IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

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-01
Introductions of Learning Technique and Prediction

Abstract*:* In this chapter our main motive to introduce about implementation beginning. Artificial Intelligence is one and only way to define everything such as man power, machine capacity, honesty, accuracy, and medical science related diagnosis, treatment, precautions, preventions, future on visions. AI based disease prediction technology (DPT-AI) can give the enormous route to give healthy life. Initially we will execute our disease prediction technology in manual version. The dependency on computer-based technology has resulted in storage of lot of electronic data in the health care industry.
As a result of which, health professionals and doctors are dealing with demanding situations to research signs and symptoms correctly and perceive illnesses at an early stage. However, Machine Learning technologies have been proven beneficial in giving an immeasurable platform in the medical field so that health care issues can be resolved effortlessly and expeditiously. Disease Prediction is a Machine Learning based system which primarily works according to the symptoms given by a user. The disease is predicted using python and many supportive programming language algorithms & comparison of the datasets with the symptoms provided by the user.

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.

1. INTRODUCTION

Accurate and exact data analysis of any health-related problem is important for the prevention and treatment of the illness. The traditional way of diagnosis may not be sufficient in the case of a serious ailment. Developing a medical diagnosis system based on machine learning (ML) algorithms for prediction of any disease can help in a more accurate diagnosis than the conventional method. It then provides for the appropriate diagnosis and treatment of the disease based on its constantly updated database, which can be developed as an application-based or website-based platform. We have designed a disease prediction system using multiple ML algorithms [10]. The data set used had more than 100 diseases for processing. Based on the health, fitness, food routines, daily routines, symptoms, age, and gender of an individual, the diagnosis system gives the output as the disease that the individual might be suffering from [11]. From the clinical, hospital perspective, delivering appropriate care requires access to patient information in a manner that is useful for beginning diagnosis and treatment. These diagnoses and treatments need to be dynamic and personalized. The weighted KNN algorithm gave the best results as compared to the other algorithms. The accuracy of the weighted KNN algorithm for the prediction will feature based on weak learner and strong learner. Our diagnosis model can act as a doctor for the early diagnosis of a disease to ensure the treatment can take place on time and lives can be saved [12][9].

2. DATA INDEXES

To implement these projects we need to arrange more valuable and specific information, knowledge including raw material areas [8]. This is the most powerful information because it is our manual version AI Disease Prediction Technology (DPT-AI). Once we create DPT-AI will give 100% accurate and valuable results [13]. Before execution or implementation we have to collect information, war data of the following sectors i.e.

1. Food Routines.
2. Drink Routines.
3. Gym Routines.
4. Breathe Routines.
5. Tech Routines.
6. Sleep Routines.
7. Speak Routines cum Habit.
8. Reading Habits.
9. Hair Sample cum Information's.
10. Bone Samples cum Information's.
11. Soil Information's.
12. Leaving Location Information's.
13. Physical Relationship Information's.
14. See on Watch Habits.
15. Healthy Fitness Information's.
16. Age Death Information's.
17. Blood Group Information's.
18. Brain cum Psychic Information's.
19. Moral, Ethical Information's.
20. Air Information's.

On the based on above information The use of deep learning and machine learning (ML) in medical science is increasing, particularly in the visual, audio, and language data fields [14]. We aimed to build a new optimized ensemble model by blending a DNN (deep neural network) model with two ML models for disease prediction using laboratory test results. 100 of attributes (diagnostic/tests) were selected from datasets based on value counts, clinical importance-related features, and missing values/data's [7].

We collected sample datasets on 10000 cases, with 50000 or more laboratory or clinical test/information cum results. We investigated a total number of 100 specific diseases based on the International Classification of Diseases including World Health Organization data's [15]. The deep learning and ML models collected differences in predictive power and disease classification patterns. We used a confusion/probability matrix and analyzed feature importance using the Supervised and Unsupervised learning based values. Our advance ML model achieved high efficiency of disease prediction through classification of diseases. This dataset will be useful in the prediction and diagnosis/precautions of diseases. ML algorithms are used in a broad range of domains, including chemical, biological, geological-genomics Deep learning (DL) is a subset of ML that differs from other ML processes in many ways [16].

Mostly ML models perform well due to their custom-designed representation and input features. Using the input data generated through that process, advance ML learns algorithms, optimizes the weights of each feature, and optimizes the final prediction. DL attempts to learn multiple levels of representation using a hierarchy of multiple layers [17].

3. META-DATA/BRIEFING

To give accurate results we need to explore and arrange that necessary and valuable information by allocated department and from web based information's cum raw materials [18].

Food Routines

Electronically data (1000 different food routines) arranged in stative way.

Drink Routines

Electronically data (1000 different drink routines) arranged in stative way.

Gym Routines

Electronically data (1000 different gym routines) arranged in stative way.

Bath Routines

Electronically data (500 different bath routines) arranged in stative way.

Tech Routines

Electronically data (3000 different tech uses routines) arranged in stative way [6].

Sleep Routines

Electronically data (200 different sleeping routines & Positions) arranged in stative way.

Speak Routines cum Habit

Electronically data (100 different speak routines and habit) arranged in stative way.

Reading Habits

Electronically data (100 different reading routines and habit) arranged in stative way.

Hair Sample cum Information's

Electronically data (1000 different type's hair) arranged in stative way.

Bone Samples cum Information's

Electronically data (100 different results of bone sample) arranged in stative way.

Soil Information's

Electronically data (100 different results of soil sample) arranged in stative way.

Leaving Location Information's

Electronically data (50 different leaving/residential details) arranged in stative way.

Physical Relationship Information's

Electronically data (100 different way of physical relationship habit) arranged in stative way.

See on Watch Habits

Electronically data (100 different seen, staring and watch habit) arranged in stative way.

Healthy Fitness Information's

Electronically metadata based on WHO standard health information to be arranged in stative way.

Age Death Information's

Electronically metadata (Based on past death scene or way) arranged in stative way including postpartum [20].

Blood Group Information's

Electronically metadata based on WHO standard information of blood group to be arranged in stative way.

Brain cum Psychic Information's

Electronically metadata based on WHO standard healthy or general mental information to be arranged in stative way.

Moral, Ethical Information's

Electronically metadata (Decision technique or based on moral, ethical information should be more than 100 ways) arranged in stative way.

Air Information's

Electronically metadata based on Environment & geological standard air information to be arranged in stative way.

4. DATA SIMULATIONS

Above meta-data will be simulate on the machine learning pattern to give manual accurate results by using Machine Learning Algorithms and its scientific theorems [5][4]. It's a purely inspired by Mean Median Mode, Standard Deviation, Percentile, Data Distribution, Normal Data Distribution, Scatter Plot, Linear Regression, Polynomial Regression, Multiple Regression, Scale, Scalars, Train/Test, Decision Tree, Confusion Matrix, Hierarchical Clustering, Logistic Regression, Grid Search, Categorical Data, K-means, Bootstrap Aggregation, Cross Validation, AUC - ROC Curve, K-nearest neighbors. Some of the theorems are here [21].

A simple scatter plot to understand disease identifying pattern:

```
import matplotlib.pyplot as plt
import numpy as np
x1 = np.array([9,7,8,7,2,17,2,9,4,13,12,5,6])
y1 = np.array([101,86,87,88,99,86,103,87,94,78,77,85,86])
plt.scatter(x1, y1)
plt.show()
```

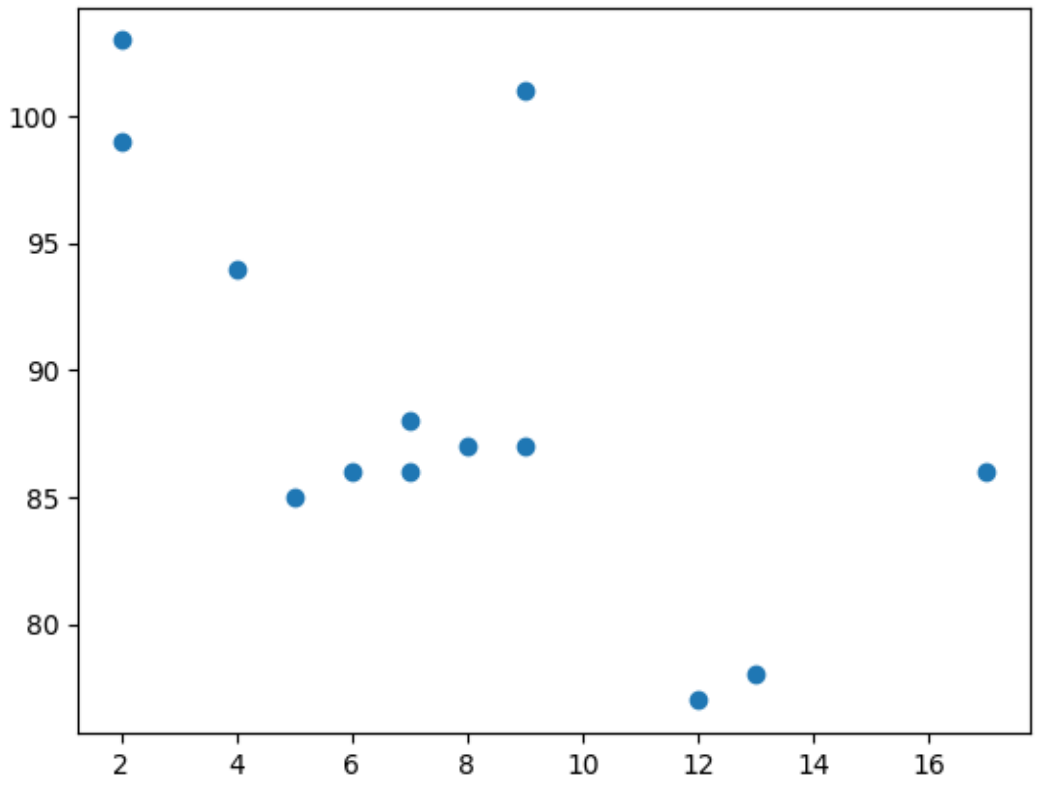


Fig.: A simple scatter plot

Compare Plots & Draw two plots on the same figure to be a relationship between people and age, but what if we plot the observations from another day as well? Will the scatter plot tell us something else[3]?

```
import matplotlib.pyplot as plt
import numpy as np
#day one, the age and age of 13 peoples:
x = np.array([5,7,8,7,2,19,2,9,4,13,12,9,6])
y = np.array([98,76,87,88,101,86,105,87,94,78,77,75,96])
plt.scatter(x, y)
#day two, the age and age of 15 peoples:
x = np.array([2,2,8,1,15,8,12,9,7,3,11,4,7,14,12])
y = np.array([105,100,84,105,90,99,90,95,94,105,79,113,91,80,95])
plt.scatter(x, y)
plt.show()
```

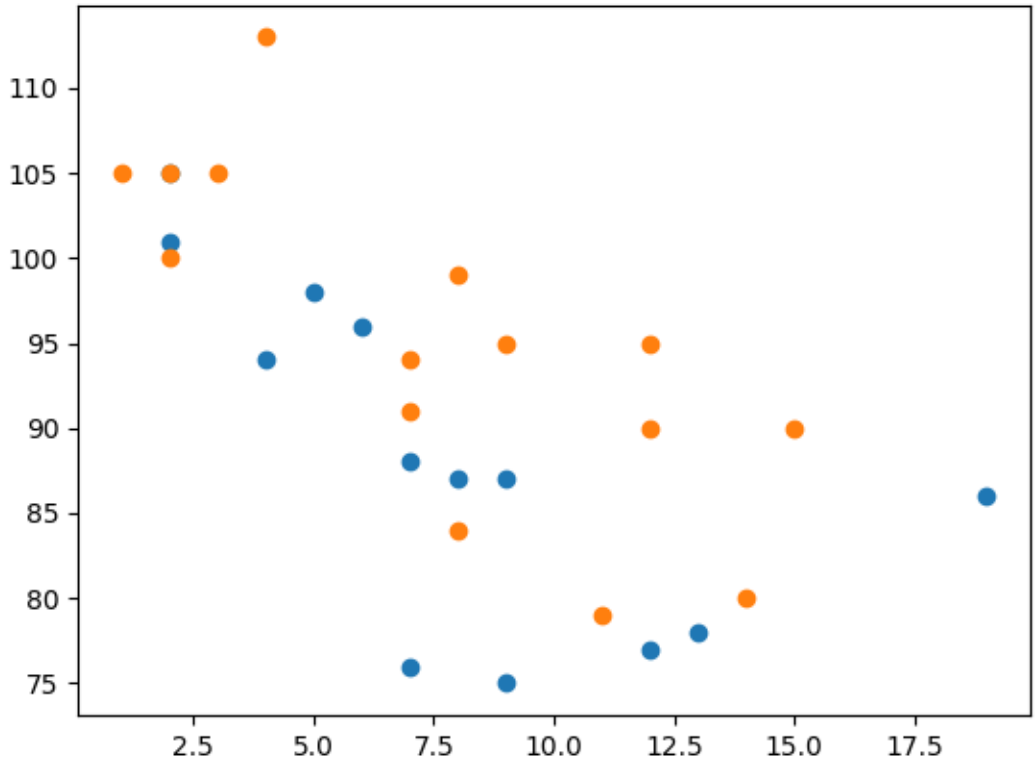


Fig.: Relationship between people and age

The two plots are plotted with two different colours, which are denoted by blue and orange. By comparing the two plots, I think it is safe to say that they both give us the same conclusion [22][23].

5. METHODS AND IMPLEMENTATIONS

We applied DL and ML models to laboratory data (features). For DL, we used a neural network with two hidden layers, with Relu/Sigmoid as the activation function for the input and hidden layers and Soft-max as the activation function for the output layer [2]. We tried to improve performance through a hyper-parameter optimization process. In general, deepening the neural network layer caused the serious problem of gradient vanishing. In this study, the use of three or more hidden layers caused gradient vanishing and over-fitting problems, such as validation loss, to increase. It's here we need to data collection and pre-processing [22].

Here we are plotting as an argument to differentiate between two things i.e. age and people.

```
import matplotlib.pyplot as plt
import numpy as np
x = np.array([5,7,6,7,2,19,2,9,4,13,12,9,6])
y = np.array([98,86,87,88,101,86,103,87,94,78,77,85,86])
plt.scatter(x, y, color = 'green')
x = np.array([2,2,8,1,15,8,12,9,7,3,11,4,7,14,12])
y = np.array([101,115,84,105,90,99,90,93,94,100,79,112,91,81,95])
plt.scatter(x, y, color = 'purple')
plt.show()
```

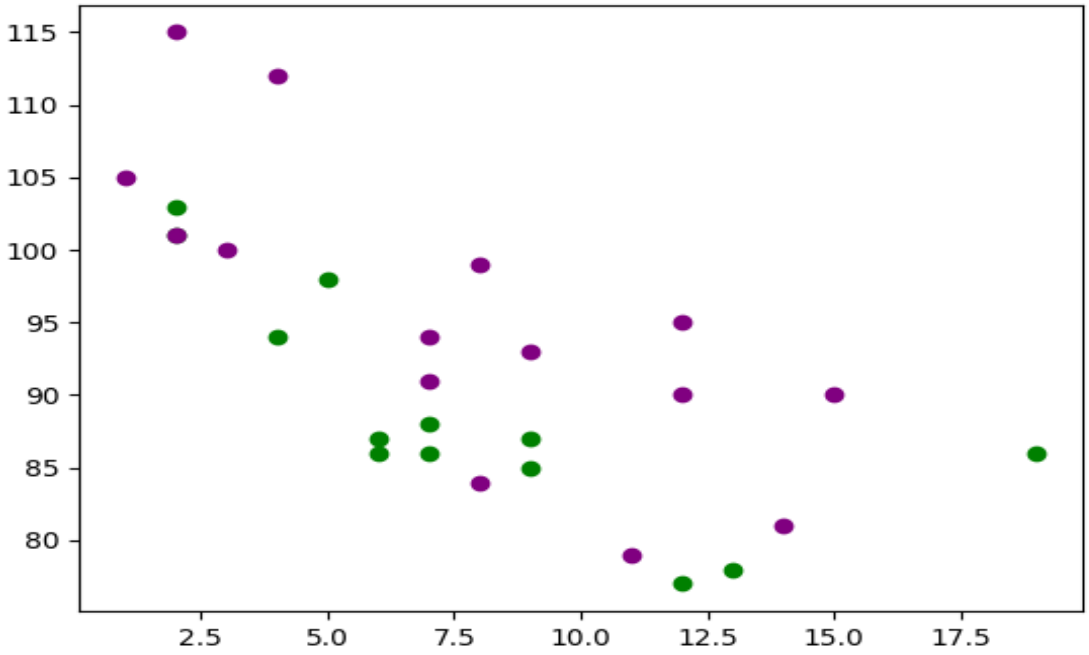


Fig.: Differentiate between two things

To differentiate three different things by multiple colours we can even set a specific colour for each dot by using an array [23][24].

```
import matplotlib.pyplot as plt
import numpy as np
x1 = np.array([5,7,9,7,2,17,2,9,4,13,12,9,6])
y1 = np.array([98,85,87,88,101,86,105,87,94,78,77,95,85])
colors = np.array(["red","green","blue","yellow","pink","black","orange","purple","beige","brown","gray","cyan","magenta"])
plt.scatter(x1, y1, c=colors)
plt.show()
```

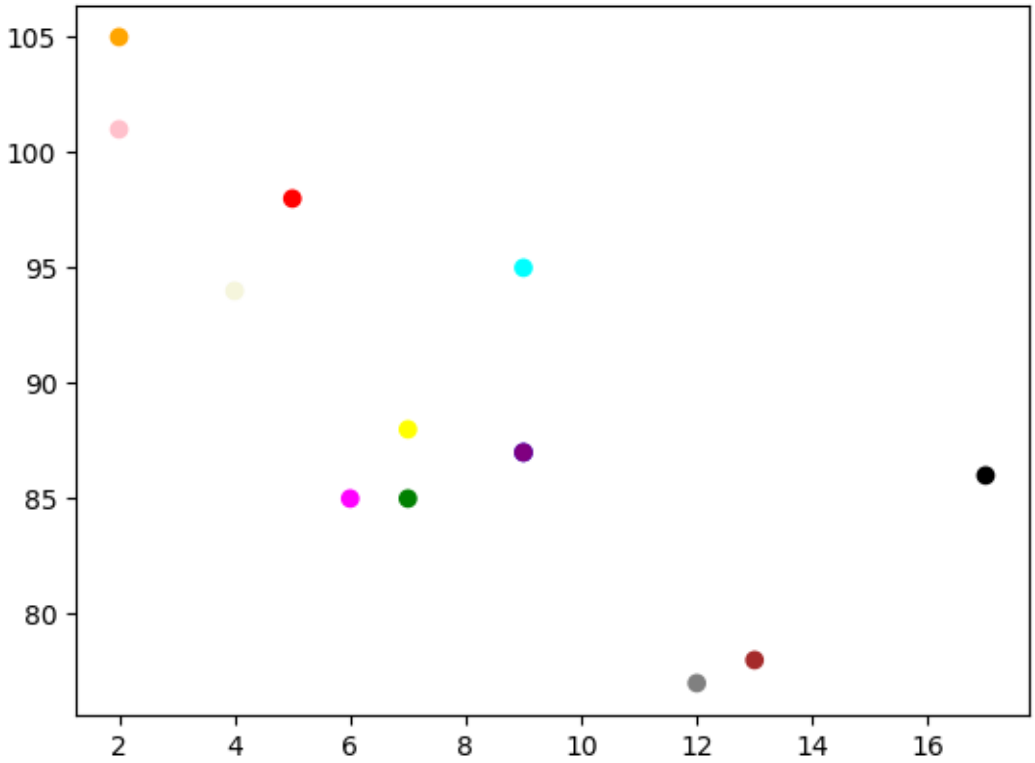


Fig.: set a specific colour for each dot by using an array.

Now we will try to understand by it is by colours, colour-map drawing of thee different things is here [24][1].

```
import matplotlib.pyplot as plt
import numpy as np
x1 = np.array([8,7,6,7,2,19,2,9,4,13,12,9,6])
y1 = np.array([98,86,87,88,101,86,105,87,94,78,77,95,85])
colors = np.array([0, 10, 20, 30, 40, 45, 50, 56, 60, 70, 80, 90, 103])
plt.scatter(x1, y1, c=colors, cmap='viridis')
plt.colorbar()
plt.show()
```

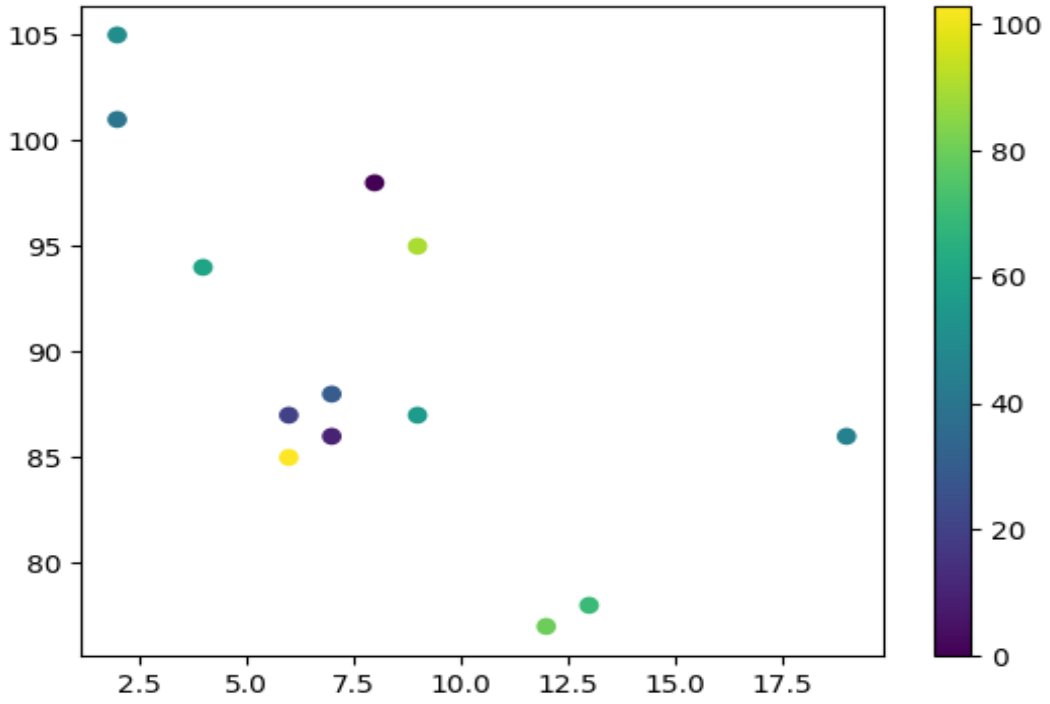


Fig.: Colour-map drawing of thee different things

6. ALGORITHMS

Regression is the term of used when you try to find the relationship between variables. In Machine Learning, and in statistical modelling, that relationship is used to predict the outcome of future events. Linear regression is the one of the regression techniques and its uses the relationship between the data-points to draw a straight line through all them. This line can be used to predict future values. In Machine Learning, predicting the future is very important like the x1-axis represents age, and the y1-axis represents peoples. We have registered the age and peoples of 13 peoples as they were passes after 80 years. Let us see if the data we collected could be used in a linear regression. Here linear regression would not be the best method to predict future values is it true? Let's see it is good fit or bad fit. To predict the values and results we have so many methods to get accurate and trustable results [25].

```
from scipy import stats

x1 = [5,7,8,7,2,17,2,9,4,11,12,9,6]
y1 = [99,86,87,88,111,86,103,87,94,78,77,85,86]
slope, intercept, r, p, std_err = stats.linregress(x1, y1)
def myfunc(x1):
  return slope * x1 + intercept
age = myfunc(10)
print(age)
```

OUTPUT: 85.59 (Predicted age at 85.59, which we also could read from the diagram as below.)

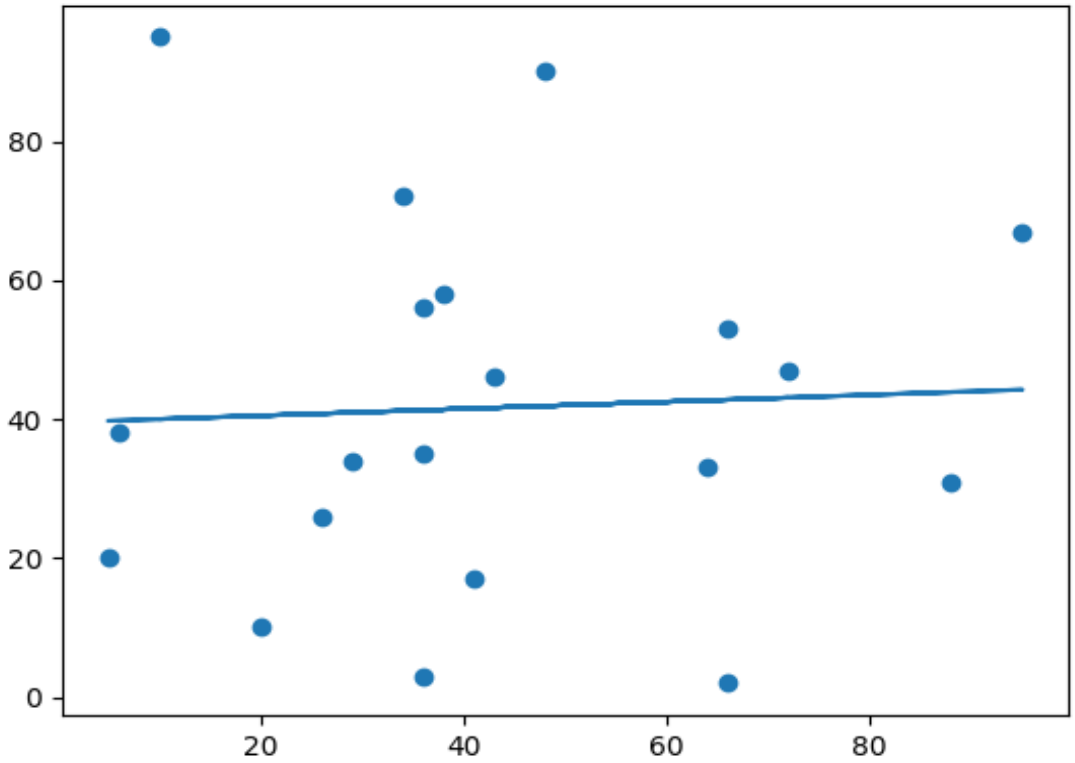


Fig.: These values for the x- and y-axis should result in a very bad fit for linear regression

7. USES OF PROGRAMMING LANGUAGES & METHODS

Using python, Kaggle, GitHub and so many resources have to implement disease prediction methods. Here we are using following methods and techniques. We are classifying in three groups to the uses programming languages and its methods. We have so many but in this chapter we are describing that only. The accuracy of each predicted value is measured by its squared residual vertical distance between the point of the data set and the fitted line.

Group-A

1. append() :Adds an element at the end of the list.
2. clear() :Removes all the elements from the list.
3. copy() :Returns a copy of the list.
4. count() :Returns the number of elements with the specified value.
5. extend() :Add the elements of a list (or any iterable), to the end of the current list.
6. index() :Returns the index of the first element with the specified value.
7. insert() :Adds an element at the specified position.
8. pop() :Removes the element at the specified position.
9. remove() :Removes the first item with the specified value.
10. reverse() :Reverses the order of the list.
11. sort() :Sorts the list [29][30].

Group-B

12. clear() :Removes all the elements from the dictionary.
13. copy() :Returns a copy of the dictionary.
14. fromkeys() :Returns a dictionary with the specified keys and value.
15. get() :Returns the value of the specified key.
16. items() :Returns a list containing a Tuple for each key value pair.
17. keys() :Returns a list containing the dictionary's keys.
18. pop() :Removes the element with the specified key.
19. popitem() :Removes the last inserted key-value pair [28][29].

Group-C

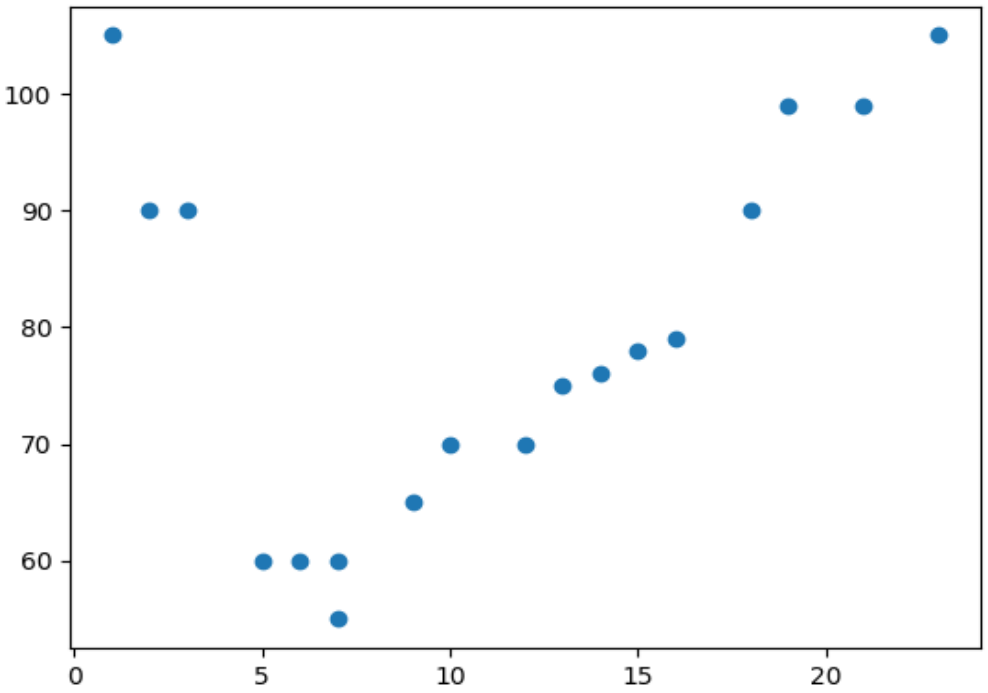
20. setdefault() :Returns the value of the specified key (If the key does not exist, Insert the key, with the specified value).
21. update() :Updates the dictionary with the specified key-value pairs.
22. values() :Returns a list of all the values in the dictionary.
23. count() :Returns the number of times a specified value occurs in a Tuple [27][26].

8. METHEMATICAL SIMULATION

Simple linear regression & polynomial regression aims to find a linear relationship to describe the correlation between an independent and possibly dependent variable. The regression line can be used to predict or estimate missing values, this is known as interpolation. The calculation is based on the method of least squares [31]. The idea behind it is to minimise the sum of the vertical distance between all of the data points and the line of best fit. If your data points clearly will not fit a linear regression a straight line through all data points, it might be ideal for polynomial regression. Polynomial regression, like linear regression, uses the relationship between the variables x and y to find the best way to draw a line through the data points. Python has methods for finding a relationship between data-points and to draw a line of polynomial regression. We will show you how to use these methods instead of going through the mathematic formula.

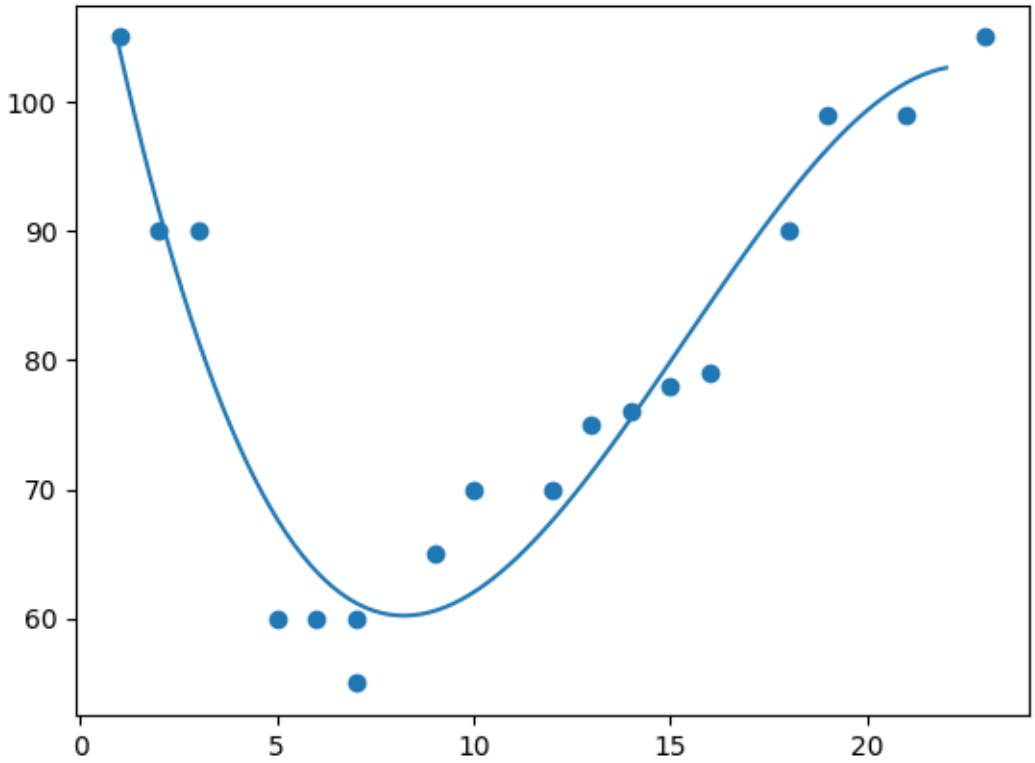
Start by drawing a scatter plot:

```
import matplotlib.pyplot as plt
x = [1,2,3,5,6,7,7,9,10,12,13,14,15,16,18,19,21,23]
y = [105,90,90,60,60,55,60,65,70,70,75,76,78,79,90,99,99,105]
plt.scatter(x, y)
plt.show()
```



To better understand polynomial regression we will import **NumPy** and **Matplotlib** then draw the line of Polynomial Regression [32]:

```
import numpy
import matplotlib.pyplot as plt
x = [1,2,3,5,6,7,7,9,10,12,13,14,15,16,18,19,21,23]
y = [105,90,90,60,60,55,60,65,70,70,75,76,78,79,90,99,99,105]
#Create the arrays that represent the values of the x and y axis
mymodel = numpy.poly1d(numpy.polyfit(x, y, 3))
#NumPy has a method that lets us make a polynomial model
myline = numpy.linspace(1, 23, 105)
#Then specify how the line will display, we start at position 1, and end at position 23
plt.scatter(x, y)
#Draw the original scatter plot
plt.plot(myline, mymodel(myline))
#Draw the line of polynomial regression
plt.show()
#Display the diagram
```



Now we will Import the modules as per we need. So create the arrays that represent the values of the x and y axis based on above data we can use the information we have gathered to predict future values for health parameter for example **breath bph at 18:00pm** like here [33].

```
import numpy
from sklearn.metrics import r2_score
x = [1,2,3,5,6,7,7,9,10,12,13,14,15,16,18,19,21,23]
y = [105,90,90,60,60,55,60,65,70,70,75,76,78,79,90,99,99,105]
mymodel = numpy.poly1d(numpy.polyfit(x, y, 3))
health = mymodel(18)
print(health)
```

OUTPUT: 92.79bph
On the basis of above values for the x- and y-axis should result in a very bad fit for polynomial regression look at the below graph [34].

```
import numpy
import matplotlib.pyplot as plt
x = [89,43,36,36,95,10,66,34,38,20,26,29,48,64,6,5,36,66,72,50]
y = [23,46,3,35,67,95,53,72,58,10,26,34,90,33,38,20,56,2,47,17]
mymodel = numpy.poly1d(numpy.polyfit(x, y, 3))
myline = numpy.linspace(3, 95, 101)
plt.scatter(x, y)
plt.scatter(x, y, color = 'purple')
plt.plot(myline, mymodel(myline))
plt.show()
```

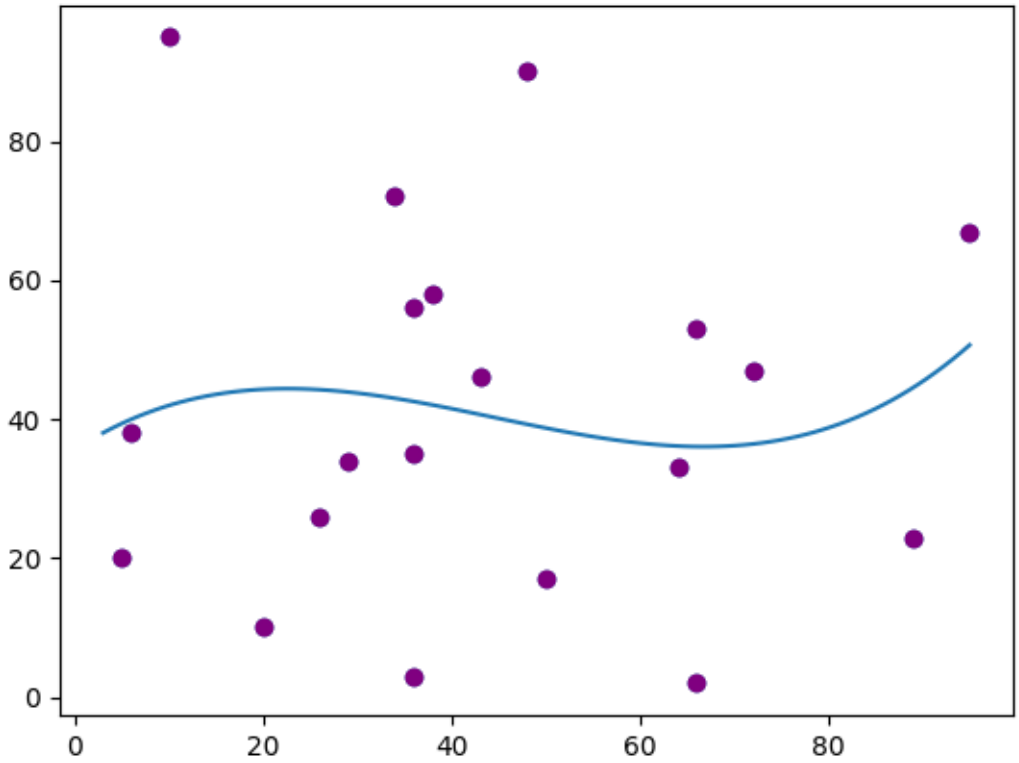


Fig.: Examples to understand very badly fit for polynomial regression

REFERENCE

1. T. M. Mitchell, “Machine learning WCB”: McGraw-Hill Boston, MA:, 1997.
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.
27. Casella, G. and Berger, R. L. (2002), "Statistical Inference" (2nd Edition), Cengage, ISBN 978-0-534-24312-8, pp. 558–559.
28. Valliant, Richard, Jill A. Dever, and Frauke Kreuter. Practical tools for designing and weighting survey samples. New York: Springer, 2013.
29. Scott A. Czepiel, "Maximum likelihood estimation of logistic regression models: theory and implementation", Conference Proceedings, 2002.
30. Xin Yan and Xiaogang. Su, Linear regression analysis: theory and computing, World Scientific, 2009.
31. P. R. J. Campbell and K. Adamson, "Methodologies for Load Forecasting", 2006 3rd International IEEE Conference Intelligent Systems, pp. 800-806, 2006.
32. Mario Oliveira, D. Marzec, G. Bordin, Arturo Bretas and D. Bernardon, "Climate change effect on very short-term electric load forecasting", 2011 IEEE PES Trondheim PowerTech: The Power of Technology for a Sustainable Society POWERTECH 2011, 2011.
33. T. Vantuch, A. G. Vidal, A. P. Ramallo-González, A. F. Skarmeta and S. Misák, "Machine learning based electric load forecasting for short and long-term period", 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), pp. 511-516, 2018.
34. Stanislav Eroshenko, K. Vinokurov and A. Smolina, Electrical load forecasting, vol. 190, pp. 299-305, 2014.

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 **5th semester 21th 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.