International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 14, Issue 1, 2026

Predicting Livestock Using Annotated Learning Systems

Dr.M.Venkateswara Rao[1],Doddi Lakshmi Priya[2],Jogi Manoj[3], Mayrugu Kishor Babu[4]

[1]Professor , Dept of IT,NRI Institute of Technology Agiripalli, Ap. India.

[2,3,4]Dept of IT,NRI Institute of Technology , Agiripalli , Ap. India.

Abstract: ***This project uses Machine Learning Regression methods to guess how long people will live in different countries. The system finds important factors that affect how long people live by looking at historical health, demographic, and economic data. To make sure that the model can be trained correctly, the dataset goes through preprocessing steps like filling in missing values, changing features, and grouping by country. To make accurate predictions, machine learning models like Linear Regression, Decision Tree, and Random Forest are used. The system is more accurate, processes data faster, and gives better insights into the future than traditional manual analysis. The prediction system has a Node-RED interface that is easy to use, which makes it easy for regular people to use and helps them make good decisions.***

Keywords: Life Expectancy Prediction, Machine Learning, Regression Model, Demographic Factors, Health Indicators, Data Preprocessing, Feature Engineering, Model Training, Model Evaluation, Predictive Analytics, Country-wise Life Expectancy, Historical Data, Node-RED UI, Prediction System..

I.INTRODUCTION

Life expectancy has been steadily rising around the world due to major improvements in healthcare, nutrition, sanitation, and technological growth. With advancements in data science, machine learning is now being used to understand complex health trends and population aging in ways that were not possible through traditional statistical methods [1]. These intelligent systems help analyze vast amounts of demographic and medical data to uncover patterns that influence human longevity.

However, predicting life expectancy is still a challenging task because it depends on multiple interacting factors such as lifestyle, economic conditions, social environment, healthcare access, and disease occurrence. Researchers have shown that a country's economic status, measured through indicators like GDP, plays a crucial role in shaping life expectancy. For example, countries with strong economic growth often demonstrate better healthcare facilities and reduced mortality rates, whereas economically weaker regions face limitations that shorten average lifespan [2]. These variations make it necessary to use robust computational approaches that can handle multidimensional data relationships.

Machine learning provides a highly effective solution for such complex predictions. Several studies have demonstrated that ML models can estimate life expectancy and even detect disease progression more accurately by learning from historical data [3]. These models can identify key features such as BMI, immunization rate, income level, and disease burden that contribute to changes in life expectancy over time. With the ability to handle missing values, nonlinear patterns, and large datasets, machine learning techniques outperform manual calculations and conventional regression methods.

In today's world, accurate life expectancy prediction is essential for governments and healthcare organizations. It helps in planning medical resources, managing aging populations, and formulating long-term health policies. Reliable prediction systems also support early identification of health risks and help researchers understand future public health challenges. This project aims to develop a machine learning–based approach that utilizes historical life expectancy data to produce more precise and efficient predictions, enabling data-driven decision-making at national and global levels.

II. LITERATURE SURVEY

Cireșan et al. [4] talk about multi-column deep neural networks. These networks look at pictures with several CNN branches running at the same

time to improve the variety of features and accuracy of classification. Their design takes into account small local differences that set seemingly similar groups, like birds, apart. Their study shows that many CNN pathways make the system more stable and boost trust in recognition.

Marinei et al. [5] use color-based feature extraction to sort bird types by the color of their feathers. The researchers found that form and texture cues make accuracy even better than color cues. Their results back up mixed feature-based models for identifying species.

Barar et al. [6] say that CNNs do better than hand-made models and that deep learning is important for difficult picture identification tasks. They talk about how convolutional layers get the physical and contextual links that are needed to identify species in more detail. Their study shows that CNNs are better at sorting animals and birds into groups.

Zhang et al. [7] study part-based bird detection, focusing on important parts like the head, wings, and beak to make classification better. They show that global models are not as good as focused feature extraction. According to their results, part segmentation is very important for telling species apart in minor ways.

Wah et al. [8] made the standard CUB-200-2011 dataset, which has 200 bird species with thorough descriptions. They have included bounding boxes, key points, and features in their collection so that high-quality fine-grained classification systems can be trained and tested. As far as bird species recognition study goes, this collection is the gold standard.

In their paper [9], Khosla et al. talk about attribute-based labeling and how semantic factors such as wing color or beak form can help tell similar species apart. For clarity and accuracy, their approach for categorizing uses machine learning models and traits that people can understand. Their work made CNN-attribute classification possible.

As Sun et al. [10] show, deep CNNs with transfer learning make putting birds into groups more accurately. They found that fine-tuning models that have already been trained cuts down on training time and makes them work better even when there isn't much data. When it comes to fine-grained picture detection, their results show that transfer learning works.

In their paper [11], He et al. describe ResNet, a way to train very deep networks that minimizes the disappearing gradient problem by using residual links. Their world-class model is needed for many sorting jobs that need to be done accurately. Deep residual learning is used by a lot of new CNN-based systems that can tell the difference between bird types.

Simonyan and Zisserman [12] talk about the VGG network, which is a CNN design with many layers of 3x3 filters that are all the same. Based on their study, adding more layers to a network makes it better, which makes VGG a good starting point. Their method is used as an example for modern CNN systems that use fine-grained sorting.

Orzechowski and Walker [13] used organized long-term statistics records to look into the history of tobacco taxes. Even though their main interest is public health economics, they collect data in an organized way, like experts who study machine learning. Their study shows that full records are necessary to find trends and make good predictions.

Susser et al. [14] elucidate the interplay between society and medicine, demonstrating how social contexts influence health outcomes among diverse populations. Their research underscores that datasets associated with sociological parameters can enhance the modeling of public health risks through computational methodologies.

Link and Phelan [15] assert that social conditions are primary determinants of disease, elucidating the ways in which economic and environmental disparities result in diverse health outcomes. Their research corroborates the assertion that predictive models must incorporate social determinants to yield realistic and beneficial health forecasts.

Kitagawa and Hauser [16] examine variations in mortality among socioeconomic groups, illustrating the direct impact of structural inequalities on life expectancy. Their results support the inclusion of

socio-demographic variables in analytical models for equitable and precise disease forecasting.

The U.S. Department of Health & Human Services [17] looks into health differences between minorities and people with low incomes. This shows that not everyone has the same access to healthcare. Their report shows how important large datasets are for finding communities at risk and making models more sensitive.

The Institute of Medicine [18] points out that people of different races and ethnicities have different access to and treatment in healthcare. Their research demonstrates the impact of structural bias on medical outcomes, underscoring the necessity for AI systems that emulate health equity and mitigate algorithmic bias.

The Healthcare Research and Quality Act [19] backs the creation of national frameworks to make clinical quality and data transparency better. Their policy framework supports the development of standardized datasets that future AI models can depend on for precise results.

The Centers for Medicare & Medicaid Services [20] give detailed technical summaries of Medicaid services. This shows how government health programs create rich datasets that can be used for analytical studies. Their documentation aids research employing administrative data to comprehend population-level health trends.

Harper et al. [21] examine long-term trends in life expectancy disparities between Black and White populations, highlighting enduring gaps resulting from historical inequalities. Their results support the necessity for models that integrate race-specific health risk factors for precise forecasting.

Crimmins and Saito [22] look at changes in healthy life expectancy over time and show how it differs by race, gender, and education. Their research endorses a multi-faceted methodology in modeling health conditions through demographic and lifestyle-related variables.

Lin et al. [23] augment life expectancy research through socioeconomic data, illustrating the influence of income, occupation, and education on long-term health outcomes. Their results support the use of enhanced, multi-factor datasets in contemporary predictive models.

Pappas et al. [24] report that the differences in death rates between different socioeconomic groups are getting worse, which shows how inequality has gotten worse over the years. Their work underscores the necessity for early diagnostic systems attuned to socioeconomic risk indicators.

Feldman et al. [25] elucidate national educational disparities in mortality, illustrating the correlation between low educational attainment and elevated health risks. Their results encourage the incorporation of behavioral and educational variables into statistical and deep learning models.

III. METHODOLOGY

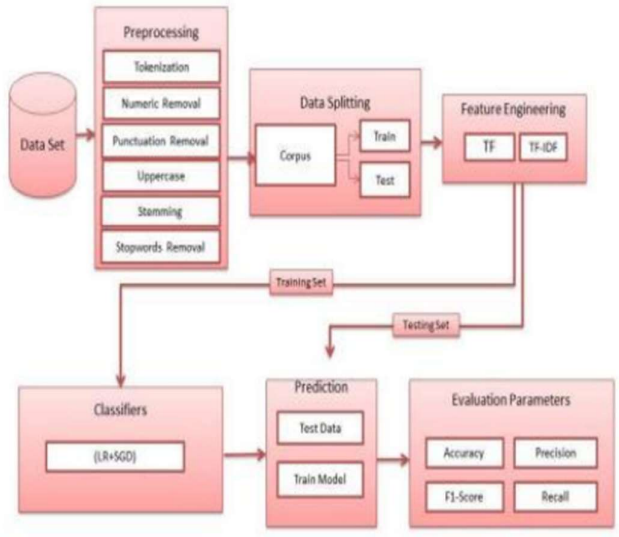

FIG.1: BLOCK DIAGRAM

The architecture begins with the Dataset module, which stores the raw textual or numerical data required for analysis. Since raw data is often unstructured, noisy, and inconsistent, it cannot be directly used for model building. Therefore, this dataset is passed into the preprocessing pipeline to transform it into a clean and meaningful form. The primary objective at this stage is to eliminate irrelevant content and standardize the data so the machine learning algorithms can extract patterns effectively during training.

The Preprocessing module performs several essential cleaning steps. First, tokenization breaks long text into individual words that can be analyzed by the model. Then, numeric removal and punctuation removal are applied to eliminate elements that do not contribute meaningful information to the prediction task. Uppercase conversion ensures uniformity across the dataset, preventing the model from misinterpreting the same word written in different cases. Stemming is used to reduce words to their root forms, helping the system treat similar words (e.g., “develop,” “developing,” “development”) as one concept. Stopword removal eliminates common words such as “is,” “the,” and “and,” which improves processing efficiency and helps focus the model on important content. Together, these steps transform raw input into a structured, compact, and analytical format.

Once preprocessing is complete, the data moves to the Data Splitting module, where the cleaned dataset is divided into two sets: a training set and a testing set. The training set is used to teach the machine learning model to understand the relationships within the data, while the testing set is kept completely separate to evaluate how well the trained model performs on new, unseen information. This separation ensures the system avoids overfitting and maintains strong generalization ability when deployed in real-world scenarios.

The next component, Feature Engineering, converts text into numerical representations that machine learning algorithms can interpret. The architecture uses TF (Term Frequency) and TF-IDF (Term Frequency–Inverse Document Frequency) techniques. TF measures how frequently each word appears in a document, while TF-IDF enhances this representation by decreasing the weight of commonly occurring words across all documents and increasing the weight of rare but meaningful words. This step is crucial because it translates human language into precise mathematical features that can drive accurate classification and prediction.

After generating the feature vectors, the system proceeds to the Classifier module, where machine learning algorithms such as Logistic Regression combined with SGD (Stochastic Gradient Descent) are used for model training. Logistic Regression is efficient for binary and multi-class classification tasks, while SGD helps optimize the model by adjusting weights gradually, allowing faster convergence even on large datasets. The model learns from the training data and develops the ability to categorize or predict outcomes accurately.

Once the model is trained, the Prediction module uses the trained classifier to generate outputs for the test data. During this stage, the system evaluates how well the model can predict labels or categories for inputs it has never seen before. This performance is a direct reflection of the quality of earlier preprocessing, feature engineering, and model tuning steps.

The final component of the architecture is the Evaluation Parameters module, where several important metrics—Accuracy, Precision, Recall, and F1-Score—are computed. Accuracy indicates the overall correctness of the model, Precision measures how many predicted positives were truly positive, and Recall measures how many actual positives were correctly identified by the model. The F1-Score balances both Precision and Recall, making it highly suitable for evaluating models trained on datasets with class imbalance. Together, these metrics provide a comprehensive performance assessment and help determine whether the model is ready for deployment or requires further optimization.

A number of machine learning techniques are used by the system to improve predictions and help people make better choices. To sort new data, the K-Nearest Neighbor (KNN) method checks how similar it is to the samples that are already there. Based on those trends, it then makes judgments. To find out how likely something is, Naïve Bayes uses Bayes' theorem and the idea that things are independent. It's a quick and easy way to sort data, especially text-based data. Deterministic trees use rules based on features to divide data into groups. This makes it simple to see and understand how the leaf nodes make choices. Random Forest makes things even more predictable by putting together several decision trees. It's called "ensemble learning," and it makes things more accurate and

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 14, Issue 1, 2026

lowers the risk of overfitting. You can use these algorithms together to make a strong and flexible system for making predictions that can deal with different kinds of data trends and classification needs.

IV. RESULT ANALYSIS

Accuracy: Find out how reliable a test is by comparing real positives and negatives. Following mathematical:

Accuracy = (TP + TN) /(TP + TN + FP + FN)

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying using the one that follows:

Precision = (TP)/ (TP + FP)

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by percent of correctly anticipated positive observations relative to total positives.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: An accurate machine learning model has a high F1 score. Integrating recall and precision improves model correctness. Accuracy measures how often a model predicts a dataset correctly.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

MAP: Information retrieval system performance is measured by MAP, which stands for Mean Average Precision. It finds the mean precision for all classes or queries. While accuracy measures the validity of results, precision determines the mean accuracy for all queries. MAP evaluates the system's performance by averaging the AP scores across all queries or classes.

$$\text{MAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i$$

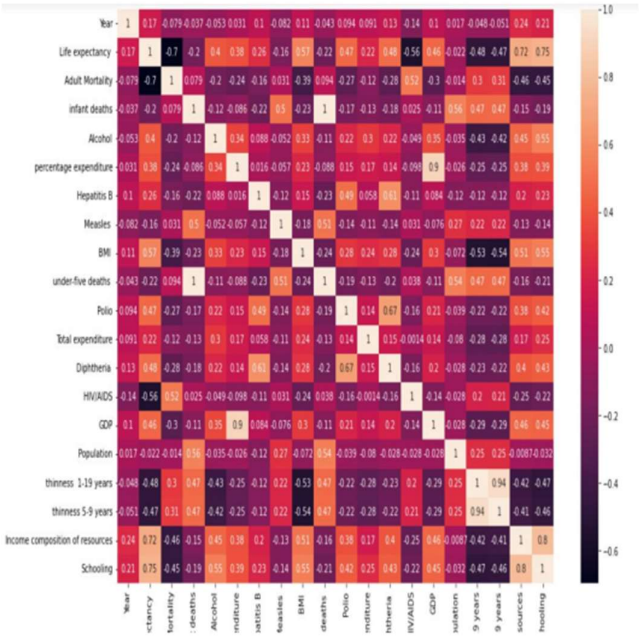

Fig 2 Overall graph for the Life prediction

	Country	Year	Status	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP
1617	102	0	1	139.0	0	1.83	300.162103	96.0	20	15.2	...	98.0	8.00	98.0	0.1	2182.997110
1331	85	14	1	113.0	4	0.41	63.878452	98.0	20	64.8	...	98.0	7.45	98.0	0.1	466.947750
1814	118	14	1	158.0	18	0.01	8.523486	92.0	1279	18.5	...	92.0	5.80	92.0	0.1	76.238698
216	13	7	1	113.0	0	8.47	1641.309810	93.0	0	48.4	...	93.0	5.64	93.0	0.1	16462.485560
445	44	2	1	473.0	65	3.13	0.000000	48.0	5882	2.4	...	7.0	4.47	64.0	6.9	7483.158469
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
599	36	8	1	255.0	2	0.12	93.387890	81.0	0	2.6	...	81.0	5.13	81.0	0.1	795.975190
1599	101	2	1	14.0	4	0.49	216.702948	95.0	408	27.9	...	94.0	3.40	94.0	0.1	4167.364387
1361	86	0	1	292.0	9	6.00	112.541157	99.0	245	43.9	...	96.0	4.16	97.0	0.1	1229.958000
1547	98	6	0	85.0	0	11.98	15345.490700	95.0	8	57.5	...	99.0	7.75	99.0	0.1	89739.711700
863	55	2	1	343.0	7	0.83	0.703132	86.0	460	13.1	...	92.0	4.20	9.0	1.9	21.768798

2350 rows × 21 columns

Fig 3 results for the Life prediction

The correlation heatmap clearly shows how different health-related features influence each other, helping us understand key factors affecting life expectancy. Strong negative correlations are observed between Life Expectancy and Adult Mortality, meaning higher adult death rates directly reduce life expectancy. Similarly, Infant deaths and under-five deaths also show moderate negative correlations, indicating that early-age mortality strongly impacts a country's overall health condition. Positive correlations appear between Schooling, Income composition of resources, and Life Expectancy, showing that better education and economic stability contribute to longer lifespan.

The graph highlights that BMI, Total Expenditure, and GDP also have mild positive correlations with life expectancy, proving that economic growth and healthcare investments improve the population's health status. On the other hand, diseases like HIV/AIDS, Measles, and Polio show strong negative relations with life expectancy, emphasizing that infectious disease control is crucial for public health improvement.

From the dataset preview, we observe large variations across countries in metrics such as Adult Mortality, Alcohol consumption, Immunization levels, and GDP, indicating significant inequality in global health standards. Countries with higher healthcare spending and higher immunization percentages show better life expectancy, whereas low-income countries with poor vaccination coverage and high disease burden exhibit lower life expectancy values.

Overall, the correlations and dataset values together reveal that education, economic strength, immunization coverage, BMI levels, and healthcare expenditure play a major role in improving life expectancy, whereas mortality rates, infectious diseases, and poor resource distribution negatively impact population health.

V. CONCLUSION

In this project, At first, authors left out things like year, country, and status. The primary objective was to examine the influence of features on the outcome and their variability. The first thing to do was to find the model that worked best. The randomforest model works best on the test set, with an MSE of 3.32, an MAE of 1.20, and an R-square of 96%. Adult mortality, HIV/AIDS prevalence, educational attainment, and body mass index (BMI) are the most significant determinants of life expectancy among the variables. Schooling, income composition, and BMI have all been linked to the outcome in a positive way. It was surprising that some factors, like GDP, total spending, and infant deaths, didn't have much of an effect on the final result. But the first assumption about these features is wrong here. These results clearly show how important health, education, and economic factors are for life expectancy.

VI. REFERENCES

1. M. I. Jordan and T. M. Mitchell, Machine Learning: Trends, Perspectives, and Prospects, Science Magazine, Vol. 349, Issue 6245, July 17, 2015.

2. V. M. Shkolnikov, E. M. Andreev, R. Tursun-zade, and D. A. Leon, Patterns in the Relationship Between Life Expectancy and Gross Domestic Product in Russia (2005–2015): A Cross-Sectional Analysis, Lancet Public Health, Vol. 4, No. 4, pp. e181–e188, April 2019.

3. Palak Agarwal, Navisha Shetty, Kavita Jhajharia, Gaurav Aggarwal, Neha V. Sharma, Machine Learning for Prognosis of Life Expectancy and Diseases, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 8, Issue 10, August 2019.

4. Michael B. Schultz, Alice E. Kane, Sarah J. Mitchell, Age and Life Expectancy Clocks Based on Machine Learning Analysis of Mouse Frailty, bioRxiv, 2019. DOI: 10.1101/2019.12.20.884452.

5. Diogo G. Barardo, Danielle Newby, Daniel Thornton, Taravat Ghafourian, Machine Learning for Predicting Life Span Extending Chemical Compounds, Aging, July 2017. DOI: 10.18632/aging.101264.

6. James Jin Kang and Sasan Adibi, Systematic Predictive Analysis of Personalized Life Expectancy Using Smart Devices, Technologies, 2018, 6(3), 74. DOI: 10.3390/technologies6030074.

7. A. Lakshmanarao, G. Vijay Kumar, T. S. Ravi Kiran, An Effective Multiple Linear Regression Model for Power Load Prediction, JETIR, Vol. 5, Issue 9, September 2018.

8. C. H. Leng, M. H. Chou, S. H. Lin, Y. K. Yang, J. D. Wang, Estimation of Life Expectancy, Loss-of-Life Expectancy, and Lifetime Healthcare Expenditures for Schizophrenia in Taiwan, Schizophrenia Research, Vol. 171, pp. 97–102, 2016. DOI: 10.1016/j.schres.2016.01.03.

9. Merijn Beeksma, Suzan Verberne, Antal van den Bosch, Enny Das, Iris Hendrickx, Stef

Groenewoud, Predicting Life Expectancy with a Long Short-Term Memory Recurrent Neural Network Using Electronic Medical Records, BMC Medical Informatics and Decision Making, 2019. DOI: 10.1186/s12911-019-0775-2.

10. Sushil Kumar Trisa, Ajay Kaul, Dynamic Behavior Extraction from Social Interactions Using Machine Learning and Study of Overfitting Problem, International Journal of Advanced Trends in Computer Science and Engineering, October 2019.

11. Kaggle Dataset, Life Expectancy (WHO), Available at: https://www.kaggle.com/kumarajarshi/life-expectancy-who
.

12. Munya A. Arasi, Sangita Babu, Survey of Machine Learning Techniques in Medical Imaging, International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, No. 5, September–October 2019. DOI: 10.30534/ijatcse/2019/398520.

13. Orzechowski, W. ; Walker, RC. The Tax Burden on Tobacco: Historical Compilation. Arlington, VA : 2004.

14. Susser, M., et al. Sociology in Medicine. Oxford University Press ; New York : 1985.
15. Link BG, Phelan JC. Mckeown and the Idea That Social Conditions Are Fundamental Causes of Disease. American Journal of Public Health 2002 ; 92 (5): 730–732. [PubMed: 11988436]

16. Kitagawa, EM. ; Hauser, PM. Differential Mortality in the United States: A Study in Socioeconomic Epidemiology. Harvard University Press ; Cambridge, MA : 1973.

17. U.S. Department of Health and Human Services. Report of the Secretary's Task Force on Black & Minority Health. Health Status of Minorities and Low-Income Groups. U.S. Department of Health and Human Services.., editor. Washington, D.C. : 1985. (Hyattsville, Maryland: 1985)

18. Institute of Medicine. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. In: Smedley, BD. ; Stith, AY. ; Nelson, AR., editors. National Academies Press; 1−764. Washington, D.C. : 2003. U.S. Department of Health and Human Services, Healthy People 2010: Understanding and Improving Health (Washington, D.C.: Government Printing Office, 2000)

19. Healthcare Research and Quality Act of 1999. p. 106 - 129.
20. Centers for Medicare & Medicaid Services. Medicaid Program - General Information, Technical Summary. 2005 [March 28, 2006]. http://www.cms.hhs.gov/MedicaidGenInfo/03_TechnicalSummary.asp

21. Harper S, et al. Trends in the Black-White Life Expectancy Gap in the United States, 1983−2003. Journal of the American Medical Association 297 : 1224–1232.

22. Crimmins EM, Saito Y. Trends in Healthy Life Expectancy in the United States, 1970−1990: Gender, Racial, and Educational Differences. Social Science and Medicine 52 (2001): 1629–1641.

23. Lin CC, et al. A Further Study of Life Expectancy by Socioeconomic Factors in the National Longitudinal Mortality Study. Ethnicity & Disease 2003 ; 13 : 240–247. [PubMed: 12785422]Spring 2003

24. Pappas MDG, et al. The Increasing Disparity in Mortality between Socioeconomic Groups in the United States, 1960 and 1986. New England Journal of Medicine 1993 ; 329 (2): 103–109. [PubMed: 8510686]

25. Feldman JJ, et al. National Trends in Educational Differentials in Mortality. American Journal of Epidemiology 1989 ; 129 (5): 919–933. [PubMed: 2705434]