

|  |  |
| --- | --- |
| Sustainability Index | |
|  |  |
| Sushrut Shitoot28th April 2023 |  |

**Table of contents**

|  |  |  |
| --- | --- | --- |
| S. No. | Section | Page No. |
| 1 | Introduction |  |
| 2 | Methodology |  |
| 2.1 | Data collection |  |
| 2.2 | Variables |  |
| 2.3 | Data processing |  |
| 2.4 | Statistical methods |  |
| 3 | Results |  |
| 3.1 | Variable weights |  |
| 3.2 | Sustainability index |  |
| 4 | Discussion, recommendations and conclusion |  |

**Section 1: Introduction**

Sustainability is the ability to establish and preserve conditions over extended periods of time that maybe conducive to co-existence of all species with a special emphasis on how to make judicious use of natural resources that are exhaustible and are currently used extensively.

While the concept of sustainability is talked about extensively, there exist limited measures to quantify its levels in terms of certain benchmarks accepted worldwide. There is a need to standardize the assessment of level of sustainability across different contexts.

Particularly in the field of education, institutes propagate the concept of sustainability as a part of the curriculum. It would do wonders to have practical demonstrations where institutes walk the talk about sustainability. Energy consumption across all mass gatherings, especially educational institutions where kids and young adults who do not fathom the gravity of the energy situation and the need for sustainable outlook towards the world spend a large portion of their day, it becomes imperative to impart the values that will showcase the importance of sustainability right from their formative years. The concept of sustainability is indispensable at these locations.

Data pertaining to about 30 features that impact sustainability as a concept was made available. The feature set corresponds to attributes of schools that may have an impact on the sustainability. However, no concrete measure of sustainability exists as of now. There is a need to create an index that may reflect the level of sustainability prevalent at each of these schools represented as different observations within the dataset.

A mix of heuristic, logical and advanced analytical measures were adopted to devise and deploy a methodology that helped create a measure of sustainability across these schools. The following sections in the document elicit the base data used, logic and concept behind the calculation, the exact methodology that was used, the outcomes and triangulation of these outcomes with multiple approaches, interpretation of these outcomes and eventually some highlights and limitations of the methodology itself.

**Section 2: Methodology**

**Section 2.1: Data Collection:**

The base data that was made available held attributes of a list of 54 schools / educational institutes that describe their degree of energy consumption, measure put in place to conserve energy, facilities that assist in achieving this goal and certain other criteria that may assist in understanding how sustainable is the premises.

A few sample variables from this dataset include (non-exhaustive list):

1. Energy consumption (KWh/m2)
2. Presence / absence of solar panels
3. Water consumption (L / person)
4. Green area percentage
5. Presence / absence of rain water tank

**Section 2.2: Variables:**

Here is the complete list of variables used in the analysis

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Feature** | **Description** |
| 1 | Energy Consumption KWh/m² | Energy consumption per year per school |
| 2 | Solar Panel | 1= available; 0 = not available |
| 3 | Insulation in walls and ceiling | 1= available; 0 = not available |
| 4 | Shading in South elevation | 1= available; 0 = not available |
| 5 | Shading in East elevation | 1= available; 0 = not available |
| 6 | Shading in West elevation | 1= available; 0 = not available |
| 7 | Water Consumption L/person | Water consumption per person in this year |
| 8 | Rainwater Tank | 1= available; 0 = not available |
| 9 | Sanitation | 1= available; 0 = not available |
| 10 | Sanitation Efficiency | 1= available; 0 = not available |
| 11 | Green Area % | Percentage of green area as part of the school campus |
| 12 | Shading Area % | Percentage of shaded area as part of the school campus |
| 13 | Parking to Teacher % | proportion of parking spots available per teacher on the school premises. |
| 14 | Closest to Public transportation | Closest public transportation point to the school campus in meters |
| 15 | Accessibility | 1= easy access; 0 = difficult access |
| 16 | Heat Island Effect | 1= available; 0 = not available |
| 17 | Waste Separation | 1= available; 0 = not available |
| 18 | Maintenance cost NIS/yr | Cost of maintenance per year |
| 19 | New Teachers # | Number of newly hired teachers this year |
| 20 | New Staff # | Number of newly hired staff this year |
| 21 | Social activities | 1= social activities; 0 = no social activities |
| 22 | Low-income students % | Percentage of low-income students in the school |
| 23 | Disabled students % | Percentage of disabled students at the school |
| 24 | Teacher Participation (%) | Percentage of teacher participation in decision making at the school for this year |
| 25 | Parents Participation (%) | Percentage of parents participation in decision making at the school for this year |
| 26 | Students Participation (%) | Percentage of students participation in decision making at the school for this year |
| 27 | Attendance Rates (%) | Percentage of students attendance at the school for this year |
| 28 | Sick Leave (%) | Percentage of students who took sick leave from the school for this year |
| 29 | Success Rate (%) | Percentage of students who succeeded their study this year |
| 30 | Dropout rate (%) | Percentage of students who left their study/left the school this year |

**Section 2.3: Data processing**

The base data was available across multiple data types (classes), predominantly numeric, integer and character. For the ease of processing, based on the real-world interpretation of the variables, data types (classes) were converted to only 2 – numeric and factor (to be processed in R)

R was used as the preferred choice of programming language and the IDE used for the project was RStudio. A combination of RStudio and Microsoft Excel was used to arrive at the final result and create documentation as well.

1. Basic calculations including range transformation, dummy coding, variable encoding, creation of a pseudo-index and, finally, weight calculation and creation of the sustainability index was carried out in Microsoft Excel
2. Advanced calculations including univariate analysis, correlation checks, feature selection techniques (Boruta), linear models for validation and tree-based algorithm (Random Forest) and respective calculations for variable importance were carried out in RStudio

Here are a few highlights and observations from the data processing:

1. It was observed that the different numeric variables in the dataset had different ranges (minimum and maximum values)
2. It was imperative to bring the variables in the same range to make them comparable, create an index based on them and also investigate them statistically
3. A range transformation was used to bring all the numeric variables to the range of 0 to 1
4. All categorical variables were binary variables with levels o and 1 indicating presence or absence of certain attributes
5. **Assumption:** For the purpose of data evaluation and consumption, categorical variables with the 2 levels as 0 and 1 were considered as such as dummy numeric variables to include them in calculations and were reverted to a yes / no state for the purpose of the modeling
6. Since no measures of sustainability existed in the dataset, it was not possible to estimate the impact of each of the variables on sustainability. The project called for an estimated measure of sustainability in the form of a pseudo-index for sustainability
7. **Assumption:** This pseudo-index was calculated using all the variables in the dataset giving them equal weights to reflect the possible contribution of all attributes
8. **Disclaimer:** While for the purpose of the creation of pseudo-index, all variables were weighted equally as the first step, it is understood that different variables have different impacts on sustainability and this was used only as a base assumption to calculate a quantified index against each observation. The modeling exercise provided the estimates for weights / importance of each of the variables considered

**Section 2.4: Statistical Methods**

**For the purpose of this discussion and the project, all available data and features were considered to be independent variables that individually and cumulatively have an impact on the sustainability pseudo-index (dependent variable)**

The statistical analysis comprises of a range of simple to complex statistical techniques:

1. Univariate analysis: mean, median, mode, levels, standard deviation, missing values
2. Correlation: Pearson’s correlation coefficient between for independent variables with each other as well as all independent variables with the dependent variable i.e. the sustainability pseudo-index
3. Feature selection techniques (Boruta): A random forest-based wrapper algorithm that uses shadow features i.e. randomly shuffled values of the same variables to calculate the variable importance
4. Linear regression: Linear model was used to validated the variables deemed important by Boruta to understand what percentage of variation in the dependent variable was explained by them
5. Random Forest: A tree-based algorithm was used to again identify the most important variables to use a different approach to identifying the most important variables
6. Finally, the variable importance calculated through Boruta was used to create a weight for each variable that was then used for calculating the sustainability index

**Section 2.5: Approach**

1. The overall approach is based on the assumption that the contribution of variables to the final sustainability score that needs to be calculated based on these metrics is varied and is based on the importance of each of the variables.
2. Range transformation: All variables were brought to the same range of 0 to 1 using a range transformation.

*New value = (Old value – Minimum value) / (Maximum value – Minimum value)*

1. Calculation of the sustainability pseudo-index: Since no measure of sustainability existed in the dataset, a pseudo-index was created by taking a sum of all the transformed variables. It captures the additive nature of different initiatives towards increasing sustainability
2. Further, the sustainability pseudo-index was considered to be the dependent variable and the variables in the base data were considered to be the independent variables
3. Perform univariate analysis: Measures of central tendency, range etc.
4. Exploratory data analysis:
   1. Correlation of independents with the dependent variable
   2. Correlation matrix for all independent variables
5. Feature selection: Boruta, which is a wrapper algorithm based on random forest, was used to calculated feature importance and also to identify which were the most importance features vs which features cannot be deemed important
6. Model Building: Linear regression and random forest models were built to validate findings and outcomes from Boruta. Variables shortlisted from Boruta, when passed through a linear regression model, indicate a ~70% explanation of variance in the dependent variable.

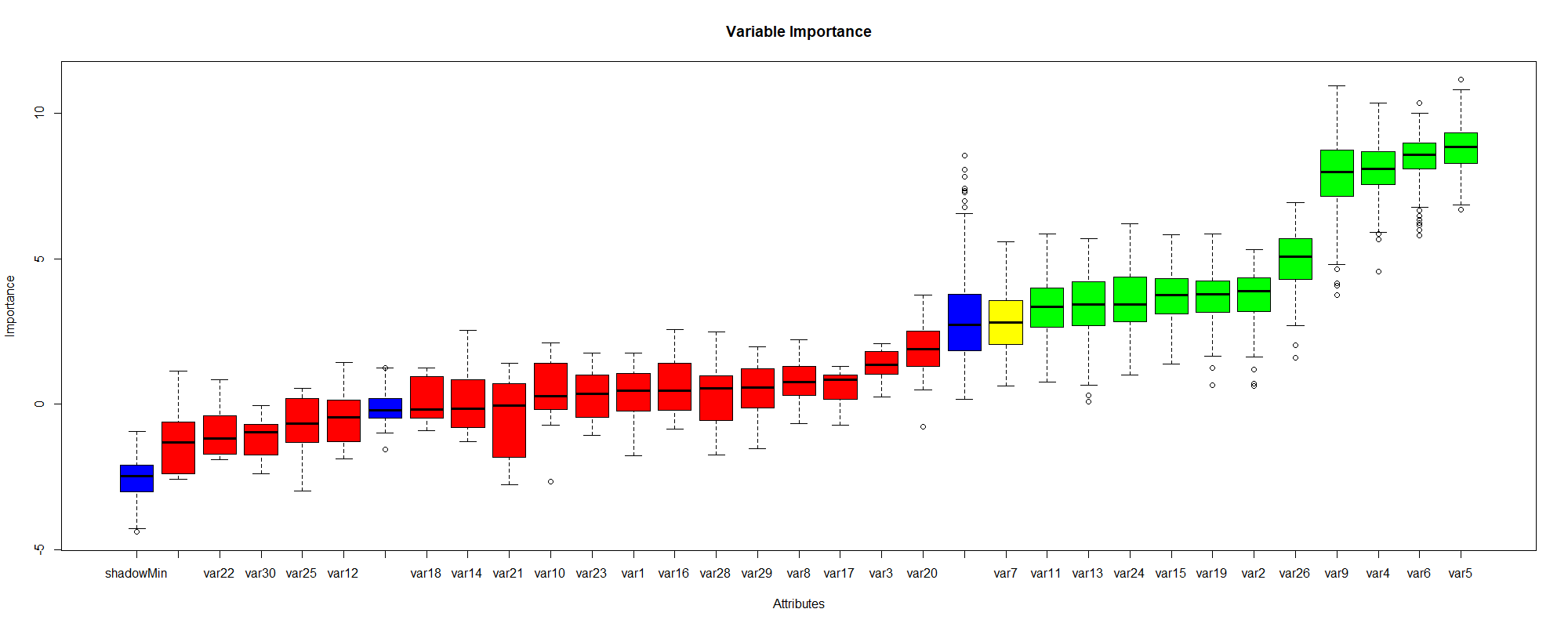
**Section 3: Results**

**Section 3.1: Variable weights**

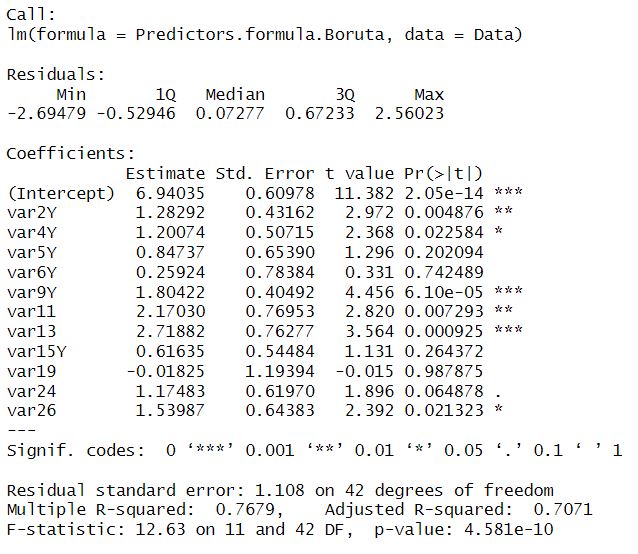
Following table indicates the variable importance and decision around whether a variable must be considered important or not:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Code | Description | Decision | Importance | Weight |
| var5 | Shading in East elevation | Confirmed | 8.79 | 14% |
| var6 | Shading in West elevation | Confirmed | 8.50 | 13% |
| var4 | Shading in South elevation | Confirmed | 8.05 | 12% |
| var9 | Sanitation | Confirmed | 7.89 | 12% |
| var26 | Students Participation (%) | Confirmed | 4.99 | 8% |
| var2 | Solar Panel | Confirmed | 3.75 | 6% |
| var15 | Accessibility | Confirmed | 3.70 | 6% |
| var19 | New Teachers # | Confirmed | 3.66 | 6% |
| var24 | Teacher Participation (%) | Confirmed | 3.58 | 6% |
| var13 | Parking to Teacher % | Confirmed | 3.42 | 5% |
| var11 | Green Area % | Confirmed | 3.37 | 5% |
| var7 | Water Consumption L/person.yr | Tentative | 2.88 | 4% |
| var20 | New Staff # | Rejected | 1.86 | 3% |
| var3 | Insulation in walls and ceiling | Rejected | 1.36 | 2% |
| var8 | Rain water Tank | Rejected | 0.78 | 1% |
| var16 | Heat Island Effect | Rejected | 0.64 | 1% |
| var17 | Waste Speration | Rejected | 0.53 | 1% |
| var29 | Success Rate (%) | Rejected | 0.47 | 1% |
| var10 | Sanitation Efficiency | Rejected | 0.38 | 1% |
| var28 | Sick Leave (%) | Rejected | 0.33 | 1% |
| var23 | Disabeld students % | Rejected | 0.32 | 0% |
| var1 | Energy Consumption KWh/m² | Rejected | 0.31 | 0% |
| var18 | Maintinance cost NIS/yr | Rejected | 0.15 | 0% |
| var14 | Closest to Public transportation m | Rejected | 0.10 | 0% |
| var12 | Shading Area % | Rejected | -0.43 | -1% |
| var21 | Social activities | Rejected | -0.46 | -1% |
| var25 | Parents Participation (%) | Rejected | -0.71 | -1% |
| var22 | Low income students % | Rejected | -0.96 | -1% |
| var30 | Drop out rate (%) | Rejected | -1.15 | -2% |
| var27 | Attendance Rates (%) | Rejected | -1.30 | -2% |

**Variable Importance chart with selected, tentative and rejected variables**



**Findings from Boruta were confirmed using Linear regression model:**



**Section 3.2: Sustainability Index**

Using the weights calculated through Boruta, a weighted sum was calculated on the raw data to arrive at the final sustainability index.

With Boruta, all variables were not deemed important since their contribution to the variability in the assumed sustainability in the pseudo index was marginal. The ones that were deemed important explained ~70% proportion of variance in the sustainability pseudo-index thereby refuting our base assumption that all variables must be weighed equally for the calculation of the sustainability index.

The validation in the form of the linear model also showed statistical significance at multiple levels for the shortlisted variables (p < 0.001, p < 0.01, p < 0.05 and p < 0.1.

The reported R squared value was 76.79% and the adjusted R squared value was 70.71%.

**Section 4: Discussion**

The study showcased certain interesting observations and inferences that may be drawn. For instance, the base assumption that all variables contribute equally to the sustainability was refuted meaning variables have preferential and dissimilar contribution to the sustainability index.

The dataset was close to ideal with no missing values and was supposedly considerably clean compared to real-world data collection. As a result, efforts towards data cleaning and reconciliation were saved.

One extremely decisive factor to findings and inference is reliability of the dataset. The current dataset only has 54 observations rendering it unreliable statistically. Larger the dataset, more the number of datapoints for the models to learn from and hence more reliable the results. Since the current findings are based on limited datapoints, it would be unwise to generalize the findings in the real world unless a more robust study is undertaken.

Linear and random forest model was optimized on root mean square error (RMSE) as the preferred error metric of choice.