Assignment Date: October 28th, 2024

Summative Assignment

Module: Applied Artificial Intelligence

Executive Summary

The objective of the assignment, set by the Non-Governmental Organization (NGO), was to use statistical evidence for a grant funding application to overcome child malnutrition. Which was achieved using datasets obtained from the World Health Organization (WHO) containing statistical data on infant mortality rates, ranging from ages 0 to 5 years, and nutritional data for the respective areas and timepoints.

Through a series of sequential steps, both nutrition and mortality datasets were preprocessed and cleaned using Python libraries, more specifically, Pandas, to handle and manipulate the datasets and NumPy to address numerical operations in Simulated Annealing (SA), and Linear Regression methods. This was to ensure the data remained consistent addressing issues such as, naming conventions, removing missing values, conversion and formatting of column values. After which the datasets were merged creating a single and comprehensive dataset suitable for further analysis.

After the data was processed and prepared, SA, an optimization search method, was selected due to its balance between explorative and exploitative characteristics, and ability to locate global optima. The algorithm was then used optimize the search and selection throughout the merged dataset for the best set of weights, for the intercept and coefficient, which in turn minimized the mean squared error (MSE) for the predicted and actual mortality rates in the linear regression model.

The final step after optimization was implementing supervised learning model, linear regression, to exhibit the accuracy of the predicted and actual relationship between breastfeeding and mortality rates. Linear regression was chosen based on the simplicity of the dataset, ease of implementation, and the assumption of an anticipated linear relationship between the variables.

The results obtained displayed a negative correlation between breastfeeding rates and child mortality rates. However, a high Mean Squared Error (MSE) value was also obtained, indicating breastfeeding is not the most, or only significant factor, affecting child mortality rates.

Future direction regarding this project suggests incorporating more correlating variables that influence mortality rates and can better improve MSE values, such as economic indicators and access to healthcare. In addition, the inclusion of more complex machine learning models such as neural networks can better capture both linear and non-linear relationships between all variables which would be better assist in future grant funding endeavours.

Introduction

Malnutrition is a continued global concern, that has derived attention and intervention efforts from many organizations including non-governmental organization (NGOs). This holds especially true when concerning malnutrition and famine in children and infants, focusing on correlating variables such as breastfeeding percentages and child mortality rates. Proving these relationships provides a strong foundation to secure grant funding to develop programs that will further assess, and potentially address, the issue at hand.

AI has been a revolutionary development that has been seen throughout many industries. Using algorithms and through the development of complex models resulting predictions, decisions, and pattern identification can be made through collected data. Similarly, we can take note the application of AI in healthcare and research and in clinical settings as one of these industries, creating opportunities for advancements and optimizing general protocols, and large-scale data analysis for sustainability and development goals [1]. Results derived increase accuracy, and efficiency in both time and resources, producing effective and significant results improving healthcare through improved diagnostics and therapeutic research in real time and with less resources [2].

The association between Artificial Intelligence and the objective of this assignment consists of providing a strong argument, with statistical evidence, that a causal factor of global child mortality rates can be attributed to nutrition, more specifically breastfeeding. Previous studies have advocated for the importance of increased breastfeeding to limit child mortality, and other physical developmental issues that can be derived from malnutrition such as infections, development of co-morbidities and constraining of cognitive and physical development [3]. Research has also pointed that breastfeeding benefits the mother as well, lowering rates of breast and ovarian cancer [4].

The datasets presented in this scenario contains abundant but inconsistent data. The use of AI methods, and other algorithms will be used to clean, organize and optimize the data for prediction and verification analysis for correlating trends. Using supervised machine learning models, and optimization algorithms, the results obtained will accurately and actively highlight the quantitative evidence needed in support of strengthening the argument for the grant application process and for future endeavours.

The proposed solution in this case consists of steps of data processing, optimization using SA, and linear regression. Data preprocessing is the initial step where missing values will be addressed and the datasets will be aligned according to country and year, ensuring the data is uniform and organized appropriately. SA was used as an optimization technique to search and select the best weights, for the intercept and coefficient, for a minimized MSE. A linear regression model was used as a prediction model for the dependent and independent variables to confirm the relation between the datasets. Where results will be used as empirical evidence for to strengthen the grant application.

Literature Review

Global health data analysis faces numerous challenges, one common issue is inconsistency and incompleteness which can be attributed to differing data collection standards, lack of proper data integrity, missing values, and time gaps impacting drawing reliable conclusions [5]. These issues significantly impede the ability to generate accurate health insights and prediction analysis, particularly in low-resource or crisis-affected regions. The implantation of AI is an active method in solving these issues, through techniques and methods for data manipulation addressing missing and misaligned data, optimization and selecting data points, and pattern recognition for accurate correlation prediction analysis.

AI has coevolved methods and applications that are in line with the common problems faced. For instance, the issue of misaligned datasets can be addressed through optimization techniques such as, genetic algorithms, SA, and Tabu search creating a comprehensive dataset [6]. AI further addresses these challenges by implementing machine learning methods such as linear regression, neural networks and decision trees. These have been used for predictive analysis for informed decision-making, program development, and contingency planning [7].

For this assignment, linear regression was used to determine provide outcomes between independent variable, breastfeeding, and the dependent, child mortality. This method appropriate due to the simplicity of the data and the anticipated linear relationship. Linear regression has also been applied in the focuses of flu season prediction, disease monitoring and tracking epidemics [8]. Using these methods will allow for early confirmation of diagnosis, assess morbidity or mortality risks, and through this create an effective health response plan comparatively to other non-machine learning models [1].

Python was chosen as the tool due to its ease of use in data manipulation and machine learning through its many available libraries and tools. Pandas libraries were used for to handle and manipulate the datasets, and NumPy allowed for mathematical operations to be conducted in SA optimization and linear regression steps [9]. The criteria used for tool selection was based on ability to achieve successful interpretation of the given datasets. In which scalability, ease of implementation, and accuracy in organizing, optimizing and analyzing the data as requirements. In addition, Python's simplicity provides flexibility when troubleshooting dataset or algorithm errors.

Linear regression can be noted as a simple yet robust model in the aspect of interpretation for predictive outcomes. However, issues arise when it encounters unfit data, more specifically, when the relationship between the variables is non-linear or if there are multiple variables that interact with one another. In this case, a different mode of approach must be taken, using more flexible models such as neural networks or decision trees would be more appropriate for these situations [10]. However direct and simple interpretation may not be as simple comparatively in those cases [10]. From an optimization perspective, SA is highly flexible and provides explorative and exploitative approaches, providing solutions closer to global optima. Other methods such as genetic algorithms, while effective and powerful, are more meticulous in fine-tuning parameters and are also costly in terms of computational power, making it a less appropriate option in the context of this assignment.

Many research efforts have already indicated that there is clear and definitive relationship between breastfeeding data and child mortality rates especially in middle to lower income countries [11]. The research has shown that additional multitude variable impacts child mortality and enhance the accuracy of prediction models, in conjunction with breastfeeding, such as socio-economic variables, access and quality of healthcare [12]. Current research has also shown the importance of breastfeeding data in different studies such as the comparison between commercial milk formula and breastfeeding and the impact on child development [13]. In addition to these new variables and datasets, more complex models can be implemented to handle more complex datasets in terms of organization and optimization, preparing for predictive interpretation. Drawing from these

research sources will allows us to anticipate the results from the datasets provided by the WHO, in that there will be an evident correlation between breastfeeding data and child mortality rates, providing a strong backing for the grant application. It will also highlight the need to obtain information on additional variables, opening discussions regarding future funding opportunities.

Research Design

Prior to the development of the research design several assumptions were made to strengthen the results. First, breastfeeding is not the only significant variable affecting child mortality and furthermore are not included in the dataset such as access to healthcare. Second, the data provided is sufficient to confidently conduct analysis and draw conclusions, even after data processing. The final assumption is that there is an anticipated linear relationship between the variables making linear regression the best representative model.

Pre-processing the data was necessary for the data to be suitable for analysis. Data manipulation and normalization was undertaken in which incomplete and inconsistent data was removed, and any misaligned year ranges were also addressed. Manual intervention to address inconsistencies in naming conventions mainly data columns that contained information on country and year were converted to ‘Country’ and ‘Year’ to maintain consistency and prevent errors when merging. NumPy was used to perform the numerical operations efficiently in both SA and linear regression, whereas Pandas was used to merging and manipulation of the datasets [14][15].

SA was the optimization technique applied to search the dataset for optimal weights to be incorporated later in the supervised linear regression. SA would begin with an initial solution, and finds the next solution, in an iterative manner, starting off in a greedy accepting a potentially worse solution initially, and as the simulations pass and the algorithm cools it becomes more selective. This approach ensures the solution does not get stuck in local optima using both explorative and exploitative methods. In this case this method was used to the results of a new solution were based on MSE improvement, but initial, worse solutions could also be accepted, depending on the parameters set for SA.

Linear regression was the supervised machine learning technique selected for this assignment due to its direct and simple output in defining the relationship between the two variables and ability to convey the significance of incremental changes. Where the independent variable was the merged data of for the column 'Infants exclusively breastfed for the first six months of life (%) and the dependent variable is merged data under column category 'Under-five mortality rate (per 1000 live births) (SDG 3.2.1) for both sexes. Based on the datasets presented the low number and uncomplex variables presented can be simply analyzed using this model. This is ideal as the results and relationships must be conveyed simply yet effectively to the grant committee.

The evaluation technique of the linear regression model was evaluated using the MSE which measured the average squared difference between predicted and actual mortality rates, and R-squared values. The MSE value indicates the accuracy of the supervised learning method to predict the results, the higher the value more indicative it is of the variability in mortality rates. SA was determined successful in its evaluation based on its ability to align and minimize the data loss while creating a unified and consistent dataset for the machine learning analysis.

Algorithmic parameters set for SA consisted of setting the initial temperature at 10.0, with a cooling rate of 0.99 a gradual cooling rate that will narrow the selection and focus and with the number of iterations, perturbation step size was set to 0.1, to ensure changes were handled incrementally. The number of iterations for the algorithm was set to 10,000 providing sufficient time for the algorithm to converge without high computational expenses. There were no specific standards for the linear regression, no regularization was applied because overfitting was not a concern due to the noncomplexed dataset [16].

Experimental Results and Analysis

The results displayed a correlation between the two variables in that an increase in breastfeeding rates corresponds to a decrease in under-five mortality rates. Where the independent variable was the merged data of for the column 'Infants exclusively breastfed for the first six months of life (%) and the dependent variable is merged data under column category 'Under-five mortality rate (per 1000 live births) (SDG 3.2.1) for both sexes. More specifically increasing breastfeeding rates by 1% is associated with a decrease in child mortality by approximately 0.495 deaths per 1,000 live births. This was derived from the output of the from the linear regression model where we noted that the intercept value was 84.51722157, the coefficient was -0.49544432 and the best MSE from all iterations was 2503.63137980853.

These results can be attributed to the SA search algorithm and the linear regression supervised learning techniques that were used. The SA algorithm allowed for an exploration and exploitation approach, avoiding being trapped in local minima, which proved helpful in searching through different weight configurations and converging on a solution that minimized MSE [17].The gradual cooling of the algorithm allowed for thorough exploration in the early stages, with fine-tuning toward the end, ensuring that the model did not become trapped in suboptimal solutions.

Linear regression was successful in displaying the anticipated positive correlating relationship between the two variables where the relationship was expected to be linear, this method provided a high accurate analysis where results were direct and easy to interpret [18]. Linear regression provided the insight needed between breastfeeding rates and child mortality but was limited in the sense that these two variables are only able to explore linear relationships as expressed through the high MSE value. High MSE indicates that it could be beneficial to include more variables and in addition to these variables more robust machine learning systems as higher complex variable correlation would need to be actualized and predicted.

Through these the results the NGO can develop a strong argument for their grant application funding backed by concrete evidence. The data clearly shows the impact breastfeeding has on mortality rates, which can also be supported by external scientific literature as well. Furthermore, the high MSE value opens the discussion for introducing more robust data collecting methods for current and different variables that can be anticipated to be further influencing mortality rates. Findings in the literature show breastfeeding is an important factor to consider when discussing child mortality rates, malnutrition, infections, and other preventable causes [19]. Research also indicates other variables can further determine what other factors can influence child mortality rate

and can be focused upon to improve child survival rates. This holds particularly true for areas such as Sub Saharan Africa, where the inclusion of breastfeeding has a profound impact in improving under 5 child mortality rate overcoming high rates of malnutrition and preventable diseases [20] [21].

Future endeavours should express incorporating other machine learning models, other than linear regression, to handle multi-variable and complex data for accurate and precise conclusions. Overall, these results are easily translatable and will also open avenues for many potential future endeavours for more complex or comprehensive data collection for another global effort to understand the other variables that may be impacting child mortality rates and not just breastfeeding

Conclusion

This report detailed the application of AI search and machine learning methods to analyze the statistical data provided by the WHO on breastfeeding rates and child mortality. The data was cleaned, organized and merged, afterwards, SA was used searching through different weight configurations and converging on a solution that minimized MSE. The processed data was then inputted into the supervised machine learning model to predict the association between breastfeeding rates and child mortality rates, displaying a significant correlation between the two variables.

The results show that there is an evident coorelation between the two variables in which when a 1% increase in breastfeeding rates decreases child mortality rates by 0.495 deaths per 1,000 live births. This data further strengthens the argument of the NGO for their application process and through the development of their promotion programs. However, due to the high MSE value of 2503.6313798085, results suggest further analysis should be taken, with the inclusion of other variables such as healthcare access and economic situation, to further improve the accuracy of the predictions for the variables.

The limitation is the focus on a single variable, breastfeeding, for mortality rate prediction, as this oversimplifies the prediction analysis. The complexity of child mortality involves multiple variables, and linear regression may attribute to inaccurate prediction outcomes. Future efforts should focus on comprehensive optimization and machine learning methods to confidently predict trends.

Future recommendations should focus on collecting and integrating all data and other variables to as well as introducing complex machine learning algorithms in addition to these variables such as neural networks and decision trees. Overall, this would better capture the true relationship that would otherwise not work in a linear relationship, providing a wholistic perspective of the interconnecting factors impacting child mortality.

References

[1] N. Schwalbe and B. Wahl, "Artificial intelligence and the future of global health," *The Lancet*, vol. 395, no. 10236, pp. 1579–1586, May 2020, doi: 10.1016/S0140-6736(20)30226-9.

[2] F. Jiang *et al.*, "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, p. 230, Jun. 2017, doi: 10.1136/svn-2017-000101.

[3] A. M. Prentice, "Breastfeeding in the Modern World," *Annals of Nutrition and Metabolism*, vol. 78, no. Suppl. 2, pp. 29–38, Jun. 2022, doi: 10.1159/000524354.

[4] K. North, M. Gao, G. Allen, and A. C. Lee, "Breastfeeding in a Global Context: Epidemiology, Impact, and Future Directions," *Clinical Therapeutics*, vol. 44, no. 2, pp. 228–244, Feb. 2022, doi: 10.1016/j.clinthera.2021.11.017.

[5] G. U. Orlu *et al.*, "A Systematic Review of Literature on Sustaining Decision-Making in Healthcare Organizations Amid Imperfect Information in the Big Data Era," *Sustainability*, vol. 15, no. 21, Art. no. 21, Jan. 2023, doi: 10.3390/su152115476.

[6] Z. Kang, Y. Guan, J. Wang, and P. Chen, "Research on Genetic Algorithm Optimization with Fusion Tabu Search Strategy and Its Application in Solving Three-Dimensional Packing Problems," *Symmetry*, vol. 16, no. 4, Art. no. 4, Apr. 2024, doi: 10.3390/sym16040449.

[7] K. Wang, "Optimized ensemble deep learning for predictive analysis of student achievement," *PLOS ONE*, vol. 19, no. 8, p. e0309141, Aug. 2024, doi: 10.1371/journal.pone.0309141.

[8] Z. Ertem, D. Raymond, and L. A. Meyers, "Optimal multi-source forecasting of seasonal influenza," *PLOS Computational Biology*, vol. 14, no. 9, p. e1006236, Sep. 2018, doi: 10.1371/journal.pcbi.1006236.

[9] A. Sapre and S. Vartak, "Scientific Computing and Data Analysis using NumPy and Pandas," vol. 07, no. 12, 2020.

[10] M. Arifuzzaman, M. R. Hasan, T. J. Toma, S. B. Hassan, and A. K. Paul, "An Advanced Decision Tree-Based Deep Neural Network in Nonlinear Data Classification," *Technologies*, vol. 11, no. 1, Art. no. 1, Feb. 2023, doi: 10.3390/technologies11010024.

[11] A. Engelhart *et al.*, "Sustainability of breastfeeding interventions to reduce child mortality rates in low, middle-income countries: A systematic review of randomized controlled trials," *Front Health Serv*, vol. 2, p. 889390, Aug. 2022, doi: 10.3389/frhs.2022.889390.

[12] R. E. Azuine, J. Murray, N. Alsafi, and G. K. Singh, "Exclusive Breastfeeding and Under-Five Mortality, 2006-2014: A Cross-National Analysis of 57 Low- and-Middle Income Countries," *Int J MCH AIDS*, vol. 4, no. 1, pp. 13–21, Jul. 2015, doi: 10.21106/ijma.52.

[13] R. Pérez-Escamilla *et al.*, "Breastfeeding: crucially important, but increasingly challenged in a market-driven world," *The Lancet*, vol. 401, no. 10375, pp. 472–485, Feb. 2023, doi: 10.1016/S0140-6736(22)01932-8.

[14] "Pandas Tutorial." Accessed: Oct. 20, 2024. [Online]. Available: https://www.w3schools.com/python/pandas/default.asp

[15] "NumPy Tutorial." Accessed: Oct. 20, 2024. [Online]. Available: https://www.w3schools.com/python/numpy/default.asp

[16] C. Toh and J. P. Brody, “Applications of Machine Learning in Healthcare,” in *Smart Manufacturing - When Artificial Intelligence Meets the Internet of Things*, T. Yen Kheng, Ed., IntechOpen, 2021. doi: 10.5772/intechopen.92297.

[17] D. Delahaye, S. Chaimatanan, and M. Mongeau, “Simulated Annealing: From Basics to Applications,” in *Handbook of Metaheuristics*, vol. 272, M. Gendreau and J.-Y. Potvin, Eds., in International Series in Operations Research & Management Science, vol. 272. , Cham: Springer International Publishing, 2019, pp. 1–35. doi: 10.1007/978-3-319-91086-4_1.

[18] Q. An, S. Rahman, J. Zhou, and J. J. Kang, “A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges,” *Sensors (Basel, Switzerland)*, vol. 23, no. 9, p. 4178, Apr. 2023, doi: 10.3390/s23094178.

[19] M. J. Sankar *et al.*, “Optimal breastfeeding practices and infant and child mortality: a systematic review and meta-analysis,” *Acta Paediatrica*, vol. 104, no. S467, pp. 3–13, 2015, doi: 10.1111/apa.13147.

[20] C. E. Pretorius, H. Asare, H. S. Kruger, J. Genuneit, L. P. Siziba, and C. Ricci, “Exclusive Breastfeeding, Child Mortality, and Economic Cost in Sub-Saharan Africa,” *Pediatrics*, vol. 147, no. 3, p. e2020030643, Mar. 2021, doi: 10.1542/peds.2020-030643.

[21] G. Scarpa *et al.*, “Socio-economic and environmental factors affecting breastfeeding and complementary feeding practices among Batwa and Bakiga communities in south-western Uganda,” *PLOS Global Public Health*, vol. 2, no. 3, p. e0000144, Mar. 2022, doi: 10.1371/journal.pgph.0000144.