Data Mining

CSI 4142 - Fundamentals of Data Science

Winter 2022

School of Electrical Engineering and Computer Science

University of Ottawa

Course Coordinator: Dr. Herna L Viktor

Teaching Assistants: Anuhbav Chharbra, Nicolas Fleece, Paul Mvula

Group 25:

Lilian Ly, 8262186

Jonathan Brar, 8209351

Logan Rose, 300059034

Submission Date: April 8, 2022

Table of Contents

Github Repository

Data Mining Folder: https://github.com/Logan-Rose/Data-Science-Project/tree/main/data-mining

Tableau File: https://github.com/Logan-Rose/Data-Science-Project/tree/main/tableau

Repository Link: https://github.com/Logan-Rose/Data-Science-Project

Part A. Data summarization, data preprocessing and feature selections

Data Summarization

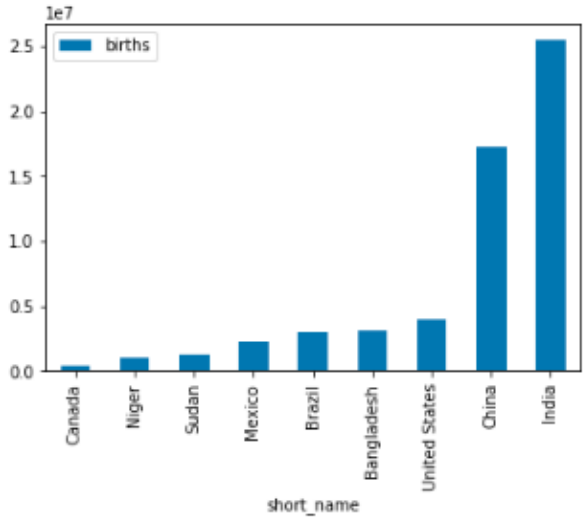

Figure 1. Number of births per country

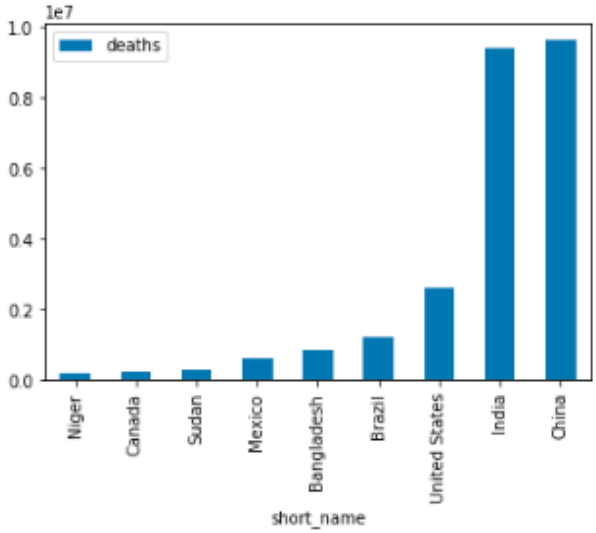

Figure 2. Number of deaths per country

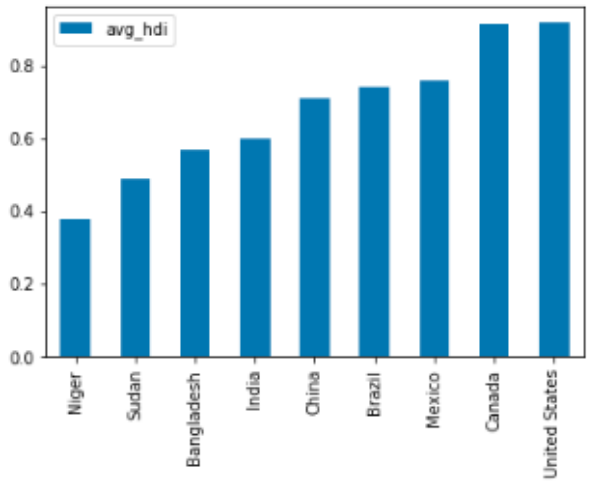

Figure 3. Average HDI per country

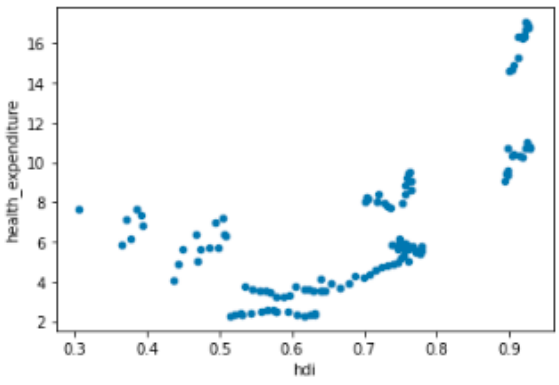

Figure 4. Health expenditure compared to HDI

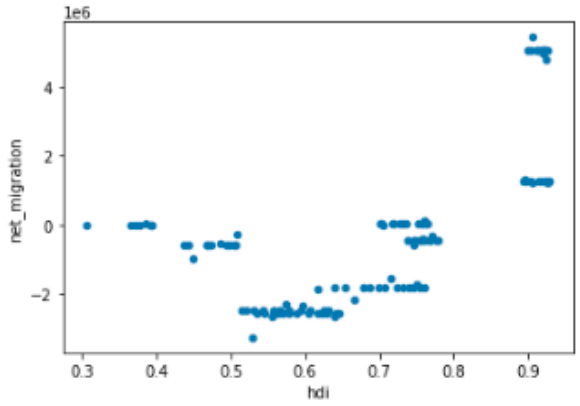

Figure 5. Net migration compared to HDI

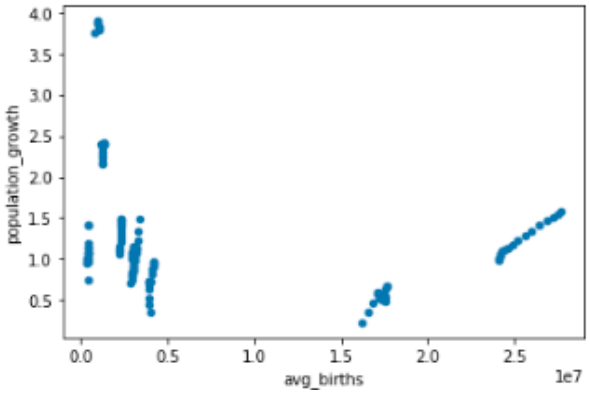

Figure 6. Population growth compared to average number of births

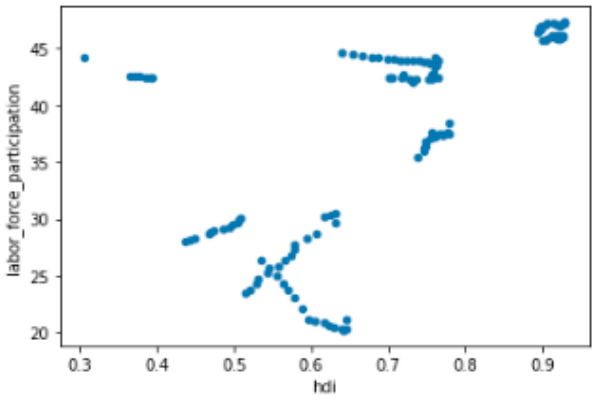

Figure 7. Labor force participation compared to HDI

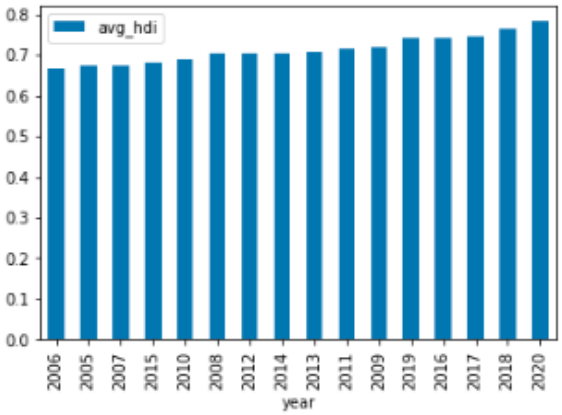

Figure 8. Average HDI between 2005 and 2020

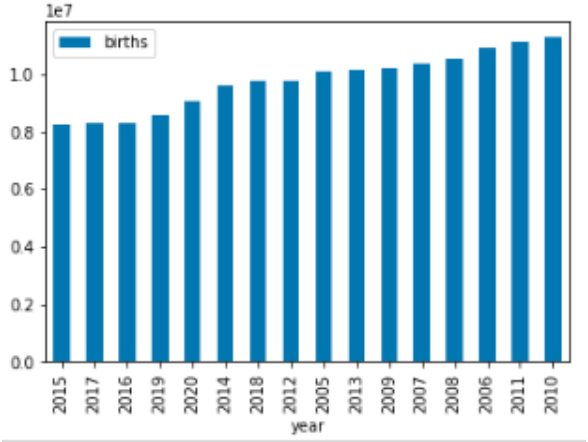

Figure 9. Number of births between 2005 and 2020

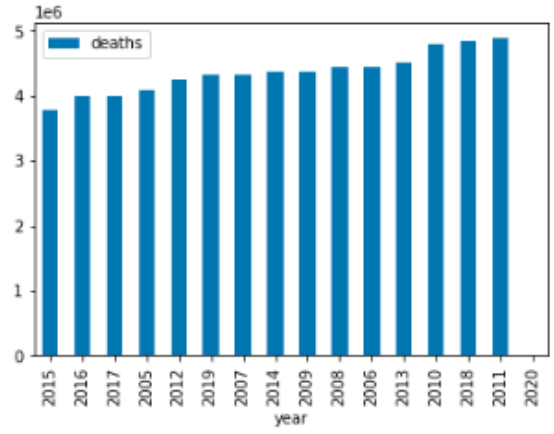

Figure 10. Number of deaths between 2005 and 2020

Data Transformation

Education Dimension, Health Dimension, and Quality of Life Dimension

Our data transformation process for the information from the world bank involved two strategies; one for fields that were missing a small amount of data, and another for fields that were missing most/all data. For fields that were missing a small amount of data, the values were imputed using scikit-learn KNNImputer. For fields that were missing most/all values, we manually filled in the information from external sources found online. In some cases where data could not be found, we took the mean values of previous years and used that for data from 2005 - 2020. The following list is the different attributes that were missing data for the specific country.

- Canada
  - Literacy Rate
    - Canada Literacy Rate 1990-2022 | MacroTrends
  - Poverty headcount
    - Canada - Poverty Headcount Ratio At $1.25 A Day (PPP) (% Of Population) - 2022 Data 2023 Forecast 1971-2017 Historical (tradingeconomics.com)
  - Total Primary Completion Rate
    - Canada - Primary Completion Rate, Total, Based On Completers - 2022 Data 2023 Forecast 2012-2016 Historical (tradingeconomics.com)
  - Female Primary Completion Rate
    - Canada - School Enrollment, Primary, Female (% Gross) - 2022 Data 2023 Forecast 1971-2019 Historical (tradingeconomics.com)
  - Male Primary Completion Rate
    - Canada - School Enrollment, Primary, Male (% Gross) - 2022 Data 2023 Forecast 1971-2019 Historical (tradingeconomics.com)
- USA
  - Poverty headcount
    - United States - Poverty headcount ratio (indexmundi.com)
- Brazil
  - Total Primary Completion Rate
    - Mean value of previous years used
  - Poverty headcount
    - Brazil - • Brazil - poverty headcount ratio 2005-2015 | Statista

Events Dimension (Disaster, Terrorism)

Two external sources were used to create the Events dimension; quality issues were handled differently for each source:

- **Disaster events** (*emdat_public_2022_02_31.csv)*:

The *emdat_public_2022_02_31.csv* contained missing values in the total affected and total deaths columns. We did some research and concluded that the empty spaces meant that there was either 0 total affected or 0 total deaths. So if there was a value greater than 0 in total affected and no value in the total deaths, then the total deaths value was set to 0. If there is a value greater than 0 in total deaths and no value in total affected, then the total affected value is equal to the total deaths value. If there was no value in either column, then both were set to 0.

The second issue with this data set is that some disasters were missing information in the date columns. If this data was missing in a row, that entire row of data was removed from the seed file. We had plenty of disaster events and felt it was fine to remove a few rows if they were missing date information.

- **Terrorist events** (*globalterrorismdb_0221dist_filtered.csv)*:

The large data source contained missing and null values under the attack type (*attacktype1_txt*), property damage (*propextent_txt*), number of wounded (*nwound)* and number of deaths (*nkill*). As only 10 events are required per country, and there were a total of 44 thousand terrorist events found for the 9 countries, such events that had any missing or null values were ignored and dropped from the data set. The data was further narrowed down by choosing only events that had an impact on the population number by taking events that had at least 1 death or wounded person.

Integrating the data to form one single event dimension from the two data sources stated above was done by first renaming all columns to match from both sources. The columns were then consolidated into being: Key, Seed_id, Country, Start_year, Start_month, Start_day, End-year, End_month, End_day, Event_type, Total_affected, Total_deaths. By doing so, both sources were merged through concatenation into one single *Event_seed.csv*.

Further data transformation was performed to remove redundant events. Since there were multiple events pointing to the same month, year and country, further preprocessing using a script was used to select only major events. In this case, major events were selected by taking the disaster or terrorist event with the largest affected number of people. That way, the fact table had less issues with the keys with no duplicate events.

Human Development Index (HDI) Measure

One of the measures is the Human Development Index (HDI) taken from the data source *HDI.csv*. Since the data time frame is taken between 2005-2020 (inclusively), there was missing data for all of the year 2020. The missing data was handled by taking the last most updated HDI for each country which was provided in 2019 to be the same for 2020.

Part B. Classification (Supervised Learning)

Table 1. Accuracy, precision, recall and time to construct results

Metric	**Gradient Boost**	**Random Forest**	**Decision Tree**
Accuracy	92%	100%	100%
Precision (weighted average)	95%	100%	100%
Recall (weighted average)	93%	100%	100%
Time to construct model	0.018s	0.029s	0.004s

Summary & Knowledge nuggets

- Ultimately It is difficult to form a strong conclusion about the accuracy of each model based on the small data sample. Perhaps if more countries, a more granular date specification, or a broader array of statistics were used, more data points could be generated, and the differences between the various different models of machine learning could be better compared and contrasted.
- We found that while all dimensions were reasonably accurate at predicting HDI, health was the least accurate but still overall very good, predicting with an 85 % accuracy
- Due to the unrealistically high accuracy of our model, we concluded that our model has likely been overfitted to the data, and if this project had a larger scope, we would be interested in investigating this further, and applying various techniques to help the model adapt to a variety of data sets.
- There is an interesting correlation between labor force participation rate and Human development Index. Countries with a low human development index have a high labor participation rate, countries in the lower-mid range have a much lower labor participation rate, and then the more developed countries once again have a high labor force participation rate. We posit the reason for this is likely that less developed countries have a high rate of employment among females out of necessity. In moderately developed countries, patriarchy and related causes push down the labor force participation rate. In the most developed countries, the labor force participation rate is higher due to social progress.

Team Work Plan

Deliverable checklist	Responsible team member(s)	Expected completion date	Actual completion date	Estimated time (hours) to complete	Actual time (hours) to complete
Data preprocessing					
Data summarisation	Jonathan/Lilan	April 1, 2022	April 1, 2022	2 hours	2 hours
Visualization of attributes	Logan	April 1, 2022	April 1, 2022	1 hour	1 hour
Data transformation	Lilian/Jonathan	April 1, 2022	April 1, 2022	4 hours	6 hours
Missing values	Jonathan/Lilian	April 1, 2022	April 1, 2022	4 hours	6 hours
Categorical data	Logan	April 1, 2022	April 2, 2022	1 hour	2 hours
Numeric data	Logan	April 1, 2022	April 2, 2022	1 hour	2 hours
Feature selection	Logan	April 1, 2022	April 2, 2022	1 hour	2 hours
Data mining - Classification					
Decision tree	Logan	April 3, 2022	April 5, 2022	2 hours	2 hours
Gradient Boosting	Logan	April 3, 2022	April 5, 2022	2 hours	2 hours
Random Forests	Logan	April 3, 2022	April 5, 2022	2 hours	2 hours
Comparison of results	Logan	April 3, 2022	April 5, 2022	2 hours	2 hours
Summary	Logan	April 3, 2022	April 5, 2022	1 hour	1 hour
Other tasks - please specify					
Report	Jonathan/Lilian	April 6th	April 8th	3 hours	2 hours
Submission	Lilian	April 6th	April 8th	15 mins	15 mins