Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 8, August : 2023

Comparative Analysis of different Regression Algorithms for Disease prediction

[1*] Dr. B.V. Ramana, Professor, Information Technology,
Aditya Institute of Technology & Management, Tekkali, India,
ramana.bendi@gmail.com.
[2] B. Manideep, Computer Science and Engineering,
Aditya Institute of Technology & Management, Tekkali, India,
bendimanideep@gmail.com
[3] Jayavardhanarao Sahukaru, Assistant Professor, Computer Science and Engineering
Aditya Institute of Technology & Management, Tekkali, India,
jayavardhanarao.mca@gmail.com
[4] Sarmista Nayak, Information Technology,
Aditya Institute of Technology & Management, Tekkali, India,
sarmistanaya02@gmail.com

Abstract

Predictive analysis has become a critical component for future prediction as the area of Machine Learning has advanced. We know that regression algorithms play an important part in future prediction. However, it is uncertain which regression model generates the best accurate predictions. We have used a variety of regression models in the research, including linear regression, polynomial regression, and logistic regression. to conduct a comparative analysis for the prediction of the most common diseases in India they are COVID-19, Cardiovascular Disease, Diabetes, and Liver Cancer. This analysis indicates which regression model is best suitable for accurately forecasting the particular disease and as a result, this prediction also aids in implementing adequate measures to avoid the diseases in the future.

Index Terms: Linear Regression, Polynomial Regression, Logistic Regression, R^2, MSE, MAE, RMSE.

1. Introduction

A systematic method for comparing objects and identifying similarities and differences is comparative analysis. A comparative study explains whether data or procedures differ and are related with one another. This provides context for the analysis, allowing you to observe the distinctions and parallels in the relationships among data sets more easily. An auto manufacturer might, for instance, evaluate the safety specifications of two or more kinds to figure out how they influence sales or whether certain elements need to be upgraded. Such a study might include comprehensive information on each feature as well as historical information to contrast the functionality of each feature. An effective comparative study also helps a business come up with strong justifications for implementing the comparison. Comparative research may compare indirect and direct rivalry in order to develop a comprehensive understanding of a market. This method uses quantitative data to provide complete data gathered from an extensive population. Examples of comparative analysis include:

- **Pattern analysis:** To make predictions or uphold probability, one must be able to recognize patterns of behavior or trends.
- **Data filtering:** Analyzes group data to find and extract data subsets.
- **Decision tree:** Analyzes the benefits and drawbacks of a decision by examining its impacts and risks.

Performance evaluation indicators were employed to conduct comparison study. Accuracy on training data is crucial, but obtaining a true and approximative answer for unknowable data is just as crucial; else, the system is useless. So, in order to build and deploy a generic model, we must evaluate it against a variety of metrics. This allows everyone to maximise performance, fine-tune the model, and obtain superior results. To Regression analysis use statistical methods to determine the relationship among between one or more “independent variables” and just one particular “dependent variable”. The criteria's calculated value is created by successively merging the predictors. The three most popular regression models in machine learning are shown here and used for prediction. They are Linear Regression, Polynomial Regression and Logistic Regression.

The word "prediction" has many different meanings and is not usually associated with predicting a future event. In other cases, it might mean looking for trends in previous data and making judgements based on that research. Many other fields, including banking, healthcare, and education, can benefit from this. Organizations could use analytics to identify similarities in existing statistics and use those patterns to inform decisions about their operations, clients, and product. Hence, "prediction" can refer to both making educated guesses about what might occur in the future as well as examining historical data to learn further about past events. The term "prediction" refers to the result of a machine learning technique which was trained on a previous dataset and then used with fresh data to determine that likelihood of a definite conclusion, like if a client would leave in 30 days. Because future events are fundamentally unpredictable, providing guaranteed exact information about the future is impossible. Prediction can be useful in developing plans for prospective developments.

Due to the status of the environment today as well as the way life exists, they are now subject to an extensive variety of disorders. To prevent such disorders from getting worse, it is essential to identify and predict them in the earliest stages. It aids in the detection of trends and the prevention of disease transmission. Predictive analytics in healthcare may enhance healthcare quality, collect more clinical data for personalized treatments, and correctly identify an individual patient's medical condition.

Despite of commendable development in the healthcare industry, there are many dangerous diseases in India that are still prevalent. By predicting these diseases, we can reduce the disease in the future by taking proper measures.

- COVID-19 (Corona Virus)
- Cardiovascular Diseases (CVD)
- Diabetes.
- Liver Cancer

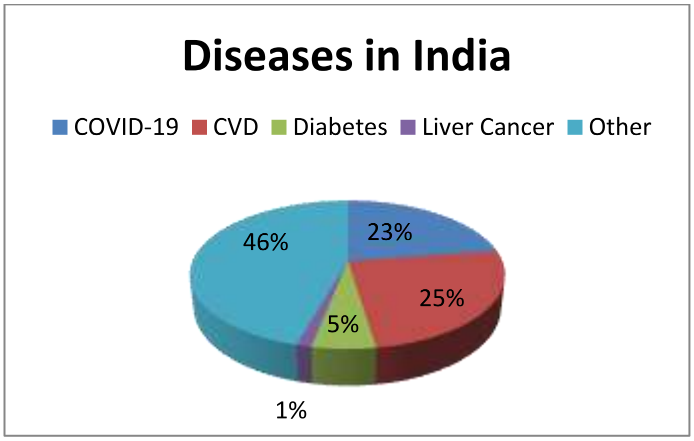


Fig: Pie chart analysis

2. Related Work

We looked into papers that used diverse algorithms to predict illness, such as KNN, linear regression, SVM, Nave Bayer, polynomial regression, logistic regression, and Random Forest. When all of these methods are compared, the linear regression, polynomial regression, and logistic regression algorithms produce the most accurate results.

We also conducted study on the dataset collection from our Literature Survey, and we discovered that the majority of the datasets are linear, resulting in lower accuracy. So, in our study, we acquired non-linear datasets from Kaggle, which offered us improved illness prediction performance. We also encountered that most performance evaluations were based on accuracy, which might not always be accurate.

As a consequence, in our project, we considered the following performance evaluation metrics for disease prediction, as by considering the most appropriate algorithms for various diseases, “R-square”, “Average Absolute Error”, “Root Mean Square Error”, and “Root Mean Square Error provided us with the results.

The performance evaluation measures allow us to make more accurate predictions. Even if the accuracy is significant, the error rate is extremely high, indicating that the model employed is unsuitable for disease prediction.

Based on MAE, MSE, RMSE and R-square performance evaluation metrics, we can forecast which regression model is most suited for disease prediction.

3. System Design & Implementation

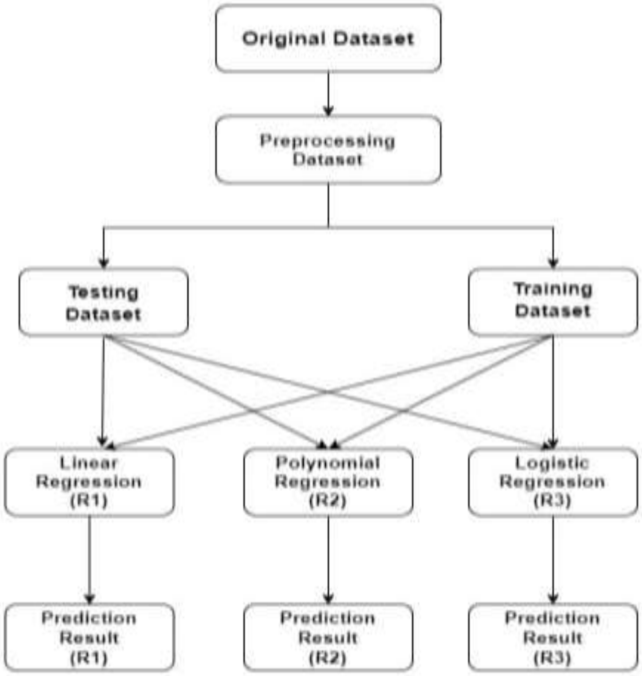


Fig: Work flow Diagram

Dataset:

First we have to collect the relevant data according to our problem, after that we have to filter the necessary data which is suitable for our experiment. We have to select the useful attributes in the data set.

Preprocessed data:

After obtaining the data from the database, we must preprocess it before training it. In this stage, we use several preprocessing approaches to remove any noisy or null values from the data.

- **Training dataset:** The machine learning algorithm uses this portion of our original data to find and learn patterns. This assists in moulding our model.
- **Testing dataset:** After creating your machine learning algorithm (utilizing data for training), you'll need to test it using information that's never been seen before. Testing data can be used to evaluate the success and growth of the training of your algorithms, as well as to change or improve them for better results.

Models Implementation:

Several models are trained as part of our endeavor to conduct comparative analysis for the identification of the most prevalent diseases in India. We looked at widespread regression Use methods like logistic regression, polynomial regression, and linear regression for comparative analysis.

1. Linear Regression

Linear regression is the most basic and often used machine learning method. Predictive analysis is performed using statistics. For continuous, real, or quantitative variables like sales, wages, age, and price of items, among others, linear regression forecasting is one option.

A statistical technique known as "linear regression" demonstrates a linear relationship between one or more independent (y) variables and a dependent (y) variable. Since there is a linear relationship between the two variables, linear regression is used to calculate the value of the dependent variable's change in relation to the independent variable's value. A linear regression can be conceptualized mathematically as:

$$a= z_0+z_1t+ e$$

a = “Dependent Variable” (Target Variable)

t = “Independent Variable” (predictor Variable)

z_0 = “Intercept of the line” (Gives an additional degree of freedom)

z_1 = “Linear regression coefficient” (scale factor to each input value).

e = “Random error value”

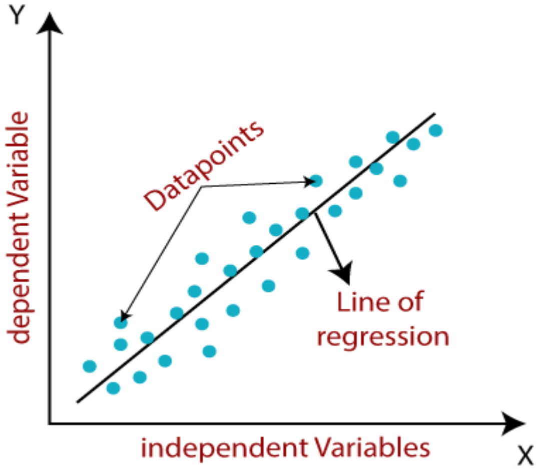


Fig: Linear Regression

2. Polynomial Regression

An nth degree polynomial is used in the regression technique known as "polynomial regression" to illustrate the relationship among a dependent (y) and an independent variable (x). The equation for polynomial regression is as follows:

$$\mathbf{a= z_0+z_1x_1+ z_2x_1^2+ z3x1+\ldots\ldots z_nx_1^n}$$

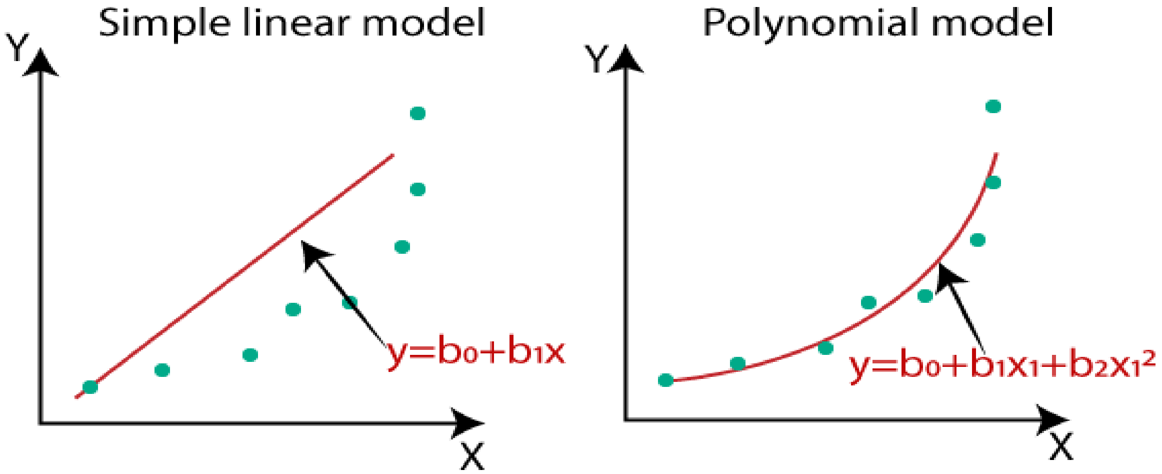


Fig: Polynomial Regression

Multiple Linear Regression converted into Polynomial Regression by adding additional polynomial terms. It is a linear model that has undergone modifications to increase precision. The polynomial regression training set is a non-linear dataset and whereas linear regression model uses non-linear functions and datasets. The original features are transformed into polynomial features in polynomial regression.

3. Logistic Regression

Logistic regression is one of the most popular supervised machine learning algorithm. Based on the independent factors, categorical variables are predicted and outcome is a categorical value.
Logistic Regression is depicted in equation as follows:

$$\mathbf{\log[a/1-a]= z_0+z_1x_1+ z_2x_2+ z_3x_3+\ldots\ldots+ z_nx_n}$$

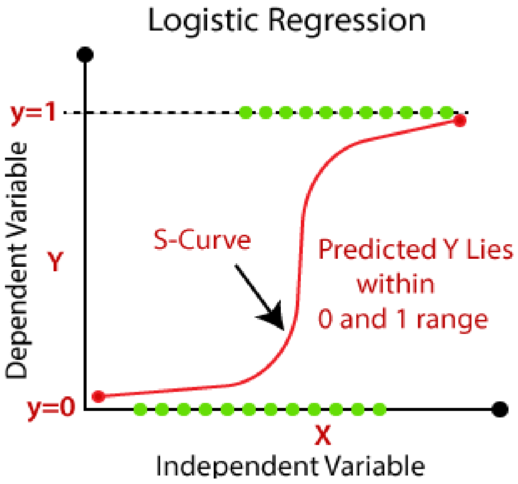


Fig. Logistic Regression Graph

Here, the odd ratio is (y/1-y). The probability of achievement will always be greater than 50% when the logarithm of odd ratio is shown to be positive.

Logistic regression fitted to "S" curved logistic function, and it forecasts two maximum values, (“0 or 1”).

Prediction Result:

Following training, we will assess the altered data to assess the “accuracy” of the algorithm. Comparison study of the regression based on the anticipated values. Accuracy for trained data is crucial, but it's just as important to have an accurate answer for unlabeled data; else, the model is useless. Hence, we must evaluate the model on a number of criteria before building and deploying a general model. This allows us to improve the performance, fine-tune the model, and obtain better results based on the performance evaluation metrics that are “Mean Square Error”, “Root Mean Square Error”, and “R Squared Error”.

1. Mean Absolute Error (MAE)

It is simple to calculate the actual disparity between the actual and expected values using the MAE metric. calculating the mean error and absolute error (the discrepancy between both the actual and projected numbers). To calculate MAE, calculate sum all the errors and divided them by both the number of observations in total.

Notation: $\mathbf{MAE = (1/x) * \sum|a_i - b_i|}$

- Σ: “Summation Symbol”
- a_i: “Actual value of the ith observation”
- b_i: “Calculated value of the ith observation”
- x: “Total number of observations”

2. Mean Squared Error (MSE)

The MSE is used to calculate the squared variance between the actual and projected values. The benefit of MSE is that we were able to avoid cancelling negative words by using square.

Notation: MSE = $(1/a) * \Sigma(ai-bi)^2$

Where:

- a_i: “Actual value of the ith observation”
- bi: “Calculated value of the ith observation”
- a: “Total number of observations”

3. Root Mean Squared Error (RMSE)

The square root operation is done on the MSE.

Notation: RMSE = $\sqrt{[\ \Sigma(a_i - b_i)^2 / a\ ]}$

where:

- Σ: Summation Symbol
- a_i: “Predicted value” of the i^{th} observation
- b_i: “Observed value” of the i^{th} observation
- a:sample size

4. R Squared (R^2)

The R^2 score is a parameter that shows “how well your model fared instead of the precise loss” for the number of wells it performed. We have shown that MAE and MSE are both context-dependent, whereas R2 score is context-neutral. R-squared values range from zero to one and can be expressed as percentages between zero and one hundred.

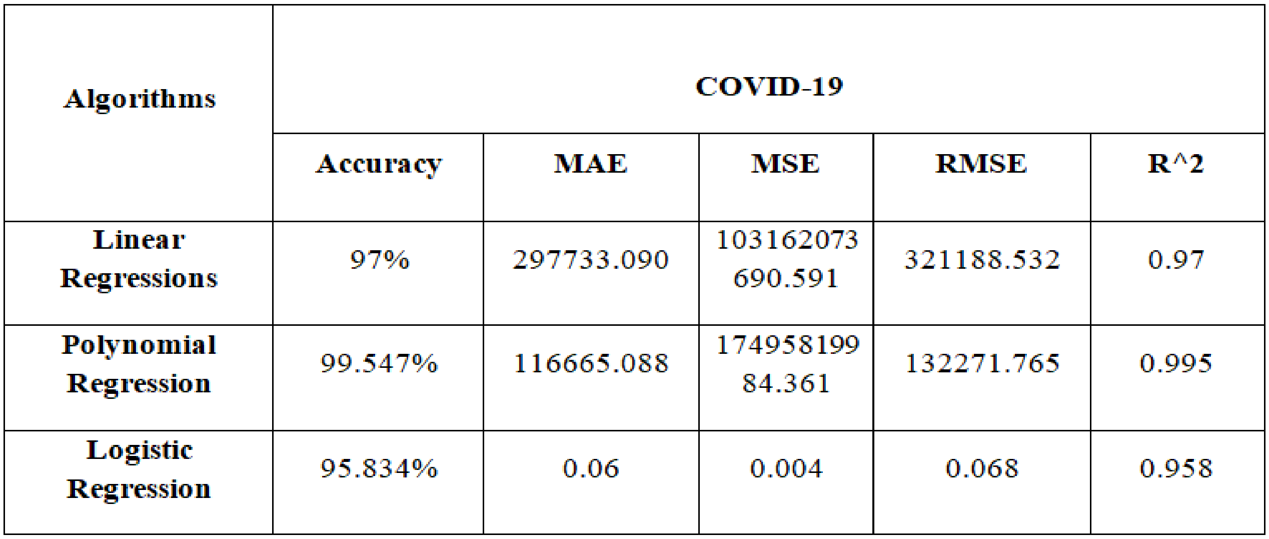
Notation: R^2=1−(ASS/BSS)

Where:

- ASS: Sum of Squares due to Regression.
- BSS: Total Sum of Squares.

4. Results

Comparative analysis of Regression algorithms for COVID-19 prediction



Algorithms	COVID-19				
	Accuracy	MAE	MSE	RMSE	R^2
Linear Regressions	97%	297733.090	103162073 690.591	321188.532	0.97
Polynomial Regression	99.547%	116665.088	174958199 84.361	132271.765	0.995
Logistic Regression	95.834%	0.06	0.004	0.068	0.958

Table 1: The Results obtained for COVID–19 disease

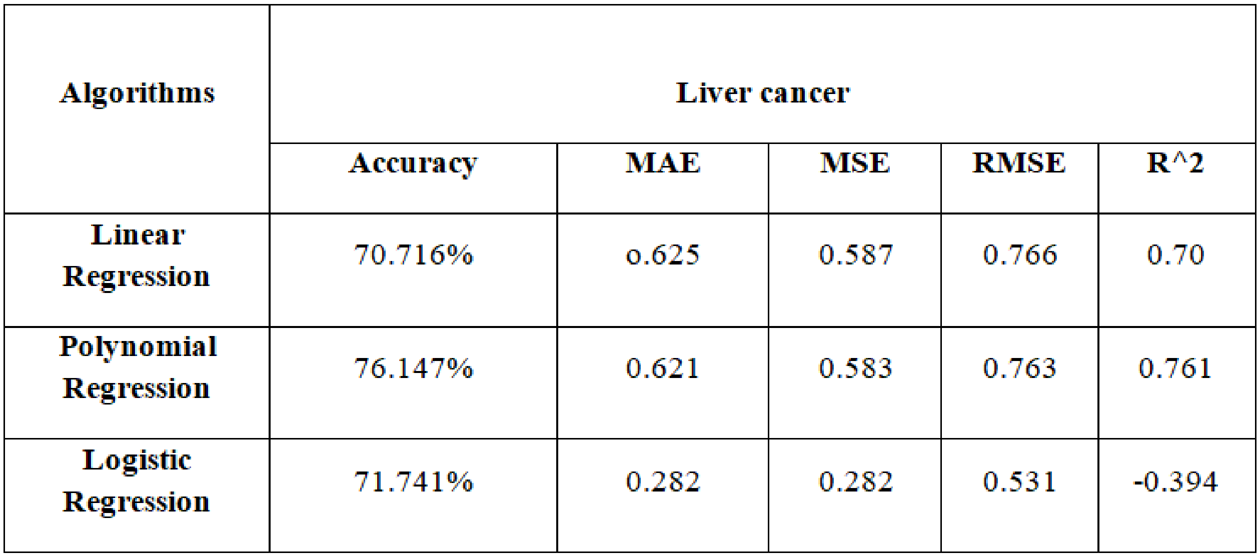
We can infer from the data above that the Logistic Regression provides reliable predictive performance for the COVID-19 illness.

Because, when compared to other regression methods, Logistic Regression has the lowest error rate. Although the accuracy was poor in comparison to other algorithms, the error rate was also low.

As a consequence, the results provided by this algorithm have the best chance of being correct. Polynomial Regression has the best accuracy of 99.54% when compared to the other regression methods. However, the error rate is relatively high when compared to Logistic Regression.

The research also indicates that the R - squared for Logistic Regression is 0.958, which is closer to one. This shows that our model is the best fit for predicting COVID-19.

Comparative analysis of Regression algorithms for Liver Cancer prediction



Algorithms	**Liver cancer**				
	Accuracy	**MAE**	**MSE**	**RMSE**	**R^2**
Linear Regression	70.716%	o.625	0.587	0.766	0.70
Polynomial Regression	76.147%	0.621	0.583	0.763	0.761
Logistic Regression	71.741%	0.282	0.282	0.531	-0.394

Table 2: The Results obtained for Liver cancer disease

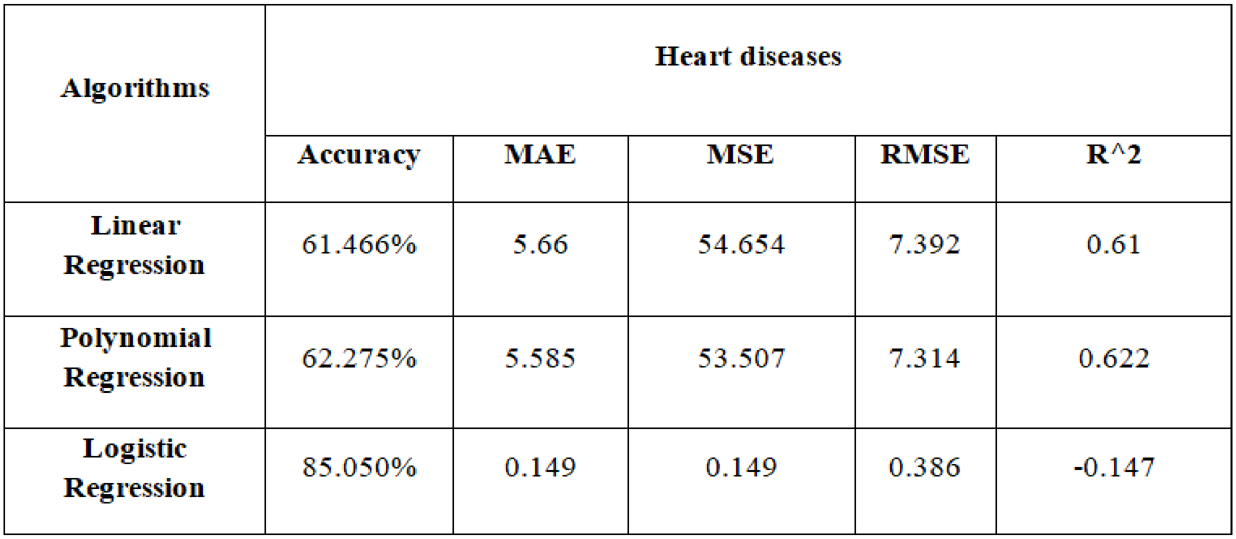
We can conclude from the comparative analysis shown above that polynomial regression yields reliable forecasts for liver cancer.

In contrast to the other regression methods, Polynomial Regression predicts Liver Cancer with the best accuracy of 76.147%, followed by logistic regression with 71.741% and linear regression with 70.716%.

This is due to the fact that the Liver Cancer data was non-linear. Our study suggests that the R2 value for Polynomial Regression is 0.761, which is closer to one.

This indicates that the Polynomial Regression method is most suited for forecasting Liver Cancer. MAE, RMSE and MSE are all closer to zero when compared to other regression procedures. The error rate for logistic regression was also relatively low when measured against other linear and polynomial regression methods.

Comparative analysis of Regression algorithms for Cardiovascular prediction



Algorithms	Heart diseases				
	Accuracy	MAE	MSE	RMSE	R^2
Linear Regression	61.466%	5.66	54.654	7.392	0.61
Polynomial Regression	62.275%	5.585	53.507	7.314	0.622
Logistic Regression	85.050%	0.149	0.149	0.386	-0.147

Table 3: The Results obtained for heart disease

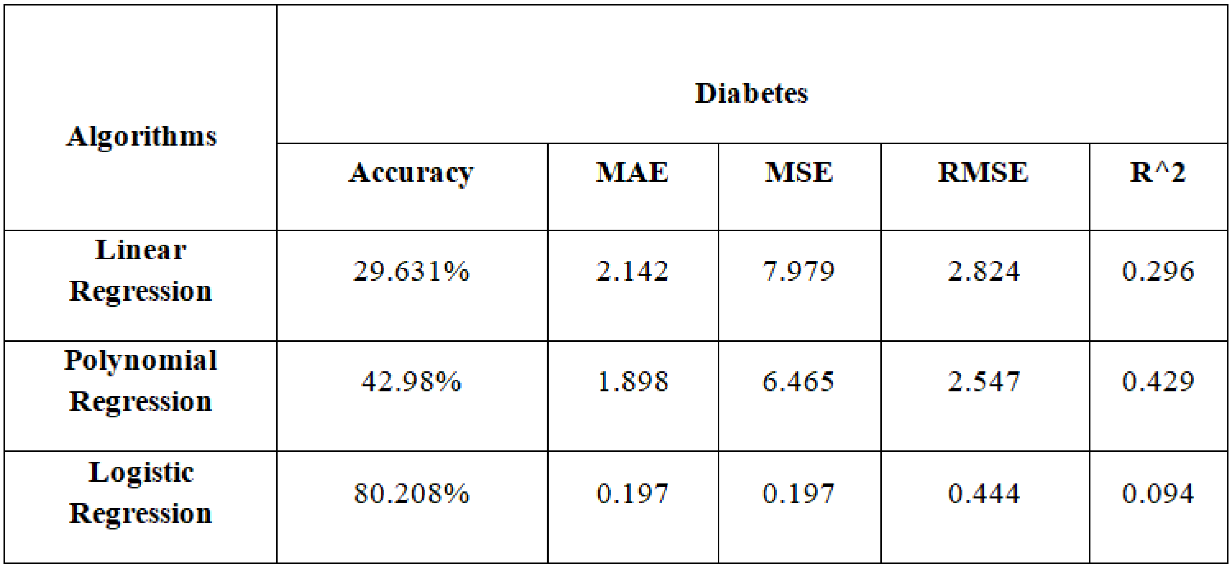
We may conclude from the comparative analysis shown above that Polynomial Regression offers accurate findings for heart disease prediction.

In contrast to the other regression techniques, Polynomial Regression has the best accuracy of 76.147% for Liver Cancer prediction, followed by logistic regression at 71.741% and linear regression at 70.716%.

This is solely because the Liver Cancer data was non-linear. Our research shows that the R2 value for Polynomial Regression is 0.761, which is closer to one.

This implies that the Polynomial Regression technique is most suited for predicting Liver Cancer. When compared to other regression techniques, the values of MAE, MSE, and RMSE are closer to zero. When compared to other Linear and Polynomial Regression techniques, Logistic Regression had a relatively low error rate.

Comparative analysis of Regression algorithms for Diabetes prediction



Algorithms	Diabetes				
	Accuracy	MAE	MSE	RMSE	R^2
Linear Regression	29.631%	2.142	7.979	2.824	0.296
Polynomial Regression	42.98%	1.898	6.465	2.547	0.429
Logistic Regression	80.208%	0.197	0.197	0.444	0.094

Table 4: The Results obtained for Diabetes disease

We can conclude from the data that above Logistic Regression offers accurate diabetes prediction results.

In contrast to the other regression techniques, Logistic Regression predicts Diabetes with the best accuracy of 80.208%, followed by Polynomial Regression with 42.98% and Linear Regression with 29.631%.

Our research shows that the R square for Logistic Regression is 0.094, which is closer to zero. As a result, Logistic Regression is the best match for diabetes prediction. When compared to other regression methods.

Logistic Regression has exceptionally low Values of 0.197, 0.197, and 0.444 for the MAE, MSE and RMSE. Thus, the best fit for diabetes prediction is logistic regression.

5. Conclusion & Future Scope

By the analysis of Table 1, we can conclude that the Logistic Regression gives accurate prediction results for the COVID-19 disease. Because, the Logistic Regression has the lowest error rate compared to other regression algorithms. Even though the accuracy was low when compared to other regression algorithms, the error rate was also relatively low. As a result, the outcomes produced by this algorithm have the highest possibility of producing accurate results. In comparison with the other regression algorithms Polynomial Regression gives highest accuracy as 99.54%, but the error rate is very high compared to Logistic Regression. The analysis also reveals that Logistic Regression had R^2 value is closer to 1 which is 0.958. This indicates that our model is the best fit for COVID-19 prediction.

By the analysis of Table 2, we can conclude that the Polynomial Regression gives accurate prediction results for the Liver Cancer. In comparison with the other regression algorithms Polynomial Regression gives highest accuracy as 76.147% for Liver Cancer prediction followed by logistic regression as 71.741% and linear regression as 70.716%. This is only because of the Liver Cancer data was non-linear data. Our analysis reveals that Polynomial Regression had R^2 value is closer to 1 which is 0.761. This indicates that Polynomial Regression algorithm is the best fit for Liver Cancer prediction. The values of MAE, MSE, RMSE are closer to 0 compared to other regression algorithms. The error rate was also very low for Logistic Regression while comparing with other Linear and Polynomial Regression algorithms.

By the analysis of Table 3, we can conclude that the Logistic Regression gives accurate prediction results for the cardiovascular disease. In comparison with the other regression algorithms, Logistic Regression gives highest accuracy as 85.050% for cardiovascular disease prediction followed by Polynomial Regression as 62.275% and Linear Regression as 61.466%. In addition, our analysis shows that the Logistic Regression algorithm has the lowest error rate, with MAE, MSE, and RMSE of 0.149, 0.149, and 0.386 respectively, when compared to the Linear Regression & Polynomial regression algorithms. According to the analysis, the R2 value for such a logistic regression is negative but also gets closer to 1. By the analysis of Table 4, we can conclude that the Logistic Regression gives accurate prediction results for the Diabetes. In comparison with the other regression algorithms, Logistic Regression gives highest accuracy as 80.208% for Diabetes prediction followed by Polynomial Regression as 42.98% and Linear Regression as 29.631%. Our analysis reveals that Logistic Regression had R^2 value is closer to 0 which is 0.094. This indicates that Logistic Regression is the best fit for diabetes prediction. As compared to other regression algorithms, the MAE, MSE, and RMSE values for Logistic Regression are similarly quite low at 0.197, 0.197, and 0.444 respectively. This indicates that the greatest match for diabetes forecasting is logistic regression.

6. References

1. Intertumoral heterogeneity and clonal evolution in liver cancer https://doi.org/10.1038/s41467-019-14050-z Losic, B. and Craig, A.J., 2020. Villacorta-Martin C et al.

2. A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms, https://doi.org/10.1023/A:1010933404324 , F. M. Javed Mehedi Shamrat et .al, 11 September 2019.

3. Impact of a deep learning assistant on the histopathologic classification of liver cancer, https://doi.org/10.1038/s41746-020-0232-8 , Amirhossein Kiani et .al , 23 March 2020

4. Development and validation of personalized risk prediction models for early detection and diagnosis of primary liver cancer , https://doi.org/10.1101/2022.01.24.2226975 , Weiqi Liao et .al, 25 January 2022

5. Body Surface Area and Body Weight Predict Total Liver Volume in Western Adults , https://doi.org/10.1080/09720502.2020.1833458 , Jean-Nicolas Vauthey et .al , 19 February 2021.

6. Prediction of COVID-19 pandemic measuring criteria using support vector machine prophet and linear regression models in Indian scenario, https://doi.org/11.10466/09720502.2020.1346895 , Amit Kumar Gupta et .al, 9 December 2019.

7. Logistic Regression Analysis to Predict Mortality Risk in COVID-19 Patients from Routine Hematologic Parameters , http://dx.doi.org/10.1101/2020.03.17.20037572 , Sudhir Bhandari et .al, 15 August 2020.

8. Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model , http://dx.doi.org/10.1183/13993003.00524 , Mohd Saqib , 24 October 2021.

9. Predictive modeling of COVID-19 confirmed cases in India http://dx.doi.org/10.1016/j.diabres.2020.10813 , Benedita B.Aladeitanb, July 2022.

10. Determine of COVID-19 pandemic measuring criteria using prophet and linear regression models in Indian scenario, http://dx.doi.org/10.1101/2020.03.17.2003757 , Priya Mathur , 23 October 2020

11. The clinical significance of interleukin-6 in heart failure: results from the BIOSTAT-CHF study,. http://doi.org/10.1109/ACCESS.2020.2968608 , John G. Cleland et .al, 6 May 2019.

12. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics, https://doi.org/10.1007/978-981-13-8798-2 , Girish Dwivedi et .al, 19 April 2020.

13. Machine Learning Prediction of Mortality and Hospitalization in Heart Failure with Preserved Ejection, https://doi.org/10.1007/s00542-018-4119-4 , Rohan Kera et .al, 10 May 2022.

14. Heart Disease prediction using data mining techniques, https://doi.org/10.1016/S0019-9958(65)90241 , N Sridev et .al, 10 March 2019 Kunreuter, H.C. and Michel-Kerjan, E.O., 1965. Paper prepared for the University of Pennsylvania law conference on climate change, 16–17 November 2006 LA Zadeh.

15. Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model, https://doi.org/10.3349/ymj.2019.60.2.191 , Rachel et .al, 16 January 2020.

16. Early detection of type 2 diabetes mellitus using machine learning-based prediction models, http://doi:/10.1002/ehf2.12419 , Gregor Stiglic et .al, 2 November 2020.

17. Handling Irregularly Sampled Longitudinal Data and Prognostic Modeling of Diabetes Using Machine Learning Techniques, https://doi.org/10.1016/j.compbiomed.2020.103949 , Amjad Rehman et .al, 4 January 2020.

18. Computer Vision and Machine Intelligence in Medical of using regression models, https://doi.org/10.1007/s00477-020-01827-8 , Sambhunath Biswas et .al, 24 June 2019.

19. Machine Learning for the Prediction of New-Onset Diabetes Mellitus during 5-Year Follow-up in Non-Diabetic Patients with Cardiovascular Risks, https://doi.org/10.1101/2020.04.13.20063461 , Yung-Kyun Noh et .al , 10 September 2020.

20. Improved logistic regression model for diabetes prediction by integrating PCA and K-mean techniques, https://doi.org/10.1101/2020.03.30.20047308 , Wenfang Fengb et .al, 22 June 2020.

21. Analysis and Prediction of diabetes using Regression Models, https://doi.org/10.1101/2020.02.27.20028027 , S.SaruS.Subashree, 4, April 2019.

22. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using linear regression model, https://doi.org/10.1504/IJAIP.2019.098592 , .Alok Ranjan Tripathy, 5, September–October 2020.

23. Predictive regression modeling of COVID-19 confirmed cases in India, DOI:10.1109/BigMM50055.2020.00062 , Benedita B.Aladeitanb et .al , 20 August 2020.

24. A Comparative Analysis of Different Regression Models on Predicting the Spread of Covid-19 in India , https://doi.org/10.1016/j.dsx.2020.07.045 , Ujjwal Maulik et .al , Oct 30-31, 2020.

25. A Comparative Study On Liver Disease Prediction Using regression models for Machine Learning Algorithms , http://doi.org/10.1016/j.ajp.2020.102089 , A.K.M Sazzadur Rahman , Nov-28, 2020.

26. A logistic regression and knn model for early prediction of diabetes, https://doi.org/10.1155/2017/4827171 , Talha Mahboob Alama , 5 June 2021.

27. Detection of lung cancer with electronic nose and logistic regression analysis,

https://doi.org/10.1016/j.chaos.2020.109942 , Madara Tirzīte, 21 september - 4 November 2020.

28. Prediction of Heart Diseases using polynomial regression models, https://doi.org/10.1016/j.dsx.2020.07.045 , Mohammad Nadeem, 9 April 2021.

29. Efficient diabetes mellitus prediction with grid based random forest classifier in association with natural language processing, https://doi.org/10.1264/2017/537995 , Dipak Kumar Bose, 15 July 2020.