Detecting the Factors Affecting Carbon Emissions in Food Recipes using Regression Models and Explainable Artificial Intelligence

Eman Ahmed[1,5,*,†], Mamdouh Gomaa[2,5,†], Ashraf Darwish[3,5,†] and Aboul Ella Hassanien[1,5,†]

[1]Faculty of Computers and Artificial Intelligence, Cairo University, Egypt
[2]Computer Science Department, Faculty of Science, Minia University, Egypt
[3]Faculty of Science, Helwan University, Egypt
[5]Scientific Research School of Egypt (SRSEG), https://egyptscience-srge.com/

Abstract

In this paper, we would like to investigate what factors affect Green House Gas (GHG) emissions in different food recipes from various cuisines. Feature selection is performed using correlation analysis. After that six different regression models are implemented including linear models such as linear regression, ridge regression and lasso regression models, and non-linear regression models including decision trees, random forest and gradient boosting. Explainable Artificial Intelligence (XAI) is applied by using the SHAPely method to study the impact of each feature on carbon emissions. Results show that high priced categories have high GHG emissions and vice versa.

Keywords

Correlation, Explainable AI, Food, GHG emissions, regression, SHAP

1. Introduction

The food industry is a major contributor to global GHG emissions, with impacts at each stage of the food production, processing, and distribution process [1] [2]. During food processing, GHG emissions are a result of the energy required to transform raw ingredients into completed food items. This covers cooking, packaging, refrigeration, and other processes that frequently use fossil fuels [3]. GHG emissions increase during the transportation of food from fields to processing plants, retail locations, and ultimately consumers, particularly when long-distance shipping is involved. The mode of transportation affects emissions, with air freight typically having the highest carbon footprint [4].

Another source of GHG emissions is wasted food. When food decomposes in landfills, it emits methane. Moreover, producing unconsumed food wastes all resources including land, water, and energy.

A lot of food items include packaging, which takes resources and energy to make. Additional GHG emissions are caused by the manufacture and disposal of packaging, particularly plastics, particularly when these processes are not handled responsibly [5].

You can cook the same recipe in a variety of ways by mixing the ingredients. Furthermore, there are countless methods to prepare meals using those items because they can be prepared in different ways. Numerous recipes are accessible online, and they include a vast amount of material that enables both amateurs and experts to explore different components in various cuisines. The researchers are able to identify both the similarities and variances among the various cuisines [6].

In this paper, we investigate the main factors that result in high GHG emissions to be able to provide recommendations on how to minimize the emissions. Different cuisines with various ingredients are

The 2024 Sixth Doctoral Symposium on Intelligence Enabled Research (DoSIER 2024), November 28–29, 2024, Jalpaiguri, India
*Corresponding author.
†These authors contributed equally.
e.ahmed@fci-cu.edu.eg (E. Ahmed); mamdouh.gomaa@mu.edu.eg (M. Gomaa); ashraf.darwish.eg@ieee.org (A. Darwish); aboitcairo@cu.edu.eg (A. E. Hassanien)
0000-0003-3122-0164 (E. Ahmed); 0009-0000-4426-5965 (M. Gomaa); 0000-0002-4604-1436 (A. Darwish); 0000-0002-9989-6681 (A. E. Hassanien)

tested. Regression models are employed to predict the amount of emissions given a set of chosen features. The set of the selected features are obtained using correlation analysis.

This paper is structured as follows: section 2 will explore the related work food industry and GHG emissions. In section 3, the basics of the used regression models are explained. Section 4 will introduce the methodology and experimental results details are discussed in Section 5. Finally, the conclusions and future work are presented in section 6.

2. Related Work

This section covers the literature review on using machine learning and deep learning in the food industry with a focus on topics related to carbon emissions.

The food business can benefit from a number of opportunities presented by AI integration along the supply chain [7]. This can entail outlining waste reduction tactics for the retail industry. Reducing food loss and waste, which account for 8%–10% of anthropogenic GHG emissions [8] [9] and one-third of the food produced for human consumption, is in line with waste prevention of the EU Waste Framework Directive's waste hierarchy priority [10] and UN Sustainable Development Goal (12.3) [11].

This study [12] develops a tracking system for carbon footprint utilizing image recognition using convolutional neural network specifically, using Inception-V3 model to investigate and enhance the benefits of environmentally sustainable eating practices. The suggested model's accuracy of 94.79%, according to the results, shows that it can successfully identify different kinds of food. The tracking device was tested for two weeks, during which time the study participants measured overall carbon footprint decreased by 22.25%.

A comprehensive life cycle analysis of "Foodforecast" is presented in [13]. It proposes a machine learning (ML) cloud service that optimizes sales forecasting to minimize food waste in bakeries. It addresses the effect of four factors including global warming, cumulative energy demand, abiotic resource depletion, and freshwater eutrophication. Using real-world case study data, the evaluation covers the indirect advantages of avoiding bakery returns in comparison to conventional ordering techniques, as well as the direct environmental effects of the used ML model and the hardware that underlies the system. According to sales estimates, it reduced bakery returns, mostly of bread and rolls, by an average of 30% in 2022. Across impact categories and return utilization scenarios, the associated environmental benefits greatly exceeded the direct consequences of the system by an order of magnitude.

3. Dataset

The used dataset can be found in [14]. There are 47 fields in the dataset, they can be divided into four categories: Environmental Impact Analysis, Detailed Ingredient and Nutritional Information, Nutrition and Keywords, and General Information. It consists of 388 recipes from 5 different recipe cuisines. Each recipe has a set of features including cooking time (min), the number of servings, nutritional composition of each dish, the weights of each ingredient in the dish, price needed to buy the ingredients utilized in the dish, calories (Kcal). This is on top of nutrients like energy in kilocalories (kcal), fat in gram (g), protein in (g), and carbohydrates in (g). Vitamins are then assessed, particularly vitamin A in micro-gram (μg), vitamin C in milli-gram (mg), and vitamin E in (mg). The amounts of Folic Acid in (μg), Calcium in (mg), Dietary Fiber in (mg), Iron in (mg), Zinc in (mg), Magnesium in (mg), Potassium in (mg), Saturated Fats in (g), Salt Equivalent in (g) and Cholesterol in (mg) were evaluated in the mineral category.

Dishes are classified into 11 main categories based on the ingredients of each dish, the categories are beef, pork, chicken, minced meat, fish, grain, processed meat, bean, mushroom, vegetables, and egg. Food loss calculated from (leftovers, direct waste, excessive removal) is included. The carbon footprints of the following processes: production, cooking, sales, and disposal are all included in the total quantity of greenhouse gas emissions.

4. Materials and Methods

In this paper, we use correlation analysis for feature selection then regression models are trained and tested for predicting the amount of carbon emissions for a given food recipe. After that, the performance of the models is compared. SHAP method will be applied to the model that results in minimum squared error to further interpret the impacts of different features on the prediction of the amount of emissions.

Figure 1 demonstrates the proposed model to identify the features that affect GHG emissions. Initially, data preprocessing is applied then features are selected based on correlation analysis with GHG emissions. Next, the training samples with the selected features are used to train different regression models to predict the amount of GHG emissions. The hyper-parameters of each regression model are chosen using grid search with 5 fold cross validation. After that, the trained models are tested using a test set. Shapely (SHAP) method is then used to detect the most important features that affected the decisions of each of the regression models. This allows us to identify the key features that have impact on the GHG emissions in different food recipes and various food cuisines. Each of the model steps are presented in detail below.

4.1. Data Preprocessing

4.1.1. 1- Log transformation to reduce skewness

Reduces skewness in data, particularly for features with high skew (i.e., features with a skewness greater than 0.5). This transformation is effective for variables with long-tailed distributions, making them closer to a normal distribution and potentially improving model performance [15].

4.1.2. 2- Outlier Removal Using Z-Score

Removes data points considered outliers to prevent them from skewing the model's training. Outliers are identified as values with Z-scores greater than 3 or less than -3. The Z-score standardizes data points by calculating how standard deviations are away from the mean. Data points with absolute Z-scores above 3 are typically considered outliers and are removed from the dataset. The Z-score is calculated by the following equation [16] [17].

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

where Z is the Z-score, X is the value being calculated, μ is the mean, and σ is the standard deviation.

4.2. Feature Selection

After the preprocessing phase, feature selection was performed using correlation analysis [18] which effectively reduced the final selection to some key features. These selected features have correlation less than -0.3 or greater than 0.3, they included 'price', 'zinc in (mg)', 'Protein in (g)', 'Energy (g)', 'calories in (Kcal)', 'Magnesium in (mg)', 'Potassium in (mg)', 'Saturated fat in (g)', 'iron in (mg)', 'Fat in (g)', 'cholesterol in (g)', 'cholesterolContent in (mg)', 'leftover', 'Pork', 'carbohydrates in (g)', 'carbohydrateContent in (g)', 'Salt equivalent in (g)', 'cooking_time', 'Vitamin E in (mg)', 'Seasonings', 'direct_disposal', 'Beverage', 'Grain', 'Chicken', 'dish', 'category'.

4.3. Regression Models

The dataset is splitted into training and test sets. The training set with the selected features are used to train the regression models. After that, the test set is used to test the models. Six regression models including Linear models including linear regression, Ridge Regression [19] and Lasso Regression [20], and non-linear models including Decision Tree [21], Random Forest [22], and Gradient Boosting. Grid search with 5-fold cross validation is used to select the hyper-parameters of the models.

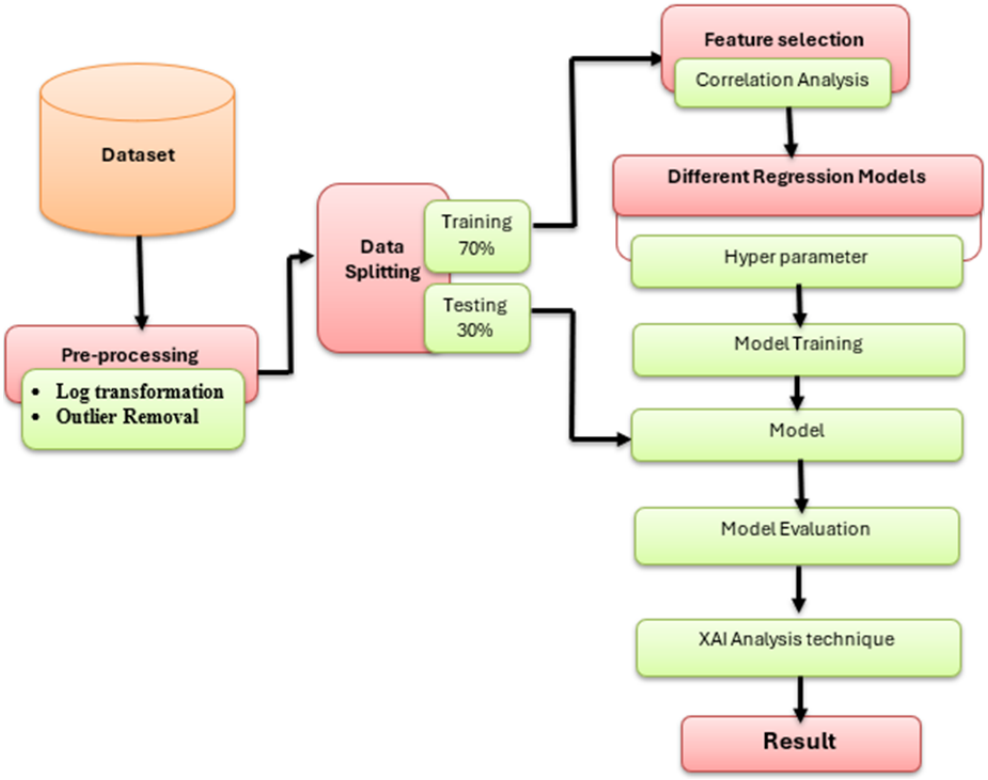

Figure 1: The proposed model.

5. Experimental Results

We have used 6 regression models including Linear, Ridge Regression, Lasso Regression, Decision Tree, Random Forest, and Gradient Boosting. Grid search with 5-fold cross validation is used to select the hyper-parameters of the models. For Ridge regression, L1 norm coefficient is chosen from 0.1 to 10 with step 1. For Lasso regression, L2 norm coefficient is selected from 0.01 to 1 with step 0.1. In case of decision trees, the maximum depth was chosen from 10 or 20 or 30. For random forest, the number of trees had a lower bound of 100 and upper bound of 200 and the maximum depth was either 10 or 20. For gradient boosting, the number of estimators is chosen from 100 to 200 and the learning rate has a lower value of 0.01 and upper value of 1 searched with step 0.1. Mean Squared Errors have been calculated on the test set and are shown in Figure 2.

It can be seen that linear regression models perform better than non-linear regression models. Lasso regression has the minimum mean squared. Accordingly, we will get the SHAP values for Lasso regression to get more interpretation on which features impacted its decisions. Figure 3 shows the SHAP summary plot. It is noticed that the price is the most feature having impact on the amount of GHG emissions. When price has high values (red), it is associated with positive SHAP values, which means that it results in increasing the amount of GHG emissions.

On the other hand, the low values of price (blue) have low SHAP values, which indicate that it results in low GHG emissions. With more inspection of the summary plot, it is noticed that categories that have high price affect the prediction of the amount of GHG emissions.

Figure 4 shows the average GHG emissions by cuisine category and identifies the cuisines with the highest and lowest average GHG emissions, from this figure we can observe that the beef category has high emissions and egg category has low emissions. Also, Figure 5 depicts the average prices of each category showing that Beef is the most expensive while egg is the least. Accordingly, these analyses support the results from SHAP summary plot of the Lasso regression model.

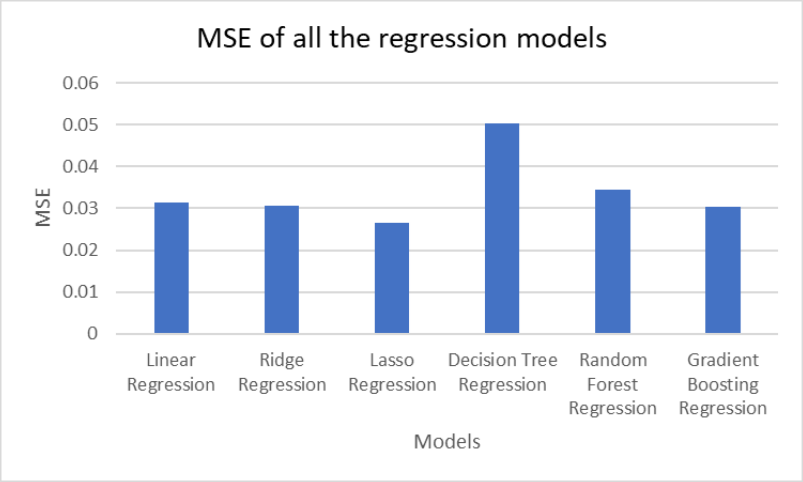

Figure 2: MSE of all the regression models.

6. Conclusion and Future Work

In this paper, different regression models are applied to predict the amount of GHG emissions using features selected using correlation analysis. Lasso regression model obtained the minimum mean squared error; hence, XAI technique (SHAP) is used to interpret the most features impacted the prediction. Price was the feature that had most impact on the prediction of the amount of GHG emissions. Hence, we conclude that food recipes that have ingredients belonging to expensive categories result in higher GHG emissions. In the future, we would experiment with larger datasets to derive more relations between the ingredients and GHG emissions.

Declaration on Generative AI

During the preparation of this work, the author(s) used QuillBot in order to paraphrase. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] I. Shabir, K. K. Dash, A. H. Dar, V. K. Pandey, U. Fayaz, S. Srivastava, N. R., Carbon footprints evaluation for sustainable food processing system development: A comprehensive review, Future Foods 7 (2023) 2666–8335. doi:10.1016/j.fufo.2023.100215.

[2] H. N. Afrouzi, J. Ahmed, B. M. Siddique, N. Khairuddin, A. Hassan, A comprehensive review on carbon footprint of regular diet and ways to improving lowered emissions, Results in Engineering 18 (2023). doi:10.1016/j.rineng.2023.101054.

[3] A. Ladha-Sabur, S. Bakalis, P. J. Fryer, E. Lopez-Quiroga, Mapping energy consumption in food manufacturing, Trends in Food Science & Technology 86 (2019) 270–280. doi:10.1016/j.tifs.2019.02.034.

[4] W. Wakeland, S. Cholette, K. Venkat, Food transportation issues and reducing carbon footprint, in: J. Boye, Y. Arcand (Eds.), Green Technologies in Food Production and Processing, Springer, Boston, MA., 2012. doi:10.1007/978-1-4614-1587-9_9.

[5] N. L. K., U. A. U., O. E. N., Z. R., B. I. N., Environmental impact of food packaging materials: A review of contemporary development from conventional plastics to polylactic acid based materials, Materials (Basel) 13 (2020) 4994. doi:10.3390/ma13214994.

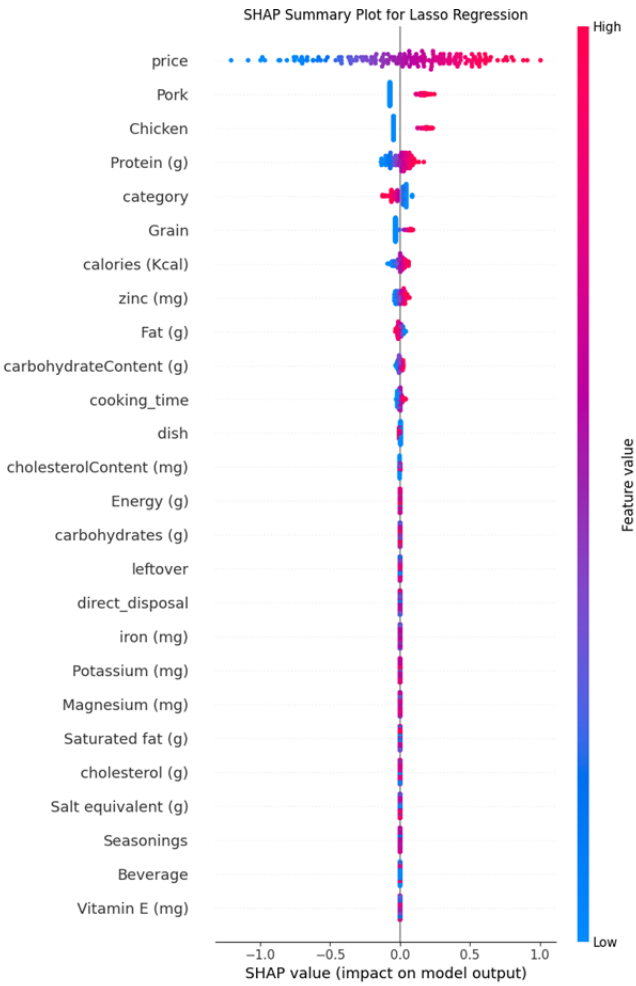

Figure 3: The SHAP summary plot for Lasso regression.

[6] A. I. I., L. E. G., M. I. Y., S. V. K., Results designing and analysis when introducing new beverage identification criteria, Food Systems 3 (2020) 4–7. doi:10.21323/2618-9771-2020-3-3-4-7.

[7] I. Kumar, J. Rawat, N. Mohd, S. Husain, Opportunities of artificial intelligence and machine learning in the food industry, Journal of Food Quality (2021) 1–10. doi:10.1155/2021/4535567.

[8] Food wastage footprint: Impact on natural resources, Summary report, Food and Agriculture Organization of the United Nations (FAO), 2013.

[9] UNEP Food Waste Index Report, Technical Report, UNEP, Nairobi, 2021.

[10] Directive 2008/98/EC of the European Parliament and of the Council of 19 November 2008 on waste and repealing certain directives text with EEA relevance, Technical Report, European Commission (EC), 2024.

[11] Transforming our world: The 2030 Agenda for Sustainable Development, Technical Report, United Nations (UN), 2015.

[12] M.-C. Chiu, Y.-L. Tu, M.-C. Kao, Applying deep learning image recognition technology to promote environmentally sustainable behavior, Sustainable Production and Consumption 31 (2022) 736–749. doi:10.1016/j.spc.2022.03.031.

[13] N. Hübner, J. Caspers, V. C. Coroamă, M. Finkbeiner, Machine-learning-based demand forecasting against food waste: Life cycle environmental impacts and benefits of a bakery case study, Journal of Industrial Ecology 28 (2024) 1117–1131. doi:10.1111/jiec.13528.

[14] Y. Long, L. Huang, R. e. a. Fujie, Carbon footprint and embodied nutrition evaluation of 388 recipes, Sci Data 10 (2023) 794. doi:10.1038/s41597-023-02702-1.

[15] C. Feng, H. Wang, N. Lu, T. Chen, H. He, Y. Lu, X. M. Tu, Log-transformation and its implications

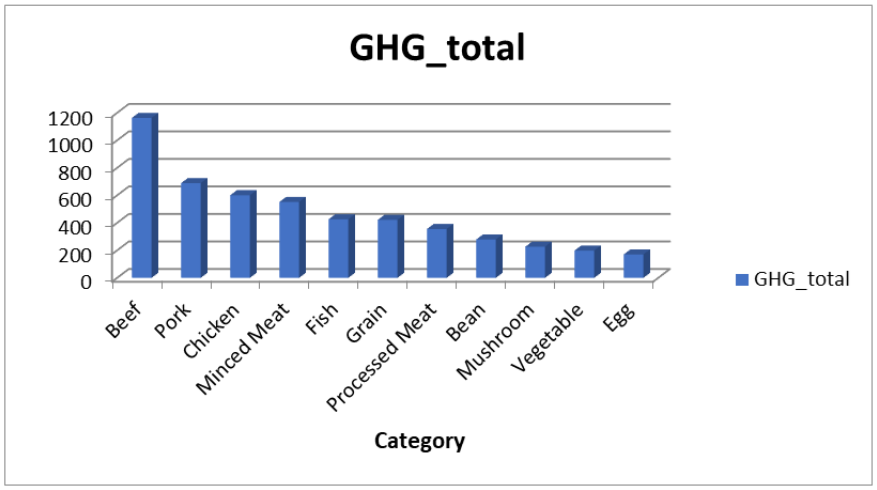

Figure 4: The amount of GHG emissions of different categories.

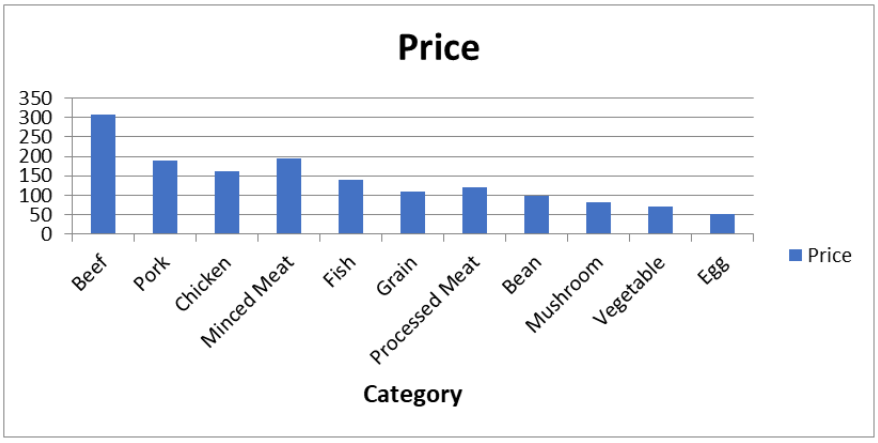

Figure 5: The prices of different categories.

for data analysis, Shanghai archives of psychiatry 26 (2024) 105–109.

[16] P. V. Anusha, C. Anuradha, P. C. Murty, C. S. Kiran, Detecting outliers in high dimensional data sets using z-score methodology, International Journal of Innovative Technology and Exploring Engineering 9 (2019) 48–53.

[17] V. Aggarwal, V. Gupta, P. Singh, K. Sharma, N. Sharma, Detection of spatial outlier by using improved z-score test, in: Proceedings of 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 788–790. doi:10.1109/ICOEI.2019.8862582.

[18] S. Ibrahim, S. Nazir, S. A. Velastin, Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis, Journal of Imaging 7 (2021) 225. doi:10.3390/jimaging7110225.

[19] G. C. McDonald, Ridge regression, WIREs Comp Stat 1 (2009) 93–100. doi:10.1002/wics.14.

[20] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Series B (Methodological 58 (1996) 267–288.

[21] L. Rokach, O. Maimon, Decision Trees, Data Mining and Knowledge Discovery Handbook, Springer, Boston, MA., 2005. doi:10.1007/0-387-25465-X_9.

[22] A. Cutler, D. R. Cutler, J. R. Stevens, Random forests, in: C. Zhang, Y. Ma (Eds.), Ensemble Machine Learning, Springer, New York, NY, 2012. doi:10.1007/978-1-4419-9326-7_5.