Impact of Country Indicators on Screen Time

Giuliano Rojas
Faculty of Social Sciences
Radboud University
Nijmegen Netherlands
giuliano.rojasmurillo@ru.nl

Sandro Kvrgić
Faculty of Social Sciences
Radboud University
Nijmegen Netherlands
sandro.kvrgic@ru.nl

ABSTRACT

Our paper focuses on investigating prediction models for screen time through country indicators. The research conducted uses linear regression and support vector machine models to best predict countries' screen time based on known indicators. The dataset used to train the models is a combination of three datasets coming from screen time by country, a country indicators dataset, and a world happiness dataset from recent years 2022-2023. From the 4 models we tested the second linear regression model resulted the best. It was found that the best attributes to predict the screen time of a country were through IQ rate and GDP.

INTRODUCTION

One of the silent contemporary problems of society as a whole is screen time. High screen time negatively affects people, causing various physical and psychological problems [1]. As this is a global social issue, we have decided to investigate and analyze how country indicators contribute to screen time. This consequently leads us to the aim of this study, which is to be able to predict screen time based on country indicators for the countries that do not have any data on their screen time. As we know some country data is hardly acquirable, because of the political situations, economic status, and many other reasons [2]. Therefore, being able to estimate an attribute based on country indicators which are mostly easily accessible holds immense value not only for screen time but also for any other attribute of interest. We decided to use three data sets containing up-to-date country indicators for various countries around the world. Using various methods which will be explained in detail later. We ended up with one data set containing 39 samples and 14 attributes, on which we implemented linear regression models and Support Vector Machines with Radial Basis Kernel. In the end, we will compare both methods and draw a conclusion about which one performs better.

LITERATURE REVIEW

There are no studies on which we continued our research, but there are some similar related studies. A study by Nagata et al. (2021), focuses on the correlation between sociodemographics and screen time with children aged nine and ten [3]. They implemented multiple linear regression models, using various sociodemographic attributes. Results of this study show that screen time is correlated with ethnicity, sex, and income of the household, as there have been consistently different results based on those modalities. Another study by Sourtilij et al. (2019), which was conducted on 322 children under the age of five, shows that screen time negatively affects sleep duration in children, while also having a positive correlation with body mass index, meaning that it has a significant impact on both of these attributes [4].

DATA SET & DATA COLLENCTION

We started our research with three data sets, country indicators, world happiness, and screen time, which were all obtained from the Kaggle website [5][6][7]. The country indicators data set contained 116 countries with 12 attributes describing various basic country indicators. Containing 137 countries and 21 in-depth attributes is the world happiness dataset. Most of the attributes from this data set held no value for our research. Lastly, the screen time data set, containing only 44 countries and 15 attributes which were important to our research. Our initial preprocessing of the data was to create a collection of variables from each of the datasets to create our own merged dataset. This refined dataset would consist of 39 countries in common that were shared among the three different datasets. The dataset would then be reduced by getting rid of unrelated measurements or variables we believed gave redundant data. For example, we removed the measurements of 'Internet via Mobile' and 'Internet via Computer' because they were already accounted for in 'Total Time Spent on Devices'. Other variables that were not of relevancy were also removed, like variables that tracked the explicability of other variables instead of directly indicating measurements of the countries. In the end, we managed to reduce the number of attributes from 48 to only 14. We made sure that the dataset consisted of no missing values. The final merged data set consisted of 39 samples with 14 attributes, out of which 10 were predictor variables.

	Country object	health_index float64	obese_rate float64	suicide_rate float64	democracies_rate f...	hdi_rate float64	life_Exp float64	iq_rate
0	Argentina	54.4	28.3	8.4	6.81	84.5	76.064	
1	Australia	71.1	29	12.5	8.9	94.4	83.579	
2	Austria	56.9	20.1	14.6	8.07	92.2	82.412	
3	Belgium	59.3	22.1	18.3	7.51	93.1	82.293	
4	Brazil	51.2	22.1	6.9	6.86	76.5	73.425	

Figure 1: Final merged complete data set

PREPROCESSING METHODS ON FINAL DATA SET

After the making of the data set, we had to use various methods to be able to work with the data. After careful examination of the data, we searched for outliers and did not find any that would require us to take measures. In time-related columns, we converted all of the hours into minutes, and in some rate columns we adjusted the scales so that they were on the same scale as the other rate columns, it was done by multiplying them by 100. The iq rate was also rescaled by subtracting 10 from every row so that the maximum Iq rate is 100. After we had scaled all of the data, we standardized it. Every column was standardized separately, because of the difference in theme, distribution, and values, ensuring that attributes can not impact each other's standardized values.

Country object	health_index float64	obese_rate float64	suicide_rate float64	democracies_rate f...	hdi_rate float64	life_Exp float64	iq_rate fl
Argentina	-0.1730220306	0.786772452	-0.38817331	-0.07334883719	-0.1705558588	-0.4606541767	-0.59
Australia	1.574098102	0.864707459	0.302248406	0.9866979216	0.9949594881	0.8900078885	0.78
Austria	0.08852289939	-0.1261804876	0.655879041	0.5657224145	0.7359560776	0.6802643582	0.69
Belgium	0.3396060322	0.09649096109	1.278942541	0.2816907471	0.8419120183	0.6588786289	0.59
Brazil	-0.5077995411	0.09649096109	-0.6407666207	-0.04798886688	-1.112386442	-0.9349585267	-0.9

Figure 2: Standardized final merged data set

EXPLORATION OF RELATIONSHIPS IN DATA

With the data standardized it was time to try and find some pattern in the data. Using a scatterplot and plotting independent variables against dependent variables, thus allowing us to see what kind of relationship they have. By plotting each possible pair of attributes, we ended up with a large number of scatterplots, but that way we could see if there were any patterns present. As a result, we could see that there was some linear relationship in most of the data.

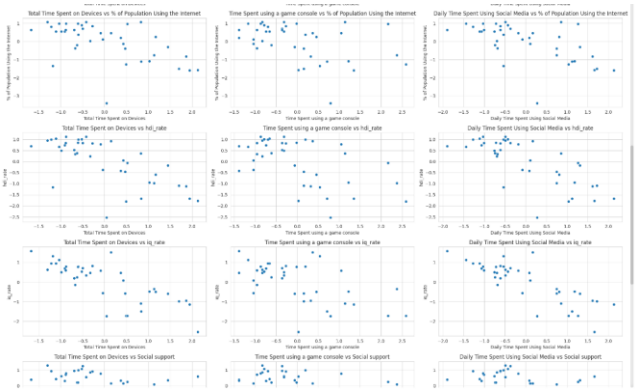

Figure 3: Some of the scatterplots used for visual inspection

After the visual inspection of the scatter plots, we had to confirm our conclusions and we did it by doing a correlation analysis. We conducted the correlation analysis for the same relationships we used in scatterplots, which resulted in confirming our suspicions that the data indeed had a linear element to it. All of the relationships had a negative correlation except for one attribute called obesity rate which consistently had a small positive correlation with every dependent attribute.

```
Correlation Coefficients for Total Time Spent on Devices:
health_index                        -0.441553
life_Exp                            -0.712793
gdp_usd                             -0.635191
democracies_rate                    -0.394651
% of Population Using the Internet  -0.550521
Happiness score                     -0.506092
hdi_rate                            -0.692310
iq_rate                             -0.777709
Social support                      -0.394244
obese_rate                           0.126739
suicide_rate                        -0.330029
```

Figure 4: Correlation analysis between all of the independent attributes and one of the dependent attributes

We also made a correlation analysis on the relationships between independent variables, to try and avoid any multicollinearity in any future implementations of our data.

```
Correlation Coefficients for life_Exp:
health_index                          0.620819
life_Exp                              1.000000
gdp_usd                               0.711351
democracies_rate                      0.457634
% of Population Using the Internet    0.750932
Happiness score                       0.698858
hdi_rate                              0.869230
iq_rate                               0.807827
Social support                        0.559995
obese_rate                            0.000025
suicide_rate                          0.096145
```

Figure 5: Correlation analysis between all of the independent attributes and one of the independent attributes

After the inspection of our correlation results we hand-picked the best attribute correlations that would help us build linear regression models and Support Vector Machines, keeping in mind multicollinearity, thus sometimes choosing an attribute with a smaller collinearity score, but with a small multicollinearity score, which would help to avoid any multicollinearity issues in our implementations. For the dependent variable, we chose the total time spent on devices, because we believe that it aligns best with our idea of screen time. For our independent variables, we chose three attributes that have low collinearity in between each other and a moderate to high negative correlation to the total time spent on devices. Those three variables are GDP, suicide rate, and IQ rate.

LINEAR REGRESSION MODEL

For predicting total time spent on devices we have decided to use a linear regression model. We split up the data into a training set and a test set, with a ratio of 0.3, meaning that 30% of the whole data is going to be used for testing, and 70% is going to be used for training the model. We decided to build two models, with the first containing three independent variables as mentioned above, and the second one containing two variables, where we did not use an attribute of suicide rate. The reason we decided to do two different models is because we wanted to test which one would perform better. We also implemented a cross-validation method for both models, allowing us a better evaluation. The evaluation of both models was split into two parts, one without cross-validation and one with cross-validation. Model coefficients, Mean Squared Error, and R-squared are evaluation methods that were used for the evaluation of a model without cross-validation. The same evaluation methods were used in cross-validation except for model coefficients. Model coefficients (MC) is an evaluation method that allows us to know how much impact on the overall model some variable has, and in which direction impact goes. Mean Squared Error (MSE) is a measure of the average squared difference between the actual and predicted values. It quantifies the overall accuracy of the model's predictions, allowing us to evaluate how precise the model is. R-squared (R^2) is a score of how well the independent variables explain the variability in the dependent variable, letting us know about the quality of our model.

ANALYSIS OF LINEAR REGRESSION MODELS

```
Model 1 Coefficients: [ 0.08704257 -0.84826976 -0.21769466]
Mean Squared Error: 0.847906531739743
R-squared: -0.26108961099752004
Model 2 Coefficients: [-0.22063538 -0.78075441]
Mean Squared Error: 0.7692301186324189
R-squared: -0.14407434635907101
```

Figure 6: Evaluation scores of both models

We will first analyze models without the use of cross-validation. For the first model with 3 predictor variables, we can see with the scores of MC that the IQ rate has the biggest impact on the model. MC scores for the second model with two predictor variables also indicate that the biggest impact is held by the IQ rate, it's worth noting that both MC scores are negative, meaning as the value for the Iq rate decreases time spent on devices increases. The MSE score for the first model is bigger than the MSE score for the second model, while the R^2 score is smaller for the first model compared to the second model. Both of these comparisons suggest that the second model has a better performance. As the differences in values are marginal, so is the improvement in the performance. To confirm this conclusion, we will also compare models after cross-validation, which should give us better insight into the performance of our two models.

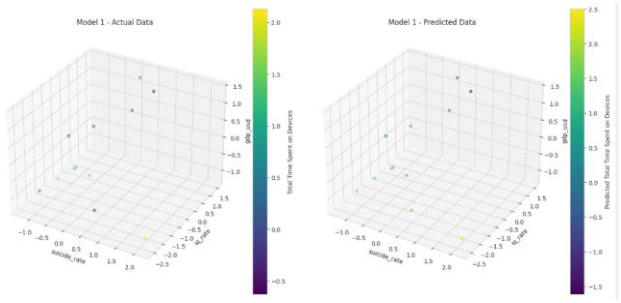

Figure 7: Two 3D scatterplots of the first model with three predictor variables, left one being actual data and the right one being predicted data

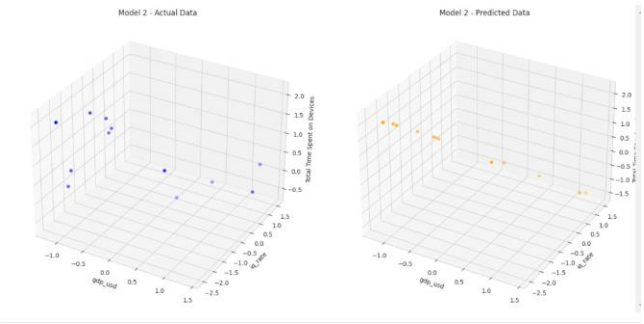

Figure 8: Two 3D scatterplots of the second model with two predictor variables, left one being actual data and the right one being predicted data

```
Model 1 Cross-Validation Results:
Mean Squared Error: [0.37364462 0.16142587 0.62707807 0.17731828 0.14817913 0.12111845
 0.09942519 0.60612922 0.83752252 0.22586489]
Average MSE: 0.33777062341122416
R-squared: [ 0.38808434 -0.48599706  0.53249262  0.60889371  0.86126087  0.70961702
  0.75333495  0.6844946  -0.70663849  0.35213306]
Average R-squared: 0.3697675616512036

Model 2 Cross-Validation Results:
Mean Squared Error: [0.39462171 0.1611078  0.54113113 0.17310903 0.14578514 0.09803728
 0.10944985 0.61943728 0.71783576 0.2107286 ]
Average MSE: 0.31712435663849525
R-squared: [ 0.35373028 -0.4830691   0.5965689   0.61817795  0.86350235  0.7649544
  0.72846464  0.67756742 -0.46275008  0.39554974]
Average R-squared: 0.4052696511594448
```

Figure 9: Evaluation scores of both models after cross-validation

After the use of cross-validation on both of our models, which is essentially just a simulation of our models multiple times, in our case 10 times, we should be able to make concrete conclusions about our models. After the inspection of evaluation scores, we can see that the first model has a slightly bigger average MSE, while also having a slightly smaller average R^2 score than the second model. This indeed does confirm our conclusion from earlier that the second model has a slightly better performance than the first model. But, as we can see from the 3D scatter plots above, and the mean evaluation scores, both models are far away from being close to optimal, and there is a lot of room for improvement.

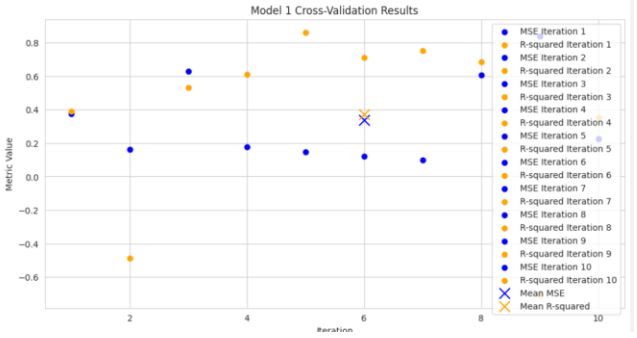

Figure 10: Evaluation scores plot of cross-validated first model

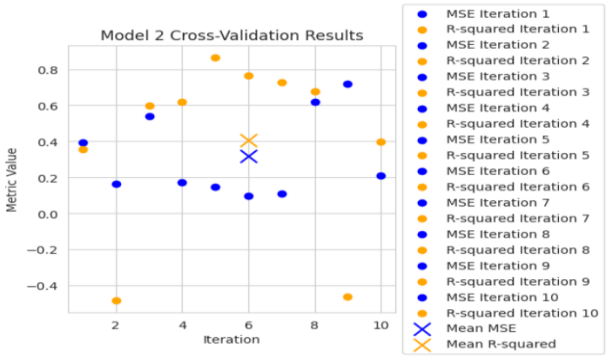

Figure 11: Evaluation scores plot of cross-validated second model

SUPPORT VECTOR MACHINES WITH RADIAL BASIS FUNCTION KERNEL

Support Vector Machines (SVMs) is a supervised machine learning algorithm that can be used either for classification or regression tasks. In our case, we are going to use it for a regression task, and later on, we are going to compare it to our linear regression model, to see which one performs better on a regression task. We are going to use SVMs with a Radial Basis Function (RBF) kernel which is commonly used with non-linear data because of its effectiveness. We are going to use the same standardized data set as we did in the linear model, as well as the same dependent and independent variables. That means that we are also going to make two SVM models one with three independent variables and one with two independent variables. Evaluation metrics are also going to stay the same, all of these steps will make it easier for comparison between the two methods.

ANALYSIS OF SUPPORT VECTOR MACHINES WITH RADIAL BASIS FUNCTION KERNEL

```
SVM Model 1 Test Set Results:
Mean Squared Error: 0.4258306464379293
R-squared: 0.5620796310126247

SVM Model 2 Test Set Results:
Mean Squared Error: 0.9323369566541295
R-squared: -0.38666280536031605
```

Figure 12: Evaluation scores of both SVM models

We will first analyze models without the use of cross-validation. For the first model, the MSE score is smaller than the MSE score

for the second model, while the R^2 score is bigger for the first model compared to the second model. Both of these comparisons suggest that the first model has a better performance. The differences in values are significant, so the performance of the first model is also significantly better. To confirm this conclusion, we will again compare the models after cross-validation, which should give us better insight into the performance of our two models.

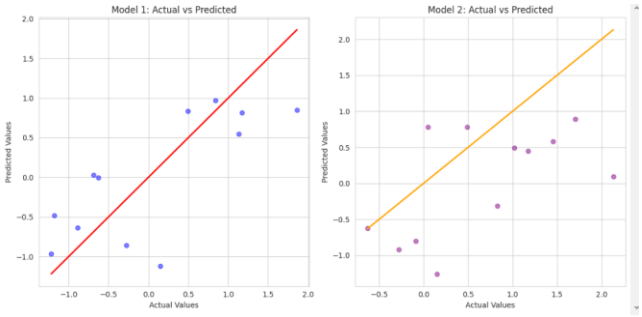

Figure 13: 2 scatterplots of both models and their performance

```
SVM Model 1 Cross-Validation Results:
Mean Squared Error: [0.20853499 0.78883166 0.71212119 0.38143028 0.08036519 0.68672805
 1.97686183 0.71495104 0.72857348 0.66741505]
Mean R-squared: [  0.72508733   0.47129863  -2.04181703  -0.41265594   0.04497463
   0.31922969  -0.91546457 -14.44331394  -2.84446547  -1.77552379]
Average Mean Squared Error: 0.694581274109732
Average R-squared: -2.0872650457154367

SVM Model 2 Cross-Validation Results:
Mean Squared Error: [0.34925069 0.20264081 0.54129535 0.28472724 0.2111819  0.17802672
 0.3033359  0.75745837 0.08790206 0.36903217]
Mean R-squared: [ 0.42803414 -0.86539903  0.59644647  0.37198458  0.8022718   0.57317873
  0.24745061  0.605724    0.82088    -0.0585255 ]
Average Mean Squared Error: 0.3284851205495789
Average R-squared: 0.35220457866715393
```

Figure 14: Evaluation scores of both SVM models after cross-validation

After the use of cross-validation on both of our SVM models, we can see that the first model has a bigger average MSE, while also having a lot smaller average R^2 score than the second model. This is an opposite conclusion from what we got without using cross-validation, and it means that the second model performs a lot better than the first model. The first model performs badly, while the second model is on the right track, and has relatively good MSE and R^2 values for an SVM model.

RESULTS

After we have analyzed both linear regression model and SVM model we can now compare the results after cross-validation and decide which method is better for our data. As we have 4 models, two for each method, we are going to compare the best model from each method. The best model for linear regression was the second model and the best model for SVM with Radial Basis Function kernel was also the second model. As we have used the same evaluation metrics we can now compare the two models.

```
Model 2 Cross-Validation Results:
Mean Squared Error: [0.39462171 0.1611078  0.54113113 0.17310903 0.14578514 0.09803728
 0.10944985 0.61943728 0.71783576 0.2107286 ]
Average MSE: 0.31712435663849525
R-squared: [ 0.35373028 -0.4830691   0.5965689   0.61817795  0.86350235  0.7649544
  0.72846464  0.67756742 -0.46275008  0.39554974]
Average R-squared: 0.4052696511594448
```

Figure 15: Evaluation scores of cross-validated second linear regression model

```
SVM Model 2 Cross-Validation Results:
Mean Squared Error: [0.34925069 0.20264081 0.54129535 0.28472724 0.2111819  0.17802672
 0.3033359  0.75745837 0.08790206 0.36903217]
Mean R-squared: [ 0.42803414 -0.86539903  0.59644647  0.37198458  0.8022718   0.57317873
  0.24745061  0.605724    0.82088    -0.0585255 ]
Average Mean Squared Error: 0.3284851205495789
Average R-squared: 0.35220457866715393
```

Figure 16: Evaluation scores of cross-validated second SVM model

Based on the evaluation scores provided above we can see that the average MSE values are slightly smaller for the linear regression model than for the SVM model, and we can also notice that the values for R2 are slightly higher for the linear regression model than for the SVM model. These results mean that the linear regression model is slightly better than the SVM model.

CONCLUSION

We conclude that we should use a linear regression model with two independent variables which are GDP and IQ rate to predict our dependent variable total time spent on devices. The model still has a lot of room for improvement.

LIMITATIONS

The biggest limitation of our study is the small sample size. For regression problems it is better to have a big data set with lots of samples, and our data set with 39 samples is certainly not big enough. For a future improvement a larger sample of countries would be better. Another limitation to our research is the lack of a time spectrum, meaning that the data consists of only recent indicators, but an interesting future improvement would be to build the models based on data from various years.

REFERENCES

[1] 2020. Negative Effects of Too Much Screen Time. *Valleywise health*.. Retrieved December 22, 2023 from https://valleywisehealth.org/blog/negative-effect-of-screen-time-adults-children/

[2] 2023. Why do countries have no data?. Retrieved December 22, 2023 from https://ts2.space/en/why-do-countries-have-no-data/

[3] Jason M. Nagata, Kyle T. Ganson, Puja Iyer, Jonathan Chu, Fiona C. Baker, Kelley Pettee Gabriel, Andrea K. Garber, Stuart B. Murray, and Kirsten Bibbins-Domingo.. 2021. Sociodemographic Correlates of Contemporary Screen Time Use among 9-10-Year-Old Children. 240, (August 2021), 213–220.e2.. DOI:https://doi.org/10.1016/j.jpeds.2021.08.077

[4] Hossein Sourtiji, Seyed Ali Hosseini, Mehdi Rassafiani, Amir Kohan, Mehdi Noroozi, and Mohammad Esmaeil Motlagh.. 2019. The Associations Between Screen Time, Sleep Duration, and Body Mass Index (BMI) in Under Five-Year-Old Children. 6, 1 (December 2019).. DOI:https://doi.org/10.5812/ans.81229

[5]]2022. Countries of the world 2022. *Kaggle*.. Retrieved December 22, 2023 from https://www.kaggle.com/datasets/moekhaledx/countries-of-the-world

[6] 2023. World Happiness Report 2023 Dataset. *Kaggle*.. Retrieved December 22, 2023 from https://www.kaggle.com/datasets/atom1991/world-happiness-report-2023

[7] 2023. Average Screen Time and Usage by Country 2023. *Kaggle*.. Retrieved December 22, 2023 from https://www.kaggle.com/datasets/prasertk/average-screen-time-and-usage-by-country

[8] Sandro Krvgić, Giuliano Rojas. 2023. Merged Dataset. MergedDataset (1).xlsx

[9] sklearn.linear_model.LinearRegression. Retrieved December 22, 2023 from https://scikit-learn/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[10] 2023. 1.4. Support Vector Machines. Retrieved December 22, 2023 from https://scikit-learn/stable/modules/svm.html

[11] sklearn.gaussian_process.kernels.RBF. Retrieved December 22, 2023 from https://scikit-learn/stable/modules/generated/sklearn.gaussian_process.kernels.RBF.html

APPENDIX

We separated the sections of our code into the different steps that lead us to our results. The first step was setting up our own data, which we needed to import various python modules, and use the 3 source datasets to make our own merged dataset. Then we formatted the dataset until it contained the necessary values to conduct our research. Following the structured dataset, we proceeded to standardize the data to prepare for the models. Then the data was analyzed using correlational methods. Lastly, it was assessed in order to determine what variables would best suit the models. The next step was to apply the data on the models. In our case, we opted for a linear regression model and a support vector machine model that uses a Radial Basis Function kernel. From this we gather the results from each of the models and arrived at our conclusion.