Comparison of performances of regression model-based prediction of meteorological conditions.

VEJENDLA, R.S., PAVULURI, B.L., VENIGANDLA, N., TINNAVALLI, D. and BANO, S.

2022

Comparison of Performances of Regression Model-Based Prediction of Meteorological Conditions

Ramya Sree Vejendla[1*], Bhagya Lakshmi Pavuluri[1], Nandini Venigandla[1], Deepika Tinnavalli[1], and Shahana Bano[1]

1 Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India
*** Corresponding author**

Abstract This paper clearly illustrates the working of different machine learning algorithms to determine the weather conditions. This involves the prediction of temperature by training with the pre-existing dataset of weather conditions on each day for around 40 years. This trained data is tested to evaluate the temperature of a certain day in the upcoming calendar by date or year. This illustration describes the comparison between different algorithm results and determines the most efficient algorithm. The algorithm involved were Linear Regression, Logistic Regression, and Clustering. These three algorithms involve different mechanisms such as predicting based on mean, probability, and grouping based on similar constraints. The model helps to select the most efficient algorithm which gives the approximate values nearer to accurate values. Though all the techniques involved in previous analysis are mostly based on mean analysis the result is almost approximate but under logistic regression, it either gives almost the accurate result or the wrong result. Here we introduce clustering since the date or year could be grouped under a certain condition where either based on the temperature of a certain year or the season.

Keywords:- Unsupervised data; Linear regression; Logistic regression; My Logit function

1 Introduction

Here in this work, the algorithms considered were Regression analysis implementing both linear regression and logistic regression and also clustering technique i.e. K-means clustering. Weather prediction[5] is all a different concept as we in the first place could predict that usually due to global warming and all the temperature rise from year to year increases. But the thing matters is at what rate does it affect and the range where it could lie and also the seasonal differences. Apart from this if

sudden calamities occur how it could affect the weather i.e., in terms of temperature, humidity, precipitation, etc. To predict and analyse certain adverse effects we implemented three different algorithms that were linear regression, logistic regression, and clustering.

1.1 Linear Regression

Linear regression is the linear representation of two different scalar values of which one is a dependent variable while the other is an independent variable. Here we predict[1] the unknown variable with respect to the other known variables. Here in linear regression, the linear procedure helps to predict the rate of change of a dependent variable and the regression describes the point values or the approximate values. Moreover, the linear regression is linear[3] not because of the rate of change is linear that is x is constantly dependent on y but because the theta or the relation is linear. There are different regression analysis[8] besides linear based on the number of dependent variables and independent variables such as multiple linear regression, ordinal regression, or multinomial regression.

1.2 Logistic Regression

Logistic Regression is used to predict[12,13] the data and explain the relationship between one variable with one or more ordinal and nominal variables. It is a predictive analysis[2] and also a static model that uses a logistic function to determine a binary dependent variable, in logistic regression, the outcome is like continuous with many possible values but the outcome has only a limited number of possible values as results. Logistic regression implements the Logit function that is used in the method of classification.

1.3 K-Means Clustering

Clustering is like the most popular partitioning algorithms in clustering is the K-means cluster analysis. It tries to cluster data based on their similarity. Also, there exist specific clusters and the data is grouped into the same clusters. There are a lot of clustering techniques like k-means clustering, hierarchical clustering, and fuzzy clustering[11]. Of all these k-means is most popular as it is easy to implement it over unknown groups of data from complex data sets. And from these unlabelled data the centroids of the k clusters as to label new data.

2 Procedure

2.1 Dataset Extraction

Here the dataset is obtained from the real-world weather conditions and this dataset consists of about 19 variables and thousands of records. The dataset consists of both dependent and independent values to implement three different algorithms as each algorithm is applied to the different number of independent variables. Hence the dataset is accomplished satisfying all these factors.

2.2 Linear Regression

In linear regression analysis the complete working depends on a linear curve that deals with one dependent variable and one independent variable. Here in this apart from fitting the linear line it also performs[10] different steps namely analysing the correlation between the variables and the directionality of data, It estimates the model that is fitting the line for the values as per the axial dimensions and finally, it validates the usefulness of the fitting model. Linear regression also utilizes a linear formula.

$$Y = a + bx$$

where Y is the dependent variable as it depends on X,
X is the independent variable.

2.3 Logistic Regression

Logistic regression determines the relation between a dependent variable and one or more independent variables. These independent variables could be of any value like nominal value, ordinal value, or sometimes a ratio value. This is also similar to the multiple linear regression[7] model. This is obtained based on probability values that are unlike in the linear regression here it uses an exponential function. We implement a maximum likelihood estimation function to obtain the best fit.

$$\ln \subset \frac{p}{1-p} \supset = a + bx$$

$$\frac{p}{1-p} = e^{a+bx}$$

$$p = \frac{e^{a+bx}}{1+e^{a+bx}}$$

2.4 Clustering

Clustering as mentioned divides the unlabelled data into clusters. Initially, we specify the number of clusters and now shuffle the dataset by deriving the centroid of each cluster. Now shuffle the dataset again and again until there is no change observed in the clusters ad data points. And finally, the goal of the k-means clustering is to find groups of data under a label i.e., unlabelled to labelled data.

$$objective\ functiona\ J = \sum_{J=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

where k = Number of clusters, n = Number of cases, xi = case i, cj = Centroid for cluster j.

3 Flow Chart

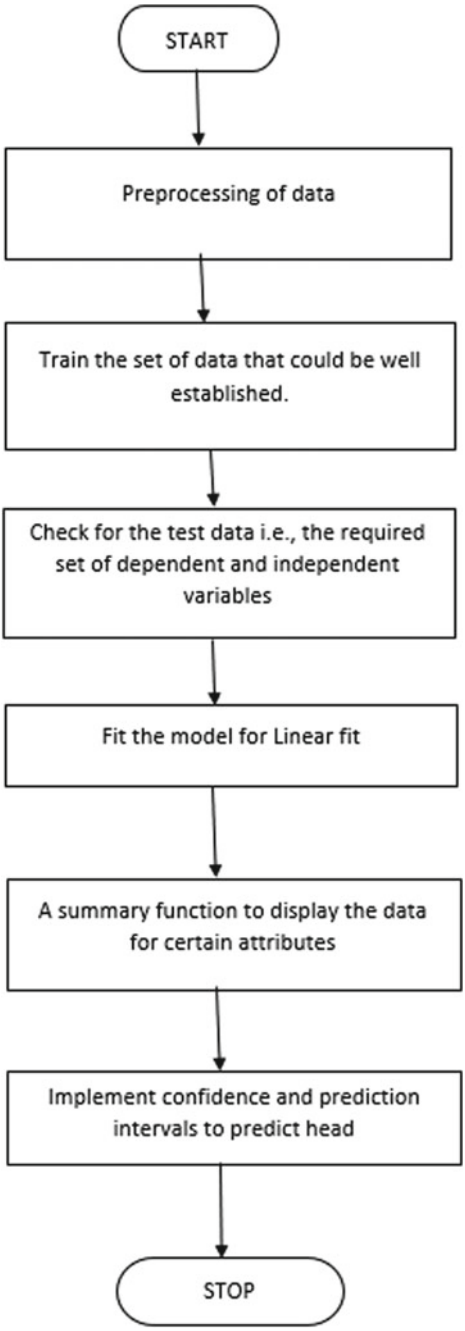
3.1 Linear Regression

To perform certain linear regression analysis initially view the dataset and pre-process the weather dataset. To perform regression analysis[9] the data needs to have both dependent and independent variables and ensure that there is no relation between the two dependent and independent variables. Train the dataset with selected records of data that could be well equipped between different sets of values. Now to get into working fit the data model to linear model that is in the linear pattern. Now to display and check the certain values with a summary function. To predict the head, implement confidence and prediction intervals. Now terminate the function (Fig. 1).

3.2 Logistic Regression

For the working of logistic regression analysis extract the weather dataset[4] and now obtain and analyse the dates and analyse the dates and temperature conditions. As a part of data pre-processing apply data visualization techniques and remove the misclassified data now similar to linear regression utilize the mylogit function to apply the regression analysis and determine the date and temperature analysis. Also, unlike linear regression which uses the simple linear term, it utilizes exponential coefficients with confidence interval fitting. Now to interpret the results by plotting the graph with results (Fig. 2).

Fig. 1 Linear regression flowchart



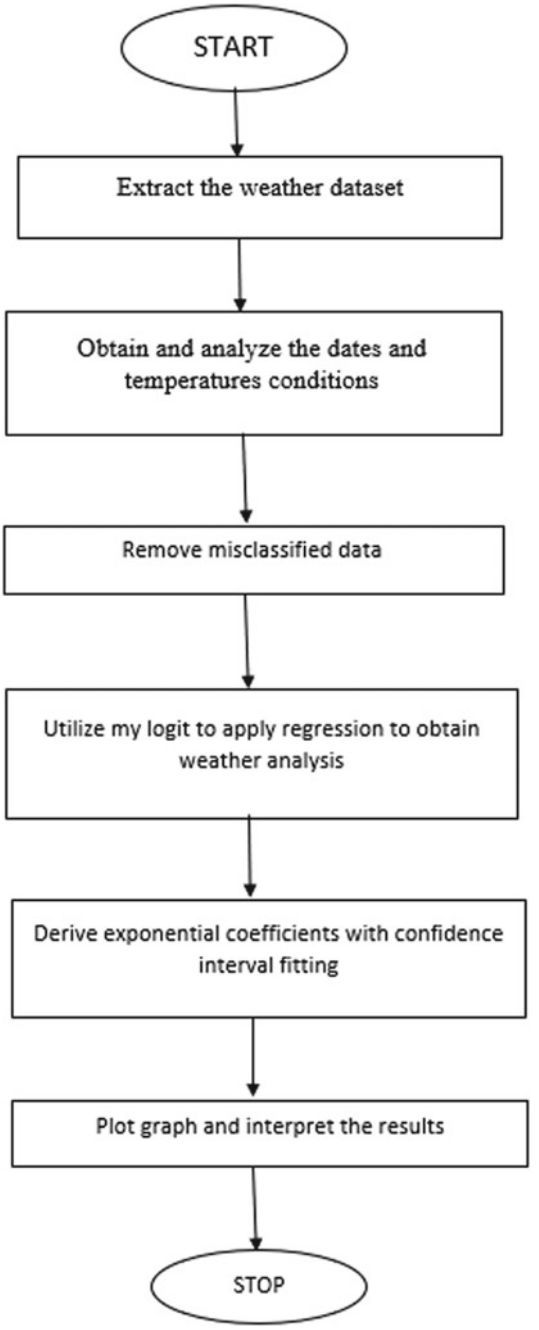


Fig. 2 Logistic regression flowchart

3.3 Clustering

To implement the clustering techniques whichever the algorithm was involved initially install the libraries and view the dataset. Now pre-process the dataset and visualize the data to remove the null variables if any exists. Apply K-means and select any of the well-established algorithms such as Hartigan, Lloyd…. Now if the algorithm exists to fit the methodology proceed with next step otherwise, repeat the application of the algorithm in otherwise conditions. Now obtain the column which you want to select. And now view the obtained clusters or plot the graph and analyse different temperatures and various constraints behind and now terminate the code (Fig. 3).

4 Results

4.1 Linear Regression

The main essence of the regression technique is the data set and here the dataset is a real-time dataset. I considered this dataset because it has a vast number of tuples predicting all the possibilities of a climate. In execution the dataset is completely read and displays the tuple data Now a set of data is set to be trained and the left is to be tested once the data is trained one can fix the data apart from which is trained is to text dataset. Now apply linear regression analysis i.e., lm to the tuples of a dataset whose relation is to be described and now here we get the summary as a quadrant representation with 4 quartiles Describing the minimum, maximum, first quartile, third quartile, and median values also with coefficient error values that is the deviation from expected results end the linear regression analysis to view this aa a plot based on the above-described analysis we plot them with respect to both the tuples selected earlier to fit. A linear fit is obtained describing the relation in a linear pattern (Figs. 4 and 5).

4.2 Logistic Regression

The dataset is the same as that used in linear regression analysis initially the dataset is being pre-processed and being checked for any missing values or out of boundary values. Consider the data and summarize it where u find the quartile values of each attribute involved. Besides I chose x for months from the dataset and then to proceed with the function that returns all the possible combinations of fog or fog temperature and more. And then a set of tuples that is years, months, and days of a year altogether. To interpret the logistic regression we consider the above tuples for a generalized linear model that on summarization gives us deviance values with respect to all

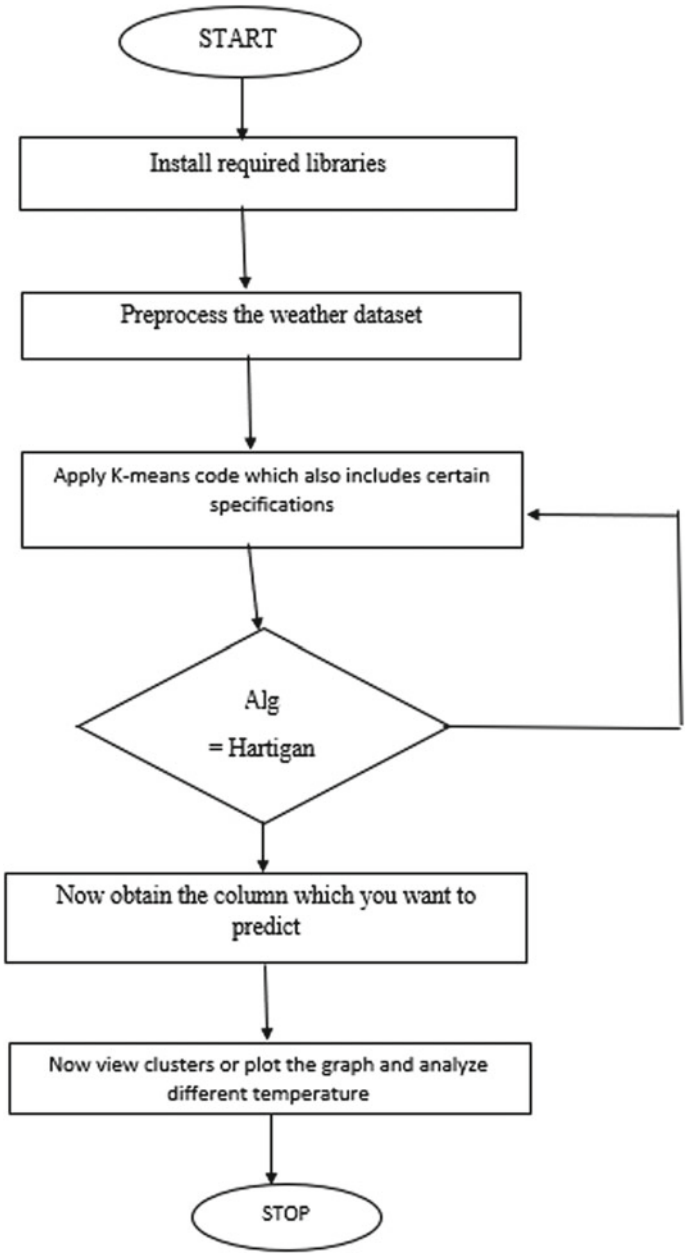


Fig. 3 Clustering flowchart

```
Call:
lm(formula = year ~ high_temp, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5046 -0.4935 -0.4743  0.5065  0.5344

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 2.016e+03  4.096e-02 49225.456   <2e-16 ***
high_temp   4.097e-04  5.572e-04     0.735    0.462
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5001 on 2739 degrees of freedom
Multiple R-squared:  0.0001974, Adjusted R-squared:  -0.0001677
F-statistic: 0.5407 on 1 and 2739 DF,  p-value: 0.4622

> head(predict(lm.fit, test, interval = "confidence"), 5)
        fit      lwr      upr
1  2016.491 2016.472 2016.511
5  2016.492 2016.473 2016.511
8  2016.495 2016.475 2016.515
10 2016.493 2016.474 2016.511
14 2016.494 2016.475 2016.513
> head(predict(lm.fit, test, interval = "prediction"), 5)
        fit      lwr      upr
1  2016.491 2015.511 2017.472
5  2016.492 2015.511 2017.473
8  2016.495 2015.514 2017.476
10 2016.493 2015.512 2017.473
14 2016.494 2015.514 2017.475
```

Fig. 4 Linear regression analysis

different months and also the quartile values for data interpretation. A confint() function gives various data with normality and co-efficiency. Then it describes the exponential result analysis and views the plot of the values in a logistic pattern which might be an s-shaped graph including all the deviations of errors on the analysis (Figs. 6, 7 and 8).

4.3 Clustering

Clustering involves different algorithms of which k-means is most widely used due to its well approximate clustering. Also, in k-means, the initial phenomenon is a bit similar to linear regression where the data is trained. Initially, the data is pre-processed here the data is viewed and the data is viewed and several random sets are selected as a part of set data. We could select the number of clusters required

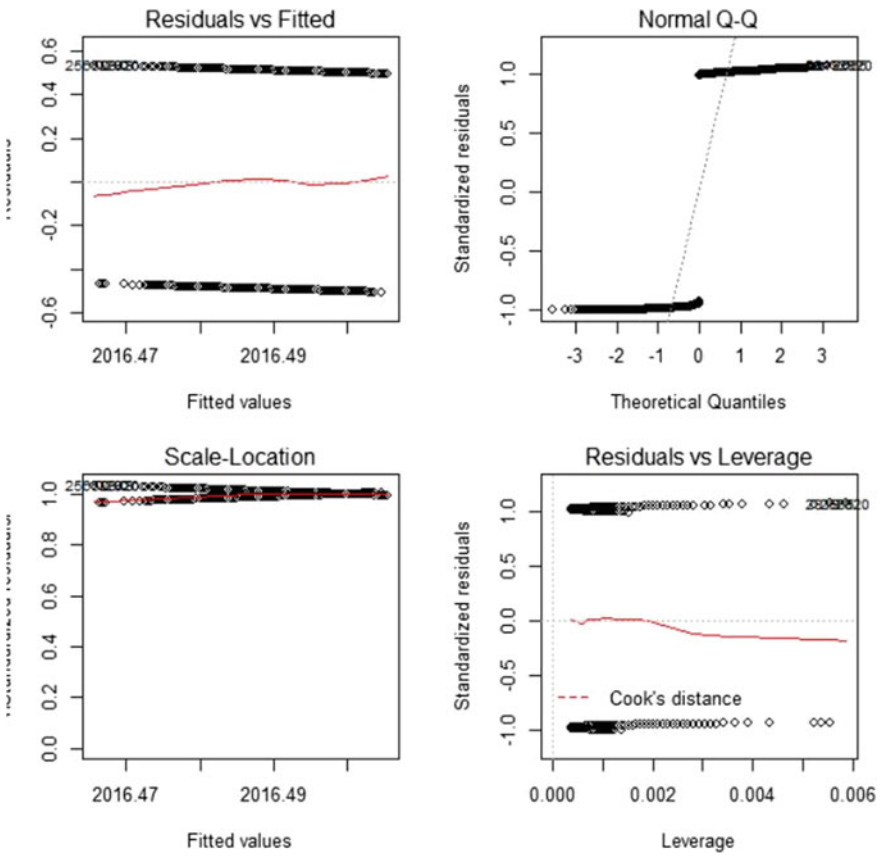


Fig. 5 Plots based on linear regression analysis years versus maximum temperature

in the asset of instruction along with certain k-means methodologies that are Lloyd, Hartigan, and others. Now based on a different technique that cluster interpretation differs and the cluster levels and values could be chosen with regard to certain iterations and all. The cluster records could be selected again and be evaluated with respect to several interpreted methodologies and the clusters are classified (Fig. 9; Table 1).

5 Conclusion

I would like to conclude the whole work in terms of accuracy. So, I performed all the different algorithm analysis on the same dataset with the same number of tuples and records. We implemented Linear regression, Logistic Regression, and K-means Clustering to predict the different range of temperatures i.e., high, medium, and

```
> confint.default(mylogit)
                  2.5 %       97.5 %
(Intercept) -837.55419860 724.72467881
month2        -1.23884332   0.66283679
month3        -0.96564877   0.98233463
month4        -0.83254204   1.18441854
month5        -0.82441257   1.00921423
month6        -0.58905039   1.23751525
month7        -0.37142737   1.49713804
month8        -0.11532125   1.96934994
month9        -0.21936199   1.86878585
month10       -0.98706440   0.74517039
month11       -1.38724105   0.43437541
month12       -1.72704225  -0.02951531
day           -0.01164503   0.03250412
year          -0.35831670   0.41642864
> plot(mylogit)
> exp(coef(mylogit))
 (Intercept)       month2       month3       month4       month5       month6       month7       month8
3.157774e-25 7.497591e-01 1.008378e+00 1.192364e+00 1.096804e+00 1.382969e+00 1.755678e+00 2.526953e+00
      month9      month10      month11      month12          day         year
2.281224e+00 8.860809e-01 6.209946e-01 4.154975e-01 1.010484e+00 1.029482e+00
> exp(cbind(OR = coef(mylogit), confint(mylogit)))
Waiting for profiling to be done...
                    OR     2.5 %   97.5 %
(Intercept) 3.157774e-25 0.0000000      Inf
month2      7.497591e-01 0.2877213 1.978369
month3      1.008378e+00 0.3807192 2.751325
month4      1.192364e+00 0.4387980 3.411961
month5      1.096804e+00 0.4330783 2.777864
month6      1.382969e+00 0.5479568 3.490974
month7      1.755678e+00 0.6849673 4.562578
month8      2.526953e+00 0.9096642 7.627880
month9      2.281224e+00 0.8198710 6.895693
month10     8.860809e-01 0.3638583 2.099856
month11     6.209946e-01 0.2456355 1.552503
month12     4.154975e-01 0.1724799 0.958320
day         1.010484e+00 0.9884654 1.033163
year        1.029482e+00 0.6987137 1.518992
>
```

Fig. 6 Logistic regression analysis

low. Hence the analysis and its results on invariant data determine the error rate and accuracy where logistic regression is most accurate. Here we consider logistics as a good way to optimize and predict the weather or temperature conditions mostly from unlabelled as well as labelled also sometimes.

```
> mydata$month <- factor(mydata$month)
> mylogit <- glm( events~ month + day+ year, data = mydata, family = "binomial")
> summary(mylogit)

Call:
glm(formula = events ~ month + day + year, family = "binomial",
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6355   0.2939   0.3792   0.4402   0.6961

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -56.414760 398.547853  -0.142   0.8874
month2       -0.288003   0.485131  -0.594   0.5527
month3        0.008343   0.496944   0.017   0.9866
month4        0.175938   0.514540   0.342   0.7324
month5        0.092401   0.467771   0.198   0.8434
month6        0.324232   0.465969   0.696   0.4865
month7        0.562855   0.476684   1.181   0.2377
month8        0.927014   0.531814   1.743   0.0813 .
month9        0.824712   0.532701   1.548   0.1216
month10      -0.120947   0.441905  -0.274   0.7843
month11      -0.476433   0.464707  -1.025   0.3053
month12      -0.878279   0.433051  -2.028   0.0425 *
day           0.010430   0.011263   0.926   0.3544
year          0.029056   0.197643   0.147   0.8831
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 801.55  on 1437  degrees of freedom
Residual deviance: 774.47  on 1424  degrees of freedom
  (2217 observations deleted due to missingness)
AIC: 802.47

Number of Fisher Scoring iterations: 5
```

Fig.7 Deviances based on regression analysis

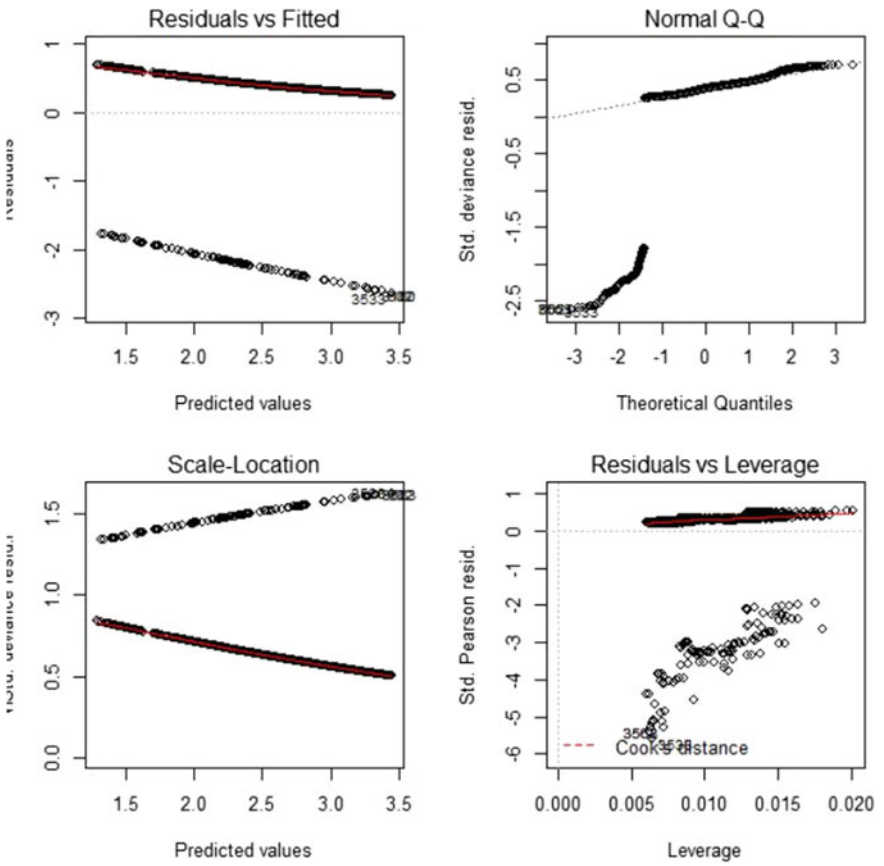


Fig. 8 Plots based logistic regression years versus maximum temperature

```
> kmeans(x,2,iter.max = 10,nstart=1,algorithm=c("Hartigan-Wong","Lloyd","Forgy","MacQueen"),trace=FALSE)
K-means clustering with 2 clusters of sizes 5, 5

Cluster means:
        [,1]       [,2]
1 0.3972067 0.01419765
2 1.1891598 1.18307225

Clustering vector:
 [1] 1 1 1 1 1 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 0.8198241 0.4264660
 (between_SS / total_SS =  80.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
 "iter"
[9] "ifault"
> km=kmeans(iris[1:4],3)
> colnames(x)<-c("x","y")
> x<-rbind(matrix(rnorm(10,sd=0.3),ncol=2),matrix(rnorm(10,mean=1,sd=0.3),ncol=2))
> cl=kmeans(x,2)
> m=kmeans(x,5,nstart=25)
> m=kmeans(x,5,iter.max = 10)
> cl1=kmeans(x,5,algorithm=c("Lloyd"))
> cl1=kmeans(x,5,algorithm=c("MacQueen"))
> cl1
K-means clustering with 5 clusters of sizes 1, 1, 4, 1, 3

Cluster means:
          [,1]       [,2]
1  1.33630701  0.7055838
2  0.47538945  0.3155323
3 -0.03964517 -0.2634254
4  0.99155648  1.0973310
5  0.75428740  0.7864129

Clustering vector:
 [1] 2 3 3 3 3 4 5 5 5 1
```

Fig. 9 Cluster-based on K-means algorithm

Table 1 Comparison of data distribution [6] over different quartiles in all three algorithms

S. No.	Results	Decision tree	Random forest	K-NN
1	Median values	-0.4743	-0.3792	-0.4094
2	Residual error or deviance	55%	80%	61.20219%
3	Accuracy error	50%	80%	0
4	Data handling	Handles simple linear data	Handles nonlinear [3] and unlabelled data	Handles unsupervised and unlabelled data
5	Benefits	Easy to understand	Best results and deals with multiple data	Deals with any group of data i.e., ungrouped or unlabelled data

References

1. Comparative Analysis of Temperature Prediction using Regression Methods and Back Propagation Neural Networks Survey on Weather Forecasting Using Data Mining, Proc. IEEE Conference on Emerging Devices and Smart Systems (ICEDSS 2018) 2–3 March 2018, 978–1–5386–3479–0/18.

2. Weather Analysis to predict rice cultivation time using multiple linear regression to escalate farmers exchange rate, 2017 International Conference on Advanced Informatics, Concepts, Theory and Applications.
3. Cloud based flight delay prediction using logistic regression, 2017 International Conference on Intelligent Sustainable Systems.
4. Numerical weather prediction using nonlinear auto regressive network for the Manaus region Brazil 2017 Innovations in power and Advanced Computing Technologies.
5. Weather monitoring using Artificial Intelligence 2016 2nd International Conference on Computational Intelligence and Networks.
6. Weather Visibility Prediction Based on Multimodal Fusion 2019 IEEE access year.
7. Prediction of Climate Variable using Multiple Linear Regression 2018 4th International Conference on Computing.
8. Rainfall Prediction using Regression Model by IJRTE 12016.
9. Comparative study of different weather forecasting models 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS).
10. Haze weather recognition based on multiple features and Random forest, 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC).
11. Analysis of Weather Prediction using Machine Learning and Big DATA 2018 International Conference on Advances in Computing and Communicational Engineering.
12. Weather Forecasting Using Artificial Neural Network, Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE, 978–1–5386–1974–2/18.
13. Rainfall Prediction based on Deep Neural Network: A Review, Proceedings of the Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020) IEEE.
14. Rainfall Forecasting in Bandung Regency using C4.5 Algorithm, 2018 6th International Conference on Information and Communication Technology (ICOICT).
15. Dynamic Line Rating Using Numerical Weather Prediction and Machine Learning, Year: 2017 | Volume: 32, Issue: 1 | Journal Article | Publisher: IEEE.