Article

Machine Learning for Automatic Rule Classification of Agricultural Regulations: A Case Study in Spain

Borja Espejo-Garcia [1], Jorge Martinez-Guanter [2], Manuel Pérez-Ruiz[3,2], Francisco J. Lopez-Pellicer[1], F. Javier Zarazaga-Soria[1]

1. Computer Science and System Engineering Department, Universidad Zaragoza, María de Luna 1, 50018, Zaragoza, Spain

2. Universidad de Sevilla. Área de Ingeniería Agroforestal Dpto. de Ingeniería Aeroespacial y Mecánica de Fluidos, Spain

3. Corresponding Author's contact information:
   Ctra. Utrera, km 1
   Sevilla, 41013, Spain
   Phone: (954) 481-389
   Fax: (954) 486-436
   E-mail: manuelperez@us.es

Abstract: Currently, pest management practices require modern equipment and the use of complex information, such as regulations and guidelines. The complexity of regulations is the root cause of the emergence of automated solutions for compliance assessment by translating regulations into sets of machine-processable rules that can be run by specialized modules of farm management information systems (FMIS). However, the manual translation of rules is prohibitively costly and therefore, this translation should be carried out with the support of artificial intelligence techniques.

In this paper, we use the official Spanish phytosanitary products registry to empirically evaluate the performance of four popular machine learning algorithms in the task of correctly classifying pesticide regulations as prohibitions or obligations. Moreover, we also evaluate how to improve their performance with the preprocessing of the texts with natural language processing techniques. Finally, due to the specific characteristics of the texts found in pesticide regulations, resampling techniques are also evaluated. Experiments show that the combination of the machine learning algorithm Logic regression, the natural language technique part-of-speech tagging and the resampling technique Tomek Links is

the best performing approach with an F_1 score of 68.8%, a precision of 84.46% and a recall of 60%. Experimental results are promising and shows that this approach can be applied to develop a computer-aided tool for transforming textual pesticide regulations into machine-processable rules. To the best of our knowledge, this is the first study that evaluates the use of artificial intelligence methods for the automatic translation of agricultural regulations into machine-processable representations.

Keywords: Rule extraction, Natural language processing, Smart precision agriculture, Integrated pest management

---

1. Introduction

In modern agriculture, production is governed by a variety of standards that restrict farming practices that could be harmful (Nikkilä et al., 2012). For example, different regulations and programs, such as Integrated Pest Management (IPM), have been developed to control the use of phytosanitary products and prevent unauthorized uses (Lozano et al., 2010). IPM can be implemented as a module of a farm management information system (FMIS), which collects, exchanges and stores a large amount of exploitation data and provides decision support for tailoring farm operation to the specific demands of stakeholders (Sørensen et al., 2010). Fountas et al. (2015) extend this model by defining a complex information ecosystem established around the farm machinery named Farm Machinery Management Information System (FMMIS), based on the Soft System Methodology (SSM). The interrelations presented in this extended model, some such as GNSS positioning data, real-time crop and soil data generated by airborne or terrestrial sensors, and input consumption and inventory management databases, have been increasing studied and adopted by farmers (Miller et al., 2017). Making farmer's decision process easier is essential.

One of the FMMIS challenges is the active support during the decision-making process, which could ensure that agricultural tasks such as fertilization and spraying are conducted according to safety and quality standards. To achieve this goal, it is necessary to translate standards and regulations into a machine-readable representation, such as formal rules, that can be executed within an FMIS.

According to Nash et al. (2011) and our own experience working with the official Spanish phytosanitary products registry, agricultural regulations consist of rules that can be mainly classified as prohibitions and obligations. Thus, each of these rules can be evaluated as true or false, with the conclusion of compliance or violation of the regulation.

The manual translation of regulations into machine-processable representations is prohibitively costly in terms of time, labour and knowledge (Wyner and Governatori, 2013). Another barrier to the actual situation encompassing European farming is that most of the data and information are unstructured, fragmented and difficult to use (Fountas et al., 2015).

To avoid these bottlenecks, techniques related to artificial intelligence, such as information retrieval, natural language processing (NLP) and machine learning (ML), can be used to identify syntactical patterns in the rules and partially automate the translation of regulations into formal rules that can then be provided to the FMIS. Moreover, in recent years, some promising results have been obtained in extracting rules from regulations in several domains (e.g., Soria et al., 2005; Wyner and Peters, 2011; Maat and Winkels, 2008). In the agricultural domain, these techniques must prove that they are highly accurate because non-compliance caused by an extraction error may carry a considerable economic penalty (Davies and Hodge, 2006). Automatic rule extraction from regulations is a complex process that requires different components. One of these components should be a rule classifier that allows regulations to be categorized as prohibitions or obligations (Figure 1). This step is critical because an error implies that the meaning of the rule is inverted. For example, a rule such as "*Do not apply to crops with fruits that must be* *preserved*" could be interpreted as "*Apply to crops with fruits that must be preserved*". Moreover, this classification could facilitate the modelling conditions and rule constraints that represent the meaning of rules and retain consistency with the original text. There are different ways to build rule classifiers, but the state-of-the-art approach includes the use of ML algorithms. Moreover, these algorithms are often enriched with linguistic knowledge that is automatically extracted by using NLP techniques and improved by using preprocessing techniques such as resampling.

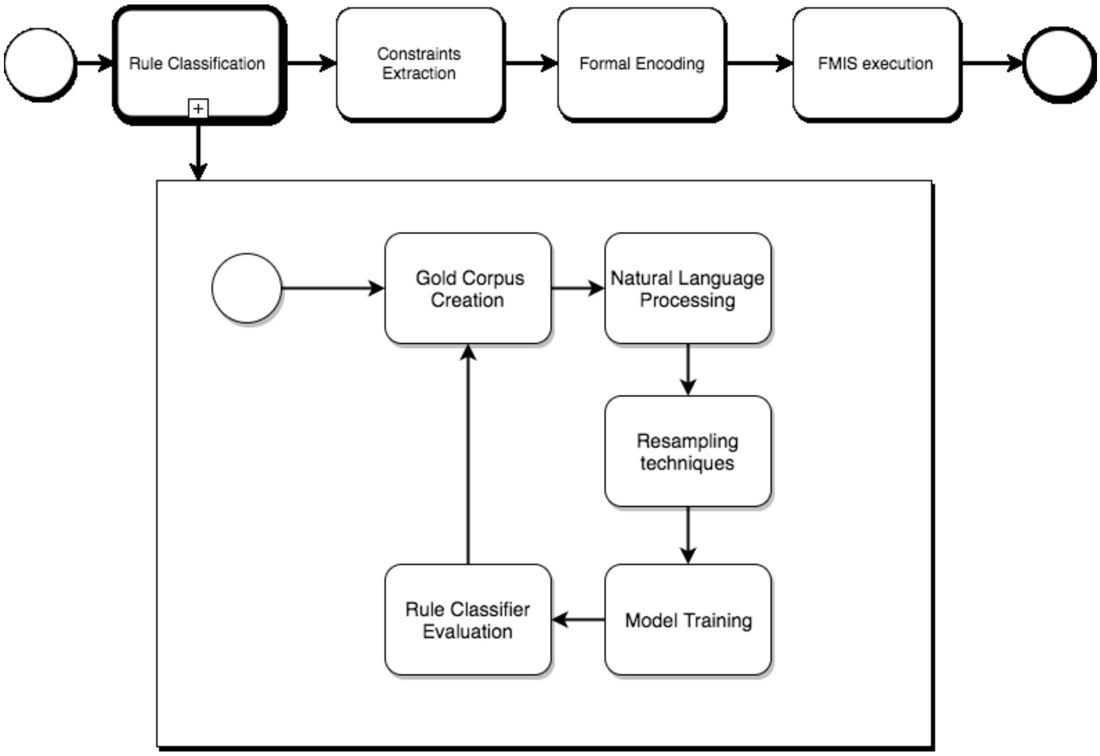


Figure 1. Processes for translating a rule into a machine-readable format. Rule classifier development is an interactive process. If a regulation changes, it may be necessary to retrain the classifier with new data. Developing a rule classifier is a complex task that requires the combination of NLP techniques, resampling methods and model training.

This work evaluates the applicability of NLP, resampling and ML techniques for building a rule classifier that can automatically discern between prohibitions and obligations in the agricultural domain using documents from the official Spanish phytosanitary product registry. *We have preprocessed these documents to extract only the parts of the text that represent the rules. Then, we have manually annotated them to create a gold corpus where the ML algorithms will find the patterns that allow the* *distinction between prohibitions and obligations. This gold corpus will also be used as a benchmark to evaluate the performance of the different techniques evaluated in this paper. It is important to note that we have created our own corpus because, as far as we know,* there is no available gold corpus focused on phytosanitary regulations in contrast *to other research domains such as spam classification or news categorization.* In this paper, we provide insights into the possibilities and limitations of existing ML, resampling and NLP techniques for usage in agriculture to support the development of decision support systems and the FMIS. Moreover, the objective of the approach presented in this paper is to provide a basis for the future automatic extraction of rules and their spatiotemporal components. As noted by Nikkilä et al. (2012), we believe that the fully automated translation of regulations is not currently feasible but building knowledge repositories and software components that gradually solve the rule translation problems, will benefit in future studies.

This article is structured as follows, Section 2 presents the materials and methods. where the complexity of the FMIS is presented and the developed methodology is detailed. Section 3 shows the results of the evaluation of the techniques analysed and Section 4 presents a discussion of its implementation in FMISs. Finally, Section 5 presents the conclusions and future directions for the integration of these techniques in modern FMISs.

2. Materials and Methods

2.1. Commercial FMIS structure and data sharing enhancements

In modern agrisystems, many devices including tractors, tractor implement, field sensors, airborne devices, etc, are used on farms. The information generated and required by these devices must be understandable to optimize collaboration efforts. To simplify the abovementioned interconnection of different farm elements and provide a unified data platform, the commercial solution Agroplanning was created. Agroplanning is a modular cloud-based FMIS that treats the tractor as a centralized connected platform for data generation and reception. The aim of the system is to incorporate the tractor-centric approach defined by Fountas et al. (2015), and equipping agricultural service companies, farmers, cooperatives and machinery manufacturers with the tools to generate the first advanced precision farming services, improve efficiency and increase the precision of agricultural management.

Information regarding the real-time position of a tractor or routes, agricultural tasks performed with an implement and decisions made by growers is not easy to obtain if these processes are not properly recorded. To permit data collection and provide a number of intelligent services, a novel hardware module, ISOBUS compliant, which provides GNSS and GPRS connectivity and up to ten I/O digital pins, has been developed. This device sends the data packages to the cloud server every 10 second and can be mounted on any agricultural vehicle. This module was created to enhance data interoperability and tailor existing systems to farmers' needs. The vehicle monitoring data are combined in the core of the Agroplanning cloud FMIS with a variety of soil, crop and climate data from wireless in-field sensor networks; and other data to improve knowledge of field conditions.

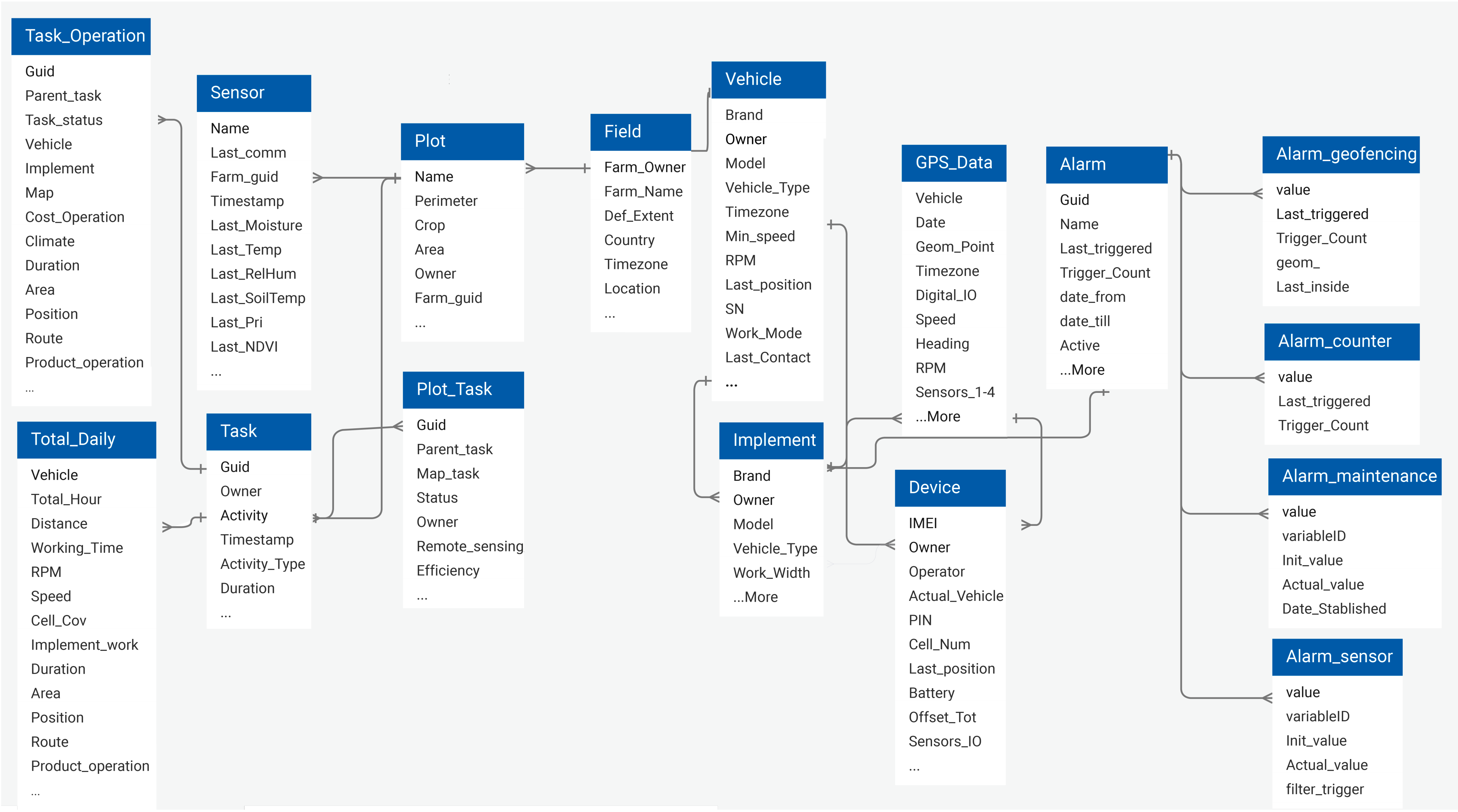
All the data is automatically transferred to a cloud platform built on Azure Web Services (Microsoft, Redmond, USA). This cloud platform uses the database systems SQLServer and NoSQL Azure Tables, were the information provided is stored into relational and non-relational databases. A comprehensive diagram of the relational entities built up on the actual commercial FMIS is provided on figure 2.

The modular architecture is reflected in the user interface, that can be seen on figure 3 below, which has been divided into interconnected blocks as follows.

- Vehicles: This block provides advanced agricultural fleet management in real time. In a visual interface with map base in OpenLayers, all information of the connected machinery, routes, implements, daily activities, alarms, etc. is shown.

- Farms: This module manages the different plots within a farm defined by the user from shapefiles or administrative information.
- FieldBook: This module combines the two previous modules to automatically generate reports of what has been done in each plot by the vehicles.
- Sensors: This module provides an analytical visualization of the growing, soil and climate conditions in each plot through wireless sensor networks.

Figure 2. Database relation diagram between entities and some example parameters on Agroplanning's commercial FMIS solution



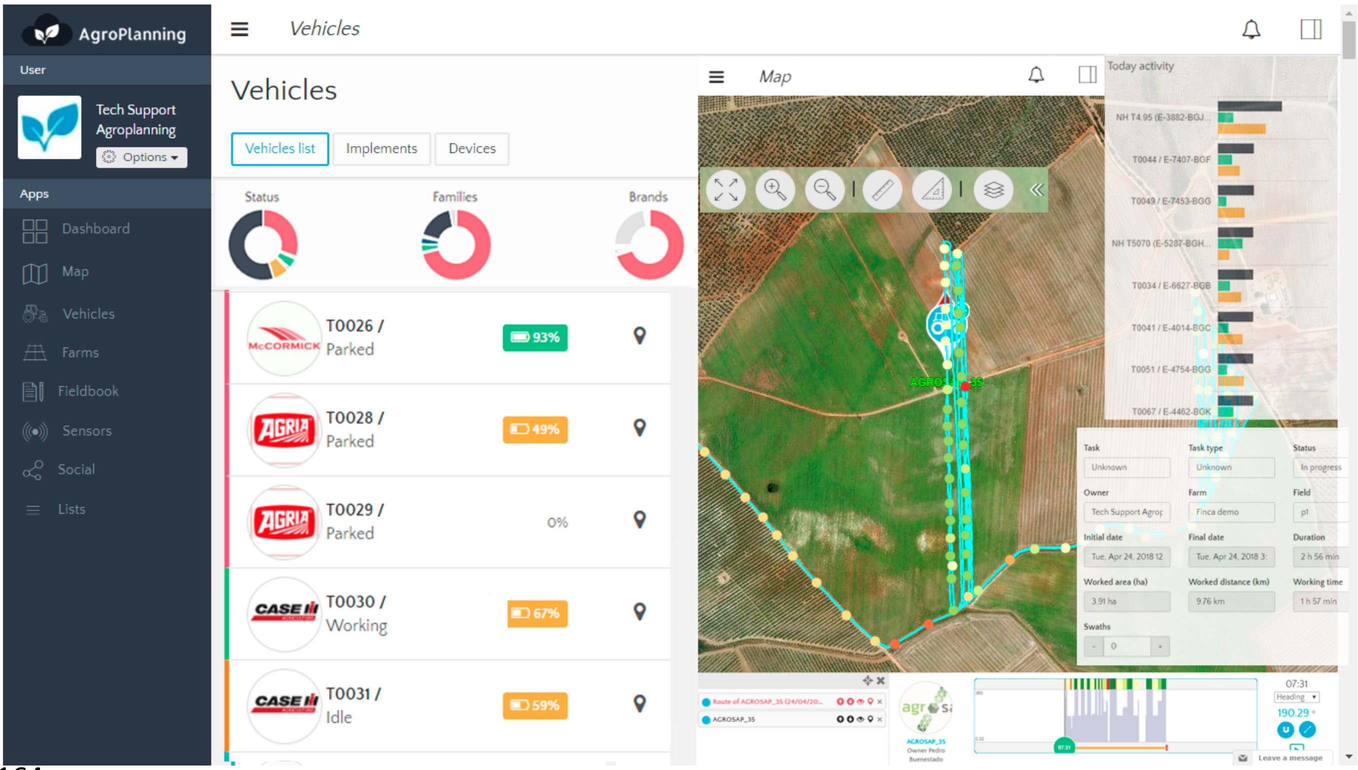


Figure 3. User interface of actual FMIS commercial solution.

The actual platform combines the different modules in such a way that all knowledge, agronomic algorithms with local components, and elements of decision support are merged to create a decision-making level that includes strategic, operational and evaluation aspects in the form of an automated report. The adoption of this integrative platform is being improved by the addition of new modules and functionalities, but its use is currently low. Specifically, approximately 20 producers are using it, mainly agricultural service providers in southwestern Spain, and the total number of connected vehicles is close to 200 (Agroplanning, Sevilla, Spain).

In addition to the actual commercial stage of the FMIS, a novel conceptual feature provides the ability to incorporate third-party actors into the system. For pesticide applications in IPM systems, national and regional administrations have various open data sources (most of them unstructured as data repositories) that include information on allowed active substances, legal application doses and safety periods.

The novelty associated with this approach of combining administrative data and connected in-field elements is the possibility of automatically generating pesticide task recommendations according to both the standards of the administration and the agronomic algorithms adapted to local conditions. In this approach, automatic prescriptions can be generated based on where a vehicle is located, the crop and variety within the exploited domain (registered in official documents). The phenological state of the crop, the actual crop needs, and what active substances are allowed in that location (along with all the other regulatory information).

In addition, these "official" prescriptions would be automatically sent to an electronic controller on the tractor or implement. On the vehicle on-board screen, the user is allowed to accept or reject the prescription. If the former happens, the user will be assured that the task will be performed according to the required safety and quality criteria. In addition, automatic task registration will ensure that this task complies with regulations and provides traceability for the performed actions. Details of the chemical amounts, frequency of tasks and pesticides used will be given through automatic reporting to both the producer and administration.

Within this conceptual framework, which can be integrated in an FMMIS to improve the decision-making capacity, links can be created among producers, companies and administrations to allow end users to make informed decision that adhere to standards regarding the use of agricultural inputs though promoting data sharing and open data access. At this point, we consider important to point out that the direct involvement of the administration in the proposed conceptual model can lead to a data privacy conflict. In the near future, and on the basis of data protection regulations, this should be resolved with their explicit consent, by means of methods of user anonymization, using as far as possible data aggregates, and even with new methods based on digital technologies such as blockchain's smart contracts between two parties (user and administration), which should include aspects such as data ownership and the generation of valuable information from them.

The goal of such a system is to increase the integration and interoperability of agricultural information and involve the administration to ensure that crop protection tasks are efficiently and cost-effectively performed and comply with all safety standards and regulations.

To achieve this target scenario, one barrier to overcome is the automatic incorporation of all the regulatory information associated with these applications. In this approach the information can be used to take actions and can be automatically incorporated into the FMIS (figure 4). This issue and the relevant details at the Spanish national level are addressed in the following section based on ML techniques and NLP.

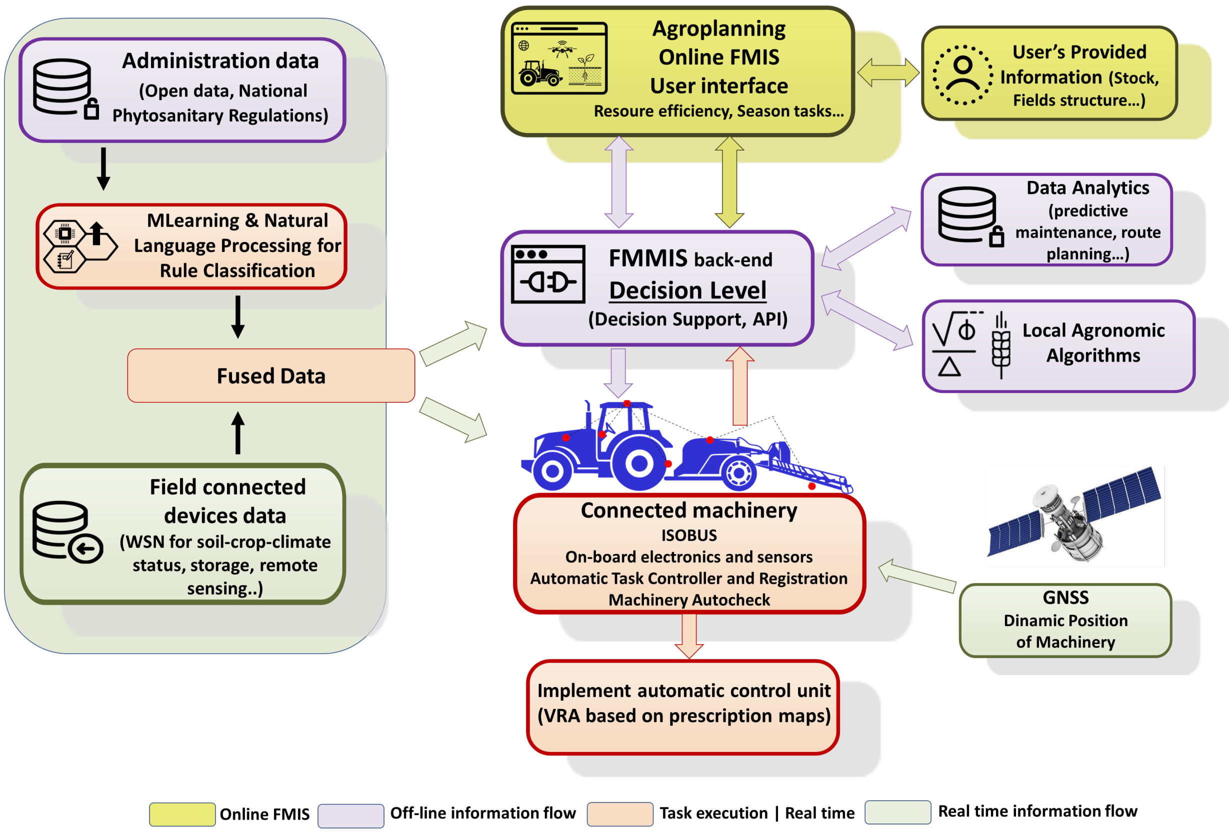


Figure 4. Complete vision of commercial stage and future enhancements on developed FMIS

2.2 Local structure of agricultural standards in Spain

Documents containing information regarding the allowed phytosanitary products and how to apply them in Spain are published in the official Spanish phytosanitary product registry, which currently contains 2,426 documents in PDF format. Figure 5 shows an example of one of these documents. The part denoted by the dark blue box shows that this document is official and published by the Government of Spain. The light blue box contains a table with information regarding how the pesticide must be applied. This table includes two parts. The first part (green box) shows the structured portion of the regulation. This part could be easily transformed into a machine-readable format using different heuristics. Here, we can find information regarding the usage, crop and dose. The second part (red box) is formatted with unstructured manual language, and its translation into a formal rule is the motivation of this research. In this part, we can find different spatiotemporal constraints that cannot be easily extracted. Each of these constraints can be categorized as an obligation or a prohibition.

Some examples of rules (translated into English) that appear in these documents with and their categorization (obligation/prohibition) are given in the following text. These rules will be used to train and evaluate the ML techniques that are the basis of the final rule classifier.

- "Apply only until flowering" (Obligation)
- "Treat from time the stalk develops until the ear emergence" (Obligation)
- "Do not apply to crops with fruits that must be preserved" (Prohibition)
- "Never apply after 10 leaves" (Prohibition)

Nº REGISTRO: 11826

APHOX

Uses and authorized doses

Use	Agent	Dose %	Method and time of application (specific conditions)
Artichoke	Fleeting	0.1	Make a single application with a maximum volume of 300 l/ha, starting from the vegetative state of 9 or more leaves not folded.

Figure 5. English translation of part of the official documents that regulate the use of pesticides in Spain.

2.3 Gold corpus creation

A gold corpus is a set of annotated texts that serves as a basis for the training and evaluation of ML algorithms. In addition, a gold corpus can be seen as the benchmark where a community research evaluates their algorithms and obtain state-of-the-art results. For example, research fields related to text classification such as spam classification or newswire categorization have their own gold corpus. Currently, to the best of our knowledge, there is no available gold corpus focused on phytosanitary regulations; therefore, we have developed our own corpus. The corpus is a monolingual Spanish corpus consisting of 2,426 PDFs collected from the official Spanish phytosanitary product registry. We manually annotated 1,135 rules in natural language as obligations or prohibitions when the text conveys such meaning related to the application of a phytosanitary product. Some examples are shown in the previous section. The corpus statistics are shown in Table 1. We believe that the is of adequate size for the evaluation of algorithms due to the small number of distinctive rules and the standardized nature of the phytosanitary vocabulary.

It is important to note that the processing of these documents is not trivial because they are published with PDF format and information extraction is subject to errors. In this part of the work, these errors have been manually fixed because our gold corpus dataset is relatively small (1,135 rules).

Table 1. Corpus statistics

Corpus Statistics

No. of rules	1,135
No. of obligations	1,119
No. of prohibitions	16
Average rule length	22 words
No. of words	25,420
No. of unique words	2,689

2.4 Natural language processing

A preprocessing step using NLP techniques is necessary to extract the most important words or groups of words from inside the rules and improve the performance of the classifier. As Collobert et al. (2000) explained, the choice of the optimal text preprocessing technique is an empirical process that is mainly based on linguistic intuition followed by trial and error. We used the following NLP techniques to improve the ML process by adding linguistic knowledge:

i. Part-of-speech (POS) tagging
ii. Stemming
iii. N-grams: unigrams and bigrams

POS tagging is the process of marking a word in a text as corresponding to a particular part of speech based on both its definition and its context (Brill, 1992). We used the Stanford POS tagger in this study (Toutanova et al., 2003). Stemming consists of removing any attached suffixes and prefixes from words because singular and plural forms of a noun or different verb forms are semantically the same in many contexts and they increase redundancy and complexity in the model. We used the Porter algorithm for stemming (Porter, 1980). N-grams attempt try to solve the problem of information loss when transforming a document into a set of independent words because sometimes word context matters. Single tokens are known as unigrams and pairs of tokens are known as bigrams. In this work, we use both types of N-grams.

Moreover, stop words and punctuation are removed by default in our evaluation. These steps remove words that are not relevant such as some articles (e.g., “the” and “a”), pronouns, etc. It is important to note that there is no single universal list of stop words, and they depend on the context. Finally, to provide a weight for each word or group of words in the corpus we use the term frequency-inverse document frequency (tf-idf) (Raschka, 2014) because it decreases the weights of words that are not relevant and not in the list of stop words.

2.5 Resampling techniques

Additional challenges come from the usage of ML techniques. It has been reported that one of these aspects is related to class imbalance, in which examples in training data associated with one class heavily outnumber the examples from other classes (Japkowicz and Stephen, 2002; Chawla et al., 2004). In our corpus, this problem arises because, as reported in Table 2, we have many more obligations than prohibitions. In this situation, the ML system may have difficulties learning the concepts related to the minority class (prohibition in our case). Despite its shortcomings, one of the procedures that has been applied in many studies is resampling (He and Garcia, 2010). Resampling is performed by oversampling or undersampling data to change the frequency of classes in the training data extracted from the gold corpus in proportion to a cost model. Resampling is only applied to the training set because the test set must be kept in its original state. In this work, we perform a broad experimental evaluation involving five different resampling methods:

i. Random oversampling (ROS),
ii. Random undersampling (RUS),
iii. SMOTE,
iv. ADASYN, and
v. Tomek Links.

In ROS, the minority class is randomly replicated to force the learning algorithm to correctly classify instances of that class, whereas RUS involves the random deletion of examples of the most frequent class to yield obtaining the opposite result. SMOTE is an advanced method of oversampling developed by Chawla et al. (2002). This approach aims to enrich the minority class boundaries by creating artificial examples in the minority class than replicating existing examples to avoid the problem of overfitting. ADASYN is another method of oversampling that was developed by Skalidis (2016). The essential concept is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data are generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn. Tomek links is a method of undersampling that searches for instances of closest neighbours that do not share the same class label (Tomek, 1976). When this relationship is identified, the Tomek link is removed from the data set, and the process is repeated until no more Tomek links can be found.

2.6 Model training

After preprocessing using the NLP techniques, we must apply different ML algorithms to obtain a rule classifier that can discriminate between prohibitions and obligations. The term ML refers to the automated detection of meaningful patterns in annotated data. The specific methods used in this paper include support vector machines, logistic regression, naive Bayes and random forests methods. The three first methods are chosen because they generate linear models that generally yield good results in high dimensional sparse problems, such as text classification, that overcome the issue of dimensionality (Bellman, 1961). A random forest method is chosen due to its effectiveness when applied to different problems, and contrary to linear classifiers, it can learn complex models that are sometimes necessary to correctly describe a classification problem. If the performance of linear and nonlinear classifiers is the same, linear classifiers are typically selected because they are simpler than nonlinear classifiers. We used the following ML algorithms in our experiments:

i. Support vector machines (SVM),
ii. Logistic regression,
iii. Naive Bayes, and
iv. Random forest (RF) methods.

SVM algorithms (Cortes and Vapnik, 1995) provide state-of-the-art text classification models because of their robustness to high dimensionality problems. An SVM model treats examples (in this work, the rules after preprocessing) as points in space, and these points are mapped so that the examples of different categories are separated by a gap that are as wide as possible. Because of the excellent results that SVM algorithms have achieved in a wide variety of domains, including in the agricultural field (Zhou et al., 2014), they have rapidly gained popularity. Logistic regression arises from the desire to model the posterior probabilities of classes (in this work, obligation and prohibition) via linear functions in the feature space (in this work, the words after preprocessing) while ensuring that the probabilities sum to one and remain in the range of [0,1] (Friedman et al., 2008). This model is also a representation of examples as points in space that are mapped as described above; however, contrary to SVM, the gaps between classes of points are as wide as possible. The naive Bayes classifier (Langley and John, 1995) is based on the popular Bayes probability theorem. It is known for creating simple yet effective linear models. For example, this approach yielded excellent results when applied for spam classification and disease prediction (Saad et al., 2012). The main difference between naive Bayes and logistic regression is that the former optimizes the joint probability and the latter optimizes the posterior probability. RF methods use decision trees (i.e., a forest) with random independently sampled vectors, and all trees in the forest have the same distribution (Breiman, 2001). They are popular algorithms in the ML community and have been recently used in the agricultural field (e.g., Brillante et al., 2015; Görgens et al., 2015).

2.7 Rule classifier evaluation

Evaluation techniques measure the correspondence between the results that the classifier generates and those of the gold standard. There is no single evaluation metric that is appropriate for all classification problems. In practice, different classification models should be compared based on a particular dataset and different metrics. Moreover, it is important to consider the high-level goal of the application: The FMIS where the rule classifier could be integrated must accurately classify the maximum number of rules to reduce the risk of prescribing the wrong pesticide or application. This goal can be evaluated with thee metrics: recall, precision and a combination of the two deemed the F_1 score.

Recall is a widely used ML metric. In our work, it is defined as the fraction of “true” prohibition rules that are effectively classified as prohibitions ($n_{pr\to pr}$). Thus, it provides a measure of the “completeness” of the system (Eq. 1). Recall decreases if the number of prohibitions misclassified as obligations ($n_{pr\to ob}$) increases. If recall is 100%, no prohibitions have been classified as obligations.

$$Recall = \frac{n_{pr\to pr}}{n_{pr\to pr}+n_{pr\to ob}} \quad (1)$$

Precision is another widely used metric and provides a measure of the “soundness” of the system. Specifically, it is the proportion of the rules correctly classified as prohibitions ($n_{pr\to pr}$) to the total number of rules classified as prohibitions ($n_{pr\to pr}$ + $n_{ob\to pr}$), as shown in Eq. 2. The precision decreases if the number of obligations misclassified as prohibitions ($n_{ob\to pr}$) increases. In this work, if the precision is lower than 100%, some obligations are classified as prohibitions and a rule such as “Apply this pesticide in the spring” could be interpreted as “Do not apply this pesticide in the spring”.

$$Precision = \frac{n_{pr\to pr}}{n_{ob\to ob}+n_{ob\to pr}} \quad (2)$$

High recall and precision values indicate good performance; however, it is important to note that there is a trade-off between optimizing recall and optimizing precision. Thus, while precision and recall are very important metrics, considering only one of them will not provide the full picture. Finally, the F_1 score combines precision and recall to provide a single metric for algorithms comparison, as shown in Eq. 3. In this work, this measure is used to identify the most balanced algorithm that is likely the best approach for categorizing rules.

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \quad (3)$$

A standard measure of classification performance is the classification accuracy. However, for datasets with skewed distributions, this measure can be misleading.

3. Results

This section shows the experimental results of the 96 different combinations achieved by evaluating 4 ML algorithms, 6 resampling methods and 4 NLP techniques to build the rule classifier. All of them are the averages of 30 runs. In each of the runs, we use stratified 10-fold cross-validation to find the best hyperparameter settings used in the ML algorithms (Table 2). This statistical technique provides good performance estimates with minimal assumptions and makes results less prone to random variation. The main disadvantage of cross-validation is the associated increased computational cost, but in this phase of the research, it is more important to obtain accurate estimates. It is important to note that optimal hyperparameter settings often differ for different datasets. Therefore, they should be tuned for each dataset.

Table 2: Parameter specification for the algorithms

Algorithm	Parameters
Naive Bayes	-
SVM	Kernel = Linear
	C = 10
	Tolerance = 0.001
	Shrinking = true
Random forest	Estimators = 20
	Pruned = false
	Impurity = Gini
Logistic regression	Penalty =12
	C= 10

Tolerance = 0.0001

The ML algorithms were implemented based on the method published by scikit-learn (Pedregosa et al., 2011), which is one of the best known and most widely used ML libraries. This package, which is written in Python, includes the implementation of many popular ML algorithms and has preprocessing and evaluation capabilities. The version of scikit-learn used in this work is 0.19.1. We investigate the learning algorithms in combination with different NLP and resampling techniques to find the combination that allows the most accurate rule classification for prohibitions and obligations. Many algorithms and NLP techniques exist that are beyond the scope of this work, but in future experiments they should be studied to potentially identify better approaches. In Table 3, we can observe the top 10 combinations of NLP, resampling and ML techniques that yielded the best precision in recognizing prohibitions. These combinations minimized the false positive error ($n_{ob\rightarrow pr}$), i.e., the number of obligations classified as prohibitions. Conversely, they exhibit low recall values, which means that some prohibitions are "lost" and incorrectly classified as obligations $n_{pr\rightarrow ob}$. POS tagging is the best technique for achieving high precision. Otherwise, more diversity is provided by other resampling techniques and ML algorithms. Logistic regression could potentially be considered the best approach because the top results use this algorithm. Table 4 shows the top 10 combinations of NLP, resampling and ML techniques that yield the best recall in recognizing prohibitions. These combinations minimize the false negative error ($n_{pr\rightarrow ob}$), i.e., the number of prohibitions classified as obligations.

Table 3: Summary of the algorithms with the best precision in prohibition classification

NLP	Resampling	Algorithm	Precision (%)	Recall (%)
POS	None	Logistic	85.00	58.57
POS	Tomek links	Logistic	84.46	60.00
POS	ROS	RF	84.04	47.85
POS	Tomek links	RF	81.54	40.00
POS	None	RF	78.72	34.28
POS	SMOTE	RF	75.73	50.71
POS	ADASYN	RF	74.25	47.14
Bigrams	SMOTE	SVM	72.12	61.42
POS	None	SVM	67.72	52.14
Bigrams	ROS	SVM	67.15	70.00

However, these methods yield low precision, which means that many obligations are "lost" and classified as prohibitions ($n_{pr \rightarrow ob}$). The stemming and unigrams methods are the specific NLP techniques that produce the best recall performance. It is important to note, that best results are always achieved by resampling techniques, specifically, oversampling techniques. This finding is expected because resampling techniques are implemented to improve the ability of ML algorithms to recognize prohibitions. The problem with these approaches is that because there are so many obligations, if an algorithm is biased in classifying rules as prohibitions, precision can significantly decrease (the best precision is 23.09%).

Finally, Table 5 shows the top 10 combinations of NLP, resampling and ML techniques that exhibit the best F_1 values for recognizing prohibitions. These results represent the most balanced approach. Thus, if we have no preference regarding the type of error and misclassifying obligations and prohibitions is equally important, this combination should be chosen. The most balanced combination yielded a 68.08% F_1 score and included POS tagging, Tomek links and logistic regression.

The remainders of the results suggest that POS tagging is implemented in the top three methods, and logistic regression is used to achieve the top two results.

To confirm that best performing result is not due to chance, we conducted a statistical significance test using the second-best (POS tagging, ROS and logistic regression). The test was performed using Welch's t-test (Welch, 1951) with a confidence level of 0.01.

Table 4: Summary of the algorithms with best prohibition recall

NLP	Resampling	Algorithm	Recall (%)	Precision (%)
Unigrams	ROS	Logistic	100	23.09
Stemming	ROS	Logistic	100	21.04
Unigrams	ADASYN	Logistic	100	20.71
Unigrams	SMOTE	Logistic	100	20.58
Stemming	ADASYN	Logistic	100	19.83
Stemming	SMOTE	Logistic	100	19.67
Unigrams	RUS	Logistic	100	6.85
Stemming	RUS	Logistic	100	6.55
Unigrams	RUS	Bayes	100	5.63
Stemming	ROS	Bayes	100	5.54

According to the test, statistical significance exists between the approaches; therefore, we can confirm that the correct selection of NLP, resampling and ML algorithms is important for developing the most accurate rule classifier.

It is also important to note that in the three tables, logistic regression is the best ML algorithm. The rationale behind these results is that simple linear models can obtain good results in combination with different resampling and NLP techniques. To determine which are the techniques that work best for rule classification, we visualize the results after aggregating all the F_1 values for all NLP, resampling and ML techniques.

Figure 6 shows a comparison of NLP techniques without considering the rest of the classification components. Notably, POS tagging exhibits the best performance. The rest of the NLP techniques yield similar results; therefore, we can infer that stemming and bigrams have little influence on the F_1 score.

In Figure 7, the behaviours of the different resampling techniques used during the experiments can be observed. ROS exhibits the most stability, although in some experiments, it yields poor results. The other oversampling techniques (ADASYN and SMOTE) displayed similar behaviours but poor performance. Undersampling techniques exhibited the worst overall performance. However, it is important to note that in particular cases, undersampling can produce high performance, such as in the case of Tomek links in combination with POS tagging and logistic regression.

Table 5: Summary of the algorithms with the best F_1 score

NLP	Resampling	Algorithm	F1 (%)
POS	Tomek links	Logistic	68.08
POS	ROS	Logistic	67.72
POS	None	RF	67.04
Bigrams	ROS	RF	66.64
Unigrams	ROS	RF	66.54
POS	SMOTE	RF	65.91
Bigrams	SMOTE	RF	63.41
Stemming	ROS	SVM	60.53
POS	ROS	SVM	58.55
POS	SMOTE	SVM	57.39

Finally, Figure 8 illustrates the performance of the different ML algorithms. SVM and logistic regression exhibit the best performance in general. In the case of logistic regression, this result was expected after reviewing the previous results. On the other hand, SVM exhibits good performance but never produces excellent results. Thus, we can state that SVM is a robust approach that should be studied further in the future to determine if it can yield results similar to those of logistic regression.

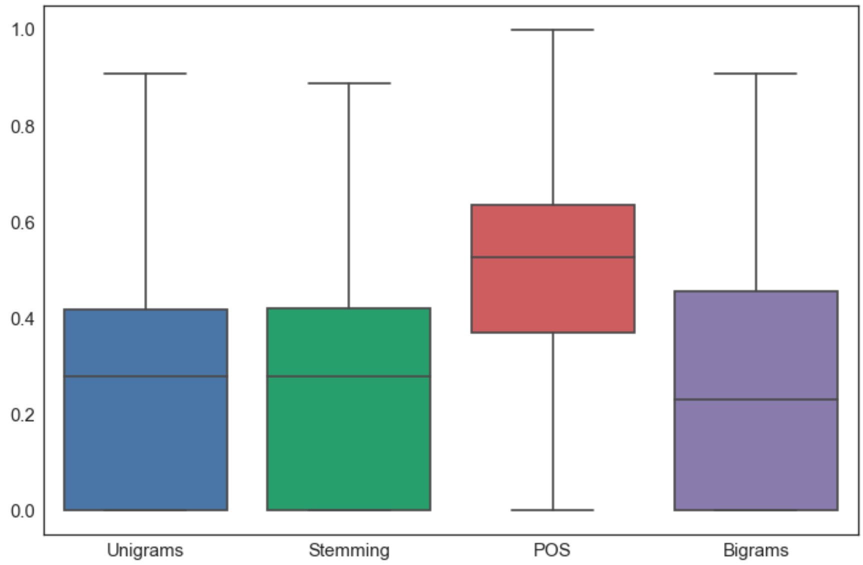


Figure 6. Performance comparison of the different NLP techniques used. While POS tagging achieves the highest F_1 score, the others obtain a similar lower performance.

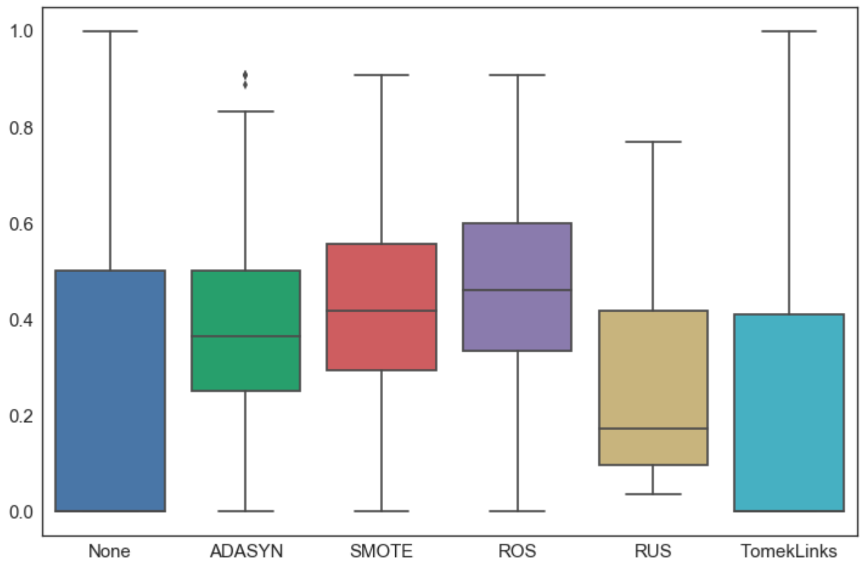


Figure 7. Performance comparison of the different resampling techniques evaluated. Over-sampling techniques such as ROS, ADASYN and SMOTE achieve the highest performances.

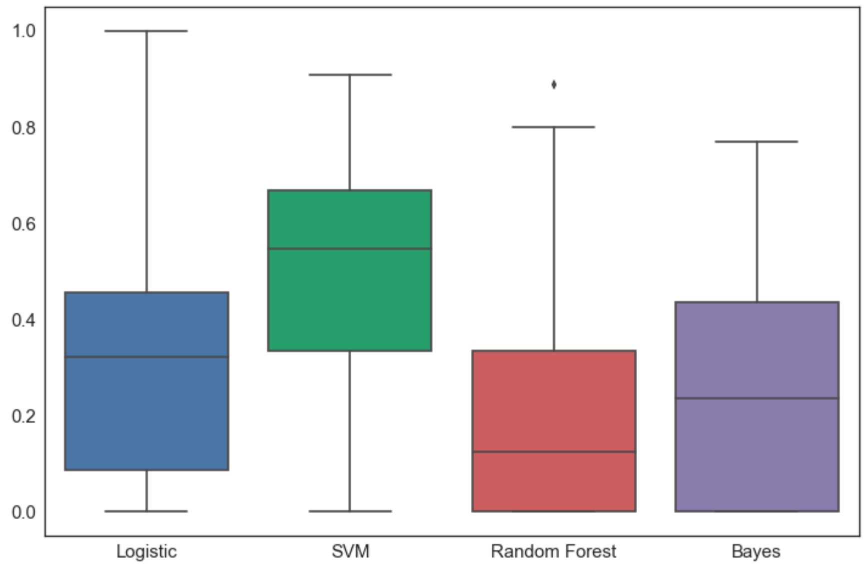


Figure 8. Performance comparison of the different machine learning algorithms evaluated. SVM achieves the highest F_1 score.

4. Discussion

Since health-related risks due to the provision of an incorrect rule are possible, it is critical that the phytosanitary rule classifier provides information to the FMIS with the maximum potential accuracy. The best approach identified in our experiments is a rule classifier that combines POS tagging, Tomek links and Logistic regression. The method yielded and F1 score of 68.8%, precision of 84.46% and recall of 60%. Although the ideal result would be 100% for all three metrics, this is unrealistic, and the literature no real automatic system can achieve this level of functionality. A human annotator could achieve this performance, but due to the abundance of regulations, it would be difficult to consider all the information that an automatic system could process. In addition, based on the automatic extraction of rules, the information provided by the FMIS would rarely be outdated. Although the idea of using artificial intelligence techniques is to bound and optimize human intervention, due to the dynamics of agricultural production, the feedback provided by humans to retrain an old rule classifier with more information is an important part of the system. Moreover, as Nash et al. (2011) noted, until new algorithms and approaches are researched, the original text of the rule must be provided to the farmer, and if the automatic translation is not working correctly, a report with the detected problems could be generated. This could be seen as a Human in the Loop (HIL) DSS (Pinto et al., 2015). It is also important to note that machine learning models make a stationary assumption, but this is not true in practice. This means that the distribution of the data will drift from what the model was originally trained upon. Distribution drift invalidates the model and, therefore, it needs to be updated.

In addition, this approach could be used as a computer-aided tool that human annotators could use to translate regulations into a formal semantic representation that could be executed within the FMIS. Therefore, this system could be seen as part of a semiautomatic rule extraction framework with an increased automation role based on inputs from future NLP, resampling and ML advances. However, although there are multiple language constructs for each sentence type, these methods are limited. Perhaps, some heuristic or post-processing methods could improve the performance of such algorithms. However, we prefer to use only ML and NLP for automatic rule translation. Finally, we agree with Nash et al. (2011) that obligations and prohibitions are good starting points for transforming rules into a machine-readable format and next step should include the extraction of information contained within the rules that represents the actions that are required or prohibited. To achieve this goal, it would be necessary to extend this approach by using external knowledge to model more complex rules. This knowledge could be based on different agricultural ontologies such as crop taxonomies proposed through open data initiatives and standards (Charvat et al., 2014). Moreover, if we add complexity to the model, the classifier should consider parts of the text whose category is not clear, and therefore, to request for human expert decision. This human expert could discard the rule because it does not contain relevant information for a specific FMIS requirement. In addition, new concepts related to law formalization such as permission, penalty and definition could be used to model new parts of the phytosanitary regulations.

5. Conclusions

In this article, we have evaluated whether it is possible to use ML techniques in combination with NLP and resampling techniques to classify rules involving prohibitions and obligations and, consequently, the applicability of these techniques in a module that can be integrated within an FMIS that supports decision making based on regulations and production standards. To the best of our knowledge, this is the first attempt to combine different automatic rule classification approaches in the agricultural domain. The best approach found in our experiments was the combination of POS tagging, Tomek links and Logistic regression. This combination yielded an F_1 score of 68.8% a precision of 84.46% and a recall of 60%. Thus, it provides promising results that will be improved with advances in ML and NLP research. The rule classifier obtained can be used as a computer-aided tool that human annotators can use to translate regulations into a formal language that could be executed within the FMIS.

Future research will use different algorithms and NLP techniques. Moreover, by introducing new techniques for information extraction, the spatiotemporal constraints could be automatically extracted and integrated within the FMIS. Therefore, an end-to-end system would be operative and regulations written in natural language could be automatically translated into machine-readable formats.

6. Acknowledgment

This work has been partially supported by the Government of Aragon and the European Social Fund. This work has also been partially supported by the Spanish Government (projects RTC-2016-4790-2 and TIN2017-88002-R).

7. References

1. Bellman, R., 1961. Adaptive control processes: A guided tour. Princeton University Press 28, 1–19. URL http://arxiv.org/abs/1302.6677
2. Breiman, L., 2001. Random Forests. Machine learning 45.1, 5–32. Brill, E., 1992. Rule-Based Part of Speech. In: Proceedings of the third conference on Applied natural language. pp. 152–155.

3. Brillante, L., Gaiotti, F., Lovat, L., Vincenzi, S., Giacosa, S., Torchio, F., Segade, S. R., Rolle, L., Tomasi, D., 2015. Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physicalmechanical characteristics in wine grapes. Computers and Electronics in Agriculture 117, 186–193. URL http://dx.doi.org/10.1016/j. compag.2015.07.017

4. Charvat, K., Esbri, M. A., Mayer, W., Charvat, K., Campos, A., Palma, R., Krivanek, Z., 2014. FOODIE - Open data for agriculture. 2014 IST-Africa Conference Proceedings (MAY 2014), 1–9.

5. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357. URL http://www.jair.org/papers/ paper953.html

6. Chawla, N. V., Japkowicz, N., Drive, P., 2004. Editorial: Special Issue on Learning from Imbalanced Data Sets Aleksander Kolcz. ACM SIGKDD Explorations Newsletter 6 (1), 1–6.

7. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P., 2000. Natural Language Processing (almost) from Scratch. Journal of Machine Learning Research 1, 1–48.

8. Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Machine Learning 20 (3), 273– 297.

9. Davies, B. B., Hodge, I. D., 2006. Farmer's preferences for environmental policy instruments: determining the aceptability of cross compliance for biodiversity benefits. Journal of Agricultural Economics 57 (3), 393–414.

10. Fountas, S., Carli, G., Sørensen, C., Tsiropoulos, Z., Cavalaris, C., Vatsanidou, A., Liakos, B., Canavari, M., Wiebensohn, J., Tisserye, B., 2015. Farm management information systems: Current situation and future perspectives. Computers and Electronics in Agriculture 115, 40–50.

11. Friedman, J., Hastie, T., Tibshirani, R., 2008. The Elements of Statistical Learning. Elements. URL http://www-stat.stanford.edu/ {~}tibs/book/preface.ps

12. Görgens, E. B., Montaghi, A., Rodriguez, L. C. E., 2015. A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics. Computers and Electronics in Agriculture 116, 221–227.

13. He, H., Garcia, E. A., 2010. Learning from Imbalanced Data Sets. IEEE Transactions on knowledge and data engineering 21 (9), 1263–1264. URL http://www.aaai.org/ Papers/Workshops/2000/WS-00-05/ WS00-05-003.pdf

14. Japkowicz, N., Stephen, S., 2002. The class imbalance problem: A systematic study. Intelligent Data Analysis Journal 6, 429–450.

15. Langley, P., John, G. H., 1995. Estimating continuous distributions in Bayesian classifier. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Montreal, Quebec, pp. 399–406.

16. Lozano, R., Diánez, F., Camacho, F., 2010. Evolution of the phytosanitary control system in the intensive horticulture model of high yield in Almería (2005-2008). Journal of Food, Agriculture and Environment 8 (2), 330–338.

17. Maat, E. D., Winkels, R., 2008. Automatic Classification of Sentences in Dutch Laws. Proceedings of the 2008 conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference 6036, 207–216.

18. Nash, E., Wiebensohn, J., Nikkilä, R., Vatsanidou, A., Fountas, S., Bill, R., 2011. Towards automated compliance checking based on a formal representation of agricultural production standards. Computers and Electronics in Agriculture 78 (1), 28–37.

19. Nikkilä, R., Wiebensohn, J., Nash, E., Seilonen, I., Koskinen, K., 2012. A service infrastructure for the representation, discovery, distribution and evaluation of agricultural production standards for automated compliance control. Computers and Electronics in Agriculture 80, 80–88.

20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

21. Pinto, R., Tobias, M., Taisch, M, 2013. Managing supplier delivery reliability risk under limited information: Foundations for a human-in-the-loop DSS. *Decision Support Systems, 54*(2), 1076-1084.

22. Porter, M., 1980. An algorithm for suffix stripping. Program, 130–137.

23. Raschka, S., 2014. Naive Bayes and Text Classification I - Introduction and Theory, 20. URL http://arxiv.org/abs/1410.5329

24. Saad, O., Darwish, A., Faraj, R., 2012. A survey of machine learning techniques for Spam filtering 12 (2), 66–73. Skalidis, S., 2016. Adaptive Synthetic Sampling Approach for Imbalanced Learning (3), 1322–1328.

25. Sørensen, C.G., Fountas, S., Nash, E., Pesonen, L., Bochtis, D., Pedersen, S.M., Basso, B., Blackmore, S.B., 2010. Conceptual model of a future farm management information system. Computers and Electronics in Agriculture. 72, 37–47. doi:10.1016/j.compag.2010.02.003

26. Soria, C., Bartolini, R., Lenci, A., Pirrelli, V., 2005. Automatic extraction of semantics in law documents. Tomek, I., 1976. Two Modification of CNN. IEEE Transactions on Systems, Man, and Cybernetics 6 (11), 769–772.

27. Toutanova, K., Klein, D., Manning, C. D., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03 1, 173– 180. URL http://portal.acm.org/citation. cfm?doid=1073445.1073478

28. Welch, B. L., 1951. On the Comparison of Several Mean Values: An Alternative Approach. URL http://www.jstor.org/stable/2332579?origin=crossref

29. Wyner, A., Governatori, G., 2013. A study on translating regulatory rules from natural language to defeasible logic. CEUR Workshop Proceedings 1004.

30. Wyner, A., Peters, W., 2011. On rule extraction from regulations. Frontiers in Artificial Intelligence and Applications 235, 113–122.

31. Zhou, R., Kaneko, S., Tanaka, F., Kayamori, M., Shimizu, M., 2014. Image-based field monitoring of Cercospora leaf spot in sugar beet by robust template matching and pattern recognition. Computers and Electronics in Agriculture 108, 58–70. URL http://dx.doi.org/10.1016/j. compag.2015.05.020