sustainability

MDPI

Article

Implementing Artificial Intelligence Techniques to Predict Environmental Impacts: Case of Construction Products

Anish Koyamparambath [1,2], Naeem Adibi [2], Carolina Szablewski [2], Sierra A. Adibi [3] and Guido Sonnemann [1,*]

[1] Institute of Molecular Sciences, University of Bordeaux, Centre National de la Recherche Scientifique, Bordeaux INP, ISM, UMR 5255, 33400 Talence, France; anish.koyamparambath@u-bordeaux.fr
[2] WeLOOP, 254 Rue de Bourg, 59130 Lambersart, France; n.adibi@weloop.org (N.A.); c.szablewski@weloop.org (C.S.)
[3] William E Boeing Department of Aeronautics and Astronautics, University of Washington, Seattle, WA 98195, USA; sierra.adibi@gmail.com
* Correspondence: guido.sonnemann@u-bordeaux.fr

Abstract: Nowadays, product designers, manufacturers, and consumers consider the environmental impacts of products, processes, and services in their decision-making process. Life Cycle Assessment (LCA) is a tool that assesses the environmental impacts over a product's life cycle. Conducting a life cycle assessment (LCA) requires meticulous data sourcing and collection and is often time-consuming for both practitioner and verifier. However, predicting the environmental impacts of products and services can help stakeholders and decision-makers identify the hotspots. Our work proposes using Artificial Intelligence (AI) techniques to predict the environmental performance of a product or service to assist LCA practitioners and verifiers. This approach uses data from environmental product declarations of construction products. The data is processed utilizing natural language processing (NLP) which is then trained to random forest algorithm, an ensemble tree-based machine learning method. Finally, we trained the model with information on the product and their environmental impacts using seven impact category values and verified the results using a testing dataset (20% of EPD data). Our results demonstrate that the model was able to predict the values of impact categories: global warming potential, abiotic depletion potential for fossil resources, acidification potential, and photochemical ozone creation potential with an accuracy (measured using R^2 metrics, a measure to score the correlation of predicted values to real value) of 81%, 77%, 68%, and 70%, respectively. Our method demonstrates the capability to predict environmental performance with a defined variability by learning from the results of the previous LCA studies. The model's performance also depends on the amount of data available for training. However, this approach does not replace a detailed LCA but is rather a quick prediction and assistance to LCA practitioners and verifiers in realizing an LCA.

Keywords: life cycle assessment; environmental product declaration; artificial intelligence; machine learning; environmental performance

Citation: Koyamparambath, A.; Adibi, N.; Szablewski, C.; Adibi, S.A.; Sonnemann, G. Implementing Artificial Intelligence Techniques to Predict Environmental Impacts: Case of Construction Products. *Sustainability* **2022**, *14*, 3699. https://doi.org/10.3390/su14063699

Academic Editors: Erwin M. Schau and Eva Prelovšek Niemelä

Received: 28 December 2021
Accepted: 8 March 2022
Published: 21 March 2022

1. Introduction

Buildings account for 40% of primary energy consumption and 36% of Greenhouse Gas emissions across Europe. Decreasing the environmental impacts of buildings is key to realizing EU 2020 objectives of a 20% reduction of GHG and energy consumption. While the building sector understands the importance of energy efficiency, the environmental impacts of the building products remain less known.

These environmental impacts combined with regulation policies by countries have prompted us to investigate the life cycle of each of these products. Life Cycle Assessment (LCA) is among the most powerful analytical tools for evaluating the environmental impacts of a product, process, or service over its entire life cycle [1]. The EU has built a framework, Level(s), to integrate LCA to create a sustainable framework as a part of the EU's transition towards net carbon neutrality. Although Level(s) is not a certification scheme, a common

way of communicating the LCA results to the customers and stakeholders is to use the ISO 14025 type III Environmental Product Declaration (EPD) [2]. An EPD provides information about the environmental impacts of varied materials and products. It can relate to products manufactured by one or several manufacturers.

Conducting an LCA study is time-intensive and requires meticulous analysis and systematic investigation to model the product system. With the increase in the need to understand the environmental impacts of the evolving building products, businesses and stakeholders are searching for a framework to estimate the environmental impacts to near accurate values. In such cases, artificial intelligence (AI) comes into play. However, there are currently no options available to businesses and stakeholders to estimate their product's environmental impacts and hotspots quickly and accurately. In addition, predicting the environmental impacts of a product or a product system requires detailed information.

Various methods were employed to predict the environmental impacts of products and product systems, such as the game theory (GT) agent-based model (ABM). GT is a widely used scientific tool in economics, biology, social sciences, and policy. This tool presents scenarios where several players choose actions from a given set of strategies. Jose et al. modeled stakeholders in the LCA as a player, an individual, a group, or a corporation. Each strategy permutation is associated with a payoff, and the game aims to maximize the payoff [3].

ABM and LCA have been used together in multiple cases [4]. Eric Bonabeau, in his article, defines ABM as a type of microscopic modeling that uses individual decision-makers called agents. The agents assess a situation based on a predefined ruleset [5,6] and use the Agent-Based Model (ABM) to quantify the environmental impacts of a non-established emerging system [6]. The decision of the agents is determined using Bayesian probability. A case study of switchgrass cultivation in the United States of America was used as an emerging energy crop during the study [6]. A comparative study between ABM and GT was also conducted favoring both the methods for prediction depending on certain circumstances. Micolier et al. presented how ABM can contribute to LCA by reducing the uncertainties in foreground inventory data [7].

AI is a discipline that envelopes everything that makes a machine intelligent. Machine learning (ML), a subset of AI, refers to mathematical and statistical algorithms designed to learn from existing datasets to improve future performance and is used widely in many real-world applications [8]. In this paper, our reference to AI is always related to the use of ML. A review on the application of ML for LCA of buildings by Barros and Ruschel showed that related research has been increasing in recent years [9]. Their study reviewed 15 articles and identified that artificial neural networks (ANN), support vector machine (SVM), bayesian network (BN), and genetic algorithm (GA) are among the used ML techniques in the selected articles. The studies focused on optimizing the performance of LCA, supporting decision-making, and impact prediction [9]. Almost 47% of the articles used ANN as their primary ML technique in their study [9].

ANN is a learning paradigm comprising a network of node layers. The concept is based on the human brain, with each node representing a neuron. A typical neural network consists of an input and an output layer with a hidden layer, which does the math. They are utilized in various applications, including LCA. Nabavvi-Pelesaraei et al. used ANN to forecast paddy production's environmental indicators and energy output with energy consumption as inputs [10]. In their subsequent study, they were able to predict the environmental indicators of sugar cane production with excellent prediction accuracy (>90%) [11]. The data for both the studies were energy input/output (human labor, machinery, fuel fertilizers, etc.) in various operations involved in paddy and sugarcane productions [10,11].

Although ANN is known for its high prediction accuracy, various parameters influence its results. One of the critical parameters is the required size of data. The LCA database of products is extensive but not harmonized, including various methods used to conduct an LCA and several assumptions. EPDs, on the other hand, are harmonized and follow a

set of predefined rules, which makes it easier to implement AI. The fact that it is publicly available is also an advantage.

ANN is a computing-intensive algorithm that would require much time for training, provided the parameters are ideal. Furthermore, the analysis and calculation between the layers are hidden and cannot be controlled, making it difficult to control the prediction. On the other hand, mathematical and statistics-based algorithms such as multiple linear regression, Bayes classifier, and decision tree regression are among the widely used prediction methods. The main advantage of these algorithms is controlling the quality of prediction. Of course, these algorithms have their limitations regarding prediction accuracy, and in most cases where the data was huge, ANN outperformed these algorithms [12,13]. The argument for choosing a mathematical algorithm over ANN is the availability of data and the ability to control the prediction. In the application of AI in LCA, a study by Hou et al. used the machine learning models K nearest neighbors (KNN), SVM, neural networks (NN), random forest (RF), adaptive boosting (Adaboost), and gradient boost machine (GBM) to predict the characterization factor of ecotoxicity. They concluded that RF was the best ML algorithm for predictive performance among the benchmarked methods [14].

Existing studies mostly correspond to using AI techniques within the scope of LCA. The inputs used in such studies are unique and uncommon, which involves additional data gathering. An AI-based instant prediction model that requires minimal product or service data is missing. Our article aims at explaining and highlighting a newly developed method to predict the environmental impacts of a product or service by learning from EPDs of construction products. Based on our knowledge, our method is the first attempt to estimate the values of four impact categories; global warming potential, abiotic depletion potential for fossil resources, acidification potential, and photochemical ozone creation potential, based on previous LCA studies and results. These four impact categories are selected among the seven total indicators available in the database for their different level of robustness according to the European commission EF 3.0 [15].

This paper is organized as follows. The Materials and Methods section presents the sources of the data used and the Artificial Intelligence (AI) techniques applied; the Results and Discussion section present the predicted values of the training, including the limitations of the method and recommendations for future use and developments.

2. Materials and Methods

2.1. Data Source: EPD

An EPD can be elaborated for any product; today, they are available on a large scale for construction products following the EN 15804: A1 standard [16]. EPDs are implemented to provide quantifiable environmental information about the product's life cycle, enabling the user to assess the environmental impacts [17,18]. EPDs are built on the guidance provided by the Product Category Rules (PCRs) for unbiased comparison of products of the same function [19,20]. In 2019, the standard EN 15804 was aligned towards the Product Environmental Footprint (PEF), proposing a new set of indicators, published under the second amendment of the EN 15804: A2 (2019) [21].

EPDs are commonly published on websites governed by policies set by stakeholders, such as governments, industry associations, or NGOs. EPD data based on EN 15804: A1 is available to the public, enabling the user to consider the environmental impacts of construction products and buildings [22]. In Germany, an association of building product manufacturers called Institut Bauen und Umwelt e.V (IBU) [23] publishes the EPDs. IBU is approved by the Federal Ministry of the Interior Building and Community to publish EPDs. EPD data based on EN 15804: A1 [16] are available online in a standardized database, the ÖKOBAUDAT platform [23]. Similarly, European countries like France have their sector-specific databases of EPDs called INIES, also available to the public [24].

Our paper used EPD results from construction products available in ÖKOBAUDAT database. The data in all EPDs are harmonized since they follow the EN 15804: A1 standard. The use of harmonized data sources is a critical factor in our study, and the

EPDs in ÖKOBAUDAT are available to download into Extensible Markup Language (XML) files [25].

The ÖKOBAUDAT platform provides EPD datasets in XML and Hypertext Markup Language (HTML). The EPD data format in ÖKOBAUDAT is compliant with EN 15804: A1 and the International Reference Life Cycle Data System (ILCD) [26]. These EPD results are derived using GaBi [27] background database [28]. The EPDs of the database have categorical and descriptive information about the construction products, and these elements are used as a basis to classify and assess these products [17].

Each EPD contains vital information describing the data: the name of the process data set, the location of the EPDs, the classification levels and description of the product/service, and the quantitative reference used to study the product/service. They also contain information about the source, owners, and developers of the EPD. In addition, the environmental impact categories, calculated as part of the life cycle impact assessment, are provided individually for each life cycle stage [28].

2.2. Data Collection and Pre-Processing

The EPDs were downloaded using an automated web-scraping tool developed in python using Selenium 3.1 [29]. The XML files were then parsed into a consolidated database using SQLite 3 [30]. The information from the EPD used for our method is the name of the product/service, classification of the product/materials, geographic location of the study, quantitative reference of the study, and the impact assessment results for a given impact category. This descriptive and categorical information was used as inputs to the ML algorithm that predicts the impact assessment results for a given category.

The ML algorithm is trained iteratively to predict the results of impact categories. The data collected is split into two subsets: the training and testing datasets. The training dataset is used to train the ML model, and the model is then validated by comparing the predicted and actual values from the testing dataset. Finally, the ML algorithm's hyperparameters are manipulated to improve results [31].

Data collection is a well-organized procedure that involves a lot of precision and accuracy. A collection of compiled data based on some criteria, known as corpora, can be typically extracted from several sources, while a corpus is a collection from a single source [32]. Text cleaning and encoding is the main task of data collection. In the EPD, the name and description are unique data corresponding to a product where categorization is impossible.

Data was collected from 1188 EPDs available on the ÖKOBAUDAT platform. EPDs contained the LCA data for both products and services. The scope of the assessment results published in the EPDs is not all from cradle to grave. Few of the EPDs are focused on the product's end of life. When counted, at least 90% of the downloaded EPDs contain the LCA results of the production stage (A1 to A3). At the end of processing, usable data after removing duplication and null values were around 980 EPDs. The processed dataset has 980 EPDs with 7 vital information: Name/description, location, 3 classification levels, functional unit, values of selected impact category. The descriptive information (i.e., "Name/description") must be characterized to be used in the algorithm. The characterization of the information is done using algorithms from a field of AI called Natural Language Processing (NLP).

2.3. Natural Language Processing (NLP)

NLP, a subfield of AI, concerns processing a large amount of text data by converting them into features that can be used in different machine learning algorithms [33]. The procedure for text processing is normalization, lemmatization, and encoding. Text normalization is a procedure to convert the text into a standardized form. It involves removing unnecessary characters, expanding abbreviations, and redaction of stop words, such as 'a', 'of', 'from', etc., from the sentences. Lemmatization groups together different words with the same root. For example, reseal and sealing are reduced to "seal" [34]. Finally, the tokenized sentences classified into a bag of words provide a matrix with the count of words

in each sentence. Table 1 represents the encoded matrix of three sentences and the count of words in each sentence.

Table 1. Encoded matrix of the bag of words.

	Product	Seal	Component	Treat	Cool	Steel	Metal	Work	Hot
Name 1	1	1	2	1	0	0	0	1	0
Name 2	1	0	0	1	1	1	1	0	0
Name 3	0	0	1	0	0	1	1	0	1

In our method, the descriptive information from the EPD is processed using NLP to form a corpus of words, which is then counted for the occurrence of words per EPD. The resultant matrix consists of 980 rows to 1353 columns. The remaining categorical data without the values of the selected impact category is encoded to binary variables resulting in a matrix of 980 rows to 243 columns. The combined data is then encoded and stored in a database. As part of the life cycle impact assessment, the result of an environmental impact category is also stored in the database along with the encoded information. The procedure is represented in Figure 1.

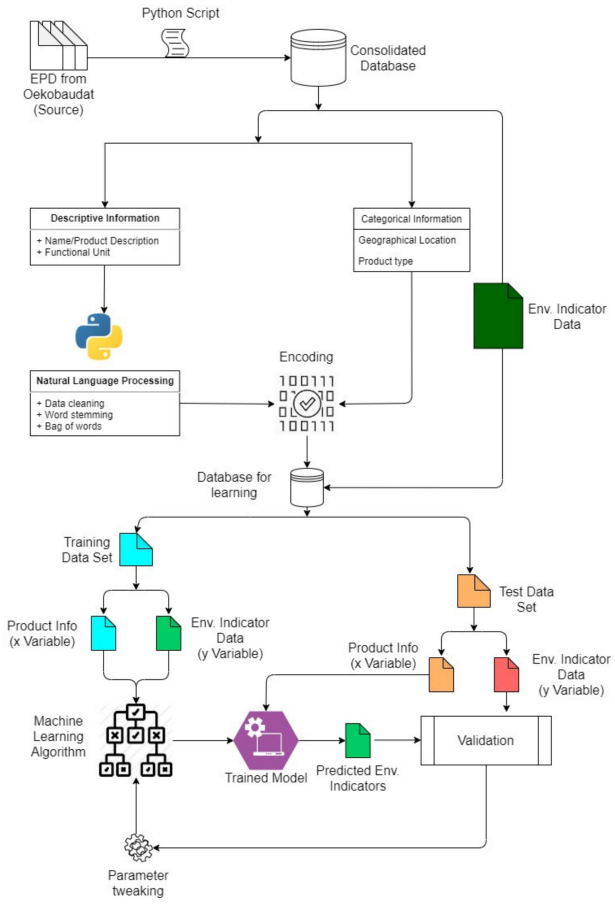

Figure 1. A flow diagram representation of the ensemble method.

Feature selection is an important aspect that determines prediction accuracy. How the selected input features (name, quantification unit, geographical location, classification) correlate with the output is vital to making a choice. More than 95% of the EPDs published in ÖKOBAUDAT have Germany as their geographical location. While considering this an essential input in a homogenous collection of EPDs, this feature (geographical location)

does not complement the output values in this scenario. The IBU has three levels of classification upon which products and services are classified. Although there was no correlation between the individual classification stage and the results, a combination showed a positive correlation.

Only one impact category can be predicted in one instance. Therefore, the values of the impact category are separated from the dataset, and data for one impact category for all the data points are stored as the 'Y' variable. Indicators are transformed using logarithmic transformation to obtain more precise results, and the remaining encoded data matrix is stored as the 'X' variable. The variables are split into two subsets: a training dataset with 80% of the data ('X_{train}' and 'Y_{train}') and a testing dataset ('X_{test}' and 'Y_{test}').

The ML model fits the 'X_{train}' and 'Y_{train}' variables, and the model is then tested using the corresponding 'X_{test}' and 'Y_{test}' data from the testing dataset. The amount of available training data influences the performance of the model. Figure 1 represents the machine learning process flow model using the random forest algorithm.

2.4. Tree-Based Algorithm

A tree-based algorithm splits the dataset based on criteria until an optimal result is obtained. A Decision Tree (DT) is a classification and regression tree-based algorithm, which logically combines a sequence of simple tests comparing an attribute against a threshold value (set of possible values) [35]. It follows a flow-chart-like tree structure, where each node denotes a test, and each branch represents an outcome of the test. The node representing the results is the Leaf node [36]. The algorithm involves two major phases: the growth phase, which partitions the given nodes to fit each class of the data, and the pruning phase, aiming to generalize the DT to avoid overfitting [35]. The training data fed into the algorithm will train the model and fit each node to a test, and DTs are sensitive to data and more prone to overfitting. Overfit is a concept that represents when an ML model is overly familiarised with the training data and cannot generalize the new dataset, and is thereby unable to predict efficiently [37].

2.5. Random Forest Regression

Random Forest (RF) is an ensemble learning method for classification and regression that constructs many decision trees [38]. They are a combination of tree predictors where each tree depends on a random vector's values sampled independently [39]. RF generates additional data for training from datasets using repetition to produce multisets of original data. In addition, RF is a bagging technique where the generated decision trees learn and predict in parallel and then aggregate (mean prediction). The aggregation is done with modifications by limiting the number of features split on each node, resulting in relying on all features instead of one particular feature.

There are more than 1500 features in the dataset, which could cause the model to overfit. Therefore, the hyperparameters that are modified in our method are max_features: number of features to be considered before splitting, max_depth: defines the maximum depth of the tree, min_samples_split: number of minimum samples required before splitting, min_samples_leaf: the minimum number of samples present at the leaf node [31].

Python has several modules for manipulating the hyperparameters to improve the model's prediction performance. The most common method used to select the optimal set of hyperparameters is the k-fold cross-validation method. It involves splitting the training dataset into k-folds where k-1 folds are used for training, and 1-fold is used to validate the training. The model's performance is computed and repeated with a diverse set of hyperparameters, and the performance of each set is compared to result in an optimal set of hyperparameters. Grid search and random search algorithms are widely used for tuning the model's hyperparameters [31]. In grid search, a set of hyperparameters values are declared. Then, each combination is evaluated and scored using k-fold cross-validation, a resampling procedure used to evaluate models using a limited data sample. In our method, we have used the three-fold cross-validation method [40].

The trained model predicts the chosen indicator's values with the 'X_{test}' variable from the testing dataset. The predicted values are compared with the values of the chosen indicator from the testing dataset. Mean squared error and the R^2 value (the percentage of dependent variable variation a regression model explains) are calculated and discussed in the following section [41].

3. Results and Discussion

Due to the availability of adequate EPD data in the ÖKOBAUDAT platform, AI techniques are used here to predict the environmental impacts of construction products as a case study. In principle, the method developed applies to any product group for which EPDs have been prepared based on agreed-upon PCRs.

As a first part of the analysis, the database of EPDs created after processing the data is split into two datasets. Then, only the input variables are used to predict the values of the selected impact categories. Figure 2 represents the data input to the algorithm and the prediction made by the algorithm.

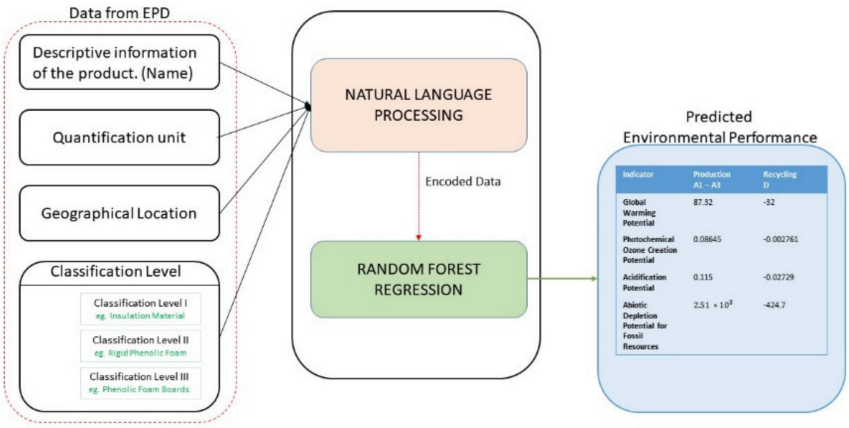

Figure 2. Inputs and outputs of the Machine Learning model.

Results and Performance of the Model

Out of 980 EPDs, 80% (784 EPDs) in random selection were used to train the model in iterations. The grid search algorithm tuned the model's hyperparameters to search for the optimal selection for each iteration. Each indicator has a different numerical range posing a challenge to select an optimal set of hyperparameters. The model's performance is studied by R^2 analysis and mean squared error. The R^2 measure is based on the Kullback–Leibler divergence method, which defines the goodness of fit measure for regression models and is the coefficient of multiple determination for multiple regression [41]. R^2 defines the performance of the model. For example, R^2 equal to 100% confirms that all the predicted results are around their means. Table 2 provides the performance of the model for different impact categories. The table also provides the error calculated between the actual and predicted values.

Table 2. Mean squared error (MSE), R-squared analysis (R^2), and the number of data points of the predicted impact categories.

Impact Category	Number of Data Points	Mean Squared Error	R-Squared Results
Photochemical Ozone Creation Potential	196	0.07	70%
Abiotic depletion potential for fossil resources	196	0.01	77%
Global warming potential	196	0.28	81%
Acidification potential	196	1.12	68%

A statistical method called mean squared error (MSE) measures the average squared difference between the predicted and actual values. As the unit of MSE is higher than the actual error value, typically, the root of MSE, also known as root mean, squared error (RMSE), is used to evaluate the model. A smaller value of MSE indicates a better model, and it is sensitive to outliers, while R^2, on the other hand, is not so sensitive to outliers; it is based on the correlation between the actual value and predicted value.

Regression models are sensitive to outliers. As seen in Figure 3, outliers are below each impact category, the points away from the cluster cloud of points. Not all outliers are errors, and few of them contain meaningful information. However, their existence affects the entire regression model. Our case study identified data points as outliers from the lower number of EPDs of specific categories. For instance, 53 EPDs were categorized as 'metals' as the first classification category. While 33 EPDs were categorized under 'steel and iron', there are fewer EPDs for 'aluminum' (3), 'lead' (1), and others. Insufficient data to learn limits the ability of the model to predict data accurately.

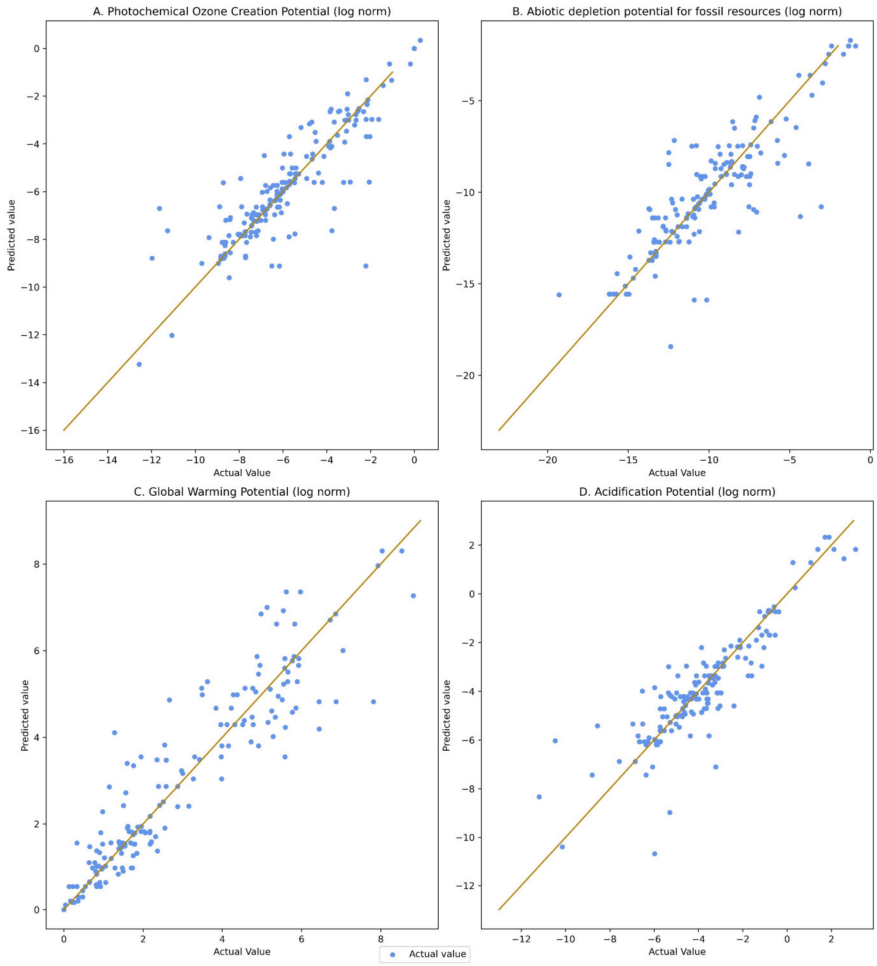

Figure 3. Visualization of prediction plot for the entire dataset for all the impact categories: (**A**) Photochemical Ozone Creation Potential, (**B**) Abiotic Depletion for Fossil Resources, (**C**) Global Warming Potential, (**D**) Acidification Potential.

The R^2 result of the Global Warming Potential (GWP), Photochemical Ozone Creation Potential (POCP), Acidification Potential (AP), and Abiotic Depletion Potential for Fossil Resources (ADPF) indicate the model's performance above 60%. The results of GWP, ADPF, and POCP are considered reasonable compared to the training data size used.

The visualization of regression results in Figure 3 shows that the actual values are close to the prediction line, except for a few outliers. Several studies show that a good

model can have a low R^2 value, and a biased model could have an excellent R^2 value [41]. Assessing the residual plot is one way to cross-verify R squared analysis limitations.

An application of the model can be demonstrated by predicting the values of a product. An EPD is selected amongst the testing dataset to demonstrate the application. From the randomly split testing dataset (20% of the complete database), an EPD of "reinforcement steel wire" with a reference flow of 1 kg is selected. This EPD was not used in the training dataset, and the model does not learn the values and is unknown. The is EPD is classified under the hierarchy of "metal" to "steel and iron" to "steel reinforcement mesh" and represents inventory for cradle to grave. This preliminary information is provided to the model, which combines it with the entire database for characterizing and encoding and is again extracted for prediction. The result of the prediction is tabulated in Table 3.

Table 3. Predicted results and actual values of the seven environmental indicators from the EPD "reinforcement steel wire".

Environmental Impact Indicators	Original Values	Units	Predicted Values
Photochemical Ozone Creation Potential (POCP)	0.000266	kg Ethene eq.	0.00019152
Abiotic depletion potential for fossil resources (ADPF)	7.627	MJ	6.102
Global warming potential (GWP)	0.6834	kg CO2 eq.	0.564
Acidification potential (AP)	0.001282	kg SO2 eq.	0.00071792

The predicted values are close to a few indicators' actual values as analyzed from the results of the testing dataset in Table 2. However, to use this model by a practitioner requires modifications and improvement of accuracy. One observation from the results is that the prediction accuracy of specific indicators is much higher than the others due to the data quality. Moreover, the number of features used is much higher than the size of the database due to the bagging and encoding of the descriptive data. In such cases, a more extensive database increases the accuracy of the model. Overall, our method's results demonstrated the ability to use regression analysis using qualitative information of the product implemented for the first time.

4. Conclusions

Increasing demand to know the environmental impacts of products and services prompts an AI-based model to predict them with minimal time, data, and modeling requirements. However, an AI-based model has extensive data requirements to predict a product's environmental impacts accurately. This article presents a working AI-based prediction model using an existing database of EPDs, which is a form of publishing results of LCA in a harmonized format. At the current stage of development and given the limited number of EPDs available, our method is intended as a check to predict the environmental impacts of a product or service quickly and is not a replacement for a detailed LCA study.

Construction products are used as a case study due to the availability of an adequate EPD database. Although existing studies pointed us to use ANN as the ML method, our choice to use the RF algorithm stems from the fact that ANN performs best with a huge database, and RF is an ensemble tree-based algorithm that performs better with more features. Using LCA studies results published in an EPD as our data source, a large amount of descriptive data must be processed and characterized using NLP. The characterized data is then fit into the RF regressor model. The trained model will predict the results of the environmental impact categories by providing information about the product as input to the model.

We have shown that the model developed can predict four impact categories with more than 65% R^2 value for our case of construction products. These results demonstrate the ability of the model to use regression analysis to predict environmental impact categories using qualitative product information. ÖKOBAUDAT is a major EPD publisher by IBU, Germany, while INIES host one of the largest database of EPD with over 3000 EPDs. In

addition to RF regression, we could examine the possibility of using ANN to apply to our problem with the INIES database. We also intend to benchmark our model with different ML methods on their performance. Our future work will also focus on implementing this method to other product groups and non-aggregated datasets of LCA results. The more EPDs available, the more accurate the results are. Therefore, combining multiple EPD databases for construction products of several countries, first at the European and then at the international level, as an enlarged data source might be an exciting way forward.

Author Contributions: Conceptualization, A.K., N.A. and G.S.; methodology, A.K. and N.A.; formal analysis, A.K., N.A. and G.S.; resources, N.A.; data curation, A.K.; writing—original draft preparation, A.K.; writing—review and editing, N.A., C.S., S.A.A. and G.S. All authors have read and agreed to the published version of the manuscript.

Funding: The Ph.D. thesis of the first author at the University of Bordeaux in the context of the TripleLink project funded by EIT Raw Materials was the opportunity to finalize this publication. The Ph.D. thesis is supported by EIT, a body of the European Union.

Acknowledgments: WeLOOP initiates the idea for this project. This publication was developed during an internship of the first author at WeLOOP in the north of France.

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

Glossary

LCA	Life Cycle Assessment
ISO	International Standard Organization
EPD	Environmental Product Development
EU	European Union
PCR	Product Category Rules
PEF	Product Environmental Footprint
NGO	Non-Government Organizations
AI	Artificial Intelligence
GT	Game Theory
ABM	Agent-Based Model
ANN	Artificial Neural Network
ML	Machine Learning
SVM	Support Vector Machine
BN	Bayesian Network
GA	Genetic Algorithm
KNN	K-Nearest Neighbor
Adaboost	Adaptive Boosting
GBM	Gradient Boosting Machine
NN	Neural Network
XML	Extensible Markup Language
HTML	Hypertext Markup Language
ILCD	International Reference Life Cycle Data System
NLP	Natural Language Processing
DT	Decision Tree
RF	Random Forest
IBU	The Institut Bauen und Umwelt e.V.
GWP	Global Warming Potential
POCP	Photochemical Ozone Creation Potential
AP	Acidification Potential
ADPF	Abiotic Depletion Potential for Fossil Resources
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
R^2	R Squared

References

1. Clift, R.; Doig, A.; Finnveden, G. The application of Life Cycle Assessment to Integrated Solid Waste Management. Part 1—Methodology. *Process Saf. Environ. Prot.* **2000**, *78*, 279–287. [CrossRef]
2. *ISO 14025*; Environmental Labels and Declarations—Type III Environmental Declarations—Principles and Procedures. International Organization for Standardization: Geneva, Switzerland, 2010. Available online: http://www.cscses.com/uploads/2016328/20160328110527052705.pdf (accessed on 7 March 2022).
3. Jose, F.; Benjamin, E.; Shelie, A. Developing LCA Techniques for Emerging Systems: Game Theory, Agent Modeling as Prediction Tools. In Proceedings of the 2010 IEEE International Symposium on Sustainable Systems and Technology, Arlington, VA, USA, 17–19 May 2010; pp. 1–6. [CrossRef]
4. Halog, A.; Manik, Y. Advancing integrated systems modelling framework for life cycle sustainability assessment. *Sustainability* **2011**, *3*, 469–499. [CrossRef]
5. Bonabeau, E. Agent-based modeling: Methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci. USA* **2002**, *99* (Suppl. S3), 7280–7287. [CrossRef]
6. Miller, S.A.; Moysey, S.; Sharp, B.; Alfaro, J. A Stochastic Approach to Model Dynamic Systems in Life Cycle Assessment. *J. Ind. Ecol.* **2012**, *17*, 352–362. [CrossRef]
7. Micolier, A.; Loubet, P.; Taillandier, F.; Sonnemann, G. To what extent can agent-based modelling enhance a life cycle assessment? Answers based on a literature review. *J. Clean. Prod.* **2019**, *239*, 118123. [CrossRef]
8. Das, K.; Behera, R.N. A Survey on Machine Learning: Concept, Algorithms and Applications. *Int. J. Innov. Res. Comput. Commun. Eng.* **2017**, *5*, 1301–1309.
9. Barros, N.N.; Ruschel, R.C. Machine Learning for Whole-Building Life Cycle Assessment: A Systematic Literature Review. In Proceedings of the 18th International Conference on Computing in Civil and Building Engineering, São Paulo, Brazil, 18–20 August 2020; Springer: Cham, Switzerland, 2021; pp. 109–122.
10. Nabavi-Pelesaraei, A.; Rafiee, S.; Mohtasebi, S.S.; Hosseinzadeh-Bandbafha, H.; Chau, K.w. Integration of artificial intelligence methods and life cycle assessment to predict energy output and environmental impacts of paddy production. *Sci. Total Environ.* **2018**, *631–632*, 1279–1294. [CrossRef]
11. 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*, 1005–1019. [CrossRef]
12. Ahmad, M.W.; Mourshed, M.; Rezgui, Y. Trees vs. Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy Build.* **2017**, *147*, 77–89. [CrossRef]
13. Stangierski, J.; Weiss, D.; Kaczmarek, A. Multiple regression models and Artificial Neural Network (ANN) as prediction tools of changes in overall quality during the storage of spreadable processed Gouda cheese. *Eur. Food Res. Technol.* **2019**, *245*, 2539–2547. [CrossRef]
14. Hou, P.; Jolliet, O.; Zhu, J.; Xu, M. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. *Environ. Int.* **2020**, *135*, 105393. [CrossRef]
15. European Commission. Environmental Footprint. European Platform on Life Cycle Assessment. Available online: https://eplca.jrc.ec.europa.eu/EnvironmentalFootprint.html (accessed on 7 March 2022).
16. Passer, A.; Lasvaux, S.; Allacker, K.; De Lathauwer, D.; Spirinckx, C.; Wittstock, B.; Kellenberger, D.; Gschösser, F.; Wall, J.; Wallbaum, H. Environmental product declarations entering the building sector: Critical reflections based on 5 to 10 years experience in different European countries. *Int. J. Life Cycle Assess.* **2015**, *20*, 1199–1212. [CrossRef]
17. Allander, A. Successful certification of an environmental product declaration for an ABB product. *Corp. Environ. Strateg.* **2001**, *8*, 133–141. [CrossRef]
18. Bovea, M.D.; Ibáñez-Forés, V.; Agustí-Juan, I. Environmental product declaration (EPD) labelling of construction and building materials. In *Eco-Efficient Construction and Building Materials. Life Cycle Assessment (LCA), Eco-Labelling and Case Studies*; Woodhead Publishing: Cambridge, UK, 2013; pp. 125–150. [CrossRef]
19. Del Borghi, A. LCA and communication: Environmental Product Declaration. *Int. J. Life Cycle Assess.* **2013**, *18*, 293–295. [CrossRef]
20. Minkov, N.; Schneider, L.; Lehmann, A.; Finkbeiner, M. Type III Environmental Declaration Programmes and harmonization of product category rules: Status quo and practical challenges. *J. Clean. Prod.* **2015**, *94*, 235–246. [CrossRef]
21. Durão, V.; Silvestre, J.D.; Mateus, R.; de Brito, J. Assessment and communication of the environmental performance of construction products in Europe: Comparison between PEF and EN 15804 compliant EPD schemes. *Resour. Conserv. Recycl.* **2020**, *156*, 104703. [CrossRef]
22. Adibi, N.; Mousavi, M.; Escobar, M.M.; Glachant, R.; Adibi, A. Mainstream Use of EPDs in Buildings: Lessons Learned from Europe. In *ISBS 2019*; IntechOpen: London, UK, 2019; p. 890.
23. Institut Bauen und Umwelt e.V. Available online: https://ibu-epd.com/ibu/ (accessed on 7 March 2022).
24. Lasvaux, S.; Habert, G.; Peuportier, B.; Chevalier, J. Comparison of generic and product-specific Life Cycle Assessment databases: Application to construction materials used in building LCA studies. *Int. J. Life Cycle Assess.* **2015**, *20*, 1473–1490. [CrossRef]
25. Pardede, E.; Rahayu, J.W.; Taniar, D. XML data update management in XML-enabled database. *J. Comput. Syst. Sci.* **2008**, *74*, 170–195. [CrossRef]

26. Recchioni, M.; Mathieux, F.; Goralczyk, M.; Schau, E.M. *ILCD Data Network and ELCD Database—Current Use and Further Needs for Supporting Environmental Footprint and Life Cycle Indicator Projects*; Report EUR 25744 EN; European Commission: Luxembourg, 2013.
27. Baitz, M.; Makishi Colodel, C.; Kupfer, K.; Florin, J.; Schuller, O.; Kokborg, M.; Köhler, A.; Thylmann, D.; Stoffregen, A.; Schöll, S.; et al. *GaBi Database & Modelling Principles 2014*; PE International AG: Stuttgart, Germany, 2014.
28. Oekobaudat. Available online: https://www.oekobaudat.de/datenbank/browser-oekobaudat.html (accessed on 7 March 2022).
29. Selenium Selenium Client Driver Documentation. Available online: https://www.selenium.dev/selenium/docs/api/py/ (accessed on 7 March 2022).
30. Sqlite3 sqlite3—DB-API 2.0 interface for SQLite databases. Available online: https://docs.python.org/3/library/sqlite3.html (accessed on 7 March 2022).
31. Probst, P.; Wright, M.N.; Boulesteix, A.L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
32. Zeroual, I.; Lakhouaja, A. Data science in light of natural language processing: An overview. *Procedia Comput. Sci.* **2018**, *127*, 82–91. [CrossRef]
33. Joseph, S.R.; Hloman, H.; Letsholo, K.; Sedimo, K. Natural Language Processing: A Review. *Int. J. Res. Eng. Appl. Sci.* **2016**, *6*, 1–8.
34. Vallbé, J.-J.; Martí, M.A.; Blaž, F.; Jakulin, A.; Mladenič, D.; Casanovas, P. Stemming and lemmatisation: Improving knowledge management through language processing techniques. In *Trends in Legal Knowledge: The Semantic Web and the Regulation of Electronic Social Systems*; European Press Academic Publishing: Florence, Italy, 2007; pp. 1–16. Available online: https://www.researchgate.net/publication/228704765 (accessed on 7 March 2022).
35. Kotsiantis, S.B. Decision trees: A recent overview. *Artif. Intell. Rev.* **2013**, *39*, 261–283. [CrossRef]
36. SoleimanianGharehchopogh, F.; Mohammadi, P.; Hakimi, P. Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study. *Int. J. Comput. Appl.* **2012**, *52*, 21–26. [CrossRef]
37. Fawagreh, K.; Gaber, M.M.; Elyan, E. Random forests: From early developments to recent advancements. *Syst. Sci. Control Eng.* **2014**, *2*, 602–609. [CrossRef]
38. Ho, T.K. Random decision forests. In Proceedings of the3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 278–282. [CrossRef]
39. Umarani, V.; Rathika, C. Predicting Safety Information of Drugs Using Data Mining Technique. *Int. J. Comput. Eng. Technol.* **2019**, *10*, 83–90. [CrossRef]
40. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
41. Cameron, A.C.; Windmeijer, F.A.G. An R-squared measure of goodness of fit for some common nonlinear regression models. *J. Econom.* **1997**, *77*, 329–342. [CrossRef]