Using Machine Learning to Identify Solid Biofuels

Mónica V. Martins[1] (✉), Luiz Rodrigues[1], and Valentim Realinho[1,2]

[1] Polytechnic Institute of Portalegre (IPP), Portalegre, Portugal
mvmartins@ipportelegre.pt

[2] VALORIZA - Research Center for Endogenous Resource Valorization, Portalegre, Portugal

Abstract. To achieve the transition from a linear economy to a circular bioeconomy, a full implementation of the biorefinery concept is needed. The sustainability of biorefineries requires knowledge of the characteristics of sitespecific biomasses and residual biomasses. However, the complexity, variability and seasonality of biomasses pose huge challenges to the conception, assessment and management of biorefinery installations. In this context the classification of biomasses based on machine learning approaches and using proximate analysis data may be helpful in the decision-making process and management of biorefinery units. On the other hand, automatic identification of the sources of solid biofuel material might be advantageous in the context where the material has already been processed and must be used in energy generation or as raw materials of a biorefinery scheme, or when the elementary contents of the materials might not be existent or easily available. In this work, we use a public dataset to build classification models to predict the solid biofuel type from their fixed carbon, volatile matter, and ash contents. The dataset aggregates 585 examples of solid biofuels classified as one of four different types: coals, woods, agriculture residues, and manufactured biomass. Since the dataset presents a strong imbalance towards one of the classes, an algorithm to promote class balancing with synthetic oversampling is used. Then, six different models for biofuel classification are built, tested, and validated. The analysis of the relative contribution to each of the features for the final model is performed. The results show that it is possible to achieve an overall classification accuracy of 90%.

Keywords: Solid biofuels · Biorefinery · Machine learning · Multi-class classification · Random forest · Extreme gradient boosting

1 Introduction

To achieve the transition from a linear economy, based mainly on fossil resources, to a circular bioeconomy, based on renewable resources, a full implementation of the biorefinery concept is needed. Biorefinery is the sustainable integration of processes

P. S. Duque de Brito et al. (Eds.): ICoWEFS 2022, *Proceedings of the 2nd International Conference on Water Energy Food and Sustainability (ICoWEFS 2022)*, pp. 78–85, 2023.
https://doi.org/10.1007/978-3-031-26849-6_9

to transform biomasses into a portfolio of biofuels, bioenergy, and bioproducts, while minimizing or zeroing residues generation [1]. But the ecological, economic, and technological sustainability of biorefineries will require the knowledge of the characteristics of site-specific biomasses and residual biomasses, since ideally generation of energy and production of goods should be decentralized [2]. However, variation in complexity, geographic availability and seasonality of biomasses, and respective characteristics poses huge challenges to the conception and assessment of proposed or and management of running biorefinery installations.

Indeed, the nature and properties of biomasses are of critical importance in the selection of the processing technologies, particularly for its use as fuel. Depending on the nature of biomasses, biofuels and biorefineries may be classified as first generation, if the use of its biomass raw material competes in any form with food or feed production, including land occupation. Second generation biofuels and biorefineries use lignocellulosic, particularly residual, biomasses as row materials. Third generation biofuels and biorefineries applies algae as staring materials [1]. In terms of processing technologies those installations may be classified in thermal, chemical, or biochemical "platforms".

Option for one or more of these platforms depends undoubtedly on the final main products to be obtained, but also on the characteristics of raw biomasses. Evidently a deep knowledge of the starting material, as detailed as possible, particularly, information on the ultimate or elemental analysis may be helpful in conception and implementation of conversion scheme. However ultimate analysis requires expensive equipment and may be frustrating and inefficient [3]. On the other hand, the so-called proximate analysis, i.e., the characterization of biomasses based on its water, fixed carbon, volatile matter, and ashes contents requires less expensive equipment and, therefore, is widely used by researchers in various modelling and prediction studies, also for process designing proposes. For instance, proximate analysis was extensively used to predict higher heating value (HHV), an extremely important parameter for the characterization of fuels, showing that it is a reasonable alternative to ultimate analysis [3].

In this context the classification of biomasses based on machine learning approaches and using proximate analysis data may be helpful in the decision-making process on and management of biorefinery or biofuels producing units. In this work, the methodologies and results for the identification of biofuels from their fixed carbon, volatile matter, and ashes contents using machine learning is presented.

2 Methods

In this section the dataset used is presented, along with the methods used to deal with the unbalanced nature of the data, and the methodology used to train and evaluate the classification models.

2.1 Data

The public dataset presented in [4] was used to build classification models to predict the solid biofuel type from their fixed carbon, volatile matter, and ash contents. The dataset was built from raw fuel composition data presented in a number of scientific publications that were categorized as four different types of solid biofuels: coals, woods, agriculture residues, and manufactured biomass. The raw fuel elements categorized in each class are presented in Table 1.

Table 1. Categorization of raw fuels. From [4].

Class name	Raw fuels
Coals	Coals, charcoals, chars
Woods	Woods, shell, pruning
Agriculture Residues	Seed, husk, leaves, grass, bark, straw, stalk
Manufactured Biomass	Municipal solid waste, RDF, sludge, briquettes

The dataset comprises a total of 585 samples, with unequal distribution among the four classes of biofuels, as represented in Fig. 1.

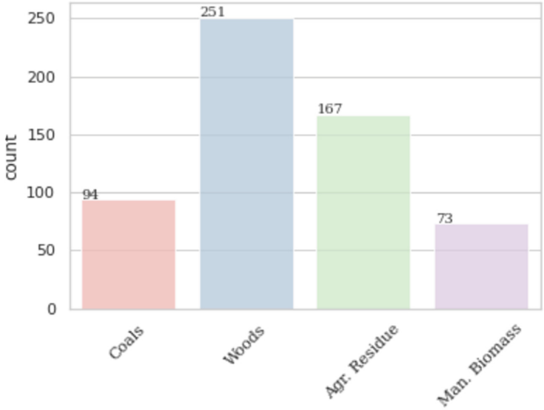

Fig. 1. Distribution of the dataset samples among the four solid biofuel classes.

The features available in the dataset for each sample the contents on fixed carbon, volatile matter, and ash. The kernel density estimate (kde) plots presented in Fig. 2 show the distribution of the contents according to the biofuel class.

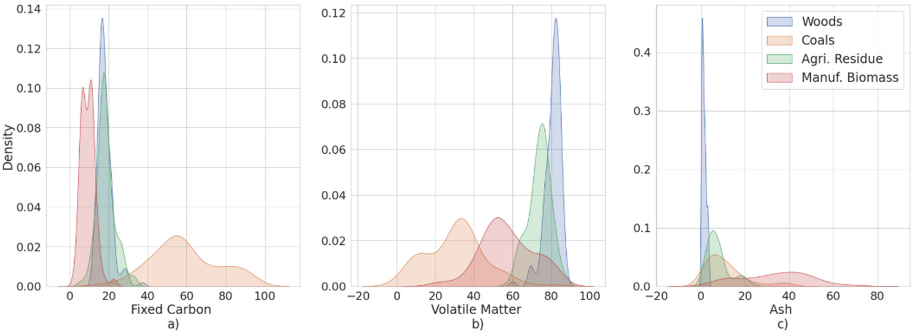

Fig. 2. Density plots for a) Fixed Carbon content, b) Volatile Matter content and c) Ash content. Different colors represent the different biofuel classes.

The kde plots show that both Manufactured Biomasses and Coals present to the naked eye with a good degree of separability from the other contents at least for one of the features, while Woods and Agriculture Residues are mostly overlapping.

The goal is to build classification models that allow to identify the biofuel category of the samples from the three mentioned features. Additionally, there is an appreciable degree of data unbalance, where the most represented class (Woods) represents 43% of the samples, whereas the less represented class (Manufactured Biomass) represents only 12% of the samples.

2.2 Data Sampling

The fact that the dataset presents an appreciable degree of class imbalance might constitute a challenge to the classification task. This happens because the machine learning models will tend to be more biased towards the majority class (or classes), and perform worse for the minority class (or classes) [5].

There are two main approaches to deal with the issue of data unbalance at the data level: either under-sampling or over-sampling. Under-sampling techniques, where some data are eliminated from the majority classes, have the disadvantage of reducing the size of the dataset. Over-sampling techniques can be used by generating new synthetic data. This was the approach followed in this work, where the SMOTE – Synthetic minority over-sampling technique [6] was used. For synthetizing new samples, the SMOTE algorithm starts by identifying neighbor examples from the minority classes in the feature space and then synthetizes a new example in the space between those neighbors. It repeats this procedure as many times as needed to create a balance between the number of samples in the classes.

In this work, the SMOTE implementation available at the *imbalanced-learn* module in the *scikit-learn* library [7] was used, which allows dealing with multiple minority classes, such as the present case.

2.3 Classification Models

Six different classification algorithms were chosen to train the classifiers: Logistic Regression(LR) [8] and Support Vector Machines (SVM) [9] were chosen due to their simplicity, ease to tune, and difficulty to overfit; Decision Trees (DT) [10] model was used because it provides more complex models, although at the expense of a tendency to overfit; Random Forests (RF) [11] and Extreme Gradient Boosting (XGB) [12], two ensemble techniques, although more difficult to tune, usually provide good performance models that are less likely to overfit than Decision Trees. To train these models the implementations available at the Scikit-learn library in Python [13] were used.

2.4 Model Training, Evaluation Metrics

Following the usual procedure, data were divided into a training set (80% of the samples) and a test set (20% of the samples). Then, for each model, a 5-fold cross validation procedure was used to avoid overfitting. This means that the training data set was divided into 5 blocks, and the training of each model was done with 4 of the blocks, with the remaining one being used for validation purposes. The process was repeated 5 times, once for each block, thus enabling the maximization of the total number of observations used for validation. The best average cross-validation estimator score was elected. Then, the overall performance of each elected model was assessed with the test set.

Accuracy (Eq. 1) is the global metric often used for evaluating classification models, computed as the number of correct predictions divided by the total number of predictions. It might however become an unreliable metric in the case of unbalanced data. In such cases, single class metrics are more adequate [14], since they allow for a better understanding of how the models behave for each class. In this work the f1-score (Eq. 4) is used, which accounts for the trade-off between precision (Eq. 2) and recall (Eq. 4), which also represent single class metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$f1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

For each trained model, f1-score, precision and recall were computed for each class, as well as the accuracy, and the average-f1score.

In the previous equations, TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative. For each model, these values were also plotted in a confusion matrix, a table that summarizes the performance of the classification model for each class.

For each trained model, f1-score, precision, and recall were computed for each class, as well as the accuracy, and the average f1-score.

3 Results

The confusion matrices obtained for each of the trained models are presented in Fig. 3**.** The dark green values along the main diagonal represent the percentage of True Positive values for each class. For each row, the values outside the diagonal represent the percentage of False Negative values for the respective class; for each column, the values outside the diagonal represent the percentage of False Positive values for the respective class.

From these matrices, it's possible to identify that both Coals and Woods are easier to correctly classify, regardless of the classifier used, than the other two biofuel classes. Coals is also one of the minority classes. The fact that every classification model presents high f1-score for this class is an indication that the data sampling technique achieved its goals. Agricultural Residues are often misclassified as Woods, which was expectable, given their content superposition, and also misclassified as Manufactured Biomasses. On the other hand, Manufactured Biomasses can be misclassified as Agriculture Residues.

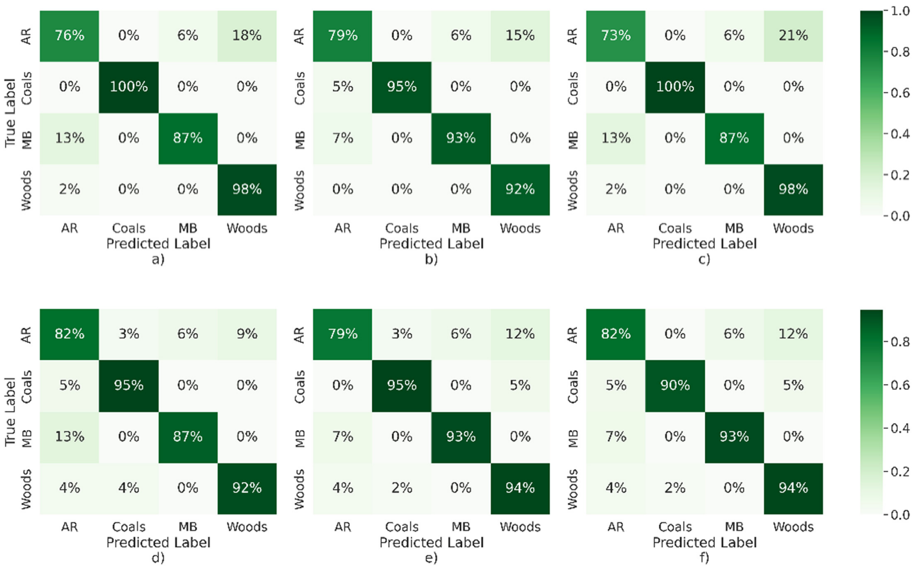

Fig. 3. Confusion matrix for the models a) Logistic Regression; b) k-Nearest Neighbors; c) Support Vector Machines; d) Decision Tree; e) Random Forest; e) Extreme Gradient Boosting.

The values for f1-score obtained with the six classification models for each class are presented in Table 2.

The f1-score values for the individual classes confirm that, in general, all models classify Coals and Woods with high f1-score (above 0.91), and that Agricultural Residues and Manufactured Biomasses seem to be harder to correctly classify, with lower f1-scores for those classes (below 0.91).

Finally, Table 3 presents the global metrics obtained for each of the trained models: accuracy, and the average values of f1-score, precision and recall.

Table 2. f1-score values for the individual classes, for the several trained models.

Biofuel class	LR	kNN	SVM	DT	RF	XGB
AR	0.82	0.80	0.80	0.83	0.84	0.84
Coals	1.00	0.97	1.00	0.90	0.93	0.92
MB	0.87	0.90	0.87	0.87	0.90	0.90
Woods	0.93	0.91	0.92	0.93	0.92	0.92

Table 3. Global metrics obtained with the trained models: average f1-score, average precision, average recall, and accuracy.

Model	f1-score	Precision	Recall	Accuracy
Logistic Regression	0.89	0.89	0.88	0.87
k-Nearest Neighbors	0.90	0.90	0.90	0.89
Support Vector Machines	0.88	0.89	0.88	0.88
Decision Tree	0.84	0.89	0.83	0.86
Random Forest	0.90	0.90	0.90	0.90
Extreme Gradient Boosting	0.90	0.90	0.90	0.90

Among the six trained classifiers, Random Forest and Extreme Gradient Boosting stand with the highest performance metrics, followed by the kNN model.

4 Discussion

In this work, a public dataset was used to build six classification models to predict the type of biofuel. The problem is addressed as a four-category classification task, in which there's an imbalance towards two of the classes. The SMOTE algorithm is used to promote class balancing with synthetic oversampling and six machine learning classification models are trained. The results show that both Random Forest and Extreme Gradient Boosting have the highest classification accuracy and that the Agriculture Residues is the biofuel type that is harder to correctly identify.

References

1. Cherubini, F.: The biorefinery concept: using biomass instead of oil for producing energy and chemicals. Energy Convers. Manag. **51**, 1412–1421 (2010). https://doi.org/10.1016/j.enconman.2010.01.015
2. Irena: BOOSTING BIOFUELS Sustainable Paths to Greater Energy Security Acknowledgements (2016)
3. Elmaz, F., Yücel, Ö.: Data-driven identification and model predictive control of biomass gasification process for maximum energy production. Energy **195** (2020). https://doi.org/10.1016/j.energy.2020.117037

4. Elmaz, F., Büyükçakır, B., Yücel, Ö., Mutlu, A.Y.: Classification of solid fuels with machine learning. Fuel 266 (2020). https://doi.org/10.1016/j.fuel.2020.117066
5. Ali, A., Shamsuddin, S.M., Ralescu, A.L.: Classification with class imbalance problem: a review. Int. J. Adv. Soft Comput. its Appl. **7**, 176–204 (2015)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002). https://doi.org/10.1613/jair.953
7. Lema, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. **40**, 1–5 (2015)
8. Hastie, T.J., Pregibon, D.: Generalized linear models. In: Statistical Models in S (2017)
9. Cortes, C., Vapmik, V.: Support-Vector Networks. Mach. Learn. **20**, 273–297 (1995). https://doi.org/10.1111/j.1747-0285.2009.00840.x
10. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**, 81–106 (1986). https://doi.org/10.1007/bf00116251
11. Breiman, L.: Random forests. Random For. **45**, 5–32 (2001). https://doi.org/10.1201/9780367816377-11
12. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
13. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **85**, 2825 (2011)
14. Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., Asadpour, M.: Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data **7**(1), 1–47 (2020). https://doi.org/10.1186/s40537-020-00349-y