Machine Learning For Prognosis of Life Expectancy and Diseases

Palak Agarwal, Navisha Shetty, Kavita Jhajharia, Gaurav Aggarwal, Neha V Sharma

Abstract: ***Longevity depends on various facets such as economic growth of the country, along with the health innovations of the region. Along with the prophecy of existence, we also figure out how sensitive a particular mainland is to few chronic diseases. These factors have a robust impact on the potential life span of the population. We study the biological and economical aspects of continents and their countries to predict the life expectancy of the population and to perceive the probability of the continent possessing long standing diseases like measles, HIV/AIDS, etc. Our research is conducted on the theory that exhibits the dependency or correlation of life expectancy with the various factors which includes the health factors as well as the economic factors. Two Machine learning algorithms simple linear regression, multiple linear regression are used for predicting the expectancy of life over different continents, whereas, decision tree algorithm, random forest algorithm, and were applied to classify the likelihood of occurrence of the disease. On comparing and contrasting various algorithms, we can infer that, multiple linear regression produces the most accurate results as to what the average life expectancy of the population would be given the current features of the continent like the adult mortality rate, alcohol consumption rate, infant deaths, the GDP of the country, average percentage expenditure of the population on health care and treatments, schooling rate, and other such features. On the other hand, we study five diseases namely, HIV/AIDS, measles, diphtheria, hepatitis B and polio. The experiment concluded that, on majority, random forest produces better results of classification based on the economic factors of the combination of various countries of different continents.***

Keywords: Machine Learning, life expectancy, Diseases, life measurement, Regression, Random forest, Decision tree.

I. INTRODUCTION

Machine learning is a field of computer science which has experienced exponential growth from the past few years. Almost every aspect of life is being changed by big data and machine learning. Health informatics sphere poses a great challenge to this domain. The ultimate aim of applying machine learning is to develop algorithms which can be trained well and make improvements over time [1] [2]. Life expectancy is one of the most important measure in terms of population's health in a country and is used as an indicator by many policy makers and researchers to complement economic measures of prosperity such as GDP etc. Prognosis of life depicts the average age that the members of a particular population group will be when they die. Life expectancy varies with developed and developing countries, ratio of birth to death, mortality rates of different countries and ratio of literate to illiterate population, all affect the survival time in one way or the other [3]. The country's growth, advancements and accessibility of resources all are the factors of affect living rate of population. The life expectancy is calculated as the average survival time which indicates the median age of population where some might live till then, some might live more time span, some might live less but on an average the predicted value is the lifetime of that continent [4].

Prognosis of life is not only instrumental in predicting living rate 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 the 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. 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.

Therefore, machine learning is the suitable method to predict and classify. Regression, classification predictive algorithms can be used in various ways to achieve the desired output. For prediction of life expectancy, regression algorithms, linear regression and multiple regression are applied, whereas, for classification of diseases occurrence, application of classification algorithms, decision tree, random forest algorithm and k-nearest neighbor algorithm are applied to obtain the desired results.

II. MATERIALS AND METHODS

Although there has been study on factors affecting life expectancy considering some demographic features, income composition and mortality rates. However, as per author's information immunization and human development index factors have not been considered till now for any predictions. Important immunization such as Hepatitis B,

Revised Manuscript Received on August 05, 2019

Palak Agarwal, Dept. of Information Technology, Manipal University Jaipur, Jaipur, India.
Navisha Shetty, Dept. of Information Technology, Manipal University Jaipur, Jaipur, India.
Kavita Jhajharia*, Dept. of Information Technology, Manipal University Jaipur, Jaipur, India.
Gaurav Aggarwal, Dept. of Information Technology, Manipal University Jaipur, Jaipur, India.
Neha V Sharma, Dept. of Information Technology, Manipal University Jaipur, Jaipur, India.

Diphtheria, polio, HIV/AIDS, and measles considered in the dataset. The study majorly focuses on immunization factors, mortality rates, economic factors, social features, and other health related information [5]. It is easier for countries to predict which factor is contributing in low life expectancy in particular and then continent as a whole. The dataset used for research is of different countries and publically available.

A. Linear Regression

Linear regression is one of the simplest and easiest algorithm to understand and apply. It is a predictive machine learning algorithm which belongs to both, statistics, as well as machine learning. Linear Regression is used to find the linear relationship between the label and its one or more features. Thus, the features play the role to help predict the label [6]. Data points are plotted which generates a regression line. We get the best fit line which is the most estimated value with minimum distance between the predicted value and the actual value. It is used mainly for determining the strength of predictors, forecasting an effect and trend forecasting. Linear regression is categorized into two parts: Simple Linear regression and Multiple Linear regression.

- **Simple Linear Regression**

Simple linear regression has one independent variable and one dependent variable. It is useful for finding the relationship between 2 continuous variables. The goal is to find the best fit line which leads to prediction. So, the equation of line ($y = mx + c$) where m, c are the constants. Thus, plot the regression line. With the help of the equation we draw the regression line and obtain the predicted data points. The basic idea is to get the best fit line with maximum data points lying on the line. The model could even be under fitted or over fitted. An over fitted model is the one in which all the data points lie on the best fit line. And, the case of under fitting occurs when less than average number of data points fall on the regression line. How it works in computers is that the computer would perform n number of iterations for all the possible values of m. After the completion of every iteration, predicted values are calculated according to the line and compared with the actual values [7].

- **Multiple Linear Regression**

Multiple linear regression contains one dependent variable but multiple independent variables. The main objective is to model the linear relationship between the dependent and independent variables and predict the outcome [8]. The regression line is drawn with the consideration of all the features. Multiple Linear Regression functions on some assumptions and they are as follows:

- There exists a linear relationship between the independent and dependent variables.
- The degree of correlation between the independent variables shouldn't be too high.
- Observations are selected randomly and independently from the population.
- The residuals should be distributed normally with mean 0 and a variance sigma.

Another term that is used extensively in regression analysis is the coefficient of determination or the R squared value. It the metric that is used to determine the variation or deviation of the predicted value from the actual value. It is also called as the score, that is, the score of accuracy of the result. The value of R^2 increases with the increase in the number of variables in multiple linear regression [9].

B. Polynomial Regression

Sometimes, when we plot the scatterplot of residuals against the features in a dataset, nonlinear relationship is observed. In that case use polynomial regression to obtain higher accuracy level which we may obtain due to the better and more accurate model or the best fit line [10]. Therefore, in polynomial regression, which is a form of regression analysis, the relationship between the features 'x' and labels 'y' is represented in the form of a nth degree polynomial in x. It fits or models the non-linear relationship between the features 'x' and labels 'y' by taking the value of x and the conditional mean of y which is the corresponding value to x. Polynomial regression is also considered to be a special case scenario of multiple linear regression because, even though it models a nonalienated model to the data, it is still linear as a statistical estimation problem as the regression function is linear in the estimated unknown parameters from the data. Thus, using a polynomial regression increases the efficiency and accuracy of the model. That is, it increases the score or the R squared value.

C. Decision Tree

Decision tree is a supervised machine learning. This algorithm can be used for regression as well as classification. The aim of decision tree is to predict the value of target variable by the decision rules which are trained to the model by creating some training data. This algorithm is easy to understand as compared to other classification algorithms. It uses tree representation to solve the problem, wherein, each internal node corresponds to the attribute and each leaf node corresponds to a class label. The primary cause of concern in decision tree is to select an attribute for root node from the dataset. Handling this concern is known as attribute selection, which, can be done by two methods, one is by information gain method and second is Gini impurity method [11]. In this research, decision tree is used to predict whether or not a continent is having a particular disease. It uses various economic factors of each country in the continent to predict the occurrence of the disease over the continent. It predicts diseases such as Hepatitis B, diphtheria, HIV/AIDS, polio and measles on the factors adult mortality, infant deaths, percentage expenditure, total expenditure, GDP, population, schooling, and average BMI of people and alcohol consumption in the countries of that continent.

Decision tree is easy to explain and it follows the same concept as humans follow while making the decisions themselves. Along with pros come the cons which are that there is high probability of overfitting in this algorithm. Calculations become complex with multi class labels [12].

D. Random Forest

Random forest algorithm is also a supervised classification machine learning algorithm. It randomly creates the forest with n number of trees. In general, more the number of tress in the forest, the more robust the forest looks like. In the similar manner, in random forest classifier, the higher number of trees in the forest gives the high accurate results. Random forest algorithm manages the problem of missing values and does not over fit the model when we have more number of trees present. It can be used for categorical data as well [13]. There are two stages in random forest algorithm, one is random forest creation, the other is to make a prediction from the random forest classifier created in the first stage [14]. In this research, random forest algorithm is used to predict the likelihood of happening of disease over the continent by considering the economic conditions of all the countries in a particular continent. Predicting for five most concerning diseases like hepatitis B, diphtheria, HIV/AIDS, polio, measles.

E. K-Nearest Neighbors

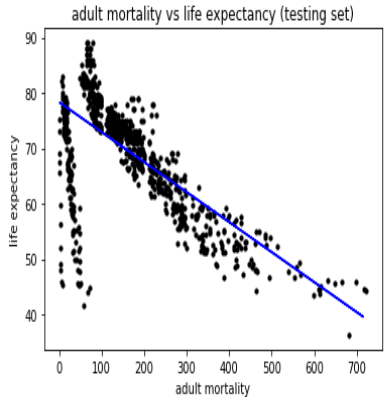
K nearest neighbors (KNN) is a non-parametric, lazy learner algorithm that stores all available use cases and classifies new cases based on the similarity measure like distance functions. The model structure is determined from the data. In this algorithm, there is no explicit training phase or it is very minimal, if required. This algorithm is also based on feature similarity that means that how closely the samples resemble with each other in the dataset. During classification, the object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. We use Euclidean distance (preferred when input variables are similar) to know which among the K instances in the training dataset are most similar to the new set of data for prediction [15]. In this research, KNN is used so that we can predict the occurrence of disease in a continent. It also classifies on the basis of combining countries, economic, social status that are present in that continent and predict whether or not there are chances of spreading of a particular disease like Hepatitis B, diphtheria, HIV/AIDS, Polio and finally measles.

III. RESULTS AND DISCUSSION:

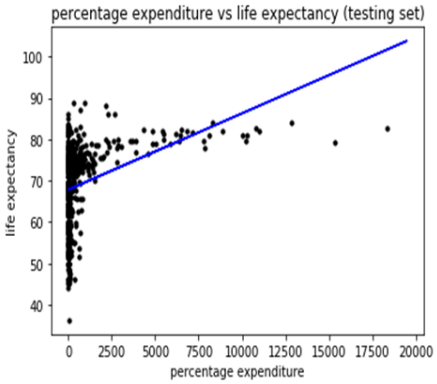
A. Regression Results:

- **Linear Regression Results**

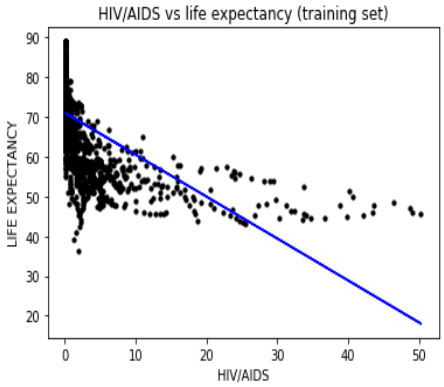
The linear regression model is applied on the dataset and keeping the label as life expectancy constant, the features changed and the following graphs have been produced.



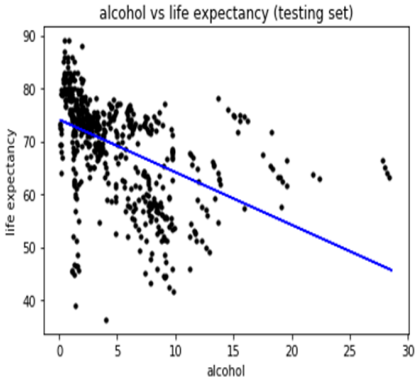
(a)



(b)



(c)



(d)

Fig 1: Representing corelation of various factors with life expectancy, (a) Adult mortality VS Life Expectancy, (b) percentage expenditure VS Life Expectancy,

(c) HIV/AIDS VS Life Expectancy, (d) Alcohol VS Life Expectancy.

Similarly, the best fit line is produced for every feature against life expectancy. Life expectancy has a positive relationship with percentage expenditure, total expenditure, GDP, Schooling, Income composition of resources and BMI and a negative correlation with Adult mortality rate, hepatitis B, under five deaths, Measles, Population, HIV/AIDS and alcohol consumption rate.

- **Multiple And Polynomial Results:**

Table 1: Predicting Life Expectancy Using Regression Analysis Over The Various Continents Of The World

CONTINENTS	MULTIPLE LINEAR REGRESSION		POLYNOMIAL REGRESSION	
	PREDICTED VALUE	SCORE (%age)	ACTUAL VALUE	PREDICTED VALUE
WORLD DATA	70.81 years	0.82	65.0 years	57.46 years
ASIA	79.21 years	0.75	79.7 years	70.29 years
AFRICA	52.03 years	0.81	47.9 years	45.59 years
AUSTRALIA	71.50 years	0.84	72.3 years	70.46 years
EUROPE	73.27 years	0.64	69.8 years	74.03 years
AMERICA	77.12 years	0.78	75.6 years	75.47 years

- **Classification results:**

Table 2: Results Of Predicting The Likelihood Of Occurrence Of The Disease Over Various Continents Of The World Using Classification Analysis.

CONTINENTS	DISEASES	DECISION TREE			RANDOM FOREST			K-NEAREST NEIGHBOURS		
		WRONG SAMPLE	SCORE	CONFUSION MATRIX	WRONG SAMPLE	SCORE	CONFUSION MATRIX	WRONG SAMPLE	SCORE	CONFUSION MATRIX
ASIA	HEPATITIS B	1	92.31	$\begin{bmatrix}3&0\\1&9\end{bmatrix}$	0	100.00	$\begin{bmatrix}3&0\\0&10\end{bmatrix}$	1	92.30	$\begin{bmatrix}3&0\\1&9\end{bmatrix}$
	DIPHTHERIA	2	84.62	$\begin{bmatrix}3&0\\2&8\end{bmatrix}$	1	92.30	$\begin{bmatrix}3&0\\1&9\end{bmatrix}$	2	84.61	$\begin{bmatrix}3&0\\2&8\end{bmatrix}$
	HIV/AIDS	4	55.56	$\begin{bmatrix}4&4\\0&1\end{bmatrix}$	3	66.62	$\begin{bmatrix}5&3\\0&1\end{bmatrix}$	3	66.66	$\begin{bmatrix}6&2\\1&0\end{bmatrix}$
	POLIO	1	92.31	$\begin{bmatrix}4&0\\1&8\end{bmatrix}$	0	100.00	$\begin{bmatrix}4&0\\0&9\end{bmatrix}$	0	100.00	$\begin{bmatrix}4&0\\0&9\end{bmatrix}$
	MEASLES	1	92.31	$\begin{bmatrix}4&0\\1&8\end{bmatrix}$	0	100.00	$\begin{bmatrix}4&0\\0&9\end{bmatrix}$	0	100.00	$\begin{bmatrix}4&0\\0&9\end{bmatrix}$
AFRICA	HEPATITS B	4	73.33	$\begin{bmatrix}4&0\\4&7\end{bmatrix}$	4	73.33	$\begin{bmatrix}4&0\\4&7\end{bmatrix}$	2	86.66	$\begin{bmatrix}4&0\\2&9\end{bmatrix}$
	DIPHTHERIA	0	100.00	$\begin{bmatrix}4&0\\0&11\end{bmatrix}$	0	100.00	$\begin{bmatrix}4&0\\0&11\end{bmatrix}$	3	80.00	$\begin{bmatrix}4&0\\3&8\end{bmatrix}$
	HIV/AIDS	2	80.00	$\begin{bmatrix}3&1\\1&5\end{bmatrix}$	2	80.00	$\begin{bmatrix}3&1\\1&5\end{bmatrix}$	1	90.00	$\begin{bmatrix}3&1\\0&6\end{bmatrix}$
	POLIO	2	86.66	$\begin{bmatrix}4&2\\0&9\end{bmatrix}$	2	86.66	$\begin{bmatrix}4&2\\0&9\end{bmatrix}$	2	86.67	$\begin{bmatrix}4&2\\0&9\end{bmatrix}$
	MEASLES	0	100.00	$\begin{bmatrix}4&0\\0&11\end{bmatrix}$	0	100.00	$\begin{bmatrix}4&0\\0&11\end{bmatrix}$	3	80.00	$\begin{bmatrix}4&0\\3&8\end{bmatrix}$
AUSTRALIA	HEPATITIS B	0	100.00	$\begin{bmatrix}3&0\\0&2\end{bmatrix}$	0	100.00	$\begin{bmatrix}3&0\\0&2\end{bmatrix}$	0	100.00	$\begin{bmatrix}3&0\\0&2\end{bmatrix}$
	DIPHTHERIA	0	100.00	$\begin{bmatrix}3&0\\0&2\end{bmatrix}$	0	100.00	$\begin{bmatrix}3&0\\0&2\end{bmatrix}$	0	100.00	$\begin{bmatrix}3&0\\0&2\end{bmatrix}$
	HIV/AIDS	1	66.67	$\begin{bmatrix}2&0\\1&0\end{bmatrix}$	3	40.00	$\begin{bmatrix}2&0\\1&0\end{bmatrix}$	2	33.33	$\begin{bmatrix}1&1\\1&0\end{bmatrix}$
	POLIO	0	100.00	$\begin{bmatrix}3&0\\0&2\end{bmatrix}$	0	100.00	$\begin{bmatrix}3&0\\0&2\end{bmatrix}$	0	100.00	$\begin{bmatrix}3&0\\0&2\end{bmatrix}$
	MEASLES	3	40.00	$\begin{bmatrix}2&0\\3&0\end{bmatrix}$	3	40.00	$\begin{bmatrix}2&0\\3&0\end{bmatrix}$	3	40.00	$\begin{bmatrix}2&0\\3&0\end{bmatrix}$
EUROPE	HEPATITIS B	4	63.64	$\begin{bmatrix}4&1\\3&3\end{bmatrix}$	4	63.64	$\begin{bmatrix}3&2\\2&4\end{bmatrix}$	4	63.64	$\begin{bmatrix}3&2\\2&4\end{bmatrix}$
	DIPHTHERIA	5	54.55	$\begin{bmatrix}4&3\\2&2\end{bmatrix}$	4	63.64	$\begin{bmatrix}4&3\\1&3\end{bmatrix}$	4	63.64	$\begin{bmatrix}4&3\\1&3\end{bmatrix}$
	HIV/AIDS	1	87.50	$\begin{bmatrix}7&0\\1&0\end{bmatrix}$	1	87.50	$\begin{bmatrix}7&0\\1&0\end{bmatrix}$	1	87.50	$\begin{bmatrix}7&0\\1&0\end{bmatrix}$
	POLIO	2	81.82	$\begin{bmatrix}5&0\\2&4\end{bmatrix}$	1	90.90	$\begin{bmatrix}5&0\\1&5\end{bmatrix}$	1	90.91	$\begin{bmatrix}5&0\\1&5\end{bmatrix}$
	MEASLES	2	81.82	$\begin{bmatrix}5&1\\1&4\end{bmatrix}$	1	90.90	$\begin{bmatrix}5&1\\0&5\end{bmatrix}$	1	90.91	$\begin{bmatrix}5&1\\0&5\end{bmatrix}$
NORTH AMERICA	HEPATITIS B	2	71.43	$\begin{bmatrix}1&1\\1&4\end{bmatrix}$	2	71.42	$\begin{bmatrix}1&1\\1&4\end{bmatrix}$	3	57.14	$\begin{bmatrix}1&1\\2&3\end{bmatrix}$
	DIPHTHERIA	1	85.71	$\begin{bmatrix}1&1\\0&5\end{bmatrix}$	1	85.71	$\begin{bmatrix}1&1\\0&5\end{bmatrix}$	1	85.71	$\begin{bmatrix}1&1\\0&5\end{bmatrix}$
	HIV/AIDS	2	60.00	$\begin{bmatrix}0&0\\2&3\end{bmatrix}$	2	60.00	$\begin{bmatrix}0&0\\2&3\end{bmatrix}$	3	40.00	$\begin{bmatrix}0&0\\3&2\end{bmatrix}$
	POLIO	0	100.00	$\begin{bmatrix}2&0\\0&5\end{bmatrix}$	0	100.00	$\begin{bmatrix}2&0\\0&5\end{bmatrix}$	1	85.71	$\begin{bmatrix}2&0\\1&4\end{bmatrix}$

	MEASLES	2	71.43	$\begin{bmatrix}3 & 2\\0 & 2\end{bmatrix}$	3	57.14	$\begin{bmatrix}3 & 2\\1 & 1\end{bmatrix}$	2	71.43	$\begin{bmatrix}4 & 1\\1 & 1\end{bmatrix}$
SOUTH AMERICA	HEPATITIS B	1	75.00	$\begin{bmatrix}1 & 0\\1 & 2\end{bmatrix}$	1	75.00	$\begin{bmatrix}1 & 0\\1 & 2\end{bmatrix}$	0	100.00	$\begin{bmatrix}1 & 0\\0 & 3\end{bmatrix}$
	DIPHTHERIA	1	75.00	$\begin{bmatrix}1 & 0\\1 & 2\end{bmatrix}$	1	75.00	$\begin{bmatrix}1 & 0\\1 & 2\end{bmatrix}$	0	100.00	$\begin{bmatrix}1 & 0\\0 & 3\end{bmatrix}$
	HIV/AIDS	2	33.33	$\begin{bmatrix}1 & 2\\0 & 0\end{bmatrix}$	2	33.33	$\begin{bmatrix}1 & 2\\0 & 0\end{bmatrix}$	2	33.33	$\begin{bmatrix}1 & 2\\0 & 3\end{bmatrix}$
	POLIO	1	75.00	$\begin{bmatrix}1 & 0\\1 & 2\end{bmatrix}$	1	75.00	$\begin{bmatrix}1 & 0\\1 & 2\end{bmatrix}$	0	100.00	$\begin{bmatrix}1 & 0\\0 & 3\end{bmatrix}$
	MEASLES	1	75.00	$\begin{bmatrix}1 & 0\\1 & 2\end{bmatrix}$	1	75.00	$\begin{bmatrix}1 & 0\\1 & 2\end{bmatrix}$	0	100.00	$\begin{bmatrix}1 & 0\\0 & 3\end{bmatrix}$

Figure 1 describes the dependency of various factors on the life expectancy parameter. As few factors shown affect the life expectancy in positive or negative manner, there are a lot more factors in the dataset which are included for predicting the correlation. Table 1 explains the accuracy and the predicted life expectancy over the various continents. Table 2 contains the classification report on the prediction of probability of occurrence of the several diseases as mentioned above over various continents.

IV. CONCLUSION

By the application of linear regression, we conclude that according to the given dataset, the features that affect the life expectancy the most are Adult mortality rate, Percentage expenditure and total expenditure on healthcare and treatments, Hepatitis B, Polio, Under 5 death rate, Measles, Population, GDP, HIV/AIDS, Schooling, Income composition, BMI, and Alcohol consumption rate. Application of multiple linear regression to the dataset has produced successful results in predicting the life expectancy of the population in the world as well as the continents in the world. The accuracy score has been good which indicates that the predicted result is almost accurate. Further the application of the polynomial regression generates results which were less accurate than the results produced by the multiple linear regression model except for the results of a few continents where the prediction was closer to the actual value by the use of polynomial regression. Thus, multiple linear regression is a better and more accurate model that produces better results and almost accurate prediction.

From classification algorithm, according to the given dataset and the model that is applied, for continent Asia, random forest turns out to be a better algorithm to be implemented among three algorithms, which enlighten that for the diseases like Hepatitis B, Polio and Measles, the predicted value will be the most accurate. Further, for continent Africa, all the three algorithms delivers nearly same results but with minor difference decision tree and random forest is more suitable for Diphtheria, Measles and Polio; for Hepatitis B and HIV/AIDS, KNN is better result producer. Considering the continent Australia, among all the algorithms decision tree produces more accurate results for all the diseases. For continent, Europe, random forest and KNN nearly produces the same amount of accurate results for the five diseases and decision tree will produce less accurate as compared to them. For continent, North America, decision tree provides the most accurate results for all the five diseases, and for continent, South America, KNN is the best algorithm to predict whether or not there is likelihood of getting affected by a particular disease. Every classification and regression algorithm produces accurate results in one context or the other.

REFERENCES

1. M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” Science, vol. 349, no. 6245, pp. 255–260, Jul. 2015.
2. A. L. Beam and I. S. Kohane, “Big Data and Machine Learning in Health Care,” JAMA, vol. 319, no. 13, p. 1317, Apr. 2018.
3. 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.
4. D. M. J. Naimark, “Life Expectancy Measurements,” in International Encyclopedia of Public Health, H. K. (Kris) Heggenhougen, Ed. Oxford: Academic Press, 2008, pp. 83–98.
5. H. Ouellette-Kuntz, L. Martin, and K. McKenzie, “Chapter Six - A Review of Health Surveillance in Older Adults with Intellectual and Developmental Disabilities,” in International Review of Research in Developmental Disabilities, vol. 48, C. Hatton and E. Emerson, Eds. Academic Press, 2015, pp. 151–194.
6. R. B. Darlington, Regression and linear models. McGraw-Hill New York, 1990.
7. K. H. Zou, K. Tuncali, and S. G. Silverman, “Correlation and Simple Linear Regression,” Radiology, vol. 227, no. 3, pp. 617–628, Jun. 2003.
8. K. J. Preacher, P. J. Curran, and D. J. Bauer, “Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis,” J. Educ. Behav. Stat., vol. 31, no. 4, pp. 437–448, Dec. 2006.
9. D. F. Andrews, “A Robust Method for Multiple Linear Regression,” Technometrics, vol. 16, no. 4, pp. 523–531, Nov. 1974.
10. H. Theil, “A Rank-Invariant Method of Linear and Polynomial
11. Regression Analysis,” in Henri Theil’s Contributions to Economics and Econometrics: Econometric Theory and Methodology, B. Raj and J. Koerts, Eds. Dordrecht:Springer Netherlands, 1992, pp. 345–381.
12. J. R. Quinlan, “DECISION TREES AS PROBABILISTIC CLASSIFIERS,” in Proceedings of the Fourth International Workshop on MACHINE LEARNING, Elsevier, 1987, pp. 31–37.
13. C. W. Olanow and W. C. Koller, “An algorithm (decision tree) for the management of Parkinson’s disease: treatment guidelines,” Neurology, vol. 50, no. 3 Suppl 3, pp. S1–S1, 1998.
14. L. Breiman, “Random Forest,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

15. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," J. Chem. Inf. Comput. Sci., vol. 43, no. 6, pp. 1947–1958, Nov. 2003.
16. I. S. Abramson, "Adaptive Density Flattening--A Metric Distortion Principle for Combating Bias in Nearest Neighbor Methods," Ann. Stat., vol. 12, no. 3, pp. 880–886, Sep. 1984.

AUTHORS PROFILE

Palak Agarwal is an undergraduate who is currently pursuing her Bachelors of Technology in Information Technology from Manipal University Jaipur, Rajasthan, India. Her areas of interest in academics is towards the domain of Data Science and Analytics, wherein, she is inclined towards working on building predictive models for different kind of dataset like healthcare data or marketing data. Exploring data and working on it has always been thought-provoking and she is always keen on drawing conclusions and extrapolating the data. Hence, this is her first research paper which involves working with the medicinal data.

Ms. Navisha Shetty is a student at Manipal University Jaipur who is pursuing her Btech degree in the stream of Information Technology and is currently in the final year of her engineering college. Academically, she has a keen interest in the field of data science and has worked on a few projects based on machine learning as well. Her research interests include sentiment analysis using NLP and other data science related topics. She is also an enthusiast in extracurricular activities in and around college.

Kavita Jhajharia Was Born in Jhunjhunu, Rajasthan, India, in 1992. She completed her B.Tech Degree from Rajasthan Technical University, India, in 2013 from department of Information Technology, and the M.tech Degree from SRM University, Sonepat, India, in 2016. She is pursuing her PhD from Manipal university jaipur, Rajasthan. She is Assistant Professor in Information Technology Department at Manipal University Jaipur since 2016. She is member of ACM and IEEE. Her Research domains are Machine Learning and Software Engineering.

Gaurav Aggarwal received his B.Tech degree in Instrumentation from University Science and Instrumentation Centre, Kurukshetra University, India in 2006. He received his M.Tech degree in Computer Science and Engineering from Department of Computer Science and Application from Kurukshetra University, India in 2008. He is pursuing his PhD from The NorthCap University, Gurgaon, India in the area of machine learning. He is working as Assistant Professor for department of Information Technology in Manipal University Jaipur, India. His area of interests includes signal processing, machine learning, and cognitive sciences. He has published several technical papers and reports in the above research areas at national and International platform.

Ms. Neha V Sharma is working as an Assistant Professor in the department of Information Technology MANIPAL University Jaipur since 2016. She has an overall experience of 9 years in teaching and industry. She has done her M.Tech in 2014. She is currently pursuing her Phd. from Manipal University Jaipur. Her research area is secured Computer Networks and Machine learning. She has authored many research papers and book chapters in reputed journals and conferences.