Decision Tree Classifier Based Model for Disease Prediction

Conference Paper · July 2021

CITATIONS
0

READS
59

6 authors, including:

Arianna Fuoco
Oakland University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Adam Komeshak
Oakland University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Mohammed Mahmoud
University of Jamestown
118 PUBLICATIONS **37** CITATIONS
SEE PROFILE

DECISION TREE CLASSIFIER BASED MODEL FOR DISEASE PREDICTION

[1]Andrew Alhaj, [2]Arianna Fuoco, [3]Adam Komeshak, [4]Camron Farida, [5]Caleb White, [6]Mohammed Mahmoud

[1]andrewalhaj@oakland.edu, [2]afuoco@oakland.edu, [3]arkomeshak@oakland.edu, [4]cfarida@oakland.edu, [5]calebwhite@oakland.edu, [6]mahmoud2@oakland.edu

[1-6]Department of Computer Science and Engineering, Oakland University, Rochester, MI, USA

Abstract: According to the Centers for Disease Control and Prevention (CDC), diseases such as malaria, diarrheal infections, meningitis and HIV/AIDS heavily affect developing countries. Research conducted previously used many different approaches; some relying on multiple Machine Learning models, while others relied on using a single model. In this research, we analyzed the dataset using the Decision Tree Classifier model in Python. Our goal is to see if there is any correlation found between age and diseases. We wanted to know which diseases had the biggest effect on each age group. With our findings, we were successfully able to identify that disease outbreaks were very prevalent in children 10 and under.

Keywords: Machine Learning (ML), Decision Tree Classifier and Python.

---

1. INTRODUCTION

In developing countries, there are increasing issues with diseases and infections spreading throughout the population. Diseases have always played a major role in human society, and one crucial element that has always emerged is the age group they affect the most. Diseases tend to affect nearly everyone at some point in their life. But, many of the diseases are certainly avoidable by merely recognizing when a person is most likely to contract a particular disease. If a person is aware of when these diseases are most likely to occur, it makes it a lot easier to prepare oneself and attempt to avoid said disease. Diseases discussed within this research include cholera, diarrhea, Ebola, malaria, Marburg virus, measles, meningitis, rubella mars, viral hemorrhagic fever and yellow fever. Furthermore, to better understand the diseases researched, a brief introduction of each will follow. It is also worth noting the relevance of the diseases in regards to the current pandemic the world is actively experiencing today, with the novel coronavirus.

Cholera has been a global public health challenge since 1817 [1]. Cholera is an infection caused by the bacterium vibrio cholerae and is transferred through contaminated food or water. If a country's public healthcare system is poverty-stricken, this disease can often become fatal. Though the disease is predictable, the needed resources to prevent the disease are essential. Both Ethiopia and Sudan are cholera endemic regions, which have experienced several cholera epidemics, including recent outbreaks in 2019, 2020 and 2021 [1]. Diarrhea is a preventable disease and it can be treated. Diarrheal disease is the second-largest cause of death in children under the age of five and contributes to the death of approximately 525,000 children each year [2]. Diarrhea is one of the most common endemic diseases in Indonesia [2]. Similar to the results of these statistics, the following research incorporates the Decision Tree algorithm to project results. Given the dataset provided, this research attempts to find a correlation between age and disease using Python. In addition, ML predictions are implemented to predict disease outcomes based on a specific gender and a particular age.

Moreover, another disease looked at is the Ebola virus disease (EVD). EVD is a severe, highly contagious and often fatal systemic disease in human and non-human primate [3]. Many are aware of the relevance of Ebola having common roots near the equator. For example, in West Africa, Ebola is very prevalent. Ebola outbreaks require an emergency response, pre-existing knowledge and understanding of exposure patterns and their interplay with gender-associated risk factors in order to provide fundamental assistance with planning said emergency response [4]. Malaria has remained one of the major global public health challenges for the last two decades, especially in low and medium-income countries; putting nearly half of the world population at risk of infection [5]. Though cases may differ depending on location, it is noted that cases within children under five years of age are relevant with research showing that one child dies from malaria every two minutes [5].

Rubella mars, also referred to as the German measles; though, it is not the same thing as the measles itself. There are similarities between the symptoms that can occur, like a red rash. Rubella is not as severe as the measles. The delayed implementation of childhood vaccines will be expected to significantly impact the disease burden of vaccine-preventable diseases (VPDs) among children [6]. Meningitis is a life-threatening inflammation of the membranes covering the brain and spinal cord [7]. As mentioned previously, it is crucial to account for disease occurrences and higher probability rates within countries of low income. Additionally, regional impacts of locations also produce higher rates of infection depending on how a specific virus thrives. For example, yellow fever is common in tropical areas within the world, such as Africa and South America. In Africa, it is the third most commonly reported type of disease outbreak [9].

Data was first analyzed and grouped using pandas. Confirmed contracted diseases were analyzed, grouped and then sorted. Age range was also analyzed, grouped and then sorted.

After a correlation was confirmed from the data analysis and visualization, ML prediction using sklearn Decision Tree Classifier was implemented in an attempt to predict disease based on the individuals age and gender. The Decision Tree Classifier algorithm was used for our predictions. Our goal is to attempt to see if there is a correlation between age, gender and disease. i.e., X ages/gender are likely to have Y diseases.

2. THE PROPOSED MODEL

Several Python libraries have been imported to aid in the completion of this research. The first is Numpy, which is a package for scientific computing. Next, Pandas is used for data analysis and is a manipulation tool. Followed by Matplotlib which offers assistance in visualizing your data. And lastly, Seaborn which is a Python data visualization library that provides more advanced ways to visualize your data.

The Decision Tree Classifier [8] [14] was imported from sklearn. Classification involves two main steps, a learning and then prediction step. The learning step is developed through the data itself. In the prediction step, the model then attempts a prediction using the given data.

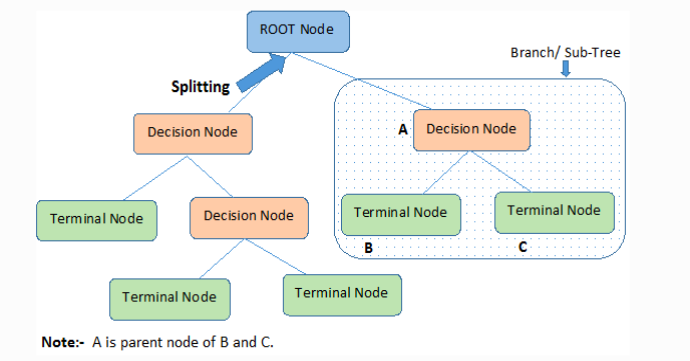

Figure 1. The Standard Decision Tree Classifier

3. RESULTS

Data analysis showed a total of 10 different diseases within the dataset. Cholera with a count of 28,590 cases, diarrhea with a count of 28,550 cases, Ebola with a count of 28,179 cases, malaria with a count of 28,535 cases, Marburg virus with a count of 28,430 cases, measles with a count of 28,471 cases, meningitis with a count of 28,362 cases, rubella mars with a count of 28,539 cases, viral hemorrhagic fever with a count of 28,401 cases and yellow fever with a count of 28,428 diseases.

```
groupby_disease = ds.groupby(["disease"]).size()
print("COUNTS OF EACH DISEASE")
print("")
print(groupby_disease)
a=ds["disease"]
print("##############################################################################")
print("DISEASES IN A LIST")
print("")
print(a)
```

```
COUNTS OF EACH DISEASE

disease
Cholera                 28589
Diarrhoea               28550
Ebola                   28179
Malaria                 28535
Marburg Virus           28430
Measles                 28471
Meningitis              28362
Rubella Mars            28539
Viral Haemmorrhaphic Fever    28401
Yellow Fever            28428
```

Figure 2. Grouping Diseases

Data analysis for the ages within the dataset showed a range from 0 to 78 years old, 0 included individuals under the age of 1.

```
groupby_age = ds.groupby("age").size()
print("COUNTS OF EACH AGE")
print("")
print(groupby_age)
print("##############################################################################")
print("AGES IN A LIST")
print("")
b=ds["age"]
print(b)
```

```
COUNTS OF EACH AGE

age
0      77
1    4276
2    7492
3    6935
4    6628
     ...
74   1730
75   1469
76   1057
77    749
78    362
Length: 79, dtype: int64
##############################################################################
```

Figure 3. Grouping Age

Data was then visualized using the Seaborn's Jointplot. Results of the joint plot showed a concentration of diseases in individuals aged 10 and under.

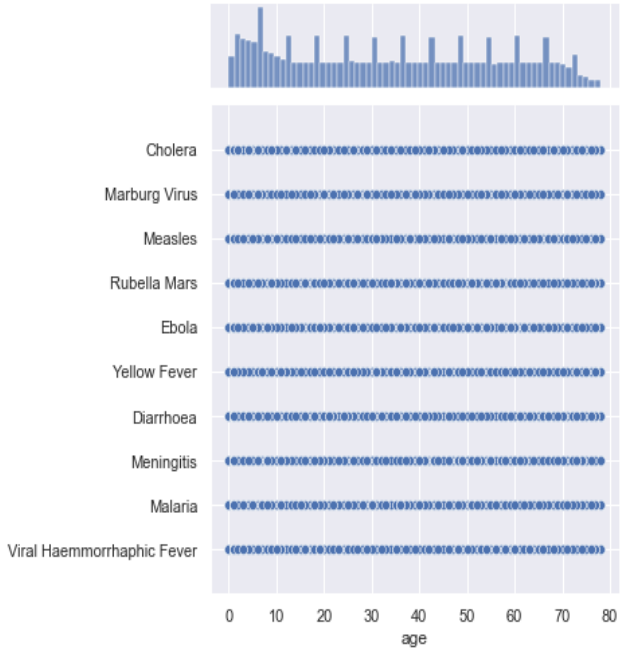

Figure 4. Diseases by Age

Machine learning predictions using the Decision Tree Classifier algorithm were able to successfully predict which diseases individuals under 10 are likely to contract from the diseases in the dataset, based on the age and gender of the individual.

```
# First value in a1: 1 = Male   2 = Female
# Second value in a1: Age

learningds = pd.read_csv("disease.csv")
X = learningds.drop(columns=["disease"])
y = learningds["disease"]

model = DecisionTreeClassifier()
model.fit(X,y)
a1=model.predict([[2, 3]])

print("Predicited disease for this invidual is " + a1)
```

```
['Predicited disease for this invidual is Cholera']
```

Figure 5. Disease Prediction

4. CONCLUSION

Diseases have, and will always, play a role in our lives. Throughout our history as a species, even now more so than ever, numerous diseases have plagued us. However, being able to understand how ML can help predict the likelihood various diseases can affect certain age and geographical locations, we might be able to prevent many unnecessary deaths.

5. FUTURE WORK

As of current, the dataset is concentrated on the aspect of third-world countries and disease correlation. However, when it comes to improvements for the future, many additions could help improve the dataset

An example of such improvement would be seeing if the dataset would work with a broader base of countries. Expanding our dataset to demonstrate the correlation between age and disease in terms of every individual region would be extremely advantageous. This improvement, accompanied with the ability to view a worldwide dataset, as well as the specified dataset of the region. Having a global view of age to disease correlation is advantageous in its own way by revealing the correlation between the diseases that can be found anywhere in the world; though, there are some diseases that are only found within specific regions.

When looking at a dataset that is showing the third-world country correlation between age and diseases contracted, people do not need to worry about a specific disease because it is not commonly found in their region. To provide some perspective, malaria is the fourth leading cause of death in Nigeria according to the CDC [10]. For example, if a citizen residing in the United States were to look at a dataset showing malaria as being one of the top diseases to be contracted within the age range of 10 to 20, it may not be very relevant to them. Aside from some researchers, anyone finding this information would have virtually no use for it because it does not affect them. In the United States, malaria is nowhere near the top 10 leading

diseases causing deaths [11]. Although, from a researcher's standpoint, being able to view a dataset from a third-world country could be useful, especially if research is being conducted on malaria or diseases in Africa. Showing data only from third-world countries may be useful in some situations, but having the ability to view all regions in the world would allow us to broaden our target demographic to much greater lengths.

One other potential improvement is to increase the age range from 79 to 90 years. This means that if the dataset were to be expanded to include an age range of 0-90 years then the importance of the dataset could increase. In addition, it would be more useful because a larger range of people could utilize it. Having the age range at 79 years does cover a greater percentage of the population as most countries have an average life expectancy of around 81 years of age, with the United States being around 78.6 years and Japan at 84.2 years [12]. However, it's great to keep in mind that these ages are averages as there are plenty of people who live beyond 79 years of age who would benefit from such a dataset. Increasing the age range would be more beneficial, along with the addition of an increased amount of regions in order to view the disease correlations at a higher perspective. For example, according to the United States Census from 2011, there are more than 4.7% of people who live to be 90 or higher [13]. Therefore, having more regions available to view provides the possibility of a much larger perspective, given an increased age range.

Since many third-world countries tend to have a lower life expectancy, having the age range at 79 years is acceptable because it is accurate to third-world country life expectancies; though, other regions do have a much higher life expectancy. To conclude, if the dataset were to provide a larger variety and increased set of regions, the need for an increased age range would also likely be necessary.

REFERENCES

[1] Hassan, O. B., & Nellums, L. B. (2021). Cholera during COVID-19: The forgotten threat for forcibly displaced populations. *EClinicalMedicine*, *32*. https://doi.org/10.1016/j.eclinm.2021.100753

[2] Saputra, A. M. W., & Wijayanto, A. W. (2021). IMPLEMENTATION OF ENSEMBLE TECHNIQUES FOR DIARRHEA CASES CLASSIFICATION OF UNDER-FIVE CHILDREN IN INDONESIA. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), 6(2), 175-180. https://doi.org/10.33480/jitk.v6i2.1935

[3] Adekanmbi, O., Ilesanmi, O., & Lakoh, S. (2021). Ebola: A review and focus on neurologic manifestations. Journal of the Neurological Sciences, 117311. https://doi.org/10.1016/j.jns.2021.117311

[4] Nkangu, M. N., Olatunde, O. A., & Yaya, S. (2017). The perspective of gender on the Ebola virus using a risk management and population health framework: a scoping review. Infectious diseases of poverty, 6(1), 135. https://doi.org/10.1186/s40249-017-0346-7

[5] Obasohan, P. E., Walters, S. J., Jacques, R., & Khatab, K. (2021). A Scoping Review of Selected Studies on Predictor Variables Associated with the Malaria Status among Children under Five Years in Sub-Saharan Africa. International Journal of Environmental Research and Public Health, 18(4), 2119. https://doi.org/10.3390/ijerph18042119

[6] Shimizu, K., Teshima, A., & Mase, H. (2021). Measles and Rubella during COVID-19 Pandemic: Future Challenges in Japan. International Journal of Environmental Research and Public Health, 18(1), 9. https://doi.org/10.3390/ijerph18010009

[7] Durrheim, D. N., Andrus, J. K., Tabassum, S., Bashour, H., Githanga, D., & Pfaff, G. (2021). A dangerous measles future looms beyond the COVID-19 pandemic. Nature Medicine, 1-2. https://www.nature.com/articles/s41591-021-01237-5

[8] Du, W., Du, W., Zhan, Z., & Zhan, Z. (2002). Building decision tree classifier on private data. *Proceedings of the IEEE International Conference on Privacy, Security and Data Mining-Volume 14*, 1–8. http://portal.acm.org/citation.cfm?id=850784

[9] Gaythorpe, K. A., Hamlet, A., Jean, K., Ramos, D. G., Cibrelus, L., Garske, T., & Ferguson, N. (2021). The global burden of yellow fever. Elife, 10, e64670. https://elifesciences.org/articles/64670

[10] CDC. (2018). *Global Health - Nigeria*. Retrieved from https://www.cdc.govm: https://www.cdc.gov/globalhealth/countries/nigeria/default.htm

[11] CDC. (2021). *Leading Causes of Death*. Retrieved from https://www.cdc.gov: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

[12] Kamal, R. (2019, December 23). *How does U.S. life expectancy compare to other countries?* Retrieved from https://www.healthsystemtracker.org/: https://www.healthsystemtracker.org/chart-collection/u-s-life-expectancy-compare-countries/#item-le_life-expectancy-at-birth-in-years-2017_dec-2019-update

[13] Bernstein, R. (2011, November 17). *Census Bureau Releases Comprehensive Analysis of Fast-Growing 90-and-Older Population*. Retrieved from https://www.census.gov/:

https://www.census.gov/newsroom/releases/archives/aging_population/cb11-194.html#:~:text=Because%20of%20increases%20in%20life,likely%20to%20reach%2010%20percent.

[14] Safavian, S. R., & Landgrebe, D. (1991). A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems, Man and Cybernetics*, *21*(3), 660–674. https://doi.org/10.1109/21.97458