sciendo

VOLUME 17, N° 1 2023
Journal of Automation, Mobile Robotics and Intelligent Systems

Feature Selection for the Low Industrial Yield of Cane Sugar Production Based on Rule Learning Algorithms

Submitted: 10th August 2022; accepted: 19th October 2022

Yohan Gil Rodríguez, Raisa Socorro Llanes, Alejandro Rosete, Lisandra Bravo Ilisástigui

DOI: 10.14313/JAMRIS/1-2023/2

Abstract:
This article presents a model based on machine learning for the selection of the characteristics that most influence the low industrial yield of cane sugar production in Cuba. The set of data used in this work corresponds to a period of ten years of sugar harvests from 2010 to 2019. A process of understanding the business and of understanding and preparing the data is carried out. The accuracy of six rule learning algorithms is evaluated: CONJUNCTIVERULE, DECISIONTABLE, RIDOR, FURIA, PART and JRIP. The results obtained allow us to identify: R417, R379, R378, R419a, R410, R613, R1427 and R380, as the indicators that most influence low industrial performance.

Keywords: *Feature selection, Rule learning, Data mining, CRISP-DM, Industrial yield.*

1. Introduction

The increase in the volume and variety of information that is computerized in digital databases and other sources has grown significantly in recent decades. Much of this information is historical, that is, it represents transactions or situations that have occurred. Apart from its function of organizational memory, historical information is useful for explaining the past, understanding the present, and predicting future information. Most of the decisions of companies, organizations, and institutions are also based on information about past experiences extracted from very diverse sources. In addition, since the data can come from different sources and may belong to different domains, the imminent need to analyze them to obtain useful information for the organization seems clear [17].

In many situations, the traditional method of turning data into knowledge involves manual analysis and interpretation. This way of acting is slow, expensive, and subjective. In fact, manual analysis is impracticable in domains where the volume of data is growing: the enormous abundance of data overwhelms the human capacity to understand it without the help of powerful tools. Consequently, many important decisions are made, not on the basis of the large amount of data available, but rather following the user's own intuition as they do not have the necessary tools. This is the main task of data mining: to solve problems by analyzing the data present in the databases [17].

In the Cuban sugar industry there is a large database that needs to be used effectively to guide productive development towards more profitable scenarios. The correct use of this information would help decision-making with objective bases. The Cuban sugar sector needs to implement methods that allow people to quantify with greater precision the influence of the technological variables of the process on industrial performance. It is necessary to foresee the behavior of its production process in order to plan and optimize the use of technical, human, and financial resources to improve those technological variables that have the greatest weight on industrial performance [27].

At present, as an important step in data preprocessing, feature selection has become a popular research direction [29]. It also allows one to remove redundant/irrelevant features and keep some important features in the data. In view of this, we can improve the classification accuracy and speed up the model building procedure [15].

In this work, the characteristics of the process that influence the low industrial performance are determined using data mining techniques. The CRISPDM methodology and the KNIME tool are used for the development of this research.

The article is structured in five sections that are described below. Related works are reviewed in Section 2. In Section 3, we carry out an understanding of the business, analyze and prepare the data used, as well as the details of the proposed methods. Then, we carry out the modeling and discussion in Section 4. Finally, the conclusions appear in Section 5.

2. Related Works

As a result of the bibliographic study carried out, there are some works on feature determination and the use of prediction techniques that are related to ours. Among these works are the following:

In [6] the authors explain that machine learning techniques benefit performance models. They applied protocols through the entire model development process: splitting data for expected sets, feature selection, cross-validation for model fitting, and model evaluation. They used three different machine

learning techniques to create models in each protocol: Regression Trees (BRT), Random Forest (RF), and Support Vector Regression (SVR).

In [16] the authors explain the hierarchical importance of the factors that influence the yield of sugarcane. They use three different machine learning techniques: Random Forest (RF); Boosting and Support Vector Machine (SVM), for which they initially propose to identify and order the main variables that condition the yield of sugarcane, according to their relative importance.

On the other hand, in [5] the authors propose that Random Forests (RF) can cope with the generation of a prediction model when the search space of predictor variables is large, because there are many different combinations of climatic, seasonal variables, climate prediction indices, and crop model outputs that could be useful in explaining the size of the sugarcane crop.

In [31] the authors use the C4.5 algorithm to find out the climatic parameter that most influences the yield of selected crops in districts of Madhya Pradesh.

In [23] the authors identify the most important risk factors from a highly dimensional data set that helps in the accurate classification of heart diseases with fewer complications. The identification of the most relevant medical features aids in the prediction of heart disease using a filter-based feature selection technique. Different ML classification models such as Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), Multi Layer Perceptron (MLP), are used in the data sets to identify the models. suitable for the problem.

In [18] the authors propose a feature selection algorithm based on association rules and an integrated classification algorithm based on random equilibrium sampling. The experimental results show that the association rule-based feature selection algorithm is better than the CART, ReliefF, and RFE-SVM algorithms in terms of classification accuracy and feature dimension. The proposed integrated classification algorithm based on random equalization sampling is superior to the comparative SMOTE-Boost and SMOTE-RF algorithms in macro accuracy, full macro speed, and macro F1 value, representing the robustness of the algorithm.

In [34] the authors propose an improved filter function selection method to select effective functions to predict the listing statuses of Chinese-listed companies. Models based on C4.5 and C5.0 decision trees are employed and compared with several other widely used models. To assess the robustness of the models over time, the models are also tested under moving time windows. The empirical results demonstrate the efficacy of the proposed feature selection method and the C5.0 decision tree model.

In [30] the authors present a novel oil spill feature selection and classification technique, based on a forest of decision trees. The work seeks the minimization of the input features used and, at the same time, the maximization of the general test classification accuracy. Examination of the robustness of the above result showed that the proposed combination achieved higher classification accuracy than other well-known statistical separation indices. Furthermore, comparisons with previous findings converge on classification accuracy (up to 84.5%) and number of features selected, but differ on actual features. This observation leads to the conclusion that there is no single optimal combination of characteristics.

In [4] the authors state that the determination of the quality and authenticity of food and the detection of adulterations are problems of growing importance in food chemistry. The objective of this study was to consider parameters that contribute to the differentiation of the beer according to its degree of quality. Chemical (e.g., pH, acidity, dry matter, alcohol content, CO2 content) and sensory feature (e.g., bitter taste, color) were determined in 70 beer samples and used as variables in decision tree techniques. These pattern recognition techniques applied to the data set allowed us to extract useful information to obtain a satisfactory classification of the beer samples according to their quality grade.

In [2] the authors state that the inductive learning of a fuzzy rule-based classification system (FRBCS) is hampered by the presence of a large number of features that increase the dimensionality of the problem to be solved. The difficulty comes from the exponential growth of the fuzzy rule search space with the increase in the number of features considered in the learning process. In this work, we present a genetic feature selection process that can be integrated into a multistage genetic learning method to obtain, more efficiently, FRBCS composed of a set of comprehensible fuzzy rules with high classification capacity. The proposed process fixes, a priori, the number of selected characteristics and, therefore, the size of the candidate fuzzy rule search space. The experimentation carried out, using the Sonar example base, shows a significant improvement in the simplicity, accuracy, and efficiency achieved by adding the proposed feature selection processes to the multistage genetic learning method or to other learning methods.

According to [19], in many systems, such as fuzzy neural networks, language labels (such as large, medium, small, etc.) are often adopted to split the original function into several fuzzy functions. To reduce the computational complexity of the system after feature fuzzification, the optimal fuzzy feature subset should be selected. They propose a new heuristic algorithm, where the criterion is based on the min-max learning rule and a fuzzy extension matrix is designed as a search strategy.

In [26] the authors propose a new feature selection method based on the bee colony and gradient boosting decision tree with the aim of addressing issues such as efficiency and informative quality of selected features. This method achieves global optimization of the decision tree inputs using the bee colony algorithm to identify informative features.

According to [33], to improve the accuracy of the classification, a preprocessing step is used to prefilter some redundant data or irrelevant features before the construction of the decision tree. The authors

propose a new decision tree algorithm based on feature weight. The experimental results show that the proposed method performs better for the measures of precision, recall, and F1 score. Furthermore, it can reduce the time required in the construction of the decision tree.

For their part, the authors in [7] state that rough sets have proven to be effective in developing machine learning techniques, including methods for discovering classification rules. In this work, they present an algorithm to generate classification rules based on similarity relationships, which allows it to be applicable in cases where the traits have a discrete or continuous domain. The experimental results show a satisfactory performance compared to other algorithms such as C4.5 and MODLEM.

In [10] the authors state that existing rule-based classification algorithms tend to generate a number of rules with a large number of conditions in the antecedent part. However, these algorithms fail to demonstrate high predictive accuracy while balancing coverage and simplicity. Therefore, it becomes a challenging task for researchers to generate an optimal rule set with high predictive accuracy. They propose a biogeography-based optimization (BBO) method. The performance of the proposed algorithm is compared with a variety of rule miners such as OneR, PART, JRip, Decision Table, Conjunctive Rule, J48, Random Tree, among others.

For their part, the authors in [24] develop two hybrid machine learning models, AdaBoost-DT and Bagging-DT based on Decision Table as a classifier for evaluating and mapping of susceptibility of flood risks for Quang Nam.

The authors of [12] state that the Fuzzy Unordered Rules Induction Algorithm (FURIA) is a recent algorithm, proposed by Huhn and Hullermeier, responsible for creating fuzzy logic rules from a given database, and for classifying it using the generated rules. In this work they intend to analyze the effectiveness of FURY as a classification method applied in different contexts. It was found that for databases with a greater number of instances, quantitative or qualitative, this algorithm presented better performance; and in most cases resulted in a good coefficient agreement.

Based on the previously revised literature, the most used feature selection method is rule-based and the most used classification algorithm is decision trees followed by random forests.

3. Data and Methods

3.1. Data

Business Understanding Currently, given the large amount of data that is collected and stored in the harvest database, traditional data management tools and statistical tools are not adequate to extract useful, understandable, and previously unknown knowledge, that is why it is necessary to apply data mining techniques to the historical records of the sugar harvest.

The computerization of the processes of the sugar industry generates abundant data. At present, the application of the programs of the existing agro-industrial platform in the AZCUBA[1] group has guaranteed the speed and quality of the harvest information. The platform is made up of several systems, including the IPlus[2] system. This is the group's harvest information system that enables the connection of the operational results of the agro-industrial process. It is displayed at different management levels. The influence that some technological variables have on industrial performance is known, either by empirical knowledge or by scientific research, such as that of [27], where the annual values of previously selected technological variables in a three-year harvest period are analyzed to predict industrial performance.

At present, it is necessary to know, based on the historical behavior of the production process, interesting relationships between the technological variables that have greater weight in the low industrial performance. From the analysis of the historical information, solid rules, unknown or as confirmation of the relationships currently used, will be identified.

Understanding Data The data of the historical behavior of the production process of the sugar harvest used in this work are provided by the AZCUBA Group's Information Technology, Communications and Analysis Department. There is a database (MS-SQL Server) for each year, corresponding to the period from 2010 to 2019. The databases have the following dimensions:

- **Number of Records:** The number of records is on average more than 4 million. The database with the fewest number of records is from 2011 with 2,369,119 records and the one with the most is from 2019 with 6,652,282 records (Fig. 1).

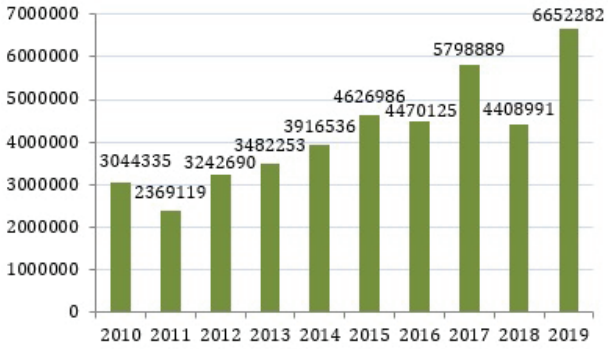

Figure 1. Number of records per year

- **Number of Indicators:** The number of indicators managed by the system is 3,605 on average, but only 578 on average are stored in the records in each database. The database that has the fewest number of indicators is that of the year 2010 with 518 indicators and the one that has the most is that of 2019 with 676 indicators (Fig. 2).

An initial exploration of the available data sources is carried out, where interesting information is revealed about the behavior of the indicators of the sugar harvests in the country. Some of the problems detected are the following:

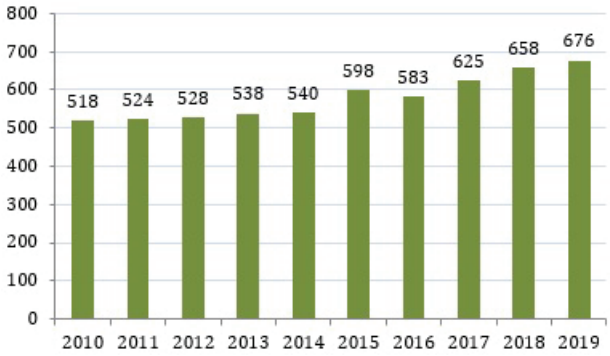

Figure 2. Number of indicators per year

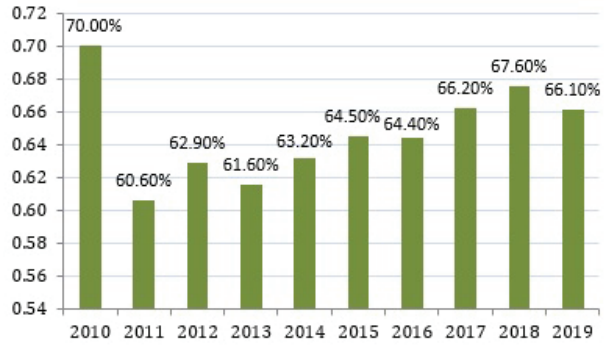

Figure 3. Percentage of records at zero per year

- The indicators increase over the years, which implies that the first data warehouses have fewer indicators than the last ones. This does not constitute an inconsistency in the coding since they adapt to the needs as time goes by. These indicators vary by addition or deletion from one year to the next.
- The 65.03% of transactional records have zero value. This can be interpreted in some specific cases as a real value, but most of them are data from unmanaged indicators. The zero data is related to the configuration of each sugar mill. The database with the least amount of zero transactional records is the one from 2011 with 60.60% of the records and the one with the most is from 2010 with 70% of the records (Fig. 3).

Data Preparation The selection of the attributes or characteristics of interest for the current investigation is carried out in the database where all the information regarding the values of the indicators analyzed in the sugar harvest is stored. The Indust_Daily_Indicator table attributes are very useful, the attributes are detailed in Table 1.

Transformations are made to the original transactional data set, obtained by means of SQL query, with a view to obtaining the mineable view.

- From the ID_INDICATOR attribute and its value contained in the VALUE_DAY attribute, a new attribute is generated with numerical values for each different ID_INDICATOR. This new attribute will be named according to the indicator's description. These new attributes are generated from the process of transposing rows into columns. It is done for all the indicators, generating attributes with the form i10, i11, R325a, R42a, etc. This process is carried out in each data source, to carry out this action it is necessary to use the Pivoting Node – KNIME. Transposing rows into columns transforms transactional records into mineable records.
- Several data sources are available, corresponding to one for each year of the sugar harvest. The basic addition method is used to integrate two or more data sets with similar attributes, but different records. Applying the KNIME tool, a workflow is designed where a Concatenate Node - KNIME is used.
- From the value of the indicator "Yield_Reported", the categorical attribute Performance Evaluation (EVAL_LOWYIELD) is generated, which can take the following ordinary values:
  - **Low:** For R295 < 10, assigning the value 1
  - **Not Low**: For R295 >= 10, assigning the value 0

  To carry out this action, it is necessary to use a Math Formula Node – KNIME, which, through the if(x,y,z) function, will allow assigning a numerical value to the previously defined ranges and conditions. Then it is necessary to apply a Cell Replacer Node – KNIME to replace the numerical values (0, 1) with the ordinary values (Low, Not Low) respectively.
- A process of filtering rows for missing values is carried out, this action is carried out using the Row Filter Node – KNIME, with which 10.24% of the records of the data set are eliminated minable.
- Due to the large number of missing values, 65.03% referring to transactional records, the omission of these attributes is taken as an alternative to mitigate missing data. This action is performed using the Missing Value Column Filter – KNIME node, where all attributes with less than 10% missing values are selected, thus omitting 83.22% of the attributes.
- Outliers are detected for each of the attributes individually. This action is performed using the Numeric Outliers – KNIME node, generating a numerical outlier model. This is used by the Numeric Outliers(Apply) node – KNIME, to treat outliers in the input data according to the parameters of the model input.

Once this process is done, the minable data set is obtained, which is saved for later use in a .CSV file.

3.2. Methods

In this session, machine learning-based methods used in trait selection for low-yield industrial cane sugar production are presented. Next, a brief description of rule learning algorithms and methods for identifying informative features is given.

Inductive rule learning is one of the most traditional fields in machine learning [1]. Inductive rule learning solves a classification problem through the induction of a set of rules or a list of decisions. The

Table 1. Attributes of The Indust_Daily_Indicator

Name	Data Type	Detail
ID_INDICATOR	INT	Main, identifies the analyzed indicator through a relationship with Iplus_Indicator.
ID_ENTITY	INT	Excluded, unimportant for the present study.
DATE_LOAD_DATA	DATETIME	Excluded, unimportant for the present study.
VALUE_DAY	NUMERIC	Main, stores the values needed for the investigation.
VALUE_HF	NUMERIC	Excluded, unimportant for the present study. It is accumulated value.
VALUE_WEEK	NUMERIC	Excluded, unimportant for the present study. It is accumulated value.

main approach is the so-called spread-and-conquer or cover algorithm, which learns one rule at a time, successively eliminating covered examples. The individual algorithms within this framework differ mainly in the way they learn individual rules [8]. Rule-based methods are useful and well known in machine learning because they are capable of creating interpretable models [25]. However, noisy examples and outliers can harm the performance of the final model [9].

The data in many real-world applications may have many dimensions, and the characteristics of that data are often highly redundant. The identification of informative features has become an important step for data mining, not only to circumvent the curse of dimensionality, but also to reduce the amount of data for processing [26].

The algorithms analyzed for the selection of characteristics are the following:

CONJUNCTIVERULE: It implements a single conjunctive learning rule from the comparison of a validated data set [21]. A rule consists of several antecedents together and the value of the class for classification. The next thing the algorithm does is to distribute the available classes to the middle term for a numeric value. If the experimental instance is inconspicuous for this rule, then it is predicted using a predetermined distribution of the class on the data covered by the learning rule [22].

DECISIONTABLE: Decision tables are classification models used for prediction [24]. Decision tables are one of the simplest forms of knowledge representation within the field of classification. Its most basic form of use is the storage of the occurrences of the most relevant attributes on each of the classes. The accuracy of this classifier depends largely on the attribute selection process that is carried out in its first stage. It is generally used as an evaluation function for the selection of attributes to the accuracy of the decision table itself using the cross-validation process [28].

RIDER: It is based on the Ripple Down Rule algorithm. It generates a default (default) rule and then takes a set of rules that predict classes for the default rule with the least error. Then it generates the best set of rules until the error is reduced. It performs a tree-like expansion of exceptions. Exceptions are a set of rules that predict classes other than the default ones [21].

FURIA: It learns fuzzy rules instead of conventional rules, and unordered rule sets instead of rule lists. Furthermore, to deal with uncovered examples it makes use of an efficient stretching rule. The experimental results presented show that FURIA significantly outperforms the original RIPPER algorithm, as well as other classifiers such as C4.5, in terms of classification accuracy [3]. The main difference between a fuzzy rule and a conventional rule is that the fuzzy rule tends to cover more, so it has an advantage over the conventional rule [25].

PART: It generates a list of decision rules in hierarchical order. In essence, it builds a rule, removes the instances it covers, and continues recursively creating rules for the final instances until there are no instances left [21]. It is considered an industry standard as a classification algorithm. It is considered a much improved algorithm in terms of prediction accuracy [25]. The algorithm uses pessimistic pruning. The algorithm generates a decision tree and the tree building and pruning operations are combined to produce the subtree that cannot be expanded further. A rule is derived from a partial tree [11].

JRIP: It based on the RIPPER algorithm (Repeated Incremental Pruning for Error Reduction). It uses several comparisons at the same time, builds a set of rules separately and then performs comparisons between them [21]. It is a learning algorithm that is based on different rules that it uses to create a set of rules that is responsible for identifying the possible classes, while minimizing the number of errors. The error is defined by the number of training examples misclassified by the rules. The algorithm assumes that the data with which it has been previously trained is similar in some way to the unseen data on which it will perform the calculations to obtain the different rules [25]. It uses sequential coverage algorithms to create ordered lists of rules. The algorithm goes through four stages: Growth of a rule, Pruning, Optimization, and Selection [14].

The training model for classification is defined with the KNIME tool and Weka nodes for rule-based classification algorithms. A workflow is designed where the six rule learning algorithms are applied. The model generated by the algorithm is taken and classified with the test data and a series of precision statistics are calculated. Subsequently, a comparative analysis between the algorithms is carried out.

The data described in the previous section were explored in the experiments. The models were built using 70% of the data for the training set and 30% for the test set. A stratified sampling is applied where, out

Table 2. Data set partitioning

	Training Set		Test Set	
	Nr. Records	%	Nr. Records	%
Total	10 904	70.00	4 674	30.00
Low	5 596	51.32	2 399	51.33
Not Low	5 308	48.68	2 275	48.67

of a total of 15,578 records, 10,904 are used for the training set and 4,674 for the test set. The partition of the data set is detailed in Table 2.

Then, in a similar way, a flow is performed to automate the selection of features from the most appropriate attribute subsets to explain a target attribute, in the sense of supervised classification, that is, to explore which attribute subsets are the best to classify the instance class. The objective attribute to explain is the EVAL_LOWYIELD categorical attribute. At the beginning of the feature selection cycle, all the features of the input data set that will be taken into account for the construction of the model are selected, as well as those that will be kept fixed in the selection process.

4. Results and Discussion

When evaluating the algorithms, the precision statistics for each one are obtained. They are detailed in Table 3.

The **Recall** metric measures how good the model is at detecting positive events [32]. It is obtained that the algorithm that is the best to identify poor performance is DECISIONTABLE with (1.0), followed by RIDOR with (0.93).

The **Precision** metric measures how good the model is for assigning positive events to the positive class [32]. It is obtained that the algorithm that presents the most precision for the training carried out to classify low performance is the CONJUNCTIVERULE with (0.94), followed by JRIP with (0.92).

The **Sensitivity** metric measures how apt the model is to detect events in the positive class [32]. It is obtained that the algorithm that presents the most sensitivity for the training carried out to classify low performance is the CONJUNCTIVERULE with (0.94), followed by the JRIP with (0.92).

The **Specificity** metric measures how exact the assignment to the positive class is [32]. It is determined that the algorithm that presents the most specificity for the training carried out to classify low performance is CONJUNCTIVERULE with (0.87), followed by FURIA with (0.62).

The **F-measure** metric is the harmonic mean of recovery and precision [32]. It is determined that the algorithm that presents the best precision and recovery to classify low performance is DECISIONTABLE with (0.92), followed by RIDOR with (0.91).

The **Cohen's Kappa Coefficient (κ)**, a concordance statistic between two researchers that corrects for chance [13], shows that the most reliable algorithm for the training performed is FURIA with (0.35), followed by JRIP with (0.33).

The **Accurancy** metric measures the percentage of cases that the model has been correct [20]. It is obtained that the algorithm that presents the best precision and recovery to classify low performance is DECISIONTABLE and RIDOR with (0.85), followed by PART with (0.8).

The algorithm that selected the highest number of attributes was Ridor with 75 attributes, while the one that selected the least amount was JRIP with 52 attributes. For its part, the algorithm that presented the lowest prediction error was PART, while the one with the highest was CONJUNCTIVERULE, as shown in Table 4.

As a result of the number of rules generated by these algorithms, the process is automated for feature selection of the most appropriate attribute subsets to explain the target attribute. Table 5 presents

Table 4. Statistics of Feature Selection Filter

Algorithms	Statistics	
	Error	Nr. of Features
CONJUNCTIVERULE	0.057	65
DECISIONTABLE	0.012	53
RIDOR	0.004	75
FURIA	0.004	62
PART	0.003	69
JRIP	0.005	52

Table 3. Acurrancy Statistics

Accuracy Statistics	CONJUNCTIVERULE		DECISIONTABLE		RIDOR		FURIA		PART		JRIP	
	Low	NotLow	Low	Not Low	Low	Not Low	Low	Not Low	Low	Not Low	Low	Not Low
True Positives	1792	751	5061	8	4744	278	4179	537	4451	276	4194	504
False Positives	113	3283	856	14	586	331	327	896	588	624	360	881
True Negatives	751	1792	8	5061	278	4744	537	4179	276	4451	504	4194
False Negatives	3283	113	14	856	331	586	896	327	624	588	881	360
Recall	0.35	0.87	1.00	0.01	0.93	0.32	0.82	0.62	0.88	0.32	0.83	0.58
Precision	0.94	0.19	0.86	0.36	0.89	0.46	0.93	0.37	0.88	0.31	0.92	0.36
Sensitivity	0.35	0.87	1.00	0.01	0.93	0.32	0.82	0.62	0.88	0.32	0.83	0.58
Specificity	0.87	0.35	0.01	1.00	0.32	0.93	0.62	0.82	0.32	0.88	0.58	0.83
F-measure	0.51	0.31	0.92	0.02	0.91	0.38	0.87	0.47	0.88	0.31	0.87	0.45
Acurrency		0.43		0.85		0.85		0.79		0.8		0.79
Cohen's kappa		0.09		0.01		0.29		0.35		0.19		0.33
Nr. Rules		1		8119		91		45		948		64

Table 5. Frequency of appearance

Frequency of Appearance	Nr. of Attributes	Selected Attributes
6 out of 6 100%	6	R740a, R230a, R365a, R638, R390, R421
5 out of 6 83.33%	16	R613, R346, R334, R3d, R375, R1a, R345, R313d, R350, R349, R160b, R314, R299e, R591, R371, R457a
4 out of 6 66.67%	41	R379, R378, R419a, R410, R380, R336, R170, R544, R545, R347a, R1426, R476a, R3a, R419, R464, R313c, R333a, i115, R4, R5a, R434, R572, R351, R594, R364, R547, R703, R282c, R160, R145, i64, R296, R370, R365, R462, R588, R534d, R, R463, R628, R457
3 out of 6 50%	20	R613a, R1427, R230, R190, R3g, R344, R434a, R1d, R574, i146, R333, R337a, R167, R365c, R364a, i113, i42a, R6, R420, R1524
2 out of 6 33.33%	13	R417, R418, R381, R497, R345a, R567, R311, R324, R744b, R5d, R282e2, R458, R529
1 out of 6 16.67%	4	R548, R551, R590, R583

a summary of the analysis carried out with the frequency of appearance of the attribute in each of the algorithms, which allows us to know those that predominate. The most predominant attributes according to their frequency are 6 attributes are present in 100% of the algorithms, as well as 16 attributes are present in 83.33% of the algorithms. 41% of the attributes analyzed, the largest number, are present in 66.67% of the algorithms.

From analysis carried out, the attributes with the highest frequency of appearance are described:

- Last Juice Extracted Total Brix (R740a): It is the amount of dissolved solids in the juice that the bagasse contains when it leaves the mills.
- Sugar 96 in Operation (R230a): It is the amount of sugar that remains in the technological equipment.
- Juices Total Pol (R365a): It is the sucrose content in the juice.
- Cubic Meters Mass Cooked A / t Cane (R638): It is an indicator of the relationship between the mixture of sugar and mother liquor discharged from the tank, with the tons of cane.
- Recovered % Pol Cane (R390): Percentage of sucrose extracted from the cane juice in the manufacturing process.
- Total Cane % Pol (R421): It is the percentage of sucrose content in the total cane.

A ROC curve analysis is performed to select the possibly optimal attributes that most influence poor performance. It creates the column that contains the two classes: EVAL_LOWYIELD and it sets the value

Table 6. Area Under Curve

Selected Attributes	Description	Area Under Curve
R417	Pol Bagasse % Pol Cane	0.87
R379	Bagasse Loss % Pol Cane	0.81
R378	Final Honey Loss % Pol Cane	0.79
R419a	Boiler House Losses % Pol Cane	0.75
R410	Total Final Honey % Pol Cane	0.74
R613	Hours Removal without extractions	0.73
R1427	Total Foreign Matter %	0.70
R380	Filter Cake Loss % Pol Cane	0.70

to which the high probabilities are assigned: Low. In Table 6, the attributes that are above the random estimation line (diagonal) are listed in descending order, representing the good classification results for the selected class.

It is interesting to point out that the attribute R417, which is the one closest to the perfect classification point, in the previous analysis is only selected as a characteristic in two algorithms. On the other hand, the attributes R379, R378, R419a, R410 are only selected as characteristic in four algorithms. While R613a is selected as characteristic in five algorithms despite having a lower area under the curve.

Based on the analysis carried out, the attributes that are considered for this study as the attributes that most influence low industrial performance are described thus:

- Pol Bagasse % Pol Cane (R417): Parts of Pol that must come out with the bagasse for every 100 parts of bagasse. It is used to measure the efficiency work in grinding.
- Bagasse Loss % Pol Cane (R379): Expresses the value of the pol in the bagasse produced by the sugar mill for every 100 parts of pol that entered with the cane.
- Final Honey Loss % Pol Cane (R378): Expresses the value of the Pol in the final molasses produced by the mill for every 100 parts of Pol that entered with the cane.
- Boiler House Losses % Pol Caña (R419a): The losses in the boiler house are made up of the final molasses, filter cake and indeterminate.
- Total Final Honey % Pol Cane (R410): Same as R378 but covers the losses produced in streams of products that leave the process through extractions such as rich honey and different juices. Reason why it says Total.
- Hours Removal without extractions (R613): It expresses the time it takes for the sugar that the cane brings to travel through the entire process until it comes out as a final product and as losses. Normal values are between 24 and 36 hours. The lower the values, the lower the indeterminate losses.
- Total Foreign Matter % (R1427): Foreign matters are parts of the cane plant that contain substances

that are harmful to the process. It is expressed as the % of the weight of the cane that is ground that is made up of these impurities: earth, bud, green leaves, dry leaves and others.

- Filter Cake Loss % Pol Cane (R380): It expresses the value of the Pol in the filter cake produced by the mill as residue for every 100 parts of Pol that came in with the cane.

5. Conclusion

- The work allows for a broad understanding of the business, an understanding of the data, as well as a preparation to carry out the modeling of different techniques. Many of the dataset's attributes were found to be worthless.
- The work allowed for comparison of different algorithms of rules and for carrying out an automated process for the selection of characteristics that allow identifying those that best fit the stated objectives.
- The following were identified: R417, R379, R378, R419a, R410, R613, R1427 and R380, as the indicators that most influence the classification of low industrial performance.
- The work constitutes a starting point for the evaluation and deeper validation of the rules and characteristics obtained.

Notes

[1] Grupo Azucarero AZCUBA, https://www.azcuba.cu
[2] Industrial Plus, https://www.datazucar.cu/?featured_item=iplus

AUTHORS
Yohan Gil Rodríguez* – ESI DATAZUCAR, AZCUBA, Avenida 23 No.171/ N y O, Vedado, Plaza de la Revolución, La Habana, Cuba, e-mail: yohan.gil@datazucar.cu, www: https://orcid.org/0000-0002-8239-4124/~ORCID.
Raisa Socorro Llanes – CUJAE, Calle 114 No.11901/ Ciclovía y Rotonda, Marianao, La Habana, Cuba, e-mail: raisa@ceis.cujae.edu.cu, www: https://orcid.org/0000-0002-2627-1912/~ORCID.
Alejandro Rosete – CUJAE, Calle 114 No.11901/ Ciclovía y Rotonda, Marianao, La Habana, Cuba, e-mail: rosete@ceis.cujae.edu.cu, www: https://orcid.org/0000-0002-4579-3556/~ORCID.
Lisandra Bravo Ilisástigui – CUJAE, Calle 114 No.11901/ Ciclovía y Rotonda, Marianao, La Habana, Cuba, e-mail: lbravo@ceis.cujae.edu.cu, www: https :// orcid.org/0000-0002-8209-4121/~ORCID.
*Corresponding author

References

[1] F. Beck, and J. Fürnkranz. "An Empirical Investigation Into Deep and Shallow Rule Learning", *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[2] J. Casillas, O. Cordón, M. J. Del Jesus, and F. Herrera. "Genetic feature selection in a fuzzy rule-based classification system learning process for high-dimensional problems", *Information Sciences*, vol. 136, no. 1, 2001, 135–157, doi: 10.1016/S0020-0255(01)00147-5.

[3] J. Coto Palacio, Y. Jiménez Martínez, A. Nowé, J. Coto Palacio, Y. Jiménez Martínez, and A. Nowé. "Aplicación de sistemas neuroborrosos en la clasificación de reportes en problemas de secuenciación", *Revista Cubana de Ciencias Informáticas*, vol. 14, no. 4, 2020, 34–47, Publisher: Universidad de las Ciencias Informáticas.

[4] B. Dębska, and B. Guzowska-Świder. "Decision trees in selection of featured determined food quality", *Analytica Chimica Acta*, vol. 705, no. 1, 2011, 261–271, doi: 10.1016/j.aca.2011.06.030.

[5] Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber. "Accurate prediction of sugarcane yield using a random forest algorithm", *Agronomy for Sustainable Development*, vol. 36, no. 2, 2016, 27, doi: 10.1007/s13593-016-0364-z.

[6] M. A. Ferraciolli, F. F. Bocca, and L. H. A. Rodrigues. "Neglecting spatial autocorrelation causes underestimation of the error of sugarcane yield models", *Computers and Electronics in Agriculture*, vol. 161, 2019, 233–240, doi: 10.1016/j.compag.2018.09.003.

[7] Y. Filiberto, R. Bello, Y. Caballero, and M. Frías. "Algoritmo para el aprendizaje de reglas de clasificación basado en la teoría de los conjuntos aproximados extendida", *DYNA*, vol. 78, no. 169, 2011, 62–70, Publisher: 2006, Revista DYNA.

[8] J. Fürnkranz. "Rule Learning". In: C. Sammut and G. I. Webb, eds., *Encyclopedia of Machine Learning*, 875–879. Springer US, Boston, MA, 2010.

[9] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, volume 72 of *Intelligent Systems Reference Library*, Springer International Publishing: Cham, 2015, doi: 10.1007/978-3-319-10247-4.

[10] P. K. Giri, S. S. De, S. Dehuri, and S. Cho. "Biogeography based optimization for mining rules to assess credit risk", *Intelligent Systems in Accounting, Finance and Management*, vol. 28, no. 1, 2021, 35–51, 10.1002/isaf.1486.

[11] Gnanambal, S., Thangaraj, M., Meenatchi, V. T., and Gayathri, V., "Classification Algorithms with Attribute Selection: an evaluation study using WEKA", *International Journal of Advanced Networking and Applications*, vol. 9, no. 6, 2018, 3640–3644.

[12] E. A. d. M. Gomes Soares, L. C. Leite Damascena, L. M. Mendes de Lima, and R. Marcos de Moraes. "Analysis of the Fuzzy Unordered Rule Induction Algorithm as a Method for Classification", 2018.

[13] J. J. T. Gordillo, and V. H. P. Rodríguez. "Cálculo de la fiabilidad y concordancia entre codificadores

de un sistema de categorías para el estudio del foro online en e-learning", vol. 27, 2009, 17.

[14] A. Gupta, A. Mohammad, A. Syed, and M. N.. "A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA", *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 7, 2016, doi: 10.14569/IJACSA.2016.070753.

[15] I. Guyon, and A. Elisseeff. "An introduction to variable and feature selection", *The Journal of Machine Learning Research*, vol. 3, 2003, 1157–1182.

[16] R. G. Hammer, P. C. Sentelhas, and J. C. Q. Mariano. "Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models", *Sugar Tech*, vol. 22, no. 2, 2020, 216–225, doi: 10.1007/s12355-019-00776-z.

[17] J. Hernández Orallo, M. J. Ramárez Quintana, and C. Ferri Ramírez. *Introducción a la Minería de Datos*, Pearson Educacion. S.A: España, 2004, OCLC: 933368678.

[18] C. Huang, X. Huang, Y. Fang, J. Xu, Y. Qu, P. Zhai, L. Fan, H. Yin, Y. Xu, and J. Li. "Sample imbalance disease classification model based on association rule feature selection", *Pattern Recognition Letters*, vol. 133, 2020, 280–286, doi: 10.1016/j.patrec.2020.03.016.

[19] Y. Li and Z.-F. Wu. "Fuzzy feature selection based on min–max learning rule and extension matrix", *Pattern Recognition*, vol. 41, no. 1, 2008, 217–226, doi: 10.1016/j.patcog.2007.06.007.

[20] J. Martinez Heras. "Precision, Recall, F1, Accuracy en clasificación", October 2020. Section: machine learning.

[21] V. B. Núñez, R. Velandia, F. Hernández, J. Meléndez, and H. Vargas. "Atributos Relevantes para el Diagnóstico Automático de Eventos de Tensión en Redes de Distribución de Energía Eléctrica", *Revista Iberoamericana de Automática e Informática Industrial RIAI*, vol. 10, no. 1, 2013, 73–84, doi: 10.1016/j.riai.2012.11.007.

[22] R. A. V. Ortega, and F. L. H. Suárez. "Evaluación de algoritmos de extracción de reglas de decisión para el diagnóstico de huecos de tensión", 2010, 127.

[23] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev. "Analyzing the impact of feature selection on the accuracy of heart disease prediction", *Healthcare Analytics*, vol. 2, 2022, 100060, doi: 10.1016/j.health.2022.100060.

[24] B. T. Pham, C. Luu, T. V. Phong, H. D. Nguyen, H. V. Le, T. Q. Tran, H. T. Ta, and I. Prakash. "Flood risk assessment using hybrid artificial intelligence models integrated with multi-criteria decision analysis in Quang Nam Province, Vietnam", *Journal of Hydrology*, vol. 592, 2021, 125815, doi: 10.1016/j.jhydrol.2020.125815.

[25] F. M. Pérez. "Estudio y análisis del funcionamiento de técnicas de minería de datos en conjuntos de datos relacionados con la Biología", 35.

[26] H. Rao, X. Shi, A. K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, X. Yuan, and L. Gu. "Feature selection based on artificial bee colony and gradient boosting decision tree", *Applied Soft Computing*, vol. 74, 2019, 634–642, doi: 10.1016/j.asoc.2018.10.036.

[27] M. Ribas García, R. Consuegra del Rey, and M. Alfonso Alfonso. "Análisis de los factores que más inciden sobre el rendimiento industrial azucarero", vol. 43, no. 1, 2016, 10.

[28] A. Rivas Méndez. "Estudio experimental sobre algoritmos de clasificación supervisada basados en reglas en conjuntos de datos de alta dimensión", 2014, Accepted: 2019-07-09T15:50:17Z Publisher: Universidad de Holguín, Facultad Informática Matemática, Departamento de Informática.

[29] M. Schiezaro, and H. Pedrini. "Data feature selection based on Artificial Bee Colony algorithm", *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, 2013, 47, doi: 10.1186/1687-5281-2013-47.

[30] K. Topouzelis, and A. Psyllos. "Oil spill feature selection and classification using decision tree forest on SAR image data", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 68, 2012, 135–143, doi: 10.1016/j.isprsjprs.2012.01.005.

[31] S. Veenadhari, B. Misra, and C. Singh. "Machine learning approach for forecasting crop yield based on climatic parameters". In: *2014 International Conference on Computer Communication and Informatics*, Coimbatore, India, 2014, 1–5, doi: 10.1109/ICCCI.2014.6921718.

[32] M. Widmann. "From Modeling to Scoring: Confusion Matrix and Class Statistics", May 2019.

[33] H. Zhou, J. Zhang, Y. Zhou, X. Guo, and Y. Ma. "A feature selection algorithm of decision tree based on feature weight", *Expert Systems with Applications*, vol. 164, 2021, 113842, doi: 10.1016/j.eswa.2020.113842.

[34] L. Zhou, Y.-W. Si, and H. Fujita. "Predicting the listing statuses of Chinese-listed companies using decision trees combined with an improved filter feature selection method", *Knowledge-Based Systems*, vol. 128, 2017, 93–101, doi: 10.1016/j.knosys.2017.05.003.