IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

(DPT-AI-CH2)

AI Based Disease Prediction Technology

[1]Mr. Buddh Bhagwan Sahu, [2]Assistant Professor

Dept. of Computer Science & Engineering, Columbia Institute of Engineering and Technology,
Near Vidhan Sabha, Tekari-493111, Raipur (C.G.), India[3]

CHAPTER-02
Learning Prediction Technique and Data Science

Abstract*:* Learning Prediction Technique and Data Science (LPT-DS) Is one of the way to simulate disease prediction technique. When we organize the data between method and machine will fetch the data from metadata. Machine always gives more accurate results based on the collected data from metadata. All metadata will follow to the linked and trained data. The dataset is a collection of metadata with the combination of dataset and trained data will gives more desirable results including learning techniques i.e. Regression, Clustering, Time-Series Forecasting, Sigmoid model, and Support Vector Machine. When past life stories appear, the metadata has been referred to as past life regression. However, regression technique does not try and prove the truth of these stories; rather, it uses the client's interpretation as a trained list, it means of providing a powerful transpersonal experience for execution. When traumatic memories, brain or mental health based on the source of the problem, regression technique includes.

Keywords: Prediction, Personalization, Participation, Database, Non-communicable chronic diseases, Proactive, Healthcare innovation and Prevention, Dataset. Linear, Non-Linear learning techniques, support vector machines, Sigmoid and Restricted Boltzmann Methods.

INTRODUCTION

Regression analysis is the statistical method used to determine the structure of a relationship between two variables (single linear regression) or three or more variables (multiple regressions). "Regression allows us to gain insights into the structure of that relationship and provides measures of how well the data fit that relationship, "One way to think of regression is by visualizing a scatter plot of your data with the independent variable on the X-axis and the dependent variable on the Y-axis. [23], [24] The regression line is the line that best fits the scatter plot data. The regression equation represents the line's slope and the relationship between the two variables, along with an estimation of error. Physically creating this scatter plot can be a natural starting point for parsing out the relationships between variables. [25], [26] Linear regression and logistic regression are two types of regression analysis techniques that are used to solve the regression problem using machine learning. They are the most prominent techniques of regression. But, there are many types of regression analysis techniques in machine learning, and their usage varies according to the nature of the data involved. Regression analysis is a predictive modelling technique that analyzes the relation between the target or dependent variable and independent variable in a dataset. The different types of regression analysis techniques get used when the target and independent variables show a linear or non-linear relationship between each other, and the target variable contains continuous values. The regression technique gets used mainly to determine the predictor strength, forecast trend, time series, and in case of cause & effect relation.

Type of Regression Analysis Techniques:

There are many types of regression analysis techniques, and the use of each method depends upon the number of factors. These factors include the type of target variable, shape of the regression line, and the number of independent variables. [21], [22]

The different regression techniques:

1. Linear Regression.
2. Logistic Regression.
3. Ridge Regression.
4. Lasso Regression.
5. Polynomial Regression.
6. Bayesian Linear Regression.

Linear regression analysis is used to predict the value of a variable based on the pre-defined value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. This type of statistical model is often used for classification and predictive analytics. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a log-it transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Log-it (pi) = 1/ (1+ dpt (-pi))

In (pi/ (1-pi)) = beta_0 + beta_1*X_1 + … + b_k*b_k

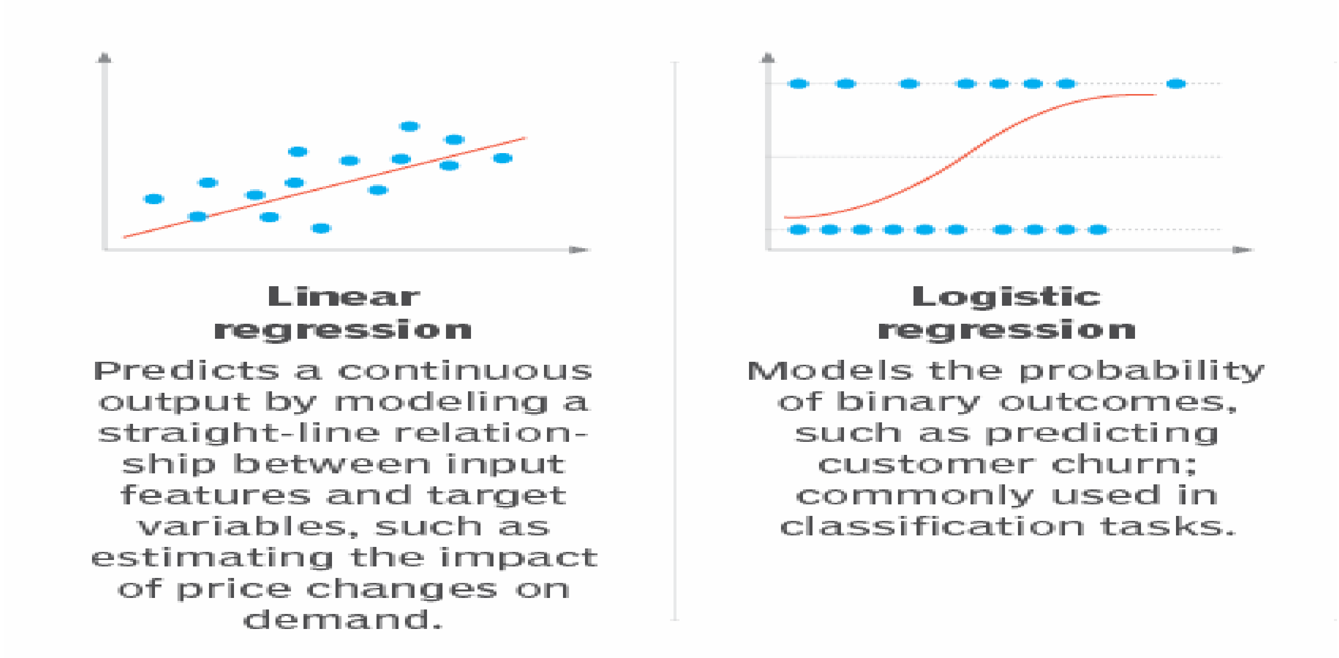
In this logistic regression equation, log-it (pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). [19], [20].This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability.

Logistic Regression: Disease prediction: In medicine, this analytics approach can be used to predict the likelihood of disease or illness for a given population. Healthcare organizations can set up preventative care for individuals that show higher propensity for specific illnesses.

Ridge Regression: In machine learning, ridge regression helps reduce over-fitting those results from model complexity. Model complexity can be due to a model possessing too many features. [17], [18] Features are the model's predictors and may also be called "parameters" in machine learning. Online tutorials often recommend keeping the number of features below the number of instances in training data sets. Such is not always being feasible however.

1. Feature possessing too much weight.
2. Feature weight refers to a given predictor's effect on the model output.
3. A high feature weight is equivalent to a high-value coefficient.

Simpler models do not intrinsically perform better then complex models. Nevertheless, a high degree of model complexity can inhibit a model's ability to generalize on new data outside of the training set. Because ridge regression does not perform feature selection, it cannot reduce model complexity by eliminating features. But if one or more features too heavily affect a model's output, ridge regression ridge regression can shrink high feature weights (i.e. coefficients) across the model per the L2 penalty term. This reduces the complexity of the model and helps make model predictions less erratically dependent on any one or more feature. [15], [16]



Lasso Regression: In machine learning, ridge regression helps reduce over-fitting those results from model complexity. Model complexity can be due to a model possessing too many features. Features are the model's predictors and may also be called "parameters" in machine learning. Online tutorials often recommend keeping the number of features below the number of instances in training data sets. Such is not always being feasible however. [13], [14] LASSO regression offers a powerful framework for both prediction and feature selection, especially when dealing with high-dimensional datasets where the number of features is large. By striking a balance between simplicity and accuracy, LASSO can provide interpretable models while effectively managing the risk of over-fitting.

Polynomial Regression: There are some relationships that a researcher will hypothesize is curvilinear. Clearly, such types of cases will include a polynomial term. Inspection of residuals, If we try to fit a linear model to curved data, a scatter plot of residuals (Y-axis) on the predictor (X-axis) will have patches of many positive residuals in the middle. Hence in such a situation, it is not appropriate. An assumption in the usual multiple linear regression analysis is that all the independent variables are independent. In the polynomial regression model, this assumption is not satisfied. [9], [10] Polynomial regression is a type of regression analysis used in statistics and machine learning when the relationship between the independent variable (input) and the dependent variable (output) is not linear. While simple linear regression models the relationship as a straight line, polynomial regression allows for more flexibility by fitting a polynomial equation to the data.

Bayesian Linear Regression: Bayesian decision making using posterior probabilities and a variety of loss functions. We discussed how to minimize the expected loss for hypothesis testing. Moreover, we introduced the concept of Bayes factors and gave some examples on how Bayes factors can be used in Bayesian hypothesis testing for comparison of two means. We also discussed how to choose appropriate and robust priors. When there is no contumacy. We applied Markov Chain Monte Carlo simulation to approximate the posterior distributions of parameters of interest. [11], [12]

General Modification-:

Continuous Target Variable: Regression deals with predicting continuous target variables that represent numerical values like.

a. Include predicting house prices.
b. Forecasting sales figures.
c. Estimating patient recovery times.

Measurement of Error: Regression models are evaluated based on their ability to minimize the error between the predicted and actual values of the target variable. Common error metrics includes. [7], [8]

a. Mean Absolute Error.
b. Mean Squared Error, and
c. Root Mean Squared error.

Complexity of Model: Regression models range from simple linear models to more complex nonlinear models. The choice of model complexity depends on the complexity of the relationship between the input features and the target variable.

Overfitting/Underfitting: Regression models are susceptible to Overfitting and Underfitting.

Interpretability: The interpretability of regression models varies depending on the algorithm used. Simple linear models are highly interpretable, while more complex models may be more difficult to interpret.

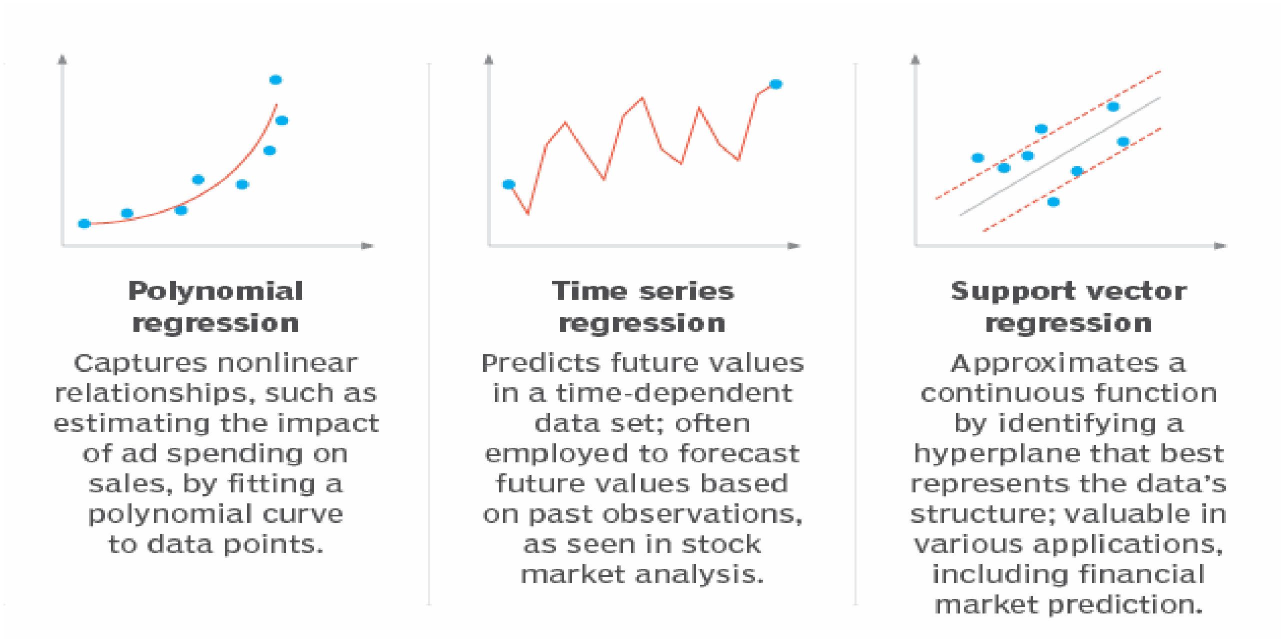
Expected Outcomes:

import the necessary libraries [5], [6]

```
from sklearn.datasets import load_blood_cancer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

X, y = load_ blood_cancer(return_X_y=True)
# split the train and test dataset
X_train, X_test,\
    y_train, y_test = train_test_split(X, y,
                                       test_size=0.20,
                                random_state=23)
# LogisticRegression
clf = LogisticRegression(random_state=0)
clf.fit(X_train, y_train)
# Prediction
y_pred = clf.predict(X_test)
acc = accuracy_score(y_test, y_pred)
print("Logistic Regression model accuracy (in %):", acc*100)
import pandas
from sklearn import linear_mode
df = pandas.read_csv("data.csv")
X = df[['Weight', 'Volume']]
Y = df['CO2']
regr = linear_model.LinearRegression()
regr.fit(X, Y)
predictedCO2 = regr.predict([[3300, 1300]])
print(predictedCO2)
```

O/P: [114.75968007]



Polynomial regression

Captures nonlinear relationships, such as estimating the impact of ad spending on sales, by fitting a polynomial curve to data points.

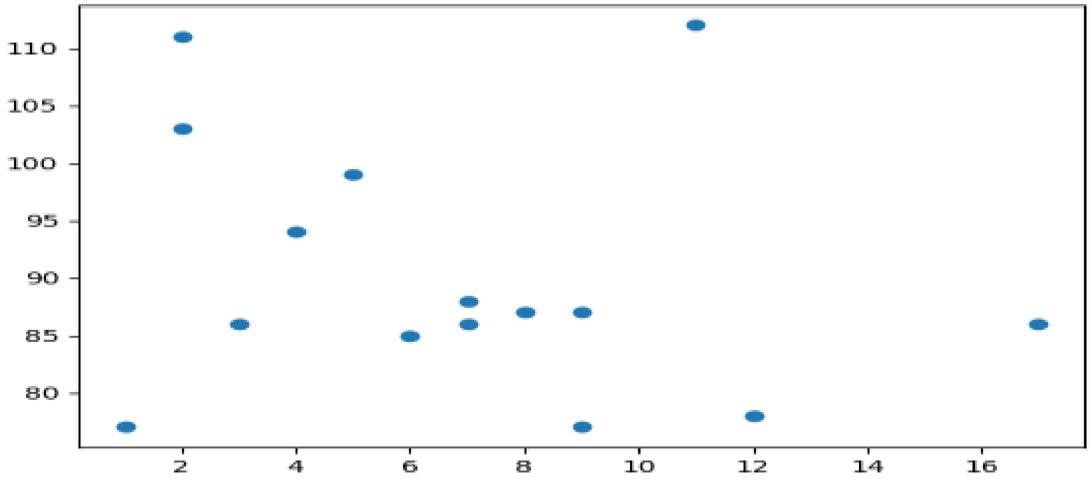
Time series regression

Predicts future values in a time-dependent data set; often employed to forecast future values based on past observations, as seen in stock market analysis.

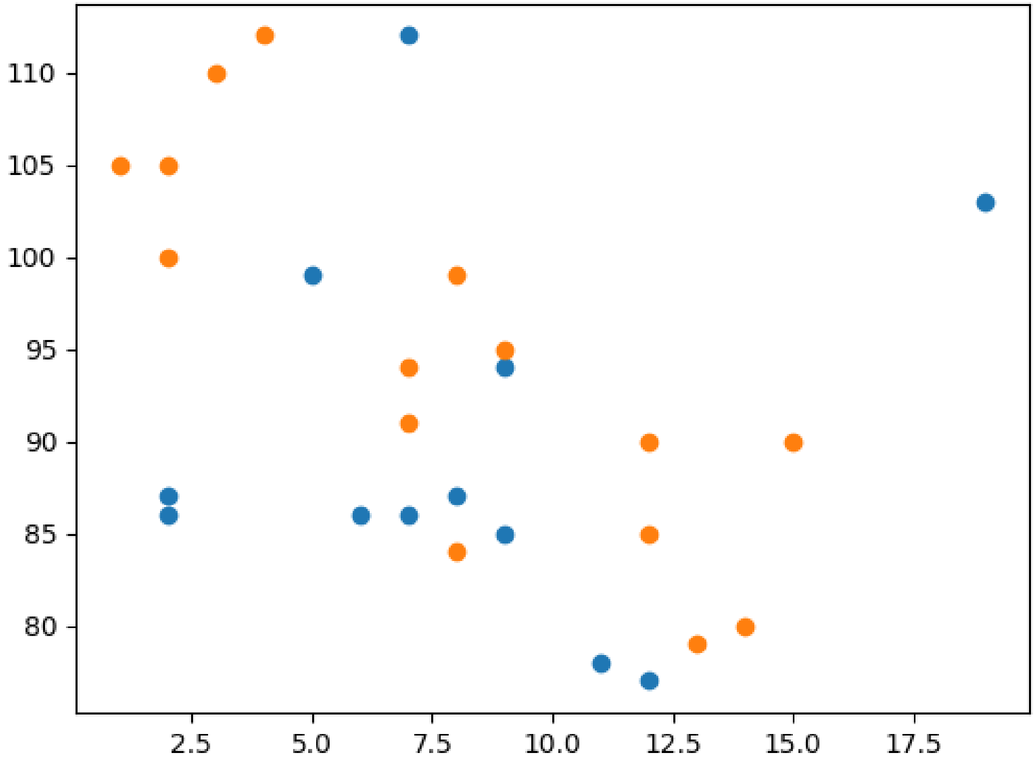
Support vector regression

Approximates a continuous function by identifying a hyperplane that best represents the data's structure; valuable in various applications, including financial market prediction.

```
import numpy [1], [4]
import matplotlib.pyplot as plt
x = numpy.random.normal(5.0, 2.0, 2000)
y = numpy.random.normal(10.0, 3.0, 2000)
plt.scatter(x, y)
plt.show()
```

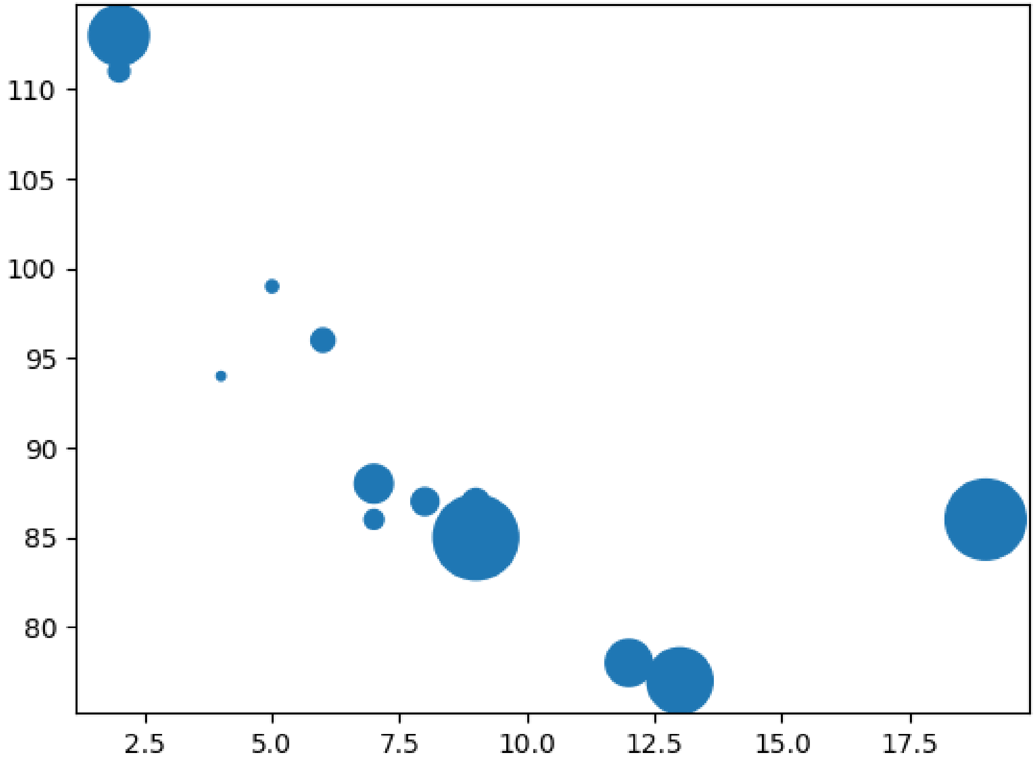


import matplotlib.pyplot as plt [2], [3]
import numpy as np
#day one, the age and speed of 13 cars:
x = np.array([5,7,8,7,2,19,2,9,11,12,9,6])
y = np.array([99,86,87,112,86,103,87,94,78,77,85,86])
plt.scatter(x, y)
#day two, the age and speed of 15 cars:
x = np.array([2,2,8,1,15,8,12,9,7,3,13,4,7,14,12])
y = np.array([100,105,84,105,90,99,90,95,94,110,79,112,91,80,85])
plt.scatter(x, y)
plt.show()



```
import matplotlib.pyplot as plt
import numpy as np

x = np.array([5,7,8,7,2,19,2,9,4,12,13,9,6])
y = np.array([99,86,87,88,113,86,111,87,94,78,77,85,96])
sizes = np.array([20,50,100,200,500,900,60,90,10,300,600,1000,75])
plt.scatter(x, y, s=sizes)
plt.show()
```



REFERENCE

1. T. M. Mitchell, “Machine learning WCB”: McGraw-Hill Boston, MA:, 1997.& matplotlib Scatter w3schools.com.
2. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surveys (CSUR). 2002;34(1):1–47.
3. Sinclair C, Pierce L, Matzner S. An application of machine learning to network intrusion detection. In: Computer Security Applications Conference, 1999. (ACSAC’99) Proceedings. 15th Annual; 1999. p. 371–7. IEEE.
4. Sahami M, Dumais S, Heckerman D, Horvitz E. A Bayesian approach to filtering junk e-mail. In: Learning for Text Categorization: Papers from the 1998 workshop, vol. 62; 1998. p. 98–105. Madison, Wisconsin.
5. Aleskerov E, Freisleben B, Rao B. Cardwatch: A neural network based database mining system for credit card fraud detection. In: Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997; 1997. p. 220–6. IEEE.
6. E, Kim W, Lee Y. Combination of multiple classifiers for the customer's purchase behavior prediction. Decis Support Syst. 2003;34(2):167–75.
7. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. In: Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on; 2008. p. 108–15. IEEE.
8. Joachims T. Making large-scale SVM learning practical. SFB 475: Komplexitätsreduktion Multivariaten Datenstrukturen, Univ. Dortmund, Dortmund, Tech. Rep. 1998. p. 28.
9. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106.
10. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Informat. 2006;2:59–77.

11. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. In: Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on; 2008. p. 108–15. IEEE.
12. Joachims T. Making large-scale SVM learning practical. SFB 475: Komplexitätsreduktion Multivariaten Datenstrukturen, Univ. Dortmund, Dortmund, Tech. Rep. 1998. p. 28.
13. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Informat. 2006;2:59–77.
14. I. Rish, "An empirical study of the naive Bayes classifier," in IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, vol. 3, 22, pp. 41–46: IBM New York.
15. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5(4):115–33.
16. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. FASEB J. 2008;22(2):338–42.
17. Borah MS, Bhuyan BP, Pathak MS, Bhattacharya P. Machine learning in predicting hemoglobin variants. Int J Mach Learn Comput. 2018;8(2):140–3.
18. Ayer T, Chhatwal J, Alagoz O, Kahn CE Jr, Woods RW, Burnside ES. Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics. 2010;30(1):13–22.
19. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IEEE Access. 2017;5:8869–79.
20. Seltman, Howard J. (2008-09-08). Experimental Design and Analysis (PDF). p. 227.
21. "Statistical Sampling and Regression: Simple Linear Regression". Columbia University. Retrieved 2016-10-17. When one independent variable is used in a regression, it is called a simple regression;(...)
22. Lane, David M. Introduction to Statistics (PDF). p. 462.
23. Zou KH; Tuncali K; Silverman SG (2003). "Correlation and simple linear regression". Radiology. 227 (3): 617–22. doi:10.1148/radiol.2273011499. ISSN 0033-8419. OCLC 110941167. PMID 12773666.
24. Altman, Naomi; Krzywinski, Martin (2015). "Simple linear regression". Nature Methods. 12 (11): 999–1000. doi:10.1038/nmeth.3627. ISSN 1548-7091. OCLC 5912005539. PMID 26824102. S2CID 261269711.
25. Kenney, J. F. and Keeping, E. S. (1962) "Linear Regression and Correlation." Ch. 15 in Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, pp. 252–285.
26. Douglas C Montgomery, Elizabeth A Peck and G Geoffrey Vining, Introduction to linear regression analysis, John Wiley & Sons, vol. 821, 2012. Kaggle, GitHub, IEEE, https://www.w3schools.com/python/python_ml_polynomial_regression.asp.

BIOGRAPHY

I, **Buddh Bhagwan Sahu**, **B.Tech/M.Tech in Computer Technology** and **Application** in the **Computer Science and Engineering** at present I am working as a **assistant professor** at **Columbia Institute of Engineering and Technology**, Raipur-Chhattisgarh. With the support cum Simulation/Implementation and teamwork's are active **6th semester 9th students** (CS&E, Dept.). We are all hard working to execute on timely our project. That project is purely research and guided by **Buddh Bhagwan Sahu** (More than 8 years excellent teaching experience) and well suited team.