Machine learning algorithms for the prediction of land ecological footprint

1st Daniella Santos Munoz
Department of Computer Science
University of Ottawa
Ottawa, Canada

***Abstract*—Machine learning algorithms can be used to solve climate change problems. A contributor to the climate change problem is the ecological footprint a country may have, which is connected to how much impact a country has on the increase in carbon. Connecting land use with regions and GDP and population can give information on the carbon increase. Being able to predict the impact of land use on the environment can help countries to manage their land use in terms of curbing climate change. Machine learning algorithms based on tree, linear, distance, ensemble and rule models were used to predict the carbon offset based on several dataset features. Regression based tree and bagging ensemble models, along with multiple regression linear models, were found to be the best carbon predictors.**

I. Introduction

Climate change is a growing problem in the world. A contributor to this problem is the ecological footprint a country may have, which is connected to how much impact a country has on the increase in carbon. How land is used increases or decreases the carbon emissions. Being able to predict the impact of land use on the environment can help countries to manage their land use in terms of curbing climate change. Machine learning can be used to make these predictions so that countries can make pre-emptive decisions, helping the environment and saving money. Finding the best algorithm for the national ecological footprint dataset will lead to better predictions.

II. Case Study

The data used in this project was produced by the Global Footprint Network and is the 2018 National Footprint Account of 196 countries [1]. The ecological footprint (ecological services consumed by humans) and biocapacity (biologically productive area) of areas are calculated to assess the sustainability of the 196 countries. This area is recorded in global hectares (gha).

The data covers the year span of 1961-2014. The 196 countries are reported with their corresponding ISO alpha-3 code and UN region and subregion they pertain to. For each country, Percapita GDP (2010 USD) and population (FAO estimate) are reported. The type of land for each country and year were recorded (Table III) and their connection is denoted in (1). These landtypes are composed of crop land, grazing land, forest land, fishing ground and built-up land, and the total land area of these is included in the data. The carbon offset to the total land composition is the global hectares of forest needed to cut off carbon emissions. In this project we focus on carbon as the target.

$$EFCons = EFProd + EFImports - EFExports \quad (1)$$

III. Experimental Setup and Evaluation

A. Data preprocessing

1) Data transformation: Data transformation was carried out in Excel and RStudio [2]. The discretization of nominal features to discrete values was carried out in RStudio: ISO alpha-3 code, UN region (Table I), UN subregion (Table II), and record (Table III). The transformation was done with the as.numeric function applied (lapply function) to the dataframe.

The calibration and normalization of numerical values was done in Excel. The interval of the year feature was shifted from 1961-2014 to 0-53. Since the carbon and total features are related to one another and some areas can be much larger than others, carbon and total were min-max normalized to 0-21,000,000,000. Since the crop land, grazing land, forest land, fishing ground and built-up land features make up the total feature, the land features were calibrated to be a fraction of the total. The percapita GDP feature was min-max normalized to 0-115,000. The population feature was min-max normalized to 0-7,500,000,000.

TABLE I
UN regions of the dataset and their levels used for discretization.

Level	UN region
1	Africa
2	Asia
3	Europe
4	Latin America and the Caribbean
5	North America
6	Oceania
7	World

2) Data reduction: Data reduction was done in two ways: row reduction and column deletion. The data creators kept the number of missing values minimal, but they allowed for the landtypes along with carbon to be blank as long as the total was included. The rows were deleted that had missing values for all of crop land, grazing land, forest land, fishing

TABLE II
UN SUBREGIONS OF THE DATASET AND THEIR LEVELS USED FOR DISCRETIZATION.

Level	UN subregion
1	Australia and New Zealand
2	Caribbean
3	Central America
4	Central Asia
5	Eastern Africa
6	Eastern Asia
7	Eastern Europe
8	Melanesia
9	Micronesia
10	Middle Africa
11	North America
12	Northern Africa
13	Northern Europe
14	Polynesia
15	South America
16	South-Eastern Asia
17	Southern Africa
18	Southern Asia
19	Southern Europe
20	Western Africa
21	Western Asia
22	Western Europe
23	World

TABLE III
RECORD TYPES FOR LANDS RECORDED FROM EACH COUNTRY AND YEAR. BIOCAPACITY (BIOCAP) AND ECOLOGICAL FOOTPRINT (EF) WERE RECORDED. ECOLOGICAL FOOTPRINT OF CONSUMPTION (CONS), PRODUCTION (PROD) AND TRADE (IMPORTS AND EXPORTS) ARE NOTED. FOR EACH TYPE THERE ARE BOTH TOTAL GLOBAL HECTARES (TOTGHA) AND GLOBAL HECTARES PER CAPITA (PERCAP).

Level	Record Type
1	BiocapPerCap
2	BiocapTotGHA
3	EFConsPerCap
4	EFConsTotGHA
5	EFExportsPerCap
6	EFExportsTotGHA
7	EFImportsPerCap
8	EFImportsTotGHA
9	EFProdPerCap
10	EFProdTotGHA

ground, built-up land and carbon attributes. The row number was reduced by 27%, from 87022 rows to 63530. The country feature was removed from the data because it was redundant with the feature ISO alpha-3 code, reducing the features from 15 to 14. The list of features used in this project can be seen in Table IV.

The statistics of the continuous features are included in Table V. The transformations can be seen in the values of the table. Correlation of the transformed data is shown as a heatmap in Figure 1.

B. Model construction

1) Linear classifier: The linear method used was multiple linear regression. The LinearRegression function of the linear model from Scikit-learn [3] was used to model, fit and predict the data. Carbon was set as the target. The multiple linear regression method was chosen because the target is continuous, and the other features are a mix of continuous and discrete. The results are shown in Figure 2.

TABLE IV
FEATURES AND THEIR LEVELS FROM THE ORIGINAL DATA AND AFTER TRANSFORMING DATA.

Original level	Feature	Project level
1	country	
2	ISO alpha-3 code	0
3	UN region	1
4	UN subregion	2
5	year	3
6	record	4
7	crop land	5
8	grazing land	6
9	forest land	7
10	fishing ground	8
11	built-up land	9
12	carbon	10
13	total	11
14	percapita GDP	12
15	population	13

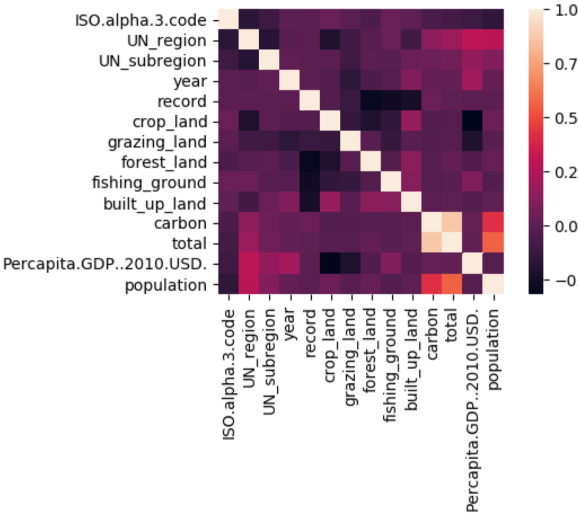

Fig. 1. Dataset feature correlation heatmap. Positive correlation is present between total and carbon, population with carbon and total, and UN region with percapita GDP and population. Negative correlation is present between UN region and crop land; record with forest land, fishing ground and built-up land; and percapita GDP with crop land and grazing land.

2) Tree-based method: The tree method used was regression. The DecisionTreeRegressor function of the tree model from Scikit-learn [3] was used to model, fit and predict the data. Carbon was set as the target. The decision tree regression method was chosen because the target is continuous, and the other features are a mix of continuous and discrete. The results are shown in Figure 3.

3) Distance-based method: The distance method used was feature agglomeration. The data was normalized first with the StandardScaler function because of the mix of discrete and continuous values. The FeatureAgglomeration function of the clustering model from Scikit-learn [3] was used to model,

TABLE V

STATISTICS OF CONTINUOUS FEATURES AFTER DATA TRANSFORMATION AND DATA REDUCTION.

Feature	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Crop land	0	0.1317	0.2282	0.268	0.3683	0.9919
Grazing land	0	0.01853	0.05791	0.12703	0.15884	0.98011
Forest land	0	0.03736	0.11043	0.18261	0.2661	0.98923
Fishing ground	0	0.008318	0.027809	0.073562	0.079261	0.994233
Built-up land	0	0	0.007863	0.018095	0.02644	0.720516
Carbon	0	0	0	0.001234	0.0000437	0.5961728
Total	0	0	0	0.0031948	0.0004989	0.9810435
Percapita GDP	0	0.004175	0.016938	0.082634	0.101593	0.988539
Population	0.0000121	0.0004629	0.0011799	0.01095	0.0033624	0.9687715

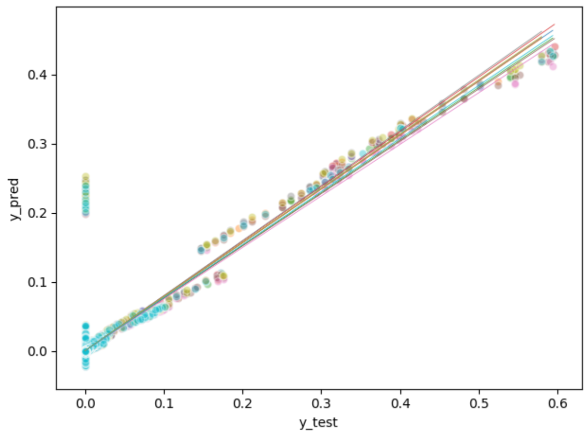

Fig. 2. Multiple linear regression algorithm results. The target test results are plotted against the predicted target results. The values of the 10-fold cross validation are included and the line of best fit of each fold is shown.

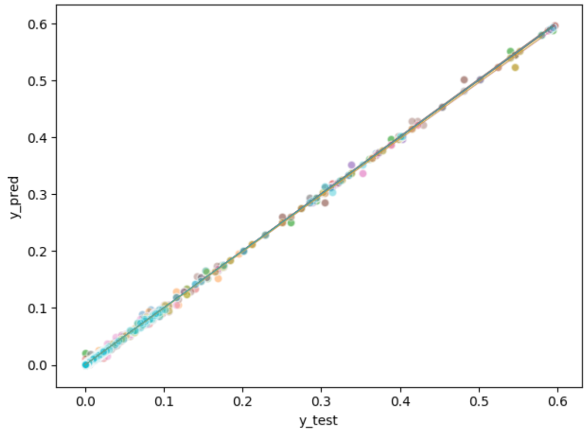

Fig. 3. Regression decision tree algorithm results. The target test results are plotted against the predicted target results. The values of the 10-fold cross validation are included and the line of best fit of each fold is shown.

fit and transform the data. The transformation of the model reduced the features, and then this reduction was inverse transformed to apply the clusters to the data. Feature agglomeration was used because of the large number of features and the sample size to decrease the number of features with clusters. The results of the 3-5 clusters are shown in Figure 4 as a heatmap.

4) Rule-based method: Problems were encountered with rule models because they did not fit the data well. Ordered and unordered rule methods could not be used because there cannot be any continuous data present. Association methods can be used with continuous features, but the target cannot be continuous. The target for our project data was continuous so the association method did not work well either.

The association method was tried by using the orange-contrib.associate.fpgrowth package [4] and Orange [5]. The data had information added to the header to be decoded by the Orange.data.Table package: D# denotes a discrete feature (continuous was default) and c# denotes the class variable (target). OneHot encoding was performed to transform the data to boolean, while decoding was performed to apply the data to the map. Frequent item sets could be produced but no results were produced with the association rules.

5) Ensemble: The ensemble method used was bagging. Since the regression tree algorithm had good results and the data is complex with multiple features, bagging was the best option. The BaggingRegressor function from the ensemble model from Scikit-learn [3] was used to model, fit and predict the data. The DecisionTreeRegressor function of the tree model was used as the estimator for the bagging model. Carbon was set as the target. The bagging regression method was chosen because the target is continuous, and the other features are a mix of continuous and discrete, and the tree used is also regression. The results are shown in Figure 5.

C. Model evaluation

1) Train-and-test paradigm: The regression trees were tested with train-test sizes of 80-20, 70-30 and 66-34. 80-20 was best as it had a lower error and higher R score. The dataset was split 10 times randomly to perform 10-fold cross-validation on the regression linear, tree and ensemble methods. Since the clustering method was unsupervised the data was not split.

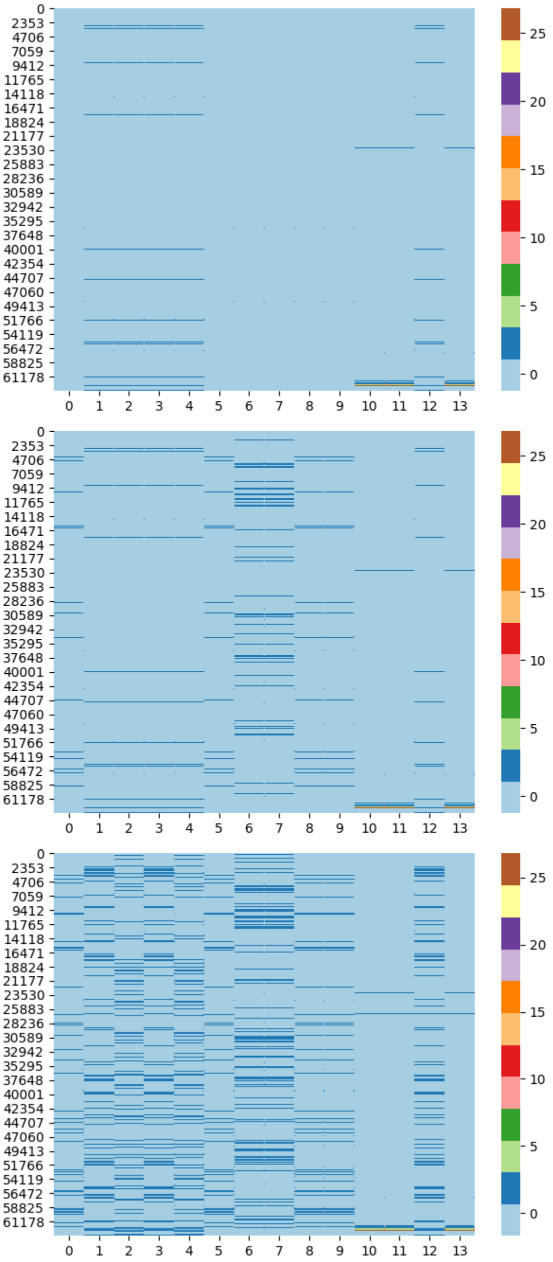

Fig. 4. Cluster heatmap of the feature agglomeration distance algorithm on the dataset features. The0 number of clusters used are three (top), four (middle) and five (bottom). For three clusters, the clusters are {0, 5, 6, 7, 8, 9}, {1, 2, 3, 4, 12} and {10, 11, 13}. For four clusters, the clusters are {0, 5, 8, 9}, {1, 2, 3, 4, 12}, {6, 7} and {10, 11, 13}. For five clusters, the clusters are {0, 5, 8, 9}, {1, 3, 12}, {2, 4}, {6, 7} and {10, 11, 13}. Dataset rows are on the left and features are on the bottom. The features are: ISO alpha-3 code (0), UN region (1), UN subregion (2), year (3), record (4), crop land (5), grazing land (6), forest land (7), fishing ground (8), built-up land (9), carbon (10), total (11), percapita GDP (12), and population (13).

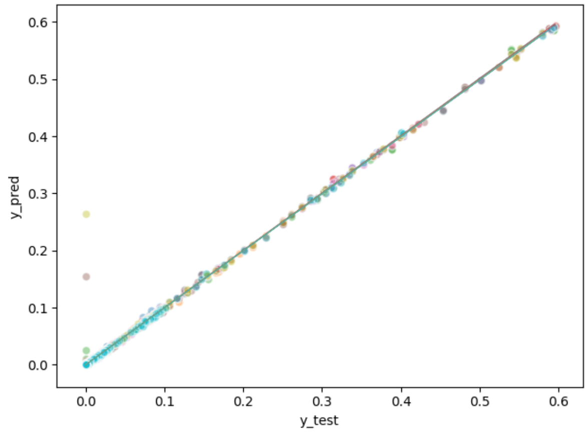

Fig. 5. Regression decision tree bagging ensemble algorithm results. The target test results are plotted against the predicted target results. The values of the 10-fold cross validation are included and the line of best fit of each fold is shown.

2) Evaluation criteria: To evaluate the regression linear, tree and ensemble methods, the tested target values were compared to the predicted target values. The metrics used to compare these methods were mean absolute error, median absolute error and R^2 regression score. Both mean absolute error and median absolute error are values to be minimized, where the lower the error and the closer it is to zero, the better the model fits. Regression is the opposite. The R^2 regression score is value to be maximized, where the higher the value and the closer it is to one, the better the model fits. The metrics of the regression tree, linear, and ensemble methods are reported in Table VI, VII and VIII.

TABLE VI

REGRESSION DECISION TREE ALGORITHM STATISTICS. THE DATA WAS RANDOMLY SPLIT 10 TIMES. THE MEAN ABSOLUTE ERROR AND MEDIAN ABSOLUTE ERROR ARE TO BE MINIMIZED. THE R^2 SCORE IS TO BE MAXIMIZED, WITH 1 BEING CORRECTLY FITTED. THE MEAN AND STANDARD DEVIATION (SD) OF EACH TYPE OF VALUE IS REPORTED.

Fold	**Mean Abs Error**	**Median Abs Error**	**R^2 Score**
1	9.4823E-05	3.9387E-05	9.9946E-01
2	9.4351E-05	3.9230E-05	9.9949E-01
3	1.0030E-04	3.9600E-05	9.9940E-01
4	9.5534E-05	3.9570E-05	9.9959E-01
5	1.5450E-04	8.2310E-05	9.9915E-01
6	1.0362E-04	3.9409E-05	9.9938E-01
7	1.0574E-04	3.8955E-05	9.9931E-01
8	9.5795E-05	3.6981E-05	9.9950E-01
9	9.9877E-05	3.9412E-05	9.9936E-01
10	9.2966E-05	3.9665E-05	9.9961E-01
mean	1.0375E-04	4.3452E-05	9.9942E-01
sd	1.8324E-05	1.3676E-05	1.3819E-04

3) Statistical testing: The metrics of Tables IX, X and XI were compared to one another for each of the regression tree, linear, and ensemble algorithms. Paired t-test was used because different algorithms are compared to each other over the 10-fold cross validation of a dataset. An α of 0.05 was used along with 9 degrees of freedom because cross validation was

TABLE VII

MULTIPLE LINEAR REGRESSION ALGORITHM STATISTICS. THE DATA WAS RANDOMLY SPLIT 10 TIMES. THE MEAN ABSOLUTE ERROR AND MEDIAN ABSOLUTE ERROR ARE TO BE MINIMIZED. THE R^2 SCORE IS TO BE MAXIMIZED, WITH 1 BEING CORRECTLY FITTED. THE MEAN AND STANDARD DEVIATION (SD) OF EACH TYPE OF VALUE IS REPORTED.

Fold	Mean Abs Error	Median Abs Error	$\mathbf{R^2}$ Score
1	1.2390E-03	5.8935E-04	7.0983E-01
2	1.1990E-03	5.8953E-04	7.7515E-01
3	1.3334E-03	5.7178E-04	7.7191E-01
4	1.3014E-03	5.7218E-04	7.3193E-01
5	1.2291E-03	5.8706E-04	7.5349E-01
6	1.3986E-03	5.7323E-04	7.9335E-01
7	1.2908E-03	6.1106E-04	8.5538E-01
8	1.2995E-03	5.7474E-04	6.7857E-01
9	1.3130E-03	5.5819E-04	7.1044E-01
10	1.2465E-03	6.0129E-04	7.7207E-01
mean	1.2850E-03	5.8284E-04	7.5521E-01
sd	5.8367E-05	1.5785E-05	5.0477E-02

TABLE VIII

REGRESSION DECISION TREE BAGGING ENSEMBLE ALGORITHM STATISTICS. THE DATA WAS RANDOMLY SPLIT 10 TIMES. THE MEAN ABSOLUTE ERROR AND MEDIAN ABSOLUTE ERROR ARE TO BE MINIMIZED. THE R^2 SCORE IS TO BE MAXIMIZED, WITH 1 BEING CORRECTLY FITTED. THE MEAN AND STANDARD DEVIATION (SD) OF EACH TYPE OF VALUE IS REPORTED.

Fold	Mean Abs Error	Median Abs Error	$\mathbf{R^2}$ Score
1	1.0496E-04	5.8164E-05	9.9955E-01
2	9.2193E-05	4.9520E-05	9.9978E-01
3	1.1195E-04	5.9132E-05	9.9947E-01
4	1.0080E-04	5.5365E-05	9.9972E-01
5	1.1576E-04	6.4502E-05	9.9955E-01
6	1.0933E-04	5.1554E-05	9.9504E-01
7	1.0849E-04	5.7299E-05	9.9978E-01
8	1.0142E-04	5.2805E-05	9.9961E-01
9	1.3203E-04	6.1804E-05	9.7919E-01
10	9.4774E-05	5.2052E-05	9.9975E-01
mean	1.0717E-04	5.6220E-05	9.9715E-01
sd	1.1439E-05	4.8259E-06	6.4732E-03

performed 10 times.

IV. INSIGHT

A. Linear classifier

Linear models can be simple, but this can also cause problems if the data used is complex, so not giving us the whole picture. However, the more variables are included in multiple linear regression, the harder it is to predict the target. With a low number of variables, linear models can underfit the target. Underfitting is good for estimation but not if an accurate prediction is needed. Linear models have high variance and low bias, so they are much better for smaller datasets and can be improved with bagging ensemble models [7]. Since the dataset used in this project is large, the linear model with its underfitting did not perform as well as the tree and ensemble model. The model is fixed so there are no differences in model composition, compared with trees and rule which have models that can change, leading to less differences in implementations [7].

TABLE IX

PAIRED T-TEST OF THE MEAN ABSOLUTE ERROR OF THE REGRESSION TREE, MULTIPLE LINEAR REGRESSION AND BAGGING ENSEMBLE ALGORITHMS. THE ALGORITHMS ARE COMPARED AGAINST ONE ANOTHER AND THE MEAN, STANDARD DEVIATION (SD), STANDARD ERROR (SE), ONE-TAILED T DISTRIBUTION VALUE FOR α OF 0.05 (T) AND CORRESPONDING P-VALUE ARE INCLUDED.

Fold	Tree-Linear	Tree-Ensemble	Linear-Ensemble
1	-1.1442E-03	-1.0142E-05	1.1340E-03
2	-1.1047E-03	2.1585E-06	1.1068E-03
3	-1.2331E-03	-1.1655E-05	1.2214E-03
4	-1.2059E-03	-5.2611E-06	1.2006E-03
5	-1.0746E-03	3.8738E-05	1.1134E-03
6	-1.2950E-03	-5.7053E-06	1.2893E-03
7	-1.1850E-03	-2.7510E-06	1.1823E-03
8	-1.2037E-03	-5.6221E-06	1.1980E-03
9	-1.2131E-03	-3.2151E-05	1.1809E-03
10	-1.1536E-03	-1.8086E-06	1.1518E-03
mean	-1.1813E-03	-3.4199E-06	1.1779E-03
sd	6.4242E-05	1.7523E-05	5.4858E-05
SE	2.0315E-05	5.5412E-06	1.7347E-05
t	-5.8147E+01	-6.1719E-01	6.7898E+01
p-value	3.3139E-13	2.7620E-01	8.2354E-14
$p - value \leq \alpha$	TRUE	FALSE	TRUE

$\alpha = 0.05$
$df = 9$

TABLE X

PAIRED T-TEST OF THE MEDIAN ABSOLUTE ERROR OF THE REGRESSION TREE, MULTIPLE LINEAR REGRESSION AND BAGGING ENSEMBLE ALGORITHMS. THE ALGORITHMS ARE COMPARED AGAINST ONE ANOTHER AND THE MEAN, STANDARD DEVIATION (SD), STANDARD ERROR (SE), ONE-TAILED T DISTRIBUTION VALUE FOR α OF 0.05 (T) AND CORRESPONDING P-VALUE ARE INCLUDED.

Fold	Tree-Linear	Tree-Ensemble	Linear-Ensemble
1	-5.4996E-04	-1.8777E-05	5.3118E-04
2	-5.5030E-04	-1.0290E-05	5.4001E-04
3	-5.3218E-04	-1.9532E-05	5.1265E-04
4	-5.3261E-04	-1.5795E-05	5.1682E-04
5	-5.0475E-04	1.7808E-05	5.2255E-04
6	-5.3382E-04	-1.2145E-05	5.2167E-04
7	-5.7210E-04	-1.8344E-05	5.5376E-04
8	-5.3776E-04	-1.5825E-05	5.2194E-04
9	-5.1878E-04	-2.2391E-05	4.9639E-04
10	-5.6162E-04	-1.2387E-05	5.4924E-04
mean	-5.3939E-04	-1.2768E-05	5.2662E-04
sd	1.9865E-05	1.1380E-05	1.7372E-05
SE	6.2818E-06	3.5988E-06	5.4934E-06
t	-8.5865E+01	-3.5478E+00	9.5863E+01
p-value	9.9844E-15	3.1186E-03	3.7087E-15
$p - value \leq \alpha$	TRUE	TRUE	TRUE

$\alpha = 0.05$
$df = 9$

B. Tree-based

Trees have low time complexity, compared to other machine learning algorithms, due to their divide-and-conquer nature. However, trees can vary greatly with the dataset used.

Simple trees can be easy to interpret; however, the more complex the data and the more nodes need to be used, the harder it can be to interpret the tree. Features with more discrete values can greatly increase the number of child nodes, especially if a branch is created for each separate value.

TABLE XI

PAIRED T-TEST OF THE R^2 SCORE OF THE REGRESSION TREE, MULTIPLE LINEAR REGRESSION AND BAGGING ENSEMBLE ALGORITHMS. THE ALGORITHMS ARE COMPARED AGAINST ONE ANOTHER AND THE MEAN, STANDARD DEVIATION (SD), STANDARD ERROR (SE), ONE-TAILED T DISTRIBUTION VALUE FOR α OF 0.05 (T) AND CORRESPONDING P-VALUE ARE INCLUDED.

Fold	**Tree-Linear**	**Tree-Ensemble**	**Linear-Ensemble**
1	2.8963E-01	-8.6008E-05	-2.8972E-01
2	2.2434E-01	-2.9524E-04	-2.2463E-01
3	2.2749E-01	-7.5400E-05	-2.2757E-01
4	2.6766E-01	-1.2594E-04	-2.6779E-01
5	2.4566E-01	-4.0026E-04	-2.4606E-01
6	2.0602E-01	4.3386E-03	-2.0169E-01
7	1.4393E-01	-4.7605E-04	-1.4441E-01
8	3.2093E-01	-1.1535E-04	-3.2104E-01
9	2.8893E-01	2.0172E-02	-2.6875E-01
10	2.2754E-01	-1.4069E-04	-2.2768E-01
mean	2.4421E-01	2.2796E-03	-2.4193E-01
sd	5.0516E-02	6.4490E-03	4.9201E-02
SE	1.5974E-02	2.0393E-03	1.5559E-02
t	1.5288E+01	1.1178E+00	-1.5550E+01
p-value	4.7820E-08	1.4630E-01	4.1246E-08
$p - value \leq \alpha$	TRUE	FALSE	TRUE

$\alpha = 0.05$
$df = 9$

Pruning can be done to remove the number of nodes, but it can only be performed on classification trees. ROC curves are good for visualizing how well the tree performs, but this can again be only used on classification.

Regression trees are best for continuous targets but are susceptible to overfitting [7].

C. Distance-based

Depending on the clustering algorithm used, the algorithm may be supervised or unsupervised. In the case of this project, the clustering algorithm used, feature agglomeration, is unsupervised. Since unsupervised algorithms do not use a target comparing results against predictions, they cannot be evaluated the same as the other algorithms.

A problem that arises with agglomeration is that the number of clusters needs to be stated in advance, causing there to be more trial-and-error before getting adequate results. Due to their complexity, clustering algorithms can have a high time complexity and may not be adequate for large datasets [7].

D. Rule-based

Rule-based methods are not good with continuous data and this was seen in this project. Ordered and unordered rules are sensitive to continuous features and will not work. Association rules can use continuous features, but the targets cannot be continuous. The number of features used in this dataset can also cause problems in creating rules. The number of rules may become unmanageable and may not give good results.

E. Ensemble

Ensemble methods can improve on base models, but this can increase the time complexity. If the base model gives good results, it will still be better than the ensemble method. Bagging can reduce variance so works best with high variance and low bias, and boosting can reduce bias so works best with low variance and high bias [6].

Bagging is a good ensemble method to use on trees [7]. Bagging was found to be the best ensemble method to use because the regression decision tree model was found to be good for the data. Since bagging uses subsampling spacing, this can cause an increase the space complexity [7].

V. SYNTHESIS

Regression based algorithms were best suited for the dataset because the dataset was a mix of discrete and continuous features and with a continuous target. The tree and ensemble algorithms performed better than the linear model because they are much better at handling the large number of features.

Although bagging gave good results along with the tree algorithm, bagging is not always needed. The tree algorithm worked as well as the bagging algorithm, and even better in terms of median absolute error. Since there was no statistical difference in R^2 score between the tree and ensemble algorithm, the tree algorithm is optimal. The tree algorithm performs well while taking less computational time, while bagging had a worse error and took a lot longer in computational time.

The regression methods were better suited for predicting carbon because the models gave a better cause-and-effect than the clustering method, but clustering can give different results that can be interpreted along with the regression model results.

The cluster method gave information about other connections (Fig. 4). Carbon, total and population were clustered together in all of the clustering algorithms. The correlation between these features can be seen in Figure 1. UN region, subregrion, year, record and percapita GDP were clustered together in both the 3- and 4-cluster algorithms, while this cluster is split up in the 5-cluster algorithm. ISO alpha-3 code, crop land, fishing ground and built-up land were clustered together in all of the clustering algorithms, and UN region and subregion were added in in the 3-cluster algorithm. Grazing land and forest land were clustered together in all of the clustering algorithms.

VI. CONCLUSION AND FUTURE WORK

The best predictor of carbon was found to be the regression decision tree algorithm, followed closely after by the regression decision tree bagging ensemble algorithm, and then by the multiple linear regression algorithm. The feature agglomeration clustering algorithm gave other insights into the dataset, such as the connection between carbon, total and population. Rule-based methods were not useful for the dataset used.

Future work can be done on the dataset to find the best predictors for other features, and more clustering algorithms could be done on the data set to visualize more clusters and connections between features. If other features pertaining to climate change are included in the data, more can be learnt about the variables affecting climate change.

Machine learning algorithms were found to be useful for making predictions on climate change related data.

References

[1] Global Footprint Network. (2018). National Footprint Accounts 2018: The Ecological Footprint of 196 Countries, Version 8. Retrieved January 16, 2019 from https://www.kaggle.com/footprintnetwork/national-footprint-accounts-2018.

[2] RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

[3] Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830.

[4] J. Han, J. Pei, Y. Yin, R. Mao. (2004). Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery 8, pp. 53-87. Retrieved April 17, 2019 from https://orange3-associate.readthedocs.io/en/latest/scripting.html.

[5] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B. (2013). Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research 14, pp. 2349-2353.

[6] A. Ravanshad. (April 27, 2018). Ensemble methods. Retrieved from https://medium.com/@aravanshad/ensemble-methods-95533944783f.

[7] P. Flach. (2012). Machine Learning: the art and science of algorithms that make sense of data. Cambridge.