Final Project
MATH 6333 90L: Statistical Learning
Dr. Tamer Oraby

Harrinson Arrubla
harrinson.arrubla01@utrgv.edu

School of Mathematical and Statistical Sciences
University of Texas Rio Grande Valley
Edinburg, Texas, USA

December 8, 2021
Fall 2021

Introduction

The data information were collected on cardiotocographies (CTG) by experts obstetricians used to monitor fetal heartbeat (fetal heart rate - FHR) and uterine contractions during pregnancy and labor. According to US National Library of Medicine National Institutes of Health up to 50% of the CTG reports evidence not reliable pathological patterns which might be classified as false positives. Additionally, there are some factors that would affect CTG such as:

Maternal	**Fetoplacental**	**Fetal**	**Exogenous**
Physical activity	age of gestation	movement	noise
Posture	umbilical cord compression	fetal behavioral states	medication
Uterine activity	placental insufficiency	stimulation to wake the fetus	smoking
Body temperature (fever)	chorioamnionitis	hypoxemia	drugs
Fluctuations in blood pressure			

Indeed, by hand I have been closely related to CTGs, and FHR analysis. Interviewing some doctors at the hospital taking advantage of my long stay, they mention that it is usual pregnant women assist there because of movement reduction which according to the following data is an explanatory variable in all the presented models in this report.

According to the collected data, pregnant women usually do not know how to correctly check fetal movement, and that is one of the reasons they assist to the hospitals for a fetal medical review which usually starts with a FHR analysis. CTGs are not 100% medical processes to check not only FHR but also fetal health in general, and that might be due to the fact the fetal growth is constantly changing, and for that reason when there is some issue regarding to the fetal health the prenatal OBGs ask for a constant follow up which consist in at least two to three checks per week.

Aims

In this report, I present a wide description about the data using statistical methods in two sections: regression and classification analysis. The regression analysis might shows us statistics about the data, and classification for characterizing the data using the explanatory variables using advance statistical leaning methods to analysis the dataset to make inferences.

1 Methods

In this section, all the methods for the multilinear regression analysis over the white wine are presented in order. This dataset is available for anyone on-line wine quality. The data information were collected based on the red and white variants of the Portuguese "Vinho Verde" wine. The dataset contains objective tests data such as:

- Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates, Alcohol.

and sensory data such as:

- Quality.

This DATASET is highly versatile. Indeed, not only the DATASET source recommend multilinear regression but also related papers introduce some other statistical methods to analyze, extract and present data from this DATASET such as classification. Nevertheless, our aim will be to implement multilinear regression using the quality as the output variable.

In each method, I decided to use it as numerical from 1 to 3, and as characters "normal", "suspect", and "pathologic". First, I take a quick look into the data using "head" function, and "name" function to validate that the data is well spelled.

At the end I will show the results for each method.

1.1 Data Cleaning and Enhancement

In this dataset all variables are abbreviated for better analysis process table 1, and grouped in two sections to perform the statistical analysis: explanatory and response variables. "CLASS" as the response variable, and the rest as explanatory variables. The data might not be transformed into numbers or characters due to the fact each process might be using the response variable differently.

1.1.1 Statistics. To start building any model, the variable analysis is crucial. First, taking a quick look into the data using "head" function we check some values about the dataser. The "name" function validates that the data is well spelled. The usual way to analyze the data is statistics and correlations.

Before, I remove non relevant variables such as:

- Width, Min, Max, Nmax, Nzeros, Mode, Mean, Median, Variance and Tendency.

These variables are based on other variables from the original data which might interfere with the analysis. Indeed, the "summary" or "stat.desc" function (*pastecs library*) present compressed that data. Those are the functions that I load to analyze the data. The analysis results are:

- There is no missing values in the data.

- In general there are 2126 observations.
- All the variables do not have minimum variables except from the tendency. Nevertheless, that variable is not a basis variable but one resulting from others.
- The minimum baby fetal heart beats base is 106 per minute, and the maximum 160 per minute. According to Johns Hopkins Medicine the fetal health rate average is in between 110 and 160. So, the fetal health max beats base is 160, and matches with the literature. On the other hand, the minimum dataset fetal heart beat base min value is 106 which is not in the average range. This case is named fetal bradyarrhythmia.
- All the explanatory variables are continuous.

Additionally, implementing basic grouping selection in the data shows that there is a high (77.85%) of cases where the fetal state is normal, (13.88%) wealthy suspect fetal, and unfortunately (8.28%) fetal in pathological state. This implies that the data is not uniformly grouped which might not allow some statistical methods fit as we expect to.

Lastly, I performed duplication analysis which might guide to a biased analysis because of data collection error. Nevertheless, the number of repetitions is not significant (14 out of 2126) which shows that there is not evidence to affirm that the data repetitions is because of mismanagement of the cardiograms.

1.1.2 Correlation. The "GGally" package describes a widely complete correlation, density and frequency table all in one included in the same plot. For that reason, we select this rather than other packages. Correlation plot shows not only correlation in between the variables but correlation in each variable with the response variable and order by strong positive (top), not or medium (center), and negative (bottom) correlations. The results of the analysis are:

- In the data we found that there is not strong correlation in between the variables. The variables analysis are classified as quantitative inputs. There are not transformation of quantitative inputs or basis expansions.
- There are negative and positive correlations. For instance, the Number of deceleration per second with the mean value of short term variability. Why? The cardiogram shows results show that on average with a range of 3-5 BPM the fetal heart rate speeds up slightly and the slows down slightly.
- There is an important number of outliers for each of the explanatory variables presented in the data. So, performing linear regression might be not a good and optimal idea. Nevertheless, the concentrated data look normally distributed which might influence to the well fitting model.

1.1.3 Training and testing sets Before starting the entire process we select the eighty percent of the actual data as the training set to perform the analysis, and the testing set to test the model. This is usually done to enhance the time processing model, and set the algorithm not only for this dataset but for general related datasets. I do not

include validation data because I could not find another dataset related to FHR from a reliable source.

1.2 Multinomial Log-Linear Model

Regression models, as it is described in the introduction, are not "black boxes" that do not described the variable relations and do not show much more description about the process but the result. I decided to use one powerful regression model belonging to Generalized Additive Models as being an extension for the traditional logistic regression model. The model will run for 3 possible classifications 2 independent binary logistic regression model. To perform the model, I used "nnet" package and "multinom" function.

The output represents the relation in between the inputs and the log odds i.e. the log of the ratio probabilities which might be interpret as the probability of selecting suspect fetal rate health vs normal fetal health rate, and the probability of selecting pathologic fetal health rate vs normal fetal health rate.

1.2.1 Output This model-running output includes some iteration history and include the final negative log-likelihood 497.01. The double of this value is the model's residual deviance equal to 994.02.

1.2.2 P - Values Here, I decided to perform manually the p-values to measure the significance level for all the variables in each of the models. The method is using the coefficients and standard errors from the multinomial regression, and we calculate one minus the "normalized" results. As it can be observed, the first log-linear equation (pathologic vs normal) has almost all variables significant respect to the log odds response variable but the FHR baseline and mean value of long term variability. On the other log-linear equation (suspect vs normal) has two not significant variables: mean value of short term variability and mean value of long term variability.

1.2.3 Best Multinomial Log-Linear Model After calculating the p-values that evidence the significance values, I decided to perform a better version of this model removing the non significant variables from the model. The non significant values might be removed by two options:

- Removing the non significant values manually. The result is:

$$CLASS \sim LB + AC + FM + UC + DL + DS + DP \\ + ASTV + MSTV + ALTV + MLTV$$

  with an AIC and residual deviance of 1122.859 and 1082.85 respectively.

- The "step" method might be useful in this case. Reading the manual this specific case belongs to the acceptable modeling regression that fit into the implementation using the "backward" direction and $k = log$. So, the resultant model is:

$$\text{odds}(path.\ VS\ nor.) \sim \beta_0 + \beta_1 LB + \beta_2 AC + \beta_3 FM + \beta_4 UC + \beta_5 DL + \beta_6 DP + \beta_7 ASTV + \beta_8 MSTV + \beta_9 ALTV + \beta_{10} MLTV$$

$$\text{odds}(susp.\ VS\ nor.) \sim \beta_0 + \beta_1 LB + \beta_2 AC + \beta_3 FM + \beta_4 UC + \beta_5 DL + \beta_6 DP + \beta_7 ASTV + \beta_8 MSTV + \beta_9 ALTV + \beta_{10} MLTV$$

with an AIC and residual deviance of 1024.175 and 980.1753 respectively.

1.2.4 Final Multinomial Log-Linear Model After the best multinomial logistic model setting, I decided that the "step" model is the best choice. This model-running output includes some iteration history and include the final negative log-likelihood 490.087. The double of this value is the model's residual deviance equal to 980.1753.

$$\beta_{pathologic} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \\ \beta_9 \\ \beta_{10} \end{bmatrix} = \begin{bmatrix} 2.449794 \\ 0.016689 \\ 0.01113 \\ 2.839 \\ 0.04175 \\ 0.02987 \\ 0.00814 \\ 0.01561 \\ 0.2006 \\ 0.00723 \\ 0.0436 \end{bmatrix} \qquad \beta_{suspect} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \\ \beta_9 \\ \beta_{10} \end{bmatrix} = \begin{bmatrix} 1.7022 \\ 0.01128 \\ 0.01300 \\ 1.9904 \\ 0.01509 \\ 0.01152 \\ 0.00431 \\ 0.00855 \\ 0.1954 \\ 0.00556 \\ 0.02401 \end{bmatrix}$$

1.2.5 Estimates interpretation

- For one-unit increase in the variable fetal movement is associated with the increase in the log odds of being in suspect fetal rate health vs normal fetal health rate in the amount of 1.9904.
- For one-unit increase in the variable accelerations per second is associated with the decrease in the log odds of being in pathologic fetal rate health vs normal fetal health rate in the amount of 0.01113.
- For one-unit increase in the variable mean value of short term variability is associated with the increase in the log odds of being in suspect fetal rate health vs normal fetal health rate in the amount of 0.1954.
- For one-unit increase in the variable uterine contractions per second is associated with the increment in the log odds of being in pathologic fetal rate health vs normal fetal health rate in the amount of 0.0417.

1.2.6 Performance The model performance has an acceptable precision due to the confusion matrix results and the specificity, sensitivity, and accuracy values for training and testing data. Nevertheless, there is an important miss-classification that might be avoid in each of the classes.

Training			
Acurracy = 87.65%			
	normal	suspect	pathologic
Sensitivity	95.41%	58.0%	62.68%
Specificity	70.54%	94.74%	98.46%

Testing			
Acurracy = 86.85%			
	normal	suspect	pathologic
Sensitivity	95.69%	55.93%	61.90%
Specificity	67.33%	95.09%	98.69%

This above table evidence that each of those training and testing present an important percentage of miss-classification in each of the classes.

1.3 KNN

I decided to start the classification analysis using one of the most used and accurate classification methods: KNN. This supervised machine learning algorithm usually is used not only for classification problems but for regression analysis. Additionally, there is no need to build a model because it uses the input data calculating neighbors and classifying the data.

To perform this model, I used "purrr" and "class" libraries using the created function in Statitical Learning class at UTRGV [1] to obtain the best K for the classification. My decision was based on validating similar models to the one created in class selecting that one as the most accurate [2].

After applying the best K (best possible K for this dataset is 3) function the model was ready to be initialized. This model will use the inputs attributes to classify the training data. As I mention before, this model is simple and accurate when the best K is selected. Then, I applied that into the training and testing data.

1.3.1 Performance The KNN performance is measured not by the estimates interpretation but the confusion matrix and the MSE in both: training and testing data.

[1] For more information please email tamer.oraby@utrgv.edu.
[2] Other methods use R libraries performing training functions.

Training				Testing			
MSE= 0.058				MSE = 0.072			
Acurracy = 94.18%				Acurracy = 92.72%			
	normal	suspect	pathologic		normal	suspect	pathologic
Sensitivity	97.52%	79.24%	87.31%	Sensitivity	96.62%	79.66%	79.66%
Specificity	84.32%	97.95%	99.29%	Specificity	82.18%	97.55%	98.95%

1.3.2 Confusion matrix interpretation To check the KNN performance classifying the data we use the confusion matrix. The confusion matrix shows that based on the training data (testing data) the KNN classify the entire data with an accuracy of 94.35% (92.7%). The model predicted 1298 (314) to be classified as normal FHR out of 1700 training data (426 testing data), with a misclassification of 32 (101). The model predicted 118 (34) to be classified as suspect FHR out of 1700 training data (426 testing data), with a misclassification of 118 (35). Lastly, the model predicted 188 (47) to be classified as pathologic FHR out of 1700 training data (426 testing data), with a misclassification of 16 (5).

Regarding to the specificity and sensitivity we might notice that:

1. Normal FHR.
    - This model has a precision of 84.3% when it makes the classification predictions, and it is correct 84.3% of the time. Additionally, it predict successfully 97.6% of the classifications.
2. Suspect FHR.
    - This model has a precision of 97.9% when it makes the classification predictions, and it is correct 97.9% of the time. Additionally, it predict successfully 79.24% of the classifications.
3. Pathologic FHR.
    - This model has a precision of 99.2% when it makes the classification predictions, and it is correct 99.2% of the time. Additionally, it predict successfully 87.3% of the classifications.

Remark. The caret package using the "confusionMatrix" library has an interesting feature called: Cohen's Kappa Statistic which is commonly used to provide measure of how good two evaluators can classify data. Here the Kappa value is 0.83 which according to the strength table is a good classification.

1.4 Tree Classification

The tree statistical methods used for studying data are set to perform classification and regression. Classification trees perform a structural 1 and 0 (yes and no) decisions that

guide the method to classify the data based on that decisions with two important parts: branches representing attributes in the data, and leaves representing decisions.

To perform this analysis I used "rpart" library (function at the same time) setting the R routine to treat our DATABASE as a categorical classification setting the method by "class".

The summary expose each step showing the number of nodes, complexity parameter, class counts with their classification probabilities, and present the splits counts. For instance, analyzing the second node for 1397 observations the class count was: (class=Normal) 1244, (class=Suspect) 80, (class=Pathological) 73. Each of those with probabilities: 0.89, 0.057, and 0.052 respectively.

1.4.1 Performance This method split the original root node dropping a relative error from 1.0 to 0.33784 where the root node is 0.33784. Next, analyzing the complexity parameter we might find the best complexity parameter value (CP) for this tree analysis. To select the optimal CP, I used "printcp" and "plotcp" function to check and visualize the CP with its relative error per node. Additionally, I check the RSQ loading "rsq.rpart" function. The CP plot evidence that the optimal number of nodes is 8 with an optimal CP equals to 0.01.

Additionally, I present a tree development plot representing all branches and leaves attributes by plotting not only linear classification process but also three dimensional plot (each node step as the third dimension). This type of plot allow the reader to check the convergence in each node, and some important relations about the classified variables. Lastly, the confusion matrix:

Training			
MSE= 0.924			
Acurracy = 94.18%			
	normal	suspect	pathologic
Sensitivity	97.89%	66.52%	83.58%
Specificity	75.68%	98.83%	98.59%

Testing			
MSE = 0.899			
Acurracy = 89.91%			
	normal	suspect	pathologic
Sensitivity	96.62%	62.71%	76.19%
Specificity	70.30%	99.18%	97.39%

1.4.2 Variable Importance and implementation plot Classification trees perform the analysis using the data variables on at a time, decide and split the data. Then, the variable importance refers about the model variable use and accuracy over that variable. Usually, tree classification methods presents gini or information gain plots visualizing all the process locating the most important variable in the top.

The decision plot shows that the variable importance in classifying the data plotting the mean value of short term variability in first place. So, basically the tree classification method used this variable as the principal variable to grouped the data.

The implementation plot using the "plotmo" library is a nice tool that represents the

linear relations in between the involved variables in the classification together with the tree dimensional plot that might be used to analyze the speed classification analysis or the number of nodes in the group selection.

1.4.3 Recommendations After reading the documentation about rpart plots, and related packages I could not find a nice rpart plot when the data have many inputs. So, I could find two solutions: quick solution and developing solution. The quick solution is gathering all the data on R, and using "graphviz" package on python which has more visual dependencies, and features. The developing option is creating better dependencies and libraries in R to visualize tree plots.

1.5 Supported Vector Machine

Here, I tried to use supported vector machine because of one important feature: overlapping data. The SVM method is optimal for generalizing the hyperplanes separation. How do we check for overlapping data? Well, the really descriptive data analysis plot allow us to check this type of overlapping behavior for two variable, and there would be importantly more overlapping characteristics for more than two variables.

To perform SVM analysis, I load the "e1071" R package with the "svm" function training a support vector machine and making the classification analysis. Before starting the process, as the classification variable was not numerical I made the conversion and used the "tune" implementation over a sequence of size 10, and cost exponentially big.

1.5.1 Performance The resulting colorful analysis plot describes the best model using color coding where darker regions imply better performance (accuracy) in the implementation. This is a visual way to explore the tuning implementation for SVM. So, we can observe that the best epsilon runs from 0 to 1, while the optimal cost is in between 110 and 40.

According to the tune method, the best model is:

SVM-Type:	C-classification
SVM-Kernel:	radial
Cost:	64

1.5.2 Confusion matrix interpretation To check the SVM performance classifying the data we use the confusion matrix. The confusion matrix shows that based on the training data (testing data) the SVM classify the entire data with an accuracy of 96.47% (91.08%). The model predicted 1317 (311) to be classified as normal FHR out of 1700 training data (426 testing data), with a misclassification of 13 (14). The model predicted 128 (38) to be classified as suspect FHR out of 1700 training data (426 testing data), with a misclassification of 108 (21). Lastly, the model predicted 195 (39) to be classified as pathologic FHR out of 1700 training data (426 testing data), with a misclassification of 61 (3).

Training			
MSE= 0.035			
Acurracy = 96.47%			
	normal	suspect	pathologic
Sensitivity	99.02%	82.63%	95.52%
Specificity	89.19%	98.7%	99.93%

Testing			
MSE = 0.089			
Acurracy = 91.08%			
	normal	suspect	pathologic
Sensitivity	95.69%	66.10%	90.47%
Specificity	78.22%	96.45%	99.21%

2 Results and conclusions

I decided to stop the analysis, due to the fact the other implementations were not optimal. I tried the neural network, LDA, and k-means method for classification ant the MSE results were significantly high. I stop the classification analysis, and atfer implement the methods we find the following results:

Method	MSE Training	MSE Testing
Multinomial Log-Linear	0.1235	0.1314
KNN	0.058	0.0727
Tree Classification	0.924	0.899
SVM	0.0352	0.0892

Finally, I select two models: multinomial logistic regression and SVM. The SVM method above all the other methods choice was because of the distribution of the data. This type of analysis has a better performance for overlapping datasets.

Additionally, I present this project as part of my code repository at GitHub due to the fact one of the aims this project was set is to be a cover letter to posterior studies.

References

[1] An Introduction to Statistical Learning with Applications in R, *G. James, D. Witten, T. Hastie, R. Tibshirani*, Springer Text in Statistics.

[2] Ayres de Campos, *A Program for Automated Analysis of Cardiotocograms*. SisPorto 2.0, Matern Fetal Med 5:311-318, 2000.

[3] Ggplot images, R repository - ggplot2 package.

[4] Introduction to regression modeling, *B. Abraham, J. Ledolter*, Springer Series in Statistics.

[5] Knn implementation, R repository - Class package.

[6] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, *Modeling wine preferences by data mining from physicochemical properties*. Science direct, Volume 47, Issue 4, November 2009, Pages 547-553.

[7] US National Library of Medicine National Institutes of Health, https://www.ncbi.nlm.nih.gov/pmc/

Appendix

Tables

1. Descriptive table about the variable abbreviations in the modeling.

Abbreviations	Description
LB	FHR baseline (beats per minute)
AC	Number of accelerations per second
FM	Number of fetal movements per second
UC	Number of uterine contractions per second
DL	Number of light decelerations per second
DS	Number of severe decelerations per second
DP	Number of prolonged decelerations per second
ASTV	percentage of time with abnormal short term variability
MSTV	mean value of short term variability
ALTV	percentage of time with abnormal long term variability
MLTV	mean value of long term variability
Width	width of FHR histogram
Min	minimum of FHR histogram
Max	Maximum of FHR histogram
Nmax	Number of histogram peaks
Nzeros	Number of histogram zeros
Mode	histogram mode
Mean	histogram mean
Median	histogram median
Variance	histogram variance
Tendency	histogram tendency
CLASS	FHR pattern class code (1 to 10)

2. P-values using the multinomial log-linear model.

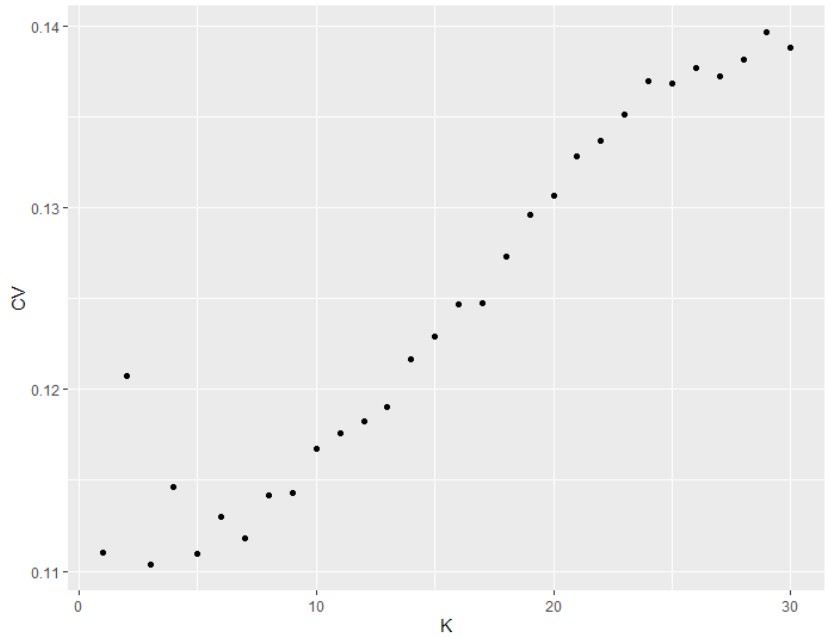
	pathologic	suspect
(Intercept)	2.625e-06	2.220e-16
LB	7.922e-01	4.432e-13
AC	0.000e+00	0.000e+00
FM	7.442e-04	1.008e-05
UC	0.000e+00	0.000e+00
DL	0.000e+00	0.000e+00

DS	0.000e+00	0.000e+00
DP	0.000e+00	0.000e+00
ASTV	0.000e+00	2.581e-07
MSTV	2.612e-07	6.954e-01
ALTV	5.018e-14	1.523e-03
MLTV	9.395e-01	8.459e-01

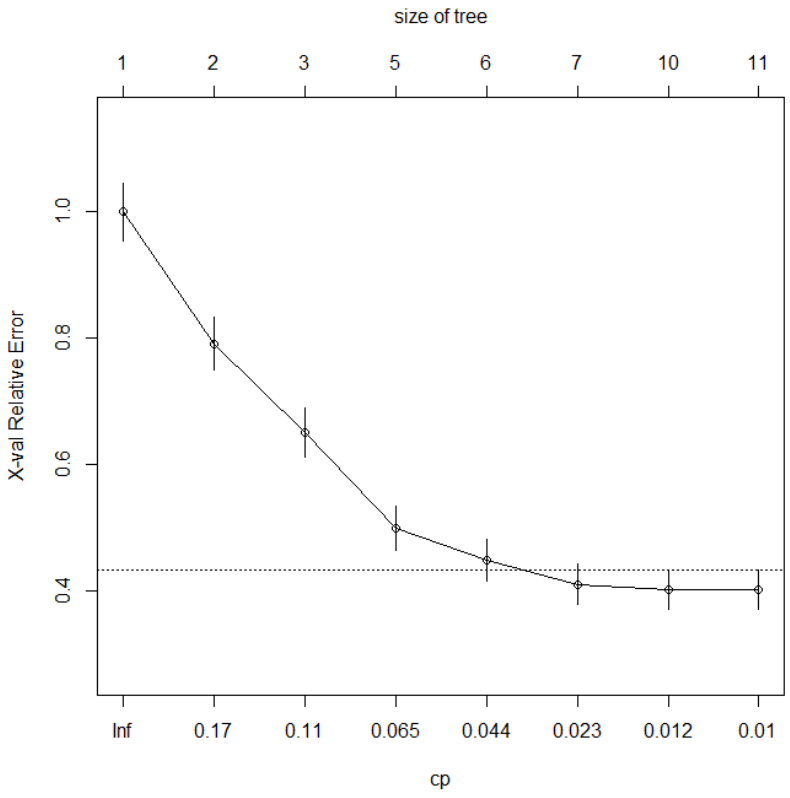
Plots

1. Data Graphical Analysis (Categorical and Continuous Data). The density axis represent the percentage of the total amount of that specific quantity.

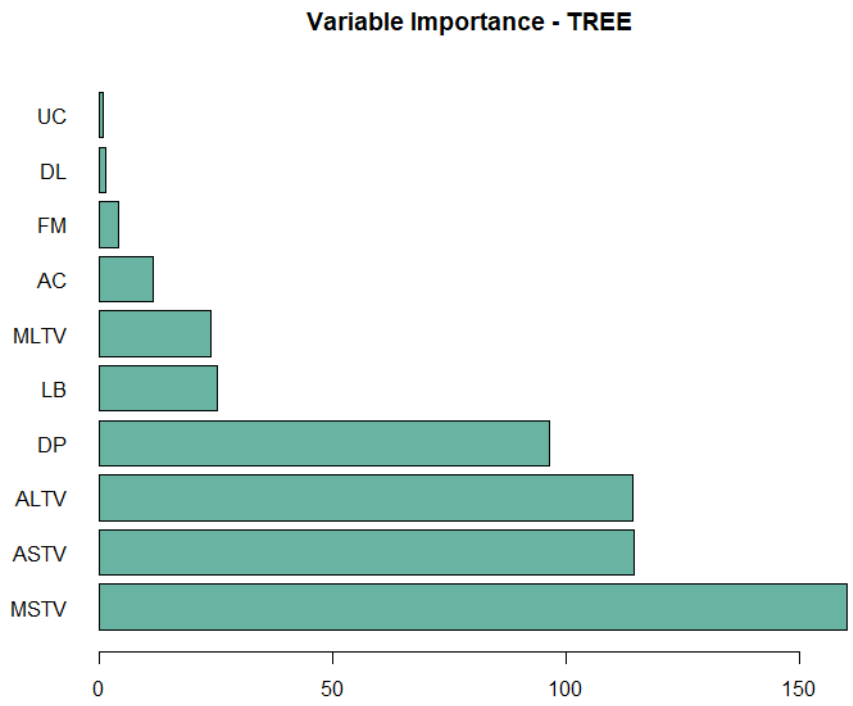
2. Best K for KNN method.



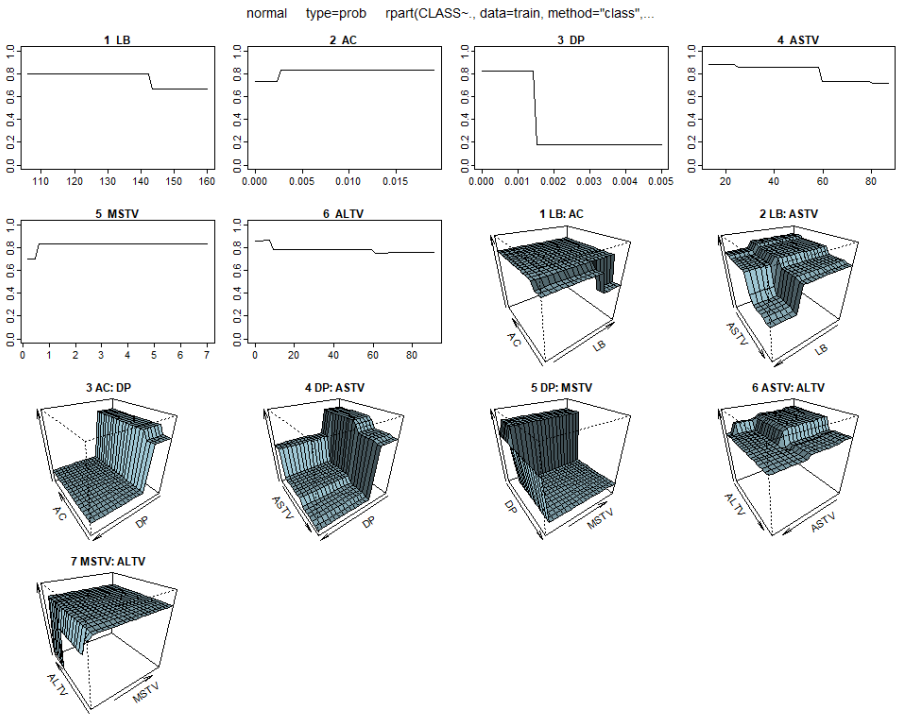
3. CP performance during the tree classification.



4. Variable tree importance during the method.



5. Variable tree variable performance in the tree implementation method.



6. Colorful performance SVM method.