ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse45942020.pdf
https://doi.org/10.30534/ijatcse/2020/45942020

Machine Learning Techniques for Life Expectancy Prediction

Kasichainula Vydehi[1], Keerthi Manchikanti[2], T.Satya Kumari[3], SK Ahmad Shah[4]
[1]Senior Assistant Professor, Aditya Engineering College, India, vydehik9@gmail.com
[2]Assistant Professor, Keshav Memorial Institute of Technology, India, keerthi1212@gmail.com
[3]Senior Assistant Professor, Aditya College of Engineering, India, satyakumari.t@gmail.com
[4]Assistant Professor, Aditya College of Engineering & Technology, India, ahmad.alisha01@gmail.com

ABSTRACT

The human life span relies upon different features like the financial development of the nation, wellbeing developments of the people. In this paper, we proposed a machine learning model to predict the life expectancy of a person. We conducted our experiments on a dataset taken from Kaggle (WHO life expectancy dataset). The dataset contains 2938 rows and 22 features. We applied various regression algorithms to predict life expectancy. We also applied various classification algorithms by dividing life expectancy into five different ranges. On investigating different models, we can infer that random forest regression produces the most exact outcomes concerning life expectancy prediction. In classification models, random forest classification is given accuracy of 88%. We used Python for implementing all our experiments.

Key words : Life Expectancy, WHO, Machine Learning, Python.

1. INTRODUCTION

Machine Learning is a field of Artificial Intelligence[12] that has experienced exponential development from the previous years. Pretty much every part of life is being changed by enormous information and machine learning. The selection of a good dataset is a challenging task in the design of the machine learning model. The goal of applying Machine Learning model is to create a well-trained model which makes improvements over time [1]. Life expectancy is one of the most significant measures as far as populace's wellbeing in a nation and is utilized as a pointer by numerous arrangement creators and scientists to supplement financial proportions of flourishing, for example, GDP and so on. The anticipation of life portrays the normal age that the individuals from a specific populace gathering will be the point at which they pass on. Life hope shifts with created and creating nations, the proportion of birth to death, death paces of various nations furthermore, the proportion of proficient to ignorant populace, all influence the endurance time in one manner or the other [2]. The nation's development, headways, and availability of assets all are elements of the effective living pace of the populace. The life anticipation is determined as the normal endurance time which shows the middle time of populace where some may live till at that point, some may live additional time range, some may live less however on a normal the anticipated worth is the lifetime of that landmass. Prognosis of life is not only instrumental in predicting living rates but also helps in deciding whether there is a tendency of occurrence of disease in a continent. Along with the prediction of life, classification of disease is another aspect of research. Disease prediction is done by considering the economic, social factors of different countries in a particular continent and then we combine that data to predict it over a continent. The growth of the country affects the occurrence of disease in the country. The development rate of the country is dependent on, GDP, population awareness, illiteracy rate to literacy rate, birth to death ratio, all the factors have a combined effect on the striking of a disease. So, here machine learning algorithms play a major role.ML models can predict or classify.

2. LITERATURE SURVEY

Applying machine learning models for health applications is not a new era. Several researchers applied machine learning models for health-related issues. For predicting life expectancy also, various researchers proposed different machine learning models. Palak Agarwal [3] et.al proposed machine learning regression and classification models for predicting life expectancy. They applied multiple linear regression and random forest regression techniques and achieved good results. Michael B Schultz [4] et.al applied various machine learning models for predicting the lifetime of the mouse. They proposed a random forest regression that was trained on FI components for chronological age to generate the FRIGHT (Frailty Inferred Geriatric Health Timeline) clock, a strong predictor of chronological. Diogo G. Barardo [5] et.al proposed Machine Learning techniques for Predicting Lifespan-Extending Chemical Compounds. With their model, they achieved a prediction accuracy of 80%. James Jin Kang[6] proposed a Predictive Analysis of Personalized Life Expectancy with Smart Devices.Leng C.H [8] et.al proposed a simple linear regression technique with the logit model -transformed survival ratio between the

cohort, gender and age combination referents through simulation from the national life table. Merijn Beeksma[9] applied a deep learning model with LSTM (Long Short-Term Memory) using electronic medical records. They have shown that their model with natural language processing techniques given better results for predicting life expectancy. They shown better performance in terms of precision and recall.

3. RESEARCH METHODOLOGY

We collected dataset form Kaggle.After that, we check for missing values. Some of the features like Adult Mortality, Alcohol, bmi, Polio, Total expenditure, diphtheria, GDP, Income composition of resources, Schooling. We replace all missing values with mean value of that particular column. We also dropped some columns like Year','Status','Hepatitis B', 'Population',' thinness 1-19 years',' thinness 5-9 years.

3.1 Dataset

The dataset used for our life expectancy problem is collected from Kaggle [11] repository. The Global Health Observatory (GHO) information archive under the World Health Organization (WHO) monitors the wellbeing status just as numerous other related elements for all nations.
The datasets are made accessible to open with the end goal of wellbeing information investigation. The dataset identified with the future, wellbeing factors for 193 nations have been gathered from a similar WHO information archive site and its relating monetary information was gathered from the United Nation site. Among all classes of wellbeing related factors, just those basic elements were picked which are increasingly agents. It has been seen that in the previous 15 years, there has been enormous advancement in the wellbeing part bringing about the progress of human death rates particularly in the creating countries in contrast with the previous 30 years. Subsequently, in this work, we have considered information from the year 2000-2015 for 193 nations for additional examination. The individual information documents have been consolidated into a solitary dataset. On introductory visual review of the information indicated some missing qualities. As the datasets were from WHO, we found no obvious mistakes. The outcome demonstrated that a large portion of the missing information was for the populace, Hepatitis B, and GDP. The last combined file (final dataset) comprises 22 Columns and 2938 lines which implied 20 anticipating factors.

3.2. Regression models

We applied various regression models on our dataset. The performance of the model can evaluate by various metrics. Some of the metrics used for regression models are Mean Absolute Error (MAE), Mean Squared Error (MSE), RMSE (Root Mean Squared Error), R-Squared-squared value is the most widely used metrics for comparing various models. The value of R-Squared lies between 0 and 1. If the value near to 1 means model performance is good.

The equation for MLR is as follows:

$$Y=\alpha 0+ \alpha\ 1X1+ \alpha\ 2X2+\ldots\ldots\ \alpha kXk + \beta$$

Here, Xi is independent features, α 0 is y-intercept (constant term), α k is slope coefficient for dependent variables, β is model error term, Y is dependent feature.

3.2.2: Decision Tree Regression

Decision tree regression is one of the most widely used regression technique, which is basically based on a tree structure model. The leaf nodes of the decision tree contain values which are the predicted values of the model. Decision trees can be used for both classification and regression tasks is shown in Figure 1.

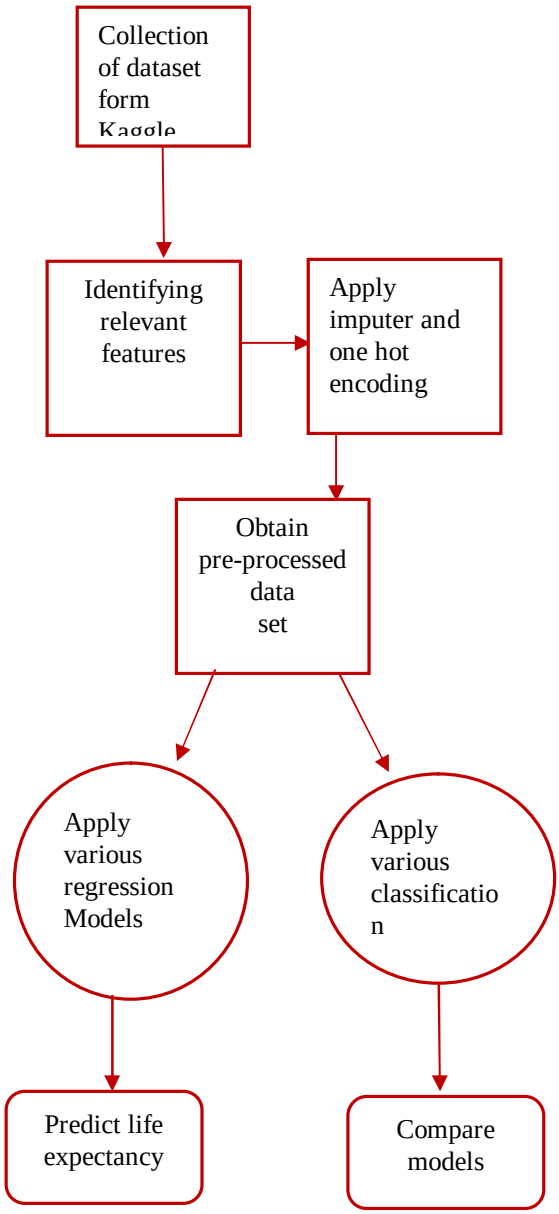

Figure 1: Proposed model

3.2.3: Random Forest Regression

Random Forest Regression is also called an ensemble model. In the ensemble model, more than one ML model is used to predict the final outcome. Random Forest uses several decision trees to predict the final outcome. So, it is more powerful model.

4. EXPERIMENTATION AND RESULTS

We collected dataset form Kaggle.After that, we check for missing values. Some of the features like Adult Mortality, Alcohol, bmi, Polio, Total expenditure, diphtheria, GDP, Income composition of resources, Schooling. We replace all missing values with mean value of that particular column. We also dropped some columns like Year','Status','Hepatitis B','Population',' thinness 1-19 years',' thinness 5-9 years.

4.1.1 Apply regression models

We applied the above regression models on the life expectancy dataset. The results are tabulated (figure 1).

Table 1: Comparison of regression models

ML model	R-squared	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Multiple Linear Regression	0.9472	1.4404	4.704	2.168
Decision Tree Regression	0.9574	1.2217	3.8008	1.9495
Random Forest Regression	0.958	1.206	3.741	1.9341

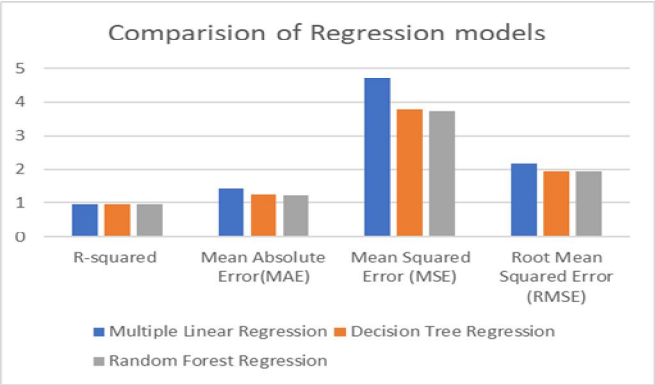

Figure 2: Comparison of Regression Models

After applying three regression models, we observed that random forest regression given better results as shown in Figure 2.

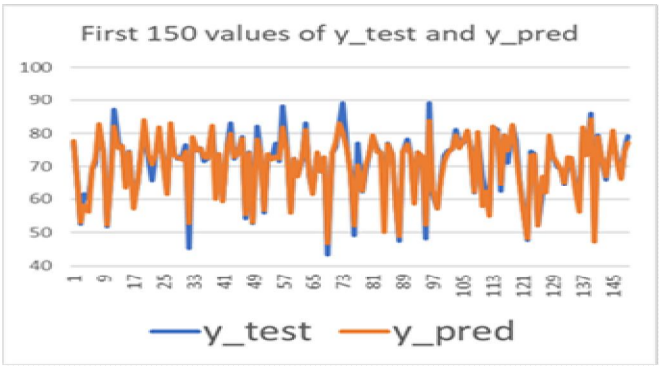

Figure 3: Comparison of y_test,y_pred values

4.1.2 Applying backward elimination

To identify the best feature which mostly affects the life expectancy, we perform regression analysis on the given dataset. Regression analysis can be done in backward elimination or forward selection procedure. In Backward elimination, initially we start with all independent features and eliminate the non-relevant features one by one and finally ended with best features.For eliminating non relevant feature, we used a measure called "P-value". If the p-value of the feature is less than 0.05, then the feature is relevant, otherwise the feature is treated as non relevant[7]. In Forward selection, initially we started with empty feature list, add one by one to the list. We implemented backward elimination only. We eliminated the country column from the dataset in backward elimination technique.

Initial list={ 0,1, 2, 3, 4,5,6,7,8,9,10,11,12,13,14}

coef	std err	t	P>|t|	[0.025	0.975]
const	53.1716	0.503	105.773	0.000	52.186	54.157
x1	-0.0204	0.001	-25.639	0.000	-0.022	-0.019
x2	0.0999	0.008	12.014	0.000	0.084	0.116
x3	0.1419	0.024	6.035	0.000	0.096	0.188
x4	0.0002	8.46e-05	1.921	0.055	-3.35e-06	0.000
x5	-1.995e-05	7.7e-06	-2.590	0.010	-3.5e-05	-4.85e-06
x6	0.0490	0.005	10.492	0.000	0.040	0.058
x7	-0.0752	0.006	-12.212	0.000	-0.087	-0.063
x8	0.0265	0.004	5.915	0.000	0.018	0.035
x9	0.1024	0.034	3.024	0.003	0.036	0.169
x10	0.0337	0.004	7.558	0.000	0.025	0.042
x11	-0.4754	0.018	-26.982	0.000	-0.510	-0.441
x12	3.393e-05	1.3e-05	2.601	0.009	8.35e-06	5.95e-05
x13	6.1713	0.639	9.665	0.000	4.919	7.423
x14	0.6738	0.042	16.029	0.000	0.591	0.756

Here, the p-value for all the features is less than 0.05. So, all the features are useful for predicting life expectancy. But consider the magnitude of regression coefficients. The regression coefficient value of x13 is 6.17 as shown in Figure 4. So, the most predominant feature that positively affects life expectancy is the "Income composition of resources" in the given dataset. But its standard error value is more. The standard error value of x1 is 0. 001 as shown in Figure 5. The x1 feature is "Adult Morality". The negative magnitudes of the regression coefficient represent that these features are negatively affecting the life expectancy prediction.

Regression coefficient values of all features

Figure 4: Comparison of Regression Coefficient values

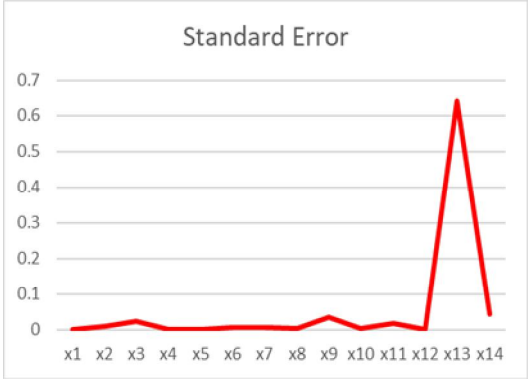

Figure 5: Comparison of Standard Values

4.2 Applying classification algorithms

After applying different regression models, we finally find the best predictor model with random forest regression. Later we applied different classification models on our dataset. In classification problems, the dependant feature must be categorical. But the values of life expectancy in the given dataset are continuous. So, we converted values continuous of life expectancy in the given dataset into five ranges namely above 80, above 70, above 60, above 0, above 40.Now, we have five class labels. So, the dataset is ready for applying classification algorithms. As this is a multiclass classification problem, now ML model needs to perform more calculations [10]. We applied three classification models K-NN, Decision tree classifier, Random Forest classifier and there comparison is shown in Table 2 and Figure 6.

4.2.1: K- Nearest Neighbor Classification

K-Nearest Neighbor is a simple and efficient machine learning classifier. It is a supervised learning model, where the class labels are known. In K-NN, K indicates the count of nearest neighbors. Initially, we select some k value and divide the data points into k groups based on distance metrics. If the new data point is close to a certain class, then divide that data point to that class.

4.2.2: Decision Tree Classification:

Like decision tree regression, decision tree classification is also the most widely used ML model. It is also a supervised learning model. For constructing decision tree, we need to find the root node at every step. It can be done by two metrices, combination of information gain and entropy or gini index.

4.2.3: Random forest classification:

Random Forest classifier is an ensemble learning model, where it combines different decision trees into a single model. It generates a forest with much number of trees. The number of trees can be decided by us. It is a powerful classifier, because it combines several decision tree classifiers.

After applying all three models, we observed that Random Forest given an better results regarding precision recall, f1-score and accuracy. Decision tree classifier also performed well with an accuracy of 87.4%.

Table 2: Comparison of classification models

ML model	K-NN	Decision Tree Classification	Random Forest Classification
Precision(0- class)	0.89	0.63	0.96
Recall(0-class)	0.70	0.82	0.55
f1-score(0-class)	0.78	0.71	0.70
Precision(1- class)	0.79	0.83	0.86
Recall(1-class)	0.81	0.81	0.90
f1-score(1-class)	0.80	0.82	0.88
Precision(2- class)	0.82	0.87	0.95
Recall(2-class)	0.75	0.79	0.82
f1-score(2-class)	0.78	0.83	0.88
Precision(3- class)	0.87	0.89	0.91
Recall(3-class)	0.89	0.90	0.97
f1-score(3-class)	0.88	0.90	0.94
Precision(4- class)	0.77	0.74	0.96
Recall(4-class)	0.74	0.84	0.87
f1-score(4-class)	0.75	0.79	0.91
Accuracy	**81.2%**	**87.4%**	**89%**

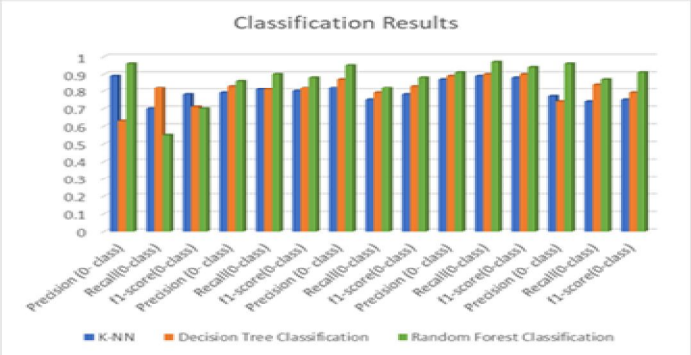

Figure 6: Comparison of classification models

4.2.4: Cross Validation

Cross validation is a technique where data samples are shuffled perfectly before going to learning phase.This technique is helpful for reducing overfitting[9].There are various number of cross validation techniques available. We applied 3-fold Repeated cross validation technique.We repeated 3-fold cross validation 2 times. In 2936 rows (we removed two rows with life expectancy less than40 from original dataset),1958 rows are used for training and 978 rows are used for testing.

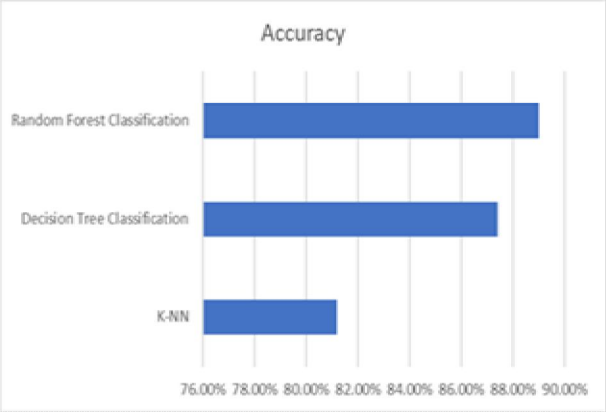

Figure 7: Comparison of classification accuracies

5. CONCLUSION

In this paper, we applied machine learning methods for life expectance prediction. As original dataset contain missing values and categorical data, first we applied data preprocessing techniques. Later we applied regression models for predicting life expectancy. The random forest has given better r-squared value. We applied backward elimination to find the features that most affecting life expectancy. We also applied classification algorithms by grouping the life expectancy values into five groups. In classification models, random forest classifier had given the best accuracy as shown in Figure 7.

REFERENCES

1. M. I. Jordan and T. M. Mitchell, **Machine learning: Trends,perspectives, and prospects**, sciencemag.org, *17 JULY 2015 • VOL 349 ISSUE 6245.*
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 in 2005–15: a cross-sectional analysis,***Lancet Public Health, vol. 4, no. 4, pp. e181–e188, Apr. 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) ISSN: 2278-3075, Volume-8 Issue-10, August 2019.*
4. Michael B Schultz, Alice E Kane1, Sarah J Mitchell, **Age and life expectancy clocks based on machine learning analysis of mouse frailty**, *https://doi.org/10.1101/2019.12.20.884452.*
5. Diogo G. Barardo, Danielle Newby, Daniel Thornton, Taravat Ghafourian,**Machine learning for predicting lifespan□extending chemical compounds**, *www.aging□us.comdoi: 10.18632/aging.101264, July 2017.*
6. James Jin Kang and Sasan Adibi, **Systematic Predictive Analysis of Personalized Life Expectancy Using Smart Devices**, *Technologies 2018, 6, 74; doi: 10.3390/technologies,6030074,www.mdpi.com/journal/technologies.*
7. A.Lakshmanarao,G.Vijay Kumar,T.S.Ravi Kiran, **An Effective Multiple Linear Regression Model For Power Load Prediction,***JETIR September 2018, Volume 5, Issue 9,2018*
8. Leng, C.H, Chou, M.H.; Lin, S.-H; Yang, Y.K.; Wang, J.D, **Estimation of life expectancy, loss-of-life expectancy, and lifetime healthcare expenditures for schizophrenia in Taiwan,** *Schizophr. Res. 2016,171, 97–102, DOI: 10.1016/j.schres.2016.01.03.*
9. Merijn Beeksmal , Suzan Verberne, Antal van den Bosch, Enny Das, Iris Hendrickx and Stef Groenewoud, **predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records**,*Beeksma et al. BMC Medical Informatics and Decision Making, https://doi.org/10.1186/s12911-019-0775-2,2019.*
10. Sushil Kumar Trisa, Ajay Kaul, **Dynamic Behavior Extraction from Social Interactions Using Machine Learning and Study of Over Fitting Problem**, *International Journal of Advanced Trends in Computer Science and Engineering, October-2019.*
11. *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, of Advanced Trends in Computer Science and Engineering, Volume 8, No.5, September - October 2019.*

https://doi.org/10.30534/ijatcse/2019/39852019