Black Sea Journal of Agriculture
doi: 10.47115/bsagriculture.1087820

Open Access Journal
e-ISSN: 2618 – 6578

Research Article
Volume 5 - Issue 3: 314-322 / July 2022

A COMPARATIVE RESEARCH ON DATA ANALYSIS WITH FACTORIAL ANOVA, LOGISTIC REGRESSION AND CHAID CLASSIFICATION TREE METHODS

Ömer AKBULUT[1]*, Ali KAYGISIZ[2], İsa YILMAZ[3]

[1]*Giresun University, Institute of science, Department of Bioprocess Engineering, 28100, Giresun, Türkiye*
[2]*Kahramanmaraş Sütcü Imam University, Agriculture Faculty, Department of Animal Science, 46000, Kahramanmaraş, Türkiye*
[3]*Muş Alparslan University, Faculty of Applied Sciences, Department of Animal Science Production and Technologies, 49000, Muş, Türkiye*

Abstract: When the data structure is large and complex, the extraction of information hidden within the data is called data mining. In the context of data mining, there are numerous methods developed for statistical data analysis. When these methods are classified as conventional-classical methods and current methods, factorial ANOVA (FANOVA) and Logistic Regression (LR) methods are shown as conventional methods, while decision trees called Classification Tree (CT) and Regression Tree (RT) can be shown as current methods. The method to be used in statistical data analysis is directly related to the researcher's hypothesis (i.e. purpose) and variable type. Therefore, the choice of data analysis method is important. In this regard, studies in which methods are examined comparatively are guiding. In this study, a dataset on which inferences could be made by ANOVA, LR, and CT methods was analyzed. With this dataset, the relationship between the birth type (single-twin) as dependent variable and the yield year and maternal age as independent variables in an Awassi sheep flock was examined. The findings of each method were interpreted in its own specific way. The methods were compared in terms of explaining the similarities and differences of the information they presented and the relationship between dependent and independent variables.It was concluded that each method offered different inferences based on purpose and perspective. It is believed that it is the right approach for researchers to determine the data analysis method appropriate to their goals by taking into account the data structure.

Keywords: ANOVA, Binary logistic regression, Classification tree algorithm, Awassi sheep, Birth type

***Corresponding author:** Giresun University, Institute of science, Department of Bioprocess Engineering, 28100, Giresun, Türkiye
E mail: omer.akbulut@giresun.edu.tr (Ö. AKBULUT)
Ömer AKBULUT https://orcid.org/0000-0002-8860-3513
Ali KAYGISIZ https://orcid.org/0000-0002-5302-2735
İsa YILMAZ https://orcid.org/0000-0001-6796-577X

Received: March 14, 2022
Accepted: June 17, 2022
Published: July 01, 2022

Cite as: Akbulut Ö, Kaygısız A, Yılmaz İ. 2022. A comparative research on data analysis with factorial ANOVA, Logistic regression and CHAID classification tree methods. BSJ Agri, 5(3): 314-322.

1. Introduction

Data constitutes the raw material of scientific research. Data can be obtained under controlled conditions through experimental studies as well as it consists of information formed in their natural environment and collected in the relevant data center. In experimental research, data is obtained by simulating the actual event. Obtaining data in this way is difficult, but analysis processes are easy.

After the data obtained depending on their actual occurrence is collected in the relevant centers, these data can reach a large data size in terms of volume, diversity, and rate of occurrence. These data can also be in a complex structure consisting of a large number of dependent and independent variables. An important part of scientific research today is comprised of the extraction of hidden information in this large and/or complex data. With a more clear expression, a dependent variable being studied is formed in a complex structure as a result of the effects of a large number of independent variables (factors). By examining the effect(s) of an independent variable(s) on the dependent variable, the researcher may aim to determine the significance, magnitude, or direction of these effects.

There are many statistical methods developed to extract information hidden in complex data. Statistical methods used for this purpose are generally called data mining.

There are many methods in data mining. These methods are widely used in fields such as economics, health, education and agriculture. The best known of these methods are Naïve Bayesian Classifiers (NBC), Artificial Neural Networks (ANN), k-Nearest Neighbor Approach, Support Vector Machines (SVM), and Decision Trees (Alev Çetin and Mikail, 2016). The use of these methods in animal husbandry was discussed by Alev Çetin and Mikail (2016) in a comprehensive review study and examples of studies in this field were presented.

The decision tree method, one of the data mining methods, contains a large number of algorithms. The major types of these algorithms are as follows: CHAID

(chi-squared Automatic Interaction Detector), Exhaustive CHAID, CART (Classification and Regression Trees) SLIQ (Supervised Learning in Quest), MARS (Multivariate Adaptive Regression Splines), SPRINT (Scalable Parallelizable Induction of Decision Trees), and QUEST (Quick, Unbiased, Efficient Statistical Tree) (Vupa Çilengiroglu and Yavuz, 2020).

To be able to make inferences from a data set based on the purpose of the research, appropriate statistical methods are used. In some cases, a data set can be analyzed with different methods for the same purpose. The most important issue here is whether the data set is in accordance with the assumptions (data volume, variable type, normality, etc.) that the statistical method to be used for its analysis considers necessary.

Another issue in choosing the appropriate analysis is the accuracy of the information produced by the method. Choosing the appropriate statistical method for the current dataset is also related to the researcher's "statistical literacy". To shed light for researchers, studies that comparatively examine the information produced by statistical methods applied to the same dataset and their reliability and the selection criteria of the correct statistical method have been conducted. Şata and Çakan (2018) examined Logistic Regression (LR) and CHAID methods comparatively in educational sciences, while Vupa Çilengiroğlu and Yavuz (2020) examined LR and CART methods comparatively on life satisfaction data. Kurt *et al*. (2008) comparatively examined Logistic Regression, Classification Trees (CT), Regression Tree (RT), and Artificial Neural Networks (ANN) methods on coronary artery diseases, while Kuyucu, (2012) comparatively examined Logistic Regression Analysis, ANNs, CART classification, and Regression Tree methods in the medical field. Kacar and Karakoc (2020) examined LR, CT, and RT methods comparatively on housing prices. CART, CHAID, and Exhaustive CHAID decision tree algorithms were studied comparatively on animal husbandry data by Tatlıyer (2020).

One of the methods of multi-factor data analysis is factorial (multi-factor) analysis of variance (FANOVA). Although the use of this method in the analysis of complex data is older than other methods, it is still widely used today. This method is used if there are effects of a large number of factors, whose subgroup numbers are equal or different, on the dependent variable (Bek and Efe, 1989; Yıldız and Bircan, 1994). Because the method is old and its use is widespread, it is a method that is better known by researchers in terms of interpreting analysis inferences. Whereas there are studies more commonly examining LR, CT, RT, or ANN algorithms comparatively in the literature, studies comparing the factorial ANOVA (FANOVA) method with the above-mentioned methods have not been encountered. FANOVA method is a more conventional and classical method than other methods. Although the inferences of the FANOVA method differ in one dimension from other methods, it also offers similar inferences. Therefore, it would be useful to compare the FANOVA method with the current methods.

The objectives of this study are: 1) interpreting the inferences produced by FANOVA, LR, and CT methods specific to all three algorithms, 2) discussing information produced by the methods in terms of their similarities, differences, or superiority by examining the inferences of the methods comparatively.

2. Material and Methods

2.1. Animal Materials

The data of the study belongs to an Awassi flock in Ceylanpinar Agricultural Enterprise located in Sanliurfa province in Türkiye. Data were obtained from the yield and breeding records kept in the flock between 2006 and 2010. In this study, all available records regarding the sex, year of yield, type of birth and maternal age (dam age) of Awassi lambs were used. Data material of the study consisted of data about 5454 head sheep that gave birth between 2006 and 2010. Mating of sheep in the enterprise is held in June and births begin in November and are completed by the end of that year. Therefore, the year in which pregnancy is provided and the yield year usually occur in the same year. The conclusion of birth as twins is a desirable condition in sheep flocks. Giving birth twin is a result of the genetic structures of the animals in addition to environmental factors such as care-feeding, pasture status, and climate in the flock during the year of pregnancy. The inheritance level of the property of being twin as birth type is low (Notter, 2008; Vatankhah and Talebi, 2008; Cottle et al., 2016); it occurs as a result of environmental factors.

2.2. Statistical Analysis

In this study, the birth type (single or twin) was considered as a dependent variable, the yield year and the maternal age as independent variables. In the enterprise from which the dataset of the study was taken, the yield year and the maternal age factors, which are thought to be effective on the birth type, were archived. Some other factors, such as the season in which pregnancy is achieved and the genetic groups formed in the flock, can be considered effective on twin births. But since they were not archived, these factors could not be studied in the data analysis. The data was analyzed by three different data analysis methods. These were FANOVA, LR and CT methods. Since the dependent (predicted) variable (birth type) is denoted by Y, and the independent (predictor) variables (yield year and maternal age) are denoted respectively by X1 and X2, the functional relationship between the dependent variable and the independent variable is written with the matrix form as follows (Equation 1):

$$Y=X\beta+\varepsilon \quad (1)$$

Where; Y is the vector of the dependent variable, X is the fixed effect matrix, β is the coefficient matrix of fixed effects, ε is the independent error vector.

2.3. Factorial ANOVA Method

In variance analysis, the dependent variable *Y* must be continuous and the independent variables *X* (factor) must be discrete. But variables exhibiting binomial distribution fit the normal distribution assumption if the volume of data (*n*) is too large (Yıldız et al., 2020). In this case, the analysis of the data can be done using the ANOVA approach, which prioritizes the normality assumption. In this context, the biometric model used in data analysis for the FANOVA method in this study is written as follows (in this model, the linear biometric model was used), (Equation 2):

$$Y_{ijk} = \mu + a_i + b_j + ab_{(ij)} + e_{ijk} \quad (2)$$

where;

Y: Observation vector of birth type

μ: Population mean of the birth type

a_i: Fixed effects of levels belonging to the variable of yield year

b_j: Fixed effects of levels belonging to the variable of maternal age

$ab_{(ij)}$: Interaction effect of yield year and maternal age

e_{ijk}: Random residuals (random error); e_{ijk} ~N; $(0,\sigma^2_e)$.

2.4. Chi-Square Independence Test and Logistic Regression (LR) Method

Chi-Square independence test is preliminary test for LR analysis. The independent variables found to be significant according to chi-square independence test are included in the LR analysis model. The analysis process related to Chi-Square independence test was explained by Yıldız *et al*. (2020).

Simple and multiple linear regression methods give accurate results under the assumptions that the dependent variable and independent variables are continuous variables with normal distribution, the independent variables are measured without error, and the error term of dependent variable is $e \sim N(0,\sigma^2)$ (Özdamar, 1999). If the dependent variable is discrete, the appropriate data analysis method for the relationship between the independent variables (can be discrete or continuous) and the dependent variables is logistic regression. If the dependent variable has two results (binomial), the method is called "Binary Logistic Regression"

When a binary dependent variable is denoted by Y_i=(0,1) and the independent variables by $X = (X_1, X_2, \ldots X_p)$, the regression model of the binary variable is written as (Equation 3):

$$Y_i = \beta_0 + \beta_1 X_i \quad (3)$$

Here, since the Y_i categorical dependent variable shows the Bernoulli distribution, the expected value of *Y* is $0 \le E(Y_i) = \pi \le 1$; when the logit transformation of it is applied, the final model for the binary logistic regression is as follows (Bircan, 2004; Vupa Çilengiroğlu and Yavuz, 2020), (Equation 4):

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \quad (4)$$

or

$$\pi(x_i) = [1 + \exp(-\beta_0 - \beta_1 X]^{-1}$$

In the logistic regression model, the estimation of the coefficients of the variables is obtained using the "maximum likelihood" method. The significance of these estimated coefficients is determined by the "G statistic" or the "Wald test" (Çokluk, 2010).

Exp(β) values included in the logistic regression summary tables are exponential logistic regression coefficients. This value is also the Odds ratio (OR) calculated for each variable. For OR, $0 < OR < \infty$ can be written and it is the ratio of two values to each other, such as "occurrence rate" and "non-occurrence rate". In other words, OR logarithm does not take a negative value (Çokluk, 2010; Vupa Çilengiroğlu and Yavuz, 2020).

OR is a metric measurement that bases on a level for each of the discrete variables, and by accepting this level "1", it refers to other levels as a multiple of this.

In logistic regression analysis, the model fit reported with the measures of the Cox and Snell R^2 statistics, the Nagelkerke R^2 statistics, the Hosmer and Lemeshow test, and the overall χ^2 test results. The Cox and Snell R^2 statistics and the Nagelkerke R^2 statistics tend to take small values. Therefore, reporting these R^2 values is not recommended (Alpar, 2011; Şahin, 2017). This R^2 values between 0.20 and 0.40 indicates that the accuracy of the model is high (Şenel and Alatlı, 2014). If the probability value of the Hosmer and Lemeshow test is P>0.05, it is an indication that the model is fit.

In LR analysis, the dependent variable is natural-class or must be turned into the natural-class position. Independent variables can be discrete or continuous. This analysis method requires to have large sample sizes (at least 15; ideal 20 and above) in each subgroup of each independent variable. There are no other assumptions that restrict the method other than these two assumptions. Therefore, LR is a much preferred data analysis method in the analysis of the relationship between the dependent variable and the independent variables. Another important reason why the method is preferred is that LR also does not require the relationship between the independent and dependent variables to be linear. The functional relationship can be exponential or polynomial. LR can produce non-linear models by assuming that there is a logit relationship between dependent and independent variables. LR analysis is a useful method that performs logarithmic transformations to bring the relationship to a linear form by preserving the nonlinear relationship in cases where the relationship between the dependent and independent variables is nonlinear (Şata and Çakan, 2018).

A detailed explanation of the statistics produced by this

method was made by Çokluk (2010), while other reasons for choosing the method were explained by Çokluk (2010), Şahin (2017) and Şata and Çakan (2018). Analysis findings that should be reported in studies using LR analysis were summarized in a review study conducted by Şenel and Alatlı (2014). For further information on LR analysis, these sources can be referred.

2.5. Decision Tree and CHAID algorithm

In the context of data, another method applied for determining and analyzing the relationship structure between dependent and independent variables is the decision tree. The aim of decision trees is to estimate the outcome values of datasets by developing a model based on data mining (Güner, 2014). Multiple regression and LR analyses are considered classical methods in relationship analyses (Gacar and Karakoç, 2020). The decision tree method is an up-to-date method and has been widely used in data analysis in recent years.

The structure of decision trees is similar to the natural tree structure, that is, it is in the form of roots, branches and leaves. Decision trees begin with the root, which covers all observations in the dataset, and are divided into branches that divide the data into subgroups. In the tree structure, separated from the root to the branches, each knuckle is named "node" (Pehlivan, 2006; Gaçar and Karakoç, 2020). The test process for each divided node is performed, and the branching process continues consecutively to the last node. After the separation process is finished, inferences are made based on the ratios belonging to the divided nodes and the categories within the last branch (group).

In the decision tree method, heterogeneous datasets are divided into homogeneous subgroups depending on the dependent variable. According to Dangeti (2017), the separation process is carried out by examining values such as entropy, Chi-Square, variance reduction criterion, and homogeneity structure in nodes (Özgür and Doğanay Erdoğan, 2020). Using these techniques, homogeneity measurements are carried out from the root node to the terminal nodes. The resulting values on the terminal node are the values estimated for the dependent variable. A large number of algorithms are used to create a decision tree. The main ones of these are CHAID, exhaust CHAID, CART, SLIQ, MARS, SPRINT and QUEST algorithms. In decision tree algorithms, the method is called a Classification Tree (CT) if the dependent variable is discrete, and a Regression Tree (RT) if the dependent variable is continuous (Breimann et al., 1984; Özkan, 2012; Koç, 2016; Eyduran et al., 2016).

Because the dependent variable discussed in this study is discrete, the decision tree created will be the classification tree. It is also reported that the CHAID algorithm works better in discrete data (Şata and Çakan, 2018). In this direction, the CHAID algorithm was selected to create the classification tree.

In this research, the results of the analysis in all three methods will be interpreted separately. In addition, the following considerations will be examined in relation to all three methods:

a) The information they provide and the level of model-specific fit
b) Significance states of the independent variables
c) Compatibility of similar statistics offered by methods
d) Information specific to the methods (i.e. inferences found in one method not found in the other method).

3. Results and Discussion

The relationship between the independent variables (yield year and maternal age) and the dependent variable (birth type) was analyzed by FANOVA, LR, and CT methods, and the findings are summarized below.

3.1. Results of the Factorial ANOVA

If the number of observations is too large, the variables in the binomial (binary) property show a normal distribution. The analysis of variance (ANOVA) results belonging to yield year and maternal age variables, which are thought to have an effect on the binary birth type (single and twin) variable, and the interaction of these two variables are summarized in Table 1.

According to the results of factorial (multi-factor) analysis of variance (FANOVA), the effect of yield year and maternal age on birth type (single or twin) was found to be significant ($P<0.001$) (Table 1). In other words, in terms of birth type, there were statistical differences between birth years and maternal ages. Also in terms of birth type, the interaction of yield year and maternal age was not statistically significant ($P=0.071$). However, the observed probability was very small. This indicates that there may be differences between some ages in terms of birth years in further analysis. Since the aim of this study was to compare analysis methods (FANOVA, LR and CT), no further evaluation related to the interaction was performed. It was satisfied with the evaluation of the main variables.

Table 1. ANOVA results related to the birth type

Source	df	Mean of squares	F	P	η^2	Power
Yield year	4	6.215	32.149	<0.001	0.023	1.000
Maternal age	3	3.220	16.657	<0.001	0.009	1.000
Yield year x Maternal age	12	0.320	1.653	=0.071	0.004	0.862
Error	5434	0.193				

R^2= 0.895 (Adjusted R^2= 0.895).

Duncan's multiple comparison test was used to determine the statistically significant factors' subgroups one of which is different from the other. When multiple comparison test results were evaluated in terms of yield year (Table 2), it was determined that the highest twin birth rate occurred in 2010, and it was followed by 2008. While the lowest twin birth rate occurred in 2009, the difference between 2006 and 2007 was not statistically significant. When the confidence interval is evaluated, the twin birth rate can go down to 25%, or up to 45% in the 95% confidence interval. Also when confidence limits for years are evaluated, while the upper and lower limits for similar years (2006-2007) overlap, the limits for different years are separate. For example, when evaluating the years 2007 and 2008, it is seen that the upper limit of 2007 (1.30) is statistically different from the lower limit of the following 2008 year (1.33). A similar situation is also observed when other years are evaluated. In research studies, confidence interval findings present additional information about between what values the point estimate (mean or ratio) will be.

When evaluating the twin birth rate in terms of maternal age, this rate is the lowest in 3-year-old dams. The twin birth rate increased with age and occurred at the highest rate (about 34%) in 6-year-old dams. However, in terms of dams aged 4, 5, and 6, the differences between twin birth rates were not statistically significant. In these age groups, twin births occurred at rates close to each other (between 23% and 32%). When confidence limits were evaluated, the upper limit value of 3-year-old dams (1.25) was found to be different from the lower limit value of 4-year-old dams (1.28). However, in the statistically similar age groups of 4, 5 and 6, the lower and upper limits did not differ from each other. This suggests that the difference in age groups is not significant. In summary, it is necessary to express that presentation of confidence interval values as well as point estimates (mean, ratio, etc.) makes a significant contribution to statistical inferences. Therefore, confidence interval values should also be presented as findings.

3.2. Chi-Square and LR Results

Dataset this study meets the assumptions of binary LR analysis one-on-one. The dependent variable (birth type) is binomial variable, single and twin. In LR, independent variables can consist of a combination of discrete and continuous variables. Here the independent variables are discrete. The LR method is sample size-sensitive, and it is necessary to have at least 15 (ideal 20 and above) observations in subgroups of each factor variable. The volume of observations in this study is quite large. If the independent variable or variables are categorical, performing Chi-Square (χ^2) independence analysis as a preliminary analysis of LR can be helpful in creating the LR model. Chi-Square analysis results values for yield year and birth type are presented in Table 4, and Chi-Square analysis results for maternal age and birth type are presented in Table 5.

Table 2. Birth type statistics by years

Yield Year	Mean	SEM	CI: 95%	
			Lower	Upper
2006	1.25[c]	0.017	1.22	1.28
2007	1.27[c]	0.016	1.24	1.30
2008	1.36[b]	0.016	1.33	1.39
2009	1.19[d]	0.016	1.16	1.22
2010	1.42[a]	0.016	1.39	1.45

[a, b]Means marked with different letters are different with an error of P<0.05.

Table 3. Birth type statistics by maternal age

Maternal Age	Mean	SEM	CI: 95%	
			Lower	Upper
3	1.23[b]	0.009	1.21	1.25
4	1.30[a]	0.012	1.28	1.32
5	1.32[a]	0.015	1.29	1.35
6	1.34[a]	0.020	1.30	1.38

[a, b]Means marked with different letters are different with an error of P<0.05.

Table 4. Distribution of birth type by yield years, and χ^2 test results

Birth Type	N=5454	Yield Year					Total
		2006	2007	2008	2009	2010	
Single	n	856	745	719	853	761	3934
	%	76.7	73.7	66.1	83.6	62.4	72.1
Twin	n	260	266	368	167	459	1520
	%	23.3	26.3	33.9	16.4	37.6	27.9

Pearson Chi-square test results χ^2 =159.999; df=4; P<0.001; Actual P=6.43 x10^{-33} R^2=0.167.

Table 5. Distribution of birth type by maternal age, and χ^2 test results

Birth Type	N=5454	Maternal age (dam age)				Total
		3	4	5	6	
Single	n	1950	1024	631	329	3934
	%	76.8	68.4	68.4	66.2	72.1
Twin	n	589	472	291	168	1520
	%	23.2%	31.6	31.6	33.8	27.9

Pearson Chi-square test results χ^2 =52.604; df=3; P<0.001; Actual P=2.23 x10^{-11} , R^2 =0.098.

Chi-Square independence test results for the relationship between the birth type variable and both independent variables were found significant (p<0.001). When the actual probability value (P) is evaluated, it is seen that the relationship between the yield year and twin birth and the measure of this relationship (R^2) is higher. According to the results of Chi-Square analysis, the R^2 value for the yield year and birth type correlation is 0.167, and for the maternal age and birth type correlation is 0.098. According to these findings, each factor can be involved in the LR model and their effects are expected to be significant.

3.3. Binary Logistic Regression Analysis Results

The LR analysis results of the relationship between dependent variables and independent variables examined in the dataset handled as defined in the method section are summarized in Table 6. In the presentation of the findings, the criteria proposed by Şenel and Balatlı (2014) were taken into account.

Table 6. Binary Logistic regression analysis results

Factors	β	SEM	Wald	df	P	Exp (β)	For EXP(β) CI 95%	
							Lower	Upper
Yield year (a)			153.940	4	<0.001			
2007	0.152	0.102	2.242	1	=0.134	1.165	0.954	1.422
2008	0.520	0.096	29.229	1	<0.001	1.683	1.393	2.032
2009	-0.433	0.111	15.180	1	<0.001	0.649	0.522	0.806
2010	0.710	0.094	57.619	1	<0.001	2.034	1.693	2.443
Maternal age(b)			53.907	3	<0.001			
4	0.386	0.074	27.251	1	<0.001	1.472	1.273	1.701
5	0.476	0.087	30.059	1	<0.001	1.610	1.358	1.909
6	0.570	0.108	27.905	1	<0.001	1.769	1.431	2.186
Constant	-1.446	.081	321.584	1	<0.001	0.235		

(a) Reference year: 2006; (b) Reference maternal age (3 year old); 2LL= 6239.051; Cox and Snell R^2 =0.039; Nagelkerke R^2 =0.056; $\chi^2(7)$= 209.488 and P<0.001; Hosmer and Lemeshow test: P=0.124.

When the results of LR analysis were examined in the dataset, the effects of both yield year and maternal age on the dependent variable (birth type) were found to be significant (P<0.001), as in the FANOVA method.

In terms of yield year, the twin-birth rates of 2006 are not different from 2007 but different from other years. When β and Exp(β) coefficients are examined, it is observed that 2009 was the year that reduced the twin-birth rates. In addition, the highest twin birth rate occurred in 2010. The twin birth rate in 2010 is nearly double that of 2006 (Exp(β)=2.034). These findings in the LR analysis are quite similar to the findings obtained by the FANOVA method.

When the findings related to maternal age are evaluated, it is observed that the effects of old ages (4, 5, 6) on twin-birth rates are positive. Increase in twin birth rate regularly increased with increasing age, and the rate of twin birth in 6 older mothers was about 1.8 times more than 3 years-old mothers (Exp(β)=1.769). These findings related to maternal age are also consistent with the findings of FANOVA.

In binary LR analysis, the results related to the model fit and related to what extent the model explains the variation in the dependent variable should be evaluated first. In this context, when Cox & Snell and Nagelkerke "pseudo" R^2 statistics for model performance (Şenel and Alatlı, 2014) are evaluated, it is observed that both values are quite small (0.039 and 0.056, respectively). Şenel and Alatlı (2014) report that for good model performance, these statistics must be between the ranges of 0.20 and 0.40. Another criterion that expresses to what extent the model explains the variation in the dependent variable in LR analysis is the Hosmer and Lemeshow statistics. This statistic is a probability value, and P> 0.05 indicates a good model fit (Çokluk, 2010). Here, since P=0.124 for the Hosmer and Lemeshow test statistic, it can be said that the model fit is good.

3.4. Classification Tree Results

The results of this method are presented as figure. In the dataset studied by the classification tree method, root, branch and leaf formation and nodes belonging to

homogeneous subgroups formed depending on birth type are shown in Figure 1. As seen in Figure 1, the effects of both studied factors on the birth type (twin born rate) are significant (P<0.001); while yield year is the first-degree effective factor, the maternal age is the second-degree effective factor. According to the CHAID algorithm, the data are summarized under 10 homogeneous subgroups (nodes) in terms of the twin-birth rate. According to age factor, 2007 and older years (here 2006 and 2007) were collected in the same node. The years of 2008, 2009 and 2010 are homogeneous within themselves in terms of years and heterogeneous groups between them. 2007 ended as the last node. However, other years were divided into two branches depending on maternal age. In terms of maternal age, these branches are in the form of 3-year-old dams and dams older than 3-years-old. The significance status of the examined independent variables (yield year and maternal age), and the differences belonging to subgroups of these factors respectively show similarity to the results of the FANOVA and LR Analyses.

The characteristic that this method offers differently from other methods and is not found in other methods is that homogeneous groups in terms of twin birth rate and the lowest and highest twin birth rate groups are determined. The total twin-birth rate in the flock is 27.9%. In 2009, the twin-birth rate was lowest with 12.3% in 3-year old dams (node 7). The highest twin-birth rate was in dams aged 4 and older (4, 5, and 6) in 2010 with 44.2%.

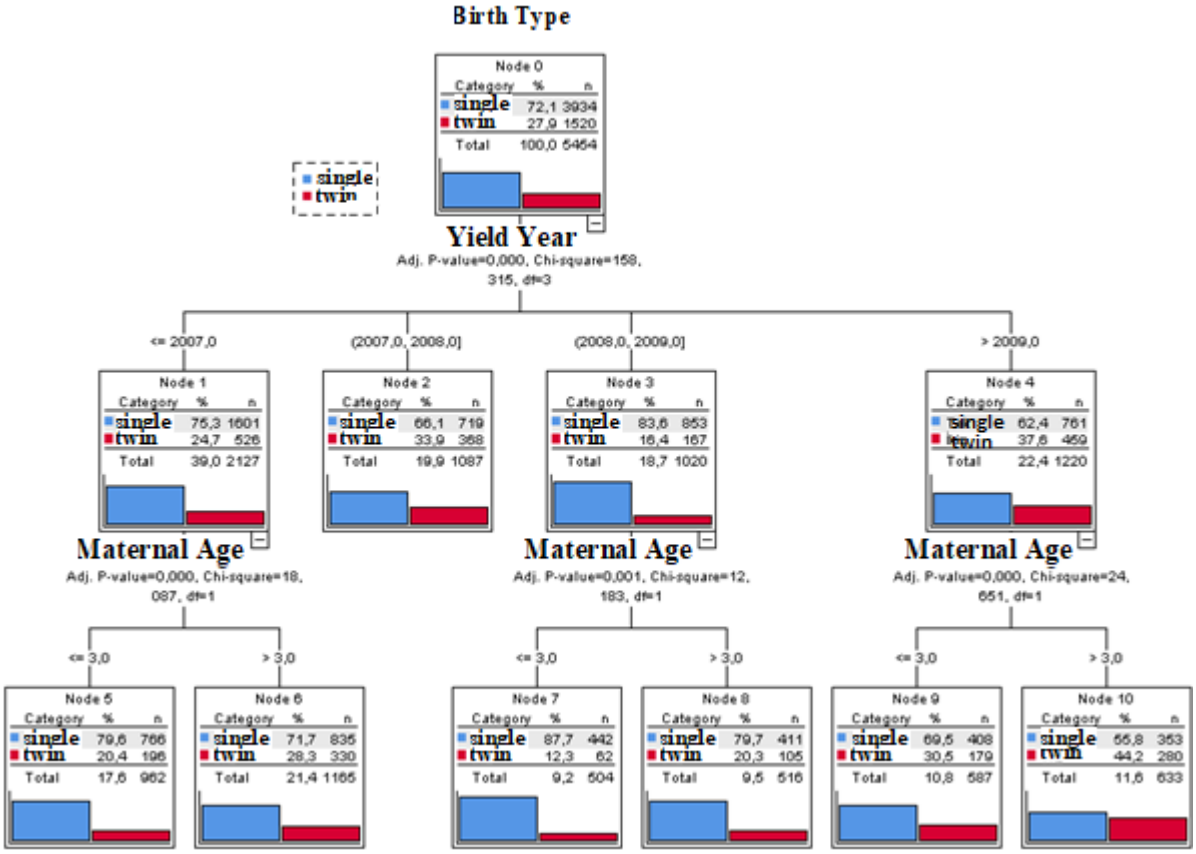

Figure 1. Classification Tree of factors that affect the birth type.

3.5. Comparative Study of the Methods

To be able to compare the studied methods, the information presented by the models is summarized in Table 7. In all three methods, the significance (P) of factors and the effect orders are similar. The most effective variable in all three methods is yield year, while the second is maternal age.

When the different inferences of the methods are evaluated, it is seen that FANOVA additionally offers information about the interaction of factors. This information is not presented directly by the other methods. However, when the definition of interaction is taken into account, interaction can be inferred from the CT structure. In the CT structure, although the maternal age in all other years except 2008 is grouped as 3 years and other ages, the termination of the node in this year can be interpreted as the interaction of maternal age and the yield year.

The LR method does not offer an inter-factor interaction information. However, based on a level defined for each factor, it presents the values, which the dependent variable will receive at other levels, as a layer of this level. This inference is not directly presented in other methods. The CHAID CT algorithm classifies different groups and presents homogeneous subgroups as more descriptive. In addition to a visual design, it also offers significance results and subgroup statistics.

In FANOVA and LR methods, confidence intervals of point

estimates can also be shown. Classification trees do not offer an inference in this direction.

FANOVA and LR methods provide R^2 statistics. However, the CT algorithm does not report these statistics. In this research, R^2 statistic is 0.895 for FANOVA. Cox and Snell R^2 is 0.039 and Nagelkerke R^2 is 0.056, for LR. The FANOVA and LR models are not comparable in the magnitude of the R^2 statistic. FANOVA R^2 statistics and LR R^2 statistics are evaluated on their own. While the FANOVA R^2 statistic is very close to one, Cox and Snell R^2 and Nagelkerke R^2 statistics are well below the 0.20 to 0.40 range.

In the literature, any study comparing the ANOVA method and the LR and CT methods was not encountered. The results of some studies examining LR and CT methods are summarized below in terms of the method proposing.

In their study in which the different forms of LR analysis and CT methods are examined comparatively, Vupa Çilengiroğlu and Yavuz (2018) reported that the method explaining the dependent variable best was the LR c=0.4 form. In a study in which they studied CHAID analysis and LR analysis comparatively, Şata and Çakan (2020) considered the use of CHAID analysis more appropriate in classification studies because CHAID analysis gave more detailed and understandable results than logistic regression analysis and explained the common effect between independent variables.

In this study, when the inferences of the methods are evaluated generally, it is seen that the results of all three methods are similar in terms of the significance of the independent variables, their significance order, and explaining the dependent variable. In addition, each method has its own inferences that are unique (and not in other methods). In this context, while FANOVA classifies interaction between factors, the LR method classifies layer values belonging to other subgroups based on one level of the factor. The CT, on the other hand, classifies data by presenting homogeneous groups at the last ends (nodes).

Table 7. Summary results of the models

Factors	sd	Factorial ANOVA		Binary Logistic Regression		Classification tree
		Test Statistic F	P	Wald Test Statistic	P	FR
Yield Year	4	32.149	$1.6E^{-26}$ [a]	153.940	$2.91E^{-32}$	1st Factor
Maternal age	3	16.657	$9.0\ E^{-11}$	53.907	$1.17E^{-11}$	2nd Factor
Interaction	12	1.653	=0.071			
Model	→	2318.460	≅0.0 (df=20)	209.448	$1.12\ E^{-41}$ (df=7)	-
Measures model fit		R^2=0.895[b]		R^2 =0.039[c] R^2 =0.056[d]	P=0.124[e]	

[a]=1.6 x 10^{-26}; [b]= very close to zero; [c]= Cox and Snell R^2=0.039; [d]= Nagelkerke R^2=0.056 ; [e]= Hosmer and Lemeshov probability P=0.124; Note: [b] vs [c] and [d] are not comparable, FR= factor ranking.

4. Conclusion

In conditions where the independent variables are discrete or continuous, but the dependent variable is binomial (binary), the appropriate method that analyzes the relationship between variables in a data structure of sufficient size is the binary LR method. If the number of observations is too large, binomial variables can also be analyzed with ANOVA. If the dependent variable is categorical, and the aim of the researcher is to summarize the data in homogeneous subgroups in the independent variable in terms of the dependent variable, the appropriate method is the classification tree. In this study, a dataset on which inferences could be made by all three methods was analyzed. In this context, the relationship between the dependent variable (birth type-single or twin) and the independent variables (yield year and maternal age) in an Awassi sheep flock was examined.

When the methods were examined comparatively, the following conclusions were reached:

1) In all three methods, the significance (P) of factors and the effect orders are similar. In research, the yield year is more effective in all three methods, while the second is the maternal age.
2) When evaluating the different inferences of the methods, it is seen that FANOVA additionally offers information on the interaction of factors.
3) However, taking into account the definition of interaction, the researcher can obtain information about the existence of the interaction from the CT structure.
4) Based on a level defined for each factor, LR presents the values, which the dependent variable will receive at other levels, as a layer of this level. This inference is not directly presented in other methods.
5) The CHAID CT algorithm classifies different groups and presents homogeneous subgroups as more descriptive. In addition to a visual design, it also offers significance results and subgroup statistics.
6) In FANOVA and LR methods, confidence intervals of point estimates can also be presented. Classification trees do not offer an inference in this direction.
7) R^2 statistics, which is a measure for the explanation level of a dependent variable by independent variables, are presented in the FANOVA and LR methods, but the CT algorithm does not report these statistics. However, FANOVA R^2 statistics and LR R^2 statistics (Cox -Snell R^2 and Nagelkerke R^2 statistics)

are not comparable in terms of fit of FANOVA and LR models.

Author Contributions

A.K. (50%) and I.Y. (50%) designed the study and collected the data, critically reviewed. Ö.A. (100%) analyzed the data and wrote the article. All authors reviewed and approved final version of the manuscript.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

Acknowledgments

The authors thank the Ceylanpinar Agricultural Enterprise employees who archive the data and the administrators who allow the use of the data.

References

Alev Çetin F, Mikail N. 2016. Data mining aplications in livestock. Turk J Agric Res, 3: 79-88.

Alpar R. 2011. Applied multivariate statistical methods. Detay Publishing, Ankara, Türkiye, 6th ed., pp: 858.

Bek Y, Efe E. 1989. Research and application methods I. 1th ed., Çukurova University, Agriculture Faculty, Textbook. Publication No 71. Adana, Türkye, pp: 395.

Bircan H. 2004. Logistic regression analysis: An application on medical data. Kocaeli Univ J Social Sci Institute, 2: 185-208.

Breiman L, Friedman JH, Olshen RA, Stone CF. 1984. Classification and regression tree. Wadsworth International Group, Belmont, California, US, pp: 3-7.

Cottle DJ, Gilmour AR, Pabiou T, Amer PR, Fahey AG. 2016. Genetic selection for increased mean and reduced variance of twinning rate in Belclare ewes. J Anim Breed Genetics, 133: 126-137.

Çokluk Ö. 2010. Logistic regression analysis: Concept and application. Educ Sci Theor Pract, 10: 1357-1407.

Dangeti P. 2017. Statistical for machine learning. 1th ed., Packt Publishing Ltd, Birmingham, UK, pp: 442.

Gacar BK, Kocakoç ID. 2020. Regression analyses or decision trees? Manisa Celal Bayar Univ J Social Sci, 18: 251-260.

Güner ZB. 2014. Cart and logistic regression analysis in data mining: An application on pharmacy provision system data. Soc Secur Profes Assoc J Soc Secur, 6: 59-61.

Koç Y, Eyduran E, Akbulut Ö. 2016. Application of regression tree method for different data from animal science. Pakistan J Zool, 49: 599-607.

Koç Y. 2016. Application of Regression Tree Method for Different Data from Animal Science. MSc thesis, Iğdır University, the Institute of Science and Technology, Iğdır, Türkiye, pp: 75.

Kurt İ, Türe M, Kurum AT. 2008. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst Appl, 34: 366-374.

Kuyucu YE. 2012. Comparison of logistic regression analysis (LRA), artificial neural networks (ANN) and classification and regression trees (C&RT) methods and an application in medicine. MSc thesis, Gaziosmanpasa University, Institute of Health Sciences. Tokat, Türkiye, pp: 112.

Notter DR. 2008. Genetic aspects of reproduction in sheep. Reprod Domestic Anim, 43: 122-128.

Özdamar K. 2004. Statistical data analysis with package programs II. Multivariate Analysis. 5th ed., Kaan Publishing House, Eskisehir, Türkiye, pp: 649.

Özgür EG, Doğanay Erdoğan B. 2020. Regression tree approach in computer adaptive testing (BUT) applications: Evaluation of standard CAT algorithm using a psychometric model with regression decision trees on artificial data. J Ankara Health Sci, 9(1): 161-167.

Özkan K. 2012. Modelling ecological data using classification and regression tree technique (CART). Süleymen Demirel Üniv Fac Forest J, 13: 1-4.

Şahin O. 2017. Determining the important risk factors in preferring Ayvalık for touristic purpose using the method of logistic. Electronic J Soc Sci, 16(61): 647-660.

Şata M, Çakan M. 2018. Comparison of results of CHAID analysis and logistic regression analysis. Dicle Univ J Ziya Gökalp Fac of Educ, 33: 48-56.

Şenel S, Alatlı B. 2014. A review of articles used logistic regression analysis. J Measur Eval Educ Psychol, 5: 35-52.

SPSS 2011. SPSS for Windows, Version 20, SPSS Inc., Chicago, US.

Tatlıyer A. 2020. The effects of raising type on performances of some data mining algorithms in lambs. KSU J Agric Nat, 23: 772-780.

Vatankhah M, Talebi MA. 2008. Heritability estimates and correlations between production and reproductive traits in Lori-Bakhtiari sheep in Iran. South African J Anim Sci, 38: 110-118.

Vupa Çilengiroğlu Ö, Yavuz A. 2020. Comparison of predictive performance of logistic regression and CART methods for life satisfaction data. European J Sci Tec, 18: 719-727.

Yıldız N, Akbulut Ö, Bircan H. 2020. Introduction to statistics, 14th ed., Culture and Education Foundation Publishing House. Erzurum, Türkiye, pp: 326.

Yıldız N, Bircan H. 1994. Research and application methods in statistics. 2th ed., Agriculture Faculty Publication No: 697. Erzurum, Türkiye, pp: 266.