DS 4002: Human Capital Index (HCI) Project

Predicting GDP Per Capita Using 2020 HCI Data And Visualizations

Group 7 - Matthew Heeter, John Le, and Andrew Nguyen

Question and Background Information

The World Bank's Human Capital Project is an initiative that aims to increase the productive capacity of people in developing countries through investments in health, education, and training. The project aims to improve human capital outcomes, such as health outcomes, educational achievements, and labor market outcomes, in order to accelerate economic growth and reduce poverty. It also aims to promote gender equality and inclusivity.

The project uses data and evidence to inform policy decisions and interventions at the national and subnational levels, and works with governments, civil society organizations, and other stakeholders to design and implement effective policies and programs. The project also focuses on the measurement of human capital, including the development of the Human Capital Index (HCI), which ranks countries (from a scale of 0 to 1, with 1 being the highest rank) based on their investments in health, education, and training.

The HCI is used to inform policy decisions and track progress over time. Overall, the goal of the Human Capital Project is to help countries achieve their development goals and create a more prosperous future for all their citizens.

Research Goals

1. Predicting health outcomes based on investments in health, education, and training: This project could involve using machine learning algorithms to predict health outcomes (such as infant mortality rates or life expectancy) based on data on investments in health, education, and training. The goal of this project would be to identify the most effective interventions for improving health outcomes in developing countries.
2. Predicting educational achievements based on investments in education and training: This project could involve using machine learning algorithms to predict educational achievements (such as literacy rates or enrollment rates) based on data on investments in education and training. The goal of this project would be to identify the most effective interventions for improving educational outcomes in developing countries.
3. Predicting labor market outcomes based on investments in education and training: This project could involve using machine learning algorithms to predict labor market outcomes (such as employment rates or wage levels) based on data on investments in education and training. The goal of this project would be to identify the most effective interventions for improving labor market outcomes in developing countries.

Related Research - HCI vs. GDP per capita seem to have a strong, positive, and linear relationship as seen in the 2016 Our World in Data website. Source: https://ourworldindata.org/grapher/human-capital-index-vs-gdp

Questions

1. What's the most optimal machine learning model for accurately predicting a country's GDP per capita using HCI (Human Capital Index) socio-economic and demographic data?
2. How do we effectively portray the relationships between HCI and GDP per capita?

Exploratory Data Analysis

Dataset: World Bank Human Capital Index Dataset (2020)

Data Cleaning

▶ Show the code

	Country Name	Region	Income Group	Probability of Survival to Age 5	Expected Years of School	Harmonized Test Scores	Learning-Adjusted Years of School	Adult Survival Rate	HUMAN CAPITAL INDEX 2020
0	Afghanistan	South Asia	Low income	0.937724	8.901891	354.758789	5.052838	0.787741	0.400284
1	Albania	Europe & Central Asia	Upper middle income	0.991177	12.889381	434.127594	8.953018	0.929366	0.634251
2	Algeria	Middle East & North Africa	Lower middle income	0.976518	11.848035	374.089081	7.091553	0.909282	0.534556

	Country Name	Region	Income Group	Probability of Survival to Age 5	Expected Years of School	Harmonized Test Scores	Learning-Adjusted Years of School	Adult Survival Rate	HUMAN CAPITAL INDEX 2020
3	Angola	Sub-Saharan Africa	Lower middle income	0.922835	8.120066	325.965485	4.234978	0.729359	0.362405
4	Antigua and Barbuda	Latin America & Caribbean	High income	0.993559	12.967560	406.997437	8.444422	0.897208	0.595704

- Main dataset: 2020 Human Capital Index
- Derived from the World Bank database
- Data from September 2020 for 174 countries with each row representing a specific country
- Includes columns on WB code, region, and income group for each country
- Also contains columns on specific Human Capital Index (HCI) information, such as Probability of Survival to Age 5, Expected Years of School, Harmonized Test Scores, Learning- Adjusted Years of School, Fraction of Children Under 5 Not Stunted, Adult Survival Rate, calculated HCI score with its lower bound and upper bound for each country
- Removed lower bound and upper bound of HCI 2020, Fraction of Children Under 5 Not Stunted column (because there were too many missing/null values), and WB code (unnecessary categorical variable)

Dataset: World Bank GDP Per Capita ($) Dataset (1965-2021)

▶ Show the code

	Country Name	2020
0	Aruba	24487.863560
1	Africa Eastern and Southern	1353.769160
2	Afghanistan	516.866552
3	Africa Western and Central	1683.436391
4	Angola	1603.993477

- Also derived from the World Bank database
- Data from 1965 - 2021 on countries' GDP per capita
- Each row represents a country and the columns are the year
- Data from 266 markets - including counties, regions and territories

Merge two datasets

▶ Show the code

	Country Name	Region	Income Group	Probability of Survival to Age 5	Expected Years of School	Harmonized Test Scores	Learning-Adjusted Years of School	Adult Survival Rate	HUMAN CAPITAL INDEX 2020	2020
0	Afghanistan	South Asia	Low income	0.937724	8.901891	354.758789	5.052838	0.787741	0.400284	516.866552
1	Albania	Europe & Central Asia	Upper middle income	0.991177	12.889381	434.127594	8.953018	0.929366	0.634251	5332.160475
2	Algeria	Middle East & North Africa	Lower middle income	0.976518	11.848035	374.089081	7.091553	0.909282	0.534556	3337.252512
3	Angola	Sub-Saharan Africa	Lower middle income	0.922835	8.120066	325.965485	4.234978	0.729359	0.362405	1603.993477
4	Antigua and Barbuda	Latin America & Caribbean	High income	0.993559	12.967560	406.997437	8.444422	0.897208	0.595704	14787.635780

Summary Stats

▶ Show the code

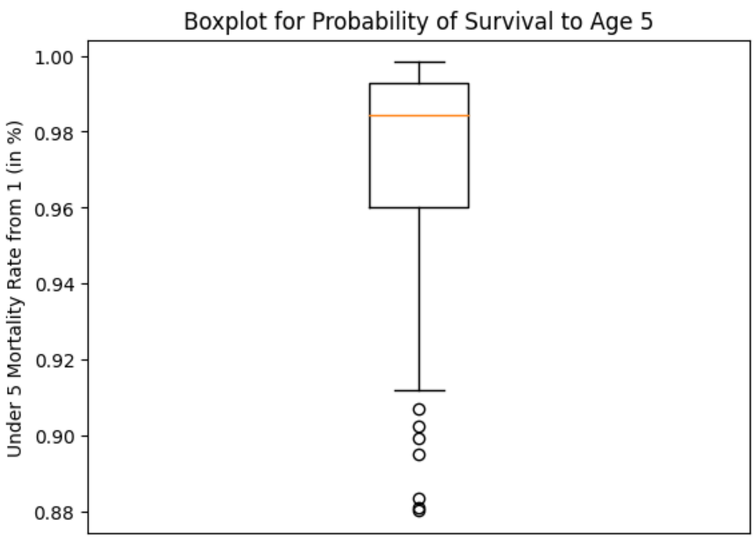
```
/var/folders/nj/705wgkn91bx4dynjyw516f840000gn/T/ipykernel_66173/750947143.py:6: FutureWarning:

this method is deprecated in favour of `Styler.format(precision=..)`
```

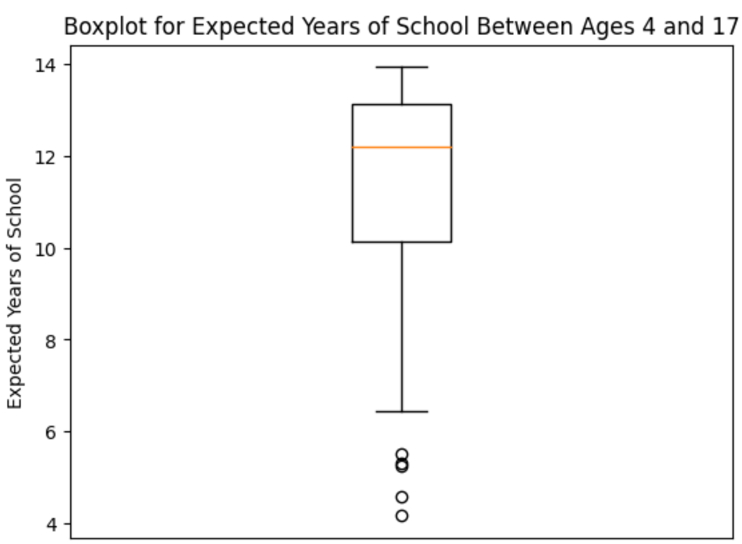
	Probability of Survival to Age 5	Expected Years of School	Harmonized Test Scores	Learning-Adjusted Years of School	Adult Survival Rate	HUMAN CAPITAL INDEX 2020	2020

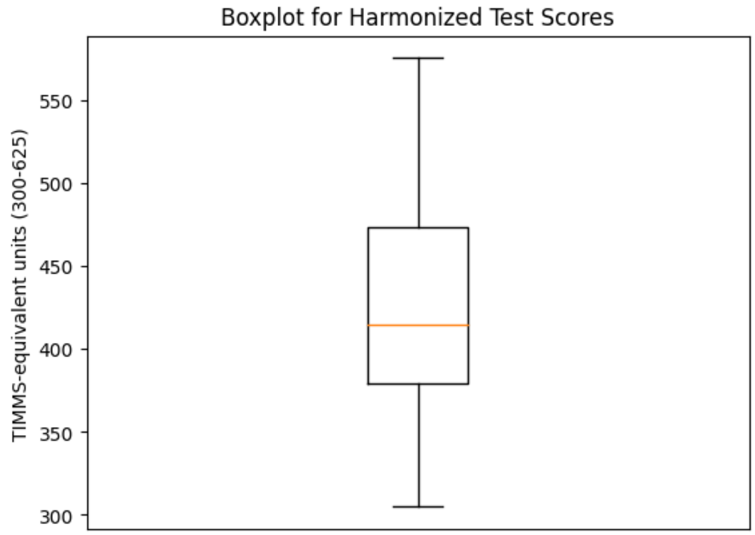
	Probability of Survival to Age 5	Expected Years of School	Harmonized Test Scores	Learning-Adjusted Years of School	Adult Survival Rate	HUMAN CAPITAL INDEX 2020	2020
count	168.000000	168.000000	168.000000	168.000000	168.000000	168.000000	168.000000
mean	0.973054	11.379255	423.947414	7.869514	0.850771	0.564158	14157.843802
std	0.027431	2.276297	62.781017	2.429186	0.083304	0.137317	19679.216603
min	0.880086	4.156989	304.922241	2.206502	0.522544	0.291632	216.826741
25%	0.960007	10.144179	379.442459	6.085780	0.802397	0.453250	1899.932748
50%	0.984219	12.201100	414.321365	7.941399	0.861430	0.566201	5342.754270
75%	0.992954	13.121243	473.094193	9.763350	0.916337	0.666993	16945.454083
max	0.998305	13.936425	575.272156	12.813290	0.961434	0.879126	117370.496900

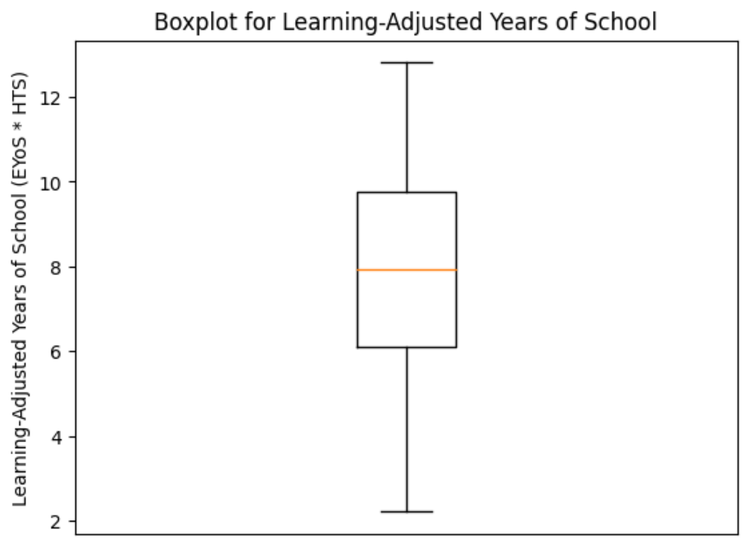
Boxplots

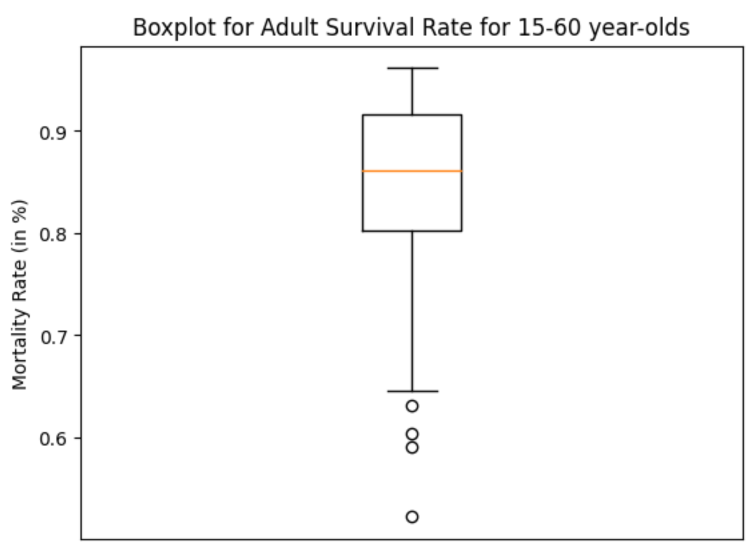
▶ Show the code

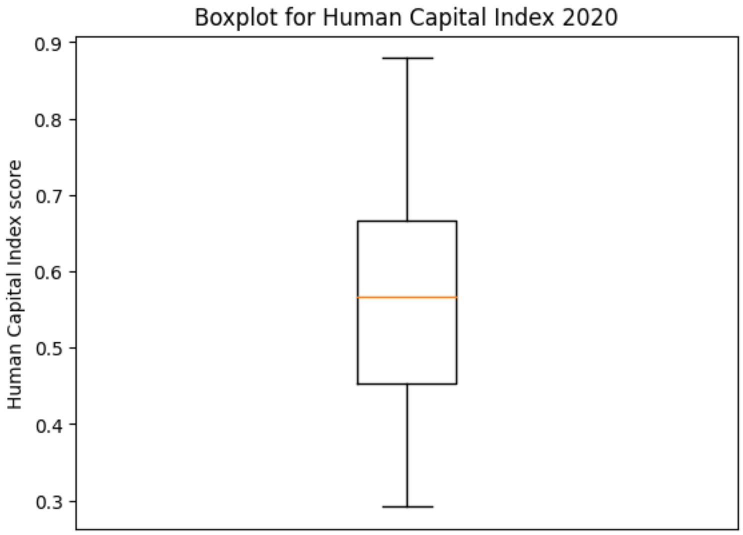





Boxplot for Harmonized Test Scores
TIMMS-equivalent units (300-625)
300
350
400
450
500
550


Boxplot for Learning-Adjusted Years of School
Learning-Adjusted Years of School (EYoS * HTS)
2
4
6
8
10
12


Boxplot for Adult Survival Rate for 15-60 year-olds
Mortality Rate (in %)
0.6
0.7
0.8
0.9



Data Processing

▶ Show the code

	Region	Income Group	Probability of Survival to Age 5	Expected Years of School	Harmonized Test Scores	Learning-Adjusted Years of School	Adult Survival Rate	HUMAN CAPITAL INDEX 2020	2020	Region_encoded	Income Group_encoded
0	South Asia	Low income	0.937724	8.901891	354.758789	5.052838	0.787741	0.400284	516.866552	5	1
1	Europe & Central Asia	Upper middle income	0.991177	12.889381	434.127594	8.953018	0.929366	0.634251	5332.160475	1	3
2	Middle East & North Africa	Lower middle income	0.976518	11.848035	374.089081	7.091553	0.909282	0.534556	3337.252512	3	2
3	Sub-Saharan Africa	Lower middle income	0.922835	8.120066	325.965485	4.234978	0.729359	0.362405	1603.993477	6	2
4	Latin America & Caribbean	High income	0.993559	12.967560	406.997437	8.444422	0.897208	0.595704	14787.635780	2	0

One Hot Encoding

▶ Show the code

▶ Show the code

	Country Name	Probability of Survival to Age 5	Expected Years of School	Harmonized Test Scores	Learning-Adjusted Years of School	Adult Survival Rate	HUMAN CAPITAL INDEX 2020	2020	Region_encoded	Income Group_encoded	...	Region_E & Central
index												
0	Afghanistan	0.937724	8.901891	354.758789	5.052838	0.787741	0.400284	516.866552	5.0	1.0	...	0.0
1	Albania	0.991177	12.889381	434.127594	8.953018	0.929366	0.634251	5332.160475	1.0	3.0	...	1.0
2	Algeria	0.976518	11.848035	374.089081	7.091553	0.909282	0.534556	3337.252512	3.0	2.0	...	0.0
3	Angola	0.922835	8.120066	325.965485	4.234978	0.729359	0.362405	1603.993477	6.0	2.0	...	0.0
4	Antigua and Barbuda	0.993559	12.967560	406.997437	8.444422	0.897208	0.595704	14787.635780	2.0	0.0	...	0.0

5 rows × 21 columns

Methods

Removing Outliers Using Z-Scores

▶ Show the code

	Probability of Survival to Age 5	Expected Years of School	Harmonized Test Scores	Learning-Adjusted Years of School	Adult Survival Rate	HUMAN CAPITAL INDEX 2020	2020	Region_encoded	Income Group_encoded	Region_East Asia & Pacific	Region_Eur & Central A
index											
0	0.937724	8.901891	354.758789	5.052838	0.787741	0.400284	516.866552	5.0	1.0	0.0	0.0
1	0.991177	12.889381	434.127594	8.953018	0.929366	0.634251	5332.160475	1.0	3.0	0.0	1.0
2	0.976518	11.848035	374.089081	7.091553	0.909282	0.534556	3337.252512	3.0	2.0	0.0	0.0
3	0.922835	8.120066	325.965485	4.234978	0.729359	0.362405	1603.993477	6.0	2.0	0.0	0.0
4	0.993559	12.967560	406.997437	8.444422	0.897208	0.595704	14787.635780	2.0	0.0	0.0	0.0

Modeling

- After removing the outliers, we wanted to split the data into a training and testing data set with a 80/20 split, respectively because we felt an 80/20 split was appropriate since we did not have that much data, so we wanted to make sure our model was well-trained (hence the high 80% on training set).
- We wanted to run three models: Linear Regression (base model), Support Vector Machines (SVM), and K-nearest neighbors (KNN).
- Linear Regression is a base model. We started with the assumption that there is a linear relationship between HCI and GDP.
- Support Vector Reggression was used to try to find a more accurate model. We chose SVR because the relationship between HCI and GDP may be non-linear and SVR will account for that, unlike linear regression.
- KNN model was chosen because how it can effectively handle a small number of features and be less sensitive to hyperparameters.

▶ Show the code

```
▾ SVR
SVR()
```

▶ Show the code

```
▾ KNeighborsRegressor
KNeighborsRegressor(n_neighbors=8)
```

R-Squared Comparisons

We wanted to use R-squared, which is the proportion of variation in dependent variable explained and predicted by the independent variable, as the primary performance metric to see which models were most effective in predicting GDP per capita using HCI data. We found Linear Regression to have the highest R-Squared out of the three models of 0.8406, so we used the Linear Regression model to represent our predicted GDP per capita score through an interactive map visualization. The KNN models R-Squared score was 0.4937 and the SVR R-Squared score was -0.1696. It is important that the SVR R-Square score was negative here, indicating that this model was not a good fit for our dataset.

▶ Show the code

```
Linear Regression MSE: 34615379.5099 R2: 0.8406
SVR MSE: 254030694.3569 R2: -0.1696
KNN MSE: 109963918.1055 R2: 0.4937
```

Linear Regression Predictions

Here, we are using our best model, Linear Regression, find the R-Squared score and the predicted GDP per capita for the whole dataset. Before, we only found the values for the 20% test dataset.

▶ Show the code

```
0.7802252904685397
```

▶ Show the code

2020	Predicted

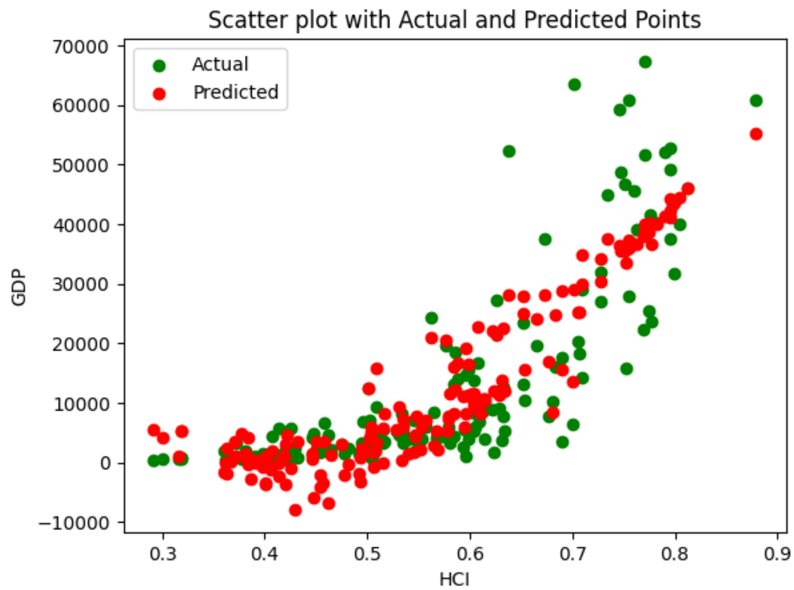
	2020	Predicted
0	516.866552	1048.507017
1	5332.160475	12036.039461
2	3337.252512	7421.897684
3	1603.993477	-1732.400091
4	14787.635780	19192.659353

Evaluation of the model

Linear Regression Scatterplot

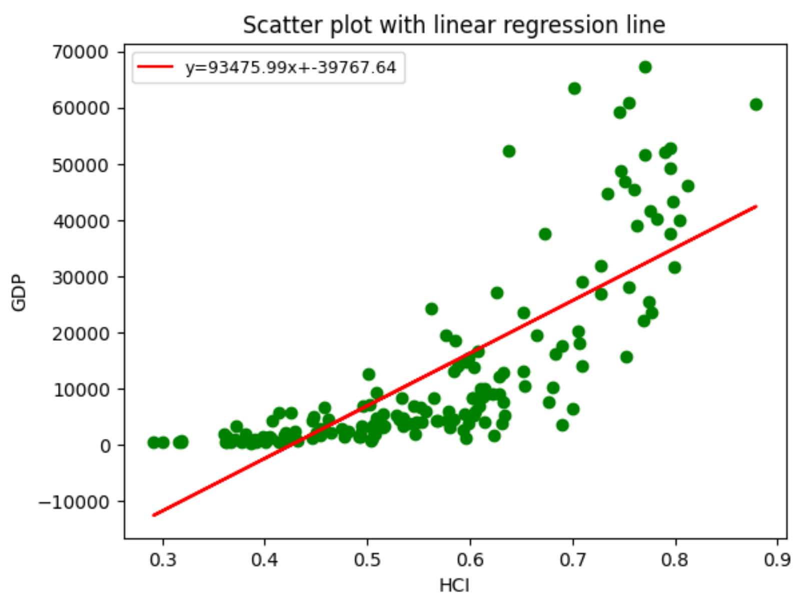
We wanted to efffectively represent the relationship between HCI and GDP per capita, so we decided to use a scatterplot because scatterplots are known to show whether or not a relationship exists between a multitude of data points.

▶ Show the code



This is a scatterplot of the HCI and GDP per capita actual and predicted values, where the predicted values are calculated from our Linear Regression model. There seems to be a strong, positive, and linear relationship between HCI and GDP per capita. Each data point represents a specific country in our dataset. As you can see, our predicted values from our Linear Regression model was not that far off from the actual values.

▶ Show the code



This is a scatterplot of the actual HCI and GDP per capita values from our dataset that shows a strong, positive, and linear relationship between HCI and GDP per capita with a calculated regression like of 93475.99x + -39767.64.

World Map Visualizations

▶ Show the code

```
Unable to display output for mime type(s): application/vnd.plotly.v1+json
```

This interactive world map shows the HCI score (from 0 to 1) across all the countries in our dataset.

▶ Show the code

```
Unable to display output for mime type(s): application/vnd.plotly.v1+json
```

This interactive world map shows the GDP per Capita (in US $) across all the countries in our dataset.

▶ Show the code

```
Unable to display output for mime type(s): application/vnd.plotly.v1+json
```

```
<function __main__.filter_data(min_gdp)>
```

▶ Show the code

```
Unable to display output for mime type(s): application/vnd.plotly.v1+json
```

This interactive world map shows the percentage error between our predicted GDP per capita values (using our Linear Regression model) and the actual GDP per capital values across all the countries in our dataset.

Conclusions

- In answering our first question of whether or not we could create a good machine learning model to predict GDP per capita using HCI information, we were able to create a decent machine learning model, using Linear Regression. We tested three different models: Linear Regression, SVM, and KNN, and their R-squared values were 0.8406, -0.1696, and 0.4937, respectively. Therefore, the Linear Regression model had the highest R-squared value out of the three models, so we decided to use that model to visually represent the predicted GDP per capita values.
- In answering our second question of whether or not we could effectively portray the relationship between HCI and GDP per capita, we were able to come up with a variety of visualizations, including scatterplots, interactive world maps, and boxplots. We noted that countries with a higher HCI score tended to also have a higher GDP per capita, hence the linear relationship between those two variables.
- We also learned how to develop the Flask application in Python, which allows us to upload our visualizations on an external web framework.
- Limitations: One limitation of our project was definitely our dataset. We only had 162 countries (or datapoints) in our dataset after removing the missing and null values, which was not a good sample size for this project. With a sample size of 162 countries, we had limited predictive power. Unfortunately, we were only able to find 2020 HCI values online because we are assuming that it takes many years to gather HCI information. Also, HCI is a fairly new concept, given that it came out in 2018, so that’s another reason why there is not a lot of data out there. However, we were really interested in this topic because we believe there were a lot of real-world applications, such as predicting health outcomes, educational achievements, and labor market outcomes. In our case, we focused on the labor market outcomes, and we noticed that countries with a higher GDP per capita have a higher HCI score, so countries should strive to increase their GDP per capita to ultimately increase their country’s overall productivity in the next generation of workers. Countries can increase their GDP per capita by focusing on four key areas: personal consumption expenditures, business investment, government spending, and net exports of goods and services.

Future Work

- Conduct the same project in the future when more HCI data becomes available
- Use more performance metrics (besides R-squared)
- Test more stacked classifiers and deep learning models
- Find a different feature than HCI to predict GDP per capita
- Create more filters within the interactive world map visualizations

Sources

https://www.worldbank.org/en/publication/human-capital https://databank.worldbank.org/indicator/NY.GDP.PCAP.CD/1ff4a498/Popular-Indicators#
https://quarto.org/docs/computations/python.html#vs-code https://ourworldindata.org/grapher/human-capital-index-vs-gdp
https://plotly.com/python/choropleth-maps/ https://towardsdatascience.com/k-nearest-neighbors-94395f445221
https://www.thebalancemoney.com/components-of-gdp-explanation-formula-and-chart-3306015 https://ipywidgets.readthedocs.io/
https://getbootstrap.com/docs/4.0/components/navbar/
https://docs.tibco.com/pub/spotfire/6.5.1/doc/html/3d_scat/3d_scat_what_is_a_3d_scatter_plot.htm#:~:text=3D%20scatter%20plots%20are%20used,%2C%20Y
https://quarto.org/docs/output-formats/html-code.html#:~:text=%23%23%20Hiding%20Code%20For%20many,the%20document%20%60execute%60%20options.

Thank you for listening! Any questions?