Comparing Random Forests Regression with Logistic Regression to Determine the Most Important Factors Affecting Congenital Malformations For Newborn Babies in Iraq

Dr.Haitham Yaqub Yousif [1], Arshad Hameed Hassan[2], Shaima Mohammed Ahmed[3]

Abstract

Congenital malformation is a structural defect in one or more parts of the body from birth , The causes and sources of birth defects may be genetic or caused by a non-genetic event before birth , Some congenital malformations may be caused by taking drugs or sometimes the causes are unknown. In recent years, the rate of congenital malformations among newborn children in Iraq has increased and to identify this problem and identify the most important factors affecting it. A sample of children with congenital malformations was taken to the maternity hospitals of the Health Department of Baghdad / Rusafa and Karkh - Department of preterm infants of 2504 *births , To identify the most important factors affecting the congenital malformations using artificial intelligence techniques and machine learning including random forests regression and logistic regression decline as these techniques are one of the most advanced techniques used in the case of big data , We concluded from this research, which aimed to find the best model for estimating the data of congenital anomalies in Iraq through the use of two types of machine learning models (artificial intelligence) and as these types are regression models at the same time and after estimating each model we compared using the mean squares error criterion and it was the best A model is a random forest model regression*.

Keywords*: Tree Regression, Random Forest, Binary Logistic regression, Congenital malformation, Mean Square Error.*

I. Introduction

Data sets consist of different dimensions and infrastructure. The relative overall performance of gadget studying algorithms on information sets with variable records traits is not nicely recognized. Greatest available effort likens fashionable general presentation among multiple replicas on a solitary data set instead of measuring the overall version overall performance for datasets consisting of different, multi-linear dimensions, kinds of enter features (e.G. Incessant and segmental), and deliveries of arithmetical variables. The presentation of mechanism knowledge procedures significantly influences the particular algorithm for execution. For example, if the board mutable isn't linearly divisible in n-dimension area, whether or not non-stop or definite, a additional multifaceted model can be required to reap advanced forecast notches. Multifaceted fashions inclusive of selection bushes or different nonparametric algorithms will have decision limitations with large variance in forecasts nonetheless little prejudice, which regularly leads to over appropriate if not well adjusted. Over fitting is the consequence of a perfect that attained a excessive rating in a education set with negative generalization of information sets from the sample. On the other pointer, parameter-based replicas inclusive of logistic reversion are much fewer multifaceted, ensuing in a lined selection restrict, however container consequence in better prejudice. Furthermore, this will interpret into inadequacy due to the fact the version nose-dives to study styles inside the information sufficiently to gain correct predictions if no longer nicely tuned. Balancing the bias as opposed to variance change-off is pushed through the complexity of the set of rules, which is important to propagating successful models of realistic applications.

Depending at the structure of the dataset, decoding the set of rules to be deployed with a purpose to gain peak performance continues to be an ad hoc procedure. This increases the queries, in what situations fixes one version instigate outperforming additional? For example, whilst the quantity of sound and descriptive variables will increase in a data set, at what opinion does the comparative version overall presentation start to skew amid the replicas? To response those queries, our effort includes constructing an logical instrument that fakes the complexities of numerous records to display the organization overall presentation of two device getting to know procedures by be around metrics for 1,000 chance generations of particular, multivariate records units. For interoperability and calculation period for perfect exercise, we taken into consideration one parametric and nonparametric device getting to know perfect for two class, logistic reversion, and chance forests, correspondingly.

Logistic reversion and chance forests are two actual popular models of machine learning which have been extensively studied. Machine getting to know is the procedure of studying exact procedures' styles or tendencies in formerly logged facts notes after which brands a forecast or class. In this effort, we study lone two type (for example Y = 1,0), that is a shape of oversaw knowledge in which an procedure pursuits to categorise the magnificence to which an input belongs. Supervised gaining knowledge of may be labeled as captivating an contribution course consisting of n functions and assigning it to an related board fee or lesson label. The time period "moderated" arose from the notion that schooling and check records sets comprise a reaction label and that the algorithm video display units the input vector and attempts to find out the possibility distribution to predict a given "y" of "x" [6]. Algorithms analyze a sequence of contribution heaviness limits that control how the enter characteristic course touches the forecast. The aim of the procedure is to discover a usual of masses on a subsection of facts that minimizes the error or loss between the underlying reality and the predicted price of accurately classifying the enter to the associated label. Rating scales consisting of accuracy, true and mistakes wonderful costs, and the place below the curve on the part of facts that turned into preserved during training are examined to assess how nicely the model grades the enter feature vector.

To examine the binary type version among random woodland and logistic regression, our paintings targeted on 4 wonderful simulation records sets: (1) growing the variance within the descriptive and sound variables, (2) cumulative the wide variety of sound variables, (3) growing the wide variety of descriptive variables, (four) growing the variety of Notes. To degree rating ratings and evaluate them amid stochastic wooded area regression and logistics, metrics including accuracy, area underneath the curve, true wonderful charge, false high-quality charge, and accuracy have been analyzed. To provide a statistical quantitative estimate of whether the distinction in version performance is essential enough to provide an explanation for the enormous difference or if the found distinction is with the aid of random hazard, a -pattern t-check is also performed on the stop of every simulated case look at.

II. Machine Learning Algorithms

The gadget knowledge procedures deliberate in this effort encompass chance woodland regression and logistic reversion. Together fashions had been applied on a large scale with success in numerous disciplines for category and regression functions [4]. The logistic regression function, parameter-primarily based version, and random forests, which is a non-parametric version, are abridged inside the subsequent segment.

2.1 Random Forest Regression

Random wooded area is a set-founded totally gaining knowledge of algorithm which include n clusters of uncorrelated selection timber [7]. It is based totally on the idea of bootstrap meeting, that is a technique of reconfiguration with substitution to reduce comparison. Random Forest makes use of more than one plants to rate (reversion) or to calculate mainstream ballots (rating) in fatal bulges while creating a forecast. Based on the impression of selection bushes, chance wooded area fashions led to tremendous enhancements in forecast correctness compared to a unmarried sapling by way of increasing the quantity of bushes; All tree inside the exercise usual is randomly tested deprived of substitution [4]. Choice trees honestly contain of a tree-like construction anywhere the pinnacle bulge is the foundation of the sapling that is again and again divided in a chain of choice bulges after the foundation pending the fatal bulge or selection node is touched.

The choice sapling set of rules is a top-down 'grasping' technique that divides the information set hooked on lesser subsections. Avaricious procedures are people who income the only answers as opposed to the most fulfilling one, that's frequently greater complicated. The pinnacle of the choice tree is called the origin bulge and this agrees to the pleasant prediction mutable. In all choice bulge, the capabilities are break up hooked on twigs and this procedure is recurrent pending the terminal bulges are touched, that are rummage-sale to brand the very last forecast.

The advantages of the use of a tree-like mastering set of rules permit exercise replicas on big data sets in adding to measurable and qualitative contribution variables. Additionally, tree-founded totally fashions may be resilient to redundant or noticeably correlated variations which could lead to customization in other getting to know procedures. Trees too must only a insufficient limits to great-music once education the version and perform fantastically properly with outliers or missing values inside the facts set. Though, trees are disposed to to negative forecast overall presentation; The choice timber themselves have a tendency to have elevated sound within the exercise usual eventually leading to consequences by tall variability. In additional phrases, because of this the perfect can correctly expect the same educated records but won't have the equal performance on statistics sets with out similar styles and differences inside the schooling set. Even mature decision timber are infamous for overprocessing and do not simplify healthy to non-visual statistics; Chance woodland resolves the riddle of overprocessing by means of the usage of a aggregate of a "set" of selection timber wherein the values inside the tree are random and independent samples.

Random sample by substitution is recognized as packing and this creates a special tree to educate on; Calculating common consequences from the variety of "n" timber will reduce variability and create smoother choice obstacles [7]. For instance, whilst the usage of the chance woodland for type, every sapling determination stretch an approximation of chance of lesson naming, the percentages determination be around ended "n" bushes and the very best harvests of anticipated elegance naming. In addition to filling or potting assembly, to reduce variability in resolution obstacles, trees need to be completely unrelated, and the potting method on my own isn't always enough. Bremann added the concept of randomly sampling the variety of functions in each tree splitting choice as a method to beautify trees in a chance woodland procedure [3].

2.1 The Random Forest Regression Model

Random Forest Regression (RF) refers to clusters of reversion timber [10] in which a T-based totally institution of non-remoted regression timber is made founded totally on bootstrap samples from the unique education records. For every node, the top-quality cut up of the node

The feature is decided from a hard and fast of functions which are randomly decided on from the whole of M's features. For $m < M$, Selecting the node split characteristic from a random set of capabilities reduces the correlation among one of a kind trees, and as a consequence the common reaction of more than one regression timber could be predicted to have less variance than that of person regression trees. Large sizes can improve the predictive electricity of man or woman timber, however they can also increase correlation among bushes and nullify any profits from the average of a couple of predictors. Reconfiguring the facts to train each tree increases independence between bushes.

2.1.1 Process of Spiting a Node

Let $x_{tr}(i,j)$ and $y(i)(i = 1,2, \ldots., n, j = 1,2, \ldots, M)$ Denotes training expectation features and output response samples, respectively. In any η_P node, we aim to determine the j_S feature from a random set of m features and a Z threshold to divide the node into two _L subnodes (left node with pathological samples $x_{tr}(I \in \eta_P, j_S) \leq \mathcal{Z}$) and η_L (Right node with satisfactory samples $x_{tr}(i \in \eta_P, j_S) > Z$).

We do not forget the node price the sum of the squared variations:

$$D(\eta_P) = \sum\nolimits_{i \in \eta_P} (y(i) - \mu(\eta_P))^2$$

Where μ (η_P) is the anticipated fee of y (i) in node η_P. Thus the discount inside the cost of section γ at node η_P is:

$$C(\gamma, \eta_P) = D(\eta_P) - D(\eta_L) - D(\eta_R)$$

The partition $\gamma *$ that maximizes $C(\gamma, \eta_P)$ for all feasible walls it's far unique for node η_P. Note that for the non-stop feature with n samples, the overall variety of n sections ought to be verified. Thus the arithmetic difficulty of every node cleavage is $O(mn)$. During the tree creation system, a node containing education samples much less than size n does not break up more than that

2.1.2 Forest Prediction

Using the random function choice manner, we suit the tree based totally on the bootstrap pattern $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ Generated from training facts.

Let us consider a prediction primarily based on a pattern x test of the tree Θ. Let $\eta(x, \Theta)$ be the partition that incorporates x, The tree reaction takes the shape [5]:

$$y(x, \Theta) = \sum_{i=1}^{n} w_i(x, \Theta)\, y(i)$$

where the weights $w_i(x, \Theta)$ are given by:

$$w_i(x, \Theta) = \frac{1_{[x_{tr}(i) \in \eta(X, \Theta)]}}{\{r: x_{tr}(i) \in \eta(x_{tr}(r), \Theta)\}}$$

Let the T trees of the Random Forest be denoted by $\Theta_1, \ldots, \Theta_T$ and let $w_i(x)$ indicate the average weights over the forest, that is:

$$w_i(x) = \frac{1}{T} \sum_{j=1}^{T} w_i(x, \Theta_j)$$

The Random Forest prediction for the check sample x is then given through:

$$\hat{y}(x) = \sum_{i=1}^{n} w_i(x)\, y(i)$$

2.2 Logistic Regression

Logistic reversion evaluation educations the association among the definite reliant on mutable and a hard and fast of self-governing (explanatory) variables. Logistic reversion is used for a name while the dependent variable contains only two values, which includes zero and 1 or sure and no. Usually the call of the polynomial logistic regression is reserved for the case whilst the dependent variable has three or extra unique values, consisting of married, single, divorced, or widowed. Although the form of information used for the based variable differs from the sort of more than one regression, the sensible use of the system is comparable.

Logistic regression competes with the function analysis as a way for analyzing express response variables. Many statisticians sense that logistic regression is greater versatile and higher applicable to modeling maximum conditions than discriminatory analysis. This is because logistic regression does not assume that the independent variables are commonly dispensed, as does the discriminant evaluation.

Logistic regression is used in a ramification of fields, which include system getting to know, greatest medicinal arenas, and the communal disciplines. For instance, the Trauma and Injury Severity Scale (TRISS), which is broadly rummage-sale to expect humanity in hurt patients, changed into at first evolved by way of Boyd et al. By logistic reversion. [2] Several other scientific balances rummage-sale to measure patient harshness the usage of logistic regression must been industrialized. [9][1][12][11]

By looking at a logistic model with positive parameters, then see how transactions may be estimated from the facts. Consider a model with predictors, x1 and x2, and one binary reaction variable (Bernoulli) Y, which we check with as p = P (Y = 1). We assume a linear relationship among the expectation variables and the log-odds of the occasion that Y = 1. This linear relationship can be written in the following mathematical shape (wherein is the logarithm possibilities, e is the logarithm base, $\beta_i$are the version parameters):

$$l = log_e \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

We can recover the odds by exponentiating the log-odds:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

By simple algebraic manipulation, the probability that Y=1 is:

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} + 1} = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$$

The above method indicates that after β_i is fixed, we are able to without problems calculate the logarithms probabilities of Y = 1 for a given observation, or the opportunity that Y = 1 for a given statement. The essential use case for logistic model is to present word x1, x2 and estimate the chance y that Y = 1.

And the general logistic regression model for k explanatory variables is:

$$log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad \dots (2)$$

Where The β 's represents unidentified limits to be anticipated.

2.2.1 Parameter Estimation

The aim of logistic regression is to estimate k + 1 for the unknown parameters β within the equation. 2. This is finished by using estimating the maximum opportunity which necessitates finding a hard and fast of parameters in which the probability of the located statistics is more. The most probability equation is derived from the probability distribution of the based variable. Since every Yi represents a binomial wide variety inside the variety i th, the not unusual opportunity density characteristic of Y is:

$$p = prob(Y = g|X_{k+1}) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{\sum_{j=1}^{k} e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}$$

The probability for a sample of n observations is then given by using:

$$l = \prod_{i=1}^{n} \prod_{j=1}^{k} \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{\sum_{j=1}^{k} e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}$$

And the log-likelihood, L, is given by:

$$L = \ln(l) = \sum_{i=1}^{n} \sum_{j=1}^{k} Y_{gi} \ln\left(\frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{\sum_{j=1}^{k} e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}\right)$$

The maximum chance estimates for β's are the ones values that maximize this log likelihood equation. This is completed by using computing the partial derivatives and putting them to zero and after that, we can solve these equations by using the numerical method as newton Raphson method.

2.2.2 Evaluation of explanatory of the logistic regression model

Can be used the coefficient of determination for the quality of the estimated logistic regression version to explain the relationship among reaction variable and explanatory variables through the use two statistics are $R^2_{Cox\&snell}$ and $R^2_{Nagelkerkl}$ so the statistics can be calculated as follows:

$$R^2_{Cox\&snell} = 1 - \left[\frac{L_0}{L_1}\right]^{\left(\frac{2}{n}\right)}$$

Where the L_0 signify the maximum probability purpose when the perfect involved intercept only and L_1 signify the maximum probability purpose when the model involved all explanatory variables, n is a sample size.

2.2.3 The goodness of fit tests for the logistic regression model

2.2.3.1 wald test

Wald's test is a way of finding out if explanatory variables in a version are important. Important approach adding something to the version; Variables that don't upload whatever can be deleted with out affecting the version in any significant way. The take a look at can be used for lots distinct fashions which include those with binary variables or continuous variables. So the hypothesis is null:

$$H_0: \beta_j = 0$$

Vs

$$H_1: \beta_j \neq 0$$

That means significant test for each parameter of the logistic reversion perfect so if the p-value fewer than 0.05 than nasty to castoff the null hypothesis

And can compute the wald statistic according to the following formula:

$$Wald = \left[\frac{\hat{\beta}_j}{S.E(\hat{\beta}_j)}\right]^2$$

Where $\hat{\beta}_j$ estimates of parameters. So $S.E(\hat{\beta}_j)$ is the standard error of estimates of parameters.

2.2.3.2 Hosmer And Lemeshow Test

This test is used to find out whether or not the model represents data well. The Chi-squared test is used for good conformity assessment to assess the differences between observation and expected values and to test the following hypothesis:[8]

$$H_0: Observed\ cases\ are\ equal\ to\ predicted\ cases$$

$$(the\ model\ represents\ data\ well)$$

$$H_1: The\ observed\ cases\ are\ not\ equal\ to\ the\ predicted\ cases$$

$$(the\ model\ does\ not\ represent\ data\ well)$$

The decision to accept the null hypothesis is if the probability worth of the Chi-four-sided number is better than the specified level of significance.

III. Application study and Results

A random sample of 2504 births has been approved in the various governorates of Iraq, where these data were collected through a form prepared by the Department of Premature Hospitals in Maternity and General Hospitals for the year 2017.

Where a variable binary response value and takes value 1 the presence of deformation and valley value 0 no deformation and valley either illustrative variables include x1 residential address and x2 sex and x3 weight, and x4 age of the mother, x5 career mother, x6 the age of the father, x7 career father, x8 is The degree of consanguinity between the parents and x9 is Presence of a previous childhood disability, x10 is type of birth present,x11 is Current birth and x12 is The presence of chronic diseases so x13 is During pregnancy, the mother was exposed to radiation, x14 is Number of previous projections, x15 is Housing type.

Statistical analysis of the data will be carried out using the statistical programming language R.

3.1 Logistic Regression Model Results

The following results were obtained:

Table (3-1) Show Dependent Variable Encoding

Original Value	Internal Value
Child Malformation	0
No Child Malformation	1

The table (3-1) shows encoding for the response variable that represents congenital deformation, where it expresses(1) for No Child Malformation and (0) for Child Malformation.

Table (3-2) shows the model summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3373.520	0.18	0. 24

The table(3-2) shows the values coefficients of determination (Cox & Snell R Square) and (Nagelkerke R Square) where the value of (Cox & Snell R Square=0.18) that meaning that the explanatory variables comprised in the perfect clarify 18% of the changes in the response variable while the value of

Nagelkerke R Square equal to (0.24) that meaning that the explanatory variables comprised in the perfect clarify 24% of the changes in the response mutable.

Table (3-3) Show Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.906	8	.547

The table(3-3) demonstrations the consequences of a Hosmer and Lemeshow test where the statistic showed that the value of sig is less than 0.05 and this incomes we receive the null hypothesis which conditions that there is no significant change amid the experiential values and the expected values

Table (3-4) show Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	x1	.016	.008	3.485	1	.062	1.016
	x2	.088	.078	1.256	1	.262	1.092
	x3	.002	.000	.228	1	.633	1.000
	x4	.002	.006	.160	1	.689	1.002
	x5	-.525	.195	7.205	1	.007	.592
	x6	-.004	.007	.379	1	.538	.996
	x7	-.032	.040	.658	1	.417	.968
	x8	.119	.082	2.118	1	.146	1.127
	x9	.032	.125	.064	1	.800	1.032
	x10	.445	.281	2.514	1	.113	1.561
	x11	.058	.089	.428	1	.513	1.060
	x12	1.032	.323	10.201	1	.001	2.806
	x13	.119	.075	2.474	1	.116	1.126
	x14	.256	.100	6.500	1	.011	1.292
	x15	.146	.102	2.072	1	.150	1.158

	Constant	-1.172	.611	3.675	1	.055	.310
a. Variable(s) entered on step 1: x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15.							

The table(3-4) shows descriptive variables comprised in the perfect and the first column represent the logistic regression model parameters estimation therefore the equation of the logistic regression model estimate can be written as the following :

$$\hat{y}_i = -1.172 + 0.016x_1 + 0.088x_2 + 0.002x_3 + 0.002x_4 - 0.525x_5 - 0.004x_6 - 0.032x_7 + 0.119x_8 + 0.032x_9 + 0.445x_{10} + 0.058x_{11} + 1.032x_{12} + 0.119x_{13} + 0.256x_{14} + 0.146x_{15}$$

As for the second, third, fourth and fifth columns, it shows the standard error for each parameter and the wald test and its degree of freedom and it significant respectively. And we notice from wald test the significance of (x_5,x_{12},x_{14}).

As for the sixth column, it represents the odds ratio. This ratio designates the amount of alteration that occurs in the chances ratio of the occurrence of the event (congenital deformation) when a change in the values of the explanatory variables occurs.

3.2 Random Forest Regression Model Results

As we mentioned previously, random forest regression is one of the machine learning algorithms, and it is one of the methods of nonparametric regression that depends on estimating a non-parametric function and that this estimate is called prediction of the concept of artificial intelligence and this type of regression models will be estimated using R programming based on the package (randomForest) and was Estimation random forest regression as following :

Table (3-4) show estimates (prediction) of random forest regression

No of Observations	$\hat{y}(x)$
1	0.7610192
2	0.8967483
3	0.5288707
4	0.7240669
5	0.6882969
6	0.6195665
7	0.5972924

8	0.7609679
9	0.7978261
10	0.6825731
11	0.5635306
12	0.5967287
.	.
.	.
.	.
2504	0.24978063

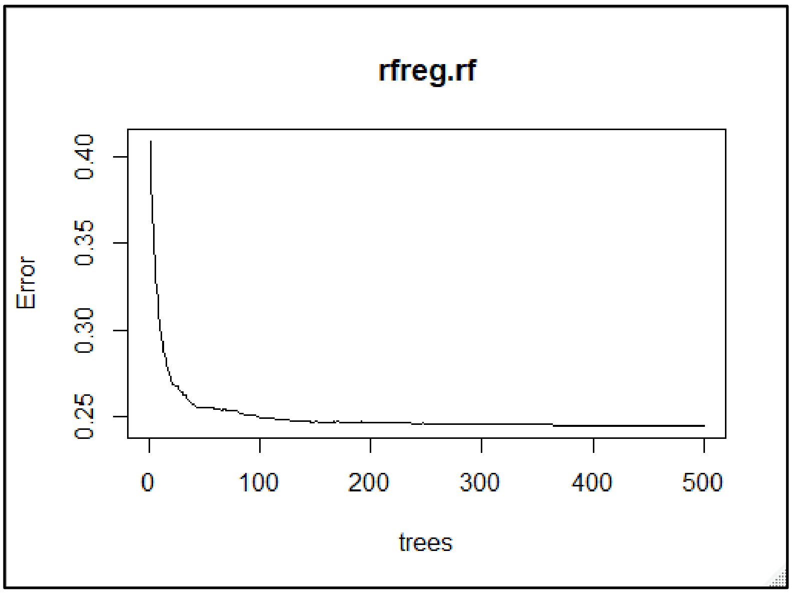

Figure 1: Errors of trees

Figure 1 shows the higher the number of trees on which the chance woodland regression perfect depends, the less error and the model becomes more accurate. When the random forest regression model was estimated for birth defects data, several 500 trees were adopted.

3.3 Compares of Models

The models will be compared based on the nasty four-sided mistake criterion that is used to judge the accuracy of the models because it depends on calculating the error values between the real values and the estimated values and is a formula as in the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where y_i are the real values of the reply variable and $\hat{y}_i$ are the estimated values of the response variable. then the result of the MSE criterion as the following table :

Table (3-5) show the MSE criterion of models

Models	Logistic Regression	Random Forest
MSE	0.6496884	0.2451726

Form table (3-5) we notice that then MSE of Random Forest Regression less than of MSE of Logistic regression, therefore, the Random Forest Regression is better than Logistic regression to estimating the Data model of congenital anomalies of newborn children in Iraq.

IV. Conclusions

We concluded from this research, which aimed to find the best model for estimating the data of congenital anomalies in Iraq through the use of two types of machine learning models (artificial intelligence) and as these types are regression models at the same time and after estimating each model we compared using the mean squares error criterion and it was the best A model is a random forest model regression, as it has the value of the criterion less than the logistic regression model. As we have concluded, the most important factors affecting birth defects are the mother job variable (x_5), the presence of chronic diseases variable (x_{12}), and the number of previous projections (x_{14}).

V. Recommendations

The researcher recommends using a chance woodland reversion perfect to forecast a birth defect in Iraq founded on what the practical study produced, and the researcher recommends comparing this model with other regression models.

References

1. Biondo, S.; Ramos, E.; Deiros, M.; Ragué, J. M.; De Oca, J.; Moreno, P.; Farran, L.; Jaurrieta, E. "Prognostic factors for mortality in left colonic peritonitis: A new scoring system". Journal of the American College of Surgeons. 191 (6): 635–42. (2000). doi:10.1016/S1072-7515(00)00758-4. PMID 11129812.
2. Boyd, C. R.; Tolson, M. A.; Copes, W. S. "Evaluating trauma care: The TRISS method. Trauma Score and the Injury Severity Score". The Journal of Trauma. 27 (4), (1987).: 370–378. doi:10.1097/00005373-198704000-00005. PMID 3106646
3. Breiman, L. Random Forests, Machine Learning, 2001, Volume 45, Issue 1, pp 5–32.

4. Couronn_e, Raphael. Probst, Philipp.Boulesteix, Anne-Laure. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 2018
5. G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 98888 (2012)1063–1095.
6. Goodfellow, Ian; Bengio,Yoshua; Courville, Aaron. Deep Learning. MIT Press. 2016
7. Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning: data mining, inference and prediction. Springer. 2009
8. Hosmer, D.W. "A comparison of goodness-of-fit tests for the logistic regression model".(1997). Stat Med. 16 (9): 965–980.
9. Kologlu, M.; Elker, D.; Altun, H.; Sayek, I.. "Validation of MPI and PIA II in two different groups of patients with secondary peritonitis". Hepato-Gastroenterology. 48 (37), (2001): 147–51. PMID 11268952.
10. L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
11. Le Gall, J. R.; Lemeshow, S.; Saulnier, F. "A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study". JAMA. 270 (24): 2957–63. (1993). doi:10.1001/jama.1993.03510240069035. PMID 8254858.
12. Marshall, J. C.; Cook, D. J.; Christou, N. V.; Bernard, G. R.; Sprung, C. L.; Sibbald, W. J. "Multiple organ dysfunction score: A reliable descriptor of a complex clinical outcome". Critical Care Medicine. 23 (10): (1995). 1638–52. doi:10.1097/00003246-199510000-00007. PMID 7587228
13. Rokach, Lior. Maimon, Oded. Data Mining with Decision Trees; Theory and Applications. 2nd ed. World Scientific Publishing Co.