Article

Classifying the Level of Energy-Environmental Efficiency Rating of Brazilian Ethanol

Nilza Duarte da Silva Lima ©, Irenilza de Alencar Nääs ©, João Gilberto Mendes dos Reis *© and Raquel Baracat Tosi Rodrigues da Silva ©

Postgraduate Program in Production Engineering, Universidade Paulista - UNIP, Dr. Bacelar Street 1212, 04026002 São Paulo, Brazil; nilsa.lima@stricto.unip.br (N.D.d.S.L.); irenilza.naas@docente.unip.br (I.d.A.N.); raquel.silva@stricto.unip.br (R.B.T.R.d.S.)

* Correspondence: joao.reis@docente.unip.br; Tel.: +55-11-5586-4145

Received: 28 February 2020; Accepted: 14 April 2020; Published: 21 April 2020

Abstract: The present study aimed to assess and classify energy-environmental efficiency levels to reduce greenhouse gas emissions in the production, commercialization, and use of biofuels certified by the Brazilian National Biofuel Policy (RenovaBio). The parameters of the level of energy-environmental efficiency were standardized and categorized according to the Energy-Environmental Efficiency Rating (E-EER). The rating scale varied between lower efficiency (D) and high efficiency + (highest efficiency A+). The classification method with the J48 decision tree and naive Bayes algorithms was used to predict the models. The classification of the E-EER scores using a decision tree using the J48 algorithm and Bayesian classifiers using the naive Bayes algorithm produced decision tree models efficient at estimating the efficiency level of Brazilian ethanol producers and importers certified by the RenovaBio. The rules generated by the models can assess the level classes (efficiency scores) according to the scale discretized into high efficiency (Classification A), average efficiency (Classification B), and standard efficiency (Classification C). These results might generate an ethanol energy-environmental efficiency label for the end consumers and resellers of the product, to assist in making a purchase decision concerning its performance. The best classification model was naive Bayes, compared to the J48 decision tree. The classification of the Energy Efficiency Note levels using the naive Bayes algorithm produced a model capable of estimating the efficiency level of Brazilian ethanol to create labels.

Keywords: biofuel policy; efficiency rating; ethanol; data mining

1. Introduction

Developing renewable energy is one of the leading global interests in promoting sustainability and environmental quality, including modern electricity grids worldwide, which have begun to rely more heavily on renewable energy sources [1–3]. Ethanol is a renewable fuel produced by the fermentation of sugarcane extract and molasses. The product has a lower carbon footprint, is biodegradable, and has greater energy-environmental efficiency (renewable energy) compared to oil due to its sustainability in the production chain with better use of natural resources [4–10]. Ethanol is one of the main biofuels consumed in Brazil.

Biofuel partially (or entirely) replaces fossil fuels in engines (flex vehicles) [11]. The addition of 27% ethanol in gasoline (Cgasoline, with the addition of anhydrous ethanol fuel) has been mandatory in Brazil since 2015 [8,12–14]. Such an initiative expanded the country’s consumption of biofuels and increased the energy matrix’s sustainability. It also supported the goal of reducing 37% of GHG emissions by 2025, compared to 2005 levels [2,12,15,16]. Differently, some European Union countries also include electromobility (electric cars) in their GHG emission reduction forecasts [17].
The production of biofuel from sugarcane is supported by the National Agency of Petroleum, Natural Gas and Biofuels’ (ANP) sustainable development programs to achieve goals to reduce greenhouse gas emissions. Therefore, annual national targets were developed by the Brazilian National Biofuel Policy (RenovaBio) [12,15]. These goals are established in units of Decarbonization Credits (CBios) calculated from annual mandatory targets defined individually for each producer and distributor according to their participation in the fossil fuel market [12,16,18–21].

The certification of biofuel production assigns different marks to each biofuel producer and importer, in a value inversely related to the carbon amount of biofuel produced. The note reflects the individual contribution of each producing agent to mitigate a specific amount of GHG concerning its fossil substitute (in tons of CO₂ equivalent). The total emission is compared with that of the equivalent fossil fuel (gasoline, for ethanol), resulting in a final score (Energy-Environmental Efficiency Rating), characterizing the mitigation of emissions. This note generates CBios for biofuel producers and importers; with the decarbonization of the Brazilian energy matrix, there is a mechanism for the commercialization of these CBios linked to the carbon intensity of biofuels [2,12,16,18].

The incentive to reduce pollutant emissions in the biofuel chain goes far beyond the use of flex vehicles by consumers [9,11,22], but it is directly linked to decarbonization credits given to biofuel producers and distributors, because although renewable, it depends on how sugarcane is produced [5,8,23–26]. Alkimim and Clarke [27] showed that the carbon debt of deforestation in Brazilian biomes for ethanol production was equivalent to 608 Mg CO₂ ha⁻¹ in the Amazon, 142 Mg CO₂ ha⁻¹ in the Cerrado, and 212 Mg CO₂ ha⁻¹ to the Atlantic Forest with the respective return time of 62, 15, and 22 years. In this context, it is essential to integrate dissemination strategies, clear and comprehensive, on what is the level of energy-environmental efficiency of biofuel producers and distributors [22,28,29], in order to raise awareness and increase the number of certificates at RenovaBio, the membership of which is not mandatory.

Predictive models are essential for the biofuel chain for both the consumer and the distributor as this classification of the level of efficiency can generate labels of their energy-environmental performance, increasing transparency and environmental responsibility concerning a product such as ethanol. The use of data mining to assess the performance of classifiers in the sugarcane sector is quite broad. It ranges from classifiers for mapping sugarcane planting [30] to deep learning techniques for the detection of sugarcane diseases [31] and classification of crop yield characteristics with neural networks used both in the recognition and in the grading of satellite images of sugarcane plantations [32].

The present study aimed to assess and classify energy-environmental efficiency levels to reduce greenhouse gas emissions in the production, commercialization, and use of biofuels certified by the Brazilian National Biofuel Policy (RenovaBio).

2. Material and Methods

2.1. Characterization of the Study

The study considered certification data for the production or efficient import of biofuels from the Brazilian National Biofuel Policy (RenovaBio) regulated by the National Agency of Petroleum, Natural Gas and Biofuels [12,33,34]. The Certificate of Efficient Production of Biofuels includes an Energy-Environmental Efficiency Note, resulting from the technical profile informed by a calculation validation spreadsheet (RenovaCalc) [2,12,33,34], and is linked to the volume of biofuel produced and marketed, generating the Decarbonization Credits (CBio) of the RenovaBio to reduce carbon emissions and improve the performance of biofuels. In addition to offering CBio, it is generated by the difference between the CO₂ equivalent emission of fossil fuels (baseline) and its biofuel substitute [12,18]. The biofuels evaluated in the study were anhydrous ethanol and hydrous ethanol. Only the database of the certificates of the efficient production or import of biofuels approved in 2019 was evaluated [12].
2.2. Classifying the Biofuel Energetic-Environmental Efficiency

The classification levels of the biofuel environmental efficiency were performed based on the Energy-Environmental Efficiency Rating. Such a rating is the result of the certificates of the production or efficient import of biofuels informed through the calculation (RenovaCalc), linked to the volume of biofuel produced and commercialized, generating Decarbonization Credits (CBio) in the RenovaBio’s program.

The energy-environmental efficiency level parameters were standardized according to Table 1. The dataset was categorized according to the Energy-Environmental Efficiency Rating (RenovaBio) [12], the rating scale varying between lower efficiency and high efficiency + (highest efficiency), and an example of the label’s design of the energy-environmental efficiency for the performance of a certificate is given (Figure 1).

Table 1. Parameters of the level of the Energy-Environmental Efficiency Rating.

| Energy-Environmental Efficiency Rating | EC | Description               |
|---------------------------------------|----|---------------------------|
| 81–100                                | A+ | High efficiency + (highest efficiency) |
| 61–80                                 | A  | High efficiency            |
| 51–60                                 | B  | Average efficiency        |
| 31–50                                 | C  | Standard efficiency       |
| 1–30                                  | D  | Lower efficiency          |

1 Based on the RenovaBio; 2 Efficiency Classes.

Figure 1. Energy-environmental efficiency label design for performance certificates.

The calculation and indication of the environmental energy efficiency score (E-EER) of the certification for efficient production of biofuel, made available by the ANP, was discretized into categories of levels (classes) (pre-processing of the dataset) for the classification (data mining) and to create one of the labels of the energy-environmental performance (post-processing) of Brazilian ethanol. This labeling system allowed it to be classified into five classes (A+, A, B, C, and D) to provide consumers with a differentiation of the ethanol consumed from different producers, regions, or states.

The RenovaBio Program allows producers and importers to be able to declare the energy-environmental efficiency of their product, which is economically attractive for decarbonization and the competitiveness of biofuels in the oil market, with a complex and solid structure (Figure 2) [2,12,18,33,34]. The label may be shown at fuel pumps to consumers with a validity of one to three years, a validity that is applied to the Certification of Efficient Production of Biofuels when approved by the ANP. It can also endorse the information and increase transparency in the biofuel market at the consumer level, helping to make a purchase decision. The objective of energy-environmental labeling is to encourage Brazilian sugar and alcohol industries to develop innovations and improvements beyond the minimum levels of efficiency. However, it is expected that
more ethanol producers will be able to adhere to the ANP certifications of RenovaBio [2,12,19,33,34], and consequently, the labeling system can be improved with the inclusion of more data on the platform.

![Plan of the emission and trade of decarbonization credits (CBios) created by the RenovaBio Program, adapted by Klein et al. [19].](image)

**Figure 2.** Plan of the emission and trade of decarbonization credits (CBios) created by the RenovaBio Program, adapted by Klein et al. [19].

### 2.3. Classification of Model Prediction

Data mining applies to this study, through techniques (algorithms), for the classification of the levels (classes) of energy-environmental efficiency in the search of strategic information that allows the extraction of implicit information existing in the databases, contributing to the process of identifying and classifying new patterns [35–37]. The steps of the data mining method were selection, pre-processing, data mining, and post-processing (knowledge filtering, interpretation and explanation, evaluation, and knowledge integration) for knowledge discovery from the classifiers [36,37]. The results obtained could be used in information management, information processing, decision making, and process control.

The data contained in the databases could be used to learn a specific target concept [35–38]. The tasks performed by data mining techniques and machine learning, the classification, build models that can be applied to unclassified data to categorize them into classes, to relate the meta attribute (whose value will be predicted) and a set of forecasting attributes [35–38].

The data were assessed in the ANP database for the registration of certificates of the production or efficient import of biofuels approved and included in the RenovaBio program in 2019 [12,33,34]. We considered only anhydrous ethanol and hydrated ethanol products, generating two products for the same biofuel producer and importer. The data pre-processing was performed in Excel spreadsheets for further processing in the data mining software Weka (Waikato Environment for Knowledge Analysis) Version 3.8.4 [39–42]. The attributes used to build the predictive model were: “biofuel-type”, “state”, “eligible-volume (%)”, “emission factor”, “Energy-Environmental Efficiency Rating”, and “LER” (Level of Efficiency Rating). Figure 3 presents the modeling process used to classify the Energy-Environmental Efficiency Rating.
Figure 3. Schematic of the modeling process used to classify the Energy-Environmental Efficiency Rating of Brazilian ethanol. E-EER, Energy-Environmental Efficiency Rating; LER, Level of Efficiency Rating.

A classifier is a mapping from unlabeled instances to (discrete) classes. Classifiers have a form (classification tree) plus an interpretation procedure (including how to handle unknown values). Most classifiers also can provide probability estimates (or other likelihood scores), which can be thresholded to yield a discrete class decision, thereby taking into account a cost/benefit or utility function [43,44].

During the pre-processing of the data, the dataset was extracted from the RenovaBio Program (ANP) platform, selecting only the data on the ethanol product and organized in a spreadsheet. The implementation of the supervised filter Resample was applied to maintain the distribution of classes in the subsample and to reach a uniform distribution for comparing the data not submitted to the filter (noResample). The filter Resample produces a random subsample of a dataset using either sampling with replacement or without replacement. The filter was made to maintain the class distribution in the subsample or to bias the class distribution toward a uniform distribution.

The J48 decision tree classification algorithms (“weka.classifiers.trees.J48 - C0.25 - M2” without the filter and in relation to “data - biofuel - weka.filters.supervised.instance.Resample - B0.0 - S1 - Z100.0” using the resample filter) and the naive Bayes algorithm (“weka.classifiers.bayes.NaiveBayes” without the filter and in relation to “data - biofuel - weka.filters.supervised.instance.Resample - B0.0 - S1 - Z100.0” using the resample filter) were used. The model validation was done using cross-validation (10 folds) applied to the dataset of 27 instances and six attributes (“biofuel-type”, “state”, “eligible-volume (%)”, “emission-factor”, “energy-environmental-efficiency-rating”, and “LER” (Level of Efficiency Rating)). For each sample, the known class label was compared with the prediction of the learned class model.

In supervised learning, each data input object is preassigned a class label. The main task of supervised algorithms (J48 and naive Bayes) is to learn a model that produces the same labeling for the provided data [43–45]. The decision tree algorithm is a widely used algorithm for classification, which uses attribute values to partition the decision space into smaller subspaces in an iterative manner, and the decision processes can be represented graphically as a tree. Each possible decision is covered and represented as a branch, and a complete decision process is essentially a path or branch from the root node to a leaf [43,44]. Naive Bayes is a classification algorithm widely used in
problems due to its simplicity, effectiveness, and robustness, being a probabilistic approach based on assumptions that resources are independent of each other and that their weights are equally important [46]. They can better represent the complex relationships between input variables found in real problems [46]. Probabilistic inference can be studied as an approach based on the assumption that decision variables follow probable distributions. The essence of a Bayesian classifier is to estimate the probabilities of all alternative models or hypotheses, given data as evidence, and then to find the most probable classification to be assigned to each new input [44,46].

In the step of the post-processing for filtering, interpretation and explanation, evaluation, and knowledge integration generated by the algorithms for knowledge discovery from the classifiers, the metrics of the performance of the algorithms were used. Post-processing procedures usually include various pruning routines, rule quality processing, rule filtering, rule combination, model combination, or even knowledge integration [47].

The last step is to evaluate the prediction, and such an analysis was made based on the performance values obtained through the test of the prediction model [48]. The performance evaluation measures of the prediction models used were the confusion matrix, accuracy, precision, and recall and the correlation coefficient between classes (Matthews Correlation Coefficient (MCC)) for testing with the resample filter. The Kappa statistic measured the learning capacity of the algorithm.

The confusion matrix presented a matrix with results obtained during the test phase of the model, and it was used in models that used classification algorithms. Considering a confusion matrix of a hypothesis, it offered an adequate measure of the classification model, by showing the number of correct classifications versus the predicted rankings for each class, over a set of instances. The number of correct answers, for each class, was located on the main diagonal of the matrix, and the other elements represented errors in the classification.

The precision represents what has been classified correctly. The values obtained in correctly classified instances and incorrectly classified instances are determinant for predicted accuracy, since they display the values of correct classification and incorrect classification obtained by the algorithm (Equation (1)).

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

where $TP =$ True Positive; $FP =$ False Positive . The sensitivity (recall) signifies the proportion of wrong classifications or the occurrence of defects. In addition to accuracy and precision, its value varies from 0 to 1, with values closer to 1 being indicators of a good performance prediction model obtained by Equation (2).

$$\text{Sensitivity (Recall)} = \frac{TP}{(TP + FN)}$$

where $TP =$ True Positive; $FN =$ False Negative.

The Kappa statistic is a metric that compares an observed precision with an expected precision (random chance). It is a measure used to deal with multi-class and unbalanced class problems. The Kappa statistic can be defined as a measure of the degree of agreement between two categorized datasets. The Kappa result varies between 0 and 1. The higher the Kappa value, the stronger the bond [49] (Equation (3)).

$$\text{Kappa statistic} = \frac{PO - PF}{1 - PE}$$

where $PO =$ proportion of observed agreements; $PF =$ proportion of agreements expected by chance.

The Matthews Correlation Coefficient (MCC) is a correlation coefficient between the dependent classes and represents a measure of quality. Unlike accuracy, precision, and sensitivity, its value ranges from $-1$ to 1, where values closer to $-1$ are indicators of a poor prediction model, values equal
to 0 indicate that the prediction model is entirely random, and values closer to 1 are indicators of a prediction model with good performance (Equation (4)).

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$  \hspace{1cm} (4)

where TP = True Positive; TN = True Negative; P = False Positive; FN = False Negative.

3. Results

3.1. Classification of the Energy-Environmental Efficiency Level for Biofuels

The If-Then classification rules are presented related to the energy-environmental efficiency level for biofuels.

The decision tree generated by the J48 algorithm presented the following classification rules (Figure 4): If the energy-environmental efficiency deficiency (E-EER) was higher than 60.3, then the classification was A (high efficiency). If the Energy-Environmental Efficiency Rating (E-EER) was less than or equal to 60.3, then the rating depended on the state where the ethanol was produced. If the state was Goiás (GO), then the classification was C (standard efficiency). If the state was São Paulo (SP), then the classification was B (average efficiency). If the state was Mato Grosso do Sul (MS), then the rating was B (average efficiency). The results indicated that the model classified with precision above 60.

![Figure 4. Decision tree for classifying the level of the energy-environmental efficiency of biofuels.](image)

The performance of the classifiers in predicting the classes of the E-EER level using the J48 decision tree algorithm (Table 2) showed 74.07% of instances classified correctly and 25.93% for those classified incorrectly, and the learning capacity of the algorithm was 0.56 for the Kappa statistic. Class A showed the best 100% True Positive (TP) rate, with a 0.07 False Positive (FP) rate, a precision of 0.93, and high recall. However, for Class C, the precision was zero, not classifying any.
Table 2. Classifier performance (J48 decision tree) for scale class prediction models of the level of the energy-environmental efficiency of biofuel.

| Metrics | Classifier Model | Correctly Classified Instances (%) | Incorrectly Classified Instances (%) | Kappa Statistic | Test Mode |
|---------|------------------|------------------------------------|-------------------------------------|----------------|-----------|
|         | J48 Decision Tree| 74.07                              | 25.93                               | 0.56           | 10-Fold Cross-Validation |

Detailed Accuracy by Class-Classifier Performance

| Class | TP Rate | FP Rate | Precision | Recall | MCC |
|-------|---------|---------|-----------|--------|-----|
| A     | 1.00    | 0.07    | 0.93      | 1.00   | 0.93|
| B     | 0.70    | 0.23    | 0.64      | 0.70   | 0.46|
| C     | 0       | 0.09    | 0         | 0      | 0.12|
| WA    | 0.740   | 0.130   | 0.680     | 0.740  | 0.600|

WA = Weighted Average. TP rate = True Positive rate. FP rate = False Positive rate. MCC = Matthews Correlation Coefficient.

The performance of the naive Bayes algorithm showed 81.48% of instances correctly classified and 18.52% for those classified incorrectly, with a learning capacity of 0.70 for the Kappa statistic. It classified Classes A and B with high performance, with an accuracy of 1.00 and 0.73, respectively (Table 3). Comparing the two algorithms, naive Bayes presented better performance indexes concerning the J48 decision tree.

Table 3. Naive Bayes classifier performance for scale class prediction models of the level of the energy-environmental efficiency of biofuel.

| Metrics | Classifier Model | Correctly Classified Instances (%) | Incorrectly Classified Instances (%) | Kappa Statistic | Test Mode |
|---------|------------------|------------------------------------|-------------------------------------|----------------|-----------|
|         | Naive Bayes      | 81.48                              | 18.52                               | 0.7            | 10-Fold Cross-Validation |

Detailed Accuracy by Class-Classifier Performance

| Class | TP Rate | FP Rate | Precision | Recall | MCC |
|-------|---------|---------|-----------|--------|-----|
| A     | 0.92    | 0       | 1.00      | 0.92   | 0.93|
| B     | 0.80    | 0.18    | 0.73      | 0.80   | 0.61|
| C     | 0.50    | 0.09    | 0.50      | 0.50   | 0.41|
| WA    | 0.81    | 0.08    | 0.82      | 0.81   | 0.73|

WA = Weighted Average. TP rate = True Positive rate. FP rate = False Positive rate. MCC = Matthews Correlation Coefficient.

The decision tree generated by the J48 algorithm for the dataset with the application of the resample filter (Figure 5) presented the following classification rules (Figure 4). If the Energy-Environmental Efficiency Rating (E-EER) is higher than 60.1, then the classification is A (high efficiency). If the Energy-Environmental Efficiency Rating (E-EER) is less than or equal to 60.1, then the classification depends on the eligible volume (%) of the biomass. If the eligible volume is higher than 97.43, then the rating is B (average efficiency). If the eligible volume is less than or equal to 97.43, then the rating is C (standard efficiency) (Figure 5). The results indicated that the use of the resample filter had higher weight for the appropriate attribute volume during the evaluation of the E-EER and that ethanol production depended directly on this volume.
Figure 5. Decision tree for the energy-environmental level classification of biofuels using the Resample filter. GO, Goiás; SP, São Paulo; MS, Mato Grosso do Sul.

The performance of the classifiers in the prediction of the classes of the E-EER level using the J48 decision tree algorithm and with the application of the resample filter (Table 4) presented 81.48% of instances correctly classified and 18.52% for those classified incorrectly, and the algorithm learning ability was 0.70 for the Kappa statistic. Classes A and B showed high prediction with values of 1.00 and 0.78, respectively. However, the forecast for the C class was 0.40.

Table 4. Classifier performance for scale class prediction models of the level of the environmental efficiency of biofuel with the resample filter.

| Classifier Model | Correctly Classified Instances (%) | Incorrectly Classified Instances (%) | Kappa Statistic | Test Mode | Filters         |
|------------------|------------------------------------|--------------------------------------|-----------------|-----------|-----------------|
| J48 decision tree| 81.48                              | 18.52                                | 0.70            | 10-fold   | supervised instance resample |

**Detailed Accuracy by Class-Classifier Performance**

| Class | TP Rate | FP Rate | Precision | Recall | MCC  |
|-------|---------|---------|-----------|--------|------|
| A     | 1.00    | 0       | 1.00      | 1.00   | 1.00 |
| B     | 0.70    | 0.12    | 0.78      | 0.70   | 0.85 |
| C     | 0.50    | 0.13    | 0.40      | 0.50   | 0.79 |
| WA    | 0.81    | 0.06    | 0.83      | 0.81   | 0.91 |

WA = Weighted Average. TP rate = True Positive rate. FP rate = False Positive rate. MCC = Matthews Correlation Coefficient.

The performance of the naive Bayes algorithm showed 77.78% of instances classified correctly and 22.22% for those classified incorrectly, with a learning capacity of 0.64 for the Kappa statistic (Table 5). It classified Classes A and B with high performance, with an accuracy of 0.77 and 0.90, respectively (Table 5). The J48 decision tree showed better performance indexes when compared to naive Bayes.
Table 5. Classifier performance (naive Bayes) for scale class prediction models of the level of the environmental efficiency of biofuel using the resample filter.

| Classifier Model | Correctly Classified Instances (%) | Incorrectly Classified Instances (%) | Kappa Statistic | Test Mode | Filters |
|------------------|------------------------------------|-------------------------------------|----------------|-----------|---------|
| Naive Bayes      | 77.78                              | 22.22                               | 0.64           | 10-fold cross-validation | supervised instance resample |

Detailed Accuracy by Class-Classifier Performance

| Class | TP Rate | FP Rate | Precision | Recall | MCC |
|-------|---------|---------|-----------|--------|-----|
| A     | 0.77    | 0       | 1         | 0.77   | 0.80|
| B     | 0.90    | 0.23    | 0.69      | 0.90   | 0.64|
| C     | 0.50    | 0.09    | 0.50      | 0.50   | 0.41|
| WA    | 0.78    | 0.10    | 0.81      | 0.78   | 0.68|

WA = Weighted Average. TP rate = True Positive rate. FP rate = False Positive rate. MCC = Matthews Correlation Coefficient.

The confusion matrix for the J48 decision tree algorithm presented results with positive gains for the classes when the resample filter was applied, going from 20 to 22 strikes, specifically for Class C (Table 6). This gain in the resample increased the normal balance or distribution of the dataset, shown in Table 4. The results of the Matthews Correlation Coefficient (MCC) proved this gain with values of 1.00 for A, 0.85 for B, and 0.79 for C.

Table 6. Confusion matrix of the classification of the level of energy-environmental efficiency with the J48 decision tree.

| Classifier Model No Instance Filter-J48 Decision Tree | A | B | C | Total | Classified as |
|------------------------------------------------------|---|---|---|-------|---------------|
| A                                                   | 13| 0 | 0 | 13    | A             |
| 1                                                   | 7 | 2 | 10| 13    | B             |
| 0                                                   | 4 | 0 | 4 | 8     | C             |
|                                                      | 14| 11| 2 | 27    |               |

| Classifier Model with the Supervised-Filter Instance: | A | B | C | Total |
|------------------------------------------------------|---|---|---|-------|
| 13                                                   | 0 | 0 | 0 | 13    | A     |
| 0                                                    | 7 | 3 | 10| 13    | B     |
| 0                                                    | 2 | 2 | 4 | 8     | C     |
|                                                      | 13| 9 | 5 | 27    |       |

The confusion matrix for the naive Bayes algorithm also showed results with positive gains for the classes when the resample filter was applied, going from 20 to 21 hits, specifically for Class C, with two hits (Table 7). However, this gain did not have good accuracy when we observed the results of Matthews Correlation Coefficient (MCC) (Table 5), which presented values of 0.80 for A, 0.64 for B, and 0.41 for C, when compared to the J48 decision tree algorithm in the same analysis condition, which performed better.
Table 7. Confusion matrix of the classification of the level of energy-environmental efficiency with naive Bayes.

| Classifier Model No Instance Filter-Naive Bayes | A | B | C | Total | Classified as |
|------------------------------------------------|---|---|---|-------|---------------|
| 13                                             | 0 | 0 | 0 | 13    | A             |
| 1                                              | 7 | 2 | 0 | 10    | B             |
| 0                                              | 4 | 0 | 0 | 4     | C             |
| 14                                             | 11| 2 | 2 | 27    |               |

| Classifier Model with the Supervised-Filter Instance: Resample | Classified as |
|---------------------------------------------------------------|---------------|
| 10 13 A                                                        |               |
| 0   9 1 B                                                      |               |
| 0   2 2 C                                                      |               |
| 10 13 4                                                        |               |

The results showed that the classification using the naive Bayes algorithm was better than J48, in the approach without the resample filter. It classified the minority class (C) better and presented a higher degree of agreement (Kappa statistic). It also indicated high performance by the Matthews correlation coefficient concerning the J48 decision tree algorithm. However, the use of the resample filter in both algorithms improved the distribution of classes in the confusion matrix (Tables 6 and 7), especially Class C.

4. Discussion

The decision tree signalized the production state importance. The E-EER value was higher for the most productive states (SP, MS). However, Brazilian ethanol producers’ certification was still small concerning the high number of states and the volume of ethanol produced in Brazil. The classification of the energy-environmental efficiency of Goiás state did not show its status as one of the states that produces the most ethanol in Brazil [50]. The number of certified producers was still less than it should be. Such a public policy (RenovaBio) was implemented in 2018, so far with little adhesion by the producers and importers in producing states.

The comparison of the J48 decision tree and naive Bayes algorithms, applied to the dataset without using filters, showed differences in the performance indicators. The naive Bayes algorithm, when compared to the J48 decision tree, presented better performance results, predicted Class C better (50% higher in the TP rate), and had better learning capacity (higher Kappa statistic value), indicating that the prediction model had a good performance with this algorithm. However, the sensitivity of the model decreased. The application of the Resample filter in the pre-processing of data for classification with the J48 decision tree algorithm was the one that showed the best performance compared to naive Bayes, as the Kappa statistic was high and the MMC for all classes was higher, indicating that the prediction model performed well.

The data mining technique is also applied in the area of environmental impacts of sugarcane production, for predicting the energy produced and the environmental impacts. Artificial intelligence methods, artificial neural networks, and adaptive neural fuzzy inference system models are also used to predict the environmental impacts of the life cycle and energy output of sugarcane production on planted farms [51].

Integrated approaches based on complex systems for forecasting the growth of sugarcane based on meteorological parameters using extreme machine learning and neural networks were able to show a more generalized model of forecasting for the growth of sugarcane, bringing benefits to industry and the community [52].

Brazil is a major producer of sugarcane and a major consumer of ethanol, with intensive production in order to meet the demands of biofuel (clean energy) and to reduce the use of fossil
fuel (oil). However, several efforts in public policy must come into synergy regarding the growing consumption of biofuel and the sustainable development of the sugarcane chain, mainly regarding its energy-environmental efficiency, use of inputs, and agricultural processes [5]. The ethanol energy-environmental performance labels are applicable in this context, as they would help consumers understand how the chain is evaluated mainly in terms of environmental impact and also collaborating with the transparency of public sector policies.

The most evaluated and used energy efficiency labels are for buildings, appliances, equipment, and lamps. These labeling systems are part of public energy-saving policies; since their implementation, there have been improvements in the evaluation standard, and they have also impacted consumption behavior [53–60]. The implementation of energy efficiency measures can guarantee a sustainable economy. In this context, the energy efficiency labeling program for buildings is generally designed with performance processes and standards, with a rigorous database. Eco-labeling is another system with an approach based on the environmental performance of products that also influences consumer choice [61]. Batista et al. [57] investigated the contribution of labeling to reducing the electricity consumption of buildings and noted that conventional buildings that adopt measures such as painting walls and ceilings white, in addition to using smoked glass, were sufficient to raise the rating to an A level.

The evaluation of buildings generally includes energy classification schemes and shows the difference between the Brazilian scheme and those applied by other countries to improve the labeling method [62]. Other studies have included a review of international building energy efficiency codes and labeling schemes to establish standards for the assessment and classification of buildings in terms of energy performance [63]. Lopes et al. [60] reviewed studies on energy efficiency policies and regulations for buildings, highlighting how the Brazilian program can be improved compared to the American and Portuguese programs; however, this labeling system does not inform about the reduction of GHG emissions.

Another study evaluated two new proposals for an energy efficiency label and a new method for assessing the energy efficiency of public lighting systems. The main difference between the proposals was the number of parameters evaluated. However, the current evaluation system evaluates only one parameter (energy efficiency index), and the study’s proposals recommend five parameters: lamps, energy efficiency index, light pollution, renewable energy contribution, and light dimming [64].

All of these studies demonstrated that the labeling system can be implemented and improved for consolidation and active contribution to consumer behavior. However, it could also increase the contribution of the ethanol production sector with greater participation with the goal of reducing greenhouse gases and reducing energy use. The biofuel sector could benefit from the implementation of an energy-environmental labeling system. With a successful approach, it could increase adherence to the RenovaBio program and consequently increase the sector’s decarbonization credits.

5. Conclusions

The Brazilian National Biofuel Policy (RenovaBio) is one of the main strategies to encourage the reduction of pollutants in the renewable energy sector from sugarcane. For this reason, the implementation of simple labeling can impact consumer behavior and increase the transparency of the incentive program to reduce environmental impacts.

After testing two classifiers, the best model evaluated was naive Bayes without the use of the resample filter compared to the J48 decision tree, also with the use of the resample filter. The classification of the Energy Efficiency Note (RenovaBio) levels using a Bayesian classifier, the naive Bayes algorithm, produced a model capable of predicting the efficiency level of Brazilian ethanol producers and importers certified to create labeling. The rules generated by the models were capable of estimating the classes according to the scale discretized into high efficiency (Classification A), average efficiency (Classification B), and standard efficiency (Classification C), with more accurate
forecasts for the observed classes. These results could generate an ethanol energy-environmental efficiency label for the end consumers and resellers of the product.

However, RenovaBio’s database of ethanol was small, concerning the records of efficiency scores registered in the program, as adherence to the program is still voluntary, and the implementation is recent, hindering deeper learning in the classification of labels. Besides, a more in-depth analysis could improve the model’s forecast in the generation of energy-environmental labels for biofuels.

Author Contributions: Conceptualization, N.D.d.S.L. and J.G.M.d.R.; methodology, N.D.d.S.L. and I.d.A.N.; software, N.D.d.S.L. and I.d.A.N.; validation, N.D.d.S.L., I.d.A.N., J.G.M.d.R., and R.B.T.R.d.S.; formal analysis, N.D.d.S.L. and I.d.A.N.; investigation, N.D.d.S.L., J.G.M.d.R., and R.B.T.R.d.S.; resources, I.d.A.N. and R.B.T.R.d.S.; data curation, N.D.d.S.L. and J.G.M.d.R.; writing, original draft preparation, N.D.d.S.L. and I.d.A.N.; writing, review and editing, I.d.A.N., J.G.M.d.R., and R.B.T.R.d.S.; visualization, I.d.A.N., J.G.M.d.R., and R.B.T.R.d.S.; supervision, I.d.A.N. and J.G.M.d.R.; project administration, N.D.d.S.L. and R.B.T.R.d.S.; funding acquisition, I.d.A.N. All authors read and agreed to the published version of the manuscript.

Funding: This research was funded by Brazilian funding organizations: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Capes Grant Number: 88882315049/2019-01, and Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPQ Grant Number: 159842/2019-0

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cabral, O.M.; Freitas, H.C.; Cuadra, S.V.; de Andrade, C.A.; Ramos, N.P.; Grutzmacher, P.; Galdos, M.; Packer, A.P.C.; da Rocha, H.R.; Rossi, P. The sustainability of a sugarcane plantation in Brazil assessed by the eddy covariance fluxes of greenhouse gases. *Agric. For. Meteorol.* 2020, 282–283, 107864. [CrossRef]

2. Grassi, M.; Pereira, G. Energy-cane and RenovaBio: Brazilian vectors to boost the development of Biofuels. *Ind. Crop. Prod.* 2019, 129, 201–205. [CrossRef]

3. Tucki, K.; Orynychz, O.; Wasiak, A.; Święc, A.; Dybaś, W. Capacity Market Implementation in Poland: Analysis of a Survey on Consequences for the Electricity Market and for Energy Management. *Energies* 2019, 12, 839. [CrossRef]

4. Filoso, S.; Carmo, J.B.d.; Mardegan, S.F.; Lins, S.R.M.; Gomes, T.F.; Martinelli, L.A. Reassessing the environmental impacts of sugarcane ethanol production in Brazil to help meet sustainability goals. *Renew. Sustain. Energy Rev.* 2015, 52, 1847–1856. [CrossRef]

5. de Andrade Junior, M.A.U.; Valin, H.; Sotterroni, A.C.; Ramos, F.M.; Halog, A. Exploring future scenarios of ethanol demand in Brazil and their land-use implications. *Energy Policy* 2019, 134, 110958. [CrossRef]

6. Cardozo, N.P.; de Oliveira Bordonal, R.; La Scala, N. Sustainable intensification of sugarcane production under irrigation systems, considering climate interactions and agricultural efficiency. *J. Clean. Prod.* 2018, 204, 861–871. [CrossRef]

7. Kumar, A.; Ogita, S.; Yau, Y.Y. (Eds.) *Biofuels: Greenhouse Gas Mitigation and Global Warming*; Springer: New Delhi, India, 2018. [CrossRef]

8. Sarangi, P.K.; Nanda, S.; Mohanty, P. (Eds.) *Recent Advancements in Biofuels and Bioenergy Utilization*; Springer: Singapore, 2018. [CrossRef]

9. Isabella, G.; Yu, A.S.O.; Silva, A.M.d.; Pegetti, A.L. Another driver of the Brazilian fuel ethanol supply chain: The consumers’ preferences. *Rev. De Adm.* 2017, 52, 304–316. [CrossRef]

10. Souza, S.P.; Seabra, J.E. Environmental benefits of the integrated production of ethanol and biodiesel. *Appl. Energy* 2013, 102, 5–12. [CrossRef]

11. Santos, A.S.; Gilio, L.; Halmenschlager, V.; Diniz, T.B.; Almeida, A.N. Flexible-fuel automobiles and CO₂ emissions in Brazil: Parametric and semiparametric analysis using panel data. *Habitat Int.* 2018, 71, 147–155. [CrossRef]

12. ANP. Política Nacional de Biocombustíveis. Available online: http://www.anp.gov.br/producao-de-biocombustiveis/renovabio (accessed on 10 December 2019).

13. Brito, T.L.F.; Islam, T.; Stettler, M.; Mouette, D.; Meade, N.; Moutinho dos Santos, E. Transitions between technological generations of alternative fuel vehicles in Brazil. *Energy Policy* 2019, 134, 110915. doi:10.1016/j.enpol.2019.110915. [CrossRef]
33. ANP. Resolução ANP Nº 758 de 2018—Regulamenta a Certificação da Produção ou Importação Eficiente de Biocombustíveis de que trata o art. 18 da lei Nº 13.576, de 26 de Dezembro de 2017, e o Credenciamento de Firmas Inspetoras. Available online: http://legislacao.anp.gov.br/?path=legislacao-anp/resol-anp/2018/novembro&item=ranp-758-2018 (accessed on 10 December 2019).

34. Brazil. Decreto Nº 9.308, 15 de Março de 2018. Metas Compulsórias Anuais de Redução de Emissões de Gases Causadores do Efeito Estufa para a Comercialização de Combustíveis de que trata a Lei nº 13.576, de 26 de Dezembro de 2017. Available online: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/Decreto/D9308.htm (accessed on 18 March 2019).

35. Larose, D.T.; Larose, C.D. Data Mining and Predictive Analytics, 2nd ed.; Wiley Series on Methods and Applications in Data Mining; John Wiley & Sons Inc: Hoboken, NJ, USA, 2015.

36. Frank, E.; Hall, M.; Holmes, G.; Kirkby, R.; Pfahringer, B.; Witten, I.H.; Trigg, L. Weka-A Machine Learning Workbench. In Data Mining and Knowledge Discovery Handbook; Maimon, O., Rokach, L., Eds.; Springer: Boston, MA, USA, 2010; pp. 1269–1277. doi:10.1007/978-0-387-08923-4_66. [CrossRef]

37. Ngai, E.W.T.; Xiu, L.; Chau, D.C.K. Application of data mining techniques in customer relationship management: A literature review and classification. Expert Syst. Appl. 2009, 36, 2592–2602. doi:10.1016/j.eswa.2008.02.021. [CrossRef]

38. Naik, A.; Samant, L. Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. Procedia Comput. Sci. 2016, 85, 662–668. doi:10.1016/j.procs.2016.05.251. [CrossRef]

39. Holmes, G.; Donkin, A.; Witten, I.H. WEKA: A machine learning workbench. In Proceedings of the ANZIS’94—Australian New Zealand Intelligent Information Systems Conference, Brisbane, Australia, 29 November–2 December 1994; pp. 357–361. doi:10.1109/ANZIS.1994.396988. [CrossRef]

40. Witten, I.H.; Frank, E.; Trigg, L.E.; Holmes, G.; Cunningham, S.J. Weka Practical machine learning tools and techniques with Java implementations. In Working Paper; University of Waikato, Department of Computer Science: Hamilton, New Zealand, 1999.

41. Markov, Z.; Russell, I. An introduction to the WEKA data mining system. ACM SIGCSE Bull. 2006, 38, 367–368. doi:10.1145/1140123.1140127. [CrossRef]

42. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. ACM SIGKDD Explor. Newsl. 2009, 11, 10. doi:10.1145/1656274.1656278. [CrossRef]

43. Provost, F.; Fawcett, T. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, 1st ed.; O’Reilly: Beijing, China, 2013.

44. Yang, X.S. Introduction to Algorithms for Data Mining and Machine Learning; Academic Press: London, UK, 2019.

45. Pérez-Ortiz, M.; Jiménez-Fernández, S.; Gutiérrez, P.; Alexandre, E.; Hervás-Martínez, C.; Salcedo-Sanz, S. A Review of Classification Problems and Algorithms in Renewable Energy Applications. Energies 2016, 9, 607. doi:10.3390/en9080607. [CrossRef]

46. Arar, M.F.; Ayan, K. A feature dependent Naive Bayes approach and its application to the software defect prediction problem. Appl. Soft Comput. 2017, 59, 197–209. doi:10.1016/j.asoc.2017.05.043. [CrossRef]

47. Maimon, O.; Last, M. Post-Processing of Data Mining Results. In Knowledge Discovery and Data Mining; Series Title: Massive Computing; Springer: Boston, MA, USA, 2001; Volume 1, pp. 53–59. doi:10.1007/978-1-4757-3296-2_4. [CrossRef]

48. Bowes, D.; Hall, T.; Gray, D. Comparing the performance of fault prediction models which report multiple performance measures: recomputing the confusion matrix. In Proceedings of the 8th International Conference on Predictive Models in Software Engineering—PROMISE ’12, Lund, Sweden, 21–22 September 2012; pp. 109–118. doi:10.1145/2365324.2365338. [CrossRef]

49. Sim, J.; Wright, C.C. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Phys. Ther. 2005, 85, 257–268. doi:10.1093/ptj/85.3.257. [CrossRef]

50. ANP. Dados Estatisticos. Available online: http://www.anp.gov.br/dados-estatisticos (accessed on 20 March 2020).

51. Kaab, A.; Sharifi, M.; Mobli, H.; Nabavi-Pelesaraei, A.; Chau, K.W. Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production. Sci. Total. Environ. 2019, 664, 10105–10119. doi:10.1016/j.scitotenv.2019.02.004. [CrossRef] [PubMed]
52. Taherei Ghazvinei, P.; Hassanpour Darvishi, H.; Mosavi, A.; Yusof, K.B.W.; Alizamir, M.; Shamshirband, S.; Chau, K.W. Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network. Eng. Appl. Comput. Fluid Mech. 2018, 12, 738–749. doi:10.1080/19942060.2018.1526119. [CrossRef]

53. Vine, E.; du Pont, P.; Waide, P. Evaluating the impact of appliance efficiency labeling programs and standards: process, impact, and market transformation evaluations. Energy 2001, 26, 1041–1059. doi:10.1016/S0360-5442(01)00053-6. [CrossRef]

54. Wiel, S.; McMahon, J.E. Energy-Efficiency Labels and Standards: A Guidebook for Appliances, Equipment, and Lighting, 2nd ed.; Technical Report LBNL–45387—2nd-Edition, LBNL-45387, 877316; Collaborative Labeling and Appliance Standards Program (CLASP): Washington, DC, USA, 2005. doi:10.2172/877316. [CrossRef]

55. Lee, S.E.; Rajagopalan, P. Building energy efficiency labeling programme in Singapore. Energy Policy 2008, 36, 3982–3992. doi:10.1016/j.enpol.2008.07.014. [CrossRef]

56. Mills, B.; Schleich, J. What’s driving energy efficient appliance label awareness and purchase propensity? Energy Policy 2010, 38, 814–825. doi:10.1016/j.enpol.2009.10.028. [CrossRef]

57. Batista, N.d.N.; Rovere, E.L.L.; Aguiar, J.C.R. Energy efficiency labeling of buildings: An assessment of the Brazilian case. Energy Build. 2011, 43, 1179–1188. doi:10.1016/j.enbuild.2010.11.010. [CrossRef]

58. Lopes, M.N.; Iwamoto, G.; Riella, H.L.; Lamberts, R.; Campos, L.M.S. The building simulation as an alternative method in brazilian energy efficiency labeling of commercial buildings. In Proceedings of the Building Simulation 2011 Conference, Sydney, Australia, 14–16 November 2011.

59. Newell, R.G.; Siikamäki, J. Nudging Energy Efficiency Behavior: The Role of Information Labels. J. Assoc. Environ. Resour. Econ. 2014, 1, 555–598. doi:10.1086/679281. [CrossRef]

60. Lopes, A.d.C.P.; Oliveira Filho, D.; Altoe, L.; Carlo, J.C.; Lima, B.B. Energy efficiency labeling program for buildings in Brazil compared to the United States’ and Portugal’s. Renew. Sustain. Energy Rev. 2016, 66, 207–219. doi:10.1016/j.rser.2016.07.033. [CrossRef]

61. Banerjee, A.; Solomon, B.D. Eco-labeling for energy efficiency and sustainability: A meta-evaluation of US programs. Energy Policy 2003, 31, 109–123. doi:10.1016/S0301-4215(02)00012-5. [CrossRef]

62. Scalco, V.A.; Fossati, M.; de Souza Versage, R.; Sorgato, M.J.; Lamberts, R.; Morishita, C. Innovations in the Brazilian regulations for energy efficiency of residential buildings. Archit. Sci. Rev. 2012, 55, 71–81. doi:10.1080/00038628.2011.641731. [CrossRef]

63. Fossati, M.; Scalco, V.A.; Linczuk, V.C.C.; Lamberts, R. Building energy efficiency: An overview of the Brazilian residential labeling scheme. Renew. Sustain. Energy Rev. 2016, 65, 1216–1231. doi:10.1016/j.rser.2016.06.048. [CrossRef]

64. Gutierrez-Escolar, A.; Castillo-Martinez, A.; Gomez-Pulido, J.M.; Gutierrez-Martinez, J.M.; González-Seco, E.P.D.; Stacic, Z. A review of energy efficiency label of street lighting systems. Energy Eff. 2017, 10, 265–282. doi:10.1007/s12053-016-9454-7. [CrossRef]