**Habiba Aziz | Final Project**

**Introduction to Computational Social Science**

**Can HDI and GDP levels be used as criteria to check which countries are more susceptible to cancer incidences/deaths? Can text data obtained from web scraping provide us an overview of public sentiment on cancer?**

**Abstract**

Cancer has no frontiers and no prejudices; it can happen to anyone when a single cell in a person’s body becomes abnormal (becomes genetically mutated). Scientists are still trying to pinpoint the root causes of this disease as well as a viable treatment. This study was conducted to look at the time series analysis of cancer incidences and deaths around the world based on GDP and HDI levels and to look at the present-day public perception about it. Data from ‘our world in data’ revealed that cancer incidences are unbiased of country, but the death rates are higher in developing countries. Additionally, 20 countries from the dataset, 10 developed and 10 developing were picked to see the incidence rates. Decision tree classifier on cancer death rate by ages identified that the data was picked randomly. Text data scraped from google scholar articles demonstrated that from 2002 – 2017, only a handful of studies linking cancer with HDI and GDP were performed. Tweets from the present were used to understand the current public opinion on cancer.

**Introduction**

Human Development Index (HDI) level is a scale that consists of a compound degree of pointers along three magnitudes: life expectancy, educational accomplishment, and availability of resources desirable for an adequate way of living. Gross Domestic Product (GDP) is mainly used to measure the healthiness of a country’s economy, and it is described as the financial value of merchandises and amenities produced in a country’s vicinity during an explicit period of time. Cancer is a disease that causes most of that cases of morbidity and mortality among patients in the world and these numbers keep on increasing as the years pass by (Bray et al. 2012). Disparities in the weight of cancer have been well-documented, and prior studies were focused on variances within countries, with an uneven load of prevailing cancers in countries categorized as high HDI, whereas infection related cancers remain to outweigh in countries with lower levels of HDI (Arnold et al.2017). Socioeconomic status is associated with cancer patients’ survival and it is vital to expand awareness of risk influences and primary detection (Ghoncheh and Mirzaei and Salehiniya2012**).**

**Analytical approach**

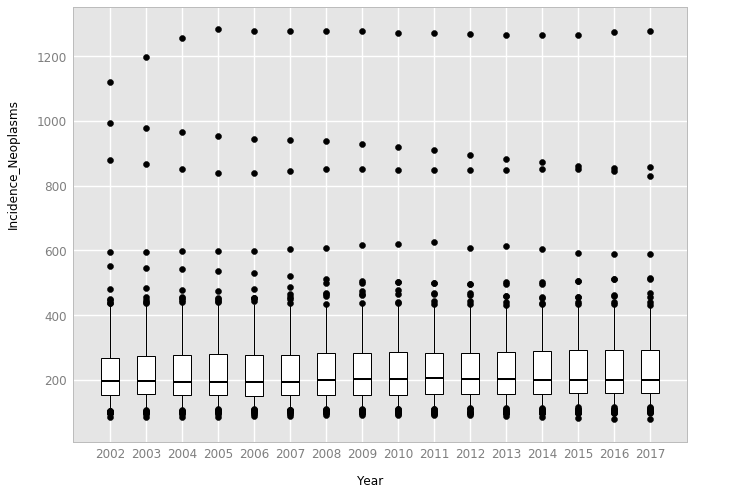
For this project, my objective was to check whether cancer incidences and death rates were related to HDI and GDP around the globe. Data sets were gathered from the website ‘Our world in data’, and later few datasets were merged using pandas dataframe in jupyter notebook (kernel python 3.6). Precautions were taken while merging the data sets, and data was cleaned for duplicate values in order to meet the criteria for time series analysis. Cancer Incidences per 100,000 cases of neoplasm (A new and abnormal growth of tissue in some part of the body, which is considered a characteristic of cancer) from 2002 - 2017 was used to describe the population of the study. Furthermore, GDP, HDI, and cancer deaths for these years were analyzed using various visualization techniques that are discussed in the results section. Additionally, to get a clear idea on the relationship between cancer and HDI/GDP, data from 10 developed countries (Norway, Switzerland, Ireland, Germany, Australia, Iceland, United Kingdom, United States, Finland, and Japan.) and 10 developing countries (Pakistan, Yemen, Liberia, Guinea, Congo, Mozambique, Afghanistan, Zimbabwe, Syria, and Iraq) was picked from the dataset based on UN HDI ranking-2019 to support the study. Moreover, correlation was applied on the desired dataset.

The dataset on cancer deaths by age was used to see the classification of cancer deaths for the 20 countries from 2002 - 2017. A supervised machine learning technique named Decision Tree Classifier module from sklearn.tree was used to check the categorization of cancer deaths by age around those 20 countries over the period of 15 years.

Web scraping techniques were used to mine data contained, which provided information about attributes based on a theme (Huanga, Morillo and Ferri 2019). About 100 pages of text data from google scholar ranging from 2002 - 2017 was mined using BeautifulSoup module, for research topics based on Cancer incidence/death and its relationship to HDI and GDP. Different modules from natural language toolkit (nltk) were used to clean the text data. LDA (Latent Dirichlet Allocation) topic model was applied on the filtered text data, number of topics, K were initially set as 20 with alpha value as 0.01 per 10 words and 20 topics were weighted.

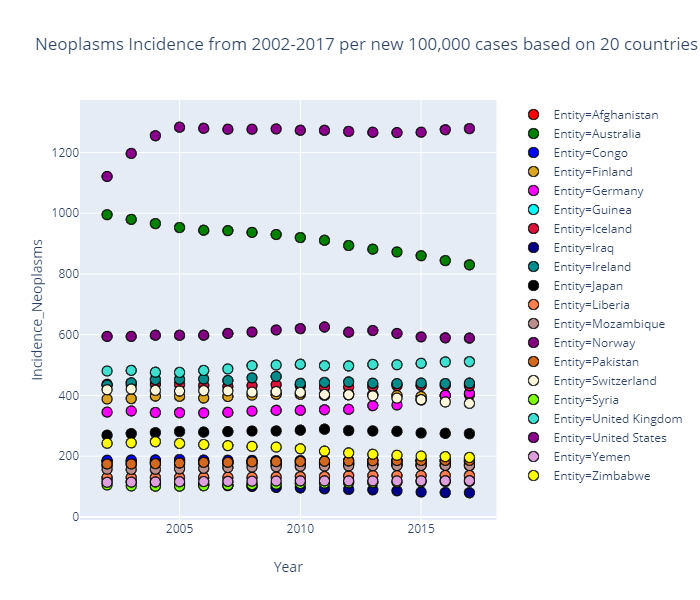
SNA (Social Network Analysis) was carried out using tweets from Twitter using hashtags on cancer. Tweets were filtered using nltk modules, and a list of words with the most recurrence (co-occurring words/bigrams) was created. After this step, tweets were used to visualize network of bigrams. Furthermore, sentiment analysis on cancer tweets was observed to see cancer relationship with GDP and HDI.

**Results**

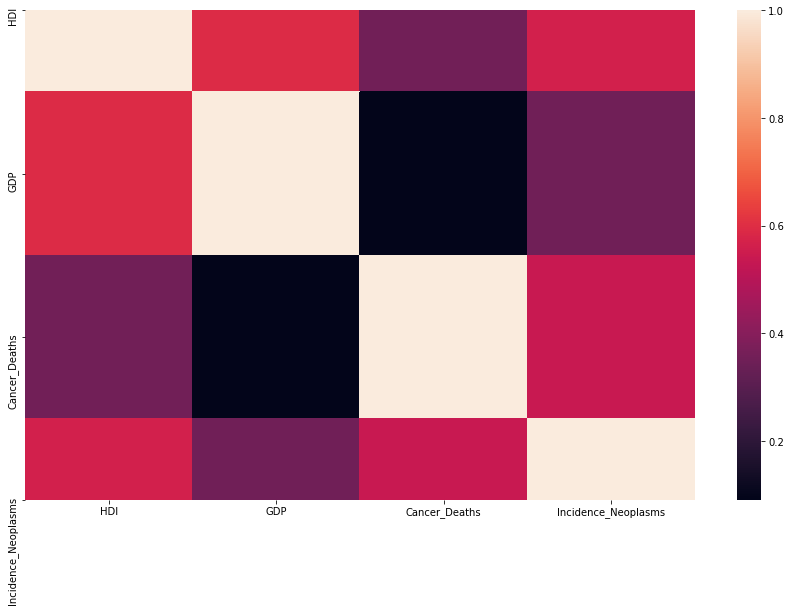


**Figure 1:** **Neoplasm Incidence yearly per new 100,000 cases**

Summary statistics for cancer incidence of neoplasm age, standardized for both genders per new 100,000 cases from year 2002 - 2017 were visualized using ggplot. Every year from 2002 - 2017 voices a story about incidence level of new cancer cases and varies for different countries. The minimum number of cases around the world through this time series is less than 200 per year, while there is an increment in the maximum number of cases, i.e., maximum number of cases in 2002 were < 1200, and this was less than that the maximum number of cases per year for all the following years – the maximum number of new cases were reported in the year 2005 (Figure 1).

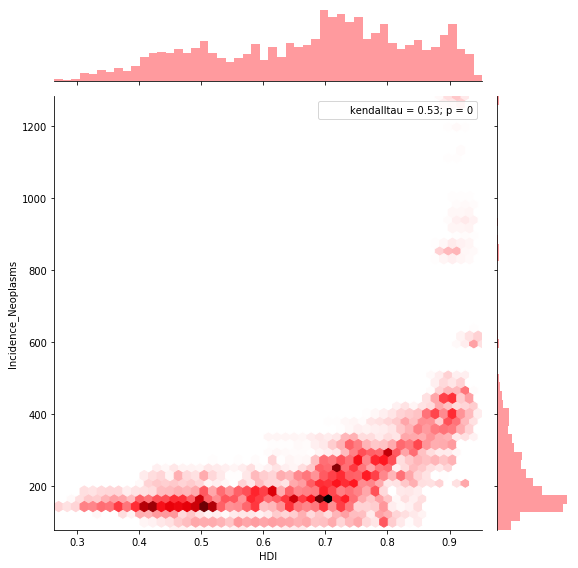


**Figure 2:** The incidences of cancer per new 100,000 cases based on 20 countries from the data set. America had the maximum number of incidences throughout the time series. Even though Australia had the second highest number of new cases, it kept declining over the years. Iraq being a developing country had the least number of new cases from 2002 to 2017.



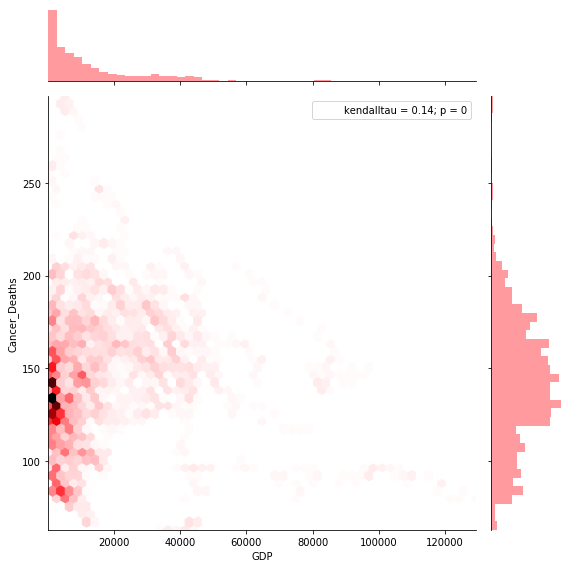
**Figure 3: Correlation matrix of cancer deaths, incidence of neoplasm, HDI, and GDP.**

Eachrow is shaded purple or orange for the sign of correlation, where dark purple color indicates negative correlation and the diagonal line specifies sign of flawless correlation. The figure clearly shows that Cancer deaths and incidences of neoplasm are negatively correlated with GDP. Whereas, HDI has a considerable amount of positive correlation with incidences of neoplasm. Corresponding to Figure 1 & 2, it can be said that cancer incidences are unrelated to the HDI.



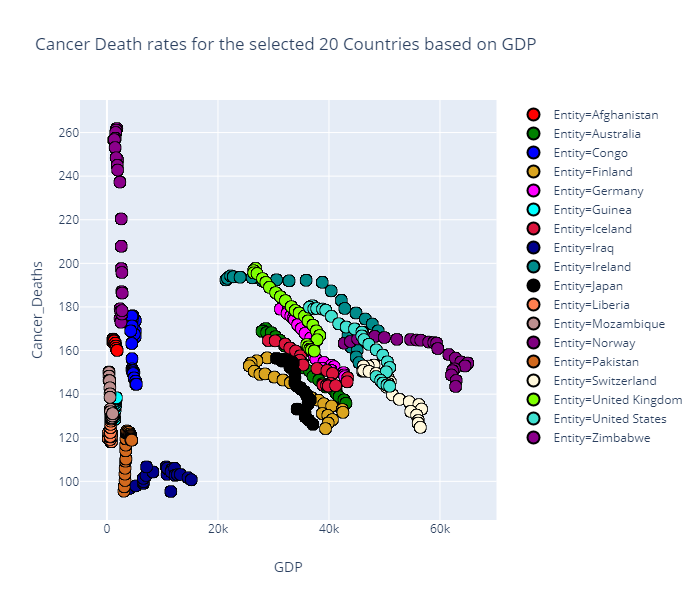
**Figure 4: Joint plot of new cancer incidences and HDI.**

The above joint plot shows that there is a correlation between HDI and cancer deaths. With kendaltu = 0.53, It is evident from figure that countries with higher GDP values have less deaths recorded than that of countries with lower GDP values.

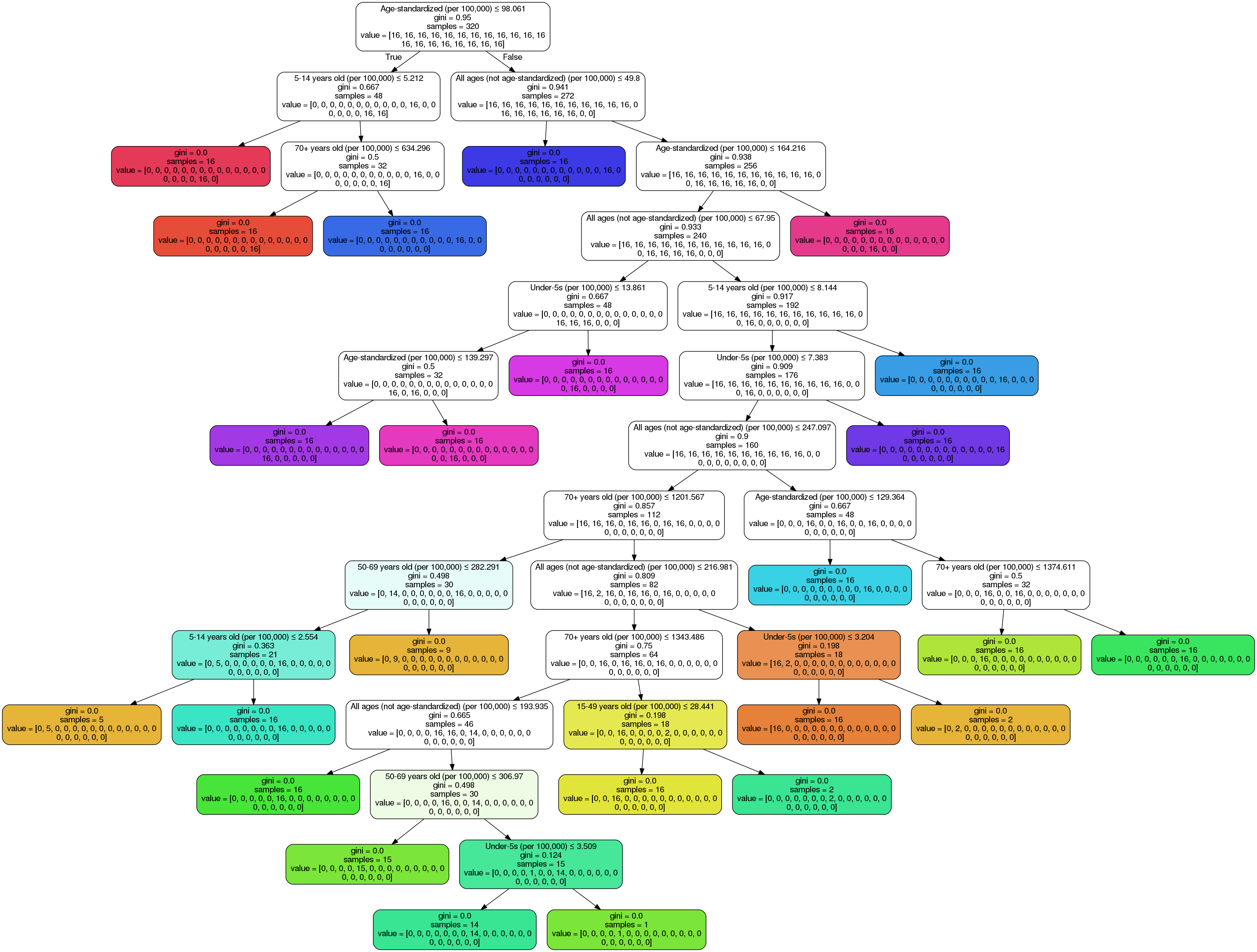


**Figure 5: Joint plot of new cancer incidences and GDP.**

This figure shows that GDP and cancer deaths are correlated, i.e., low GDP equals more cancer deaths. Additionally, Figure 5 depicts that new cancer incidences are reported more in developed countries than that of developing countries, which can be linked to the fact that developed countries have advanced in medicine and have good resources in terms of cancer treatments. To check for the evidence of these two figures, i.e., relationship between HDI/GDP and cancer, I further looked into the trends of the 20 selected countries.



**Figure 6:** This figure represents **Plotly** visualization of the death rates of 10 developed and 10 developing countries with respect to GDP and HDI (the size of these plots is based on HDI levels). Here the it can be seen that countries with High HDI level or GDP (developed) had reported less numbers of cancer deaths from 2002 to 2017. While there was a variation in number of cancer deaths for developing countries during that time.



**Figure 7:** **Classification of cancer deaths by age from 2002 and 2017 for 20 countries.**

The top root node of this tree in Figure 7 starts the classification of cancer deaths based on age standardized per 100,000 deaths with a Gini index of 0.95, which means that this is randomly distributed across the class (Scikit uses gini index/impurity for classification). This top root node level shows a trend that the cancer death rates among the ages present in this dataset were randomly picked or it can be said that the number of cancer deaths reported were randomly distributed. Accuracy score was applied to this model, and it was significant.

Classification report

precision recall f1-score support

0 1.00 1.00 1.00 16

1 1.00 1.00 1.00 16

2 1.00 1.00 1.00 16

3 1.00 1.00 1.00 16

4 1.00 1.00 1.00 16

5 1.00 1.00 1.00 16

6 1.00 1.00 1.00 16

7 1.00 1.00 1.00 16

8 1.00 1.00 1.00 16

9 1.00 1.00 1.00 16

10 1.00 1.00 1.00 16

11 1.00 1.00 1.00 16

12 1.00 1.00 1.00 16

13 1.00 1.00 1.00 16

14 1.00 1.00 1.00 16

15 1.00 1.00 1.00 16

16 1.00 1.00 1.00 16

17 1.00 1.00 1.00 16

18 1.00 1.00 1.00 16

19 1.00 1.00 1.00 16

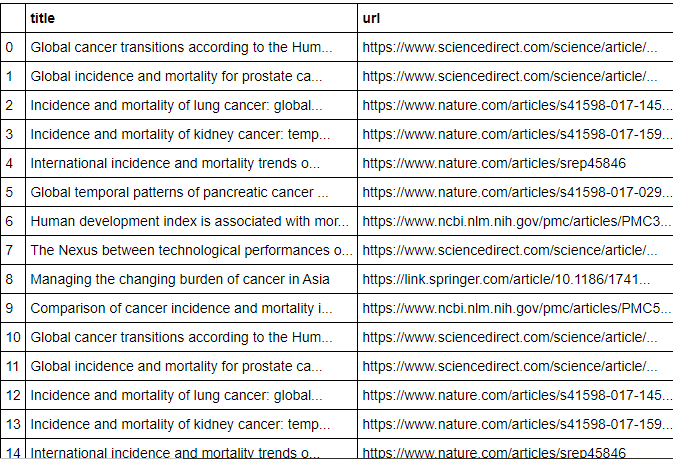
accuracy 1.00 320

macro avg 1.00 1.00 1.00 320

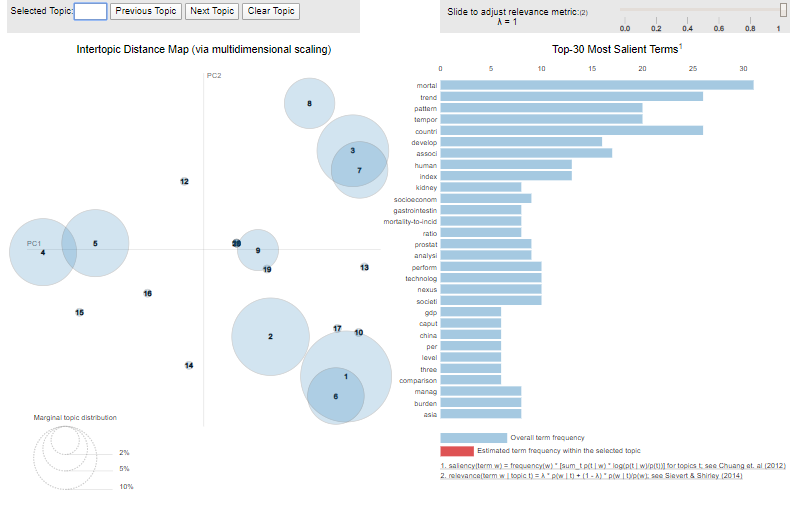
weighted avg 1.00 1.00 1.00 320

**Figure 8:** **Model Classification report, 0 – not randomly classified and 1 – random classification**

In Figure 8. it can be seen that this model has accurate score for the confusion matrix, which further implies that the classification of deaths based on age in the 20 countries selected is picked randomly.



**Figure 9:** **The text data scraped from google scholar.**

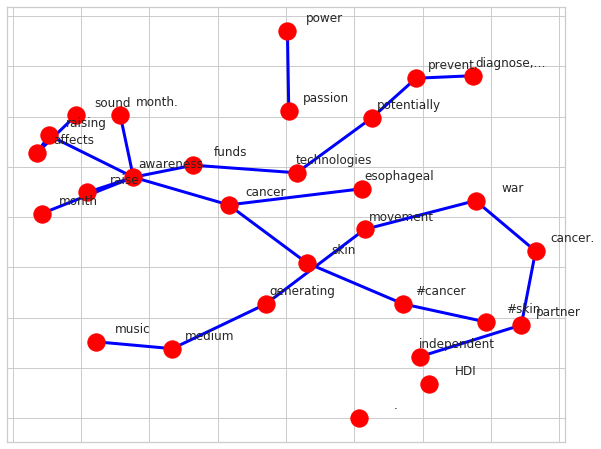


**Figure 10: LDA topic model.**

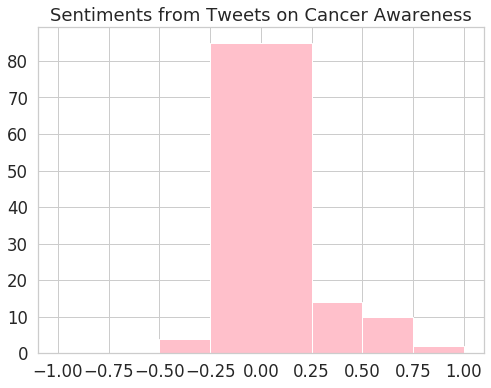
Text data was scraped from google scholar using Beautifulsoup and the data was cleaned (Figure 9). The clean text data was converted into a list and later into a corpus. The number of topics initially set was 20 and 10 words were picked from 20 topics to check if cancer was associated with GDP and HDI and was it a concern of the publishers during those years (2002-2017). Figure 10 shares a snapshot of the LDA module, which is more interactive and livelier in jupyter notebook. Estimated term frequency of topics 1, 6, and 10 is about ‘GDP’, ‘mortality’, ‘kidney’, ‘prostat’, ‘population bas’ and so on. For topics 4 and 5, the estimated term frequency is ‘human’, ‘development’, and ‘index’; and for topic 7, the words ‘lung’, and ‘socioeconomic status’ have more estimated frequency. There were evidences of studies conducted based on GDP and HDI for different diseases, but not enough studies have been conducted on the relationship between cancer and HDI/GDP.

Based on historical findings on previous research topics, I got curious about the present moment and general public awareness towards cancer and its relationship with GDP and HDI. The 1000 most recent tweets on ‘cancer awareness’ from Twitter were scraped. Co-occurring words in tweets i.e., bigrams from nltk were used to create a network graph. Using this network graph, HDI node (a hashable node) was added to the network to check for an association between the tweets used in this network graph. Figure 11 shows the network graph where HDI node does not show any connectivity between the tweeted words.

To identify the orientation of tweets on cancer awareness, sentiment analysis based on polarity was applied on the scraped tweets, and it was noticed that the nature of tweets on cancer awareness was neutral from 0 to 0.25 and -0.25, i.e., 80% of tweets were unbiased (Figure 12). There were very few tweets which had negative sentiments, and positive tweets were present, but not that justifiable to show a connection between cancer and HDI/GDP.



**Figure 11:** **The networks of co-occurring words in tweets on cancer awareness.**



**Figure 12:** **Sentiment on tweets.**

**Discussion and conclusion**

While going through the time series data, previous research topics, and current Twitter data, I discovered that there are evidences of relationship between cancer and HDI/GDP but the evidences may vary based on the source. From the data analysis of ‘our world in data’, it was found that developed countries with high HDI and GDP had the maximum number of new cancer cases from 2002 to 2017, but the death rates for those countries were significantly less than that of developing countries. Furthermore, the GDP data from 10 developed and 10 developing countries showed that there was undoubtedly high number of new cases in developed countries, but the death rates are higher in developing countries, which indicates the difference of resources and the advancement of medicine between the two country groups. A few researches have previously been performed establishing the relationship between the association of cancer and HDI/GDP, but those studies were limited in numbers than that of the other diseases. Finally, the present-day public opinion on cancer revealed that the public sentiment towards cancer awareness was neutral and the tweets usually don’t relate cancer with HDI/GDP.

**References**

Cuomo, Raphael E. and Tim K. Mackey. 2017. “The Availability of Essential Cancer Medication: An Analysis of National Formularies.” *Journal of Cancer Policy* 12:49–54.

Ghoncheh, Mahshid, Maryam Mirzaei, and Hamid Salehiniya. 2016. “Incidence and Mortality of Breast Cancer and Their Relationship with the Human Development Index (HDI) in the World in 2012.” *Asian Pacific Journal of Cancer Prevention* 16(18):8439–43.

Are, Chandrakanth et al. 2017. “Global Epidemiological Trends and Variations in the Burden of Gallbladder Cancer.” *Journal of Surgical Oncology* 115(5):580–90.

Vallejo-Huanga, Diego, Paulina Morillo, and Cèsar Ferri. 2019. “A Dataset of Attributes from Papers of a Machine Learning Conference.” *Data in Brief* 24:103836.

Bray, Freddie, Ahmedin Jemal, Nathan Grey, Jacques Ferlay, and David Forman. 2012. “Global Cancer Transitions According to the Human Development Index (2008–2030): a Population-Based Study.” *The Lancet Oncology* 13(8):790–801.

Anon. n.d. “Our World in Data.” *Our World in Data*. Retrieved May 22, 2020 (https://ourworldindata.org/).

Ritchie, Hannah. n.d. “How Many People in the World Die from Cancer?” *Our World in Data*. Retrieved May 22, 2020 (https://ourworldindata.org/how-many-people-in-the-world-die-from-cancer).