Research Methodology for predicting Life Expectancy using Machine Learning

Olgerta Idrizi†
Business Informatics Department
Mediterranean University
Tirana, Albania
idriziolgerta@gmail.com

Miranda Harizaj
Department of Automation
Polytechnic University of Tirana
Tirana, Albania
miranda.harizaj@fie.edu.al

ABSTRACT

Life expectancy (LE) models have an immense effect on the social and financial structures of many countries around the world. For this reason, we should be focused on creating technics, algorithms and methodologies to predict LE. This paper is a review to find some of efficient algorithms and methodologies to predict it. Artificial intelligence techniques and machine learning algorithms can analyze large amounts of data so they can be used to predict life expectancy.

Life expectancy depends on various variables like mortality rate, life expectancy from the past years, alcohol consumption rate, infant death, covid mortality and other illness that affect life expectancy rate. In the last years SARS-Covid has had a big impact on mortality rate and also on life expectancy. For this reason, we need to find a way to predict LE after Covid with the best model.

This analysis is a review of different machine learning algorithms that have achieved better accuracy based on pertinent features of the datasets. Based on the analyze of simple linear regression and k-nearest neighbors (KNN) algorithms, machine learning techniques were applied in order to develop an accurate prediction solution for life expectancy with the effect COVID-19. By comparing these machine learning algorithms, it is analyzed which among them is more accurate to predict life expectancy based on other variables. On the end we will discuss the results and conclusion for the better methodology which fit.

KEYWORDS

Life expectancy, Methodology, Machine Learning Algorithms, Linear Regression, KNN, etc.

1. Introduction

Everything in this world has a limited life expectancy. Humans also have a limited life span to survive. Life span prediction has a greater impact in the modern society because of food habits, different types of diseases, environmental conditions and other factors. After Covid 19, unfortunately, life expectancy was decreased. It has been the impact of the increase of mortality rate in this last years, because of Covid 19 pandemic.

Life expectancy is always defined statistically as the average number of years remaining at a given age. During the last years, life expectancy at birth has risen rapidly, due to many factors [1]. During the last century, statistics show the continuous increase in life expectancy at birth, resulted because of economic development and the improvement lifestyles, progresses in healthcare and medicine etc. There are a lot of factors that affect in life expectancy and mortality rate like gender, genetics, lifestyle, hygiene, access to health care, diet, exercise etc. Evidence-based studies indicate that longevity is based on two major factors, genetics, and lifestyle choices [2].

In this paper it is presented some of the machine learning methodologies that can help in predicting life span on the future, based on the data model. With the use of machine learning algorithms and data analytics can be prognosticated and examined the life span of the individual human being and can be used different classification algorithms for this prediction to accomplish higher accuracy.

2. Methodology

The forecasting is an important and different scenario from other statistics methodologies, so for this reason it is good to know the objectives and essential theory behind the problem and how it's going to be implemented practically.

The objective is to predict the result of the number of dependent variables in the comparison to number of independent variables. We can use various Machine Learning techniques for solving these problems. Some of the techniques will be discussed below.

2.1 Linear Regression

A correlation heat map is a graphical representation of correlation matrix representing the correlation between different variables and this helps understanding the linear dependencies of variables. The range of Correlation is (-1,1) and is calculated between two variables. Correlation value near to zero means the two variables are unrelated and close to 1 means the two variables are perfectly related.

As first step it is required adapt a supervised regression algorithm that fits the task requirements. Today there are a bunch of algorithms for regression tasks, and among them each has its pros and cons. The algorithm can result in superior outcomes compared to others but might deficiency in terms of explain ability. Even that, the deployment of such complex algorithms is not an easy task. There is a trade-off between accuracy, model complexity and model explain ability.

Linear Regression is a comparatively simple and explainable algorithm. Deployment of Linear Regression requires minimal efforts, but on the contrary, it lacks accuracy when the data is non-linear. Complex algorithms perform better on non-linear datasets, but then the model lacks explain ability [3].

LR is a regression algorithm with a linear approach and can be used to predict a continuous value of a given data point by simplifying the given data. The linear part indicates the linear approach for the generalization of data.

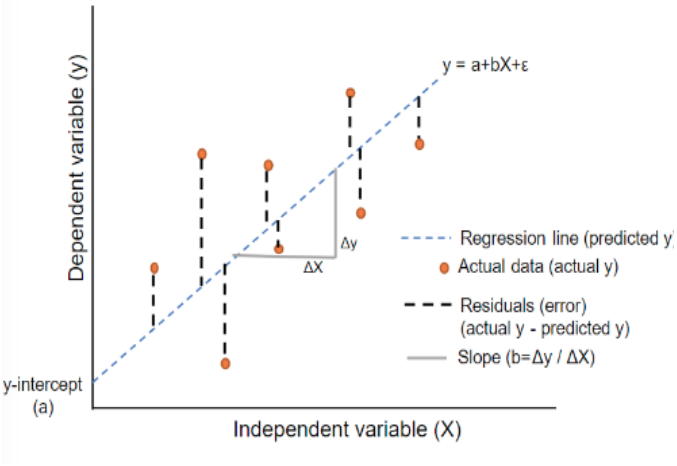

Figure 1. Linear regression [3]

In this algorithm it is predicted the dependent variable (y) using a given independent variable (x). A residual error is measured with the difference of the real value from the regression line that is the difference between a predicted value and the observed value.

If the coefficient of determination R-square came out to be closer to 1, this can indicate that the model optimally predicts the Life expectancies.

2.2. KNN Algorithm

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning method first developed by Evelyn Fix and Joseph Hodges in 1995, and later expanded by Thomas Cover. It is one of the simplest ML Algorithms and is used for classification and regression. In both cases, the input consists of the k closest training examples in a data set. The output depends on whether k-NN is used for classification or regression:

It is needed to determine which k data points out of training are closer to the data point in order to get a prediction for. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. It is supposed that the data are the lowing:

$$\begin{matrix} X_{11}, X_{12}, X_{13}, X_{1m}, Y_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ X_{n1}, X_{n2}, X_{n3}, X_{nm}, Y_n \end{matrix} \quad (1)$$

In the above array with n rows and m+1 columns, the first m columns are the attributes used to predict assuming that all attribute values x are numerical while the label values y are categorical. Then it is calculated the distance between data points. The Euclidean distance is a good choice for such a distance function if the data is numerical.

$$d(s, x_j) = \sqrt{(s_1 - X_{j1})^2 + \cdots .. (s_m - X_{jm})^2} \quad (2)$$

In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors. If k = 1, then the output is simply assigned to the value of that single nearest neighbor and is the reason why calculating training errors are useless.

The k-NN algorithm is a type of classification where the function is only approximated locally and all computation is accepted until function evaluation. If the features represent different physical units or come in much different scales then normalizing the training data can improve its accuracy dramatically, since this algorithm relies on distance for classification [4].

A useful methodology can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones for both, classification and regression.

A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

3. Results and Conclusion

In this paper it is analyzed how the artificial intelligence techniques and machine learning algorithms can predict life expectancy. By employing data through datasets, the correlation between attributes like mortality rate, life expectancy from the past years, alcohol consumption rate, Infant death, Covid mortality and other illness are monitored.

In this paper are reviewed different machine learning algorithms that have achieved better accuracy based on pertinent features of the datasets. Machine learning techniques were applied

in order to develop an accurate life expectancy after COVID-19 based on those variables.

Advantages:

Linear Regression

- It performs good for linearly separable data
- It is simple to implement, interpret and efficient to train
- It handles overfitting pretty well using dimensionally reduction techniques, regularization, and cross-validation
- extrapolation beyond a specific data set

KNN Algorithm

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large [3].

Disadvantages:

Linear Regression

- The assumption of linearity between dependent and independent variables
- It is often quite prone to noise and overfitting
- Linear regression is quite sensitive to outliers
- It is prone to multicollinearity

KNN Algorithm

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples [4].

The inclusion and dependency of these suggested features on life expectancy is still a matter of debate and a future part of research in this particular domain. Furthermore, the future enhancement can be made by using deep learning algorithm which may give better solution. Depending on the dataset, and other variables, the best adaptation is made.

REFERENCES

[1] Scholey, J., Aburto, J. M., Kashnitsky, I., Kniffka, M., Zhang, L., Jaadla, H., Dowd, J. & Kashyap, R. (2022). Life expectancy changes since COVID-19. Avaiable: https://www.nature.com/articles/s41562-022-01450-3

[2] Bali, V., Aggarwal, D., Singh, S., & Shukla, A. (2021). Life Expectancy: Prediction & Analysis using ML, ICRITO.

[3] Predicting Life Expectancy using Liner Regression. Avaiable: https://www.enjoyalgorithms.com/blog/life-expectancy-prediction-using-linear-regression

[4] K-Nearest Neighbor (KNN) Algorithm for Machine Learning. Available: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning