Research on Life Expectancy Prediction Based on Logistic Regression and KNN

Wentao Ji*

Department of Economics, University of Connecticut, Connecticut, USA

*Corresponding author: wentao.ji@uconn.edu

Abstract: The author checks the factors affecting life expectancy by reviewing the literature, and then displays the correlation graph to check the multicollinearity. Second, a training set (70%) and a test set (30%) are created from the dataset collected in this paper. The accuracy of their forecasts is then checked using two different ways—Logistic Regression and KNN before dropping the variable with high correlation with others and slight statistical significance. The accuracy for each model Logit (1), Logit (2), KNN (1) and KNN (2) is 0.8936,0.8723,0.8511 and 0.8723, respectively. The author's conclusions are as follows: (1) For Logistic Regression Prediction, a lack of information is a major factor that affects accuracy; (2) For KNN Prediction, removing one or more highly linked explanatory variables can improve prediction; (3) Overall, Logistic Regression Prediction has slightly higher accuracy than KNN. Perhaps this is due to the fact that KNN requires a bigger sample size to prevent misclassification, and that the best K are chosen based more on cross-validation experience than the sound statistical theory.

Keywords: Life expectancy; prediction; logistic regression; KNN.

1. Introduction

Life expectancy is always a heatedly debated topic in public because everyone (on the presumption of the hypothesis of rational man) is in pursuit of longevity. As for the factors that affect life expectancy, experts from a wide range of disciplines have provided us with many theories. When analyzing large amounts of data, the author may obtain the results which are out of expectations.

For starters, economic factor is a critical factor concerning life expectancy. From a macro perspective, the life expectancy of a region is closely related to economic development. Aisa and Pueyo pointed out that governmental investment in the field of health care can increase endogenous longevity [1]. The analysis conducted by Cervellati et al. also showed that during the period of rapid economic growth in the Western Countries in the 18th century, people's life expectancy also increased rapidly, and the mortality rate dropped significantly [2].

From a micro perspective, factors such as income and the socioeconomic status of the family also have important effects on human life. Breyer et al.'s research pointed out that health care expenditures per capita increase with aging population growing [3]. A PhD team led by Chetty found through computing US data from 2001 to 2014 that life expectancy increased with an increasing income level in general. But the association between those two varied in different fields [4]. Shaw proved in the research that life expectancy is strongly correlated with one's medical expenditure in the middle age. And those who spend more on health items can be expected to live longer [5]. Wilkinson obtained similar result that reasonable income distribution raises health condition [6]. To sum up, both for the countries and individuals, the increase both in income and health expenditure has an important impact on life expectancy.

Secondly, environment (geographical location) also has a huge impact on people's longevity. Hamidi found that urban sprawl may result in an increase in death toll in the US. That is because a widely distributed urban system tends to bring higher travelling speed, lower quality of life, fewer health care units, and less convenience of accessing health food [7]. Fabio and Marini's research showed that environment is positively related to life expectancy [8]. Arden calculated through a model and concluded that better air quality leads to a prominent increase in life expectancy [9]. When Araki et al. analyzed the factors affecting life expectancy in Japan, they found that as for men and

women aged from 20 to 40, the home at birth was the main factor affecting life expectancy. As for women aged 40, newborns and 20-year-old men as well as all 65-year-old people, the proportion of the elderly to young people has a significant impact on life expectancy [10]. Carr revealed that a person's early family circumstances and economic status can have an impact on their personality, emotions, health, and other aspects of their lives. That can subsequently have an impact on their quality of life and physical well-being [11].

Mirowsky indicated that life expectancy is 7 years longer with an extra year of education [12]. Marioni found that education is related to several health indicators, including high longevity attention based on epidemiological study on gene fragment [13]. Torssander showed that children's educational level can be a crucial factor determining one's health condition and life expectancy [14]. The research conducted by Meara et al. also found that life expectancy increases by 30% as education levels increases [15].

2. Methodology

2.1 Data Source

The data was obtained from Kaggle. There are 193 countries listed in the author's dataset from 2000 to 2015. Perhaps analyzing that data from the perspective of panel data is a better option. However, due to the author's specific analysis, the author will select the most recent year 2010 to set the dataset as cross-section data.

2.2 Variable Selection

First, the author will select the general government expenditure on health as a percentage of total government expenditure as the author's second independent variable. To be more precise, more people would be healed for if the government had been spending more in the national health system. The expenditures include training the doctors or nurses, creating more effective drugs, and so forth.
For the fourth independent variable, the author will use GDP per capita (Gross Domestic Product per capita (in USD)). GDP per capita can accurately reflect people's consumption abilities, whereas GDP indicators cannot. If a person has more wealth, he or she is more likely to take care of his or her own health condition by getting a medical check, spending more money on practicing the regimen, and so on, thereby increasing their own longevity. Concerning the variable 'status' (which includes both developed and developing countries), this variable is very closely related to GDP, so the author will choose to exclude it from the model analysis.

With respect to the income issue, another index called HDI (Human Development Index) can also well reflect the income composition of resources. Because that index (ranging from 0 to 1) can illustrate the ease with which people develop themselves through natural resources. From that perspective, it compensates for the drawbacks of the GDP per capita which ignores the quality of one's income. Accordingly, the author will take income composition of resources (ICR) as the author's fifth independent variable.

According to academics who have studied the subject, a greater level of education also increases a person's longevity. The variable schooling in the author's dataset is an excellent indication of educational level. Educated individuals are more likely to pay attention to basic health principles. For instance, such groups are more likely to eat healthier foods, exercise frequently and pay more attention to the latest health information. So, the author takes 'scho' as sixth variable.

Finally, when it comes to some serious diseases, once a person is infected with it, he or she is more likely to get into trouble. Observing the data set, the author will identify the most common disease—HIV as the author's third independent variable [16]. Furthermore, alcohol is a significant cause of decreasing human life expectancy. As a result, it is critical that the author will incorporate the alcohol addiction into his subsequent analysis. And 'Alc' can be as the first independent variable [17].

Because the author has already explained why he chose those independent variables, the author will list those variables (including abbreviations) and their corresponding explanations (Table 1) as follows.

Table 1. Data Description

Variables	Explanation
Life Expectancy (LE)	Life Expectancy in Age
Alcohol (Alc)	Alcohol Recorded Per Capital (15+) Consumption (in Litres of Pure Alcohol)
Total Expenditure (TE)	General Government Expenditure on Health as a Percentage of Total Government Expenditure
HIV/AIDS(HIV)	Death Per 1000 Live Birth/AIDS (0-4 years)
GDP (Per Capita)(GDPPC)	Gross Domestic Product Per Capita (in USD)
Income Composition of Resources(ICR)	Human Development Index inTerms of Income Composition of Resources (Index Ranging from 0 to 1)
Schooling (Scho)	Number of Years of Schooling(Years)

Table 2. Statistical Description

Statistic	le	alc	te	hiv	gdppc	icr	te
Min	36.30	0.01	0.92	0.10	8.38	0.00	4.50
1st Qu	63.80	1.31	4.09	0.10	687.96	0.52	10.55
Median	73.00	4.54	5.82	0.10	2187.79	0.71	12.80
Mean	70.25	4.94	6.01	1.36	7502.57	0.66	12.57
3rd Qu	76.20	7.92	7.91	0.50	5842.13	0.80	14.60
Max	89.00	14.97	11.87	21.60	87646.75	0.94	20.30

Looking closely at Table 2, the author notices that the icr minimum is 0.000. That indicates people in some nations or regions have relatively few natural resources and therefore their chances of developing themselves by exploiting these resources are slim. In sharp contrast, the maximum of icr is 0.9360. That means that people who are in easy-to-get-natural-resources countries have much easier access to leading a high-quality life.

Additionally, the range between gdppc is large, which indicates there is a yawning gap between the affluent and the destitute. Furthermore, le and gdppc need to be logarithmized because their absolute values are rather large.

2.3 Method

After theoretically and quantitatively describing the variables, the author does a correlation plot analysis by computing the correlation coefficients. The author can evaluate the highly associated variables and eliminate one or more of them based on the results of the correlation plot in order to avoid multicollinearity.

Next, a random selection of 70% of observations are used as training data and the remaining 30% are used as test data after computing the median of Life Expectancy (le) and setting the numbers 1 if le1 > Median (le) and 0 if le1< Median (le). Following that, R runs 'Logistic Model Regression' using the training data. Then, it is possible to examine the model's performance using test data. After that, to fill up the prediction with all 1 and to set the prediction with a probability of less than 0.50 as "0" can be accessible. The accuracy is the average of the quantity of comparable test and prediction data [18].

When implementing the Logistic Regression model, the author also places a high value on the statistical significance of each explanatory variable. At the same time, the author considers dropping

the independent variable(s) with the worst statistical significance (while also considering high-correlated factors and economic theories), runs the new "Logistic Model Regression," and then follows the next set of steps to see if the accuracy of the second implementation has been improved. As for KNN, a ten-fold cross-validation was utilized by the author. It divides into 10 sub-samples 70% of the training set for life expectancy and its related independent variables, which the author had previously chosen [19]. The remaining 9 samples are utilized for training, and one sub-sample is set aside as data for the model's validation. One cross-validation was performed ten times, once for each subsample, for a total of ten times on average.

The six explanatory factors that affect life expectancy are then regressed using the KNN method using this training set as the basis. At this point, it is possible to determine the K value that corresponds to the greatest accuracy. The remaining 30% of the test set is then utilized to forecast life expectancy based on the aforementioned results. In order to determine the accuracy of the KNN algorithm's predictions, the author can evaluate how closely the predicted values of life expectancy match the values in the life expectancy test set. The author then repeats the preceding procedures while using the explanatory variables from the second Logistic Regression to regress KNN, in order to determine whether the new accuracy of the final KNN algorithm has increased.

In a nutshell, the author will then thoroughly compare the accuracies of Logistic Regression prediction regarding life expectancy and KNN algorithm prediction before and after dropping some explanatory variable(s); compare the accuracies of Logistic Regression prediction and KNN algorithm before and after dropping some explanatory variable(s); and compare the accuracies of Logistic Regression prediction and KNN algorithm.

3. Results and Discussion

3.1 Variables Correlation Visualization

According to Fig. 1 (correlation plot), ICR and SCHO have a pretty high correlation coefficient of 0.79. One explanation for this outcome is that those who work in relatively easy-to-get-natural-resources and comfortable environments may be paid more, and those who are well-paid devote more time to the education of their children or themselves. However, if those two variables are one of the regressors, they may give rise to the collinearity. Simultaneously, checking the statistical significance of those two regressors is also crucial, and then whether one of them or two should be dropped can be pinned down.

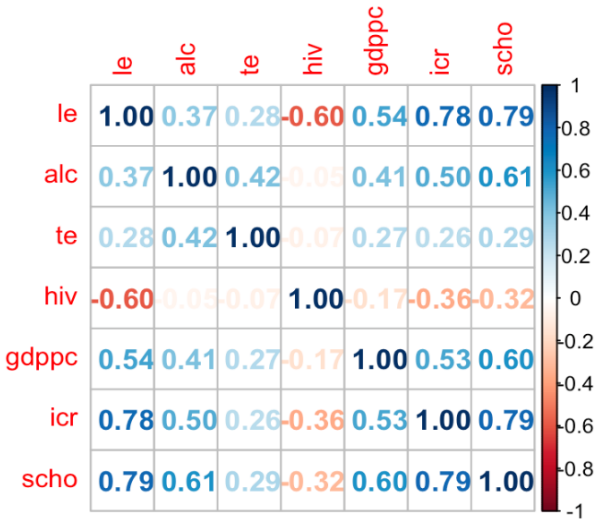

Fig. 1 Correlation Plot

In particular, the relationship between the AIDS variable 'hiv' and the life expectancy variable is negative, and the correlation coefficient's absolute value is higher than 0.5, which corresponds to the number of newborn deaths from AIDS per 1,000 live births. Human life expectancy has a significant negative association with that variable, which is consistent with health economics theory's predictions.

3.2 Logistic Regression Model Discussion

$$L_i = \ln(\frac{P_i}{1-P_i}) = \beta_0+\beta_1 \text{ALC}_i+\beta_2 \text{TE}_i+\beta_3 \text{HIV}_i+\beta_4 \text{GDPPC}_i+\beta_5 \text{ICR}_i+\beta_6 \text{SCHO}_i+u_i \quad (1)$$

Table 3. Logistic Regression (1) Deviance Residuals Statistic

Statistic	Deviance Residuals
Min	-2.3372
1Q	-0.1302
Median	0.0501
3Q	0.4184
Max	1.6357

With respect to the Table 3, the median of deviance residuals is 0.0501; the range of deviance residuals is 1.6357-(-2.3372)=3.9729.

Table 4. Logistic Regression (1) Coefficient Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-19.2805	5.3186	-3.6250	0.0003***
alc	-0.2469	0.1401	-1.7620	0.0780.
te	0.5398	0.2374	2.2740	0.0230*
hiv	-0.7456	1.4653	-0.5090	0.6109
gdppc	-0.1794	0.2613	-0.6870	0.4922
icr	48.8193	14.9269	3.2710	0.0011**
scho	-1.1652	0.5364	-2.1720	0.0298*

Sign.: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

As for Table 4, the estimate signals of gdppc and scho are negative respectively, which are out of our expectation. Especially for the gdppc, one possible explanation is that some nations clearcut a lot of agricultural areas and exploit them for high-value commercial lands, disregarding the limits of natural resources and the ecological environment's carrying capacity in order to rapidly build their own economies. At the same time, people's physical and mental health are ignored, and their life expectancy is shortening as a result of poor environmental circumstances and rising stress. In addition, the P value of Z-statistic for gdppc is 0.4922, higher than 0.05, indicating that the gdppc variable is not a good indicator, which perhaps includes other variables' information. Plus, the correlation coefficient between gdppc and scho is 0.6000, which is fairly high. Considering on those factors, gdppc can be dropped. As regards for the signal of scho, one possible explanation is that the longer a person attends school, the higher their educational level will be; after graduation, these highly educated persons have a propensity to select mentally taxing employment, which will reduce their longevity. Meanwhile, the P value of Z-statistic for scho is 0.0298, which is lower that 0.05, hence that variable can be retained. Moreover, although the estimated signal of hiv is correct, the P value of Z-statistic for hiv is 0.6109 that much higher than 0.05, manifesting that discard this statistically insignificant variable hiv is wise.

Table 5. Prediction Table Logistic Regression (1)

Prediction	0	1
0	20	4
1	1	22

According to table 5, when the author has a close look at that above table where there is also a list of rates that are often computed from a confusion matrix for a binary classifier, the author has 20/(20+1)=0.9524 opportunity to predict that the life expectancy is lower than the median of life expectancy--73 successfully. In other words, these people are predicted to be relatively non-long-lived individuals. Meanwhile, the author has 22/(4+22)=0.8462 opportunity to predict that the life expectancy is longer than the median of life expectancy--73 successfully. Overall, the accuracy of predicting life expectancy for the first logistic regression model is (20+22)/(20+4+1+22)=0.8936.

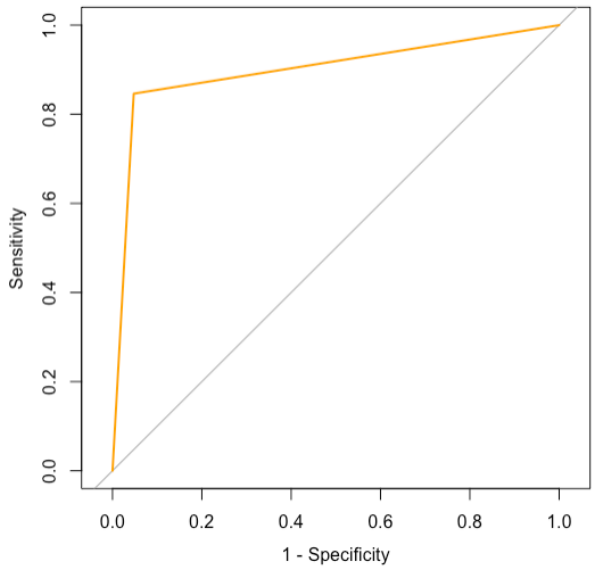

Fig. 2 ROC Curve: Logistic Regression (1)

The area under the ROC curve, as shown in Fig. 2, is 0.8993, which is higher than 0.5. This indicates that the prediction power is reasonably solid.

$$L_i=\ln(\frac{P_i}{1-P_i})=\beta_0+\beta_1 \text{ALC}_i+\beta_2 \text{TE}_i+ \ +\beta_3 \text{ICR}_i+\beta_4 \text{SCHO}_i+\mu_i \quad (2)$$

Table 6. Logistic Regression (2) Deviance Residuals Statistic

Statistic	Deviance Residuals
Min	-2.0242
1Q	-0.1112
Median	0.0000
3Q	0.2300
Max	1.7456

With respect to the Table 6, the median of deviance residuals is 0.0000 which is lower than 0.0501 in Table 3. The range of deviance residuals is 1.7456-(-2.0242)=3.7698, which is lower than 3.9729 in Table 3, indicating that the regression fitness is better than the first logistic regression model.

Table 7. Logistic Regression (1) Coefficient Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-20.8132	5.1201	-4.0650	4.8e-05***
alc	-0.2446	0.1319	-1.8550	0.0636.
te	0.4684	0.2140	2.1890	0.0286*
icr	49.4918	14.3519	3.4480	0.0006***
scho	-1.1700	0.5002	-2.3390	0.0193*

As shown in Table 7, all of the Pr(>|z|)s for explanatory variables are statically significant. The P value corresponding to its Z statistic is very low, which means that when other variables remain unchanged, if icr increases by 1 unit, the estimated logistic increase on average is about 49.4918 units, indicating that the two have a positive relationship.

Table 8. Prediction Table for Logistic Regression (2)

Prediction	0	1
0	20	5
1	1	21

According to table 8, if the author carefully examines the above table, which also includes a list of rates frequently calculated from a confusion matrix for a binary classifier, the author has a 20/(20+1) = 0.9524 chance of correctly predicting that the life expectancy is lower than the median life expectancy of 73. These individuals are therefore expected to live relatively short lives. The author has a chance of 21/(5+21) = 0.8077 of correctly predicting that the life expectancy will be greater than the median life expectancy of 73. The overall accuracy of the second logistic regression model in predicting life expectancy is (20+21)/(20+1+5+21) = 0.8723.

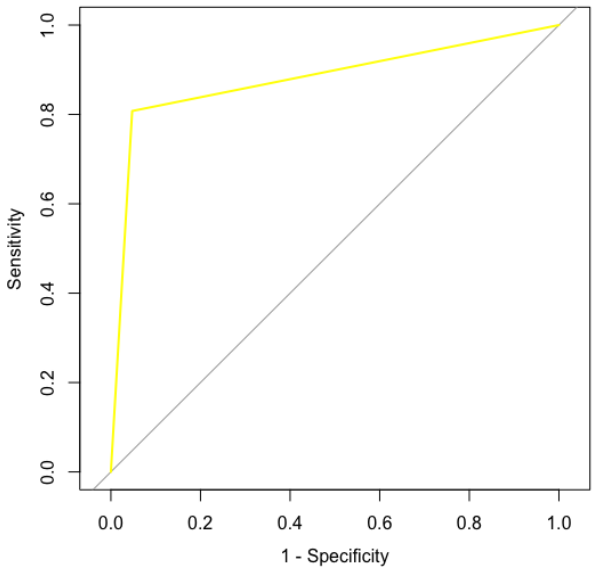

Fig. 3 ROC Curve: Logistic Regression (2)

According to Fig. 3, the area under the ROC curve is 0.8800, which is greater than 0.5. This shows that the prediction power is comparatively strong.

3.3 KNN Discussion

As displayed in Fig. 4, the cross-validation accuracy for neighbors K-5 is 0.77777, for neighbors K-7 it is 0.7786, and for neighbors K-9 it is 0.7795. Consequently, 9 is chosen as the ideal k in the model.

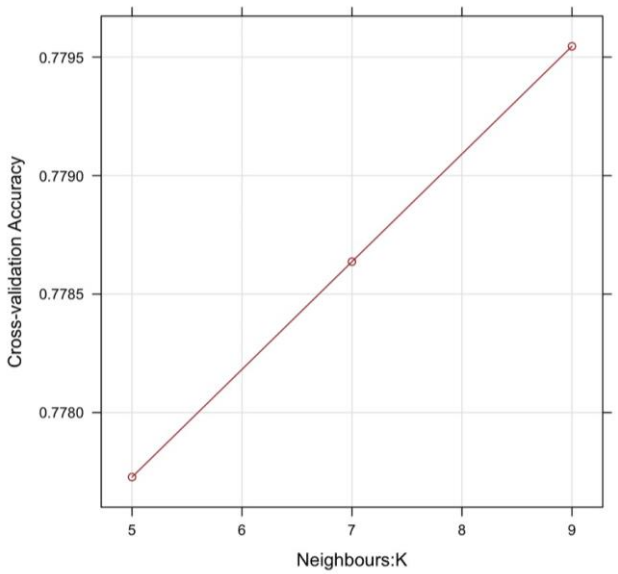

Fig. 4 Cross-Validation Accuracy for Neighbors: K (1)

Table 9. Prediction Table for KNN (1)

Prediction	0	1
0	16	2
1	5	24

According to table 9, if the author carefully examines the above table, which also includes a list of rates frequently calculated from a confusion matrix for a binary classifier, the author has a 16/(16+5) =0.7619 chance of correctly predicting that the life expectancy is lower than the median life expectancy of 73. These individuals are therefore expected to live relatively short lives. The author has a chance of 24/(2+24) = 0.9231 of correctly predicting that the life expectancy will be greater than the median life expectancy of 73. The overall accuracy of the first KNN model in predicting life expectancy is (16+24)/(16+2+2+24) =0.8511.

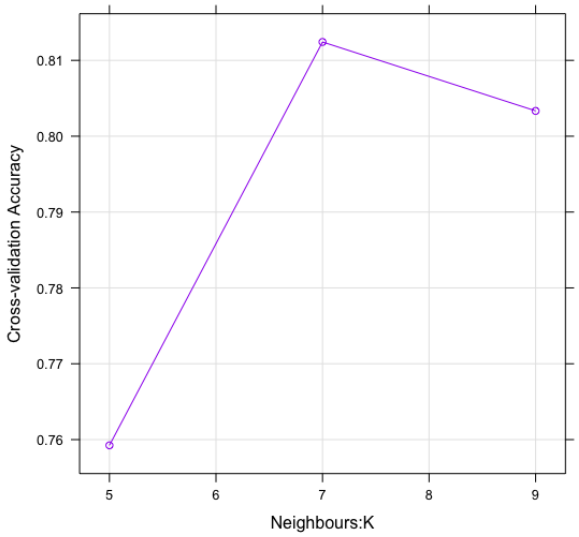

Fig. 5 Cross-Validation Accuracy for Neighbors: K (2)

The cross-validation accuracy for neighbors K——5, K——7, and K——9 is 0.7592, 0.8124, and 0.8033, respectively, as depicted by Fig. 5. As a result, the model's ideal k is set at 7.

Table 10. Prediction Table for KNN (2)

Prediction	0	1
0	18	3
1	3	23

According to table 10, the author has an 18/(18+3) =0.8571 chance of correctly predicting that the life expectancy is lower than the median life expectancy of 73 if people carefully review the above table, which also includes a list of rates frequently calculated from a confusion matrix for a binary classifier. As a result, it is anticipated that these people would have relatively brief lives. The odds of the author accurately forecasting that the life expectancy will be higher than the median life expectancy of 73 are 23/ (3+23) = 0.8846. The second KNN overall accuracy in estimating life expectancy is (18+23)/ (18+3+3+23) = 0.8723.

Table 11. Prediction Accuracy Comparison

Model	Accuracy
Logistic (1)	0.8936
Logistic (2)	0.8723
KNN (1)	0.8511
KNN (2)	0.8723

4. Conclusion

Table 11 illustrates the accuracy of Logistic Regression prediction regarding life expectancy before dropping ‘hiv’ and ‘gdppc’ is 0.8936 and the accuracy of that after discarding ‘hiv’ and ‘gdppc’ is 0.8732. One possible explanation for that discrepancy is that the accuracy of the prediction decreases slightly after removing one of the variables with a high correlation coefficient and a variable with insignificant statistical results.

Prior to skipping "hiv" and "gdppc," the accuracy of the KNN method is 0.8511; after skipping those two dependent variables, the accuracy of the KNN algorithm is 0.8723. With respect to the result of KNN (1), one possible explanation is that the explanatory variables are highly correlated, and the prediction accuracy for uncommon categories is relatively low.

Overall, the average accuracy of Logistic (1) and Logistic (2) is (0.8936+0.8723)/2=0.8830; the average accuracy of KNN (1) and KNN (2) is (0.8511+0.8723)/2=0.8617. The average prediction accuracy of KNN is slightly lower than expected based on the results, which may be related to the KNN Algorithm's propensity for misclassification in the class domain with a limited sample size (the observations here are 155.). Another possibility is that rather than using rigorous and established statistical theory, the best value of k is frequently chosen by K-fold cross-validation.

The authors may decide to use a dataset with a larger sampling size for statistical prediction in the future, if time permits. Additionally, the author can collect more variables that have an impact on life expectancy and can use more sophisticated techniques for classifying independent variables, increasing the prediction life model's accuracy even more.

References

[1] Aisa, Rosa, and Fernando Pueyo. Government health spending and growth in a model of endogenous longevity.Economics letters, 2006, 90(2): 249-253.

[2] Cervellati, Matteo, and Uwe Sunde. Human capital formation, life expectancy, and the process of development. American Economic Review, 2005, 95(5): 1653-1672.

[3] Breyer, Friedrich, and Stefan Felder. Life expectancy and health care expenditures: a new calculation for Germany using the costs of dying. Health Policy,2006, 75(2): 178-186.

[4] Chetty, Raj, et al. The association between income and life expectancy in the United States, 2001-2014. Jama,2016, 315(16): 1750-1766.

[5] Shaw, James W., William C. Horrace, and Ronald J. Vogel. The determinants of life expectancy: an analysis of the OECD health data. Southern Economic Journal, 2005, 71(4): 768-783.

[6] Wilkinson, Richard G. Income distribution and life expectancy. BMJ: British Medical Journal,1992, 304(6820):165.

[7] Hamidi, Shima, et al. Associations between urban sprawl and life expectancy in the United States. International journal of environmental research and public health, 2018, 15(5):861.

[8] Mariani, Fabio, Agustin Pérez-Barahona, and Natacha Raffin. Life expectancy and the environment." Journal of Economic Dynamics and Control ,2010, 34(4): 798-815.

[9] Pope III, C. Arden, Majid Ezzati, and Douglas W. Dockery. Fine-particulate air pollution and life expectancy in the United States. New England Journal of Medicine, 2009, 360(4): 376-386.

[10] Araki, Shunichi, and Katsuyuki Murata. Factors affecting the longevity of total Japanese population. The Tohoku Journal of Experimental Medicine, 1987,151(1): 15-24.

[11] Carr, Deborah. Early-Life lnfluences on later life well-Being: innovations and explorations. The Journals of Gerontology: Series B, 2019, 74(5): 829-831.

[12] Mirowsky, John, and Catherine E. Ross. Socioeconomic status and subjective life expectancy. Social Psychology Quarterly, 2000: 133-151.

[13] Marioni, Riccardo E., et al. Genetic variants linked to education predict longevity.Proceedings of the National Academy of Sciences , 2016, 133(47): 13366-13371.

[14] Torssander, Jenny. From child to parent? The significance of children's education for their parents' longevity. Demography, 2013, 50(2): 637-659.

[15] Meara, Ellen R., Seth Richards, and David M. Cutler. The gap gets bigger: changes in mortality and life expectancy, by education, 1981–2000. Health affairs, 2008,27(2): 350-360.

[16] May, Margaret T., et al. Impact on life expectancy of HIV-1 positive individuals of CD4+ cell count and viral load response to antiretroviral therapy. AIDS (London, England), 2014, 28(8): 1193.

[17] Trevisan, Maurizio, et al. Drinking Pattern and Mortality: The Italian Risk Factor and Life Expectancy Pooling Project. Annals of epidemiology, 2001, 11(5): 312-319.

[18] Gareth, James, et al. An introduction to statistical learning: with applications in R. Spinger, 2013:130-137.

[19] Huang, Jianglin, et al. Cross-validation based K nearest neighbor imputation for software quality datasets: an empirical study. Journal of Systems and Software, 2017 ,132: 226-252.