**ABSTRACT:**

Health literacy is a discrete form of literacy and becoming an increasingly important aspect for social, economic, and health development. Health literacy is already seen as a crucial tool for the prevention of non-communicable disease with investments in education and communication. The dominant explanation for this trend is increased change in clinical and public health sectors of literacy. Greater knowledge and accessible information regrading noncommunicable diseases are allowing citizens to take the necessary precautions and strive for living a healthier life style. Today with the rapid development of coronavirus disease 2019 (COVID-19), there has been a need for people to acquire and apply health information, and adapt their behavior at a fast pace. Health communication intended to educate people about precautionary measures to take for getting or spreading the infection has become widely available. However, there is also misinformation, and individuals are considered to be able to acquire, understand, and use this information in a sound and ethical manner, or in other words to be health literate. We used data from United States Census Bureau, Internal Revenue Service (IRS), Centers of Medicare and Medicaid Services, and National Science Board, to measure literacy and health care access across The United States of America. With the help of machine learning models, we were able to analyze the correlation between literacy and COVID-19 case count across the country. We saw with an increase white collar jobs, education, and health care access, there was a decrease in number of cases across the country. These states with a higher literacy rate, also showed lower chance of the state COVID-19 case count being above the national average.

1. **INTRODUCTION:**

A 29-year-old Black woman with three days of abdominal pain and fever was brought to an emergency room by her family. After a brief evaluation, she was told that she would need an exploratory laparotomy. She subsequently became agitated and demanded to have her family take her home. When approached by staff, she yelled, “I came here in pain and all you want is to do is an exploratory on me! You will not make me a guinea pig!” She refused to consent to any procedures and later died of appendicitis. (<https://www.ncbi.nlm.nih.gov/books/NBK216035/>)

A exploratory laparotomy is surgery to open up the abdomen in order to find the cause of the symptoms, that testing could not diagnose. If the cause of the problems is discovered, treatment is often done at the same time. This procedure was misunderstood as a laboratory experiment by the patient, showing a lack of health literacy. (<http://myhealth.ucsd.edu/3,40432>)

The first use of the phrase “health literacy” occurred in 1974, began appearing in academic peer-reviewed literature in the early 1990s, and has experience exponential growth ever since. This has indicated a growing internationalization of the field of health literacy – a field that has been dominated by the United States.

However, as the field of health literacy has expanded in scope and depth, the term “health literacy” has come to take different meaning to various audiences, becoming a source of confusion and debate. In 1999, the American Medical Association’s Ad Hoc Committee on Health Literacy, defined the term as “constellation of skills, including the ability to perform basic reading and numerical tasks required to function in the health care environment” including “the ability to read and comprehend prescription bottles, appointment slips, and other essential health related materials”. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1831571/>) The definition used by Healthy People 2010, Institute of Medicine (IOM), and Network of the National Library of Medicine were similar: “The degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions”. (<https://www.ncbi.nlm.nih.gov/books/NBK216035/>) These definitions refer to health literacy as a set of individual capacities, which are both the innate potential of the individual, as well as his or her skills. An individual’s health literacy is mediated by education, and its adequacy is affected by culture, language, and the characteristics of health-related settings. Creating health literacy as a shared function of cultural, social, and individual factors. Showing that the causes and remedies for limited health literacy rest with out cultural and social framework, the health and education systems that serve it, and the interactions between these factors. (<https://www.ncbi.nlm.nih.gov/books/NBK216035/>)

The first domain within health literacy is individual capacity, which is the set of resources that a person has to deal effectively with health information, health care personnel, and the health care systems. Individual capacity has two components: reading fluency and prior knowledge. Reading fluency is the ability to mentally process written materials and form new knowledge. The National Adult Literacy Study (NALS) defined reading fluency through 3 skill sets: the ability to read and understand text, the ability to locate and use information in documents, the ability to apply arithmetic operations and use numerical information in printed materials. Prior knowledge is composed of vocabulary and conceptual knowledge (e.g. what is cancer and how does it injure the body). Therefore, individuals will understand written and spoken communication better if they are familiar with the words and concepts presented. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1831571/>)

The second domain in health literacy is culture and society. The term “culture” refers to the shared ideas, meaning, and values acquired by individuals as members of society. Cultural, social, and family influences are critical in shaping attitudes and beliefs. In health literacy, cultural and society influence is how people interact with the health system and help determine the adequacy of health literacy skills. Culture and society are defined through native language, socioeconomic status, gender, race, and ethnicity, along with influence of mass media as represented by news, publishing, advertising, marketing, and any health information available through electronic sources. These are conditions over which the individuals have little or no control but they influence the ability to participate fully in a health-literate society. (<https://www.ncbi.nlm.nih.gov/books/NBK216035/>)

The third domain in health literacy is the education system. In the United States, the education system consists of K-12 system, adult education programs, and higher education. K-12 education is charged with the development of literacy and numeracy skills in English, which form the foundation for complex comprehension and application in the later grades. Adult education programs provide opportunities for individuals who drop out of K-12, who completed high school but did not acquire strong skills, who did not have full school opportunities, or for immigrants who may never had access to education so wish to learn, speak, read, and write English. (<https://www.ncbi.nlm.nih.gov/books/NBK216035/>)

The last domain in health literacy is the health care system, which includes all people performing activities such as health-related messages and action plans, rights and responsibilities are shaped, research initiates are begun, health-promoting recommendations are developed and supported, access is monitored, and regulations are enforced. The is consistent evidence supporting the notion that health literacy affects the interaction of individuals with health contexts and the health-care system, and may further affect health status and outcomes. (<https://www.ncbi.nlm.nih.gov/books/NBK216035/>)

In summary, according to findings in Health Literacy: A Prescription to End Confusion, the committee states health literacy is defined based on the interaction of individuals’ skills with health contexts, the health-care system, the education system, and broad social and cultural factors at home, at work, and in the community. Moreover, the committee concludes that the links between education and health outcomes are strongly established. (<https://www.ncbi.nlm.nih.gov/books/NBK216035/>)

Today health literacy is already a crucial tool for the prevention of non-communicable disease with investments in education and communication sought to be sustainable, long-term measures starting early in the life course. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7156243/>) However, the question exists on the same efficiency with communicable diseases. These are illnesses caused by viruses or bacteria that people spread to one another through contact with contaminated surfaces, bodily fluids, blood products, insect bites, or through the air. Some examples of communicable diseases include HIV, hepatitis A, B, and C, measles, and salmonella. Most common forms of spread include fecal-oral, food, sexual intercourse, insect bites, contact with contaminated fomites, droplets, or skin contact. (<https://www.ncbi.nlm.nih.gov/books/NBK470303/>)

Today the world is seeing a unprecedented pandemic caused by the communicable disease called Coronavirus (COVID-19). An infection caused by a newly discovered coronavirus, which causes mild to moderate respiratory illness. The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes. Elderly and individuals with underlying medical conditions such as cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness. (<https://www.who.int/health-topics/coronavirus#tab=tab_1>)

Rapid development of the virus into a pandemic has called for people to acquire and apply health information, and adapt their daily life at a fast pace. Sporting events were canceled, classes and occupations went virtual, and retails were closed. Health communication intended to educate people about the severe acute respiratory syndrome coronavirus 2(SARS-CoV-2), prevention, and symptoms became widely available. With most valuable information distributed as easy-to-understand manner that offer simple and practical solutions, such as washing hands, maintaining social distance, and where to find latest news, recommendations, and advice. Unfortunately, there has also been complex, contradictory, and false information spreading in todays technology driven world. Individuals have to acquire, understand, and use the information in a sound and ethical manner to be health literate. However, COVID-19 has highlighted that poor health literacy among a population is a global public health crisis. In Europe, nearly half of adults reported having problems with health literacy and not having relevant competencies to take care of their health and that of others. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7156243/>)

With COVID-19 two aspects of health literacy have emerged. First globally, health literacy is as important for the prevention of communicable disease as it is for non-communicable diseases. Second, along with system preparedness, individual preparedness is key for solving complex real-life problems. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7156243/>)

1. **MATERIALS AND METHODS:**

The final goal of the research was to develop a machine learning model for predicting case count per state and predictive classification of each state as higher or lower than the national average case count according to the socioeconomic and sociodemographic analysis of each state. This goal was achieved through primary exploratory data analysis (EDA), spearman correlation, and ordinary least square regression (OLS). Furthermore, predictive analysis was done through data preprocessing and fitting in linear, support vector machines (SVM), random forest, XGBoost, and logistic regression models.

* 1. Spearman Correlation:

The Spearman correlation is a nonparametric measure of the monotonicity of the relationship between two datasets. It does not assume that both datasets are normally distributed and varies between -1 and +1, with 0 implying no correlation. Correlations of -1 or +1 imply an exact monotonic relationship. A monotonic relationship is defined as one of the following: as the value of the independent variable increases, so does the dependent variable, or as the value of independent variable increases, the dependent variable value decreases. (<https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php>)

A monotonic relationship is not a strict assumption of Spearman’s correlation, one can run a Spearman’s correlation on a non-monotonic relationship to determine if there is a monotonic component to the association. The test is used for either ordinal variables or for continuous data that has failed the assumptions necessary for conducting the Pearson correlation. (<https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php>)

* 1. Ordinary Least Squares Regression (OLS):

The Ordinary Least Square Regression (OLS) is more commonly named linear regression. The OLS method calculates the best-fitting line for the observed data by minimizing the sum of square differences between the observed and predicted values. This minimization leads to the following estimator:

β is the vector of the estimators of the βi parameters, X is the matrix of the explanatory variables preceded by a vector of 1s, y is the vector of the n observed values of the dependent variable, p\* is the number of explanatory variables to which we add 1 if the intercept is not fixed, wi is the weight of ith observation, and W is the cum of the wi weights, and D is a matrix with the wi weights on its diagnol. (<https://www.xlstat.com/en/solutions/features/ordinary-least-squares-regression-ols>)

* 1. Linear Regression and Support Vector Machines (SVM):

Linear Regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered as an explanatory variable, and the other is considered to be dependent variable. One should determine whether or not there is a relationship between the variables of interest. This does not imply there has to be a cause and effect relationship between the variables, but there is a significant association between the two. (<http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>)

Support vector machines (SVM) are a set of supervised learning methods used for classification, regression, and outlier detection. The advantages of SVMs are: effective in high dimensional spaces and where number of dimensions is greater than the number of samples, uses a subset of training points in the decision function (called support vectors), and versatile as different kernel functions can be specified for the decision function. For the goal of this project, Support Vector Regression is used, which produced by support vector classification depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. It also depends only on a subset of the training data, because the cost function ignores samples whose prediction is close to their target. (<https://scikit-learn.org/stable/modules/svm.html#svm-regression>)

* 1. Random Forest and XGBoost:

Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve particular problems. They are primarily used to improve the performance of a model or reduce the likelihood of an unfortunate selection of a poor one. This learning is often used for classification, prediction, or function approximation. Other applications of ensemble learning include assigning a confidence of the decision made by the model, selecting optimal features, data fusion, incremental learning, nonstationary learning, and error-correcting

An ensemble-based system is obtained by combining diverse models, thus such systems are also known as multiple classifier systems, or just ensemble systems.

<https://dl.acm.org/doi/pdf/10.1145/2939672.2939785>

<https://xgboost.readthedocs.io/en/latest/tutorials/model.html>

<http://www.scholarpedia.org/article/Ensemble_learning#:~:text=Ensemble%20learning%20is%20the%20process,%2C%20function%20approximation%2C%20etc.)>

<https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/springerEBR09.pdf>

<https://www.researchgate.net/profile/Andy_Liaw/publication/228451484_Classification_and_Regression_by_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression-by-RandomForest.pdf>

<https://escholarship.org/uc/item/35x3v9t4>

<https://scikit-learn.org/stable/modules/ensemble.html#forest>

1. **RESULTS:**
2. **DISCUSSION:**
3. **REFERENCES:**