. More recently,deep learning techniques have been used to determine the risk trough the food chain. This was made recurring to Long Term Memory neural networks with very promising results.

With the knowledge of the literature, different machine learning techniques were implemented using a dataset of food inspections in Chicago in order to classify the risk level. This preliminary work achieved an accuracy of 0.8752 for the XGBoost model,resulting in an explainable and reproducible methods to be achieved.

As a primary work, the results were quite positive. However, there is still a some margin for improvement.

Resumo

A cadeia alimentar pode ser definida como a troca de alimentos entre vários atores desde da colheita dos alimentos até chegar ao consumidor final. Estas trocas têm que obedecer a conjunto de regras para não existirem perigos para a saúde publica ou infrações económicas.

Deste modo em 2005, o governo Português criou a Autoridade de Segurança Alimentar e Económica com o intuito de fiscalizar e prevenir o cumprimento da legislação das atividades económicas, nos setores alimentar e não alimentar. A ASAE também assegura a avaliação e comunicação dos riscos na cadeia alimentar, sendo o organismo nacional de ligação com as entidades congéneres, a nível europeu e internacional.

Devido ao grande volume de agentes económicos a operar no país, esta instituição necessita de um sistema de planeamento de inspecções baseada na estimativa de risco para cada agente económico.

O sistema deverá se focar em matrizes de risco tendo em atenção o volume de consumo, a taxa de incumprimento, o serviço e outros fatores. Para uma melhor compreensão do assunto em questão, foi recolhido e detalhadamente analisado um conjunto de obras literárias.

Através desta análise foi possivel conhecer técnicas que fazem gestão alimentar como os Hazard Analysis and Critical Control Point, que permitem identificar faltas de segurança do produto alimentar que ocorrem na cadeia alimentar, e de Risk Based Systems, que identificam os riscos mais prováveis e a sua gravidade de acontecer, mostram ter a capacidade de agir preventivamente, estabelecendo medidas de forma a minimizar estes riscos.

Na literatura o uso de sistemas de aprendizagem computacional mostraram ser sistemas automáticos bastante fiáveis para determinar risco em diferentes áreas. Sendo que algoritmos de combinação de classicadores mostraram ser mais vantajosos em relação aos outros. Mais recentemente, tem se utilizado técnicas de deep learning, redes neuronais Long Short Term Memory para proceder à estimativa de risco com resultados bastante prometedores.

Tendo todo esta pesquisa em consideração, foram implementadas diferentes técnicas, tendo sido alcançadas exatidões de 0.875 para o modelo XGBoost,permitindo criar um modelo explicável e reprodutível no que toca a determinar a frequencia minima de inspeção. Sendo um trabalho primário no campo, os resultados foram bastante positivos.

No entanto, ainda existe uma margem para melhorias.

Acknowledgements

Em primeiro lugar, quero expressar os meus profundos agradecimentos à Professora Alexandra Alves Oliveira pela sua orientação, disponibilidade e as suas valiosas ideias, sem as quais a dissertação não poderia ser concluída. O agradecimento engloba, ainda, a Professora Brígida Mónica Faria e todos os elementos do LIACC pertencentes ao projeto, pelas indicações e pela amabilidade com que me incluíram no grupo. Um enorme agradecimento à minha família, que durante toda a vida me apoiou. Sinto que apenas palavras não sejam capazes de o descrever. Um especial agradecimento àqueles que desde que me lembro estão sempre ao meu lado, em momentos de felicidade ou de tristeza, de boémia ou de seriedade ou simplesmente a rematar uma bola sem considerar a distanciamento de segurança, fruto da pandemia. Finalmente, um especial agradecimento a todos os que me acolheram quando entrei neste território desconhecido que era a Faculdade de Engenharia do Porto. A todos que fizeram destes 5 anos um conjunto de memórias que, se descritas, dariam um livro. A todas as partilhas, a todas as gargalhadas, a todas tardes passadas a procrastinar, a todos os cafés, a todos os acidentes culinários, a todo o lavar de roupa suja, a todos os desafios, a todos aqueles que gostam de ver o mundo em chamas. Muito Obrigado! E um especial obrigado ao Covid-19, por não me ter permitido criar ainda mais memórias.

José Dias

"Home is behind, the world ahead,
and there are many paths to tread
through shadows to the edge of night,
until the stars are all alight."

J.R.R.Tolkien

Contents

List of Figures

List of Tables

Abbreviations

ASAE	Autoridade de Segurança Alimentar e Económica
BSE	Bovine Spongiform Encephalopathy
EFSA	European Food Safety Authority
LIACC	Laboratório de Inteligência Artificial e Ciência de Computadores
POPFAA	Plano Operacional Práticas Fraudulentas na Área Alimentar
QUAR	Quadro de Avalição e Responsabilizaçao
HACCP	Hazard Analysis and Critical Control Points
CCP	Critical Control Points
Uk	United Kingdom
RBI	Risk Based System
RC	Risk Control
FSMS	Food safety Management System
Eu	Euopean Union
ISO	International Organization for Standardization
LR	Logistic Regression
SVM	Support Vector Machine
k-nn	K-nearest-neighbours
GBDT	Gradient Boosting Decions Tree
RF	Random Forest
AC	Accuracy
LSTM	Long Short Term Memory
XGBoost	eXtreme Gradient Boosting INE
National Statistical Institute	

Chapter 1

Introduction

1.1 Context

In the late 1990s, a series of food related crises such as Salmonella, Bovine spongiform encephalopathy (BSE), commonly know as mad cow disease [26], lead to the establishment of the European Food Safety Authority (EFSA) in 2002 [1]. This agency works independently of the European legislative and executive institutions and has the mission to communicate risks associated with the food chain and provides scientific advice [1].

In Portugal, a similar agency was created in November of 2005 with the objective of inspect and prevent non - compliance in economic and non-economic activities. This agency denominated Autoridade de Segurança Alimentar e Economica (ASAE), also ensures risk assessment and communication in the food chain. In the preventive aspect of ASAE's work, the food system and the national sampling plan takes into account the estimated food risk of the samples from the previous years [8].

ASAE in collaboration with Artificial Intelligence and Computer Science Laboratory (LIACC) proposed a project in which promotes the public health and safety of the consumer regarding food and ensures the compliance of rules between economic operators, recurring to machine learning and artificial intelligence techniques. So, this project focus in creating a method that improves the procedure of inspections through the analysis of risk matrix and prioritizing inspections. This method will enhance food and economic security and the trust of consumers in the economic agents. Also, will ensure that imported and domestically produced food are correctly handled in all steps of the food supply chain. The method will be based in risk matrix estimation [18]. In the matrix the level of risk will be categorized in two criteria: probability of happening and the impact of the risk, attributing colors according to the type of risk [50]. The matrix will be computed based on the consumer volume, service/product, default rate and other factors. The base of the system will be the data collected in the previous years.

1.2 Motivation

Every year, unsafe food leads to 600 million food-borne diseases resulting in 420000 deaths, worldwide [19]. In 2003, Portugal was afflicted with a food related crisis, when were found antibiotic nitrofurans in chicken. These compounds reported carcinogenic and genotoxic behaviour, becoming illegal to use in animals destined for human consumption. The use of this antibiotic can be implicated in the presence of human salmonella's, also might be linked to dissemination of multi resistant Salmonella Typhimurium throughout the country [23]. However, this is not the unique case in the Portuguese reality. In March of 2009, an outbreak of listeria affected 30 people, 27 in Lisbon and 3 in Vale do Tejo region. The health and food security authorities deemed a type of cheese as the responsible of the illness, reporting a fatality rate of 36.7 [48]. More recently, was detected in all Europe undeclared inconsistencies in the labelling of a type of burgers composition where horse meat was undeclared or improperly declared, constituting in some cases 100% of the meat in the burgers [51]. So a active inspection of the food chain and analysis of the food is vital to better public health and to prevent food-borne illness outbreaks.

Food related diseases entails beyond social impact, they have a huge economic impact representing high burdens in the economy. In the United States, food-borne illness registered an economic burden of 15.6 billion dollars annually [6], also food safety implies a cost of 7 billion dollars to perform all the procedures since the notification of the consumers to the payment of damages from lawsuits [39]. In 2018, ASAE performed 43,105 inspections resulting in 11,873,230 euros seized [16]. These inspections will result in an improved security in the food along the food chain, but also will help reduce the economic burden of food-borne diseases.

To achieve, the pretended result it is necessary to establish an efficient inspection process.So, every three-years, a document is developed with the aim of defining the objectives of Authority action. Also, it is established a global schedule of the inspections and the priority sectors that will be inspected. This plan is made with strict coordination with human and logistic resources available. The inspections are performed according to he risk characterization, acquired experience, recommendations and from findings from internal controls [10]. Implementing a process based on evidences is crucial to expand the efficiency and the capability of the methodology. Violations and non-compliances reports is useful information for risk assessment at the task level. The methodology will provide assurance that the safety practitioners will take into account the likelihood and the severity of registered violations or non-compliance. This information will lead to preventives actions against the mechanism that higher estimated risk [30]. Applying a similar though to the food chain and economic agents, the building of an information basis system will allow a better used of the human and logistical resources and the implementation of preventive actions that will minimize the burden of inappropriate safety practices.

1.3 Objectives

This work presents the state of art and preliminary studies aiming the building of a continuous system for analysing the variables and factors that contribute to food security risk assessment during the dissertation phase. The system will be based on the knowledge accumulated over the years by ASAE specialists with the generation of global risk matrices and on evidence provided by consumption volume, default rate, product and service characteristics as well as other factors.

1.4 Structure

The remainder of this work is structured as follows: chapter 2 presents the reviewed literature regarding the knowledge of risk and the state-of-art methods to make risk assessment.

Chapter 3 presents some preliminary results based on the algorithms presented on chapter 2, while chapter4 will focus on the conceptualization and development of the methodology used in this dissertation . Finally in the chapter 5 some conclusions are drawn by all the work developed until the moment.

Chapter 2

State of the art

2.1 Food chain and Economic agents

In Biology, food chain can be defined as the feeding relations between organisms of particular species with organisms from another. This chain is formed by producers, consumers and decomposers. Throughout this chain there are exchanges of matter and energy. It always occurs from primary producers to consumers, having nutrients recycled by the decomposers [29]. A similar process is made in the food supply chain, where the products are transferred between actors, starting in the harvest process and ending in the consumer, following the pipeline 2.1 we can see all the actors in the process [37].

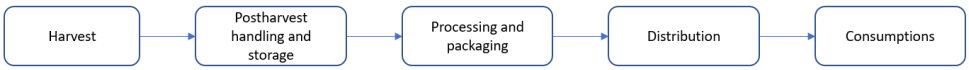


Figure 2.1: Pipeline of Food chain supply[37]

Through the food chain is registered a waste in the different steps of several billion USD. Also, if some contaminated gods enter the food chain it can reach the consumers causing public health problems and endangerment's to people. In order to avoid this problem it is necessary to intervene in the food chain in different steps considering the stage of development of the country. If it is a country in development it is necessary an action in the early steps of the food chain, however, if it is a developed country the changes will be made in final steps [37]. The actors in the food supply chain can be denominated as economic agents. These are individuals, institutions or groups of institutions that play a role in an economic circuit through their investments and decisions. They play different roles in the chain such as production, investment or consumption, establishing essential economic relationships with each other [5].

All the actors in food supply chain can commit two types of infractions: food fraud and food safety. Although food safety refers to food related issues in the food along the food chain, if the food isn't contaminated or if it is transported at the appropriate temperature, food fraud analyzes economic infractions that can be profitable.Usually, it happens when the potential profit is high

with a low risk to be detected. Food fraud encompasses falsification of foodstuffs, fraud on goods, counterfeiting and others, mislabel foods to increase the profit from a determined products. In Portugal, the food chain inspections are made recurring to the Plano Operacional Práticas Fraudulentas na Área Alimentar (POPFAA) will be able to verify the existence of fraudulent practices in foodstuffs throughout the commercial circuit allowing to ensure free practice and fair competition between operator, safeguarding the consumer interests and protection[12].

In the harvest step, 84% of total of the global food is from corps and 1 billion poorest people depend of the livestock. So, if a contamination occurs in this step a public health problem will arise contaminating several people. Therefore, an inspection is important to prevent harm to the people and protecting the consumer interest, safeguarding. Also, during the processing and packing will an inspection will ensure that the best conditions are being met. As well as, will ensure that the products are corrected labelled allowing the consumers to buy the products they want and not creating a suppliers monopoly, establishing market rules. This will allow a free competitions between the suppliers resulting in a more choices to the consumer [2].

2.2 Autoridade de Segurança Alimentar e Económica

In November of 2005, Autoridade de Segurança Alimentar e Económica (ASAE) was created with the mission of supervising and prevent compliance with the regulatory legislation, governing the conduct of economic activities in the food and non-food sectors, as well as to assess and communicate risks in the food chain. It is the entity responsible to communicate with its counterparts at an international level. This institution is governed by scientific independence, precaution, credibility, transparency and confidentiality.

ASAE aims to maintain as the reference entity in consumer protection, public health, safeguarding market rules and free competition, through providing a public service of excellence. To do it , it cultivates values such as integrity (honesty and ethic), quality (accuracy and efficiency), commitment (responsibility and engagement), independence (impartiality and transparency) and credibility (reliability and trust) (Fig. 2.2) [8].

To ensure that all territory is monitored, there are three decentralized units called Regional Units: North, Center and South. Each regional units has several operational units leading to a total of 12 in country. The formal structure of the organization can be seen in the fig. 2.3 and fig. 2.3 [9].

ASAE is ruled by a code of conduct and ethics to allow a better performance and a better contribute to society. The code of conduct can be established in 4 points: common rules, inspections area, scientific and laboratory area and procedure decisions area. All employees must act with the public interest in mind, acting with ethics and assuring justice. Also, they must not discriminate against any citizen based on sex, race or others, must have cooperate with the citizens and other organizations, provide legal and accurate information when questioned or redirect to organizations that can clarify the doubts and they can not require more information of the citizens than the essential minimum to perform the activity. However, depending the activity some rules can be more specific, for example for inspections area the inspectors can only use equipment, vehicles

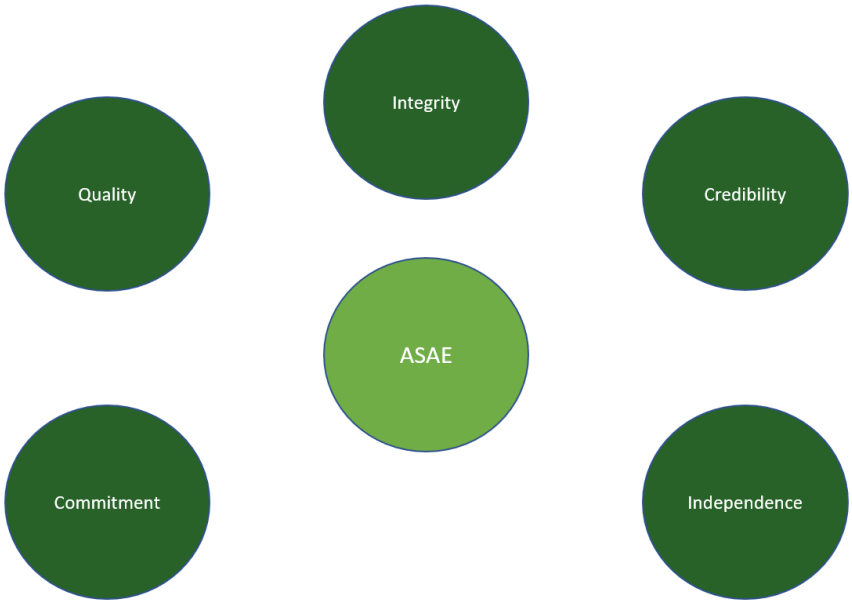


Figure 2.2: Values of ASAE adapted from [8].

and installations to professional use, in the scientific and laboratory area that can not communicate the risk in a public manner and in decisions area the collaborators must be committed to improve academically and personally in order to better do their work [4].

Annually, ASAE define a plan of activities where it is described the goals to be achieve during that year. For example in 2019, to ensure regional inspections was used the formula:

$$\frac{Number\,of\,executed\,regional\,operations}{Number\,of\,planned\,regional\,operations} * 100, \tag{2.1}$$

with the perspective of reaching 85% [10]. This plan is evaluated using the "Quadro de Avaliação e Responsabilização" (QUAR), which are an evaluation matrix of performance in services, goals, performance indicators given the available resources allowing the identification of the deviations and the and their causes at the end of the management cycle[15]. Also, ASAE has created “Plano Nacional de Fiscalização Alimentar” in order to establish a frequency of regular and risk-proportional control. This inspection plan must be made not only recurring to the general inspectors expertise but also using the operational results of previous years and operators’ past non-compliance’s history, the recommendations of the commission, commitments, and protocols, the previous critical or serious violations in each of the chain phase and sector and number of economic operators in each geographical area, and much more relevant variables [11]. More over this inspection plan must be clear and explainable to others organizations

As part of the mission, ASAE has been developing an institutional policy of social and environmental concern in the interventions with the interested parts. The policy acts in 3 axes: preservation of human resources, social responsibility and environmental variable. In the first one, the major concern is proper use of public funds. Social responsibility regards the appreciation of

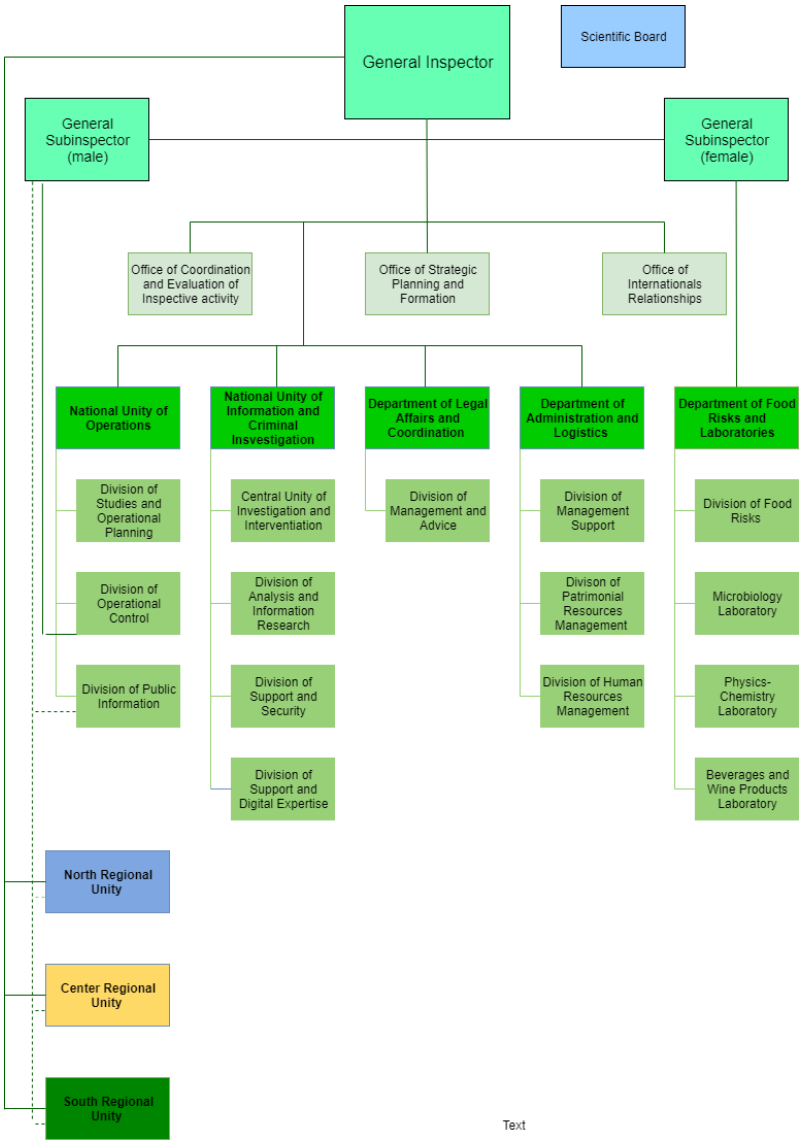


Figure 2.3: Chart of ASAE adapted from [9]

the people in the organization and the support of people, acting in the society through donations, in this factor is registered the most effort. Finally, reusing seized material, donations, and managing properly other waste confers a environmental aspect to this policy[17].

2.3 Definitions of risk and assessment processes

Risk can be defined as a probability of a negative outcome occur caused by external or internal vulnerabilities that can be avoid trough preemptive action [34]. When applied this definition regarding the food industry, can be described as a hazard present in food products that cause harm of a certain magnitude [22].

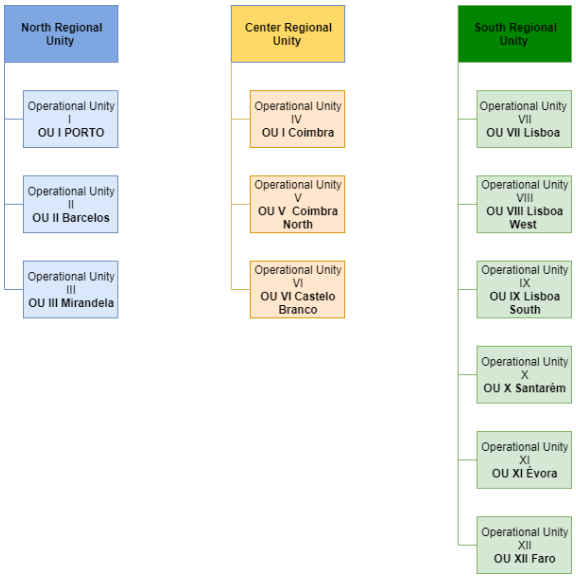


Figure 2.4: Chart of Operation Units From ASAE [9]

The risk assessment process contemplates three phase: the risk identification, the risk analysis and the risk evaluation. The risk analysis phase can be divided in determination of the consequence, the probability occurrence and the intrinsic risk level [40].

In the late 1950, in the United States in order to develop secure food to the mission Appolo, part of a spacial program, it was implemented a Hazard Analysis and Critical Control Points (HACCP) system by Pilsbury Company with Nasa.

The HACCP system is a preventive system that ensures the quality of the food products. This system can be used in all phases of the food chain, and identifies specific dangers that affect the consumption. This type of system allows the implementation of preventive measures to avoid contamination of the product and to identify the Critical Control Points (CCP) that need to be maintained under surveillance, and thus preventing food related accidents and improving the public health [20]. To implement this type of system some principles should be followed. First, it is necessary to identify dangers and preventive measures, then identify the critical control points, establish critical limits for each measure associated with CCP, control each CCP, establish corrective measures for each case of deviating limit, establish verification procedures and finally, create a registration system for all checks carried out [7]. In the 80s, was registered an increase in cases of food contamination's leading to the implementation of HACCP based control systems in 1990 in the United Kingdom (UK).

The systems set, along with a series of measures taken by the government, such as the training of professional health officers and buying equipment's to create the HACCP system, allowed a better defence in food safety problems. Finally, in 1993 the HACCP system were required trough the implementation of European Union (EU) directives in non-animal and animal products. In non-animal products a hazard analysis must be made, but there is no need to apply all HACCP principles or documentation, contrary to animal products. The new regulation originated a series

of Codes of Practice for enforcers. More specifically, the code number 9 [3] refers to food inspections stating the two primary goals and contains a rating scheme to prioritizing and to calculate the frequency of inspection, based on potential hazards, level of compliance, confidence in control systems. Also, was provided guidance regarding the assessment of adequacy of HACCP systems. In order to do a competent assessment of risk, the enforcers must consider history of problems/complaints, the severity and imminence of hazard and the critical customers group. Through a enforcement of this type of systems, an increase focus in early stages of discussion with the owner on the procedures and hygiene system have been registered [27].

In 2006, the European Parliament obligates all operators of the food sector to create and applied one or more permanent processes based on the principles of the HACCP. HACCP systems can serve as tool to assess risk. The risk assessment can be divided in multiple task as previously explained.

They are highly recommended to perform risk identification, determine the consequences and to perform risk evaluation. However, these systems can not be applied in the risk analysis in order to estimate the probability of a hazard occurrence and the level of risk. The system is capable of placing controls in all relevant parts of a process to protect against the hazards previously identified, providing quality, reliability and safety to a product. The best feature of the system is the capability to minimize risks recurring to controls throughout the process rather then inspecting the end product. The systems can be described by the following pipeline 2.5 [40]:

The system requires 3 factors as inputs, the risk associated with the hazards, the ways that hazards can be controlled and information about the hazard. Then the HACCP applies the principles mentioned in the beginning of the section obtaining a hazard analysis worksheet and the HACCP plan.

More recently, Risk Based Control (RBI) systems have been implemented. The system can be described in two phases, the risk categorization phase where is prioritized hazard food combinations and the risk based surveillance where is defined the frequency of inspections.

The risk categorization phase is an important phase for governments to understand where should spent their resources. So, several studies have been made in order to prioritize risk control through the evaluation of costs, illness and disease burden. The methods to prioritize risk can take into account opinions of experts in the area, recur to risk matrices or to computerized models based on risk assessment.

In Australia, a risk assessment tool to prioritize microbial food safety hazards has been developed, this tool uses the probability of specific hazards to occur and their effects. This tool can be uses to asses the relative risk using different combinations of data such as products, pathogens and processing steps, in a rapid and efficient way. A similar tool has been developed in the Netherlands, called swift QMRA [35]. Allowing more possibilities to change the input values in order to prioritize risk through comparison [57].

In 2014, due to the many occupational accidents in the United States, Azadeh in [25] proposes a RBI system to reduce work-related injuries, illness and death in the work place recurring to data from different stakeholders. The system assessed the workers accident severity grade using tree

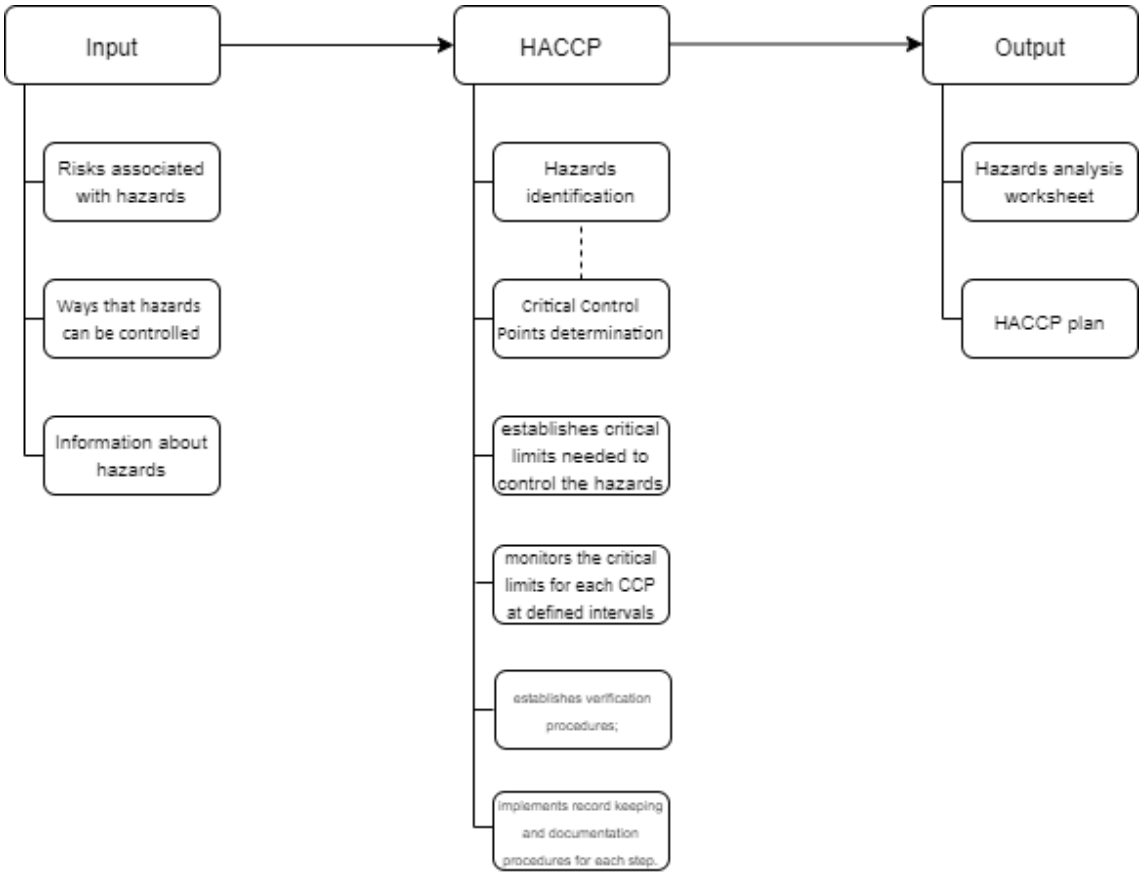


Figure 2.5: HACCP pipeline [40]

dimensional risk assessment matrix. In order to increase the information in the process the system used a risk score. This requires detailed data from workers and worker place in order to be the most efficient as possible. This process allow a real-time monitoring of severity, facilitating the implementation of hazard controls to reduce the most severe accidents. The system utilized data from different industries to improve the risk estimation.

Also, as stated in [30] initial analysis of reported accidents while performing tasks has proven to be useful in safety practice. It was applied a semi-quantitative methodology, where it was estimated the likelihood of occurring an accident and the resulting severity by taking into account the actual distribution of accident mechanisms in each of the tasks. The semi-quantitative approach consists in using likelihood of occurrence of accident and severity, recurring to data from the accident mechanisms from that time. The procedure will be explained with more detailed in the next paragraphs.

In risk prediction different approaches have been made regarding machine learning and deep learning techniques. For example, with the objective of assessing the risk in marine oil pipes, a RBI system was proposed in [41]. It consisted in the determination of the consequence of failure and its probability of occurrence and the estimation of the risk ranking leading to an inspection plan. The methodology also uses risk matrices. The knowledge acquisition, which is the base of the entire system, was made recurring to machine learning techniques (classification techniques,

association rules) and a more human analysis process trough interviews with observation and protocol analysis. To obtain explicit knowledge from practical experiences a conceptual model with problem-solving methods is suggested in the fig.2.6. The models consisted in analysing a particular context event, developing a solution and inquiring about the lesson learned from this particular experience leading to explicit knowledge. The data and the collected information were

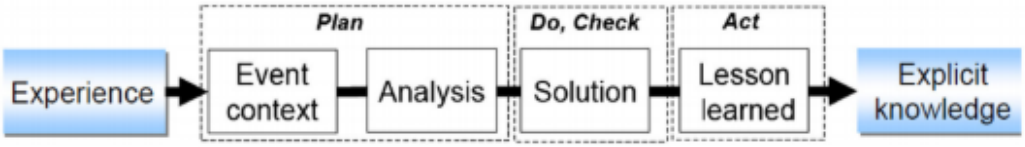


Figure 2.6: Conceptual model to obtain explicit knowledge [41]

the input of the risk based inspections. Then it was calculated the risk assessment recurring to the probability of failure and the consequence of failure. After that, it was computed the risk ranking in risk matrix, leading to inspection plan. Finally, occurs a mitigation phase (in case of need) and re-assessment. This process is explained in the fig. 2.7 This process allows the determination of

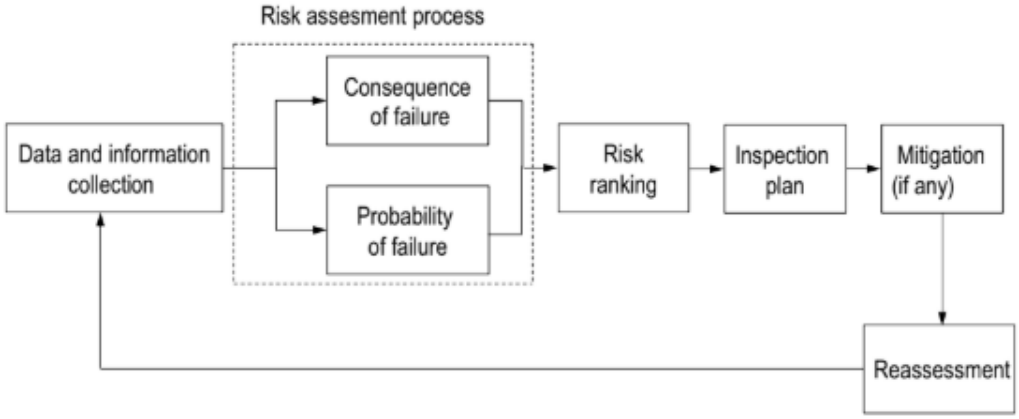


Figure 2.7: Risk assessment process from [41]

the optimal interval of time necessary to perform the inspections [41].

Risk matrices can be use as another method to perform risk ranking, through the combination of the probability of occurrence of risky events (on the x-axis) and the severity of the hazard (on the y-axis) (fig. 2.8) [57]. To create risk assessment monograms the probability is divided in likelihood of occurrence and manufacturing control, then is multiplied for the severity in order to reach a qualitative indication of risk, allowing to score the items from high to low. The different hazards are scored on a scale from high to low [57]. In the UK, UK Food Standards Agency developed the most comprehensive monogram where incorporates consumers information, including consumption pattern and affected populations.

As defined in ISO 31010, risk matrices are ideally used to identified the risk, are important to determine the consequence, the probability and the level of the risk. The matrices can be applied

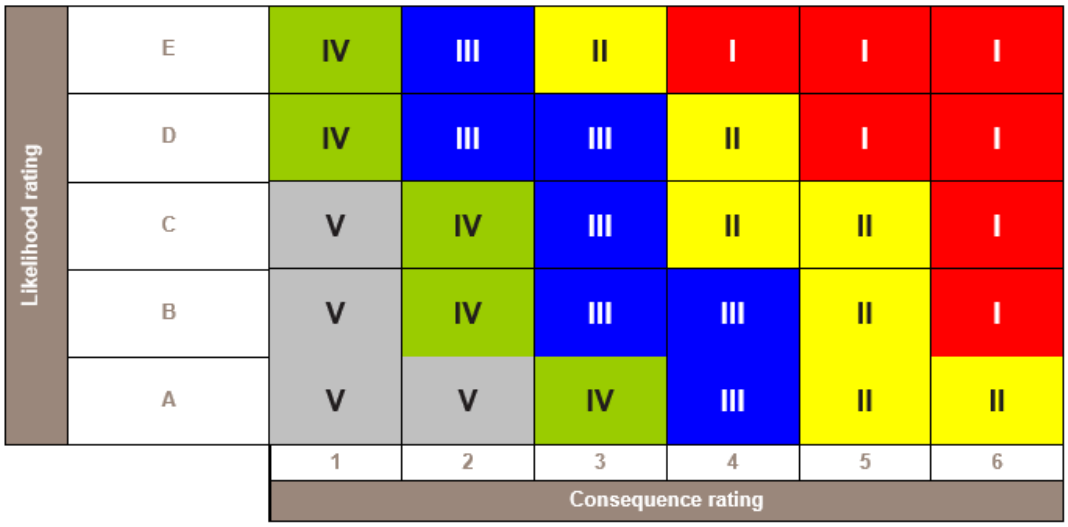


Figure 2.8: Example of a Risk Matrix[40]

in risk evaluation although not in large extent. The matrices creation and process can be described in the pipeline 2.9:

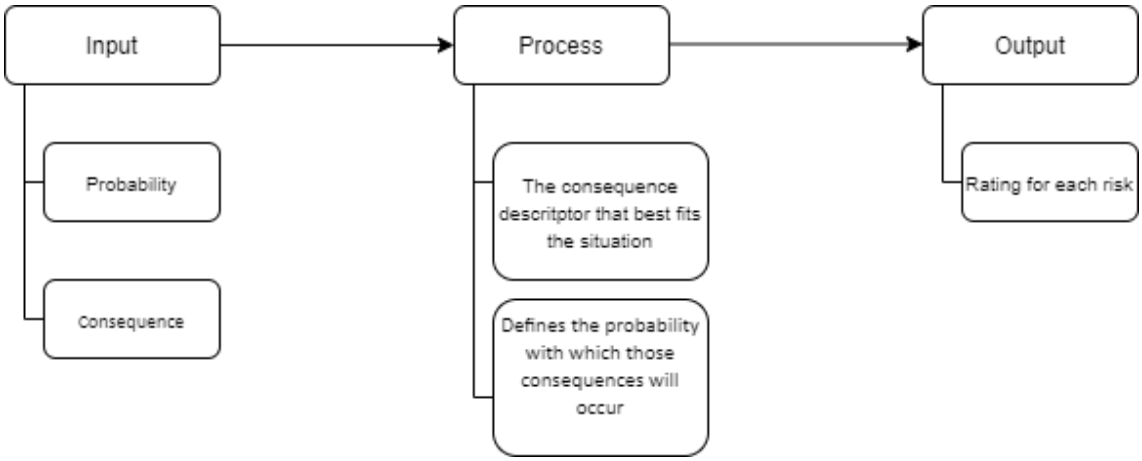


Figure 2.9: Matrices pipeline by ISO31010 [40]

The matrix will be constructed from the probability of the hazard to occur and the its consequence. Then a description of the consequence is made and is calculated the probability of occurrence and the consequence giving the rating of each risk.

Many risk events may have a range of outcomes with different associated probability, therefore the user must choose the ranking that is more suitable to the situation.

The level of risk defined by the matrix may be associated with a decision rule such as to treat or not to treat the risk. This process is easy to use and provides a rapid ranking of risks into different significance levels. Since, a matrix is designed to be appropriate for a certain circumstances it may be difficult to have a common system applied across a range of circumstances relevant to other organization. It's use is very subjective and it tends to have significant variation between raters,

it is difficult to combine or compare the level of risk for different categories of consequences and risks cannot be aggregated [40].

Once the most important hazard and products are identified, the RBI systems as to be set up. The RBI in order to elaborate the frequency of inspection uses the Risk Categorization (RC) results and the company characteristics [57]. The companies can be classified based on historical data, the effectiveness of Food Safety Management System present and its socio-economic behaviour. In order to estimate risk of non-compliance, some information must be provided such as past records of business operator, own checks of the national food safety authority, basically "any information that might indicate non-compliance", reaching a compliance profile (Figure 2.10). This profiles indicates the compliance of each target group, if they are prone to violation or to compliance.

Historical data can be used as an indicator of good practices, however this is not guaranteed since errors might still occur [57]. Aside the available food safety data, the quality of the FSMS is important to ensure a better perform of the RBI system. So, the effectiveness of the FSMS, must be incorporated into the RBI system. Other systems and activities such as HACCP can be judged using scores to key criteria and can be assessed within an FSMS [57].

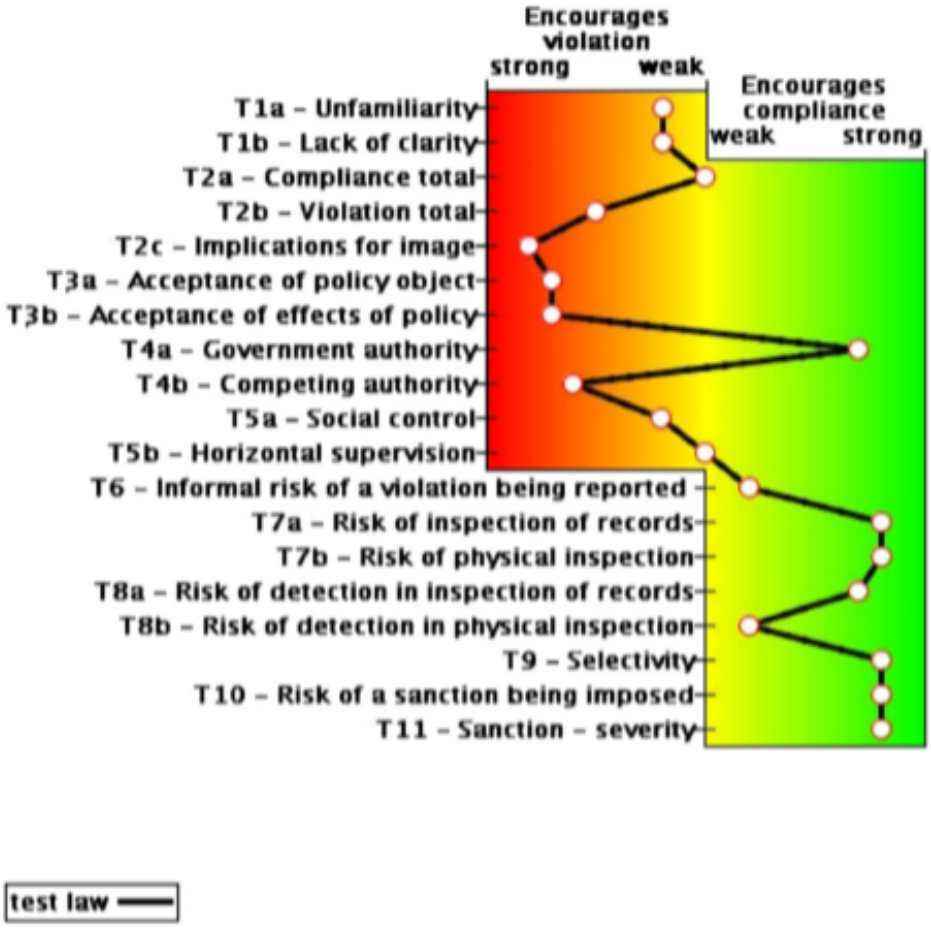


Figure 2.10: Example of compliance profile[57]

So, in order to compute the RBI systems previously described, it's necessary the calculation

of the risks criteria that the systems will have to obey. The ASAE takes into to account the planning of the unit, the seasonality referent to economic activities and regional specificity as relevant criteria. This criteria can be described in a more detailed way in the following criteria:
Consume volume(V) - this criteria is the food consume volume and to be implemented must resort to data from official references
Performance(D) - refers to the total of process of the previous year and by activity sector. Measures the level of risk for public health or economic safety for the consumer, with precautionary measures associated.
Product risk(PS) - it is computed from the frequency that reports and complaints are received, from the estimate risk, from the origin of the product and trough communication with other entities both national and international.

Regarding the number of operators that need to be inspected defined in QUAR, a priority order for risk-based operator selection is established, based on prior inspections and on the irregularities presented as well as its degree of severity. The operators can act in the alimentary area and economic area, in different phases of food chain and the economic chain described in Figure 2.11.

Figure 2.11: Operators' areas of inspections adapted from[15]

2.4 Risk Based Algorithms

In the literature, different approaches to risk assessment were made in different areas. These approaches vary according to the analyzed system and the type of risk that must be predicted. So, for this work, only the most relevant part of the literature was inquired.

2.4.1 Machine Learning

Regarding machine learning algorithms, different techniques were evaluated in different situations.

Equipment failure causes unexpected and undesirable events in great proportions in the oil and gas industry. So, it is necessary to perform inspection in all the equipment, however this process is not cost-effective. To resolve this problems Rachman in [53] describes different machine learning approaches to perform RBI screening assessment. The system will be used prevent equipment breakdowns recurring to past risk detailed assessments and knowledge transfer. The RBI system recurs to the risk of failure of the equipment to prioritize the inspection. The analyzed methodology possesses two types of assessment: screening and detailed assessment. The screening phase is prone to human error, so a machine learning methodology is crucial to improve the system. The proposed workflow can be decomposed in the following steps: Feature Selection, Data pre-processing and Algorithm Selection.

Feature Selection is an important step in the workflow. Initially, in a preliminary selection all the features with missing values and that were duplicated were removed. Following, a more knowledgeable selection was applied resulting in an exclusion of features that caused information leakages and features that require extensive data gathering and assessment. Finally, a filter feature selection technique was used to remove multicollinear and zero and near zero variance features, as represented in the fig.2.12.

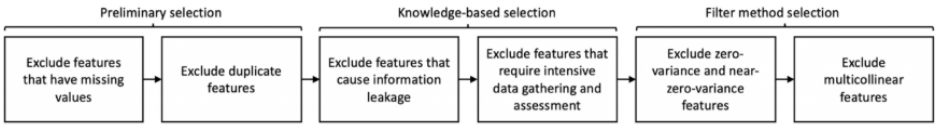


Figure 2.12: Feature selection pipeline [53]

Data cleaning: includes imputing empty values, correcting data format, removing unnecessary features, resolving inconsistencies, and identifying and removing outliers. The Data preprocessing can be described in three steps: Data integration, Data reduction, Data transformation. The steps consisted of merging multiple databases, reducing the volume of data without sacrificing the quality of data and, at last, smoothing and aggregation of data. In the final step of the proposed system, different machine learning algorithms were evaluated to compare one to another. The algorithms used can be divided into normal classifiers and ensemble classifier. The normal classifiers consisted in Logistic Regression (LR), Support Vector Machines (SVM) and K-nearest-neighbours (k-nn). The ensemble classifiers used were Gradient Boosting Decision Trees algorithm (GBDT) and Random Forest (RF).

Logistic regression is classifier that utilizes a mathematical approach that relate a set of independent variables of interest, X_s to a dependent variable. The model is described by the function represented in the equation2.2.

$$P(x) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \tag{2.2}$$

where β_i and α represents constant terms to unknown parameters [43].

Support Vector Machines are easy and approachable method, that can be applied in the different classifications problems on the real world. This algorithms, aiming pattern recognition, makes

use of a training data with N-dimensional patterns x_i and class labels y_i to construct a function of type:

$$f : R^N \rightarrow \{\pm 1\} \tag{2.3}$$

This will allow f to correctly classify new examples (x,y) from the test set, so that $f(x) = y$. This solution is generated from the same underlying probability distribution $P(x,y)$ as the training set [38].

K-nearest-neighbours is a classifying technique that separates the entities into classes or groups. The algorithm uses distances functions to calculate the similarity between patterns. The classifier is based on the Euclidean distance between the training samples and the test samples. The Euclidean distance can be computed as :

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + ... + (x_{ip} - x_{lp})^2} \tag{2.4}$$

where x_i is a input sample with p features, number of features, n represents the total number of inputs.

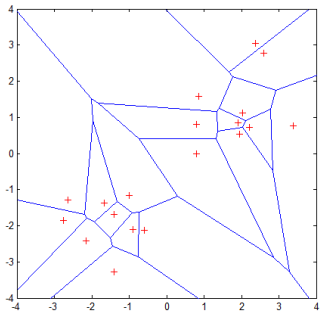


Figure 2.13: Voronoi tessellation showing Voronoi cells of 19 samples marked with a "+"[52]

Analysing the fig. 2.13, that represents a graphic depiction of the nearest neighbour. Where the Voronoi cell,R, surrounding the positive sample. R_i is referent to the Voronoi cell from sample x_i, and x represents all possible points within. Based on the Voronoi characteristic that the nearest sample is determine by the closest Voronoi cell edge, k-nearest-neighbours labels the test samples with the class of its k nearest training samples. K is normally odd to avoid ties [52].

Gradient Boosting Decision Trees algorithm is a type of algorithm that consists in an ensemble of decision trees. The space of regression tree in this algorithm can be computed as:

$$F = f(x) = wq(x)(q : Rm \rightarrow T, w\varepsilon RT) \tag{2.5}$$

T is the number of leaves in the tree. The q represents the structure of each tree that maps an example to the corresponding leaf index. The independent tree structure and leaf weights w are represented by f_k [31]. The model will give the final prediction based on the sum of the score in the corresponding leaves(w) from the sum of predictions from each tree. The tree models possess a relationship between the response and the predictors can be modelled by locally constant fits.

The constant will be fitted in the divided input space, region R_1 to R_N being N the number of regions. The model can be defined as:

$$f(x) = \sum_{m=1}^{M} \theta_m I(x \varepsilon R_n) \tag{2.6}$$

where R_N defines the constant fit in the region.

Finally, the algorithm processes the boosting. Boosting is a combined technique of weak learners and that is capable of improved accuracy. The model outcomes are based on the outcomes of previous instants. The outcomes that are wrongly classified are weighted higher when compared to the outcomes corrected classify. The boosting fits models and can be expressed in the form:

$$f(x) = \theta_0 + \sum_{m=1}^{M} \theta_m \phi_m(x) \tag{2.7}$$

The model class of the basis function must be specified. The algorithm finds the optimal linear combination of N basis functions from this class. The basis functions are learned using a base learner and then are added in a sequential manner, during the fitting[31].

The Random Forest classifier can be described has a combination of tree classifiers, where each classifier is generated using a random vector with samples independents from the input vector. Also, each tree gives a unit vote to the most popular class to classify an input vector. The random forest to a grow a tree can use randomly selected features or a combination of features at each node. These algorithms will reduce the risk of overfitting to the training set. To evaluate the performance of the classifiers was applied nested cross-validation and performing evaluation metrics, following the pipeline 2.14.

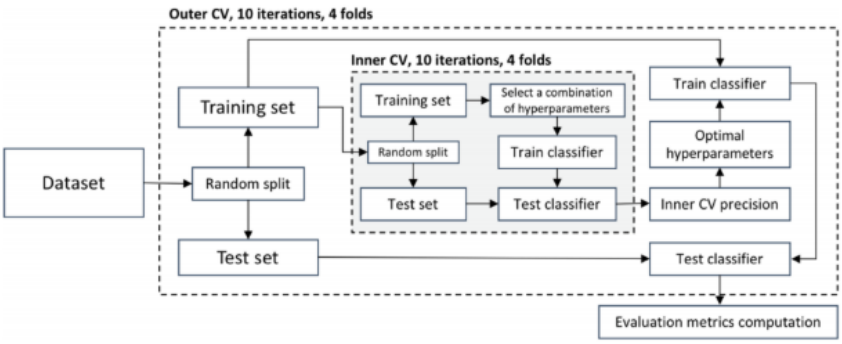


Figure 2.14: Model selection and evaluation using nested CV[53]

From the metrics evaluation was possible to conclude that the worst performance for simple classifiers where LR, SVM and the k-nn had precision of 75.34%, 79.64% and 82.36%, respectively. The k-nn algorithm outperformed the others simple classifiers.

Concerning ensemble classifiers Random forests and Gradient Boosting Decision Trees were the most efficient, although the preferred algorithm was GBDT because it outperforms the RF

algorithm on ever evaluation metric except in Precision presenting the value of 82.96% inferior to 85.21% presented in the RF. The ensemble methods outperform single methods. Trough this analysis is possible to conclude that all the algorithms can be highly applicable to calculate the probability to a hazard occur.

Wang [58] describes a pre-warning system to help food manufacturers find food risk in advance and to offer information to maintain food quality and safety. The system proposed can be separated on different stages and uses data from the food chain supply. After analysing the data, a pre-processing was applied to avoid abnormalities in the data. Following this step, a rules association mining technique was implemented to find a relationship between the items within the dataset. The relationship can be represented by:

$$X \Rightarrow Y \tag{2.8}$$

where the X represents a itemset and Y the suggested food safety assurance setting. The rule is valid if the pairs attribute-value are true for the particular case. Consider that T is a set of transactions of a given database, the percentages of of cases appear in the dataset recurring to the formula 2.9. This formulation is denominated the support rule.

$$supp(x) = \frac{|\{t \in T; X \subseteq t\}|}{|T|} \tag{2.9}$$

The generated rules can be measure using two different methods. To calculated the number of times that the rule was truth in dataset the confidence can be computed using the equation2.10

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X) \tag{2.10}$$

The association rules were evaluated taking into account the following equation:

$$RI(X \Rightarrow Y) = \frac{P(XY)}{P(X)P(Y)} \tag{2.11}$$

Encouraging results were obtained, with Transit time $= 05:00 +$ Season $=$ Winter $\Rightarrow$ Product type $=$ Yogurt obtaining the higher value of RI, 4.65 being the most suitable association rule. However it takes to long to extract the association rules, furthermore, in association rule analysis small factors have sharp boundaries that are different from the elements near the boundary, so it is suggested the application of fuzzy sets. These types of approaches are being implemented in health research. The association are highly applicable to proceed to risk identification and to determine the risk consequence

Bhatla in [28] makes use of the fuzzy logic and machine learning algorithms to reduce the number of attributes used in the determination of the risk to possess an heart disease, resulting in a reduction of the tests taken by the patients. The suggested system must efficiently diagnose the presence of heart disease in an individual. This was made recurring to fuzzy logic. A fuzzy set is a collection of elements with a varying degree of compatibility to features or properties that are distinctive to the collection, these characteristics are value in interval [0,1]. Assuming the fuzzy

rules established, the diagnose is classified in Normal, Low Risk, Medium Risk and High Risk. To achieve this classification different machine learning methods were used.

A decision tree algorithm is a classifier that uses a series of questions about the features associated with the items. The node in the tree represents a question and the internal nodes appoint to a child node for each possible answer. In a simple way, a decision tree can be represented by the fig. 2.15, were the questions are yes or no leading to respective child node [42].

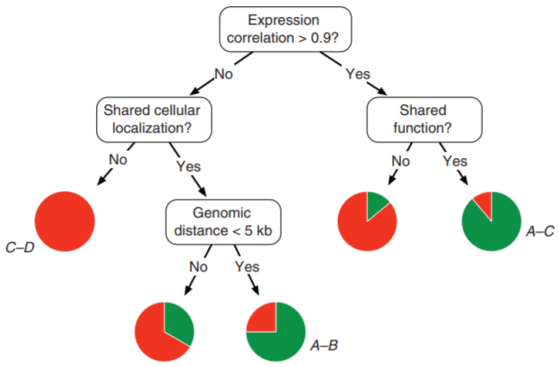


Figure 2.15: Example of hypothetical decision tree [42]

The class is assigned to the item by following the path from the topmost node, the root, to a node without children, a leaf, considering the answers in the previous nodes and the leaf that is reached. Each leaf contains the probability distribution over the classes, and estimates the conditional probability that an item reaching the leaf belongs to a given class. The answer of the trees can be computed efficiently, however the questions can be complicated [36].

The Naïve Bayes Classifier uses the Bayesian theorem and is appropriated when the input dimensionality is high.

The classification via Clustering uses two machine learning techniques (classification and Clustering) where the clustering forms groups and structures in the data, and then the general structure is applied to new data to obtain a label. This method can be explained in the fig.2.16.

The proposed method, fuzzy logic combined with machine learning techniques achieved results of 100% to predict the risk of heart diseases, simply using four attributes. To compare the efficiency between the number attributes, the techniques were tested with a different number of attributes, although the results were quite similar about 99.62% using 15 attributes with decision trees, the method proposed surpasses the previous ones. The decision trees classifier presented an error of 0 while the Naive Bayes classifier had a 0.72% error. So, all the algorithms can be highly applicable to identify the risk level.

To classify Credit Risk, Rahayu in [54] recurs to kernel logistic regression, that is a non-linear form of Logistic Regression. This technique constructs the model using high-dimensional features. This method revealed to be must more efficient in determining the percentage of correct predictions when compared with SVM, reaching accuracy of 78%, when the SVM algorithm only

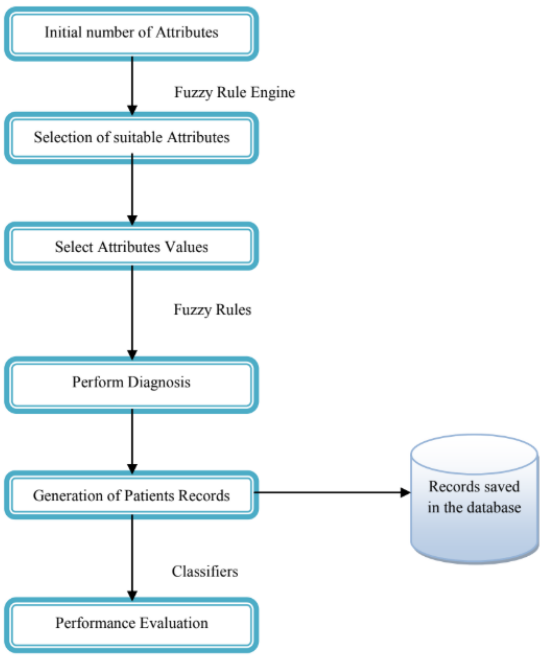


Figure 2.16: Pipeline of the system[28]

reached 67% recurring to the accuracy formula:

$$AC = \frac{a+d}{a+b+c+d} \tag{2.12}$$

where, a represents the true negatives, b the false positives, c the false negatives and finally d the true positives. Also, was compute the accuracy of the algorithms to determine the proportion of positive cases using the equation2.13.

$$ACII = \frac{d}{c+d} \tag{2.13}$$

However, the SVM has a better efficiency into determining the positives of the system (Table 2.1).

Indicators	KLR	SVM
AC	78%	67%
ACII	92%	96%

Table 2.1: Comparison of credit risk classification performance between KLR and SVM[54]

The kernel logistic regression will be highly applicable to Risk identification.

Li in [46] proposes a system to determine a generic project risk, using data mining techniques recurring to data transmission.A generic project risk element transmission theory is based on the fact that if a sub risk factor is reasonably divided, must exist a quantities mathematical relation between the overall risk element and the sub risk element.Based on this knowledge the data mining

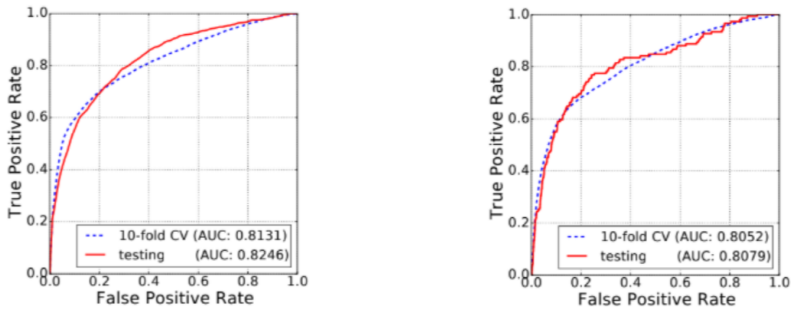


Figure 2.17: Comparison of ROC curve between Random Forest(Left) and SVM(Right)

method was used to acquire the risk transmission matrix from historical datasets analysis in order to solve the quantitative calculation. Also, a data mining frame was constructed based on the knowledge, the data mining method was used to acquire the risk transmission matrix from the historical databases analysis in order to solve the quantitative calculation. The controlled risk element was also given with the controlling risk degree. To obtain the transmission matrix it will necessary to divide the risk element information by work-breakdown structure or other techniques. After that, it is necessary to treat the information, through filtering. Through the analysis of the data the transmission matrix is built, dividing the risk states and calculating the transmission probability between each state. Once the risk element transmission matrix is acquired, it is possible to chose the risk element that control the project. The method is highly applicable to identify risk.

Madaio in [47] explains a system used in Atlanta to predict fire risk and to prioritize fire risk inspections. This system used several data from different sources. This resulted in a total of 19397 new commercial institutions to inspect. After the merging of the data, a pre-processing step was done. In this step, a binary feature was added, allowing to identify the properties that had missing data. In this properties the missing values were replace with 0. The merging of datasets resulted in 252 variables. Some variables were manually removed and then forward and backward feature selection processes were applied to see the contribution of the variables to the model and then removed those who did not contribute to higher accuracy. Then SVM and Random Forest were applied and compared.The True Positive Rate represents the number of fire that were predicted, allowing to save lives.Comparing the two algorithms the SVM had higher True Positive Rate, 71.36%, than Random Forest that only had 69.28%. The Random Forest perform better when each tree had a depth of 10, was used 200 trees. However, Random Forest has a higher under the curve area accomplishing a better distinction between classes (Fig. 2.17). The algorithms can be applicable to proceed to risk identification.

To evaluate the efficiency of risk-based inspections, Moura in [33] proposes a Multi-Objective Genetic Algorithm or MOGA. The system proposed is a combination of previous RBI systems information with MOGA algorithm 2.18. The RBI system was previous explained in the section 2.3, so in this section will only be explored the algorithm and the results. As stated in Konak [44] this type of algorithm provides only feasible inspections programs as outcomes. The MOGA explores a reduced search space, preventing from getting stuck into an unfeasible part of the search

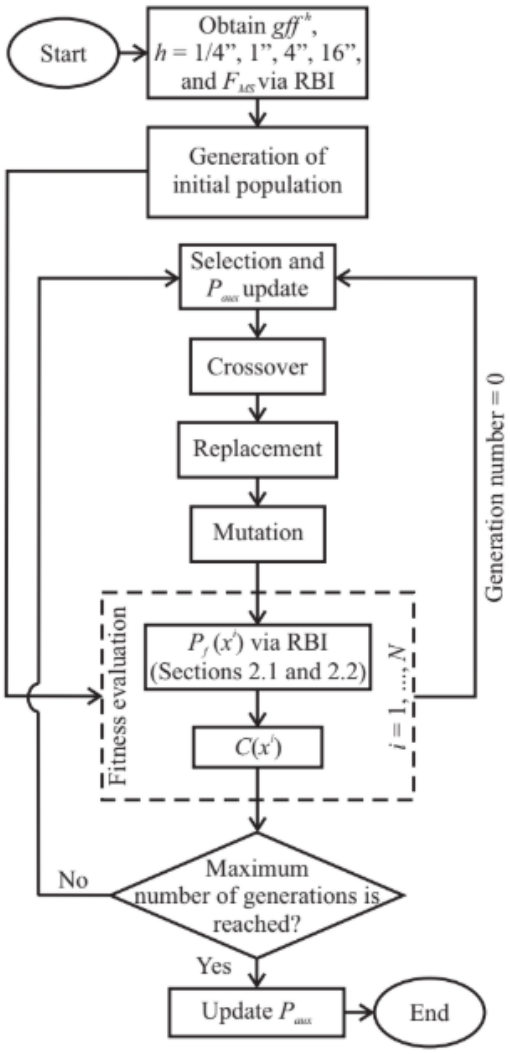


Figure 2.18: Pipeline of the RBI system in combination with MOGA [33]

space. If X is a solution space, so a chromosome is a binary vector solution with $x \varepsilon X$. The algorithm revolves around the mapping between the solution space and the chromosome. A group of solutions vector is a denominated population. The population is randomly initialized and new solutions are generated from previous ones recurring to cross-over and mutations. The cross-over method combines two chromosomes, with the best genes, more fitted to form new chromosomes, offspring. The mutation will insert random characteristics into the chromosomes, altering values in vector solution. This will assist in search of local optimums. The reproduction is made trough the selection of chromosomes for the next generation, the fitness of the individual determines the probability of its survival in the next generation. The fitness of an individual, the solution is determined recurring to fitness functions. The fitness functions can be expressed in a weighted sum approaches, where is assigned weights to each normalized objective functions. [44]. The weight can be embedded within the chromosome of solution x_i or utilizing *priori* approach, where the user provides the weights. Another approach is the Pareto-Ranking, where the population

Parameter	Value
Population size	250
Number of generations	500
Probability of crossover	0.95
Number of cut points	0.8
Probability of mutation	0.05

Table 2.2: Parameters used in MOGA[33]

is ranked according to a dominance rule, and then each solution is assigned fitness value on its rank in the population. SPEA is a method that assigns better fitness values to the non-dominated solution at underrepresented regions of X. This whole process will converge into an overall good solution. Using a population size of 250, with 500 generations, with 95% probability of crossover, 8 cut points and a probability of mutation 5%, the MOGA was capable of encounter 98% of the solutions from the Pareto space, optimal space as seen in the table 2.2 and in the table 2.3.

This algorithm combined with a RBI system allows providing information on how the inspection budget should be more efficiently spent. This algorithm will be highly applicable to determine the consequence, the probability and the risk evaluation in the risk assessment process.

2.4.2 Deep Learning

Mao, in 2018 describes a system where utilizes blockchain, which ensures food safety combined with a deep learning network named Long Short Term Memory (LSTM) [49]. The system was made to evaluate the credit of the different stakeholders in the food supply chain, leading to better effectiveness of supervision and management. The process uses Hyperledger blockchain to satisfy the requirements for permissions needed to the different roles and to authenticate in the food supply chain. This makes that the traders can be held accountable for the evaluation credit process while remaining anonymous. First, information about the transition and the credit evaluation was collected. Then it was implemented the process to evaluate the credit for regulation. There are a merge system responsible for combining the blockchain technology and the deep learning model. The model analysis the credit evaluation text and generate at least one credit evaluation result. The blockchain provides credit evaluation information and trades transaction to act as an input to the model. The model is composed of multiple layers and the layers are composed of several cells that are explicitly designed to store information for a certain period.

Statitics	Number of solutions	Number of exact solutions
Minimum	43	42
Median	46	44
Maximum	46	45
Mean	45.57	42.28
Std.dev	0.5730	0.5333

Table 2.3: Results from MOGA [33]

In the most simple LSTM, the cells consist of three gates (input, forget, output).Each value of the cell is saved by the three gates that permit the modification of the value of the cell. The final output of LSTM is given by:

$$h_j^i = o_j^i \odot tanh(c_j^i) \tag{2.14}$$

The LSTM can be presented in different layers as in fig. 2.19. The output of credit evaluation

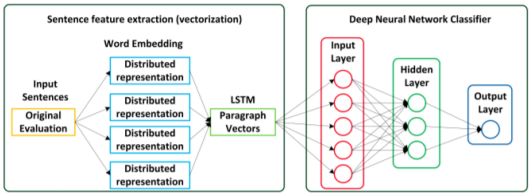


Figure 2.19: Pipeline of LSTM [49]

is given in two ratings ("positive" and "negative") according to the inputs received. The LSTM gives more importance to content-based features rather than local information[49]. The epoch value obtained after 5 tests was the best value of epoch. Training the model is a crucial stage, in order to find the parameters, weights in neural network, that minimize the loss functions in this case the binary cross entropy mathematical represented by:

$$f(G,O) = -G * log(O) + (1-G) * log(1-O)) \tag{2.15}$$

Where G is a target and O represents a networ output.

This system proved to be more accurate than the SVM and naive Bayes model as indicated by the accuracy and F1-score described in the fig.2.20. So, the LSTM can be highly applicable to

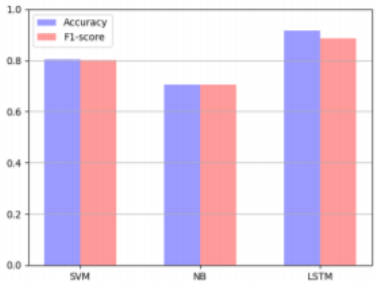


Figure 2.20: Comparison between LSTM and different methods to credit evaluation [49]

identify risk in the risk assessment process.

2.5 Conclusions

Different techniques can be applied in order to determinate different procedures. In the table.A presents in the appendixA, is demonstrated the utility of each technique. Each technique was

evaluated referring to the phases of the risk assessment that contemplates the risk identification, the risk analysis and the risk evaluation. The risk analysis phase can be divided consequence, the probability to occur and the risk level. The techniques were classified has applicable or not applicable conform the situation. The objective were determined based on the previous section.

Chapter 3

Preliminary Results

In this section it will be described the implemented methodology followed by the presentation and discussion of the obtained preliminary results.

3.1 Methods

The dataset used to train and evaluate the models refers to the results of annual Chicago restaurant inspections in order to ensure continued compliance with City ordinances and regulations as well as to respond to complaints. The variable are described in subsection 3.1.1. The methodology proposed to train this data obeys the following pipeline 3.1.

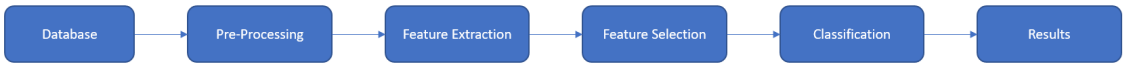


Figure 3.1: Methodology pipeline

As seen in the pipeline, the methodology can be decomposed in 6 phases, the database described in 3.1.1, the pre-processing, the feature extraction and construction, the feature selection, classification and finally evaluation of the models.

3.1.1 Datasets

The dataset consisted in information from inspections of restaurants and others food establishments in Chicago from January 1,2010. The inspections were performed recurring to a standardized procedure by the staff form the Chicago Department of Public Heatlh's Food Protection Program. All data from database is reviewed and approved by the State of Illinois Licensed Environmental Health Practitioner (LEHP).

This dataset can be decomposed in twenty different columns:

1. Inspection ID: represents the id of the each inspection. This id is unique for each inspections and it is represented by a integer (example:2346133).

2. DBA Name: Name under which each establishment operates, is a categorical variable and it is represented by strings (example: LAS TRADICIONALES).
3. AKA Name: Name of each establishment is known,is a categorical variable and it is represented by strings (example: LAS TRADICIONALES).
4. License #: It is a integer variable each allows to discriminate each establishment since the type of license is specific to establishment.
5. Facility Type: Type of facility of each establishment, this is considered a categorical variable and in the dataset is represented using string (example: restaurant).
6. Risk: This variable classifies the risk in 3 levels (1-high,2-medium, 3-low). This variable sis represented using a string and trough this infer that is a multi classification problem (three classes).
7. Address: Refers to the Address of the establishment, each address represents one establishment and is represented by strings.
8. City: Refers to the City of the establishments, this variable is the same to all the Establishment(Chicago)
9. State: Refers to the Sate of the establishments, this variable is the same to all the Establishment(Ilinois).
10. Zip code: Refers to the Zip code .
11. Inspection Date: Refers the to the date that the inspection occur, allowing to establish a time relation between them. This variable is a string (example:2019-10-17T00:00:00.000).
12. Inspection Type: Refers to the type of inspection performed. This variable is represented using string and can be divided into 106 different types(example: License).
13. Result: Refers to the results obtained by the inspections. This variable is a string type variable and possesses different outcomes: Pass, Fail, Pass with Conditions, the other will be removed.
14. Violations: Refers to the type of violations presented in the inspection. Being a string type variable is represented more than one type of violations in each cell, needing to be separated.
15. Latitude: Refers to the latitude in which the establishment is localized. This variable is a integer type (example:41.909910705082176).
16. Longitude: Refers to the longitude in which the establishment is localized. This variable is a integer type (example:-87.71438551021203).
17. Location-Combines the Latitude and Longitude variables. The other columns were empty so were not refered.

3.1.2 Data Preprocessing

The first step was preprocess our data, so a clean dataset could be used to implement the classification techniques. Initially, all the data from 2018 forward was removed because in the year 2018 the code that identify the type of violations was altered, resulting in misinformation. Then was necessary to eliminate the duplicates that existed in the dataset to allow a better performance of the classifiers. Once all the duplicates were removed was necessary to eliminate the outliers present in the different variables. To archive this results the features were evaluated to see the weight of each label of the feature. For example in the fig.3.2, can be seen the labels that had a lot less cases when comparing to the others.

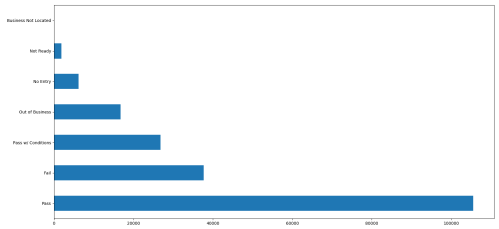


Figure 3.2: Example of the distribution of labels in the Results feature

This labels were considered outliers and removed. This process was applied to all the variables resulting in the removal of the least significant labels. In the Facility Type and to the Results were only consider the 26 most significant labels and 3 most significant labels, respectively.

Finally, the data was standardized. The standardized allows to give less importance to outliers.

3.1.3 Classification

In this step , various machine learning approaches were adopted.The main objective of classification is creating a classifier given a training set with class labels. The classifier will attribute a class label to an example. The algorithms implemented were Naive-Bayes. SVM, k-nn, Decision Trees, Random Forest XGBoost and LSTM.

The next section will the describe the parameters used in each of the previous mentioned algorithm.

3.1.3.1 Parameters

The algorithms were implemented using python with the implementation of sckit.learn libraries. All the algorithms were tested using teh default settings of this libraries with the exception of RF, the XGBoost and the LSTM.

The Random Forest was trained with 100 trees and with the nodes of the trees expanded until all leaves are pure.

The XGBoost is a type o gradient boosting algorithm and was computed using the parameters in Table 3.1 The neural network constructed in the methodology consisted in 3 layers. The firsts

Parameters	
Max depth	5
Number of estimators	1000
Number of classes	9

Table 3.1: Parameters used in the XGBoost

layer was a LSTM with 100 neurons and one hidden layer with 8 neurons. The hidden layer uses a rectifier activation function which is a good practice.Finally, was added one output layer with 3 neurons. The use of this deadset obliges that output layer must create 3 output values, one for each class. The use of a activation function 'softmax' in the final layer was to ensure the output values are in the range of 0 and 1 and may be used as predicted probabilities.

3.1.4 Feature Selection and Construction

In this section to build and select the most effective features was necessary to analysis the data.

Through a analysis to the variables, was concluded that most of the variables was categorical and possessed a string type. So, to this variables a label was assigned a new numeric label to allow the application of the classifiers. Then, was registered null values in the Facility Type and Inspection Type that were replace by a representative numeric label.

Through a analysis of the Results of the inspection tree new columns were necessary to characterize the type of violations that existed. The new columns added were denominated Minor, Serious and Critical containing the number of this type of violations for each inspection.

To pick the best features to improve the performance of the model a filter feature selection approach and Principal Component Analysis.

A filter feature selection approach is an approach that is independent of an induction algorithm and serve to filter irrelevant features, this method is computational less expensive. The filter ranks each feature using a uni-variate metric and then selects the highest-ranking features.

The metric variance is used to remove constant and almost constant features, the chi-square is used to classification problems,similar to the problem exposed, determines the dependency of two variables recurring to statistical test of independence. It can remove duplicate features and is capable of determine the ability of the independent feature to predict the target variable. However, filter methods only analyze individual features for identifying its importance, leading to the loss of important influencer feature when combined with the others.Through this technique a graphic was obtained.

In the fig.3.3, is possible to determine that the features months and inspection type do not influence enough the classifiers. This was the method selected due to the fact that the methodology. This was the method selected due to the fact that the methodology used relied on large amount of data, making the use of wrapper methods very computational expensive.Since, this analysis uses a

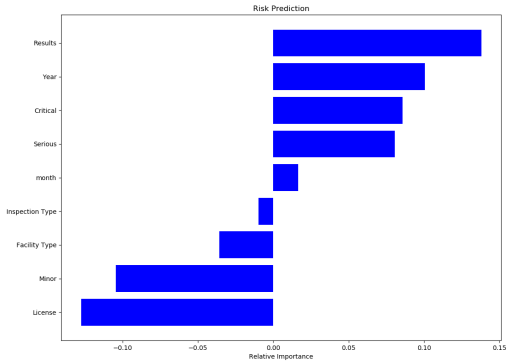


Figure 3.3: Importance of the features using a filter method

greedy search to encounter the optimal features to an specific machine learning algorithm. Principal Component Analysis, refers to a analyse where is chosen the number of principal components to be retained. This technique was implemented in the most different algorithms.

3.1.5 Evaluation

A 5-fold-cross validation was computed, to evaluate the classification performance, splitting the data in 5 parts at random and uses one of these parts as test and the others as training. The process is repeated fives times, assuring that the classifier is not tested with samples that were used in the train.In the process the test data is always different[45]. This ensures that the classifier is not tested samples that were used in the training and also decreases the deviations in the results. The predictions resulting from the five folds are aggregated and the accuracy is computed as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.1}$$

Where TP,TN,FP,FN corresponds to true positives, true negatives, false positives and false negatives ,respectively. The true positives corresponds to the inspections that were corrected classified, while the true negatives corresponds to inspections that were corrected identified but do not had any type of risk. The false positives and false negatives correspond to the cases were the inspection were incorrectly classified with type of risk.

3.2 Results and Discussion

In this section, a comparison will be established between the different algorithms implemented during the methodology. The methodology applied recurred to all the the algorithms previously described. The difference between the performances of the different algorithms can be seen in the following table:

This results are referent to results before the implementation of the techniques of features selection. As seen in the table 3.2 ,the best algorithm approach was XGBoost with a accuracy

Table 3.2: Accuracy results for the different implementations before feature selection.

	Accuracy
Simple Classifiers	
SVM	0.72
Naive Bayes	0.30
Decision Trees	0.42
K-nearest-neighbours	0.82
Ensenble Classifiers	
Random Forest	0,42
XGBoost	0.83
Neural Networks	
LSTM	0.73

of 0.83. The multi-classification Naive-Bayes and decision trees algorithms reported accuracy of 0.30 and 0.42, respectively. This indicates that these methods classify the inspections in a random way, indicating that should not be used in this type of the classification. This is also seen in the random forest algorithm.Although, SVM and k-nearest-neighbours were not ensemble methods had proven useful in risk classification, with accuracy of 0.72 and 0.8200 ,respectively, but being outperformed by ensemble classifiers. The LSTM demonstrated a reasonable accuracy taking into account the number of features utilized. After this first analysis, were implemented two types of feature selection has mentioned in 3.1.4. The results can be seen in table3.3.

Table 3.3: Accuracy results for the different implementations after feature selection.

	Accuracy
Simple Classifiers	
SVM	0.72
Naive Bayes	0.30
Decision Trees	0.57
K-nearest-neighbours	0.82
Ensemble Classifiers	
Random Forest	0,54
XGBoost	0.82
Neural Networks	
LSTM	0.73

Making the feature filter selection the accuracy of some algorithms decrease and others improved. The XGBoost registered a decline in the accuracy, this indicates that to this type of classifiers features that are not independent from other has more relevance than in decision trees classifier or the the decisions trees classifiers suffer from overfitting to the training data. The method XGBoost had an accuracy of 0.82. This means that the two features removed were not independent from others. The best improvement was the decision trees algorithm registering an

accuracy of 0.57.The wrapper feature selection is suggested to the ensemble classifiers in order to improve the accuracy, although it is more computational expensive. The LSTM registered the same accuracy, this could be to fact that the number of features were not sufficient in first place and reducing the same features did not improved the algorithm. Then was tested the PCA technique, however the results were not considered relevant to the exercise.

So, trough this analysis the best methods for classify risk utilizing inspections the ensemble classifiers outperform the others. The best technique was the XGBoost algorithm with an accuracy of 0.83. However, the SVM and Knn algorithm also presented promising results despite being simple algorithms. Finally, was expected that the neural network LSTM could accomplished the best results, but this was not verified. This fact is related to the number of features that were used revelling insufficiency to achieve better accuracy.

3.3 Conclusion

The importance of inspections and risk assessment has been proven a never-ending challenge that will contribute to a better health of the population. The algorithm approach necessities of some adjustments. The adjustments passes to recognized the type of violations trough the text present in the feature, utilizing text mining techniques such neural networks. Also, calculating more relevant features could lead to an increase in the accuracy of the algorithms. Adapting the data of 2018 and forward could also help to improve the efficiency of the algorithms. Associating this data with more data from the Chicago City databases could relate more factors such as garbage collection allowing to increase the number of features.

Chapter 4

Inspection frequency estimation based on risk

Nowadays, ASAE's methodology to identify entities to be inspected is based on risk they present to Public Health and Food Security, or to Commercial Practices and Industrial property or to Safety and Environment, and built on a qualitative model. However, with the increasing volume of collected data such as the number of complaints and the number of economic agents that are constantly opening, leads to the necessity to review the strategy and implement a quantitative approach. The main goal to of this approach is to determine the minimal inspection frequency,which mean the classes priority risk in each sector by month and district.

In order to fulfil the ASAE's principles of scientific independence, precaution, credibility, transparency and confidentiality described in [8], the developed method must be reproducible and explainable method recurring to a traditional machine learning pipeline based on the state-of-art methods. The pipeline demonstrated is the same as the pipeline described in 3.1.

The upcoming section details the available data and all the implementation procedures in order to test a quantitative inspection planning model based on risk assessment. In order to transform a qualitative inspection model in a quantitative inspection model several steps had to be done. The first one was to understand the qualitative model in which ASAE's bases their planning and the identification of the underlying features 4.1. Then was performed an analysis of the dataset which is presented in the sections 4.2 and 4.3. Sections 4.4 and 4.5 refer to the search for new features and the feature selection. Section 4.6 refers to the machine learning algorithms used for classification selected according to the criteria described in chapter 2 and in Appendices A. This table evaluates the algorithms described in literature according to the objective and applicability to each of the step of the risk assessment process according to ISO 31000 norm. This gives valuable information about the best method that should be implemented to fulfill the task at hands. Finally, section 4.7 discusses the obtained results.

4.1 Inspection Plan Qualitative Model

In order to conduct their inspections, ASAE first creates an Inspection Plan where priority-setting criteria were identified by the previously mentioned three major areas: the Public Health and Food Security, the Commercial Practices and Industrial property and at last the Safety and Environment described in [24].

To ensure that the plan is fulfilled this organization developed an inspection program based on qualitative risk matrices, through the use criteria and the identification of risk indicators. So, ASAE define 3 basic criteria to determine risk. These criteria are Consumption Volume (V), Performance (D) and Product risk (PS).

The Consumption Volume must be segmented by the activity sector and are from external but trustworthy sources as for example, INE or academic intuitions.

The Performance criteria is closely related to to the legal aspects of the process. The performance is determined by the total number of process of the previous year by sector. Also, is described using the degree of non-conformity, the level of risk to public health or economic security for the consumer and the associated precautionary measures taken.

Finally, the Product Risk can established using the frequency of the complaints received, the estimate of risk prepared by the Food Risks Division of ASAE for the Food Area is related to the origin of the product and, at last, the communications between international agencies.

The risk estimate by the Food Risks Division is associated with the representativeness of the operators' activity, the area of activity and the non-compliance related with activity which includes an analysis of the trends in the results of the National Sample Collection Plan (PNCA) an external laboratory. Also, one important criteria is the seasonality of the economic activities well as the regional specificity.

The features that contribute to each of this three criteria are qualitatively assessed and segmented into different "levels of risk" or severity.

The criteria have the purpose to help infer the minimal annual frequency of inspection by area of activity in order to elaborate the Inspection Plan.

In order to determine the minimal annual frequency inspection recurring to machine learning models the qualitative process must be transformed to a quantitative process to obtain more diligent results. Taking into consideration the three basic criteria defined by ASAE and their associated risk indicators models will trained to estimate the minimal annual frequency inspection for each sector accommodating the specificities of seasonality and regionality. Two output approaches were considered when determining the minimal annual frequency inspection. The approach was trough the calculation by how many times a sector has been inspected in a month in each district creating classes of inspection priority based on the magnitude of the values. It was considered that the higher volume of inspections represents a higher workload and therefore a context of higher risk.

This output will be used to determine the features utilised in the classification models.

4.2 Dataset

ASAE provided information regarding inspected institutions during a timeline of 14 years, beginning in 2005. The provided information was in a raw state being necessary to input in a database. As this thesis fits in a much bigger project this implementation was made by LIACC investigators involved in creating the prototype. The information was implemented in a database using MYSQL workbench platform to facilitate its handling, then information was extracted to python to proceed its treatment. The information contained into the database can be described as following columns:

1. **ID of the Operational Units**: represents the Operational unit that conducted the inspection, is a categorical variable and it is represented by strings.
2. **Inspection Type**: represents the type of inspection, if is a planned or non planned inspection. It is represented by a string.
3. **Date of inspection**: represents the date in each inspection was made. This variable allows to establish a timeline.
4. **Date when a complaint is submitted**: represents the date on which one complaint was made if any have been made.
5. **Number of initialized proceedings**: refers to the number of initialized proceedings after each inspection.
6. **Number of arrests**: refers to the number of arrests initiated to that particular inspection.
7. **Number of closed establishments**: refers to the number of closed facilities to that particular inspection.
8. **Has administrative offenses**: It is a boolean variable indicating if a particular inspection has administrative offenses. It is also, considered a categorical variable.
9. **Has crimes**: It is a boolean variable indicating if a particular inspection registered a crime. It is also, considered a categorical variable.
10. **Number of proceeding with administrative offenses**: refers to the number of infractions that result in administrative offenses.
11. **Number of proceeding with crimes**: refers to the number proceedings that result in crimes.
12. **The inspection has non-compliance procedures** : It is a boolean variable indicating if a particular inspection registered non-compliance procedures. It is also, considered a categorical variable.
13. **State of the inspection**: refers to the state of the inspection. The inspection can take two possible states, if completed (c) or incomplete (p). It is a boolean variable.

14. **Number of notices with administrative offenses**: refers to the number of notices that result in administrative offenses by inspection.

15. **Number of notices with crime**: refers to the number of notices that result in crime.

16. **Number of notices with administrative offenses**: refers to the number of notices that result in administrative offense.

17. **Number of infractions with administrative offenses**: refers to the number of infractions result in administrative offenses.

18. **Number of infractions with crime**: refers to the number of infractions result in crimes.

19. **Number of proceedings with no administrative offenses**: refers to the number of proceedings that don't have administrative offenses by inspection .

20. **Number of proceedings with no crimes**: refers to the number of proceedings that don't have administrative offenses by inspection.

21. **Number of by inspections in a partial state** : refers to the number of inspection in a partial state.

22. **Has infractions**: It is a boolean variable, that refers to the fact if a inspection had infractions.

23. **Sector ID**: refers to an identifier number that distinguish different sectors relatively to the commercialized or produced products. Is is a categorical variable.

24. **Sector Description**: refers to a description of the previously mentioned column.

25. **District**: refers to the District of the establishments using strings.It is represented by a sting.

26. **County**: refers to the County of the establishments using strings.

27. **Number of Samples**: refers to the number of samples collected by inspection.

28. **Notification**: it is a boolean variable, referring if a inspection was notified or not.

29. **Suspension**: it is a boolean variable, referring if the economic agents was suspended.

The database will suffer further transformations in order to produce the risk matrix used. These transformations will be described in the next sections.

4.2.1 Data Preprocessing

The first step was to eliminate the duplicates, meaning if the same inspections was registered in the database more than once that inspection was only accounted once. This allowed to achieve a better performance of the classifiers.

4.3 Analysis of the entities inspections dataset

As previous stated the inspections are referent to economic operators either in the food industry as in the economic chain. All the processes implemented will follow the previous mentioned design in order to obtain the minimal frequency of inspection each year for the sector of activity in a specific district, all features were analysed, so as to infer their behavior throughout time and the distribution of the labels in the features.

The first step was to analyse the numbers of motives that triggered the inspections, planned or non-planned, trough the years as seen in the fig. 4.1. From the analysis of the graph, it can be seen that the number of reactive inspections have been increasing. Although the underlying generated process we are trying to estimate refers to the planned inspections, the reactive or non-planed inspections represents a risk based reality that is take into account. Therefore, they were included.

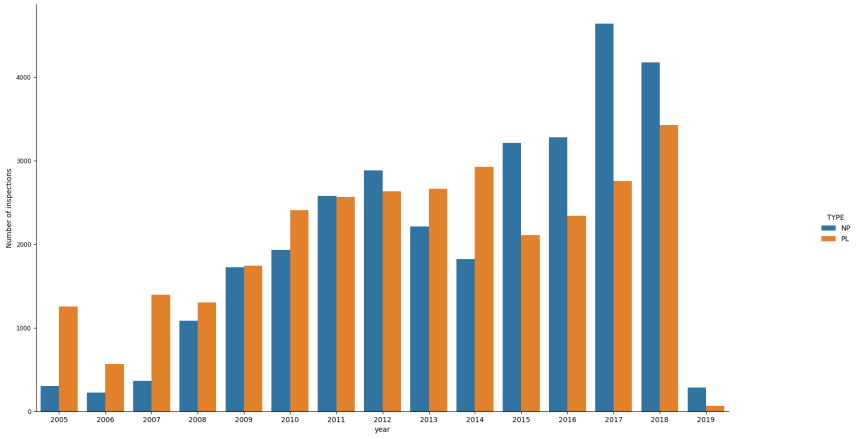


Figure 4.1: Number of planned and reactive inspections throughout the years.

Since seasonality is an important criteria, the analysis of the distribution of type of inspections through the months is important. Due to the large volume of data, we restricted the data from 2017 until 2018, to illustrate the variability of the feature. The results are plotted in the fig. 4.2. There can be seen that the planned inspections have the local maximum are roughly every six months, so the most expected frequency of inspection is biannual.

Subsequently, the features that refered to the number of processes was represented in fig. 4.3. The number of proceedings has risen since 2005 reaching its peak in 2018.

When examining the number of arrests in the same timeline, fig. 4.4, they reach the peak in 2012 and have much lower values in 2018, supporting that the more severe form of infraction has been decreasing in opposition to the number of processes. Both important features when determining the risk indicators of the performance of the economic agents criteria and therefore the minimal frequency inspection.

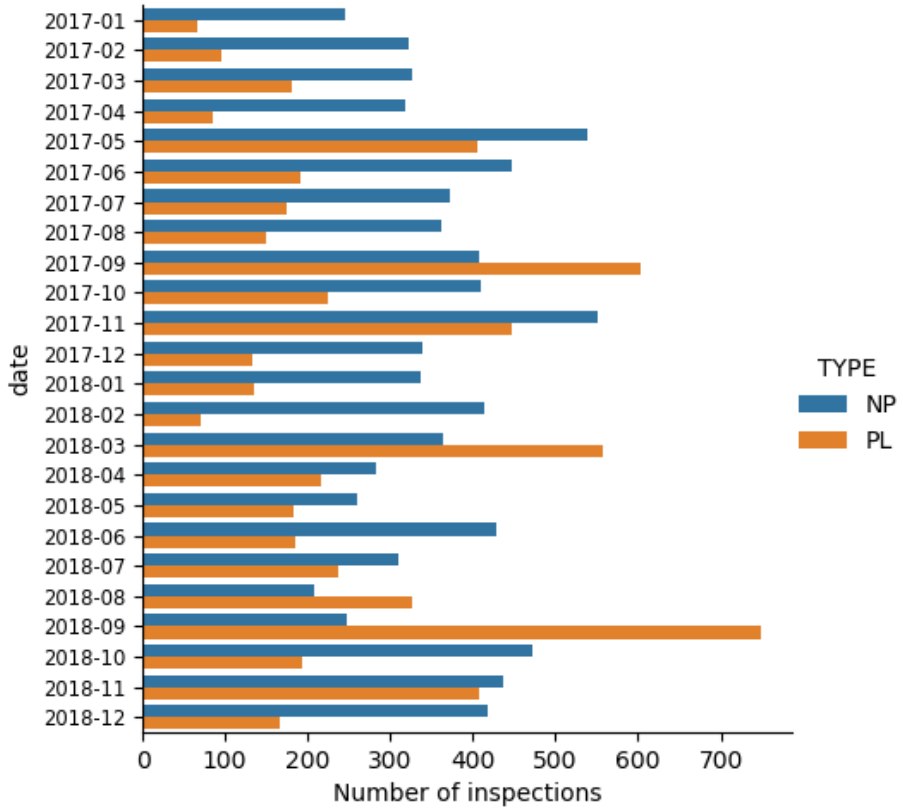


Figure 4.2: Number of planned and reactive inspections from 2017 to 2018 by month, where NP represent Non-planned inspections and the PL, planned inspections.

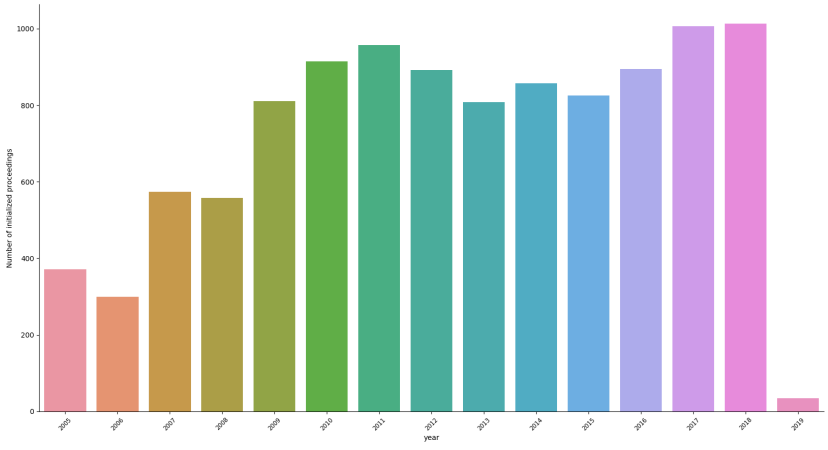


Figure 4.3: Number of initiated proceedings during the course of the years.

Also, the number of establishments that have been closed after the inspections has reached the maximum value in 2008, however in 2018 we have another local maximum (fig. 4.5).

Next, the number of process, notices and infractions with crime and administrative offenses

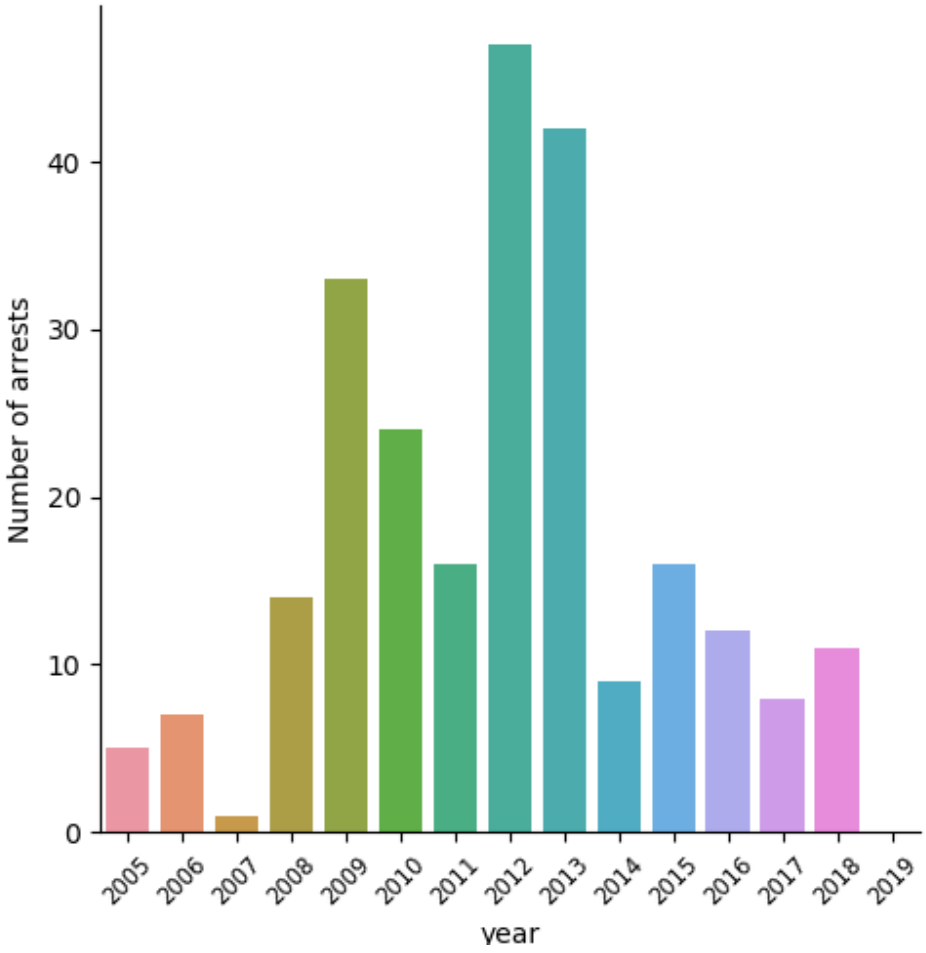


Figure 4.4: Number of arrests proceedings during the course of the years.

were studied in the figures 4.6 and 4.7, respectively. The number of administrative offenses were far superior to crimes, following the same tendency as the previous mentioned graph the number of arrests is decreasing. The maximum value registered in these 3 different features was in 2016, but in the same year this was not the maximum for the arrest feature and therefore crimes committed were not severe enough to emit an arrest, as seen in the Fig. 4.6. Also, the 3 elements presented similar behaviour regarding the number of crimes and administrative offenses as seen in Fig. 4.6 and Fig. 4.7. So, all the features are important to analysed from a juridic point of view the inspection. It must be noticed that the scale is different depending on the graphic that is analysed.

The higher values of administrative offenses when compared to the crimes is also supported when considering the inspections that had crimes and administrative offenses represented in Fig. 4.8. As can be seen the number of administrative offenses is far superior to the ones with crimes.

A final analysis was made to see if the number of processes that not had nor administrative offenses nor crimes (Fig.4.9). The number of processes that result in no crimes and no administrative offenses are inferior to those in which these condition was verified, the higher value reported

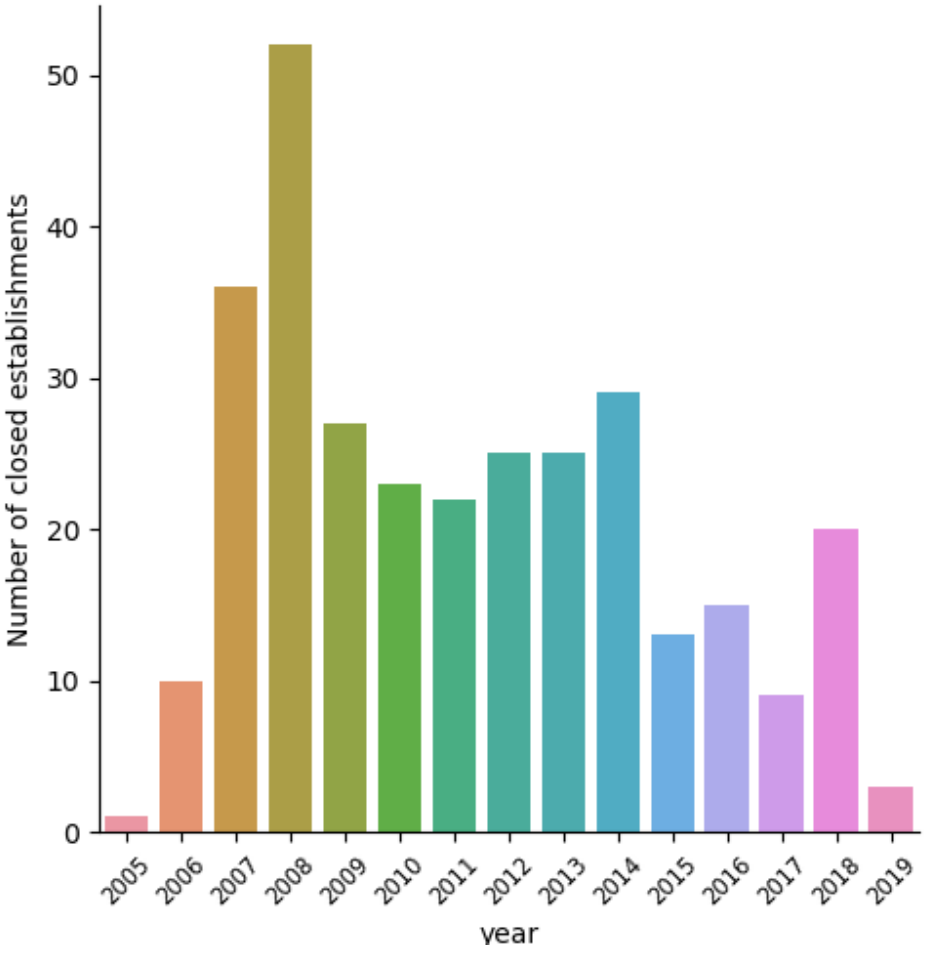


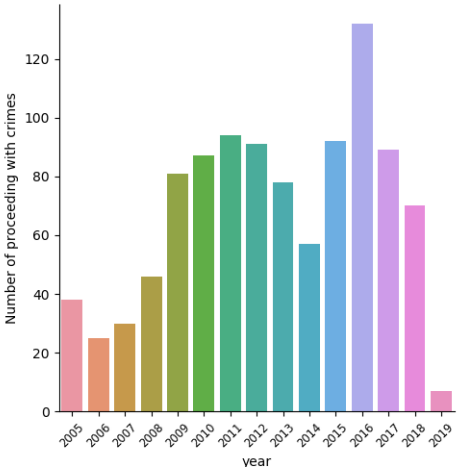
Figure 4.5: Number of closed establishments during the course of the years.

for processes with no crimes was in 2012, while the highest values reported for processes with no administrative offenses was in 2007 4.6.

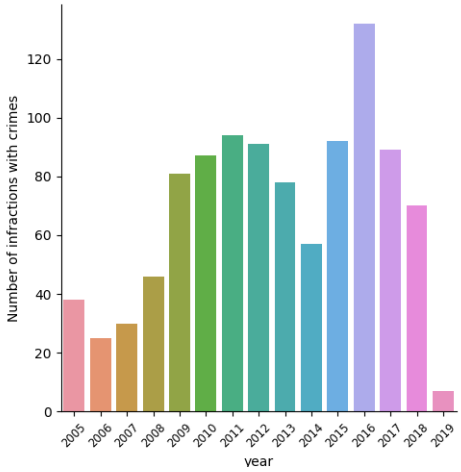
Afterwards, an analysis to the number of collected samples and the presence of complaints was made. As seen in the Fig. 4.10, the number of collected samples reached its maximum value in 2011. Trough this analysis can be seen that the data before 2009 was not collected. However, that represents a considerate portion of information in the others columns so it was not disregard, as can be seen in the previous figures. It reaches its maximum value in 2011 and thereafter the number of samples has been decreasing until 2017.

Regarding the presence of complaints can be seen that is raising, supporting the rise of the number of non-planned inspections, reaching its maximum value in 2011. Therefore, translating a reality that must be captured. Yet, from this year forward the data of complaint was not registered, but represented nonetheless important information in other columns. To analyse this feature was considered each data when a complaint was made as a complaint made.

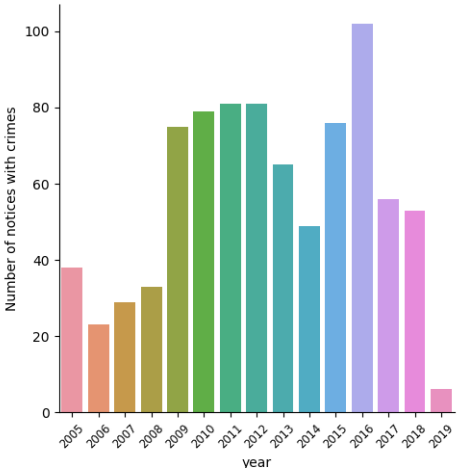
Then, it was necessary to analyse the number of inspections by sector to obtain a perception



(a) Processes with crimes

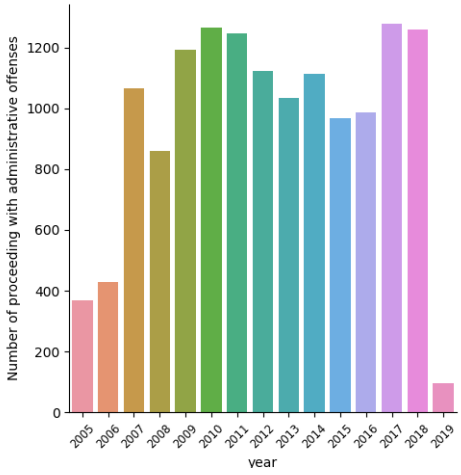


(b) Infractions with crimes

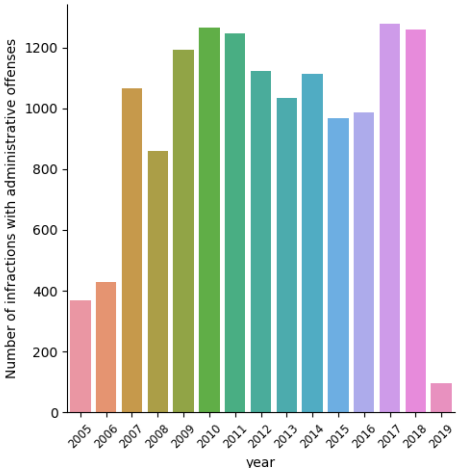


(c) Notices with crimes

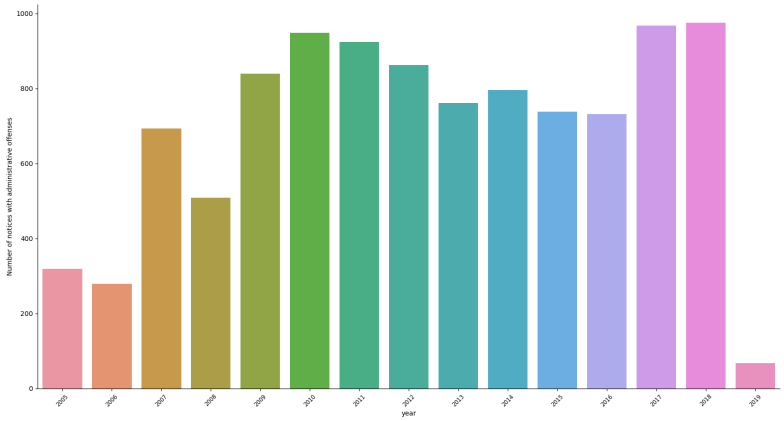
Figure 4.6: Number of processes, infractions and notices with crimes trough the years.



(a) Processes with administrative offenses

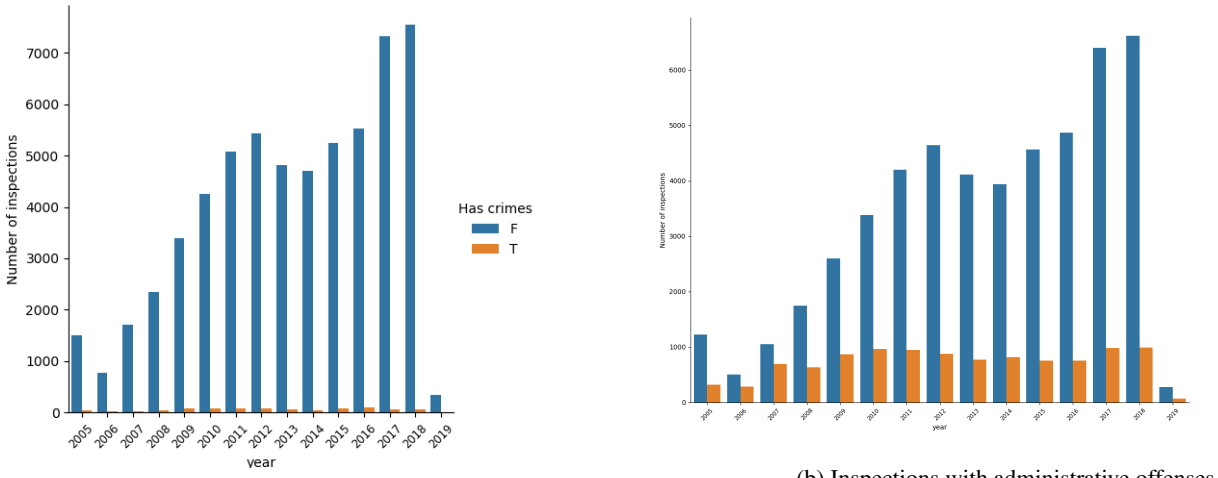


(b) Infractions with administrative offenses



(c) Notices with administrative offenses

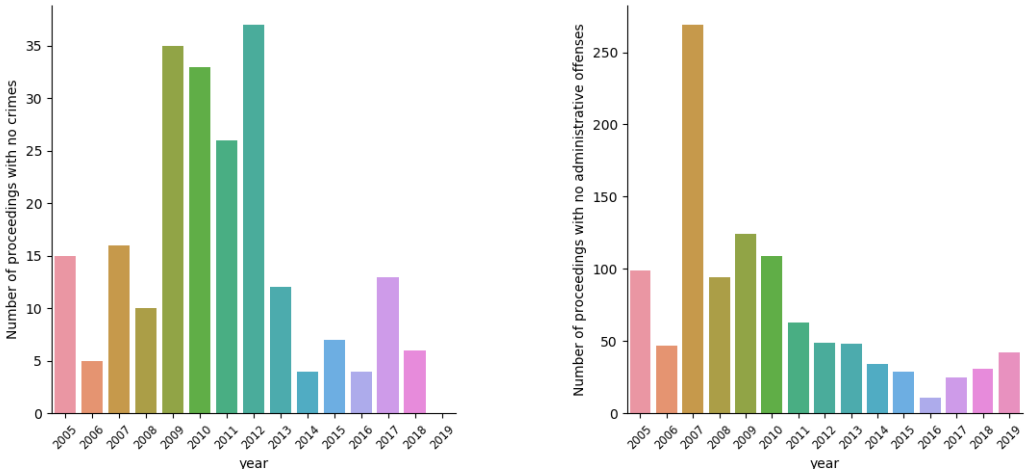
Figure 4.7: Number of processes, infractions and notices with administrative offenses trough the years.



(a) Inspections with crimes

(b) Inspections with administrative offenses

Figure 4.8: Inspections with crimes and administrative offenses trough the years.



(a) Processes with no crimes

(b) Processes with no administrative offenses

Figure 4.9: Processes with no crimes and no administrative offenses trough the years.

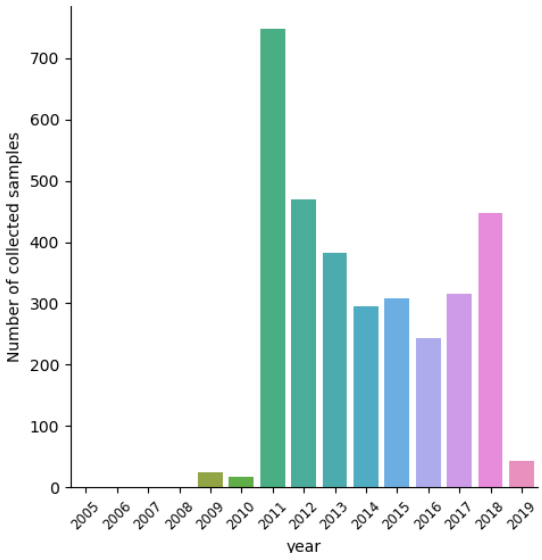


Figure 4.10: Total number of collected samples trough the years.

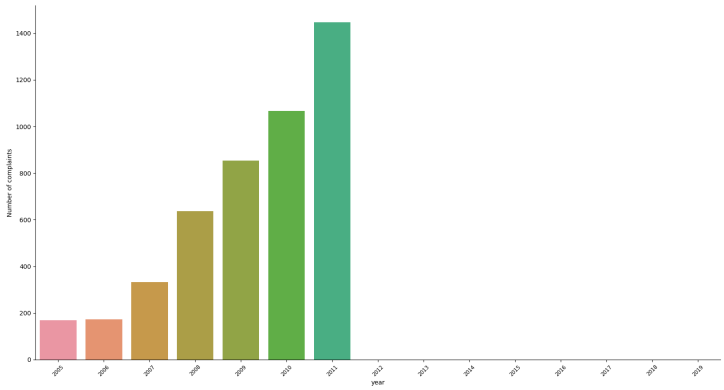


Figure 4.11: Total number of complaints trough the years.

how it evolves over time and if the ground truth could be calculated from it. The results are shown in the Fig. 4.12. In order to improve visual clarity, since the number of different sector were very high, the representations was only made to the six most frequent sectors. The sectors represented are the retail sale of beverages in specialized stores (n = 4725), road freight transport (n = 4941), retail sale of watches and jewellery in specialized stores (n = 4777), retail sale of footwear and leather goods in specialized stores (n = 4772), other land passenger transport (n = 4939), land, urban and suburban passenger transport (n = 4931). Analysing the graph, it can be noticed a more or less constant time space between the peaks for each sector. However, some of the peaks do not appear in the same month in different years. This can be due to effect of the specific season products such as the olive oil. So, to minimise the variation of value of month was proposed to identify if a particular sector should be inspected in a trimester starting in December.

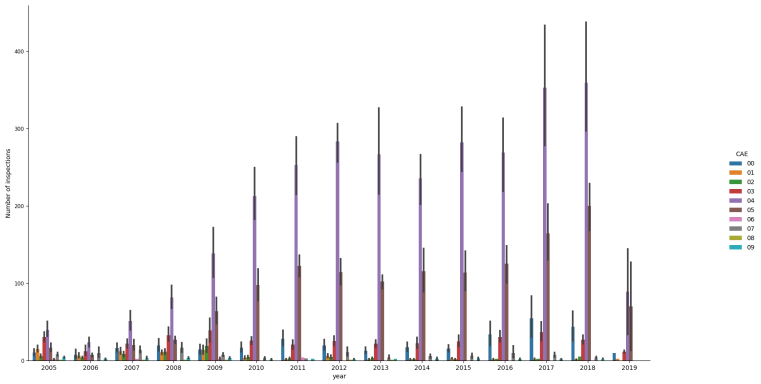


Figure 4.12: Comparison the number of inspections between the sectors during a year timeline.

Afterwards, was made a comparison between the number of notifications that a inspection received. Examining the image 4.13,the number of inspections that received only one notification is far superior in all years than the inspections that received more than one. Then, it was analysed

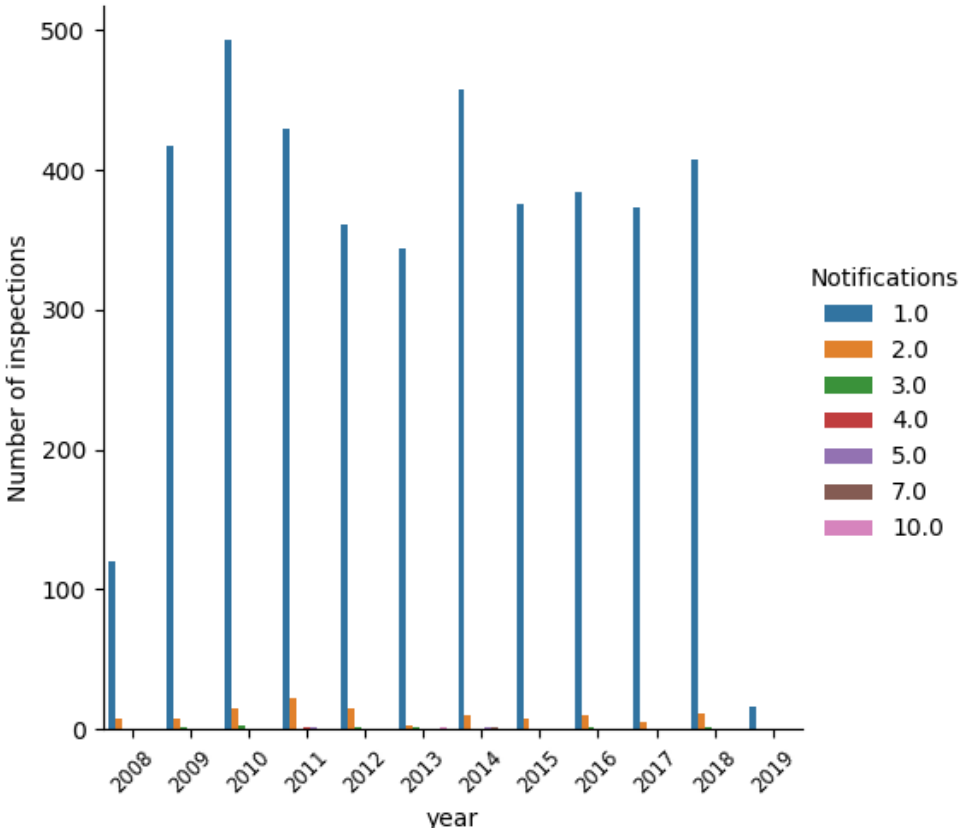


Figure 4.13: Comparison the number of Notifications trough the years.

the districts in which occur the most inspections trough out the years. This district corresponds

to Lisboa, following by Porto district and Setúbal. Which being the biggest cities of Portugal, concentrates the higher number of economic agents

This results are expressed in the Fig. 4.14

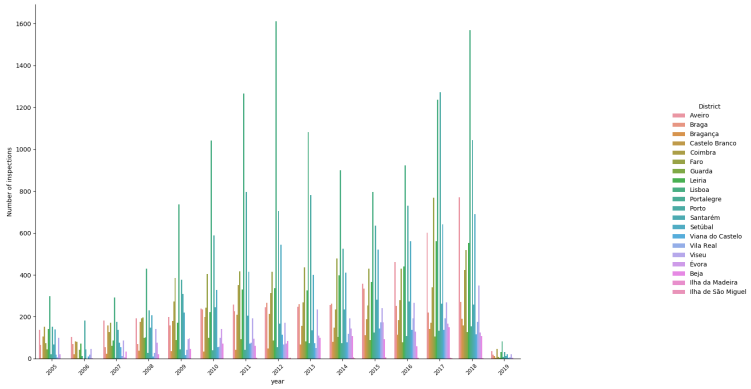


Figure 4.14: Comparison the number of inspections by districts trough the years.

Following the analyses of the inspections by district it was important to check which of the organizational units conduct more inspections. The results are expressed in the Fig.4.15. The organizational unit that perform more inspections was the UO1 in 2018, located in Porto,supporting the previous presented results. Next, the number of inspections that were suspended was repre-

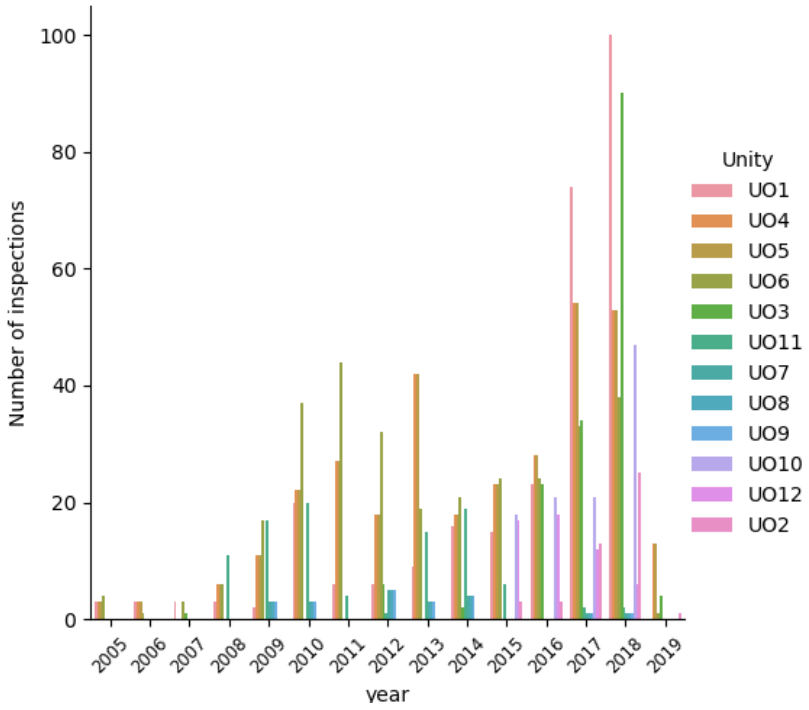


Figure 4.15: Comparison the number of inspections realized by organisational units trough the years

sented in Fig. 4.16. It can be seen that the number of partial (p) inspections have been decreasing, however it registered a rise in the year of 2018. The 'T' represents inspections that are finished.

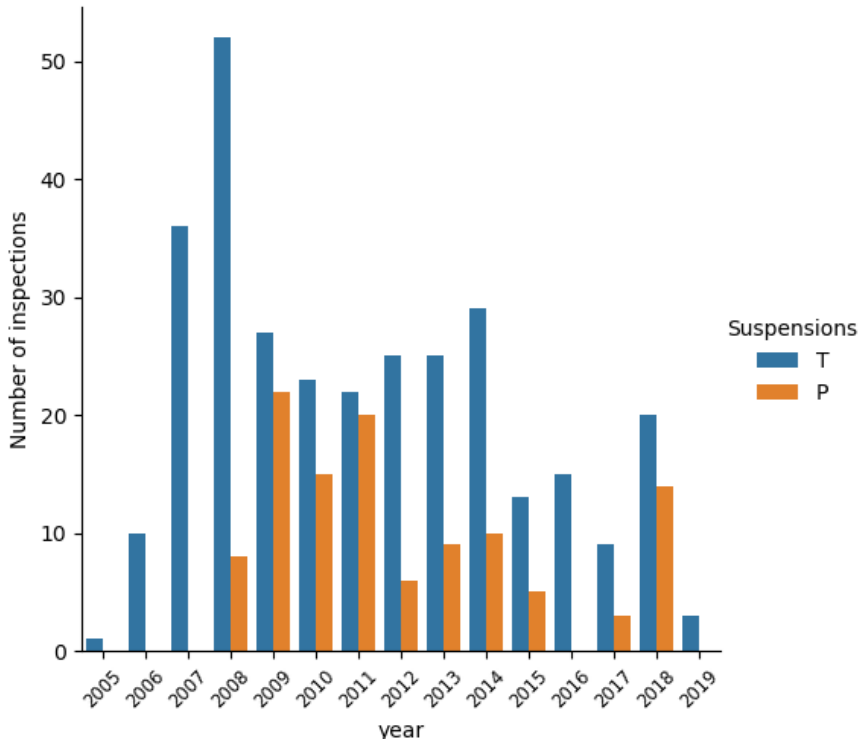


Figure 4.16: Comparison of inspections state trough the years.

Finally, the results of the number of inspections in a partial state was compared during the years, Fig. 4.17. The values reached it maximum in 2013, but has a noticeable increases in 2018 when compared to the other years.

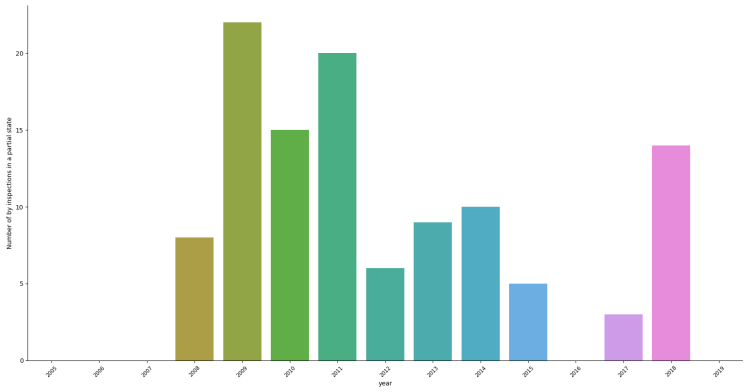


Figure 4.17: Number of inspections in a partial state trough the years.

4.4 Feature Extraction

To build a successful risk based inspection model, all the selected features must describe as best as possible the three basic risk criteria. The risk criteria were already explained in previous section 4.1 taking into consideration [24]. The criteria was composed by the performance, consumption value and the service product. The associated risk indicators are described in table 4.1.

Table 4.1: Qualitative criteria source[24]

Features	**Risk Indicators**
Performance	Number of cases from the previous year, segmented by sector of activity Degree of non-conformity Level of risk to public health or economic security for the consumer Associated precautionary measures
Consumption volume	Data from INE
Product Risk	Complaints received at ASAE (frequency) Risk estimate prepared by the Food Risks Division for the Food Area Product origin Communications by other national and international entities

The risk indicators will be approximated by some of the available features.

4.4.1 Performance

As previous state the performance criteria can be defined by several risk indicators. To cover this criteria, as accurately as possible, several columns of the database were used and they are represented in the table 4.2.

Table 4.2: Performance criteria represented by features.

Number of initialized proceedings	Number of arrests
Has administrative offenses	Has crimes
Number of proceeding with crimes	Number of notices with crime
Number of infractions with crimes	Number of infractions with administrative offenses
Number of proceedings with no crimes	Number of closed establishments
Number of proceeding with administrative offenses	Number of proceedings with no administrative offenses
Number of notices with administrative offenses	

The proposed features suffered some transformations before could be used as part of the risk model. These transformations centered around the calculation of the average number for each feature.

4.4.2 Consumption volume

To obtain information about the consumption volume, was necessary to consul another information sources of information such as the Portuguese food scale and several indicators of Portuguese

Economic Activity. Through research of information that could be used to create the features and with the time, regional and activity sector needed granularity, the most reliable one was collect from Instituto Nacional de Estatística (INE) recurring to the use of Application Programming Interface (API) following the instructions described in [13]. The information extracted from the API, that fulfilled the needed requirements, contained the number of companies and the business volume by region, by sector and by year. Although, several others statistics were found such as Food Scale [14]. However, the information had not the same granularity as the information from the ASAE. So, it was necessary to reduce the granularity of the the region to districts. This lead to the districts of the same region to have the same value. The criteria can be described by the features represented in the table4.3.

Table 4.3: Consumption volume criteria represented by features.

number of companies	volume of business

4.4.3 Product Risk

The Product Risk criteria contemplates several points to address. In order to be used in the risk model the average number of samples was calculated.The number of samples used can give some information about the risk estimated and be a good complement to the number of legal cases. Then, to estimate the frequency of complaints, first, was estimated the number of complaints by sector, month and district considering that a complain is represented by a date in 'Date of Complaint'. Finally, was computed the average values of Infractions, Notifications and Suspensions for each sector in a specific month by district. However, there was one indicator that could not be computed due to the lack of information on communication between national and international entities. The product origin was not determined, since there was no information regarding this, although it is intimately related with the sector and the district. Also, the seasonality was estimated recurring to the Date of inspection extracting into a new feature date the month and the year. The criteria recurred to the features described in table 4.4.

Table 4.4: Service/Product criteria represented by features.

Date of inspection	Date when a complaint is submitted
The inspection has non-compliance procedures	Has crimes
State of the inspections	Sector ID
District	Number of Samples
Notification	Suspension

4.4.4 Risk Model

Finally, in order to compute the minimal inspection frequency and to determine if the models that will be used were accurate, the ground truth was created. Firstly, as stated in 4.1, was computed

how many times a sector has been inspected in a month in each district creating classes of inspection priority based on the magnitude of the values. These results are will be later compared in the results section 4.7.

4.5 Feature Selection

The number of features that resulted from the feature extraction stage could lead to effects of high-dimensional data in the next steps. So, a feature selection method was necessary. The method is divided into filter and wrapper methods. As stated in the section3.1.4 , the filter method only evaluated the intrinsic properties of the data disregarding the interaction with the classifier. On the contrary, the wrapper method selects features based on the performance of the underlying classifier model. Therefore, computational more demanding than the filter method, therefore less suitable to big data sets [32]. Therefore, it was applied an filter method, more specifically a ReliefF algorithm described in the next section.

4.5.1 ReliefF

ReliefF is filter feature selection technique that is computationally efficient. The filter assigns a weight to each feature according to the values of k neighbouring samples. The algorithm selects an instance at randomly and search its k neighbours from the same class and the k neighbours of each of the classes, assigning weights according the distance between the instance and its neighbours classes and rewarding those of who give distant values to neighbours from different classes, allowing to indirectly consider feature interaction making it particularly sensitive feature dependencies and interactions [55]. The algorithm allows to rank the weights, enabling the choice of the best features and the elimination of those with a low discriminative power. The ReliefF is a faster and scalable method, contrarily to multivariate methods.

The algorithm was implemented in python with value of k as 10 to rank and select the N best features according to Urbanowicz in [56] as the most commonly used . The best features were analysed in the4.7.

4.6 Classification

Based on the literature, it was applied three types of different classifiers. The simple classifiers include the SVM, the Naive Bayes, Decision Trees and K-Nearest-Neighbours, as the ensemble classifiers were composed by the Random Forest and XGBoost classifiers. Finally, a Long-Short Term Memory Neural Network was constructed. In the Simple classifiers was used an error-correcting output codes (ECOC) classifier, that reduce the problems of classification with three or more classes to a set of binary classifiers in a one-vs-one approach[21]. This was utilised because its a multiclassing problem. As previous stated, all the methods were implemented in python with the package of keras.

After, a grid search was applied to improve the classifiers with the most suitable hyperparameters, regarding the classification mean accuracy over all the trimester or the annual frequency of inspection.

For the Decision trees for the max depth of three, the grid search ranged from 4 to 100 in increments of 1 until the value 10 and thereafter, in increments of 10.

The k-nn classifier was optimized using increments 1 from 1 to 11 for the values of k, also it was analyzed the best metric between the manthatan and the eucledian distance and the weight distribution, uniform or distance.

Concerning the SVM, the polynomial kernel order ranging from 1 to 6 and a radial basis function(RBF) kernel were compared, the value of gamma was automatically computed by the implementation.

Regarding the Random forest ensemble classifier, several hyperparameters were compared in order to obtain the best performance of the classifier. Using a grid search technique, this ranged from 50 to 100 in increments of 10 concerning the max depth of each tree. Regarding the max features per leaf a search with increments 5 from 10 to 25. The minimal samples per leaf and the minimal split of the samples were incremented by 1 starting in 2 and 7 and ending in 4 and 9 , respectively. Finally, the number of trees in the classifier were ranged from 100 to 1000 in increments of 100.

In respect of the XGBoost, were compared the number of trees starting at 500 in increments of 100 to 1000, the learning rate between the 0.01 and 0.2, the max depth of the trees starting at 3 with increments of 3 to 9. The minimal child weight were also tested between 3 values 10, 11, 12, the fraction of observations to be randomly samples for each tree was also compared between 0.5 and 0.7. Finally, the grid search method was not applied to the LSTM.

This method allowed a construction of best and more efficient algorithms, through the comparison of different hyperparemeters to improve the classification.

Which means a 5-fold-cross validation was computed. As previous explained, a 5-fold-cross validation was computed to evaluate the classification performance, splitting the data into 5 parts at random and uses one of these parts as test and the others as training. The process is repeated five times with the test data changing at each iteration, guaranteeing that samples used in trains are not used in the test [45]. The method ensures that the data is not biased separated, decreasing the deviation in the results. The predictions resulting from the five folds are aggregated and the accuracy is computed as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.1}$$

Where TP, TN, FP, FN corresponds to true positives, true negatives, false positives and false negatives, respectively. The true positives corresponds to the sectors that were corrected classified, while the true negatives corresponds to sectors that were corrected identified but do not had any type of risk. The false positives and false negatives correspond to the cases were the sectors were

incorrectly classified with type of risk or with no risk, respectively. The accuracy can interpreted with the fraction of corrected predictions by the model.

4.7 Results

4.7.1 Hyperparameters

In the previous section, it was described the methods to ensure an improvement in the overall behaviour of the system through the calibration of the different parameters.

Utilizing the grid search technique explained in that section, the models were optimised to the best parameters. The best parameters for each model are described in the following table 4.5. These were the models utilized in the remaining of the work.

Table 4.5: Best hyperparemters for each classifier.

Classifier	Hyperparameters
Simple Classifiers	
Decision Tress	criteria=gini max_depth=40
K-NN	metric=manhattan n_neighbours=2 weights=distance
SVM	kernel=polynomial gamma=3 degree=3
Ensemble Classifiers	
Random Forest	max_depth= 70 max_features= 15 min_samples_leaf = 2, min_samples_split = 8, n_estimators= 700
XGBoost	learning_rate=0.2 max_depth=6 gamma=2 min_child_weight=11 subsample=0.7 n_estimators=700

4.7.2 Number of selected features

After define the hyerparameters of the classifiers, was tested the ideal number of features selected by RelifF algorithm to see how much the course of dimensionality influences the proposed classifiers.

The mean classification accuracy was computed for the N best features, with N ranging from 1 to 27 in increments of 1. The results are plotted in Figure 4.18. All the classifiers with optimal

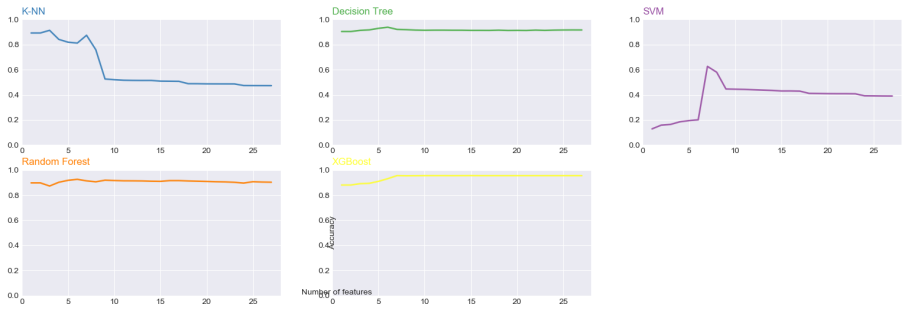


Figure 4.18: Selection of the best number of features in increments of 1.

hyper parameters were considered when comparing the average accuracy between methods. All the classifiers experienced a slight increase in the performance as the number of selected features also increased. However, the SVM and K-NN classifiers show a notable decrease in the accuracy with the increased number of features. This phenomenon is described in machine leaning as the Curse of Dimensionality. Taking into account a fixed number of samples, with the increase of the dimensionality, the volume of the feature space increases exponentially, which makes the samples sparse and the classification more difficult.

The others classifiers seems not be afflicted with this problems. As the criteria mentioned in the section 4.5 must be obeyed, the classifiers affected by the Curse of Dimensionality were discard because all the features must be used.

In the next section only the SVM, Decision Trees, the Random Forest and XGBoost classifiers were addressed and the LSTM neural networks.

4.7.3 Minimal frequency inspection Classification

To achieve the best explainable and reproducible model to determine the minimal inspection frequency, trough the estimation of the class of inspection priority by month in each district in each sector, an analysis of the performance of the different classifiers was essential. This allows to determine the best classifier to be implemented.

As seen in the previous section the accuracy was very similar in the classifiers not affected by the Curse of Dimensionality so it was necessary a more complete analysis using different metrics. The classifies will be evaluated according to metrics extracted from the confusion matrix. Therefore, for each class, the accuracy, precision, recall and F1-score were computed in each

classifier following Equations 4.2 until 4.4.

$$Precision = \frac{TP}{TP+FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP+FN} \tag{4.3}$$

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{4.4}$$

This metrics offer a different perspective of the performance of the system. Precision refers to the proportion of the positive that was actually corrected, the number of true positives divided by all the positives. The recall refers to the positive identifications that was correct, the number of true positives divide the true positives and the false negatives. Usually, exists an exchange when referring to the the improving of the Precision and the Recall, meaning that if one improves the other decreases. So, both metrics have to be considered when evaluating the results. The f1 metric allows a perception between this two metrics trough the computation of the harmonic mean of precision and recall. The results in Table 4.6 confirm that the model is balanced and achieves overall satisfactory performance. The Random Forest represent the worst results of all

Table 4.6: Results of the minimal inspection frequency approach

Classifiers	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Simple Classifiers				
Decision Trees	0.7262	0.7271	0.7290	0.7280
Ensemble Classifiers				
Random Forest	0.8658	0.8663	0.8668	0.8665
XGBoost	0.8752	0.8752	0.8752	0.8752
Deeplearning				
LSTM	0.5652	0.5643	0.5651	0.5647

the classifiers, however as a best relation between the recall and the precision than the Decision Trees. The XGBoost presents the best results with an accuracy of 0.8752 accuracy. Also, this classifier is the best detecting the class of inspection priority risk. The classifier gives information if a sector has a high priority risk in a particular month and in a particular district, allowing a better allocation of resources. This is demonstrated by the high values of precision and recall. The LSTM performed very poorly when compared with the other classifiers, with the lowest of the metrics of them all. This metrics were so low that can be used in the problem-solving contradicting what was previous assumed based on the literature.

4.7.4 Final Remarks

In this chapter, two different approaches to infer the minimal frequency inspection were proposed. The results reveal a good overall performance identifying the minimal frequency of inspection taking into account the ground truth was not given and had to be calculated. The results reveal

a good performance, with accuracy of 87,52% to identify the class priority risk of each sector in a month and district . The results of some of the models described in the literature in response to similar problems to some degree usually, presented worse results. When leading with real data some problems arose such lack of information or difficulties in conceptualize the problem to validate the used methodology. However, these problems were mitigated achieving a good tool that can be implemented in this real-life problem, bearing in mind that is a first approach to the problem and can be further improved.

Chapter 5

Conclusions and Future Work

The main goal to this dissertation was to develop an efficient tool capable of analyse ASAE's data and determine the minimal frequency of inspection, concretely the estimation of the class risk inspection in each sector by district in each month. The results obtained corroborated the use of this tool in a real context and as a solution for the problem. An accuracy of 0,8752 was achieved to determine the minimal frequency inspection by sector considering all the inspections, planned and non-planned. So, the method reproducible and explainable was achieve with medium success. This register an improvement relatively to the methods tested in the section3.2. Meanwhile, must be kept in mind that the dataset used in both determinations were different.

The biggest challenges of this work were the conceptualization of the problem and some prior lack of information. These challenges are related to a totally new problem that arose. The work devised in the thesis seeked to supplant some of the first difficulties when creating new methodologies. The results revealed that the information added, extracted from trustworthy sources and the features utilized to described the 3 criteria demanded by ASAE, Performance, Consumption Value and Product are quite suitable and can be used as base to further improve the tool.

Since it a first approach exists some room of improvement. As future work is propose to assign different weights can be done to crimes related features since it is more provable to occur an inspection in this sector. Determine the minimal frequency of inspection not by sector but by economic operator can be crucial into to build a more desirable tool. Have access to an output define by ASAE could also help to better validate the results. Also, It is suggested to implemented regression-based algorithms to further improve the methodologies. Also, in order to accommodate some fluctuations in the values that can occur because of moving holidays such as Easter its recommended to infer the if a sector in a district is inspected in a specific trimester. Besides all the previous mentioned limitations and suggestions, future work should explore in a deeper manner the LSTM classifier that although reported the worst results in a highly versatile method to analysed sequences. These suggestions could help to improve the work made on this thesis and carve the path to a more regulated and crime free society.

Appendix A

Table

The following table presents a summary of the study of the algorithms according to the objective and applicability to the steps of the risk assessment process. |p1.8cm|p3cm|p4cm|p1.2cm|p1.2cm|p1.2cm|p1.2cm|p

		Risk Assessment Process			
Type	Description	Objective			Accuracy
Risk		Risk analysis		Risk	
Identification	Consequence	Probability	Level of Risk	evaluation	

Table A.0 – continued from previous page

		Risk Assessment Process			
Type	Description	Objective			Accuracy
Risk		Risk analysis		Risk	
Identification	Consequence	Probability	Level of Risk	evaluation	

Continue in the next page

Naive-Bayes Using a training set, a classifier is constructed based on the bayesian distribution, where the attributes are independent from one another. The classifier assigns a label to the input. This type of classifier can be used with different datasets(example: heart disease). Used to assigns a label,class,to an example. Should not be used Highly recommended Should not be used Should not be used Should not be used Approximately, 0.70. However, to determine risk of heart disease this method represented an accuracy of 1.00.

K-NN Using a training set, a classifier is constructed. This classifier can be used in regression or classification and the assigned label is dependent of the label of the nearest neighbours. RBI screening assessment reutilizing information from RBI systems. The dataset utilized was from oil pipelines Highly recommended Should not be used Should not be used Should not be used

Should not be used 0.8787 to screening assessment in a RBI system

SVM Support vector machines is classifier that separates the input space, utilizing linear or non-linear kernels as separate functions. This type of classifier can be used with different datasets.Used to assigns a label,class,to an example. Can be used Highly recommended Should not be used Highly recommended Should not be used Depending the situation is utilized it can be different accuracies. It presents 0.8489 in a screening assessment,0.67 in credit risk, True positive rates of 0.7136 in prediction of risk of fire, also it a achieved an accuracy of 0.80 in credit risk of stakeholders in the food chain

KLR Kernel Logistic regression classifier is very similar to support vector machines that can be generalized to multiple classes using trough kernel multi-logit regression. Utilized to determine credit risk.The data utilized was from the german credit data. This data had 15 attributes and 1000 samples. Highly recommended Highly recommended Should not be used Highly recommended Should not be used 0.78 to classify credit risk

Association rules Machine learning algorithm that allows to determine strong relations between variables in databases. Utilize to pre-determine risk in the food-chain. A case study was made in Sanyuan were a supply chain system was created.accountability and empowerment, which enabled company to ensure product quality Highly recommended Highly recommended Should not be used Should not be used Should not be used -

Decision Trees Decision tree is an algorithm that uses condition into is nodes to reach one classification. Used with different datasets. Used to predict risk or in problems of classification. For example predict the risk of a person has heart disease Should not be used Highly recommended Highly recommended Can be used Can be used Accuracy of 1.00 with 0 error to predicting the risk of a person has heart disease.

Multi-Objective Genetic Algorithm An algorithm that utilizes concepts of genetic and gene transmission and applies to computation. This algorithms will converge to the best solution approximating of the Pareto Space. Risk evaluation recurring to RBI system information Should not be used Highly recommended Highly recommended Can be used Highly recommended -

Random Forest/Decision Trees Ensemblers Algorithms Random Forest is a ensemble algorithm that consists in multiple decision trees in the input and output Used to determine the risk of fires to better plan the inspections. Can be used Highly recommended Highly recommended Can be used Can be used 0.9228

LSTM LSTM is an artificial neural network. This network is capable of feedback connections allowing to process entire sequences of data Utilize to evaluate the credit of stakeholders in the food chain. Highly recommended Highly recommended Highly recommended Highly recommended Highly recommended Approximately, 1.00. Outperforms other algorithms as SVM and Decision Trees

References

[1] About efsa | european food safety authority. http://www.efsa.europa.eu/en/aboutefsa.

[2] Averting risks to the food chain. http://www.fao.org/3/b-i6538e.pdf.

[3] Code of practice no 9. http://www.reading.ac.uk/foodlaw/uk/cop9.pdf.

[4] Código de conduta e Ética. https://www.asae.gov.pt/asae1/instrumentos-de-gestao/codigo-de-conduta-e-etica.aspx.

[5] Economic agents.

[6] Economic burden of major foodborne illnesses acquired in the united states. https://www.ers.usda.gov/webdocs/publications/43984/52807_eib140.pdf.

[7] Haccp. https://www.asae.gov.pt/seguranca-alimentar/haccp.aspx.

[8] Missão, visão e valores. https://www.asae.gov.pt/asae20/missao-visao-e-valores.aspx.

[9] Organograma. https://www.asae.gov.pt/asae20/organograma.aspx.

[10] Plano de atividades. https://www.asae.gov.pt/asae20/instrumentos-de-gestao/plano-de-atividades.aspx.

[11] Plano estratégico. https://www.asae.gov.pt/asae20/instrumentos-de-gestao/plano-estrategico-.aspx.

[12] Plano operacional. https://www.asae.gov.pt/inspecao-fiscalizacao/fraude-alimentar/plano-operacional.aspx.

[13] Portal do ine. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_api&INST=322751522&xlang=pt.

[14] Portal do ine. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=289818234&PUBLICACOESmodo=2&xlang=pt.

[15] Quar. https://www.asae.gov.pt/asae20/instrumentos-de-gestao/quar.aspx.

[16] Resultados operacionais. https://www.asae.gov.pt/inspecao-fiscalizacao/resultados-operacionais.aspx.

[17] Sustentabilidade. https://www.asae.gov.pt/asae20/sustentabilidade.aspx.

[18] What is fmea? failure mode & effects analysis | asq. https://asq.org/quality-resources/fmea.

[19] Who_fos_15.02_eng.pdf. https://apps.who.int/iris/bitstream/handle/10665/200046/WHO_FOS_15.02_eng.pdf?sequence=1.

[20] Anabela Afonso. Metodologia haccp. *Segurança e qualidade alimentar*, 1:12–15, 2006.

[21] Erin L Allwein, Robert E Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of machine learning research*, 1(Dec):113–141, 2000.

[22] Wayne A Anderson. The future relationship between the media, the food industry and the consumer. *British Medical Bulletin*, 56(1):254–268, 2000.

[23] P Antunes, J Machado, and L Peixe. Illegal use of nitrofurans in food animals: contribution to human salmonellosis?, 2006.

[24] ASAE. Critérios base para a estimativa de risco das atividades a fiscalizar.

[25] Nasibeh Azadeh-Fard, Anna Schuh, Ehsan Rashedi, and Jaime A Camelio. Risk assessment of occupational injuries using accident severity grade. *Safety science*, 76:160–167, 2015.

[26] Diána Bánáti. European perspectives of food safety. *Journal of the Science of Food and Agriculture*, 94(10):1941–1946, 2014.

[27] John Barnes and RT Mitchell. Haccp in the united kingdom. *Food Control*, 11(5):383–386, 2000.

[28] Nidhi Bhatla and Kiran Jyoti. A novel approach for heart disease diagnosis using data mining and fuzzy logic. *International Journal of Computer Applications*, 54(17), 2012.

[29] Frederic Briand and Joel E Cohen. Environmental correlates of food chain length. *Science*, 238(4829):956–960, 1987.

[30] Jesús A Carrillo-Castrillo, Juan Carlos Rubio-Romero, Jose Guadix, and Luis Onieva. Risk assessment of maintenance operations: The analysis of performing task and accident mechanism. *International journal of injury control and safety promotion*, 22(3):267–277, 2015.

[31] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[32] Sanmay Das. Filters, wrappers and a boosting-based hybrid for feature selection. In *Icml*, volume 1, pages 74–81, 2001.

[33] Márcio das Chagas Moura, Isis Didier Lins, Enrique López Droguett, Rodrigo Ferreira Soares, and Rodrigo Pascual. A multi-objective genetic algorithm for determining efficient risk-based inspection programs. *Reliability Engineering & System Safety*, 133:253–265, 2015.

[34] Maxx Dilley and Tanya E Boudreau. Coming to terms with vulnerability: a critique of the food security definition. *Food policy*, 26(3):229–247, 2001.

[35] Eric G Evers and Jurgen E Chardon. A swift quantitative microbiological risk assessment (sqmra) tool. *Food Control*, 21(3):319–330, 2010.

[36] Dewan Md Farid, Li Zhang, Chowdhury Mofizur Rahman, M Alamgir Hossain, and Rebecca Strachan. Hybrid decision tree and naïve bayes classifiers for multi-class classification tasks. *Expert systems with applications*, 41(4):1937–1946, 2014.

[37] Jenny Gustavsson, Christel Cederberg, Ulf Sonesson, Robert Van Otterdijk, and Alexandre Meybeck. *Global food losses and food waste*. FAO Rome, 2011.

[38] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[39] Malik Altaf Hussain and Christopher O Dawson. Economic impact of food safety outbreaks on food businesses. *Foods*, 2(4):585–589, 2013.

[40] International Organization for Standardization. ISO/IEC 31010:2009 Risk management - Risk assessment techniques. *Risk Management*, 31010:92, 2009.

[41] Bernard Kamsu-Foguem. Information structuring and risk-based inspection for the marine oil pipelines. *Applied Ocean Research*, 56:132–142, 2016.

[42] Carl Kingsford and Steven L Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–1013, 2008.

[43] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[44] Abdullah Konak, David W Coit, and Alice E Smith. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering & System Safety*, 91(9):992–1007, 2006.

[45] Damjan Krstajic, Ljubomir J Buturovic, David E Leahy, and Simon Thomas. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):1–15, 2014.

[46] Cun-bin Li and Jian-jun Wang. Model of generic project risk element transmission theory based on data mining. *Journal of Central South University of Technology*, 15(1):132–135, 2008.

[47] Michael Madaio, Shang-Tse Chen, Oliver L Haimson, Wenwen Zhang, Xiang Cheng, Matthew Hinds-Aldrich, Duen Horng Chau, and Bistra Dilkina. Firebird: Predicting fire risk and prioritizing fire inspections in atlanta. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 185–194, 2016.

[48] R Magalhães, Gonçalo Almeida, V Ferreira, I Santos, Joana Silva, MM Mendes, J Pita, G Mariano, I Mâncio, MM Sousa, et al. Cheese-related listeriosis outbreak, portugal, march 2009 to february 2012. *Eurosurveillance*, 20(17):21104, 2015.

[49] Dianhui Mao, Fan Wang, Zhihao Hao, and Haisheng Li. Credit evaluation system based on blockchain for multiple stakeholders in the food supply chain. *International journal of environmental research and public health*, 15(8):1627, 2018.

[50] Adam S Markowski and M Sam Mannan. Fuzzy risk matrix. *Journal of hazardous materials*, 159(1):152–157, 2008.

[51] PJ O'mahony. Finding horse meat in beef products—a global problem. *QJM: An International Journal of Medicine*, 106(6):595–597, 2013.

[52] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[53] Andika Rachman and RM Chandima Ratnayake. Machine learning approach for risk-based inspection screening assessment. *Reliability Engineering & System Safety*, 185:518–532, 2019.

[54] SP Rahayu, SW Purnami, and A Embong. Applying kernel logistic regression in data mining to classify credit risk. In *2008 International Symposium on Information Technology*, volume 2, pages 1–6. IEEE, 2008.

[55] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

[56] Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018.

[57] ED Van Asselt, P Sterrenburg, MY Noordam, and HJ Van der Fels-Klerx. Overview of available methods for risk based control within the european union. *Trends in food science & technology*, 23(1):51–58, 2012.

[58] Jing Wang and Huili Yue. Food safety pre-warning system based on data mining for a sustainable food supply chain. *Food Control*, 73:223–229, 2017.