# Predicting a Country's Sustainability Using Artificial Neural Networks and Public Datasets

AI and Sustainability Special Interest Group 2

# 1. INTRODUCTION

In this rapidly expanding world, the environment and sustainability are pressing concerns. Climate change is endangering human society: rising sea levels threaten to overtake coastal cities, changing weather reduces crop yields, and natural disasters are cropping up more often (Impacts on Society, n.d.). These climate issues are largely caused by unsustainable practices, meaning that people use more materials than what nature can provide. Unsustainability can also result in social issues such as job loss, bad water quality, and poverty (Brova, 2009). To prevent unsustainable practices, data and analytical models can be used to guide planning and decision-making. Unfortunately, collecting sustainability data often involves surveying various entities and individuals, which are expensive and time-consuming processes. Furthermore, there are often missing values in survey data because of oversights, logistics issues, uncooperative subjects, etc. Standard statistical models need many data entries and very clean data to accurately predict future sustainability metrics (Bzdok, Altman, & Krzywinski, 2018). This project proposes to use artificial neural networks (ANN) to predict whether a country is sustainable or unsustainable. Because ANNs can learn hidden correlations between features, they can possibly infer the label based on fewer data points (Wilson, Dougherty, & Davenport, 2019). The results of this project will potentially assist in sustainable decision-making.

## *1.1 General artificial intelligence*

The degradation of the natural environment and the climate crisis is a serious issue in today's world that requires the most advanced and innovative solutions. This is where artificial intelligence (AI) comes into play. AI is a branch of computer science that allows computers to make predictions and decisions. It takes in large amounts of data to then interpret the world around them, digest and learn from information, make decisions based on what they have

learned, and finally take appropriate action. AI is revolutionizing all sorts of fields from healthcare to agriculture to automation. In terms of sustainability, AI can help transform crop yields, carbon pollution, and the wastefulness of resources. By better monitoring and managing environmental conditions and crop yields, as well as reducing fertilizer and water, farmers can significantly improve crop yields. Similarly, companies such as Stem, ClimaCell, and Foghorn Systems are trying to manage the demand and supply of renewable energy through the use of deep predictive capabilities and intelligent grid systems in AI (Gow, 2020). AI is also being used in transportation to reduce traffic congestion, improve the transport of cargo, and allow for more autonomous driving capability. It can even be seen in the manufacturing industry where it is used to reduce waste and energy use in production facilities. AI has the capabilities to do great things, one of them being giving humans a chance to live in a sustainable future.

*1.2 Artificial Neural Networks*

Neural Networks (Artificial Neural Networks/ANNs) are a machine learning tool for computers to learn to perform tasks after being fed training examples. For example, for the task of object recognition, the neural network would be fed labeled images of houses, cars, objects, etc. The ANN would learn from this data and find patterns in the images that relate them to their labels. ANNs can also predict continuous values. ANN's mimic the human brain in the sense that they have nodes arranged in layers, assign weights to incoming connections, and fire signals when activated (Malik, 2019).

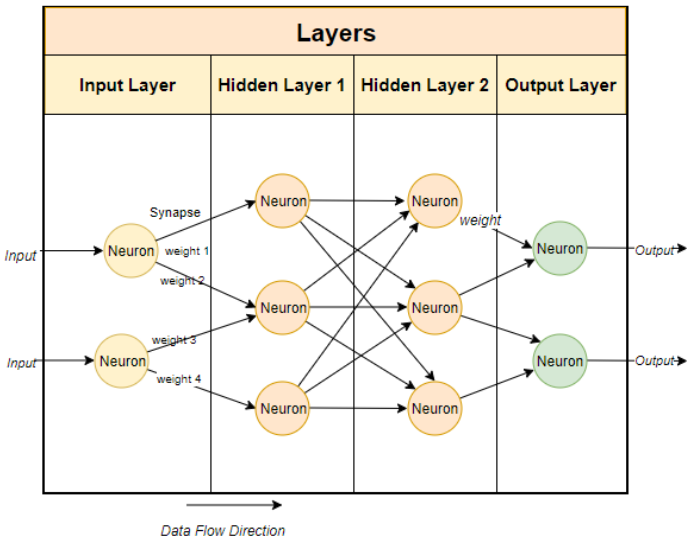


Figure 1: General ANN structure

As shown in Figure 1, a Multi-Layer Perceptron Neural Network has the following structure:

1. Input Layer has nodes that pass information to the next layer (hidden layer)
2. Hidden Layers of nodes connected to all nodes in the layer before it and to all nodes in the next layer. Data flows in the forward direction only. Each node assigns a weight to its incoming connections and if the sum of product of the weights with the connection value exceeds a threshold value, the node “fires” a signal (with this sum as data) to all its outgoing connections. This is illustrated in Figures 2 and 3.
3. An activation function or transfer function defines the output of a node given an input or set of inputs.
4. Weights and Thresholds are initially set to random values. Training data is fed to the nodes in the input layer and it passes through each hidden layer (where it is multiplied with weights and summed) eventually reaching the output layer. These weights and thresholds are constantly adjusted until the model is accurate- ex. It consistently recognizes an image of a dog as a dog.

5. Back propagation is used to calibrate the weights to be used in the network to minimize the margin of error (Babs, 2020).

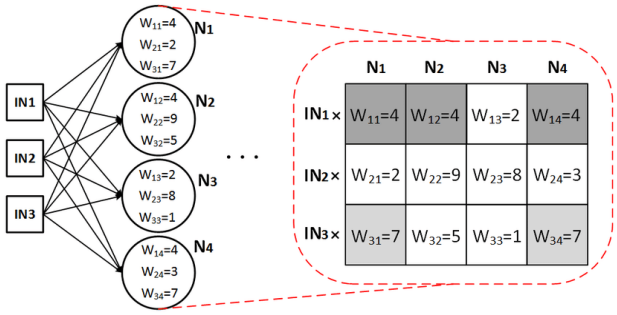


Figure 2: Example ANN Calculations

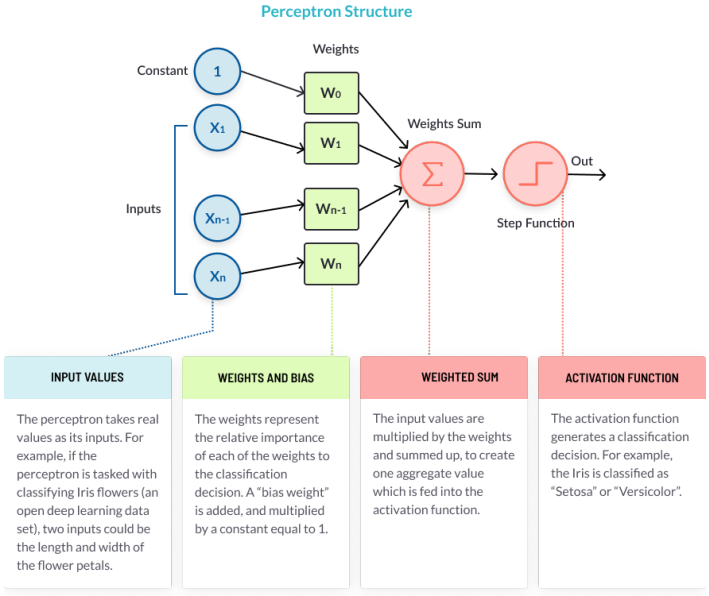


Figure 3: One node/perceptron of an ANN(Perceptrons & Multi-Layer Perceptrons, n.d.)

*1.3 Current Work*

The proposed research is inspired by the research work "Using Machine Learning Tools to Classify Sustainability Levels in the Development of Urban Ecosystems" . This paper looks at the United Nations' Sustainable Development Goals that apply to Bogota, Columbia. Although the paper accurately demonstrates the sustainability of this region, there are still a few limitations

that can be expanded on in the future (Molina-Gómez, Rodríguez-Rojas, Calderón-Rivera, Díaz-Arévalo, & López-Jiménez, 2020).

Although it does include factors like poverty, hunger, education, and economic growth, the majority of the goals do not focus on environmental factors.(The 17 goals, 2020). Instead, by focusing only on the environment, sustainability classifiers are more likely to be accurate because climate change is the cause for most of these problems (The Effects of Climate Change, 2020). Secondly, the sustainable development goals for the environment do not go into as much depth as the World Development Indicators dataset that is provided by the World Bank. Widening the scope by looking at multiple goals is not an effective way to have more data. To understand these indicators, and their impacts, there should be more concentrated data on each goal: a limitation of the sustainable development goals. Thus, focusing on fewer indicators may produce more accurate results in future research.

To compensate for the reduced amount of data in each sustainable development goal, the paper focuses on Bogota, Columbia rather than trying to implement their model to a global scale. However, to understand the global implications of sustainability, it is necessary to expand past the city regions. Both datasets in this research paper look at sustainability on the international level.

## 2. METHODS

The aim of this project is to use ANNs to predict if a country is sustainable or unsustainable. Two free, public datasets were selected from Kaggle.com to be used in this project.

### *2.1 World development indicators dataset*

The “World Development Indicators by Countries” database, updated ten months ago on the website *Kaggle,* was utilized for the research. The database contains information from the World Bank, which fights poverty while also prioritizing sustainability. As the World Bank website describes, the data was collected from “officially-recognized international sources,” and the database presents the “most current and accurate” information relating to global development (Who we are, 2020). Data on a total of 214 countries with varying economies is stored in the database, and the data has been collected since 1960. Information was updated quarterly by the World Bank. The data was collected about a variety of topics and some of them are briefly discussed as follows.

Freshwater data relating to each country was one focus within the database. The information involves: yearly freshwater usage, taking into account many circumstances; the percentage of yearly freshwater use for resources and industry in 2014, and for agriculture and domestic means in 2015; and the percent of the population with access to drinking water services, with a further focus on urban and rural population in 2018.

Emissions were additionally observed with a focus on energy dependency, efficiency, and carbon dioxide (CO2) emissions. The emission data was similarly compiled in a file, which specifically contains: the percentage of total energy use in 1990 and 2015; the gross domestic product (GDP) of energy use in 1990, reported in the 2011 equivalent of dollars per kilogram of oil; the gross domestic product (GDP) of energy use in 2015, reported in the 2011 equivalent of dollars per kilogram of oil; carbon dioxide emissions in 1990 and in 2014, focusing on the burning of fossil fuels; the carbon dioxide emissions in 1990 and 2014, specifically emissions from burning coal; and the carbon dioxide emissions per capita from 1990, focused on burning fossil fuels and cement creation.

Pollution, although not a specific file in the database, could be observed through other information. Surrounding particulate matter 2.5 (PM2.5) within the sustainability file depicts the presence of air pollution in the studied countries. The data on PM2.5 reports on concentrations of PM2.5 in urban and rural areas.

Furthermore, energy access data for each observed country was also included in the database, focusing specifically on sustainable energy. The information reports on: the percent of the population with access to energy in 2000 and 2016; the percent of urban and rural populations with access to energy in 2016; the percent of the populations in 2000 and 2016 with access to clean fuel and technologies for cooking; the percentage of renewable energy consumed in 2000 and 2015 out of the total energy consumed; and the percent of renewable electricity produced out of total electricity produced.

*2.2 Ecological footprint dataset*

The "National Footprints Accounts 2018" database (2018), updated three years ago on Kaggle, was also utilized in the project. This database is an annual production from Global Footprint Network, an international nonprofit organization that works to help end ecological overshoot by making ecological limits central to decision-making (Global Footprint Network, n.d.). Every year, the organization works hard to combine and synthesize over 30 datasets to calculate the Ecological Footprint and biocapacity of countries across the world. The overall dataset that we are using shows the results for 196 countries in the world from 1961 through 2014. For our research purposes, we only looked at data values from the "total" column regarding biocapacity and effective total consumption in the year 2014.

The "total" column in the dataset is a sum of all the other columns' values that represent the number of global hectares of different land types (crop land, grazing land, forest land, etc.)

either required to support consumption or production, or that are supported by biological capacity (biocapacity). Biocapacity refers to the ability of an ecosystem to produce useful biological materials and to absorb its spillover wastes. The Ecological Footprint of consumption indicates the consumption of biocapacity by a country's inhabitants and can be calculated by adding the ecological footprint of production and the net ecological footprint of trade together. When the ecological footprint is greater than biocapacity, the country is unsustainable and when it is less than biocapacity, the country is sustainable.

*2.3 Training details*

Data was processed using pandas (for creating dataframes) and numpy (array operations), two Python packages commonly used for data analysis. To clean data, the data points with NaN (not a number) values were dropped. The different features from the World Development Indicators dataset were matched up based on country (i.e., all the United States's data were put in one dataframe row).

The data used in the experiments was selected based on three criteria. First, the data had to be relevant to sustainability and biocapacity. Second, there could not be many NaN values for the data. Since the dataset is not big to begin with, NaN values should be avoided. Lastly, the data must have been collected around 2014. That way, these data will be relevant to the effective consumption and biocapacity data used as labels.

Given a country's feature set, the ANN will classify it as sustainable or unsustainable. To determine sustainability, the national footprint dataset was used. The total effective consumption of a country in 2014 was subtracted from the country's total biocapacity in 2014. If the difference is negative, the country is unsustainable and vice versa. 2014 was picked to be the year because it is more recent and the World Development Indicator dataset had the most data entries around

2014. There two sets of data combinations were used in the investigation. The data entries contained in the two sets are summarized in Table 1.

Table 1: Different feature sets used as inputs for 2 ANN models

| Set 1 | Set 2 |
|---|---|
| Total greenhouse gas emissions % change 1990-2012 | Access to electricity % of population 2017 |
| Access to electricity % of population 2016 | Renewable energy consumption % of total final energy consumption 2015 |
| Carbon dioxide emissions per capita metric tons 2014 | Carbon dioxide emissions per capita metric tons 2014 |
| Renewable energy consumption % of total final energy consumption 2015 | Agricultural land % of land area 2014-16 |
| Fertilizer consumption kilograms per hectare of arable land 2014-16 | Agricultural employment % of total employment 2014-16 |
| 'Annual freshwater withdrawals % of internal resources 2014 | Carbon dioxide emissions kilograms per 2011 PPP $ of GDP 2014 |
| Carbon dioxide emissions per capita metric tons 201 | Rural population % of total 2018 |
| Annual freshwater withdrawals % of internal resources 2014 | Land use Forest area % of land area 2016 |
| | Arable land hectares per person 2016 |
| | Renewable electricity output % of total electricity output 2015 |
| | Access to electricity % of urban population 2016 |

The ANN regressor and classifier was built using sklearn's multilayer perceptron regressor and classifier, respectively. The GridSearchCV function was used to exhaustively search different hyperparameters such as activation function, solver, and network structure. For a given network, it was trained and scored 25 times with different training testing splits in order to

obtain consistent results. Also, the regressor and classifier for each feature set used the same ANN structure so as to maintain consistency.

## 3. RESULTS

The ANN regression model that predicted the numerical value of the consumption was tested on feature set 1 and achieved an average accuracy score of -0.021. The standard deviation of the scores obtained in the 25 runs is 0.037. The negative score indicates that the regression model has a very poor performance on predicting the numerical values of consumption. The ANN regressor model used in the experiment has a single hidden layer with 20 nodes. The solver was Adam, and the activation function was ReLu.

The ANN model that classified countries as sustainable or unsustainable achieved an average accuracy score of 0.746. The standard deviation of the scores from 25 runs is 0.061. Evidently, the classifier is much better than the regressor. Figure 4 shows the training and validation accuracy curves during the training with one split of data set 1.

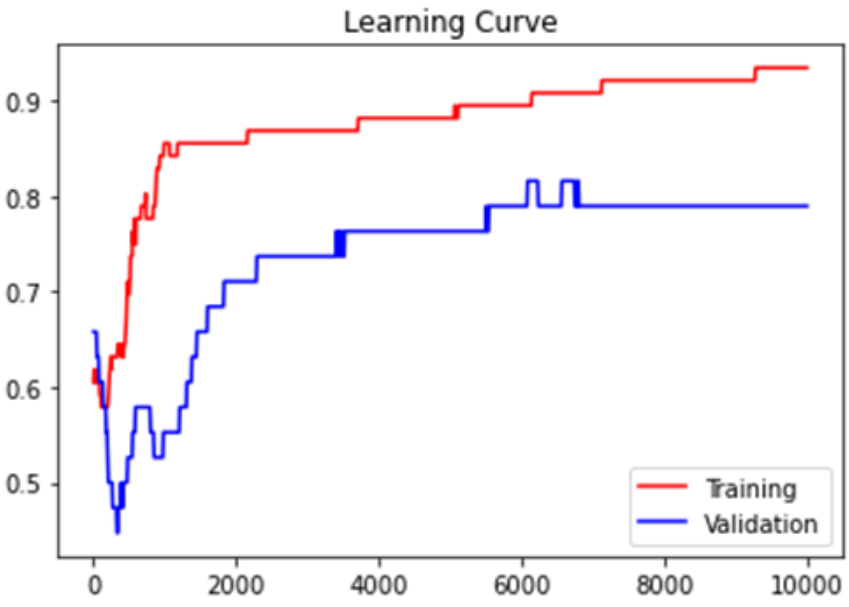


Figure 4: Feature set 1 network training and validation curve

The ANN regressor for feature set 2 achieved an average accuracy score of -0.0085 and the standard deviation of the accuracy scores is 0.020. The classifier achieved an average accuracy score of 0.765 with standard deviation of 0.058. In the experiment with data set 2, the ANN has two hidden layers, which have 16 and 5 nodes, respectively. Once again, Adam was the solver, and ReLu was the activation function. Figure 5 shows the training and validation accuracy curves during the training with one split of data set 2.

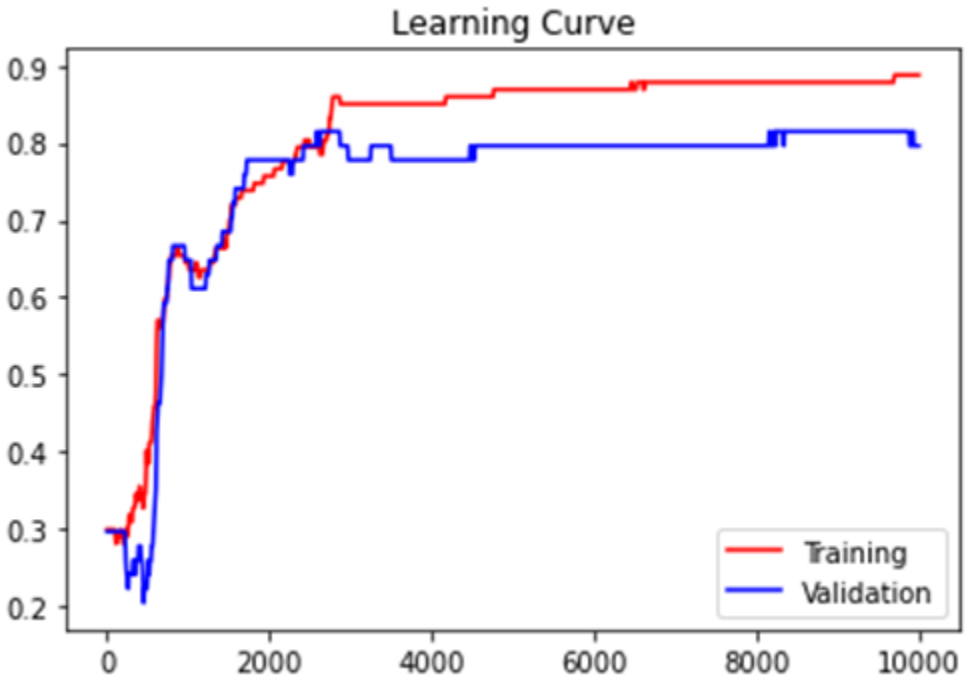


Figure 5: Feature set 2 network training and validation curve

The experimental results indicated that the regression models always had much worse performance compared to the classifier model. Also, the dataset was very small, which negatively affected the model accuracy. Ideally, there should be one data point for one country. However, the selected datasets also had some null values, so certain countries were dropped. Unfortunately, this is the nature of working with free, publicly sourced datasets involving countries.

# 4. CONCLUSION

In this project, ANN models were implemented to predict a country's sustainability. Two datasets were used: World Development Indicators and National Footprint Accounts. Both datasets are free, public, and sourced from Kaggle.com.

Both ANN regression and classification models were explored to predict countries as sustainable or unsustainable. The inputs of the models were a set of sustainability related features organized by country. Experiments were conducted with two separate feature sets.

The ANN regressor achieved average accuracy scores of -0.021 with data set 1 and -0.0085 with data set 2 (see Table 1 for specific feature sets). This means that the ANN regressor's predictions had larger errors, which is partially because the correlation between input and output is very weak. This is reasonable because every country's situation is different and cannot be fully characterized by the limited data entries in the datasets.

The ANN classifier achieved accuracy scores of 74.6% with data set 1 and 76.5% with data set 2. Thus, the classifier can classify countries as sustainable or unsustainable with moderate accuracy. The small accuracy difference between feature sets 1 and 2 may be attributed to the fact that feature set 2 had more data entries than set 1. With a larger dataset, it is reasonable to assume that the inference accuracy will further increase for both trials. The ANN classifier was able to perform much better than the regressor since classification is a simpler task. The classifier was able to learn subtle correlations and use them to draw overarching conclusions. However, these subtle correlations are not enough to obtain a numerical value.

The main limitation of this project was that the world development dataset was small, unsystematic, and contained many missing values. In the future, this project can be improved by using larger, more complete datasets. Also, in this project, the features were selected based on

their relevance to sustainability, the date, and the amount of NaN values. Due to the limited time, only two sets of feature combinations were examined. In the future, Python programs will be developed to systematically explore different feature combinations to identify the best combination with the highest prediction accuracy. The utility package developed in the project for extracting and pairing data entries will be very helpful for this future effort.

REFERENCES


The 17 goals: Sustainable development. (2020). Retrieved February 24, 2021, from https://sdgs.un.org/goals

Babs, T. (2020, August 16). The mathematics of neural networks. Retrieved February 25, 2021, from https://medium.com/coinmonks/the-mathematics-of-neural-network-60a112dd3e05

Brova, G. (2009). The Emergence of Environmental and Social Sustainability. *BU Arts and Sciences Writing Program, 1*. doi:https://www.bu.edu/writingprogram/journal/past-issues/issue-1/brova/

Bzdok, D., Altman, N., & Krzywinski, M. (2018, April 03). Statistics versus machine learning. Retrieved January 23, 2021, from https://www.nature.com/articles/nmeth.4642

The effects of climate change. (2020, December 23). Retrieved February 25, 2021, from https://climate.nasa.gov/effects/

Global footprint network. (n.d.). Retrieved February 26, 2021, from http://www.footprintnetwork.org/

Gow, Glenn. "Environmental Sustainability And AI." Forbes, 21 Aug. 2020, www.forbes.com/sites/glenngow/2020/08/21/environmental-sustainability-and-ai/?sh=7f61bc867db3

Impacts on Society. (n.d.). Retrieved January 23, 2021, from https://www.globalchange.gov/climate-change/impacts-society

Malik, F. (2019, May 18). Neural networks bias and weights. Retrieved February 25, 2021, from https://medium.com/fintechexplained/neural-networks-bias-and-weights-10b53e6285da

Molina-Gómez, N. I., Rodríguez-Rojas, K., Calderón-Rivera, D., Díaz-Arévalo, J. L., & López-Jiménez, P. A. (2020). Using machine learning tools to classify sustainability levels in the development of urban ecosystems. *Sustainability, 12*(8), 3326. doi:10.3390/su12083326

National footprint accounts 2018. (2018, July 11). Retrieved February 26, 2021, from https://www.kaggle.com/footprintnetwork/national-footprint-accounts-2018

Perceptrons & Multi-Layer Perceptrons: The artificial neuron. (n.d.). Retrieved February 25, 2021, from https://missinglink.ai/guides/neural-network-concepts/perceptrons-and-multi-layer-perceptrons-the-artificial-neuron-at-the-core-of-deep-learning/

Who we are. (n.d.). Retrieved February 26, 2021, from https://www.worldbank.org/en/who-we-are

Wilson, H., Dougherty, P., & Davenport, C. (2019, August 21). The Future of AI Will Be About Less Data, Not More. Retrieved January 23, 2021, from https://hbr.org/2019/01/the-future-of-ai-will-be-about-less-data-not-more

World development indicators. (2020). Retrieved February 26, 2021, from https://datacatalog.worldbank.org/dataset/world-development-indicators